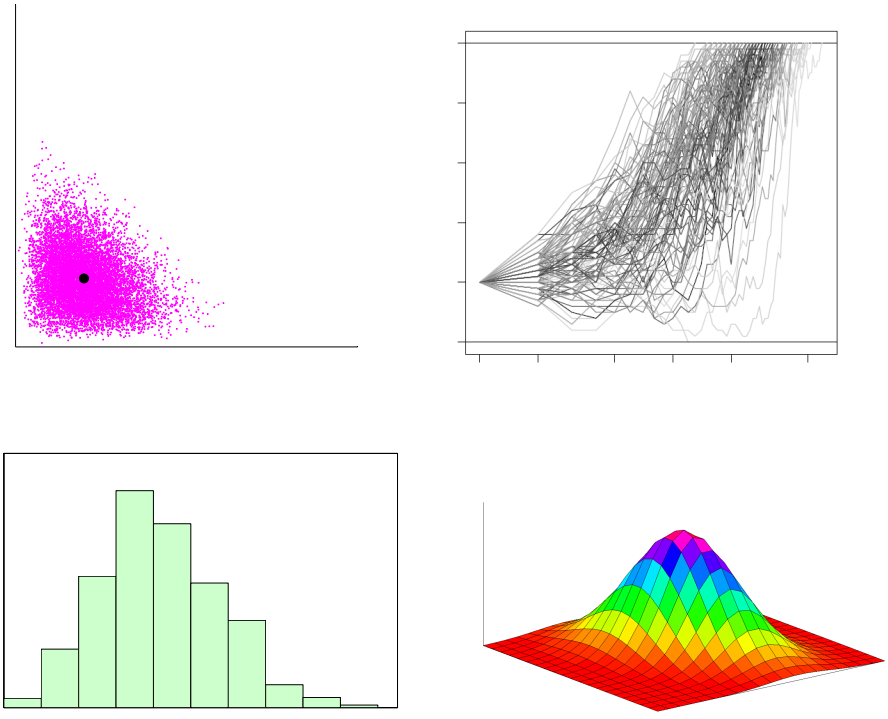
Philipps Universität Marburg

Discussion Papers on Statistics and Quantitative Methods

Statistical Matching for Complex Samples

Karlheinz Fleischer & Heiko Groenitz

5 / 2015



Download from:
http://www.uni-marburg.de/fb02/statistik/forschung/discpap

Coordination: Prof. Dr. Karlheinz Fleischer • Philipps-University Marburg
School of Business and Economics • Research group Statistics
Universitätsstraße 25 • D-35037 Marburg
E-Mail: k.fleischer@wiwi.uni-marburg.de

Statistical Matching for Complex Samples

Karlheinz Fleischer & Heiko Groenitz[1]

13.11.2015

Abstract

Statistical Matching (or equivalent data fusion) is the conflation of two original data sets with different units to a final overall data set. One original data set contains characteristics X and Z, the other Y and Z. The final data consist of values for X, Y, and Z. The merging of the input data sets is based on the common characteristic Z. Well-working fusions allow researchers to combine single databases and discover new relations by applying methods for complete data. Furthermore, long, expensive, and tiring surveys can be avoided. There are plenty variants of matching procedures and applications. However, theoretical properties of fusion data for original samples drawn by more complex sampling schemes, which are very relevant in practice, remain unclear. Therefore, we consider such more complex input samples in this article. We mathematically derive the after-fusion density and identify situations in which the densities before and after fusion are equal. Estimations based on fusion data are also addressed. Our analytic results are based on some idealizing assumption. Hence, we finally investigate their stability in simulations. Our findings sensitize data fusion users that successful fusions depend on quite strong requirements.

Zusammenfassung

Statistisches Matchen, auch Datenfusion genannt, ist das Zusammenfügen von zwei Ausgangsdatensätzen mit unterschiedlichen Einheiten zu einem finalen Gesamtdatensatz. Der eine Ausgangsdatensatz beinhaltet Merkmale X und Z, der andere Y und Z. Der resultierende Gesamtdatensatz besteht aus Werten für X, Y und Z. Die Verschmelzung der ursprünglichen Datensätze basiert auf dem gemeinsamen Merkmal Z. Funktionierende Fusionen erlauben es Forschern, einzelne Datenbanken zusammenzufügen und neue Beziehungen mit Methoden für vollständige Daten zu suchen. Darüber hinaus, könnte man auf lange, teure und ermüdende Umfragen verzichten. Es existieren vielfältige Varianten von Fusionsmethoden und zahlreiche Anwendungen in der Literatur. Jedoch bleiben theoretische Eigenschaften von Fusionsdaten bei komplexeren, in der Praxis sehr relevanten Ziehungsverfahren für die Ausgangsstichproben unklar. Daher betrachten wir solche komplexeren Ausgangsstichproben in diesem Artikel. Wir leiten die Dichte nach der Fusion mathematisch her und identifizieren Situationen, in denen die Dichten vor und nach der Fusion übereinstimmen. Die Schätzung aus Fusionsdaten wird ebenfalls adressiert. Unsere analytischen Resultate basieren auf einer idealisierenden Annahme. Aus diesem Grund untersuchen wir anschließend die Stabilität der Resultate mit Hilfe von Simulationen. Unsere Erkenntnisse sensibilisieren Nutzer von Datenfusionen dafür, dass erfolgreiche Fusionen von recht starken Forderungen abhängen.

KEYWORDS: data fusion; data quality; imputation; missing values; sampling theory

[1]Institution of both authors: Philipps-University Marburg, School of Business and Economics, Research group Statistics, Universitätsstraße 25, D-35037 Marburg, Germany. E-mail: k.fleischer@wiwi.uni-marburg.de and groenitz@staff.uni-marburg.de.

1 Introduction

Starting point for statistical matching (also known as data fusion) are two data sets from two independent samples. For the first sample, data on characteristics X and Z are gathered while values for characteristics Y and Z are recorded for the other sample. X, Y, and Z are possibly multivariate. Z is present in both samples and is called common characteristic. Typically, Z comprises socio-demographic attributes. For units in the first sample, we have missing values for Y whereas X is unobserved for units in the second sample. Consequently, we have a special missing data pattern. Notice, according to the independence of the original samples, the sets of sample units of the first and second data set are usually different. This distinguishes our situation from a record linkage situation, in which different information on the same units of a population is available in different data sources like official registers.

Now, statistical matching means to search for each individual in the first sample, a "similar" unit in the second sample and impute the corresponding Y value to the unit from the first sample. The similarity is measured via Z or some function of it. In this way, we obtain an overall data set containing values for all characteristics X, Y, and Z, where the values of X and Z come from the first sample and the Y values are from the second sample. Hence, statistical matching is the conflation of our two original data sets into an overall data set. This final data set serves as basis for statistical investigations. According to the described imputation procedure, the names recipient sample (R sample) and donor sample (D sample) are often used for the first and second sample, respectively.

Successful fusions enable investigators to combine multiple available databases to one big database, which can be analyzed by standard complete-data methods. Furthermore, for planned huge surveys, long and all-embracing interviews can be replaced by multiple surveys with less questions and data fusion afterwards. This is especially advantageous, because long interviews are expensive and tire the respondents resulting often in answer refusals.

There are divers variants of fusion methods and evaluations of fusion data in the existing literature. Sims (1972) remarks the importance of conditional independence of the characteristics not jointly observed given the common characteristic for a successful fusion. Kadane (1978) considers normally distributed (X, Y, Z) and suggests to derive the valid values of the covariance of X and Y and base the fusion on these. In Rodgers (1984), evaluations of statistical matching based on special, fixed data sets as well as alternatives to statistical matching can be found. Rubin (1986) proposes file concatenation with adjusted weights and multiple imputation. Multiple imputation means that matching is operated under different assumptions to show variability of the results. Goel and Ramalingam (1989) provide an empirical evaluation of different matching strategies. Kovacevic and Liu (1994) assess fusion methods in simulations, where the effect of auxiliary data is also addressed. Rässler and Fleischer (1998) address mathematically and in simulations the distribution and distribution parameters after the fusion for simple random sampling with replacement (SRSWR) for both original samples. Moriarity and Scheuren (2001) modify the work of Kadane (1978). Moriarity and Scheuren (2003) show improvements on Rubin (1986). Rässler (2004) presents a non-iterative Bayesian approach to statistical matching (NIBAS). Gilula et al. (2006) suggest a direct estimation of the joint distribution of the not jointly observed characteristics without using a matched data set. Conti et al. (2008) consider a different data pattern with less variables as well as partly complete data and compare some matching procedures with imputation techniques which are based on the estimation of the regression function. Stuart (2010) gives an overview how matching can be used to estimate causal effects. D'Orazio et al. (2012) compare different matching methods using simulations, in which the sample units of an EU-SILC survey in Italy are treated as population, one input sample is drawn by stratified simple random sampling, and stratified one stage cluster sampling is considered for the other input sample. These comparisons, however, do not involve the estimation quality for the distribution of the characteristics not jointly observed. Reiter (2012) improves the standard multiple imputation variance estimator for

the data fusion situation. Conti et al. (2013) measure estimation uncertainty for a statistical matching situation. For more fusion methods or more details, we refer to the monographs Rässler (2002) and D'Orazio et al. (2006).

