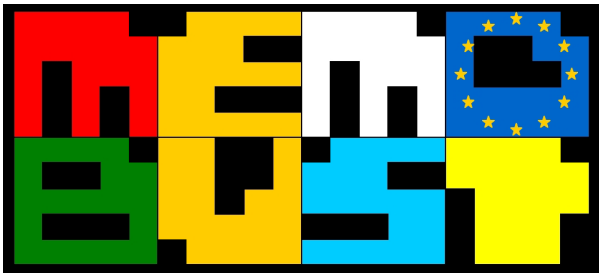

This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Manual Integration

Contents

General section

1. Summary

Macro-integration is the process to integrate data from different sources on an aggregate level, to enable a coherent analysis of the data. Although this definition looks very clear it raises many questions. In this module on macro-integration or balancing as it is often called, macro-integration is illustrated using the supply use system of the national accounts as an example. In this case macro-integration means resolving inconsistencies between independent source data which are the base for estimates of the individual cells of the supply use tables (SUT). Inconsistencies become apparent by violating the identities of the system.

Macro-integration is detection of, search for causes and definition of solutions for the reconciliation of inconsistencies within the limits of identities which have to be fulfilled and a plausible outcome.

At least part of macro-integration needs to be done manually because the causes of inconsistencies have often a non-statistical character and cannot be caught in robust rules.

2. General description

Macro-integration is the process to integrate data from different sources on an aggregate level, to enable a coherent analysis of the data. Although this definition looks very clear it raises many questions. Characterising macro-integration leads to something like *the reconciliation of inconsistent statistical data on a high level of aggregation*. Then immediately questions arise like 'What is meant by inconsistency?' and 'How are those inconsistencies revealed?'

In order to detect inconsistencies in statistical data, one needs a framework consisting of definitions of variables and relations between variables (a.o. identities). Having such a framework, a more accurate 'definition' could read: *Balancing is an activity required when inconsistent statistical information from independent sources is brought together in an 'accounting' framework consisting of well-defined variables, accounting identities on combinations of variables and less strict relations between the sets of variables.*

Macro-integration is an activity that is well known for its application in the compilation of national accounts, where data from many independent statistical sources are brought together in an accounting framework as described in the System of National Accounts (SNA) and European System of Accounts (ESA). The accounting framework reveals inconsistencies in the source data. In national accounts often the term balancing is used instead of macro-integration. In this module, the terms will be used interchangeably. In this module, macro-integration will be illustrated using the supply and use system of the national accounts as an example.

2.1 Supply-use tables

Three basic identities are the basis for the supply and use system. In this section they are presented in a simplified form.

(1) $P + M = IC + C + I + E$

where

P = production of goods and services

M = imports of goods and services
IC = intermediate consumption
C = consumption of households and government
I = investment (including changes in inventories)
E = exports of goods and services

All variables are well-defined in such a way that equation (1) is per definition true and thus an identity.

Identity (1) claims that all goods and services sold (left hand side of (1)) are also bought (right-hand-side of (1)), which is per definition true. The left hand side of this identity says that all goods and services sold are either domestically produced (P) or imported (M). The right hand side of the identity says that all goods and services bought are used for intermediate consumption of industries, consumption of households and government, fixed capital formation and changes in inventories and exports.

Within an accounting framework the definition of the variables must be very precise. If an activity is defined as production, there must also be a use, else there cannot exist equality between sales and purchases. So exhaustiveness of definitions as to what should be included and what not and exhaustiveness of measurement of variables are important conditions for this identity.

The second identity in the SUT is the definition of 'value added':

(2) $Y = P - IC$

where

Y = value added

Total value added for a country is gross domestic product (GDP), the growth of which is the most important and widely used indicator for judging the performance of an economy.

Combining equations (1) and (2) gives another way of deriving GDP:

(3) $Y = C + I + E - M$

A third way of estimating GDP is based on incomes.

(4) $Y = W + OS$

where

W = wages and salaries (including social premiums)
OS = operating surplus / mixed income

Operating surplus is a residual item, i.e., what remains of the revenues of production after deduction of the costs of intermediate consumption and wages. The term mixed income is added for the income of self- employed persons. Their income has the character of both wages and operating surplus.

