

Estimating detailed frequency tables from registers and sample surveys

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

2015 | 03

Jacco Daalmans

Repeated weighting is a method for the consistent estimation of multiple frequency tables from registers and sample surveys. Statistics Netherlands uses repeated weighting for the compilation of the Dutch census. The application of the method is not without its problems. Estimation problems may occur, especially for detailed tables, of which some of the cells are covered by only few (or even none) observations.

This paper reviews existing solutions for methodological problems and proposes new solutions when necessary. The problems and solutions are illustrated with experiences of the Dutch 2011 Census compilation. The general message of this paper is that repeated weighting can be applied to very complex estimation problems, although it still has its limitations.

Keywords: Repeated weighting, Census, Data integration, Combining registers and surveys, Detailed estimation, IPF, Empty cell problem

Contents

1. Introduction	4
2. Repeated weighting	5
2.1 Conditions and assumptions	5
2.2 Steps in repeated weighting	6
2.3 Adjustment of the weights	8
2.4 Adjustment of table estimates	9
3. Implementation problems	9
3.1 Computational problems	9
3.2 Edits	10
3.3 Zero cell problem	10
3.4 Conflicting marginal totals	13
3.5 Order dependence	14
4. Summary and conclusions	15
References	16
Appendix A. Repeated weighting	18
Appendix B. A special case of the conflicting marginal total problem	20
Appendix C. The epsilon method	22
Appendix D. Conflicting marginal totals	24
Appendix E. Identifying estimation problems	25
Acknowledgements	28

1. Introduction

Statistics can be compiled from a growing amount of data sources. The rich availability of data sources is both a blessing and an ordeal. On the one hand, detailed figures can be produced. On the other hand, there is a risk of inconsistent results: if a population estimate is independently estimated from different sources, the outcomes will usually be different. This numerical inconsistency is often considered undesirable, because it can confuse users. Therefore, there is a need for statistical methods that can be used for the consistent estimation from multiple data sources.

This paper considers the consistent estimation of multiple contingency tables. Contingency tables display the number of times that categories occur, for example the frequency distribution of the Dutch population by age and sex. For contingency tables consistency means that all common marginal totals in different tables have to correspond to each other. For example, in every table containing sex and occupation, the number of male managers has to be identical. Throughout this paper it will be assumed that different tables have some variables in common, otherwise consistency would not be a problem.

In general, a distinction can be made between data sources that are based on integral observation and data sources that cover a subset of the units of a population. Here, these two types of data sources will be called registers and sample surveys respectively. This distinction is just made for ease of explanation, in practise registers do not always cover the entire target population.

In the literature several methods are available for consistent estimation of contingency tables from registers and sample surveys, for example repeated weighting, repeated imputation, mass imputation and macro integration. A comparison will be given in De Waal (2014).

The current paper deals with repeated weighting only: a method developed by Statistics Netherlands and described in a number of papers, e.g. Renssen and Nieuwenbroek (1997), Nieuwenbroek et al. (2000), Renssen et al. (2001), Houbiers et al. (2003), Knottnerus and Van Duin (2006). Repeated weighting was applied in the Dutch 2001 and 2011 Censuses. An essential property of this method is that there is a clear relationship between the table estimates and the data sources from which these estimates are made: all estimates can be obtained by weighting the observed micro data.

As mentioned in Houbiers et al. (2003), there are some known complications of repeated weighting, especially for large, detailed tables.

The first complication is computational problems: long computation time and out of memory problems.

A second problem is the existence of so-called edits: relationships between different variables that need to be obeyed. Because different tables can be estimated from different sources, relationships between variables in different tables are not automatically satisfied, even if these relations are satisfied within each source.

A third complication is the zero-cell problem: estimates have to be made from a sample survey that – due to the sample mechanism – may not cover some of the categories that are known to exist in the population. For example, a sample survey that does not include 86 year-old men of an ‘other than EU’ nationality.

A fourth complication concerns estimation problems, which are caused by the impossibility to satisfy all constraints. The number of consistency constraints grows with every table that is estimated. After estimating some tables, it may be no longer possible to satisfy all constraints simultaneously. An easy example of this is that one cell value needs to be 100 and 200 at the same time, in order to be consistent with all previously estimated tables.

A fifth problem is order-dependence. In repeated weighting, tables are estimated one by one. Each table has to be consistently estimated with all previously estimated tables. The estimation order matters: a different order gives a different result. In theory, it is better to estimate all tables simultaneously. Obviously, there is no order problem in that case. In practice however, simultaneous estimation will often be infeasible: an extensive weighting model is needed, that is technically hard to solve. By estimating the tables one by one, the large weighting model is divided into a number of smaller, easier to handle, problems. All five complications are especially relevant for detailed tables, of which some of the cells are covered by only few (or even none) observations. In the literature several solutions have been proposed. The aim of this paper is to review existing solutions and to propose new solutions when necessary. Therefore, this paper is especially useful for statisticians who consider using repeated weighting for complex estimation problems.

We illustrate the problems and solutions by experiences of the Dutch 2011 Census compilation. For the application several detailed contingency tables had to be produced with socio-economic and demographic variables. The Census takes place every ten years in almost all countries. In the past in many countries traditional censuses were conducted, meaning that data were collected by the use of Census questionnaires. Presently, nine countries in the world, including the Netherlands, conduct a virtual census, based on data sources that are already available at the Statistical office. The set of tables to be produced for the 2011 Census is much more detailed than for the preceding 2001 Census. Therefore, estimation problems occurred very often. Although the solutions for the estimation problems in this paper have specifically been chosen for the Dutch 2011 Census, this paper may also be valuable for other applications in which detailed frequency tables have to be compiled from registers and sample surveys.

The organization of this paper is as follows. Section 2 explains the method of repeated weighting. In Section 3 several existing and new solutions are described for the aforementioned estimation problems. Section 4 summarizes the conclusions.

2. Repeated weighting

This section gives a brief description of the method of repeated weighting. For a more extensive description we refer to one of the papers mentioned in the introduction.

2.1 Conditions and assumptions

A first condition is that all target tables relate to the same target population, e.g. all Dutch inhabitants. This means for example that all tables necessarily have to add up to the same total.

Another condition is that weights have to be available for each record in each data source. These weights are often based on inverse inclusion probabilities of the records in their data source. In addition, these weights may also reflect response selectivity (Houbiers, 2004 and Van Duin and Snijders, 2003). For example, a weight of 12 means that a certain person in the sample represents 12 persons, of whom 11 are not selected in the sample. The weights of all register units are one, since it is assumed that registers cover the entire target population (see above).

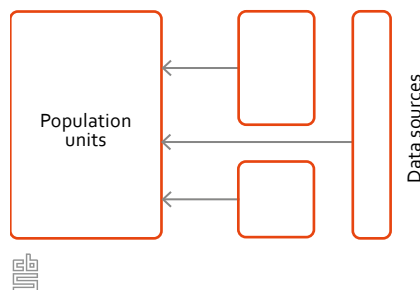
Further, it will be assumed that, within each data source, there are no obvious errors, in particular we assume that edit rules within each data source are satisfied. For example, there cannot be any 5 year-old professors. Inconsistencies need to be resolved in a preceding data editing process, otherwise inconsistent results may occur in the final table estimates. A second assumption is that the different registers are mutually consistent. If two registers contain the same population unit, the common information about that unit has to be the same in both registers. As we will see in Section 2.3, repeated weighting will not lead to consistent results, if this assumption is not satisfied.