Data fusions are applied in practice in various fields. For instance, in media analysis, data on television viewing behavior are often merged with data on purchasing behavior with the aim to locate the optimal slot for advertisement of a product. For example, in Germany, such fusions are conducted by the institution AGMA. Details on many applications of statistical matching can be found in the books Rässler (2002, Chapter 3) and D'Orazio et al. (2006, Chapter 7.1). Rässler (2002) first lists fusions in Europe, afterward, fusions in the USA and Canada are addressed. D'Orazio et al. (2006) describe a variety of applications in microsimulation, market research, and official statistics.

Although there are a lot of matching schemes and practical matching applications, theoretical properties of fusion data are insufficiently explored in the literature so far. Indeed, for SRSWR for both original sample, it is known that conditional independence of X and Y given Z is a central requirement for a successful fusion (e.g., Rässler and Fleischer (1998), Rässler (2002, Chapter 2.3)). We review this case in Section 2. However, for other, more complex sampling designs of the original samples, which are often applied in practice, the distribution after the fusion and criteria under which the distributions before and after fusion are equal remain unclear.

This article addresses these open points. For more complex sampling schemes of the input samples, we analytically derive the density after the matching process (Section 3). The after-fusion density consists of two factors, one depends on the R sample, the other on the D sample. In Section 4, we compare the density after fusion with the density before the fusion and identify situations in which the densities before and after matching coincide. In other words, when do fusion data possess a good quality? We also explain how to estimate from fusion data. For our explicit formulas for the after-matching density, some idealizing assumption on one original sample is made. For this reason, we eventually analyze the robustness of our analytic results with respect to this assumption in simulations (Section 5).

2 Some notation and the case of SRSWR for both samples

Let $G = \{e_1, ..., e_N\}$ be a finite population of N units, X, Y, and Z possibly multivariate characteristics on this population, as well as x_i, y_i, and z_i $(i = 1, ..., N)$ the ith population unit's value for X, Y, and Z, respectively. We consider two independent samples from G. For the R sample, data on X and Z are gathered while values for Y and Z are recorded for the D sample. For each individual in the R sample, we search a unit in the D sample such that the Z values of these two units are identical or at least closest to each other. If multiple units from the D sample are possible, one of these is selected randomly. When such a unit from the D sample was detected, we impute the corresponding Y value to the unit from the R sample.

Let n_R and n_D be the size of the R and D sample, respectively. Set $n = n_R + n_D$ for the total sample size. If n_R is fixed, i.e., not random, we define X_i, Y_i, Z_i, U_i, $\tilde{X}_i$, $\tilde{Y}_i$, $\tilde{Z}_i$ for $i = 1, ..., n_R$ as follows. X_i, Y_i, and Z_i describe the ith R sample unit's outcome of X, Y, and Z, respectively. $U_i \in \{1, ..., N\}$ indicates which population unit is selected as ith sample unit in the R sample, e.g., $U_1 = 105$ means that individual e_{105} is the first unit in the R sample. Furthermore, $\tilde{X}_i$, $\tilde{Y}_i$, and $\tilde{Z}_i$ are random variables representing the ith data row in the matched data set, that is, $\tilde{X}_i = X_i$, $\tilde{Z}_i = Z_i$, and $\tilde{Y}_i$ is a Y value imputed from a donor unit. If n_D is also fixed, we additionally introduce random variables X_i, Y_i, Z_i, U_i for $i = n_R + 1, ..., n$. Here, X_{n_R+l}, Y_{n_R+l}, and Z_{n_R+l} $(l = 1, ..., n_D)$ represent the lth D sample unit's value of X, Y, and Z, respectively, and $U_{n_R+l} \in \{1, ..., N\}$ describes which population unit is

the lth unit in the D sample.

Densities of attributes on the universe or of random variables are denoted by f with associated index. Since we consider a finite population, the densities are with respect to the counting measure, but may be approximated by a continuous one on occasion. E.g., $f_{X,Y}(x,y)$ is the relative frequency how often $(X,Y)=(x,y)$ appears in G, $f_{Z_i}(z)$ is the probability that random variable Z_i equals z. An important task in the matching context is to investigate under which criteria the distribution after the fusion equals the distribution before the fusion. That is, when does

$$(\tilde{X}_i, \tilde{Y}_i, \tilde{Z}_i) \sim (X_i, Y_i, Z_i) \tag{1}$$

$(i = 1, ..., n_R)$ hold? For the case of SRSWR for both input samples, this question has already been answered in the literature. E.g., Rässler and Fleischer (1998) comprehensively treat this case and explicitly derive the distribution and distribution parameters after the fusion for the idealized situation in which

the distributions of the vector (Y,Z) in the D sample and in the population are equal. (2)

In particular, it can be shown that the vectors $(\tilde{X}_i, \tilde{Y}_i, \tilde{Z}_i)$, $i = 1, ..., n_R$, are iid, and

$$f_{\tilde{X}_i,\tilde{Y}_i,\tilde{Z}_i}(x,y,z) = f_{X,Z}(x,z) \cdot f_{Y|Z}(y|z) \tag{3}$$

holds implying that (1) is true if and only if X and Y are conditionally independent given Z. Several identities on marginal distributions and moments (e.g., $f_{\tilde{Y}_i} = f_{Y_i}$, $\mathbb{E}(\tilde{Y}_i^l) = \mathbb{E}(Y_i^l)$ for $l \in \mathbb{N}$), as well as

$$\mathbb{E}(Cov(X,Y|Z)) = Cov(X,Y) - Cov(\tilde{X}_i, \tilde{Y}_i) \tag{4}$$

can also be established. Notice, when we apply the symbol $\mathbb{E}$ to characteristics on the population, the corresponding population mean is meant. The mean $\mathbb{E}(Cov(X,Y|Z))$ represents a quality measure for the fusion, because values near zero indicate that the covariance between X and Y in the population is close to the covariance after matching. Furthermore, if the reproduction of $Cov(X,Y)$ by the fusion is of interest, (4) implies that the requirement that X and Y are on average conditionally uncorrelated suffices, conditional independence is not necessary. Rässler and Fleischer (1998) also investigate by simulations the accuracy of estimates (e.g., for means, variances, covariances) computed from the fusion data set, where nearest neighbor (NN) matching is applied. It turns out that the results for NN matching match the theoretical results (that is, the results under (2)) quite well.

3 Distribution after fusion for complex sampling

In practice, sampling designs that are more complex than SRSWR are often applied. For this reason, we derive the after-fusion distribution for other combinations of sampling schemes for the R and D sample.

3.1 General considerations

For complex sampling schemes, it is often adequate to introduce further characteristics T_R and T_D on G being relevant for the sampling procedure of the R sample and D sample, respectively. For instance, these attributes may describe the stratum a respondent belongs to, the single draw selection probability, or the inclusion probability. For $i = 1, ..., N$, the ith population unit's value of T_R and T_D is denoted by t_i^R and t_i^D, respectively. Furthermore, let $T_1, ..., T_{n_R}$ be the values of T_R for the R sample units, i.e., $T_i = t_{U_i}^R$ for $i = 1, ..., n_R$. We further introduce the random variables $T_i = t_{U_i}^D$ for $i = n_R + 1, ..., n$, that is, these variables describe the values of T_D for the units in the D sample.

