Combining Unlinkable Data By Utilising Statistical Matching and Assessing its Quality

Goudie, Francesca

Universiteit
Leiden
The Netherlands

Combining Unlinkable Data By Utilising Statistical Matching and Assessing its Quality

Francesca Goudie

Thesis advisor: Dr Ton de Waal, CBS

Thesis advisor: Dr Arnout van Delden, CBS

Thesis advisor: Dr Mark de Rooij, Institute of Psychology, Leiden University

Defended on 29th September, 2022

MASTER THESIS

STATISTICS AND DATA SCIENCE

UNIVERSITEIT LEIDEN

Contents

Abstract

Statistical Matching (SM) seeks to combine two datasets that have few or no overlapping units through establishing the relationship between variables based on a set of common variables. Little research has been done to establish a procedure looking at the quality of a SM method on a particular dataset. Therefore, this thesis proposes a bootstrapping method for this, which estimates the variance and bias of the matched data. This method can be used on two datasets with categorical target variables, which have a small overlap in units. This procedure was tested in a simulation study, which looked at different data conditions including proxy variables, and implemented in a case study from Statistics Netherlands. Two different SM methods were used. The results of this study are promising because within the different simulations the estimates for bias and variance were relatively close to the true values, although the method has its limitations.

1 Introduction

Surveys are a crucial way to collect data, where participants answer questions related to variables of interest. However, sometimes statisticians are interested in the relationship between variables which have been collected in two different sample surveys. In this situation, the main option is to utilise direct linkage of overlapping units in the datasets containing the variables of interest. In direct linkage observations about the same person or unit are linked to each other based on an identifying variable to form a new dataset, see D'Orazio et al. (2001, p. 433). Identifying variable(s) can be based on a national identification number or a collection of information, such as the combination of date of birth, name and address (Schnell, 2021, p. 3).

A desire to publish outputs based on variables in different datasets is a situation that occurs in the context of National Statistical Institutes (NSIs). NSIs are the independent public statistical agencies within countries and the main publisher of official statistics, which relate to their country (European Statistical System Committee, 2017). NSIs carry out large numbers of surveys to retrieve information on the population. Although some NSIs have access to large administrative datasets, the scope of administrative data is not exhaustive and it is not uncommon that researchers want to know the relationship between variables that are available only in different surveys. In an effort to decrease the response burden on participants of surveys, NSIs will ask different people to answer different surveys. This means, there is only a small cross-over of participants between surveys. Under such conditions, direct linkage is unhelpful here as there are no or very few overlapping units. This means that statistical analyses on this linked data will not have a lot of power and few inferences can be made. In this situation Statistical Matching (SM) can be used instead.

SM was developed to establish the relationship between two variables of interest, referred to as target variables, which have been recorded in two separate datasets with no or few overlapping units. SM seeks to estimate this relationship based on a set of matching variables which have observations in both datasets (Donatiello et al., 2016, p. 2).

1.1 Statistical Matching

Let Y and Z be the target variables. Let Y be in dataset A, of size n_A, with observations $y_1, ..., y_{n_A}$. Let Z be in dataset B, of size n_B, with observations $z_1, ..., z_{n_B}$. Then, let $\mathbf{X}$ refer to a group of matching variables which are present in both datasets A and B. This is the most simplistic SM scenario and is represented in Figure 1.

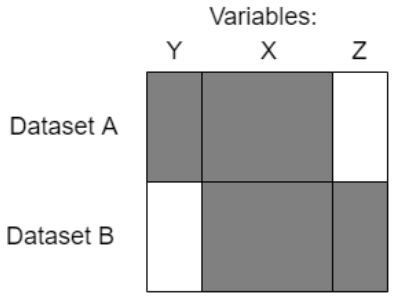

Figure 1: Main layout of Statistical Matching scenario. Where Y and Z refer to the two target variables, and $\mathbf{X}$ refers to the group of matching variables. The shaded grey sections signify that there are observed values for those variables in the datasets, and where white sections signify there are not observed values for those variables in those sections.

SM has two main types of methods, micro and macro approaches. In the micro approach the objective is to create a synthetic data set containing both target variables, Y and Z. On the other hand, the macro approach aims to estimate the joint distribution between the target variables by constructing a model of the distribution on all the available data (D'Orazio et al., 2006, p.2). In practice, the micro approach is used more frequently as the synthetic dataset can be used for further analyses, and the output of this approach will be used in this paper.

1.1.1 The Conditional Independence Assumption

In order to estimate the joint distribution of Y and Z based on two different datasets assumptions often need to be made. Most methods of SM assume the Conditional Independence Assumption (CIA). CIA assumes that there is independence between the target variables, Y and Z, conditionally on the matching variables, $\mathbf{X}$. In other words, the CIA states that the relationship between Y and Z can be explained entirely through

X. CIA is a very limiting assumption and rarely holds in practice (Donatiello et al., 2016, p. 3). There are several ways to limit the reliance on CIA. The main way to try and satisfy or attempt to satisfy CIA is through some auxiliary information, such as an additional dataset, possibly an overlap in units, or a proxy variable (Donatiello et al., 2014, p. 53). A proxy is a contaminated version of one of the target variables, answering the same question but with additional noise, and is therefore expected to have a similar distribution as its respective target variable (Kim et al., 2016, p. 29; D'Orazio et al., 2006, p.67). The proxy variable may be older, or may contain older information, or asked in a slightly different context. This is illustrated in Figure 2. With a proxy variable the CIA can be maintained since one of the matching variables, **X**, is closely related to one of the target variables, Y or Z (Donatiello et al., 2016, p.5). The use of proxy variables has not been tested very often in literature and it would therefore be useful to look at the use of proxy variables when assessing SM methods.

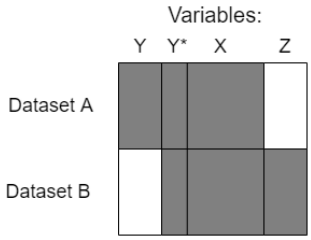

Figure 2: Statistical Matching scenario with a proxy variable. Where Y and Z refer to the two target variables, Y* refers to the proxy variable of Y and **X** refers to the group of matching variables. Both **X** and Y* are observed in both datasets A and B. The shaded grey sections signify that there are observed values for those variables in those datasets, and white sections signify there are not observed values for those variables.

1.1.2 Parametric versus Non-parametric approaches

SM approaches are further split into parametric, non-parametric approaches and mixtures of the two. In parametric approaches, SM estimates a parametric model that defines the relationship between the target variables Y and Z, based on the matching variables. Macro-parametric methods assume a family of parametric distributions across each of the variables, such as univariate normal, multinomial or multinormal. From this assumption the joint distribution between the variables can be determined mathematically (D'Orazio et al., 2006, pp. 14-23). Micro-parametric approaches estimate a parametric model and then generate a synthetic dataset from this model (D'Orazio et al., 2006, pp.25-26).

Macro non-parametric approaches stemmed from the limitations of macro parametric approaches. When the data is categorical and discrete, macro parametric approaches can utilise a multinomial distribution, which can be very flexible. However, when the

data is continuous it is harder to restrict the data to a distribution, which is where nonparametric approaches are relevant (D'Orazio et al., 2006, p. 31). These approaches tend to estimate relationships between Y and Z and $\mathbf{X}$ based on the data itself, see D'Orazio et al. (2006, pp. 31-34). Non-parametric micro approaches create a synthetic dataset through non-parametric methods, either through conditional mean matching based on a non-parametric regression function of the variables, see D'Orazio et al. (2006, pp.26-29;35) or through random draws. For random draws, most situations use hot deck procedures, which replace the missing values with observed values in the dataset. There are many variations on hot deck procedures. All have one dataset that is seen as a 'donor' dataset, say A, and the other dataset as the 'recipient' dataset, B. Values from A's target variable are matched to the observations in dataset B. Which value is chosen is based on the particular procedure used, e.g. randomly drawn from groups defined by the matching variables or through the distance based on a distance measure between the matching variables of the recipient record and the potential donor records (D'Orazio et al., 2006, pp. 37-42).

Mixed methods utilise a mixture of both parametric and non-parametric methods. They usually estimate the parameters of a parametric model and then use a non-parametric technique such as a hot deck conditionally on the estimated model (D'Orazio et al., 2006, p. 47). There are many variations of these mixed methods. In this project, two non-parametric micro approaches are used.

1.2 Quality of Statistical Matching Methods

Within the field of SM there has been research looking into uncertainty of the SM process, mostly through computing uncertainty bounds, see for example Conti et al. (2012) and D'Orazio (2019). However, these ideas of uncertainty do not look at the specific SM method being applied and instead look at uncertainty based on the available data. There has therefore been little research into the quality of the output after a SM method has occurred, specifically focusing at the SM methods performance. The main research has been done by Rässler (2002, pp.29-32; 2004, pp. 156-159), who talks about, what she calls the validity of the matching procedure. Validity has four levels (1) "the true (but unknown) values of the Y variable for the recipient observations are reproduced.", can only be checked without simulation if Z directly determines Y (Rässler, 2002, p.30). (2) "the true joint distribution of all variables is reflected in the statistically matched dataset" (Rässler, 2002, p.30), this requires checking assumptions, for example, whether CIA holds if it has been assumed (Rässler, 2004, p.158). (3) "the correlation structure and higher moments are preserved after statistical matching" (Rässler, 2002, p.30). (4) "the marginal and joint distributions of the variables in the donor sample are preserved

in the statistically matched" dataset (Rässler, 2002, p.30). In this final level, one can check the marginal distributions by comparing the matched dataset with the original datasets (Rässler, 2004, pp. 158-159). The only way to check levels 1-3 of a SM method is through a simulation study (Leulescu and Agaftei, 2013, p. 20; Rässler, 2002, p.30).

A desire to assess the quality of a SM method on a particular dataset without the ability to perform a simulation study forms the biggest problem when looking at the quality of SM methods. As with other methods in statistics, the SM method will work better on some datasets over other datasets, based on the true relationship between the target variables and what auxiliary information is available. This means one is unable to gauge how good the quality of output is from SM on the dataset you are using, as you do not have all information available, otherwise you would not be using SM. A possible solution to this, specifically for categorical target variables with datasets that have an overlap, would be to make use of the contingency tables of proportions for these two variables in the overlap. Then to look at the differences in these proportions compared to that of the matched data, in order to understand the quality of the matching process. This could be done through looking at the variance and bias of each cell in the contingency table of proportions through a bootstrapping process, suggested in this paper, which has not been suggested in literature before.

1.3 Research aim

The main aim of this project is to find a procedure to establish the quality of a SM method on a particular dataset by estimating the variance and bias that occurs through the SM method. The proposed quality measure procedure is restricted to categorical target variables, focusing on the contingency tables of proportions created between the two matched target variables and requires an overlap. This overlap is where there are observations for the same person or unit in both datasets, where observations are available for both target variables for these units, this is shown in Figure: 3. The overlap will be used to estimate the true proportions of the cells in the contingency table. We test the proposed quality measure procedure on two different SM methods in a simulation study. We also look at the impact that different proxy variables have on the quality of the outcome of the SM methods. We then also apply the procedure in a case study situation with data from Statistics Netherlands, within the context of an NSI.

