Discussion Paper

Mass imputation for census estimation

2017 | 04

Jacco Daalmans

Content

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 LFS's are based on sample surveys, comprising approximately 300,000 persons. 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 10 million people. Although coverage of EAF is continuously expanded, a selective part of the population is still uncovered. This paper investigates the applicability of mass imputation for estimation of unknown educational levels at person level, in particular, focusing on technical and methodological aspects.

Keywords

Census; Virtual Census; Mass Imputation; data processing; highest completed education level

1. Introduction

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. An important variable of the Population and Housing Census is the highest level of education attained. This variable was taken from the Dutch Labour Force Survey (LFS) for the 2011 Census. Educational attainment data are however also available from the more comprehensive Educational Attainment File (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 data base 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. 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. 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 Labour Force Survey (LFS) information for several years, 2004 and upwards. 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. Two estimation methods can be used for this purpose, weighting and mass-imputation (see e.g. De Waal, 2016). 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. 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. 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. For several reasons, mass imputation is an attractive option for Dutch Census compilation. 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. 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. 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.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 project "Improvement of the use of administrative sources" (ESS.VIP ADMIN WP6 Pilot studies and applications and has received EU funding under the grant agreement 07112.2016.004-2016.593.

Technically, the imputation method needs to be appropriate for an application to millions of records. Methodologically, the imputation method must be capable to take the selectivity of different data sources into account. This intermediate report describes our first results. The final version of this imputation method will later be described in a separate methodological report. Firstly, we propose a mass imputation method. Secondly, we compare results of an application to 2011 data with the Census results at aggregate level. A more comprehensive comparison at the level of detailed Census tables will be provided in the final report.

2. EAF and other data sources

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.

Subsection 2.1 explains the structure of the EAF. Subsection 2.2 gives an overview of the information available in our constructed data set. Subsection 2.3 describes census data that were used as a benchmark for our study. Subsection 2.4 deals with a particular problem for our data; the conversion of internal education codes to Census definitions.

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.
Subsection 2.1 explains the structure of the EAF. Subsection 2.2 gives an overview of the information available in our constructed data set. Subsection 2.3 describes census data that were used as a benchmark for our study. Subsection 2.4 deals with a particular problem for our data; the conversion of internal education codes to Census definitions.

2.1 Structure of the EAF data

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.
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 2.1.1.

2.1.1 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>

Categories A, B and C will be used throughout this paper.
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.
It can be seen in 2.1.1 that approximately half of the target population is observed in EAF registers. Sample survey observations are available for about 5 percent of cases for which no register information is available.
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.
We will 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.
For the future it can be expected that there will be a sheer increase of the share of Category A, due to the inclusion of new registers. On the one hand, this is desirable, because of the larger share of the population that is integrally observed. On the other hand, estimates may become less accurate, because the relative size of Category C may become smaller, meaning that fewer data are available as a basis for the estimates in Category B.
When constructing the data set for our application, the problem occurred that the reference day of the EAF differs from Census day. The reference day of EAF is October 1, whereas Census day is January 1. To solve this problem, EAF data were converted; the EAF for October 2011 were merged to the population frame of the 2011 Census. Of all 13,748,724 persons in the Census population frame, 13,569,189 could be matched to the EAF (98.6%). The unmatched records belong to people that were registered the Population Register at the beginning of 2011, but, due to mortality and emigration, not anymore on 1 October 2011. Because educational attainment is unavailable for that relatively small group of records, the unmatched records are assigned Category B in 2.1.1, meaning that educational levels of these people are estimated by imputation.

2.2 Information in the EAF data set

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

- Educational attainment (if available, that is: for parts A and C in 2.1.1);
- Source for educational attainment within EAF (register, sample survey, or none);

– Weights for sample survey observations; obtained from EAF data[1]. The weights are intended to make inferences about Parts B and C ("the remaining part") from Part C;

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

– 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.
– '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.

A comparison with the variables used in the study of Scholtus and Pannekoek (2016), shows the following two differences between their and our application:

– Scholtus and Pannekoek (2016) did not consider all census variables in their study, because census estimation was not their main purpose. However in their study, they included other variables from SSD, that are closely related to the census variables used in our study;
– Scholtus and Pannekoek (2016) made use of education from the Public Employment Service Register (PESR). It was mentioned, that there is a very strong association between PESR education and educational attainment. In our study we do not include PESR as an additional data source, because information from PESR is already standardly used for the determination of educational attainment in the current version of EAF.