Moreover, we define $\tilde{T}_i$ $(i = 1, ..., n_R)$ to be the value of T_R corresponding to the ith data row in the matched data file, where, of course, $\tilde{T}_i = T_i$ holds. Since we have quite a lot necessary quantities, we provide an overview in Table 1.

quantities on the population	
$\{e_1, ..., e_N\}$	population G
$x_1, ..., x_N$	values of X
$y_1, ..., y_N$	values of Y
$z_1, ..., z_N$	values of Z
$t_1^R, ..., t_N^R$	values of T_R
$t_1^D, ..., t_N^D$	values of T_D

variables for two original samples				
X_1	Y_1	Z_1	U_1	T_1
		$\vdots$		
X_{n_R}	Y_{n_R}	Z_{n_R}	U_{n_R}	T_{n_R}
X_{n_R+1}	Y_{n_R+1}	Z_{n_R+1}	U_{n_R+1}	T_{n_R+1}
		$\vdots$		
X_n	Y_n	Z_n	U_n	T_n

variables for fusion data			
$\tilde{X}_1$	$\tilde{Y}_1$	$\tilde{Z}_1$	$\tilde{T}_1$
		$\vdots$	
$\tilde{X}_{n_R}$	$\tilde{Y}_{n_R}$	$\tilde{Z}_{n_R}$	$\tilde{T}_{n_R}$

Table 1: Summary of quantities on the population, random variables for R and D sample, and random variables representing the data after fusion. Notice, e.g., the density $f_{X,Z}$ can be different from f_{X_i,Z_i}.

For the distribution after matching, we generally have for $i = 1, ..., n_R$

$$f_{\tilde{X}_i,\tilde{Y}_i,\tilde{Z}_i,\tilde{T}_i}(x, y, z, t) = f_{\tilde{X}_i,\tilde{Z}_i,\tilde{T}_i}(x, z, t) \cdot f_{\tilde{Y}_i|\tilde{X}_i,\tilde{Z}_i,\tilde{T}_i}(y|x, z, t) = f_{X_i,Z_i,T_i}(x, z, t) \cdot f_{\tilde{Y}_i|Z_i}(y|z), \quad (5)$$

where the latter equality holds, because the X, Z, and T_R values come directly from the R sample and the fusion is based only on the Z value. To obtain a convenient expression for the second factor in (5), idealizing assumptions are required. In particular, we will assume that

the distribution of the vector (Y, Z) in the D sample equals its expected distribution. (6)

Notice, for SRSWR, (6) and (2) are equivalent. Typically, assumption (6) implies that exact matches are possible. Then, $f_{\tilde{Y}_i|Z_i}(y|z)$ is the probability to randomly select a unit with Y value y among the donor sample units with the Z value of the ith R sample unit. Examples for the concrete computation of the two factors in (5) are given in the next subsections.

3.2 The first factor

The first factor in (5) depends on the sampling scheme for the R sample, but not on the sampling procedure for the D sample. We consider some concrete sampling methods for the R sample in the following and compute this factor.

3.2.1 General with-replacement (GWR) sampling for *R* sample

Here, the n_R recipient units are selected with replacement and independent of each other where we denote the probability that population unit j is selected in the ith draw $(i = 1, ..., n_R)$ by p_j^R

($p_j^R \in (0,1)$, $\sum_{j=1}^N p_j^R = 1$). That is, the p_j^R are single draw selection probabilities and we have

$$\mathbb{P}\left(\bigcap_{i=1}^{n_R}\{U_i = j_i\}\right) = \prod_{i=1}^{n_R} p_{j_i}^R, \quad j_i \in \{1,...,N\}.$$

Such GWR sampling reduces to SRSWR when every $p_j^R = 1/N$. The p_j^R can be viewed as values of a characteristics P_R. Furthermore, let P_i denote the single draw selection probability of sample unit i, i.e., $P_i = p_{U_i}^R$ for $i = 1,...,n_R$. Finally, let $\tilde{P}_i$ ($i = 1,...,n_R$) be the single draw selection probability corresponding to the ith row in the matched data set. Notice, P_R, P_i and $\tilde{P}_i$ play the role of T_R, T_i, and $\tilde{T}_i$ from Subsection 3.1. For the first factor in (5), it is true that

$$f_{X_i,Z_i,P_i}(x,z,p) = \sum_{j=1}^{N} p_j^R \cdot 1_{\{(x,z,p)\}}(x_j, z_j, p_j^R). \tag{7}$$

3.2.2 Stratified SRSWR for *R* sample

Stratified sampling means that the universe G is divided into a number of subpopulations (strata) and a sample is drawn from each subpopulation. With respect to the R sample, say G is divided into s_R strata $G_1^R,...,G_{s_R}^R$ with G_l^R having size N_l^R ($N_1^R + ... + N_{s_R}^R = N$). From stratum G_l^R, a simple random sample with replacement of size n_l^R ($n_1^R + ... + n_{s_R}^R = n_R$) is drawn where these s_R samples are independent. Define for $l = 1,...,s_R$ the index set

$$I_l^R = \{i \in \{1,...,n_R\} \ : \ \text{sample unit } i \text{ belongs to stratum } l\},$$

i.e., I_l^R indicates which sample units were drawn from the lth stratum. Let S_R be an attribute on the population G (with possible values $1,...,s_R$) representing the stratum an individual belongs to. For $i = 1,...,n_R$, let S_i denote the stratum the ith sample unit belongs to and define $\tilde{S}_i$ to be the stratum associated with the ith data row in the matched data file ($\tilde{S}_i = S_i$ holds). S_R, S_i, $\tilde{S}_i$ correspond to T_R, T_i, $\tilde{T}_i$ from Subsection 3.1. Notice that we have $S_i = l$ for $i \in I_l^R$. Consider an index $i \in I_l^R$, that is, we look at an R sample unit selected from stratum G_l^R and have $S_i = l$. For such i, we have

$$f_{X_i,Z_i,S_i}(x,z,l) = f_{X_i,Z_i}(x,z) = f_{X,Z|S_R}(x,z|l). \tag{8}$$

3.2.3 General without-replacement sampling for *R* sample

In this subsection, we assume that the R sample is drawn by a without-replacement scheme and that the sample size n_R is fixed, that is, not random. Set $\pi_{ij} = \mathbb{P}(U_i = j)$ for $i = 1,...,n$ and $j = 1,...,N$, i.e., π_{ij} is the probability that the jth population unit is selected in the ith draw. Define π_j^R to be the inclusion probability of the jth population unit ($j = 1,...,N$) with respect to the R sample. Consequently, $\pi_{1j} + ... + \pi_{n_R,j} = \pi_j^R$ holds. To give an example, for simple random sampling without replacement (SRSWOR), we have $\pi_j^R = n_R/N$ and $\pi_{ij} = 1/N$. The π_j^R are values of a characteristic Π_R. Set $\Pi_i = \pi_{U_i}^R$, that is, the random variable Π_i represents the inclusion probability for the R sample of the ith R sample unit. Let $\tilde{\Pi}_i$ be the inclusion probability for the R sample corresponding to ith row in the data set after fusion. Thus, Π_R, Π_i, and $\tilde{\Pi}_i$ correspond to T_R, T_i, and $\tilde{T}_i$ in Subsection 3.1. Then, the first factor in (5) equals

$$f_{X_i,Z_i,\Pi_i}(x,z,\pi) = \sum_{j=1}^{N} 1_{\{(x,z,\pi)\}}(x_j, z_j, \pi_j^R) \cdot \pi_{ij}. \tag{9}$$

3.3 The second factor

In this subsection, we calculate the second factor from (5) for some concrete sampling schemes for the D sample.

3.3.1 GWR sampling for D sample

Adapting the notation from Subsection 3.2.1, p_j^D $(j = 1, ..., N)$ is the single draw selection probability of population unit j with respect to the D sample and the p_j^D can be viewed as values of a characteristic P_D. To derive a convenient expression of the second factor in (5), we assume that the distribution of (Y, Z) in the D sample equals its asymptotic sampling distribution. According to the strong law of large numbers (SLLN),

$$n_D^{-1} \cdot \sum_{j=n_R+1}^{n_R+n_D} 1_{\{(y,z)\}}(Y_j, Z_j)$$

converges almost surely (a.s.) to

$$\mathbb{P}(Y_{n_R+1} = y, Z_{n_R+1} = z) = \sum_{j=1}^{N} p_j^D \cdot 1_{\{(y,z)\}}(y_j, z_j)$$

as $n_D \to \infty$, that is, the SLLN guarantees at least an approximate validity of this assumption for a large D sample. This assumption is equivalent to (6) and implies that exact matches of recipient and donor units with respect to the Z value are possible. Then, we obtain

$$f_{\tilde{Y}_i|Z_i}(y|z) = \frac{\sum_{j=1}^{N} p_j^D \cdot 1_{\{(y,z)\}}(y_j, z_j)}{\sum_{j=1}^{N} p_j^D \cdot 1_{\{(z)\}}(z_j)}. \tag{10}$$

3.3.2 D sample stratified by Z

For the D sample, say G consists of s_D strata $G_1^D, ..., G_{s_D}^D$ with sizes $N_1^D, ..., N_{s_D}^D$. The strata are defined via outcomes of Z. Let S_D be a characteristic on G with values in $\{1, ..., s_D\}$ describing to which stratum an individual belongs. Consequently, S_D is a function of Z. We draw s_D independent simple random samples with replacement from G where the lth subsample is of size n_l^D and selected from G_l^D. We assume (6) implying that exact matches are possible. Consider a value z such that an individual with Z value equal to z belongs to stratum G_l^D. Then, for $i = 1, ..., n_R$,

$$f_{\tilde{Y}_i|Z_i}(y|z) = f_{Y|Z,S_D}(y|z, l) = f_{Y|Z}(y|z) \tag{11}$$

holds.