In national accounts terminology the three ways of estimating GDP are denoted by:

(A) The production method ($Y = P - IC$)

(B) Expenditure method ($Y = C + I + E - M$)

(C) Income method ($Y = W + OS$)

Estimating GDP using these three methods, not only data on GDP become available, but also data on its components like exports, household consumption, value added (per industry). The three ways of estimating GDP are combined in the SUT and it are these three ways that make balancing necessary. Populating the equations with data from various, independent, dedicated sources, leads in general to three different estimates of GDP. Macro-integration or balancing leads to only one estimate of GPD by reconciliation of inconsistencies between the source data.

Balancing can be done on the level of the equations (A) – (C) above. This, however, leads to non-optimal results. Causes of inconsistencies between source data are usually not clear, which may lead to wrong adjustments. More detailed information will lead to better estimates of GDP and its components.

A first extension that the leads to the supply use framework is the breakdown of identity (1) to n types of goods and services (commodities).

(1.1) $P_1 + M_1 = IC_1 + C_1 + I_1 + E_1$

(1.2) $P_2 + M_2 = IC_2 + C_2 + I_2 + E_2$

.

.

.

(1.n) $P_n + M_n = IC_n + C_n + I_n + E_n$

An appropriate choice of the commodity classification of the supply use system facilitates balancing. A useful criterion is to limit the number of possible producers and users of a commodity. The analysis for reconciliation is then limited to the data of those producers and users instead of 'the whole economy'.

A second breakdown concerns equation 2 which is split into m industries (NACE-classes).

(2.1) $Y_1 = P_1 - IC_1$

(2.2) $Y_2 = P_2 - IC_2$

.

.

.

(2.m) $Y_m = P_m - IC_m$

Also in the case of industries an appropriate choice of the industry classification facilitates balancing.

With these extensions a (still simplified) system of supply and use tables is constructed. Figure 1 gives a schematic reflection of a supply-use framework.

A system as presented in Figure 1 can be used for balancing the three methods for estimating GDP. A balanced system of supply and use tables, meaning that the identities are fulfilled on the commodity and industry level, leads to only one estimate of GDP.

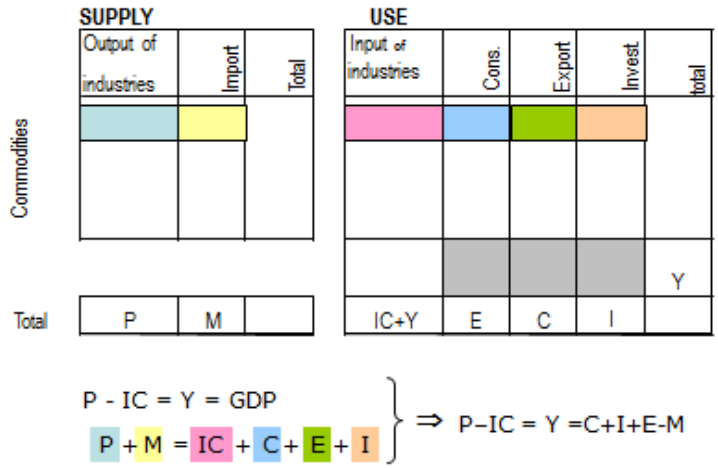

Figure 1. Supply use framework

Next to the identities of the SUT, less strict relations between the variables of the system exist which can improve the quality of the GDP-estimates. The most important ones concern volume and price changes. The concept of constant prices adds a lot of additional information in the SUT which is extremely useful in the detection and reconciliation of inconsistencies between source data.

In constant price estimation the year-to-year value changes are decomposed in a volume and a price change

$$Value\_change = \frac{\sum P_t * Q_t}{\sum P_{t-1} * Q_{t-1}} = \frac{\sum P_t * Q_t}{\sum P_{t-1} * Q_t} * \frac{\sum P_{t-1} * Q_t}{\sum P_{t-1} * Q_{t-1}}$$

The indicators for the volume and price change happen to be a Laspeyres volume index and a Paasche price index. As advised in SNA and ESA, in the national accounts the previous year is mostly used as the base year for the price and volume indices.

