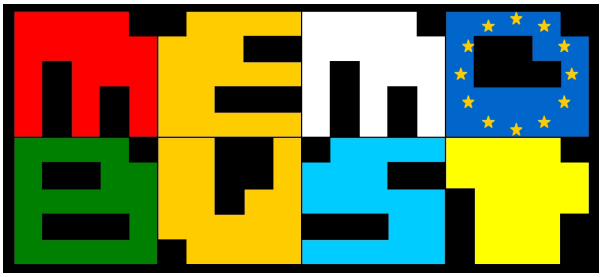

This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Reconciling Conflicting Microdata

Contents

General section

1. Summary

In data fusion we consider microdata consisting of records that are composed of information from different sources. Such composite records may consist of several combinations of sources (see the module "Micro-Fusion – Data Fusion at Micro Level"). Records may be a combination of values obtained from a register with values obtained from a survey for the same units (obtained by record linkage). Records may also combine information from several surveys with non-overlapping units, in which case a unit from one source is matched with a similar (but not identical) unit from another source. In addition, records with values obtained from different sources can also arise as a consequence of item non-response and subsequent imputation in which case the two sources are the directly observed values versus the values generated by the imputation method.

In all these cases the composition of a record by combining information obtained from different sources may give rise to consistency problems because the information is conflicting in the sense that edit rules that involve variables obtained from the different sources will often be violated.

The purpose of reconciling conflicting microdata is to solve the consistency problems by making slight changes or adjustments to some of the variables involved. Apart from the choice of variables to be adjusted, an adjustment method should also be specified since there are a number of methods to handle the adjustment problem. In this module three different approaches to the reconciliation problem will be described and the properties of the solutions will be discussed.

2. General description of the method

2.1 Composite records arising in micro-fusion and imputation

In this module we are concerned with the task of reconciling conflicting information in statistical microdata that may arise if (some of) the individual records are composed of data obtained from different sources. In the module "Micro-Fusion – Data Fusion at Micro Level" two general cases have been described that give rise to such composite units: record linkage (see also the theme module "Micro-Fusion – Object Matching (Record Linkage)") and statistical matching (see also the theme module "Micro-Fusion – Statistical Matching"). In addition, imputation for non-response (see the topic "Imputation") also creates a composite record. Thus we have the following three situations in which composite records can arise:

<u>Record linkage</u>

This type of data fusion, which is a common and increasing practice in the production of business statistics, concerns the linkage of (usually) a sample survey to a register. In this case the linked records consist of register information combined and enriched with survey information on the same units. Both sources will usually also have a few variables in common, apart from the variables used to identify the unit that are necessary for the linking process. In business statistics the main administrative source today is the tax register, providing information on at least the total turnover, which will be a common variable since it will also be measured in the survey. It should be noted that such common variables may have different values in the register and the survey.

Statistical matching

The second case concerns the integration of two (or more) sample surveys which have some variables in common while others are specific for each of the sources. Let the set of common variables be denoted by X and the sets of specific variables by Y and Z. Usually the samples will be (almost) non overlapping and therefore there will be no units with all sets of variables observed. In this case synthetic records can be constructed from one of the sources, say with Y observed, by filling in or imputing the variables Z. These imputations can be obtained by a regression model relating Z to X, which can be estimated using the other source where both Z and X are observed. Alternatively a hot-deck imputation method can be used where values for Z are obtained from a similar record from the other source, found by matching on the common variables X (see Figure 2 and the accompanying text in the module "Micro-Fusion – Data Fusion at Micro Level" or D'Orazio et al., 2006). In the case of hot-deck imputation the composite record consists of values obtained from different but similar units.

Imputation

Records with values obtained from different sources can also arise as a consequence of item non-response and subsequent imputation. In this case one of the sources of the composite record consists of observed values and the other of imputed values derived from a parametric or nonparametric imputation model. This situation is similar to the one arising from statistical matching since in both cases the composite record consists of observed and imputed values. The difference is, however, that the synthetic records in statistical matching all have the same variables imputed, while in the item non-response case the non-response pattern and hence the variables requiring imputation, can be different for each record.

