Valérie Gares[1] / Chloé Dimeglio[2] / Grégory Guernec[3] / Romain Fantin[4] / Benoit Lepage[5] / Michael R. Kosorok[6] / Nicolas Savy[7]

On the Use of Optimal Transportation Theory to Recode Variables and Application to Database Merging

[1] University of Rennes, INSA, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France, E-mail: valerie.gares@insa-rennes.fr
[2] University of Toulouse III - INSERM, UMR 1027 - CHU Toulouse, Toulouse, France, E-mail: chloe.dimeglio@univ-tlse3.fr
[3] INSERM, UMR 1027, Toulouse, France, E-mail: gregory.guernec@inserm.fr
[4] University of Toulouse III, Toulouse, France, E-mail: rom.fantin@wanadoo.fr
[5] University of Toulouse III - INSERM, UMR 1027 - CHU Toulouse, Toulouse, France, E-mail: benoit.lepage@univ-tlse3.fr
[6] Department of Biostatistics, University of North Carolina at Chapel Hill - Chapel Hill, NC, USA, E-mail: kosorok@unc.edu
[7] Toulouse Institute of Mathematics UMR C5583, Toulouse, France, E-mail: Nicolas.Savy@math.univ-toulouse.fr

Abstract:
Merging databases is a strategy of paramount interest especially in medical research. A common problem in this context comes from a variable which is not coded on the same scale in both databases we aim to merge. This paper considers the problem of finding a relevant way to recode the variable in order to merge these two databases. To address this issue, an algorithm, based on optimal transportation theory, is proposed. Optimal transportation theory gives us an application to map the measure associated with the variable in database A to the measure associated with the same variable in database B. To do so, a cost function has to be introduced and an allocation rule has to be defined. Such a function and such a rule is proposed involving the information contained in the covariates. In this paper, the method is compared to multiple imputation by chained equations and a statistical learning method and has demonstrated a better average accuracy in many situations. Applications on both simulated and real datasets show that the efficiency of the proposed merging algorithm depends on how the covariates are linked with the variable of interest.

DOI: 10.1515/ijb-2018-0106
Received: October 15, 2018; **Revised:** July 17, 2019; **Accepted:** August 8, 2019

1 Introduction

Nowadays, sharing and producing information from heterogeneous sources becomes a major issue and is an important and ubiquitous challenge in the Big Data era. This question is now widely found not only in medical field but also in spatial data processing, finance, robotics, and in many other fields where the need of global and quality knowledge is required to make a better decision. The main issue when merging databases is to associate, mix and include databases from different sources in order to provide an enriched synthetic database. The underlying idea is more information is extracted from merged database than we would obtain from using the databases separately [1, 2].

In the field of database fusion, different techniques are widely used to produce combinations of heterogeneous data from different sources [3], especially probabilistic models [4], Hidden Markov Models [5], technique of least square, multi-agent systems [6], logical reasoning [7] and, probably the best known, the Bayes rule [8, 9].

In this paper, one focuses our attention to a specific issue related to database fusion: variable recoding problem. When two databases have to be merged, it is usual and problematic to observe a categorical variable that is not coded in the same scale in both databases. This problem may occur in many situations: for example, a change in the associated collection questionnaire for asking the same information between two waves of recruitment (for different subjects) or two waves at different ages (for same subjects) in two different studies, or different questionnaires for asking the same information (for different subjects).

The motivation of this investigation comes from the analysis of a french longitudinal cohort of children: ELFE study [10]. A variable of interest is the answer of the question: "how would you rate your overall health?".

Valérie Gares is the corresponding author.

During the first baseline data collection wave (January to April 2011), the different possible answers are proposed in a five point ordinal scale: "excellent", "very well", "well", "fair", "bad" and during the second baseline data collection wave (May to December 2011), there are a five other point ordinal scale: "very well", "well", "medium", "bad" and "very bad". This difference in coding information yields to difficulties to compare these two waves. A preliminary step of recoding appears to be an appealing strategy.

The problem can be formalized in terms of two databases A and B: the first contains the observations of $P+Q$ covariates $(X^1,\dots,X^{P+Q})$ measured on $n_A$units, the second the observations of a subset of P covariates $(X^1,\dots,X^P)$ measured on n_B units. Consider a variable Y observed by means of Y^A on database A and by means of Y^B on database B (see Table 1).

Table 1: Statement of the database merging problem.

Database A					Database B				
X^1	...	X^{P+Q}	Y^A	Y^B	X^1	...	X^P	Y^A	Y^B
1			Observed	Unobserved	1			Unobserved	Observed
...					...				
...					...				
n_A					n_B				

To make inference and analysis of the merged database, it is therefore necessary to find a common scale of assessment. The objective is thus to complete Y^Aon database B and/or complete Y^Bon database A.

This issue can be seen from different points of view involving different families of techniques:

1. **Statistical matching**. The problem could be seen as a specific data integration problem which could be assimilated to record linkage or more particularly to statistical matching. Indeed, by supposing that Y^A and Y^B are two distinct variables observed in two different databases A and B. This problem refers to statistical matching and several methods exist to solve it [11, 12]. At the so-called micro level, the aim is to generate a common database in which all the variables are filled-in by imputing Y^A in database B (and/or vice versa) by means of information extracted from the set of common covariates $\mathbf{X}$. However, these methods are based on imputation procedures (declination of hot-deck approaches) and require the conditional independence assumption (independence between Y^A and Y^B given $\mathbf{X}$). This assumption remains a strong drawback for their applicability, while the use of external auxiliary information, as described in [13], still remains a seldom usable alternative. In our specific context, Y^A and Y^B are samples representing the same information, the conditional independence assumption is obviously not satisfied and using these procedures nevertheless could lead to important estimation bias.

2. **Imputation procedures**. The variable recoding issue could be viewed as a classical missing data problem. In this context, the missingness process is clearly considered missing at random (MAR) following Rubin's classification [14]. This problem has been widely studied in the literature and many existing methods for imputing missing data exist. A popular approach for multiple imputations, known for its flexibility and its ability to generate plausible values, is MICE algorithm (Multivariate Imputation by Chained Equations, [15]) which generates multiple imputations for incomplete datasets by itering conditional densities using Gibbs sampling (fully conditional specification).

3. **Supervised learning procedures**. Classification learning could be also considered to solve this problem. Indeed, considering a first step consisting in predicting the outcome for example Y^A in database A from common covariates and a second step consisting in predicting Y^A in database B with the same covariates using parameters estimated in the first step [16, 17].

4. **Methods for latent variables**. As Y^A and Y^B refers to the same information Y this can be interpreted as a latent variable. The objective of those methods is to model a common latent variable Y[18, 19].

By construction, many methods from these families listed below only account for the information contained in database A to complete Y^B and contained in database B to complete Y^A. The information contained in Y^A on database A (resp. Y^B on database B) may be better exploited. Assuming that the distribution of Y^A (resp. Y^B) is the same in database A and B, the theory of optimal transportation (see [20] for a survey) exhibits a map that pushes the distribution of Y^A forward to the distribution of Y^B. Using that map and the link between covariates and outcome, new algorithm of recoding, called the OT-algorithm (Optimal Transportation algorithm) can be constructed. To do so, we have to assume that the covariates explain the outcomes Y^A and Y^B similarly in the

two databases. Compared to this family of methods, the algorithm proposed presents the advantage to consider all the information contained in the two databases in only one model. In the authors' knowledge, this is the first attempt to use optimal transportation theory in this context.

In order to challenge OT-algorithm, various methods, chosen from the previously introduced families, have been taken as comparators: the polytomous logistic regression (PR) [21], supervised learning procedures which generalizes logistic regression to multiclass outcomes and MICE as the current reference method among the imputation procedures. The aim of the other methods evoked is slightly different.

