United Nations ECE/CES/ GE.41/2017/3

Economic and Social Council

Distr.: General
12 July 2017

Original: English
English and Russian only

Economic Commission for Europe

Conference of European Statisticians

Group of Experts on Population and Housing Censuses

Nineteenth Meeting
Geneva, 4–6 October 2017
Item 2 of the provisional agenda
Innovations in census methodology and use of new data sources

Mass imputation for Census estimation

Note by Statistics Netherlands[1]

Summary

An important variable of the Population and Housing Census is the highest level of education attained. For the 2011 Census, this variable was observed from Dutch Labour Force Surveys (LFS). The combined LFS's are based on a sample survey, comprising approximately 300 000 persons in total. For the upcoming 2021 Census, Statistics Netherlands plans to use a more extensive data source, the Educational Attainment File (EAF). The EAF includes data from several registers and sample surveys and has a coverage of more than 6 million of people. Although coverage of EAF is continuously expanded, a selective part of the population is still uncovered. This document investigates the applicability of mass imputation for estimation of unknown educational levels at person level, focusing in particular on technical and methodological aspects.

[1] Prepared by Jacco Daalmans

GE.17-11632(E)
1711632

Contents

I. Introduction

1. In the Netherlands a so-called virtual Population and Housing Census is conducted (see for instance Schulte Nordholt, 2014). This means that results are produced by combining available data that are not primarily collected for the census. Register data are used as much as possible whenever these are available and of sufficient quality. Supplementary sample survey information is used for variables that are not (yet) fully available from registers.

2. An important variable of the Population and Housing Census is the highest level of education attained. This variable was taken from the LFS for the 2011 Census. Educational attainment data are however also available from the more comprehensive EAF. Recently, much effort has been spent on the (further) development of the EAF. The EAF contains data derived at a certain reference day from the Educational Archive (EA), a longitudinal database with information from several sources. Currently, educational attainment is known in the EAF for more than 10 million people out of 17 million inhabitants. Therefore it is very attractive to use this information for the upcoming 2021 Census.

3. The EA sources include registers and sample surveys. The registers include amongst others the Exam Results Register, the Central Register for Enrolment in Higher Education, see Linder et al. (2011) for more details. The amount of data that is observed from a register steadily grows, due to the continuous inclusion of new registers. Since registers have only come into existence in recent years, starting from the 80s, these do not include persons who completed their education before that time. Hence, coverage of registers is selective.

4. For the part of the Dutch population without available register data, supplemental sample survey information is included in the EAF. More in particular, the current EAF contains LFS information for several years, 2004 and upwards.

5. In addition, there is still quite a large group of persons that is neither covered by registers nor by sample survey observations (around 6 million people). Hence, deriving results for the entire target population relies on estimation.

6. Two estimation methods can be used for this purpose, weighting and mass-imputation (see e.g. De Waal, 2016).

7. Mass imputation means that an educational attainment level is filled in for each person with missing educational data. This approach leads to a rectangular data set with values for all variables and all population units. Scholtus and Pannekoek (2016) studied the suitability of mass imputation for the EAF for generic purposes.

8. An important drawback of mass imputation is that imputed values may be used for different purposes than intended. Imputed values can be mistakenly considered as observed values. A researcher who wants to study the relation between two variables may draw wrong conclusions if the imputation model does not take this relation into account. A famous example is the relation between having a dog as a pet and spending money on dog food. Using an imputation model for having a dog or not without using the amount of money spent on dog food as covariate may lead to the erroneous result that many people without a dog spend money on dog food. For the aforementioned reason it was decided that mass imputation is not appropriate for generic purposes.

9. Nevertheless, as mentioned in Scholtus and Pannekoek (2016), mass imputation can still be an appropriate method for specific applications. The Dutch virtual Census was explicitly mentioned as one of these potential applications.

10. For several reasons, mass imputation is an attractive option for Dutch Census compilation.

11. Firstly, weighting would imply that the EAF weights need to be combined with the weights of other sample surveys that are used for the Dutch census. It is unclear how this

can be done from a methodological point of view. This is also the main reason why EAF was not used for the 2011 Census.