2.2 Introduction to the micro-level consistency problem

To illustrate the consistency problem at micro level, we consider the following situation that arises in business statistics (cf. Pannekoek, 2011). There is information on some key variables available from reliable administrative data. Let these variables be the total turnover *(Turnover)*, the number of employees *(Employees)* and total amount of wages paid (*Wages*). These variables are used to compile the short term economic statistics (STS) and are published quarterly as well as yearly. The yearly structural business statistics (SBS), requires much more detail and this more detailed information is not available from registers. Therefore, a sample survey is conducted to obtain the additional details. After linking the sample data to the register, the situation arises that for the key variables, two sources are available for each responding unit in the sample: the register value and the survey value and for the other variables only survey values are obtained. To be consistent with already published STS figures on *Turnover* and possibly other key variables, the register values are used for the key variables and the survey values for the other variables. Thus we create composite records based on two sources: register and survey. This is illustrated in table 1 below. The column *Survey values* displays the survey values of the eight variables for a responding unit. In the column *Composite (I)* the values of the composite record are shown; the survey values for the key variables are replaced by the register values (in bold). As an alternative we also consider, for illustrative purposes, the situation that we only have *Turnover* available from administrative sources resulting in the values in the column *Composite (II)*.

Business records generally have to adhere to a number of accounting rules and logical constraints. These constraints are widely employed for checking the validity of a record and are, in this context,

referred to as edit rules (see “Statistical Data Editing – Main Module”). For the example record above, the following three edit rules are formulated:

e_1: $x_1 - x_5 + x_8 = 0$ (*Profit = Turnover – Total Costs*)

e_2: $-x_3 + x_5 - x_4 = 0$ (*Turnover = Turnover main + Turnover other*)

e_3: $-x_6 - x_7 + x_8 = 0$ (*Total Costs = Wages + Other costs*)

Notice that these edits are connected by the variables *Turnover* and *Total Costs*, which is true for many of the edits used in business statistics and has consequences for adjustment for consistency.

Table 1. Example Business record with data from two sources

Variable	*Name*	*Survey values*	*Composite (I)*	*Composite (II)*
x_1	Profit	330	330	330
x_2	Employees (Number of employees)	20	**25**	20
x_3	Turnover main (Turnover main activity)	1000	1000	1000
x_4	Turnover other (Turnover other activities)	30	30	30
x_5	Turnover (Total turnover)	1030	**950**	**950**
x_6	Wages (Costs of wages and salaries)	500	**550**	500
x_7	Other costs	200	200	200
x_8	Total costs	700	700	700

Both composite records lead to violation of the edit rules, which we refer to as the micro-level consistency problem. In particular, composite record (I) violates all three edit rules and composite record (II) violates the two edit rules involving *Turnover*. To obtain a consistent record some of the values have to be changed or “adjusted". Since the register values are considered reliable and already used in publications, the survey values are an obvious choice in this case.

When the data are obtained from a single source, e.g., a single survey questionnaire, the violation of hard edit rules that describe relations between variables, such as the balance edits, indicate that a response error has occurred. When data are from different sources, edit rules that describe relations between variables can also be violated by (slight) differences in definitions of variables or time differences between the two sources. In such cases the cause of the violation need not be a response error and is therefore termed an inconsistency between the sources.

The example above is just a simple illustration, in practice the number of variables as well as the number of edit rules can be much larger. The structural business statistics (SBS) are an example with a large number of variables and edit rules. An SBS questionnaire can be divided in sections. It contains, for instance, sections on employees, revenues, costs and results. In each of these sections a total is broken down in a number of components that can again be broken down in sub-components. Components of the total number of employees can be part-time and full-time employees and components of total revenues may be subdivided in turnover and other operating revenues. The total costs can have as components: purchasing costs, depreciations, personnel costs and other costs. Each of these breakdowns of a (sub)total corresponds to what is called a balance edit. SBS questionnaires also contain a profit and loss section where revenues are balanced against the costs to obtain the

