Discussion Paper

Multi-source Statistics: Basic Situations and Methods

The views expressed in this paper are those of the author(s) and do not necessarily reflect the policies of Statistics Netherlands

2017 | 12

Ton de Waal
Arnout van Delden
Sander Scholtus

Content

Summary

National Statistical Institutes (NSIs) all over the world are moving from single-source statistics to multi-source statistics. By combining data sources more detailed statistics can be produced. By utilizing a combination of already available data sources NSIs can also produce more timely statistics and respond more quickly to events in society as one does not have to wait until these data have been collected. By combining survey data with already available administrative data and Big Data NSIs can save data collection and processing costs, without having to place extra burden on respondents. However, multi-source statistics come with new problems that need to be overcome before the resulting output quality is sufficiently high and before those statistics can be produced efficiently. What complicates the production of multi-source statistics is that they come in many different varieties as data sources can be combined in many different ways. Given the rapidly increasing importance of producing multi-source statistics in Official Statistics there has been considerable research activity in this area over the last few years. Some frameworks have been developed for multi-source statistics in recent years. Useful as these frameworks are, they generally do not give guidelines to which method could be applied in a certain situation arising in practice. In this paper we aim to fill that gap, and structure the world of multi-source statistics and its problems and provide some guidance to suitable methods for these problems.

Keywords

Administrative data, Data integration, Multi-source statistics, Statistical methods, Survey data

1. Introduction

National Statistical Institutes (NSIs) all over the world are moving from single-source statistics to multi-source statistics. On the one hand, this is due to higher quality demands with respect to the statistics produced: more detailed data, more timely data, and a general demand for a faster response from NSIs to events in society. On the other hand, many NSIs face budget cuts that make large-scale surveys too costly to set up and maintain.

NSIs traditionally have produced single-source statistics, where basically only data from a single data source are utilized to produce these statistics. Often, also other data sources are used in this process, but only as additional data, for instance as auxiliary data to calibrate or improve estimates or as supplemental data to validate the statistics produced. In most cases the single data sources are surveys, although nowadays administrative data are more and more used as single data sources and also Big Data are starting to be used.

By combining data sources more detailed statistics can be produced. By utilizing a combination of already available data sources NSIs can also produce more timely statistics and respond more quickly to events in society as one does not have to wait until these data have been collected. By combining survey data with already available administrative data and Big Data NSIs can save data collection and processing costs, without having to place extra burden on respondents.

Moving from single- to multi-source statistics therefore seems the way to go. However, this transition is not an easy one. Multi-source statistics come with new problems that need to be overcome before the resulting output quality is sufficiently high and before those statistics can be produced efficiently. A first step that is often taken when one wants to use multiple data sources for producing multi-source statistics is micro-integration (see, e.g., Bakker, 2011). In the micro-integration step, for instance, variables and units in different data sources are harmonized as well as possible and measurement errors in different data sources are corrected for as well as possible. However, micro-integration cannot solve all the problems that arise in the context of multi-source statistics.

What complicates the production of multi-source statistics is that they come in many different varieties as data sources can be combined in many different ways. Every variety seems to come with its own problems for which tailor-made solutions are needed. It often feels like for every new multi-source statistics one has to reinvent the wheel again.

Given the rapidly increasing importance of producing multi-source statistics in Official Statistics there has been considerable research activity in this area over the last few years. Some frameworks have been developed for multi-source statistics in recent years, see, for instance, Bakker and Daas (2012) and Zhang (2012), who focus on processing steps and error sources in multi-source statistics. Useful as these frameworks are, they generally do not give guidelines to which method could be applied in a certain situation arising in practice. In this paper we aim to fill that gap, and structure the world of multi-source statistics and its problems and provide some guidance to suitable methods for these problems.

In the current paper we do not strive to offer an all-encompassing theoretical framework of some kind, such as a framework describing all error types that could arise in a multi-source context or a framework attempting to describe all possible situations. Instead, this paper has a much more pragmatic aim. Our goal is to provide practical guidelines for producers of multi-source statistics on which problems may be encountered and which kinds of methods can be applied to overcome these problems for a number of important basic situations that can arise in practice.

In order to identify the most important research questions with respect to multi-source statistics, we propose a breakdown into eight basic situations that seem to be most commonly encountered in practice. A given situation in practice may well involve several basic situations at the same time.

The remainder of the paper is organized as follows. Section 2 discusses some characteristics of multi-source statistics. These characteristics can be used to identify basic situations for multi-source statistics. Section 3 describes such basic situations, as well as corresponding methodological challenges and methods to overcome these challenges. For each basic situation, we also discuss a real-world example. Section 4 concludes the paper with a discussion.

2. Characteristics of multi-source statistics

The characterization of multi-source data is complicated due to the inherent heterogeneous nature of the sources. We discuss some aspects that seem relevant in many situations.

The first aspect that is relevant is the aggregation level of the data. Do the data sources consist of only micro data, only aggregated data, or are micro data from some sources to be combined with aggregated data from other sources? Obviously different kinds of methods are likely to be needed for these cases. Likewise, it is also important whether cross-sectional or longitudinal data sources are to be combined with each other.

When data sources are to be combined, (some of) the units in these data sources may be overlapping or none of the units may be overlapping. In some cases the data sources may even contain information on different kinds of units, such as enterprises in survey data and legal units in administrative data (Alajääskö and Roodhuijzen, 2016), or persons in one data source and enterprises in another data source (Ruotsalainen, 2005). If (some of the) units are overlapping, one has more information for these overlapping units than for the other units. This creates its own opportunities and methodological challenges.

Similarly, when data sources are to be combined, (some of) the variables in these data sources may be overlapping or none of the variables may be overlapping. If variables are overlapping in data sources, they can have different values in these sources due to measurement errors. If so, statistical methods can be used to estimate the true values of the overlapping variables.

Also, “population issues” form an important aspect of multi-source statistics. There are several of such population issues. First, in some cases the population is known, i.e. we have a register containing all units in the population. In other cases such a population register is lacking. The absence of a population register obviously complicates the estimation process. In this paper we focus on the situation where we do have a population register. Second, there can be under- or overcoverage in (some of) the data sources. That is, selective groups of units may be missing in a sample or in a supposedly complete enumeration of the population (undercoverage) or units may unwittingly be duplicated in a sample or in a complete enumeration of the population (overcoverage). Finally, a data source can contain a complete enumeration of its target population, it can be selected by means of probability sampling from its target population, or it can be selected by non-probability sampling from its population. For the estimation process, the first case, i.e. a complete enumeration, is the easiest case to deal with. The second case, where the data are selected by means of probability sampling, can often be dealt with by means of weighting or imputation techniques. The last case, where the data are selected by means of non-probability sampling, is the hardest case to handle.

An important and fundamental problem when combining different data sources is timeliness of these sources. Different data sources may be available at different moments and the quality of the data sources may vary over time. In particular, the progressiveness of administrative data, i.e. the fact that administrative data sources generally contain more and/or higher quality data as time passes, often presents a problem for early estimates. Progressiveness of administrative data is a fundamental problem in the sense that there is not much we can do about it, apart from estimating the effect of progressiveness after more and/or higher quality data have become available, and predicting (better) values for the data that are still missing and/or are of lower quality while producing early estimates (see also Zhang, 2014).

3. Basic situations and their methods

In this section we present eight basic situations that we consider to be the most important ones in practice. We propose and elaborate these basic situations with respect to the aspects mentioned in Section 2. Many practical situations can be built on these basic situations.

In our discussion we assume that the main aim of multi-source statistics is to produce high quality estimates on an aggregated level. We assume that the construction of combined micro data, while important for research purposes, is considered to be less important than the production of estimates on an aggregated level.

We distinguish the following eight basic situations:

1. Data sources with full population coverage, non-overlapping in variables;
2. Data sources with full population coverage, non-overlapping in units;

3. Overlapping variables but non-overlapping units
4. Overlapping variables and overlapping units
5. Undercoverage
6. Aggregated data only
7. Micro data and aggregated data
8. Longitudinal data

Below we discuss each of these basic situations in detail.

3.1 Data sources with full population coverage, non-overlapping in variables

The first basic situation concerns multiple cross-sectional data sources covering the target population where the different data sets contain different target variables (see Figure 1). We refer to this as the “split-variable” case. Provided that the data are error-free, the data can simply be linked to produce output statistics.

Concerning the illustrations in this document note that:

1. The rectangle of white blocks to the left represents the population frame;
2. Different blue colours represent different input data sources
3. Orange/brownish colours represent derived output statistics
4. Shaded blocks represent aggregated data, bright blocks represent micro data.

Figure 1. Combining micro data sources non-overlapping in variables, without coverage problems

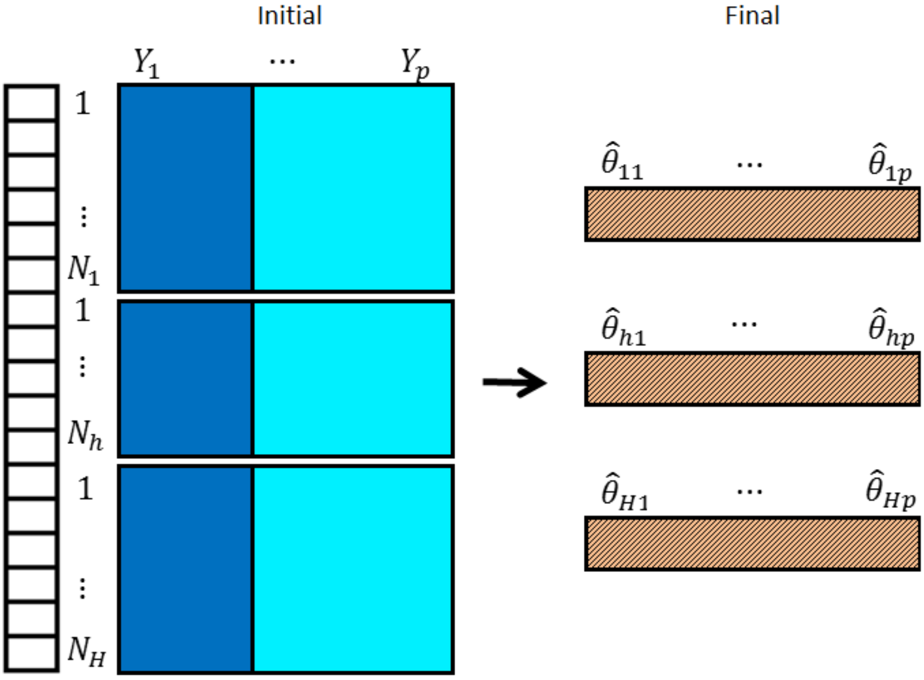
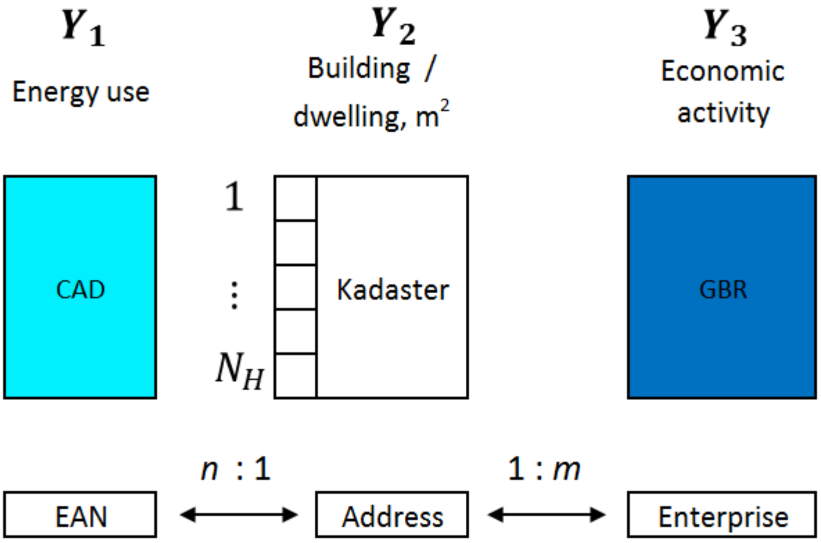


