On the use of optimal transportation theory to recode variables and application to database merging

Valérie Garès, Chloé Dimeglio, Grégory Guernec, Romain Fantin, Benoit Lepage, Michael R Kosorok, Nicolas Savy

On the use of optimal transportation theory to recode variables and application to database merging.

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

[1] *Université de Toulouse III - INSERM, UMR 1027 - CHU Toulouse, Toulouse, chloe.dimeglio@univ-tlse3.fr*
[2] *Univ Rennes, INSA, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France, valerie.gares@insa-rennes.fr*
[3] *INSERM, UMR 1027, Toulouse, gregory.guernec@inserm.fr*
[4] *Université de Toulouse III, Toulouse et rom.fantin@wanadoo.fr*
[5] *Université de Toulouse III - INSERM, UMR 1027 - CHU Toulouse, Toulouse, benoit.lepage@univ-tlse3.fr*
[6] *Department of Biostatistics, University of North Carolina at Chapel Hill - Chapel Hill, North Carolina, kosorok@unc.edu*
[7] *Institut de Mathématiques de Toulouse ; UMR5219 - Université de Toulouse ; CNRS - UPS IMT, F-31062 Toulouse Cedex 9, France, Nicolas.Savy@math.univ-toulouse.fr*
**both authors contribute in the same ratio in the preparation of the paper.*

Abstract

To merge 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 has demonstrated a better performance 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.

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 a strong knowledge database. This allows us to extract more information from merged database than we would obtain from using the databases separately ([4, 8]). Different techniques are already widely used to produce combinations of heterogeneous data from different sources ([1]), especially probabilistic models ([19]), the best known of which are probably the Bayes rule ([24, 6]), Hidden Markov Models ([16]), the technique of least square, multi-agent systems ([9]) and logical reasoning ([7]).

In this paper, one focuses our attention to a specific issue related to database merging, the recoding problem. Indeed, it is usual and problematic when two databases have to be merged, to observe a categorical variable that is not coded in the same scale in both databases. This problem can occur in many situations: for example, in a epidemiology survey, this can be 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, this can be a different questionnaire for asking the same information (for different subjects).

The problem can be formalized in terms of two databases A and B: the first contains the observations of $P + Q$ variables measured on n_A units, the second of the observations on a subset of P variables for 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). To make inference and analysis of the merged database, it is therefore necessary to find a common scale of evaluation. The objective is thus to complete Y^A on database B and/or complete Y^B on database A.

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

Table 1: Statement of the database merging problem.

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

This variable recoding issue could be treated as a classical missing data problem. In this context, the missingness process is considered missing at random. This problem has been widely studied in the litterature ([13]) and many existing methods for treating missing data could be used. Moreover, Y^A and Y^B refers to the same information Y which can be interpreted as a latent variable. Methods of prediction of this latent class could also be applied (class latent analysis or trait latent analysis) ([3, 18]). Finally, methods for classification learning, enable to explain, for example Y^A in database A from covariates and then to predict Y^A in database B, in a second step ([22]).

Methods 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 exploited. Indeed, assuming that the distribution of Y^A (resp. Y^B) is the same in database A and B, the theory of optimal transportation ([23]) 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. In the authors knowledge, this is the first attempt to use optimal transportation theory in this context.

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 the new algorithm based on Optimal Transportation. The assessments of the performances 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 in-

trinsic performances of the OT-algorithm. Indeed, this algorithm is based on an estimation procedure which necessitates a sufficiently large sample size 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 performance of the OT-algorithm with the correlation between covariates and outcome. The performances of the OT-algorithm are compared with multiple imputation technique ([17]). 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 ([14]).

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 ([11]) 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 ([23]) 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 ([10]). 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 equation (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:

$$\sum_{r=1}^{R}\sum_{s=1}^{S} \gamma_{r,s}\ c(p_r,q_s), \tag{3}$$

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 $\mathbf{C} = (C^1, \ldots, C^P)$ the set of P covariates observed in both databases A and B and $\mathbf{C}_i$ (resp. $\mathbf{C}_j$) the values of $\mathbf{C}$ 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 $\{p_1, \ldots, p_R\}$ (resp. $\{q_1, \ldots, q_S\}$). We denote by ind(A) $= \{1, \ldots, n_A\}$ and ind(B) $= \{1, \ldots, n_B\}$.

