UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS Work Session on Statistical Data Editing

(den Haag, Netherlands, 24-26 April 2017)

Topic (ii): Implementation of standards – Generic Statistical Data Editing Models, Validation and Transformation Language (VTL), etc.

A new GSDEM of multisource data for multiple statistics

Prepared by L.-C. Zhang (lcz@ssb.no) and S. Jentoft (susie.jentoft@ssb.no), Statistics Norway

Abstract

The editing process of the A-ordningen (of contractual payments, social and health benefits, etc.) in Norway is different compared to the earlier ways in which statistical data are processed from multiple administrative register data. It is characterised by implementing as many common statistical editing function as possible, before the process and data sprout into different subject domains. One may refer to this as coordinated editing, as a new GSDEM of multisource data for multiple statistics. The aim is to clarify and secure the coherence between different statistics, which in turn can lead to new statistical outputs that cannot be achieved using the traditional approach. The key elements include the identification and creation of the common base statistical unit (BaSE) that can serve the different output needs, and to let micro-data editing be actively driven from a macro accounting perspective. We believe this new GSDEM is applicable to many situations of multisource statistics.

I. Introduction

1. It is becoming increasingly common to use administrative data sources to produce statistics. This is part of the modernisation process at many national statistical offices (NSOs) whereby the goal is to make best use of available data, reduce the response burden and production costs. In order to make the statistics relevant and to improve quality, usually several sources need to be combined. There arises thus an issue of ***editing multisource data for producing multiple statistics***.

2. In Version 1.0 of Generic Statistical Data Editing Models (GSDEM, UNECE, 2015), Statistical Data Editing (SDE) flow models were used to describe the sequencing and conditional flow logic among different SDE sub-processes. One of the aims is to provide a common language for international collaboration on statistical data editing. Another is to provide a theoretical framework for assessment of the current processes. In particular, Scenario e is the flow model that explicitly addresses SDE of multisource statistics. Under this model, datasets are collected from different sources, on which micro-integration SDE functions of units and/or measurements are performed, and the data is funnelled through the SDE processes towards to the target statistical output. Here the target statistics may well be the same as previously based on sample surveys, and may be envisaged as 'one' end product prepared for a single release, e.g. one for employment, one for education, one for annual income, etc. Two examples of this 'traditional' approach at Statistics Norway are register-based yearly Employment and Household Income statistics, respectively.

3. For producing *multiple* statistics from the *same* set of input data, a straightforward implementation of this model typically leads to what we call ***parallel editing*** (Figure 1, left), where after transformation and linkage the data is split into separate processes parallel to each, each of which is designed and operated for the purpose of a specific statistics. However, it would most likely result in

end products that are not as coherent or comparable with each other as one would expect, given the fact that they are all derived form the same set of raw data.

4. The editing process of the A-ordningen (of contractual payments, social and health benefits, etc.) in Norway is different. Instead of parallel editing, the process for A-ordningen is characterised by implementing as many common SDE functions as possible, before the process and data sprout into different subject domains. One may refer to this as ***coordinated editing*** (Figure 1), which represents a new GSDEM of *multicourse data for multiple statistics*. The aim of such a flow model is to clarify and secure the coherence between different statistics, which in turn can lead to new statistical outputs that cannot be achieved using the traditional approach. The key elements of the approach include the identification and creation of the common base statistical unit (BaSE) that can serve the different output needs, and to let micro-data editing be actively driven from a macro accounting perspective. In the case of A-ordningen, coordinated editing enables us to produce a Labour Force Account (LFA) that explicitly connects Employment and Wage statistics, which is of a higher quality than what was possible previously. We believe this new GSDEM is applicable to many similar situations.

5. Below we first elaborate and compare parallel and coordinated editing in slightly more formal terms, as depicted in Figure 1. More details of coordinated editing in A-ordningen are then given to illustrate the key elements of the approach and the new LFA is briefly outlined.

