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

Raphaël Jauslin[a] and Yves Tillé[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 correlations. 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.
Key words: auxiliary information, balanced sampling, data integration, distance, unequal probability sampling

[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. It is on this last point that we focus in this paper. While 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). However, massive imputations also allow merging data from different sources.

In this paper, we propose an efficient method of 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 therin).

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 be accurate by integrating information from both sources. The matching should 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 shown in an unpublished 2021 technical report available from Jauslin, R. et al., (arXiv:2101.05568) implemented in the `R` package 'StratifiedSampling' proposed by Jauslin et al. (2021). 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 $\mathbf{x}_k \in \mathbb{R}^p, k \in U$ are measured on both units selected in S_1 and S_2. The vectors of variables $\mathbf{y}_k \in \mathbb{R}^q, k \in U$ are measured only on the units selected in S_1. The vectors of variables $\mathbf{z}_k \in \mathbb{R}^r, k \in U$ 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.

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

$$\mathbf{X} = \sum_{k \in U} \mathbf{x}_k.$$

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

$$\widehat{\mathbf{X}}_{v1} = \sum_{k \in S_1} v_{1k} \mathbf{x}_k, \quad \widehat{\mathbf{X}}_{v2} = \sum_{\ell \in S_2} v_{2\ell} \mathbf{x}_\ell. \tag{1}$$

Following the same idea on the variables of interest, the totals can be estimated

$$\mathbf{Y} = \sum_{k \in U} \mathbf{y}_k \text{ and } \mathbf{Z} = \sum_{k \in U} \mathbf{z}_k,$$

using the following estimators:

$$\widehat{\mathbf{Y}}_{v1} = \sum_{k \in S_1} v_{1k} \mathbf{y}_k, \quad \widehat{\mathbf{Z}}_{v2} = \sum_{\ell \in S_2} v_{2\ell} \mathbf{z}_\ell.$$

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 $\mathbf{x}_k$ the variables $\mathbf{y}_k$ and $\mathbf{z}_k$ are independent, the relationships between the variables $\mathbf{y}_k$ and $\mathbf{z}_k$ can be analyzed. For example, the variables $\mathbf{y}_k$ and $\mathbf{z}_k$ are dummy variables with respectively p and r categories, we could be interested in the estimation of the contingency table

$$\mathbf{N}_{yz} = \sum_{k \in U} \mathbf{y}_k \mathbf{z}_k^\top.$$

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

$$\mathbf{\Sigma}_{yz} = \text{cov}(\mathbf{Y}, \mathbf{Z}).$$

The developments below are also possible if one of the two sources is the whole population. For example, if $S_1 = U$ then $v_{1k} = 1, k \in U$. Then $\widehat{\mathbf{X}}_{v1} = \mathbf{X}$ and $\widehat{\mathbf{Y}}_{v1} = \mathbf{Y}$.

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). Meaning that we are looking for a new weighting system

$$w_{1k}, k \in S_1 \text{ and } w_{2\ell}, \ell \in S_2, \tag{2}$$

such that we have the following results:

$$\widehat{\mathbf{X}}_{w1} = \sum_{k \in S_1} w_{1k}\mathbf{x}_k = \widehat{\mathbf{X}}_{w2} = \sum_{\ell \in S_2} w_{2\ell}\mathbf{x}_\ell.$$

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 $n_{12} = \#(S_2 \cap S_2)$ denote the size of the intersection of the two samples. Guandalini and Tillé (2017) analyse estimators of the form $\widehat{\mathbf{X}}_\alpha = \alpha\widehat{\mathbf{X}}_{v1} + (1-\alpha)\widehat{\mathbf{X}}_{v2}$. They showed that to best estimate $\mathbf{X}$ using both S_1 and S_2, the value of α must be equal to

$$\alpha^{\text{opt}} = \frac{\text{var}(\widehat{\mathbf{X}}_{v2}) - \text{cov}(\widehat{\mathbf{X}}_{v1}, \widehat{\mathbf{X}}_{v2})}{\text{var}(\widehat{\mathbf{X}}_{v1}) + \text{var}(\widehat{\mathbf{X}}_{v2}) - \text{cov}(\widehat{\mathbf{X}}_{v1}, \widehat{\mathbf{X}}_{v2})}.$$

