An Efficient Approach for Statistical Matching of Survey Data Through Calibration, Optimal Transport and Balanced Sampling

Raphaël Jauslin\textsuperscript{a} and Yves Tillé\textsuperscript{a}

Abstract

Statistical matching aims to integrate two statistical sources. These sources can be two samples or a sample and the entire population. If two samples have been selected from the same population and information has been collected on different variables of interest, then it is interesting to match the two surveys to analyse, for example, contingency tables or covariances. In this paper, we propose an efficient method for matching two samples that may each contain a weighting scheme. The method matches the records of the two sources. Several variants are proposed in order to create a directly usable file integrating data from both information sources.

Keywords: auxiliary information, balanced sampling, data integration, distance, unequal probability sampling

\textsuperscript{a}Institute of statistics, University of Neuchatel, Av. de Bellevaux 51, 2000 Neuchatel, Switzerland

(E-mail: raphael.jauslin@unine.ch)
1 Introduction

Integrating data from different sources represents a major challenge in statistics. Yang and Kim (2020) and Kim and Tam (2020) discussed a set of methods for data integration. Statistical matching is the field of statistics that deals with the best way to merge two different files by matching units based on a group of common variables (D'Orazio et al., 2006; D'Orazio, 2019).

Renssen (1998) distinguishes between two types of analysis. The macro approach focuses on estimating a full joint distribution between the respective variable of interest in the two samples. This can be, for example, a covariance matrix, a correlation or a contingency table. The objective of the micro approach intends to complement one file with information from the other, imputation to correct non-response is an example related to this approach (Haziza, 2009; Chen and Haziza, 2019). Kim et al. (2016) use the technique of fractional imputation to perform statistical matching.

In this paper, we propose an efficient method for statistical matching. Units are matched based on the proximity of a group of variables measured in both surveys. Moreover, both sources can either have common units or have an empty intersection. One of the two sources may even contain the entire population. In addition, we impose a set of constraints in order to take advantage of all the information available in the two sources. This method can also be used to realize imputations and so can also be used to micro approach analyses.

Both sources of information may contain a weighting system that allows the files to be extrapolated to the full population. These weights are usually calculated to take into account inclusion probabilities, non-response treatment and calibration. In official statistics, the calibration methods have been proposed by Deville and Särndal (1992) and Deville et al. (1993) to adjust survey data on census data or a register. Calibration can also be used to adjust or harmonize several surveys from different sources (see Guandalini and Tillé, 2017, and references therein). Dudoignon (2018) through an unpublished paper, proposed to use the idea of optimal transport on internet audience data.

We have set out a series of recommendations that a matching method should follow: The method should match common units as a priority. The result of the matching must integrate information from both sources. The matching should also take into account the weighting system. After matching, the estimated totals of the variables common to both sources must be identical to the totals before matching. An optimal matching should take advantage of all the information available in both sources.

The proposed methods therefore allow the matching of two data files but also the imputation of one file on another. First, calibration theory is used to harmonise the two samples. Then a linear program is used to perform an optimal matching while taking into account the weights. This program can be written as an optimal transport problem. Finally, the values to be matched can be selected using a balanced stratified sampling technique as presented in Jauslin et al. (2021b) and implemented in the R package ‘StratifiedSampling’ (Jauslin et al., 2021a). The methods either perform matching based on a division of weights, produce a prediction, or impute a value from one source to another.
2 Problem and notation

Consider a population $U = \{1, \ldots, k, \ldots, N\}$ of size $N$ from which two random samples $S_1$ and $S_2$ of size respectively equals to $n_1$ and $n_2$ have been selected. It is assumed that three groups of variables can be measured on the population units. The vectors of variables $x_k \in \mathbb{R}^p, k \in U$ are measured on both units selected in $S_1$ and $S_2$. The vectors of variables $y_k \in \mathbb{R}^q, k \in U$ are measured only on the units selected in $S_1$. The vectors of variables $z_k \in \mathbb{R}^r, k \in U$ are measured only on the units selected in $S_2$.

Generally in survey sampling, samples are designed with complex weighting systems $v_1k, k \in S_1$ and $v_2\ell, \ell \in S_2$. These weights can take into account the inverses of the inclusion probabilities, a possible re-weighting to compensate questionnaire non-response and a possible calibration. Frequently, a statistical match is made when one of the two file is seen as recipient file while the other one is seen as donor file. Throughout this manuscript, we will suppose, without loss of generality, that the sample $S_1$ is the recipient file while $S_2$ is the donor file. Generally we have $n_2 > n_1$, note that one of the two sample might be the whole population.