12. Secondly, the compilation of results for certain subpopulation is easier, as this is simply a matter of counting (imputed) values. Thus, detailed census tables can be easily produced and questions with respect to the education level for certain sub populations can be answered rapidly.

13. For the Census a set of mutually consistent tables need to be compiled from the data sources. Several techniques are available to achieve numerical consistent results, like repeated weighting and macro integration, see e.g. Daalmans (2016). The application of these techniques on imputed data does not seem to be a problem. We will however not further discuss this issue in this report.

14. Statistics Netherlands currently studies the suitability of mass imputation for the compilation of Dutch Census 2021. The work is carried out as part of the Eurostat project "Improvement of the use of administrative sources" (ESS.VIP ADMIN WP6 Pilot studies and applications).

15. This report describes our first results. Firstly, we propose a mass imputation method. Secondly, we compare results of an application to 2011 data with the Census results at aggregate level.

II. EAF and other data sources

16. To test the feasibility of mass imputation a data set was constructed. This data set was derived from an EAF and enriched with other data sources.

17. Part II Section A explains the structure of the EAF. Part II Section B gives an overview of the information available in our constructed data set. Part II Section C describes census data that were used as a benchmark for our study.

A. Structure of the EAF data

18. For our study, an EAF-based data set was constructed with reference day January 1, 2011, official 'Census day'. The target population includes 13 748 724 persons who are 15 years or older, which is exactly the same number as published in the Dutch 2011 Census. The population younger than 15 years is not considered, since educational attainment for individuals younger than 15 years are imputed as 'not applicable' in the census.

19. For each person, the data set includes an EAF register observation for educational attainment, if available. If no register information is available, an observation is taken from one of the sample surveys included in the EAF, i.e. a LFS for one of the past eight years (2004 and later). If sample survey information is absent as well, no information on educational attainment is presented. A schematic overview of the data set is provided in Figure 1.

Figure 1
Schematic overview of the EAF-based data

<table>
<tr><td rowspan="2">(A)
Registerpart
N = 6 456 834</td><td>(B)
Remaining part –
LFS data unavailable
N = 6 951 418
(to be estimated)</td></tr>
<tr><td>(C)
Remaining part –
LFS data available
N = 340 472</td></tr>
</table>

20. Categories A, B and C will be used throughout this document.

21. A main distinction can be made between data with and without a register observation, called the register part (A) and remaining parts (B and C). The remaining parts can be further subdivided into parts with and without sample survey information, denoted by C and B respectively.

22. It can be seen in Figure 1 that approximately half of the target population is observed in EAF registers. Sample survey observations are available for about 5% of cases for which no register information is available.

23. In regular EAF production, the information in part C is used to estimate the educational attainment for part B. The information in part A cannot be used, because this part is selective.

24. We adopt a similar approach in our study. The important difference with regular EAF production is that imputation is applied instead of weighting. In our approach, population estimates for educational attainment are obtained by adding the observed counts for Parts A and C to the imputed counts for the persons in part B.

B. Information in the EAF data set

25. This Section summarizes the information that is included in the data set that is used for the current project. From EAF, we know for each person:

(a) Educational attainment (if available, that is: for parts A and C in Figure 1);

(b) Source for educational attainment within EAF (register, sample survey, or none);

(c) Weights for sample survey observations; obtained from EAF data. The weights are intended to make inferences about Parts B and C (“the remaining part”) from Part C.

26. The EAF-based data set was enriched with information of other data sources, that are included in the system of Social Statistical Datasets (SSD):

(a) Census variables used for Census compilation observed from registers, like age, sex and citizenship and industry of work, see Appendix A for an overview. Data on these variables are available for all 13 748 724 persons in the target population.

(b) ‘Percentile of income’.

Data on personal gross income is available for a large majority of cases.

Analogous to Scholtus and Pannekoek (2016), income percentile was converted into a categorical variable with 6 categories: five quintiles, i.e. bottom 20 percent, next 20 percent and so on, and unknown/not available.

III. Methodology

27. This Part describes the mass imputation method that will be considered in the remainder of the report.

A. Theory

28. In a previous study on mass imputation, Scholtus and Pannekoek (2016) compared two imputation methods, random hot deck imputation and logistic regression imputation.