2.2 Steps in repeated weighting

We continue explaining repeated weighting. Roughly speaking, repeated weighting consists of three steps; the focus in this paper will be on the last step.

In the first step, data linkage, all data sources are linked at unit level (see Figure 2.2.1). This results in a database that contains one record for each unit in the population. For more details, we refer to Bakker et al. (2014).

2.2.1 Data linkage



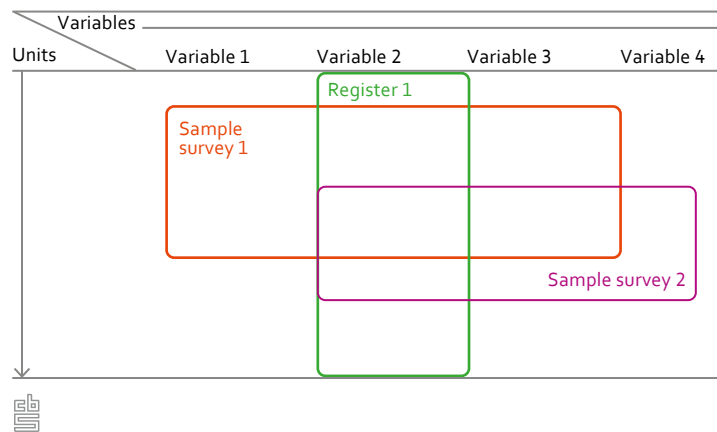
The second step, micro integration, aims at reducing inconsistencies at unit level. If a person's occupation is 'manager' according to one data source and 'clerical support worker' according to another source, a correction needs to be made to resolve the inconsistency. Usually, predefined decision rules are used for this purpose, see for example Bakker (2011).

A motivation of this step will be presented later in this section.

The third step of repeated weighting is the actual estimation of target tables. In order to estimate each table as accurately as possible, each estimate will be based on the largest possible number of records. For the estimation of tables so-called data blocks are compiled. A data block consists of all records that have a maximum set of variables in common. Each table is compiled from one data block. Therefore, all variables that are included in one table have to be available in one data block. For example, if a table needs to be compiled with 'sex', 'occupation' and 'current activity status' there has to be at least one data block that contains all these three variables.

In order to obtain accurate estimates, it is important that each data block is filled with as much as records as possible. To make this clear, we consider an example that is illustrated in Figure 2.2.2 below.

2.2.2 Schematic view of data blocks



Data blocks that are covered by one data source, have to be filled from that source. For example, in the illustration above, a data block consisting of Variable 1, 2 and 3, is covered by Sample survey 1. That data block needs to be filled with all records of that data source.

Data blocks that are covered by more than one data source need to be filled with all records of all these data sources. In the illustration above, a data block, consisting of the Variables 2 and 3 is covered by Sample surveys 1 and 2. In order to make use of as much as information as possible, the data block has to be filled with all records that appear in Sample survey 1, in Sample Survey 2 or in both surveys. Finally, data blocks, that are not covered by any single data source, are filled from a combination of data sources, that together cover all variables of a data block. All common records of the different data sources are used to 'fill' the data block. In the illustration above there is no single data source that covers Variable 1 and Variable 4. To estimate tables containing Variables 1 and 4 all common records of Sample Survey 1 and 2 are used, that is: the records that belong to the same population units.

In order to derive population estimates from a data block, weights are needed: so-called starting weights. Just as for the sampling weights of each data source, starting weights correct for sampling error and nonresponse error. Starting weights are derived from the weights of the data sources. Analogous to the data source weights, all starting weights of register observations necessarily have to be one.

For the estimation of a table, one first determines the most suitable data block from which this table can be compiled. If a table can be compiled from multiple data blocks, the data block with most records is chosen, because this choice leads to the most reliable results.

A first table estimate, a so-called *initial table estimate*, can be derived by aggregation of the starting weights. That is: the value of a cell is estimated as the sum of the starting weights of all units that contribute to that cell.

These initial table estimates are not necessarily consistent. In order to achieve consistency there are two approaches. The approach used in the original repeated weighting papers is to adjust the starting weights of the individual units (see Section 2.2). Alternatively, the problem can also be solved by adjusting the initial table estimates at the macro level (see Section 2.3). It is shown in Boonstra (2006) that both techniques can lead to the same results. In the remainder of this section, both approaches will be described, because both will be used in the solutions for the problems of repeated weighting that are given in Section 3.

2.3 Adjustment of the weights

In the original approach, the calibration property of the GREG-estimator (Särndal et. al., 1992) is repeatedly applied. The calibration approach means that the weights of some sample are adjusted, so that one or more auxiliary variables in the sample become consistent with some pre-specified targets. For example, calibration to 'sex' means that the sample is (re) weighted, such that the number of males and females become consistent with their known population numbers. In this simple example there is one marginal total of one dimension. In reality the problem is often more complicated, as the estimated tables need to fit to multiple marginal totals, that can include more than one variable. Not all weights can be adjusted; the weights of register records necessarily have to remain one, since registers are assumed to be integral. Register-based tables can be directly obtained by counting from a register; there is no estimation involved. Because register based cell values cannot be adjusted, repeated weighting implicitly assumes that the different registers are already mutually consistent (as noted in Section 2.1). In a practical application, possible inconsistencies among registers have to be removed, in a preceding micro-integration process (the above-mentioned second step). As a result of calibration, for each table a set of 'corrected' weights is obtained, that can be used to derive a table estimate from a data block. The sets of corrected weights are usually different for each table.

In repeated weighting tables are estimated sequentially. Each table is compiled, such that it becomes consistent with all previously estimated tables and all registers. For each table it is determined which marginal totals it has in common with the register(s) and with all previously estimated tables. Subsequently, the table is estimated, calibrating on these common marginal totals. Example 1 shows how the calibration works for two tables.

Example 1

Two tables have to be produced:

Table 1: 'age × sex × educational attainment'

Table 2: 'age × geographic area × educational attainment'

These tables are estimated from two data sources: a register, containing 'age', 'sex' and 'geographic area' and a sample survey that covers all four variables. Two data blocks are defined that contain the variables from the register and the sample survey respectively. Table 1 is the first table to be estimated. It needs to be consistently estimated with the register. The common marginal total of Table 1 and the register is 'age x sex'. To achieve consistency, Table 1 is calibrated on this marginal total. Table 2 needs to be consistently estimated with the register and with Table 1. For that purpose Table 2 has to be calibrated on 'age × geographic area' and 'age × educational attainment', the two marginal totals Table 2 has in common with the register and Table 1.

An important property of repeated weighting is that combination of categories, that do not occur in any data source, will be estimated as zeroes in all tables. On the one hand this is a very desirable property: provided that the data sources are cleaned from errors, it precludes the possibility of nonzero counts for categories that cannot exist in practice, for example 5 year-old professors. On the other hand, as shown in Section 3.3, this property is a source of estimation problems.