3.3.3 D sample stratified by Y

Consider the situation of Subsection 3.3.2 except that S_D is now a function of Y and not of Z. We again assume (6). Consider y corresponding to stratum G_l^D. This means that an individual having Y value y belongs to stratum G_l^D. Then, we have

$$f_{\tilde{Y}_i|Z_i}(y|z) = \frac{n_l^D \cdot f_{Y,Z|S_D}(y, z|l)}{n_1^D \cdot f_{Z|S_D}(z|1) + ... + n_{s_D}^D \cdot f_{Z|S_D}(z|s_D)}. \tag{12}$$

3.3.4 General without-replacement sampling for D sample

Let us first extend the notation of Subsection 3.2.3. The quantity π_j^D is the jth population unit's inclusion probability for the D sample. The values π_j^D are outcomes of a characteristic Π_D defined on G. Moreover, for $j = 1, ..., N$, let H_j^D be a random variable with possible values in $\{0, 1\}$ where $H_j^D = 1$ if the jth population unit is selected in the D sample and $H_j^D = 0$ else. Subsequently, we assume as in (6) that the distribution of (Y, Z) in the D sample equals its expected distribution. Then, exact matches are possible. For the expected number how often $(Y, Z) = (y, z)$ appears in the D sample,

$$\mathbb{E}\left(\sum_{j=1}^{N} 1_{\{(y,z)\}}(y_j, z_j) \cdot H_j^D\right) = \sum_{j=1}^{N} 1_{\{(y,z)\}}(y_j, z_j) \cdot \pi_j^D$$

holds. This implies

$$f_{\tilde{Y}_i|Z_i}(y|z) = \frac{\sum_{j=1}^{N} 1_{\{(y,z)\}}(y_j, z_j) \cdot \pi_j^D}{\sum_{j=1}^{N} 1_{\{z\}}(z_j) \cdot \pi_j^D}.$$

4 Comparison of distribution before and after fusion

In this section, we address concrete combinations of sampling schemes for the R and the D sample and investigate when the distribution before the fusion equals the distribution after the fusion.

4.1 General with-replacement sampling for both samples

We consider the settings of Subsections 3.2.1 and 3.3.1 and assume that

$$p_j^R = p_j^D (=: p_j) \tag{13}$$

holds for $j = 1, ..., N$, i.e., we have the same single draw selection probabilities for both samples and $P_R = P_D (=: P)$. For instance, in the special case where the single draw selection probabilities are proportional to some function g of Z (i.e., pps sampling if $g(Z)$ can be interpreted as "size"), (13) is fulfilled. Then, we have that the vectors $(\tilde{X}_i, \tilde{Y}_i, \tilde{Z}_i, \tilde{P}_i)$, $i = 1, ..., n_R$, are iid and the density after fusion is

$$\begin{aligned} f_{\tilde{X}_1,\tilde{Y}_1,\tilde{Z}_1,\tilde{P}_1}(x, y, z, p) &= f_{X_1,Z_1,P_1}(x, z, p) \cdot f_{Y_{n_R+1}|Z_{n_R+1}}(y|z) \\ &= f_{X_1,Z_1,P_1}(x, z, p) \cdot f_{Y_1|Z_1}(y|z) \\ &= f_{Z_1}(z) \cdot f_{X_1,P_1|Z_1}(x, p|z) \cdot f_{Y_1|Z_1}(y|z). \end{aligned}$$

As a consequence,

$$f_{\tilde{X}_1,\tilde{Y}_1,\tilde{Z}_1,\tilde{P}_1}(x, y, z, p) = f_{X_1,Y_1,Z_1,P_1}(x, y, z, p),$$

i.e., the densities before and after fusion are equal, if and only if

$$(X_1, P_1) \text{ and } Y_1 \text{ are independent given } Z_1. \tag{14}$$

Under (14), we can unbiasedly estimate the mean of functions of (X, Y, Z) from the fusion data using the Hansen-Hurwitz estimator. That is, the mean of a characteristic $h(X, Y, Z)$ with a measurable function h is estimated by

$$\hat{\mathbb{E}}(h(X,Y,Z)) = \frac{1}{n_R}\sum_{i=1}^{n_R}\frac{h(\tilde{X}_i,\tilde{Y}_i,\tilde{Z}_i)}{N\cdot\tilde{P}_i} \tag{15}$$

with $\mathbb{E}(\hat{\mathbb{E}}(h(X,Y,Z))) = \mathbb{E}(h(X,Y,Z))$. We give three examples in which (14) holds:

(i) On the population, X and Y are conditionally independent given Z and, furthermore, P is equal to $1/N$.

(ii) P_1 and (X_1, Y_1, Z_1) are independent and X_1 and Y_1 are independent given Z_1.

(iii) P is a function g of Z (that is, $p_j = g(z_j)$ for $j = 1, ..., N$) and X and Y are conditionally independent given Z on the population.

In (i), we have SRSWR. In (i) and (ii), the claim follows by standard calculations with (conditional) densities. Moreover, regarding (ii) notice if we have on the population that P and (X, Y, Z) are independent and that X and Y are independent given Z, this does not automatically imply requirement on the random variables stated in (ii) due to the GWR sampling scheme. The case (iii) is very important, because it covers the situation of single draw selection probabilities proportional to some function of Z (pps sampling). We now proof that the claim follows from (iii). It is true that

$$\begin{aligned} f_{X_1,Y_1|Z_1}(x,y|z) &= \frac{\sum_{j=1}^N g(z_j)\cdot 1_{\{x\}}(x_j)\cdot 1_{\{y\}}(y_j)\cdot 1_{\{z\}}(z_j)}{\sum_{j=1}^N g(z_j)\cdot 1_{\{z\}}(z_j)} \\ &= \frac{\sum_{j=1}^N g(z)\cdot 1_{\{x\}}(x_j)\cdot 1_{\{y\}}(y_j)\cdot 1_{\{z\}}(z_j)}{\sum_{j=1}^N g(z)\cdot 1_{\{z\}}(z_j)} \\ &= f_{X,Y|Z}(x,y|z) = f_{X|Z}(x|z)\cdot f_{Y|Z}(y|z). \end{aligned} \tag{16}$$

We can show similarly $f_{X_1|Z_1}(x|z) = f_{X|Z}(x|z)$ and $f_{Y_1|Z_1}(y|z) = f_{Y|Z}(y|z)$. For (14), we have to establish

$$f_{X_1,P_1|Z_1}(x,p|z)\cdot f_{Y_1|Z_1}(y|z) = f_{X_1,P_1,Y_1|Z_1}(x,p,y|z).$$

For $p \neq g(z)$, both sides of the equation are zero. For $p = g(z)$, this equation reduces to $f_{X_1|Z_1}(x|z)\cdot f_{Y_1|Z_1}(y|z) = f_{X_1,Y_1|Z_1}(x,y|z)$, which is shown above. Thus, the proof is completed.

4.2 SRSWR (*R* sample) and stratification by *Z* (*D* sample)

With Subsection 3.3.2, we have for z such that an individual with Z value equal to z belongs to the stratum G_l^D and $i = 1, ..., n_R$,

$$f_{\tilde{X}_i,\tilde{Y}_i,\tilde{Z}_i}(x,y,z) = f_{X,Z}(x,z)\cdot f_{Y|Z}(y|z).$$

Hence, for $i = 1, ..., n_R$, the equivalences

$$(\tilde{X}_i,\tilde{Y}_i,\tilde{Z}_i)\sim(X_i,Y_i,Z_i) \iff \quad (\tilde{X}_i,\tilde{Y}_i,\tilde{Z}_i)\sim(X,Y,Z) \iff \quad X \text{ and } Y \text{ are independent given } Z$$

are true. In other words, we have the same condition under which the densities before and after matching are equal as in Section 2 for twice SRSWR.