The thesis will first outline the proposed quality measure (Section 2.1) and the SM methods it will be tested on (Section 2.2). Followed by discussing the simulation study used to test the proposed quality measure (Section 3), and the data used for this (Section 3.1). Next, the results of the simulation study will be discussed (Section 3.2). Then the case study situation will be outlined (Section 4), with the data being explained (Section

4.1) and the results of this case study stated (Section 4.2). Finally, the results of the thesis will be discussed overall (Section 5).

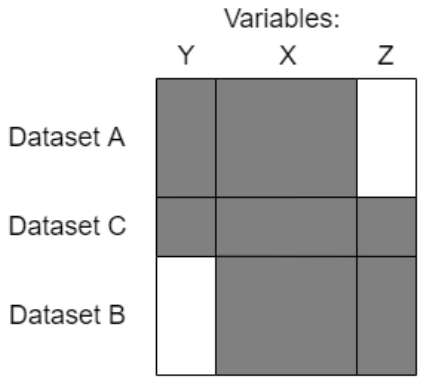

Figure 3: Statistical Matching scenario with a cross-over dataset. Y and Z refer to the two target variables, and **X** refers to the group of matching variables. **X** variables are observed in all three datasets A, B and C. The shaded grey sections signify that there are observed values for those variables in those datasets, and white sections signify values that are not observed and are missing from the datasets. So dataset C is a complete data containing all variables

2 Methods Theory

In this section, the methods used in this project will be outlined. Firstly, in section 2.1 we outline the proposed quality measure to assess a particular SM method on a particular dataset. Then in section 2.2 the two SM methods used to test the quality measure will be explained.

2.1 Quality Measure

To explain the proposed quality measure, firstly, the situation will be defined. Let A be a dataset of size n_A, which contains observations for target variable Y and auxiliary variables **X**. Also, let B be a dataset of size n_B, which contains target variable Z and auxiliary variables **X**. Finally, let C be a dataset of overlapping units from dataset A and B, a not uncommon situation where a small number of the same participants respond to both surveys A and B, this scenario is shown in Figure 3. Let C have a size of n_C which contains all the variables Y, Z and **X**, with n_A and n_B already including n_C. So, the total number of observations across the datasets is $n = n_A + n_B$. Variable Y and Z are categorical variables with T_Y and T_Z categories respectively.

In the procedure, bootstrap samples are drawn from the datasets A and B, including C, the overlap between them. Each bootstrap iteration has a bootstrap sample for A and

a bootstrap sample for B. In each bootstrap iteration the number of observations drawn from A is fixed to size n_A, but with a probability of coming from dataset A excluding C of n_A^*/n_A, where $n_A^* = n_A - n_C$, and the probability of it coming from dataset C of n_C/n_A. The number of values drawn from C is defined as d_C and then $n_B - d_C$ values are drawn from B. This is to make sure the situation is similar to reality where there would be a small overlap between two samples A and B of fixed sizes, while still keeping some randomness with respect to the size of this overlap. Within each bootstrap iteration, after the bootstrap samples are drawn, the SM method is performed on the bootstrap samples, which produces a synthetic dataset that contains at least Y and Z. A contingency table of proportions for Y and Z on the matched bootstrapped samples is then calculated. This is then repeated over R bootstrap iterations, where R is a large number. From the contingency tables of proportions over all the bootstrap estimates, the variance, bias, relative bias and relative standard error are estimated. The bias and relative bias are estimated against the estimated contingency table from the overlap. This is the basis of the quality measure procedure, and is described in more detail in Algorithm: 1.

2.2 Statistical Matching methods

We tested the quality measure procedure on two SM methods, focusing on categorical data for both the target and matching variables. Both methods are versions of a hot deck procedure. As already mentioned, in a hot deck procedure, one dataset is considered the 'donor' dataset, say dataset A, and the other dataset is considered the 'recipient' dataset, so dataset B. 'Donor' values for the target variable, Y, not in dataset B are chosen at random for the observations in B from dataset A, based on the corresponding matching variable values, $\mathbf{X}$, (D'Orazio et al., 2006, p. 36). The first method used is a modified random hot deck procedure. In random hot deck procedures, the 'donor' observation is chosen at random from a group of observations, usually based on the matching variables (D'Orazio et al., 2006. p.37). In this project exact matching was used, this is where the group of observations to be drawn from is established through exact matches between the values (or categories) of the 'recipient' observation's matching variables and those of the matching variables of the 'donor' dataset. This is possible as the matching variables are categorical. When no exact matching donor is available, one of the matching variables will be dropped. The list of matching variables are imputed into the function by the user in a vector, this vector dictates the order that the matching variables will be dropped, with the first one listed being dropped first. When there is still no exact matching donor, a next matching variable will be dropped. This procedure will be repeated until a donor can be found. If there are still no observations after trying to remove each of the variables

Algorithm 1 Quality measure procedure using a bootstrap method

Input: Dataset A (with Y and X) of size n_A, dataset B (with Z and X) of size n_B and a known overlap of A and B, referred to as dataset C (with Y, Z and X) of size n_C, which is included in the n_A and n_B. Then let n_A^* and n_B^* refer to the size of datasets A and B excluding the observations from dataset C.

1: Use dataset C to estimate the true proportions of the population, p_{ij}, for each cell (i, j) of the contingency table of proportions for Y $\times$ Z. This estimate is referred to as $\hat{p}_{ij}$, with row probabilities given by $\hat{p}_{i.}$ and column probabilities given by $\hat{p}_{j.}$.
repeat R times

2: Draw n_A values of the categories of Y from A (including C) with the probability of a particular value of Y being drawn based upon $\hat{p}_{i.}$, which is the estimated proportions of each value of a category in Y.

3: Draw complete observations from A (including C) to correspond to these values of Y drawn in the previous step. Do this with a probability of an observation coming from A excluding C being n_A^*/n_A and the probability that the observation comes from C being n_C/n_A. From these samples, let d_C represent the number of observations drawn from C, so the number of observations drawn from A (excluding C) was $d_A = n_A - d_C$.

4: The let $d_B = n_B - d_C$ be the number of values to be drawn from B excluding C. Draw d_B categories of Z values from B excluding C with the probability equal to $\hat{p}_{j.}$, which are the estimated proportions of each category in Z. Then sample observations from B excluding C for these respective Z values.

5: Apply the statistical matching method being tested to the constructed bootstrapped samples of A (including C) and B (including C).

6: Create a contingency table of the proportions of the statistically matched data for $Y \times Z$. This table is defined as $c_{ij}^{(r)}$, for each iterations r, with $r = 1, ...R$.
End of iterations

7: Calculate the average for each of the cells in the proportions contingency table across all bootstrapped samples. This is referred to as $\bar{c}_{ij}$. i.e.

$$\bar{c}_{ij} = \frac{1}{R} \sum_r c_{ij}^{(r)}$$

8: Then calculate the bias per cell ($\hat{b}_{ij}$), relative bias per cell ($\hat{b}_{\text{rel}_{ij}}$), variance per cell ($\hat{v}_{ij}$) and relative standard error ($\hat{s}_{\text{rel}_{ij}}$). Formally, this was found by:

$$\hat{b}_{ij} = \bar{c}_{ij} - \hat{p}_{ij}$$

$$\hat{b}_{\text{rel}_{ij}} = \frac{\bar{c}_{ij} - \hat{p}_{ij}}{\hat{p}_{ij}}, \text{ for } \hat{p}_{ij} > 0$$

$$\hat{v}_{ij} = \frac{1}{R-1} \sum_r \left(c_{ij}^{(r)} - \bar{c}_{ij} \right)^2$$

$$\hat{s}_{\text{rel}_{ij}} = \frac{\sqrt{\hat{v}_{ij}}}{\bar{c}_{ij}}, \text{ for } \bar{c}_{ij} > 0$$

Output: Bias per cell, $\hat{b}_{ij}$, relative bias per cell, $\hat{b}_{\text{rel}_{ij}}$, variance per cell, $\hat{v}_{ij}$ and relative standard error per cell, $\hat{s}_{\text{rel}_{ij}}$.

on their own, then combinations of two variables will be dropped. All combinations of two will be exhausted and if still not possible then combinations of three and so on until all variables have been dropped, and the donor value is picked at random from dataset A. In doing this a 'donor' observation can always be chosen for all observations in the 'recipient' dataset B. Therefore, always leading to an output dataset, in this case the output will have all values from Y and Z, as well as, the related $\mathbf{X}$.

The second method will be another hot deck procedure, however, this time the 'donor' observation will be chosen based on their distance to the respective 'recipient' observation based on the matching variables, $\mathbf{X}$. As the matching variables are categorical the distance measure chosen in this situation is the Gower distance. The Gower distance is a method created to look at the dissimilarities between variables, in which not all the variables are continuous, see Gower (1971, p.859). The observation with the smallest Gower distance between the matching variables, and therefore the closest, is then chosen. This method was used with the help of the function NND.hotdeck in the package StatMatch in R (D'Orazio, 2020). All methods and procedures were implemented in R (R Core Team, 2020).

3 Simulation Study

In this section the simulation study will first be outlined, followed by the data used in the simulation study (Section 3.1) and then the results of the simulation study (Section 3.2). A simulation study was performed to evaluate how close the values of the proposed quality measure are to their true counterparts. So, for instance, we compared the estimated bias with the estimated true bias.

The procedure taken in the simulation study follows an iterative procedure that is performed S times, where S is a large number. At each iteration, s, two samples are drawn from the population in each iteration, and are called datasets A and B. The sizes of datasets A and B are fixed to n_A and n_B, the overlap, dataset C, between them does not have a fixed size and is random within each pair of draws. To put into other words, the draw sizes were fixed for the size of A (including C) and B (including C). This was chosen as this is most realistic situation, as samples A and B tend to be drawn independently of each other from the population. Therefore, any overlap (C) that exists between them is random. The desired SM method is then performed on the two samples, and the contingency table of proportions is then found based on the synthetic data created with the SM method. From these S contingency tables one estimate for the true bias, variance, relative bias and relative standard deviation are estimated. The samples in each iteration then have the quality measure procedure applied to them, leading to S bootstrap estimates of bias, variance, relative bias and relative standard error. The

true values and the estimated quality measure values can then be compared. The full simulation study is outlined in Algorithm: 2, and some inspiration for this was taken from Scholtus and Daalmans (2021).

The values calculated after the simulations were used to assess the quality of the proposed procedure. This was done as follows: the bootstrapped estimates from the quality measure for variance, bias, relative bias and relative standard error were compared to their true estimates, for each cell (i, j) of the contingency table of proportions, which has just been estimated. This was done for the true estimate θ_t, for example for bias b_{ij}, comparing with the bootstrap estimates, $\theta_b^{(s)}$, for example the bootstrap bias estimate, $\hat{b}_{ij}^{(s)}$.