This article is organized as follows: a brief review of Optimal Transportation theory together with the application to the variable recoding problem is described in Section 2. Section 3 details OT-algorithm based on Optimal Transportation. The assessments of the average accuracies of the algorithm are investigated in Section 4 by means of simulation studies. The first simulation study is based on a "deterministic decision rule" in order to investigate the intrinsic average accuracies of the OT-algorithm. Indeed, this algorithm is based on an estimation procedure which necessitates sufficiently large sample sizes for databases A and B. The minimal size is evaluated in Section 4.1. The second simulation study in Section 4.2 is based on a "stochastic decision rule" in order to link the average accuracy of the OT-algorithm with the correlation between covariates and outcome. The average accuracies of the OT-algorithm are compared with multiple imputation technique and to polytomous logistic regression. Section 5 is the application of OT-algorithm on a real dataset. Finally, some concluding remarks are given in Section 6.

2 Optimal transportation

Consider a pile of sand distributed with density f, that has to be moved to fill a hole (of the same total volume) according to a new distribution, whose density is prescribed and is g. Consider a map T describing this movement, $T(x)$ represents the destination of the particle of sand originally located at x. The Optimal Transportation problem consists in finding a map T such that the average displacement is minimal (a cost function c measuring the displacement from x to y has to be introduced at this point). This is the original statement of the Tranportation problem due to Gaspar Monge [22].

2.1 Abstract statements of the optimal transportation problem

Consider $\mathbb{X}$ and $\mathbb{Y}$ two Radon spaces. Given μ a probability measure on $\mathbb{X}$, ν a probability measure on $\mathbb{Y}$ and $c : \mathbb{X} \times \mathbb{Y} \longrightarrow [0, \infty]$ a Borel-measurable function (the cost function), Monge's formulation of the optimal transportation problem consists in finding a map (transport map) $T : \mathbb{X} \to \mathbb{Y}$ that realizes the infimum:

$$\left\{ \int_{\mathbb{X}} c(x, T(x)) \, \mathrm{d}\mu(x) \;\middle|\; T_*(\mu) = \nu \right\}, \tag{1}$$

where $T_*(\mu)$ denotes the so-called push-forward measure of μ (the image measure of μ by T).

A map T that attains this infimum is called an "optimal transportation map". Monge's formulation of the optimal transportation problem may be ill-posed, because sometimes there is no T satisfying $T_*(\mu) = \nu$. This happens for example when μ is a Dirac measure but ν is not. Monge's formulation of the transportation problem is a strongly non-linear optimization problem and to find a solution requires rigid assumptions on the regularity of T and on the cost function.

Kantorovich's formulation [23] consists in finding a measure $\gamma \in \Gamma(\mu, \nu)$ that realizes the infimum:

$$\left\{ \int_{\mathbb{X}\times\mathbb{Y}} c(x, y) \, \mathrm{d}\gamma(x, y) \;\middle|\; \gamma \in \Gamma(\mu, \nu) \right\}, \tag{2}$$

where $\Gamma(\mu,\nu)$ denotes the set of measures on $\mathbb{X} \times \mathbb{Y}$ with marginals μ on $\mathbb{X}$ and ν on $\mathbb{Y}$. This is related to optimal coupling theory. Kantorovich's formulation plugs the problem in a linear setting and the solution is achievable thanks to compacity argument. It can be shown [20] that a minimizer for this problem always exists as soon as the cost function c is lower semi-continuous.

2.1.1 The discrete case on the line

In the discrete case, the Optimal Transportation problem is known as Hitchcock's problem [24]. The measures are defined by weighted Dirac measures (δ_x denotes Dirac measure at point x):

$$\mu = \sum_{r=1}^{R} a_r \delta_{p_r} \qquad \text{and} \qquad \nu = \sum_{s=1}^{S} b_s \delta_{q_s}$$

where $\{p_1, \dots, p_R\}$ (resp. $\{q_1, \dots, q_S\}$) are the locations of point masses of measure μ (resp. ν) and a_r (resp. b_s) are the weights verifying $\sum_{r=1}^{p} p_r = \sum_{s=1}^{q} q_r = 1$.

The Optimal Transportation problem in this setting consists in finding a measure γ which satisfies eq. (2). In this context, γ is a $S \times R$ matrix and for any r and any s, $\gamma_{r,s}$ represents the joint probability $(p_r, q_s) \to \mathbb{P}(X = p_r, Y = q_s)$, where $X \sim \mu$ and $Y \sim \nu$ and can be seen as a map from modality p_r of X to modality q_s of Y. The cost function is, in this setting, a $S \times R$ matrix $(c(p_r, q_s),\ r = 1, \dots, R\ ;\ s = 1,\ \dots, S)$. The problem consists in finding γ that minimizes:

$$\sum_{r=1}^{R} \sum_{s=1}^{S} \gamma_{r,s}\, c(p_r, q_s),$$

under the following constraints, for any r and any s,

$$\gamma_{r,s} \geq 0, \qquad \sum_{r=1}^{R} \gamma_{r,s} = b_s \qquad \text{and} \qquad \sum_{s=1}^{S} \gamma_{r,s} = a_r.$$

2.2 Application to database merging

In the sequel, our attention focuses on the discrete setting which is the most common and the hardest to handle setting.

2.2.1 General considerations

Consider two databases A and B we aim to merge. The same covariates are assessed on both databases. Denote $\mathbf{X} = (X^1, \dots, X^P)$ the set of P covariates observed in both databases A and B and $\mathbf{X}_i^A$ (resp. $\mathbf{X}_j^B$) the values of $\mathbf{X}$ observed for patients i of database A (resp. j of database B). Our attention focuses on a variable Y evaluated in both databases but not assessed on the same variable. Denote Y^A the assessment of Y on database A and Y^B the assessment of Y on database B. For example Y could be measured by a three-category discretization on A and by a four-category discretization on B. Table 1 with $Q = 0$ illustrates the appearance of the databases we are describing. In order to merge those databases, we have to complete Y^A on database B and/or complete Y^B on database A. Note that the problem is not reversible when the number of modalities is not the same. Let μ be the distribution of Y^A and ν the distribution of Y^B. Distribution μ (resp. ν) is assumed discrete with modalities $\{p_1, \dots, p_R\}$ (resp. $\{q_1, \dots, q_S\}$). We denote by ind(A) $= \{1, \dots, n_A\}$, ind(B) $= \{1, \dots, n_B\}$ and ind(A $\cup$ B) $= \{1, \dots, n_A + n_B\}$.

2.2.2 Assumptions

In order to properly plug our problem in an Optimal Transportation framework, two assumptions have to be fulfilled.

- Assumption 1 :
  - $(Y_k^A, k \in \text{ind(A} \cup \text{B)})$ are i.i.d with same distribution μ,
  - $(Y_k^B, k \in \text{ind(A} \cup B))$ are i.i.d with same distribution ν.

  Assumption 1 imposes that the unobserved valued of Y^A (resp. Y^B) on database B (resp. A) comes from the same distribution as Y^A (resp. Y^B) on database A (resp. B).

– Assumption 2 : $(Y_k^A|\mathbf{X}_k^A, k \in \text{ind}(\text{A} \cup \text{B}))$ (resp. $(Y_k^B|\mathbf{X}_k^B, k \in \text{ind}(\text{A} \cup \text{B}))$) are i.i.d with same distribution as $Y^A|\mathbf{X}^A$ (resp. $Y^B|\mathbf{X}^B$).

Assumption 2 demands that the covariates explain the outcomes Y^A and Y^B similarly in both databases. Notice that Assumption 2 cannot be verified from the data. That allows us to define a relevant cost function in Section 2.2.3 below.