4.3 SRSWR (*R* sample) and stratification by *Y* (*D* sample)

With Subsection 3.3.3, we have for y corresponding to stratum G_l^D

$$f_{\tilde{X}_i,\tilde{Y}_i,\tilde{Z}_i}(x,y,z) = f_{X,Z}(x,z)\cdot\frac{n_l^D\cdot f_{Y,Z|S_D}(y,z|l)}{n_1^D\cdot f_{Z|S_D}(z|1)+...+n_{s_D}^D\cdot f_{Z|S_D}(z|s_D)}.$$

For instance, in the special case with

$$s_D = 2, n_1^D = n_2^D, f_{S_D}(1) = f_{S_D}(2) = 1/2, \tag{17}$$

the second factor reduces to $f_{Y|Z}(y|z)$ and we have:

$$\text{distribution before and after fusion are equal} \iff X \text{ and } Y \text{ independent given } Z. \tag{18}$$

However, in general, the second factor is unequal to $f_{Y|Z}(y|z)$ and equivalence (18) does not hold.

4.4 Stratified SRSWR (*R* sample) and SRSWR (*D* sample)

The results from Subsection 3.2.2 are useful here. For $i \in I_l^R$, that is, for an R sample unit from the lth stratum, we have $S_i = l$ and the density before matching is $f_{X_i,Y_i,Z_i,S_i}(x,y,z,l) = f_{X_i,Y_i,Z_i}(x,y,z) = f_{X,Y,Z|S_R}(x,y,z|l)$ while the density after matching equals

$$f_{\tilde{X}_i,\tilde{Y}_i,\tilde{Z}_i,\tilde{S}_i}(x,y,z,l) = f_{X,Z|S_R}(x,z|l) \cdot f_{Y|Z}(y|z).$$

Thus, we have

$$f_{\tilde{X}_i,\tilde{Y}_i,\tilde{Z}_i,\tilde{S}_i}(x,y,z,l) = f_{X_i,Y_i,Z_i,S_i}(x,y,z,l) \iff (S_R, X) \text{ and } Y \text{ independent given } Z. \tag{19}$$

Under (19), the mean of a characteristic $h(X,Y,Z)$ can be estimated without bias from the after-fusion data by

$$\hat{\mathbb{E}}(h(X,Y,Z)) = \sum_{l=1}^{s_R} \frac{N_l^R}{N} \cdot \left(\frac{1}{n_l^R} \cdot \sum_{i \in I_l^R} h(\tilde{X}_i, \tilde{Y}_i, \tilde{Z}_i) \right). \tag{20}$$

The condition (19) is satisfied if X and Y are conditionally independent given Z and additionally one of the following cases applies:

(i) $S_R = 1$.

(ii) S_R is a function g of Z, i.e., $S_R = g(Z)$.

(iii) S_R is a function g of X, i.e., $S_R = g(X)$.

Case (i) means that we have ordinary SRSWR for the R sample. We are then in the situation of Section 2. In (ii), we have stratification by Z and (19) follows, because

$$f_{S_R,X|Z}(l,x|z) \cdot f_{Y|Z}(y|z) = f_{S_R,X,Y|Z}(l,x,y|z) \iff f_{X|Z}(x|z) \cdot f_{Y|Z}(y|z) = f_{X,Y|Z}(x,y|z)$$

is true for $l = g(z)$. In situation (iii), in which stratification by X is present, (19) can be shown similarly.

4.5 Stratified SRSWR for both samples

The density before the fusion is for $i \in I_l^R$

$$f_{X_i,Y_i,Z_i,S_i}(x,y,z,l) = f_{X,Y,Z|S_R}(x,y,z|l).$$

For the density after matching, we have for $i \in I_l^R$

$$f_{\tilde{X}_i,\tilde{Y}_i,\tilde{Z}_i,\tilde{S}_i}(x,y,z,l) = f_{X,Z|S_R}(x,z|l) \cdot f_{\tilde{Y}_i|Z_i}(y|z).$$

Regarding the second factor, let us first consider a D sample stratified by Y. Due to (12), we have

$$f_{\tilde{Y}_i|Z_i}(y|z) = \frac{n_l^D \cdot f_{Y,Z|S_D}(y,z|l)}{n_1^D \cdot f_{Z|S_D}(z|1) + ... + n_{s_D}^D \cdot f_{Z|S_D}(z|s_D)}.$$

For a special case with $f_{\tilde{Y}_i|Z_i}(y|z) = f_{Y|Z}(y|z)$, for instance, in (17), the distributions of (X_i, Y_i, Z_i, S_i) and $(\tilde{X}_i, \tilde{Y}_i, \tilde{Z}_i, \tilde{S}_i)$ are equal if and only if

$$(S_R, X) \text{ and } Y \text{ independent given } Z. \tag{21}$$

The requirement (21) is fulfilled if conditional independence of X and Y given Z and furthermore $S_R = 1$ or $S_R = g(Z)$ (i.e., R sample stratified by Z) or $S_R = g(X)$ (i.e., R sample stratified by X) hold, compare with Subsection 4.4.

Let us now study a D sample stratified by Z. Then, $f_{\tilde{Y}_i|Z_i}(y|z) = f_{Y|Z}(y|z)$ follows from Subsection 3.3.2 and $(\tilde{X}_i, \tilde{Y}_i, \tilde{Z}_i, \tilde{S}_i) \sim (X_i, Y_i, Z_i, S_i)$ is true if and only if (21) holds again.

4.6 SRSWOR (R sample) and inclusion probabilities proportional to $g(Z)$ (D sample)

We need notation and results from Subsections 3.2.3 and 3.3.4. We have $\pi_j^D = c \cdot g(z_j)$ with a constant c and $j = 1, ..., N$. That is, we have πps sampling if $g(Z)$ possesses an interpretation as "size". As a consequence, we have the simplification

$$f_{\tilde{Y}_i|Z_i}(y|z) = f_{Y|Z}(y|z)$$

and, because Π_R is constant,

$$\begin{aligned} & f_{X_i,Y_i,Z_i,\Pi_i}(x,y,z,\pi) = f_{\tilde{X}_i,\tilde{Y}_i,\tilde{Z}_i,\tilde{\Pi}_i}(x,y,z,\pi) \\ \Longleftrightarrow & f_{X,Y,Z,\Pi_R}(x,y,z,\pi) = f_{X,Z,\Pi_R}(x,z,\pi) \cdot f_{Y|Z}(y|z) \\ \Longleftrightarrow & f_{X,Y,Z}(x,y,z) = f_{X,Z}(x,z) \cdot f_{Y|Z}(y|z) \end{aligned}$$

The latter condition is equivalent to conditional independence of X and Y given Z. We remark that this subsection also covers the case of twice SRSWOR. In this case, we have $\pi_j^D = n_D/N = n_D/N \cdot g(z_j)$ for the constant function $g = 1$.

4.7 Inclusion probabilities proportional to $g(Z)$ for both samples

Here, we present approximate considerations. On the one hand, our without-replacement samples are described by $(U_1, ..., U_{n_R}, U_{n_R+1}, ..., U_n)$. On the other hand, we consider with-replacement samples described by $(U_1^*, ..., U_{n_R}^*, U_{n_R+1}^*, ..., U_n^*)$ where we have that $U_1^*, ..., U_{n_R}^*$ are independent, $U_{n_R+1}, ..., U_n$ are independent, and that

$$\begin{aligned} \mathbb{P}(U_i^* = j) &= \frac{\pi_j^R}{n_R} = const. \cdot g(z_j), \quad i = 1, ..., n_R \\ \mathbb{P}(U_i^* = j) &= \frac{\pi_j^D}{n_D} = const. \cdot g(z_j), \quad i = n_{R+1}, ..., n \end{aligned}$$

When n is small in comparison with N, the properties of the without-replacement situation will often differ not that much from the properties of the with-replacement setup. For the with-replacement constellation, we have seen in Subsection 4.1 that the distributions before and after fusion are equal if X and Y are independent given Z. Hence, there will often be no large difference between the distribution

before and after matching for the without-replacement situation if X and Y are independent given Z. Then, the Horvitz-Thompson estimator based on fusion data

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{n_R} \frac{h(\tilde{X}_i, \tilde{Y}_i, \tilde{Z}_i)}{\tilde{\Pi}_i} \tag{22}$$

will provide at least an approximately unbiased estimation of the mean of the characteristic $h(X, Y, Z)$ with a measurable function h.