Figure 1 illustrates the situation that we are interested in: estimating a set of p target parameters, denoted by $\hat{\theta}_{h1}, \dots, \hat{\theta}_{hp}$, for a set of domains $h = 1, \dots., H$ within the population. The target parameters within each domain are based on the corresponding variables $Y_{h1}, \dots, Y_{hp}$ that are observed for units $1, \dots, N_h$ in the case of

a full enumeration of the population, or observed for units $1, \dots, n_h$ with $n_h \leq N_h$ in the case of a sample. In the latter case weighting procedures are often used to obtain estimates for population parameters, such as totals and means (see, e.g., Särndal, Swensson and Wretman, 1992). The sampling case will probably not arise often in Situation 1, but it is common for the basic situations that are discussed later.
We assume that the data sets also contain a set of background variables $\boldsymbol{Z}$, where $\boldsymbol{Z} = (Z_1, \dots, Z_k)'$, for instance variables that are used to link the data sets to the population register. Background variables are omitted from the figures unless they are a crucial part of the estimation procedure of the target parameters.
An example of Situation 1 is the population census where in some countries part of the variables are collected through a census survey and the remaining variables are obtained from administrative sources (UN/ECE, 2014, pp. 11–12). Another example is the integration of different administrative data sources on economic performance of businesses. For instance, in the Netherlands administrative data on profit and loss are sometimes combined with administrative data on personnel costs.
Although the estimation of the target parameters in itself seems straightforward, there are at least two potential problems that may also apply to the other basic situations to be discussed below. First of all, the units in the different sources may need to be harmonized. This may for instance be the case when administrative data sources on businesses are combined. In the Netherlands for instance, administrative units for Value Added Tax (VAT) data may differ from administrative units for profit and loss data. In turn, those administrative units may differ from the statistical units for which the target population is defined.
Second, we need a record linkage step to link the units in the sources to the population register. When unique unit identifiers, such as unique personal identification numbers, are present in the sources, deterministic linkage can be used (Herzog, Scheuren and Winkler, 2007). When the same non-unique identifier variables, such as names and addresses, are present in both sources, probabilistic linkage might be used. The classical paper on probabilistic linkage is by Fellegi and Sunter (1969).
An example of unit type differences and linkage issues occurred in the integration of various administrative data sources at Statistics Netherlands (SN) to compute energy use per m2 for dwellings and for businesses or institutions (Figure 2). These outcomes are further stratified by type of dwelling and by type of economic activity. The central data concern administrative client energy data sets (CAD) obtained from gas and electricity distributors, which consist of the complete volume of energy delivery in the Netherlands. The CAD is linked to a central register on addresses and buildings (Kadaster), which contains building / dwelling type and their area. It is also linked to a general business register (GBR) to identify business activities and to find the economic activity. Various other administrative data sets are used to refine and improve the stratifications (not shown in Figure 2).

Figure 2. Some of the data (and unit types) for compiling energy use per square metre by economic activity



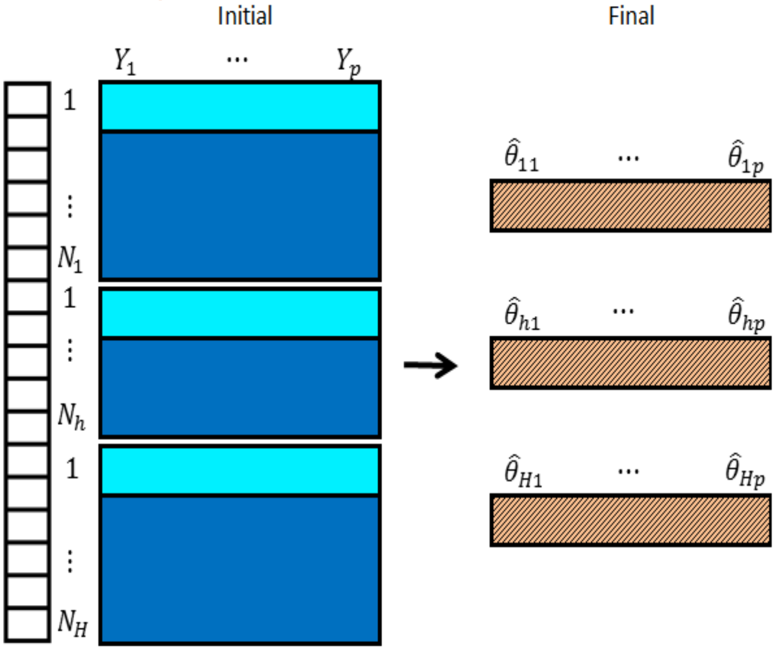
The unit type within the CAD is the “energy (gas or electricity) connection point”, identified by a unique energy connection point number (Dutch: EAN). The EAN is related to an address (postal code, house number and house number suffix) and client name. This address information is also found in the Kadaster data and in the GBR data. The linkage by address is not always unique. One address may contain multiple energy connection points, which can be solved by adding up the energy use of the different EANs. In addition, one may also have one EAN that is linked to a building that contains multiple activities / enterprises. In this case one appoints the energy use to the dominant activity in that building, which is not an ideal approach. Linkage issues occur due to spelling and format variations in the address, which need to be harmonized. The original address information led to a matching rate of 80%, after harmonization of the spelling variations this increased to 96% (ECSM, 2017, Chapter 4b).

3.2 Data sources with full variable coverage, non-overlapping in units

The second basic situation also concerns multiple cross-sectional data sources covering the target population, but in this case the different data sources contain different units (see Figure 3). We refer to this as the “split-population” case. In this situation there might be differences in the conceptual definitions of the variables. Provided the data are in an ideal error-free state and the concepts are identical, the different data sources are complementary to each other in this case, and likewise to Situation 1 they can be simply “added” to each other in order to produce output statistics.

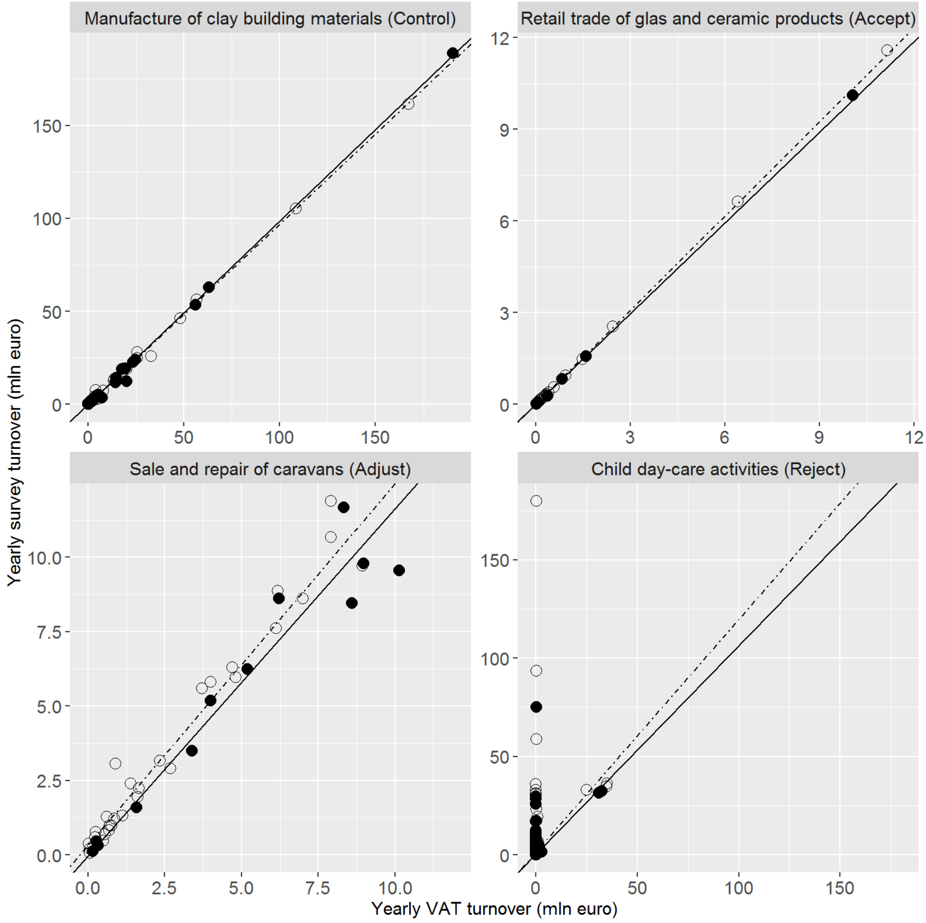
An example of Situation 2 is the estimation of quarterly turnover at SN. The turnover data are available from a combination of census data and administrative VAT data, and both are linked to the GBR (Van Delden and De Wolf, 2013). The VAT data are available for fiscal administrative units, and they can be uniquely linked to the enterprises in the GBR only for the small and medium sized enterprises. The complementary group of large and complex enterprises receives a census survey. Statistics New Zealand (Chen, Page and Stewart, 2016) use a very similar approach, where sub-annual sales data are obtained from administrative Goods and Service Trade data, complemented by survey data for the large and complex units.

Figure 3. Combining non-overlapping micro data sources without coverage problems


Initial
Final
Y_1
$\cdots$
Y_p
1
$\vdots$
N_1
1
$\vdots$
N_h
1
$\vdots$
N_H
$\hat{\theta}_{11}$ $\cdots$ $\hat{\theta}_{1p}$
$\hat{\theta}_{h1}$ $\cdots$ $\hat{\theta}_{hp}$
$\hat{\theta}_{H1}$ $\cdots$ $\hat{\theta}_{Hp}$

A problem that is sometimes encountered in Situation 2 is that the target variables in the sources may need to be harmonized. In the quarterly turnover example at SN, the variable in the administrative source may differ from the one in the survey. Derivation rules, as in micro-integration (Bakker, 2011), may be used to derive the target variables from those present in the input sources. Currently, such rules are often relatively simple and based on expert judgement. More complicated methods of harmonization, which aim to estimate true values of variables, require multiple measurements of the same variables on the same units. This will be treated in Situation 4.

Figure 4. Regression of yearly survey turnover against VAT turnover. Symbols: filled circles and solid line (2009 data), open circles and dot-dashed line (2010) (redrawn from Van Delden et al., 2016)

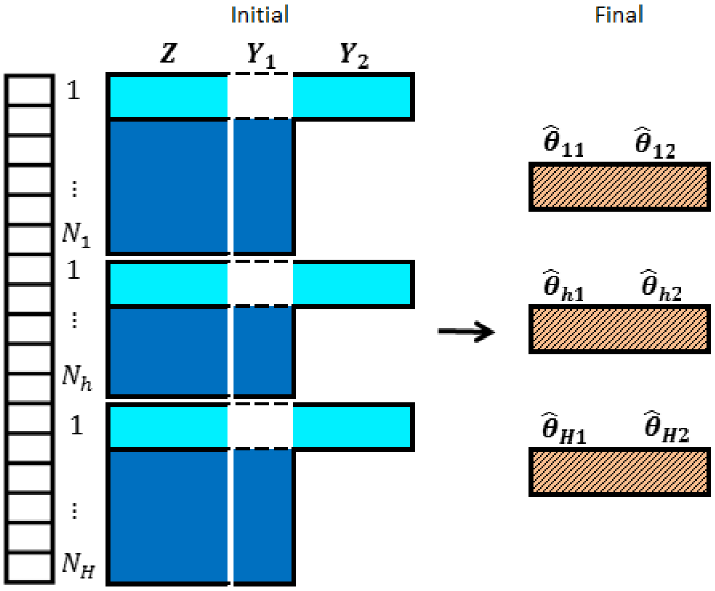


A method to harmonize variables based on multiple sources and that relies on the assumption that one source can be used as the 'gold standard' is given in Van Delden et al. (2016). They analysed the relation between the metadata and the data of annual survey turnover and VAT in 2009 and 2010, for more than 300 domains of economic activity. These domains were used to compute the outcomes of all different publication cells. Van Delden et al. (2016) divided the domains into four groups. The Control group concerns domains where there are no differences in the definitions of survey and VAT turnover. The example in Figure 4 shows a linear relationship with values very close to (0,1) for intercept and slope respectively. The Accept group concerns domains with conceptual differences but only small numerical differences; see the example in Figure 4. The Adjust group concerns domains with conceptual differences and systematic numerical differences. Figure 4 shows the domain "Sale and repair of caravans" where VAT values are systematically smaller than the survey values because in that domain the purchase costs of second hand goods may be subtracted from the declared VAT turnover values. For the units in this domain a correction factor can be applied. The final group, Reject, concerns domains with conceptual differences and large non-systematic numerical differences. The example "Child day-care activities" concerns a domain with units that have a derogation to declare VAT for certain economic activities.

