

## Calibrated imputation for multivariate categorical data

Authors	de Waal, T.; Daalmans, J.
Published in	Asta-advances in Statistical Analysis
DOI	<a href="https://doi.org/10.1007/s10182-023-00481-z">10.1007/s10182-023-00481-z</a>
Publication Date	2023
Document Version	publishersversion
Link	<a href="https://research.tilburguniversity.edu/en/publications/630ec08f-61dd-4730-926a-6b55a95066e9">https://research.tilburguniversity.edu/en/publications/630ec08f-61dd-4730-926a-6b55a95066e9</a>
Citation	de Waal, T & Daalmans, J 2023, 'Calibrated imputation for multivariate categorical data', Asta-advances in Statistical Analysis. <a href="https://doi.org/10.1007/s10182-023-00481-z">https://doi.org/10.1007/s10182-023-00481-z</a>
Download Date	2026-05-17 11:44:26
Rights	<p>General rights</p> <p>Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.</p> <ul style="list-style-type: none"> <li>- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.</li> <li>- You may not further distribute the material or use it for any profit-making activity or commercial gain</li> <li>- You may freely distribute the URL identifying the publication in the public portal"</li> </ul> <p>Take down policy</p> <p>If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.</p>



# Calibrated imputation for multivariate categorical data

Ton de Waal<sup>1,2</sup> · Jacco Daalmans<sup>1</sup>

Received: 28 November 2022 / Accepted: 15 September 2023

© The Author(s) 2023

## Abstract

Non-response is a major problem for anyone collecting and processing data. A commonly used technique to deal with missing data is imputation, where missing values are estimated and filled in into the dataset. Imputation can become challenging if the variable to be imputed has to comply with a known total. Even more challenging is the case where several variables in the same dataset need to be imputed and, in addition to known totals, logical restrictions between variables have to be satisfied. In our paper, we develop an approach for a broad class of imputation methods for multivariate categorical data such that previously published totals are preserved while logical restrictions on the data are satisfied. The developed approach can be used in combination with any imputation model that estimates imputation probabilities, i.e. the probability that imputation of a certain category for a variable in a certain unit leads to the correct value for this variable and unit.

**Keywords** Non-response · Edit rules · Fully conditional specification · Mass imputation

## 1 Introduction

Non-response is a major problem for anyone collecting and processing data, such as National Statistical Institutes (NSIs). When left untreated, non-response can lead to biased estimates or invalid results from statistical analyses. Non-response can be subdivided into item non-response, where some values from otherwise observed units are missing, and unit non-response, where entire units are not observed.

A commonly used technique to deal with missing data is imputation (see, e.g., Rubin 1987, Schafer 1997, Little and Rubin 2002, De Waal et al. 2011 and Van Buuren 2012). In imputation, missing values are estimated and filled in into the dataset. Imputation is particularly used often for item non-response. It is sometimes

---

✉ Ton de Waal  
t.dewaal@cbs.nl

<sup>1</sup> Statistics Netherlands, PO Box 24500, 2490 HA The Hague, the Netherlands

<sup>2</sup> Tilburg University, PO Box 90153, 5000 LE Tilburg, the Netherlands

also used for unit non-response, although weighting is a more common technique for correcting for unit-nonresponse.

Imputation can become challenging if the variable to be imputed has to comply with a known total. This situation occurs often if the total has been published before. Deviation from an earlier published total is deemed undesirable by many NSIs as this may lead to possible confusion among users due to conflicting results for the same phenomena. Even more challenging is the case when several variables in the same dataset need to be imputed. In addition to the known totals there can be so-called edit rules (or edits for short) that have to be satisfied by the data. An example of an edit is that a baby cannot have completed primary school.

We illustrate the problem with an example. Statistics Netherlands publishes information on the highest educational level attained. In the Netherlands, the first results on the highest educational level attained are based on weighting the so-called Education Attainment File (EAF). Later, information based on the highest educational level attained is combined with other data sources to construct a virtual population census, i.e. a population census that is mainly based on administrative data covering the entire population. In the case of the Dutch Population Census, all variables except highest educational level attained and occupation are based on integral administrative data. After construction of the virtual population census, we can break down information on the highest educational level attained into detailed groups of the population by using the background information available in the population census. To facilitate the estimation process for the virtual population census, highest educational level attained is mass imputed in the virtual population census, i.e. highest educational level attained is imputed for all population units for which no value has been observed. In that way, a complete dataset for the entire Dutch population is constructed, which can be used for multiple estimation purposes. Daalmans (2017) proposes a method for mass imputation of highest educational level attained based on logistic regression that can be used for the Dutch Population Census. However, the results for highest educational level attained based on the population census will deviate from the earlier published results based on weighting the EAF if standard imputation techniques, such as logistic regression, are used. Besides highest educational level attained, other variables, in particular occupation, have missing values and need to be imputed.

As far as we are aware, only one method has thus far been proposed in the literature that allows one to impute categorical data with missing values for multiple variables such that previously published totals are preserved and specified edits are satisfied (De Waal et al. 2017). However, that method is very time-consuming, and can only be applied to relatively small problem instances. In some cases, the method also has to “backtrack”, i.e. a previously imputed variable may need to be imputed in a different way. As noted by De Waal et al. (2017), this would lead to an even more time-consuming and extremely complicated process.

In the current paper, we propose an imputation approach that can be used for a broad class of imputation methods for multivariate categorical data such that previously published totals are preserved by the imputed data while edits are satisfied,

and that can handle much larger problem instances than the approach by De Waal et al. (2017). This imputation approach is based on adding a calibration step to standard imputation techniques. It can be used in combination with any imputation model that estimates imputation probabilities, i.e. the probability that imputation of a certain category for a variable in a certain unit leads to the correct value for this variable and unit. Examples of imputation models that estimate such imputation probabilities are multinomial models and logistic regression models.

Our proposed imputation approach generalizes an approach by Favre et al. (2005). In their approach, only one categorical variable is to be imputed subject to edits and known totals. We generalize this to the multivariate case where multiple categorical variables have to be imputed subject to edits and known totals. We achieve this by adopting a fully conditional specification approach (see Subsect. 3.2) that takes the previously published totals into account in combination with a Fellegi–Holt approach to satisfy all edits (see Fellegi and Holt 1976 and the Subsect. 3.1 of the current paper). Whereas Favre et al. (2005) approximate imputation probabilities given that known totals have to be preserved in only one way, we also examine two alternative approximations (see Sect. 3.3). In this paper, we will assume that all population units will be imputed, i.e. that mass imputation is used, and that therefore no weighting is necessary to obtain estimates for population totals.

Section 2 of this paper first discusses the approach developed by Favre et al. (2005) for a single categorical variable with missing data. Section 3 discusses our proposed generalization to multivariate categorical missing data. Section 4 describes the evaluation study that we carried out to assess the proposed generalization, while Sect. 5 examines the results of this study. Section 6 examines the estimation of imputation variance by means of a pseudo-population bootstrap approach. Section 7 concludes the paper with a short discussion.

## 2 Approach by Favre, Matei and Tillé for univariate missing data

The approach of Favre et al. (2005) for imputation of univariate missing categorical data subject to edits and known totals consists of four steps. In the first step, user-specified edits are used to find structural zeroes for the variable to be imputed, i.e. for each record in the dataset the categories that are not allowed according to the observed values in combination with the specified edits are determined. In the second step, for each record the imputation probabilities of the categories that are allowed are estimated. In the implementation of Favre et al. (2005) this is done by assuming a multinomial logistic model, taking the structural zeroes into account. In the third step, these probabilities are calibrated so that, for each category, they sum up to the corresponding known total and, for each record, to one. This is achieved by using iterative proportional fitting (IPF). In the fourth step, Cox' controlled rounding algorithm (see Cox 1987) is used to fix one of the probabilities for the allowed categories per record to one and the probabilities of the other allowed categories to zero. The category for which the probability is set to one for a certain record is imputed in that record.

**Table 1** Units to be imputed and totals per category

	Category $c_1$	Category $c_2$	Category $c_3$
Unit 1	0	*	*
Unit 2	*	*	*
Unit 3	0	*	*
Unit 4	*	*	*
Unit 5	*	0	*
Unit 6	*	*	0
Unit 7	*	*	*
Unit 8	0	*	*
Total	4	1	3

**Table 2** Imputation probabilities

	Category $c_1$	Category $c_2$	Category $c_3$
Unit 1	0.0	0.4	0.6
Unit 2	0.6	0.2	0.2
Unit 3	0.0	0.3	0.7
Unit 4	0.5	0.2	0.3
Unit 5	0.8	0.0	0.2
Unit 6	0.7	0.3	0.0
Unit 7	0.5	0.1	0.4
Unit 8	0.0	0.2	0.8

We illustrate the basic ideas of the approach by Favre et al. (2005) by means of an example. In this example one variable with three categories is to be imputed in eight units. In the first step of the approach, the observed values for the other variables are filled in into the edits to find the structural zeroes for the variable to be imputed. For convenience, we assume that this step has already been carried out for all units in the dataset.

The units to be imputed are given in Table 1. The known totals are given in the last row.

The cells with a “\*” are allowed to be imputed. The zeroes in units one, three, five, six and eight in Table 1 denote values that are not allowed to be imputed due to the specified edits, i.e. the structural zeroes. As explained above, imputation means that in each unit one of the “\*” is replaced by a one, and the other “\*” by zeroes.

The value to be imputed in a certain unit is essentially found by randomly drawing one of the allowed values using the imputation probabilities obtained from the imputation model. We assume that, taking into account that certain categories are not allowed in certain units, the imputation probabilities obtained by the imputation model are as in Table 2 (Step 2).

In Table 2, we see, for instance, that, according to the assumed imputation model, the probability that the actual value of the variable to be imputed in unit

one equals  $c_2$  is 0.4. In Step 3 of their approach, Favre et al. (2005) apply IPF to the imputation probabilities in Table 2 so that each row sums up to one and each column to the corresponding total. This step is needed to take the known totals into account. The adjusted probabilities are given in Table 3, together with the totals per category in the last row.

Using IPF to adjust the imputation probabilities actually only leads to an approximation for the exact imputation probabilities that take known totals and edits into account. In general, it is very complicated and/or time-consuming to compute the exact probabilities. We will return to this point in Subsect. 3.3.