For every entry in the SUT a so-called six pack of data is then available consisting of level estimate for the year T in prices of T (Current prices, CUP) , year T in prices of T-1 (Constant prices, COP) and T-1 in prices of T-1 (Current prices of T-1, T-1) The ratio of the first two giving the price index and the ratio of the latter two give the volume index, which together with the value index complete the six pack illustrated in Figure 2.

The choice of index formulae assures that in terms of levels the identities of the SUT hold also in previous years' prices. The price and volume changes are mainly used for plausibility checks.

An example: in a competitive economy price changes for all producers and users are expected to be more or less the same. If the price change for a certain entry in the row of the SUT differs from all the others, this in a signal that there might be a mistake.

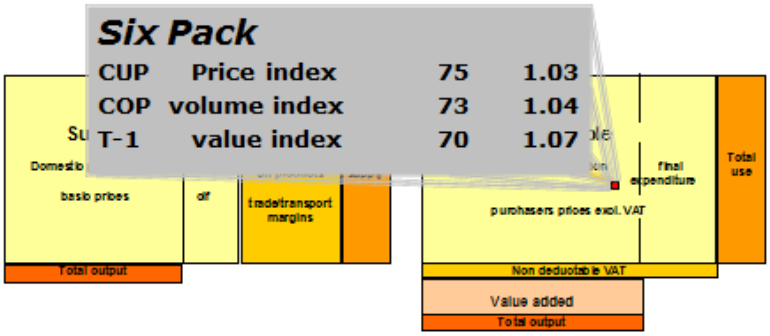

Figure 2. Constant price estimation and the six pack

Another example: if one produces 1 kilo of cheese, one needs a certain amount of milk; if one produces two kilos of cheese, one needs twice the amount of milk. The volume change of the production of cheese is closely linked to the volume change of the (intermediate) input of milk. If the volume changes do not match, this is a signal that there might be a mistake.

When adding labour data to the SUT in terms of full time equivalents or hours worked, changes in labour productivity can be calculated, which can be used as an indicator for the plausibility of the results

The supply use tables are a balancing framework which helps to detect inconsistencies and implausibilities in source data. The next step is to balance the framework in order to get unique and plausible estimates for GDP and its components.

2.2 Causes of inconsistencies

Working with statistical data based on samples and questionnaires and influenced by non-response etc. means working with margins of error. Even when samples are perfect and response is 100% there will be inconsistencies. The cause is then a statistical one. In such a case balancing could be done automatically using the inverse of the margins of error of the statistics concerned as weights. Methods for automated balancing described in other modules of the "Macro-Integration" topic are based on this principle.

However, statistics are never ideal and inconsistencies are not only caused by sampling etc. but have causes of a more or less non-statistical nature. It is these causes of inconsistencies that make manual integration or manual balancing necessary as a preliminary step prior to automated balancing.

Causes of inconsistencies are manifold and inconsistencies arise at various stages of collecting and processing of data. Some examples:

i. Causes of inconsistencies in data at the unit level

For the collection of data on sales and purchases mostly statistical units like enterprises, establishments or kind of activity units are defined (see the topic "Statistical Registers and Frames" for more details on units). These statistical units consist of sets of legal units. In the simplest case the statistical unit is the same as the legal unit, but often the statistical unit consists of more than one legal unit. Having a well-defined statistical unit does not necessarily mean that it corresponds to, for example, units used by the company concerned for their tax declaration. In case the respondent follows his bookkeeping or tax records, the reporting unit is not the same as the statistical unit. This can lead to missing data of certain legal units or double counting. This risk increases when data are collected by different agencies for example the Statistical Office, the Central Bank or tax authorities.

A second and widespread cause of inconsistencies is globalisation. When a unit in a country is the (economic) owner of all goods (and services) purchased and sold, it will report its worldwide figures in business statistics, even when the goods concerned never enter the country of residence of the company. On the other hand foreign trade statistics on goods are based on goods crossing borders, so goods that never enter the country of residence are missing. In this case there is an inconsistency between business statistics and foreign trade statistics which both serve as a source for the supply use system.

Other causes of inconsistencies at the unit level are 'mistakes'. For a significant number of companies the bookkeeping year differs from the calendar year used in the national accounts (and other annual statistics). Entering the bookkeeping data in the questionnaire causes inconsistencies in the SUT when these data are confronted with other statistics.