3.3 Overlapping variables but non-overlapping units

A slightly different situation occurs when, besides having non-overlapping units as in Situation 2, we also have a number of overlapping variables and some target variables that are available in only one of the sources. We call this Situation 3 (see Figure 5). We still would like to join the non-overlapping variables for one part of the units to the other units. For this, statistical matching techniques are available.

Figure 5. Combining non-overlapping micro data sources with part of the variables is in a single source, without coverage problems



In Italy, the main data sources available for estimating household income and expenditure are the Household Budget Survey conducted by ISTAT (the Italian National Institute of Statistics) and the Survey on Household Income conducted by the National Bank of Italy. Unfortunately, there is no single data source available that contains data on both household income and expenditure. In order to examine the effects of policy changes on the relation between household income and expenditure one therefore resorts to using statistical matching (see Conti, Marella and Neri, 2015). Statistical matching can be carried out on the micro level or on the macro level.
When statistical matching is carried out on the micro level, one combines data from individual units in the different data sources to construct synthetic records with information on all variables. In particular, in the micro level approach information from one data set, the donor data, is used to estimate target values in the other data set, the recipient data. In this way one constructs a complete synthetic micro data set containing values for all variables of the recipient units. Which of the two data sets is to be selected as the recipient data set is a matter of choice.
When statistical matching is carried out on the macro level, one constructs a parametric model for all the data, for instance a multivariate normal model for numerical data or a multivariate multinomial model for categorical data, and then estimates the parameters of this model. These parameters are subsequently used to estimate the population parameters one is interested in. For an overview of methods

for statistical matching on both the macro level and the micro level we refer to D'Orazio, Di Zio and Scanu (2006).

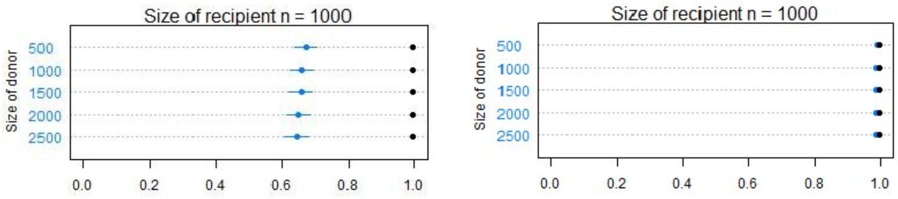
In Figure 5 we have two data sources. Data source 1 (dark blue) contains variables $\boldsymbol{Y}_1$ and $\boldsymbol{Z}$ and data source 2 (light blue) $\boldsymbol{Y}_2$ and again $\boldsymbol{Z}$. Variables $\boldsymbol{Z}$ are the common (background) variables that are used to statistically match the data sources. When statistical matching is carried out on the micro level, variables $\boldsymbol{Z}$ are used to match individual units in data source 1 – the recipients – to individual units in data source 2 – the donors.

The fundamental issue of statistical matching is that the relationship between the target variables $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ cannot be estimated directly, but only indirectly. In order to do so, one has to rely on untestable assumptions, i.e. untestable from the data sources themselves, about this relationship. The most common assumption is the Conditional Independence Assumption (CIA), which says that conditional on the values of background variables $\boldsymbol{Z}$ the target variables $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ are independent. In other words, the CIA says that the relationship between target variables $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ can be entirely explained by the values of the background variables $\boldsymbol{Z}$. As an alternative to the CIA the so-called Instrumental Variable Assumption (IVA) has recently been proposed (see Kim, Berg and Park, 2016). It is unclear yet whether this new assumption is more often justified than the CIA or vice versa.

When neither the CIA nor the IVA holds true, one could attempt to use additional information from a related data source containing (proxies for) both $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$. This additional data source can then be used in a two-step hot deck procedure (see D'Orazio, Di Zio and Scanu 2006). This approach has been studied by Den Ouden (2016) on data from the Dutch Structural Business Statistics from 2007. In particular she used data for the wholesale industry for statistical matching and data from the retail industry as additional information. The data set for the wholesale industry was split into two parts, containing common background variables and a non-common target variable each. Those two parts were subsequently statistically matched, using the common background variables as matching variables.

The results are rather mixed. In one scenario Den Ouden (2016) used “operating revenue” as variable $\boldsymbol{Y}_1$ and “operating expense” as variable $\boldsymbol{Y}_2$. She statistically matched 1,000 records in one part of the data set (the recipient records) with 500, 1,000, 1,500, 2,000, respectively 2,500 donor records in the other part, using no additional information and using additional information from 100 records from the retail industry in which “operating revenue” and “operating expense” were observed. In the true data the correlation between “operating revenue” and “operating expense” is very high (0.993). In the statistically matched data sets without using additional information, the correlations are much lower (a bit over 0.6; see the left panel in Figure 6). In the statistically matched data sets with using additional information, the correlations are almost identical to the correlation in the true data (see the right panel in Figure 6). In this scenario, using additional information leads to excellent results with respect of preservation of the correlation between the target variables.

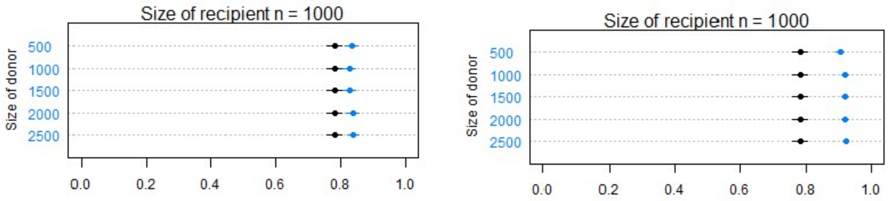
Figure 6. Correlation between "operating revenue" and "operating expense" in statistically matched data without the use of additional information (left panel) and with the use of additional information (right panel) (black = with additional information, blue = without additional information)



In another scenario she used "operating expense" as variable $\boldsymbol{Y}_1$ and "personnel costs" as variable $\boldsymbol{Y}_2$. Again, she statistically matched 1000 recipient records with 500, 1,000, 1,500, 2,000, respectively 2,500 donor records, using either no additional information or additional information from 100 records in the retail industry in which "operating expense" and "personnel costs" were observed. In the true data the correlation between "operating expense" and "personnel costs" is high (0.796). The correlations in the statistically matched data sets without additional information are a bit higher (ranging from 0.824 to 0.842; see the left panel in Figure 7). However, the correlations in the statistically matched data sets with additional information are even higher (ranging from 0.897 to 0.908; see the right panel in Figure 7). So, in this scenario, using additional information leads to an overestimation of the correlation between the target variables.

The main conclusion that can be drawn from Den Ouden's study is that it depends on to what extent the additional information resembles the true data whether the use of additional information is beneficial or not.

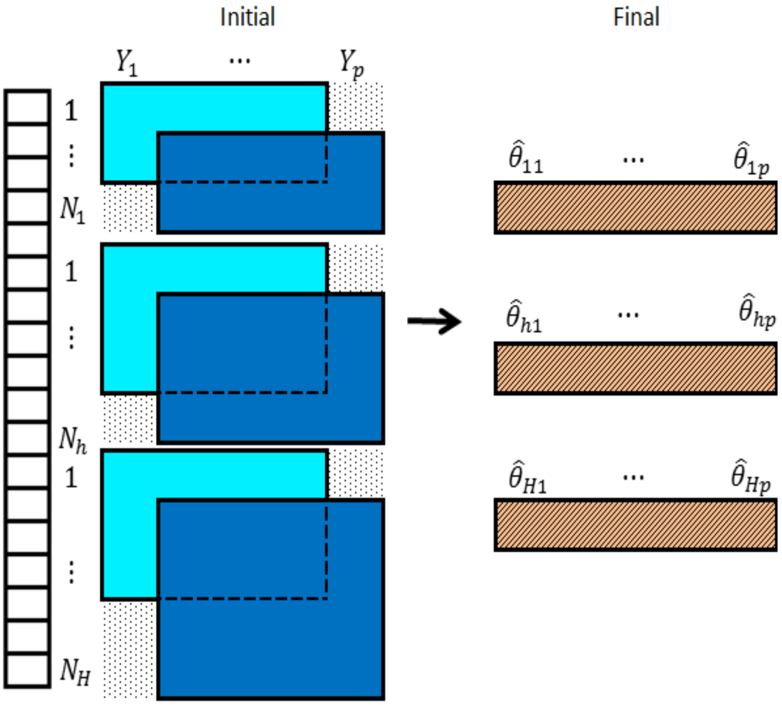
Figure 7. Correlation between "operating expense" and "personnel costs" in statistically matched data without the use of additional information (left panel) and with the use of additional information (right panel) (black = with additional information, blue = without additional information)



3.4 Overlapping variables and overlapping units

Situation 4 (see Figure 8) is characterized by a deviation from Situation 2, by which there exists overlap concerning both units and measurements between the different data sources.

Figure 8. Combining overlapping micro-data sources without coverage problems



In this situation, at least for a subset of the units in the population we have multiple measurements of the same target variable(s), coming from different data sources. Due to measurement errors, these observed variables from different sources will usually not agree exactly for all units.

An example of Situation 4 arises in education statistics in the Netherlands. There exist both administrative and survey data on the education level of Dutch people. The administrative data have a shorter history and do not cover people who completed their education before the time at which registration began. The survey data cover the entire current population but have missing data due to sampling. Some persons can be found in both sources and the respective education level measurements do not always agree with each other. In fact, both sources may contain measurement error. The education level that is derived from administrative data may be wrong when a part of someone's educational career was not registered, e.g., because they studied abroad. In the survey, respondents can make mistakes when reporting their education level.

When the same phenomenon is observed for the same units in multiple data sources, one can utilize the multiple observations to identify and correct residual errors. An approach that is often used in practice at NSIs is micro-integration (Bakker, 2011). In addition to the harmonization step described for Situation 2, in the present situation micro-integration also involves comparing the available observations for each overlapping unit to determine which of the data sources is most likely to contain the best approximation of the true value for that unit. Often, deterministic correction and derivation rules are used for this. In many applications, some form of micro-editing is also needed to obtain consistency between different target variables observed in different sources (Di Zio and Luzi, 2014). In the above example of Dutch education statistics, micro-integration is used to combine the available information from administrative and survey data into a single education level for each person (Linder, Van Roon and Bakker, 2011).

Alternatively, it may be possible to find an appropriate model for the measurement errors in the observed variables. In that case, model-based estimates can be obtained of the underlying true values of the target variable(s), either at the individual level or directly at the level of the target parameters. The true value itself is (usually) not observed; this is called a latent variable. The precise relation between the latent true value and the observed values depends on the type of model. In their basic form, most measurement error models assume that the errors are independent across observed variables, given the underlying true value; this is known as a conditional (or local) independence assumption. For some applications, this assumption may be unrealistic.

