Statistics Netherlands

Discussion Paper

General Approaches for Consistent Estimation based on Administrative Data and Surveys

2015 | 11

Ton de Waal

September

Content

Abstract

National statistical institutes (NSIs) fulfill an important role as providers of objective and undisputed statistical information on many different aspects of society. To this end NSIs try to construct data sets that are rich in information content and that can be used to estimate a large variety of population figures. At the same time NSIs aim to construct these rich data sets as efficiently and cost effectively as possible. This can be achieved by utilizing already available administrative data as much as possible, and supplementing these administrative data with survey data collected by the NSI. In this paper we will focus on obtaining consistent population estimates from such a combination of administrative data sets and surveys. We will sketch general approaches based on weighting, imputation and macro-integration, and discuss their advantages and drawbacks.

1. Introduction

National statistical institutes (NSIs) fulfil an important role as providers of objective and undisputed statistical information on many different aspects of society. To this end NSIs try to construct data sets that are rich in information content and that can be used to estimate a large variety of population figures. At the same time NSIs aim to construct these rich data sets as efficiently and cost effectively as possible.
This can be achieved by utilizing already available administrative data as much as possible, and supplementing these administrative data with survey data collected by the NSI. These survey data themselves are often collected by a mix of modes, such as web surveys, paper questionnaires, face-to-face interviews and telephone interviews in order to ensure data quality and reduce data collection costs.
Utilizing available administrative data obviously holds many opportunities for NSIs, simply because these data do not have to be collected again, which saves NSIs a lot of data collection and processing costs, without having to place extra response burden on respondents. Moreover, these administrative data sets often contain information that NSIs are unable to collect themselves, such as wages for all individuals in certain subpopulations or turnover of all enterprises in a certain branch of industry. Collecting such information on so many units in the population is generally impossible for an NSI. Administrative data may also contain data on variables that would otherwise be unavailable for NSIs, such as detailed and precise information on the medical records of individual persons.
Unfortunately, administrative data do not only offer opportunities for NSIs, they also present challenges. One of these challenges is to process available administrative data so they can be used for statistical purposes. This often involves correcting for differences in definitions between variables in the administrative data and target variables (see Section 2).
In this paper we will mainly focus on another challenge: obtaining consistent population estimates from a combination of administrative data sets and surveys. Here, and in the rest of the paper, the terms “consistent” and “consistency” refer to numerical consistency between estimates from different tables, not to “consistency” in the usual meaning in mathematical statistics. There are several general approaches for achieving numerical consistency. In this paper we will sketch general approaches based on weighting, imputation and macro-integration, and discuss their advantages and drawbacks. All approaches we will discuss, except model-based macro-integration (see Section 6.1), can handle both categorical and numerical data.
In the late 1990s Statistics Netherlands decided to develop what was then called the Social Statistical Database; it is nowadays referred to the System of social statistical data sets (SSD). This SSD should eventually contain all data on persons and households available in administrative data sources or collected by Statistics Netherlands itself. In order to produce statistical figures from such an SSD new methodology had to be developed.
Options discussed then were roughly the same as discussed in the current paper, although the technical aspects of the approaches have substantially improved since then, opening up new opportunities and possibilities for utilizing these approaches.

An exception is CERISE (Consistent Estimation using Repeated Imputation Satisfying Edits), see Section 5.2, for which the basic ingredients, i.e. powerful imputation algorithms, were not available at the time. These powerful imputation algorithms, which can satisfy edits rules and preserve known or previously estimated totals, have only been developed in recent years. CERISE is a new option for combining administrative data and surveys. With respect to macro-integration only rudimentary versions based on linear programming were known at the time at Statistics Netherlands. Since the 1990s macro-integration techniques have been developed a lot further at Statistics Netherlands, and our understanding of such techniques has improved substantially.
This paper is organized as follows. Section 2 discusses what we are trying to measure and what we actually observe, and ways to reconcile the differences between those. Section 3 describes the problem of obtaining consistent estimates from a combination of administrative data and survey data in some detail. Sections 4 to 6 sketch the general approaches for combining data we consider in this paper. Methods based on weighting are discussed in Section 4, methods based on imputation in Section 5, and methods based on macro-integration in Section 6. Section 7 gives an overview of the pros and cons of the most promising approaches. In a sense Table 4 in Section 7, which summarizes these pros and cons, may be seen as the main product of this paper. Finally, Section 8 concludes the paper with a brief discussion.

2. The data and the truth

NSIs publish statistical information about certain phenomena, such as the number of unemployed people and the total revenue in a certain branch of industry. The first step towards publishing such statistical information is designing a conceptual definition of the phenomenon to be estimated. There are several aspects to be considered when defining a certain phenomenon. First of all, one has to define the phenomenon itself. For example, in the case of unemployed people one has to define when one considers a person as unemployed. Does one consider someone who is working for only 8 hours per week as employed even though he wants to work more hours per week? Does one consider someone who is not working and does not want to work as unemployed? How about people who are not able to work: are they unemployed? Note that these questions do no arise while processing the data, but only while establishing the definition. While processing the data, Statistics Netherlands, just like other NSIs, uses an official definition of employment.
For many variables one also has to define the moment or period for which they are to be measured. Persons may, for instance, be requested to answer whether they were married on a particular date of specify their gross income over a certain year.
Finally, one also has to define the population one wants to measure the phenomenon in. In the case of the Netherlands, does one aim to count all persons living in the Netherlands when measuring the number of unemployed people, or only the Dutch

citizens? Does one count all people, or only people with a certain minimum and/or maximum age?

Combined, the conceptual definition of the variable, the moment or period of measurement and the conceptual definition of the population describe the conceptual phenomenon one is trying to measure. The value of this conceptual phenomenon is considered to be the true value for this phenomenon.

Once one has defined the phenomenon to be measured on a conceptual level, the next step is to operationalize the conceptual phenomenon. For example, one could design a questionnaire to measure the number of unemployed persons. In our example, one then has to think about how one is going to ask whether someone is unemployed according to the conceptual definition. One also has to operationalize the population. For example, if one decides to send interviewers to potential respondents, some persons in the population may not be reached, such as homeless people or people who no longer live at the registered address. Or, if one decides to use available administrative data to estimate the number of unemployed people, the units in the administrative data source may differ from the conceptual population.

During data collection measurement errors may be made. When a questionnaire is used, the respondents may make a mistake, they may give an incorrect answer on purpose, or mistakes may be made at the NSI while processing the data. Likewise, administrative data may also contain mistakes, for instance because they are outdated. Also, some units one wanted to observe may be missing in an administrative data source.

No matter how one collects the data, it is very likely that the observed data – either survey data, administrative data, or from a Big Data source (see Section 3.2 for some information on Big Data in general) – differ from the conceptual, i.e. true, values of the phenomena one aimed to measure.

At NSIs the observed data are processed in an attempt to reconstruct the true value as well as possible. One tries to correct for differences in definition between the conceptual definition and the operational definition (see Bakker 2011). One also tries to correct for differences between the conceptual population and the operational population.

As a next step one usually tries to correct for measurement errors (see De Waal, Pannekoek and Scholtus, 2011). This can be done by utilizing logical relations in the data, for example that someone who claims to have no job cannot have a wage from a job at the same time. Another way of correcting for measurement errors is by comparing observed values from a certain unit to values of similar units. Large deviations may indicate that data of the unit under consideration are erroneous.

After cleaning the observed data for differences due to definitions and measurement errors, some residuals errors due to remaining differences in definition or measurement errors one was not able to correct for will be left in the data. That is, even after the cleaning process, the observed values will generally differ from the true value.

In this paper we will consider the truth, i.e. the variable that gives the true value of the conceptual phenomenon, as a latent variable. The observed data – either survey data, administrative data, or from a Big Data source – provide information about the true value.

To obtain an estimate for the true value of a unit for a certain phenomenon we model the true value using the available cleaned observed data, from survey(s), administrative data or Big Data sources as auxiliary data.
The models used in practice are often quite simple. If a single data source is available and conceptual differences and measurement errors have been corrected for as well as possible, one generally has no other option than to consider the corrected observed value as a – hopefully good – approximation of the true value.
If more data sources are available, more opportunities for estimation of the true value arise. Often one determines which of the data sources is generally best for a certain subpopulation, and then uses the corrected values in that data source as approximation of the true value for that subpopulation. For numerical data, one could in some cases also use the average value of the data observed in the various data sources as approximation of the true value. One can also use more complex models, for instance based on latent class analysis. Related models in the context of combining data sources that do not provide microdata but only data on an aggregated level have been developed and studied by Bakker (2012), Pavlopoulos and Vermunt (2013), Scholtus and Bakker (2013) and Scholtus (2014).
Modeling the true values as a function of observed values in several data sources – either by simple or more complex techniques – can be seen as a method for reconciling the differences between the observed values and find the best estimates for the true values.
In this paper we assume that for each variable a model for estimating the true value of the phenomenon using the available corrected observed data has been developed. Whenever we refer to the value of a variable in the remainder of this paper, we will mean the (estimated) true value of the phenomenon under consideration.
Zhang (2012) provides a more extensive discussion of the potential differences between the true, i.e. the conceptual target value, and the observed value.