In Step 4 of their approach Favre et al. (2005) apply Cox's controlled rounding algorithm (Cox 1987) to ensure that each record with a missing value for the target variable is imputed and the known totals are preserved. Cox' controlled rounding algorithm is a stochastic procedure. In the rounding algorithm, a rounding base  $b$  is specified. All entries in the table, including the marginal totals, are rounded to integer multiples of  $b$ . In the approach by Favre et al. (2005), the rounding base  $b$  equals 1. Since all internal entries are (adjusted) imputation probabilities that lie between zero and one, all internal entries are rounded to either zero or one. Entries that already are multiples of  $b$  are not changed by Cox' algorithm. This implies that the marginal totals, which are non-negative integers, are not changed by Cox' method. An appealing property of Cox's controlled rounding algorithm is that it preserves additivity of the table. That is, the rounded internal entries sum up to the rounded marginal totals. This implies that, since the entries in each row sum up to one, exactly one entry per row will be rounded to one, and will hence be imputed. All other entries in a row will be set to zero. Another appealing property of Cox' controlled rounding algorithm is that it is unbiased. That is, if we were to repeat the rounding algorithm an infinite number of times, the averages of the rounded cell values over the repetitions would be equal to the original unrounded cell values, i.e. the adjusted imputation probabilities given in Table 3. In other words, the categories that are imputing are drawn according to the adjusted imputation probabilities in Table 3. If we apply Cox' controlled algorithm to Table 3, we may, for instance, obtain Table 4.

**Table 3** Adjusted imputation probabilities and totals per category

	Category $c_1$	Category $c_2$	Category $c_3$
Unit 1	0.000	0.308	0.692
Unit 2	0.798	0.081	0.121
Unit 3	0.000	0.223	0.777
Unit 4	0.717	0.087	0.196
Unit 5	0.898	0.000	0.102
Unit 6	0.885	0.115	0.000
Unit 7	0.702	0.043	0.255
Unit 8	0.000	0.143	0.857
Total	4	1	3

**Table 4** Imputed units

	Category $c_1$	Category $c_2$	Category $c_3$
Unit 1	0	0	1
Unit 2	1	0	0
Unit 3	0	0	1
Unit 4	0	1	0
Unit 5	1	0	0
Unit 6	1	0	0
Unit 7	1	0	0
Unit 8	0	0	1
Totals	4	1	3

### 3 Generalization to multiple variables with missing data

Our generalization of the approach of Favre et al. (2005) to multiple categorical variables with missing data consists of two different phases:

- A start-up phase where we impute all variables for the first time. This phase serves two different purposes. First, it gives us “rough” imputations, which are later improved upon in the second phase. Second, and more importantly, the imputed dataset after this phase satisfies all edits. A dataset that satisfies all edits is a prerequisite for the method in the second phase, which achieves consistency with the totals. In the start-up phase we do not preserve known totals yet.
- The actual imputation phase where we iteratively re-impute all data that were originally missing. This phase also serves two purposes. First, in this phase we improve the imputations after the first phase. Second, in this phase we preserve known totals by calibrating the imputations to these totals.

#### 3.1 The start-up phase

In the start-up phase we use sequential imputation, i.e. we impute each record in turn and within each record we impute each variable in turn. For each record with missing values, we apply the following steps for each variable to be imputed.

1S. For the current variable to be imputed, we use a – usually rather simple – imputation model, for instance, a multinomial imputation model involving only a few auxiliary variables, to estimate imputation probabilities for its categories.

2S. For the current variable to be imputed, we derive all allowable categories, given the observed and already imputed variables in the record under consideration and the edits.

3S. We impute the current variable to be imputed by drawing categories using the imputation probabilities from Step 1S until we draw an allowable category.

Step 2S is a fundamental step as it ensures that after the start-up phase we will have a fully imputed dataset satisfying all specified edits, which we need as a

starting point for the actual imputation phase. It is also a non-trivial step, and may in fact be the most complicated step of our entire imputation approach from a technical point of view.

We first illustrate the problem with sequentially imputing data having to satisfy edits. Suppose we have a dataset with three variables: *Marital status*, *Age* and *Relation to head of household*. The possible values of *Marital status* are “Married”, “Unmarried”, “Divorced” and “Widowed”, of *Age* “< 16 years” and “≥ 16 years”, and of *Relation to head of household* “Spouse”, “Child” and “Other”. Suppose we have two user-specified edits: the first edit saying that someone who is less than 16 years cannot be married, and the second one that someone who is not married cannot be the spouse of the head of household. The first edit involves variables *Age* and *Marital status*, and the second edit variables *Marital status* and *Relation to head of household*. Now suppose that both *Marital status* and *Age* are missing in a certain record, the observed value of *Relation to head of household* in that record is “Spouse”, and that *Age* is the current variable to be imputed. Neither of the two user-specified edits involves both the observed variable, *Relation to head of household*, and the current variable to be imputed, *Age*. That is, neither of these two edits prevents us from imputing the value “< 16 years” for *Age*. However, if we were to do that, we would later notice – while trying to impute *Marital status* – that there is no value for *Marital status* that will satisfy both edits, since a person younger than 16 years cannot be married while someone who is a spouse of the head of household has to be married. So, the problem is that we have to take into accounts edits involving variables to be imputed later while imputing the current variable.

We will now sketch the procedure in Step 2S to overcome this problem and then illustrate this procedure by means of the above example. First of all, we fill in the values of the observed and already imputed variables (if any) in the record under consideration into the user-specified edits. This leads to a set of edits for the remaining variables to be imputed. The main idea of the approach in Step 2S is to eliminate all remaining variables to be imputed except the current variable to be imputed from this set of edits by means of the Fellegi–Holt elimination method (see Appendix A). When a variable is eliminated by means of the Fellegi–Holt elimination method, a new set of edits is obtained that has to be satisfied by all remaining variables to be imputed. By repeated elimination, we obtain a set of edits that only involves the current variable to be imputed. A set of edits for a single categorical variable simply defines a set of allowed values for that variable. So, from the derived set of edits for the current variable to be imputed we can immediately see the allowable values for that variable.

The Fellegi–Holt elimination method has the nice and very important property that if and only if the new set of edits for the remaining variables to be imputed obtained after elimination of a variable can be satisfied, a value for the eliminated variable exists such that the set of edits before elimination can also be satisfied (for details and proofs see Fellegi and Holt 1976 and De Waal and Quere 2003). By repeated application of this property, we find that if and only if the current variable to be imputed is imputed such that the set of edits obtained after elimination of all other variables remaining to be imputed is satisfied, all eliminated variables can be imputed such that all user-specified edits can be satisfied (again see De Waal and

Quere 2003). In turn this implies that we can impute the variables with missing values in a record sequentially, i.e. one at the time, while still ensuring that all edits can be satisfied.

In mathematical logic, the Fellegi–Holt elimination method is known as (multi-valent) resolution (see, e.g., Chandru and Hooker 1999 and Hooker 2000). Resolution can be used to check whether a set of propositions (the edits in our case) can be satisfied.

We now briefly illustrate the procedure in Step 2S by means of our example. We first introduce some notation. We denote the number of variables by  $n$ . In the case of categorical data, an edit  $k$  is usually written in so-called normal form, i.e. as a Cartesian product  $F_1^k \times F_2^k \times \dots \times F_n^k$  of non-empty sets  $F_s^k$  ( $s = 1, 2, \dots, n$ ), meaning that if for a record with values  $(v_1, v_2, \dots, v_n)$  we have  $v_s \in F_s^k$  for all  $s = 1, 2, \dots, n$ , then the record fails edit  $k$ , otherwise the record satisfies edit  $k$ .

In normal form the edit saying that someone who is less than 16 years cannot be married can be written as

$$\{\text{Married}\} \times \{< 16 \text{ years}\} \times \{\text{Spouse, Child, Other}\}, \quad (1)$$

and the edit saying that someone who is not married cannot be the spouse of the head of household as

$$\{\text{Unmarried, Divorced, Widowed}\} \times \{< 16 \text{ years, } \geq 16 \text{ years}\} \times \{\text{Spouse}\} \quad (2)$$

We fill in the value “Spouse” for *Relation to head of household* into edits (1) and (2) and obtain the edits

$$\{\text{Married}\} \times \{< 16 \text{ years}\} \quad (3)$$

and

$$\{\text{Unmarried, Divorced, Widowed}\} \times \{< 16 \text{ years, } \geq 16 \text{ years}\} \quad (4)$$

for the variables to be imputed, *Marital status* and *Age*. (For notational convenience, in edits, we do not mention variables whose values have been substituted into a set of edits nor variables that have been eliminated).

Since *Age* is the current variable to be imputed in our example, we have to eliminate variable *Marital status* from (3) and (4). We obtain the edit

$$\{< 16 \text{ years}\} \quad (5)$$

for variable *Age* (see Appendix A).

Edit (5) implies that only the value “ $\geq 16$  years” is allowed to be imputed for variable *Age*. Indeed, if we later impute the value “Married” for *Marital status*, we obtain an imputed record satisfying both user-specified edits (1) and (2).

### 3.2 The actual imputation phase

In the actual imputation phase we iteratively (re-)impute all data that were originally missing, using the other (partly imputed) data as auxiliary data. We essentially apply a fully conditional specification (FCS) approach for imputation of multivariate missing data (see, e.g., Raghunathan et al. 2001, Van Buuren and Groothuis-Oudshoorn 2011 and Rubin 2003). In FCS, one specifies a separate multivariate imputation model for each variable to be imputed. In principle, all other variables can be used as auxiliary variables in the imputation model for a certain variable. In the estimation process of the model parameters, as well as in the actual imputation process, previously imputed values of the auxiliary variables are used.

FCS is usually applied for Bayesian versions of multiple imputation, where one specifies a prior distribution for each imputation model, derives the posterior distribution given the prior distribution and the observed data, and then draws multiple imputations from the posterior distribution. In this paper, we will apply FCS in a frequentist context, i.e. without specifying prior distributions for the imputation models. Our approach is similar to the approach by Siddique and Belin (2008). However, whereas Siddique and Belin (2008) use a hot-deck imputation method without an explicit imputation model for each variable to be imputed, we will use multinomial imputation models in our simulation study.

In the actual imputation phase we start with the fully imputed dataset obtained after the start-up phase. We iteratively (re-)impute the data that were originally missing, where we impute each variable in turn. For each variable to be imputed we apply the following steps.

1I. For the current variable to be imputed, we use an imputation model, for instance, a multinomial imputation model with in principle all other variables as auxiliary data, to estimate imputation probabilities for its categories.