For categorical data, models based on latent class analysis can be used (e.g., Hagenaars and McCutcheon, 2002). Application of latent class models to measurement errors in statistical data are considered by, among others, Biemer (2011), Si and Reiter (2013), Pavlopoulos and Vermunt (2015), Oberski (2017), and Boeschoten, Oberski and De Waal (forthcoming). For instance, Boeschoten, Oberski and De Waal (forthcoming) use a latent class model to model the true value of a variable that is observed (with error) in multiple sources. That latent class model is then used to estimate a distribution of true values for that observed variable, from which multiple values are drawn and imputed. By using multiple imputation, one can obtain not only estimated true values, but also a measure of the uncertainty of these values.

To model measurement errors in numerical data, one may use a structural equation model (e.g., Bollen, 1989) or a finite mixture model (e.g., McLachlan and Peel, 2000). Within official statistics, structural equation models have traditionally been used for questionnaire design (Saris and Andrews, 1991) and econometrical analyses (Bound, Brown and Mathiowetz, 2001). Recently, applications to multi-source statistics have been considered by Bakker (2012) and Scholtus, Bakker and Van Delden (2015). For the same context, finite mixture models have been developed by Meijer, Rohwedder and Wansbeek (2012) and Guarnera and Varriale (2015, 2016). Under such a model, units are supposed to belong to two or more components, where each component has a different distribution of observed values. Guarnera and Varriale explicitly consider the case that measurement errors are 'intermittent': part of the observed values are correct and the remaining values contain errors.

Scholtus, Bakker and Van Delden (2015) used a structural equation model for the estimation of quarterly turnover at SN (see Situation 2). In the approach by Van Delden et al. (2016) that was mentioned above, VAT turnover was compared to turnover from a survey, under the assumption that the survey turnover was correct. In the application by Scholtus, Bakker and Van Delden (2015), a structural equation model was used in which three observed turnover variables – VAT turnover, survey turnover and turnover from a different administrative data source, the Profit Declaration Register (PDR) – were all seen as error-prone measurements of a latent variable "true turnover". Furthermore, for a small subsample of the original data set, an additional manual editing effort was made to obtain the true turnover value. For the units in this "audit sample", a fourth turnover measurement was included in the model which was assumed to be equal to the latent variable. The inclusion of this audit sample was necessary to identify all parameters of the structural equation model.

Table 1. Parameter estimates for turnover from a structural equation model for two publication cells. Standard errors are shown in parentheses. (Source: Scholtus, Bakker and Van Delden, 2015.)

Parameter	Retail trade of passenger cars			Specialized repair of motor vehicles		
	VAT	survey	PDR	VAT	survey	PDR
intercept	–0.04 (0.04)	–0.01 (0.05)	0.00 (0.05)	–0.04 (0.05)	–0.04 (0.08)	–0.02 (0.08)
slope	0.79 (0.01)	1.01 (0.01)	1.02 (0.01)	1.29 (0.19)	1.21 (0.19)	1.23 (0.20)
validity	0.98	0.99	1.00	0.99	0.98	0.99

Table 1 shows a selection of the results for two different publication cells. For each observed variable, the model contains a linear regression on the latent true turnover, of which the estimated intercept and slope parameters are shown in the table. For instance, the estimated relation between true turnover and VAT turnover for the publication cell “Retail trade of passenger cars” is:

$$\text{VAT turnover} = -0.04 + 0.79 \times \text{true turnover},$$

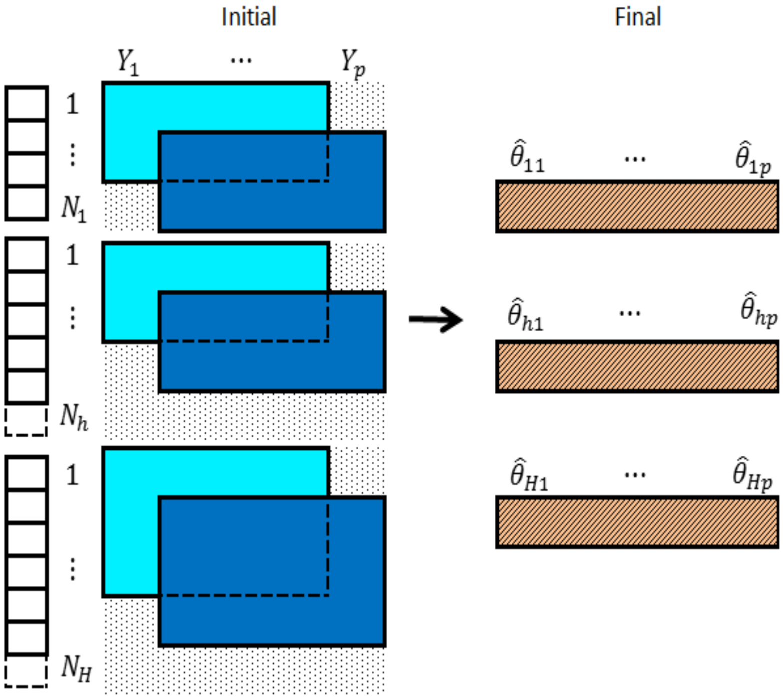
so the observed VAT turnover is about 20% too small on average. By contrast, for the publication cell “Specialized repair of motor vehicles”, the observed VAT turnover is apparently too large on average (estimated slope is 1.29). This illustrates that the relationship between true turnover and VAT turnover differs between publication cells. In particular, it follows that a direct tabulation of the total VAT turnover of each publication cell may give a wrong impression of their relative contributions to the Dutch economy, as some publication cells actually have a larger or smaller turnover from a statistical point of view.

Table 1 also shows the estimated validity of each observed turnover variable, i.e., its correlation to the underlying true turnover. It is seen that these correlations are all close to 1 in this example. This means that the influence of random measurement errors on these observed variables is small. In this case, it may be possible to obtain accurate predictions of the underlying true turnover values from the estimated model, given one or more of the observed values (see also Meijer, Rohwedder and Wansbeek, 2012). The model can thus be used to obtain a correction formula that adjusts the observed VAT turnover to the scale of true turnover, similar to the correction factor that was derived by Van Delden et al. (2016) for the Adjust group (see Situation 2).

3.5 Undercoverage

Situation 5 is characterized by a further deviation from Situation 4, by which the combined data entail undercoverage of the target population, even when the data are in an ideal error-free state (see Figure 9). In this situation, the total population size (and the size of each domain in particular) is not known.

Figure 9. Combining overlapping micro-data sources with undercoverage



Producers of official statistics are often interested in estimating the unknown size of a population. In particular, an important problem in a population census is to estimate the number of persons in the target population who were missed by all data sources used in the census. So-called capture-recapture methods are often used to solve this problem (Fienberg, 1972; Bishop, Fienberg and Holland, 1975; IWGDMF, 1995).

The simplest application of the capture-recapture method is based on two independent samples from the target population. One can imagine a 2 × 2 contingency table of the expected population counts of persons being included or excluded in the first and second sample. Let m_{11} denote the expected number of persons in the overlap of the two samples, m_{10} and m_{01} the expected numbers of persons observed in the first sample but not the second sample and vice versa, and m_{00} the expected number of persons that are not observed in either sample. Similarly, let n_{ij} denote the corresponding observed counts in the realized samples ($i = 0,1; j = 0,1$). By definition, one observes $n_{00} = 0$: a structural zero. If the samples are independent, an estimate for m_{00} can be obtained from the other observed counts as follows (e.g., Bishop, Fienberg and Holland, 1975):

$$\hat{m}_{00} = \frac{\hat{m}_{10}\hat{m}_{01}}{\hat{m}_{11}} = \frac{n_{10}n_{01}}{n_{11}}.$$

An estimate for the total population size, including the part that was missed by both samples, is then given by $\hat{N} = n_{11} + n_{10} + n_{01} + \hat{m}_{00}$. Formally, the capture-recapture method can be derived from a log-linear model for the above contingency table (Bishop, Fienberg and Holland, 1975). This approach is also referred to as Dual System Estimation (Ding and Fienberg, 1994).

An example of Situation 5 where the capture-recapture method can be applied concerns a population census followed by a post-enumeration survey (Wolter, 1986; Brown, Abott and Diamond, 1999 and Brown et al., 2006). Here, the post-enumeration survey is conducted with the specific aim of estimating the undercount in the original population census. The capture-recapture method can also be applied by NSIs that conduct a census based on administrative data (Van der Heijden et al., 2012; Baffour et al., 2013; Gerritse, 2016). In this case, data from at least two administrative sources are linked together and each source is considered as an independent sample from the population.

Gerritse et al. (2016) applied a capture-recapture method to estimate the amount of undercoverage in the population size estimate of the 2011 virtual census in the Netherlands. The census itself was based on the Dutch population register. For the estimation of undercoverage, two additional registers were linked to the population register: an employment register and a crime suspects register. One complicating factor is that the census aims to count the number of “usual residents”, where persons are classified as usual residents if they have lived at least 12 months in the Netherlands or intend to do so at the time of the census. For persons who were found in the additional registers but not in the population register, it is not always clear whether they should be counted as usual residents.

Gerritse et al. (2016) used probabilistic linkage to link the three registers. To handle missing values on the “usual resident” status, two different approaches were used: an EM algorithm and imputation by predictive mean matching. The latter approach was found to be more flexible and therefore preferred by the authors. Due to limitations of the available data, Gerritse et al. (2016) could estimate the amount of undercoverage only for persons between 15 and 65 years. For this age-group, they concluded that there were

- about 33 thousand additional usual residents not covered by the population register but covered by the employment register;
- about 1 thousand additional usual residents not covered by the population or employment register but covered by the crime suspects register;
- between 54 and 151 thousand additional usual residents not covered by any of the three registers.

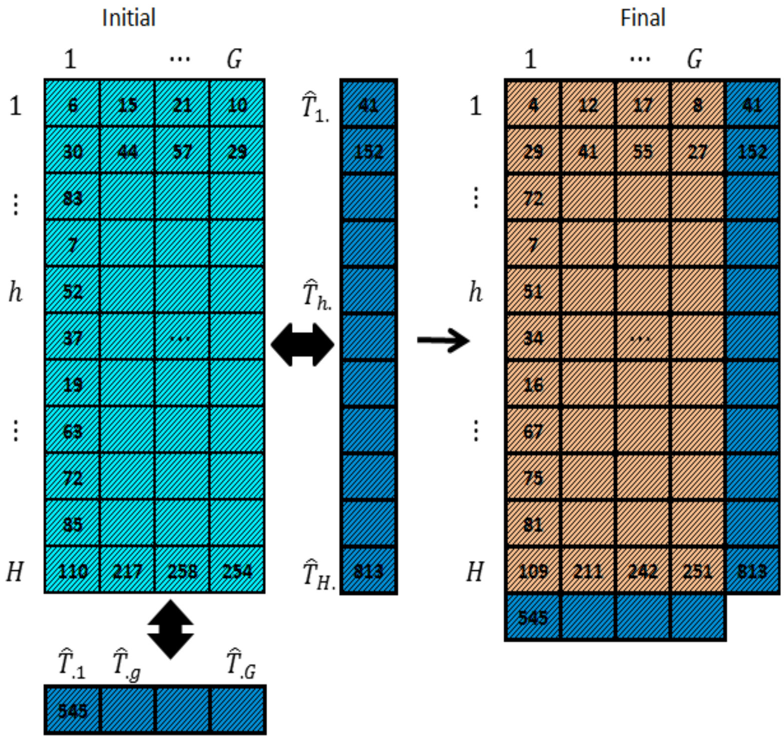
The latter range of estimates was obtained by applying the capture-recapture method under various scenarios. Thus, the total undercoverage in the population register with respect to usual residents was estimated to lie between 88 and 185 thousand. In terms of the total number of usual residents in the Dutch population register, according to Gerritse et al. (2016) this amounts to undercoverage of 0.5% to 1.1%.

