Obtaining numerically consistent estimates from a mix of administrative data and surveys

de Waal, A.G.

Published in:
Statistical Journal of the IAOS

DOI:
10.3233/SJI-150950

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

Statistical Journal of the IAOS 32 (2016) 231–243
DOI 10.3233/SJI-150950
IOS Press

Obtaining numerically consistent estimates from a mix of administrative data and surveys

Ton de Waal[a,b]
[a]*Statistics Netherlands, PO Box 24500, 2490 HA The Hague, The Netherlands*
[b]*Tilburg University, Warandelaan 2, 5037 AB, Tilburg, The Netherlands*
Tel.: +31 70 337 4930, +31 13 466 2395; E-mail: t.dewaal@cbs.nl, a.g.dewaal@uvt.nl

Abstract. 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. In this paper we focus on one of the challenges when using a mix of administrative data sets and surveys, namely obtaining numerically consistent population estimates. We will sketch general approaches based on weighting, imputation and macro-integration for solving this problem, and discuss their advantages and drawbacks.

Keywords: Administrative data, data integration, imputation, macro-integration, weighting, survey data

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. 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. 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 obtaining numerically consistent estimates for population totals based on a mix of administrative data sets and surveys. Here, and in the rest of the paper, the terms "consistent" and "consistency" refer to numerical equality of estimates in different tables, not to "consistency" in the usual meaning in mathematical statistics.

At Statistics Netherlands, obtaining numerically consistent estimates is especially a problem for the Dutch Census. In the Netherlands the Census is not based on a complete enumeration of the Dutch population. To produce the tables required for the Dutch Census, Statistics Netherlands instead combines available administrative data and survey data (see [23]). This would lead to inconsistent estimates if standard weighting techniques were to be applied as we will illustrate in Section 2 of this paper.

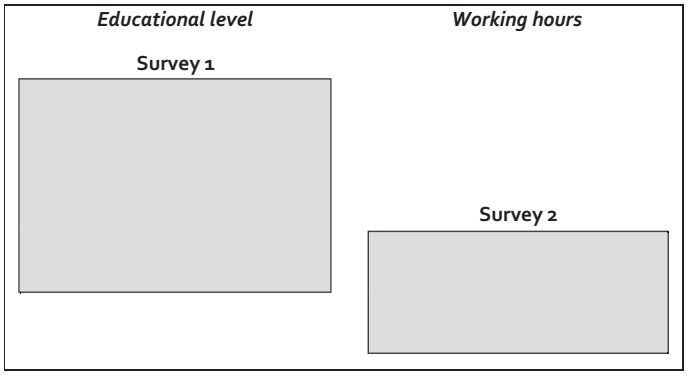

Fig. 1. Illustration of the problem, part 1.

In this paper we will sketch general approaches for obtaining numerically consistent estimates for population totals 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 5), can handle both categorical and numerical data.

The world of official statistics is, at least up to now, fundamentally a design-based one rather than a model-based one, if only because NSIs do not want to be accused of subjectivity in their choice of models. Statistical models are used, but mainly to assist the design-based framework. A common aspect of the approaches we will discuss in this paper is therefore that they can all be applied in a design-based or model-assisted framework. This holds true even for imputation, where one can use a "design-based" method such as hot deck imputation rather than posit an explicit model for the missing data, and for macro-integration, where a model is only used to reconcile design-based estimates.

This paper is organized as follows. Section 2 describes the problem of obtaining numerically consistent estimates from a mix of administrative data and survey data in some detail. Sections 3 to 5 sketch the general approaches for combining data we consider in this paper. Methods based on weighting are discussed in Section 3, methods based on imputation in Section 4, and methods based on macro-integration in Section 5. Section 6 gives an overview of the pros and cons of the most promising approaches. In a sense Table 11 in Section 6, which summarizes these pros and cons, may be seen as the main result of this paper. Finally, Section 7 concludes the paper with a brief discussion.

2. The consistent estimation problem

Different estimates for the same phenomenon could lead to confusion among users of these figures. Many 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.

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 problem with a small fictitious example where we aim to combine the estimates of only two samples. In this example we have a population of 10,000 persons from which we draw two surveys by means of simple random sampling without replacement. Survey 1 contains information on the educational level of 2,000 persons in three categories: low, medium and high. Each person in Survey 1 has a survey weight of 5. Survey 2 contains information on "working hours" of 1,000 persons in two categories: fulltime (more than 35 hours per week) and part-time (at most 35 hours per week). Each person in Survey 2 has a survey weight of 10. Two hundred persons are selected in both Survey 1 and Survey 2, and each of those persons has a survey weight of 50. The situation is shown in Fig. 1.