The population totals on the common auxiliary variables are equal to:

$$X = \sum_{k \in U} x_k.$$  

It can be estimated either by the sample $S_1$ or the sample $S_2$, which are mathematically written:

$$\hat{X}_{v_1} = \sum_{k \in S_1} v_{1k}x_k, \quad \hat{X}_{v_2} = \sum_{\ell \in S_2} v_{2\ell}x_\ell.$$  

(1)

Following the same idea on the variables of interest, the totals

$$Y = \sum_{k \in U} y_k \text{ and } Z = \sum_{k \in U} z_k,$$

can be estimated using the following estimators:

$$\hat{Y}_{v_1} = \sum_{k \in S_1} v_{1k}y_k, \quad \hat{Z}_{v_2} = \sum_{\ell \in S_2} v_{2\ell}z_\ell.$$ 

In the micro approach, the two samples $S_1$ and $S_2$ are merged into a single usable file, while, the macro approach focuses on the joint distribution of the variables of interest. Under the usual hypothesis that, conditionally to the variables $x_k$, the variables $y_k$ and $z_k$ are independent, the relationships between the variables $y_k$ and $z_k$ can be analyzed. For example, if the variables $y_k$ and $z_k$ are dummy variables with respectively $q$ and $r$ variables, we could be interested in the estimation of the contingency table

$$N_{yz} = \sum_{k \in U} y_kz_k^\top.$$ 

If the variables of interest are continuous, we could be interested in computing the covariance matrix of the totals

$$\Sigma_{yz} = \text{cov}(Y, Z).$$
3 Harmonization by calibration

Since we are working with two different samples, we firstly harmonize the sampling weights \( v_{1k}, k \in S_1 \) and \( v_{2\ell}, \ell \in S_2 \) in order to have same totals given in Equations (1). That means we are looking for a new weighting system

\[
\begin{align*}
  w_{1k}, k \in S_1 \quad \text{and} \quad w_{2\ell}, \ell \in S_2,
\end{align*}
\]

such that we have the following results:

\[
\begin{align*}
  \hat{\mathbf{X}}_{\omega 1} = \sum_{k \in S_1} w_{1k} \mathbf{x}_k = \hat{\mathbf{X}}_{\omega 2} = \sum_{\ell \in S_2} w_{2\ell} \mathbf{x}_\ell.
\end{align*}
\]

One aspect that must be taken into account is the intersection between \( S_1 \) and \( S_2 \). Several cases can occur, the sample \( S_1 \) can be included in \( S_2 \) or vice versa, the intersection can also be empty. Let \( n_{12} = \#(S_1 \cap S_2) \) denote the size of the intersection of the two samples. Guandalini and Tillé (2017) analyse estimators of the form \( \hat{\mathbf{X}}_\alpha = \alpha \hat{\mathbf{X}}_{\omega 1} + (1 - \alpha)\hat{\mathbf{X}}_{\omega 2} \). They showed that, when \( p = 1 \), to best estimate \( \mathbf{X} \) using both \( S_1 \) and \( S_2 \), the value of \( \alpha \) must be equal to

\[
\alpha_{\text{opt}} = \frac{\text{var}(\hat{\mathbf{X}}_{\omega 2}) - \text{cov}(\hat{\mathbf{X}}_{\omega 1}, \hat{\mathbf{X}}_{\omega 2})}{\text{var}(\hat{\mathbf{X}}_{\omega 1}) + \text{var}(\hat{\mathbf{X}}_{\omega 2}) - \text{cov}(\hat{\mathbf{X}}_{\omega 1}, \hat{\mathbf{X}}_{\omega 2})}.
\]

This optimal value minimizes the variance of \( \hat{\mathbf{X}}_\alpha \). However, it depends on unknown variances and on a covariance that must be estimated. Since variance estimators are particularly unstable, we may find ourselves far from the optimal estimator. Guandalini and Tillé (2017) suggest to use a proxy value for \( \alpha_{\text{opt}} \) that only depends on the sample sizes and the size of the overlapped sample:

\[
\alpha^* = \frac{n_1 - n_{12}}{n_1 + n_2 - 2 n_{12}}.
\]

One can then construct the estimator \( \hat{\mathbf{X}}^* = \alpha^* \hat{\mathbf{X}}_{\omega 1} + (1 - \alpha^*)\hat{\mathbf{X}}_{\omega 2} \). In particular, if \( S_2 \subset S_1 \), then \( \alpha^* = 1 \) and \( \hat{\mathbf{X}}^* = \hat{\mathbf{X}}_{\omega 1} \). Moreover, if \( S_1 \cap S_2 = \emptyset \), then \( \alpha^* = n_1 / (n_1 + n_2) \).