To compare estimates of $\theta_b^{(s)}$ to θ_t, the average absolute differences over the S bootstrap estimates was calculated. The average absolute differences over the S bootstrapped estimates was given by:

$$\text{AAD} = \frac{1}{S} \sum_{s=1}^{S} |\theta_b^{(s)} - \theta_t|$$

Additionally, in order to look at whether the sampling within the iterative scheme caused large differences in the contingency tables used to estimate the true bias and true relative bias we calculated a corrected estimator. These corrected estimators for true bias and true relative bias were found referred to as the true estimation error and true relative estimation error of a given sample. This was calculated as the difference between the samples contingency table after the SM was applied to the samples and the true proportions contingency table of the combined samples A and B:

$$\varepsilon^{(s)} = e_{ij}^{(s)} - t_{ij}^{(s)}$$

The average over the S samples is defined as $\bar{\varepsilon}$. For the relative estimation error, this was given by the same difference divided by the the true contingency table of the samples:

$$\varepsilon_{\text{rel}}^{(s)} = \frac{e_{ij}^{(s)} - t_{ij}^{(s)}}{t_{ij}^{(s)}}$$

With the average of the S samples for this given by $\bar{\varepsilon}_{\text{rel}}$. $\varepsilon^{(s)}$ and $\varepsilon_{\text{rel}}^{(s)}$ were then plotted to show the variation that exists within each cell on the contingency table estimates based on the sampling in the simulation study, and if the sampling makes a difference to the estimate of true bias and true relative bias.

Algorithm 2 Simulation study to assess quality measure procedure performance

A dataset is available for a complete population, with complete observations for all variables **X**, Y and Z. The contingency table of proportions for Y × Z for the whole population is referred to as p_{ij}.

Repeat S times

1: Draw samples A_s and B_s, of a fixed size, n_A and n_B respectively (with $s = 1, ..., S$ for each pair in the separate iterations) from the population at random without replacement. The random overlap between A_s and B_s is defined as C_s. The size of which varies in each iteration.

2: The true contingency table of proportions over samples A and B combined is referred to as $t_{ij}^{(s)}$. For sample A_s (excluding C_s) remove variable Z and for B_s (excluding C_s) remove variable Y, so that it fits the situation for statistical matching. All observations are available for sample C_s.

3: Inital steps to calculate the true bias, true relative bias, true variance and true standard deviations.

(a) Apply the SM method to A_s and B_s, including the overlap C_s.

(b) Create a contingency table of proportions for Y and Z based on these results, with each cell value referred to by $e_{ij}^{(s)}$, for cell (i, j). This proportional contingency table should be stored, and these values are then used to create an estimate for true bias and variance over all iterations, outlined after the iteration procedure.

4: Apply the proposed bootstrap approach (Algorithm 1) on A_s and B_s. Resulting in bootstrap estimates for variance ($\hat{v}_{ij}^{(s)}$), bias ($\hat{b}^{(s)}ij$), relative bias ($\hat{b}_{rel,ij}^{(s)}$) and relative standard error ($\hat{s}_{rel,ij}^{(s)}$) for each cell (i, j) of the contingency table.

End of iterations

5: Then work out the true bias (b_{ij}), mean across estimated proportions ($\bar{e}_{ij}$), the true relative bias ($b_{rel,ij}$). true variance (σ_{ij}^2) and true relative standard deviation ($\sigma_{rel_{i,j}}$) all per cell, as follows:

$$b_{ij} = \frac{1}{S}\sum_{s=1}^{S} e_{ij}^{(s)} - p_{ij}$$

$$\bar{e}_{ij} = \frac{\sum_{s=1}^{S} e_{ij}^{(s)}}{S}$$

$$b_{rel,ij} = \frac{\bar{e}_{ij} - p_{ij}}{p_{ij}}, \text{for } p_{ij} > 0$$

$$\sigma_{ij}^2 = \frac{1}{S-1}\sum_{s=1}^{S}(e_{ij}^{(s)} - \bar{e}_{ij})^2$$

$$\sigma_{rel,ij} = \frac{\sqrt{\sigma_{ij}^2}}{\bar{e}_{ij}}, \text{for } \bar{e}_{ij} > 0$$

3.1 Data for Simulation Study

For the simulation study, data was used from Statistics Netherlands on the 2016 Health Monitor (Public Health Monitor, 2016). The dataset was taken to be the whole population. The possible target variables consisted of categorical variables with less than five categories, and possible matching variables were all categorical variables focusing on common background variables. In order to decide the best target and matching variables, the similarity between variables was looked at. We used Cramer's V as a measure of similarity for this. Cramer's V between two categorical variables is defined as

$$\text{Cramer's V} = \frac{(X^2/n)}{\min(c-1, r-1)}$$

where X^2 is the Chi-squared test statistic, n is the sample size, c is the number of categories of one variable and r is the number of categories of the other. This means that the statistic ranges from 0 to 1, with zero indicating no association and one perfect association (McHugh, 2018, pp. 417 - 418).

Initially, possible matching and target variables were established from the dataset based on the categorical criteria, and where there were small amounts of missing data in the variable. The possible target variables were chosen based on their Cramer's V to each other and the matching variables. These variables were income quintiles (5 categories, 1 - 0-20%, 2 - 20-40%, 3 - 40-60%, 4 - 60-80%, 5 - 80-100%), referred to as Y, which has 2981 missing values. The other chosen target variable was General Health, which had participants rank their general health on a scale from Very good to Very bad. In order to decrease the number of categories in the Y $\times$ Z contingency table, this variable was re-coded, where very good and good, as well as, bad and very bad, were merged into two categories, resulting in three categories for the re-coded variable. General Health had 5786 missing values for this variable.

The matching variables were chosen based on their similarities to the target variables. Four matching variables were chosen. These were: Age (14 classes in 5 year age groups, from 19 to 85 or older), Ethnicity (categorical variable, with 13 categories including indigenous, various types of non-western foreign and various types of western foreign), Sex (categorical variable with 1 - Man and 2 - Woman) and Education Level (4 categories, 1 - low (LO), 2 - middle 1 (MAVO, LBO), 3 - middle 2 (HAVO, VWO, MBO), 4 - high (HBO, WO)). Their similarity to the target variables given through Cramer's V can be seen in Table 1. After the target and matching variables were established, all units with missing observations in any of the variables were removed from the dataset, leaving 421,226 observations from an original size of 457,232.

A synthetic proxy variable was then made for variable Y. This was done in order to investigate at how well the proxy variable helped in maintaining the CIA assumption,

Table 1: The similarity between the simulation study's matching variables to target variables

Matching Variables	Age	Ethnicity	Sex	Education Level
Cramer's V to Income	0.596	0.321	0.318	0.197
Cramer's V to Health	0.393	0.004	0.379	0.165

and to see if this had any impact on the quality procedures results. The proxy variable was made using a transition matrix. Where the probability of a certain category of Y changing to another category is given by each line. The transition matrix used is given by:

$$\begin{bmatrix} 0.7 & 0.075 & 0.075 & 0.075 & 0.075 \\ 0.075 & 0.7 & 0.075 & 0.075 & 0.075 \\ 0.075 & 0.075 & 0.7 & 0.075 & 0.075 \\ 0.075 & 0.075 & 0.075 & 0.7 & 0.075 \\ 0.075 & 0.075 & 0.075 & 0.075 & 0.7 \end{bmatrix}$$

So, for example in this situation the category '0-20%', will stay as category '0-20%' in the new proxy variable with the probability of 0.7, and will change to one of the other categories all with the probability of 0.075. The Cramer's V between this proxy variable and the target variable Y is 0.612. A value for close to 0.6 was chosen as this is approximately the proxy variable's Cramer's V to target variable in the case study situation (Section 4).

It was also decided to test another synthetic proxy variable with a Cramer's V of around 0.4. A lower Cramer's V was chosen as in theory the CIA will hold less well, this was something that was interesting to confirm in practice but also to investigate if this had any impacts on the estimates of the quality measure. This was tested only on the distance hot deck procedure for time reasons. In this situation the transition matrix was given by:

$$\begin{bmatrix} 0.525 & 0.11875 & 0.11875 & 0.11875 & 0.11875 \\ 0.11875 & 0.525 & 0.11875 & 0.11875 & 0.11875 \\ 0.11875 & 0.11875 & 0.525 & 0.11875 & 0.11875 \\ 0.11875 & 0.11875 & 0.11875 & 0.525 & 0.11875 \\ 0.11875 & 0.11875 & 0.11875 & 0.11875 & 0.525 \end{bmatrix}$$

The Cramer's V between this proxy variable and the target variable Y is 0.397.

In the simulation study the size of the samples n_A and n_B were chosen to be 10500, as 10500 is approximately the same proportion of the final dataset size (421226) that this dataset (421226) is of the Dutch population in 2016 (16979120 (CBS, 2021)), meaning samples A and B are approximately 0.0248 the size of the populations they are looking

Table 2: The true proportions of the contingency table for Y × Z for the data used in the simulation study.

	General Health		
Income quintiles	Good	Okay	Bad
0-20%	0.046	0.027	0.008
20-40%	0.113	0.067	0.015
40-60%	0.156	0.055	0.010
60-80%	0.189	0.047	0.007
80-100%	0.217	0.038	0.005

at.

The decision for the parameters that were used, i.e. R and S, in the simulation study and in the case study can be found in Appendix B. It was decided to use R = 200 and S = 100, as at both of these values the estimates stabilised.

3.2 Simulation Study Results

From the data in the simulation study, the true proportions of the contingency table for Y × Z can be seen in Table 2. Table 2 shows that the highest proportion of people reported having good health, followed by okay and then the smallest number reporting having bad health. The proportion of people reporting good health increased with income quintiles.

Table 3, shows the true bias (b_{ij}), the average of the bootstrap estimates, $\hat{b}_{ij}^{(s)}$ over S iterations and the absolute average difference (AAD) between the bootstrap estimates and the true bias. In each of the three sections the rows refer to the income quintiles and the columns the general health of the participants. The results on the left refer to the simulation study results using distance hot deck procedure, with a proxy variable that has a Cramer's V of approximately 0.6 to the target variable, income quintiles. The middle results refer to the simulation study results using the random hot deck procedure, also with a proxy variable that has a Cramer's V of approximately 0.6 to the target variable, income quintiles. Finally, results in the right column refer to the distance hot deck procedure, but this time with a proxy variable with a Cramer's V of approximately 0.4 to the target variable. When looking at the true bias per cell, the distance and random hot deck procedures with a proxy Cramer's V of 0.6 have similar values to each other across the cells, with absolute values ranging from 0.093 $\times 10^{-3}$ to 7.989 $\times 10^{-3}$. The distance hot deck procedure with a Cramer's V of approximately 0.4 had a wider range of values values for true bias across the cells, ranging from 0.018 $\times 10^{-3}$ to 9.322 $\times 10^{-3}$. The average estimated biases from the bootstrap procedure in each iteration ranged from an absolute value of 0.255 $\times 10^{-3}$ to 9.842 $\times 10^{-3}$ across the different SM procedures and

datasets, following similar patterns in each. The absolute difference between the true bias and the bootstrap bias estimates ranged from 3.148 $\times 10^{-3}$ to 13.453 $\times 10^{-3}$. When it came to the largest differences in bias, these were seen in the cells which have the largest true proportions in the population: primarily in the first column with the exception of Cell (1,1).

Table 3: The true bias, b_{ij}, given for each cell of the contingency table of proportions, calculated through iterations of the simulation study. Followed by the average estimated bootstrap biases over the simulation iterations, i.e the average of $\hat{b}_{ij}^{(s)}$ over the S iterations. This is followed by the average absolute difference between the bootstrap bias estimates in the simulation study and the true bias (AAD). All values given in $\times 10^{-3}$.