Suppose we want to estimate the tables "educational level", "working hours" and "educational level x work-

Table 1
Observed numbers in Survey 1

Educational level	Observed number
Low	350
Medium	1,000
High	650

Table 2
Population estimates for "educational level"

Educational level	Estimated total
Low	1,750
Medium	5,000
High	3,250

Table 3
Observed numbers in Survey 2

Working hours	Observed number
Fulltime	600
Part-time	400

Table 4
Population estimates for "working hours"

Working hours	Estimated total
Fulltime	6,000
Part-time	4,000

ing hours". We use Survey 1 to estimate the table "educational level". Table 1 contains the observed values in this survey. The population estimates for "educational level", obtained by multiplying the numbers in Table 1 with the survey weights, are given in Table 2.

We use Survey 2 to estimate the table "working hours". Table 3 contains the observed values in this survey. The population estimates for "working hours", obtained by multiplying the numbers in Table 3 with the survey weights, are given in Table 4.

Finally, we estimate the table "educational level x working hours" by means of all units for which we have observed both "educational level" and "working hours", i.e. the overlap of the two surveys (see the shaded parts in Fig. 2). Table 5 contains the observed values in this overlap. The population estimates for "educational level x working hours", obtained by multiplying the numbers in Table 5 with the survey weights, are given in Table 6.

The tables with population estimates, Tables 2, 4 and 6, illustrate the problem of obtaining numerically consistent estimates. Consider, for instance, the number of persons with a high educational level. According to Table 2 this number is estimated as 3,250, whereas according to the more detailed Table 6 this number is estimated as 3,000. Analogously, according to Table 4 the number of person working fulltime is estimated as

Table 5
Observed numbers in the overlap of the two surveys

	Fulltime	Part-time	Observed total
Low	30	20	50
Medium	50	40	90
High	30	30	60
Total	110	90	200

Table 6
Population estimates for "educational level x working hours"

	Fulltime	Part-time	Estimated total
Low	1,500	1,000	2,500
Medium	2,500	2,000	4,500
High	1,500	1,500	3,000
Total	5,500	4,500	10,000

6,000, whereas according to Table 6 this number is estimated as 5,500. Estimates for the same phenomenon hence differ in different tables.

Without taking special precautions, one will obtain different estimates for "educational level" and "working hours", depending on the units on which the estimates are based.

In this example we had only two surveys. In practice we often have more data sources, not only sample surveys but also administrative data sources. Naturally, this further complicates the problem.

3. Weighting-based approaches

In this section we examine approaches for obtaining numerically consistent estimated from several data sources based on weighting the data.

3.1. The traditional weighting approach

The traditional way in survey sampling to estimate population totals is by assigning a survey 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, counts for 24 units with the same values for the target variables in the population, of whom 23 were not selected in the sample. 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.

To obtain survey weights per sample unit, one usually 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 re-

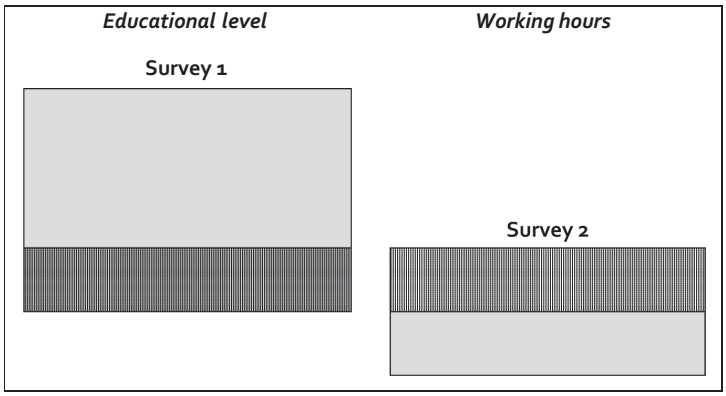

Fig. 2. Illustration of the problem, part 2.

spond, sample weights are often adjusted slightly before they are used to calculate weighted totals. 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 are 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.

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.

A strong point of the traditional weighting approach in the case of a single survey is that relationships between variables are automatically maintained.

In the traditional weighting approach, after estimating and analysing the relevant population totals, one generally retains the survey weights in the data set as this allows one to later analyse the data in more detail. For example, at first one may be interested only in analysing 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 analyse these correlations.