In order to compute the two new weighting systems (2) close to \( v_{1k}, k \in S_1 \) and \( v_{2\ell}, \ell \in S_2 \), the two samples are calibrated \( \hat{\mathbf{X}}^* \). If \( G_k(w_{1k}, v_{1k}) \) is one of the pseudo-distance defined in Deville and Särndal (1992), one can search the weighting systems that solve the following problem:

\[
\begin{align*}
\left\{ \begin{array}{ll}
\text{minimize} & \sum_{k \in S_1} G_k(w_{1k}, v_{1k}) + \sum_{\ell \in S_2} G_\ell(w_{2\ell}, v_{2\ell}) \\
\text{subject to} & \sum_{k \in S_1} w_{1k} \mathbf{x}_k = \sum_{\ell \in S_2} w_{2\ell} \mathbf{x}_\ell = \hat{\mathbf{X}}^*, \\
& \hat{\mathbf{N}}^* = \sum_{k \in S_1} w_{1k} = \sum_{\ell \in S_2} w_{2\ell} = \alpha^* \sum_{k \in S_1} v_{1k} + (1 - \alpha^*) \sum_{\ell \in S_2} v_{2\ell}.
\end{array} \right.
\end{align*}
\]

The calibration problem must ensure that the new weights remain positive. This can be obtained, for example, by taking as pseudo-distance the divergence of Kullback-Leibler, i.e. \( G_k(w_{1k}, v_{1k}) = w_{1k} \log w_{1k}/v_{1k} \). Thus, the new weights obtained have the same sum:

\[
\sum_{k \in S_1} w_{1k} = \sum_{\ell \in S_2} w_{2\ell}.
\]
They also allow us to define new and more coherent estimators for $X$, $Y$ and $Z$

\[
\hat{X}_{w1} = \sum_{k \in S_1} w_{1k} x_k = \hat{X}_{w2} = \sum_{\ell \in S_2} w_{2\ell} x_{\ell},
\]

\[
\hat{Y}_{w1} = \sum_{k \in S_1} w_{1k} y_k \text{ and } \hat{Z}_{w2} = \sum_{\ell \in S_2} w_{2\ell} z_{\ell}.
\]

Figure 1: Representation of the statistical matching with intersection. Hatched area are unknown quantities in each samples.

4  Renssen’s methods

A method for estimating contingency table is developed in Renssen (1998) and more recently presented in D’Orazio et al. (2006). The general idea consists of harmonizing the weighting systems as explained in the Section 3 and then uses the matching variables $x_k$ to create linear models to get an estimated contingency table. At the first step, regression coefficients $\beta_{yx}$ and $\beta_{zx}$ are computed from the samples $S_1$ (respectively $S_2$) by using the weights, $w_{1k}$, $k \in S_1$ (respectively $w_{2\ell}$, $\ell \in S_2$). Using a weighted linear model, the following coefficients are obtained:

\[
\hat{\beta}_{yx} = \left( \sum_{k \in S_1} w_{1k} y_k x_k^{\top} \right) \left( \sum_{k \in S_1} w_{1k} x_k x_k^{\top} \right)^{-1},
\]

\[
\hat{\beta}_{zx} = \left( \sum_{\ell \in S_2} w_{2\ell} z_{\ell} x_{\ell}^{\top} \right) \left( \sum_{\ell \in S_2} w_{2\ell} x_{\ell} x_{\ell}^{\top} \right)^{-1}.
\]
The contingency table is then estimated using the matrix product:

\[ \hat{N}_{yz} = \hat{\beta}_{yz} \left( \alpha^* \sum_{k \in S_1} w_{1k} x_k x_k^\top + (1 - \alpha^*) \sum_{\ell \in S_2} w_{2\ell} x_\ell x_\ell^\top \right) \hat{\beta}_{zx}^\top, \]

where \( \alpha^* \) is the coefficient (3) that depends on the value \( n_{12} \). Renssen’s method can be easily generalized to continuous case, but some assumptions must be satisfied on variables \( y_k, k \in S_1 \) and \( z_\ell, \ell \in S_2 \). For more information, we refer the reader to the article of Renssen (1998) and the book written by D’Orazio et al. (2006).

5 Matching by optimal transport

The main idea of our method uses the optimal transport to perform a statistical matching. Optimal transport is an old mathematical problem that consists of finding the best solution to minimize the cost of transporting some quantities of goods from a given set of locations to a given set of destinations. In its simple case, the optimal transport problem can be solved with a linear program. However, it has been a very fruitful topic in statistics for the past 10 years and it is strongly related to the notion of Wasserstein distance. We refer the reader to Panaretos and Zemel (2020) for more information on optimal transport and the Wasserstein distance.

In the particular case of statistical matching, we use the optimal transport problem to match the units of two different samples. By reminding that \( S_1 \) denotes the recipient sample while \( S_2 \) stands for the donor sample, the idea is then to see which units of \( S_2 \) can be associated to a particular unit \( k \in S_1 \). We start by computing a \( n_1 \times n_2 \) matrix \( D \) containing the distances between the units of \( S_1 \) and the units of \( S_2 \). We can for example use the usual Euclidean distance or a Mahalanobis distance defined as follows:

\[ d^2(k, \ell) = (x_k - x_\ell)^\top \hat{\Omega}_{xx}^{-1} (x_k - x_\ell), \]

where

\[ \hat{\Omega}_{xx} = \frac{1}{N^*} \left\{ \alpha^* \sum_{k \in S_1} w_{1k} (x_k - \hat{X})(x_k - \hat{X})^\top + (1 - \alpha^*) \sum_{\ell \in S_2} w_{2\ell} (x_\ell - \hat{X})(x_\ell - \hat{X})^\top \right\}, \]

and

\[ \hat{X} = \frac{\hat{X}^*}{N^*}. \]