29. These are methods that are technically appropriate for large-scale applications (millions of imputations) and that are able to deal with selectivity of observations.

30. Both methods rely on so-called auxiliary variables. Auxiliary variables are assumed to be available for the entire target population. The imputation methods exploit the association between the target variable and the auxiliary variable(s).

31. Random hot deck imputation basically means that for each so-called recipient, i.e. a record to be imputed, a donor is searched for with the same scores on all auxiliary variables. Missing values of a recipient are replaced by the corresponding values of a donor record. If multiple donors are found for one record, imputation depends on donor choice.

32. On the contrary, it may also happen that no donor can be found with exactly the same scores on all auxiliary variables; a problem that is more likely to occur if many auxiliary variables are applied. Because of this problem of random hot deck imputation, and because so-called nearest-neighbour hot deck imputation is likely to be (too) slow for imputing millions of records, Scholtus and Pannekoek (2016) concluded that logistic regression imputation is more appropriate for problems with many auxiliary variables.

33. With logistic regression, the relation between the imputation variable and the auxiliary variables is estimated by means of a logistic regression model. This model includes main effects of explanatory variables on the target variable. Because the model does not account for interaction terms, the above mentioned problem that too little data are available to fit the model is less likely to occur. A drawback is however that estimates may be less accurate. The regression approach produces for each record to be imputed probabilities that the imputation variable belongs to a certain category. These estimated probabilities are used as a basis for the assignment of categories for the imputation variable. This assignment is based on a stochastic process, meaning that different results are obtained after a repeated application of the method.

34. In standard logistic regression the target variable is assumed to have two categories. However, for the census, the target variable educational attainment, is classified according to eight categories. To solve this complication, Scholtus and Pannekoek (2016) proposed the so-called continuation ratio model, a method that was earlier described by Agresti (1990, Section 9). The continuation ratio model gives rise to a sequential process. In each step the probability for one education category is estimated by means of a standard logistic regression model. Suppose that the number of categories is denoted by C. In Step i the probability is estimated for category i ($i < C$), given the assumption that the category is not in $\{1,\ldots, i–1\}$, or in other words the probability that the category is i rather than $\{i + 1,\ldots, C\}$. As proven in Agresti (1990, Section 9) this sequential process leads to the

same results as a more complicated approach in which all probabilities are estimated at once.

35. In the logistic regression approach stratification can be applied, which means that a problem is broken down into sub problems according to the categories of one or more stratification variable(s). For example, stratification with respect to sex means that men outside the sample are imputed by using data from men and the same applies to women.

36. An advantage of stratification is that smaller problems are obtained which may be technically easier to deal with. Another advantage is that more accurate results may be obtained. Stratification is especially useful if the stratification variables are highly associated with the target variable.

37. Scholtus and Pannekoek (2016) applied a continuation level model to estimate educational attainment according to three categories (Low, Middle, High). These are different categories than in the census, where eight categories are defined. A conclusion of their application is that logistic regression does not yield very accurate results at micro level, but that results are more accurate at macro level.

38. Another important conclusion is that results of a multi-dimensional table, in which education level is broken down by other variable(s) can be accurately estimated, provided that the other variable(s) is/are included in the regression model. Thus, one can conclude that all variables that are relevant for the Dutch census need to be incorporated in the regression model, or more precisely, at least all variables that appear in the same tables as educational attainment.

B. Model specification

39. A first choice that needs to be made in the application of the model is the choice of the variables used for stratification. As mentioned in Part II, several variables are available; a variety of variables that are published in the census and income. It was explained in Part III Section A that stratification variables should preferably be highly associated with educational attainment. To determine the degree of association with educational attainment we will use Cramer's V measure below. This measure gives a value in the range from 0 to 1; zero means no association, one means maximal association.

40. The results in Table 1 show that income has the largest association with educational attainment. Hence, that variable was chosen as stratification variable.