3. The consistent estimation problem

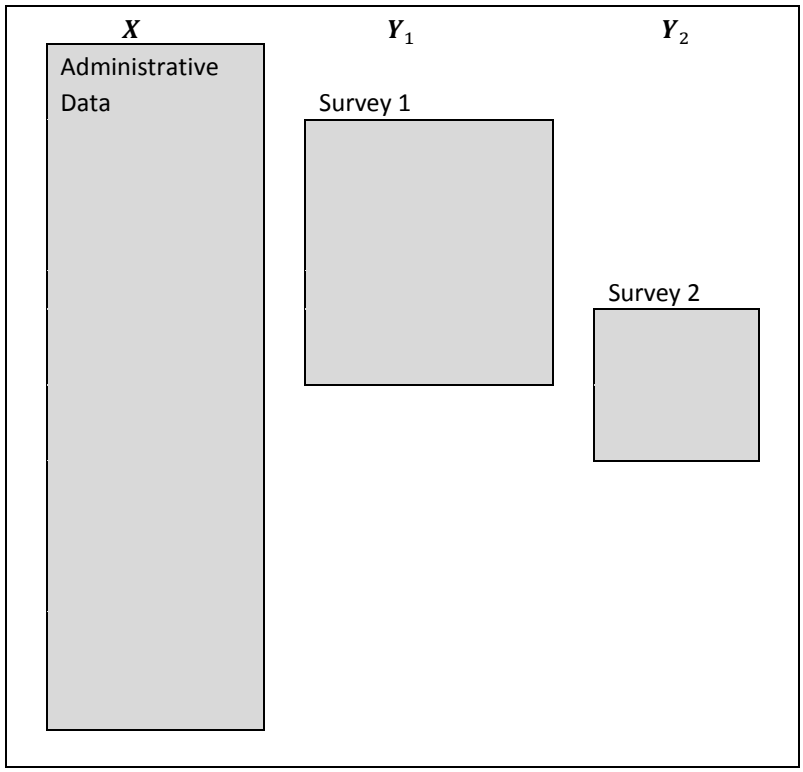
3.1 The problem

Different estimates for the same phenomenon could lead to confusion among users of these figures. Many other NSIs, such as Statistics Netherlands, have therefore adopted a one-figure policy. According to this one-figure policy, estimates for the same phenomenon in different tables should be equal to each other, even if these estimates are based on different underlying data sources.
When using a mix of administrative data sources and surveys to base estimates upon, the one-figure policy becomes problematic as for different (combinations of) variables data on different units, e.g. different persons, may be available. This means that different estimates concerning the same variable may yield different results, if one does not take special precautions. In principle, these differences are merely

caused by “noise” in the data, such as sampling errors. So, in a strictly statistical sense, different estimates concerning the same variables are to be expected and are not a problem. However, different estimates would violate the one-figure policy and form a problem from this point of view.

We illustrate the basic problem by means of a simple example with one administrative data source and two surveys. We assume all variables in these three data sources measure different phenomena. We denote the variables in the administrative data source by $\boldsymbol{X}$, the variables in Survey 1 by $\boldsymbol{Y}_1$, and the variables in Survey 2 by $\boldsymbol{Y}_2$ (see Figure 1). For example, $\boldsymbol{X}$ in the administrative data source may contain information on the place of residence of persons, $\boldsymbol{Y}_1$ in Survey 1 information on their occupation, and $\boldsymbol{Y}_2$ in Survey 2 on their education. In such a case we may be interesting in the (estimated) number of persons with a certain occupation, education and place of residence and combinations hereof.

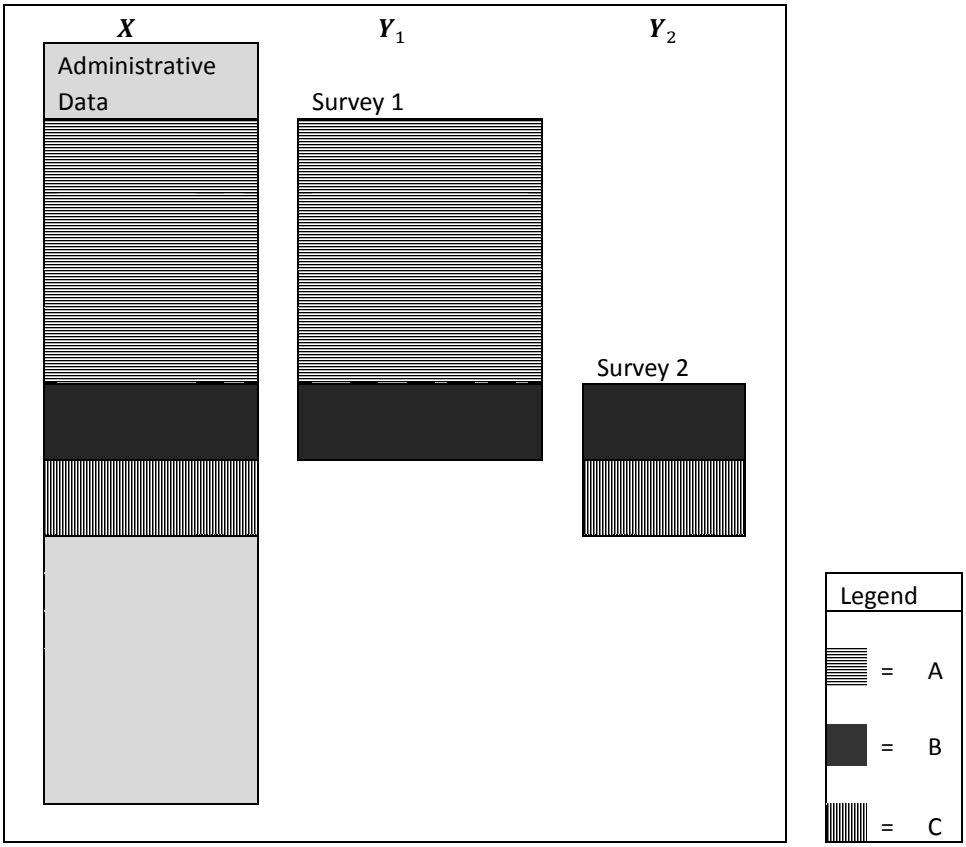
3.1.1 Figure 1: Illustration of our basic problem, part 1



A basic assumption we make throughout this paper, and have indeed already made implicitly, is that the units in the various data sources to be combined can be linked to each other. We assume that the administrative data source contains values for all units in the target population, whereas Surveys 1 and 2 only contain values for samples of this population. We are interested in estimating, on the population level, $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ in relation to $\boldsymbol{X}$, and the joint relation of $\boldsymbol{Y}_1$, $\boldsymbol{Y}_2$ and $\boldsymbol{X}$. For example, if $\boldsymbol{X}$, $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ are categorical variables, we want to estimate the cross-tables $\boldsymbol{X} \times \boldsymbol{Y}_1$, $\boldsymbol{X} \times \boldsymbol{Y}_2$, and $\boldsymbol{X} \times \boldsymbol{Y}_1 \times \boldsymbol{Y}_2$ for the population. If $\boldsymbol{X}$ and $\boldsymbol{Y}_1$ are categorical data and $\boldsymbol{Y}_2$ is a numerical variable, we may want to estimate $\boldsymbol{X} \times \boldsymbol{Y}_1$ for the population,

population totals of $\boldsymbol{Y}_2$ for groups defined by $\boldsymbol{X}$, and population totals of $\boldsymbol{Y}_2$ for groups defined by the cross-tabulation of $\boldsymbol{X} \times \boldsymbol{Y}_1$.
In this paper we assume that some units occur in both Survey 1 and Survey 2. Variables in different data sources may either be (essentially) the same or may be different. Whatever the situation, we assume that the observed data are first used to estimate the true values of the conceptual target variables (see Section 2).
To estimate these quantities, it makes sense to use as much relevant data as possible. That is, in order to estimate $\boldsymbol{Y}_1$ in relation to $\boldsymbol{X}$, one would like to use the data of the units of parts A and B of the administrative data source and Survey 1 (see Figure 2). In order to estimate $\boldsymbol{Y}_2$ in relation to $\boldsymbol{X}$, one would like to use parts B and C of the administrative data source and Survey 2. Finally, in order to estimate the joint relation of $\boldsymbol{Y}_1, \boldsymbol{Y}_2$ and $\boldsymbol{X}$ one would like to use part B of the administrative data source, Survey 1 and Survey 2.
Now, we are ready to explain the basic problem. For simplicity, we will assume that all three data sources each contain only one categorical variable. Using parts A and B of the administrative data source and Survey 1 to estimate $\boldsymbol{X} \times \boldsymbol{Y}_1$ leads to population estimates for the variables in $\boldsymbol{Y}_1$ and for the variables in $\boldsymbol{X}$ as $\boldsymbol{X}$ and $\boldsymbol{Y}_1$ are marginals of this table. However, population figures for the administrative variables $\boldsymbol{X}$ are already known: one can simply count the values in the administrative data source. The basic problem is that the estimates from parts A and B of the administrative data source and Survey 1 for $\boldsymbol{X}$ will generally differ from the already known figures for $\boldsymbol{X}$. Similarly, using parts B and C of the administrative data source and Survey 2 to estimate $\boldsymbol{X} \times \boldsymbol{Y}_1$ leads to population estimates for $\boldsymbol{X}$ that will generally differ from the already known figures for $\boldsymbol{X}$.
Moreover, using part B of the administrative data source, Survey 1 and Survey 2 to estimate $\boldsymbol{X} \times \boldsymbol{Y}_1 \times \boldsymbol{Y}_2$ will generally lead to different population estimates for $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ than using parts A and B of the administrative data source and Survey 1 to estimate $\boldsymbol{Y}_1$ and using parts B and C of the administrative data source and Survey 2 to estimate $\boldsymbol{Y}_2$.
Without taking special “precautions” one will end up with different estimates for $\boldsymbol{X}$, $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$, depending on the units one bases the estimates upon. Our aim is to develop methods for taking such precautions and ensure consistency between estimates.
We have described the problem for one administrative data source and two surveys. The problem obviously can also occur in the general situation with m administrative data source and n surveys.
In this paper we mainly focus on the situation where we aim for consistency between estimated (or known) population totals. This includes estimated population means, as these are just estimated population totals divided by the number of population units, and, in principle, also estimated correlations, as these are estimated sums of products.