II. Two flow models of multisource data for multiple statistics

6. A common process flow is to combine multisource data through micro-integration SDE functions of units and/or measurement, where the input data is funnelled through the processes towards the target statistical output. Formally speaking, the data matures and moves through various so-called ***steady states*** (Rennsen & van Delden, 2009). ***Raw data*** is the first state of data received/collected by the NSO. This may be in a variety of formats and is cleaned and transformed in a way to have one or several common units, to which various measures (or values) are associated. Afterwards the data reaches the 2nd state called ***micro-data,*** at Statistics Norway and in this paper. A defining characteristic of micro-data is that unit-level linkage is now possible with another data sets. The data is then edited and processed further to reach the state which we call ***statistical data,*** which forms the basis of describing properties of real-world objects, i.e. producing statistics, the statistical units of interest are unequivocally delineated and the target population definitively formed. Most of the substantive editing and weighting takes place between micro-data and statistical data. Finally, seasonal adjustments or other techniques may be used to create the statistical output, i.e. ***disseminated data***, which can be macro statistics in the form of table, figure, etc. or unit-level dataset.

7. Traditionally, for multiple statistics that are all based on the *same* set of raw data, editing is split into separate parallel processes between the micro-data and the multiple statistical data. Each of these parallel processes is designed and implemented for the purpose of the corresponding end product. This has the advantage that each statistics can tailor editing specifically to its own outcomes, and is able to define its own population and derives its own variables in a *convenient* way. However, it almost always results in outcome statistics that are not directly compatible with each other. Thus, while each end statistical data may be reused in a new life cycle for a different statistical output, it is e.g. usually impossible to present a macro account, which explicitly relates, say, two statistics that are both derived from the same input data set by means of parallel editing.

8. If two statistical datasets refer to the same real-world objects (or units), then ideally the same identification should be associated with these objects (Renssen & van Delden, 2009). This ensures the compatibility and coherence of the processed statistical and disseminated datasets. Very often, however, the same socio-economic phenomenon can be described in terms of *different* objects. An example is A-ordningen, to be described in more details later on, where the various payment data can be summarised in terms of Employment statistics as well as Wage statistics. While person is the unit of employment, job is the unit of wage, so that the two target populations of Employment and Wage

statistics, respectively, consist of different statistical units. Similar discrepancy in measures can also arise. For instance, while wage can possibly be calculated on an hourly basis, employment is usually described in percentage of full-time equivalent (FTE). Therefore, when the same micro-data is subjected to parallel editing towards different target statistics, the data will undergo different transformations, and the resulting statistical outcomes will become incongruent.

9. In the design of coordinated editing, however, common unit and variable harmonisation and re-classification becomes an explicit concern and receives special attention. The relevant SDE functions are formed to ensure the coherence and compatibility of the resulting distinct statistics, which in return allows one to relate the different statistics explicitly through accounting relationships that can be referred to as macro accounts. An example of LFA will be outlined later on. A key element of coordinated editing is therefore the identification and creation of a suitable common ***base statistical unit (BaSE)***, on which one can organise the various units and measures relevant for the different statistics. In the case of Employment and Wage statistics based on A-ordningen, a so-called work relation is created from the micro-data as the BaSE, during the editing processes leading up to the ***intermediate statistical data*** (Figure 1). This allows e.g. a person to have several work relationships with one or several employers.

10. At some point, SDE functions that are specific for each statistics will become necessary, in order to arrive at the respective statistical units and the associated measures. In Figure 1, intermediate statistical data marks the point, before which all the SDE functions are common to the different statistics included in the design of coordinated editing, and after which the SDE functions are specific to each statistics. We make two observations. First, depending on the situation, the intermediate statistical data may or may not be physically stored as such for posterity. However, even when the data is not physically preserved as such, it should still be possible to identify a point in coordinated editing, where the data corresponds to the intermediate statistical data in concept. Second, coordination is still necessary in the separate editing beyond the intermediate statistical data (marked in Figure 1), in order to ensure coherence and compatibility of the different statistics, and to be able to construct macro accounts that explicitly relate them to each other. Some illustrations will be given later on.

11. Compared to parallel editing, coordinated editing has several advantages: 1) it promotes a single editing flow (i.e. towards the intermediate statistical data) which avoids duplication of SDE functions and potentially the creation of incompatible functions in a parallel design, 2) it ensures that multiple statistical outputs are harmonised and compatible with each other, 3) it provides the possibility for generating new statistics that would not have been possible by parallel editing.

III. Example: A-ordningen from Statistics Norway

12. Statistics Norway has access to around 80 administrative registers and works actively to ensure good cooperation with their owners. With a production of around 1000 statistics annually, the emphasis on using these registers effectively is paramount. An important example is the newly operational A-ordningen. This is a cross-departmental administrative system for reporting employment, wage and other payment data. Since 2015, it has replaced the 5 previous reporting systems with a *single point of entry* that is updated monthly. The data is above all used by two major governmental departments (Norwegian Tax Department and Norwegian Labour and Welfare Administration) and shared with Statistics Norway. The system collects data on salary payments, hours worked, taxes, social benefits, etc. The reporting is organised at the employer level. The raw data is then distributed to the two governmental departments and Statistics Norway.