Table 1
Cramer's V
(results are shown in decreasing order of association)

Variable	*Cramer's V*
Income	0.184
Industry / branch of economic activity (IND)	0.177
Current activity status (CAS)	0.159
Status in Employment (SIE)	0.151
Age (AGE)	0.121
Sex (SEX)	0.116
Location place of work (LPW)	0.108
Country Place of birth (POB)	0.098
Country of citizenship (COC)	0.067
Year of arrival in the country (YAE)	0.067

Variable	*Cramer's V*
Household status (HST)	0.056
Locality / Size of locality (LOC)	0.048
Place of usual residence / geographical area (GEO)	0.032
Place of usual residence one year prior to the census (ROY)	0.020

41. A further choice that needs to be made is which variables to choose as auxiliary variables in the regression model. It was decided to include all census variable as auxiliary variables, since it was already mentioned in Part III Section A, that accurate results of a breakdown of education by other variables can only be obtained for variables that are included as auxiliary variables in the imputation model.

42. A last issue is whether to include weights when fitting a regression model. It was decided to take the EAF-weights into account, because the weighted data can be assumed to be more representative than unweighted data. The weights correct, amongst others, for the fact that certain persons have higher probabilities of selection in a survey than others.

IV. Results at population level

43. A first conclusion is that the logistic regression approach of Part III was successfully applied for the estimation of the missing educational levels. This confirms the result in Scholtus and Pannekoek (2016) that imputation of 6 951 418 records is not a problem from a technical point of view.

Table 2
Educational attainment levels; census versus imputed EAF

Education	*Census*	*%*	*EAF-imputed EAF*	*%*
1	223 688	1.6	328 166	2.4
2	1 150 028	8.4	1 231 546	9.0
3	3 424 182	24.9	3 314 753	24.1
4	4 765 748	34.7	5 121 639	37.3
5	390 840	2.8	491 166	3.6
6	3 544 570	25.8	3 231 535	23.5
7	65 169	0.5	29 919	0.2
Unknown	184 498	1.3	0	0.0
Total	**13 748 724**		**13 748 724**	

44. Table 2 compares the Census results and the fully imputed EAF at population level.

45. The results in Table 2 are noticeable different. Finding full explanations for the differences is beyond the scope of the current project; this will be done in the sequel of the project.

46. Because differences in results for the two different parts of the population, parts A (EAF registers) and B & C (EAF imputations), may appear for different reasons, we will divide the results in Table 2 into these two parts.

A. Category A (register part of EAF)

47. This Section focuses on part A in Figure 1, the "register" part of the EAF.

48. In the approach proposed in this report, educational levels of part A are derived by directly counting from the registers.

Table 3
Education levels; EAF-registers

Education	*EAF-part A*	*%*
1	145 053	2.2
2	478 204	7.4
3	1 315 198	20.4
4	2 513 370	38.9
5	148 500	2.3
6	1 854 141	28.7
7	2 368	0.0
Unknown	0	0.0
Total	**6 456 834**	

49. Most remarkable in the register totals as shown in Table 3 is that the highest level of educational attainment is hardly observed in the EAF-registers.

50. This also seems to explain the relatively low occurrence of educational category 7 for the entire Dutch population in Table 2, when compared to official Census results.

51. Table 4 below compares relative occurrence of reported education levels for a group of 182 775 persons whose data are reported both in an EAF-register and in the LFS's used for census compilation. Differences between those register-based results and the LFS observations for the same people occur due to the use of different sources and measurement errors therein.

Table 4
Percentage distribution of education levels, unweighted, selection of Part A (N= 182 775)

Education	*LFS's Census*	*EAF Registers*
1	0.8	0.5
2	6.7	4.2
3	24.4	19.6
4	35.1	39.2
5	2.0	2.8
6	30.5	33.2
7	0.5	0.5

NOTE: Category 8 (unknown) is ignored, i.e. these records do not count for the total.

52. For the same group of people, differences in results are remarkably large. In general, educational levels reported in the EAF registers are higher than in the LFS's that were used for census estimation. From this, one can conclude that there is a severe effect of measurement error.

53. One explanation is related to persons who are observed within EAF both in a register and in a LFS-survey. For these persons the highest educational level from both sources is stored in EAF. If a register displays lower educational attainment than a LFS survey, the LFS survey value is selected to mark the highest education level, although that value will still be considered a register value in the EAF. So, what happens here, is that LFS sample survey observations are used to correct register based observations. The above-mentioned also explains why educational category 7 occurs more often in Table 4 than in Table 3.