5 Simulations

In Section 4, we have derived conditions under which the distribution before the fusion equals the distribution after the fusion. These results require the idealizing assumption (6). We will now consider a more realistic setup without assumption (6) and usually without exact matches. When we can not find an exact match, the nearest neighbor (NN) method is applied. That is, for every R sample unit, we search a D sample unit whose Z value is closest to the Z value of the considered individual from the R sample. If more than one D sample unit possesses the minimum distance in Z to the considered R sample unit, one of these D sample units is selected randomly. We throughout allow that each D sample unit can be used any number of times as donor, i.e., we permit "polygamy". Moreover, we do not introduce a maximum allowed distance in Z of matched units, that is, we have a caliper equal to infinity.

The aim of this section, is to investigate whether the results from Section 4 hold also in the more realistic situation. For this, we consider various concrete specifications for the draw of R and D sample where we orientate ourselves towards Section 4 and conduct wide simulations using the software "R" (Version 3.1.3). The sample size of the R sample is always $n_R = 500$ while the D sample comprises always $n_D = 1000$ units.

Throughout this section, we consider two populations G_1 and G_2 each of size $N = 10^6$. In G_1, the vector (X, Y, Z) is approximately normally distributed with

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \sim N\left(\begin{pmatrix} 30 \\ 40 \\ 50 \end{pmatrix}, \begin{pmatrix} 25 & -3.2 & -10 \\ -3.2 & 64 & 32 \\ -10 & 32 & 100 \end{pmatrix} \right). \tag{23}$$

In G_2, we approximately have

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \sim N\left(\begin{pmatrix} 30 \\ 40 \\ 50 \end{pmatrix}, \begin{pmatrix} 25 & 30 & -10 \\ 30 & 64 & 32 \\ -10 & 32 & 100 \end{pmatrix} \right). \tag{24}$$

For instance, X, Y, and Z could represent the daily duration of watching a certain television program between 6 pm and 8 pm, the monthly expenditures for a certain product, and the person's age. For the normal distribution in (23), we have conditional independence of X and Y given Z, because

$$Cov(X, Y) = Cov(X, Z) \cdot Cov(Y, Z)/Cov(Z, Z)$$

is true. In (24), such conditional independence does not hold. For our simulations, G_1 and G_2 are obtained by drawing a sample of size N from the corresponding normal distribution. Since negative outcomes of X, Y, Z may lead to negative single stage selection probabilities or inclusion probabilities, such numbers are converted to the absolute value. This is only a tiny modification, because the chance of obtaining a negative value for above normal distributions is very small.

By combining a population (G_1 or G_2) with a scheme for generating the R sample and a sampling procedure for the D sample, we obtain several constellations. For each constellation, we conduct 1000 simulation replications where each replication proceeds as follows. First, R and D sample are generated, second, the matched data set is calculated, and third, the matched data set is evaluated by computing estimates for certain means on the population.

The way how the estimates are calculated depends of the sampling scheme of the R sample. Concerning this, let us consider the characteristic $h(X, Y, Z)$ with a measurable function h. To estimate the mean of $h(X, Y, Z)$, we apply the Hansen-Hurwitz estimator (15) for R samples drawn by GWR sampling. For a stratified R sample, it is appropriate to use (20). For without-replacement sampling for the R sample, we apply the Horvitz-Thompson estimator (22). The estimators (15), (20), and (22) reduce to the ordinary mean if we have SRSWR, only one stratum, and SRSWOR, respectively. When 1000 replications were conducted, we have 1000 realizations of each considered estimator, from which we compute the average estimate and make a comparison with the true mean on the population.

5.1 GWR sampling for both samples

In this subsection, we investigate if the results from Subsection 4.1 remain valid in our current more realistic situation without assumption (6) and with NN matching rather than exact matching. R and D sample are drawn by GWR sampling with single draw selection probabilities

$$p_j = p_j^R = p_j^D \quad (j = 1, ..., N).$$

We consider single draw selection probabilities proportional to X, Y, or Z. That is, we have $p_j = x_j / \sum_{l=1}^N x_l$ or $p_j = y_j / \sum_{l=1}^N y_l$ or $p_j = z_j / \sum_{l=1}^N z_l$. By combining two populations (G_1 or G_2) with three sampling strategies (probabilities proportional to X, Y or Z for both samples), we obtain six constellations.

The simulation results are given in Table A.1. Let us consider the upper part of this table, which corresponds to G_1, i.e., the population for which conditional independence holds. We first list the theoretical values which hold for an exact normal distribution and the actual values for G_1. Then, we provide average estimates obtained by our simulations. The lower part of Table A.1 is analogously arranged where population G_2 is considered there. Let us have a look at single stage selection probabilities proportional to Z first (see last column in the table). Here, the average estimates are close to the actual values for G_1. This coincides with Subsection 4.1. For G_2, the estimation of $\mathbb{E}(XY)$ fails. Selection probabilities proportional to Y or X were not treated in Subsection 4.1. In our simulations, we observe often biased estimates for the case with Y. In the case with X and the universe G_1, the average estimates seem to be as good as those for selection probabilities proportional to Z and G_1.

5.2 SRSWR (R sample) and stratification by Z (D sample)

With respect to the D sample, the population is divided into two strata according to the Z values. Units with $Z \leq 45$ belong to stratum 1, other units are assigned to stratum 2. We draw 400 and 600 units from stratum 1 and 2, respectively. Table A.2 shows the simulation results. For universe G_1, that is, for the case with conditional independence, the population means are well reproduced. This confirms the result of Subsection 4.2. For G_2, the estimation of $\mathbb{E}(XY)$ fails. In other words, the conditional independence can not be relinquished.

5.3 SRSWR (R sample) and stratification by Y (D sample)

Now the D sample is stratified by Y. Units with $Y \leq 40$ belong to stratum 1, while $Y > 40$ holds for the second stratum. In a first case, we draw 500 units from each stratum. In a second case, we select

750 and 250 units from stratum 1 and 2, respectively. The results are provided in Table A.3. At least the estimate for $\mathbb{E}(XY)$ is poor for G_2. For G_1 and the unequal sample sizes for the D subsamples, the estimations often involve a massive bias. For G_1 and the equal sample sizes $n_1^D = n_2^D = 500$, the considered population means are estimated well. This observation coincides with Subsection 4.3.

5.4 Stratified SRSWR (*R* sample) and SRSWR (*D* sample)

Here, we combine a stratified R sample with SRSWR for the D sample. The R sample consists of a sample of size $n_1^R = 200$ from stratum 1 and a sample of size $n_2^R = 300$ from stratum 2. In a first case, units with $X \leq 28$ belong to stratum 1 and units with $X > 28$ are assigned to stratum 2 (stratification by X). In a second case, we have $Z \leq 45$ for the first stratum and $Z > 45$ for the second stratum (stratification by Z). The output of the simulations can be found in Table A.4. For the conditional independence case (that is, for G_1), the average estimates are throughout close to the true values. This could be expected after Subsection 4.4. For the other population G_2, the average estimates for $\mathbb{E}(XY)$ are not close to $\mathbb{E}(XY)$.

5.5 Stratified SRSWR for both samples

In this subsection, the R sample is stratified by X or Z and the D sample is stratified by Y or Z. The subsample sizes are $n_1^R = 200$ and $n_2^R = 300$ as well as $n_1^D = n_2^D = 500$. For an R sample stratified by X, we have $X \leq 28$ for stratum 1. For an R sample stratified by Z, $Z \leq 45$ holds in stratum 1. For a D sample stratified by Y, $Y \leq 40$ is true in stratum 1. For a D sample stratified by Z, we have $Z \leq 52$ in the first stratum. According to Table A.5, we have expedient estimations for population G_1. Thus, we find the results from Subsection 4.5 here again. For the other universe, i.e., for G_2, the estimation of $\mathbb{E}(XY)$ does not work.