2I. For each record in which the value of the current variable to be imputed was originally missing, we fill in the other data in that record (either observed or imputed) into the edits. This may lead to some structural zeroes, i.e. categories that are not allowed to be imputed, in some records for the current variable to be imputed.

3I. For the current variable to be imputed, we approximate the correct imputation probabilities per record, i.e. we use the probabilities from the imputation model of Step 1I and adjust them, taking structural zeroes and – if applicable – known totals into account. For each record to be imputed, taking structural zeroes into account simply amounts to setting imputation probabilities for those structural zeroes to zero and rescaling the other imputation probabilities so they sum up to one. Subsect. 3.3 discusses how to adjust imputation probabilities so they take known totals into account.

4I. If the totals of the current variable to be imputed are not known, we use the adjusted imputation probabilities from Step 3I to draw imputations. If the totals of the current variable to be imputed are known, we use Cox' controlled random rounding algorithm to find the imputations.

We keep on iterating the above steps 1I to 4I for all variables until the distribution of the imputed data has converged to a final distribution.

Note that, in contrast to Step 2S of the start-up phase, Step 2I is quite simple. The reason why Step 2I is so simple from a technical point of view is that Step 2S of the start-up phase guarantees after the start-up phase we will have an imputed dataset satisfying all edits. Due to this, Step 2I in the actual imputation phase only has to ensure that we never impute a value for the current variable to imputed that is not allowed according to the other data in that record and the edits.

### 3.3 Adjusting the imputation probabilities

As we already mentioned in Sect. 2, using IPF to adjust the imputation probabilities leads to an approximation for the exact imputation probabilities that take edits and known totals into account. In this section we discuss two other approximations. We assume that the imputation probabilities from the posited imputation model already take edits into account.

In principle, the exact imputation probabilities taking known totals in account can be found by enumerating all possibilities. To illustrate how, we consider Table 5 below. We suppose that we want to impute a variable  $x$  with three categories ( $c_1$ ,  $c_2$  and  $c_3$ ) in four units. We assume that there are no structural zeroes, we know the totals of the categories and we have estimated imputation probabilities for the four units.

The exact adjusted imputation probabilities are given by  $\Pr(x_i = k | \text{known totals})$ , where  $x_i$  is the value of  $x$  in unit  $i$  ( $i = 1, \dots, 4$ ) and  $k = c_1, c_2$  or  $c_3$ . If we were to disregard the totals, there would be  $3 \times 3 \times 3 \times 3 = 81$  possible outcomes, and the exact imputation probabilities would be given by the values in Table 5. However, if we do take the known totals into account, there are only 12 possible outcomes. These are given in Table 6 below, together with their corresponding probabilities according to Table 5.

For instance,  $\Pr(\text{outcome} = 1) = \Pr(x_1 = c_1) \times \Pr(x_2 = c_1) \times \Pr(x_3 = c_2) \times \Pr(x_4 = c_3) \approx 0.0313$  according to Table 5. Using Table 6 we can easily calculate  $\Pr(x_i = k | \text{known totals})$  for  $i = 1, \dots, 4$  and  $k = c_1, c_2$  or  $c_3$ . For example,

$$\Pr(x_1 = c_1 | \text{known totals}) = \frac{0.0313 + 0.0078 + 0.0313 + 0.0156 + 0.0078 + 0.0156}{0.1445} \approx 0.7568,$$

where 0.1445 is the sum of all the probabilities over all 12 possible outcomes.

**Table 5** Estimated imputation probabilities and known totals

Unit	Category		
	$c_1$	$c_2$	$c_3$
1	0.50	0.25	0.25
2	0.25	0.50	0.25
3	0.25	0.50	0.25
4	0.25	0.25	0.50
Totals	2	1	1

**Table 6** Possible outcomes, given known totals

Unit	Outcome											
	1	2	3	4	5	6	7	8	9	10	11	12
1	$c_1$	$c_1$	$c_1$	$c_1$	$c_1$	$c_1$	$c_2$	$c_2$	$c_2$	$c_3$	$c_3$	$c_3$
2	$c_1$	$c_1$	$c_2$	$c_2$	$c_3$	$c_3$	$c_1$	$c_1$	$c_3$	$c_1$	$c_1$	$c_2$
3	$c_2$	$c_3$	$c_1$	$c_3$	$c_1$	$c_2$	$c_1$	$c_3$	$c_1$	$c_1$	$c_2$	$c_1$
4	$c_3$	$c_2$	$c_3$	$c_1$	$c_2$	$c_1$	$c_3$	$c_1$	$c_1$	$c_2$	$c_1$	$c_1$
Pr	0.0313	0.0078	0.0313	0.0156	0.0078	0.0156	0.0078	0.0039	0.0039	0.0039	0.0078	0.0078

In this very small example we can enumerate all possible outcomes. In general, enumeration is infeasible, and we have to resort to approximating the adjusted imputation probabilities. Besides the IPF approach as used by Favre et al. (2005), we have tested two alternative approximations in our simulation study.

The underlying idea of the alternative approximations is that in each unit in which the value of the current variable to be imputed is missing, we have to decide whether we are going to impute category  $c$  or not.

If we only consider a certain category  $c$  and assume that we can neglect the other categories, we would have to draw from a so-called Conditional Binomial (CB) distribution (see, e.g., Chen and Liu 1997 and Chen 1998). The CB distribution is closely related to a Poisson-Binomial (PB) distribution. Suppose we have a dataset  $S$  with  $|S|$  units and let  $\mathbf{Z} = (Z_1, \dots, Z_{|S|})$  be the outcomes of  $|S|$  independent Bernoulli trials with probabilities  $p_1, \dots, p_{|S|}$ . The probability of a total of  $k$  successes in the  $|S|$  trials is then given by a PB distribution, which we denote as  $\text{PB}(k; \mathbf{p}_S)$  where  $\mathbf{p}_S = (p_1, \dots, p_{|S|})$  is a vector of parameters of the PB distribution. The conditional distribution of  $\mathbf{Z}$  given that  $\sum_{i \in S} Z_i = k$  is a CB distribution, with parameters  $k$  and  $\mathbf{p}_S$ .

We will approximate the CB distribution for our situation. Suppose there are  $m_x$  units (numbered 1 to  $m_x$ ) in which the value of the variable  $x$  under consideration is missing and we have to impute category  $c$   $k_c$  times. Together these  $m_x$  units form a set  $S_x$ . We consider a unit  $i$  in which the value of the variable under consideration is missing. The exact adjusted probability for imputing category  $c$  in unit  $i$  is given by

$$p_{ic}^* = \frac{p_{ic} \times \Pr(\text{category } c \text{ is selected } (k_c - 1) \text{ times in the remaining } (m_x - 1) \text{ units})}{\Pr(\text{category } c \text{ is selected } k_c \text{ times in } m_x \text{ units})}$$

if  $k_c \neq 0$ , and  $p_{ic}^* = 0$  if  $k_c = 0$ , where the  $p_{ic}$  are the imputation probabilities without taking the known totals into account and the  $p_{ic}^*$  are the adjusted imputation probabilities that do take the known totals into account, i.e. the  $p_{ic}^*$  are the probabilities of our CB distribution. That is,

$$\begin{aligned} p_{ic}^* &= \frac{p_{ic} \text{PB}((k_c - 1); \mathbf{p}_{S_x \setminus \{i\}})}{p_{ic} \text{PB}((k_c - 1); \mathbf{p}_{S_x \setminus \{i\}}) + (1 - p_{ic}) \text{PB}(k_c; \mathbf{p}_{S_x \setminus \{i\}})} \\ &= \frac{p_{ic}}{p_{ic} + (1 - p_{ic}) \frac{\text{PB}(k_c; \mathbf{p}_{S_x \setminus \{i\}})}{\text{PB}((k_c - 1); \mathbf{p}_{S_x \setminus \{i\}})}} \end{aligned} \tag{6}$$

In this paper we use simple approximations for the PB distribution. As a reviewer pointed out, a PB distribution can quite efficiently be computed exactly by means of a Fast Fourier Transform (Hong 2013), which is implemented in the “poisbinom” package. We have not used this exact approach for two reasons. A pragmatic reason is that even with an efficient approach to calculate a PB distribution exactly, our simulation study – where we had to calculate/approximate a PB distribution millions of times (see Sect. 4) – is likely to take much computing time. Even with our simple

approximations our simulation study took several months. A more important theoretical reason is that using a CB distribution – and implicitly a PB distribution – is itself an approximation, since the assumption that we can neglect the other categories when drawing a category is not completely correct. The estimated imputation probabilities and known totals for the other categories affect the adjusted imputation probabilities for the category  $c$  under consideration. For instance, using the exact PB distribution in Eq. (6) gives  $\Pr(x_1 = c_1 | \text{known totals}) = 0.7596$  for Table 5, which differs slightly from the correct value obtained by enumeration mentioned above.

Our first simple approximation of  $PB(k_c; \mathbf{p}_{S_x \setminus \{i\}})$  is by a Poisson distribution  $\Pr(\lambda_{S_x \setminus \{i\}})$ , with  $\lambda_{S_x \setminus \{i\}} = \sum_{t \in S_x \setminus \{i\}} p_{tc}$  (see, e.g., Chen et al. 1994, Chen and Liu 1997, Chen 1998 and Chen 2000). Using this approximation when  $k_c \neq 0$ , we get  $\frac{PB(k_c; \mathbf{p}_{S_x \setminus \{i\}})}{PB((k_c-1); \mathbf{p}_{S_x \setminus \{i\}})} \approx \frac{\lambda_{S_x \setminus \{i\}}}{k_c}$ .

A second, alternative approximation of  $PB(k_c; \mathbf{p}_{S_x \setminus \{i\}})$  is by a binomial distribution with success probability given by  $p_{S_x \setminus \{i\}}(c) = \sum_{t \in S_x \setminus \{i\}} p_{tc} / |S_x \setminus \{i\}| = \sum_{t \in S_x \setminus \{i\}} p_{tc} / (m_x - 1)$  for category  $c$  (see, e.g., Chen and Liu 1997). Using this approximation when  $k_c \neq 0$ , we get  $\frac{PB(k_c; \mathbf{p}_{S_x \setminus \{i\}})}{PB((k_c-1); \mathbf{p}_{S_x \setminus \{i\}})} \approx \frac{m_x - (k_c - 1)}{k_c} \frac{p_{S_x \setminus \{i\}}(c)}{(1 - p_{S_x \setminus \{i\}}(c))}$ .