This optimal value minimizes the variance of $\widehat{\mathbf{X}}_\alpha$. However, it depends on unknown variances and 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 α^{opt} that only depends on the sample sizes and the size of the sample overlapped:

$$\alpha^* = \frac{n_1 - n_{12}}{n_1 + n_2 - 2\, n_{12}}. \tag{3}$$

One can then construct the estimator $\widehat{\mathbf{X}}^* = \alpha^*\widehat{\mathbf{X}}_{v1} + (1-\alpha^*)\widehat{\mathbf{X}}_{v2}$. In particular, if $S_2 \subset S_1$, then $\alpha^* = 1$ and $\widehat{\mathbf{X}}^* = \widehat{\mathbf{X}}_{v1}$. Moreover, if $S_2 \cap S_2 = \emptyset$, 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 $\widehat{\mathbf{X}}^*$. 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:

$$\begin{cases} \text{minimize} & \displaystyle\sum_{k \in S_1} G_k(w_{1k}, v_{1k}) \text{ and } \sum_{\ell \in S_2} G_\ell(w_{2\ell}, v_{2\ell}) \\ \text{subject to} & \displaystyle\sum_{k \in S_1} w_{1k}\mathbf{x}_k = \sum_{\ell \in S_2} w_{2\ell}\mathbf{x}_\ell = \widehat{\mathbf{X}}^*, \\ \text{and} & \displaystyle\widehat{N}^* = \sum_{k \in S_1} w_{1k} = \sum_{\ell \in S_2} w_{2\ell} = \alpha^* \sum_{k \in S_1} v_{1k} + (1-\alpha^*) \sum_{\ell \in S_2} v_{2\ell}. \end{cases}$$

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 $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$

$$\widehat{\mathbf{X}}_{w1} = \sum_{k \in S_1} w_{1k}\mathbf{x}_k = \widehat{\mathbf{X}}_{w2} \sum_{\ell \in S_2} w_{2\ell}\mathbf{x}_\ell,$$

$$\widehat{\mathbf{Y}}_{w1} = \sum_{k \in S_1} w_{1k}\mathbf{y}_k \text{ and } \widehat{\mathbf{Z}}_{w2} \sum_{\ell \in S_2} w_{2\ell}\mathbf{z}_\ell.$$

4 Renssen's methods

A method for estimating contingency table is developed in by 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 previous Section 3 and then use the matching variables $\mathbf{x}_k$ to create linear models to get an estimated contingency table. At the first step, regression coefficients $\boldsymbol{\beta}_{yx}$ and $\boldsymbol{\beta}_{zx}$ are computed from the samples S_1 (respectively S_2) by using the weights, w_{1k}, $k \in S_1$ (respectively $w_{2\ell}$, $\ell \in S_2$), described in Equations (2). Using a weighted linear model, the following coefficients are obtained:

$$\widehat{\boldsymbol{\beta}}_{yx} = \left(\sum_{k \in S_1} w_{1k}\mathbf{y}_k\mathbf{x}_k^\top \right) \left(\sum_{k \in S_1} w_{1k}\mathbf{x}_k\mathbf{x}_k^\top \right)^{-1},$$

$$\widehat{\boldsymbol{\beta}}_{zx} = \left(\sum_{\ell \in S_2} w_{1\ell}\mathbf{z}_\ell\mathbf{x}_\ell^\top \right) \left(\sum_{\ell \in S_1} w_{1\ell}\mathbf{x}_\ell\mathbf{x}_\ell^\top \right)^{-1}.$$