The questionnaires for business statistics are designed in such a way that data over the various industries can be compared and added. The needs of users like national accounts require specific definitions of variables in the questionnaires, which cannot always be derived directly from bookkeeping records. When a respondent uses his own definitions of variables, this may cause inconsistencies in the SUT as well.

Last but not least, a company can provide incomplete data. If, for example, data on changes in inventories are lacking, the transformation from purchases and sales to production, intermediate consumption and value added (a key variable in NA) cannot be made.

ii. Causes of inconsistencies in data at the statistical office

The processing of collected microdata by subject matter statistics can cause inconsistencies. Although procedures for grossing up are routine in a statistical office, the target population is less straightforward. An important issue in this case is the existence of units, by which is meant whether or not units were active during the whole reporting period. A second issue in this matter is outlier detection and treatment.

Small enterprises often get less detailed questionnaires, implying the necessity to break down the aggregated variables to the level of detail of the large enterprises. The assumptions made for this calculation may be incorrect. The same holds for the (further) breakdown of variables from business statistics to the commodity classification of the SUT.

The last cause of inconsistencies to be mentioned is the hidden economy. When no or insufficient estimates for hidden economy are included in the SUT, inconsistencies will arise. When, for example, consumers buy beer at the pub they usually do not know whether it is, economically speaking, a 'black' or a 'white' beer, implying that in household consumption beer is reported, while in business statistics the 'black' beer will be missing.

2.3 Detection and balancing of inconsistencies in the supply-use tables

Balancing should result in a consistent and plausible set of supply and use tables. This statement means that inconsistencies are not limited to the violation of the identities of the framework, but also concern the less strict plausibility relations. In balancing the detection of inconsistencies and implausibilities on the one hand and finding the causes on the other is most important part of the job. Having this knowledge, finding a way to resolve the inconsistency is then mostly straightforward.

Most easy and straightforward in the detection of inconsistencies is the violation of the identities of the SUT both in current and previous year's prices. The most important ones are given in an aggregated form in equations (1) and (2) of Section 2.1 and say that per commodity supply and use must be equal and per industry total output must be equal to total input (including value added).

Inequality between supply and use asks for an investigation to the exhaustiveness of the estimation on both sides. Questions like 'Are black and illegal activities included in the production for all relevant goods and services?' and 'Are estimates for household consumption of tobacco and liquor exhaustive?' are bound/likely to arise.

Part of the inconsistencies are caused by the imperfect measurement with globalised companies. In order to resolve the inconsistencies, data on the unit level must be investigated and reconciled. The way in which the company organises its production processes influences the way of recording in the SUT and directs which data should be adjusted. This part of the balancing must be done manually, because the judgement of which way to go in the SUT is mainly based on qualitative information (how the company is organised, which unit is economic owner, etc.).

A third cause of inconsistencies is an incorrect breakdown of variables from business statistics to the commodity level of the SUT. For example: the variable 'Office needs' from business statistics is broken down in 'paper', 'printer cartridges', 'pens' and 'note blocks', and in the confrontation between demand and supply it turns out that there is a shortage of paper and a surplus of cartridges. Because no (real) information is available to make the breakdown, the most obvious solution is adjusting the (assumed) distribution key of office needs.

Causes of violation of the identities are not always easy to detect and additional information can be very helpful. Adding constant (previous year's) price estimates adds a lot of information to the system of supply and use tables. Not only the identities of the SUT for the constant price estimates are at stake, also the less strict plausibility relations between variables based on price and volume changes are in the picture.

Looking at industries (columns of the SUT) one expects that the volume change of production is more or less of the same size as the volume of intermediate consumption. This relation is stronger for the output goods and the input of raw materials than for input of services. However when there is a big difference between the two volume changes this is an indication that there might be something wrong in the data and further investigation is advisable.

When combined with labour data the volume changes of value added can be used to calculate changes in labour productivity. Generally one expects that labour productivity is rising gradually every year (except perhaps with the start of a recession). A decrease or a high growth of productivity can point to a mistake in the data.