## 4 Evaluation study

### 4.1 Implementation

We implemented our code in R (see R Core Team 2020). In the start-up phase, we simply used the observed fractions of the categories of each variable to be imputed as imputation probabilities, unadjusted for known totals. For instance, if 49% of the observed people are female and 51% male, then we impute “Female” with probability 0.49 and “Male” with probability 0.51 in variable *Gender*. In the imputation phase, we have used a more complicated multinomial model. For each variable we have used all other (partly imputed) variables as auxiliary variables while estimating the imputation probabilities according to this model. We have implemented the estimation of the imputation probabilities using the function “multinom” from the R package “nnet”. This function uses neural networks to fit multinomial models given a set of auxiliary variables to the available complete data, and thus obtains estimates for the imputation probabilities for a target variable (for more details, see Ripley 2020). In our imputation approach we need to handle edits. That is, we need to substitute values into edits and we need to derive implied edits by eliminating variables (see Sect. 3.2). For this we have used the R package “editrules” (see De Jonge and Van der Loo 2012). R code for Cox’s controlled rounding algorithm was kindly provided to us by our colleague Sander Scholtus (Statistics Netherlands). We have gratefully used this code in our implementation. Our R code is available on GitHub: <https://github.com/tonwaal/Calibrated-Imputation-for-Multivariate-Categorical-Data>.

**Table 7** True totals of Educational level

Category	“0”	“1”	“2”	“3”	“4”	“5”	“9”
Total	254	578	807	1220	116	551	258

**Table 8** True totals of Occupation

Category	“1”	“2”	“3”	“4”	“5”	“6”	“7”	“8”	“9”	“999”
Total	260	336	360	251	229	34	212	147	178	1777

## 4.2 Dataset

The dataset used for our simulation study is a small part of the Dutch Population Census 2001 published by Statistics Netherlands. The dataset contains 3,784 randomly drawn individuals with 12 categorical variables. The variables and the corresponding numbers of categories are: *Gender* (2 categories), *Age* (17 categories), *Position in the household* (8 categories), *Size of the household* (6 categories), *Residential area last year* (3 categories), *Nationality* (3 categories), *Country of birth* (3 categories), *Educational level* (7 categories), *Economic status* (8 categories), *Occupation* (10 categories), *NACE code* (13 categories) and *Marital status* (4 categories). Descriptions of the categories of these variables can be found in Appendix B. Appendix C contains a description of the user-specified edits. The dataset used is complete and does not contain any missing values for any of the units. We treated this dataset as our population. In our simulation study we introduced missingness into the data and mass imputed the missing values in the population.

In our study, we focused on the variables *Educational level* and *Occupation*, since these are the only variables in the Dutch Population Census that are (partly) based on surveys, rather than on administrative data covering (almost) the entire population as is the case for the other variables. In the simulation study, we assumed that only the categories of variable *Educational level* have to sum up to known totals for our population. This mimics the situation for the Dutch Population Census, where estimated totals for *Educational level* are known from the Educational Attainment File and mass imputation is planned to be used. The true totals of *Educational level*, respectively *Occupation*, are given in Tables 7 and 8.

## 4.3 The simulation study

For our simulation study we introduced missingness in individual data items in the dataset described in Subsect. 4.2. In general, three kinds of missing data mechanisms are distinguished in the literature: Missing Completely At Random (MCAR), Missing At Random (MAR) and Not Missing At Random (NMAR) (Rubin 1976). Roughly speaking, a missing data mechanism is MCAR, if the reason for a value being missing does not depend on the value itself, nor on values of background variables. A missing data mechanism is MAR, if the reason for a value being missing

does not depend on the value itself, but does depend on values of background variables. A missing data mechanism is NMAR, if the reason for a value being missing depends on the value itself, even after correcting for the background variables. MCAR is the simplest case to deal with. Many imputation methods are able to correct for this situation. MAR is more complicated than MCAR. One can correct for a MAR mechanism by taking appropriate background variables into account in the imputation process. NMAR is the most difficult case by far. One can only correct for this case by relying on assumptions that cannot be tested from the dataset with missings itself.

In this paper we focus on MAR mechanisms, since these are the most important missingness mechanisms in practice. Sometimes we will also refer to results for MCAR mechanisms. Those results are given in the Supplementary Material to this paper. For more results for MCAR and NMAR mechanisms we refer to De Waal and Daalmans (2019).

In our simulation study, we examined two different fractions of missingness (referred to as Low and High) and three approximation methods for the correct imputation probabilities (the binomial approximation, the Poisson approximation and the IPF approximation).

In a preliminary study we found that the number of iterations in the actual imputation phase (see Sect. 3.2) affects the results as expected, and that after ten iterations our imputation approach has converged to (near) optimal results (see also De Waal and Daalmans 2019). Setting the number of iterations to ten appears to be a good trade-off between quality of the obtained results and the required computing time for our data. In our simulation study we therefore set the number of iterations in the actual imputation phase to ten.

We took into account that in practice values for some variables will be missing (far) more often than for other variables. For instance, values for *Educational level* and *Occupation* will be missing quite often in the Dutch Population Census, whereas values of, for example, *Gender* will be missing only very rarely.

The stochastic process to create missingness in a variable is independent from the missingness process for any of the other variables. For the MAR mechanisms, we examined situations where the missingness of *Educational level* or the missingness of *Occupation* depends on the age class of the person. In particular, we assumed that *Educational level* or *Occupation* are observed more often for people in a younger age class than for people in an older age class. For *Educational level* this reflects the current situation in the Netherlands, where the *Educational level* of younger people is more frequently available in administrative datasets than for older people. We defined a “young” class consisting of people up to 29 years (categories 1 to 6 of Age), a “middle” class consisting of people from 30 up to 54 years (categories 7 to 11 of Age), and an “old” class consisting of people of 55 years and older (categories 12 to 17 of Age). The “young” class consists of 1,285 persons, the “middle” class of 1,714 persons, and the “old” class of 785 persons.

For the MAR mechanisms for *Educational level*, we assumed that the missing data mechanism for each of the other variables, including *Occupation*, is MCAR. Similarly, for the MAR mechanism for *Occupation*, we assumed that the missing data mechanisms for each of the other variables, including *Educational level*,

**Table 9** Numbers of missings for the MAR mechanism for educational level and Low missingness

Variable	Young	Middle	Old	Total
Gender	6	9	4	19
Age	64	86	39	189
Position in the household	64	86	39	189
Size of the household	64	86	39	189
Residential area last year	6	9	4	19
Nationality	6	9	4	19
Country of birth	6	9	4	19
Educational level	161	471	314	946
Economic status	64	86	39	189
Occupation	321	429	196	946
NACE code	64	86	39	189
Marital status	64	86	39	189
Total	890	1452	760	3102

is MCAR. The numbers of missings that we created for the MAR mechanism for *Educational level* for each variable in all three age classes for the Low missingness scenario are given in Table 9. For *Educational level* the missingness percentages are: 12.5% for the “young” class, 27.5% for the “middle” class and 40% for the “old” class. For High missingness, the numbers of missings are twice as high as in Table 9.

The numbers of missings that we created for the MAR mechanism for *Occupation* are the same as in Table 9, except that the number of missings for *Educational level* and *Occupation* are interchanged, i.e. the numbers of missings for *Educational level* are 321 for the “young” class, 429 for the “middle” class and 196 for the “old” class, and for *Occupation* 161 (12.5%) for the “young” class, 471 (27.5%) for the “middle” class and 314 (40%) for the “old” class. Again, for High missingness, the numbers of missings are twice as high.

For all four MAR missingness mechanisms (MAR for *Educational level* with Low and High missingness and MAR for *Occupation* with Low and High missingness) we generated 250 missing data patterns and deleted the corresponding values from our dataset.

As mentioned above, we will refer to some results for MCAR mechanisms in order to compare them to the results of the MAR missingness mechanisms. In our MCAR missingness mechanisms we created the number of missings given in the last column of Table 9 (“Total”) for the Low scenario and twice those numbers for the High scenario. The difference with the MAR missingness mechanisms is that for the MCAR missingness mechanisms missingness is not influenced by the value of *Age*.

### 4.4 Quality measures

Since we introduced missingness in a dataset with known values ourselves, we are able to compare the imputed values to the actual values. The main interest of NSIs is the production of high-quality descriptive statistics, such as totals and means. To measure the quality of our imputation approach, we therefore examine to which extent totals for *Occupation* are preserved. We also examine to which extent cell totals are preserved for the cross-table of *Educational level* and *Occupation*.

We denote the categories of a variable  $x$  by  $1, 2, \dots, C_x$ , where  $C_x$  is the number of categories of variable  $x$ . For each category  $c$  ( $c = 1, 2, \dots, C_x$ ) of variable  $x$ , we calculate two quality measures,  $B_x(c)$  and  $M_x(c)$ , which are defined by

$$B_x(c) = \frac{\sum_{s=1}^S (T_{s,\text{imp}}(c;x) - T_{\text{true}}(c;x))}{S}$$

and

$$M_x(c) = \frac{\sum_{s=1}^S (T_{s,\text{imp}}(c;x) - T_{\text{true}}(c;x))^2}{S}$$

where  $S$  is the number of generated missing data patterns (250 in our case),  $T_{s,\text{imp}}(c;x)$  is the total of category  $c$  of variable  $x$  in the  $s$ -th imputed dataset ( $s = 1, \dots, S$ ), and  $T_{\text{true}}(c;x)$  is the corresponding total in the original, complete dataset.  $B_x(c)$  is the empirical bias of the imputed total of category  $c$  of variable  $x$ , and  $M_x(c)$  its empirical mean squared error.