54. The 0.5% occurrence of the highest educational attainment category can be expected to be primarily originated through LFS observations that are used to correct EAF register-based observations, as it was already observed that register observations hardly exist at all for Category 7.

B. Categories B&C (part of the EAF without register observations)

55. In this Section we present results for the Categories B and C of the population; the "remaining part" of the EAF, i.e. the part for which no register information is available.

56. We compare the imputation-based EAF results with a Census benchmark, derived from LFS records used for Census estimation, but only those records are considered that belong to parts B or C, i.e. records for which no register observation is available in the current EAF. These records are weighted by the same weights as the ones used for Census estimation.

Table 5
Estimates for part B&C – percentages of education categories

EDU	*EAF – Sample survey (C)*	*Mass imputation estimates (B&C)*	*Census Benchmark (B&C)*
1	1.4	2.5	2.2
2	8.2	10.3	10.3
3	25.1	27.4	27.6
4	38.6	35.8	35.3
5	5.4	4.7	3.7
6	21.0	18.9	20.4
7	0.4	0.4	0.4

57. In Table 4 it follows that for most categories mass imputation results are closer to the Census benchmark than the EAF sample survey observation that are used as a basis of estimation. This suggests that the mass imputation method can – at least partly – correct for selectivity of the sample survey observations within the EAF.

58. It can also be seen that the estimated education levels are fairly close to their census-based benchmarks, most in particular for Categories 1, 2, 3 and 7.

V. Robustness

59. We continue with a sensitivity analysis to examine the robustness of results. It is verified how sensitive imputation results are with regard to estimation order, model size and weights.

A. Estimation order

60. As explained in Part III, the imputation method estimates for each record probabilities of education categories in increasing order, starting with Category 1 and ending with Category 7. It was verified whether results are much affected if probabilities were estimated in reverse order (from 7 to 1). Theoretically, it can be expected that there is not a large effect. The results in Table 6 actually confirm this.

B. Model size

61. In the following exercise we compare the imputation results in Table 4, with results based on a smaller logistic regression model with fewer auxiliary variables. The simple model only includes age, industry and sex as auxiliary variables. It follows that results are not very sensitive with respect to a reduction of the number of explanatory variables. It can however be expected that differences in results are larger at a more detailed level, particularly in a breakdown of educational attainment by categories that are omitted in the reduced model.

C. Choice of weights

62. To estimate the probabilities that educational attainment belongs to certain categories a logistics regression model was estimated, based on weighted data. The original results are based on a model in which weights were taken from the official EAF-publications. Alternatively, inclusion weights of the LFS's can be used. Inclusion weights are calculated so that they can correct for unbalanced inclusion in a sample. The weights for the official EAF-publications are more advanced, these weights also correct for selectivity with respect to auxiliary variables and for differences between target populations for the publication year and the historic years for which sample survey observations are included in EAF. The results in Table 6 show that there are no significant differences for most educational categories.

Table 6
Percentage occurrence of education categories; Parts B and C (N = 7 289 890)

Education	*Original model*	*Reverse order*	*Smaller model*	*LFS weights*
1	2.5	2.4	2.2	2.3
2	10.3	10.3	10.3	10.4
3	27.4	27.5	27.5	27.5
4	35.8	35.7	35.9	36.0
5	4.7	4.7	4.7	4.7
6	18.9	19.0	19.0	18.7
7	0.4	0.4	0.4	0.4

63. To conclude, the results in Table 6 indicate that the model-based estimates are quite robust for changes in model setup

VI. Validation

64. In this Part, cross-validation method is applied to assess accuracy of imputations. This basically means that educational attainment is estimated for persons who are actually

observed in a sample survey. This provides opportunity to compare estimated and observed counts for educational categories.

65. Cross-validation is conducted as follows: the sample survey observations are randomly split into ten groups. For each of those ten groups educational attainment is estimated by means of a model that is estimated on the basis of the other nine groups.