The capture-recapture method is based on strong assumptions (see, e.g., Gerritse, 2016), including independent sampling and perfect linkage of data sources. Research has shown that estimates of population size based on the capture-recapture method can be severely biased when some of these assumptions are violated (Brown et al., 2006; Van der Heijden et al., 2012; Baffour et al., 2013; Gerritse, 2016). There is on-going research into generalizations of the capture-recapture method and alternative methods that require less strong assumptions; see, e.g., Lawless (2014, Chapter 17), Ding and Fienberg (1994 and 1996), Di Consiglio and Tuoto (2015) and Zhang (2015). In addition to undercoverage, overcoverage may also be a problem: data sources may contain units that do not belong to the target population and/or multiple copies of the same units. Overcoverage may require de-duplication processes, for example via data linkage algorithms or by developing model-based estimates to remove overcoverage probabilistically.

3.6 Aggregated data only

Situation 6 (see Figure 10) is the macro data counterpart of Situation 4: in Situation 6 only aggregated data overlap with each other and need to be reconciled.

Figure 10. Combining macro-data sources



An example of Situation 6 is provided by the National Accounts, where aggregated data from many different sources need to be reconciled with each other subject to both equality and inequality constraints.

To reconcile aggregated data macro-integration can be applied. When macro-integration is applied, only estimated figures on an aggregated level are adjusted. The underlying micro data are not adjusted or even considered in this adjustment process. The main goal of macro-integration is to obtain a more accurate, univalent and complete set of estimates for the variables of interest. Univalent estimates means that estimates for the same phenomenon have the same value in different tables. The starting point of macro-integration is a set of estimates in tabular form. When one wants to use macro-integration to reconcile the data, it is important that (an approximation or indication of) the variance of each entry in the tables to be reconciled is available or can be computed. The entries of the tables are adjusted by means of a macro-integration technique so all differences between tables are reconciled and the entries with the highest variance are adjusted the most.

In the macro-integration approach often a constrained optimization problem is constructed. A target function, for instance a quadratic form of differences between the original and the adjusted values, is minimized, subject to the constraints that the adjusted common figures in different tables are equal to each other and additivity of the adjusted tables is maintained. Inequality constraints can be imposed on these quadratic optimization problems. The resulting constrained optimization problems can be exceedingly large. Fortunately, modern solvers for mathematical optimization problems are capable of handling large problems.

In the literature also Bayesian macro-integration methods have been proposed.

Several methods for macro-integration have been developed, see, for instance, Stone, Champernowne and Meade (1942), Denton (1971), Byron (1978), Sefton and Weale (1995), Magnus, Van Tongeren and De Vos (2000), Boonstra, De Blois and Linders (2011), Mushkudiani, Daalmans and Pannekoek (2012 and 2015) and Daalmans (2015).

Mushkudiani, Daalmans and Pannekoek (2015) have examined the use of macro-integration based on solving a constrained optimization problem for the reconciliation of labour market statistics. They consider a simple example of two different sources. In this example there is a labour force population of 60,000 persons, and a monthly labour force survey of 6,000 persons is conducted to estimate the unemployment rate. The auxiliary variables in the population register are “Age” and “Gender”. In the survey two variables are observed: whether a person has a job and, if not, whether she/he is registered in the Public Employment Service (PES) register.

Mushkudiani, Daalmans and Pannekoek (2015) consider results of the survey for three months: January, February and March. From these figures they estimate the number of unemployed persons in each group of the population by multiplying the survey figures by 10.

We denote the population figures by x_{ijt} $(t = 1,2,3; i = 1,2,3,4; j = 1,2)$, where t stands for the month, and i and j denote the cells of the matrix “Age”×”Gender”, see Table 2. In parenthesis are the numbers of persons registered in the PES according to the survey, denoted by y_{ijt}. There are fewer persons registered at the PES than unemployed persons, because some unemployed people do not register at the PES, for instance because they are not eligible for social unemployment benefits.

Table 2. Weighted unemployment data x_{ijt} (y_{ijt})

Age	January		February		March	
	Female	Male	Female	Male	Female	Male
20-29	350 (250)	340 (270)	360 (250)	350 (290)	370 (330)	330 (300)
30-39	400 (380)	350 (320)	420 (370)	360 (320)	420 (350)	370 (350)
40-49	600 (500)	560 (500)	580 (510)	560 (490)	610 (580)	580 (550)
≥ 50	420 (300)	380 (250)	420 (310)	400 (310)	430 (350)	400 (380)

From the PES register data we can derive the number of persons that were registered as unemployed labour force at the end of each quarter. We denote these by R_{ijk}, where i and j again denote the cells of the matrix "Age"×"Gender" and k denotes the quarter, see Table 3.

Table 3. PES register data at the end of the first quarter R_{ijk}

Age	Gender	
	Female	Male
20-29	350	330
30-39	390	360
40-49	600	570
≥ 50	370	395

In the ideal case the values of $y_{ij\text{March}}$ and $R_{ij\text{March}}$ should be the same. This is, however, not the case. For example in March there were 330 women of age 20 – 29 registered at the PES according to the survey, and 350 according to the PES register data.

In this example we consider the PES register data to be highly reliable. These figures, i.e. the R_{ijk}, will be kept fixed in the reconciliation process. In the reconciliation process we find estimates $\hat{x}_{ijt}$ and $\hat{y}_{ijt}$ for x_{ijt} and y_{ijt}, respectively. These estimates are found by solving an optimization problem subject to the following constraints (for more detail on this optimization problem, see Mushkudiani, Daalmans and Pannekoek, 2015)

– $\hat{y}_{ij\text{March}}$ should be equal to $R_{ij\text{March}}$
– The monthly changes of the series x_{ijt} and y_{ijt} are preserved as well as possible
– The ratios $x_{ijt}\ /y_{ijt}$ are preserved as well as possible.

The results of the reconciliation process are given in Table 4.

Table 4. Reconciled unemployment data $\hat{x}_{ijt}$ $(\hat{y}_{ijt})$

Age	January		February		March	
	Female	Male	Female	Male	Female	Male
20-29	375.1 (268.0)	375.2 (298.3)	385.0 (268.2)	393.7 (319.0)	393.7 (350.0)	363.8 (330.0)
30-39	445.2 (422.0)	360.8 (329.9)	466.3 (410.9)	467.1 (329.8)	467.1 (390.0)	380.7 (360.0)
40-49	622.4 (519.0)	581.9 (519.5)	602.0 (529.4)	631.5 (509.5)	631.5 (600.0)	601.5 (570.0)
≥ 50	446.0 (318.8)	398.3 (262.5)	445.7 (329.2)	419.2 (323.7)	455.2 (370.0)	416.7 (395.0)

Note that the reconciled survey estimates for the number of persons registered at the PES, i.e. the numbers in parentheses in Table 4, in March are indeed equal to the number according to the PES register (see Table 3).

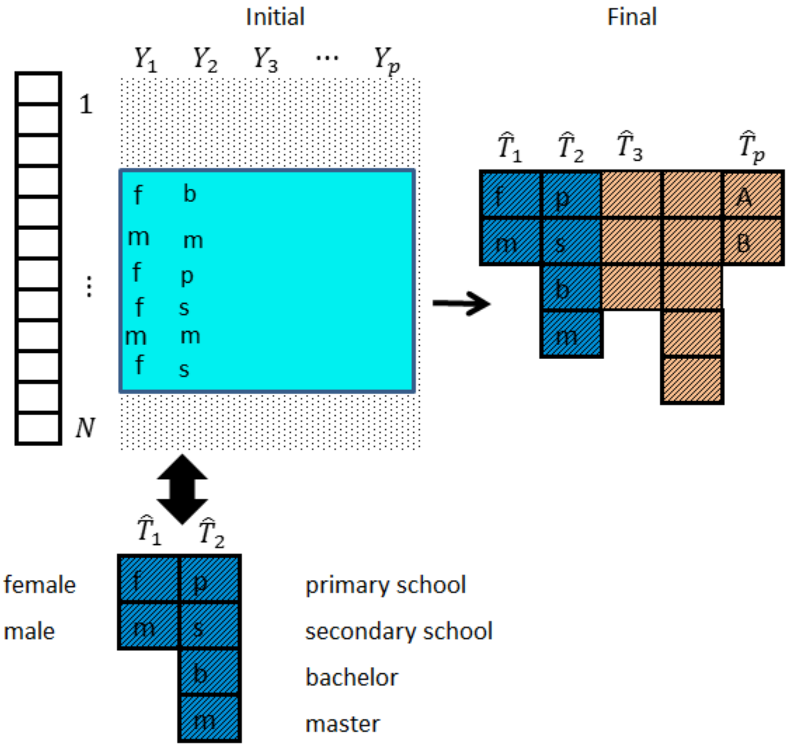
When the data have to satisfy equality constraints only, simple variance formulas are available for the macro-integration approach. However, when the data also have to satisfy inequality constraints, variance formulas become difficult to derive.

Approximations can be found in Knottnerus (2016) and in Boonstra, De Blois and Linders (2011).

3.7 Micro data and aggregated data

Situation 7 (see Figure 11) is characterized by a variation of Situation 4, by which aggregated data are available besides micro data. There is still overlap between the sources, from which there arises the need to reconcile the statistics at some aggregated level. Of particular interest is when the aggregated data are estimates themselves. Otherwise, the conciliation can be achieved by means of calibration which is a standard approach in survey sampling (see, e.g., Särndal, Swensson and Wretman, 1992).

Figure 11. Combining a micro-data source with a macro-data source



An example of Situation 7 is the Dutch virtual census, which is based on a number of administrative data sources and surveys. Population totals, either known from an administrative data source or previously estimated, are imposed as benchmarks provided they overlap with an additional survey dataset that is needed to produce new output statistics.

When micro data and aggregated data have to be estimated and reconciled, several methods are available, such as repeated weighting, repeated imputation, mass imputation and macro-integration. De Waal (2016), on which part of this section is based, discusses the pros and cons of these methods.

When repeated weighting is used, a separate set of weights is assigned to sample units for each table of population totals to be estimated. In this approach population tables are estimated sequentially and each table is estimated using as many sample units as possible in order to keep the sample variance as low as possible. The combined data from the data sources are divided into rectangular data blocks. Such a block consists of a maximal set of variables for which data on the same units has been collected. The data blocks are chosen such that each table to be estimated is covered by at least one data block.

Data from a data source covering the entire population can simply be counted. Data only available from surveys are weighted by means of regression weighting. In that case starting weights need to be assigned to all units in the block to be weighted. For a survey one usually starts with the inverse inclusion probabilities of the sample units, corrected for response selectivity. For a data block containing the overlap of two independent surveys, one usually begins with the product of the standard survey weights from each of the surveys as starting weight for each observed unit.

When estimating a new table, all cell values and margins of this table that are known or have already been estimated for previous tables are kept fixed to these known or previously estimated values. This is achieved by using regression weighting, where the starting weights are adjusted by calibrating to known or previously estimated values. This ensures univalency (see Situation 6) of the cell values and margins of the new table and previous estimates.

Repeated imputation is similar to repeated weighting. The main difference is that imputation instead of weighting is used to produce estimates. Repeated imputation is again a sequential approach where tables are estimated one by one. For some variables in a table estimates may have already been produced while estimating a previous table. These variables are then calibrated to the previously estimated totals. In order to apply repeated imputation a calibrated imputation method should be used, that is an imputation method that preserves known or previously estimated totals. Such calibrated imputation methods have been developed by Chambers and Ren (2004), Zhang (2008), Zhang and Nordbotten (2008), Pannekoek, Shlomo and De Waal (2013), Coutinho, De Waal and Shlomo (2013), Kim et al. (2014), Da Silva and Zhang (2014) and De Waal, Coutinho and Shlomo (forthcoming).