Similarly, for each combination of categories  $c$  ( $c = 1, 2, \dots, C_x$ ) and  $c'$  ( $c' = 1, 2, \dots, C_y$  with  $C_y$  the number of categories of a variable  $y$ ) of variables  $x$ , respectively  $y$ , we calculate two measures,  $B_{x,y}(c, c')$  and  $M_{x,y}(c, c')$ , which are defined by

$$B_{x,y}(c, c') = \frac{\sum_{s=1}^S (T_{s,\text{imp}}(c, c'; x, y) - T_{\text{true}}(c, c'; x, y))}{S}$$

and

$$M_{x,y}(c, c') = \frac{\sum_{s=1}^S (T_{s,\text{imp}}(c, c'; x, y) - T_{\text{true}}(c, c'; x, y))^2}{S}$$

where  $T_{s,\text{imp}}(c, c'; x, y)$  is the number of times that the combination of category  $c$  of variable  $x$  and category  $c'$  of variable  $y$  occurs in the  $s$ -th imputed dataset, and  $T_{\text{true}}(c, c'; x, y)$  is the number of times that the combination of category  $c$  of variable  $x$  and category  $c'$  of variable  $y$  occurs in the original, complete dataset. We summarize  $B_{x,y}(c, c')$  and  $M_{x,y}(c, c')$  into two quality measures  $B_{x,y}^+$  and  $M_{x,y}^+$  defined by

$$B_{x,y}^+ = \frac{\sum_{c=1}^{C_x} \sum_{c'=1}^{C_y} |B_{x,y}(c, c')|}{C_x C_y}$$

and

$$M_{x,y}^+ = \frac{\sum_{c=1}^{C_x} \sum_{c'=1}^{C_y} M_{x,y}(c, c')}{C_x C_y}$$

In the summation for  $B_{x,y}^+$ , we take the absolute values of  $B_{x,y}(c, c')$ , since we do not want positive and negative values to cancel out.  $B_{x,y}^+$  is the average absolute empirical bias over all cells in the cross-table of  $x$  and  $y$ , and  $M_{x,y}^+$  is the average empirical mean squared error over these cells.

Although generally considered to be less important by NSIs, we also look at how often the correct category is imputed for the missing values, i.e. the prediction accuracy of individual values. That is, for variable  $x$ , we calculate

$$D_x = \frac{\sum_{s=1}^S \sum_{i \in \text{Miss}_{x,s}} I(x_{i,s,\text{imp}} = x_{i,\text{true}})}{S}$$

where  $x_{i,s,\text{imp}}$  is the value of variable  $x$  in unit  $i$  in the  $s$ -th imputed dataset,  $x_{i,\text{true}}$  is the corresponding value in the original, complete dataset,  $\text{Miss}_{x,s}$  is the set of units for which the value of variable  $x$  is missing in the  $s$ -th sample, and  $I$  is the indicator function, i.e.  $I(x_{i,s,\text{imp}} = x_{i,\text{true}}) = 1$  if  $x_{i,s,\text{imp}} = x_{i,\text{true}}$  and  $I(x_{i,s,\text{imp}} = x_{i,\text{true}}) = 0$  otherwise. We will summarize the results for  $D_x$  into a single number  $D^+$  defined by

$$D^+ = \sum_x D_x$$

where the summation runs over all 12 variables. For  $D^+$ , the higher its value, the better the quality of the imputations. For the other quality measures, the smaller their values, the better the quality of the imputations.

If using the binomial or the Poisson approximation for the exact imputation probabilities taking known totals has any effect on the quality of the imputations, this should most clearly be reflected in the  $D^+$  measure, since the better the approximations of the imputation probabilities, the more missing values should be imputed correctly.

We have also compared estimated standard errors of the estimated totals for the categories of *Occupation* obtained by a bootstrap approach to their corresponding (approximate) true standard deviations. The bootstrap approach used to do this and the results thereof are described in Sect. 6.

## 5 Results

By design, our proposed imputation approach preserves totals for the categories of *Educational level* in all scenarios. Likewise, by design, our proposed imputation approach also satisfies specified edits.

**Table 10**  $B_x(c)$  and  $M_x(c)$  for Occupation for MAR mechanisms with Low missingness

		MAR mechanism in Educational level									
		“1”	“2”	“3”	“4”	“5”	“6”	“7”	“8”	“9”	“999”
Binomial	$B_x(c)$	-0.24	0.07	-0.12	0.06	0.52	0.52	0.42	0.12	0.00	-1.34
	$M_x(c)$	109.04	116.03	180.08	96.94	93.45	20.41	74.33	73.16	85.89	17.58
Poisson	$B_x(c)$	-0.94	0.04	-0.06	0.64	0.54	0.16	0.07	0.43	0.10	-0.98
	$M_x(c)$	118.16	107.32	155.95	95.84	112.55	18.92	91.27	73.63	102.92	15.57
IPF	$B_x(c)$	-0.50	0.52	-1.03	0.62	0.00	0.67	-0.10	0.60	0.43	-1.22
	$M_x(c)$	110.89	100.33	174.47	98.15	92.14	19.93	86.99	75.25	89.14	17.50
		MAR mechanism in Occupation									
		“1”	“2”	“3”	“4”	“5”	“6”	“7”	“8”	“9”	“999”
Binomial	$B_x(c)$	-0.84	0.92	-0.02	-0.19	1.06	0.57	-0.38	0.20	0.76	-2.08
	$M_x(c)$	148.32	120.28	171.68	104.18	90.27	19.02	82.84	74.518	81.318	17.92
Poisson	$B_x(c)$	-1.65	1.19	0.66	-0.00	0.30	0.57	0.61	-0.06	0.70	-2.33
	$M_x(c)$	141.82	105.25	159.88	113.84	89.60	18.23	83.70	70.60	67.95	21.75
IPF	$B_x(c)$	-1.01	0.82	0.52	-0.79	1.30	0.56	0.47	-0.40	0.88	-2.35
	$M_x(c)$	142.24	122.04	186.55	115.24	92.33	20.34	74.04	73.40	81.16	20.80

### 5.1 Univariate results

We give univariate results for *Occupation* for the two MAR mechanisms – a MAR mechanism for *Educational level* and a MAR mechanism for *Occupation* – with Low missingness. The conclusions that we can draw for High missingness are similar to those for Low missingness. For other variables, the results are similar to those of *Occupation* (see De Waal and Daalmans 2019).

Table 10 presents univariate results for *Occupation* for the MAR mechanism in *Educational level* as well as for the MAR mechanism in *Occupation* for Low missingness.

In Table 10, we see that  $B_x(c)$  is quite low compared to the totals of the categories of *Occupation* in the complete dataset (see Table 8). Also, standard deviations  $sd_x(c)$ , computed as  $sd_x(c) = \sqrt{M_x(c) - B_x^2(c)}$ , are quite low for Table 10, i.e. close to zero and much smaller than the totals of the categories of *Occupation*. So, univariate results for *Occupation* are preserved quite well for both MAR mechanisms.

### 5.2 Cross-tables

Table 11 gives the results for  $B_{x,y}^+$  and  $M_{x,y}^+$  for the cross-table of *Educational level* and *Occupation* for all four MAR mechanisms. In Appendix D we give results for  $B_{x,y}(c, c')$  and  $M_{x,y}(c, c')$  for the case of Low Missingness for each cell in the cross-table of *Educational level* and *Occupation* for our MCAR missingness

**Table 11** Cross-table of educational level and occupation:  $B_{x,y}^+$  and  $M_{x,y}^+$  for MAR mechanisms

	Low missingness		High missingness	
	$B_{x,y}^+$	$M_{x,y}^+$	$B_{x,y}^+$	$M_{x,y}^+$
<i>MAR mechanism in educational level</i>				
Binomial	22.32	1678.14	22.16	1657.88
Poisson	22.29	1671.99	22.20	1659.15
IPF	22.37	1680.57	22.23	1662.77
<i>MAR mechanism in occupation</i>				
Binomial	22.32	1676.48	22.17	1645.60
Poisson	22.37	1680.33	22.21	1648.69
IPF	22.34	1676.66	22.27	1657.29

mechanism and for the MAR missingness mechanisms for both *Educational level* and *Occupation*.

In Table 11 we might see some effects of using the binomial or Poisson approximation instead of the IPF approximation for the exact imputation probabilities taking known totals into account. However, we do not see such an effect in Table 11, suggesting that the all three approximations work about equally well.

The results for  $B_{x,y}^+$  and  $M_{x,y}^+$  in Table 11 for the MAR mechanisms are clearly higher than for MCAR mechanisms with the same number of missing values (see the Supplementary Material to this paper). This shows that cross-tables are clearly less well estimated for MAR mechanisms than for MCAR mechanisms. The results in Table 11 show that estimates for the cross-table of *Educational level* and *Occupation* are biased for our MAR mechanisms (see also Appendix D).

### 5.3 Number of correct imputations

As already mentioned, any effects of using the binomial or Poisson approximation instead of the IPF approximation should most clearly be reflected in the  $D^+$  measure. However, we see no such an effect in Table 12. This confirms the earlier

**Table 12** Results for  $D^+$  for MAR mechanisms; in brackets the percentage of correct imputations for the total number of missings

	Low missingness	High missingness
<i>MAR mechanism in educational level</i>		
Binomial	1044.97 (33.69%)	2076.28 (33.47%)
Poisson	1045.60 (33.71%)	2079.04 (33.51%)
IPF	1044.09 (33.66%)	2076.89 (33.48%)
<i>MAR mechanism in occupation</i>		
Binomial	1047.83 (33.78%)	2056.11 (33.14%)
Poisson	1045.58 (33.71%)	2056.92 (33.15%)
IPF	1046.22 (33.73%)	2057.28 (33.16%)

finding that the IPF approximation works just as well as the binomial or Poisson approximation for our dataset.

The results for  $D^+$  in Table 12 for the MAR mechanisms are clearly worse than for MCAR mechanisms with the same number of missing values (see the Supplementary Material to this paper). This shows that prediction accuracy for individual values decreases substantially for a MAR mechanism in comparison to an MCAR mechanism.

## 6 Variance estimation

Often an estimate – or at least a good indicator – for the variance of an estimator is considered very important, and that certainly holds true for estimates for a population census. In this section we therefore compare estimated standard errors for the estimated totals of the categories of *Occupation* obtained by a bootstrap approach to their (approximate) true standard deviations. Since we are dealing with a finite population that we mass impute we have used a pseudo-population bootstrap approach to estimate the standard errors.

As in the rest of this paper, in our pseudo-population bootstrap approach we assume that all units in the population are (at least partly) observed, so the inclusion probability of each population unit is one. However, in each record the values of some variables may be missing. That missingness is caused by a random missingness process, which – for computational reasons – is based on the MCAR mechanism for the Low missingness scenario in this section rather than on MAR mechanisms as in the rest of this paper.

