United Nations ECE/CES/GE.41/2013/22

Economic and Social Council

Distr.: General
3 September 2013

Original: English

Economic Commission for Europe

Conference of European Statisticians

Group of Experts on Population and Housing Censuses

Fifteenth Meeting
Geneva, 30 September – 3 October 2013
Item 2 of the provisional agenda
Census methodology

A new micro-macro method for estimating Dutch census tables[1]

Note by the Statistics Netherlands

Summary

In European countries different sources are used to compile the census tables: Census questionnaires, registers and 'recycled' sample surveys. Statistics Netherlands uses a combination of registers and 'recycled' Labour Force Surveys. The basis for the Dutch Census of 2011 is the Central Population Register. Since information on educational attainment and occupation is lacking in the available registers, Statistics Netherlands also has to rely on the Labour Force Survey (LFS). Consequently, not all Dutch census tables can be counted. For 23 of the 60 Census tables estimation techniques have to be applied. Statistics Netherlands has developed an advanced estimation method: repeated weighting. Repeated weighted was successfully applied for the Census 2001, but due to sheer increase of detail in the tables to be estimated, the method does not work for the detailed tables of the Census of 2011. The estimation is especially hampered by the many "zero cells", which are sample zeroes but no population zeroes. For the estimation of the Census of 2011 a new additional method has been developed, the so-called micro-macro method. The new method solves some of the estimation problems of repeated weighting, in particular the problem of the zero cells. The new method has been successfully applied for the 2011 Census: all required tables have been estimated.

[1] Prepared by Jacco Daalmans.

I. Introduction

1. Statistics Netherlands uses registers and 'recycled' surveys for the estimation of the 2011 Census tables. The backbone is the Central Population Register, which covers the whole population. The combined registers contain all the Census variables demanded, except educational attainment and occupation. Therefore the Labour force survey is used as an additional source. Each Labour Force Survey contains educational attainment, occupation and many other Census variables, but is available for 100,000 people only, out of a population of about 16,7 million. Labour force surveys of the three year period mid 2009 – mid 2012 are used to increase the coverage to about 300,000 people. The term "recycled survey" is used because the Census is not the main goal of the survey.

2. The simultaneous use of registers in combination with surveys is considered by National Statistical Institutes (NSIs) as an approved way to improve the quality of estimates (Houbiers, 2004).

3. All tables that do not contain educational attainment and occupation can be easily counted from the register. The focus of this paper is on the tables that contain educational attainment, occupation or both. These tables have to be estimated rather than counted. For the 2011 Census 42 tables had to be estimated. These tables are parts of hypercubes that have been defined by Eurostat. Statistics Netherlands has chosen to estimate subsets of hypercubes, rather than full hypercubes, because many computational problems can be expected when estimating the highly detailed full hypercubes.

4. The census tables need to be estimated consistently. Consistency means that all common margins of different tables have to be identical. For example, for all tables that contain age group and occupation, the number of armed forces between 25 and 29 years old has to be the exactly same.

5. Statistics Netherlands has developed the method of repeating weighting for consistently estimating a set of tables from multiple data sources (Renssen and Nieuwenbroek (1997), Nieuwenbroek et al. (2000), Renssen et al. (2001), Houbiers et al. (2003), Knottnerus and Van Duin (2006)). Unfortunately, this method cannot always be applied. In some cases estimation problems occur. The estimation is especially hampered by the "zero cell" problem, the problem that the sample is too small for the required level of detail. It may happen that, due to random mechanism of the sample, some cells are accidently not covered by the survey, so that the method of repeated weighting cannot be used.

6. Hence, there is a need for a method with a broader scope of applicability. In this paper we describe extensions of repeated weighting that solve the estimation problems experienced by repeated weighting. The new method is especially designed for the detailed tables of the 2011 Census, that include a great deal of zero valued cells.

7. In Section 2 the method of repeated weighting is described. Section 3 deals with the estimation problems of repeated weighting. In Section 4 solutions for the estimation problems are described. Section 5 concludes.

II. Repeated weighting

A. Outline

8. The technique of repeated weighting is designed to correct numerical inconsistencies, that arise from survey errors, not to correct for highly biased estimates. Repeated weighting is essentially applied for cosmetic purposes.