The contingency table is then estimated using the matrix product:

$$\widehat{\mathbf{N}}_{yz}^{REN} = \widehat{\boldsymbol{\beta}}_{yx} \left(\alpha^* \sum_{k \in S_1} w_{1k}\mathbf{x}_k\mathbf{x}_k^\top + (1-\alpha^*) \sum_{\ell \in S_2} w_{2\ell}\mathbf{x}_\ell\mathbf{x}_\ell^\top \right) \widehat{\boldsymbol{\beta}}_{zx}^\top.$$

where α^* is the coefficient (3) that depends on the value n_{12}. Renssen's method can be easily generalized to continuous case, but some assumption must be satisfied on variables $\mathbf{y}_k$, $k \in S_1$ and $\mathbf{z}_\ell$, $\ell \in S_2$. 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 Wasserstein distance.

In our case, the optimal transport problem is used to match the units from the sample S_1 to the units of the sample S_2. We start by computing a $n_1 \times n_2$ matrix $\mathbf{D}$ 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:

$$d^2(k,\ell) = (\mathbf{x}_k - \mathbf{x}_\ell)^\top \widehat{\mathbf{\Sigma}}_{xx}^{-1}(\mathbf{x}_k - \mathbf{x}_\ell),$$

where

$$\widehat{\mathbf{\Sigma}}_{xx} = \frac{1}{\widehat{N}^*}\left\{\alpha^* \sum_{k\in S_1} w_{1k}(\mathbf{x}_k - \widehat{\overline{\mathbf{X}}})(\mathbf{x}_k - \widehat{\overline{\mathbf{X}}})^\top + (1-\alpha^*)\sum_{\ell\in S_2} w_{2\ell}(\mathbf{x}_\ell - \widehat{\overline{\mathbf{X}}})(\mathbf{x}_\ell - \widehat{\overline{\mathbf{X}}})^\top\right\},$$

and

$$\widehat{\overline{\mathbf{X}}} = \frac{\widehat{\mathbf{X}}^*}{\widehat{N}^*}.$$

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:

$$\begin{cases} \text{minimize} & \displaystyle\sum_{k\in S_1}\sum_{\ell\in S_2} W_{k\ell}\, d(k,\ell) \\ \text{subject to} & \displaystyle\sum_{k\in S_1} W_{k\ell} = w_{2\ell} \text{ for all } \ell \in S_2, \\ & \displaystyle\sum_{\ell\in S_2} W_{k\ell} = w_{1k} \text{ for all } k \in S_1, \\ & W_{k\ell} \geq 0, \text{ for all } k \in S_1, \text{ and } \ell \in S_2, \end{cases}$$

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 there exist 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 that

$$\sum_{k\in S_1}\sum_{\ell\in S_2} W_{k\ell}\mathbf{x}_k = \sum_{k\in S_1}\sum_{\ell\in S_2} W_{k\ell}\mathbf{x}_\ell = \widehat{\mathbf{X}}^*.$$

The output of the linear program ends with a matrix of weights $\mathbf{W}$ of size $(n_1 \times n_2)$. The non-zero entries in the ith rows of the matrix $\mathbf{W}$ 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 proposed 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.

5.1 Matching by using prediction

We can do a prediction by computing the weighted averages of the $\mathbf{x}_\ell$ and $\mathbf{z}_\ell$ 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 $\mathbf{x}_k$ and the $\mathbf{z}_k$ $k \in S_1$,

$$\widehat{\mathbf{x}}_k = \sum_{\ell \in S_2} q_{k\ell} \mathbf{x}_\ell \text{ and } \widehat{\mathbf{z}}_k = \sum_{\ell \in S_2} q_{k\ell} \mathbf{z}_\ell, \text{ for all } k \in S_1.$$

The matching quality can be evaluated by comparing the $\mathbf{x}_k$ with the predictions $\widehat{\mathbf{x}}_k$. For the predicted values $\widehat{\mathbf{x}}_k$, the calibration is always valid. Indeed that

$$\sum_{k \in S_1} w_{1k} \widehat{\mathbf{x}}_k = \sum_{k \in S_1} w_{1k} \mathbf{x}_k = \widehat{\mathbf{X}}^*.$$