A final property is that not only the table estimates can be computed, but also the variance of the estimates. We refer to Houbiers et al. (2004) for more details on the computation of the variances.

2.4 Adjustment of table estimates

As mentioned above, it has been shown in Boonstra (2006) that the process of adjusting the starting weights is equivalent to a macro integration process, in which the cell values of the different tables are adjusted.

In the macro-integration approach a constrained optimization problem is constructed. In this problem an objective function is minimized, that measures the difference between an initial table estimate and an adjusted table estimate. The minimization is under the constraint that each table is consistently estimated with all previously estimated tables and with all registers. Analogous to the calibration approach, not all cell values can be adjusted; cell values that are based on integral observation are fixed.

Several formulations of the minimization problem are possible. It has been shown in Boonstra (2006) that the repeated weighting approach of adjusting the starting weights is equivalent to a weighted least squares (WLS) macro-integration problem. A more formal description of that optimization problem will be given in Appendix A.

3. Implementation problems

The five implementation problems introduced in Section 1 are reviewed in this section. Each problem will be described, known solutions in the literature will be given and new solutions will be proposed.

3.1 Computational problems

Problem

The computation of large, detailed tables can be problematic. The computation time can be undesirably long or the computation can even be impossible due to 'out of memory' problems.

Known solutions

As mentioned in Section 2.4, the problem of achieving consistent estimates can be formulated as a weighted least squares (WLS) problem. In this problem the cell values of the tables are minimally adjusted to achieve consistency with all previously estimated tables. Besides WLS there is a number of other optimization methods that can also be used, based on different objective functions.

In the literature comparisons are made between WLS, iterative proportional fitting (IPF) and other methods, such as Newton's method and Chi-square methods, see for example Little and Wu (1991). Boonstra (2006) suggested in an unpublished paper to apply IPF instead of WLS, mainly to avoid computational problems. The IPF algorithm is a recursive, minimal adjustment method that proportionally fits sample observations to known marginal totals. The algorithm is generally attributed to Deming & Stephan (1940), who applied the method to the 1940 American census, but the method goes by many names, depending on the field and the context (Pritchard, 2009). The computational performance of IPF is better than that of WLS, because WLS requires the computation of a matrix inverse, while IPF does not. By using Bascula, the standard weighting software of Statistics Netherlands (Nieuwenbroek and

Boonstra, 2005), several detailed Dutch 2011 Census tables that could not be estimated with WLS, can easily be produced with IPF.

A second advantage of IPF over WLS is that IPF guarantees nonnegative outcomes, while WLS does not. Negative values are not allowed for frequency tables.

A major drawback of IPF, compared with WLS, is the difficulty of estimating variances of the estimates. In Boonstra (2006) an estimator of the variance is given, but it is yet unclear how well it performs.

For the Dutch 2011 Census application it has been decided to use IPF rather than WLS, mainly due to its better computational performance.

3.2 Edits

Problem

Repeated weighting does not take into account adjustments to comply with consistency rules for different variables in different tables (so-called *edits*). An example of such a rule is that the number of people who have never resided abroad cannot exceed the number of people who were born in the home country. The reason underlying this rule is that someone who has never lived abroad, has by definition been born in the home country. This rule not only applies to the overall population, but also to all possible sub-populations (e.g. 23 year-old married men).

It is assumed that prior to repeated weighting each individual data source has been edited. Thus inconsistencies do not occur within each data source. However, as different tables can be estimated from different data sources, edits of variables in different tables are not automatically satisfied.

Known solutions

Renssen et al. (2001) solve this problem by extending all tables containing a variable of a certain edit to include all other variables that appear in that edit. In our example, 'place of birth' is added to all tables that include 'ever/never resided abroad'.

Thus, all variables subject to the same edit are estimated in the same tables. This prevents violated edits, because each table is estimated from one data block and it may be assumed that all edits are satisfied within each data block, because of the micro-integration step that is carried out prior to the estimation of the tables. However, a drawback of this solution is that more detailed tables are obtained, which may easily lead to estimation problems.

Thus, repeated weighting is not very suitable for problems with many edits. For the Dutch 2011 Census application, few edits have been identified and the above mentioned solution were successfully applied. For other applications, with many relationships between variables, new solutions have to be implemented, which requires further research.

3.3 Zero cell problem

Problem

The zero cell problem, which is also known as empty cell problem, is the problem that estimates have to be made without any underlying data. It occurs if a characteristic, that is known to exist in the population, is – due to the sample mechanism – not covered by a sample survey from which the estimates are made.

Example 2

One wants to estimate the table: 'geographic area × industry × educational attainment'. The first two variables – 'geographic area' and 'industry' – are observed in a register, which shows that 34 persons live in the 'geographic area' North-Holland and work in the mining industry. However, information on 'educational attainment' is only available from a sample survey that does not cover any of these 34 people. Consequently, the educational levels of the 'mining' people from North-Holland cannot be estimated.

In the example above consistent estimation is not possible, because characteristics that do not appear in the data sources necessarily have to be estimated as zero. According to the sample survey there would be no persons that live in North-Holland and work in the mining industry. But, according to the register there are 34 such persons. If this problem occurs the entire table cannot be estimated, even the cells for which ample source information is available.

The zero cell problem can be identified in advance. It can be established whether all categories that occur in the population also exist in each sample survey.

Known solutions

Several solutions for the zero cell problem have been proposed in the literature. Three solutions will be given below.

Beckman et al. (1996) propose a solution for iterative estimation procedures, like IPF, see Subsection 3.1.2. Their solution is to break off the estimation after a number of iterations and accept the inconsistencies that remain. However, for the Census application inconsistencies are not tolerated.

Guo and Bhat (2007) propose to merge 'small' cells into fewer, higher populated cells, so that empty cells occur less often. This solution was applied for the 2001 Dutch Census (Schulte-Nordholt et al. 2004). Tables that could not be estimated, were replaced by one or more less detailed tables. For the Dutch 2011 Census, however, this was not feasible.

Firstly, as the tables to be produced are much more detailed than for the 2001 Census, the zero cell problem occurs much more often. Secondly, aggregating a large number of tables is not acceptable, because of the loss of results, which hinders comparison with other countries.

Houbiers (2004) proposed a third solution, the so-called epsilon method. This solution consists of an application of the Pseudo Bayes estimator of Bishop et al. (1975). The solution can be applied in the macro-integration approach, as explained in Section 2.4.

First of all, a so-called *initial table estimate* is compiled by aggregation of the starting weights of a data block. All characteristics that do not appear in a data source will by definition have a zero value in all initial estimates. The zero cell problem arises because initial estimates of zero necessarily remain zero in the adjusted tables, which is a consequence of the multiplicative correction factors that are used¹⁾, see e.g. formula (A.5) in Appendix A. The epsilon method means that the initial estimates of zero are replaced by some small nonzero 'ghost' value. The micro data are not adjusted. The replacement of the zeroes enhances the applicability of the method, because initial cell estimates of zeros are fixed in the estimation, while the small ghost values can be adjusted, if necessary.

The epsilon method is a technical solution for the zero cell problem. Some examples are shown in Appendix C.1.