True Bias, b_{ij}									
	Distance Hot Deck Cramer's V of 0.6			Random Hot Deck Cramer's V of 0.6			Distance Hot Deck Cramer's V of 0.4		
Income quintiles	General Health								
	Good	Okay	Bad	Good	Okay	Bad	Good	Okay	Bad
0-20%	3.030	-5.578	-2.526	5.558	-4.439	-2.061	4.027	-6.557	-3.210
20-40%	6.463	-5.278	-2.132	7.317	-5.821	-1.892	9.180	-6.843	-2.638
40-60%	0.316	-0.093	0.678	-0.314	-0.240	0.530	0.018	-0.403	0.661
60-80%	-3.159	3.332	1.793	-4.345	3.416	1.473	-4.438	4.737	2.177
80-100%	-6.525	7.266	2.413	-7.989	6.816	1.990	-9.071	9.322	3.038
Average over the bootstrap bias estimates ($\hat{b}_{ij}^{(s)}$)									
Income quintiles	General Health								
	Good	Okay	Bad	Good	Okay	Bad	Good	Okay	Bad
0-20%	4.200	-6.171	-3.574	3.928	-4.091	-2.408	2.226	-5.633	-2.959
20-40%	5.183	-4.513	-1.199	7.035	-8.217	-1.620	9.842	-8.381	-2.031
40-60%	-0.793	1.195	0.255	-0.412	0.851	0.517	1.953	-3.104	2.656
60-80%	-3.094	3.157	2.567	-3.923	3.297	1.422	-5.498	4.801	1.755
80-100%	-5.472	6.313	1.945	-6.631	8.146	2.104	-8.523	9.540	2.951
Average absolute difference between true bias and bootstrap bias estimates, AAD									
Income quintiles	General Health								
	Good	Okay	Bad	Good	Okay	Bad	Good	Okay	Bad
0-20%	6.665	7.116	4.174	7.190	6.457	3.669	6.509	5.739	3.631
20-40%	11.275	10.409	4.651	10.238	10.270	4.382	10.061	8.882	4.309
40-60%	11.872	8.843	4.088	12.017	9.128	4.468	9.333	7.959	4.159
60-80%	11.290	9.753	3.851	13.155	8.363	3.861	9.627	8.095	3.705
80-100%	13.440	9.192	3.435	13.453	8.146	3.148	10.060	9.112	3.495

As an example the large spread in the bootstrap bias estimates, $\hat{b}_{ij}^{(s)}$, across the simulation iterations can be seen in Figure 4 using the random hot deck procedure. From this graph, it can be seen that for a given sample individual bootstrap estimates of the bias can range widely, even though the average of the bootstrap estimates is often very close to the true estimate.

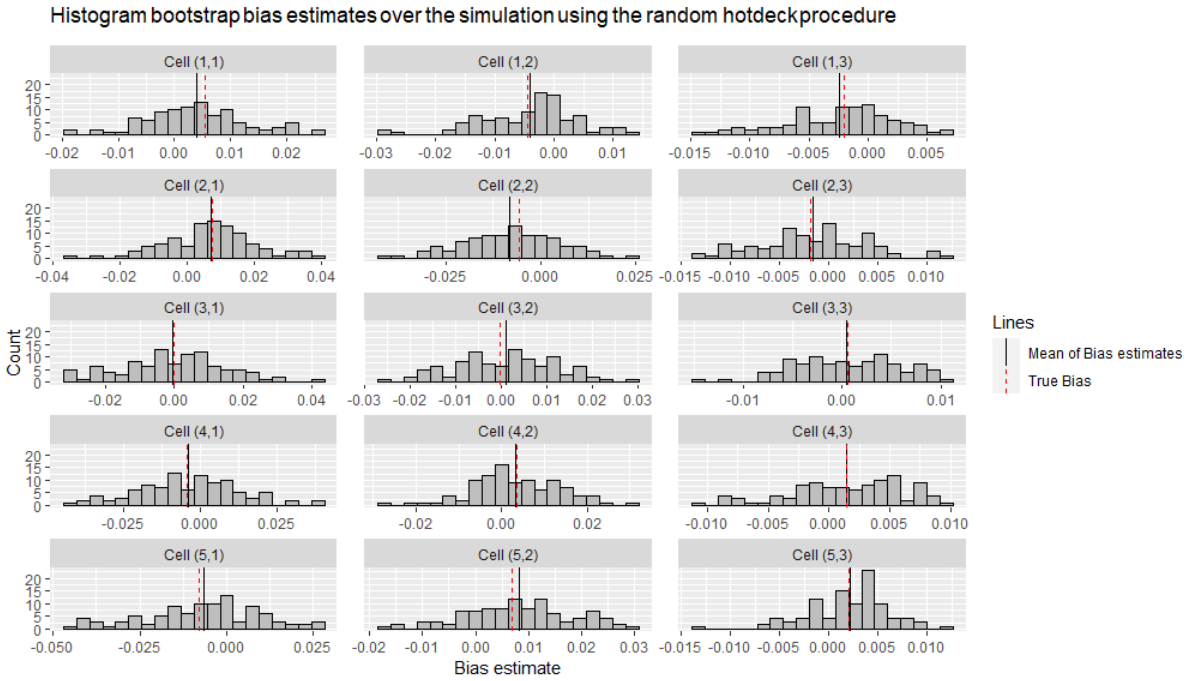

Figure 4: Histogram showing spread of bootstrap bias estimates ($\hat{b}_{ij}^{(s)}$) over the S=100 iterations of the simulation study using the random hot deck procedure, with proxy variable to income with Cramer's V of 0.6. The average over these bootstrap estimates is shown in black and the True Bias (b_{ij}) is given with a red dotted line.

Table 4 shows the true relative bias per cell ($b_{rel,ij}$), followed by the average of the bootstrap estimates for relative bias, ($\hat{b}_{\text{rel}_{ij}}^{(s)}$), over the S simulation iterations. Finally the table shows the average absolute difference (AAD) between the bootstrap relative bias estimates and the true relative bias. The table's structure is the same as bias for presenting the various SM methods tested. For true relative bias there are differences between some cells for the distance and random hot deck procedures with Cramer's V of 0.6, while other cells have similar values. Across these two SM methods with this same proxy, the relative bias ranges from 0.2% to 46.4%. This shows in practice, some of the SM methods have a true relative bias that can be considerably larger in some cells. As for bias, the relative bias has much wider range for the distance hot deck with a Cramer's V of 0.4, with most cells performing worse but some performing better. Here the relative bias ranges from 0% to 58.5%. The average bootstrap relative bias estimates over the iterations ranged from 0.7% to 36.9%, again showing a wide range across the SM

Table 4: The true relative bias, $b_{rel_{i,j}}$, given for each cell of the contingency table of proportions, calculated over the iterations of the simulation study. Followed by the average bootstrap relative bias estimates over the simulation study iterations, i.e the average of the $\hat{b}_{\text{rel}_{ij}^{(s)}}$ estimates over the S iterations. This is followed by the average absolute difference between the bootstrap relative bias estimates and the true relative bias estimates, AAD. The '.' represents where values are undefined, values are undefined when at least one of simulation study iteration's bootstrap estimate for that cell was undefined. This happened because in the overlap of that simulation sample, the proportion for that cell was zero, so for that cell ($\hat{p}_{ij}$) was zero. This means that the relative bias is not able to be defined as the value has been divided by zero.

True relative bias, $b_{rel_{i,j}}$									
	Distance Hot Deck Cramer's V of 0.6			Random Hot Deck Cramer's V of 0.6			Distance Hot Deck Cramer's V of 0.4		
Income quintiles	General Health								
	Good	Okay	Bad	Good	Okay	Bad	Good	Okay	Bad
0-20%	0.066	-0.207	-0.315	0.120	-0.164	-0.257	0.087	-0.243	-0.400
20-40%	0.057	-0.078	-0.142	0.065	-0.086	-0.126	0.081	-0.102	-0.176
40-60%	0.002	-0.002	0.071	-0.002	-0.004	0.056	0.000	-0.007	0.069
60-80%	-0.017	0.071	0.255	-0.023	0.072	0.209	-0.024	0.101	0.310
80-100%	-0.030	0.191	0.464	-0.037	0.179	0.383	-0.042	0.245	0.585
Average over the bootstrap relative bias estimates ($\hat{b}^{(s)}_{\text{rel}_{ij}}$)									
Income quintiles	General Health								
	Good	Okay	Bad	Good	Okay	Bad	Good	Okay	Bad
0-20%	0.147	-0.068	.	0.143	-0.040	.	0.116	-0.096	.
20-40%	0.065	-0.022	0.137	0.081	-0.083	.	0.107	-0.092	.
40-60%	0.008	0.081	.	0.012	0.071	.	0.022	0.031	.
60-80%	-0.007	0.169	.	-0.010	0.140	.	-0.023	0.157	.
80-100%	-0.018	0.290	.	-0.023	0.359	.	-0.036	.	.
Average absolute difference between true relative bias and bootstrap relative bias estimates, AAD									
Income quintiles	General Health								
	Good	Okay	Bad	Good	Okay	Bad	Good	Okay	Bad
0-20%	0.185	0.299	.	0.193	0.267	.	0.180	0.249	.
20-40%	0.109	0.160	0.417	0.100	0.142	.	0.105	0.120	.
40-60%	0.078	0.193	.	0.080	0.195	.	0.065	0.156	.
60-80%	0.062	0.271	.	0.069	0.222	.	0.048	0.206	.
80-100%	0.062	0.320	.	0.061	0.339	.	0.045	.	.

methods. There were also values which were undefined, values were undefined when at least one of iteration's estimated values in that cell had an estimated proportion of the population for that given cell ($\hat{p}_{ij}$) of zero. This means that the relative bias is not able to be defined as the value has been divided by zero. The average absolute difference between the bootstrap relative bias estimates and the true relative bias over all methods and datasets ranged from 4.5% to 41.6 %, with some cells undefined for the same reason as previously mentioned. Most cells, however, have an AAD of around 10 %.

When looking at the relative bias, the cells with the smallest true proportion in the population performed worst, which was the reverse of bias. These cells are mostly the cells in the third column and cell (5,2). This is not a surprising result, as although these cells had the smaller bias results, these bias results where not small enough to mitigate that their proportion in the whole population is smaller than other cells in the proportions contingency table, which leads to much larger relative bias.

Figure 5 shows the spread of relative bias bootstrap estimates across the simulation study's iterations. Similar conclusions as with bias can be drawn here that individual estimates can vary considerably, even though the average of the bootstrap estimates is often close to the true value.