Then, we search for weights \( W_{k\ell} \) for each couple \( k \in S_1, \ell \in S_2 \). To do this, we solve the following linear program:

\[
\begin{align*}
\text{minimize} & \quad \sum_{k \in S_1} \sum_{\ell \in S_2} W_{k\ell} d(k, \ell) \\
\text{subject to} & \quad \sum_{k \in S_1} W_{k\ell} = w_{2\ell} \quad \text{for all } \ell \in S_2, \\
& \quad \sum_{k \in S_1} W_{k\ell} = w_{1k} \quad \text{for all } k \in S_1, \\
& \quad W_{k\ell} \geq 0, \quad \text{for all } k \in S_1, \text{ and } \ell \in S_2,
\end{align*}
\]
where $W_{kk} = \min(w_{1k}, w_{2k})$, for all couples of identical units in $S_1$ and $S_2$. These constraints force the matching of identical units that can be selected from both samples. This linear program is nothing more than an optimal transport problem for which it exists many efficient implementations.

Most of the $W_{k\ell}$ weights are zero. It is therefore not necessary to manipulate a large matrix of data. The realized calibration is not adversely affected in the linear program. Thus we have

$$
\sum_{k \in S_1} \sum_{\ell \in S_2} W_{k\ell} x_k = \sum_{k \in S_1} \sum_{\ell \in S_2} W_{k\ell} x_{\ell} = \hat{X}^*.
$$

The output of the linear program ends with a matrix of weights $W$ of size $(n_1 \times n_2)$. The non-zero entries in the $i$th rows of the matrix $W$ contain the corresponding weights of the matched units in the sample $S_2$. We generally do not have a one-to-one match, which means that for each unit $k$ in $S_1$ we have more than one unit with weights not equal to 0 in $S_2$. The next two sections propose two different ways to obtain, from the output of the optimal transport, a file where each unit from $S_1$ has only one imputed unit from $S_2$. Without loss of generality, in the following development, we suppose that sample $S_1$ is completed by realizing a prediction from $S_2$.

**Figure 2:** Insight of the statistical matching using optimal transport.

5.1 Matching by using prediction

We can do a prediction by computing the weighted averages of the $x_{\ell}$ and $z_{\ell}$ of $S_2$. Formally, this gives the following quantity to compute:

$$
q_{k\ell} = \frac{W_{k\ell}}{\sum_{\ell \in S_2} W_{k\ell}} = \frac{W_{k\ell}}{w_{1k}}, \text{ for all } k \in S_1, \ell \in S_2.
$$
By using these new weights, we can then compute a prediction of the $x_k$ and the $z_k$
$k \in S_1$,
$$\hat{x}_k = \sum_{\ell \in S_2} q_{k\ell} x_\ell \text{ and } \hat{z}_k = \sum_{\ell \in S_2} q_{k\ell} z_\ell, \text{ for all } k \in S_1.$$ 

The matching quality can be evaluated by comparing the $x_k$ with the predictions $\hat{x}_k$. For the predicted values $\hat{x}_k$, the calibration is always valid. Indeed that
$$\sum_{k \in S_1} w_{1k} \hat{x}_k = \sum_{k \in S_1} w_{1k} x_k = \hat{X}^*.$$ 

However, the interest of the procedure is that we now have predicted values $\hat{z}_k$ for each unit of $S_1$ whereas these variables were only measured on $S_2$.

### 5.2 Matching by using stratified balanced sampling

As explained in Section 5, the output of the optimal transport has generally repetition of some units. In some cases, we are interested in obtaining a synthetic file without any repetition of some units. In this section, we propose an imputation method based on the optimal transport result. We propose to use a balanced sampling method to select from the repetition stratum of a unit $k$ of $S_1$, only one unit of the units of $S_2$. This is in line with hypothesis that $S_2$ is the donor file while the $S_1$ is the recipient file. To do this, we randomly generate a matrix of Bernoulli random variable $a_{k\ell}, k \in S_1, \ell \in S_2$, where $a_{k\ell}$ is 1 if unit $\ell \in S_2$ is imputed to unit $k \in S_1$. Since each unit $k$ of $S_1$ can only receive one imputation, we must have
$$\sum_{\ell \in S_2} a_{k\ell} = 1, \text{ for all } k \in S_1.$$ 

We now want to generate the random matrix of $a_{k\ell}$ with expectations $E(a_{k\ell}) = q_{k\ell}$ in such a way that the following system of equations is satisfied at best
\[
\begin{align*}
\sum_{k \in S_1} \sum_{\ell \in S_2} a_{k\ell} W_{k\ell} x_\ell \approx \sum_{k \in S_1} \sum_{\ell \in S_2} W_{k\ell} x_\ell = \hat{X}^*, \\
\sum_{k \in S_1} \sum_{\ell \in S_2} a_{k\ell} W_{k\ell} z_\ell \approx \sum_{k \in S_1} \sum_{\ell \in S_2} W_{k\ell} z_\ell
\end{align*}
\]
and with
$$\sum_{\ell \in S_2} a_{k\ell} = 1, \text{ for all } k \in S_1.$$ 

This sampling problem is known as ‘stratified balanced sampling’ (see Jauslin et al., 2021b; Hasler and Tillé, 2014; Jauslin et al., 2021a). Indeed, each unit $k$ of $S_1$ can be seen as a stratum for which a unit $\ell$ of $S_2$ must be selected.