2.2.3 Cost function

The problem reduces to the choice of a relevant cost function between modality p_r of μ and modality q_s of ν. To define such a cost, our attention restricts to patients satisfying modality p_r in database A and patients satisfying modality q_s in database B. A natural way to do so is to consider a cost function between a modality p_r in database A and a modality q_s in database B which is small if these modalities refer to the same individuals and which is large if these modalities refer to different individuals. As distribution of Y^A and Y^B are never observed for the same individuals, this function refers to the distance between covariates vectors of individuals in database A having modality p_r and individuals in database B having modality q_s. Thus, considering d a distance on $\mathbb{R}^P$, a relevant cost function is defined as:

$$\begin{aligned} c(p_r, q_s) &= \mathbb{E}\left[d(\mathbf{X}^A, \mathbf{X}^B)|Y^A = p_r, Y^B = q_s\right] && \text{if } \mathbb{P}[Y^A = p_r, Y^B = q_s] \neq 0 \\ &= 0 && \text{otherwise} \end{aligned}$$

where $\mathbf{X}^A$ and $\mathbf{X}^B$ are independent.

The choice of the distance d depends on the type of the covariates. This may necessitate a preliminary transformation of the covariates. For example, in the case of only categorical covariates were considered, the Hamming distance from the associated complete disjunctive tables can be used. In the case of continuous covariates, one can use directly the Euclidean or Manhattan distance. Finally, in the case of mixed types of covariates, a distance for mixed data could be used (e. g. the Heterogeneous Euclidean-Overlap Metric [25], the Value Difference Metric [26], or the Mahalanobis distance) or a distance for continuous covariates applied on the coordinates extracted from a factor analysis of mixed data [27].

3 Algorithm for variable recoding: OT-algorithm

Consider γ^{opt}, the optimal joint distribution of (Y^A, Y^B) defined, as explained in Section 2.1.1 as the solution to Hitchcock's problem, solution to the linear programming: γ^{opt} is the minimum of:

$$\gamma = \{\gamma_{r,s}, r = 1, \dots, R, s = 1, \dots, S\} \to \sum_{r=1}^{R}\sum_{s=1}^{S} \gamma_{r,s}\, c\,(p_r, q_s)\,,$$

under the following constraints:

$$\begin{cases} \sum_{r=1}^{R} \gamma_{r,s} = \mu_s, & \forall s = 1, \dots S \\ \sum_{s=1}^{S} \gamma_{r,s} = \nu_r, & \forall r = 1, \dots R \\ \gamma_{r,s} \geq 0, & \forall r = 1, \dots R, \forall s = 1, \dots S. \end{cases}$$

The keypoint of the method is to consider an estimator of γ^{opt}. To do so, it is natural to consider the empirical distributions of μ and ν given by the estimator $\hat{a}_r$ (resp. $\hat{b}_s$) defined as:

$$(\hat{a}_{n_A})_r = \frac{1}{n_A}\sum_{i=1}^{n_A} \mathbb{I}_{\{Y_i^A = p_r\}}, \quad r = 1, \dots R \tag{3}$$

$$(\hat{b}_{n_B})_s = \frac{1}{n_B}\sum_{j=1}^{n_B} \mathbb{I}_{\{Y_j^B = q_s\}}, \quad s = 1, \dots S \tag{4}$$

and to consider as estimator of γ^{opt}, a solution $\hat{\gamma}^{opt}_{n_A,n_B}$ to the Hitchcock's problem associated with an estimator of the cost function, solution to the linear programming: $\hat{\gamma}^{opt}_{n_A,n_B}$ is the minimum of:

$$\gamma = \{\gamma_{r,s}, r = 1, \dots, R, s = 1, \dots, S\} \to \sum_{r=1}^{R}\sum_{s=1}^{S} \gamma_{r,s}\, \hat{c}_{n_A,n_B}(p_r, q_s),$$

under the following constraints:

$$\begin{cases} \sum_{r=1}^{R} \gamma_{r,s} = (\hat{b}_{n_B})_s, & \forall s = 1, \dots S \\ \sum_{s=1}^{S} \gamma_{r,s} = (\hat{a}_{n_A})_r, & \forall r = 1, \dots R \\ \gamma_{r,s} \geq 0, & \forall r = 1, \dots R, \forall s = 1, \dots S \end{cases}$$

with, for any $r = 1, \dots, R$ and $s = 1, \dots S$,

$$\begin{aligned} \hat{c}_{n_A,n_B}(p_r, q_s) &= \frac{1}{\kappa_{r,s}} \sum_{i=1}^{n_A}\sum_{j=1}^{n_B} d(\mathbf{X}^A_i, \mathbf{X}^B_j) \mathbb{I}_{\{Y^A_i = p_r\,,\, Y^B_j = q_s\}} \quad \text{if } \kappa_{r,s} \neq 0 \\ &= 0 \qquad \text{otherwise} \end{aligned}$$

and $\kappa_{r,s} = \sum_{i=1}^{n_A}\sum_{j=1}^{n_B} \mathbb{I}_{\{Y^A_i = p_r\,,\, Y^B_j = q_s\}}$.

Notice that Assumption 1 insures that the estimators (3) and (4) are unbiased and Assumption 2 makes the introduction of the estimated cost relevant.

The proposed OT-algorithm splits in two parts.

Step 1. Estimation $\hat{\gamma}^{opt}$ of the optimal joint distribution of (Y^A, Y^B)

- Compute $(\hat{a}_{n_A})_r$'s and $(\hat{b}_{n_B})_s$'s the empirical distributions of μ and ν given by (3) and (4).
- Compute the matrix of distances between each pair of patients of database A and database B.
- Compute the matrix of costs for each pair of modalities (p_r, q_s) thanks to eq. (5).
- Solve the Hitchcock's problem defining $\hat{\gamma}^{opt}_{n_A,n_B}$.

Step 2. Affectation of a predicted value $\hat{Y}^B$ for each patient of database A comes from a nearest neighbor algorithm accounting for a distance constructed from covariates

- Compute, for any $r = 1,\dots, R$ and $s = 1,\dots, S$:

$$N_{r,s} = Ent(n_A \times \hat{\gamma}_{r,s})$$

  where $Ent(x)$ denote the integer part of x. $N_{r,s}$ stands for the number of subjects having modality p_r for Y^A and q_s for Y^B in database A.

- Consider, for any r and any s:

$$\begin{aligned} \mathcal{N}_{r,s} &= \left\{(i,j) | y^A_i = p_r, y^B_j = q_s\right\} \\ \mathcal{N}_r &= \cup_{s=1}^{S} \mathcal{N}_{r,s} = \left\{i | y^A_i = p_r\right\} \end{aligned}$$

  - Consider $(\tilde{r}, \tilde{s}) = \text{argmax}_{r,s}\, \mathcal{N}_{r,s}$
  - For any $i \in \mathcal{N}_{\tilde{r}}$,
    - if $\text{card}(\mathcal{N}_{\tilde{r}}) \leq \mathcal{N}_{(\tilde{r},\tilde{s})}$ then $Y^B_i = q_{\tilde{s}}$ (all the subjects are recoded in $q_{\tilde{s}}$),
    - else we have to identify which patients in $\mathcal{N}_{\tilde{r}}$ will be recoded in $q_{\tilde{s}}$. The patients selected are the ones closer to this modality in terms of average distance to modality $q_{\tilde{s}}$ defined as:

$$c_i(p_{\tilde{r}}, q_{\tilde{s}}) = \frac{1}{\sum_{j=1}^{n_B} \mathbb{I}_{\{y^B_j = q_{\tilde{s}}\}}} \sum_{j=1}^{n_B} d(\mathbf{X}^A_i, \mathbf{X}^B_j)\, \mathbb{I}\left(y^B_j = q_{\tilde{s}}\right),$$

    - Remove patient that has been recoded at this step and repeat the procedure,
  - Removed patients of modality $(p_{\tilde{r}}, q_{\tilde{s}})$ and repeat the procedure.