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

5.2 Matching by using balanced sampling

In this section, we propose an imputation method based on the optimal transport result. We propose to choose a unit ℓ of S_2 to assign the value $\mathbf{z}_\ell$ to a particular unit k of S_1. 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 $\mathrm{E}(a_{k\ell}) = q_{k\ell}$ in such a way that the following system of equations is satisfied at best

$$\sum_{k \in S_1} \sum_{\ell \in S_2} \frac{a_{k\ell} W_{k\ell}}{q_{k\ell}} \mathbf{x}_\ell \approx \sum_{k \in S_1} \sum_{\ell \in S_2} W_{k\ell} \mathbf{x}_\ell = \widehat{\mathbf{X}}^*,$$

$$\sum_{k \in S_1} \sum_{\ell \in S_2} \frac{a_{k\ell} W_{k\ell}}{q_{k\ell}} \mathbf{z}_\ell \approx \sum_{k \in S_1} \sum_{\ell \in S_2} W_{k\ell} \mathbf{z}_\ell$$

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 Hasler and Tillé, 2014; Jauslin et al., 2021). Indeed, each unit k of S_1 can be seen as a stratum for which a unit ℓ of S_2 must be selected.

The imputed values are then

$$\mathring{\mathbf{x}}_k = \sum_{\ell \in S_2} a_{k\ell} \mathbf{x}_\ell \text{ and } \mathring{\mathbf{z}}_k = \sum_{\ell \in S_2} a_{k\ell} \mathbf{z}_\ell, \text{ for all } k \in S_1.$$

Again, we have

$$\sum_{k \in S_1} w_{1k} \mathring{\mathbf{x}}_k \approx \sum_{k \in S_1} w_{1k} \mathbf{x}_k = \widehat{\mathbf{X}}^*.$$

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

If $\mathrm{E}_q(.)$ is the expectation to the $a_{k\ell}$ conditionally to S_1 and S_2. Then, for all $k \in S_1$, we have,

$$\mathrm{E}_q(\mathring{\mathbf{x}}_k) = \sum_{\ell \in S_2} \mathrm{E}_q(a_{k\ell})\mathbf{x}_\ell = \sum_{\ell \in S_2} q_{k\ell}\mathbf{x}_\ell = \widehat{\mathbf{x}}_k,$$

end

$$\mathrm{E}_q(\mathring{\mathbf{z}}_k) = \sum_{\ell \in S_2} \mathrm{E}_q(a_{k\ell})\mathbf{z}_\ell = \sum_{\ell \in S_2} q_{k\ell}\mathbf{z}_\ell = \widehat{\mathbf{z}}_k.$$

6 Analysis of the data

In order to analyse the data, there exist five possibilities:

1. One can use full result of the optimal transport problem $(\mathbf{x}_k, \mathbf{x}_\ell, \mathbf{y}_k, \mathbf{z}_\ell, W_{k\ell}, k \in S_1, \ell \in S_2)$.

2. One can use the predicted values $(\mathbf{x}_k, \widehat{\mathbf{x}}_k, \mathbf{y}_k, \widehat{\mathbf{z}}_k, w_{1k}, k \in S_1)$ by predicting the values of $k \in S_1$.

3. One can use the imputed values $(\mathbf{x}_k, \mathring{\mathbf{x}}_k, \mathbf{y}_k, \mathring{\mathbf{z}}_k, w_{1k}, k \in S_1)$ by imputing the values of $k \in S_1$.

4. One can use the predicted values $(\mathbf{x}_\ell, \widehat{\mathbf{x}}_\ell, \widehat{\mathbf{y}}_\ell, \mathbf{z}_\ell, w_{2\ell}, \ell \in S_2)$ by predicting the values of $\ell \in S_2$.