3.1.2 Figure 2: Illustration of our basic problem, part 2



Example

We give a simple numerical example based on Houbiers (2004) to illustrate the problem. In this example we want to estimate the number of jobs against some background variables: gender and educational level (seven categories) of the person having a certain job, and the hours worked per week in that job ($\leq$ 35 hours corresponds to part-time,> 35 hours corresponds to full-time). We assume we have an administrative data source containing information on gender, a Survey 1 containing information on working hours and a Survey 2 containing information on education (see Figure 2).

We use as many records as possible to estimate the table "gender $\times$ working hours", i.e. we use parts A and B of the administrative data source and Survey 1 in Figure 2. We use the calibration (or GREG) estimator to calibrate the estimates on the number of males and females in the administrative data source (see, e.g., Bethlehem 2009, or Särndal, Swensson and Wretman 1992 for more on the calibration estimator). Note that this automatically guarantees consistency with the administrative data source in this respect. This gives us the results in Table 1. Due to minor rounding errors, the number of males working part-time and the number of males working full-time do not sum up precisely to the total number of males in Table 1.

3.1.3 Table 1. Estimates of "gender × working hours"

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

We also estimate the table "gender × education", using as many records as possible, i.e. using parts B and C of the administrative data source and Survey 2. Again, we use the calibration estimator to calibrate the estimates on the number of males and females in the administrative data source. This gives us the results in Table 2.

3.1.4 Table 2. Estimates of "gender × level of education"

Gender × level of education	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

Finally, we estimate the table "gender × working hours × education", using as many records as possible and using the calibration estimator to calibrate on the number of males and females in the administrative data source. We obtain the results given in Table 3.

3.1.5 Table 3. Estimates of "gender × working hours × level of education"

Gender × working hours × level of education	Male, Part-time	Male, Full-time	Total male	Female, Part-time	Female, Full-time	Total female
Primary or less	92.2	227.4	369.6	173.3	52.1	225.4
Lower secondary general	103.5	144.0	247.5	193.3	85.6	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	76.6	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	287.0	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

Tables 1 to 3 illustrate the consistent estimation problem. Consider, for instance, the number of males that work part-time. According to Table 1 that number is estimated as 701.1, whereas to the more detailed Table 3 that number is estimated as 712.1, however. Similarly, according to Table 2 the number of males with primary education or less is estimated as 357.2, whereas according to Table 3 that number is estimated as 369.6. Estimates for the same phenomenon hence differ in different tables.

The general approaches we consider in this paper offer different solutions for the consistency problem illustrated above. In later sections we will consider (repeated) weighting where differences are reconciled by means of a weighting scheme for the available records, (repeated) imputation where differences are reconciled by means of an imputation scheme for the unobserved data, and macro-integration where differences between tables are reconciled on an aggregated level.

As a side note, in rare cases inconsistency problems can also arise if one has only one data source. An example is when one has a data source with information on the education level of people and their children of 16 years or older (if any). If one then wants to estimate the education level of parents with children of 16 years or older, one can either use the records of these parents or the records of their children. These parents and children will generally have different weights, and estimates based on these weights will lead to different results. We will not examine this kind of inconsistency problems in a single data source. In some cases one can treat the data

source as two different sets – a data set for the parents and a separate data set for the children of 16 year or older – and then apply one of the approaches considered in this paper. In many cases better approaches are likely to be available, however.

3.2 Some aspects of the problem

When combining several data sources many different situations can occur. This is true even if one wants to combine only two data sources on the same kind of units (say, persons), each having only one variable. Even in such a simple case there are four different situations: (i) the data sources refer to the same population and have (essentially) the same variable, (ii) the data sources refer to the same population but have different variables, (iii) the data sources refer to different populations but have (essentially) the same variable, and (iv) the data sources refer to different populations and have different variables. Here we say that "a data source refers to a population" when either this population is integrally observed or a sample from this population is drawn.

In this paper we examine the case that all data sources refer to the same population (above situations (i) and (ii)). This means that there are no under-coverage problems. To correct for under-coverage in situations (iii) and (iv) one can use capture/recapture techniques (see, e.g., Gerritse, Van der Heijden and Bakker 2013). We assume that over-coverage can be corrected by deleting all units that do not belong to the target population. A prerequisite for this is that these units can be identified.

In this paper variables in different data sources may either be (essentially) the same or may be different. Whatever the situation, we assume that the observed data are first used to estimate the true values of the conceptual target variables (see Section 2).

When one wants to combine data sets, but the units in these data sets do not overlap or one does not have sufficient identifying information to link the units in these data sets, one could resort to statistical matching (see D'Orazio, Di Zio and Scanu 2006). Depending on whether the statistical matching is carried out on a micro-level or on a macro-level, it may be seen as a special form of mass imputation (see Section 5.1) or as a special form of (model-based) macro-integration, respectively. Since we assume that data sets do overlap and the units can be linked – as is the case in the SSD – we will not examine statistical matching in this paper.

A recent development in the world of official statistics is the data deluge due to Big Data. Big Data are often characterized by the three V's: Volume, Velocity and Variety (see, e.g., Buelens et al. 2014). "Volume" means that Big Data sources are generally larger than regular systems can handle smoothly. "Velocity" refers to the frequency at which data become available or to the short period between the occurrence of an event and Big Data on this event becoming available. "Variety" refers to the wide diversity of Big Data sources. Especially if one wants to combine Big Data sources with other sources, one may not be able to link units, not even in principle. For example, an NSI may be interested in combining Twitter data with administrative data on persons. However, units in Twitter data are tweets that are not always

linkable to persons or enterprise, simply because the NSI is generally not able to tell which person or enterprise sent a tweet.
Thus far we focused on the situation that microdata for the administrative data sources and surveys are both available. This is not always the case. In some situations one only has survey data and population figures from an external source available. For example, at Statistics Netherlands we publish figures on movements by train by Dutch citizens. The total estimated number of kilometres travelled by train is obtained from the major Dutch railway company. In a survey Statistics Netherlands asks additional questions on background characteristics such as age and gender, questions on when people travel, and so on. The total estimated number of kilometres travelled by train from the major Dutch railway company can be used to calibrate the survey data. In a similar way, unlinkable Big Data might in some cases still be used to calibrate survey data to. Conversely, estimates obtained from Big Data might be calibrated on estimates obtained from other data sources.
In other cases, one may only have estimated population figures and no microdata at all available. In such a case, one can resort to macro-integration (see Section 6) to reconcile differences between these estimated population figures.

4. Weighting-based approaches

In this section we examine approaches for combining several data sources based on weighting the data.

4.1 The traditional weighting approach

The traditional way in survey sampling to estimate population totals is by assigning a weight to each unit in the sample and then calculate the weighted total. Roughly speaking, a unit in a sample survey with weight, say 24.86, counts for 24.86 units with the same values for the target variables in the population, of whom 23.86 are not selected in the sample. Note that fractional weights are common.
To obtain weights per sample unit, one starts with the sample weight, i.e. the inverse of the probability of selecting a unit in the sample. Due to unit non-response, where some of the units that were intended to be observed for some reason did not respond, sample weights are often adjusted slightly before they are used to calculate weighted totals. Item non-response, where only some items of otherwise responding units are missing, is usually treated by imputation. Adjusting sample weights to correct for unit-nonresponse can be done in different ways, for instance by calibrating to known population totals of auxiliary variables. To derive the final survey weights, one uses a weighting model including a target variable and relevant auxiliary variables. The parameters of this model have to be estimated based on the sample at hand.

In the traditional survey sampling context, where one has a single sample survey for estimating population totals, one applies what we will call the “traditional weighting” approach. By this we mean that one constructs a single weighting model including, in principle, all relevant variables and all relevant relations between them. The weighting model aims to correct for sample selection effects and for unit non-response effects.

The approach relies on the ability to capture all relevant variables and relevant relations between them in the weighting model, and at the same time estimate the model parameters sufficiently accurately. Given that all relevant variables and relevant relations among them can be captured sufficiently accurately by the weighting model, the traditional weighting approach is rather straightforward, if one has a single survey sample available. To estimate a population total one simply multiplies the value of the variable to be estimated in each sample unit with the survey weight of that unit, and sums these products to obtain the estimate for the population total. A strong point of the traditional weighting approach in the case of a single survey is that relationships between variables are automatically maintained.

The traditional weighting approach quickly becomes problematic when one wants to combine several administrative data sets and surveys, in particular if units in these data sources partly overlap, since not all estimates will be based on the same set of units (see Figures 1 and 2). Taking into account that estimates may be based on different sets of units is exceedingly complicated in the traditional weighting approach, and no one has thus far attempted to do so. If one does not take into account that different sets of units will be used to estimate different (combinations of) variables, inconsistencies between estimates may arise as we saw in Section 3.

In practice there is a trade-off between the variables and relations between them that one wants to include in the weighting model and the accuracy of the estimated parameters of this model. The more variables and relations between them, the more generally applicable the model becomes. For variables and relations between variables that were excluded from the model, there is no guarantee that the estimated population totals are accurate However, including too many variables and relations between them in the weighting model can lead to unstable and inaccurate model parameters.

In the traditional weighting approach, after estimating and analyzing the relevant population totals, one generally retains the survey weights in the data set as this allows one to later analyze the data in more detail. For example, at first one may be interested only in analyzing the univariate results and perhaps a few correlations, whereas later one may be interested in many more correlations. Keeping the survey weights in the data enables one to later analyze these correlations.

Owing to the complexity when wanting to combine several data sources and the impossibility to capture all relevant variables and relations in a single weighting model, the traditional weighting approach cannot really be used when one want to combine administrative data and survey data. In the remainder of this paper we will therefore not consider the traditional weighting approach anymore. We have sketched the approach here, because traditional weighting is the most often used and best understood way to obtain estimates for a sample survey at NSIs. In that sense it is a kind of stepping stone to other approaches. By contrasting other

approaches to the traditional weighting approach one better understands the essential aspects of these approaches.
Preston (2014) examines a traditional weighting approach for a combination of administrative data and survey data. His focus is on correcting for missing data due to linkage problems, however, rather than on obtaining consistent population estimates as in the current paper.

