Discussion Paper

Statistical matching: Experimental results and future research questions

2015 | 19

Ton de Waal

Content

Abstract

National statistical institutes try to construct data sets that are rich in information content by combining available data as much as possible.
Unfortunately, units, e.g. persons or enterprises, in different data sources cannot always be matched directly with each other, for example because different data sources often contain different units. In such a case one can sometimes resort to statistical matching rather than exact matching. Statistical matching can be used when different data sources contain (different) units with a set of common (background) variables. These common variables may then be used to match similar units in the data sources to each other. From March 2015 till the end of June 2015 two master students, Sofie Linskens and Janneke van Roij, at Tilburg University evaluated some methods for statistical matching, namely random hot deck, distance hot deck and statistical matching using a parametric model, on categorical data from the Dutch Population Census 2001. In this paper we describe the methods that they examined and the results they obtained.

1. Introduction

National statistical institutes (NSIs) fulfil an important role as providers of objective and undisputed statistical information on many different aspects of society. To this end NSIs try to construct data sets that are rich in information content and that can be used to estimate a large variety of population figures. At the same time NSIs aim to construct these rich data sets as efficiently and cost effectively as possible. This can be achieved by combining available data, such as Big Data, administrative data or survey data, as much as possible.

Unfortunately, units, e.g. persons or enterprises, in different data sources cannot always be matched directly with each other as different data sources often contain different units. Even when different data sources do have units in common, it may still be impossible to match these units exactly as sufficient identifying information to do so may be lacking.

In such a case one can sometimes resort to statistical matching (see, e.g., Rodgers 1984, Kadane 2001, Moriarity and Scheuren 2001, Rässler 2004, D'Orazio, Di Zio and Scanu 2006, Moriarity 2009 and Eurostat 2013) rather than exact matching. Statistical matching can be used when different data sources contain (different) units with a set of common (background) variables. These common variables may then be used to match similar units in the data sources to each other. For example, if we have a data source with information on the education level of persons, their gender, age and municipality and another data source with information on the occupation of (other) persons, their gender, age and municipality, we can use the information on gender, age and municipality to statistically match similar units in the data sources with each other. The unit type of the units in the different data sets has to be the same, for instance persons can only be statistically matched with (other) persons and not with enterprises.

The main goal of statistical matching is to estimate the relationship between the non-common variables, the target variables, in the different data sources as well as possible. In our example the goal is to estimate the true relationship between the education level and the occupation of persons in the population as well as possible. For Statistics Netherlands statistical matching is not a new topic. Several decades ago research on statistical matching has already been carried out at Statistics Netherlands (see Mokken 1984 and De Jong 1990a, 1990b, 1991 and 1997). That research has never led to major applications of statistical matching at Statistics Netherlands. The main reason probably was that there never was any urgent reason to do so: surveys and (later also) administrative data were sufficient to produce the statistical information needed.

Lately interest in statistical matching at Statistics Netherlands and other NSIs is rising again. An important reason for this is the recent data deluge due to Big Data. Big Data are often characterized by the three V's: Volume, Velocity and Variety (see, e.g.,

Buelens et al. 2014). 'Volume' means that Big Data sources are generally larger than regular systems can handle smoothly. 'Velocity' refers to the high frequency at which data become available or to the short period between the occurrence of an event and Big Data on this event becoming available. 'Variety' refers to the wide diversity of Big Data sources. Owing to this wide diversity, one may not be able to directly match units in the Big Data to other data sets.

Another reason why statistical matching is becoming more interesting for NSIs is that users of statistical information want more detailed and timelier information, such as statistical information on unexpected current events in society. Such detailed and timely information cannot be provided by surveys and administrative data sources alone any more.

Examples of situations where statistical matching techniques may be useful are matching of two non-overlapping surveys with common background variables, matching of Big Data to survey or administrative data, and finding imputation values when for certain groups of units a number of variables are missing by design. The latter situation can, for instance, when certain units send in their data automatically in an electronic way and in return do not have to fill a more extensive questionnaire. From March 2015 till the end of June 2015 two master students, Sofie Linskens and Janneke van Roij, at Tilburg University evaluated some methods for statistical matching on data from the Dutch Population Census 2001. They were supervised by the author of the current paper, who works most of his time at Statistics Netherlands and part-time at Tilburg University. In this paper we will describe the methods that they examined and the results they obtained. Section 2 of this paper describes the methods that were examined and Section 3 the results. Section 4 concludes the paper with a short discussion, including a number of key questions for future research.

2. Methods for statistical matching

2.1 Introduction to statistical matching

For convenience we will assume that statistical matching is used to combine two data sets. If there are more than two data sets to be combined, one can start by combining two of them and then add one more data set at a time until all data sets have been combined.

Statistical matching can be carried out on the micro level or on the macro level. When statistical matching is carried out on the micro level, one combines data from individual units in the different data sources to construct synthetic records with information on all variables. In particular, in the micro level approach information

from one data set, the donor data, is used to estimate target values in the other data set, the recipient data. In this way one constructs a complete synthetic microdata set containing values for all variables for the recipient units. Which of the two data sets is selected as the recipient data set is a matter of choice. Often one chooses the data set with the largest number of target variables or with the target variables that are considered to be the most important ones. The recipient data set can, however, also be selected on other criteria, such as the correlation between target variables and background variables.

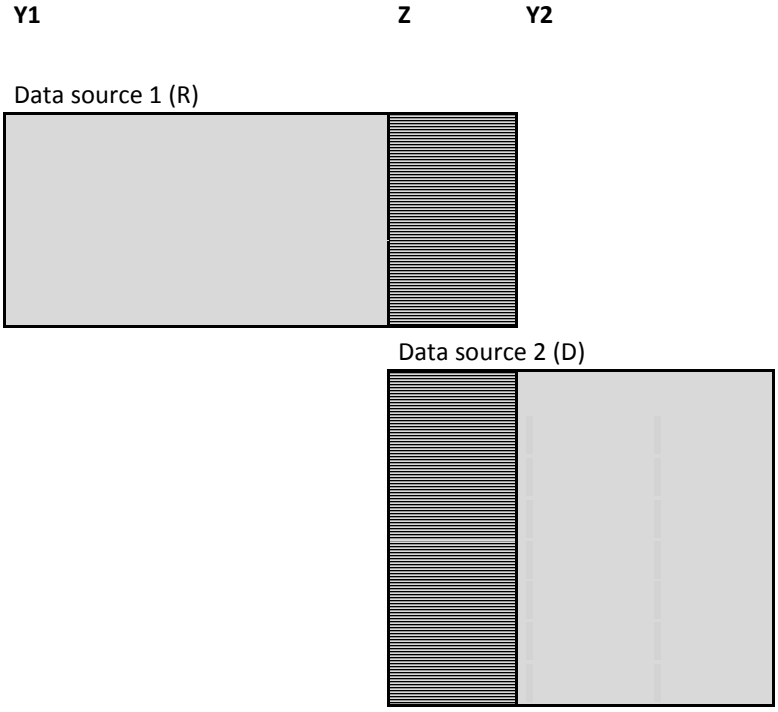
When statistical matching is carried out on the macro level, one constructs a parametric model for all the data, for instance a multivariate normal model for numerical data or a multivariate multinomial model for categorical data, and then estimates the parameters of this model. These parameters are subsequently used to estimate population parameters one is interested in. In the macro approach one does not construct a complete synthetic microdata set containing values for all variables. For an overview of methods for statistical matching on both the macro level and the micro level we refer to D'Orazio, Di Zio and Scanu (2006). Linskens (2015) and Van Roij (2015) both focused on micro level approaches.