The imputed values are then
$$x_k = \sum_{\ell \in S_2} a_{k\ell} x_\ell \text{ and } z_k = \sum_{\ell \in S_2} a_{k\ell} z_\ell, \text{ for all } k \in S_1.$$
Again, we have
\[ \sum_{k \in S_1} w_{1k} \hat{x}_k \approx \sum_{k \in S_1} w_{1k} x_k = \hat{X}^*. \]
However, the interest of the procedure is that we now have values \( \hat{z}_k \) for each unit of \( S_1 \) whereas these variables were only measured on \( S_2 \).

If \( E_q(.) \) is the expectation to the \( a_{kl} \) conditionally to \( S_1 \) and \( S_2 \), then, for all \( k \in S_1 \), we have,
\[
E_q(\hat{x}_k) = \sum_{\ell \in S_2} E_q(a_{kl}) x_{\ell} = \sum_{\ell \in S_2} q_{kl} x_{\ell} = \hat{x}_k,
\]
and
\[
E_q(\hat{z}_k) = \sum_{\ell \in S_2} E_q(a_{kl}) z_{\ell} = \sum_{\ell \in S_2} q_{kl} z_{\ell} = \hat{z}_k.
\]

6 Analysis of the data

Once the optimal transport is performed, different analysis are possible. We can either work directly with the optimal transport result or use the prediction or the imputations methods. This gives us five possibilities to analyse the data.

1. Use full results of the optimal transport problem \((x_k, x_\ell, y_k, z_\ell, W_{kl}, k \in S_1, \ell \in S_2)\).
2. Use the predicted values \((x_k, \hat{x}_k, y_k, \hat{z}_k, w_{1k}, k \in S_1)\) by predicting the values of \( k \in S_1 \).
3. Use the imputed values \((x_k, \hat{x}_k, y_k, \hat{z}_k, w_{1k}, k \in S_1)\) by imputing the values of \( k \in S_1 \).
4. Use the predicted values \((x_\ell, \hat{x}_\ell, y_\ell, z_{2\ell}, \ell \in S_2)\) by predicting the values of \( \ell \in S_2 \).
5. Use the imputed values \((x_\ell, \hat{x}_\ell, y_\ell, z_{2\ell}, \ell \in S_2)\) by imputing the values of \( \ell \in S_2 \).

For all these possibilities, the estimations of the means are completely consistent for the five possibilities. Indeed, we obtain
\[
\hat{Z} = \frac{\sum_{k \in S_1} \sum_{\ell \in S_2} W_{kl} z_\ell}{\sum_{k \in S_1} \sum_{\ell \in S_2} W_{kl}} = \frac{\sum_{\ell \in S_2} w_{2\ell} z_\ell}{\sum_{\ell \in S_2} w_{2\ell}} = \frac{\sum_{k \in S_1} w_{1k} \hat{z}_k}{\sum_{k \in S_1} w_{1k}},
\]
and
\[
\hat{Y} = \frac{\sum_{k \in S_1} \sum_{\ell \in S_2} W_{kl} y_\ell}{\sum_{k \in S_1} \sum_{\ell \in S_2} W_{kl}} = \frac{\sum_{k \in S_1} w_{1k} y_k}{\sum_{k \in S_1} w_{1k}} = \frac{\sum_{\ell \in S_2} w_{2\ell} \hat{y}_\ell}{\sum_{\ell \in S_2} w_{2\ell}}.
\]

If the variables are categorical, we can then estimate a contingency table using the results of the optimal transport matching, or the prediction on \( S_1 \) (respectively on \( S_2 \)),
\[
\hat{N}_{yz} = \sum_{k \in S_1} \sum_{\ell \in S_2} W_{kl} y_\ell z_\ell^T = \sum_{k \in S_1} w_{1k} y_k z_k^T = \sum_{\ell \in S_2} w_{2\ell} \hat{y}_\ell z_\ell^T.
\]
We can also use the imputed values that give slightly different results,
\[
\hat{N}_{yz}^1 = \sum_{k \in S_1} w_{1k} y_k \hat{z}_k^T \text{ and } \hat{N}_{yz}^2 = \sum_{\ell \in S_2} w_{2\ell} \hat{y}_\ell z_\ell^T.
\]
If the variables are continuous, we can estimate the covariances between the $y_k$ and the $z_\ell$ variables. We can also work indifferently from $S_1 \times S_2$, $S_1$ or $S_2$. Indeed, we have

$$\hat{\Sigma}_{yz} = \frac{1}{N^*} \sum_{k \in S_1} \sum_{\ell \in S_2} W_{k\ell} (y_k - \hat{Y})(z_\ell - \hat{Z})^\top$$

$$= \frac{1}{N^*} \sum_{k \in S_1} w_{1k} (y_k - \hat{Y})(z_k - \hat{Z})^\top$$

$$= \frac{1}{N^*} \sum_{\ell \in S_2} w_{2\ell} (\hat{y}_\ell - \hat{Y})(z_\ell - \hat{Z})^\top.$$

