Documenting the statistical editing process, a case study of the Luxembourg Financial Accounts

Roymans Ingber*
Banque Centrale du Luxembourg, Luxembourg, Luxembourg – ingber.roymans@bcl.lu

Abstract

We discuss a way to document the statistical editing process in the form of process tables. Work on process tables for official economic statistics has so far focussed on hi of the Finregates like Gros s National Income (GNI) but here we present them for every item of the Financial Accounts statistic, i.e. the financial part of the National Accounts. The generation of process tables forms only a part of our system for the automated generation of metadata, whose basic components we will briefly describe here. We argue that not only provide process tables the users with an important quality indicator, they also provide the compilers with an important management tool as they give insight in which part of the editing process is responsible for which part of the result. We conclude with some observations: Naturally, the design of the statistical production process will determine the ease with which useful metadata can be extracted. The use of a modular software architecture and standardized data formats greatly facilitates the aggregation and translation of both normal data and corrections data and thus the compilation of process tables. The starting point of the process tables can be chosen conveniently, but extensions beyond that point may demand the exchange of metadata on editing corrections with external data providers. Finally, we think that international metadata exchange standards, like Statistical Data and Metadata Exchange (SDMX), should facilitate the exchange of process tables.

Keywords: statistical metadata; process tables; national accounts

1. Introduction

During the statistical compilation process often corrections are made to the original data, either manual or automated, with the aim to detect and adjust errors in the collected micro data. To document the impact of this editing process on the published results can provide important information for the users (Nordbotten 2000). It should warn users of certain interpretations and false conclusions even if the published estimates are as precise as possible. It can also make a contribution to transparency, which is essential to gain public trust in official statistics (UNECE 1992).

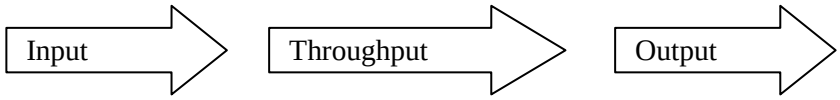
Process tables aim to give a detailed overview of the impact of the editing process (Eurostat 2005, GNI Committee 2006). They do not provide a quantitative measure of the accuracy or reliability of the estimates but instead show the extent to which the results are based upon real data and the extent to which these data are modified and adjusted. Together with other metadata on methods and sources they provide useful information for the end-users of the statistic.

Producers of official statistics therefore have to implement next to their production process dedicated systems that allow the extraction of useful metadata about the production process. In this paper we examine the case of the Financial Accounts (FA) statistics in Luxembourg. Process tables form only a part of the output from our system for the automated generation of metadata.

* The views expressed are the author's and not those of the Banque Centrale du Luxembourg.

2. The statistical compilation process

In general a statistical compilation process consists of an input, throughput and an output component:



What flows through the process is the statistical data, with source data as input and publications as final output. Source data consists ideally of observations, i.e. empirical data, for example the results of a survey. The throughput phase can consist of every thinkable data manipulation but typically involves several data editing steps. Editing consist of making corrections to the data, i.e. adding or subtracting a certain amount, adding and removing observations. Essential is that each editing step is optional in the sense that a final result could also be obtained without editing.

Process tables

Process tables aim to give an overview of the impact of the editing process on the data and on the final result. One can easily produce a process table as follows: Let there be n different editing steps. Order all the editing steps in a chronological or any other convenient order. We first define S_i (i = 1, 2, … n) as the outcome of the statistical compilation process with only those editing steps performed of rank i and lower. Also we define S_0 as the outcome in the case of completely unedited data. Then we can define the effect E_i of editing step i on the final outcome as the difference:

(1) $E_i = S_i - S_{i-1}$.

This is not the only possible way to define the effect of an editing step but it has the convenient properties that the final result can now be written as the sum of the unedited result plus the accumulatedE

$$_n = S_0 + E_1 + E_2 + \dots E_n \quad (2)$$

These formulas define our process tables. They give a measure of the numerical impact of the different data editing steps and provide a quantitative way of documenting the editing process.