When mass imputation is used, one imputes all fields in the combined data set from all the data sources for which no value was observed. This is done even for units that were intentionally not observed, for instance for units that were not included in a sample survey. Mass imputation leads to a rectangular data set with values for all variables and all units. After imputation, estimates for population totals can be obtained by simply counting or summing the values of the corresponding variables. Mass imputation has, for instance, been studied by Whitridge, Bureau and Kovar (1990), Whitridge and Kovar (1990) and Shlomo, De Waal and Pannekoek (2009). The approach relies on the ability to capture all relevant variables and relevant relations between them in the imputation model, and to estimate the model parameters sufficiently accurately. As described in De Waal (2016) capturing all relevant variables and relevant relations between them in the imputation model is often a fundamental problem, which hampers the use of mass imputation.

Macro-integration has already been described for Situation 6 in Section 3.6, and can be applied in Situation 7 too, by first transforming the micro data to aggregated data themselves.

Table 5. Estimated numbers of males/females by level of education and "working hours" (unit = 1,000), obtained by a standard weighting approach

Gender×Working hours	Male			Female		
Level of education	**Part-time**	**Full-time**	**Total male**	**Part-time**	**Full-time**	**Total female**
Primary or less	92.2	277.4	369.6	173.3	52.1	225.4
Lower secondary general	103.5	144.0	247.5	193.3	86.6	278.9
Lower secondary vocational	87.7	478.0	565.7	213.1	61.7	274.8
Upper secondary general	74.0	149.7	223.7	154.8	76.7	231.5
Upper secondary vocational	166.4	1,236.7	1,403.1	672.6	363.0	1,035.6
Vocational college	123.9	482.4	606.3	305.4	175.5	480.9
University or more	64.4	267.0	331.3	85.6	86.5	172.1
Total	**712.1**	**3,035.1**	**3,747.2**	**1,798.1**	**901.1**	**2,699.2**

Houbiers (2004) gives an example of repeated weighting applied to (fictitious) data from a Structure of Earnings Survey. In this example the frequency table "Gender"×"Working hours"×"Education" is estimated, both without and with repeated weighting. There are three important data blocks: data block 1 with information on "Gender"×"Working hours"×"Education" with approximately 50,000 observations obtained from the overlap of two surveys, data block 2 with information on Gender×"Working hours" for about half of the Dutch population[1], and data block 3 with information on "Level of education"×"Gender" with approximately 100,000 observations obtained from a survey (data block 3).

In Table 5 the estimated number of males/females is given by "Educational level" (seven categories) of the person having a certain job and the hours worked per week in that job (< 35 hours corresponds to a part-time job, ≥ 35 hours corresponds to a full-time job).This frequency table is estimated from data block 1.

To obtain the estimates in Table 5, the block weights were raised to inflate figures from the data block to the population. These block weights were calibrated on "Gender". This implies that the estimated total numbers of jobs occupied by males and females are exactly equal to the register totals, i.e. 3,747.2 thousand for the males and 2,699.2 thousand for the females.

The standard errors of the estimates in Table 5 (not reported here) can be estimated by Taylor linearization of the regression estimator (see Särndal, Swensson and Wretman 1992).

[1] The Dutch population consists of about 17 million persons.

In Houbiers' example the margin "Gender" ×"Working hours" , i.e. the last row in Table 5, can be estimated much more accurately from data block 2. Likewise, the margin "Gender" ×"Education", i.e. total male/female columns in Table 5, i.e. can be estimated more accurately from data block 3. The resulting estimates for these margins are shown in Tables 6 and 7.

Table 6. More accurately estimated numbers of males/females by "working hours" (unit = 1,000)

Gender×Working hours	Male			Female		
	Part-time	Full-time	Total male	Part-time	Full-time	Total female
Total	701.1	3,046.2	3,747.2	1,827.9	871.3	2,699.2

Table 7. More accurately estimated numbers of males/females by level of education (unit = 1,000)

Level of education×Gender	Total male	Total female
Primary or less	357.2	209.8
Lower secondary general	267.3	290.5
Lower secondary vocational	595.8	320.6
Upper secondary general	213.2	205.4
Upper secondary vocational	1,374.0	1,006.9
Vocational college	616.0	500.1
University or more	323.8	166.0
Total	**3,747.2**	**2,699.2**

Note that, except for the total numbers of males and females, the estimates in Tables 6 and 7 differ clearly from the corresponding margins in Table 5.
Univalency of the target table and the accurately estimated margins in Tables 6 and 7 can be obtained by applying repeated weighting. The result is shown in Table 8.
The estimated standard errors of the repeated weighting estimates in Table 8 (not shown here) are smaller than for the corresponding estimates in Table 5. This illustrates that repeated weighting not only achieves univalent results, but can also lead to more accurate estimates. For more details on this example we refer to Houbiers (2004).

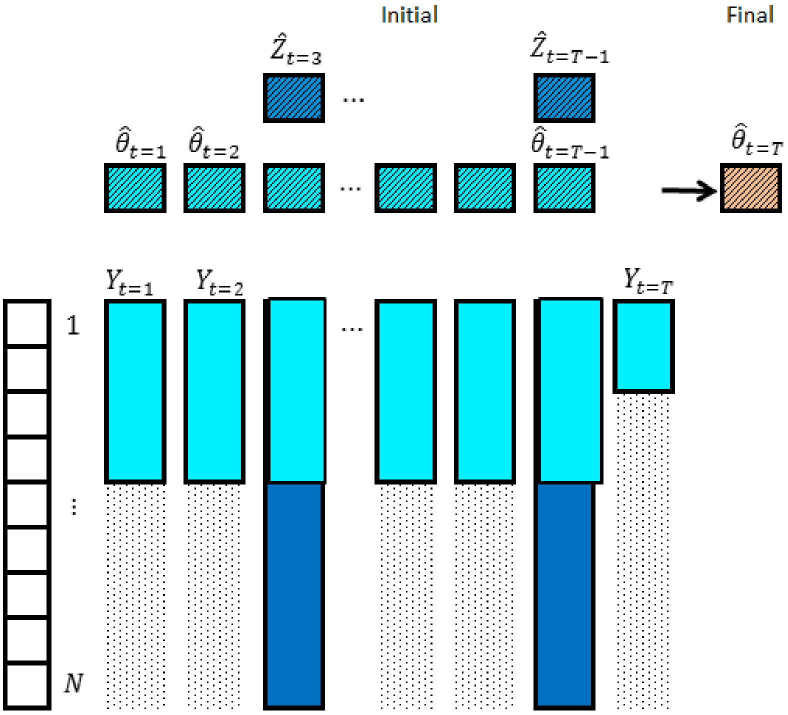
Table 8. Estimated numbers of males/females by level of education and "working hours" (unit = 1,000), obtained by repeated weighting

Gender×Working hours	**Male**			**Female**		
Level of education	**Part-time**	**Full-time**	**Total male**	**Part-time**	**Full-time**	**Total female**
Primary or less	87.3	269.9	357.2	163.0	46.8	209.8
Lower secondary general	110.2	157.1	267.3	203.9	86.6	290.5
Lower secondary vocational	90.4	505.3	595.8	250.6	70.0	320.6
Upper secondary general	69.2	143.9	213.2	139.6	65.8	205.4
Upper secondary vocational	159.0	1,215.0	1,374.0	664.0	342.8	1,006.9
Vocational college	123.3	492.7	616.0	322.4	177.7	500.1
University or more	61.5	262.2	323.8	84.4	81.6	166.0
Total	**701.1**	**3,046.2**	**3,747.2**	**1,827.9**	**871.3**	**2,699.2**

3.8 Longitudinal data

Finally, longitudinal data are introduced in Situation 8. We limit ourselves to the issue of reconciling time series of different frequencies and qualities, as illustrated in Figure 12. When time series of a high frequency (e.g. monthly) are benchmarked on those of a lower frequency (e.g. quarterly or annual), a specific set of macro-integration methods can be applied. The difference with the macro-integration in Situation 6 is that the data are now related to each other over time. The data of the low frequency series are usually considered to be exogenous and are kept fixed, since these are usually based on the most comprehensive information.
When high and low frequency series of target variables are observed, reconciliation of these series is known as benchmarking. A related problem is that of disaggregation: a series of low frequency of a target variable is disaggregated by using an indicator series of high-frequency for the target variable.

Figure 12. Combining longitudinal data sources



Situation 8 is for instance found at SN where monthly turnover based on a sample survey of enterprises is used to compute turnover indices for the short-term statistics. These indices are computed for a number of publication cells. An example of the time series of the publication cell “Manufacture of cutlery, tools and general hardware”, is given in Figure 13 from August 2010 till December 2011. These sample survey data (labelled as “source” in Figure 13) are benchmarked against quarterly turnover values. Those quarterly turnover values are largely based on VAT data supplemented by survey data, which was explained already in the example for Situation 2. These quarterly data are kept fixed, since they cover nearly the complete population. The horizontal lines in Figure 13 represent the average monthly index values per quarter of the source and the benchmark data. The differences between those two are very large in the first and second quarter of 2011. In fact those differences are too large to be reconciled by automatic benchmarking in practice, but the example nicely illustrates differences between benchmarking methods.

A wide range of methods is available for benchmarking and temporal disaggregation. Perhaps the most basic method is to preserve the original levels with pro-rating. Pro-rating means that the level estimates are adjusted with the same relative factor. Another method to preserve the original levels is that by Chow and Lin (1971). It expresses the estimation of the high-frequency values as a linear regression on the low-frequency values and finds the solution by generalized least squares.

Figure 13. Index of monthly turnover: source data and three benchmarked series: “Prorating”, “MP” (Movement Preservation) and “MP (weighted)” (see text). Month 1 = January 2010.

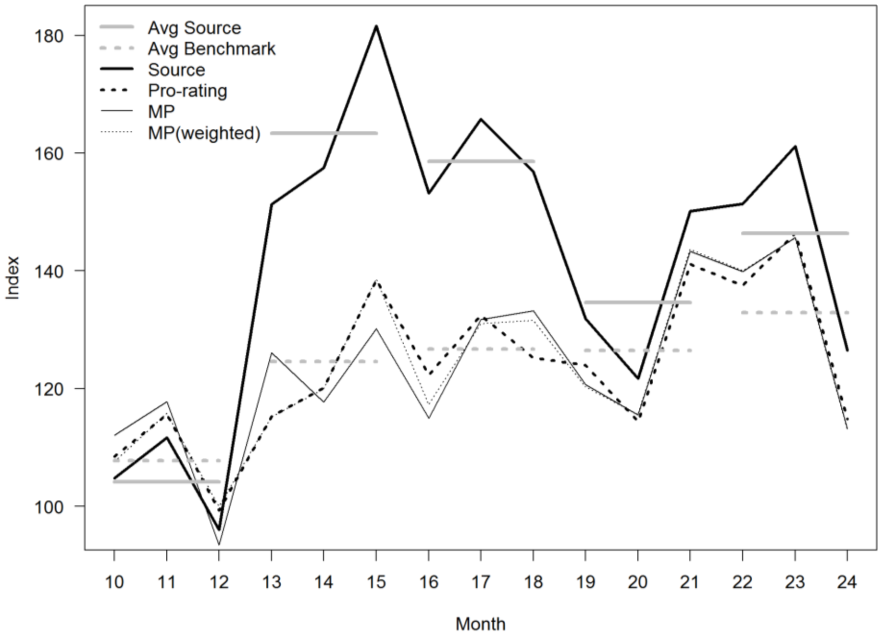


Table 9. Monthly growth rates of the series in Figure 13.