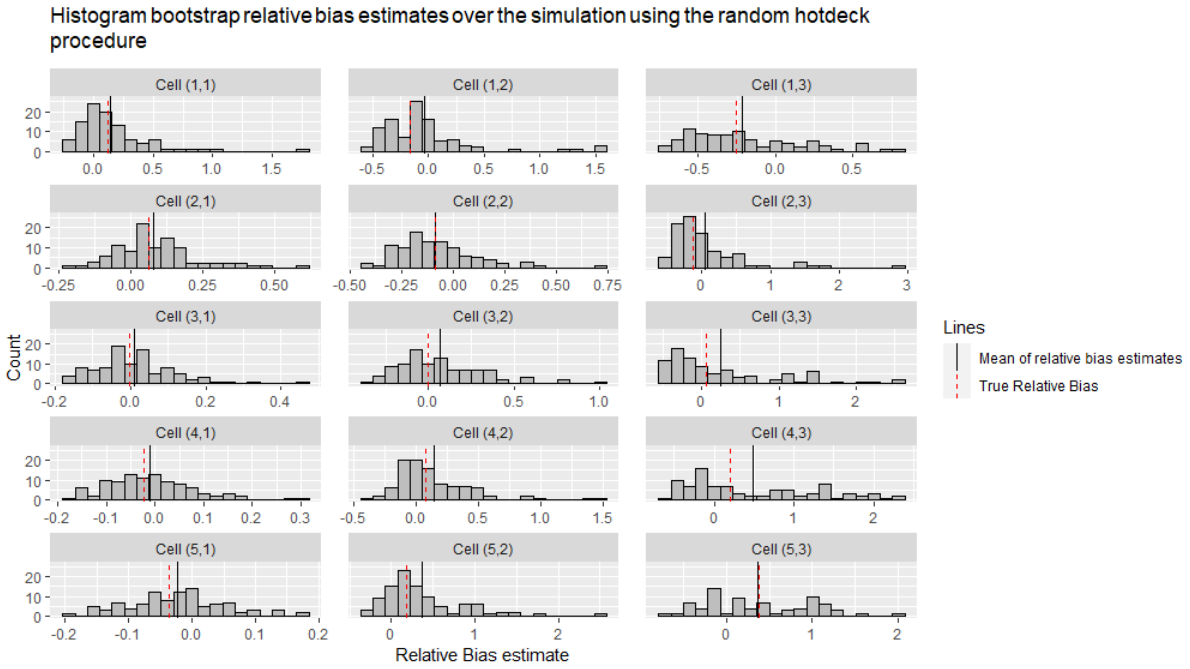

Figure 5: Histogram showing spread of relative bias bootstrap estimates ($\hat{b}^{(s)}_{\text{rel}_{ij}}$) over the S=100 iterations of the simulation study using the random hot deck procedure, with proxy variable to income with Cramer's V of 0.6. The average over these bootstrap estimates is shown in black and the True Relative Bias ($b_{rel,ij}$) is given with a red dotted line. Undefined values where removed from the histogram, and were also removed when calculating the mean value line given in this graph.

Table 5 gives the true variance for each cell, σ_{ij}^2, followed by the average over the bootstrap estimates for variance, $\hat{v}_{ij}^{(s)}$, in the S iterations. Lastly, the average absolute differences between the bootstrap variance estimates and the true variance are given. The tables are structured similarly to the previous tables. For the true variance, the two distance and random hot deck procedures with proxy variables of 0.6 to the target variable and the distance hot deck with a proxy variable of 0.4 to the target variable have similar results and trends across the cells. Over all three cases the variance ranged from 0.467 $\times 10^{-6}$ to 24.275 $\times 10^{-6}$. The average bootstrap variance estimates ranged from 0.596 $\times 10^{-6}$ to 24.594 $\times 10^{-6}$. The average absolute difference between the bootstrap variance estimates and the true variance ranged from 0.139 $\times 10^{-6}$ to 6.088 $\times 10^{-6}$.

Table 5: The true variance, σ^2_{ij} given for each cell of the contingency table of proportions, calculated over the iterations of the simulation study, followed by the average bootstrap variance estimates, i.e. the average of the estimates $\hat{v}^{(s)}_{ij}$ over the S simulation iterations. Then the average absolute difference between the bootstrap variance estimate and the true variance, AAD. All values given in $\times 10^{-6}$.

True Variance, σ^2_{ij}									
	Distance Hot Deck Cramer's V of 0.6			Random Hot Deck Cramer's V of 0.6			Distance Hot Deck Cramer's V of 0.4		
Income quintiles	General Health								
	Good	Okay	Bad	Good	Okay	Bad	Good	Okay	Bad
0-20%	5.886	2.189	0.528	8.034	1.970	0.625	7.660	2.606	0.467
20-40%	14.382	5.312	1.083	10.819	7.468	1.100	13.822	6.417	0.969
40-60%	16.381	6.117	1.036	17.338	5.449	0.718	15.302	6.051	1.062
60-80%	22.316	5.718	0.912	19.916	5.127	0.998	19.672	5.625	0.870
80-100%	24.275	4.532	0.512	19.494	5.777	0.957	18.514	4.218	0.863
Average over the bootstrap variance estimates ($\hat{v}^{(s)}_{ij}$)									
Income quintiles	General Health								
	Good	Okay	Bad	Good	Okay	Bad	Good	Okay	Bad
0-20%	7.214	2.466	0.596	7.479	2.598	0.612	7.552	2.495	0.491
20-40%	14.342	6.390	1.354	15.039	6.429	1.306	15.820	6.688	1.235
40-60%	18.546	5.839	1.073	18.279	5.818	0.994	20.405	6.192	1.014
60-80%	20.977	5.241	0.919	21.339	5.212	0.885	23.636	5.918	0.937
80-100%	22.524	4.742	0.804	21.912	4.681	0.723	24.594	5.288	0.816
Average absolute difference between true variance and bootstrap variance estimates, AAD									
Income quintiles	General Health								
	Good	Okay	Bad	Good	Okay	Bad	Good	Okay	Bad
0-20%	1.472	0.435	0.176	1.006	0.637	0.161	1.098	0.490	0.139
20-40%	1.263	1.189	0.408	4.224	1.229	0.326	2.309	0.969	0.402
40-60%	2.437	0.765	0.284	1.706	0.786	0.313	5.130	0.887	0.282
60-80%	2.234	0.836	0.229	2.186	0.625	0.244	4.150	0.869	0.242
80-100%	2.536	0.686	0.316	2.763	1.163	0.264	6.088	1.180	0.232

Table 6 gives the true relative standard deviation of each cell, $\sigma_{rel,ij}$, followed by the average of the bootstrap relative standard error estimates, $\hat{s}^{(s)}_{\text{rel}_{ij}}$ over the S iterations. Finally, the average absolute difference between the bootstrap relative standard error estimates and the average true relative standard deviation (AAD) is given. The SM methods are ordered similarly to Tables 1-3. For the true relative standard deviation, it can be seen that all methods and proxy variables had similar results for respective cells with values ranging from 2.1% to 13.6%. The average bootstrap relative standard error range from 2.2% to 16.0% over the various SM methods. The average absolute difference between the bootstrap relative standard error estimates and the average true relative standard deviation ranged from 0.1 % to 2.2 %.

Looking at the variance and relative standard deviation together, it can be seen that they follow a similar pattern to the bias and relative bias, in that the cells which have the smallest true variance, average bootstrap estimates and average absolute difference for variance, have the largest estimates for these values for true relative standard deviation and average bootstrap standard error. In the case of variance the cells, with the smallest true variance have the smallest proportion of the population. The opposite can be said for relative standard deviation, as the average over the iterations contingency tables will be smallest in these cells, this small value increases the size of the values for relative standard deviation.

In this paper, the simulation study used synthetic proxy variables, to investigate their impact in mitigating the CIA. Two different datasets were used with the distance hot deck procedure with Cramer's V values of 0.4 and 0.6 between the income quintiles target variable and the synthetic proxy. As expected over most cells the dataset with a proxy variable of Cramer's V 0.4 to target variable had higher true bias and true relative bias compared to the dataset with a Cramer's V of 0.6 to target variable. This was expected as the proxy variable of Cramer's V of 0.6 to target variable would better maintain the CIA. The variance and relative standard deviation had similar results over datasets with both proxy variables, this is an interesting result and requires more thought as to why this is the case.

Table 6: Table giving the true relative standard deviation , $\sigma_{rel_{i,j}}$, over the simulation study iterations, followed by the average over the bootstrap estimates for relative standard error, i.e. the average of the $\hat{s}^{(s)}_{\text{rel}_{ij}}$ estimates over the S iterations, then the average absolute difference between the bootstrap relative standard error estimates and the true relative standard deviation, AAD.

True relative standard deviation, $\sigma_{rel_{i,j}}$									
	Distance Hot Deck Cramer's V of 0.6			Random Hot Deck Cramer's V of 0.6			Distance Hot Deck Cramer's V of 0.4		
Income quintiles	General Health								
	Good	Okay	Bad	Good	Okay	Bad	Good	Okay	Bad
0-20%	0.049	0.069	0.132	0.055	0.062	0.133	0.055	0.080	0.142
20-40%	0.032	0.037	0.081	0.027	0.044	0.080	0.030	0.042	0.080
40-60%	0.026	0.045	0.100	0.027	0.042	0.084	0.025	0.045	0.101
60-80%	0.025	0.047	0.108	0.024	0.045	0.117	0.024	0.046	0.101
80-100%	0.023	0.047	0.094	0.021	0.054	0.136	0.021	0.043	0.113
Average over the bootstrap relative standard error estimates ($\hat{s}^{(s)}_{\text{rel}_{ij}}$)									
Income quintiles	General Health								
	Good	Okay	Bad	Good	Okay	Bad	Good	Okay	Bad
0-20%	0.055	0.076	0.142	0.053	0.074	0.140	0.057	0.079	0.160
20-40%	0.032	0.041	0.088	0.032	0.042	0.091	0.033	0.042	0.095
40-60%	0.028	0.044	0.101	0.027	0.045	0.105	0.030	0.046	0.106
60-80%	0.025	0.046	0.108	0.025	0.046	0.112	0.026	0.045	0.109
80-100%	0.022	0.048	0.115	0.022	0.050	0.122	0.024	0.048	0.116
Average absolute difference between true relative standard deviation and bootstrap relative standard error estimates, AAD									
Income quintiles	General Health								
	Good	Okay	Bad	Good	Okay	Bad	Good	Okay	Bad
0-20%	0.007	0.008	0.021	0.004	0.012	0.016	0.004	0.008	0.026
20-40%	0.002	0.005	0.012	0.004	0.003	0.013	0.003	0.003	0.017
40-60%	0.002	0.003	0.013	0.002	0.004	0.021	0.005	0.003	0.013
60-80%	0.001	0.003	0.014	0.001	0.003	0.014	0.002	0.003	0.014
80-100%	0.001	0.003	0.022	0.002	0.005	0.020	0.003	0.005	0.015