9. The method repeatedly applies the Generalized regression estimator (GREG-estimator). The GREG-estimation is a so-called calibration method. It (re)weights a sample such that one or more auxiliary variables in the sample become consistent with pre-specified targets. For example, when estimating the number of armed forces from some sample, it (re)weights the sample so that the number of males and females in the sample become consistent with their known population numbers.

10. With repeated weighting, a table is estimated from some survey, such that it becomes consistent with:

(a) all previously estimated tables;

(b) all relevant registers.

11. For each table it is determined which margins it has in common with the register(s) and with all previously estimated tables. Subsequently, the table is estimated, calibrating on these common margins. As a result a numerically consistent table estimate is obtained.

1. Example 1

12. Two tables are estimated:

Table 1: Age × Sex × Educational attainment

Table 2: Age × Geographic area × Educational attainment

13. A register is available, that contains Age, Sex and Geographic area. Educational attainment is available from a survey. Because Educational attainment appears in Table 1 and 2, both tables need to be estimated from that survey. To achieve consistency, Table 1 has to be calibrated on the margin:

Age × Sex (from the register).

14. Table 2 has to be consistently estimated with the margins:

Age × Geographic area (from the register) and

Age × Educational attainment (from Table 1).

B. Properties

15. Repeated weighting is a method that starts with input at the micro level, i.e. with initial survey weights and produces output at the macro level, i.e. consistent table estimates.

16. It is assumed that survey weights are available for each record in the survey. If the weights are inverse inclusion probabilities, the well-known Horvitz-Thompson estimator is obtained. In practise, slightly adjusted weights are used: Horvitz-Thompson weights are determined first and these are corrected for nonresponse bias and sampling fluctuation (Houbiers, 2004 and Van Duin and Snijders, 2003).

17. By using the initial weights an initial table estimate can be obtained. That table estimate does not necessarily have to be consistent with other tables and with register(s). To achieve consistency, the initial table estimates have to be adjusted.

18. It can be derived from Houbiers (2004) that a consistent table estimate can be obtained by multiplying each initial cell estimate by some cell dependent correction factor. That is:

Final estimate = Initial estimate * correction factor.

19. The correction factors of repeated weighting can be derived by solving a minimal adjustment problem (Boonstra, 2006). A consistent table estimate is searched for that is in some pre-defined way as close as possible to the initial table estimate. In repeated weighting quadratic differences are minimized. That method is known as Weighted Least Squares (WLS). However, the literature mentions some alternative choices, of which iterative proportional fitting (IPF) is especially interesting, because estimation problems occur less frequently. We will come back to that issue in Section 3.

20. Because of the multiplicative nature of correction, initial estimates of zero cannot be adjusted. This will later turn out to be an important property. We will see that is an important source of estimation problems.

21. Further, it follows that if some specific cell is not covered by the survey, the repeated weighting estimate for that cell will be zero. This is a very desirable property: it is the main reason why weighting is favoured over imputation. According to Kooiman (1998), imputation models are never sufficiently rich to account for all significant data patterns and may easily lead to oddities in the estimates, like nonzero estimates for phenomena that not occur in any of the data sources. As explained above this problem cannot occur in case of weighting.

22. Another property of repeated weighting is that the results are order dependent: a different result is obtained if tables are estimated in some different order. Houbiers (2004) mentions that this order-problem can be prevented by using the so-called “splitting-up” method. For the detailed 2011 Census tables the splitting up method is not a workable option, because it implies that thousands of tables need to be estimated, which poses a high risk of estimation problems.

23. Finally, it should be mentioned that the variance of the repeated weighting estimator can be computed (Knottnerus & van Duin, 2006): an important property for the Census application, since variances are required for some of the Census tables.

III. Estimation problems

24. In this section four estimation problems will be identified that have been experienced in the implementation of repeated weighting: 1) The zero cell problem; 2) Inconsistencies between source data and margins; 3) Inconsistent margins and 4) Computational problems.

A. The zero cell problem

25. 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 some characteristic, that is known to exist in the population, is due to sample fluctuation not covered by any of the surveys from which the estimates are made.