4.2 Repeated weighting

As a way to overcome the problems of the traditional weighting approach, the so-called "repeated weighting" (RW) approach has been developed at Statistics Netherlands in the late 1990s (see, e.g., Kroese and Renssen 2000, Houbiers et al 2003, and Houbiers 2004). In the RW approach a separate set of weights is assigned to sample units for each table of population totals to be estimated.
In the RW approach the tables to be estimated 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 administrative data sources and surveys are divided into rectangular blocks. Such a block consists of a maximal set of variables for which data on the same units has been collected (see Figure 2). The data blocks are chosen such that each table to be estimated is covered by at least one data block. Item non-response in a block is assumed to be treated beforehand by means of imputation.
How a table is estimated depends on the available data. Data from an available administrative data source covering the entire population can simply be counted. Data only available from surveys are weighted by means of regression weighting (see Särndal, Swensson and Wretman 1992). In that case weights must 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, just as in the traditional weighting approach. These weights are then further adjusted by calibrating them to previously estimated totals. For a data block containing the overlap of two surveys, one usually begins with the product of the standard survey weights from each of the surveys as starting weight for each observed unit, and then corrects these starting weights by calibrating to totals known from administrative data sources and previously estimated totals.
When estimating a new table, all margins of this table that are known or have already been estimated for previous tables are kept fixed to their known or previously estimated values, i.e. the regression weighting is calibrated on these known or previously estimated values. This ensures that the margins of the new table are consistent with previous estimates.
We will use Figures 1 and 2 to explain the basic elements of RW. In the RW approach low-dimensional tables are in principle weighted before higher-dimensional ones. So, the RW approach would, for instance, start with estimating the cross-table $\boldsymbol{X} \times \boldsymbol{Y}_1$. To estimate this table weighting is used, where one calibrates the estimates for the marginal $\boldsymbol{X}$ on the already known values from the administrative data source. Similarly, the cross-table $\boldsymbol{X} \times \boldsymbol{Y}_2$ is estimated by calibrating on the known values for $\boldsymbol{X}$. Next, the cross-table $\boldsymbol{X} \times \boldsymbol{Y}_1 \times \boldsymbol{Y}_2$ is estimated by calibrating on the estimated

cross-tables $\boldsymbol{X} \times \boldsymbol{Y}_1$ and $\boldsymbol{X} \times \boldsymbol{Y}_2$. In this way previously estimated values are preserved when making new estimates for a new table.
The RW approach was developed with the estimation of a set of related frequency census tables in mind. In principle, the same estimation strategy can also be used for tables containing quantitative variables, such as "income". Houbiers et al. (2003) provide more details on the RW approach.
Note that after estimating a table of population totals and analyzing this table, one can delete the weights that have been used to generate this table as these weights will not be valid for generating and analyzing other tables. When one wants to estimate and analyze other tables, one generates a new set of survey weights for that particular table. The only reason for retaining the survey weights in the data set would be to document which weights have been used to generate this table, but that could instead be done by documenting the procedure for calculating the weights.
The RW approach may seem simple, but is not without complications as noted by Houbiers et al. (2003) and Daalmans (2014). We briefly discuss some of these complications.
One complication is the occurrence of empty cells in high-dimensional tables, i.e. cells without any observations (the empty or zero cell problem). Empty cells lead to population estimates with value zero. "Strange", i.e. either very large or very small, weights may have to be given to other cells in order to preserve known or previously estimated totals when there are many empty cells. In some cases it may not even be possible to find suitable weights at all. This happens when a cross-tabulation has to be calibrated on some estimated marginal total, but the data from which the table must be estimated does not contain any units corresponding to this marginal. Attempts to solve the empty cell problem may lead to discrepancies between the microdata and the estimated tables, and to estimated combinations of categories that cannot occur in the population. One of the attempts to solve the empty cell problem is the so-called epsilon method. This method assumes that each cell is populated by at least one unit. If no unit is present in the data sources, a non-zero "ghost value" will be used as an initial estimate.
Another complication is that, although the approach takes known or previously estimated totals into account, it does not take more subtle edit rules into account (the edit rule problem). An example is that the number of people with a driver's license should be less than or equal to the number of people with the minimum age or older to obtain a driver's license. The former figure may be estimated based on a sample survey, whereas the latter may be derived directly from an administrative data source covering the entire population. After application of the RW approach the estimate for the number of people with a driver's license may be higher than the estimate for the number of people with the minimum age or older to obtain a driver's license (see also Kroese and Renssen 2000 and Van de Laar 2004).
In principle, the RW approach could be modified to include such more subtle edit relations (see Van der Laar 2004 and Daalmans 2014). For instance, when a variable involved in edits occurs in a table to be estimated, Daalmans (2014) extends the table by adding all variables involved in those edits. Estimating such extended tables is obviously much more demanding than estimating the original tables. It remains to be examined to what extent extending tables to be estimated in this way is a solution to the problem of satisfying edit relations. Like the traditional weighting approach, the

RW approach does ensure automatically that relationships between data items from a single data source are maintained.
Further complications of the RW approach are that for large, detailed tables computation can become problematic (computational problems), that after a number of tables have been estimated conflicting marginal totals can occur so that it becomes impossible to estimate a new table with all required marginal totals (the problem of conflicting totals), and that the results of RW depend on the order in which the tables are estimated (the order dependency problem). In particular, values that are estimated later in the process will on average deviate more from estimates based on the original data sources than values that are estimated at the beginning of the process as they have to comply with more constraints.
Houbiers et al. (2003) note that RW is not suitable when estimates on several, non-hierarchical, subpopulations have to be produced.
RW has been developed only for cross-sectional data. An extension to longitudinal data seems very hard, if not impossible, to develop.
Like the other techniques in this paper, RW is mainly applied for cosmetic purposes, namely to ensure consistency between estimated tables. However, calibrating to totals based on large sample sizes generally leads to a reduction of the sample variance for tables based on smaller sample sizes (except for cells with hardly any observations). The same reduction of sample variance when calibrating to totals based on large sample sizes also occurs for CERISE and macro-integration which we discuss later.

4.3 Measuring the quality

Under certain assumptions, for instance that selectivity of samples can be corrected for by means of an appropriate weighting model, weighting-based estimators are generally chosen so that they are (approximately) design-based unbiased. In other words, bias is negligible for such estimators. To measure the quality of weighting-based estimators one therefore focuses on estimating the variance of population estimates.
For the traditional weighting approach, for many sampling designs and weighting-based estimators variance formulas for population estimates are readily available. We refer to refer to the literature on sampling theory, such as Cochran (1977), Särndal, Swensson and Wretman (1992) and Knottnerus (2003), for more on variance formulas.
Variance formulas have also been derived for the calibration or regression estimator for the case where the control totals are estimated themselves (see Renssen and Nieuwenbroek 1997 and Berger, Muñoz and Eancourt 2009) as well as for RW (see Knottnerus 2001, Houbiers et al. 2003, Van Duin and Snijders 2003, and Knottnerus and Van Duin 2006), but only when the data do not have to satisfy inequality restrictions.
For very complex sampling designs or weighting-based estimators variance formulas are hard to derive. In such cases one can resort to resampling methods, such as the jackknife, bootstrap and balanced repeated replication (see Wolter 1985). Preston (2014) notes that in many situations replication methods are preferred to variance

formulas as the former can be applied for basically all estimators, whereas variance formulas have to be derived for each estimator separately.
For the situation where part of the data on a certain variable are obtained from an administrative data source and the rest is estimated using a survey, Kuijvenhoven and Scholtus (2011) have applied bootstrap methods to measure the accuracy of certain weighting-based estimators.

5. Imputation-based approaches

In this section we discuss approaches for combining administrative data and surveys based on imputation techniques.
Imputation is a much more “precise” instrument than weighting, where we use “precise” in a non-statistical manner. If we use the metaphor of a surgical operation, we could compare weighting to surgery with traditional medical instruments, such as a scalpel, and imputation to surgery with nanotechnology. For instance, suppose we have data on a sample of 20,000 persons from the Dutch population (about 16.8 million persons). In the traditional weighting approach one would have 20,000 survey weights that can be used for calibration and, later, estimation purposes. In the mass imputation one would have about 16.8 million (the number of persons in the Dutch population minus the 20,000 observed respondents) values for each variable that can be used for calibration, and later, estimation purposes. A danger of using mass imputation is than one is trying to achieve more than is possible.
Like surgery with traditional instruments and surgery with nanotechnology, weighting and imputation are both useful techniques. Sometimes the precision of nanotechnology is not needed, and would only cost too much time and/or budget. Likewise, sometimes imputation is not necessary for the results one wants to obtain.

5.1 Mass imputation

In the mass imputation approach, one imputes all variables for which no value was observed for all population units (see Whitridge, Bureau and Kovar 1990, Whitridge and Kovar 1990, Shlomo, De Waal and Pannekoek 2009). This leads to a rectangular data set with values for all variables and all population units. After imputation estimates of totals can be obtained by simply counting or adding the values of the corresponding variables.
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. Given that all relevant variables and relevant relations among them can be captured accurately by the imputation model, the approach is very straightforward.

In 1997 Kooiman, the then head of the Methodology Department of Statistics Netherlands, wrote an influential internal report that has become part of the Statistics Netherlands' collective memory. This paper has had a major impact on people's perception of (mass) imputation, from the time the report was written up till now. Some of the examples have become part of Statistics Netherlands's "folklore".