Our situation differs from the usual situation considered in the literature on estimating the variance of estimators based on imputed data in the sense that missingness occurs in all our variables, whereas in the literature missingness is usually assumed to occur in only one variable. Little seems to be known about applying a pseudo-population bootstrap approach when several variables contain missingness. This means that (the application of) our pseudo-population bootstrap approach is somewhat experimental.

Our pseudo-population bootstrap approach is similar to the approach in Scholtus and Daalmans (2021), and is as follows:

- The starting point is our complete population dataset without any missing values.
- For  $s = 1, \dots, N_{mis}$  we do the following
  1. Introduce missingness in our population dataset by means of the MCAR missingness procedure for the Low missingness scenario, and thus create a version,  $Pop_{mis,s}$ , of our population dataset that contains missing values.
  2. Impute the missing values by means of our mass imputation approach. This leads to estimated total  $T_{s,imp}(c; Occupation)$  for each category  $c$  of *Occupation*.
  3. Select the set of completely observed units  $Pop_{mis,com,s}$  in  $Pop_{mis,s}$ .
  4. Create one complete pseudo-population of the same size (3,784 units) as our original population. Suppose that  $Pop_{mis,com,s}$  consists of  $n_s$  completely observed values. Since we have used an MCAR missingness mechanism, each

unit is equally likely to be completely observed. We therefore start by creating  $\left\lfloor \frac{3784}{n_s} \right\rfloor$  copies of each unit in  $Pop_{mis,com,s}$ , where  $\lfloor q \rfloor$  denotes the largest integer value less than  $q$ . The total number of units created in this way is very likely to be less than 3,784. If so, we then randomly draw some extra units from  $Pop_{mis,com,s}$ , where each unit may be drawn extra only once, until we have created a pseudo-population  $Pseudo_s$  with 3,784 complete units.

5. For  $b = 1, \dots, B$  we do the following
  - a. Introduce missingness in  $Pseudo_s$  by means of our MCAR missingness procedure, so we obtain a pseudo-population  $Pseudo_{mis,s}$  with missing values.
  - b. Impute  $Pseudo_{mis,s}$  by means of our mass imputation approach.
  - c. Calculate the total  $T_{s,pseudo}(c; Occupation)$  for each category  $c$  of *Occupation* in the imputed version of  $Pseudo_{mis,s}$ .
6. For each category  $c$  of *Occupation*, calculate the bootstrap variance  $\text{var}_{s,boot}(c) = (B - 1)^{-1} \sum_{b=1}^B \left( T_{s,pseudo}(c; Occupation) - \bar{T}_{s,pseudo}(c; Occupation) \right)^2$  with  $\bar{T}_{s,pseudo}(c; Occupation) = B^{-1} \sum_{b=1}^B T_{s,pseudo}(c; Occupation)$ .
7. For each category  $c$  of *Occupation*, estimate the standard error of the estimated total of category  $c$  of *Occupation* by  $se_{s,boot}(c) = \sqrt{\text{var}_{s,boot}(c)}$ .

In our simulation study we have used  $N_{mis} = 250$  and  $B = 200$ . As noted in the literature, for variance estimation,  $B = 200$  bootstrap replicates are often considered sufficient (see Efron and Tibshirani 1993, Sect. 6.4).

In Step 3 we create only one pseudo-population. In principle, it may be better to construct several pseudo-populations. However, previous results in Chauvet (2007) and Kuijvenhoven and Scholtus (2011) suggest that creating several pseudo-populations instead of only one hardly affects the estimated variances. In other words, creating a single pseudo-population as we did, often leads to variance estimates of similar accuracy as creating several pseudo-populations.

We approximated true standard deviations for the categories of *Occupation* by introducing missingness into our original population by means of our MCAR procedure for the Low missingness scenario 2,500 times. To each of these 2,500 datasets with missing values we applied our mass imputation approach, calculated the totals for the categories of *Occupation* in the imputed datasets, and estimated the true standard deviations over the 2,500 datasets.

Our pseudo-population bootstrap approach to estimate the standard errors of the totals of *Occupation* is very time-consuming since we applied our iterative imputation method  $N_{mis} \times B = 250 \times 200 = 50,000$  times. Since we use ten iterations in our imputation method, we had to apply our imputation method 500,000 times in total. For this reason we have used a MCAR missingness mechanism, instead of MAR missingness mechanisms as we did in the rest of our paper. Even with a MCAR missingness procedure and using parallelized R code on a PC with four computing cores, it took more than two months to run this simulation study.

**Table 13** Averages of the estimated standard errors over the  $N_{mis} = 250$  populations with missing values divided by true (approximate) standard deviations of the categories of Occupation

“1”	“2”	“3”	“4”	“5”	“6”	“7”	“8”	“9”	“999”
0.95	0.97	0.96	0.95	0.96	0.89	0.94	0.96	0.96	1.06

We note that even with a Low missingness scenario the expected percentage of units with one or more missing values is about 48.66%, so only about 51.33% of the units, i.e.  $51.33\% \times 3784 \approx 1942$  units, are expected to be completely observed and can be used to construct a pseudo-population.

Table 13 gives the averages of estimated standard errors over  $N_{mis}$  populations with missing values divided by true (approximate) standard deviations for the estimated totals of the categories of *Occupation*. We see that in most cases the average of the estimated standard error divided by the corresponding true (approximate) standard deviation is close to, but a bit less than, one. This shows that for most cases the estimates for the variances based on our pseudo-population bootstrap approach are close to the true variances of the imputation approach. Exceptions are category “6” – the category with the smallest true total – which is clearly less than one, and “999” – the category with the highest true total – which is a bit larger than one.

Table 14 gives coverage rates of the estimated 95%-confidence intervals for the categories of *Occupation*. These 95%-confidence intervals are computed as

$$(T_{s,imp}(c; Occupation) - 1.96 \times se_{s,boot}(c), T_{s,imp}(c; Occupation) + 1.96 \times se_{s,boot}(c)).$$

Note that the estimated 95%-confidence intervals are centered around  $T_{s,imp}(c; Occupation)$ , i.e. around totals obtained by applying our mass imputation procedure directly to  $Pop_{mis,s} (s = 1, \dots, N_{mis})$  and before constructing a pseudo-population. This way of computing the confidence intervals was suggested by a reviewer.

The coverage rates of the estimated 95%-confidence intervals should be close to the nominal rate of 95%, which is indeed the case for all categories except category “6” as can be seen in Table 14

## 7 Discussion

In this paper we have generalized the imputation approach of Favre et al. (2005) to multiple variables to be imputed. We have also tested three approximations for the imputation probabilities taking known totals into account. We have carried out a simulation study to examine the properties of our proposed methodology in several different situations.

**Table 14** Coverage rates of the estimated confidence intervals of the categories of Occupation

“1”	“2”	“3”	“4”	“5”	“6”	“7”	“8”	“9”	“999”
95,2%	96,0%	92,4%	93,2%	97,6%	90,0%	93,6%	91,6%	92,4%	95,2%

The first conclusion that we can draw is that our proposed methodology does work in the sense that it allows us to impute multivariate missing data such that edits are satisfied and known totals are exactly preserved.

For MCAR missing data mechanisms (see the Supplementary Material to this paper), the univariate results for individual variables and results for two-dimensional cross-tables are (nearly) unbiased. For MAR missing data mechanisms, univariate results are also (nearly) unbiased. However, whereas results for cross-tables are (nearly) unbiased for MCAR data mechanisms, they are biased for MAR mechanisms. Also, the prediction accuracy for individual values is much lower for MAR mechanisms than for MCAR mechanisms. In order to preserve cross-tables better for MAR mechanisms one should use imputation models that capture statistical relations between variables better than our relatively simple multinomial imputation models do. For instance, for our MAR mechanisms we could build different multinomial imputation models for each of our three age classes (see Sect. 4.3). Testing such more advanced imputation models that better capture statistical relations between variables is a point for future research.

In Sect. 6 we examined the use of a pseudo-population bootstrap approach to estimate the standard errors of the totals of *Occupation*. For most categories of *Occupation*, the averages of estimated standard errors obtained by the bootstrap approach were quite close to the corresponding (approximate) true standard deviations. Also, coverage rates of the estimated confidence intervals of the estimated totals for the categories of *Occupation* were for most categories quite close to the nominal rate of 95%. This suggests that using our pseudo-population bootstrap approach generally gives reasonable estimates for the imputation variance.

In this paper, we focused on mass imputation as this is the most relevant situation for the Dutch Population Census. However, a modified version of the proposed imputation approach also seems useful for cases where a sample of the population is imputed, and *weighted* sums of the imputed values have to sum up to known population totals for some variables. Favre et al. (2004 and 2005) developed a variant of Cox's controlled rounding algorithm that is able to handle this situation. This variant can be included into our imputation approach by replacing the original version of Cox' controlled rounding algorithm with the variant developed by Favre et al. (2004). We leave this extension of our imputation approach to potential future work.

## Appendix A: Technical details Fellegi–Holt elimination approach

In this appendix, we describe the Fellegi–Holt elimination method itself. As in Sect. 3.1 we denote the number of variables by  $n$  and write each edit  $k$  in normal form. We denote the domain, i.e. the set of all possible values, of a variable  $s$  ( $s = 1, \dots, n$ ) by  $\text{Dom}_s$ . When  $F_s^k$  is not the domain of a variable ( $s = 1, \dots, n$ ), variable  $s$  is said to be involved in edit  $k$ . To eliminate a variable  $t$  from a current set of edits by means of the Fellegi–Holt elimination method, we start by determining all index sets  $U$  such that

$$\bigcup_{k \in U} F_t^k = \text{Dom}_t \tag{A.1}$$

and

$$\bigcap_{k \in U} F_s^k \neq \emptyset \text{ for } s \neq t \tag{A.2}$$

From these index sets we select the *minimal* ones, i.e. the index sets  $U$  that obey (A.1) and (A.2), but none of their proper subsets obey (A.1). Given such a minimal index set  $U$  we construct the implied edit

$$\bigcap_{k \in U} F_1^k \times \dots \times \bigcap_{k \in U} F_{t-1}^k \times \text{Dom}_t \times \bigcap_{k \in U} F_{t+1}^k \times \dots \times \bigcap_{k \in U} F_n^k \tag{A.3}$$

By adding the implied edits resulting from all minimal sets  $U$  to the current set of edits, and then removing all edits involving the variable  $t$  to be eliminated, we obtain a set of edits for the remaining variables.