¹⁾ In Section 3.2 it is proposed to apply IPF instead of WLS. Switching from WLS to IPF does not solve the zero cell problem, because both estimation techniques have the property that all initial zeroes are preserved in the final estimates.

However, a major drawback is that a nonzero count may be obtained for certain combinations of categories of variables that do not appear in a sample survey. For *sampling zeroes* this is not a problem. Sampling zeroes are estimates for which there is no *a priori* reason why these particular combinations of categories do not exist. *Structural zeroes* are more problematic. Structural zeroes are particular combinations of categories that cannot exist, e.g. married children. If the epsilon method is applied, it is not guaranteed that structural zeros are preserved. Thus, by using the epsilon method, estimation problems are solved, but implausible results can be obtained. A possible solution is to apply the epsilon method to sampling zeroes only. However, in practise this solution will often be not useful, because of the impossibility of distinguishing the sampling zeroes from the structural zeroes. It is often not known which combination of characteristics can occur in practise and which combinations cannot. This information could only be retrieved from a register, but the problem is that a register is unavailable for the particular cells for which the problem occurs. Another drawback of the epsilon method is that the epsilon values are somewhat artificial. Nevertheless, cell estimates that are not based on a minimum number of observations can be left out from a publication.

A further drawback of the epsilon method is that the connection between the micro data and the cell estimates is lost. By using the epsilon method it is no longer possible to obtain weights for the micro data that can be used to derive the aggregated cell counts from the micro data. This was however not a problem for the Dutch 2011 Census, since corrected micro-data weights were not required.

For the Dutch 2011 Census an extension of the epsilon method was applied. The extension prevents implausible results to a large extent. More details on the implementation issues are given in Appendix C.2. The extension of the epsilon method is described in the following subsection.

New solutions

In this subsection a solution is presented for the problem of the implausible nonzero estimates.

As mentioned in the last section, the epsilon method solves the zero cell problem, but as a side effect, implausible results may be obtained. In particular, implausible nonzero estimates can be yielded for characteristics that do not appear in any of the data sources.

The solution consists of estimating certain additional, low-dimensional auxiliary tables. These tables are not very detailed; they typically involve one or two variables and have to be chosen such that they do not involve any sampling zeros.

The auxiliary tables are estimated in advance of all other tables. Because of the absence of sampling zeroes the zero cell problem will not occur in the estimation. The original repeated weighting method is used to estimate the auxiliary tables, i.e. not using the epsilon method. As a consequence all structural zeroes are preserved in the estimates. After the auxiliary tables are estimated, all target tables follow. As target tables have to be estimated consistently with the auxiliary tables, there will be no deviation from the data sources at the low dimensional level of the auxiliary tables. Moreover, the problem of illegal nonzero results, i.e. a nonzero estimate for characteristics that cannot occur, will not occur for cells that are covered by an auxiliary table.

For example, an auxiliary tables 'education x age', will not contain any 5 years-old professors in that table, as such persons do not appear in the data sources. In addition, the use of the auxiliary table also prevents the occurrence of 5 year-old professors in all target tables, as these target tables have to be estimated consistently with all auxiliary tables.

In the implementation of the method the most important question is which auxiliary tables have to be estimated. A first requirement is that all cells that necessarily have to be zero

have to be covered by at least one of the auxiliary tables. Otherwise, a zero final estimate cannot be guaranteed. Further, the choice of the auxiliary tables is a matter of trade-off between plausibility and applicability. Estimating a great deal of auxiliary tables enhances the plausibility of the results, because of the preservation of the low-dimensional distributions that are estimated in those tables. On the other hand, estimating many auxiliary tables increases the risk of estimation problems, because each additional table imposes certain constraints on all following tables to be estimated. More details on implementation issues are given in Appendix C.3.

3.4 Conflicting marginal totals

Problem

Another problem arising in the estimation process is ‘conflicting marginal totals’. Each table estimated imposes certain constraints on all subsequently estimated tables. When a certain number of tables have been estimated, it may become impossible to estimate a new table consistently with all previously estimated tables. Appendix D.1 shows an example. Unlike the zero cell problem, the problem of conflicting marginal totals cannot be anticipated.

There is an especially high risk of conflicting marginal totals in case of two (or more) tables with a large number of common variables, as these tables have a large number of common marginal totals. Although it may be possible to estimate each table in isolation, these tables may not be estimated in sequence, because of the many consistency constraints they impose on each other.

The problem has also been recognized in the literature. Cox (2003) pointed out that, by using IPF, see Subsection 3.1.2, for three or higher dimensional tables, it can happen that for a given set of marginal totals, no table exists that fits these totals. This problem can even occur if all marginal totals are consistent, meaning that these totals imply the same aggregates, like the same grand-total, the same two-dimensional totals, the same three dimensional totals and so on. Or in other words: that all the marginal totals of the marginal totals are the same. In Subsection 3.1.2 it is proposed to use IPF. The original repeated weighting method, however, is equivalent to WLS. By using WLS estimation problems occur less often, but negative cell estimates are obtained instead. Negative cell values are however not acceptable either. Thus, for the occurrence of the problem of conflicting marginal totals, it does not matter whether WLS is used, or the alternative estimation technique IPF.

Known solutions

Estimation problems often occur due to the presence of hidden edits. In that case the solution proposed in Section 3.2.2 can be applied to solve the problem. In other cases the epsilon method may solve the problem, see Appendix B for an example.

There is however not a generic solution for the problem of conflicting marginal totals. The only way to tackle this problem is to prevent it. One way of doing this is to estimate less detailed tables, if the problem of conflicting marginal totals occurs. This is the approach used in the Dutch 2001 Census less (Schulte-Nordholt et al., 2004). A drawback however is that it leads to a loss of information.

A second solution for the problem is to estimate the tables in a different order. The order chosen for the 2011 Census was based on a plan produced by ‘trial and error’.

A third way to prevent estimation problems, proposed by Houbiers (2003), is to merge tables with a large number of common marginal totals. Instead of estimating several tables with overlapping variables (e.g. Tables A, B and C), one table is estimated (e.g. Table D), that contains the union of the variables of the original tables (Tables A, B and C).

If the Tables A, B and C are estimated in sequence, many consistency constraints need to be fulfilled. Table B has to fit to the marginal totals it shares with Table A. Table C has to be consistently estimated with Tables A and B. In estimating Table C, the marginal totals the table has in common with Tables A and B are fixed. Thus, the degree of freedom is reduced. It may even be impossible to obtain a consistent estimate.

Estimating the combined table, Table D, at once means that it does not have to fit to the marginal totals of the earlier estimated Tables A, B and C. Thus, estimation problems may be avoided. The tables A, B and C can be derived by aggregating Table D.

Example 3

Suppose that

Table 1: 'industry × geographic area × educational attainment'

cannot be estimated. It is not possible to find estimates that fit all of the marginal totals of two previously estimated tables:

Table 2: 'industry × geographic area × occupation'

and

Table 3: 'geographic area × educational attainment × occupation'

Instead of estimating Tables 1, 2 and 3, it may be possible to estimate the combined table:

Table 4: 'industry × geographic area × educational attainment × occupation'

consisting of the union of the variables in Tables 1, 2 and 3. Thereafter, Tables 1, 2 and 3 can be obtained by aggregating over Table 4.