Figure 2.1.1 illustrates the situation for statistical matching. In this figure we have two data sources. Data source 1 contains variables $\boldsymbol{Y}_1$ and $\boldsymbol{Z}$ and data source 2 $\boldsymbol{Y}_2$ and again $\boldsymbol{Z}$. Variables $\boldsymbol{Z}$ are the common (background) variables that are used to match the data sources. When statistical matching is carried out on the micro level, variables $\boldsymbol{Z}$ are used to match individual units in data source 1 – the recipient (R) – to individual units in data source 2 – the donor (D).

The fundamental issue of statistical matching is that the relationship between the target variables $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ cannot be estimated directly. When statistical matching is used, the relationship between variables $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ is estimated indirectly. In order to do so, one has to rely on untestable assumptions about this relationship. That is, these assumptions are untestable from the data sources themselves. The most common assumption is the Conditional Independence Assumption (CIA). The CIA says that conditional on the values of background variables $\boldsymbol{Z}$ the target variables $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ are independent. In other words, the CIA says that the relationship between target variables $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ can be entirely explained by the values of the background variables $\boldsymbol{Z}$.

Statistical matching is closely related to imputation. When imputation is applied to missing items in a data set, the values of these items are estimated and filled in (see, e.g., De Waal, Pannekoek and Scholtus 2011 for more on imputation). The difference between imputation and statistical matching is that imputation is used for estimating variables that are partly observed together in a single data set, whereas statistical matching is used for estimating variables that are not observed (together) in a data set but in another, related data set.

2.1.1 The situation for statistical matching



All variables in the data sets that were used for the evaluation studies were categorical. The statistical matching methods selected for the evaluation studies were therefore suitable for categorical data. Statistical matching methods that can only be applied to numerical data were not examined.

The statistical methods selected for the evaluations studies by Linskens (2015) and Van Roij (2015) were all based on hot deck techniques. Hot deck techniques replace missing values by observed data from similar units, where similarity is measured by means of background variables.

At NSIs, hot deck techniques are quite popular for imputing missing data. There are several reasons for this. First, hot deck techniques yield realistic imputation values as these are based on actually observed values. Second, imputed values based on hot deck cannot be outside the range of possible values. Third, it is not necessary to model the distribution of the missing data. Fourth, hot deck techniques are relatively simple to implement, understand and apply in practice. For the same reasons hot deck methods are also interesting to NSIs for statistical matching.

Hot deck methods for statistical matching can be 'unconstrained' or 'constrained' (see D'Orazio, Di Zio and Scanu 2006 for an overview of hot deck methods for statistical matching). A hot deck method is called unconstrained when each donor record may be used multiple times and constrained when each donor record may only be used once. Constrained hot deck requires the number of donor records to be equal to or larger than the number of recipient records.

The main advantage of constrained matching in comparison to unconstrained matching is that the marginal distribution of the target variables in the donor data set $\boldsymbol{Y}_2$ is better maintained in the matched data set. A disadvantage is that the average distance between the recipient and donor values of the matching variables, i.e. a subset of $\boldsymbol{Z}$, is likely to be larger than in the unconstrained case. Moreover, constrained matching is computationally much more demanding than unconstrained matching. In this paper we will only examine unconstrained matching.

There are several hot deck methods. Linskens (2015) examined random hot deck and distance hot deck. Van Roij (2015) examined statistical matching using a parametric model and again distance hot deck. We will describe these methods below.

2.2 Random hot deck

In its basic form, random hot deck simply consists of drawing for each record in the recipient data set a random record from the donor data set. However, random hot deck is usually applied in a slightly more advanced form, where units of both the recipient data set and the donor data set are grouped in homogeneous subsets, referred to as donation classes (see D'Orazio, Di Zio, and Scanu 2006). These donation classes are based on one or more categorical variables selected from the common background variables present in both data sets, the matching variables. For each recipient record a donor record is randomly drawn from the donation class of the recipient record. That is, only donor records that score similarly to the recipient record on matching variables may be drawn. In the unconstrained form of random hot deck, donor records may be used for multiple recipient records.

In general, the main challenge of random hot deck is the construction of good donation classes. When donation classes are too small, the same donor records may be used repeatedly, distorting, for instance, variance estimates. When donation classes are too large, donor records will be less similar to the recipient records, distorting, for instance, the accuracy of the estimates.
Random hot deck was applied using the RANDwNND.hotdeck routine from the R package StatMatch (D'Orazio 2015).

2.3 Distance hot deck

When distance hot deck is applied, one determines for each record from the recipient data set the distance to, in principle, all records in the donor data set. The donor record with the smallest distance is then selected for matching with the recipient record under consideration. In the case of ties, one of the records with the smallest distance to the recipient record is chosen at random. In distance hot deck usually no donation classes are constructed.

The distance is computed on the selected matching variables. To measure the distance between a recipient record and a potential donor record different distance

functions can be used (again see D'Orazio, Di Zio, and Scanu 2006). In the evaluation studies by Linskens (2015) and Van Roij (2015) the Manhattan distance function was used. The Manhattan distance function calculates the absolute distances between the categories of the recipient record and the categories of a potential donor record summed across the selected matching variables. Here the categories of each variable are assumed to be numbered from 1 to the number of categories of this variable. The absolute distance for a variable is then the absolute difference between the two numbers for the two records involved.

The main challenge of distance hot deck is to construct a good distance measure, i.e. to select appropriate matching variables and an appropriate distance function. Distance hot deck was applied using the NND.hotdeck routine of the StatMatch package (D'Orazio 2015).

2.4 Statistical matching using a parametric model

We developed a new method for statistical matching that we will refer to in this paper as parametric statistical matching. This novel method is inspired by predictive mean matching, which can be used for matching of numerical variables. In predictive mean matching one estimates regression models for target variables $\boldsymbol{Y}_1$ with the $\boldsymbol{Z}$ variables as predictors using the data in the recipient data set, and for target variables $\boldsymbol{Y}_2$ with the $\boldsymbol{Z}$ variables as predictors using the data in the donor data set. Once these regression models have been estimated, one can use the regression model(s) for target variables $\boldsymbol{Y}_1$ with the $\boldsymbol{Z}$ as predictors to obtain estimates $\widehat{\boldsymbol{Y}}_1$ for $\boldsymbol{Y}_1$ in both the recipient and the donor sets, use the regression model(s) for target variables $\boldsymbol{Y}_2$ with the $\boldsymbol{Z}$ as predictors to obtain estimates $\widehat{\boldsymbol{Y}}_2$ for $\boldsymbol{Y}_2$ in both the recipient and the donor sets, and then use distance hot deck to match the records $\left(\widehat{\boldsymbol{Y}}_1, \widehat{\boldsymbol{Y}}_2, \boldsymbol{Y}_2, \boldsymbol{Z}\right)$ for the donor data set to the records $\left(\widehat{\boldsymbol{Y}}_1, \widehat{\boldsymbol{Y}}_2, \boldsymbol{Y}_1, \boldsymbol{Z}\right)$ for the recipient data set. The distance between the records is computed using $\widehat{\boldsymbol{Y}}_1$ and $\widehat{\boldsymbol{Y}}_2$ rather than $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$.

By adding an extra modelling step, one hopes to 'stabilize' or 'smooth' the observed data, and obtain better matches. Since the evaluation data set contains only categorical data, predictive mean matching cannot be applied directly and we had to develop a method suitable for categorical data.