1. Example 2

26. One wants to estimate the table: "Geographic area × Industry × Educational attainment". Geographic area and Industry are observed in a register. From the register it is known that there are 34 persons living in the geographic area "North-Holland" that work in the mining industry. However, information on Educational attainment is only available in a survey, that does not cover any of the aforementioned 34 people. Consequently, no source is available for the estimation of the educational levels of the 'mining' people from North-Holland.

27. The repeated weighting method cannot deal with the problem of zero cells. The method cannot even be applied when only one zero cell is found in the table to be estimated.

28. In the example above all initial estimates of the educational levels of the 'mining workers' living in North-Holland are zero. Those initial estimates have to be adjusted such that the sum over all educational levels becomes 34. In Section 2 it was already explained that initial estimates of zero cannot be adjusted. Because multiplicative weighting is used, it is impossible to fit the initial estimates of zero to the required total of 34. As a consequence the entire table cannot be estimated, even the cells of the table for which ample source information is available.

29. Due to zero cell problems, of all 42 census tables only a few can be estimated by using the original repeated weighting method.

30. Bishop et al. (1975) treated a similar problem in the context of fitting log linear models. They applied a Pseudo-Bayes estimator to solve the problem. In Section 4.2 a similar solution will be applied to the zero cell problem of repeated weighting.

B. Inconsistencies between source data and margins

31. Another cause of estimation problems are the many consistency constraints that have to be satisfied. Each table that is estimated imposes certain constraints on all following tables. After a certain amount of tables have been estimated, it may become more and more troublesome to estimate any further tables. The following example illustrates the problem that the source data and the required margins can become inconsistent. If this problem occurs, it becomes impossible to estimate a table from the available sources in such a way that it fits all required margins.

2. Example 3

32. Our aim is to estimate two tables:

Table A: Citizenship × Geographic Area × Educational attainment

Table B: Citizenship × Sex × Educational attainment

33. 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 1 and 2 below show the register and the survey.

Table 1
Register

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

Table 2
Survey

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

34. There is no zero cell problem, because all combinations of categories from the register are present in the survey.

35. First, repeated weighting is applied to estimate Table A. Table A has to be consistent with the margin Citizenship × Geographic area of the register. 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 margin and therefore do not need to be adjusted. The estimated Table A is shown in Table 3.

Table 3
Repeated weighting outcome for Table A

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

36. Our next aim is to estimate Table B: "Citizenship × Sex × Educational attainment". Table B has to be consistent with the margins: "Citizenship × Education" of Table A and "Citizenship × Geographic Area" of the register. However, it is not possible to satisfy both constraints at the same time.

37. From Table A 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 4 shows that it is impossible to fit both margins: one survey record cannot be made consistent with two different margins.

Table 4
Margins of Table B

Education-> *Sex*	*Low*	*High*	*Margin register*
Male	1 Survey record	→	5
Female	↓	2 Survey records	5
Margin Table A	4	6	

38. There would not be any estimation problem if the survey would contain a highly educated male.

C. Inconsistent margins

39. Another estimation problem is that, after the estimation of a few tables, a set of conditions can be imposed on some new table, that can impossibly be fulfilled by reweighting. Contrary to the estimation problem of Section 3.2 the problem does not depend on the lack of survey data. The following example illustrates this problem.

1. Example 4

40. One wants to estimate the table Citizenship × Industry × Educational attainment. Citizenship and Industry are observed in a register, Educational attainment comes from a survey.

41. The register contains:

(a) 10 persons from Oceania

(b) 51 persons working in the mining industry

and both groups include 4 persons from Oceania that work in the mining industry. The following margins are derived from the previously estimated tables:

Geo. area	*Education*	*Count*
Oceania	Low	1
Oceania	High	9

Industry	*Education*	*Count*
Mining	Low	49
Mining	High	2

42. We will show that it is impossible to realize consistency with respect to these margins and the register at the same time.

43. 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.

44. 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.

45. 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 required conditions, no matter what data source is used.

46. In the problem above it is impossible to fit some table to some given set of margins. This problem can arise because the margins are derived from different sources. This problem has also been recognized in the literature. Cox (2003) pointed out that, for three or higher dimensional tables, it may happen that for some given set of margins, no table exists that fits all these margins.