2.2.2 Assumptions

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

- Assumption 1 :
  - $(Y_i^A,\ i \in \text{ind(A)} \cup \text{ind(B)})$ are i.i.d with same distribution μ
  - $(Y_i^B,\ i \in \text{ind(A)} \cup \text{ind(B)})$ 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 : $(Y_i^A|C_i,\ \text{ind(A)}\cup\text{ind(B)})$ (resp. $(Y_i^B|C_i,\ \text{ind(A)}\cup\text{ind(B)})$) are i.i.d with same distribution as $Y^A|C$ (resp. $Y^B|C$).

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. Considering that the farther these patients are (in terms

of distance between covariates), the more expensive the transportation is. If assumption 2 is fulfilled then a relevant choice for the cost function is:

$$c(p_r, q_s) = \frac{1}{\kappa_{r,s}} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d(\mathbf{C}_i, \mathbf{C}_j) \, \mathbb{I}_{\left\{Y_i^A = p_r \ , \ Y_j^B = q_s\right\}}, \tag{4}$$

where $\kappa_{r,s} = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} \mathbb{I}_{\left\{Y_i^A = p_r \ , \ Y_j^B = q_s\right\}}$ and d is the distance between vectors of covariates.

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, we can use the Hamming distance from the associated complete disjunctive tables. In the case of continuous covariates, one can use directly the Euclidean or Manhattan distance. Finally, in the case of mixed covariates, we can use a distance for mixed data (e.g. the Heterogeneous Euclidean-Overlap Metric ([2]), the Value Difference Metric ([20]), or the Mahalanobis distance) or a distance for continuous covariates on the coordinates extracted from a factor analysis of mixed data ([15]).

3 Algorithm for variable recoding: OT-algorithm

The proposed OT-algorithm splits in two parts.

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

- Compute the empirical distributions of μ and ν given by the estimation $\hat{a}_r$ (resp. $\hat{b}_s$) defined as:

$$\hat{a}_r = \frac{1}{n_A} \sum_{i=1}^{n_A} \mathbb{I}_{\left\{Y_i^A = p_r\right\}}, \qquad r = 1, \dots R$$

$$\hat{b}_s = \frac{1}{n_B} \sum_{j=1}^{n_B} \mathbb{I}_{\left\{Y_j^B = q_s\right\}}, \qquad s = 1, \dots S$$

  Notice that Assumption 1 (defined in Section 2.2.2) insures that these estimators are unbiased.

- Compute the matrix of distances between each pair of patients of database A and database B (the Euclidian distance between the transformed covariates is used by default).

- Compute the matrix of costs for each pair of modalities (p_r, q_s) thanks to equation (4).

- As explained in Section 2.1.1 a solution is given by solving Hitchcock's problem statement this means a Linear programming: $\hat{\gamma}$ is the minimum of

$$\gamma = \{\gamma_{r,s}, r = 1, \ldots, R, s = 1, \ldots, S\} \rightarrow \sum_{r=1}^{R}\sum_{s=1}^{S} \gamma_{r,s}\ c\left(p_r, q_s\right)$$

under the following constraints

$$\begin{cases} \sum_{r=1}^{R} \gamma_{r,s} = \hat{b}_s, \quad \forall s = 1, \ldots S \\ \sum_{s=1}^{S} \gamma_{r,s} = \hat{a}_r, \quad \forall r = 1, \ldots R \\ \gamma_{r,s} \geq 0, \quad \forall r = 1, \ldots R, \forall s = 1, \ldots S \end{cases}$$

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 $r = 1, \ldots, R$ and $s = 1, \ldots, S$:

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

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:

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

  - 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 $Y_i^B = q_{\tilde{s}}$ (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:

$$c_i(p_{\tilde{r}}, q_{\tilde{s}}) = \frac{1}{\sum_{j=1}^{n_B} \mathbb{I}_{\left\{y_j^B = q_{\tilde{s}}\right\}}} \sum_{j=1}^{n_B} d(C_i, C_j)\ \mathbb{I}\left(y_j^B = q_{\tilde{s}}\right),$$