4 Simulation studies

In this section the average accuracy of the algorithm defined in Section 3 are assessed by means of simulation studies. Database A of size n_A and database B of size n_B are constructed by $n_A + n_B$ random generations of the P covariates according to predefined distributions. Denote parameter $F = n_B/n_A$, the ratio between the sizes of the two databases. The construction of variables Y^A and Y^B for the $n_A + n_B$ patients depends on the generation plan. The values of Y^B for patients 1 to n_A and the values of Y^A for patients $n_A + 1$ to $n_A + n_B$ allow us to assess the average accuracies of the algorithm defined in Section 3 by comparing these values to the predicted ones $\hat{y}^B$ in database A (resp. $\hat{y}^A$ in database B).

4.1 Performance of the OT-algorithm : effect of sample size

4.1.1 Simulation design

The Optimal Transportation algorithm is based on estimated values of the parameters of the distributions of Y^A and Y^B. Obviously, the sizes of the databases are thus parameters of potential importance in the average accuracies of the algorithm. In order to investigate this question, a simulation study is performed by considering a deterministic construction of variables Y^A and Y^B. As our attention focuses on the databases sample size, P is fixed to two covariates (X^1, X^2). To construct (X^1, X^2), consider (C^1, C^2) a two-dimensional Gaussian distribution with mean (0, 0), $\text{cor}(C^1, C^2) = 0.2$, $\text{var}(C^1) = \text{var}(C^2) = 1$. X^1 is the discretization of C^1 in two modalities and is so Bernoulli-distributed $B(\pi_1)$ with $\pi_1 = 0.4$, X^2 is the discretization of C^2 in two modalities and is so Bernoulli-distributed $B(\pi_2)$ with $\pi_2 = 0.3$. The construction of Y_i^A and Y_i^B for any patient i, is defined by the following rules, which endows Y^A and Y^B with three and four modalities respectively:

$$\begin{aligned}
&\text{If } X_i^1 = 1 \text{ and } X_i^2 = 1 \quad \text{then } Y_i^A = 3 \text{ and } Y_i^B = 4,\\
&\text{If } X_i^1 = 1 \text{ and } X_i^2 = 0 \quad \text{then } Y_i^A = 2 \text{ and } Y_i^B = 3,\\
&\text{If } X_i^1 = 0 \text{ and } X_i^2 = 1 \quad \text{then } Y_i^A = 3 \text{ and } Y_i^B = 2,\\
&\text{If } X_i^1 = 0 \text{ and } X_i^2 = 0 \quad \text{then } Y_i^A = 1 \text{ and } Y_i^B = 1.
\end{aligned}$$

4.1.2 Simulation scenarios

In order to investigate the role of sample sizes n_A and n_B, different scenarios are considered. First, the ratio F is fixed as 1 (well-balanced scenarios) and n_A varies over {50, 100, 500, 1000, 5000}. Second, the size n_A is fixed as 1000 and F varies over {0.25, 0.5, 0.75} (unbalanced scenarios).

4.1.3 Methods

For the cost function involved in OT-algorithm, as categorical covariates have been considered, Hamming distance has been used.

4.1.4 Results

The assessment of the average accuracy of the OT -algorithm is evaluated by means of the parameter Perf(OT), the Average Prediction Accuracy, defined as:

$$Perf(OT) = \frac{1}{n_A}\sum_{i=1}^{n_A} \mathbb{I}_{\{\hat{y}_i^B = y_i^B\}} + \frac{1}{n_B}\sum_{i=1}^{n_B} \mathbb{I}_{\{\hat{y}_i^A = y_i^A\}} \tag{5}$$

where $\hat{y}^B$ and $\hat{y}^A$ are the predicted values from the OT-algorithm.

The results for well-balanced scenarios and results for unbalanced scenarios are collected in Table 2. The results are expressed in terms of mean over 100 independent runs of the algorithm together with the corresponding standard errors.

Table 2: Assessment of the effect of sample size on the average accuracy of the OT-algorithm from deterministic databases (mean ± standard error over 100 independent simulations runs). On the left, Well-balanced scenarios, varying n_A. On the right, unbalanced scenarios varying F for n_A fixed to 1000.

n_A	Perf(OT)	F	Perf(OT)
50	0.89 ± 0.06	0.25	0.95 ± 0.02
100	0.92 ± 0.04	0.50	0.96 ± 0.02
500	0.96 ± 0.02	0.75	0.97 ± 0.01
1000	0.97 ± 0.01		
5000	0.99 ± 0.01		

From Table 2, the average accuracy of the OT-algorithm increases as the sample size n_A and the ratio F increases. The average accuracies exceed more than 89 % in all considerated scenarios. The OT-algorithm gives better average accuracy in a well-balanced design than in an unbalanced context. The OT-algorithm demonstrates acceptable average accuracy in this deterministic context. Since we consider an estimation problem, this is not surprising: the larger the sample size (n_A and n_B) is, the better the quality of the estimates is.

4.2 Performance of the OT-algorithm: effect of association between covariates and outcome

4.2.1 Simulation design

By construction, the average accuracies of the OT-algorithm are linked to the dependence of Y^A and Y^B with the covariates. This second simulation study highlights the link between those average accuracies and the main parameters which depend on the generated databases. To do so, a more complicated simulation design is considered involving $P = 3$ covariates (X^1, X^2, X^3). Those covariates are constructed from (C^1, C^2, C^3), a three-dimensional $\mathcal{N}((0,0,0);\Sigma)$ Gaussian distribution with:

$$\Sigma = \begin{pmatrix} 1 & \rho & \delta \\ \rho & 1 & \mu \\ \delta & \mu & 1 \end{pmatrix}.$$

X^1 is the discretization of C^1 in two modalities in order to be $B(\pi_1)$ Bernoulli-distributed. $X^1 = \mathbb{I}_{\{C^1 > t_1\}}$ where t_1 is chosen such as $\pi_1 = \mathbb{P}(C^1 > t)$. X^2 is the discretization of C^2 in three modalities in order to be $\mathcal{M}(\pi_{21}, \pi_{22})$ multinomially-distributed. $X^2 = \mathbb{I}_{\{t_{21} < C^2 < t_{22}\}} + \mathbb{I}_{\{C^2 > t_{22}\}}$ where t_{21} and t_{22} is chosen such as $\pi_{21} = \mathbb{P}(t_{21} < C^2 < t_{22})$ and $\pi_{22} = \mathbb{P}(C^2 > t_{22})$. Finally, $X^3 = C^3$ and is normally-distributed.

The construction of y_i^A and y_i^B for any patient i, is defined by the following rules including an error term on the determination of Y^A and Y^B. Consider Y to be a continuous outcome defined by:

$$Y = C^1 + C^2 + C^3 + \sigma U,$$

with U following a standard normal distribution. Y^A is the discretization of Y by quartiles in database A and Y^B is the discretization of Y by tertiles in database B.

The data observed are covariates (X^1, X^2, X^3), Y^A for n_A subjects in database A and Y^B for n_B subjects in database B.

Scenarios consists in choosing values for parameters $\rho, \delta, \mu, \pi_1, \pi_{21}, \pi_{22}, \sigma$. Parameters $\rho, \delta, \mu, \sigma$ are related to the parameter R^2 which measures the association between covariates and the outcome and is defined as:

$$R^2 = \frac{\text{var}(C_1 + C_2 + C_3)}{\text{var}(Y)}. \tag{6}$$

$$\begin{aligned} &= \frac{\text{var}(C^1 + C^2 + C^3)}{\text{var}(C^1 + C^2 + C^3 + \sigma U)}, \\ &= \frac{3 + 2\rho + 2\delta + 2\mu}{3 + 2\rho + 2\delta + 2\mu + \sigma^2}. \end{aligned} \tag{7}$$

This relation (7) allows us to calibrate the model in order to obtain a given R^2 which appears to be the parameter of paramount importance for the relevancy of the algorithm.