One drawback of merging tables is that it results in more detailed tables that may be more difficult to estimate. Technical problems may also occur, for example out-of-memory-problems, slow computation etc. Therefore, merging should only be applied when strictly necessary.

Another drawback is that the problem of conflicting marginal totals only becomes clear, after some of the tables that cause the problem have already been estimated. If it is decided to merge tables, these earlier estimates have to be discarded.

Merging was applied twice in the Dutch 2011 Census. In both cases, the problem of conflicting marginal totals has been solved successfully. For other applications additional solutions may be needed.

New solutions

In Appendix E a simple method is presented to identify the occurrence of an estimation problem before the estimation of a particular table, but after the estimation of all earlier tables in the sequence. To the best knowledge of the author this method has not been mentioned somewhere else in the literature. The method does not have to be applied for each table to be estimated, its main contribution is that it helps to understand why certain estimation problem occur.

3.5 Order dependence

Problem

As mentioned in the introduction, the repeated weighting results are order dependent: the order of estimation matters for the results. Besides that ambiguous results are not desirable as such, there is also a relationship between the quality of the estimates and the order of

estimation. Tables that are estimated in the beginning of the process do not have to satisfy as much as consistency constraints as tables that are estimated late in the process. Therefore the first estimates will not deviate as much from the data sources as the last estimates.

Known solutions

The order dependence can be avoided by using the so called ‘splitting up’ method, see Houbiers (2004). This means that all lower-dimensional marginal totals of all tables are estimated, before the tables themselves are estimated. If, for example, a three dimensional table is to be estimated, first all one and two dimensional marginal totals have to be estimated. A disadvantage is that a large amount of marginal totals is obtained for high-dimensional tables. The estimation of these tables easily leads to estimation problems: in particular the problem of the inconsistent marginal totals will occur, see Section 3.4. For the Dutch 2011 Census, the application of the splitting up method would result in estimating thousands of marginal totals. Estimating all those marginal totals will certainly lead to many estimation problems. Therefore it has been decided not to apply the ‘original’ splitting up method, but a variant that is described below, in which the order-dependence of the estimation results has been partially accepted.

Instead of estimating all marginal totals of a tables, one could alternatively apply this technique to the low-dimensional marginal totals only. One could, for example, estimate all one and two dimensional marginal total, but not all three and higher dimensional marginal totals.

The result of this ‘partial’ splitting up method would be that the order-dependence is solved for the low dimensional marginal totals, but not for the higher dimensional crossings. This solution was applied to the 2011 Dutch Census. Note, that this partial application of splitting up method coincides with the technique of the auxiliary tables, which also results in the estimation of low-dimensional marginal totals, see also Section 3.3.3.

Because order-dependence cannot be easily solved within repeated weighting, it would be worthwhile to study alternative methods, of which macro-integration (Mushkudiani et al., 2012) is a promising example.

4. Summary and conclusions

Statistics Netherlands has developed the method of repeated weighting for the consistent estimation of frequency tables. These tables are estimated from one or more data sources: registers and sample surveys. An important application is the Census, but it can be more generally applied to other areas, in which consistent frequency tables have to be produced. From the literature it is known that estimation problems may occur, especially for detailed tables. In this paper several solutions have been reviewed. Pros and cons of existing solutions are given and new solutions are proposed. We summarise three problems with a solution. A first problem is the zero-cell problem. In the literature the epsilon method has been proposed as a solution. A drawback however is that implausible nonzero estimates are obtained for structural zeros: categories that intrinsically cannot exist. Here, we propose a new extension to the epsilon method that largely prevents those implausible nonzero estimates.

A second problem is that relationships between different variables need to be obeyed, so-called edits. This problem can be solved by extending tables with additional variables. However, from a practical point of view this solution imposes a new problem: more detailed tables are obtained that are more difficult or even impossible to estimate.

A similar complication occurs when solving another estimation problem: the problem of conflicting marginal totals (third problem). This problem can be solved by merging tables. But, as a side effect, more detailed tables are obtained, which are also more difficult, or even impossible, to estimate.

A fourth problem is that computational problems easily occur for detailed tables. These problems can be solved by switching the estimation technique from weighted least squares to iterative proportional fitting.

All estimation problems have been solved in the 2011 Dutch Census. Although the Census 2011 has shown that repeated weighting can be applied for the estimation of very detailed tables, there are some drawbacks that may not be easily solved in other applications. First, repeated weighting is not appropriate for problems with many edits between the different variables.

Second, the solutions for the estimation problems are not easy to apply. It is not a recipe that can be easily followed up: it requires some preparation. For example, the estimation order has to be determined beforehand. For the 2011 Census the estimation order has been based on a plan produced after long experimenting with 'trial and error'.

Third, the results of repeated weighting are order dependent: tables are estimated in sequence and a different order yields different results. Besides that ambiguous results are not desirable as such, there is also a relationship between the quality of the estimates and the order of estimation. It can be expected that tables that are estimated late in the process will deviate more from the data sources than tables that are estimated in the beginning. Houbiers (2004) proposed a splitting-up method that avoids the order-dependence, but that solution is not appropriate for detailed tables.

Hence, it would be worthwhile to study alternative methods, of which macro-integration (Mushkudiani et al., 2012) is a promising example. For a comparison of different methods we refer to De Waal (2014).

Summarizing, the contribution of this paper is to make repeated weighting applicable to complex estimation problems. Some methods have been proposed to cope with estimation problems that occur in practical applications. There are still some other problems left for further research.

References

- Bakker, B.F.M. (2011). *Micro-integration*. Statistical Methods (201108). Statistics Netherlands, The Hague/Heerlen. <http://www.cbs.nl/NR/rdonlyres/DE0239B4-39C6-4D88-A2BF-21DB3038B97C/0/2011x3708.pdf>.
- Bakker, B.F.M., J. van Rooijen and L. van Toor (2014). The System of social statistical data sets of Statistics Netherlands: an integral approach to the production of register-based social statistics, *Statistical Journal of the IAOS*, forthcoming September.

Beckman, R.J., K.A. Baggerly & M.D. McKay (1996). Creating synthetic baseline populations. *Transportation Research Part A*, 30, 415–429.

Bishop, Y., S. Fienberg & P. Holland (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

Boonstra H. (2006). Calibration of tables of estimates. Unpublished Report, Statistics Netherlands.

Cox L.H. (2003). On properties of multi-dimensional statistical tables. *Journal of Statistical Planning and Inference*, 117, 251–273.

Deming W. and F. Stephan (1940). On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal totals are Known. *Annals of Mathematical Statistics*, Vol. 11, No. 4, p. 427–444.

Duin, C. van & V. Snijders (2003). Simulation Studies on Repeated Weighting. Discussion Paper 03008. Statistics Netherlands. <http://www.cbs.nl/NR/rdonlyres/203C85C6-7075-47A0-97BA-A3B748D393FE/0/Discussionpaper03008.pdf>.

European Commission (2008). Regulation (EC) No 763/2008 of the European Parliament and of the Council of 9 July 2008 on population and housing censuses. *Official Journal of the European Union*, L218, 14–20.

Guo, J. Y. and C. R. Bhat (2007). Population synthesis for microsimulating travel behaviour. *Transportation Research Record*, 2014, 92–101.