As previously seen, it is also possible to use the imputed values that give slightly different results

$$\hat{\Sigma}_{yz}^1 = \frac{1}{N^*} \sum_{k \in S_1} w_{1k} (y_k - \hat{Y})(z_k - \hat{Z})^\top$$

and

$$\hat{\Sigma}_{yz}^2 = \frac{1}{N^*} \sum_{\ell \in S_2} w_{2\ell} (\hat{y}_\ell - \hat{Y})(z_\ell - \hat{Z})^\top.$$

Since $E_q(\hat{y}_k) = \hat{Y}_k$ and $E_q(\hat{z}_k) = \hat{Z}_k$, then $E_q(\hat{\Sigma}_{yz}^1) = E_q(\hat{\Sigma}_{yz}^2) = \hat{\Sigma}_{yz}$. The three estimators are thus very close to each other. One can thus use in an undifferentiated way $S_1 \times S_2$, $S_1$ or $S_2$.

7 Simulations

In this section we propose two simulations, the first one on a simulated normal dataset where the conditional independence assumption holds and a second one on the Austrian data EU-SILC (European Union Statistics on Income and Living Conditions).

7.1 Gaussian example

In this section we discuss a generated dataset such that the CIA is satisfied. Let a population $U$ of 10 000 units generated such that the matching variables $x_k \in \mathbb{R}^p$, $k \in U$, the variables recorded only in $S_1$, $y_k \in \mathbb{R}^q$, $k \in U$ and the variables recorded only in $S_2$, $z_k \in \mathbb{R}^r$, $k \in U$ are normally distributed with joint distribution equals to

$$f(x, y, z|\mu, \Sigma) = \frac{1}{\sqrt{2\pi}^{p+q+r} |\Sigma|} \exp \left[ -\frac{1}{2} \left( (x, y, z)^\top - \mu \right)^\top \Sigma^{-1} \left( (x, y, z)^\top - \mu \right) \right],$$

where the parameters are defined as follow

$$\mu = \begin{pmatrix} \mu_x \\ \mu_y \\ \mu_z \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} & \Sigma_{xz} \\ \Sigma_{yx} & \Sigma_{yy} & \Sigma_{yz} \\ \Sigma_{zx} & \Sigma_{zy} & \Sigma_{zz} \end{pmatrix}.$$
To ensure that the CIA holds, the quantity $\Sigma_{yz}$ must be determined by the following equality (D’Orazio et al., 2006):

$$\Sigma_{yz} = \Sigma_{yz} - \Sigma_{xx}^{-1} \Sigma_{xz}$$

Let suppose that $p = 3$, $q = 2$ and $r = 2$. We simulate 10 000 samples of recipient sample $S_1$ and donor $S_2$ by using simple random sampling without replacement with sizes fixed to $n_1 = 600$ and $n_2 = 3000$. The parameters of the matching variables are equal to

$$\mu_z = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad \Sigma_{zz} = \begin{pmatrix} 7.364 & 2.579 & -0.475 \\ 2.579 & 5.694 & -0.021 \\ -0.475 & -0.021 & 7.864 \end{pmatrix}.$$ 

By using the properties of the Gaussian distribution, if $y_k = (y_{k1} \ y_{k2})^\top \in \mathbb{R}^2, k \in U$ and $z_k = (z_{k1} \ z_{k2})^\top \in \mathbb{R}^2, k \in U$ are linear combination of the matching variables, the CIA holds and we can estimate the covariance matrix $\Sigma_{yz}$. Linear combinations are then equal to

$$y_{k1} = 0.2x_{k1} - 0.3x_{k2} + x_{k3} + \varepsilon_k,$$
$$y_{k2} = 1.2x_{k1} + 0.4x_{k2} - 0.5x_{k3} + \varepsilon_k,$$
$$z_{k1} = -0.4x_{k1} + x_{k2} - 0.3x_{k3} + \varepsilon_k,$$
$$z_{k2} = -1.4x_{k1} + 0.3x_{k2} - 0.6x_{k3} + \varepsilon_k,$$

where $\varepsilon_k \sim \mathcal{N}(0, 1)$ and the variance-covariance matrix is given by

$$\Sigma_{yz} = \begin{pmatrix} -3.667 & -5.368 \\ 2.627 & -9.772 \end{pmatrix}.$$ 