4.2.2 Simulation scenarios

In order to assess the average accuracies of the algorithm as a function of the sample size n_A, the correlation between the three covariates Σ, the association measure between the covariates and the outcome R^2, different scenarios are considered:

- Scenarios (Sn) investigate the effect of the sample size n_A by fixing $F = 1$, $R^2 = 0.5$, $\rho = \delta = \mu = 0.2$, $\pi_1 = 0.5$, $\pi_{21} = \pi_{22} = 0.3$ and varying $n_A \in \{50, 100, 500, 1000, 5000\}$.
- Scenarios (SF) investigate the effect of the ratio F between the sample sizes of the datasets A and B by fixing $n_A = 1000$, $R^2 = 0.5$, $\rho = \delta = \mu = 0.2$, $\pi_1 = 0.5$, $\pi_{21} = \pi_{22} = 0.3$ and varying F in $\{0.25, 0.5, 0.75, 1\}$.
- Scenarios (SR) investigate the effect of R^2 by fixing $n_A = 1000$, $F = 1$, $\rho = \delta = \mu = 0.2$, $\pi_1 = 0.5$, $\pi_{21} = \pi_{22} = 0.3$ and varying R^2 in $\{0.2, 0.4, 0.6, 0.8\}$.
- Scenarios (Sρ) investigate the effect of ρ by fixing $n_A = 1000$, $F = 1$, $R^2 = 0.5$, $\delta = \mu = 0.2$, $\pi_1 = 0.5$, $\pi_{21} = \pi_{22} = 0.3$ and varying ρ in $\{0.2, 0.4, 0.6, 0.8\}$.

4.2.3 Method

As discussed in the Introduction, the following methods have been selected as comparison methods with OT-algorithm:

1. Among the supervised learning methods, polytomic regression (PR) has been chosen and supposed to fit 2 different models :
   - one model for outcome Y^A adjusted on covariates $\mathbf{X}$. Parameters are estimated using individuals in databases A than predict Y^A for individuals in databases B.
   - one model for outcome Y^B on the same covariates $\mathbf{X}$. Parameters are estimated using individuals in databases B than predict Y^B for individuals in A.
2. *Among imputation models*, MICE has been selected. The algorithm generates multiple imputations for incomplete datasets by itering conditional densities using Gibbs sampling. For a given outcome, all other columns in the database were included as the default set of predictors to make the results comparable to those obtained with the OT-algorithm. Five imputed datasets were generated and the pooled results were retained to impute the appropriate targets. The structural parts of the imputation models and the error distributions have been specified according to the types of the covariates: we used the Predictive Mean Matching (pmm) method when the covariates were continuous and the polytomous regression method when the covariates were categorical.

For the cost function involved in OT-algorithm, as mixed covariates have been considered, Euclidian distance has been applied on the coordinates extracted from a factor analysis of mixed data.

Notice that the results are obtained by R version 3.2.5 and especially the packages '`MICE`' for multiple imputation by chained equation [28], '`FactoMineR`' for factor analysis of mixed data [29], '`linprog`' for simplex algorithm and '`MASS`' for polytomic regression.

4.2.4 Results

The assessment of the average accuracy of the different algorithms (MICE, PR and OT) is assessed by means of the following indicators:

- Average accuracy of method m noted Perf(m) defined by:

$$\text{Perf}(m) = \frac{1}{n_A}\sum_{i=1}^{n_A} \mathbb{I}((\hat{y}_i^B)^m = y_i^B) + \frac{1}{n_B}\sum_{i=1}^{n_B} \mathbb{I}((\hat{y}_i^A)^m = y_i^A). \tag{8}$$

  where $(\hat{y}^B)^m$ and $(\hat{y}^A)^m$ are the predicted values from the algorithm m.

– $\text{Conc}(m^1, m^2)$ defined as:

$$\text{Conc}(m^1, m^2) = \frac{1}{n_A}\sum_{i=1}^{n_A} \mathbb{I}((\hat{y}_i^B)^{m^1} = (\tilde{y}_i^B))^{m^2} + \frac{1}{n_B}\sum_{i=1}^{n_B} \mathbb{I}((\hat{y}_i^A)^{m^1} = (\tilde{y}_i^A)^{m^2}). \quad (9)$$

evaluates the concordance of predicted values between both algorithms.

The main results for simulation studies with scenarios (Sn), resp. (SF), (SR) and (Sρ) are summarized in Figure 1, resp. Figure 2, Figure 3 and Figure 4 which are the plots of the average (over the 100 simulation runs) of Perf(OT), Perf(MICE) and Perf(PR) over the coefficient n_A (resp. F, R^2 and ρ).

The results for $\text{Conc}(m^1, m^2)$ for scenarios (Sn), (SF), (SR) and (Sρ) are collected in Table 3. The results are expressed in terms of mean over 100 independent runs of the algorithm together with the standard error of the different indicators defined above.

Table 3: Estimation of the average accuracy of OT, MICE and PR algorithms together with concordance criteria. (mean ± stardard error over 100 independent simulation runs).

(a) Scenarios (Sn) varying n_A and fixing $F = 1$, $R^2 = 0.5$, $\rho = \delta = \mu = 0.2$, $\pi_1 = 0.5$, $\pi_{21} = \pi_{21} = 0.3$**.**

n^A	Conc(OT,MICE)	Conc(OT,PR)
50	0.50 ± 0.10	0.52 ± 0.06
100	0.48 ± 0.09	0.52 ± 0.04
500	0.48 ± 0.04	0.54 ± 0.02
1000	0.48 ± 0.03	0.53 ± 0.02
5000	0.48 ± 0.02	0.53 ± 0.01
10000	0.48 ± 0.01	0.53 ± 0.01

(b) Scenarios (**SF**) varying F and fixing $n_A = 1000$, $R^2 = 0.5$, $\rho = \delta = \mu = 0.2$, $\pi_1 = 0.5$, $\pi_{21} = \pi_{21} = 0.3$.

F	Conc(OT,MICE)	Conc(OT,PR)
0.25	0.52 ± 0.04	0.56 ± 0.02
0.5	0.49 ± 0.03	0.54 ± 0.02
0.75	0.49 ± 0.03	0.54 ± 0.02
1	0.48 ± 0.03	0.53 ± 0.02

(c) Scenarios (**SR**) by varying R^2 and fixing $n_A = 1000$, $F = 1$, $\rho = \delta = \mu = 0.2$, $\pi_1 = 0.5$, $\pi_{21} = \pi_{21} = 0.3$.

R^2	Conc(OT,MICE)	Conc(OT,PR)
0.2	0.36 ± 0.03	0.43 ± 0.02
0.4	0.44 ± 0.03	0.50 ± 0.02
0.6	0.51 ± 0.02	0.57 ± 0.02
0.8	0.58 ± 0.02	0.63 ± 0.02

(d) Scenarios (**Sρ**) by varying ρ and fixing $n_A = 1000$, $F = 1$, $R^2 = 0.5$, $\delta = \mu = 0.2$, $\pi_1 = 0.5$, $\pi_{21} = \pi_{21} = 0.3$.

ρ	Conc(OT,MICE)	Conc(OT,PR)
0.2	0.48 ± 0.02	0.53 ± 0.02
0.4	0.49 ± 0.03	0.55 ± 0.02
0.6	0.52 ± 0.03	0.60 ± 0.02
0.8	0.54 ± 0.02	0.61 ± 0.02

From Figure 1, the average accuracy of prediction of OT, MICE and PR algorithms, increases as the sample size n_A increases in well-balanced design situations. The OT-algorithm always provides better average accuracies (>66 %) than those obtained with the MICE algorithm and PR (<51 %). When the sample size is too small (less than 500), the average accuracies of all algorithms are unstable and reaches stability when n_A is greater than 500. Multiplying the sample size by 100 (from $n_A = 50$ to $n_A = 500$), generates a higher average accuracy gain for the OT-algorithm (10 %) than for the MICE algorithm (only 4 %) and PR (only 4 %). From Table 3(a), the concordance rates between MICE and OT stays low (a little more than 50 %) whatever the considered scenario and remains stable when the sample size n varies.