Kooiman's aim with the paper was essentially "to tame the beast", where a mass imputed data set was "the beast". He had an issue that has to be taken very seriously, namely: what can happen with a fully imputed data set when the analyses that will be carried out on the imputed data set are not known beforehand? Apart from potential statistical issues with a fully imputed data set, his principle objection to such a data set is that it may be used for purposes for which it was never intended, and, moreover, that it is hard to tell from the imputed data set itself that one is using it for unintended purposes.

Kooiman's best known example is combining the amount of money spent per month on dog food, (which may be known from a Budget Survey), with whether or not people have a dog as pet (which may be known from a Survey on Living Conditions). Including these variables, the amount of money spent per month on dog food and whether one has a dog as pet or not, in an imputation model is, except in very exceptional cases, not deemed important enough. Including information on their correlation in an imputation model is even more unlikely.

In his example Kooiman indeed assumed that these variables and information on their correlation are not included in the imputation model for mass imputation. He notes that this is not a problem at all, as long as one is aware for which purposes the imputation model was designed. His issue was that one may not be aware of this. This may especially be the case if several departments of an NSI are involved. It may also be the case if the imputed data set is used to base statistical figures upon for a longer period of time instead of only once as later statisticians may have forgotten the precise original intentions of the imputation model.

In this dog food example, if one is not aware that the imputation model was not designed to capture the relation between the amount of money spent on dog food and having a dog as pet or not, one may decide to analyze and publish the relation between these variables. In this case one may come to the rather shocking – but completely unjustified – conclusion that many people in the Netherlands who do not have a dog as pet spent money on dog food, and that conversely many people who do have a dog a pet do not buy dog food. This conclusion would make an interesting headline in a newspaper, and would lead to major problems for Statistics Netherlands and its position in the Dutch society!

The problem is the use of the imputed data set for purposes it was not intended for and for which the imputation model was not designed. This (unintended) misuse of the imputed data set is the beast that needs to be tamed. Once the beast is loose, it is impossible to control it. It needs to be tamed before it can be released. In other words, (unintended) misuse of the imputed data set should be avoided as best as we can.

In principle, one could use the mass imputation approach to obtain estimates of the relevant population totals, analyze them, and then delete the imputations that lead to these results. This would "tame the beast". However, just as in the traditional

weighting approach, where one would like to retain the weights in the data set, one would like retain the imputations in the data to allow one to analyze the data set in more detail later.
In the mass imputation approach Kooiman's beast is on the loose. It is generally impossible to capture all relevant variables and relations in the imputation model, simply because there are not enough observations to estimate all model parameters accurately, which implies that many relations in the imputed data are spurious and do not reflect the relations in the population.
Kooiman (1998) concludes that mass imputation is not a viable strategy for obtaining consistent estimates from a set of administrative data sources and surveys. In particular, there are generally not enough degrees of freedom for developing a sufficiently rich imputation model that is able to account for all relevant relations between survey and administrative variables. Owing to the lack of degrees of freedom, the imputation model will have to ignore some important relations in the data (see also Kroese, Renssen and Trijssenaar 2000).
For rich data sets with many variables, especially if not all tables to be estimated are specified beforehand, we endorse Kooiman's conclusion. We do think that for data sets with a limited number of variables and where all tables to be estimated are specified beforehand mass imputation is a viable option, and perhaps even one of the best options available. In this paper we focus on the situation where not all tables to be estimated are known beforehand. Apart from a small passage in the Discussion, we will therefore not consider mass imputation anymore in the remainder of this paper.

5.2 CERISE

Whereas mass imputation is the equivalent of traditional weighting, CERISE (Consistent Estimation using Repeated Imputation Satisfying Edits) is the equivalent of repeated weighting. The important difference is how estimates are produced: in the case of RW by means of a weighting method, in the case of CERISE by means of an imputation method. Like RW, CERISE is a sequential approach where tables are estimated one by one. For some variables in a table estimates may have already been produced while estimvating a previous table. Similar to RW, these variables are then calibrated to the previously estimated totals.
In CERISE, imputation is not seen as a way to obtain a complete data set, but as an estimation technique. To emphasize that CERISE is an estimation method rather than a way to obtain complete data, we have given it the name CERISE instead of "repeated imputation". As an estimation technique CERISE avoids some of the pitfalls of RW.
We will again use Figures 1 and 2 to explain the basic ideas of CERISE. As in RW approach, low-dimensional tables are in principle estimated before higher-dimensional ones. CERISE would, for instance, start with estimating the cross-table $\boldsymbol{X} \times \boldsymbol{Y}_1$ by imputing $\boldsymbol{Y}_1$ for all population units for which the value of $\boldsymbol{Y_1}$ is missing. The imputation model is based on parts A and B of the administrative data and Survey 1 in Figure 2. Obviously, the population total of $\boldsymbol{X}$ known from the administrative data is maintained in the imputed data as all values of $\boldsymbol{X}$ are used.

Similarly, the cross-table $\boldsymbol{X} \times \boldsymbol{Y}_2$ is estimated by imputing $\boldsymbol{Y}_2$ for all population units for which the value of $\boldsymbol{Y_2}$ is missing. The imputation model is based on parts B and C of the administrative data and Survey 2 in Figure 2. Again the known population total of $\boldsymbol{X}$ data is obviously maintained in the imputed data. Finally, the cross-table $\boldsymbol{X} \times \boldsymbol{Y}_1 \times \boldsymbol{Y}_2$ is estimated by imputing $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ for all population for which either of these values is missing, while calibrating on the estimated cross-tables $\boldsymbol{X} \times \boldsymbol{Y}_1$ and $\boldsymbol{X} \times \boldsymbol{Y}_2$. The imputation model to impute the cross-table $\boldsymbol{X} \times \boldsymbol{Y}_1 \times \boldsymbol{Y}_2$ is based on part B of Figure 2.

An advantage of using CERISE is that one can take edit rules into account on the unit level. By taking these edit rules into account one can avoid some of the inconsistencies that can in principle occur with RW. Taking edit rules into account obviously restricts the imputation models one can apply. CERISE offers more flexibility than RW does as weighting models can always be translated into an imputation model, whereas the converse is not always the case.

Note that, just as in the RW approach, there is generally no need to retain the imputations after estimating a table of population figures and analyzing the results as the imputations are only used (and valid) for producing a particular table, and are not suited for other purposes.

Kooiman's beast can hence be tamed in two ways. First of all, when estimating a table one could include all necessary auxiliary variables, including if one wishes crossings of categorical variables, in the imputation model for that particular table in order to produce accurate results. The imputations cannot be (mis)used for estimating other results. Second, if one feels the beast has not been tamed sufficiently, for example because the quality of some correlations between variables in a table cannot be guaranteed since these correlations were not included in the imputation model, one could delete all imputations and keep only the population estimates. In the CERISE approach, where imputations are only valid for one particular table, deleting these imputations after the table has been produced is no problem.

The strongest aspect of CERISE is that it does not only produce estimates for population totals, but also constructs a (synthetic) population that leads to these totals. One can check the consistency of this synthetic population (for instance, the number of persons with a driver's license should at most be equal to the number of persons with the minimal age to obtain a driver's license), and even the plausibility of this synthetic population (are there many unlikely units or not?). If one deems the constructed synthetic population to be unrealistic, one may decide to impute the data again in order to construct a new synthetic population that, hopefully, has more realistic properties. If one is unable to construct a synthetic population with realistic properties at all, this suggests that something is wrong with the imputation procedure or with the observed data one started with. This is an extra quality check that RW does not offer. Checking the plausibility of a large data set with many observations obviously needs to be done very efficiently in order to keep the costs low. Promising methods are selective editing, where only the most influential records that are also likely to contain errors are checked, and automatic editing, where microdata are checked and correcting automatically by a computer (see De Waal, Pannekoek and Scholtus 2011 for more on selective and automatic editing), and

visual techniques such as the so-called Tableplot (see, e.g., Tennekes, De Jonge and Daas 2013 and Tennekes and De Jonge 2013).
Another advantage of CERISE is that it allows one to produce consistent estimates for several, non-hierarchical classifications, for instance for non-hierarchical age groups or different classifications for branch of industry. Also in this sense CERISE is more flexible than RW classifications have to be hierarchical.
A potential advantage of CERISE is that, given sufficiently powerful imputation models, the imputed data may be used to obtain estimates for small domains, simply by summing or counting the imputed data for each small domain. To which extent this is possible remains to be examined. Preston (2014) also mentions the potential use of imputed data sets as a sampling frame for future samples.
A prerequisite for applying CERISE is an imputation method that succeeds in preserving the statistical aspects of the true data as well as possible, that is able to satisfy specified edit rules and that is able to preserve previously estimated totals. Such imputation methods have recently been developed by, for example, Coutinho, De Waal and Shlomo (2013), Pannekoek, Shlomo and De Waal (2013), and De Waal, Coutinho and Shlomo (2014).
As we saw already, in the context of mass imputation, Kooiman (1998) notes that owing to the lack of degrees of freedom the imputation model will have to ignore some important relations in the data. Kooiman (1998) also notes that this is not the case when one wants to estimate a limited number of tables with a limited number of cells by means of imputation. The imputation models for estimating these tables only need to take the relevant relations for these tables into account, and can safely ignore other relations in the data, exactly what CERISE does.
In Section 4.2 we described some complications of RW. We now briefly discuss whether such complications also occur for CERISE. As CERISE is implemented by applying one or more imputation methods that preserve edits and previously estimated population totals, we select one such imputation method for the comparison. The imputation method we select is the calibrated hot deck imputation method proposed by Coutinho, De Waal and Shlomo (2013) for categorical data. Coutinho, De Waal and Shlomo (2013) actually describe several calibrated hot deck imputation methods, depending on how hot deck is actually carried out. For the purposes of the current paper, the differences between these methods are not important, and will we discuss them as if they are the same.
In the imputation method proposed by Coutinho, De Waal and Shlomo (2013) one aims to use multivariate hot deck imputation, where several missing values in a record are imputed with values from a single donor record. If this is not possible owing to edit constraints or constraints due to previously estimated population totals, the method automatically switches to univariate hot deck imputation, where missing values in a record are imputed with values from several donor records. If even this is not possible, the method automatically switches to imputing values that are allowed according to the edit and population total constraints, but are not observed in the sample. The empty cell problem is hereby avoided. This illustrates the higher flexibility of CERISE compared to RW.
Coutinho, De Waal and Shlomo (2013) take edit rules on the unit level into account. Consistency problems between different variables in different tables, i.e. the edit rule problem, therefore cannot occur.