Table 1: Mean squared errors and its bias-variance decomposition on 10 000 simulations of the estimation of the variance-covariance matrix $\Sigma_{yz}$. Matrix notation $\mathbf{B}$ stands for the squared bias while $\mathbf{V}$ denote the variance of the estimator $\hat{\Sigma}_{yz}$.

| Method          | $\mathbf{B}^{opt}_{yz}$ | $\mathbf{V}^{opt}_{yz}$ | $\text{MSE}(\hat{\Sigma}_{yz}^{opt})$ | $\mathbf{B}^{ren}_{yz}$ | $\mathbf{V}^{ren}_{yz}$ | $\text{MSE}(\hat{\Sigma}_{yz}^{ren})$ | $\mathbf{B}^{bal}_{yz}$ | $\mathbf{V}^{bal}_{yz}$ | $\text{MSE}(\hat{\Sigma}_{yz}^{bal})$ |
|-----------------|--------------------------|--------------------------|--------------------------------------|--------------------------|--------------------------|--------------------------------------|--------------------------|--------------------------|--------------------------------------|
| Optimal transport| $\begin{pmatrix} 0.014 & 0.022 \\ 0.013 & 0.046 \end{pmatrix}$ | $\begin{pmatrix} 0.032 & 0.09 \\ 0.042 & 0.162 \end{pmatrix}$ | $\begin{pmatrix} 0.046 & 0.113 \\ 0.056 & 0.208 \end{pmatrix}$ | $\begin{pmatrix} 0 & 0 \\ 0.011 & 0 \end{pmatrix}$ | $\begin{pmatrix} 0.104 & 0.255 \\ 0.162 & 0.568 \end{pmatrix}$ | $\begin{pmatrix} 0.104 & 0.255 \\ 0.173 & 0.568 \end{pmatrix}$ | $\begin{pmatrix} 0.011 & 0.021 \\ 0.013 & 0.045 \end{pmatrix}$ | $\begin{pmatrix} 0.055 & 0.127 \\ 0.086 & 0.235 \end{pmatrix}$ | $\begin{pmatrix} 0.066 & 0.148 \\ 0.099 & 0.279 \end{pmatrix}$ |

Table 1 shows the mean squared errors of the simulations of the three different methods. The mean squared error result shows that optimal transport procedure and balanced imputation method are better. It is important to note here that the optimal transport is
slightly biased. Optimal transport and balanced imputation do not use a model prediction to create the match. In this particular case, where everything is linear, the Renssen’s method will be very powerful in creating a statistical match. Nevertheless, the mean squared error is better than using the optimal transport methods. In general, with a real dataset, the CIA does not hold, and furthermore, it is even not testable. The next section presents an example where the CIA does not hold and the optimal transport methods are better.

7.2 EU-SILC example

This section proposes a simulation study to see how the proposed method performs compared to the method proposed by Renssen (1998) on the dataset eusilc available in the R package Alfons and Templ (2013). This dataset contains 14 827 observations and 28 variables. It is based on real Austrian data EU-SILC (European Union Statistics on Income and Living Conditions). We slightly modified the dataset to remove the missing values. It represents then a dataset of 12 107 observations. Table 2 shows a summary of the different variables used for the simulations. In particular, pl030 is the categorical variable representing the economic status while eqIncome represents the continuous variable of household income. pl030 is the variable of interest recorded only in S_1 while eqIncome is recorded only in S_2. Figure 3 shows the income household by economic status, each simulation will estimate the average income by category.

We run simulations using stratified balanced sampling design (Jauslin et al., 2021b) with sample size n_1 = 1000 for each sample S_1 and n_2 = 4000 for each sample S_2. Inclusion probabilities are selected such that the design respects an optimal stratification, i.e., the number of unit selected in each stratum is proportional to the product of the stratum size and the standard error of eqIncome. In addition, the sample are balanced on the totals of the matching variables.

To measure the effectiveness of our proposed estimators, we use the mean squared error (MSE). Table 3 shows the mean squared errors of the estimation of the average income within each category based on 10 000 simulations. The table displays also the bias-variance decomposition. We can observe that the matching based on optimal transport is the best in terms of mean squared error.
Table 2: Selected variables of the eusilc dataset of the R package developed by Alfons and Templ (2013). The first five variables are the ones used for the matching while the two last ones are the variables of interest.

Matching variables

| Variable | Description |
|----------|-------------|
| hsize | The number of persons in the household. |
| db040 | The federal state in which the household is located. |
| age | The person’s age. |
| rb090 | The person’s gender. (male or female) |
| pb220a | The person’s citizenship (AT, EU and Other). |

Variables of interests

| Variable | Description |
|----------|-------------|
| p1030 | 1: working full time.  
2: working part time.  
3: unemployed.  
4: pupil, student, further training, unpaid work experience,  
in compulsory military or community service.  
5: in retirement or early retirement or has given up business.  
6: permanently disabled or/and unfit to work or other inactive person.  
7: fulfilling domestic tasks and care responsibilities. |
| eqIncome | Slightly simplified version of the household income. |