5.6 SRSWOR (*R* sample) and Sampford sampling (*D* sample)

We have without-replacement samples in this subsection. The R sample is generated by SRSWOR. The D sample is drawn by Sampford sampling with inclusion probabilities proportional to X, Y, or Z. That is, we have for $j = 1, ..., N$, the inclusion chances

$$\pi_j^D = n_D \cdot x_j \,/\, \sum_{l=1}^{N} x_l \quad \text{or} \tag{25}$$

$$\pi_j^D = n_D \cdot y_j \,/\, \sum_{l=1}^{N} y_l \quad \text{or} \tag{26}$$

$$\pi_j^D = n_D \cdot z_j \,/\, \sum_{l=1}^{N} z_l. \tag{27}$$

The simulation results are listed in Table A.6. For G_1 and inclusion probabilities for the D sample proportional to Z, the estimators work confirming the corresponding statement from Subsection 4.6.

5.7 Sampford sampling for both samples

We now analyze Sampford sampling for both samples and consider inclusion probabilities proportional to X, Y, or Z. Hence, we have (25), (26), or (27) and analog equations for π_j^R. Table A.7 contains the simulation output for this subsection. The observations are similar to those of Subsection 5.1, in which with-replacement samples were addressed. Especially for G_1 and inclusion probabilities for both samples proportional to Z, we have good estimates on average.

6 Summary and conclusions

In this article, we addressed statistical matching for several combinations of sampling strategies for the recipient sample and the donor sample. As sampling designs, stratified sampling, general with-replacement sampling and general without-replacement sampling were incorporated. In a first step, we have conducted analytical investigations for each considered pair of sampling schemes. In particular, we first derived the distribution after fusion. Second, we identified situations in which the distribution after fusion equals the distribution before matching. Third, we stated an estimator that is appropriate when the after-fusion and before-fusion distributions are equal.

To obtain a convenient expression for the second factor of the after-matching density, we have made an idealizing assumption. This assumption was that the distribution of the vector (Y, Z) in the D sample equals its expected distribution. Consequently, our analytic findings on after-fusion distributions (Section 3) and on criteria for correct after-fusion distributions (Section 4) depend on this requirement. Mathematically, the idealizing assumption can be written as an event $\mathcal{B}$ and in Section 3, we have actually derived the conditional density $f_{\tilde{X}_i,\tilde{Y}_i,\tilde{Z}_i,\tilde{T}_i}(x, y, z, t|\mathcal{B})$. The hope is then that

$$f_{\tilde{X}_i,\tilde{Y}_i,\tilde{Z}_i,\tilde{T}_i}(x, y, z, t) \approx f_{\tilde{X}_i,\tilde{Y}_i,\tilde{Z}_i,\tilde{T}_i}(x, y, z, t|\mathcal{B}),$$

i.e., that conditioning on $\mathcal{B}$ does not matter that much. In this case, the criteria from Section 4 would be robust. To explore this robustness, we have conducted simulations. Joyfully, these simulations indicate that the analytic results on correct after-fusion distributions from Section 4, which were obtained under the idealizing assumption, remain valid also without this assumption.

From our article, we can conclude the following points regarding the quality of fusion data. Conditional independence of the characteristics not jointly observed given the common characteristic is a central necessity for a successful fusion. Here, a successful fusion means that the distribution after matching equals the distribution before matching. However, this conditional independence is often not sufficient. Instead, stronger or additional requirements are needed (e.g., compare with (14), (17), (19), (21)).

Another main observation is that identical sampling schemes for both original samples are not necessary to obtain a successful fusion. For example, several combinations of SRSWR and stratification or a combination of SRSWOR and with-replacement sampling with inclusion probabilities proportional to some function of the common characteristic work as long as we have conditional independence.

This article does not intend to unrestrictedly advertise data fusions. Instead, the aim is to sensitize data fusion users that successful fusions depend on certain criteria, typically involving conditional independence and, for more complex sampling, partly additional or stronger requirements according to the sampling schemes of the input samples. Already the conditional independence of X and Y given Z is a quite strong requirement, because this means that there is no influence from Y to X (and also no influence from X to Y) when Z is fixed. For example, in the context of television and buying behavior, this means that we cannot improve the prediction of buying behavior by data on television watching when we have already conditioned on the common characteristics. As stated, for example, in Rässler (2002, p. 35), this situation will usually not hold if the common characteristics comprise only demographic and socio-economic variables. A possible remedy stated in Rässler (2002, p. 35) to mitigate this problem is that the common characteristics Z should also incorporate some variables on television and buying behavior.

Appendix: Simulation outputs

	mean of	theoretical	actual	average estimates $\propto X$	$\propto Y$	$\propto Z$
G_1	X	30.00	29.99	29.99	29.98	29.99
	Y	40.00	40.00	39.98	41.37	40.02
	Z	50.00	50.00	49.96	50.02	50.00
	X^2	925.00	924.52	924.96	924.00	924.27
	Y^2	1664.00	1663.99	1663.28	1773.15	1665.55
	Z^2	2600.00	2600.36	2597.60	2602.22	2600.51
	XY	1196.80	1196.54	1196.68	1237.54	1197.17
	XZ	1490.00	1489.79	1489.51	1490.09	1489.79
	YZ	2032.00	2032.18	2030.59	2100.48	2033.07
G_2	X	30	30.01	30.01	30.02	30.01
	Y	40.00	40.01	41.11	41.34	39.99
	Z	50.00	50.01	50.02	49.99	50.01
	X^2	925.00	925.42	925.24	925.86	925.51
	Y^2	1664.00	1665.23	1752.70	1769.94	1663.49
	Z^2	2600.00	2600.63	2601.88	2598.63	2601.30
	XY	1230.00	1230.75	1230.20	1237.62	1197.04
	XZ	1490.00	1490.64	1490.80	1490.46	1490.79
	YZ	2032.00	2033.03	2088.31	2097.06	2032.23

Table A.1: Simulated average estimates for GWR sampling for both samples, see Subsection 5.1. The column "theoretical" contains theoretical population means which occur for an exact normal distribution. The column "actual" provides the actual means on the finite population (G_1 or G_2). For example, $\propto X$ means single stage selection probabilities proportional to X.

	population G_1			population G_2		
mean of	theoretical	actual	average estimates	theoretical	actual	average estimates
X	30.00	29.99	29.99	30.00	30.01	30.01
Y	40.00	40.00	40.00	40.00	40.01	40.00
Z	50.00	50.00	50.01	50.00	50.01	50.01
X^2	925.00	924.52	924.18	925.00	925.42	925.35
Y^2	1664.00	1663.99	1663.72	1664.00	1665.23	1663.70
Z^2	2600.00	2600.36	2601.58	2600.00	2600.63	2600.96
XY	1196.80	1196.54	1196.19	1230.00	1230.75	1197.04
XZ	1490.00	1489.79	1489.77	1490.00	1490.64	1490.62
YZ	2032.00	2032.18	2032.32	2032.00	2033.03	2032.06

Table A.2: Simulated average estimates for SRSWR for the R sample and a D sample stratified by Z in comparison with the actual values, see Subsection 5.2.

	population G_1				population G_2			
mean			average estimates				average estimates	
of	theoretical	actual	$n_1^D = n_2^D$	$n_1^D \neq n_2^D$	theoretical	actual	$n_1^D = n_2^D$	$n_1^D \neq n_2^D$
X	30.00	29.99	30.01	29.99	30.00	30.01	30.01	30.01
Y	40.00	40.00	39.98	37.26	40.00	40.01	40.01	37.24
Z	50.00	50.00	49.98	49.99	50.00	50.01	50.02	49.98
X^2	925.00	924.52	925.36	924.31	925.00	925.42	925.62	925.77
Y^2	1664.00	1663.99	1661.97	1443.38	1664.00	1665.23	1664.76	1441.80
Z^2	2600.00	2600.36	2598.41	2599.09	2600.00	2600.63	2601.87	2598.27
XY	1196.80	1196.54	1196.46	1114.74	1230.00	1230.75	1197.71	1114.81
XZ	1490.00	1489.79	1489.88	1489.43	1490.00	1490.64	1491.13	1490.10
YZ	2032.00	2032.18	2030.13	1890.61	2032.00	2033.03	2033.05	1889.08

Table A.3: Simulated average estimates for SRSWR for the R sample and a D sample stratified by Y in comparison with the actual values, see Subsection 5.3. The size of the sample from stratum i is n_i^D.