As for RW, the computation of large, detailed tables can be problematic in the CERISE approach, so computational problems can occur. As for RW, after a number of tables have been estimated, it may become impossible to estimate a new table that it is consistent with all relevant previously estimated marginal totals. The problem of conflicting totals can hence also occur, although it is less likely as one has more degrees of freedom in CERISE as argued in the introduction to Section 5. Finally, as for RW, in CERISE the results are dependent on the order in which the tables are estimated. So, the order dependency problem is also an issue for CERISE.
In contrast to RW, CERISE does not automatically ensure that relationships between data items from a single source are maintained. If one wants to maintain such relationships, they should be included in the imputation model(s).
CERISE has thus far been developed only for cross-sectional data. An extension to longitudinal data is in principle possible by using longitudinal imputation techniques. Thus far the CERISE approach has only been applied only small data sets, not on large data sets arising in practice (see Coutinho, De Waal and Shlomo 2013, Pannekoek, Shlomo and De Waal 2013, and De Waal, Coutinho and Shlomo 2014). Evaluations on large data sets remain to be carried out. Software for the CERISE approach is not yet generally available.

5.3 Measuring the quality

If possible, imputation-based estimators are generally chosen so that they are (approximately) design-based unbiased. Failures of the imputation model, resulting in implausible imputations, should be checked for and prevented.
Given that failures of the imputation model have been checked for and prevented, one generally focuses on estimating the variance of population estimates to measure the quality of imputation-based estimators, just like one does for weighting-based estimators.
Estimating the variance of the population estimates is generally more difficult for imputation than it is for weighting. For imputation in general three different approaches have been proposed to estimate variances (see Chapter 10 by Haziza in Pfeffermann and Rao 2009, and Chapter 7 in De Waal, Pannekoek and Scholtus 2011).
A first approach is the analytical approach. In the analytic approach explicit formulas are derived for variance estimation after imputation. These formulas can be seen as adding a correction term to standard design-based variance formulas for weighting techniques in order to take the fact that imputation is used into account (see, e.g., Fay 1991, Särndal 1992, Deville and Särndal 1992, Rao and Sitter 1995, Shao and Steel 1999 and Beaumont 2005).
A second approach is the resampling approach. In resampling one constructs several versions of the data to be imputed. To each of these different versions one then applies the imputation method. The variation across the results provides a measure for the variance.
Resampling methods, such as the jackknife and bootstrap, have frequently been used for variance estimation in complex surveys with imputed data (see, e.g., Wolter 1985, Rao and Shao 1992, Shao and Sitter 1996 and Shao 2002). An advantage of the resampling approach over the analytic approach is that it is much more generally

applicable. Whereas analytic variance formulas need to be derived for different kinds of imputation methods separately, and can become quite complex, the resampling approach can be used in more or less the same form for a broad range of sampling designs and imputation methods.
A final approach is multiple imputation. In the multiple imputation approach, each missing value is imputed several, say M, times based on the same data, but with variations in the parameters of the imputation model. The variation between the M imputations is used to estimate the increase in variance due to nonresponse and imputation. Simple formulas exist that combine the multiple estimates to a single one and, most importantly, employ the variance between the estimates to obtain an estimator for the variance of the combined parameter estimate. Multiple imputation was meant from the outset as a method to provide not only a solution to the missing data problem but also to reflect the uncertainty in the imputations. The approach was originated by Rubin (1978, 1987). More details on multiple imputation can be found in, e.g., Schafer (1997) and Little and Rubin (2002).
Whereas in the resampling approach one has several versions of the data to be imputed and applies the (stochastic) imputation to each of these versions only once, in the multiple imputation approach one does the opposite: one applies the imputation method to the same data set with missing values several times.
While the analytical approach and the resampling approach can be used for both weighting-based techniques and imputation-based techniques, the multiple imputation approach is only available for imputation-based techniques.
For CERISE the analytic approach appears too hard to apply as variance formulas quickly become too complex to derive. The resampling approach can, however, still be applied. Moreover, some implementations of CERISE allow extensions to multiple imputation versions.

6. Macro-integration

In this section we examine approaches for combining administrative data and surveys based on macro-integration techniques.

6.1 A description of macro-integration

Macro-integration is the process of reconciling statistical figures on an aggregate level. These figures are usually in the form of multi-dimensional tabulations, obtained from different sources. When macro-integration is applied, only estimated figures on aggregated level are adjusted. The underlying microdata are not adjusted or even considered in this adjustment process. The main goal of macro-integration is to obtain a more accurate, consistent and complete set of estimates for the variables of interest. Several methods for macro-integration have been developed, such as the methods by Stone, Champernowne and Meade (1942), Denton (1971), Byron (1978), Sefton and Weale (1995) and Magnus, Van Tongeren and De Vos (2000).

Traditionally, macro-integration has mainly been applied in the area of macro-economics, in particular for compiling the National Accounts. At Statistics Netherlands macro-integration is applied to benchmark quarterly and annual estimates for the National Accounts (see Bikker, Daalmans and Mushkudiani 2013). Also, applications in other areas have been studied at Statistics Netherlands, namely for the reconciliation of tables of Transport and Trade Statistics (see Boonstra, De Blois and Linders 2011), for the Census 2011 (see Mushkudiani, Daalmans and Pannekoek 2012), and for combining estimates of labour market variables (see Mushkudiani, Daalmans and Pannekoek 2012).

The starting point of macro-integration is a set of estimates in tabular form. These can be quantitative tables, for instance a table of average income by region, age and gender, or frequency tables, for instance a cross-tabulation of age, gender, occupation and employment.

If the estimated figures in these tables are based on different sources and (some of) the tables have margins in common, these margins are often conflicting as we have already seen in the example in Section 3.

When one wants to use macro-integration to reconcile the data, the reconciliation process consists of several phases. In the first phase the data sources need to be edited and imputed (see De Waal, Pannekoek and Scholtus 2011) separately. In the next phase these edited and imputed data sources are separately used to estimate aggregated tables. In order to apply a macro-integration method later on, it is important that (an approximation or indication of) the variance of each entry in the tables to be reconciled is computed. In the final phase 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. This is, for instance, the case for the so-called Denton method (see Mushkudiani, Daalmans and Pannekoek 2012, and Bikker, Daalmans and Mushkudiani 2013). 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 in 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. At Statistics Netherlands software has been developed, using modern solvers, for the reconciliation of National Accounts tables that is able to handle problems with a large number of variables (up to 500,000) and constraints (up to 200,000).

In the literature also Bayesian macro-integration methods have been proposed based on a truncated multivariate normal distribution (see Magnus, Van Tongeren and De Vos 2000, and Boonstra, De Blois and Linders 2011). In that Bayesian framework adding inequality constraints is more complicated, although Boonstra, De Blois and Linders (2011) present an approximation method for dealing with inequalities within this framework. Calculations for the truncated multivariate normal distribution are quite complicated, making the model-based approach rather hard to apply, especially for large problems. The approach based on solving a constrained optimization

problem seems to be able to handle much larger integration problems than the model-based approach.
Macro-integration based on solving a mathematical optimization problem is different from model-based macro-integration. However, under certain conditions both kinds of approaches lead to the same results (Fernández 1981).
Boonstra (2004) has compared macro-integration to the calibration or generalized regression (GREG) estimator (see Särndal, Swensson and Wretman 1992). He concludes that the GREG estimator can be seen as a special case of macro-integration.
We will use Tables 1 to 3 to illustrate macro-integration. The macro-integration approach starts by first estimating Tables 1 to 3 separately. For each estimated figure in these tables, one also needs to derive (an indication of) its variance. Next, the estimated figures in these tables are reconciled so the adjusted common figures are equal to each other and additivity of the adjusted tables is maintained. In this reconciliation process only the figures in Tables 1 to 3 and (indications of) their variance are used, not the underlying microdata.
As explained by Mushkudiani, Daalmans and Pannekoek (2012), macro-integration has an important advantage over RW and CERISE: macro-integration can reconcile all tables simultaneously instead of table by table, as long as the number of variables or constraints does not become too large. If tables are reconciled simultaneously, a better solution may be found, requiring less adjustment than RW or CERISE. In principle, one could also apply RW or CERISE to estimate several tables simultaneously, but in practice the number of unknowns to be estimated (the number of cells in the tables or the number of population units for RW, respectively CERISE) quickly become too large to handle. For macro-integration the number of unknowns is often much smaller than for RW or CERISE.
Note that, if one wishes to do so in the macro-integration approach, one can also reconcile separate tables to a set of already estimated tables, although the advantage of finding better solutions would then be lost. Another strong point of the macro-integration approach is that some of the methods have been developed with longitudinal (numerical) data in mind instead of only cross-sectional data.
Macro-integration is a fundamentally different technique than RW and CERISE. For the sake of completeness we briefly discuss to which extent the complications for RW mentioned in Section 4.2 also arise for macro-integration.
Macro-integration methods can be subdivided into methods that lead to additive adjustment of the tables to be reconciled and methods that lead to multiplicative adjustments. With the former kind of macro-integration methods, the empty cell problem cannot occur, whereas with the latter kind the problem can occur.
Macro-integration allows one to take more kinds of edit constraints into account, for example inequality constraints, than RW can. This may lead to more accurate and more consistent final figures than when RW is applied. As for RW and CERISE: the computation of large, detailed tables can be problematic, so computational problems may arise. When tables are estimated simultaneously in the macro-integration approach, the problem of conflicting marginal totals cannot occur and the order dependency problem is not a relevant issue. If separate tables are reconciled with a set of already estimated tables, conflicting tables can arise and the order dependency problem is again an issue.