D. Computational problems

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

IV. New weighting method

48. In this section the new weighting method is explained, the method that has been applied to the 2011 Census. The differences between the new method and repeated weighting are described in the four subsections below.

A. Iterative proportional fitting

49. In repeated weighting the problem of finding consistent estimates is solved by minimizing a weighted least square objective function (WLS). Here, we propose to apply a different technique: iterative proportional fitting (IPF), mainly to avoid computational problems.

50. The IPF-algorithm was designed in a time when powerful computers were not yet available. It is a very easy algorithm to apply and to understand. 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 IPF algorithm is a recursive method that proportionally fits sample observations to known marginal totals. In the application to the 2011 Census initial tables estimates are calibrated on the margins that are required to achieve consistency among different tables.

51. As mentioned above, the most important reason for switching from WLS to IPF is to avoid computational problems. By using Bascula, the standard weighting software of Statistics Netherlands (Nieuwenbroek and Boonstra, 2005), several detailed Census tables that could not be estimated with WLS could easily be solved with IPF.

52. 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. When using WLS negative values may be obtained, but in practice these rarely occur.

53. The major drawback of IPF, compared with WLS, is that it is difficult to estimate the variances of the estimates. In Boonstra (2006) an estimator of the variance is given, but it is yet unclear how well it performs. For application in case of the 2011 Census it is important that variances can be estimated, because this is a requirement for the quality report. The problem of computing variances for the 2011 Census estimates still has to be solved.

54. Analogous to WLS, initial estimates of zero are not adjusted. As a consequence, applying IPF instead of WLS does not solve the zero cell problem.

55. In the literature comparisons are made between WLS, IPF and other methods, such as Newton's method and Chi-square methods, see for example Little and Wu (1991). It appears that IPF performs slightly better than WLS.

B. The epsilon method

1. Method

56. The zero cell problem will be solved by applying the so-called epsilon method. The epsilon method is based on the Pseudo Bayes estimator of Bishop et al. (1975) and is also known as the method of 'ghost values' (Houbiers, 2004). It does not only solve the problem of the zero cells, but also the problem of inconsistency between data source and margins, as described in Section 3.2.

57. In the first step of repeated weighting an initial table is estimated based on some survey. Thereafter an estimation process follows that achieves consistent estimates. The epsilon method means that the empty cells in the initial table are filled in with some small, nonzero "ghost value". Although, the initial table estimates are adjusted, the underlying microdata remain unchanged.

58. The replacement of the initial zero table estimates enhances the applicability of the method, because initial estimates of zeroes are fixed in the estimation, while the small ghost values can be adjusted if necessary.

59. First we return to Example 2 and illustrate that the epsilon method solves the empty 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 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.

60. We also illustrate that the epsilon solves the problem stated in Example 3.

Table 5
The problem of Example 3

Education-> *Sex*	*Low*	*High*	*Margin register*
Male	4	0	5
Female	0	6	5
Margin Table a	4	6	

61. Table 5 shows the initial table estimate, together with the required margins. 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 some 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 6.

Table 6
Possible solution of Example 3

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

2. Properties

62. The major drawback of the epsilon method is that it creates a discrepancy between the microdata and the table estimates. It is possible to estimate a positive count for some characteristic that does not appear in a survey.

63. For sampling zeroes this is not a problem. Sampling zeroes are estimates for which there is no a priori reason why that particular combination of categories does not exist. Structural zeroes are more problematic. Structural zeroes are particular combinations of categories that cannot exist. For instance, in The Netherlands the combination "younger than 18" and "having a driver's license" is a structural zero. It would be a problem if a positive estimate is obtained for such a combination.

64. Repeated weighting guarantees that structural zeroes are preserved. However, when the epsilon method is applied, this no longer holds. Thus, by using the epsilon method, estimation problems are solved, but table estimates may be implausible. In Section 4.3 it will be explained how implausible results can be avoided.

65. Another drawback of the epsilon method is that the epsilon values are somewhat artificial. Nevertheless, after estimating all tables in the table set, artificial cell counts may be combined with other cells, or left out completely from a publication. Statistics Netherlands only publishes cell estimates that are based on a minimum number of observations, the remaining estimates are suppressed.