  - 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 performance 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 performances 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 performances 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 (C^1, C^2). To construct (C^1, C^2), consider (D^1, D^2) a two-dimensional Gaussian distribution with mean $(0, 0)$, $\text{cor}(D^1, D^2) = 0.2$, $\text{var}(D^1) = \text{var}(D^2) = 1$. C^1 is the discretization of D^1 in two modalities and is so Bernoulli-distributed $B(\pi_1)$ with $\pi_1 = 0.4$, C^2 is the discretization of D^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:

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

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 Results

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

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

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 performance 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)
50	0.89 ± 0.06
100	0.92 ± 0.04
500	0.96 ± 0.02
1000	0.97 ± 0.01
5000	0.99 ± 0.01

F	Perf(OT)
0.25	0.95 ± 0.02
0.50	0.96 ± 0.02
0.75	0.97 ± 0.01

4.1.4 Conclusions

From Table 2, the average performance of the OT method increases as the sample size n_A and the ratio F increases. The average performances exceed more than 89% in all considerated scenarios. The OT-algorithm gives better performance in a well-balanced design than in an unbalanced context. The OT method demonstrates acceptable performance 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 performances 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 performances and the main parameters which depend on the generated databases. To do so, a more complicated simulation design is considered involving $P = 3$ covariates (C^1, C^2, C^3). Those covariates are constructed from (D^1, D^2, D^3), a three-dimensional $\mathcal{N}((0,0,0); \Sigma)$ Gaussian distribution with:

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

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

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

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 (C^1, C^2, C^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:

$$\begin{aligned} R^2 &= \frac{\text{var}(D_1 + D_2 + D_3)}{\text{var}(Y)} && (6) \\ &= \frac{\text{var}(D^1 + D^2 + D^3)}{\text{var}(D^1 + D^2 + D^3 + \sigma U)}, \\ &= \frac{3 + 2\rho + 2\delta + 2\mu}{3 + 2\rho + 2\delta + 2\mu + \sigma^2}. && (7) \end{aligned}$$

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 performances 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 Results

The assessment of the performance of the OT-algorithm is assessed by means of the following indicators:

- Perf(OT) defined by (5)
- Perf(MICE) defined by

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

  where $\tilde{y}^B$ and $\tilde{y}^A$ are the predicted values from the MICE algorithm. This indicator plays the role of comparator to assess the performances of MICE (Multivariate Imputation by Chained Equations) algorithm ([21]). This algorithm generates multiple imputations for incomplete datasets by 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.

- Conc(OT, MICE) defined as

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

  evaluates the concordance of both algorithms.

Notice that the results are obtained by R version 3.2.5 and especially the following packages: *MICE* for multiple imputation by chained equation ([21]), *FactoMineR* for factor analysis of mixed data ([12]) and *linprog* for simplex algorithm.

The results 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. The main results for simulation studies with scenarios (SI) (resp. (SR)) are summarized in Figure 2 (resp. Figure 1) which are the plots of the average (over the 100 simulation runs) of Perf(OT) and Perf(MICE) over the coefficient n_A (resp. R^2).

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

(a) Scenarios (**SI**) 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	Perf(OT)	Perf(MICE)	Conc(OT, MICE)
50	0.66 ± 0.12	0.46 ± 0.10	0.51 ± 0.10
100	0.73± 0.05	0.49 ± 0.08	0.51 ± 0.09
500	0.76 ± 0.02	0.50 ± 0.03	0.51 ± 0.03
1000	0.76 ± 0.01	0.50 ± 0.03	0.51 ± 0.03
5000	0.76 ± 0.01	0.50 ± 0.01	0.51 ± 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	Perf(OT)	Perf(MICE)	Conc(OT, MICE)
0.25	0.81 ± 0.02	0.52 ± 0.04	0.54 ± 0.04
0.5	0.79 ± 0.01	0.51 ± 0.03	0.53 ± 0.04
0.75	0.78 ± 0.01	0.50 ± 0.03	0.52 ± 0.03
1	0.76 ± 0.01	0.50 ± 0.02	0.51 ± 0.03