A drawback of the macro-integration approach in comparison to RW and CERISE is that there is no consistency between the microdata and the reconciled table figures. That is, one cannot re-calculate the table figures from the underlying microdata directly. This problem may be overcome by deriving weights by means of the calibration estimator, using the reconciled macro-integrated figures to calibrate the results on. Such weights do not need to exist, however, for instance owing to the occurrence of empty cells. To overcome this problem, one may develop similar approaches as has been developed for RW, although these approaches lead to their own discrepancies between the microdata and the estimated tables.

6.2 Measuring the quality

When the data only have to satisfy equality constraints variance formulas are quite easy to derive for the macro-integration approach. When the data also have to satisfy inequality constraints deriving variance formulas becomes difficult to derive, however.

Knottnerus (2013 and 2014) has derived (approximate) variance formulas for macro-integration using a quadratic target function in the optimization model if both equality and inequality constraints have to be satisfied. Under certain conditions, these formulas are exact when, besides equality constraints, only a single inequality constraint has to be satisfied. It remains to be examined how well these variance formulas perform when several inequality constraints have to be satisfied.

For implementations of the macro-integration approach by means of model-based methods approximate variance formulas are also available (see Boonstra, De Blois and Linders 2011).

For cases where no (approximate) variance formulas are available or are not accurate enough, one could, in principle, resort to applying re-sampling methods, such as the bootstrap method. As far as we are aware, the application of re-sampling methods for macro-integration remains to be examined, though.

7. Overview of pros and cons

In Table 4 we have summarized the pros and cons of the most promising approaches for obtaining consistent estimates from a mix of administrative data sources and surveys. This table may be used in two different ways: (i) to determine the most suitable approach for obtaining consistent estimates for a given mix of data sources, and (ii) to identify potential research topics for improving the approaches.
Since the properties of macro-integration based on solving a mathematical optimization problem and model-based macro-integration differ slightly, we have listed both versions of macro-integration. We have subdivided the characteristics of the methods into 6 main classes:

1. Consistency issues:
   - Can edit rules be taken into account?
   - Are microdata consistent with the reconciled totals, i.e. if we were to use the microdata to estimate the totals, would we obtain the reconciled results?
   - Can the existence of a (synthetic) population corresponding to the reconciled totals be guaranteed, and can the plausibility of such a (synthetic) population be checked?
   - Are relationships between the data items within a single data source automatically maintained?
   - Can data be checked and edited on a micro level?
   - How time-consuming is this checking process?
2. Estimation issues:
   - Can all tables be estimated simultaneously?
   - Is there an order dependency problem?
   - Can application of the approach lead to conflicting totals so that a new table cannot be reconciled anymore?
3. Computational aspects:
   - Are there computational problems?
   - Can the empty cell problem occur?
4. Additional options of the approach:
   - Can the approach be used for longitudinal besides cross-sectional data?
   - Can the approach be used to obtain consistent estimates for different, non-hierarchical subpopulations?
   - Is the approach potentially suitable for obtaining estimates for small areas?
   - Is the approach potentially suitable for constructing a sampling frame for future surveys?
5. Quality issues:
   - How can one measure the quality of the reconciled estimates? Are variance formulas available? If not, can one estimate the variance of the population estimates in an alternative manner?
6. Practical issues:
   - Can (variations of) the approach handle both categorical and numerical data?
   - How complex is the method to apply in practice?
   - How flexible is the method?

– Is there any danger of (mis)using the data by using it for purposes for which the estimation model was not designed?

7.1.1 Table 4. Overview of the pros and cons of the approaches

	Repeated weighting	**CERISE (Repeated imputation)**	**Macro-integration (optimization)**	**Macro-integration (model-based)**
Consistency issues				
Edit rules taken into account?	Not all edits rules are taken into account	Yes	Yes	Yes, but inequality edits only approximately
Consistency between microdata and estimated totals?	Yes (in a limited form), except in some cases if the epsilon method for the empty cell problem is applied. For different tables to be estimated, different sets of weights have to be generated.	Yes (in a limited form). For different tables to be estimated, different imputations have to be generated.	No, but in some cases the calibration estimator may be used to derive suitable weights	No, but in some cases the calibration estimator may be used to derive suitable weights
Plausible (synthetic) population guaranteed?	No	Yes, a (synthetic) population is guaranteed and plausibility can be checked	No	No
Relationships within data sources automatically maintained?	Yes	No, these relationships have to be added to the model	No. Not really applicable as reconciled microdata are not available	No. Not really applicable as reconciled microdata are not available

	Repeated weighting	CERISE (Repeated imputation)	Macro-integration (optimization)	Macro-integration (model-based)
Can data be checked and edited on micro level?	No	Yes	No	No
How time-consuming is the checking process?	Not very time-consuming as the options for checking are limited	Checking process requires efficient (very) checking methods	Not very time-consuming as the options for checking are limited	Not very time-consuming as the options for checking are limited
Estimation issues				
Can tables be estimated simultaneously?	No	No	Yes	Yes
Order dependency problem?	Yes	Yes	No (Yes, if separate tables are estimated)	No (Yes, if separate tables are estimated)
Possibly conflicting totals?	Yes	Yes, but not very likely	No (Yes, if separate tables are estimated)	No (Yes, if separate tables are estimated)
Computational aspects				
Computational problems?	Yes, for large detailed tables	Yes, for large detailed tables	No, only for extremely large detailed tables	Yes, for very large detailed tables
Empty cell problem?	Yes	No	No for additive methods; yes for multiplicative methods	No for additive methods; yes for multiplicative methods
Additional options of the approach				
Usable for longitudinal data?	No. It is unclear how the approach should be extended for longitudinal data	No, but the approach can be extended to longitudinal data	Yes	No, but the approach can probably be extended to longitudinal data

	Repeated weighting	CERISE (Repeated imputation)	Macro-integration (optimization)	Macro-integration (model-based)
Usable for different, non-hierarchical subpopulations	No	Yes	No	No
Usable for small area estimation?	No	Possibly	No	No
Usable for constructing sampling frames?	No	Possibly	No	No
Quality issues				
Measuring quality	Variance formulas available when the data do not have to satisfy inequality restrictions	By means of resampling or multiple imputation	(Approximate) variance formulas available	(Approximate) variance formulas available
Practical issues				
Categorical and numerical data	Yes	Yes	Yes	No, current implementations have only been developed for numerical data
Complexity	Complex method if one wants to take care of consistency issues as well as possible	Complex method	Once implemented not very complex	Complex method
Flexibility	Not very flexible	Very flexible	flexible	flexible
Danger of misuse?	No	If imputed data sets are preserved, there is a danger of misuse	No	No

8. Discussion

In this paper we have examined general approaches for obtaining consistent population estimates from a combination of administrative data sources and surveys. If we consider statistical matching as a special form of mass imputation or (model-based) macro-integration, these approaches are all general approaches for obtaining consistent estimates for a mix of administrative data and surveys we are aware of.
Of the approaches we have examined, the most promising ones are RW, CERISE and macro-integration. Macro-integration is the least ambitious of the three approaches. In this approach one "merely" aims to construct consistent population estimates after all relevant tables have been estimated separately. Especially for longitudinal data and time series, macro-integration is often an excellent tool for reconciling data over time.
RW and CERISE are more similar to each other. The choice for one of these approaches is hence more difficult to make. RW and CERISE seem about equally complex. A practical advantage of RW is that it is based on weighting, which is a very common and often used technique at NSIs. This makes RW an attractive and natural choice for NSIs.
CERISE is the most ambitious approach. For each table to be estimated, CERISE actually constructs a (synthetic) population that gives the estimated totals and allows one to check whether this population is a plausible one. This makes CERISE the easiest to understand approach from a conceptual point of view. Note that this does not mean that CERISE is also the easiest to understand from a technical point of view since the precise methodological implementation of CERISE may be quite complex. In cases where it is not needed, an imputation-based approach as CERISE may be like cracking a walnut with a sledgehammer. In other cases it is a valuable, and sometimes even indispensable, tool.
Houbiers (2004) notes that "It is not very likely that the ultimate goal of consistency between all possible estimates from the Social Statistical Database will be reached soon, if at all, but repeated weighting seems a suitable method to estimate well-defined table sets consistently. As such, the method should be seen as a tool in the toolbox of estimation methods. Depending on the particular purposes of some publication, and the estimates one wants to make for this publication, one can decide to use repeated weighting or any other suitable estimation".
The same holds for CERISE and macro-integration. All techniques, RW, CERISE and macro-integration, deserve a place in the toolbox. Depending on the precise reconciliation problem, and the complexity one is willing to allow one of these tools can be picked. If one has little time available, one usually has to resort to a relatively simple approach, such as macro-integration. If one has more time and highly skilled staff available, one may be willing to use a more complicated approach, such as CERISE. Such a more complicated approach may have the advantage that the data quality is better or can be better guaranteed, that more detailed figures can be estimated, or that consistency on a more detailed level is ensured.
RW, CERISE and macro-integration leave plenty opportunities for future research. As already mentioned in Section 7, Table 4 can be used to identify potential research

topics. Examples are determining in which cases the calibration estimator can be used to derive suitable weights to maintain the consistency between estimated totals and microdata for the macro-integration approach, extending CERISE to longitudinal data, and extending RW to non-hierarchical classifications.
Combining RW, CERISE and macro-integration is another area for future research. One could, for example use macro-integration to first obtain estimated population figures and then use RW or CERISE to calibrate the microdata to these estimated figures. In that way one might avoid the order dependency problem, and at the same time profit from the pros of RW or CERISE.
In this paper we have dismissed mass imputation as a viable option for rich data sets with many variables, especially if not all tables to be estimated are specified beforehand. However, for data sets with a limited number of variables and for which all tables to be estimated can be specified beforehand, mass imputation appears to be an excellent option. On all aspects mentioned in Table 4, except “complexity” and “danger of misuse”, the score seems to be positive. When all tables to be estimated are known beforehand, the danger of misuse can be prevented, or in any case severely limited, by not using the microdata anymore after the estimates have been produced.
A final research topic would be the development of a novel general approach that would overcome any drawbacks of the general approaches considered in this paper. For the moment we leave this task to the reader.