For this part of the evaluation study Van Roij (2015) considered only one target variable in the recipient data set and one target variable in the donor data set. In our parametric matching approach, for each record in the data sets multinomial logistic regression was used to estimate the probability of observing a certain category in the target variable. For each matching variable, this leads to a vector of estimated probabilities: one estimated probability for each category of this variable. Next, these estimated probabilities were themselves used as numerical matching variables in a distance hot deck procedure, where the Euclidean distance function was used to match a donor record to each recipient record.

We have chosen to use the Euclidean distance function for the sake of simplicity as this distance function was readily available in the StatMatch package. However, since the estimated probabilities sum to one, a distance measures for compositional data might have been a more appropriate choice. For more about distance measures for compositional data we refer to Aitchison (2003) and Pawlowsky-Glahn and Buccianti (2011).

Van Roij (2015) compared several logistic regression models for estimating $\hat{Y}_1$ and for estimating $\hat{Y}_2$ (see also Subsection 'General information' of Section 3.2).The models that fitted the data best were chosen to calculate predicted probabilities. As explained above these models were then used to estimate probabilities for the categories of both target variables in both data sets.

3. Evaluation studies

3.1 The evaluation approach

In order to be able to evaluate and compare the methods described in Section 2, a complete data set was randomly split into a donor data set and a recipient data set which both consisted of all background variables and one or more target variables. The units in the recipient data and the donor data did not have any overlap. The complete data sets, and hence also the recipient data sets and the donor data sets, for Linskens and Van Roij contained information on different population units. In the evaluation studies the fused data sets obtained after statistical matching were compared to the original data of the units in the recipient data set containing all target variables and background variables. We will refer to these data as the complete recipient data.

The complete data sets contained information from a subset of the Dutch Population Census 2001, which was protected against disclosure of confidential information by means of recoding. The selection of the two complete data sets and the splitting up of these data sets into recipient and donor data sets was done by the author of the current paper.

Background variables present in both the recipient and donor data sets were gender, age, position in the household, size of the household, living area in the previous year, nationality, mother country, and marital status. In addition, the donor data set included information on education level and the recipient data set information on economic status, occupation, and branch of industry. The recipient data set was complemented with information on education level from the donor data set. For more information on the variables and their categories, see Table 5.1.1 in the Appendix.

For practical reasons Linskens (2015) limited the number of cases to 2,000 in the recipient data set and another 2,000 in the donor data set, while Van Roij (2015) limited the number of cases to 3,000 in the recipient data set and another 3,000 in the donor data set. These records were again randomly selected.

In the evaluation studies by Linskens (2015) and Van Roij (2015) education level was treated as the most important target variable. In the evaluation studies univariate results for education level and multivariate results for education level with other variables were calculated.
All techniques and analyses were performed using R.

3.2 The evaluation results

In this section we present results of the evaluation studies. We begin in Subsection 'General information' with some information on, for instance, which background variables were used for constructing donation classes in random hot deck, which background variables were used in the distance function for distance hot deck, and which background variables were used in the models for parametric statistical matching. In Subsections 'Univariate results for education level' and 'Preserving (in)dependency of education level with background variables' we present univariate results for education level, respectively results for the relationships between education level and background variables. In a sense good univariate results for the target variables and good results with respect to preserving the relationships between target variables and background variables are a prerequisite for possible successful application of statistical matching. If the quality of univariate results for target variables or results with respect to preserving the relationships between target variables and background variables is not acceptable, statistical matching cannot be applied. The real test of statistical matching, however, is whether the relationships between target variables in the recipient data set and target variables in the donor data set are preserved. Preserving these relationships as well as possible is the main goal of statistical matching. Subsection 'Preserving relationships between target variables' presents results for how well these relationships are preserved. The results in Subsection 'General information' to 'Preserving relationships between target variables' were all calculated without using survey weights. We end this section by describing results when surveys weights are used and comparing differences between weighted and unweighted results.

General information

Random hot deck

In order to prevent empty donation classes Linskens (2015) used only the background variable age to define these classes. This resulted in donation classes consisting of 3 to 205 donor records. The average number of donor records per donation class was 138. Since the first quartile represents 113 donor records, the minimum number of donor records of 3 might be considered an outlier. Overall the donation classes seem sufficiently large to avoid distortion of variance estimates.

Distance hot deck

In both evaluation studies the matching variables for distance hot deck were selected by examining Cramérs V. In social sciences a Cramérs V value between 0 and 0.25 is often considered to indicate a weak association, a value between 0.25 and 0.35 a medium association and a value above 0.35 a strong association. In the evaluation study by Linskens (2015) the values for Cramérs V ranged from 0.11 (gender) to 0.47 (age). Based on these results the background variables age (V = 0.47), marital status (V = 0.32), living area in the previous year (V = 0.28), and position in household (V = 0.27) were selected as matching variables in that study. Van Roij (2015) selected the same background variables for matching for the same reasons.
For all results of Linskens (2015) and Van Roij (2015) we refer to Tables 3.2.1, 3.2.2 and 3.2.3.

In the evaluation study by Linskens (2015) distance hot deck resulted in a minimum distance between recipient record and donor record of zero and a maximum distance of 82. On average, the distance between the recipient record and donor record was 0.167. Since the third quartile also represents a distance of zero, the maximum distance of 82 might be considered an outlier. The number of available donors at the minimum distance for each recipient record ranged from 1 to 133, with an average of 63.6 available donors at the minimum distance.

This makes a comparison of results for random hot deck and distance hot deck not completely fair as the higher number of matching variables is likely to favour distance hot deck. We will nevertheless compare the results for random hot deck and distance hot deck as the fact that distance hot deck can take a relatively large number of matching variables into account much more easily than hot deck is an important difference between these methods.

3.2.1 Association between education level and background variables in the donor data set of Linskens (2015)

Background variable	Cramérs V
Gender	0.11
Age	0.47
position in the household	0.27
size of the household	0.16
living area in the previous year	0.28
nationality	0.13
mother country	0.14
marital status	0.32

3.2.2 Association between education level and background variables in the donor data set of Van Roij (2015)

Background variable	Cramérs V
Gender	0.10
Age	0.51
position in the household	0.27
size of the household	0.15
living area in the previous year	0.24
nationality	0.10
mother country	0.12
marital status	0.23

3.2.3 Pairwise association between economic status and background variables in the recipient data set of Van Roij (2015)

Background variable	Cramérs V
gender	0.11
age	0.55
position in the household	0.46
size of the household	0.35
living area in the previous year	0.11
nationality	0.05
mother country	0.07
marital status	0.07

Parametric statistical matching

Van Roij (2015) examined the pairwise association between variables (measured by means of Cramérs V) in the donor data set to select matching variables for parametric statistical matching. As a result of this examination she selected age, position in the household and marital status as matching variables.

To create the fused data set, Van Roij (2015) applied several multinomial regression analyses to the donor data set with education level as the dependent variable and the selected matching variables as the independent variables. The full factorial model fitted the data best according to Van Roij (2015). Probabilities for each category of education level were saved and added as variables to the donor data set. The same model was then used to predict education level in the recipient data set.

Van Roij (2015) also examined the pairwise association between variables in the recipient data set. Based on this examination, age, position in the household, and marital status were selected for multinomial regression models with economic status as dependent variable. Several multinomial regression analyses were applied on the recipient data set with economic status as the dependent variable. Again the full factorial model fitted the data best according to Van Roij, and this model was used to estimate the probabilities.

Univariate results for education level

In the study by Linskens (2015) the random hot deck method predicted found the correct scores on education level in 20.6% of the cases, whereas the distance hot deck method found the correct scores in 36.5% of the cases.