(c) Scenarios (**SR**) and (**S**ρ) 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	Perf(OT)	Perf(MICE)	Conc(OT, MICE)
0.2	0.71 ± 0.01	0.36 ± 0.02	0.37 ± 0.03
0.4	0.75 ±0.01	0.46± 0.03	0.47 ± 0.03
0.6	0.78 ± 0.01	0.55 ± 0.02	0.56 ± 0.02
0.8	0.84 ± 0.01	0.66 ± 0.02	0.66 ± 0.02

(d) Scenarios (**S**ρ) 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	Perf(OT)	Perf(MICE)	Conc(OT, MICE)
0.2	0.77 ± 0.01	0.50 ± 0.02	0.52 ± 0.03
0.4	0.77 ± 0.01	0.51 ± 0.02	0.52 ± 0.03
0.6	0.77 ± 0.01	0.51 ± 0.02	0.52 ± 0.03
0.8	0.77 ± 0.01	0.51 ± 0.02	0.52 ± 0.03

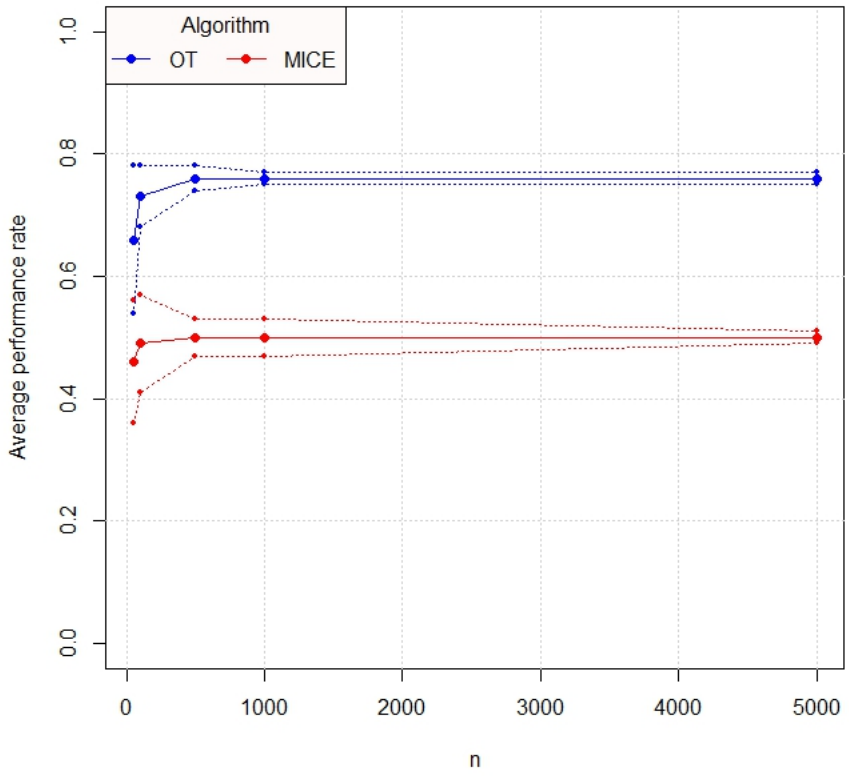

Figure 1: Performance of OT (in continuous line) and MICE (in dashed line) methods (means ± standard errors) on non determinist data $F = 1$, $R^2 = 0.5$, $\rho = \delta = \mu = 0.2$, $\pi_1 = 0.5$, $\pi_{21} = \pi_{22} = 0.3$, varying n_A.