In the example in Sect. 3.1 we have the edits

$$\{\text{Married}\} \times \{< 16 \text{ years}\} \tag{A.4}$$

and

$$\{\text{Unmarried, Divorced, Widowed}\} \times \{< 16 \text{ years, } \geq 16 \text{ years}\} \tag{A.5}$$

for *Marital status* and *Age* after filling in the observed value “Spouse” for *Relation to head of household* into the two user-specified edits (1) and (2). So, we have  $F_1^1 = \{\text{Married}\}$ ,  $F_2^1 = \{< 16\text{years}\}$ ,  $F_1^2 = \{\text{Unmarried, Divorced, Widowed}\}$  and  $F_2^2 = \{< 16\text{years, } \geq 16\text{years}\}$ , where superscript “1” refers to (A.4) and superscript “2” to (A.5). By taking  $U = \{1, 2\}$  and using (A.3), we obtain the implied edit

$$\{< 16 \text{ years}\} \tag{A.8}$$

for variable *Age*.

### Appendix B: Categories of the variables

**Age:** “1” (0–4 years), “2” (5–9 years), “3” (10–14 years), “4” (15–19 years), “5” (20–24 years), “6” (25–29), “7” (30–34 years), “8” (35–39 years), “9” (40–44 years), “10” (45–49 years), “11” (50–54 years), “12” (55–59 years), “13” (60–64 years), “14” (65–69 years), “15” (70–74 years), “16” (75–79 years), “17” (80 years or older).

**Position in the household:** “1110” (Child), “1121” (Married without children), “1122” (Married with children), “1131” (Living together without children), “1132” (Living together with children), “1140” (Alone living elderly person), “1210” (Living alone), “1220” (Different household).

**Size of the household:** “111” (1 person), “112” (2 persons), “113” (3 persons), “114” (4 persons), “125” (5 persons), “126” (6 or more persons).

**Educational level:** “0” (Pre-primary), “1” (Primary), “2” (Lower secondary), “3” (Upper secondary), “4” (Post-secondary), “5” (Tertiary), “9” (Without any education).

**Economic status:** “111” (Employee, other), “112” (Student with job), “120” (Independent worker), “210” (Unemployed), “221” (Education-related), “222” (Retired), “223” (Houseman/housewife), “224” (Other inactive)”.

**Occupation:** “1” (ISCO 1: legislators, senior officials and managers), “2” (ISCO 2: professionals), “3” (ISCO 3: technicians and assistant professionals), “4” (ISCO 4: clerks), “5” (ISCO 5: service, shop, market sales workers), “6” (Other), “7” (ISCO 7: craft and relative workers), “8” (ISCO 8: plant and machine operators and assistants), “9” (ISCO 9: elementary occupations), “999” (Not working).

**NACE code:** “111” (NACE A+B: agriculture, hunting, forestry and fishing), “122” (NACE C+D+E: mining, manufacturing and electricity), “124” (NACE F: construction), “131” (NACE G: wholesale, retail trade, repair), “132” (NACE H: hotels and restaurants), “133” (NACE I: transport, storage, communication), “134” (NACE J: financial intermediation), “135” (NACE K: real estate, renting and business activities), “136” (NACE L: public administration, defense), “137” (NACE M: education), “138” (NACE N: health, social work), “139” (NACE O: other community, social personal service activities), “200” (Not working).

**Marital status:** “1” (Unmarried), “2” (Married), “3” (Widowed), “4” (Divorced).

**Gender:** “1” (Male), “2” (Female).

**Residential area last year:** “1” (Same COROP area), “2” (Other COROP area, or outside the Netherlands), “9” (Not applicable (person less than 1 year old)).

**Nationality:** “1” (Dutch), “2” (From other countries in Europe), “3” (Other).

**Country of birth:** “1” (The Netherlands), “2” (From other countries in Europe), “3” (Other).

## Appendix C: User-specified edit rules

The user-specified edit rules that have to hold for our dataset are given below. These edits indicate impossible combinations. Variables that are not mentioned in an edit can take any value within its domain. For example, the first edit below says that if the value of *Age* is category “1”, “2” or “3”, i.e. if a person is 14 years or less, then the value of *Marital status* cannot be category “2” (married), “3” (widowed), or “4” (divorced), irrespective of the values on any of the other variables.

- ( $Age \in \{“1”, “2”, “3”\}$ )  $\times$  ( $Marital\ status \in \{“2”, “3”, “4”\}$ )
- ( $Position\ in\ the\ household \in \{“1121”, “1122”\}$ )  $\times$  ( $Marital\ status \in \{“1”, “3”, “4”\}$ )
- ( $Age \in \{“1”, “2”\}$ )  $\times$  ( $Educational\ level \in \{“1”, “2”, “3”, “4”, “5”\}$ )
- ( $Age \in \{“1”, “2”, “3”\}$ )  $\times$  ( $Educational\ level \in \{“2”, “3”, “4”, “5”\}$ )
- ( $Age \in \{“1”, “2”, “3”, “4”\}$ )  $\times$  ( $Educational\ level \in \{“5”\}$ )

- $(Age \in \{“1”, “2”, “3”\}) \times (Position\ in\ the\ household \in \{“1121”, “1122”, “1131”, “1132”, “1140”, “1210”\})$
- $(Age \in \{“1”, “2”, “3”\}) \times (Economic\ status \in \{“111”, “120”, “210”, “222”, “223”\})$
- $(Age \in \{“1”, “2”, “3”\}) \times (Occupation \in \{“1”, “2”, “3”, “4”, “5”, “6”, “7”, “8”, “9”\})$
- $(Age \in \{“1”, “2”, “3”\}) \times (NACE \in \{“111”, “122”, “124”, “131”, “132”, “133”, “134”, “135”, “136”, “137”, “138”, “139”\})$
- $(Size\ of\ the\ household \in \{“111”\}) \times (Position\ in\ the\ household \in \{“1110”, “1121”, “1122”, “1131”, “1132”, “1140”, “1220”\})$
- $(Size\ of\ the\ household \in \{“112”\}) \times (Position\ in\ the\ household \in \{“1122”, “1132”, “1220”\})$
- $(Size\ of\ the\ household \in \{“113”, “114”, “125”, “126”\}) \times (Position\ in\ the\ household \in \{“1110”, “1121”, “1122”, “1131”, “1132”, “1140”, “1220”\})$

**Appendix D: Cross-table of educational level and occupation:**  
 $B_{x,y}(c, c')$  and  $M_{x,y}(c, c')$

In this appendix we give results for  $B_{x,y}(c, c')$  and  $M_{x,y}(c, c')$  for the cells in the cross-table of *Education level* and *Occupation*. For comparison, in Table 15 we first give the true cross-table for *Education level* and *Occupation*.

In Tables 16 and 17 we give results for  $B_{x,y}(c, c')$ , respectively  $M_{x,y}(c, c')$  for our MCAR missingness mechanism, in Tables 18 and 19 the results for  $B_{x,y}(c, c')$ , respectively  $M_{x,y}(c, c')$  for the MAR mechanism for *Educational level*, and in Tables 20 and 21 the results for  $B_{x,y}(c, c')$ , respectively  $M_{x,y}(c, c')$  for the MAR mechanism for *Occupation*. We only give results for the IPF approximation and Low missingness. Results for the binomial approximation and the Poisson approximation are similar to the results for the IPF approximation.

**Table 15** True cross-table for Educational level and Occupation

Occupation	Educational level						
	“0”	“1”	“2”	“3”	“4”	“5”	“9”
“1”	0	14	44	112	14	76	0
“2”	0	1	9	60	16	250	0
“3”	0	13	28	228	24	67	0
“4”	0	17	71	124	18	21	0
“5”	1	24	82	111	5	6	0
“6”	0	2	9	16	3	4	0
“7”	1	28	65	109	3	6	0
“8”	1	33	45	63	0	5	0
“9”	3	42	75	52	1	5	0
“999”	248	404	379	345	32	111	258

**Table 16** Cross-table for Educational level and Occupation for MCAR mechanism, IPF and Low missingness:  $B_{x,y}(c, c')$ 

Occupation	Educational level						
	“0”	“1”	“2”	“3”	“4”	“5”	“9”
“1”	0.08	0.58	0.72	0.25	-0.40	-1.44	0.20
“2”	0.08	0.97	1.72	4.28	0.28	-7.55	0.21
“3”	0.11	0.80	2.22	-4.00	-0.49	1.12	0.24
“4”	0.06	0.27	-1.28	-0.64	-0.40	1.79	0.19
“5”	0.04	-0.61	-0.71	-0.74	0.27	1.55	0.20
“6”	0.02	0.07	-0.19	-0.08	-0.10	0.21	0.08
“7”	0.08	-0.19	-0.92	-0.54	0.22	1.20	0.15
“8”	0.08	-0.55	-0.79	-0.46	0.22	1.41	0.10
“9”	-0.14	-1.22	-1.80	1.32	0.32	1.34	0.17
“999”	-0.42	-0.12	1.02	0.62	0.06	0.37	-1.53

**Table 17** Cross-table for educational level and occupation for MCAR mechanism, IPF and low missingness:  $M_{x,y}(c, c')$ 

	Educational level						
	“0”	“1”	“2”	“3”	“4”	“5”	“9”
“1”	0.09	8.33	17.86	28.67	6.52	21.22	0.27
“2”	0.09	2.30	9.42	44.24	6.94	85.65	0.26
“3”	0.12	9.11	20.41	58.82	10.90	29.68	0.33
“4”	0.07	8.58	27.02	32.08	7.95	14.62	0.24
“5”	0.81	13.11	26.05	31.48	2.51	6.91	0.25
“6”	0.02	1.18	4.36	4.84	1.22	2.19	0.10
“7”	0.65	15.15	25.64	29.18	2.02	5.94	0.18
“8”	0.57	12.60	13.97	17.20	0.28	5.88	0.13
“9”	1.45	15.24	22.84	20.51	0.84	5.27	0.25
“999”	3.95	52.99	77.91	88.17	13.55	36.00	3.64

Comparing Tables 18 and 21 to Table 15, we see that bias is quite substantial for both the MAR mechanism for *Educational level* as well as for the MAR mechanism for *Occupation*. The results in Tables 18, 19, 20 and 21 for MAR mechanisms are clearly worse than the corresponding results in Tables 15 and 16 for the MCAR mechanism. This supports the conclusion in Sects. 5.2 and 7 that results for cross-tables are biased for MAR mechanisms, whereas they are (nearly) unbiased for the MCAR mechanism.