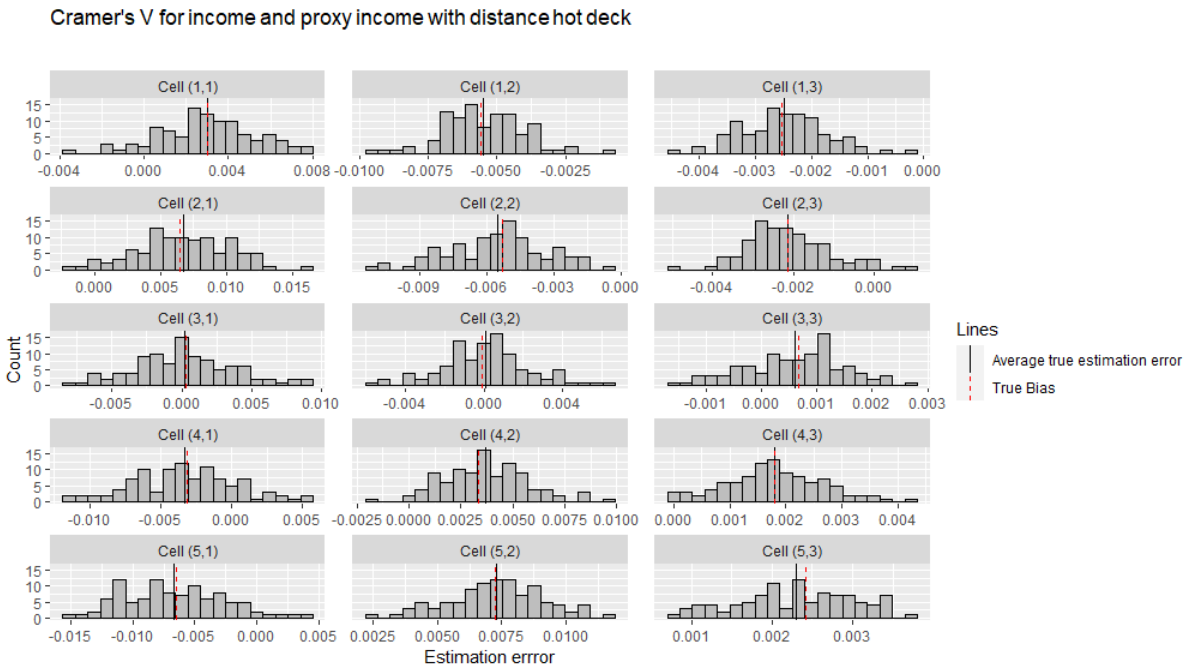

Figure 6: Histogram showing the distribution of estimation error over the S sampling replicates, given as $\varepsilon^{(s)} = e_{ij}^{(s)} - t_{ij}^{(s)}$, where (s) refers a results in a given iteration. The red dotted line shows the true bias b_{ij}, also shown in the other graphs. The black line shows average true estimation error, $\bar{\varepsilon}$. This is for the distance hot deck procedure with a Cramer's V between proxy and target variable of 0.6.

An attempt to understand the estimation error ($\varepsilon^{(s)}$) caused by the sampling within the iterations of the simulations can be seen in Figure 6. These graphs show the distribution of the estimation error for each iteration, and also the average in these estimation error ($\bar{\varepsilon}$) is given, as well as, the true bias (b_{ij}) as reference. Figure 7 shows similar graphs for the relative estimation error ($\varepsilon_{\text{rel}}^{(s)}$), with the distribution of relative estimation error shown over each iteration, and the average ($\bar{\varepsilon}_{\text{rel}}$) of these given, as well as, the true relative bias ($b_{rel_{ij}}$). These graphs show that the estimation error is in most cases very similar to the true bias, this shows that sampling error does not have too big an effect on the calculation of the true bias, however for some cells there are larger differences showing that this does have an effect overall. The relative estimation error shows a very similar pattern to this.

The random and distance hot deck method both performed similarly. This was shown through the absolute difference between the bootstrapped estimates and the true estimates were similar across all cells of the contingency tables of proportions for bias, relative bias, variance and relative standard deviation, with the same cells recording similar values in difference. As these methods are very similar to each other this is not a surprise, as both methods randomly pick observed values from one dataset based on the

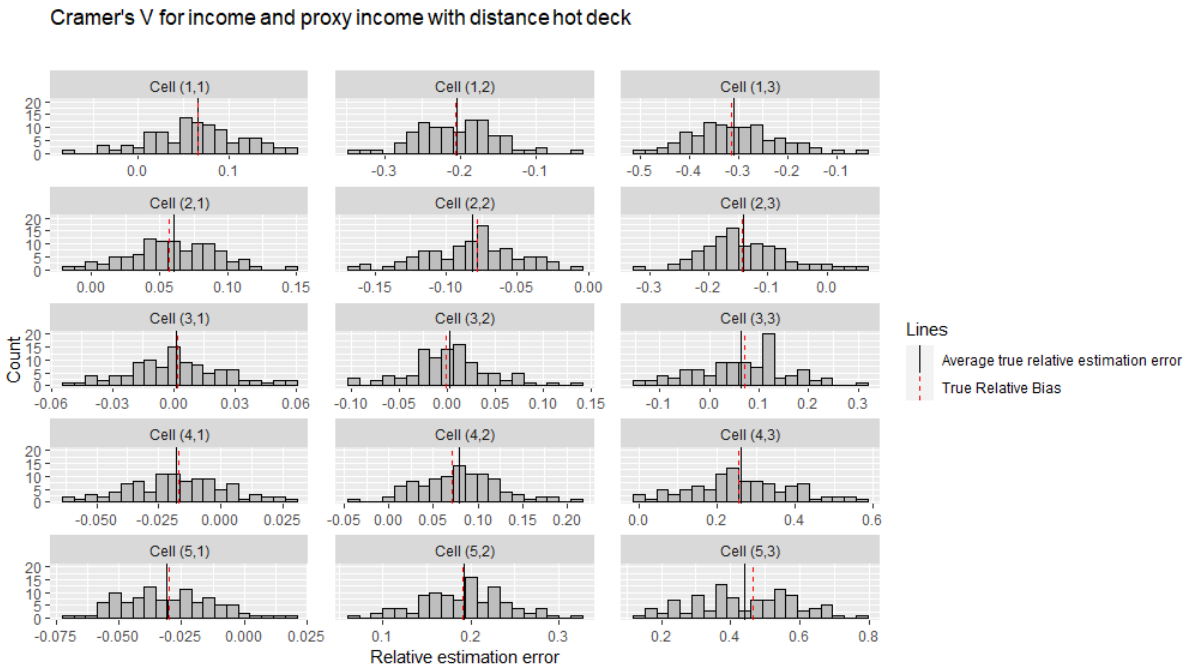

Figure 7: Histogram showing the relative estimated error over the S sampling replicates, given as $\varepsilon^{(s)} = (e_{ij}^{(s)} - t_{ij}^{(s)})/t_{ij}^{(s)}$, where (s) refers a results in a given iteration. The red dotted line shows the relative true bias $b_{rel_{i,j}}$, also shown in the other graphs. The black line shows average true relative estimation error, $\bar{\varepsilon}_{\text{rel}}$. This is for the distance hot deck procedure with a Cramer's V between proxy and target variable of 0.6.

relationship between the matching variables.

4 Case Study at Statistics Netherlands

This section will look at the quality measure procedure tested in a case study situation at Statistics Netherlands. Section 4.1 will outline the data used, and then section 4.2 will look at the results of this case study.

4.1 Data used in Case Study

The proposed quality measure was tested on a real case study from Statistics Netherlands. This was tested on the same two SM methods. The data used was the Labour Force Survey from 2016, which looks into workforce information variables, and the Health Monitor from 2016 (Public Health Monitor, 2016), which is a collection of information on the health of people in the Netherlands. In this example, we also used a proxy variable. As a method of attaining an appropriate proxy variable, both datasets were directly linked to on administrative data on jobs and wages. The administrative data is submitted by employers on data relating to their employees for each job they have.

The registry itself collects data over varying time periods for each individual, however, the registry data used here had been re-coded by the CBS so that all observations are for one person for one job over a month period. The Health Survey and Labour Force Survey data are both available for 2016. It was decided that for simplicity reasons, that observations would be linked only for the month of October in the administrative data. October was chosen as the Health Survey had taken place from September to December 2016, and the Labour Force Survey had happened over the whole course of 2016. The direct linkage between both surveys and the registry data was possible as all three datasets had a random identification number (RIN), as an identifying variable for each individual. We restricted the data to the target population that was common to all three datasets: persons that are working for an employer. For clarity, this excludes self-employed people as they are not recorded in the registry data. It was decided the variable number of hours worked in a week, would be used as a proxy, as it existed in both the registry and Labour Force Survey. It was chosen over other variables as it had the highest Cramer's V to related variables in the Labour Force Survey. In the registry data, if the same person was recorded twice in the registry data, for example because they have multiple jobs, the values for both these relevant variables were added together. After linkage, these variables were transformed so they were in the same categories over the same time period as the variables in the Labour Force data. More specifically, the variable of number of hours in the registry data was transformed to number of hours within a week, as it had been within the month. This was done by multiplying by 12 and dividing by 52. It was then split into the same categories as the variable in the Labour force survey, which were '0-11 hours', '12-23 hours', '24-34 hours' and '35+ hours'. The Cramer's V between the proxy variable and the number of hours worked in the Labour Force Survey was found to be 0.649. This proxy variable was added to the Health Monitor in the same way.

Possible matching variables were then established, these were categorical variables that were found in both datasets that were defined the same or could be re-coded to be defined in the same categories. The list of possible matching variables were: Sex, age, marital status, immigrant generation, highest education level, ethnicity and household composition. Household makeup was instantly excluded due to a very high amount of missing data (over 50%) in the Health survey. In order to establish which of these six possible matching variables would be used, the Cramer's V between them and the numbers of hours worked variable, as well as, other possible target variables was found (Table 7).

The possible target variables from the health survey were chosen based on variables in the Health monitor, which were already categorical and had few categories. There were 10 possible variables chosen from the health survey, most related to alcohol consumption,

Table 7: Relationship between possible matching variables and combination of possible matching variables against the target variables in the Case study data. These values are given as Cramer's V.

Variable Combination	Cramer's V to hour's worked	Cramer's V to BMI
Immigrant Generation	0.042	0.033
Ethnicity	0.046	0.031
Marital status	0.134	0.077
Sex	0.514	0.179
Age	0.170	0.109
Education level	0.162	0.206
Sex & Education level	0.277	0.184
Age & Education level	0.281	0.202
Age & Sex	0.370	0.129
Age & Education level & Sex	0.382	0.214

smoking and a persons Body Mass Index (BMI). Looking at the Cramer's V values across each of the possible matching variables against the possible target variables, none stood out as any better or worse than the others, all had similar Cramer's V values across the possible matching variables. It was, therefore, decided to choose the target variable based on the variable with the least missing values, BMI, which had considerably less missing values (23) than the other variables which all had several thousand. Therefore, BMI was chosen as the other target variable.

Out of the six possible matching variables, the Cramer's V was then found for each against the two target variables. Two of the matching variables had Cramer's V values of almost zero for both target variables, these were ethnicity and immigrant generation. Marital status had a reasonable Cramer's V with number of hours worked, but not with BMI. The final three possible matching variables were sex, age and education level. It was decided to possibly use these three variables, either all three or a combination of the three. In order to decide, which variation would be used, these variables were merged into variables in all combinations and the value with the highest Cramer's V was used as matching variable(s). The results of the combinations of matching variables and the Cramer's V to target variables is shown in Table 7. From this it was decided to use all of the possible matching variables together.

For ease it was decided to remove all observations with missing or unknown values in either the target variables or in the matching variables. This lead to in the end 51,577 observations in the Labour Force survey and 148,068 observations in the Health Monitor, the original datasets before removal of missing and unknown data and constriction to the working population had 125,874 and 457,232 observations respectively. The overlap of participants in the datasets of the two surveys used was very small: 67.

Table 8: Estimated Contingency table after distance hot deck and random hot deck SM methods have been implemented on the entire Case study dataset.