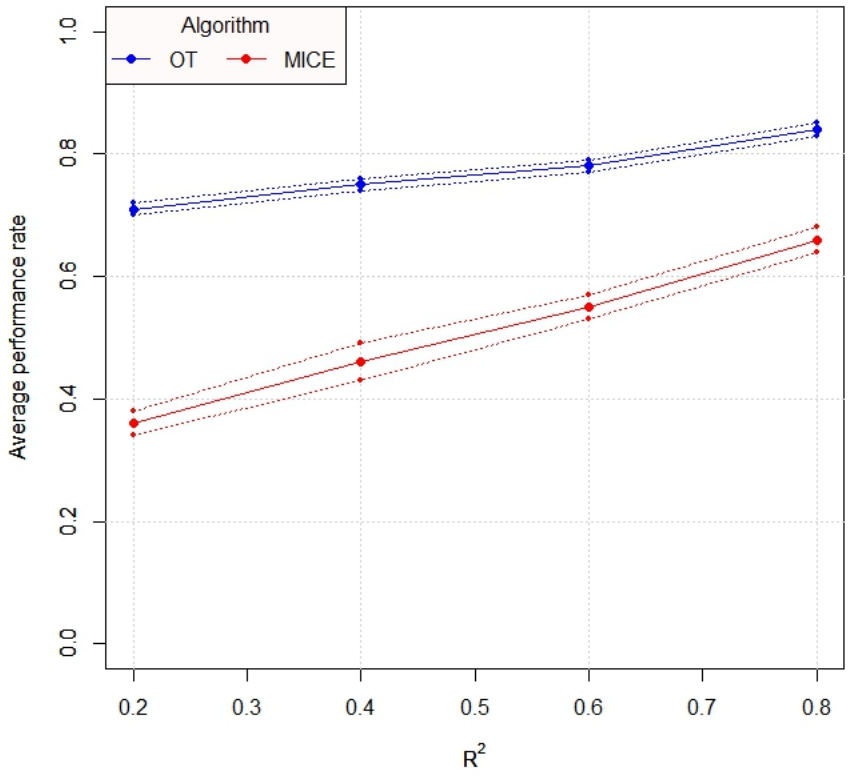

Figure 2: Performance of OT (in continuous line) and MICE (in dashed line) methods (means ± standard errors) on non determinist data $n_A = 1000$, $F = 1$, $\rho = \delta = \mu = 0.2$, $\pi_1 = 0.5$, $\pi_{21} = \pi_{22} = 0.3$, varying R^2.

4.2.4 Conclusions

From Table 3(a), the average performance of prediction of OT and MICE algorithms, increases as the sample size n_A increases in well-balanced design situations. The OT-algorithm always provides better average performances ($> 66\%$) than those obtained with the MICE algorithm ($< 51\%$). The curves in Figure 2 confirm this trend. When the sample size is too small (less than 500), the average performances of both algorithms are unstable and reaches stability when n_A is greater than 500. We can approximate this sufficient sample size to obtain reliable prediction of the average performance for the OT-algorithm in future research. Multiplying the sample size by 100 (from $n_A = 50$ to $n_A = 500$), generates a higher average performance gain for the OT-algorithm (10%) than

for the MICE algorithm (only 4%). The concordance rates between MICE and OT stays low (a little more than 50%) in each case and remains stable when the sample size n varies.

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

According to Table 3(c), the average performance of prediction of OT and MICE algorithms decreases as the R^2 increases, and the covariates better predict the outcome (13% increase with OT and 30% 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. Figure 1 shows this difference in slope between average performances when R^2 varies. We can notice that the MICE curve tends to approximate the OT performance curve. The concordance rates between the two algorithms increases as R^2 increases. When the R^2 criterion is close to 0.8, the average performances are very close to those obtained in the deterministic context, because the covariates explain a large part of the variability of the outcome.

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

To conclude, in each table, the standard errors of performance of the OT and MICE algorithms remain stable. The OT-algorithm demonstrates a better performance than the MICE algorithm 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 Application to a real-life dataset: the ELFE database

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	"passable"	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 ($n_A = 2236$) and data of the second wave is consider as database B ($n_B = 11324$). 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.

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

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
`MHS`	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

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.

The results of recoding of `MHS` in database A and database B by the OT-algorithm are given by the confusion matrix between the two completed scales and are presented in Table 6. 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.