results (profit or loss), which leads to edits of the form e_1. This last type of edit connects the edits from the costs section with the edits from the revenues section. Therefore, almost all variables are connected by edit rules and changing one variable will lead to necessary changes in most other values if the structure as laid down in the edit rules is to be preserved. In some cases there is no explicit connection between variables specifying employment in terms of the numbers of employees in different categories and the other, financial, variables. Since relations between, e.g., number of employees and wages should be preserved, adjustment methods should take care of relations not specified by edit rules and methods to accomplish this are the method described in the module "Micro-Fusion – Generalised Ratio Adjustments" and an approach discussed in the module "Micro-Fusion – Minimum Adjustment Methods" (section 2.5.2).

2.3 Overview of adjustment methods to achieve consistency

Adjustment methods change (or adjust) some of the values of some variables (the adjustable variables) in a record such that the resulting adjusted record satisfies all the specified edit constraints. Three different adjustment methods are treated in three separate modules: "Micro-Fusion – Prorating", "Micro-Fusion – Minimum Adjustment Methods", and "Micro-Fusion – Generalised Ratio Adjustments". Below we give a short overview of these methods.

Prorating is a simple ratio adjustment for balance edits (see Banff Support Team, 2008). It solves the possible inconsistencies for each constraint separately. It is an intuitively appealing method that is easy to interpret and to apply. For composite record (II) in table 1, a prorating adjustment to resolve the violation of edit-rule e_2 would entail multiplying the components of *Turnover*, x_3 and x_4, by the ratio of the register and survey values for *Turnover* (1030/950). This ratio adjustment has the effect that the ratios of the components of turnover to their total become equal to the values of these ratios obtained from the survey, but the levels of the components are consistent with the register value of the total. This reflects the availability of information in the two sources and the priority of the total from the register. A drawback of this method is that for interrelated balance edits the result is dependent on the order in which the edits are treated, which introduces arbitrariness in the solution. In practice different orders can indeed lead to substantially different solutions. Especially for the extensive systems of balance edits encountered in the SBS this can be a problem. This method is treated in more detail in the module "Micro-Fusion – Prorating".

The minimum adjustment approach is to make adjustments to the adjustable variables that are minimal in some sense, such that the adjusted record satisfies all constraints (see Pannekoek, 2011). The minimal adjustments are thus obtained by minimising a chosen distance metric subjected to the edit constraints. Since this optimisation approach treats all edits simultaneously there is no problem with the order in which the edits are handled and it leads to a single optimal solution. This solution does, however, depend on the chosen optimisation criterion. In the module "Micro-Fusion – Minimum Adjustment Methods" the optimisation approach is described and properties of the solutions for three different optimisation criteria are discussed. Some solutions are characterised by additive adjustments that preserve the differences between variables that are part of the same (set of) constraint(s) and other solutions are characterised by multiplicative constraints that preserve the ratios between variables that are part of the same (set of) constraint(s).

The third adjustment method is generalised ratio adjustment (see Pannekoek and Zhang, 2011). The method uses multiplicative adjustments, just as the methods Prorating and one of the minimum

adjustment methods (the KL-adjustments, see the module "Micro-Fusion – Minimum Adjustment Methods"). The generalised ratio adjustments method aims to make the adjustments as uniform as possible. Furthermore, and in contrary to the other methods, the method can result in adjustments to variables that are not involved in the constraints. In this sense it can solve the problem, mentioned at the end of the previous section, of preserving relations between variables that are not connected by edit rules.

3. Preparatory phase

4. Examples – not tool specific

5. Examples – tool specific

6. Glossary

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

7. References

Banff Support Team (2008), Functional Description of the Banff System for Edit and Imputation. Technical Report, Statistics Canada.

D'Orazio, M., Di Zio, M., and Scanu, M. (2006), *Statistical matching: theory and practice*. John Wiley, Chichester.

Pannekoek, J. (2011), Models and algorithms for micro-integration. In: *Report on WP2: Methodological developments*, ESSNET on Data Integration, available at http://www.cros-portal.eu/content/wp2-development-methods.