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 Figs 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 succeeded in doing so. Owing to this, the traditional weighting approach cannot really be used for obtaining numerically consistent estimates from a mix of administrative data and survey data.

In the remainder of this paper we will 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.

3.2. Repeated weighting

As a way to overcome the problems of the traditional weighting approach, the so-called "repeated weighting" (RW) approach was developed at Statistics Netherlands in the late 1990s (see, e.g. [11,12,15]). In the RW approach a separate set of weights is assigned to sample units for each table of population totals to be estimated.

Table 7
Population estimates for "educational level x working hours" after repeated weighting

	Fulltime	Part-time	Estimated total
Low	1,200	550	1,750
Medium	2,800	2,200	5,000
High	2,000	1,250	3,250
Total	6,000	4,000	10,000

In the RW approach population tables are estimated sequentially and each table is estimated using as many sample units as possible in order to keep the sample variance as low as possible. The combined data from 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. 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 [20]). 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 known or 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 cell values and margins of this table that are known or have already been estimated for previous tables are kept fixed to these known or previously estimated values, i.e. the regression weighting is calibrated on these known or previously estimated values. This ensures that the cell values and margins of the new table are numerically consistent with previous estimates.

We will use Figs 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 table for "educational level". To estimate this table we apply weighting and obtain the results in Table 2. We also estimate the table for "working hours" by means of weighting. We obtain the results in Table 4. Next, we estimate the table "working hours x educational level" by calibrating on the estimated numbers in Tables 2 and 4. The results depend on the exact weighting model, and could, for instance, be given by the numbers in Table 7. In this way previously estimated values are preserved when making estimates for a new table. After estimation of a population table, the weights used to produce this table can be thrown away since they are only valid for this particular table anyway. If one wants to estimate a new table, new weights for that table have to be derived.

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". In [12] more details on the RW approach are provided.

A strong aspect of the RW approach is that it, just like the traditional weighting approach, automatically ensures that relationships between data items from a single data source are maintained.

The RW approach may seem simple, but is not without complications as noted by [5,12]. 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 previously estimated marginal total, but the data from which the table must be estimated does not contain any units corresponding to this marginal. 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" is used as an initial estimate. However, this attempt 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.

Another complication is that, although the approach takes known or previously estimated totals into account, it does not take so-called edit rules into account

(*the edit rule problem*). An example of such an edit rule 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 [15,25]). In principle, the RW approach could be modified to include such edit rules (see [5,25]). For instance, when a variable involved in edit rules occurs in a table to be estimated, Daalmans (see [5]) extends the table by adding all variables involved in those edit rules. 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 rules.

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 [12] is noted 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 to develop.

Like the other techniques in this paper, RW is mainly applied for cosmetic purposes, namely to ensure numerical 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. The same reduction of sample variance when calibrating to totals based on large sample sizes also occurs for repeated imputation and macro-integration which we discuss later.

4. Imputation-based approaches

In this section we discuss approaches for obtaining numerically consistent estimates from several data sources based on imputation techniques.

4.1. Mass imputation

In the mass imputation approach, one imputes all variables for which no value was observed for all population units, even for units that were intentionally not observed, for instance because they were not included in a sample survey (see [22,26,27]). This leads to a rectangular data set with values for all variables and all population units. The imputations are generated by means of an (explicit or implicit) imputation model. After imputation, estimates for population totals can be obtained by simply counting or summing 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 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 [13] 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 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 was 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 in an imputation model is, except in very exceptional cases, not deemed important enough. Including information on their relation in an imputation model is even more unlikely. If the relation between these two variables is ignored, values imputed for "amount of money spent per month on dog food" does not depend on the value for "do you have a dog as pet", and vice versa.

In his example Kooiman indeed assumed that these variables and information on their relation are not in-

cluded 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 in producing estimates. 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.

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 analyse 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 as pet do not buy dog food. This conclusion would make an interesting headline in a newspaper, for example as "proof" for extreme poverty in the Netherlands, 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.

In principle, one could use the mass imputation approach to obtain estimates of the relevant population totals, analyse them, and then delete the imputations that led to these results. 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 analyse the data set in more detail later.

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. In [14] Kooiman therefore concluded that mass imputation is not a viable strategy for obtaining numerically consistent estimates from a set of administrative data sources and surveys.

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. However, in this paper we will focus on the situation where not all tables to be estimated are known beforehand. Apart from a brief remark in the Discussion, we will therefore not consider mass imputation anymore in the remainder of this paper.