Houbiers, M., P. Knottnerus, A.H. Kroese, R.H. Renssen & V. Snijders (2003). Estimating consistent table sets: position paper on repeated weighting. Discussion Paper 03005. Statistics Netherlands. <http://www.cbs.nl/NR/rdonlyres/6C31D31C-831F-41E5-8A94-7F321297ADB8/0/discussionpaper03005.pdf>.

Houbiers M. (2002). Lege cellen bij herhaald wegen. Internal report, Statistics Netherlands, Voorburg (in Dutch).

Houbiers M. (2004). Towards a social statistical database and unified estimates at Statistics Netherlands. *Journal of Official Statistics*, 20, No. 1, 2004, 55–75.

Knottnerus P. & C. van Duin (2006). Variances in Repeated Weighting with an Application to the Dutch Labour Force Survey. *Journal of Official Statistics*, Vol.22, No. 3, 2006, 565–584.

Little, R.J.A. & M.M. Wu (1991). Models for Contingency Tables with Known Marginal totals when Target and Sampled Populations Differ. *Journal of the American Statistical Association*, 86, 87–95.

Mushkudiani N., J. Pannekoek, J.A. Daalmans (2012). Macro-integration techniques with applications to census tables and labour market statistics. Statistics Netherlands. <http://www.cbs.nl/NR/rdonlyres/AD653253-647D-4FFD-AFC4-67E2BDE602EE/0/201201x10pub.pdf>.

Nieuwenbroek, N.J. & H. Boonstra (2005). Bascula 4.0 Reference manual. Statistics Netherlands.

Nieuwenbroek, N.J., R.H. Renssen & L. Hofman (2000). Towards a generalized weighting system. In: *Proceedings, Second International Conference on Establishment Surveys*, American Statistical Association, Alexandria VA.

Pritchard D.R. (2009). Synthesizing Agents and relationships for land use/ transportation modelling, University of Toronto.

Renssen, R.H. & N.J. Nieuwenbroek (1997). Aligning Estimates for Common Variables in two or more Sample Surveys. *Journal of the American Statistical Association*, 90, 368–374.

Renssen, R.H., A.H. Kroese & A.J. Willeboordse (2001). Aligning Estimates by Repeated Weighting. Research paper, Statistics Netherlands.

Särndal, C.E., B. Swensson and J. Wretman (1992). Model assisted survey sampling, Springer Verlag, New York.

Schulte-Nordholt E., M. Hartgers and R. Gircour (eds.) (2004). The Dutch Virtual Census of 2001. Analysis and Methodology. Statistics Netherlands, Voorburg/Den Haag.

Waal, T. de (2014), General approaches to combining administrative data and Surveys (forthcoming), Statistics Netherlands.

Appendix A.

Repeated weighting

In this appendix, repeated weighting is described in a more formal fashion, following Houbiers (2004).

The aim of the repeated weighting estimator (RW-estimator) is to estimate the P cells of a frequency table Y_1, \dots, Y_P . The estimates are made from a dataset, of which initial weights w_i are available for all n records. Each record contributes to one cell of the table. A dichotomous variable y_{ip} will be used which is one if record i contributes to cell p and zero otherwise.

A simple population estimator is given by

$$\hat{t}_y^w = \sum_{i=1}^n w_i \vec{y}_i \quad (\text{A.1})$$

where \vec{y}_i is a P -vector, containing the elements y_{ip} for $p = 1, \dots, P$.

A special case is the Horvitz-Thompson estimator, in which the weights w_i are the inverse of the inclusion probabilities.

The population estimator \hat{t}_y^w is independent of all other tables and registers and is not necessarily consistent with other tables. To realize consistency, a population estimate needs to be calibrated on all marginal totals that the table has in common with all registers and with all previously estimated tables. These marginal totals are denoted by the J -vector \vec{r} .

There is a relationship between the cells of a table and its marginal totals: a marginal total is a collapsed table that is obtained by summing along one or more dimensions. Each cell contributes to a specific marginal total or it does not. The relation between the P cells and the J marginal totals is expressed in an $(J \times P)$ -matrix \mathbf{L} . An element l_{jp} is 1 if cell p of the target table contributes to marginal total j and zero otherwise.

A table estimate \hat{t}_y is consistent if it satisfies

$$\mathbf{L} \hat{t}_y = \vec{r} \quad (\text{A.2})$$

The repeated weighting estimator is defined by

$$\hat{t}_y^{RW} = \hat{\mathbf{T}} \mathbf{L} (\mathbf{L} \hat{\mathbf{T}} \mathbf{L}')^{-1} \vec{r} \quad (\text{A.3})$$

where $\hat{\mathbf{T}} = \text{diag}(\hat{t}_y^w)$. It can easily be seen that the estimator \hat{t}_y^{RW} fulfils (A.2), meaning it is consistent with respect to all known marginal totals. The formula in (A.3) can also be written as

$$\hat{t}_y^{RW} = \hat{\mathbf{T}} \hat{\mathbf{c}} \quad (\text{A.4})$$

where $\hat{\mathbf{c}}$ is a P -vector defined by $\mathbf{L} (\mathbf{L} \hat{\mathbf{T}} \mathbf{L}')^{-1} \vec{r}$. For the p th element of \hat{t}_y^{RW} it follows that

$$\left(\hat{t}_y^{RW} \right)_p = \left(\hat{t}_y^w \right)_p c_p \quad (\text{A.5})$$

which states that each repeated weighting estimator can be written as a product of an inconsistent population estimator \hat{t}_y^w and a correction factor for achieving consistency. It implies that the repeated weighting estimator $\left(\hat{t}_y^{RW}\right)_p$ is zero, if $\left(\hat{t}_y^w\right)_p$ is zero, meaning that initial estimates of zero are not adapted.

In Boonstra (2006) it is shown that the repeated weighting estimate is the solution of a minimal adjustment problem

$$\text{Min}_{t_y^*} (t_y^* - \hat{t}_y^w)' \hat{T}^{-1} (t_y^* - \hat{t}_y^w) \quad (\text{A.6})$$

under the constraints

$$L\hat{t}_y = \vec{r} \quad (\text{A.7})$$

a so-called weighted least squares problem (WLS).

First, an inconsistent estimate \hat{t}_y^w is derived from the microdata in some straightforward way (see formula A.1). Then, a consistent table estimate is searched for, that is in some way as close possible to \hat{t}_y^w .

Repeated weighting is a method that starts with input at the micro level, i.e. with initial weights w_i and produces output at the macro level, i.e. the table estimate \hat{t}_y^w . Alternatively, it can also be formulated as a method which leads to output at the micro level. Knottnerus and van Duin (2006) show that a table estimate \hat{t}_y^{RW} can be translated into corrected weights w_i^* at the micro level, in such a way that the table \hat{t}_y^{RW} can be derived from these weights, i.e. so that:

$$\hat{t}_y^{RW} = \sum_{i=1}^n w_i^* \vec{y}_i \quad (\text{A.8})$$

From this formula it follows that repeated weighting estimates will be zero for cells not covered by any data source observations.

Appendix B.