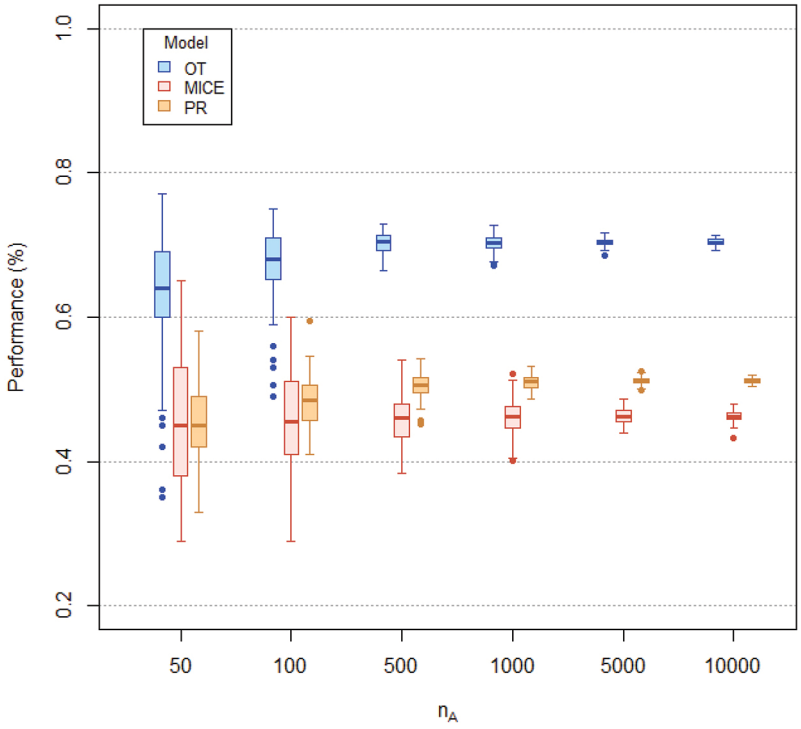

Figure 1: Boxplot of average accuracy distribution for the three methods (OT, MICE and PR) on non determinist data. $F = 1$, $R^2 = 0.5$, $\rho = \delta = \mu = 0.2$, $\pi_1 = 0.5$, $\pi_{21} = \pi_{22} = 0.3$, varying n_A.

From Figure 2, the average accuracy of prediction of OT, MICE algorithms and PR decreases as the ratio F increases (6 % decrease with OT, 5 % with MICE, 5 % with PR when F varies from 1 to 0.25). From Table 3(b), the concordance rates between MICE and OT and PR and OT stays low (a little less than 50 % for MICE and a little less than 50 % for PR) in each case but is stable across values of the ratio F.

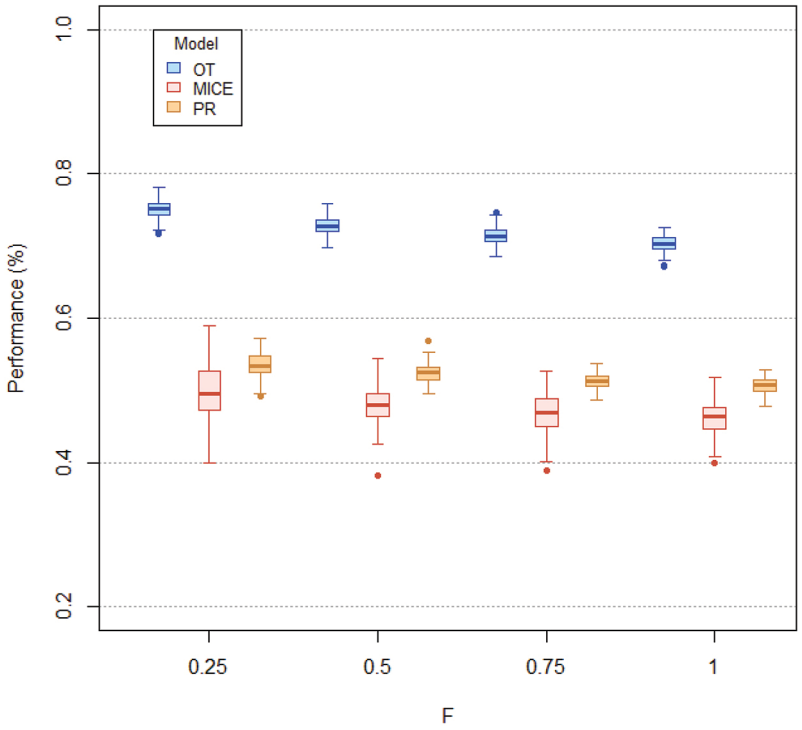

Figure 2: Boxplot of average accuracy distribution for the three methods (OT, MICE and PR) on non determinist data. $n_A = 1000$, $R^2 = 0.5$, $\rho = \delta = \mu = 0.2$, $\pi_1 = 0.5$, $\pi_{21} = \pi_{22} = 0.3$, varying F.

According to Figure 3, the average accuracy of prediction of OT and MICE algorithms decreases as the R^2 increases, and the covariates better predict the outcome (4 % increase with OT, 62 % increase with MICE and 53 % increase with MICE when R^2 varies from 0.2 to 0.8). This gives opposite results than those observed in the determinist context but is coherent with the construction of the OT-algorithm. We can notice that the MICE and PR mean tends to approximate the OT average accuracy curve. From Table 3(c), the concordance rates between the three algorithms increases as R^2 increases. When the R^2 criterion is close to 0.8, the average accuracies are very close to those obtained in the deterministic context, because the covariates explain a large part of the variability of the outcome.

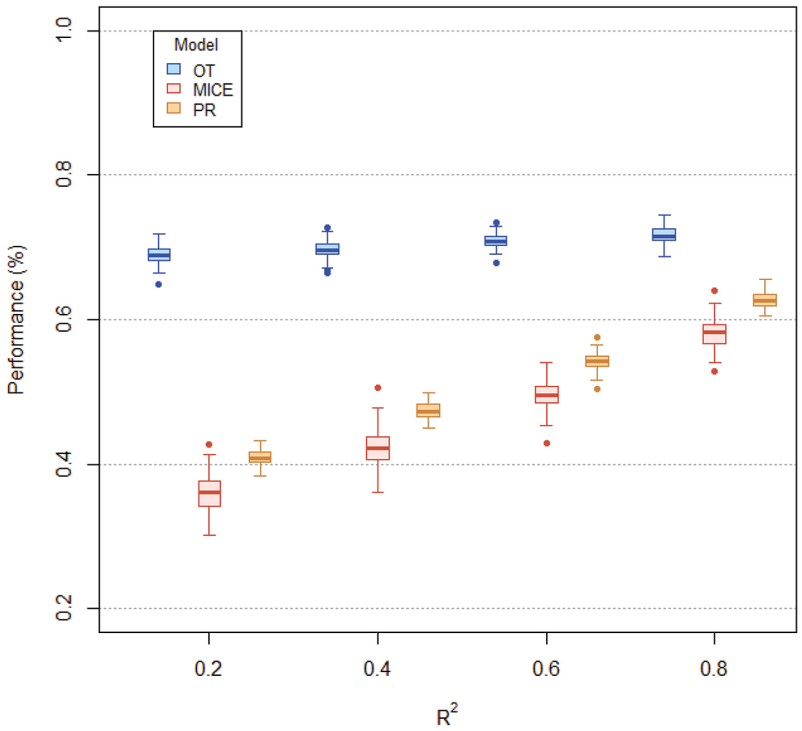

Figure 3: Boxplot of average accuracy distribution for the three methods (OT, MICE and PR) on non determinist data. $n_A = 1000$, $F = 1$, $\rho = \delta = \mu = 0.2$, $\pi_1 = 0.5$, $\pi_{21} = \pi_{22} = 0.3$, varying R^2.