Table 7
Cross validation (N=340 472)

Observed \ *Estimated*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*Total*
1	534	811	1 416	1 264	118	469	6	4 618
2	817	4 191	9 733	9 155	1 123	2 928	37	27 984
3	1 346	9 664	28 106	3 1901	3 850	10 477	125	8 5469
4	1 340	8 999	31 830	57 945	7 563	23 256	355	131 288
5	125	1 123	3 828	7 549	1 285	4 271	58	18 239
6	450	2 929	10 124	2 2930	4 298	29 746	902	71 379
7	5	25	107	355	82	878	43	1 495
Total	**4 617**	**27 742**	**85 144**	**131 099**	**18 319**	**72 025**	**1 526**	

66. Table 7 compares estimated and observed educational levels at micro level. The numbers on the diagonal correspond to correct imputations. The share of correction imputations is 36%. For 68% of cases estimates lie within one category of the observed category.

67. Fortunately, differences are much smaller at aggregate level. This can be seen from Table 8 below, which compares relative occurrence of education categories.

Table 8
Percentage occurrence of education categories in Part C (N = 340 472); unweighted

EDU	*Observed*	*Estimated*
1	1.4	1.4
2	8.2	8.1
3	25.1	25.0
4	38.6	38.5
5	5.4	5.4
6	21.0	21.2
7	0.4	0.4

68. A comparison at a more detailed level is made in Table 9. The results in that table are based on two-dimensional totals in which educational attainment is broken down by one other census variable (educational attainment x sex, or educational attainment x citizenship for example).

Table 9
Average percentage of discrepancy between estimated and observed counts*

Census variable in two-dimensional totals	*Average percentage difference*
Age (AGE)	5.6
Current activity status (CAS)	3.8
Country of citizenship (COC)	12.3
Place of usual residence/ Geographical area (GEO)	3.3
Household status (HST)	4.4
Industry / Branch of economic activity (IND)	6.6
Locality / Size of locality (LOC)	3.9
Location place of work (LPW)	4.8
Country Place of birth (POB)	6.2
Place of usual residence one year prior to the census (ROY)	12.8
Sex (SEX)	3.6
Status in Employment (SIE)	3.9
Year of arrival in the country (YAE)	7.4

* Based on all cells for which the observed count is at least 10.

69. The table shows average percentages of discrepancy between estimated and observed counts. The average over all 13 two-dimensional marginal totals is 5.8%.

70. To see what happens to the previous results if educational attainment is further specified, average discrepancy was also computed for one three-dimensional table: educational attainment x age x geographical area. The average discrepancy turned out to be 11.7% , based again on all cells with an observed count of ten or higher.

71. It is actually to be expected that average discrepancy is higher than for most of the results in Table 9, because of the higher level of detail and because the regression model that is used for estimation does not capture three-dimensional interactions.

VII. Results for a census table

72. This Part compares mass imputed EAF and Census results at the level of cells of one randomly chosen Census table, hypercube 24.2 (GEO x LPW x SEX x POB x AGE x EDU, see Appendix A for the meaning of the abbreviations). Table 10 shows averages of absolute values of relative differences between the two different results.

73. It can be seen that the differences are the largest for the cells with the smallest estimated counts. A more extensive analysis of these differences will be the topic for the sequel of the project.

Table 10
Average percentage of discrepancy between estimated and observed counts*

Estimated cell count; mass imputation	*Average relative difference – mass imputed EAF compared to 2011 Census*
1 000 ≤ 2 500	42.8%
2 500 ≤ 5 000	22.3%
5 000 ≤ 10 000	15.9%

Estimated cell count; mass imputation	*Average relative difference – mass imputed EAF compared to 2011 Census*
10 000 ≤ 25 000	11.2%
25 000 ≤ 50 000	8.2%
50 000 ≤ 100 000	9.4%
100 000 and more	9.8%

* thanks to Frank Linder.

VIII. Discussion

74. This document proposes a mass imputation approach to estimate educational attainment within the EAF. The method, based on logistic regression, takes the selectivity of observations into account. Technically, the model is suitable for the processing of millions of records. An empirical application in this document has shown that it is actually possible to estimate more than 6 million educational levels. The estimated education levels approximate Census results, at least at aggregate level. Therefore, the proposed imputation method can be deemed appropriate for Census estimation.