A special case of the conflicting marginal total problem

Although the epsilon method is not a generic solution for the 'conflicting marginal total problem', described in Section 3.3.3, in some case it does solve the problem. An example of such a case is given below.

Example B.1

Our aim is to estimate two tables:

Table I: 'citizenship × geographic area × educational attainment'

Table II: 'citizenship × sex × educational attainment'

For simplicity, we only consider table cells for the 'citizenship' category Oceania. One register is available, which contains the integral data on 'citizenship', 'geographic area' and 'sex'. The crossing of all four variables is available from a survey. In this simplified example we assume that there are two categories of 'geographic area': north and south, two categories of 'sex': male and female and two categories of 'educational attainment': low and high. Table B.1 and B.2 below show the register and the survey.

Table B.1

Citizenship	Sex	Geo.area	Count
Oceania	Male	North	5
Oceania	Female	North	1
Oceania	Female	South	4

Table B.2

Citizenship	Sex	Geo.area	Education	Count	Weight
Oceania	Male	North	Low	1	4
Oceania	Female	North	High	1	2
Oceania	Female	South	High	1	4

There is no zero cell problem, according to the definition in Section 3.3, because all register entries are present in the survey.

First, repeated weighting is applied to estimate Table I. Table I has to be consistent with the marginal total 'citizenship × geographic area' derived from the register. Thus, there have to be 6 people from Oceania in the north and 4 in the south. The initial survey weights are already consistent with that marginal total and therefore do not need to be adjusted. The estimated Table I is shown in Table B.3.



Table B.3

Citizenship	Geo. area	Education	Count
Oceania	North	Low	4
Oceania	North	High	2
Oceania	South	High	4

Our next aim is to estimate Table II: 'citizenship × sex × educational attainment'. Table II has to be consistent with the marginal totals: 'citizenship × educational attainment' of Table I and 'citizenship × sex' of the register. However, it is not possible to satisfy both constraints simultaneously.

From Table I it follows that there have to be four persons with a low educational level. From the register it can be seen that there are five males. In the survey there is one male and one person having a low educational level, both being the same. Table B.4 shows that it is impossible to fit both marginal totals: one survey record cannot be made consistent with two different marginal totals.

Table B.4

Education-> Sex	Low	High	Marginal total register
Male	1 Survey record		5
Female		2 Survey records	5
Marginal total Table A	4	6	

The distinctive feature in the example above is that there would not be any estimation problem if the survey would contain a highly educated male person. In this particular case the problem can be solved by using the epsilon method in Section 3.3.2. Again we would like to stress that, the epsilon method is not a generic solution: it does not always solve the problem of the inconsistent marginal totals.

Appendix C.

The epsilon method

C.1 Examples

First we return to Example 2 and illustrate that the epsilon method solves the zero cell problem. In Example 2, all initial estimates of the educational levels of the ‘mining workers’ from North-Holland are zero. These have to be adjusted such that their sum becomes 34. Repeated weighting does not work, because it is impossible to adjust an initial estimate of zero. The problem is solved by replacing each initial zero value by some artificial nonzero value ‘epsilon’. Contrary to the zeroes, the epsilons can be adjusted. Hence, it is made possible to calibrate to a total of 34 persons.

We also illustrate that the epsilon method solves the problem in Appendix B.1.

Table C.1

Education-> Sex	Low	High	Marginal total register
Male	4	0	5
Female	0	6	5
Marginal total Table a	4	6	

Table C.1 shows the initial table estimate, together with the required marginal totals. Again, repeated weighting does not work: it is impossible to obtain a consistent table because the initial estimates of zero cannot be adjusted. The problem can be solved, by replacing the two zeroes by a nonzero value epsilon. Consequently, all four initial estimates can be changed. As a result a consistent table can be estimated. A possible solution is shown in Table C.2.

Table C.2

Education-> Sex	Low	High	Marginal total register
Male	3	2	5
Female	1	4	5
Marginal total Table a	4	6	

C.2 Implementation issues of the epsilon method

In the implementation of the epsilon method two further questions have to be answered:

1. To which zero cells the epsilon method has to be applied?
2. What is the best value for epsilon?

In Houbiers (2002) it is advised to use as few cells with an epsilon as possible. The epsilons are somewhat artificial and their influence should not be too large.

However, in the Dutch 2011 Census the method is applied on detailed tables that contain more zero than nonzero estimates. It is difficult to determine which of the many zero cells should at least be replaced by epsilon, to avoid estimation problems. A possible solution would be to apply the epsilon method on sampling zeroes only, but because of the difficulty in distinguishing the sampling zeroes from the structural zeroes, it was decided to apply the epsilon method on each empty cell.

Houbiers (2002) states that the value of epsilon should not be taken too large. Again, the reason is to reduce the influence of the artificial epsilons as much as possible. A larger epsilon normally leads to a larger estimate. For the estimation of the Dutch 2011 Census, all empty cells are replaced by the value one for simplicity reasons. The question whether a better result is obtained if other values of epsilon are applied is open for further research.

Two measures have been implemented to reduce the influence of the arbitrarily chosen epsilon: the estimation of the auxiliary low-dimensional tables, as explained in the Sections 3.3.3 and C.3, and the publication strategy: in the 2011 Census Statistics Netherlands only publishes cell estimates based on a minimum number of observations.

C.3 Implementation of the auxiliary tables to the 2011 Dutch Census

In Section 3.3.2. it is explained that the epsilon method solves the zero cell problem.

A major disadvantage is that implausible nonzero estimates can be obtained. In the following subsection a solution has been proposed: estimating low dimensional auxiliary tables.

One aspect of the implementation of that solution is the choice of the auxiliary tables.

Estimating many auxiliary tables leads to the most plausible results, but enhances the risk of estimation problems.

For the Dutch 2011 Census application it has been decided to use one and two-dimensional auxiliary tables. Each auxiliary table includes at least one of the variables 'educational attainment' or 'occupation'. These are the only two Census variables for which no register has been used. The following one and two dimensional tables have been estimated:

- 'educational attainment';
- 'occupation';
- 'educational attainment × occupation' ;
- 'educational attainment × [register variable]';
- 'occupation × [register variable]'

For the symbol '[register variable]' all register variables can be filled in, one at a time, for example: 'sex', 'geographic area' and 'industry'.

The choice of the auxiliary tables turned out well: all the required tables have been estimated and a comparison with other publications (labour force surveys) showed that plausible results have been obtained.

As mentioned in Subsection 3.3.3, auxiliary tables have to be chosen, such that they do not include any sampling zeros. However, in some of the aforementioned one and two-dimensional auxiliary tables sampling zeros were present. For these auxiliary tables a two-step procedure was applied.

First, the proposed auxiliary table has been replaced by a less detailed auxiliary table. The less detailed table is obtained by aggregating along the variable(s) in the originally proposed table. For example, 'age' × 'occupation', where 'age' is measured in one year classes, can be replaced by a table in which 'age' is classified into five year classes.

After estimating an 'aggregated auxiliary table', the originally proposed auxiliary table is estimated, using to the same method as for all original Census tables. That is, by using the epsilon method and the IPF estimation method. This second step is needed to ensure that the estimates do not deviate too much from the data sources at the level of the one and two dimensional crossings. After this second step, all target census tables are estimated.