66. In the implementation of the epsilon method two further questions remain to be answered:

1) To which zero cells the epsilon method has to be applied?

2) What is the best value for epsilon?

67. In Houbiers (2002) it is advised to use as few epsilons as possible. The epsilons are somewhat artificial and their influence should not be too large. However, in the 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. Therefore, in the 2011 Census application, the epsilon method has been applied to each empty cell.

68. 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 epsilons normally leads to a larger estimate. For the 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.

69. The literature mentions some alternative methods for solving the empty cell problem. Guo and Bhat (2007) propose to merge 'small' cells into fewer, higher populated cells, so that empty cells occur less often. Beckman et al. (1996) break off the IPF procedure after reaching some number of iterations and accept the inconsistencies that remain. For the census inconsistencies are not tolerated and the design of the tables in the European Census of 2011 cannot be changed because of reason of comparison with other European countries. Therefore, the epsilon method is to be preferred to any other method.

C. Helptables

1. Method

70. As mentioned in the last section, the epsilon method solves the zero cell problem, but as a side effect, implausible results can be obtained. In particular, some implausible nonzero estimate can be yielded for some characteristic that does not appear in any of the data sources. In this section a solution is presented for this problem.

71. This solution consists of estimating certain additional, low-dimensional helptables. Helptables are not very detailed, they typically involve only one or two variables. The helptables are estimated in advance of all other tables to be estimated.

72. In estimating the helptables the original repeated weighting method is used, i.e. without epsilon method. As a consequence all structural zeroes are preserved in the estimates. Furthermore, it is possible to estimate the variances of those helptables.

73. After the helptables have been estimated, all target tables, i.e. the parts of the hypercubes as defined by Eurostat, follow. Thus, all target tables have to be consistently estimated with all helptables estimated before. As a consequence the low-dimensional distributions, that are measured in the helptables, have to be preserved in the target tables. This guarantees that there is no deviation from the data sources at some low level of detail.

74. Further, all structural zeroes that appear in any helptable will remain zero in all target tables. Thus, the problem of the implausible nonzero estimates can be solved, at the low dimensional level of the helptables. The problem still remains in higher dimensional crossings.

2. Properties

75. In the implementation of the method the most important question is which helptables have to be estimated. Below that question will be discussed.

76. A first requirement is that all cells, that necessarily have to be zero, have to be covered by at least one of the helptables. Otherwise, a zero final estimate cannot be guaranteed.

77. Further, the choice of the helptables is a matter of trade-off between plausibility and applicability. Estimating a great deal of helptables 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 helptables increases the risk of estimation problems, because each additional table imposes certain constraints on all following tables to be estimated.

78. For the 2011 Census it has been decided to use one and two-dimensional helptables. Each helptable includes at least one of the variables Educational attainment or Occupation, because these are the only two Census variables that are not observed in a register in the Netherlands. The following one and two dimensional tables have been estimated:

(a) Educational attainment

(b) Occupation

(c) Educational attainment × Occupation

(d) Educational attainment × [register variable]

(e) Occupation× [register variable]

79. For [register variable] all register variables can be filled in, one at a time, for example: Sex, Geographic area and Industry. The choice of the helptables turned out well: all required tables could be estimated and a comparison with other publications (the LFS) show that plausible results were obtained.

D. Merging tables

1. Method

80. In Section 3 four estimation problems have been identified. For three problems solutions have already been given. There is one problem left: the problem that it can be impossible to estimate some table, such that it corresponds with all required margins. This problem cannot be solved, because the margins derived from previously estimated tables and registers cannot be adjusted. The only way of dealing with this problem is to prevent it.

81. In particular, there is a high risk of estimation problems if there are two (or more) tables with a lot of common variables. Those tables also have many margins in common. Although it may be possible to estimate each table on itself, these tables may not be estimated in sequence, due to the many consistency constraints that these tables impose to each other.

82. The estimation problem can be circumvented by merging tables into one ‘large’, artificial table. Instead of estimating several tables with some overlapping variables (say Table A, B and C), one table is estimated (say Table D), that contains the cross section of the variables of the original tables (Table A, B and C). The many consistency constraints for achieving constancy among Table A, B and C do not apply. Thus, estimation problems may be avoided.