In her evaluation study Linskens (2015) made inferences by independent samples t-tests to examine mean and variance differences between fused data sets and the complete recipient data. For these analyses, education level was assumed to be numeric. This can be justified by the ordinal nature of this variable. The independent samples t-test shows the mean score of education level after the application of random hot deck (M = 3.11) and the mean score of education level in the complete recipient data (M = 3.02) not to be significantly different at a 5% significance level. Furthermore, the estimated values for education level resulting from the random hot deck method show to be slightly positively correlated with the true values of education level, although the correlation coefficient is not significantly deviant from zero at a 5% significance level (see also Table 3.2.4).

The independent samples t-test shows the mean score of education level after the application of distance hot deck (M = 3.15) and the mean score of education level in the complete recipient data (M = 3.02) not to be significantly different at a 5% significance level. Furthermore, the estimated values for education level resulting from the distance hot deck method show to be positively correlated with the true values of education level (significant at a 5% significance level).

As pointed out by Sander Scholtus, a paired sample t-test instead of an independent samples t-test would have been more appropriate for our situation. The results obtained by Linskens (2015) are therefore only indicative and should not be used to base any firm conclusions upon.

3.2.4 Weighted means (M) and standard deviations (SD) for education level after distance hot deck, after random hot deck and in the complete recipient data (Linskens 2015)

distance hot deck (N = 2,000)		**random hot deck (N = 2,000)**		**complete recipient data (N = 2,000)**	
M	SD	M	SD	M	SD
3.15	3.09	3.11	4.78	3.02	2.17

Cross tables of education level across data sets obtained by Van Roij (2015) are presented in Tables 3.2.5 and 3.2.6. As the p-value of the chi square test is smaller than the 5% significance level, the null hypothesis that the education level in the complete recipient data is independent of the education level in the data set obtained after distance hot deck is rejected. This means that education level in the complete recipient data is related to education level in the data set after distance hot deck. The strength of the association between these variables in the evaluation data is 0.419 (Cramér's V). The p-value of the chi square test for independence between education level in the complete recipient data and education level obtained after

parametric statistical matching is larger than the 5% significance level meaning that we do not reject the null hypothesis that the variable education level in these data sets are independent of each other. The strength of the association, measured with Cramér's V, between education level in complete recipient data and education level after parametric statistical matching is 0.045. This means that education level in the complete recipient data is hardly related to education level in the data set after parametric statistical matching.

3.2.5 Cross table of education level across data sets: complete recipient data versus data after distance hot deck (Van Roij 2015)

complete recipient data	**distance hot deck**							
Education level	0	1	2	3	4	5	9	Total
0	105	58	10	8	1	9	28	219
1	55	147	98	113	8	44	0	465
2	15	110	197	205	14	98	0	639
3	16	122	229	372	40	163	0	942
4	4	10	18	38	5	20	0	95
5	3	66	92	175	13	86	0	435
9	27	0	0	0	0	0	177	204
Total	225	513	644	911	81	420	205	2,999[1]

3.2.6 Cross table of education level across data sets: complete recipient data versus data after parametric statistical matching (Van Roij 2015)

complete recipient data	**parametric statistical matching**							
Education level	0	1	2	3	4	5	9	Total
0	26	30	44	64	5	33	17	219
1	32	74	98	153	18	65	25	465
2	39	110	132	192	17	108	41	639
3	59	156	224	262	29	147	65	942
4	7	15	15	34	0	18	6	95
5	26	78	94	132	13	62	30	435
9	13	38	30	69	9	30	15	204
Total	202	501	637	906	91	463	199	2,999

[1] Due to a minor error the total number of records sums up to 2,999 in this and other tables instead of to 3,000. This minor error does not affect the overall conclusions.

Preserving (in)dependency of education level with background variables

When two variables are dependent (or independent) in the complete data set, it is desirable to maintain the same dependency relationship in the fused data set. For all fused data sets Linskens (2015) and Van Roij (2015) therefore carried out chi-square tests between education level and all background variables to test those relationships for (in)dependency in the fused data sets. The results were compared to the equivalent chi-square tests performed using complete recipient data to examine possible distortions of dependency relationships. Found distortions in dependency relationships were statistically tested for significance.

Table 3.2.7 obtained by Linskens (2015) shows that for random hot deck the estimated values of education level are dependent on all background variables except for gender, nationality, and mother country. Since results of the complete recipient data show dependency between education level and all background and target variables, the results of random hot deck for variables gender, nationality, and mother country therefore represent possible distortions of dependency relationships.

Linskens (2015) found that the cross table for education level and gender based on the fused data set after random hot deck was not significantly different from the same cross table based on complete recipient data. However, the cross tables for education level and nationality and education level and mother country do differ significantly for the fused data set after random hot deck and complete recipient data.

For distance hot deck the estimated values of education level show to be dependent on all background variables except for nationality. Since results of complete recipient data show dependency between education level and all background variables, the result for nationality represents a possible distortion of the dependency relationship between education level and nationality as a result of distance hot deck. Linskens (2015) found that the dependency relationship between education level and nationality after distance hot deck is indeed significantly different from this dependency relationship in complete recipient data.

In the evaluation study by Van Roij (2015) slightly different results were obtained: her results (see Table 3.2.8) suggest that in all three data sets there is an association between education level and age, position in the household and marital status, i.e. these three relationships are dependent in all three data sets. In addition, a non-significant association between education level and nationality was found in all three data sets, i.e. this relationship is considered to be independent in all three data sets. In complete recipient data there was a significant association between education level and gender. However, in both synthetic data sets this association was non-significant.

For parametric statistical matching there was a discrepancy in the association between education level and size of the household, and between education level and mother country compared to complete recipient data. In complete recipient data, a

significant association was found between education level and both background variables, while after parametric statistical matching there was not.
By means of chi square tests Van Roij (2015) investigated whether the two distributions education level × size of household and education level × mother country after parametric statistical matching could have arisen by sampling from the same population as these distributions in complete recipient data. A non-significant result for size of household as well as for mother country was found, indicating that the null hypothesis of independence is not rejected. This means that it is unlikely that the distributions in the fused data set could have arising by sampling from the same population as the distribution in the complete recipient data. This result is consistent with the earlier mentioned result that the association between education level and size of household was distorted after parametric statistical matching.

Van Roij (2015) also performed chi square tests to test whether her three samples for education level and gender could have arisen by sampling from the same population despite their different associations. First, the results of distance hot deck was compared to the complete recipient data and results showed a non-significant result. The null hypothesis of independence is not rejected, thus indicating that there is no reason to assume that the samples are from the same population. The data obtained from parametric statistical matching were also compared to the complete recipient data. Results were similar. This is consistent with earlier mentioned results.
Van Roij (2015) also calculated cross tables between education level and gender for her three data sets. The results are given in Tables 3.2.9 to 3.2.11.

3.2.7 Chi-square tests testing independency between education level and background variables (Linskens 2015)

Background variable	**Distance hot deck (N = 2,000)**		**Random hot deck (N = 2,000)**		**Complete recipient data (N = 2,000)**	
	χ^2(df)	p-value	χ^2(df)	p-value	χ^2(df)	p-value
gender	19.35 (7)	0.007*	10.42 (6)	0.108	17.05 (6)	0.009*
age	3183.09 (119)	<0.001*	3074.44 (102)	<0.001*	3289.37 (102)	<0.001*
position in household	1023.09 (49)	<0.001*	958.33 (42)	<0.001*	975.80 (42)	<0.001*
size of household	315.24 (35)	<0.001*	267.46 (30)	<0.001*	293.14 (30)	<0.001*
living area in previous year	323.87 (14)	<0.001*	302.36 (12)	<0.001*	334.15 (12	<0.001*
nationality	23.36 (21)	0.325	20.86 (18)	0.287	44.08 (18)	<0.001*
mother country	38.99 (14)	<0.001*	20.42 (12)	0.060†	41.91 (12)	<0.001*
marital status	561.25 (21)	<0.001*	507.37 (18)	<0.001*	481.95 (18)	<0.001

* indicates that the variables are found to be dependent in the given data set.