From Figure 4, the average accuracy of prediction of the OT and MICE algorithms and PR remain stable as the ratio ρ increases. The variation of correlation between covariates does not influence the average accuracy whatever the used algorithm. From Table 3(d), the concordance rates between MICE and OT and PR and OT stay low (a little more than 50 %) in each case but remain stable as the coefficient of correlation ρ varies.

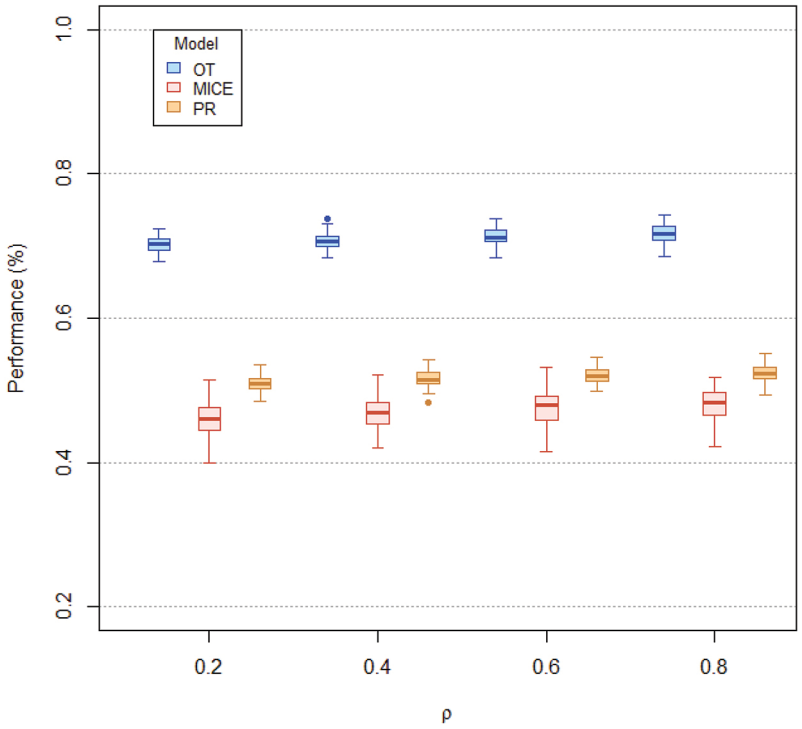

Figure 4: Boxplot of average accuracy distribution for the three methods (OT, MICE and PR) on non determinist data. $n_A = 1000$, $F = 1$, $R^2 = 0.5$, $\delta = \mu = 0.2$, $\pi_1 = 0.5$, $\pi_{21} = \pi_{22} = 0.3$, varying ρ.

To conclude, in each table, the standard errors of average accuracy of the OT and MICE algorithms and PR remain stable. The OT-algorithm demonstrates a better average accuracy than the MICE algorithm and PR overall. It always gives good predictions for more than 66 % of the simulated data in each scenario. Notice that "overlapping issues", classical problem in classification, which appears when the values of the covariates is the same for two different subjects and the value of outcomes are different. This explain the 20 % of subjects which are not well classified in the best situation $R^2 = 0.8$ and $n = 1000$.

5 ELFE database: application to a real-life dataset

The ELFE (Etude Longitudinale Francaise depuis l'Enfance) project is a nationally representative french cohort started in 2011, included more than 18 000 children, followed from birth. The aim is to explain how various contextual factors (such as perinatal conditions and environment) affect children's developmental health and well-being over time, and into adulthood. During the first baseline data collection wave (between January and April 2011), the mother's health status of the participating children was collected using a question ("How would you rate your overall health") `MHS` containing categories on a five point ordinal scale: "excellent", "very well", "well", "fair", "bad" which corresponds to the standard scale used in French Cohorts. However, during the second baseline data collection wave (May to December 2011), the health state of the mother `MHS` was collected using the same question containing categories on a different five point ordinal scale: "very well", "well", "medium" "bad" and "very bad" , the standard scale used currently in many European cohorts (see Table 4 for details).

Table 4: ELFE study. Description of the modalities of the outcome `MHS` at each wave.

MHS	**First wave**		**Second wave**	
Modality	**Coding**	**Number (%)**	**Coding**	**Number (%)**
1	"excellent"	950 (42.54)	"very well"	1834 (16.20)

2	"very well"	1047 (46.89)	"well"	4374 (38.64)
3	"well"	212 (9.49)	"medium"	4586 (40.51)
4	"fair"	22 (0.99)	"bad"	478 (4.22)
5	"bad"	2 (0.00)	"very bad"	49 (0.43)

In order to unify the database by means of a recoding of variable MHS by OT-algorithm the data of the first wave is consider as database A ($n_A = 2233$) and data of the second wave is consider as database B ($n_B = 11321$). Three covariates coded in the same way in both databases are selected for their ability to predict the outcomes:

- AGE (continuous): the mother's age at baby birth in years.
- PL (categorical with six modalities): the health state of the mother and her physical limitations reported for a duration of at least six months.
- CMH (categorical with three modalities): the chronic mother health problem at two months of baby age.
- MGH (categorical with five modalities): the mother' general health

As mixed covariates has been considered, the cost function involved in OT-algorithm is based on the Euclidian distance applied on the coordinates extracted from a factor analysis of mixed data.

The association between the outcome and the covariates are tested independently in each dataset by using standard chi-square tests of independence for categorical covariates and student tests for continuous covariates. Each obtained p-value is less than 10^{-14}. The same results hold by ascending inclusion in an ordered logistic regression.

Table 5 do not show any significant difference between covariates distribution at wave 1 and wave 2 except age. Assumption 1 is thus realistic.

Table 5: ELFE study. Description of covariates at each wave. Modalities together with the numbers at each wave (%) for each categorical covariates and mean ± standard error for continuous covariate AGE. Comparison of the distribution for each covariate by means of an adequate test. The modalities for the MHS variable are not the same at wave 1 and wave 2.

Covariate	Modalities	Wave 1	Wave 2	p-value
MGH	1	1047 (46.89)	5238 (46.27)	0.22
	2	1002 (44.87)	5159 (45.57)	
	3	170 (7.61)	861 (7.61)	
	4	12 (0.54)	58 (0.51)	
	5	2 (0.09)	5 (0.04)	
PL	Severely limited	18 (0.81)	64 (0.57)	0.20
	Limited	140(6.27)	657 (5.80)	
	No	2075 (92.92)	10600 (93.63)	
CMH	Yes	285 (12.76)	1433 (12.66)	0.99
	No	1948 (87.24)	9888 (87.34)	
AGE		30.77 ± 4.68	31.10 ± 4.80	0.002

Table 6 gives the coupling distribution given by the first step of OT-algorithm. The results of recoding of MHS in database A and database B by the OT-algorithm are given in terms of confusion matrix between the two completed scales which is the matrix G where $G_{i,j}$ denotes the number of individuals coded i in the European and recoded j in the French coding. G is presented in Table 7. The tridiagonal structure observed for this matrix reflects a good re-allocation of the values from one outcome to another. The values on the diagonal and on the first lower diagonal represents 89.2 % of the recoding.

Table 6: ELFE study results. Coupling $\hat{\gamma}$ distribution. In rows, European coding, in columns, French coding.

	"very well"	"well"	"medium"	"bad"	"very bad"
"excellent"	0.162	0	0	0	0
"very well"	0.263	0.123	0	0	0
"well"	0	0.346	0.059	0	0
"fair"	0	0	0.036	0.006	0.001
"bad"	0	0	0	0.004	0

Table 7: ELFE study results. Confusion matrix of the recoding by means of the OT-algorithm (number (%)). In rows, European coding, in columns, French coding.