Let $S(\mathbf{x})$ be a summary statistic on an input data set represented by a vector $\mathbf{x}$. In the case of the National Accounts (NA) statistics, all of our editing corrections C_i on $\mathbf{x}$ can be represented by a simple addition of data points: $C_i(\mathbf{x}) = \mathbf{x} + \mathbf{c}_i$. Further, our summary statistics are simple population totals. In this special case it holds that $S(\mathbf{x} + \mathbf{c}_i) = S(\mathbf{x}) + S(\mathbf{c}_i)$ and thus the effect of each editing step i on the outcome as given by formula 1 becomes simply $E_i = S(\mathbf{c}_i)$. Formula 2 then leads to the following process table definition:

(3) $S_n(\mathbf{x}) = S(\mathbf{x}) + S(\mathbf{c}_1) + S(\mathbf{c}_2) + \dots S(\mathbf{c}_n)$

2. Case study: The Luxembourg Financial Accounts

The FA are the financial part of the NA. They show for each sector of the economy the stocks and flows in financial assets and liabilities. The Luxembourg NA are compiled according to the guidelines of the European System of Accounts (ESA) 2010 (Eurostat 2010). By construction, the NA are subject to several internal consistencies, of which the most important are the horizontal, the vertical and the

stock-flow consistency. The horizontal consistency states that each assets of a sector has to be at the same time the liability of a counterpart sector. The vertical consistency states that every non-financial transaction has an equivalent financial transaction, while the stock flow consistency assures that the change in stocks during the accounting period equals the sum of the flows.

The compilation process of the Financial Accounts for Luxembourg is to a high extend automated, including the generation of metadata. At the hand of figure 1 we will now briefly discuss some of the basic concepts behind the system and explain how it allows for the automated generation of process tables and other metadata.