3.2.8 Chi-square tests testing independency of education level and background variables (Van Roij 2015)

Background variable	Distance hot deck (N = 3,000)		Parametric statistical matching (N = 3,000)		Complete recipient data (N = 3,000)	
	χ^2(df)	p-value	χ^2(df)	p-value	χ^2(df)	p-value
gender	6.09 (6)	0.413*	4.27 (6)	0.640	27.67 (6)	<0.001*
age	4644.19 (102)	<0.001*	142.92 (102)	0.005	4715.21 (102)	<0.001*
position in household	1451.82 (48)	<0.001*	68.56 (48)	0.027	1370.74 (48)	<0.001*
size of household	308.55 (30)	<0.001*	39.44 (30)	0.116	331.48 (30)	<0.001*
living area in previous year	518.11 (12)	<0.001*	21.72 (12)	0.041	534.87 (12)	<0.001*
nationality	12.87 (18)	0.800	16.77 (18)	0.539	24.58 (18)	0.137
mother country	21.92 (12)	<0.001*	4.42 (12)	0.975	21.92 (12)	0.038
marital status	674.51 (18)	<0.001*	31.33 (18)	0.026	756.29 (18)	<0.001

* indicates that the variables are found to be dependent in the given data set.

3.2.9 Cross table of education level and gender in complete recipient data (Van Roij 2015)

Gender	Education level							
	0	1	2	3	4	5	9	Total
Male	113	191	263	471	53	220	106	1,417
Female	106	274	376	471	42	215	98	1,582
Total	219	465	639	942	95	435	204	2,999

3.2.10 Cross table of education level and gender after distance hot deck (Van Roij 2015)

Gender	Education level							
	0	1	2	3	4	5	9	Total
male	118	224	305	430	38	199	104	1,418
female	107	289	339	482	43	221	101	1,582
Total	225	513	644	912	81	420	205	3,000

3.2.11 Cross table of education level and gender after parametric statistical matching (Van Roij 2015)

Gender	Education level							
	0	1	2	3	4	5	9	Total
male	90	253	298	433	40	217	87	1,418
female	112	248	339	474	51	246	112	1,582
Total	202	501	637	907	91	463	199	3,000

Preserving relationships between target variables

As we mentioned before, the main goal of statistical matching is to preserve the relationship between target variables in the two data sets to be matched as well as possible. Linskens (2015) and Van Roij (2015) therefore examined the relationship between education level and economic status in detail.

Linskens (2015) performed chi-square tests to test for independency between education level and the target variables economic status, occupation, and branch of industry in the recipient data set. Again, those results were compared to the equivalent chi-square tests performed using the complete recipient data to examine possible distortions of these dependency relationships. The estimated values of education level show to be dependent on all target variables in all three data sets (see Table 3.2.12).

3.2.12 Chi-square tests testing independency of education level and target variables in recipient data set (Linskens 2015)

Target variable	Distance hot deck (N = 2,000)		Random hot deck (N = 2,000)		Complete recipient data (N = 2,000)	
	χ^2(df)	p-value	χ^2(df)	p-value	χ^2(df)	p-value
economic status	1395.85 (49)	<0.001*	507.37 (42)	<0.001*	1639.43 (42)	<0.001*
occupation	444.91 (63)	<0.001*	431.44 (54)	<0.001*	1218.20 (54)	<0.001*
branch of industry	442.23 (84)	<0.001*	448.69 (72)	<0.001*	868.19 (72)	<0.001*

* indicates that the variables are found to be dependent in the given data set.

Linskens (2015) and Van Roij (2015) performed chi-square tests to test whether a cross table for education level and economic status based on a fused data set obtained by one of their methods could have arisen by sampling from sampling the same population as the corresponding cross table based on the complete recipient data.
The cross tables obtained by Linskens (2015) are presented in Table 3.2.13 for the complete recipient data, Table 3.2.14 for the data set after random hot deck, and Table 3.2.15 for data set after distance hot deck. These tables show the relationship between education level and economic status to be fairly similar across the three data sets. Chi-square tests indeed show that the joint distribution of education level and economic status in the complete recipient data is not significantly different from the joint distribution after random hot deck or from the joint distribution after

distance hot deck. Furthermore, the joint distribution of education level and economic status shows not to be significantly different across the two fused data sets. Tables 3.2.16 to 3.2.18 give the cross tables of education level and economic status per data set obtained by Van Roij (2015). The hypothesis of independence between education level and economic status in the complete recipient data was rejected at a 5% significance level. The same conclusion was drawn for the data after distance hot deck. However, the test of independence was not rejected for the data after parametric statistical matching. That is, the relationship between education level and economic status was distorted after parametric statistical matching.
When using a chi square test to test whether the joint distribution of education level and economic status after distance hot deck could have arisen by sampling from the same population as the joint distribution of education level and economic status obtained from the complete recipient data, Van Roij (2015) found that there was no significant difference between the two distributions. For parametric statistical matching she obtained the same result, i.e. no significant difference between the distribution for the complete recipient data and the data after parametric statistical matching.

3.2.13 Cross table education level and economic status in the complete recipient data (Linskens 2015)

Economic status	**Education level**								
	0	1	2	3	4	5	9	98	Total
111	1	70	171	378	48	202	0	0	870
112	0	8	29	38	3	5	0	0	83
120	0	5	13	45	1	14	0	0	78
210	0	8	7	12	1	4	0	0	32
221	142	65	37	20	3	2	45	0	314
222	3	63	48	40	9	32	0	0	195
223	3	42	67	62	2	12	0	0	188
224	2	42	37	39	3	17	100	0	240
Total	151	303	409	634	70	288	145	0	2,000

3.2.14 Cross table education level and economic status in data after random hot deck (Linskens 2015)

Economic status	**Education level**								
	0	1	2	3	4	5	9	98	Total
111	3	101	206	314	42	202	0	2	870
112	5	14	32	27	3	2	0	0	83
120	0	9	14	27	4	24	0	0	78
210	1	3	9	10	2	7	0	0	32
221	133	65	40	26	2	2	46	0	314
222	4	67	46	45	2	31	0	0	195
223	0	32	53	54	8	41	0	0	188
224	4	24	29	46	4	32	100	1	240
Total	150	315	429	549	67	341	146	3	2,000

3.2.15 Cross table education level and economic status in data after distance hot deck (Linskens 2015)

Economic status	Education level								
	0	1	2	3	4	5	9	98	Total
111	6	103	205	320	35	200	0	1	870
112	4	13	28	31	4	3	0	0	83
120	0	10	21	23	5	18	0	1	78
210	0	5	9	11	0	7	0	0	32
221	122	71	50	17	2	2	50	0	314
222	3	65	40	54	3	30	0	0	195
223	1	32	53	55	7	39	0	1	188
224	0	24	43	44	2	27	100	0	240
Total	136	323	449	555	58	326	150	3	2,000

3.2.16 Cross table education level and economic status in the complete recipient data (Van Roij 2015)

Economic status	Education level							
	0	1	2	3	4	5	9	Total
111	7	87	280	586	70	307	0	1,337
112	1	8	25	24	3	3	0	64
120	0	8	19	63	3	23	0	116
210	2	8	17	14	0	3	0	42
221	191	117	36	26	8	3	71	452
222	5	117	93	74	3	39	0	331
223	5	69	100	93	2	20	0	289
224	8	51	69	62	6	37	133	366
Total	219	465	639	942	95	435	204	2,999