	Distance Hot Deck				Random Hot Deck			
BMI	Number of hours worked							
	0-11	12-23	24-34	35+	0-11	12-23	24-34	35+
upto 18.5	0.002	0.003	0.004	0.006	0.002	0.003	0.004	0.005
18.5-20	0.006	0.008	0.015	0.019	0.005	0.009	0.016	0.018
20-25	0.030	0.067	0.129	0.237	0.029	0.065	0.129	0.239
25-30	0.016	0.044	0.085	0.201	0.017	0.045	0.087	0.202
from 30	0.006	0.021	0.035	0.066	0.007	0.021	0.033	0.064

The data was then used on both SM methods once, to get estimates of the contingency tables, as would be used in a realistic situation. The quality measure procedure was then used on the data using both the distance and random hot deck procedures. Simple random sample was assumed for simplicity, so weights were not taken into account.

4.2 Results of Case Study

In this section we look at the results of the quality measure in the case study with data from Statistics Netherlands. All tables in this section are arranged with the results from the distance hot deck procedure on the left and then the random hot deck procedure on the right. Table 8 shows the estimated contingency tables after both SM methods had been performed on the complete data once. Both methods gave very similar predictions for the proportions, with the highest proportion for both being for a BMI of '20-25' and number of hours worked of '35+', and the lowest for both being a BMI of 'upto 18.5' and number of hours worked of '0-11'. The estimated proportions range from 0.002 to 0.239 across the two hot deck procedures.

Table 9 shows the contingency table of the proportions among the overlap values (C), which is used as the estimate for the true population proportions $\hat{p}_{ij}$. Here it can be seen that the first row and the first two values of the second row all had no values in C. The other values in column one and row two also do not have large proportions, while the largest proportion is given in a BMI of '25-30' and number of hours worked of '35+'.

Table 9: Estimated contingency table from the overlap of the case study data, C.

BMI	Number of hours worked			
	0-11	12-23	24-34	35+
upto 18.5	0.000	0.000	0.000	0.000
18.5-20	0.000	0.000	0.015	0.015
20-25	0.030	0.060	0.149	0.194
25-30	0.015	0.030	0.104	0.284
from 30	0.015	0.015	0.045	0.030

Table 10: The bootstrap bias estimates ($\hat{b}_{ij}$) on the case study data, for both SM methods. All values given in $\times 10^{-3}$

	Distance Hot Deck				Random Hot Deck			
BMI	Number of hours worked							
	0-11	12-23	24-34	35+	0-11	12-23	24-34	35+
up to 18.5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
18.5-20	3.610	4.190	-2.529	-1.949	3.650	4.216	-2.479	-1.914
20-25	0.050	-12.794	-3.534	25.272	0.130	-12.851	-3.404	25.303
25-30	5.542	10.345	16.365	-46.180	5.548	10.419	16.060	-46.364
from 30	-9.167	-1.848	-10.146	22.772	-9.256	-1.787	-10.209	22.937

Table 10 gives the bootstrap bias estimate ($\hat{b}_{ij}$) for the case study data. The absolute values of estimated bias ranged from 0 to 46.550 $\times 10^{-3}$. Table 11 gives the bootstrap relative bias estimate ($\hat{b}_{\text{rel}_{ij}}$). The bootstrap relative bias estimates ranged from 0.1% to 77.3%. There are also values which are undefined for relative bias, this is because the estimated proportion of the population for that given cell ($\hat{p}_{ij}$) is zero. As there were no values in the overlap for these cells, see Table 9. This means that the relative bias is not able to be defined as the value has been divided by zero. In this situation the undefined values are the first row and the first two values of the second row. Table 12 gives the bootstrap variance estimate (σ^2_{ij}). The bootstrap variance estimates on the case study data ranged 0 to 5.937 $\times 10^{-6}$ across the two SM methods. Table 13 gives the bootstrap relative standard error estimate ($\hat{s}_{\text{rel}_{ij}}$). The bootstrap relative standard error estimates ranged from 1.0% to 8.8% across the two SM methods, there are also undefined values for relative standard error, this is because the average of the contingency tables in the bootstrap samples ($\bar{c}_{ij}$) was zero for these values.

The first row of bootstrap estimates for bias and variance is zero, this was because the first category of BMI, 'up to 18.5', has no observations in the overlap. This means that the estimated population proportions contingency table, $\hat{p}_{ij}$, seen in Table 9, are zero for all cells in the first row of the proportions contingency table. As a result of this,

no bootstrap samples will come from this category as the bootstrap method used selects values of target variable Y, BMI, proportionally based on the row sums ($\hat{p}_{i.}$). This means that the zero values for the first row of both variance and bias, stem from the fact that the proportions contingency tables in all of the matched bootstrapped samples datasets

Table 11: The bootstrap relative bias estimate ($\hat{b}_{\text{rel}_{ij}}$) on the case study data for both SM methods. The '.' represents where values are undefined, values are undefined when the estimated proportion of the population for that given cell ($\hat{p}_{ij}$) is zero, so relative bias estimate is undefined as the value has been divided by zero.

	Distance Hot Deck				Random Hot Deck			
BMI	Number of hours worked							
	0-11	12-23	24-34	35+	0-11	12-23	24-34	35+
upto 18.5	.	.	.	.	.	.	.	.
18.5-20	.	.	-0.169	-0.131	.	.	-0.166	-0.128
20-25	0.002	-0.214	-0.024	0.130	0.004	-0.215	-0.023	0.130
25-30	0.371	0.347	0.157	-0.163	0.372	0.349	0.154	-0.164
from 30	-0.614	-0.124	-0.227	0.764	-0.620	-0.120	-0.228	0.768

Table 12: The bootstrap variance estimate ($\hat{v}_{ij}$) on the case study data for both SM methods, all values given in $\times 10^{-6}$

	Distance Hot Deck				Random Hot Deck			
BMI	Number of hours worked							
	0-11	12-23	24-34	35+	0-11	12-23	24-34	35+
upto 18.5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
18.5-20	0.119	0.115	0.363	0.569	0.121	0.117	0.367	0.475
20-25	0.825	1.277	3.119	4.411	0.862	1.108	3.668	5.138
25-30	0.621	1.113	2.945	5.510	0.559	0.967	3.065	5.521
from 30	0.159	0.515	1.197	1.686	0.189	0.481	1.321	1.800

Table 13: The bootstrap relative standard error estimate ($\hat{s}_{\text{rel}_{ij}}$) on the case study data, for both SM methods. The '.' represents where values are undefined, values are undefined when the average of the proportions contingency tables in the bootstrap samples ($\bar{c}_{ij}$) is zero, and so the value has been divided by zero.

	Distance Hot Deck				Random Hot Deck			
BMI	Number of hours worked							
	0-11	12-23	24-34	35+	0-11	12-23	24-34	35+
upto 18.5	.	.	.	.	.	.	.	.
18.5-20	0.096	0.081	0.049	0.058	0.095	0.081	0.049	0.053
20-25	0.030	0.024	0.012	0.010	0.031	0.022	0.013	0.010
25-30	0.038	0.026	0.014	0.010	0.037	0.024	0.015	0.010
from 30	0.069	0.055	0.032	0.025	0.077	0.053	0.033	0.025

will be zero for this category. This also explains why there are undefined values for the relative standard error, where the mean of the matched contingency tables is zero, leading to the estimate for relative standard error that is divided by zero and therefore undefined.

As with the simulation study, both of the SM methods tested showed similar results. This is unsurprising as they are have similar methodology in randomly picking observed values based on the matching values used.

5 Discussion

The aim of this study was to find a method to estimate the quality of a SM method on a particular dataset, where simulation studies cannot be carried out. A quality measure procedure like this helps to determine if a contingency table estimated through a SM method is accurate enough for a particular dataset, i.e. have a small enough variance and bias. The proposed quality measure procedure described relied on an overlap in units between the two datasets, and was designed solely for categorical target variables. The quality procedure aimed to estimate the bias and variance of the contingency tables of proportions for categorical target variables after a SM method is applied.

This proposed quality measure was tested in a simulation study with two different SM methods, using data from Statistics Netherlands that was assumed to be a complete population. The simulation study found that the estimations for the relative bias had a average absolute difference (AAD) from the true relative bias of 5-34% over the cells, but over most cells were around 10%; this is a very promising sign in the usefulness of the method going forward. In particular it attempts to assess Rässler's point (1), where 'the true values of the Y variable for the recipient observation are reproduced' (Rässler, 2002,

p.30) without the need for a simulation study, although it is not completely accurate. Particularly considering the information available without this quality measure, where most researchers have almost nothing to go on when considering if a SM method should be used for a dataset. They only have uncertainty bounds for the results based on the data which do not vary based on SM method (D'Orazio, 2019), although it would be good to see how uncertainty bounds relate to this measure in the future.

Although the results are promising, there is a large range of difference between cells. Suggesting that for at least some cells the bootstrap estimated bias may be a considerably inaccurate representation of the true values. This is an unsurprising result as the procedure assumes that the 'true' population proportions are based on the overlap of the two datasets, which is obviously not an accurate representation of the true population. If they had been large enough samples on their own, SM would not be needed and the overlap would be used solely for analysis. However, the overlap is used here as there is no good alternative.

The bias was found to have the largest difference between true and average bootstrap estimated values in cells that had the largest true proportions in the population. This is presumably because, when the bootstrap procedure is estimating the 'true' population from the overlap, the random nature and small size of this overlap is far more likely to underestimate or overestimate these cells in the different simulation samples, leading to larger differences in the bias estimates. This is also supported by the average results over the variance estimates, where these cells with the highest absolute bias difference are also the cells which have the highest true variance and average bootstrapped estimated variance, showing that these cells vary much more than others.

By contrast the results of relative bias found that the cells with the smallest true proportion of the population performed worst. This is because although they had smaller bias, their population proportion was so small that the relative bias was much larger. There were many undefined cells for relative bias. Cells have undefined values when in at least one of the S iterations, the overlap has a zero in that cell. This emphasises the problems that can arise from a small overlap. When looking at these values in practice, it may be useful to look at the contingency table of the overlap, as to gain an indication whether a high value for relative bias is based on a small number of observations in the overlap, or whether it shows a greater problem with the SM method used instead, although this should be done with caution as it could never be truly known.

The graphs of the spread of bias and relative bias bootstrap estimates show that there is a wide range of estimates, depending on the samples chosen. Practically this can be an issue, as you will have no idea how inaccurate or accurate an individual estimate is based on your datasets, even though on average you know the bootstrap estimate would be relatively accurate.

Estimation error looked to understand the effects of the samples on estimations of true bias and true relative bias. It could be seen that the average value of the estimation error ($\bar{\varepsilon}$) was in general very similar to the true bias (b_{ij}). However, in some cells there were slight differences, showing that sampling error did play a role in this estimation. This is particularly true for Cell (5,3). This is presumably because this cell had the lowest true proportion of the population and so the sampling had the biggest impact on this cell. The same takeaways come from looking at the relative estimation error.

In both the simulation study and the case study the two hot deck procedures had similar results across estimations. This was unsurprising as the methods are themselves very similar. Firstly, they are both hot deck procedures, which randomly pick a 'donor' value based on the matching variables. The random hot deck procedure does this by finding the observations with the same values, while the distance hot deck procedure does this based on the distance between the matching values. So, for an observation in the 'recipient' dataset which have observations in the 'donor' dataset with the exact same matching variables values both hot deck procedures are randomly selecting donors from the same possible donor observations. So, the methods are very similar and it is unsurprising the results are so alike.