13. A-ordningen now provides the basis for statistics on register-based Employment and Wage statistics. Previously, Wage statistics were based on yearly and quarterly sample surveys and yearly updated administrative data, whereas register-based Employment statistics were derived yearly from a number of administrative sources, most of which are now replaced by A-ordningen. One of the main advantages of the A-ordningen, from a statistical perspective, is that it provides a greater possibility to harmonise the Employment and Wage statistics, since all the data are now available at the same time.

Clearly, by definition, the two statistics are closely connected with each other and the difference, while fundamental should not detract from the objective of harmonisation.

14. Coordinated editing of A-ordningen starts with a complex system of editing at the source. It transforms and links the 5 sources of incoming raw data into micro-data, which involves much initial automated editing. It is then transformed into a monthly file (called L2) as the intermediate statistical data in Figure 1, where the BaSE is work relation as explained above. The intermediate statistical data is further subjected to separate editing for Employment and Wage statistics, by the respective divisions responsible for the two statistics. For Employment statistics, multiple work relations (BaSEs) may be combined at the person level (i.e. the statistical unit for Employment statistics) to identify main and secondary jobs and the related measures such as job percentages. For Wage statistics, work relations (BaSEs) are transformed into jobs (i.e. statistical unit for Wage statistics) with relevant measures such as working hours and standardised amount of salary.

15. Since person and job are two different units, one would have been unable to express explicitly the connection between the two statistics, had they been processed by means of parallel editing. The creation of BaSE in A-ordningen is the necessary first step for consolidating the coherence between the two. But coordination is also necessary after the data has reached the state of intermediate statistical data. As illustrated in Table 1, the BaSEs in L2 can be partitioned into four sets:

U_{11}: included in both Employment and Wage statistics

U_{10}: only included in Employment statistics

U_{01}: only included in Wage statistics

U_{00}: not included in Employment or Wage statistics

	September 2016	September 2015
Common to both (U_{11})	2 678 586	2 653 513
Only in Employment statistics (U_{10})	235 468	132 229
Only in Wage statistics (U_{01})	21 234	17 818
Excluded from both (U_{00})	2 754 146	2 839 966
Total	**5 689 434**	**5 643 526**

Table 1. Partition of base statistical units for Employment and Wage Statistics.

Notice that the number of BaSEs is generally not equal to the number of persons, which are the statistical unit of Employment Statistics, e.g. because one person can have multiple work relations in L2. Neither is the number of BaSEs generally the same as the number of jobs, which are the statistical unit of Wage statistics, e.g. because not every work relation in L2 is admitted as a job by definition. Thus, while the summary overview in Table 1, which is made possible by the introduction of BaSE, is necessary for achieving the coherence between the two statistics, further coordination e.g. in the delineation of the respective target populations remains necessary, in order not to 'lose' the coherence in the subsequent separate editing towards Employment and Wage statistics. Coordination in this way can also highlight similarities and differences. One outcome of this study is to transfer common processes identified back into the SDE flow prior to L2. It is also expected that populations for these two statistics can be aligned more closely to further improve comparability.

16. Finally, we briefly outline here the approach to the Labour Force Account (LFA) as a new statistics, by means of coordinated editing. To start with, the LFA is divided into two parts. The first part will be based on the common BaSE sub-population (U_{11}), where an account will be constructed that connects the two real, observable totals: the total of employed persons (Y) and the total of contractual wage payment (W). More specifically, the two will be connected by two accounting equations: a) one from Y to the total number of FTEs (T), according to a suitable breakdown by NACE and profession, b) one from T to W via Wage per FTE. We refer to this as the *substantive* part

of the LFA. The second part will be based on the BaSE sub-population only included in Employment statistics (U_{10}), which will be referred to as the *normative* part of the LFA. Several possibilities and their potential uses are currently under investigation. For instance, some of these employed persons are on leave for various reasons. One may calculate the would-have-been Wage cost had these people all return to work now, which can be contrasted to the actual benefit payments they receive.