3.2.17 Cross table education level and economic status in data after distance hot deck (Van Roij 2015)

Economic status	Education level							
	0	1	2	3	4	5	9	Total
111	13	160	301	565	44	255	0	1,338
112	5	10	20	22	1	6	0	64
120	2	16	30	42	3	23	0	116
210	0	5	16	13	3	7	0	44
221	177	122	51	18	7	5	72	452
222	18	119	94	58	7	35	0	331
223	5	47	74	111	8	44	0	289
224	5	34	58	83	8	45	133	366
Total	225	513	644	912	81	420	205	3,000

3.2.18 Cross table education level and economic status in data after parametric statistical matching (Van Roij 2015)

Economic status	Education level							
	0	1	2	3	4	5	9	Total
111	94	226	283	404	36	204	91	1,338
112	5	10	11	15	2	13	8	64
120	8	20	26	31	4	21	6	116
210	0	10	7	10	2	10	5	44
221	43	63	95	142	10	72	27	452
222	12	51	80	105	14	45	24	331
223	20	45	71	87	11	43	12	289
224	20	76	64	113	12	55	26	366
Total	202	501	637	907	91	463	199	3,000

Weighted results

The results given so far were calculated without using survey weights. When survey weights are used, the conclusions are slightly different than without the use of survey weights. In this subsection we will describe the main differences with the unweighted results.

We note that in our calculations we have used the same statistical tests for weighted data as for unweighted data. This is not fully justified, especially not for the Tables 5.1.3 to 5.1.9 in the Appendix. Instead of using the standard tests one should actually use statistical tests that have been adapted for weighted survey data (see, for instance, Rao and Scott 1979 and 1987, and Holt, Scott and Ewings 1980).
When survey weights are used, independent samples t-tests performed by Linskens (2015) lead to the conclusion that random hot deck generates significantly (at a 5% significance level) more accurate mean scores on education level than distance hot deck (see Table 5.1.2 in the Appendix). When survey weights are used the conclusion is drawn that both distance and random hot deck are not accurate in estimating the true values of education level, i.e. the means in the fused data sets differ significantly from the mean in the complete recipient data, whereas when no survey weights are used the conclusion is drawn that both hot deck methods do yield accurate estimates. However, as we already noted a paired sample t-test would have been more appropriate than an independent samples t-test in our case, so the results by Linskens (2015) should be taken with a (small) grain of salt.

When comparing weighted and unweighted correlation coefficients between estimated and true values of education level, Linskens (2015) found that the use of survey weights does not have considerable impact on this correlation coefficient after the application of random hot deck: $r = 0.02$, $p = 0.475$ for unweighted cases and $r = -0.03$, $p = 0.151$ for weighted cases. However, for distance hot deck the use of survey weights increased the estimated correlation coefficient with a considerable amount, $r = 0.34$, $p < 0.001$ for unweighted cases and $r = 0.43$, $p < 0.001$ for weighted cases.

Linskens (2015) found that dependency relationships between education level and background variables were not distorted when survey weights are used for both random and distance hot deck (see Table 5.1.3). As we have already seen, one or more (significant) distortions of such dependency relationships do occur for unweighted cross tables, depending on the hot deck method of choice (see Table 5). Dependency relationships between education level and target variables in the recipient data set were not distorted irrespective of using survey weights (see Table 5.1.4) or not (see Table 3.2.12).

Tables 5.1.5 and 5.1.6 give weighted cross tables of education level across data sets obtained by Van Roij (2015). These results confirm that education level in the complete recipient data and the data after distance hot deck are associated. A significant result was also found for education level in the complete recipient data and the data after parametric statistical matching. However, the association is weak (Cramér's V = 0.063).

Weighted cross tables between the variables education level and economic status calculated by Van Roij (2015) are presented in Tables 5.1.7, 5.1.8 and 5.1.9. All tables showed significant results, meaning there is an association between education level and economic status in all three data sets. The strength of the association in the complete recipient data was 0.360 (Cramér's V). The Cramér's V of the weighted cross table based on variables after distance hot deck was 0.336, and after parametric statistical matching 0.070, i.e. a considerably weaker association than in the complete recipient data. This means that education level in the complete recipient data is hardly related to education level in the data set after parametric statistical matching.

The weighted joint distributions of education level with economic status in the fused data sets obtained by Van Roij (2015) were compared to the weighted joint distribution in the complete recipient data to investigate whether the distributions could have arisen by sampling from the same population. Results were significant for the data set after distance hot-deck, but non-significant for the data set after parametric statistical matching, This mean that the weighted joint distributions of education level with economic status after distance hot deck and in the complete recipient data could have arisen by sampling from the same population, but that it is very unlikely that the weighted joint distributions of education level with economic status after parametric statistical matching and in the complete recipient data could have arisen from the same population.
For more weighted results we refer to Linskens (2015) and Van Roij (2015).

4. Discussion and key questions for future research

Examining the results presented in this paper, it is clear that the form of parametric statistical matching implemented by Van Roij (2015) performed the worst for the evaluation data. This is further confirmed by Van Roij (2015), who has calculated the Hellinger distance (see, e.g., Bhattacharyya 1943) for the relation between education level and the background variables in the fused data sets compared to the relation found in the complete recipient data (results not reported in the current paper). Overall, the distances are larger after parametric statistical matching than after distance hot deck.

Parametric statistical matching obviously did not lead to more stable (and better results) than random hot deck or distance hot deck. In fact, the use of a model to assist the matching led to worse results in the evaluation study by Van Roij (2015). A possible explanation for this is that the logistic regression models do not capture the distribution of the data well enough, but it remains to be examined whether this possible explanation is correct.

When comparing correlation coefficients between estimated and true values of education level for random hot deck and distance hot deck, Linskens (2015) found that distance hot deck showed to generate the highest positive correlations with the complete recipient data. Thus, distance hot deck showed to generate more reliable estimates in this respect than random hot deck.

When comparing random hot deck and distance hot deck with regard to distortions of dependency relationships between education level and background variables, Linskens (2015) found that distance hot deck significantly distorted the dependency relationship between education level and nationality while random hot deck significantly distorted dependency relationships between education level and nationality, and education level and mother country (see Table 3.2.7).
When comparing both methods with respect to dependency relationships between education level and target variables of the recipient data, Linskens (2015) found that both random hot deck and distance hot deck do not seem to distort these dependency relationships (see Table 3.2.12).

Summarizing, distance hot deck results in significant positive correlations between estimated and true values of education level whereas random hot deck does not and distance hot deck results in fewer (significant) distortions of dependency relationships between the estimated target variable and background variables. Hot deck methods do not differ in distortion of dependency relationships between the

target variable in the donor data and target variables in the recipient data set (see Subsection 'Preserving relationships between target variables' of Section 3.2).
With regard to efficiency, both random hot deck and distance hot deck are incorporated in the R package StatMatch (D'Orazio 2015). Both methods can be efficiently applied to many (large) data sets. With regard to simplicity, distance hot deck requires some additional effort compared to random hot deck, because a distance function has to be determined. However, in our opinion the benefit of more accurate estimates outweighs the (minimal) additional effort of determining a distance function. Therefore, we prefer distance hot deck over random hot deck for our evaluation data.