	"very well"	**"well"**	**"medium"**	**"bad"**	**"very bad"**
"excellent"	2196 (16.2)	588 (4.3)	0 (0)	0 (0)	0 (0)
"very well"	2982 (22.0)	1666 (12.3)	773 (5.7)	0 (0)	0 (0)
"well"	0 (0)	3917 (28.9)	801 (5.9)	80 (0.6)	0 (0)
"fair"	0 (0)	0 (0)	405 (3.0)	75 (0.6)	20 (0.1)
"bad"	0 (0)	0 (0)	0 (0)	51 (0.4)	0 (0)

6 Results and discussion

In this paper, OT-algorithm is introduced. That algorithm aims to recode variables. Variable recoding is a usual issue which appears when a variable is not coded on the same scale in two different databases while merging or at two different times while comparing. OT-algorithm splits in two steps. The first step is based on optimal transportation theory specifying the optimal numbers of transitions from a scale to another and a second step, an allocation rule, based on average distance between covariates.

OT-algorithm is based on two assumptions:

– First, the distribution of the variable of interest is the same in both databases. This assumption is realistic when merging databases from two waves of recruitment but has limitations when merging two cohorts for example from different countries. This has already been studied in North American NHANES study and the French National Health Survey. The distribution of the outcome "self-rated health" is not distributed identically in the two databases. Poor self-rated health is more frequently reported in France [30].

– Second, the covariates explains the outcome in the same way in both databases. This assumption cannot be evaluated from data but example of situation where this assumption is not acceptable are numerous. For example in [30] a comparison of the outcomes "functional limitations" and "self-rated health" in these shows that "functional limitation" is more strongly associated with "poor self-rated health" for the most educated men than in the least educated in US rather than in France.

The average accuracies of OT-algorithm has been assessed by simulations studies. The results show that the method works very well. The average accuracies depend on the sample size of the databases and of the intensity of the link between covariates and the outcome of interest (essessed by R-square parameter). In any situation, OT-algorithm is more accurate than a multiple imputation algorithm.

OT-algorithm has been applied to recode a variable on real dataset where the scales of coding are different at two different times. This investigation shows the average accuracy of the OT-algorithm for practical use.

Acknowledgements

We would like to thank the scientific coordinators, IT and data managers, statisticians, administrative and family communication staff, study technicians of the ELFE coordination team, and the families who gave their time for the study. The ELFE survey is a joint project between INED (Institut National d'Etudes Démographiques), INSERM (Institut National de la Santé et de la Recherche Médicale), EFS (Etablissement Francais du Sang), InVS (Institut de Veille Sanitaire), INSEE (Institut National de la Statistique et des Etudes Economiques), the Ministry of Health (DGS, Direction Générale de la Santé), the Ministry of Environment (DGPR, Direction Générale de la Prévention des Risques), the Ministry of Health and Employment (DREES, Direction de la Recherche, des Etudes, de l'Evaluation et des Statistiques), and the CNAF (Caisse Nationale des Allocations Familiales), with the support of the Ministry of Research and CCDSHS (Comité de Concertation pour les Données en Sciences Humaines et Sociales) and the Ministry of Culture (DEPS, Département des études, de la prospective et des statistiques).

Funding

This research has received the help from "Région Occitanie" Grant RBIO-2015-14054319 and Mastodons-CNRS Grant.

References

[1] Bloch I. Fusion d'informations en traitement du signal et des images. France: Hermes Science Publication. 2003
[2] Hall D, Llinas J. An introduction to multisensor data fusion. Proc. IEEE. 1997;85:6–23.
[3] Abidi M, Gonzalez R. Data fusion in robotics and machine intelligence. United States: Academic Press. 1992
[4] Smyth P, Heckerman D, Jordan M. Probabilistic independance networks for hidden markov probability models. Technical Report MSR-TR-96-03, Microsoft Research, 1996.
[5] Rabiner L. A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE. 1989;77:257–85.
[6] Haton J, Charpillet F, Haton M. Numeric/symbolic approaches to data and information fusion. Proceedings of the International Conference on Multisource-Multisensor Information Fusion - Fusion 1998, II, 1998:888–95.
[7] Gebhardt J, Kruse R. Information source modelling for consistent data fusion. Proceedings of the International Conference on Multisource-Multisensor Information Fusion - Fusion 1998, I, 1998:27–34.
[8] Duda R, Hart P. Pattern classification and scene analysis. New York/Chichester/Brisbane/Toronto/Singapore: A Wiley interscience publication ed. 1973
[9] Xu L, Krzyzak A, Suen C. Methods of combining multiple classifiers and their application to handwriting recognition. IEEE Trans Syst. Man Cybern: A Wiley interscience publication ed. 1992;22:418–35.
[10] Vandentorren S, Bois C, Pirus C, Sarter H, Salines G, Leridon H, et al. Rationales, design and recruitment for the Elfe longitudinal study. BMC Pediatr. 2009;9:58.
[11] Okner BA. Constructing a new microdata base from existing microdata sets: the 1966 merge file. Ann Econ Soc Meas. 1972;1:325–62.
[12] Rässler S. Statistical matching. Lecture notes in statistics, vol. 168. New York: Springer-Verlag, 2002. DOI: 10.1007/978-1-4613-0053-3. A frequentist theory, practical applications, and alternative Bayesian approaches.
[13] D'Orazio M, Di Zio M, Scanu M. Statistical matching. Wiley Series in Survey Methodology. Chichester: John Wiley & Sons, Ltd., 2006. DOI: 10.1002/0470023554, theory and practice.
[14] Little R, Rubin D. Statistical analysis with missing data. NY: Wiley, 1987.
[15] van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Stat Meth Med Res. 2007;16:219–42. DOI: 10.1177/0962280206074463.
[16] Kotsiantis SB. Supervised machine learning: a review of classification techniques. Informatica (Ljubl.). 2007;31:249–68.
[17] Vapnik VN. The nature of statistical learning theory. New York: Springer-Verlag, 1995. DOI: 10.1007/978-1-4757-2440-0.
[18] Bartholomew D, Knott M, Moustaki I. Latent variable models and factor analysis: a unified approach. United States: Wiley, 3rd ed. 2011
[19] Skrondal A, Rabe-Hesketh S. Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models, Chapman and Hall/CRC ed. UK: 2004
[20] Villani C. Optimal transport, old and new. Grundlehren des mathematischen Wissenschaften. France: Springer-Verlag. 2009:338
[21] Engel J. Polytomous logistic regression. Stat. Neerlandica. 1988;42:233–52. DOI: 10.1111/j.1467-9574.1988.tb01238.x.
[22] Monge G. Mémoire sur la Théorie des Déblais et des Remblais. Hist. de l'Acad. des Sciences de Paris, 1781:666–704.
[23] Kantorovich L. On the translocation of masses. J Math Sci. 2006; 133:1381–2,the original paper was published in Dokl Akad Nauk SSSR 1942;37(7-8):227–29.
[24] Hitchcock F. The distribution of a product from several sources to numerous localities. J Math Phys Mass Inst Tech. 1941;20:224–30.
[25] Aha D, Kibler D, Albert M. Instance based learning algorithms. Mach Learn. 1991;6:37–66.
[26] Stanfill C, Waltz D. Toward mempry-based reasoning. Commun ACM. 1986;29:1213–28.
[27] Pages J. Analyse factorielle multiple appliquée aux variables qualitatives et aux données mixtes. Revue de statistique appliquée. 2002;4:5–37.
[28] van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. J Stat Softw, Art. 2011;45:1–67.
[29] Lê S, Josse J, Husson F. FactoMineR: a package for multivariate analysis. J Stat Softw. 2008;25:1–18.
[30] Delpierre C, Datta GD, Kelly-Irving M, Lauwers-Cances V, Berkman LF, Lang T. What role does socio-economic position play in the link between functional limitations and self-rated health: France vs. USA? Eur J Public health. 2012;22:317–21.