2.3 Benchmark data

For benchmark purposes, we also use LFS data that were used for the 2011 Dutch Census compilation.

For the Census three years of LFS data around Census day were used, one-and-a-half year before and one-and-a-half year after Census day, consisting of a total number of 331,968 observations.

2.4 Towards census codes for educational attainment

The EAF data contain detailed internal education codes that differ from the codes used in the census. Therefore, it is necessary to convert the internal codes to the

[1] Weights were taken from the so-called EAF prototype, a different data set than the one used for the current study. Weights were taken from a different data set because weights are unavailable for the current 2011 data set.

International Standard Code on Education (ISCED) that is used for the census. For this purpose, a recoding scheme was developed.
To test this conversion, ISCED-based educational attainment codes were compared among 130,913 records that are observed both in the EAF-based data set (Category C in 2.1.1) and in the LFS data used for 2011 Census estimation. The meaning of the education codes 1-8 is shown in Appendix B.

2.4.1 Number of records by education category, two data sets

EAF / *Census*	1	2	3	4	5	6	7	Total
1	1898	17	0	1	0	0	0	1916
2	1	10322	41	32	8	21	0	10425
3	0	16	32052	369	6	24	0	32467
4	0	3	40	50009	1342	133	0	51527
5	0	0	2	93	5329	16	0	5440
6	0	1	0	46	3	28481	1	28532
7	0	0	0	0	0	2	556	558
8	0	10	12	18	1	7	0	48
Total	1899	10369	32147	50568	6689	28684	557	

The cross-tabulation shows that the conversion to ISCED codes was reasonably successful. For 129,647 out of 130,913 cases (98.3 %), the same ISCED-code was assigned to an EAF-record as in the data used for census compilation.
There are however also notable differences. Category 8, “unknown education level”, appears in the census data, but not in the EAF data set. This difference is not very relevant, because of the relatively low number of 48 records and because the category “unknown” is not very informative.
More importantly, there is a relatively large group of 1,342 records that are classified according to category 5 in the EAF data set, and that belongs to category 4 according to the census data. This is approximately one fifth of all records that are classified category 5 in the EAF. Comparability may improve by a correction of the recoding scheme. We leave this as a potential subject for further research.

3. Methodology

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

3.1 Theory

In a previous study on mass imputation, Scholtus and Pannekoek (2016) compared two imputation methods, random hot deck imputation and logistic regression imputation.
These are methods that are technically appropriate for large-scale applications (millions of imputations) and that are able to deal with selectivity of observations.
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).
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.
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.
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.
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.
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.
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.
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.
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.

3.2 Model specification

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 Section 2, several variables are available; a variety of variables that are published in the census and income. It was explained in Subsection 3.1 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.

3.2.1 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
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

The results in 3.2.1 show that income has the largest association with educational attainment. Hence, that variable was chosen as stratification variable.
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 Subsection 3.1, 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.
The proposed regression model consists of no fewer than thirteen auxiliary variables. As mentioned in Linder *et al.* (2011) a model with many auxiliary variables may lead to an unnecessary large variance. However, after empirical research it was also concluded that a large model can still be satisfactory.
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.

4. Results

In this section we present results of the proposed mass imputation method. Because differences in results for the parts A and B & C may appear for different reasons, the discussion of results will be subdivided into two parts.

4.1 Category A

This subsection focuses on part A in 2.1.1, the “register” part of the EAF.
In the approach proposed in this report, educational levels of part A will be derived by directly counting from the registers.
Differences between those register-based results and the census results, for the same part of population, occur due to the use of different sources and measurement errors therein.
It is important to note that the differences do not depend on imputations (these only affect Category B, the part of the population for which no information on educational attainment is available in the EAF).
One explanation for differences is that education attained in the first half of 2012 is observed in the LFS’s used for Census estimation, but not in the registers in the EAF for October 2011. The effect of this can however expected to be small.

The table in 4.1.1 compares the reported education levels for a group of 182,775 people whose data are both reported in an EAF-register[2] and in the LFS's used for census compilation.