The results of random hot deck and especially of distance hot deck appear to justify an old observation by Mokken (1984): 'statistical matching gives better results than one might expect'. It is quite easy to show that in the worst case statistical matching gives very bad results. In practice, the situation seems to be considerably less bad. This is illustrated by Tables 3.2.13 and 3.2.15, and by Tables 3.2.16 and 3.2.17, and the conclusion in both evaluation studies that the distribution of education level × economic status after distance hot deck and as this distribution in the complete recipient data could have arising by sampling from the same population.
Although we have established that distance hot deck appears to be the more promising statistical matching method for our evaluation data, some questions remain to be answered. The first question is: is the quality of statistically matched data sufficiently high? The simple answer to this question is that this depends on the quality of data one is aiming for, and that hence statistical matching will not always lead to sufficiently high quality. From this we can derive two key questions for future research: how can we extract more information that is useful for statistical matching from the data sets to be matched, and what additional information, if any, is needed in order to obtain sufficiently accurate statistical results by means of statistical matching? Can information on population totals of variables or relationships between variables known from other sources be used to improve the results of statistical matching? Should we, for instance, conduct a small survey sample to obtain information on the correlation coefficients between a subset of the target variables from different sources? If so, how many and which target variables should be included in the sample survey and how large should the sample survey(s) be? Preferably we would like to limit the subset of variables and the sample size in such a survey as much as possible in order to keep response burden and costs as low as possible.

A strongly related question is: how can we measure the quality of estimates based on statistical matched data in practice, i.e. when one does not know the complete data? Suppose we conduct a small sample survey of the complete data: can we use the results of that sample survey to say something about the quality of the statistically matched data? Alternatively: can we develop a theoretical model for the data including relations between target and background variables that enables us to measure the quality?

Another question that still needs to be answered, and that is our second key question for future research, is: how can we apply statistical matching to Big Data? In theory, statistical matching can be applied to Big Data in (at least) two fundamentally different ways:

- Statistical matching may be used to combine a Big Data source with other data sources. Here Big Data are the data to be statistically matched with other data.
- Alternatively, Big Data may be used as auxiliary information for statistical matching of other data sources. For instance, data collected on the Internet may be utilized to estimate the correlation coefficients between target variables in two data sources to be statistically matched. Here Big Data are 'merely' used to facilitate the statistical matching of other data.

A more technical research question that needs to be answered is: how can we take additional information on, for instance, population totals of variables, logical relationship between variables (for instance, 'males cannot be pregnant'), and statistical relationships between variables such as correlations into account in the statistical matching process?

We hope to answer these questions in future papers on statistical matching.
The possible cases where statistical matching techniques may be useful, such as matching of two non-overlapping surveys with common background variables, matching of Big Data to survey or administrative data, and finding imputation values when for certain groups of units a number of variables are missing by design as mentioned in the Introduction, can help to structure the research on statistical matching and to improve the practical usefulness of this research.

References

Aitchison J. (2003), The Statistical Analysis of Compositional Data (reprint). The Blackburn Press. Caldwell, New Jersey, US.

Bhattacharyya, A. (1943), On a Measure of Divergence between Two Statistical Populations Defined by Their Probability Distributions. Bulletin of the Calcutta Mathematical Society 35, pp. 99-109.

Buelens, B., P. Daas, J. Burger, M. Puts and J. van den Brakel (2014), Selectivity of Big Data, Discussion paper, Statistics Netherlands.

De Jong, W.A.M. (1990a), Exakt en Synthetisch Koppelen, of: Hoe Bedrieglijk Overeenkomsten Kunnen Zijn. Report, Statistics Netherlands.

De Jong, W.A.M. (1990b), Een Handleiding voor Synthetisch Koppelen. Statistics Netherlands.

De Jong, W.A.M. (1991), Statistische Onderzoekingen: Technieken voor het Koppelen van Bestanden. Report, Statistics Netherlands.

De Jong, W.A.M. (1997), Bestandskoppeling ten behoeve van Statistisch Onderzoek. Report, Statistics Netherlands.

De Waal, T., Pannekoek, J. and Scholtus, S. (2011), Handbook of Statistical Data Editing and Imputation. John Wiley & Sons, New York.
D'Orazio, M. (2015), Statistical Matching and Imputation of Survey data with StatMatch. ISTAT, Italy.
D'Orazio, M., M. Di Zio and M. Scanu (2006), Statistical Matching: Theory and Practice. John Wiley and Sons, Chichester, UK.
Eurostat (2013), Statistical Matching: A Model Based Approach for Data Integration. Methodologies and Working papers, Eurostat, Luxembourg.
Holt, D., A.J. Scott and P.D. Ewings (1980), Chi-Squared Test with Survey Data. Journal of the Royal Statistical Society Series A 143, pp. 303-320.
Kadane, J.B. (2001), Some Statistical Problems in Merging Data Files. Journal of Official Statistics 17, pp. 423-433.
Linskens, S.J. (2015), Statistical Matching: A Comparison of Random and Distance Hot Deck. Report, Tilburg University, The Netherlands.
Mokken, R.J. (1984), 'Statistical Matching' of: Synthetische Koppeling. In: Voor Wetenschap en Praktijk, Statistics Netherlands.
Moriarity, C. (2009), Statistical Properties of Statistical Matching. VDM Verlag, Saarbrücken, Germany.
Moriarity, C. and F. Scheuren (2001), Statistical Matching: A Paradigm for Assessing the Uncertainty in the Procedure. Journal of Official Statistics 17, pp. 407-422.
Pawlowsky-Glahn, V. and A. Buccianti (2011), Compositional Data Analysis: Theory and Applications. John Wiley and Sons, Chichester, UK.
Rao, J.N.K. and A.J. Scott (1979), The Analysis of Categorical Data from Complex Sample Surveys. American Statistical Association Conference, Washington, D.C.
Rao, J.N.K. and A.J. Scott (1987), On Simple Adjustments to Chi-Square Tests with Sample Survey Data. The Annals of Statistics 15, pp. 385-397.
Rässler, S. (2004), Data Fusion: Identification Problems, Validity, and Multiple Imputation. Austrian Journal of Statistics 33, pp. 153-171.
Rodgers, W.L. (1984), An Evaluation of Statistical Matching. Journal of Business & Economic Statistics 2, pp. 91-102.
Van Roij, J. (2015), Statistical Matching: A Comparison of Distance Hot Deck and Model-Based Estimation. Report, Tilburg University, The Netherlands.

5. Appendix

5.1.1 Description of variables

Variable	Code	Label
education level	0	pre-primary
	1	primary
	2	lower secondary
	3	upper secondary
	4	post secondary
	5	tertiary
	6	no education
	98	unknown
economic status	111	employee, other
	112	in education and employed
	120	self-employed
	210	unemployed
	221	in education
	222	retired
	223	houseman/housewife
	224	other inactive
	998	unknown
gender	1	male
	2	female
	8	unknown
age	1	0-4 years
	2	5-9 years
	3	10-14 years
	4	15-19 years
	5	20-24 years
	6	25-29 years
	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
	98	unknown
marital status	1	unmarried
	2	married
	3	widowed

Variable	Code	Label
	4	divorced
	8	unknown
position in the household	1110	child
	1121	married without children
	1122	married with children
	1131	cohabitant without children
	1132	cohabitant with children
	1140	single parent
	1210	single
	1220	other
	9998	unknown
household size	111	1 person
	112	2 persons
	113	3 persons
	114	4 persons
	125	5 persons
	126	6 or more persons
	998	unknown
living area in the previous year	1	same COROP-area
	2	other COROP-area
	9	n/a
	998	unknown
nationality	1	Dutch
	2	other European
	3	other
	98	unknown
mother country	1	The Netherlands
	2	other Europe
	3	other
	8	unknown
occupation	1	ISCO 1
	2	ISCO 2
	3	ISCO 3
	4	ISCO 4
	5	ISCO 5
	6	ISCO 6
	7	ISCO 7
	8	ISCO 8
	9	ISCO 9
	998	unknown
	999	unemployed
branch of industry	111	NACE A+B
	122	NACE C+D+E
	124	NACE F
	131	NACE G
	132	NACE H