83. For example, suppose that

Table 1: Industry × Geographic area × Educational attainment

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

Table 2: Industry × Geographic area × Occupation

and

Table 3: Geographic area × Education × Occupation

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

Table 4: Industry × Geographic area × Education × Occupation

that consist of the cross section of the variables in Table 1, 2 and 3. Thereafter, Table 1, 2 and 3 can be obtained by aggregating over Table 4.

85. The problem stated in Example 4 in Section 3.3 can also be solved by merging tables. In Example 4 it was impossible to estimate the table Citizenship × Industry × Educational attainment. The solution would be to merge all tables that include those three variables into one large table.

2. Properties

86. A drawback of merging tables is that larger, detailed tables are obtained, of which the estimation may be hampered by computational problems: large computation time, out of

memory problems etc. Therefore, merging tables should only be applied when it is strictly necessary.

87. Another drawback is that the problem of inconsistent margins may become visible, after some of the tables that cause the problem already have been estimated. The estimates of those tables have to be discarded and the same holds true for all tables that are estimated after one of these tables.

88. In the application to the 2011 Census the solution of merging tables has been applied twice. In both cases, the problem of inconsistent margins has been solved successfully.

89. In one case several tables that involve 'Age' have been merged. The 95-100 year olds are relatively rare in the survey. It turned out to be impossible to satisfy all consistency constraints for this age group, when estimating tables sequentially. After merging tables the problem disappeared.

90. In the other case, tables have been merged that involve a crossing of some very detailed variables: Citizenship, Place of birth and Year of arrival. Estimation problems occurred when those tables were estimated separately, but not when those tables were merged into one big table.

91. The computation time for the two merged tables turned out to be much larger than for the subsets of hypercubes: several hours instead of several minutes. Therefore we did not merge tables, where it was not necessary, for example: estimating full hypercubes instead of subsets of them.

V. Conclusions

92. Statistics Netherlands has developed the method of repeating weighting for consistent estimating of a set of tables from multiple data sources. Repeated weighting has been successfully applied for the Census 2001, but due to the large increase of detail in the tables to be estimated, it does not suffice for the 2011 Census. In this paper four estimation problems have been identified and corresponding solutions are given, most of them based on Houbiers (2004).

93. The solutions have been implemented into an extended weighting method. For Statistics Netherlands the extended weighting method was the only alternative for estimating all required Census tables. All 42 highly detailed 2011 Census tables that cannot be directly counted from a register, have been estimated successfully. When using the original repeated weighting method only a limited number of tables can be estimated.

94. The results of the estimation process are satisfactory. The estimates from different tables are mutually consistent: all common margins are similar. Further, it is also important that the results are plausible, which turned out to be the case by comparing the estimates with other publications.

95. The method described in this paper can also be useful for other National Statistical Offices with 'recycled' surveys as a data source for estimating Census tables. The method may even be applied in other application areas in which consistent frequency tables have to be estimated from the combination of surveys and registers.

96. The most important drawback of the method is the difficulty of estimating variances. Variances can only be computed for one- or two-dimensional tables, the so-called 'helptables' of Section 4.3, but not for the higher dimensional parts of hypercubes that have to be estimated for the quality report of Census. A solution for this problem still has to be found.

97. A second ‘drawback’ is that the method is not easy to apply. It is not an recipe, that can be easily followed up: it requires a lot of preparation. For example, in Section 4.4 it is mentioned that estimation problems can be avoided by merging tables. It is not always easy to see in advance which of the tables should be merged. It is even not clear whether all estimation problems can be solved in other applications than the 2011 Census.

98. A further drawback of any repeated weighting method is the order dependence. Tables are estimated one after the other and for the results the order matters. The more tables have been estimated, the more consistency constraints apply. Therefore it can be expected that the first tables are better estimated than the last tables, in the sense that the final estimates resemble the data sources more closely. Houbiers (2004) described a so-called “splitting-up method” that avoids order-dependence, but a major disadvantage of that method is that a huge amount of tables may have to be estimated.

References

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. 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.

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.

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

Houbiers M. (2004). Towards a socials 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.

Kooiman, P. (1998). Mass imputation: Why not!? Research paper, Statistics Netherlands, Voorburg (in Dutch).

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

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.