6 Results and discussion

In this paper, OT-algorithm is introduced. That algorithm aims to recode variables. Variable recoding is an usual issues which appears when a variable is not coded on the same scale in two different databases while merging or at two

Table 6: 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"	"average"	"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)
"passable"	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)

different times while comparing. *OT*-algorithm splits in two step. 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 ([5]).
- 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 ([5]) 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 performances of OT-algorithm has been assessed by simulations studies. The results show that the method works very well. The performances 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 performance of the OT-algorithm for practical use. We have also successfully applied our methodology to a dataset on children.

References

[1] M.A Abidi and R.C. Gonzalez. *Data fusion in robotics and machine intelligence*. Academic Press, 1992.

[2] D. Aha, D. Kibler, and M. Albert. Instance Based Learning Algorithms. *Machine Learning*, 6:37–66, 1991.

[3] D.J. Bartholomew, M. Knott, and I. Moustaki. *Latent Variable Models and Factor Analysis: A Unified Approach*. (3rd ed). Wiley, 2011.

[4] I. Bloch. Fusion d'informations en traitement du signal et des images. *Hermes Science Publication*, 2003.

[5] Cyrille Delpierre, Geetanjali Dabral Datta, Michelle Kelly-Irving, Valerie Lauwers-Cances, Lisa F. Berkman, and Thierry Lang. What role does socio-economic position play in the link between functional limitations and self-rated health: France vs. usa? *European journal of public health*, 22 3:317–21, 2012.

[6] R. Duda and P. Hart. Pattern Classification and Scene Analysis. 1973.

[7] J. Gebhardt and R. Kruse. Information Source Modelling for Consistent Data Fusion. *Proceedings of the International Conference on Multisource-Multisensor Information Fusion - Fusion'98*, I:27–34, 1998.

[8] D.L. Hall and J. Llinas. An introduction to multisensor data fusion. *Proceedings of the IEEE*, 85:6–23, 1997.

[9] J. Haton, F. Charpillet, and M. Haton. Numeric/symbolic approaches to data and information fusion. *Proceedings of the International Conference on Multisource-Multisensor Information Fusion - Fusion'98*, II:888–895, 1998.

[10] F.L. Hitchcock. The distribution of a product from several sources to numerous localities. *J. Math. Phys. Mass. Inst. Tech.*, 20:224–230, 1941.

[11] L. Kantorovich. On the transfer of masses. *Dokl. Acad. Nauk. USSR*, 37:7–8, 1942.

[12] S. Lê, J. Josse, and F. Husson. FactoMineR: A Package for Multivariate Analysis. *Journal of Statistical Software*, 25(1):1–18, 2008.

[13] R.J. Little and D. Rubin. *Statistical Analysis with Missing Data*. Wiley, NY, 1987.

[14] G. Monge. Mémoire sur la Théorie des Déblais et des Remblais. *Hist. de l'Acad. des Sciences de Paris*, pages 666–704, 1781.

[15] J. Pages. Analyse factorielle multiple appliquée aux variables qualitatives et aux données mixtes. *Revue de statistique appliquée*, 4:5–37, 2002.

[16] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–285, 1989.

[17] D.B. Rubin. Multiple Imputation for Nonresponse in Surveys. *New York: Wiley & Sons*, 1987.

[18] A. Skrondal and S. Rabe-Hesketh. *Generalized Latent Variable Models.* 2005.

[19] P. Smyth, D. Heckerman, and M. Jordan. Probabilistic Independance Networks for Hidden Markov Probability Models. *Technical Report MSR-TR-96-03, Microsoft Research.*, 1996.

[20] C. Stanfill and D. Waltz. Toward mempry-based reasoning. *Communications of the ACM*, 29:1213–1228, 1986.

[21] S. van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software, Articles*, 45(3):1–67, 2011.

[22] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, second edition, New York, NY, USA, 2000.

[23] C. Villani. Optimal transport, old and new. *Grundlehren des mathematischen Wissenschaften, Springer-Verlag*, 338, 2009.

[24] L. Xu, A. Krzyzak, and C. Suen. Methods of combining multiple classifiers and their application to handwriting recognition. *IEEE Trans. on Systems, Man, and Cybernetics*, pages 418–435, 1992.