**Table 18** Cross-table for educational level and occupation for MAR mechanism for educational level, the IPF approach and low missingness:  $B_{x,y}(c, c')$

Occupation	Educational level						
	"0"	"1"	"2"	"3"	"4"	"5"	"9"
"1"	5.20	22.08	27.88	-21.14	-4.61	-29.58	0.17
"2"	1.80	39.47	68.80	89.17	-3.56	-195.90	0.22
"3"	5.99	30.74	67.71	-91.10	-9.46	-4.10	0.22
"4"	3.14	21.54	-11.27	-25.26	-6.23	17.98	0.09
"5"	2.76	8.16	-21.22	-24.66	1.66	33.14	0.16
"6"	0.98	4.63	0.04	-5.97	-2.73	3.03	0.02
"7"	2.07	3.92	-12.52	-28.07	7.37	27.09	0.13
"8"	0.24	-17.54	1.79	-9.88	6.21	19.12	0.06
"9"	0.75	-11.31	-37.29	16.34	5.13	26.32	0.07
"999"	-22.93	-101.68	-83.92	100.56	6.20	102.89	-1.12

**Table 19** Cross-table for educational level and occupation for MAR mechanism for educational level, the IPF approach and low missingness:  $M_{x,y}(c, c')$

Occupation	Educational level						
	"0"	"1"	"2"	"3"	"4"	"5"	"9"
"1"	28.84	498.88	795.08	473.07	24.65	886.97	0.20
"2"	4.12	1575.26	4757.07	7982.04	17.16	38,390.34	0.29
"3"	37.62	960.38	4615.61	8330.27	94.58	31.81	0.25
"4"	10.88	477.96	142.66	662.10	42.44	333.85	0.12
"5"	8.97	77.71	468.14	626.82	6.16	1108.69	0.20
"6"	1.13	23.41	2.50	38.50	7.69	10.49	0.02
"7"	5.24	24.33	171.27	808.77	58.04	742.52	0.14
"8"	0.69	313.55	12.45	109.96	40.57	372.69	0.06
"9"	1.84	137.71	1402.83	284.78	28.72	700.08	0.08
"999"	531.48	10,388.4	7098.97	10,170.66	50.66	10,626.34	2.44

**Table 20** Cross-table for educational level and occupation for MAR mechanism for occupation, the IPF approach and low missingness:  $B_{x,y}(c, c')$ 

Occupation	Educational level						
	"0"	"1"	"2"	"3"	"4"	"5"	"9"
"1"	5.52	21.62	27.15	-19.85	-4.75	-29.92	0.22
"2"	1.60	39.84	68.62	88.57	-3.39	-195.46	0.22
"3"	5.84	30.49	69.31	-93.13	-9.30	-3.44	0.24
"4"	3.22	21.46	-11.96	-25.26	-6.39	18.76	0.17
"5"	2.88	7.54	-21.16	-24.03	1.36	33.22	0.19
"6"	0.90	5.16	-0.03	-6.03	-2.83	2.82	0.02
"7"	1.86	3.71	-13.02	-27.47	7.52	27.27	0.14
"8"	0.56	-18.03	1.29	-9.69	6.72	19.02	0.13
"9"	0.79	-10.87	-36.96	15.56	5.04	26.24	0.21
"999"	-23.14	-100.91	-83.24	101.33	6.02	101.50	-1.55

**Table 21** Cross-table for educational level and occupation for MAR mechanism for occupation, the IPF approach and Low missingness:  $M_{x,y}(c, c')$ 

Occupation	Educational level						
	"0"	"1"	"2"	"3"	"4"	"5"	"9"
"1"	32.26	479.64	755.90	417.13	25.56	904.82	0.24
"2"	3.51	1597.40	4728.70	7875.74	16.50	38,217.13	0.30
"3"	36.16	942.27	4827.96	8702.65	91.70	26.52	0.26
"4"	11.74	471.53	161.31	659.62	45.26	360.87	0.20
"5"	9.80	67.14	465.88	599.47	4.53	1112.73	0.21
"6"	1.11	28.46	2.26	39.20	8.18	9.62	0.02
"7"	4.59	24.42	184.49	774.79	60.29	752.78	0.15
"8"	1.28	330.20	11.54	105.93	47.22	368.63	0.16
"9"	1.87	127.30	1378.78	259.12	27.50	695.04	0.27
"999"	543.98	10,224.72	6984.86	10,332.61	45.96	10,332.23	3.75

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s10182-023-00481-z>.

**Acknowledgements** This project was partly funded by Eurostat under the call "Multipurpose statistics for efficiency gains in production" (project 2018-NL-ESS.VIP.BUS).

## Declarations

**Conflict of interest** There are no financial or non-financial interests that are directly or indirectly related to the work submitted for publication.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long

as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Chandru, V., Hooker, J.N.: Optimization Methods for Logical Inference. John Wiley & Sons, New York (1999)
- Chauvet, G.: Méthodes de bootstrap en population finie. PhD Thesis, L'Université de Rennes (2007). <https://pastel.archives-ouvertes.fr/tel-00267689/document>. Accessed 16 Sep 2022
- Chen, S.X.: Weighted polynomial models and weighted sampling schemes for finite population. *Ann. Stat.* **26**, 1894–1915 (1998)
- Chen, S.X.: General properties and estimation of conditional Bernoulli models. *J. Multivar. Anal.* **74**, 69–87 (2000)
- Chen, S.X., Liu, J.S.: Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Stat. Sinica* **7**, 875–892 (1997)
- Chen, X.H., Dempster, A.P., Liu, J.S.: Weighted finite population sampling to maximize entropy. *Biometrika* **81**, 457–469 (1994)
- Cox, L.: A constructive procedure for unbiased controlled rounding. *J. Am. Stat. Assoc.* **82**, 520–524 (1987)
- Daalmans, J., Mass imputation for census estimation. Discussion paper 2017–04, Statistics Netherlands (2017). <https://www.cbs.nl/en-gb/background/2017/11/mass-imputation-for-census-estimation>. Accessed 16 Sep 2022
- De Waal, T., Quere, R.: A fast and simple algorithm for automatic editing of mixed data. *J. off. Stat.* **19**, 383–402 (2003)
- De Waal, T., Pannekoek, J., Scholtus, S.: Handbook of Statistical Data Editing and Imputation. John Wiley & Sons, New York (2011)
- De Waal, T., Coutinho, W., Shlomo, N.: Calibrated hot deck imputation for numerical data under edit restrictions. *J. Surv. Stat. Methodol.* **5**, 372–397 (2017)
- Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman & Hall/CRC, London (1993)
- Favre, A.-C., Matei, A., Tillé, Y.: A variant of the Cox algorithm for the imputation of non-response of qualitative data. *Comput. Stat. Data Anal.* **45**, 709–719 (2004)
- Favre, A.-C., Matei, A., Tillé, Y.: Calibrated random imputation for qualitative data. *J. Stat. Plan. Inference* **128**, 411–425 (2005)
- Fellegi, I.P., Holt, D.: A systematic approach to automatic edit and imputation. *J. Am. Stat. Assoc.* **71**, 17–35 (1976)
- Hong, Y.: On computing the distribution function for the Poisson-binomial distribution. *Comput. Stat. Data Anal.* **59**, 41–51 (2013)
- Hooker, J.: Logic-based Methods for Optimization. John Wiley & Sons, New York (2000)
- De Jonge, E., Van der Loo, M.: Error localization as a mixed-integer program in editrules. Discussion paper, Statistics Netherlands (2012), <https://www.cbs.nl/nl-nl/achtergrond/2014/15/error-localizati-on-as-a-mixed-integer-problem-with-the-editrules-package>. Accessed 14 Oct 2022
- Kuijvenhoven, L., Scholtus, S.: Bootstrapping combined estimators based on register and sample survey data. Discussion Paper, The Hague: Statistics Netherlands (2011). Available at: <http://www.cbs.nl/nl-nl/achtergrond/2011/39/bootstrapping-combined-estimator-based-on-register-and-sample-survey-data>. Accessed 16 Sep 2022
- Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data, 2nd edn. New York, Wiley (2002)
- Mashreghi, A., Haziza, D., Léger, C.: A survey of bootstrap methods in finite population sampling. *Stat. Surv.* **10**, 1–52 (2016)
- R Core Team: R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2020). <https://www.R-project.org/>. Accessed 16 Sep 2022

- Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J., Solenberger, P.: A multivariate technique for multiply imputing missing values using a sequence of regression models. *Surv. Methodol.* **27**, 85–95 (2001)
- Ripley, B.: Package nnet (2020). Available at <https://cran.r-project.org/web/packages/nnet/nnet.pdf>. Accessed 16 Sep 2022
- Rubin, D.B.: Inference and missing data. *Biometrika* **63**, 581–590 (1976)
- Rubin, D.B.: Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, New York (1987)
- Rubin, D.B.: Nested multiple imputation of NMES via partially incompatible MCMC. *Stat. Neerl.* **57**, 3–18 (2003)
- Schafer, J.L.: Analysis of Incomplete Multivariate Data. Chapman & Hall, London (1997)
- Scholtus, S., Daalmans, J.: Variance estimation after mass imputation based on combined administrative and survey data. *J. off. Stat.* **37**, 433–459 (2021)
- Scholtus, S.: Variances of census tables after mass imputation of educational attainment. Discussion Paper, The Hague: Statistics Netherland (2020). Available at: <http://www.cbs.nl/en-gb/background/2018/49/variances-of-census-tables-after-mass-imputation>. Accessed 16 Sep 2022
- Siddique, J., Belin, T.: Multiple imputation using an iterative hot-deck with distance-based donor selection. *Stat. Med.* **27**, 83–102 (2008)
- Van Buuren, S.: Flexible Imputation of Missing Data. Chapman & Hall/CRC, Boca Raton, Florida (2012)
- Van Buuren, S., Groothuis-Oudshoorn, K.: MICE: multivariate imputation by chained equations in R. *J. Stat. Softw.* **45**, 1–67 (2011)
- De Waal, T., Daalmans, J.: Multivariate mass imputation for the population census given known totals. Eurostat (2019). [https://ec.europa.eu/eurostat/cros/system/files/admin\\_wp6\\_2018\\_nl.pdf](https://ec.europa.eu/eurostat/cros/system/files/admin_wp6_2018_nl.pdf). Accessed 16 Sep 2022

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.