4.2. Repeated imputation

Whereas mass imputation is the equivalent of traditional weighting, repeated imputation (RI) 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 RI by means of an imputation method. Like RW, RI is a sequential approach where tables are estimated one by one. For some variables in a table estimates may have already been produced while estimating a previous table. Similar to RW, these variables are then calibrated to the previously estimated totals.

In RI, imputation is not seen as a way to obtain a complete data set, but as an estimation technique. To emphasize that RI is an estimation method rather than a way to obtain complete data, we have also given it the name CERISE (*Consistent Estimation using Repeated Imputation Satisfying Edits*) instead of "repeated imputation" (see [7]).

We will again use Figs 1 and 2 to explain the basic ideas of RI. As in the RW approach, low-dimensional tables are in principle estimated before higher-dimensional ones. RI would, for instance, start with estimating the table "educational level" by imputing the variable "educational level" in all population units for which its value is missing. The imputation model is based on Survey 1. Next, we estimate the table "working hours" by imputing the variable "working hours" in all population units for which its value is missing. The imputation model is based on Survey 2. Finally, we estimate the table "educational level x working hours". The imputation model is based on the overlap of the two surveys and we calibrate on previously estimated values. Depending on the exact imputation model used, we might get similar results as in Tables 2, 4 and 7.

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 analysing the results as the imputations are only used (and valid) for producing a particular table, and are not suited for other purposes. If one wants to estimate a new table, new imputations have to be generated for that table.

An advantage of using RI is that one can take edit rules into account on the unit level. By taking these edit rules into account one can avoid inconsistencies that can occur with RW.

A strong aspect of RI is that it does not only produce estimates for population totals, but also constructs a (synthetic) population that leads to these totals. One can easily check whether this synthetic population satisfies edit rules. One can also check 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 a quality check that RW does not offer.

Another advantage of RI is that it allows one to produce numerically consistent estimates for several non-hierarchical classifications, for instance for non-hierarchical age groups or different classifications for branch of industry. In this sense RI is more flexible than RW.

A potential advantage of RW 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. In [19] also the potential use of imputed data sets as a sampling frame for future samples is mentioned as a possible advantage.

A prerequisite for applying RI 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, see, e.g. [4,8,18].

As we saw already in the context of mass imputation, Kooiman (see [14]) notes that owing to the lack of degrees of freedom an imputation model for mass imputation will have to ignore some important relations in the data. In [14] it is also noted 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 RI does.

Kooiman's principal objection to mass imputation can hence be overcome in two ways with RI. First of all, when estimating a table, which will generally contain relatively few variables, one can include all necessary auxiliary variables in the imputation model for that particular table in order to produce accurate results. Second, one can delete all imputations after estimation of a table and keep only the population estimates.

In Section 3.2 we described some complications of RW. We now briefly discuss whether such complications also occur for RI. As RI is implemented by applying one or more imputation methods that preserve edit rules 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 [4] for categorical data. Actually, in [4] several calibrated hot deck imputation methods, depending on how hot deck is actually carried out, are described. 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 [4] 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 rules and population total constraints, but are not observed in the sample. The empty cell problem is hereby avoided.

In [4] edit rules are taken into account on the unit level. 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 RI 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 numerically 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 RI (total number of missing values) as in RW (number of weights, i.e. number of records). Finally, as for RW, in RI the results are dependent on the order in which the tables are estimated. So, the order dependency problem is also an issue for RI.

In contrast to RW, RI 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).

RI has thus far been developed only for cross-sectional data. An extension to longitudinal data is in principle possible by using longitudinal imputation techniques.

5. 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 an 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, numerically consistent and complete set of estimates for the variables of interest. Several methods for macro-integration have been developed, see, e.g. [3,6,16,21,24].

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 [1]). Also, applications in other areas have been studied at Statistics Netherlands, namely for the reconciliation of tables of Transport and Trade Statistics (see [2]), for the Census 2011 (see [17]), and for combining estimates of labour market variables (see [17]).

The starting point of macro-integration is a set of estimates in tabular form. These can be quantitative tables, for instance tables of average income by region, age and gender, or frequency tables, for instance cross-tabulations of age, gender, occupation and employment. If the estimated figures in these tables are based on different sources and (some of) the tables have cells in common, these cell values are often conflicting as we have already seen in the example in Section 2.

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 [9]) 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.

Table 8
Population estimates for "educational level" after macro-integration

Educational level	Estimated total
Low	2,150
Medium	4,750
High	3,100

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 [1,17]). 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 internal cell values of the adjusted tables sum up to the corresponding marginal totals. 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 [2,16]). In that Bayesian framework adding inequality constraints is more complicated, although [2] 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 (see [10]).