When the proxy variable had a lower Cramer's V of 0.4 to the target variable rather than a Cramer's V of 0.6 to the proxy variable, in general, the true bias and true relative bias had higher values over the cells. This is because with a closer proxy the CIA assumption will be maintained more (Donatiello et al., 2016). For variance and relative standard deviation, both datasets showed similar results across all cells, with some cells having larger variance and relative standard deviation and some lower. This is an interesting result as the variance comes as a result of the errors from the statistical matching process and the sampling error. It would, therefore, be expected that both would be equally affected by sampling error and that the data with Cramer's V of 0.6 to target variable would have less error based on the matching process, leading to lower variance overall rather than similar variance. Finally, when looking at the absolute differences between the two proxy variables bootstrap estimates and true estimates, the results were very similar for the same cells across both datasets, suggesting the quality measure procedure worked similarly across both datasets. Overall, the use of a proxy variable does show promising signs that it will decrease the bias in the statistical matching process and therefore, should be used if possible in a practical situation. The case study shows that there are practical situations where a proxy variable can be used.

In this project, the quality measure procedure was tested on data in a case study. The first thing to note is that the data in this situation had a notably small overlap, this shows how common it is for the overlap to be small. Particularly in the context of NSIs, whose aim is to minimise the burden on the respondents of surveys. This is emphasised

by the fact that these datasets where chosen precisely because they were more likely to have a larger absolute overlap. The small overlap meant some categories had no values, highlighting a problem of using the overlap.

The results of the case study data showed particularly large bias and relative bias for some values in the contingency table. An example of this are the high values for the absolute values of relative bias in cells (5,1) and (5,4) for both SM methods. These high values could be for many of reasons. It could stem from the overlap underestimating or overestimating the proportion for this cell leading to high bias, which is possible as it has an estimate of 0.015 and 0.030 respectively, where as the estimated estimates after each method was implemented once was 0.007 and 0.065. This under or over estimation is most likely down to the small size of the overlap. It could also be the case that these methods of SM are very biased for this cell. This uncertainty is a key component of the downfall of this quality measure procedure. The variance and relative standard error seemed low for both these cells, suggesting that there was little difference in estimates and more chance the high value comes from the low proportion estimate in the overlap. Based on these results the researcher in this practical situation could equally choose both methods as they show similar results. If you take into consideration the reasons above for the large values in these cells, then it could be considered that these methods perform well on this data. However, the values are still high, that I would not base my decision for this data on this. In this situation weights were also not used and simple random sampling was assumed for simplicity, for a accurate population estimate weights would need to be used, further research is needed on how to incorporate weights into the quality measure.

The clear limitation of the quality measure procedure stems from the need to have a least a small overlap in the observations. This is challenging in practice as it requires you to have some identifying variable(s). But also in many cases the overlap is so small that it would not be usable, as this leads to less accurate results in the quality procedure. The overlap size in the simulation study was random. However, it would be useful to look into fixed overlap sizes and see if they make any difference on the quality procedure's accuracy at estimating the true bias and variance. In particular to look at what would be the minimum size for the overlap to still be effective, say to look at overlaps of 5%, 10%, 20% or 30% of the smaller dataset. This would be crucial in understanding when this procedure could work in practice.

Following on from this, another clear limitation of the quality measure procedure is that it looks solely at categorical data. It is common to want to know the relationship between other types of variables. It would therefore be useful to investigate new ideas for the quality measure for both situations.

Another limitation to the quality measure is how the bootstrap samples are drawn,

where $\hat{p}_{ij}$ is used as a basis for probability drawing. This is done in an attempt to keep the bootstrap samples reflective of the population. However, it comes with one main disadvantage; if there are no observations for one category in the overlap, this category will not be drawn in any bootstrap samples. Therefore it would be important to look into new ways to draw bootstrap samples.

Furthermore, the simulation study used to test this quality measure procedure was limited to one dataset, although used with two different proxy variables. It would be important to test the measure on multiple datasets, with varying number and type of matching variables.

This method was tested with only two hot decking SM methods. It would also be important to try this method on other types of SM methods, particularly ones which do not rely on the CIA, or which attempt to mitigate CIA in another way.

It would also be useful to explore the proxy variables effect in more detail, assessing its impact when using SM. But also if it does in fact have an impact on the quality measure procedure, by investigating different levels of similarity. However, the similarity in the results of the quality procedure with the proxy of Cramer's V of 0.4 and 0.6 to target variables suggests that the results would be similar regardless of the situation.

Overall this project presented a quality measure procedure which was helpful in giving a researcher an indication of the usefulness of a SM method for particular data. Although, caution is needed as the procedure is reliant on the overlap, with many issues arising when there are no values of a certain category in the overlap. Caution also needs to be made as for some cells the AAD were large, and in a practical situation it would be unknown which cells this would relate to. However, overall the results were promising, where there is no other method to achieve this.

References

CBS. (2021). *Population; key figures* [Accessed: 16-06-2022]. https://opendata.cbs.nl/#/CBS/en/dataset/37296eng/table

Conti, P. L., Marella, D., & Scanu, M. (2012). Uncertainty analysis in statistical matching. *Journal of Offical Statistics*, *28*(1), 69–88.

Donatiello, G., D'Orazio, M., Frattarola, D., Rizzi, A., Scanu, M., & Spaziani, M. (2014). Statistical matching of income and consumption expenditures. *International Journal of Economic Sciences*, *3*(3), 50–65.

Donatiello, G., D'Orazio, M., Frattarola, D., Rizzi, A., Scanu, M., & Spaziani, M. (2016). The statistical matching of EU-SILC and HBS at ISTAT: Where do we stand for the production of official statistics [26-27th September]. *DGINS - Conference of the Directors General of the National Statistical Institutes*.

D'Orazio, M. (2019). Statistical learning in official statistics: The case of statistical matching. *Statistical Journal of the IAOS*, *35*, 435–441.

D'Orazio, M. (2020). *Statmatch: Statistical matching or data fusion* [R package version 1.4.0]. https://CRAN.R-project.org/package=StatMatch

D'Orazio, M., Zio, M. D., & Scanu, M. (2001). Statistical matching: A tool for integrating data in national statistical institutes. *Conference: New Techniques and Technologies for Statistics and Exchange of Technology and Know-how: Greece*, 443–440.

D'Orazio, M., Zio, M. D., & Scanu, M. (2006). *Statistical matching: Theory and practice*. Chichester, UK: Wiley.

European Statistical System Committee. (2017). European statistics code of practice: For the national statistical authorities and eurostat (eu statistical authority).

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, *27*(4), 857–871.

Kim, J. K., Berg, E., & Park, T. (2016). Statistical matching using fractional imputation. *Survey Methodology*, *42*(1), 19–40.

Leulescu, A., & Agaftei, M. (2013). Statistical matching: A model based approach for data integration. *Eurostat: Methodologies and Working Papers*.

McHugh, M. L. (2018). *The sage encyclopedia of educational research, measurement, and evaluation*. Thousand Oaks, SAGE Publications Inc.

Public Health Monitor 2016 of the Community Health Services, Statistics Netherlands and the National Institute for Public Health and the Environment.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org

Rässler, S. (2002). *Statistical matching: A frequentist theory, practical applications, and alternative bayesian approaches*. New York, USA: Springer.

Rässler, S. (2004). Data fusion: Identification problems, validity, and multiple imputation. *Austrian Journal of Statistics*, *33*(1&2), 153–171.

Schnell, R. (2021). Privacy preserving record linkage in the context of a national statistical institute. *German Record Linkage Center Working Paper Series*.

Scholtus, S., & Daalmans, J. (2021). Variance estimation after mass imputation based on combined administrative and survey data. *Journal for Offical Statistics*, *37*(2), 433–459.

6 Appendix A: R-code

The code used in this project is available at:
https://github.com/FranGoudie/masterthesis

7 Appendix B: Deciding parameters

In the simulation study the number of iterations, S, needed to be established, as well as the number of bootstrap iterations, R, within the the iterative loop. In order to do this samples were taken from the simulation data in the same fashion as the simulation and the quality measure was run at various intervals to see when the bias and variance converged. Figure 9 gives a graph of iterations of the bootstrap at various intervals. Based on the graphs we concluded that the bias estimates hardly changed from 200 iterations onwards, the same was true for variance, so B was set at 200. To decide S we examined how many iterations it took for the bias and variance of the true estimates to converge in the simulation study, this can be seen in Figure 8. Therefore, due to this graph and time constraints, it was decided to use S = 100 iterations.

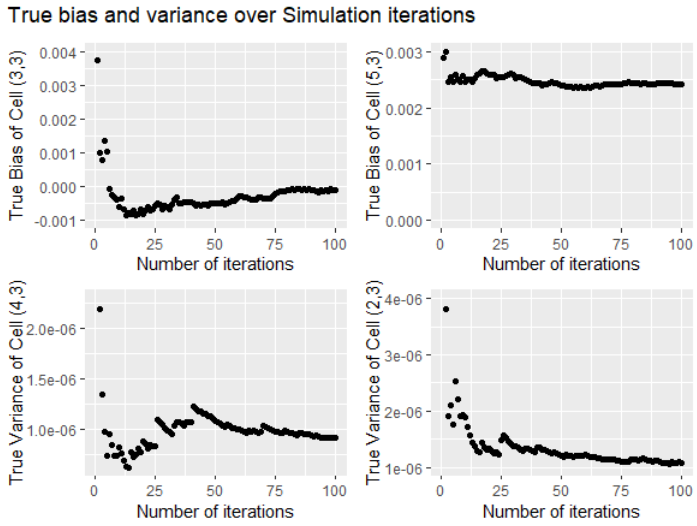

Figure 8: Plot showing the True Bias and True Variance estimate over 100 iterations of the simulation, using the distance hot deck procedure.

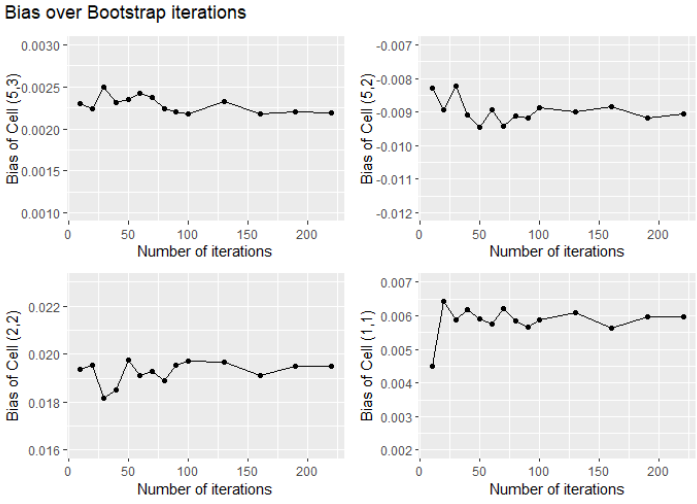

Figure 9: Plot showing the bias estimates over varying iterations of the bootstrap, using the random hot deck procedure.