Another practical aspect of the auxiliary tables is the choice of the estimation method. In this paper two estimation methods are considered: WLS and IPF. In the original repeated weighting method WLS is used. However, in Section 3.1.2 it is proposed to use IPF, mainly to avoid computation problems for detailed tables. Since auxiliary tables are not very detailed, both methods can be used for their estimation. For the Dutch Census application WLS has been used for the estimation of the auxiliary tables, whereas IPF is used for all target tables. The motivation for applying WLS for the auxiliary tables is that it offers the possibility of estimating variances, whereas IPF does not. So for the Dutch census variances were computed for the low dimensional crossings in the auxiliary tables only.

Appendix D.

Conflicting marginal totals

The example below illustrates the problem of conflicting marginal totals.

D.1 Example

One wants to estimate the table 'citizenship × industry × educational attainment'. 'Citizenship' and 'industry' are observed in a register, 'educational attainment' comes from a survey. The register contains:

- 10 persons from Oceania;
- 51 persons working in the mining industry

and both groups include four persons from Oceania that work in the mining industry. The following marginal totals have been derived from the previously estimated tables.

Table D.1

Citizenship	Education	Count
Oceania	Low	1
Oceania	High	9

Table D.2

Industry	Education	Count
Mining	Low	49
Mining	High	2

These marginal totals are obviously consistent with the register.

We will show that it is impossible to estimate the table, such that it is consistent with the marginal totals of the previously estimated tables and the marginal totals that can be derived from the register.

Firstly, we consider the lowly educated people: there is one person from Oceania and 49 ‘mining’ persons with a low educational level. This implies that the combination Oceania, mining industry & low educational level can occur once at most.

Secondly, we consider the highly educated people: nine persons from Oceania and two mining workers have a high educational level. It follows therefore that the combination: Oceania, mining industry & high educational level can occur twice at most.

By combining both results, it can be seen that the combination Oceania & mining industry can occur three times at most; there cannot be more than two highly educated people and one lowly educated person. This contradicts results from the register which states that there are four ‘mining’ persons from Oceania. Thus, it will be impossible to satisfy all the required conditions, no matter what data source is used.

Appendix E.

Identifying estimation problems

In Section 3.4.1 it is stated that estimation problems can be caused by ‘conflicting marginal totals’. It is however difficult to understand why such estimation problems occur. In this section it is explained how estimation problems can be identified in advance. Thereafter it will be shown that estimation problems are an indicator for the presence of hidden edits.

Identifying estimation problems

Below some criterion will be introduced to identify the problem of conflicting marginal totals, as explained in Section 3.4. If some marginal total exceeds some derived upper bound, the problem of conflicting marginal totals is present.

First of all, an upper bound can be derived for each cell of a table. Because frequency counts cannot be negatively valued, each cell cannot attain any value higher than each of the marginal totals to which it contributes. Therefore, an upper bound for some cell is the lowest

marginal total to which it contributes. The upper bounds on the individual table cells can be aggregated to an upper bound for a marginal total. If there is any marginal total that exceeds its derived upper bound, the set of marginal totals is infeasible and estimation problems will inevitably occur.

Example

One wants to estimate the table: 'resided abroad × place of birth × educational attainment'.

- 'Resided abroad' contains the categories: ever resided abroad (Ever) and never resided abroad (Never);
- 'Place of birth' contains the categories: born in reported country (Rep. country) and born abroad (Abroad).
- 'Educational attainment' involves: low educational level (Low) and high Educational level (High).

The following marginal totals apply to the table:

Table E.1

Resided abroad	Place of birth	Count
Ever	Rep. Country	1000
Ever	Abroad	2000
Never	Rep. Country	7000
Never	Abroad	0

Table E.2

Resided abroad	Education	Count
Ever	Low	2800
Ever	High	200
Never	Low	5000
Never	High	2000

Table E.3

Place of birth	Education	Count
Rep. Country	Low	6200
Rep. Country	High	1800
Abroad	Low	1600
Abroad	High	400

From Table E.1 it can be seen that all persons, who have never resided abroad, are born in the reporting country, which is completely obvious from a logical point of view.

Table E.4 displays all the table cells. This table also shows the upper bounds that are implied by the three marginal totals and the minimum value of these three upper bounds.

Table E.4

Resided abroad	Place of birth	Education	Upper bound implied by marginal total			
			1	2	3	min
Ever	Rep. Country	Low	1000	2800	6200	1000
Ever	Rep. Country	High	1000	200	1800	200
Ever	Abroad	Low	2000	2800	1600	1600
Ever	Abroad	High	2000	200	400	200
Never	Rep. Country	Low	7000	5000	6200	5000
Never	Rep. Country	High	7000	2000	1800	1800
Never	Abroad	Low	0	5000	1600	0
Never	Abroad	High	0	2000	400	0

Upper bounds for each marginal total can be computed from the minimum upper bounds of each cell, by aggregating along the cells of Table E.4. In the tables E.5, E.6, E.7 the actual marginal totals are compared with their derived upper bounds.

Table E.5

Resided abroad	Place of birth	Upper bound	Actual value
Ever	Rep. Country	1200	1000
Ever	Abroad	1800	2000
Never	Rep. Country	6800	7000
Never	Abroad	0	0

Table E.6

Resided abroad	Education	Upper bound	Actual value
Ever	Low	2600	2800
Ever	High	400	200
Never	Low	5000	5000
Never	High	1800	2000

Table E.7

Place of birth	Education	Upper bound	Actual value
Rep. Country	Low	6000	6200
Rep. Country	High	2000	1800
Abroad	Low	1600	1600
Abroad	High	200	400

Each of the tables E.5, E.6 and E.7 show that two marginal totals exceed their upper bounds. Thus, it can be concluded that the marginal totals are infeasible and that estimation problems will occur, regardless of the data that are used for the estimation of the table.

A further inspection of E.5, E.6 and E.7 can lead to the identification of hidden edits. For instance, in Table E.6 it can be seen that the marginal total: high educational level & never resided abroad is inconsistent with its derived upper bound.

Of the people with a high educational level, Table E.2 and E.3 show that there are 2,000 who never resided abroad and 1,800 who were born in the reporting country. This contradicts Table E.1, from which it can be seen that all people who never resided abroad are born in the reporting country. From this result the hidden edit: 'place of birth' = abroad then 'resided abroad' = ever can be revealed.

The problem arises because the three marginal totals are based on different sources. The combined marginal totals in E.2 and E.3 do not take account for the 'hidden relation' between 'resided abroad' and 'place of birth' in E.1.

In Section 3.2 it is explained how hidden edits can be dealt with in the context of repeated weighting. This solution consists of extending each table that contains variables of some edit with all other variables that appear in that edit. In our example: tables containing 'resided abroad' are extended with 'place of birth' and vice versa.

Acknowledgements

The author would like to thank Jeroen Pannekoek, Arnout van Delden, Nino Mushkudiani, Sander Scholtus, Bart Bakker, Eric Schulte-Nordholt, Jantien van Zeijl, Frank Linder, Sue Westerman, Marieke Bijl-Wageveld and Harm-Jan Boonstra for their valuable comments on earlier versions of this paper.

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 cell	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.

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 contact form: www.cbs.nl/information

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