Pannekoek, J. and Zhang, L.-C. (2011), Partial (donor) imputation with adjustments. Working Paper No. 40, UN/ECE Work Session on Statistical Data Editing.

van der Loo, M. (2012), *rspa: Adapt numerical records to fit (in)equality restrictions with the Successive Projection Algorithm*. R package version 0.1-1.

Specific section

8. Purpose of the method

The purpose of the method is to adjust the values of some variables in a data record to remove edit violations to ensure consistency of the data values obtained from different sources.

9. Recommended use of the method

1. The method should be used after detection and treatment of errors and missing values.

10. Possible disadvantages of the method

1. When inconsistencies arise due to large errors in some values, these errors may propagate to other values due to adjustment. Influential errors should therefore be treated before the method is applied.

11. Variants of the method

1. Prorating
2. Minimum adjustment methods
3. Generalised ratio adjustments

12. Input data

1. Data records with possibly inconsistent values and edit rules.

13. Logical preconditions

1. Missing values
    1. Missing values are allowed but edit rules involving variables with missing values cannot be checked and no adjustment with respect to these edit rules will take place.
2. Erroneous values
    1. Influential erroneous values should be treated before the method is applied.
3. Other quality related preconditions
    1.
4. Other types of preconditions
    1.

14. Tuning parameters

1. The amount of change applied to individual variables can be controlled by specifying weights for the variables.

15. Recommended use of the individual variants of the method

1.

16. Output data

1. The output consists of the same individual records as the input, with values adapted when needed to ensure consistency with the edit rules.

17. Properties of the output data

1. The output data are ensured to be consistent with all specified edit rules that do not involve variables with missing values.

18. Unit of input data suitable for the method

The input consists of individual records that are treated one-by-one, independently.

19. User interaction - not tool specific

1.

20. Logging indicators

1.

21. Quality indicators of the output data

1.

22. Actual use of the method

1. Adjustments of imputed values to ensure that edit rules are satisfied is used in the production process for Structural Business Statistics at Statistics Netherlands.

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Micro-Fusion – Data Fusion at Micro Level
2. Micro-Fusion – Object Matching (Record Linkage)
3. Micro-Fusion – Statistical Matching
4. Statistical Data Editing – Main Module
5. Statistical Data Editing – Editing Administrative Data
6. Imputation – Main Module

24. Related methods described in other modules

1. Micro-Fusion – Prorating
2. Micro-Fusion – Minimum Adjustment Methods
3. Micro-Fusion – Generalised Ratio Adjustments

25. Mathematical techniques used by the method described in this module

1.

26. GSBPM phases where the method described in this module is used

1. Phase 5 - Process

27. Tools that implement the method described in this module

Available software options vary for the three (classes) of methods discussed in this module: prorating, the optimisation approach and generalised ratio adjustment.

1. Statistics Canada's generalised edit and imputation software Banff, contains a routine PRORATE that provides an off-the-shelf, generalised prorating application. However, for specific applications the prorating calculations are not difficult to implement. So, without the availability of generalised prorating software, the application of prorating could be performed by an ad hoc implementation using general statistical packages with programming facilities such as R or SAS.
2. The optimisation methods can be implemented, in general, by using standard (commercially) available solvers for convex optimisations problems and the same holds for the generalised ratio approach. For the optimisation methods based on Least Squares and Weighted Least Squares a specific R-package is freely available (van der Loo, 2012).

28. Process step performed by the method

GSBPM Sub-process 5.3: Review, validate and edit

Administrative section

29. Module code

Micro-Fusion-M-Reconciling Conflicting Microdata

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	05-03-2013	first version	Jeroen Pannekoek	CBS (Netherlands)
0.2	17-04-2013	second version	Jeroen Pannekoek	CBS (Netherlands)
0.2.1	09-09-2013	preliminary release		
0.3	20-12-2013	improvements based on the EB-review	Jeroen Pannekoek	CBS (Netherlands)
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:00