5. One can use the imputed values $(\mathbf{x}_\ell, \mathring{\mathbf{x}}_\ell, \mathring{\mathbf{y}}_\ell, \mathbf{z}_\ell, w_{2\ell}, \ell \in S_2)$ by imputing the values of $\ell \in S_2$.

For all these possibilities the estimation of means is completely consistent for the five possibilities. Indeed, we obtain

$$\widehat{\overline{\mathbf{Z}}} = \frac{\sum_{k\in S_1}\sum_{\ell\in S_2} W_{k\ell}\mathbf{z}_\ell}{\sum_{k\in S_1}\sum_{\ell\in S_2} W_{k\ell}} = \frac{\sum_{\ell\in S_2} w_{2\ell}\mathbf{z}_\ell}{\sum_{\ell\in S_2} w_{2\ell}} = \frac{\sum_{k\in S_1} w_{1k}\widehat{\mathbf{z}}_k}{\sum_{k\in S_1} w_{1k}} \approx \frac{\sum_{k\in S_1} w_{1k}\mathring{\mathbf{z}}_k}{\sum_{k\in S_1} w_{1k}},$$

and

$$\widehat{\overline{\mathbf{Y}}} = \frac{\sum_{k\in S_1}\sum_{\ell\in S_2} W_{k\ell}\mathbf{y}_k}{\sum_{k\in S_1}\sum_{\ell\in S_2} W_{k\ell}} = \frac{\sum_{k\in S_1} w_{1k}\mathbf{y}_k}{\sum_{k\in S_1} w_{1k}} = \frac{\sum_{\ell\in S_2} w_{2\ell}\widehat{\mathbf{y}}_\ell}{\sum_{\ell\in S_2} w_{2\ell}} \approx \frac{\sum_{\ell\in S_2} w_{2\ell}\mathring{\mathbf{y}}_\ell}{\sum_{\ell\in S_2} w_{2\ell}}.$$

If the variables are categorical, we can then estimate a contingency table using the results of the optimal transport matching, $S_1 \times S_2$, the prediction on S_1 (respectively on S_2),

$$\widehat{\mathbf{N}}_{yz} = \sum_{k\in S_1}\sum_{\ell\in S_2} W_{k\ell}\mathbf{y}_k\mathbf{z}_\ell^\top = \sum_{k\in S_1} w_{1k}\mathbf{y}_k\widehat{\mathbf{z}}_k^\top = \sum_{\ell\in S_2} w_{2\ell}\widehat{\mathbf{y}}_\ell\mathbf{z}_\ell^\top.$$

We can also use the imputed values that give slightly different results,

$$\widehat{\mathbf{N}}^1_{yz} = \sum_{k\in S_1} w_{1k}\mathbf{y}_k\mathring{\mathbf{z}}_k^\top \text{ and } \widehat{\mathbf{N}}^2_{yz} = \sum_{\ell\in S_2} w_{2\ell}\mathring{\mathbf{y}}_\ell\mathbf{z}_\ell^\top.$$

If the variables are continuous, we can estimate the covariances between the $\mathbf{y}_k$ and the $\mathbf{z}_\ell$ variables, we can also work indifferently from $S_1 \times S_2$, S_1 or S_2. Indeed, we have

$$\begin{aligned}\widehat{\mathbf{\Sigma}}_{yz} &= \frac{1}{\widehat{N}^*}\sum_{k\in S_1}\sum_{\ell\in S_2} W_{k\ell}(\mathbf{y}_k - \widehat{\overline{\mathbf{Y}}})(\mathbf{z}_\ell - \widehat{\overline{\mathbf{Z}}})^\top \\ &= \frac{1}{\widehat{N}^*}\sum_{k\in S_1} w_{1k}(\mathbf{y}_k - \widehat{\overline{\mathbf{Y}}})(\widehat{\mathbf{z}}_k - \widehat{\overline{\mathbf{Z}}})^\top \\ &= \frac{1}{\widehat{N}^*}\sum_{\ell\in S_2} w_{2\ell}(\widehat{\mathbf{y}}_\ell - \widehat{\overline{\mathbf{Y}}})(\mathbf{z}_\ell - \widehat{\overline{\mathbf{Z}}})^\top.\end{aligned}$$