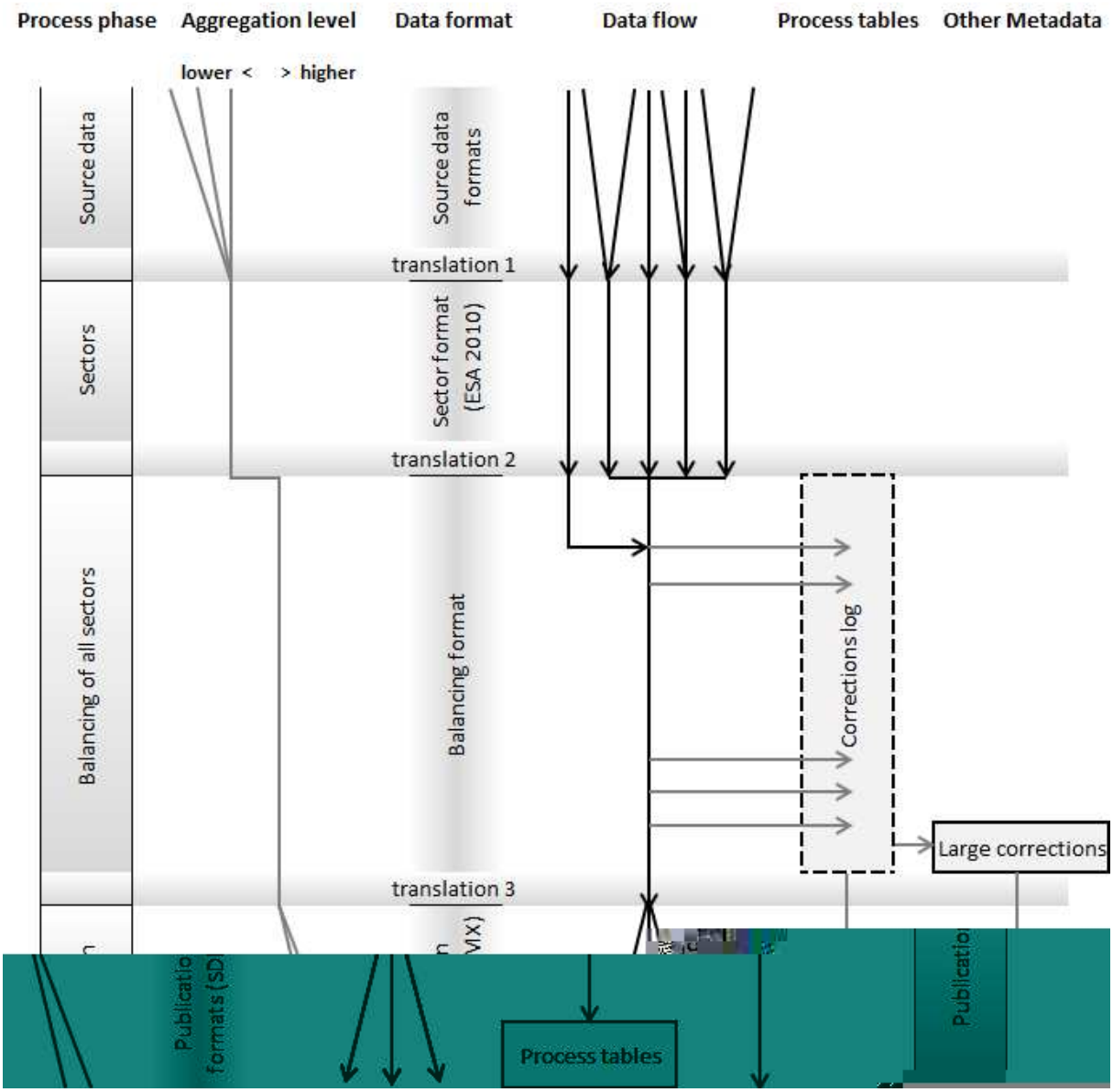


Figure 1. A schematic overview of the Financial Accounts production process.

Process phase
The “source data” and “publication” phase in the schema correspond with the input and output phase respectively. Our throughput phase is organized in two parts. In the first part the source data is transformed into complete financial accounts for each sector of the economy. In the second part the data for the different sectors is combined into financial accounts for the entire economy. This second phase we call the “balancing”.

Aggregation level
Statistical data initially consists of observations. In economic statistics, each observation will in general contain one numerical variable, often an amount in the local currency, together with a certain number of classification variables. The latter can contain either numerical or categorical values, often elements of a hierarchical classification. When data are aggregated, groups of observations are replaced with summary statistics. By aggregating, the same data is represented in a different, more compacted form. In general, the level of aggregation will increase (and the level of detail decrease) during the process, with the source data being the least, and the published data the most aggregated.

Data format
During the entire process data will have to be stored, either permanently or temporarily, in databases, or in files. Data will in general be stored in tables in a certain fixed format. It is the combination of all the variables together with the possible values that each variable can take that we will here call the “format”. It is this format that determines the aggregation process and therefore the level of aggregation or alternatively the level of detail of the data.

The source data can exist in a large number of non-standardized formats. Early in the process this data is translated to a single standardized working format. This format contains already the ESA classification codes used for our publications, while maintaining as much detail from the source data as possible. This we call the “sector” format. After the compilation of the sectors, the data is aggregated to another format used for the balancing: the “balancing” format. Finally, the publication and dissemination the use of ccur in many different formats, typically one for each publication. This leads to the use of four subsequent types of data formats and also implies the translation from one format to another at three points during the process. Translation modules are used to translate one format to another.

The use of standardized formats greatly simplifies the automation as it helps to organize the production process in a modular way. Thus, software modules that work with the same format can easily be interchanged or replaced, allowing for more efficient maintenance and reuse of existing software code.

Data flow
During the entire process we perform three basic data manipulations:

- classification: classification of the collected source data into hierarchical categories,
- making editing corrections:
    - removal of outliers,
    - benchmarking and enhancing the original source data with data from other sources
    - imputation of missing data,
    - removal of internal inconsistencies,
- compilation of summary statistics: aggregation to population totals by classification category.

Editing can occur at all levels of aggregation: at the micro level, for example to correct outliers, but also at higher level, i.e. macro-editing. All editing corrections are performed automatically, based on fixed rules laid down in the software. During the balancing phase every editing correction made is

stored in a database, together with some additional information such as the reason why it was made or the time it was made.

Process tables and other metadata

From this corrections log process tables can be extracted. For this, the corrections that are stored in the format that they have originally been made in will have to be translated and aggregated to the publication formats. This asks for standard translation and aggregation modules that work just as well with normal data as with corrections data. It is this technical challenge that is the reason that we have so far implemented only the logging of corrections during the balancing phase. As a result, the starting point or our process tables, the point that defines our "source data", coincides with the start of the balancing phase. Further, when combining data from different sources it is ambiguous what to call the "source" and what the "correction"; for this in practice a choice has to be made. Also other metadata can be generated; for example, we can publish a list of all corrections that exceed a certain threshold.

Typically, one process table is produced for each publication and for each published period. An example of a small part of our process table is shown in Fig. 2.