	population G_1				population G_2			
mean			average estimates				average estimates	
of	theoretical	actual	X-strat	Z-strat	theoretical	actual	X-strat	Z-strat
X	30.00	29.99	30.00	29.99	30.00	30.01	30.00	30.01
Y	40.00	40.00	39.99	40.02	40.00	40.01	40.03	40.04
Z	50.00	50.00	50.01	50.03	50.00	50.01	50.01	50.00
X^2	925.00	924.52	924.92	924.05	925.00	925.42	925.13	925.42
Y^2	1664.00	1663.99	1663.07	1664.34	1664.00	1665.23	1666.11	1667.17
Z^2	2600.00	2600.36	2600.89	2603.09	2600.00	2600.63	2601.40	2599.57
XY	1196.80	1196.54	1196.57	1196.37	1230.00	1230.75	1197.83	1198.07
XZ	1490.00	1489.79	1490.23	1490.29	1490.00	1490.64	1490.68	1490.21
YZ	2032.00	2032.18	2031.86	2033.46	2032.00	2033.03	2033.96	2033.86

Table A.4: Simulated average estimates for an R sample stratified by X or Z and SRSWR for the D sample, see Subsection 5.4. E.g., X-strat means an R sample stratified by$X$.

	mean of	theoretical	actual	average estimates X/Y-strat	Z/Y-strat	X/Z-strat	Z/Z-strat
G_1	X	30.00	29.99	29.99	29.99	29.99	30.01
	Y	40.00	40.00	40.00	39.99	40.00	40.01
	Z	50.00	50.00	50.01	50.01	50.00	49.99
	X^2	925.00	924.52	924.31	924.29	924.53	925.56
	Y^2	1664.00	1663.99	1663.60	1662.86	1664.14	1664.15
	Z^2	2600.00	2600.36	2600.91	2601.39	2600.01	2599.48
	XY	1196.80	1196.54	1196.12	1196.09	1196.65	1197.48
	XZ	1490.00	1489.79	1489.76	1489.86	1489.74	1490.36
	YZ	2032.00	2032.18	2032.05	2031.87	2032.16	2031.92
G_2	X	30.00	30.01	30.00	30.00	30.01	30.00
	Y	40.00	40.01	40.01	40.01	40.00	40.04
	Z	50.00	50.01	50.01	50.01	49.99	50.02
	X^2	925.00	925.42	925.06	925.01	925.73	924.89
	Y^2	1664.00	1665.23	1664.49	1665.10	1663.74	1666.96
	Z^2	2600.00	2600.63	2600.77	2600.66	2599.46	2602.05
	XY	1230.00	1230.75	1197.11	1197.15	1197.21	1197.80
	XZ	1490.00	1490.64	1490.40	1490.30	1490.38	1490.42
	YZ	2032.00	2033.03	2032.69	2032.85	2031.52	2034.44

Table A.5: Simulated average estimates for stratified sampling in both samples, see Subsection 5.5. E.g., X/Y-strat means that the R sample is stratified by X and the D sample is stratified by Y.

	mean of	theoretical	actual	average estimates $\propto X$	$\propto Y$	$\propto Z$
G_1	X	30.00	29.99	29.99	29.99	29.99
	Y	40.00	40.00	40.02	41.35	40.00
	Z	50.00	50.00	50.01	50.00	50.01
	X^2	925.00	924.52	924.49	924.54	924.48
	Y^2	1664.00	1663.99	1665.11	1771.10	1664.29
	Z^2	2600.00	2600.36	2601.35	2600.60	2601.52
	XY	1196.80	1196.54	1196.96	1237.15	1196.62
	XZ	1490.00	1489.79	1490.12	1489.88	1490.14
	YZ	2032.00	2032.18	2033.43	2098.56	2032.71
G_2	X	30.00	30.01	30.02	30.01	30.00
	Y	40.00	40.01	41.11	41.37	40.03
	Z	50.00	50.01	50.01	50.01	49.99
	X^2	925.00	925.42	926.03	925.42	925.18
	Y^2	1664.00	1665.23	1753.13	1772.77	1666.18
	Z^2	2600.00	2600.63	2601.20	2600.28	2598.89
	XY	1230.00	1230.75	1230.85	1238.17	1197.83
	XZ	1490.00	1490.64	1491.24	1490.35	1489.91
	YZ	2032.00	2033.03	2088.30	2099.64	2032.87

Table A.6: Simulated average estimates for SRSWOR (R sample) and Sampford sampling (D sample) see Subsection 5.6. E.g., $\propto X$ means that we have inclusion probabilities proportional to X for the D sample.

	mean of	theoretical	actual	average estimates $\propto X$	$\propto Y$	$\propto Z$
G_1	X	30.00	29.99	29.99	30.00	29.98
	Y	40.00	40.00	39.95	41.39	39.98
	Z	50.00	50.00	49.93	50.00	50.00
	X^2	925.00	924.52	925.31	924.64	924.18
	Y^2	1664.00	1663.99	1661.10	1773.76	1663.40
	Z^2	2600.00	2600.36	2595.74	2599.33	2601.50
	XY	1196.80	1196.54	1196.18	1238.18	1196.17
	XZ	1490.00	1489.79	1488.98	1489.51	1489.96
	YZ	2032.00	2032.18	2028.69	2099.51	2032.14
G_2	X	30.00	30.01	30.01	30.00	30.02
	Y	40.00	40.01	41.13	41.38	40.02
	Z	50.00	50.01	49.97	50.00	50.01
	X^2	925.00	925.42	925.57	925.34	925.86
	Y^2	1664.00	1665.23	1754.81	1773.41	1665.67
	Z^2	2600.00	2600.63	2598.12	2600.63	2600.23
	XY	1230.00	1230.75	1231.29	1238.66	1198.05
	XZ	1490.00	1490.64	1490.11	1490.69	1490.89
	YZ	2032.00	2033.03	2088.09	2099.86	2033.07

Table A.7: Simulated average estimates for Sampford sampling for both samples, see Subsection 5.7. E.g., $\propto X$ means that we have inclusion probabilities proportional to X for both samples.

References

[1] Conti P.L. / Marella D. / Scanu M. (2008): Evaluation of matching noise for imputation techniques based on nonparametric local linear regression estimators. Computational Statistics and Data Analysis 53, 354-365.

[2] Conti P.L. / Marella D. / Scanu M. (2013): Uncertainty analysis for statistical matching of ordered categorical variables. Computational Statistics and Data Analysis 68, 311-325.

[3] D'Orazio M. / Di Zio M. / Scanu M. (2006): Statistical Matching, Theory and Practice. Wiley.

[4] D'Orazio M. / Di Zio M. / Scanu M. (2012): Statistical Matching of Data from Complex Sample Surveys. Proceedings of the European Conference on Quality in Official Statistics - Q2012.

[5] Gilula Z. / McCulloch R.E. / Rossi P.E. (2006): A Direct Approach to Data Fusion. Journal of Marketing Research 43, 73-83.

[6] Goel P.K. / Ramalingam T. (1989): The Matching Methodology: Some Statistical Properties. Springer.

[7] Kadane J.B. (1978): Some Statistical Problems in Merging Data Files. 1978 Compendium of Tax Research, U.S. Department of the Treasury, 159-171. (Reprinted in Journal of Official Statistics 17, 423-433).

[8] Kovacevic M.S. / Liu T.-P. (1994): Statistical Matching of Survey Datafiles: A Simulation Study. Proceedings of the Section on Survey Research Methods, American Statistical Association, 479-484.

[9] Moriarity C. / Scheuren F. (2001): Statistical Matching: A Paradigm for Assessing the Uncertainty in the Procedure. Journal of Official Statistics 17, 407-422.

[10] Moriarity C. / Scheuren F. (2003): A Note on Rubin's Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations. Journal of Business & Economic Statistics 21, 65-73.

[11] Reiter J.P. (2012): Bayesian Finite Population Imputation for Data Fusion. Statistica Sinica 22, 795-811.

[12] Rässler S. / Fleischer K. (1998): Aspects Concerning Data Fusion Techniques. ZUMA Nachrichten Spezial 4, 317-333.

[13] Rässler S. (2002): Statistical Matching - A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches. Springer.

[14] Rässler S. (2004): Data Fusion: Identification Problems, Validity, and Multiple Imputation. Austrian Journal of Statistics 33, 153-171.

[15] Rodgers W.L. (1984): An Evaluation of Statistical Matching. Journal of Business & Economic Statistics 2, 91-102.

[16] Rubin D.B. (1986): Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations. Journal of Business & Economic Statistics 4, 87-94.

[17] Sims C.A. (1972): Comment on "Constructing a New Data Base from Existing Microdata Sets: The 1966 Merge File". Annals of Economic and Social Measurement 1, 343-345.

[18] Stuart E.A. (2010): Matching Methods for Causal Inference: A Review and a Look Forward. Statistical Science 25, 1-21.