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

$$\widehat{\mathbf{\Sigma}}^1_{yz} = \frac{1}{\widehat{N}^*}\sum_{k\in S_1} w_{1k}(\mathbf{y}_k - \widehat{\overline{\mathbf{Y}}})(\mathring{\mathbf{z}}_k - \widehat{\overline{\mathbf{Z}}})^\top$$

and

$$\widehat{\mathbf{\Sigma}}^2_{yz} = \frac{1}{\widehat{N}^*}\sum_{\ell\in S_2} w_{2\ell}(\mathring{\mathbf{y}}_\ell - \widehat{\overline{\mathbf{Y}}})(\mathbf{z}_\ell - \widehat{\overline{\mathbf{Z}}})^\top.$$

Since $\mathrm{E}_q(\mathring{\mathbf{y}}_k) = \widehat{\mathbf{y}}_k$ and $\mathrm{E}_q(\mathring{\mathbf{z}}_k) = \widehat{\mathbf{z}}_k$, then $\mathrm{E}_q(\widehat{\mathbf{\Sigma}}^1_{yz}) = \mathrm{E}_q(\widehat{\mathbf{\Sigma}}^2_{yz}) = \widehat{\mathbf{\Sigma}}_{yz}$. 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

This section proposes a simulation to see how the proposed method is working 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 in order to remove the missing values. It represents then a dataset of 12 107 observations. We estimate the contingency table when the categorical variable which represents the economic status (`pl030`) is crossed with a discretized version of the equivalized household income (`eqIncome`). In order to discretize the equivalized income, we have calculated percentiles (0.15,0.30,0.45,0.60,0.75,0.90) of the variable and defined the category as intervals between the values. Table 1 shows a summary of the different variables while Table 2 presents the contingency table that we are going to estimate.

We run simulations using simple random sampling with sample size $n_1 = 5000$ for each sample S_1 (respectively $n_2 = 3000$ for each sample S_2). To measure the effectiveness of our proposed estimators, we use the mean squared error (MSE). Table 3 shows the mean squared errors based on 10000 simulations of the estimation of Table 2. This means that for each entry in the contingency table, we calculate the mean squared error. Pearson's ϕ independence between the variable `pl030` and the `eqIncome` is calculated on the true contingency table and on each estimated table from a simulation. Table 4 shows the mean squared errors and their decomposition into bias and variance of the Pearson's ϕ coefficients of the different methods. The first row of the table shows the ϕ coefficient

calculated on the true contingency table. It means that the closer the ϕ coefficient of a particular method is to this referenced value, the more efficient the method is at returning the true result. All methods based on optimal transport are implemented in the `R` package 'StratifiedSampling' (Jauslin et al., 2021) while the Renssen's method can be find in the `R` package 'StatMatch' proposed by D'Orazio (2019).

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

[Table 4 about here.]

8 Conclusion

Statistical matching is set to become a valuable tool with the increasing amount of data created in this century. In this manuscript, 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 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 square error. In addition, the ϕ values are closer to the reference value. A major problem that persists in statistical matching is the assumption of conditional independence. Since in most cases this assumption is not satisfied, it is generally difficult to do anything other than assuming that this assumption is verified. Our method returns a ϕ estimate closer to reality as shown by the simulations. Thus, our results show that the proposed methods are less sensitive to a conditional independence defect, which 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).

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., and Tillé, Y. (2021). *StratifiedSampling: Different Methods for Stratified Sampling*. R Foundation for Statistical Computing, Vienna, Austria. R package version 0.1.0.

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 calibation 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.