Growth rates	13	14	15	16	17	18	19	20	21	22	23	24
Source	57.5	4.1	15.3	-15.6	8.2	-5.4	-15.9	-7.8	23.4	0.9	6.4	-21.5
Pro-rating	16.1	4.1	15.3	-11.6	8.2	-5.4	-1.0	-7.8	23.4	-2.5	6.4	-21.5
MP	35.0	-6.6	10.6	-11.7	14.6	1.1	-9.4	-4.3	24.0	-2.3	4.1	-22.3
MP (weighted)	15.2	4.2	15.5	-15.4	11.7	0.5	-8.6	-3.9	24.2	-2.5	4.0	-22.3

A disadvantage of pro-rating and of the Chow-Lin method is that it leads to the so-called step problem: when observing reconciliation adjustments of the changes between two successive high-frequency periods, disproportionally large adjustments may be observed in the transition from one low-frequency period to the next. For instance, in the turnover example, the monthly growth rate in January 2011 was 57.5 % in the source data and after applying pro-rating it was adjusted to 16.1 % due to the step-problem (Table 9). A similarly large adjustment can be seen in the growth rate of July 2011.

An alternative to level preservation is to preserve the changes in the original high-frequency series. One such method is the movement preservation method (MP) by Denton (1971), slightly modified by Cholette (1984). The movement preservation method minimizes the squared differences between adjusted and original first-order differences over the entire period of the series (Bikker, Daalmans and Mushkudiani, 2011). Therefore, the value of an adjusted monthly rate of change is not only determined by the corresponding quarters, but also by previous and next quarters. This way, a large shift in monthly changes just before and after the end of a quarter is avoided. In the turnover example, the monthly growth rate in January 2011 for the

series benchmarked by the MP was 35 % which is closer to the growth rate of the source than was the case after benchmarking with pro-rating. Also the growth rate adjustment in July 2011 was smaller after applying MP than after pro-rating.
The MP method can be refined by applying weights to the adjustments of the benchmark series. These weights should reflect the accuracy of level estimates and the accuracy of estimated growth rates of the high-level frequency series that is to be adjusted. When growth rates are estimated accurately, adjustments to these growth rates should be small. When growth rates are not measured accurately, adjustments to these growth rates may be larger. One may argue that in this latter case there is not really a step problem from a statistical point of view, but only from a 'cosmetic' point of view. Recall that there are large benchmarking differences in the first quarter of 2011 (Figure 13). For instance, the monthly growth rate of February 2011 was 4.1 % in the source data and -6.6 % after applying MP. This was considered to be unrealistic by the production staff at SN. We adjusted the weight such that MP may adjust the figures more in the first quarter of 2011. Table 9 (MP, weighted) shows that the benchmarked series are now closer to the growth rates of the source data, but the disadvantage is that the step-problem in January 2011 has now returned.
Benchmarking can also be applied to multiple time-related variables. The problem now is to deal with time constraints as well as with constraints between variables (Bikker, Daalmans and Mushkudiani, 2011). Di Fonzo and Marini (2003 and 2005) and Bikker and Buijtenhek (2006) combined the Denton method for time constraints with the method of Stone, Champernowne and Meade (1942) for handling constraints between the variables. Bikker, Daalmans and Mushkudiani (2013) extended the method to include other modelling features, like so-called soft constraints (i.e. constraints that have to be satisfied only approximately), ratio constraints and inequality constraints. Their method is used in several production processes of National Accounts in the Netherlands.
Another method to preserve the changes in the original high-frequency series is the so-called Growth Rate Preservation method (GRP; Causey and Trager, 1981), which is often recommended In the literature on longitudinal data reconciliation (e.g., Bloem, Dippelsman and Mæhle, 2001). Daalmans et al. (2016) compared Denton's method to GRP and concluded that GRP has some important practical disadvantages for official statistics.

4. Discussion

We are fully aware that the basic situations we have considered in this paper do not offer a complete description of all situations that may arise in practice and that our basic situations give a simplified view of reality. At the same time, we do feel that this paper offers useful guidelines to producers of multi-source statistics. Many situations arising in practice are variations of the basic situations that we have discussed in this paper, or combinations of such basic situations. In the discussion of the basic situations we have pinpointed important problems that can occur for these situations. This will allow producers of multi-source statistics to anticipate the problems that may occur for their specific situation. In the discussion of the basic situations we also described and gave references to important methods that can be used to overcome the problems. Hopefully, this will give the producers of multi-source statistics a flying start to overcome the problems for their own specific case. Many of the methods referred to in this paper have only recently been developed. These methods are therefore still in their infancy and will hopefully be improved upon in many different aspects in the coming years.

For Situations 1–5, two additional complications need to be mentioned. The first of these occurs when we have a subset of the variables in one source (say administrative data) and other variables in a second source (say sample survey data) and the sources contain overlapping units, but the reference periods of the two sources are different. For many variables, values differ for different reference periods. This hampers the harmonization of data sources.

A second complication occurs when data are available at different levels of aggregation. For instance, we may want to combine data on bankruptcies (available at the level of legal persons) with data on the number of jobs of employees. The latter are available at the level of enterprises, where an enterprise may be a combination of legal persons

The latter complication is related to so-called unit errors. Unit errors occur when units are defined differently in one source than in another source, when the units in available data sources are not defined according to the official definition that one wants to use at the NSI, or when units have to be constructed. For instance, “enterprises” may be defined differently between different data sources, the definition of an “enterprise” in a data source may differ from the official definition used at the NSI, and the NSI may need to construct households from available data on individual persons. In this paper we have not discussed unit errors. We refer to Zhang (2011, 2012) for more information.

Another topic we have not discussed in this paper is confidentiality. Confidentiality issues already occur for single-source statistics. This problem becomes aggravated for multi-source statistics, simply because more information becomes available. We refer to Willenborg and De Waal (2001) and Hundepool et al. (2012) for more on confidentiality issues and methods to prevent disclosure of confidential information in general.

Finally, we remark that after combining data sources, one is usually interested in estimating the accuracy of the outcomes. Different quality measures and methods to compute them for various situations are currently under development for this purpose in the ESSnet on Quality of Multi-source Statistics, which is partly funded by the EU (De Waal, Van Delden and Scholtus, 2017). In this ESSnet quality measures and methods to compute for all situations described in this paper with the exception of Situation 3 are examined and tested.

We hope that our discussion of various situations, and the above-mentioned issues we have not discussed in this paper, will inspire other researchers to do research on the highly important and highly interesting area of producing multi-source statistics.

Acknowledgement

The authors thank Harm Jan Boonstra for his useful comments on an earlier version of this paper.

References

Alajääskö. P. and A. Roodhuijzen (2016). Micro data Linking in Business Statistics. Accessed at http://ec.europa.eu/eurostat/statistics explained/index.php?title=Micro_data_linking_in_business_statistics_-_introduction&redirect=no

Baffour, B., J.J. Brown and P.W.F. Smith (2013), An Investigation of Triple System Estimators in Censuses, Statistical Journal of the International Association for Official Statistics, 29, pp. 53-68.

Bakker, B.F.M. (2011), Micro-Integration: State of the Art. Chapter 5 in: State of the Art on Statistical Methodologies for Data Integration. Report on WP1 of the ESS net on Data Integration.

Bakker, B.F.M. (2012), Estimating the Validity of Administrative Variables. Statistica Neerlandica, 66, 8-17.

Bakker, B.F.M. and P. Daas (2012), Some Methodological Issues of Register Based Research, Statistica Neerlandica, 66, pp. 2-7.

Biemer, P.P. (2011), Latent Class Analysis of Survey Error (Hoboken, New Jersey: John Wiley and Sons).

Bikker, R.P. and S. Buijtenhek (2006), Alignment of Quarterly Sector Accounts to Annual data. Statistics Netherlands, Voorburg, http://www.cbs.nl/NR/rdonlyres/D918B487-45C7-4C3C-ACD0-oE1C86E6CAFA/0/Benchmarking_QSA.pdf.

Bikker, R., J. Daalmans and N. Mushkudiani (2011), Macro-integration. Data reconciliation. Statistical Methods (201104), Statistics Netherlands.

Bikker, R., J. Daalmans and N. Mushkudiani (2013), Benchmarking Large Accounting Frameworks: a Generalised Multivariate Model. Economic Systems Research 25, pp. 390-408.

Bishop, Y., S. Fienberg and P. Holland (1975), Discrete multivariate analysis, theory and practice (New York: McGraw-Hill).

Bloem, A., R. Dippelsman, and N. Mæhle, (2001), Quarterly National Accounts Manual: Concepts, Data Sources, and Compilation. International Monetary Fund, Washington, D.C.

Boeschoten, L., D. Oberski and T. de Waal (forthcoming), Latent Class Multiple Imputation for Multiple Observed Variables in a Combined Dataset. To be published in the Journal of Official Statistics.

Bollen, K.A. (1989), Structural Equations with Latent Variables. New York: John Wiley and Sons.

Boonstra, H.J. (2004), Calibration of Tables of Estimates. Report, Statistics Netherlands.

Boonstra, H.J., C.J. de Blois and G.J. Linders (2011), Macro-Integration with Inequality Constraints an Application to the Integration of Transport and Trade Statistics. Statistica Neerlandica 65, 407-431.

Bound, J., C. Brown and N. Mathiowetz (2001), Measurement Error in Survey Data. In: Heckman and Leamer (eds.), Handbook of Econometrics, Volume 5, pp. 3705-3843, Elsevier, Amsterdam.

Bozik, J.E. and M.C. Otto (1988), Benchmarking: Evaluating methods that preserve month-to-month changes. Bureau of the Census - Statistical Research Division, RR-88/07, URL: http://www.census.gov/srd/papers/pdf/rr88-07.pdf.

Brown, J.J., O. Abott and I.D. Diamond (2006), Dependence in the 2001 One-Number Census Project. Journal of the Royal Statistical Society. Series A (Statistics in Society) 169, pp. 883-902.

Brown, J., I. Diamond, R. Chambers, L. Buckner and A. Teague (1999), A Methodological Strategy for a One-Number Census in the UK. Journal of the Royal Statistical Society. Series A (Statistics in Society) 162, pp. 247–267.

Byron, R.P. (1978), The Estimation of Large Social Account Matrices. Journal of the Royal Statistical Society A 141, pp. 359-367.

Causey, B. and M.L. Trager (1981), Derivation of Solution to the Benchmarking Problem: Trend Revision. Unpublished research notes, U.S. Census Bureau, Washington D.C. Available as an appendix in Bozik and Otto (1988).

Chambers, R. (2009), Regression Analysis of Probability-Linked Data. Official Statistics Research Series 4, Statistics New Zealand (available at http://www.statisphere.govt.nz/further-resources-and-info/official-statistics-research/series/volume-4-2009.aspx#2)

Chambers, R. L. and R. Ren (2004), Outlier Robust Imputation of Survey Data. In: ASA Proceedings of the Joint Statistical Meetings, American Statistical Association, pp. 3336-3344.

Chen, C., M. J. Page and J. M. Stewart (2016), Creating New and Improved Business Statistics by Maximising the Use of Administrative Data. Paper presented at the Fifth International Conference on Established Surveys, Geneva, Switzerland.

Cholette, P. (1984), Adjusting Sub-Annual Series to Yearly Benchmarks. Survey Methodology 10, pp. 35-49.

Chow, G.C. and A. Lin (1971), Best Linear Unbiased Interpolation, and Extrapolation of Time Series by Related Series. Rev. Economics and Statistics 53, pp. 372-375.

Conti, P.L., D. Marella and A. Neri (2015), Statistical Matching and Uncertainty Analysis in Combining Household Income and Expenditure Data. Temi di Discussione (Working papers) 1018, Banca d'Italia.

Coutinho, W., T. de Waal and N. Shlomo (2013), Calibrated Hot Deck Imputation Subject to Edit Restrictions. Journal of Official Statistics 29, pp. 299-321.

Daalmans, J. (2015), Estimating Detailed Frequency Tables from Registers and Sample Surveys. Discussion paper, Statistics Netherlands.

Daalmans J., di Fonzo, T., Mushkudiani, N. and R.P. Bikker (2016), Removing the Gap between Annual and Sub- Annual Statistics based on Different Data Sources. Proceedings of the Fifth International Conference on Establishment Surveys, June 20-23, 2016, Geneva, Switzerland.