Looking at commodities (rows of the SUT) one expects that in a competitive economy the price changes are more or less the same for all economic agents. When on the commodity level a price change deviates seriously from the average, it is an indication that there might be something wrong in the data and further investigation is advisable.

Macro-integration implies the adjustment of statistical source data. The use of independent secondary information gives a more solid base to the adjustments. Two examples:

- The sales of motorcars can be confronted with the number of newly registered number plates.
- The consumption of liquor can be compared with tax revenues on liquor from the government administration.

The causes of inconsistencies mentioned above mostly cannot be reconciled using automatic techniques. Manual integration in which the non-statistical causes of inconsistencies are investigated is therefore a necessary step in the balancing process prior to automated balancing.

3. Design issues

As described above, inconsistencies in the supply use tables may have many causes. Finding such causes can be very labour-intensive. The coordination and design of source statistics, classifications of the SUT help to limit the scope of investigations.

i. Coordination

The general business register is an important tool for co-ordinating business statistics. The (unique) definition of units of observation helps to avoid double counting and gives a view on those parts of the economy not covered by statistics (white spots) and for which additional estimates have to be made.

Harmonisation of the industry classification used in the GBR, business statistics and the SUT facilitates the search for causes on inconsistencies. The statistical process from unit data to a balanced SUT can be analysed in the assurance that at all stages the same population (part of the economy) is investigated.

Additivity of business statistics is very helpful for the compilation of supply and use tables. The first requirement for additivity is that double counting must be prevented. Second is that definitions of common variables must be the same (or convertible into the common definition) in all questionnaires and, as far as possible, use the same classification (and coding) for the (breakdown of) variables.

ii. Classifications in the supply and use tables

An appropriate choice of the industry and commodity classification in the supply-use system facilitates the search for causes of inconsistencies.

For the industry classification the following aspects should be considered:

- Link to international classifications (NACE)
- Homogeneous output and input structure
- Data availability
- Market versus non-market producers
- Exempted from VAT
- Size (avoid relative small amounts)

The more homogeneous an industry is concerning input and output structure, the stronger the link will be between the volume changes of production and intermediate consumption. Deviations point directly to possible mistakes. Changes in the output structure can be the cause of the deviations and in that case no adjustments are required. When there are no changes, analysis of microdata can lead to the cause and reconciliation of the inconsistencies.

Also an appropriate choice of the commodity classification in the SUT facilitates the search for causes of inconsistencies.

For the commodity classification the following aspects should be considered:

- link to international classifications
- homogeneous concerning taxes on products
- homogeneous concerning subsidies on products
- homogeneous concerning VAT tariffs
- homogeneous concerning trade and transport margins
- homogeneous concerning destination
- homogeneous concerning price changes
- availability of data
- size (avoid relative small amounts)

Homogeneity of commodity to elements concerning valuation (taxes, subsidies and trade margins) makes the compositions of transactions transparent and clear cut and makes analysing much easier.

Homogeneity concerning destination means that the number of users of a commodity is limited. In case there is only one producer and one user of a commodity the search for the cause of inconsistencies concerns only two source statistics. When a commodity has 20 users, the search becomes more complicated.

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate "Glossary" provided as part of the handbook.

7. References

United Nations (2008), *System of national accounts*.

Eurostat (2010), *European system of accounts* (forthcoming).

Interconnections with other modules

8. Related themes described in other modules

1. Statistical Registers and Frames – Main Module
2. Macro-Integration – Main Module

9. Methods explicitly referred to in this module

1.

10. Mathematical techniques explicitly referred to in this module

1.

11. GSBPM phases explicitly referred to in this module

1.

12. Tools explicitly referred to in this module

1.

13. Process steps explicitly referred to in this module

1. Manual reconciliation of macrodata

Administrative section

14. Module code

Macro-Integration-T-Manual Integration

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	17-04-2013	first version	Piet Verbiest	Statistics Netherlands
0.2	30-10-2013	comments by referee	Piet Verbiest	Statistics Netherlands
0.3	10-03-2014	comments by ed. board	Piet Verbiest	Statistics Netherlands
0.3.1	11-03-2014	preliminary release		
0.3.2	12-03-2014	minor change in glossary		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	26-3-2014 13:24