Table 9
Population estimates for "working hours" after macro-integration

Working hours	Estimated total
Fulltime	5,750
Part-time	4,250

Table 10
Population estimates for "educational level x working hours" after macro-integration

	Fulltime	Part-time	Estimated total
Low	1,350	800	2,150
Medium	2,650	2,100	4,750
High	1,750	1,350	3,100
Total	5,750	4,250	10,000

We will use the example in Section 2 to briefly illustrate macro-integration. The macro-integration approach starts by first estimating the population totals for "educational level", "working hours" and "educational level x working hours", i.e. Tables 2, 4 and 6, 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 internal cell values sum up to the corresponding marginal totals. In this reconciliation process only the figures in Tables 2, 4 and 6 and (indications of) their variance are used, not the underlying microdata. This may lead to the adjusted Tables 8 to 10.

As explained by [17], macro-integration has an important advantage over RW and RI: 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 RI. 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.

We briefly discuss to which extent the complications for RW mentioned in Section 3.2 also arise for macro-integration. Macro-integration methods can be subdivided into methods that lead to additive adjustments 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.

Like RI, macro-integration can take edit rules into account. As for RW and RI: the computation of large, detailed tables can be problematic. So computational problems may arise, although the problems that can be solved by macro-integration are much larger than for RW and RI. 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 RI is that there is no direct relation 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.

6. Overview of pros and cons

In Table 11 we have summarized the pros and cons of the general approaches for obtaining numerically consistent estimates from a mix of administrative data sources and surveys. In this table we consider current implementations of these general approaches or relatively simple extensions thereof. Table 11 may be used in two different ways: (i) to determine the most suitable approach for obtaining numerically consistent estimates for a given mix of data sources, and (ii) to identify potential research topics for improving the approaches. A new aspect in Table 11 that we have not yet discussed before is "Quality issues", i.e. how the variance of population estimates can be estimated. For more on this aspect of the approaches, we refer to [7].

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:

- Consistency issues:
  - Can edit rules be taken into account?
  - Are microdata directly related to the reconciled totals, i.e. if we were to use the microdata to estimate the totals, would we obtain the reconciled results?

Table 11
Overview of the pros and cons of the approaches

	Repeated weighting	Repeated imputation	Macro-integration (optimization)	Macro-integration (model-based)
Consistency issues				
Edit rules taken into account?	Not all edit rules are taken into account	Yes	Yes	Yes, but inequality edits only approximately
Consistency between microdata and estimated totals?	Yes, except in some cases if the epsilon method for the empty cell problem is applied	Yes	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 explicitly to the imputation model	No. Not really applicable as reconciled microdata are not available	No. Not really applicable as reconciled microdata are not available
Can data be checked and edited on micro level?	No	Yes	No	No
How time-consuming is the process of checking the plausibility of estimates?	Not very time-consuming as the options for checking are limited	If one wants to check the plausibility of the imputed microdata, (very) efficient methods are required	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 to longitudinal data	No, but the approach can be extended to longitudinal data	Yes	No, but the approach can probably be extended to longitudinal data
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 are available when the data do not have to satisfy inequality restrictions	By means of resampling or multiple imputation	(Approximate) variance formulas are available	(Approximate) variance are formulas available
Practical issues				
Applicable to 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 small danger of misuse	No	No

  - 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 the process of checking the plausibility of estimates?
- 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?

- Computational aspects:
  - Are there computational problems?
  - Can the empty cell problem occur?
- Additional options of the approach:
  - Can the approach be used for longitudinal besides cross-sectional data?
  - Can the approach be used to obtain numerically 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?
- 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?
- 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? That is, how many options does one have to amend estimates and take edit rules into account?
  - Is there any danger of (mis)using the data by using it for purposes for which the estimation model was not designed?

7. Discussion

In this paper we have examined several general approaches for obtaining numerically consistent population estimates from a mix of administrative data sources and surveys that can all be applied in a design-based or model-assisted framework. Of the approaches we have examined, the most promising ones are RW, RI and macro-integration. Macro-integration is the least ambitious of these three approaches. In this approach one "merely" aims to construct numerically 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 RI are more similar to each other. The choice for one of these approaches is hence more difficult to make. RW and RI 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.

RI is the most ambitious approach. For each table to be estimated, RI actually constructs a (synthetic) population that gives the estimated totals and allows one to check whether this population is a plausible one.