4.1.1 Percentage distribution of education levels, unweigthed (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.

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.
One explanation for the measurement errors is that EAF registers does not measure so called NiRWO, which includes education attained at private institutions, in foreign countries or for a doctor degree, whereas the LFS's used for Census estimation do include these categories. Zult and Scholtus (2016) describe a model-based approach to estimate the impact of this. The impact on the results in 4.1.1 is however unclear, because the results of Zult and Scholtus (2016) were derived from different data.
A further 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. What happens here, is that sample survey observations are used to correct register-based observations.
Thus, it follows that, if it were true that all records from the census LFS's would be included in the current EAF, educational levels in the LFS cannot be higher than in an EAF register. In our application it is however not true that all "Census LFS records" are also contained in the EAF, but this still holds for a substantial part, 182,775 out of 331,968 cases.
In 4.1.2 below, we compare the EAF register-based totals (Category A) with the results of the 2011 Census results for the entire Dutch population (Category A+B+C). There are clear differences for all educational categories. The most remarkable difference occurs for category 7 ("Second stage of tertiary education - ISCED level 6"). The relative occurrence of this category is 0.5% for the census and 0.0% for the EAF registers (rounded). The highest level of educational attainment is hardly observed in the EAF registers. Because the EAF registers cover approximately half of the target

[2] Including the records in the EAF for which information is available from registers and from LFS.

population, it can be expected that the low report of the highest educational level affects the entire target population.

4.1.2 Education levels; census versus EAF

Education	Census	%	EAF-part A	%
1	223,688	1.6	145,053	2.2
2	1,150,028	8.4	478,204	7.4
3	3,424,182	24.9	1,315,198	20.4
4	4,765,748	34.7	2,513,370	38.9
5	390,840	2.8	148,500	2.3
6	3,544,570	25.8	1,854,141	28.7
7	65,169	0.5	2,368	0.0
unknown	184,498	1.3	0	0.0
Total	13,748,724		6,456,834	

In conclusion, we showed that the proposed EAF-based compilation method will lead to large differences with respect to the current LFS-based census. Because these differences cannot be explained from the imputation methodology, we will not attempt to solve these in the current project.

4.2 Categories B & C

In this subsection 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.
A first conclusion is that the logistic regression approach of Section 3 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.
We will now compare the imputation based results with two benchmarks: 1) benchmark obtained from the Census and 2) benchmark from EAF sample surveys.
The 'Census' benchmark is 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.
The 'EAF' benchmark is obtained from LFS observations in the current EAF. The weights from regular EAF production are applied on these data.

4.2.1 Estimates for part B&C - percentages of education categories

EDU	EAF - Sample survey (C)	Mass imputation estimates (B&C)	Census Benchmark (B&C)	EAF benchmark (B&C)
1	1.4	2.5	2.2	2.0
2	8.2	10.3	10.3	9.6
3	25.1	27.4	27.6	26.6
4	38.6	35.8	35.3	36.1
5	5.4	4.7	3.7	4.9
6	21.0	18.9	20.4	20.3
7	0.4	0.4	0.4	0.4

In 4.2.1 it can be seen that the mass imputation results are closer to the two benchmarks than to 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.
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. The larger share of Category 5 may be explained by a potential problem in the transformation of the detailed internal educational code into the ISCED-based codes that are used in the census. As mentioned in Subsection 2.4 it can be expected that Category 5 is overestimated due to this problem. However, one must also keep in mind that although Census results were produced with care, these may not be totally free of error.
On average, the mass imputation results better approximate the Census benchmark than the EAF-based benchmark. An explanation for this is that the regression model used for mass imputation explicitly models the relation between educational attainment and all census variables, whereas the weighting method for the EAF production contains a selection of census variables and a collection of other variables that are not included in the census.
We continue this subsection with a sensitivity analysis to examine the robustness of results. It is verified how sensitive results are with regard to estimation order, model size and weights.

Estimation order
As explained in Section 3, 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. This is actually confirmed by the results in 4.2.2

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

EDU	Original order	Reverse order
1	2.5	2.4
2	10.3	10.3
3	27.4	27.5
4	35.8	35.7
5	4.7	4.7
6	18.9	19.0
7	0.4	0.4

Model size

In the following exercise we compare the results in 4.2.1 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.