Table 1: Selected variables of the `eusilc` dataset of the `R` package developed by Alfons and Templ (2013). The first five variables are the one used for the matching while the two last one are the variable of interest to compute the contingency table.

`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).
`pl030`	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`	(0,15] : income smaller than 11348.13. (15,30] : income between 11348.13 and 14722.05. (30,45] : income between 14722.05 and 17643.73. (45,60] : income between 17643.73 and 20887.35. (60,75] : income between 20887.35 and 24816.57. (75,90] : income between 24816.57 and 32641.01 . (90,100] : income larger than 32641.01.

Table 2: Contigency table computed by crossing the economical status (`pl030`) and the discretized equivalized incomce (`eqIncome`).

	(0,15]	(15,30]	(30,45]	(45,60]	(60,75]	(75,90]	(90,100]
1	409	616	722	807	935	1025	648
2	189	181	205	184	165	154	82
3	137	90	72	75	59	52	33
4	210	159	103	95	74	49	46
5	470	462	492	477	459	435	351
6	57	25	28	30	17	11	10
7	344	283	194	149	106	91	40

Table 3: Mean squared error of the contingency Table 2 computed on 10 000 simulations. The proposed methods are compared to the Renssen method. Entries in bold are those that are smaller than the Renssen mean squared error. The total sum of the mean squared error on each table are equal to 284867.6, 146975.4 and 151837.7.

	(0,15]	(15,30]	(30,45]	(45,60]	(60,75]	(75,90]	(90,100]
Renssen							
1	93431.3	26204.8	2713	1625.7	19819.7	46848.6	18065.1
2	136.4	120.1	708	237.5	104.1	159.8	404.5
3	2956	192.7	28.8	31.4	394.5	636.9	407.2
4	11871.5	1852.1	130.5	400.4	1717.8	4478	408.7
5	1317.3	1457.7	1531	1174.7	858.3	1860.2	722.9
6	840.2	14.9	8.5	15.2	97.7	206.1	34.5
7	18594.3	7056.7	146.3	1277	3682.7	4493.6	3392.6
Optimal							
1	**39931.1**	**12516.8**	**1996.3**	1939	**10579.4**	**17807.2**	**9303.7**
2	433.1	360.7	858.9	486.4	414.1	612.6	660
3	**2134.6**	307	142.8	142.6	468.5	**578.9**	**288.5**
4	**4145.5**	**757.1**	306.2	406.1	**684.6**	**1206**	**195.8**
5	3445.9	**1137.5**	**1238**	1363	1279.7	**1137.9**	909.1
6	**503.8**	46.7	53.2	53.8	103.8	**197.4**	40.2
7	**11970.3**	**4523.5**	387.9	**991.5**	**2738**	**2941.2**	**2249.3**
Balanced Sampling							
1	**40129**	**12858.7**	**2283.7**	2246.4	**10900.1**	**18140.5**	**9536.2**
2	503.1	435.8	931.3	558.7	483.1	679.2	703.7
3	**2164.8**	336.4	172.1	172.2	504.7	**612.9**	**310.2**
4	**4196.3**	**808.3**	352	454.1	**726.3**	**1247.7**	**221.9**
5	3651.3	**1336.1**	**1441.1**	1547.3	1464.8	**1298.9**	1060.1
6	**513.1**	56.8	64.1	64.9	114	208.6	46.7
7	**12030.7**	**4626.5**	465	**1061.2**	**2802.2**	**3012.3**	**2302.9**

Table 4: Pearson ϕ coefficient and its mean squared error computed on 10 000 simulations. For each methods the decomposition into bias and variance is printed. The table contains also the coefficient ϕ computed on the true contingency Table 2.

	ϕ	MSE	biais	variance
True contingency table	0.290	-	-	-
Renssen	0.070	0.04834	-0.21969	0.00007
Optimal	0.145	0.02113	-0.14487	0.00015
Balanced Sampling	0.150	0.01983	-0.14019	0.00018