References

Bakker, B. F. M. (2011), Micro-integration. Statistical Methods Series, Statistics Netherlands.
Bakker, B. F. M. (2012), Estimating the Validity of Administrative Variables. Statistica Neerlandica 66, pp. 8–17.
Beaumont J.-F. (2005), Calibrated Imputation in Surveys under a Quasi-Model-Assisted Approach. Journal of the Royal Statistical Society B 67, pp. 445–458.
Berger, Y.H., J.F. Muñoz and E. Rancourt (2009), Variance Estimation of Survey Estimates Calibrated on Estimated Control Totals – An Application to the Extended Regression Estimator and the Regression Composite Estimator. Computational Statistics and Data Analysis 53, pp. 2596-2604.
Bethlehem, J. (2009). Applied survey methods: A Statistical Perspective. John Wiley & Sons, New York.
Bikker, R., J. Daalmans and N. Mushkudiani (2013). Benchmarking Large Accounting Frameworks: a Generalised Multivariate Model. Economic Systems Research 25, pp. 390-408.
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, pp. 407-431.

Buelens, B., P. Daas, J. Burger, M. Puts and J. van den Brakel (2014), Selectivity of Big Data, Discussion paper, Statistics Netherlands.
Byron, R.P. (1978), The Estimation of Large Social Account Matrices. Journal of the Royal Statistical Society A 141, pp. 359-367.
Cochran, W.G. (1977), Sampling Techniques. John Wiley & Sons.
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. (2014), Estimating Detailed Frequency Tables from Registers and Sample Surveys. Discussion paper, Statistics Netherlands.
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.
Deville, J.-C. and Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. Journal of the American Statistical Association 87, pp. 376–382.
De Waal, T., W. Coutinho and N. Shlomo (2014), Calibrated Hot Deck Imputation for Numerical Data under Edit Restrictions. Discussion paper, Statistics Netherlands (forthcoming).
De Waal, T., J. Pannekoek and S. Scholtus (2011), Handbook of Statistical Data Editing and Imputation. John Wiley & Sons, New York.
D'Orazio, M., M. Di Zio and M. Scanu (2006), Statistical Matching: Theory and Practice. John Wiley & Sons, New York.
Fay, R.E. (1991), A Design-Based perspective on Missing Data Variance. In: Proceedings of the 1991 Annual Research Conference. Washington, D.C.
Fernández, R.B. (1981) A Methodological Note on the Estimation of Time Series. The Review of Economics and Statistics 63, pp. 471–476.
Gerritse, S., P.G.M. van der Heijden and B.F.M. Bakker (2013), Robustness of Population Size Estimates against Violation of the Independence Assumption. In: Proceedings of the 59th ISI World Statistics Conference, Hong Kong.
Houbiers, M. (2004), Towards a Social Statistical Database and Unified Estimates at Statistics Netherlands. Journal of Official Statistics 20, pp. 55-75.
Houbiers, M., P. Knottnerus, A.H. Kroese, R.H. Renssen and V. Snijders (2003), Estimating Consistent Table Sets: Position Paper on Repeated Weighting. Discussion paper, Statistics Netherlands.
Knottnerus, P. (2001), Varianties bij Herhaald Wegen. Internal note, Statistics Netherlands.
Knottnerus, P. (2003), Sample Survey Theory: Some Pythagorean Perspectives. Springer-Verlag, New York.
Knottnerus, P. (2013), Macro-integratie, Regressietheorie, Kalman-Vergelijkingen en Bayesiaanse Posteriors. Internal note, Statistics Netherlands.
Knottnerus, P. (2014), Variantieformules voor Macro-Integratie en Regressies met Ongelijkheidsrestricties. Internal note, Statistics Netherlands.
Knottnerus, P. and C. Van Duin (2006), Variances in Repeated Weighting with an Application to the Dutch Labour Force Survey. Journal of Official Statistics 22, pp. 565–584
Kooiman, P. (1997), Sociaal Statistisch Bestand: Wensdroom of Nachtmerrie. Internal note, Statistics Netherlands.

Kooiman, P. (1998), Massa-imputatie: Waarom Niet!?. Internal note, Statistics Netherlands.
Kroese, A.H. and R.H. Renssen (2000), New Applications of Old Weighting Techniques; Constructing a Consistent Set of Estimates Based on Data from Different surveys. In: Proceedings of ICES II. American Statistical Association, Buffalo NY, pp. 831-840.
Kroese, B., R.H. Renssen and M. Trijssenaar (2000), Weighting or Imputation: Constructing a Consistent Set of Estimates Based on Data from Different Sources. Netherlands Official Statistics 15, pp. 23-31.
Kuijvenhoven, L. and S. Scholtus (2011), Bootstrapping Combined Estimators Based on Register and Sample Survey Data. Discussion paper, Statistics Netherlands.
Little, R.J.A. and Rubin, D.B. (2002), Statistical Analysis with Missing Data, 2nd ed. John Wiley & Sons, Hoboken, NJ.
Magnus, J.R., 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.
Mushkudiani, N., J. Daalmans and J. Pannekoek (2012), Macro-Integration Techniques with Applications to Census Tables and Labour Market Statistics. Discussion paper, Statistics Netherlands.
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 (2014), Measuring Temporary Employment. Do Survey or Register Data Tell the Truth? Accepted for publication in Survey Methodology.
Pfeffermann, D. and Rao, C. R. (2009), Handbook of Statistics 29. Elsevier, Amsterdam.
Preston, J. (2014), Treatment of Missing Data in Statistical Data Integration. Report, Australian Bureau of Statistics.
Rao, J.N.K. and Shao, J. (1992), Jackknife Variance Estimation with Survey Data under Hot Deck Imputation. Biometrika 79, pp. 811–822.
Rao, J. N. K. and R. R. Sitter (1995), Variance Estimation under Two-Phase Sampling with Application to Imputation for Missing Data. Biometrika 82, pp. 453–460.
Renssen, R.H. and N.J. Nieuwenbroek (1997), Aligning Estimates for Common Variables in Two or More Sample Surveys. Journal of the American Statistical Association 92, pp. 368-374.
Rubin, D.B. (1978), Multiple Imputations in Sample Surveys – A Phenomenological Bayesian Approach to Nonresponse. In: Proceedings of the Section on Survey Research Methods. American Statistical Association.
Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, New York.
Särndal, C.-E. (1992), Methods for Estimating the Precision of Survey Estimates when Imputation Has Been Used. Survey Methodology 18, pp. 241–252.
Särndal, C.E., B. Swensson and J. Wretman (1992), Model Assisted Survey Sampling. New York: Springer-Verlag.
Schafer, J.L. (1997), Analysis of Incomplete Multivariate Data. Monographs on Statistics and Applied Probability 72. Chapman & Hall, London.
Scholtus, S. (2014), Explicit and Implicit Calibration of Covariance and Mean Structures. Discussion paper, Statistics Netherlands.

Scholtus, S. and B.F.M. Bakker (2013), Estimating the Validity of Administrative and Survey Variables through Structural Equation Modeling. Discussion paper, Statistics Netherlands.
Sefton, J. and M. Weale (1995), Reconciliation of National Income and Expenditure. Cambridge University Press, Cambridge, UK.
Shao, J. (2002), Replication Methods for Variance Estimation in Complex Surveys with Imputed Data. In: Survey Non-Response (eds. R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A. Little), pp. 303–324. John Wiley & Sons, New York.
Shao, J. and R.R. Sitter (1996). Bootstrap for Imputed Survey Data. Journal of the American Statistical Association 93, pp. 819-831.
Shao, J. and Steel, P. (1999), Variance Estimation for Survey Data with Composite Imputation and Nonnegligible Sampling Fractions. Journal of the American Statistical Association 94, pp. 254–265.
Shlomo, N., T. de Waal and J. Pannekoek (2009), Mass Imputation for Building a Numerical Statistical Database. UN/ECE Work Session on Statistical Data Editing, Neuchâtel, Switzerland.
Stone, R., D.G. Champernowne and J.E. Meade (1942). The Precision of National Income Estimates. Review of Economic Studies 9, pp. 111-125.
Tennekes, M. and E. de Jonge (2013), Visual Profiling of Large Statistical Datasets. Paper presented at the NTTS conference.
Tennekes, M., E. de Jonge and P. Daas (2013), Visualizing and Inspecting Large Datasets with Tableplots. Journal of Data Science 11, pp. 43-58.
Van de Laar. R. (2004), Edit Rules and the Strategy of Consistent Table Estimation. Discussion paper, Statistics Netherlands.
Van Duin, C. and V. Snijders (2003), Simulation Studies of Repeated Weighting. Discussion paper, Statistics Netherlands.
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.
Wolter, K. M. (1985), Introduction to Variance Estimation. Springer-Verlag, New York.
Zhang, L.-C. (2012), Topics of Statistical Theory for Register-Based Statistics and Data Integration. Statistica Neerlandica 66, pp. 41-63.

Explanation of symbols

.	Data not available
*	Provisional figure
**	Revised provisional figure (but not definite)
x	Publication prohibited (confidential figure)
–	Nil
–	(Between two figures) inclusive
0 (0.0)	Less than half of unit concerned
Empty cel	Not applicable
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