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

EDU	Original model	Smaller model (fewer auxiliary variables)
1	2.5	2.2
2	10.3	10.3
3	27.4	27.5
4	35.8	35.9
5	4.7	4.7
6	18.9	19.0
7	0.4	0.4

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.

Choice of weights

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.

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

EDU	Original model – (EAF weights)	Alternative weights (LFS inclusion weights)
1	2.4	2.3
2	10.3	10.4
3	27.4	27.5
4	35.7	36.0
5	4.7	4.7
6	19.1	18.7
7	0.4	0.4

The results in 4.2.4 show that there are no significant differences for most educational categories. The largest differences are observed for the categories 4 and 6. To conclude, the results in 4.2.2-4.2.4 indicate that the model-based estimates are quite robust for changes in model setup.

5. Validation

In this section, 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.
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.

5.1.1 Cross validation (N=340,472)

Obs. / Est.	1	2	3	4	5	6	7
1	534	811	1416	1264	118	469	6
2	817	4191	9733	9155	1123	2928	37
3	1346	9664	28106	31901	3850	10477	125
4	1340	8999	31830	57945	7563	23256	355
5	125	1123	3828	7549	1285	4271	58
6	450	2929	10124	22930	4298	29746	902
7	5	25	107	355	82	878	43

Obs.=observed; Est.=estimated

The table in 5.1.1 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. Fortunately, differences are much smaller at aggregate level.

This can be seen from 5.1.2 below. The observed and estimated counts correspond to the row and column totals of 5.1.1.

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

EDU	Observed		Estimated	
1	4618	(1.4%)	4617	(1.4%)
2	27984	(8.2%)	27742	(8.1%)
3	85469	(25.1%)	85144	(25.0%)
4	131288	(38.6%)	131099	(38.5%)
5	18239	(5.4%)	18319	(5.4%)
6	71379	(21.0%)	72025	(21.2%)
7	1495	(0.4%)	1526	(0.4%)

A comparison at a more detailed level is made in 5.1.3. 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).

5.1.3 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

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

It is to be expected that average discrepancy is higher than for most of the results in 5.1.3, because of the higher level of detail and because the regression model that is used for estimation does not capture three-dimensional interactions.

6. Discussion

This intermediate report 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 paper 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.

The implicit aim of the empirical application in this paper 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.

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. 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 data. More research is needed to investigate this, but this research is not planned within the current project.

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

This intermediate report has identified two reasons for discrepancy between Census results and imputed EAF: measurement error and estimation error. We showed that for our application measurement error is much more influential than estimation error. If our purpose is to assess the estimation error, it is important to separate estimation errors and measurement errors. Since estimation errors only occur for the part of population without any register information, separate results have been presented for the parts of populations with and without register information in the EAF. It is advisable to adopt a similar approach in the sequel of the project.

In the last section 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.

Acknowledgements

This paper is based on work carried out as part of the Eurostat project "Improvement of the use of administrative sources" (ESS.VIP ADMIN WP6 Pilot studies and applications). The action has received EU funding under the grant agreement 07112.2016.004-2016.593. The paper reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains

The author is grateful to Ton de Waal, Frank Linder, Eric Schulte Nordholt for useful suggestions which greatly helped to improve previous versions of this paper

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-gb/background/2016/46/divide-and-conquer-solutions-for-estimating-large-consistent-table-sets (accessed January 2017).

De Waal T. (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. URL: 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.
URL: 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.

Appendix A. 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).

Appendix B. Code labels for educational attainment

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)

Explanation of symbols

Empty cell	Figure not applicable
.	Figure is unknown, insufficiently reliable or confidential
*	Provisional figure
**	Revised provisional figure
2015–2016	2015 to 2016 inclusive
2015/2016	Average for 2015 to 2016 inclusive
2015/'16	Crop year, financial year, school year, etc., beginning in 2015 and ending in 2016
2013/'14–2015/'16	Crop year, financial year, etc., 2013/'14 to 2015/'16 inclusive

Due to rounding, some totals may not correspond to the sum of the separate figures.

Colofon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress
Statistics Netherlands, Studio BCO

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contactform: www.cbsl.nl/information