Figure 3: Boxplot of the household income by economic status.
Table 3: Results of 10,000 simulation for the estimation of average income per category. Relative root mean squared errors as well as bias-variance decomposition are calculated for each economical status. The overall mean squared error are equal to 48.304 for the optimal transport, 62.792 for the balanced imputation and 53.176 for the method of Renssen.

| Method | Economical Status | 1   | 2   | 3   | 4   | 5   | 6   | 7   |
|--------|-------------------|-----|-----|-----|-----|-----|-----|-----|
| MSE    | opt               | 0.072 | 0.058 | 0.158 | 0.151 | 0.02 | 0.247 | 0.241 |
|        | bal               | 0.074 | 0.078 | 0.178 | 0.168 | 0.035 | 0.304 | 0.25  |
|        | ren               | 0.089 | 0.027 | 0.15  | 0.227 | 0.014 | 0.23  | 0.238 |
| Bias   | opt               | 0.072 | 0.051 | 0.151 | 0.146 | 0.016 | 0.229 | 0.239 |
|        | bal               | 0.071 | 0.051 | 0.15  | 0.146 | 0.017 | 0.229 | 0.237 |
|        | ren               | 0.089 | 0.025 | 0.15  | 0.226 | 0.01  | 0.23  | 0.237 |
| Variance| opt              | 0.009 | 0.027 | 0.047 | 0.038 | 0.013 | 0.093 | 0.035 |
|        | bal               | 0.02  | 0.059 | 0.095 | 0.084 | 0.031 | 0.2   | 0.081 |
|        | ren               | 0.006 | 0.009 | 0.01  | 0.013 | 0.01  | 0.02  | 0.012 |
8 Conclusion

Statistical matching is set to become a valuable tool with the increasing amount of data created in this century. In this article, we propose new methods for matching two complex surveys. The proposed statistical matching methods are flexible depending on the type of analysis we want to perform. We can either have a one-to-one unit matching using stratified balanced sampling, or use the optimal output of the linear program, or finally use prediction using weighted averages.

Based on simulations, we observe that the proposed methods have lower cumulative mean squared error. A major assumption that persists in statistical matching is the conditional independence. Since in most cases this assumption is not satisfied, it is generally difficult to do anything other than assuming this assumption. Our method returns a mean squared error smaller in both cases where conditional independence assumption is satisfied or not. Thus, our results show that the proposed methods are less sensitive to a conditional independence defect. This suggests that they are more efficient and give a better quality statistical match.
References

Alfons, A. and Templ, M. (2013). Estimation of social exclusion indicators from complex surveys: The R package laeken. *Journal of Statistical Software*, 54(15):1–25.

Chen, S. and Haziza, D. (2019). Recent developments in dealing with item non-response in surveys: a critical review. *International Statistical Review*, 87:S192–S218.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87:376–382.

Deville, J.-C., Särndal, C.-E., and Sautory, O. (1993). Generalized raking procedure in survey sampling. *Journal of the American Statistical Association*, 88:1013–1020.

D’Orazio, M. (2019). StatMatch: Statistical Matching or Data Fusion. R Foundation for Statistical Computing, Vienna, Austria. R package version 1.4.0.

D’Orazio, M., Di Zio, M., and Scanu, M. (2006). *Statistical matching: Theory and practice*. John Wiley & Sons, Hoboken (New Jersey).

Dudoignon, L. (2018). Fusion statistique de données d’enquêtes : dernières avancées pour les mesures d’audience. Médiamétrie, 70 rue Rivay, 92532 Levallois.

Guandalini, A. and Tillé, Y. (2017). Design-based estimators calibrated on estimated totals from multiple surveys. *International Statistical Review*, 85:250–269.

Hasler, C. and Tillé, Y. (2014). Fast balanced sampling for highly stratified population. *Computational Statistics and Data Analysis*, 74:81–94.

Haziza, D. (2009). Imputation and inference in the presence of missing data. In Pfeffermann, D. and Rao, C. R., editors, *Sample surveys: Design, methods and applications*, pages 215–246, New York, Amsterdam. Elsevier/North-Holland.

Jauslin, R., Eustache, E., Panahbehagh, B., and Tillé, Y. (2021a). StratifiedSampling: Different Methods for Stratified Sampling. R Foundation for Statistical Computing, Vienna, Austria. R package version 0.4.0.

Jauslin, R., Eustache, E., and Tillé, Y. (2021b). Enhanced cube implementation for highly stratified population. *Japanese Journal of Statistics and Data Science*, 4:783–795.

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

Kim, J.-K. and Tam, S.-M. (2020). Data integration by combining big data and survey sample data for finite population inference. *International Statistical Review*, under press:1–20.

Panaretos, V. M. and Zemel, Y. (2020). *An Invitation to Statistics in Wasserstein Space*. Springer International Publishing.

Renssen, R. H. (1998). Use of statistical matching techniques in calibration estimation. *Survey Methodology*, 24(2):171–183.
Yang, S. and Kim, J. K. (2020). Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science*, 3:625–650.