75. The implicit aim of the empirical application in this document is to approximate as much as possible all two-way Census totals, consisting of educational attainment and one other census variable. The objectives for future applications are not known yet, but these may be different than the implicit objectives for the current study. To meet future objectives, the specification of the imputation model in this report can be flexibly adapted. Once, the objectives are set up, it may be desirable to reconsider the model specifications. If, for example, certain two-way totals are more important than others, less important totals had better be excluded from the imputation model, because this may improve accuracy of the more importantly considered totals.

76. It is quite possible that in future census results will be required to align with previously published results from other statistics, for instance regular EAF production.

77. The currently proposed method is not appropriate for this purpose. Nevertheless, in literature extensions of our imputation method are available that can deal with “fixed” or “semi-fixed” totals that are already known from other publications, see e.g. Favre et al. (2005). It is however unclear how these methods perform, when applied to very large amounts of data. More research is needed to investigate this, but this research is not planned within the current project.

78. In the sequel of this project, mass imputation results will be more extensively compared with Census results at the detailed level of multivariate Census tables, so-called hypercubes.

79. In the last Part of this report cross-validation is applied to evaluate model performance. Conform the findings of Scholtus and Pannekoek (2015), it was found that imputation are not very accurate at micro level, but much more accurate at aggregate level. The cross validations may be further expanded to build a criterion for the suitability of publication for (aggregate) results.

References

Agresti A. (1990), Categorical Data Analysis. John Wiley & Sons, New York.

Daalmans J.A. (2016), Divide-and-Conquer solutions for estimating large consistent table sets, Discussion paper 2016–19, Statistics Netherlands. https://www.cbs.nl/en-https://www.cbs.nl/en-gb/background/2016/46/divide-and-conquer-solutions-for-estimating-large-consistent-table-setsgb/background/2016/46/divide-and-conquer-solutions-for-estimating-large-consistent-table-sets (accessed January 2017).

De Waal T. de (2016), Obtaining numerically consistent estimates from a mix of administrative data and surveys, Statistical Journal of the IAOS, 32, 231–243.

Favre, A.-C., A. Matei and Y. Tillé (2005), Calibrated Random Imputation for Qualitative Data. Journal of Statistical Planning and Inference 128, pp. 411–425.

Linder, F., D. van Roon and B. Bakker (2011), Combining Data from Administrative Sources and Sample Surveys; the Single-Variable Case. Case Study: Educational Attainment. Report for Work Package 4.2 of the ESSnet project Data Integration. https://ec.europa.eu/eurostat/cros/content/wp4-case-studies_en (accessed January 2017).

Scholtus S. and J. Pannekoek (2015), Mass-imputation of educational levels (in Dutch), Statistics Netherlands, Internal report, The Hague/Heerlen.

Schulte Nordholt, E. (2014), Dutch Census 2011, Analysis and Methodology, Statistics Netherlands, The Hague/Heerlen
https://www.cbs.nl/NR/rdonlyres/5FDCE1B4-0654-45DA-8D7E-807A0213DE66/0/2014b57pub.pdf (accessed January 2017).

Zult D., S. Scholtus (2016), The estimation of NiRWO (in Dutch), Statistics Netherlands, Internal report, The Hague/Heerlen.

Annexes

Annex I

List of variables

The following variables appear in the demographic part of the 2011 Dutch Census:

- Age (AGE)
- Current activity status (CAS)
- Country of citizenship (COC)
- Place of usual residence / Geographical area (GEO)
- Household status (HST)
- Industry / branch of economic activity (IND)
- Locality / Size of locality (LOC)
- Location place of work (LPW)
- Country / Place of birth (POB)
- Place of usual residence one year prior to the census (ROY)
- Sex (SEX)
- Status in employment (SIE)
- Year of arrival in the country (YAE).

Annex II

Code labels for educational attainment

The following variables appear in the demographic part of the 2011 Dutch Census:

Code	*Meaning*
1	No formal education
2	ISCED Level 1. Primary education
3	ISCED Level 2. Lower secondary education
4	ISCED Level 3. Upper secondary education
5	ISCED Level 4. Post-secondary non-tertiary education
6	ISCED Level 5. First stage of tertiary education
7	ISCED Level 6. Second Stage of tertiary education
8	Not stated (of the persons aged 15 years or over)