References

Renssen, R., & van Delden, A. (2009). *Standardisation of design and production of statistics: A sevice oriented approach at Statistics Netherlands.* The Hague: Statistics netherlands.

UNECE. (2015). *Generiv Statistical Data Editing Models, GSDEMs, (version 1.0, October 2015).*

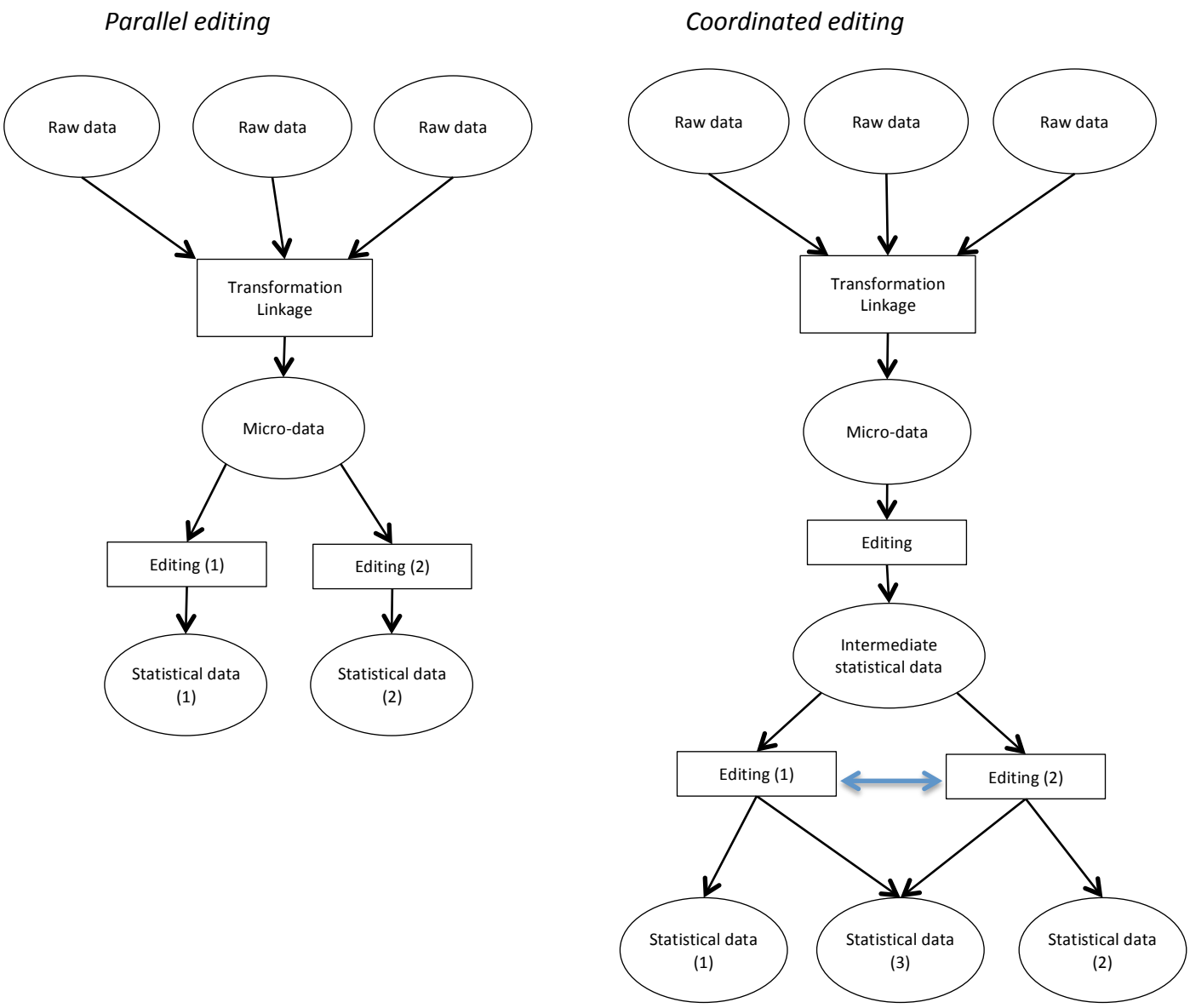

Figure 1. Alternative SDE flows. Left shows an approach where editing split after the micro-data stage and follow parallel processes towards two statistics. Alternative approach on right shows coordinated editing to form one intermediate statistical dataset, which can then be used to produce coherent multiple statistics, including macro counts based on them.