Da Silva, D.N. and L.C. Zhang (2014), Adjustments for Survey Imputed Datasets to Achieve First and Second–order Properties. Joint Statistical Meeting.

Den Ouden, M. (2016), The Use of Distance Hot Deck Statistical Matching for Structural Business Statistics. Master thesis, Utrecht University.

Denton, F.T. (1971), Adjustment of Monthly or Quarterly Series to Annual Totals: An Approach Based on Quadratic Minimization. Journal of the American Statistical Association 66, pp. 99-102.

De Waal, T. (2016), Obtaining Numerically Consistent Estimates from a Mix of Administrative Data and Surveys. Statistical Journal of the IAOS 32, pp. 231–243.

De Waal, T., W. Coutinho and N. Shlomo (forthcoming), Calibrated Hot Deck Imputation for Numerical Data under Edit Restrictions. Accepted for publication in the Journal of Survey Statistics and Methodology.

De Waal, T., A. Van Delden and S. Scholtus (2017), Output Quality of Multi-source Statistics. Paper presented at the NTTS conference 2017, Brussels.

Di Consiglio, L. and T. Tuoto (2015), Coverage Evaluation on Probabilistically Linked Data, Journal of Official Statistics 31, pp. 415–429.

Di Fonzo, T. and M. Marini (2003), Benchmarking Systems of Seasonally Adjusted Time Series According to Denton's Moving Preservation Principle. University of Padova, http://www.oecd.org/dataoecd/59/19/21778574.pdf.

Di Fonzo, T. and M. Marini (2005), Benchmarking a System of Time Series: Denton's Movement Preservation Principle vs. Data Based Procedure. University of Padova, http://epp.eurostat.cec.eu.int/cache/ITY_PUBLIC/KSDT-05-008/EN/KS-DT-05-008-EN.pdf.

Di Zio, M. and O. Luzi (2014), Theme: Editing Administrative Data. In: Memobust Handbook on Methodology for Modern Business Statistics, Eurostat, Luxembourg.

Ding, Y. and S.E. Fienberg (1994), Dual System Estimation of Census Undercount in the Presence of Matching Error, Survey Methodology 20, pp. 149–158.

Ding, Y. and S.E. Fienberg (1996), Multiple Sample Estimation of Population and Census Undercount in the Presence of Matching Errors, Survey Methodology 22, pp. 55–64.

D'Orazio, M., M. Di Zio and M. Scanu (2006), Statistical Matching: Theory and Practice. John Wiley and Sons, Chichester, UK.

ECSM (2014), Energy Compilers Statistics Manual Chapter 4b. Data Sources and Data Collection. Available at https://unstats.un.org/oslogroup/methodology/docs/escm-edited/ESCM%20Chapter%204b%20140423.docx.

Fellegi, I.P. and A.B. Sunter (1969), A Theory for Record Linkage. Journal of the American Statistical Association 64, pp. 1183-1210.

Fienberg, S. (1972), The Multiple Recapture Census for Closed Populations and Incomplete 2k Contingency Tables. Biometrika 59, pp. 409-439.

Gerritse, S.C. (2016), An Application of Population Size Estimation to Official Statistics. PhD Thesis, Utrecht University.

Gerritse, S.C., B.F.M. Bakker, P.-P. de Wolf and P.G.M. van der Heijden (2016), Undercoverage of the Population Register in the Netherlands, 2010. Published as Chapter 5 in Gerritse (2016).

Guarnera, U. and R. Varriale (2015), Estimation and Editing for Data from Different Sources. An Approach Based on Latent Class Model. Working Paper No. 32, UN/ECE Work Session on Statistical Data Editing, Budapest.

Guarnera, U. and R. Varriale (2016), Estimation from Contaminated Multi-Source Data based on Latent Class Models. Statistical Journal of the IAOS 32, pp. 537-544.
Hagenaars, J.A. and A.L. McCutcheon (eds.) (2002), Applied Latent Class Analysis, New York: Cambridge University Press.
Herzog, T.N., Scheuren, F.J. and W.E. Winkler (2007), Data Quality and Record Linkage Techniques. Springer.
Houbiers, M. (2004), Towards a Social Statistical Database and Unified Estimates at Statistics Netherlands. Journal of Official Statistics 20, pp. 55-75.
Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Schulte Nordholt, K. Spicer and P.P. de Wolf (2012), Statistical Disclosure Control. New York: John Wiley and Sons.
IWGDMF (International Working Group for Disease Monitoring and Forecasting) (1995), Capture- Recapture and Multiple Record Systems Estimation. Part 1. History and Theoretical Development. American Journal of Epidemiology 142, pp. 1059-1068.
Kim, J.K., E. Berg and T. Park (2016), Statistical Matching using Fractional Imputation. Survey Methodology 42, pp. 19-40.
Kim, H.J., J.P. Reiter, Q. Wang, L.H. Cox and A.F. Karr (2014), Multiple Imputation of Missing or Faulty Values under Linear Constraints. Journal of Business and Economic Statistics 32, pp. 375-386.
Knottnerus, P. (2016), On New Variance Approximations for Linear Models with Inequality Constraints. Statistica Neerlandica 70, pp. 26-46.
Lawless, F. (2014), Statistics in Action. A Canadian Outlook. Ontario: Apple Academic Press Inc.
Linder, F., D. van Roon and B.F.M. Bakker (2011), Combining Data from Administrative Sources and Sample Surveys; The Single Variable Case. In: ESSnet Data Integration, WP4 Case Studies, pp. 39-97, Luxembourg: Eurostat.
Magnus, J.T., J.W. van Tongeren and A.F. de Vos (2000), National Accounts Estimation using Indicator Ratios. The Review of Income and Wealth 46, pp. 329-350.
McLachlan, G.J. and D. Peel (2000), Finite Mixture Models, New York: John Wiley and Sons.
Meijer, E., S. Rohwedder and T. Wansbeek (2012), Measurement Error in Earnings Data: Using a Mixture Model Approach to Combine Survey and Register Data. Journal of Business and Economic Statistics 30, pp. 191–201.
Mushkudiani, N., J. Daalmans and J. Pannekoek (2012), Macro-Integration Techniques with Applications to Census Tables and Labour Market Statistics. Discussion paper, Statistics Netherlands.
Mushkudiani, N., J. J. Daalmans and J. Pannekoek (2015), Reconciliation of Labour Market Statistics using Macro-Integration. Statistical Journal of the IAOS 31, pp. 257–262.
Oberski, D. (2017), Estimating Error Rates in an Administrative Register and Survey Questions Using a Latent Class Model. In P. Biemer, B. West, S. Eckman, B. Edwards, and C. Tucker (Eds.), Total Survey Error. Wiley, New York.
Pannekoek, J., N. Shlomo and T. de Waal (2013), Calibrated Imputation of Numerical Data under Linear Edit Restrictions, Annals of Applied Statistics 7, pp. 1983-2006
Pavlopoulos, D. and J. K. Vermunt (2015), Measuring Temporary Employment. Do Survey or Register Data Tell the Truth? Survey Methodology 41, pp. 197–214.

Ruotsalainen, K. (2005), Combining Enterprise Data to Employment Data in Register-based Employment Statistics. Paper for the Siena Group on Social Statistics, meeting in Helsinki 2005 Session on Record-linking. Accessed at http://www.stat.fi/sienagroup2005/kaija.pdf

Saris, W.E. and F.M. Andrews (1991), Evaluation of Measurement Instruments Using a Structural Modeling Approach. In: Biemer, Groves, Lyberg, Mathiowetz and Sudman (eds.), Measurement Errors in Surveys, pp. 575-597, New York: John Wiley and Sons.

Särndal, C.E., B. Swensson and J. Wretman (1992), Model Assisted Survey Sampling. New York: Springer-Verlag, 1992.

Sefton, J. and M. Weale (1995), Reconciliation of National Income and Expenditure. Cambridge University Press, Cambridge, UK.

Scholtus, S., B.F.M. Bakker and A. Van Delden (2015), Modelling Measurement Error to Estimate Bias in Administrative and Survey Variables. Discussion paper, Statistics Netherlands.

Shlomo, N., T. de Waal and J. Pannekoek (2009), Mass Imputation for Building a Numerical Statistical Database. Neuchâtel, Switzerland: UN/ECE Work Session on Statistical Data Editing.

Si, Y. and J.P. Reiter (2013), Nonparametric Bayesian Multiple Imputation for Incomplete Categorical Variables in Large-Scale Assessment Surveys. Journal of Educational and Behavioral Statistics 38, pp. 499-521

Stone, R., D.G. Champernowne and J.E. Meade (1942), The Precision of National Income Estimates. Review of Economic Studies 9, pp. 111-125.

UN/ECE (2014), Measuring Population and Housing. Practices of UNECE Countries in the 2010 Round of Censuses. New York and Geneva: United Nations.

Van Delden, A. and P.P. de Wolf (2013), A Production System for Quarterly Turnover Levels and Growth Rates Based on VAT Data. In Proceedings of the Conferences on New Techniques and Technologies for Statistics, March 5–7, Brussels. Available at http://www.cros-portal.eu/sites/default/files//NTTS2013%20Proceedings_0.pdf (accessed December 2013).

Van Delden, A., J. Pannekoek, R. Banning and A. de Boer (2016), Analysing Correspondence between Administrative and Survey Data, Statistical Journal of the IAOS 32, pp. 569-584.

Van der Heijden, P.G.M., J. Whittaker, M. Cruyff, B. Bakker and R. van der Vliet (2012), People Born in the Middle East but Residing in the Netherlands: Invariant Population Size Estimates and the Role of Active and Passive Covariates, Annals of Applied Statistics 6, pp. 831-852.

Whitridge, P., M. Bureau and J. Kovar (1990), Mass Imputation at Statistics Canada. In: Proceedings of the Annual Research Conference, U.S. Census Bureau, Washington D.C., pp. 666-675.

Whitridge, P. and J. Kovar (1990), Use of Mass Imputation to Estimate for Subsample Variables. In: Proceedings of the Business and Economic Statistics Section, American Statistical Association, pp. 132-137.

Willenborg, L. and T. de Waal (2001), Elements of Statistical Disclosure Control. New York: Springer-Verlag.

Wolter, K.M. (1986), Some Coverage Error Models for Census Data, Journal of the American Statistical Association 81, pp. 338-346

Zhang, L.C. (2008), A Triple-Goal Imputation Method for Statistical Registers. Neuchâtel: UN/ECE Work Session on Statistical Data Editing.

Zhang, L.C. (2011), A Unit-Error Theory for Register-Based Household, Journal of Official Statistics 27, pp. 415-432.

Zhang, L.C. (2012), Topics of Statistical Theory for Register-Based Statistics and Data Integration. Statistica Neerlandica 66, pp. 41-63.

Zhang, L.C. (2014), Data Integration. Survey Statistician 70, pp. 15-24.

Zhang, L.C. (2015), On modelling register coverage errors, Journal of Official Statistics 31, pp. 381-396

Zhang, L.C. and S. Nordbotten (2008), Prediction and Imputation in ISEE: Tools for More Efficient Use of Combined Data Sources. Vienna: UN/ECE Work Session on Statistical Data Editing.

Explanation of symbols

Empty cell	Figure not applicable
.	Figure is unknown, insufficiently reliable or confidential
*	Provisional figure
**	Revised provisional figure
2014–2015	2014 to 2015 inclusive
2014/2015	Average for 2014 to 2015 inclusive
2014/’15	Crop year, financial year, school year, etc., beginning in 2014 and ending in 2015
2012/’13–2014/’15	Crop year, financial year, etc., 2012/’13 to 2014/’15 inclusive

Due to rounding, some totals may not correspond to the sum of the separate figures.

Colofon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress
Statistics Netherlands, Studio BCO

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contactform: www.cbsl.nl/information

© Statistics Netherlands, The Hague/Heerlen/Bonaire 2015.
Reproduction is permitted, provided Statistics Netherlands is quoted as the source.