Variable	Code	Label
	133	NACE I
	134	NACE J
	135	NACE K
	136	NACE L
	137	NACE M
	138	NACE N
	139	NACE O
	200	unemployed
	998	unknown

5.1.2 Weighted means (M) and standard deviations (SD) for education level after distance hot deck, after random hot deck and in the complete recipient data

Distance hot deck (N = 2,000)		Random hot deck (N = 2,000)		Complete recipient data v(N = 2,000)	
M	SD	M	SD	M	SD
3.19	3.75	3.16	4.78	3.01	2.17

5.1.3 Chi-square tests testing independency of education level and background variables on weighted cases (Linskens 2015)

Background variable	Distance hot deck (N = 2,000)		Random hot deck (N = 2,000)		Complete recipient data (N = 2,000)	
	χ^2(df)	p-value	χ^2(df)	p-value	χ^2(df)	p-value
gender	1502.92 (7)	<.001*	1430.90 (7)	<.001*	2035.64 (6)	<.001*
age	270435.40 (119)	<.001	281668.30 (119)	<.001*	298295.30 (102)	<.001*
position in household	90101.47 (49)	<.001*	83058.90 (49)	<.001*	92173.91 (42)	<.001*
size of household	35765.52 (35)	<.001*	33353.93 (35)	<.001*	35611.08 (30)	<.001*
living area in previous year	25105.57 (14)	<.001*	26399.94 (14)	<.001*	28076.34 (12)	<.001*
nationality	4223.28 (21)	<.001*	4647.01 (21)	<.001*	5105.89 (18)	<.001*
mother country	2913.57 (14)	<.001*	3114.90 (14)	<.001*	5824.00 (12)	<.001*
marital status	47961.86 (21)	<.001*	40409.54 (21)	<.001*	42260.94 (18)	<.001*

* indicates that the variables are found to be dependent in the given data set.

5.1.4 Chi-square tests testing independency between education level and target variables of recipient data set on weighted cases (Linskens 2015)

Target variable	Distance hot deck (N = 2,000)		Random hot deck (N = 2,000)		Complete recipient data (N = 2,000)	
	χ^2(df)	p-value	χ^2(df)	p-value	χ^2(df)	p-value
economic status	126629.10 (49)	<0.001*	129824.90 (49)	<0.001*	154227.90 (42)	<0.001*
occupation	38501.72 (63)	<.0001*	39797.34 (63)	<0.001*	100048.70 (54)	<0.001*
branch of industry	41001.81 (84)	<0.001*	45566.60 (84)	<0.001*	74725.72 (72)	<0.001*

* indicates that the variables are found to be dependent in the given data set.

5.1.5 Weighted cross tables of education level across data sets: complete recipient data versus distance hot deck (Van Roij 2015)

Complete recipient data	Distance hot deck							
Education level	0	1	2	3	4	5	9	Total
0	10,513	5,259	834	968	94	1,347	2,821	21,836
1	5,461	15,344	8,253	9,970	682	3,242	0	42,952
2	1,318	10,717	15,447	14,047	946	7,289	0	49,764
3	2,051	11,503	18,307	28,029	3,499	12,549	0	75,938
4	300	651	1,316	3,444	554	2,080	0	8,345
5	396	5,926	8,263	14,755	1,365	7,410	0	38,115
9	2,712	0	0	0	0	0	17,636	20,348
Total	22,115	46,885	48,173	77,059	5,699	37,825	19,607	257,298

5.1.6 Weighted cross tables of education level across data sets: complete recipient data versus parametric statistical matching (Van Roij 2015)

Complete recipient data	Parametric statistical matching							
Education level	0	1	2	3	4	5	9	Total
0	2,569	3,129	4,194	5,917	911	3,157	1,958	21,835
1	2,643	7,026	9,180	14,464	1,686	5,530	2,424	42,953
2	2,549	9,883	10,410	14,289	1,255	8,218	3,160	49,764
3	4,111	12,473	20,176	20,554	2,082	10,861	5,681	75,938
4	496	1,595	1,171	3,347	0	1,326	409	8,344
5	1,686	6,800	9,412	12,465	873	4,729	2,150	38,115
9	1,278	3,751	2,968	6,980	900	2,984	1,488	20,349
Total	15,332	44,657	57,511	78,016	7,707	36,805	17,270	257,298

5.1.7 Weighted cross table of education level and economic status in the complete recipient data (Van Roij 2015)

Economic Status	Education level							
	0	1	2	3	4	5	9	Total
111	408	5,916	18,797	43,306	5,893	25,557	0	99,877
112	109	413	1,933	2,034	403	138	0	5,030
120	0	684	1,328	5,243	192	2,132	0	9,579
210	161	616	889	1,031	0	172	0	2,889
221	18,346	10,681	2,559	3,224	993	309	7,133	43,245
222	569	14,766	12,137	9,603	209	4,505	0	41,789
223	1,240	5,672	6,811	6,455	118	1,370	0	21,666
224	982	4,205	5,310	5,043	536	3,931	13,215	33,222
Total	21,835	42,953	49,764	75,939	8,344	38,114	20,348	257,297

5.1.8 Weighted cross table of education level and economic status after distance hot deck (Van Roij 2015)

Economic Status	Education level							
	0	1	2	3	4	5	9	Total
111	793	12,651	21,071	42,671	3,615	19,138	0	99,939
112	338	746	1,584	1,976	35	352	0	5,031
120	93	1,420	2,154	3,686	345	1,881	0	9,579
210	0	283	837	1,093	251	424	0	2,888
221	17,727	11,794	3,323	1,563	609	989	7,242	43,247
222	3,381	16,125	11,545	6,076	1,096	3,564	0	41,787
223	222	3,995	5,750	7,211	517	3,970	0	21,665
224	197	2,384	6,155	6,999	672	3,600	13,215	33,222
Total	22,751	49,398	52,419	71,275	7,140	33,918	20,457	257,358

5.1.9 Weighted cross table education of level and economic status after parametric statistical matching (Van Roij 2015)

Economic Status	Education level							
	0	1	2	3	4	5	9	Total
111	5,988	17,348	22,523	31,360	2,248	13,720	6,753	99,940
112	367	727	1,009	1,067	159	1,124	579	5,032
120	618	1,616	2,213	2,462	247	1,717	705	9,578
210	0	653	344	491	378	564	459	2,889
221	4,146	6,107	9,532	13,638	1,001	6,247	2,576	43,247
222	895	7,615	10,281	13,491	1,430	5,323	2,753	41,788
223	1,720	3,351	5,389	5,953	1,107	3,495	650	21,665
224	1,599	7,239	6,222	9,615	1,136	4,617	2,795	33,223
Total	15,333	44,656	57,513	78,077	7,706	36,807	17,270	257,362

Explanation of symbols

Empty cell	Figure not applicable
.	Figure is unknown, insufficiently reliable or confidential
*	Provisional figure
**	Revised provisional figure
2014–2015	2014 to 2015 inclusive
2014/2015	Average for 2014 to 2015 inclusive
2014/'15	Crop year, financial year, school year, etc., beginning in 2014 and ending in 2015
2012/'13–2014/'15	Crop year, financial year, etc., 2012/'13 to 2014/'15 inclusive

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

Colofon

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

Prepress
Statistics Netherlands, Studio BCO

Design
Edenspiekermann

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