All techniques, RW, RI and macro-integration, deserve a place in the toolbox of an NSI. Depending on the precise reconciliation problem, and the complexity one is willing to allow one of these tools can be chosen. 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 RI. 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.

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 11, 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.

RW, RI and macro-integration leave plenty opportunities for future research. As already mentioned in Section 6, Table 11 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 a direct relation between estimated totals and microdata for the macro-integration approach, extending RI to longitudinal data, and extending RW to non-hierarchical classifications.

Combining RW, RI 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 RI 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 RI.

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

[1] R. Bikker, J. Daalmans and N. Mushkudiani, Benchmarking Large Accounting Frameworks: A Generalised Multivariate Model, *Economic Systems Research* **25** (2013), 390–408.

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

[3] R.P. Byron, The Estimation of Large Social Account Matrices, *Journal of the Royal Statistical Society A* **141** (1978), 359–367.

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

[5] J. Daalmans, *Estimating Detailed Frequency Tables from Registers and Sample Surveys*. Discussion paper, Statistics Netherlands, 2015.

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

[7] T. de Waal, *General Approaches for Consistent Estimation based on Administrative Data and Surveys*, Discussion paper, Statistics Netherlands, 2015.

[8] T. de Waal, W. Coutinho and N. Shlomo, *Calibrated Hot Deck Imputation for Numerical Data under Edit Restrictions*, Discussion paper, Statistics Netherlands, 2015.

[9] T. de Waal, J. Pannekoek and S. Scholtus, *Handbook of Statistical Data Editing and Imputation*, John Wiley & Sons, New York, 2011.

[10] R.B. Fernández, A Methodological Note on the Estimation of Time Series, *The Review of Economics and Statistics* **63** (1981), 471–476.

[11] M. Houbiers, Towards a Social Statistical Database and Unified Estimates at Statistics Netherlands, *Journal of Official Statistics* **20** (2004), 55–75.

[12] M. Houbiers, P. Knottnerus, A.H. Kroese, R.H. Renssen and V. Snijders, *Estimating Consistent Table Sets: Position Paper on Repeated Weighting*, Discussion paper, Statistics Netherlands, 2003.

[13] P. Kooiman, *Sociaal Statistisch Bestand: Wensdroom of Nachtmerrie*, Internal note, Statistics Netherlands, 1997.

[14] P. Kooiman, *Massa-imputatie: Waarom Niet!?* Internal note, Statistics Netherlands, 1998.

[15] A.H. Kroese and R.H. Renssen, 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, 2000, pp. 831–840.

[16] J.T. Magnus, J.W. van Tongeren and A.F. de Vos, National Accounts Estimation using Indicator Ratios, *The Review of Income and Wealth* **46** (2000), 329–350.

[17] N. Mushkudiani, J. Daalmans and J. Pannekoek, *Macro-Integration Techniques with Applications to Census Tables and Labour Market Statistics*, Discussion paper, Statistics Netherlands, 2012.

[18] J. Pannekoek, N. Shlomo and T. de Waal, Calibrated Imputation of Numerical Data under Linear Edit Restrictions, *Annals of Applied Statistics* **7** (2013), 1983–2006.

[19] J. Preston, *Treatment of Missing Data in Statistical Data Integration*, Report, Australian Bureau of Statistics, 2014.

[20] C.E. Särndal, B. Swensson and J. Wretman, *Model Assisted Survey Sampling*, New York: Springer-Verlag, 1992.

[21] J. Sefton and M. Weale, *Reconciliation of National Income and Expenditure*, Cambridge University Press, Cambridge, UK, 1995.

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

[23] Statistics Netherlands, *Dutch Census 2011: Analysis and Methodology*, Report, Statistics Netherlands, 2014.

[24] R. Stone, D.G. Champernowne and J.E. Meade, The Precision of National Income Estimates, *Review of Economic Studies* **9** (1942), 111–125.

[25] R. Van de Laar, *Edit Rules and the Strategy of Consistent Table Estimation*, Discussion paper, Statistics Netherlands, 2004.

[26] P. Whitridge, M. Bureau and J. Kovar, Mass Imputation at Statistics Canada, in: *Proceedings of the Annual Research Conference*, U.S. Census Bureau, Washington D.C., 1990, pp. 666–675.

[27] P. Whitridge and J. Kovar, Use of Mass Imputation to Estimate for Subsample Variables, in: *Proceedings of the Business and Economic Statistics Section*, American Statistical Association, 1990, pp. 132–137.