Process table

for the third quarter of 2014 (in million euro)

Identifier	Source data	Security by Security	Horizontal balancing	Vertical balancing	Continuity	Manual corrections	Discrepancy	Final result
Q:N:LU:W0:S11:S1:N:N:B102:F:_Z	0	1	-22	0	-1084	0	0	-1105
Q:N:LU:W0:S11:S1:N:N:B9F:F:_Z	-900	1827	344	0	0	0	0	1271
Q:N:LU:W0:S11:S1:N:N:BF90:F:_Z	-109969	4204	11603	0	0	0	0	-94163
Q:N:LU:W0:S124:S1:N:N:B102:F:_Z	0	0	0	0	0	0	0	0
Q:N:LU:W0:S124:S1:N:N:B9F:F:_Z	3026	0	0	0	0	0	0	3026
Q:N:LU:W0:S124:S1:N:N:BF90:F:_Z	-3532	0	0	0	0	0	0	-3532
Q:N:LU:W0:S128:S1:N:N:B102:F:_Z	0	0	0	0	0	0	0	0
Q:N:LU:W0:S128:S1:N:N:B9F:F:_Z	-5034	0	-14	0	0	0	0	-5048
Q:N:LU:W0:S128:S1:N:N:BF90:F:_Z	1650	0	-18	0	0	0	0	1632
Q:N:LU:W0:S129:S1:N:N:B102:F:_Z	0	0	0	0	0	0	0	0
Q:N:LU:W0:S129:S1:N:N:B9F:F:_Z	0	0	0	0	0	0	0	0
Q:N:LU:W0:S129:S1:N:N:BF90:F:_Z	0	0	-4	0	0	0	0	-4
Q:N:LU:W0:S12K:S1:N:N:B102:F:_Z	-8	1110	0	0	0	0	0	1102

Figure 2. Some records from a process table for the Financial Accounts.

The identifier column shows the unique (SDMX) code identifying each published figure, while the other columns show the subsequent editing steps in the process:

- Source data: the sector data used as input for the balancing process
- Security by Security: corrections to align the data with data from our Security-by-Security database on the holdings and issuances of securities
- Horizontal balancing: corrections made to remove any horizontal inconsistencies
- Vertical balancing: corrections made to remove any vertical inconsistencies
- Continuity: corrections made to ensure the stock-flow consistency
- Manual corrections: ad hoc manual corrections made
- Discrepancy: changes to the source data not accounted for in the above; should be zero
- Final result: the final published figures.

5. Conclusions

Although the concept of process tables as described above is rather straightforward, in practice some things have to be taken in consideration:

It goes without saying that the statistical compilation process should be designed from the start with the need to generate metadata in mind. The way the compilation process is designed decides how easy it is to extract meaningful metadata from it. Essential hereby is that each editing correction is logged. The use of a limited number of standardized formats facilitates the translation and aggregation of the corrections data using general translation modules that work as well with normal data as with cor48(e)-1.91882( )-282.926(p)-0.300048(r)-4.15818(o)-0.300048(c)8.95629(e)-1.91977(s)-2.53846(s)-2.53846( )-272

The choice of what you consider "source data", i.e. the starting point of the process table, is ambiguous and has to be clearly explained. This ambiguity might also impair the cross-country comparison of process tables. It will often be practical considerations that determine the choice of starting point. In our case, the extension of the starting point beyond the beginning of the "balancing" phase would require extra information on corrections made during the "sector" phase. This information has to come partly from external data providers, and therefore would ask for the exchange of metadata on corrections with those external providers, something that has not yet been arranged.

Process tables can provide the users of a statistic with useful metadata about the editing process. However, they are not perfect as they do not show all important metadata, like the number of corrections made or the reason why a correction was made. They provide a useful management tool for compilers; they show for example where in the process something goes wrong, while the comparison of two process tables can quickly reveal what particular step in the process was responsible for a change in the result. Therefore, we think that future implementations and developments of international metadata exchange standards like SDMX should aim to facilitate the exchange of process tables.

References

[1] Eurostat (2005). Process Tables Compilation Guide. GNI Committee 054. (Handout distributed during the 5th Meeting of the GNI Committee held in Luxembourg on 5-6 July 2005.)

[2] Eurostat (2010). European System of Accounts ESA 2010. Luxembourg

[3] GNI Committee (2006). Report from the Commission to the European Parliament and the Council on the application of Council Regulation 1287/2003 on the harmonisation of gross national income at market prices (GNI Regulation).

[4] UNECE **(**1992). United Nations Economic Commission for Europe: Fundamental Principles of Official Statistics in the UNECE region. Geneva

[5] Nordbotten S. (2000). Meta-data about editing and accuracy for end users. Proceedings from UNECE Work Session on Statistical Editing. Cardiff