Fusions de bases de données : application sur simulations

Réunion d'Unité

Chloé Dimeglio

UMR 1027 équipe 5

18/02/2016

Outline

Plan

Projet Big Data - appel d'offre Région (T. Lang et N. Savy)

Problématiques connexes

Projet européen H2020-Lifepath, Recherche Clinique....

Partenaires

IMT-IRIT-IFERISS-University of North Carolina. Intérêts : IUCT-CHU

Tâches

1. Alignement de variables (C.Dimeglio)- **avancé**
2. Gestion des données manquantes (G.Guernec) - **avancé**
3. Intégration des données en langage naturel (IRIT) - **initié**
4. Sensibilité d'une base de données (C.Dimeglio, B.Lepage)
5. Introduction de variables latentes (C.Dimeglio, B.Lepage)
6. Réflexion autour de la temporalité (équipe 5)
7. Aspects juridiques, éthiques, sociaux (équipe 4) - **projet de thèse**

Exemple

Base A		
Sexe	Age	Activité
M	30	1
M	65	0
M	63	1
F	15	0
M	3	0
F	43	1

Base B		
Sexe	Age	Activité
M	32	5
F	28	4
F	46	8
M	68	7
M	8	8
M	11	8

- Deux bases A et B, une même variable Activité codée de deux façons différentes dans chacune des bases.
- Des covariables liées à la variable d'intérêt et communes aux deux bases.

Méthode développée

Fusionner les données sur la base d'une variable commune par du **transport de mesure**.

$$\text{Activité}_{\text{Base A}} \overset{Transport}{\rightarrow} \text{Activité}_{\text{Base B}}$$

Ambrosio, L., Brenier, Y., Buttazzo, G., Caffarelli, L., Evans, L.C., Pratelli, A. et Villani, C. (2001) : Optimal transportation and applications

Villani, C. (2012) : Topics in optimal transportation

Plan

Pré-requis

Cadre

Soit A et B deux bases.
On note X et Y la variable commune d'intérêt codée de deux façons différentes sur les deux bases, telles que :

X	x1	x2	...
P(X=xi)	a1	a2	...
Y	y1	y2	...
P(Y=yj)	b1	b2	...

On note $cov(X)$ et $cov(Y)$ les covariables d'intérêt, communes aux deux bases, associées à X et Y.

Transporter les mesures

Idée générale

On suppose que deux mesures ν et μ sont associées aux distributions des deux variables X et Y.
On cherche l'application optimale T telle que $\nu = T\mu$

Cas continu

Non développé ici...

Cas discret

Toutes les applications T telles que $T : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ sont solutions, caractérisées par les matrices de transfert de la base A vers la base B

Conséquence

Nécessité de déterminer une fonction de "coût" de passage d'une échelle à l'autre pour définir l'optimalité de l'application.

Plan

Modélisation

Mesures et matrice de permutation

• Soit $\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$ la mesure associée à la base A et $\nu = \sum_{j=1}^{m} b_j \delta_{y_j}$ celle associée à la base B.

• Les plans de transfert sont alors les matrices de permutation γ telles que :

$$\gamma = \sum_{i,j} \gamma_{i,j} \delta_{(x_i, y_j)}$$

Où :

$$\sum_{j} \gamma_{i,j} = a_i$$

et

$$\sum_{i} \gamma_{i,j} = b_j$$

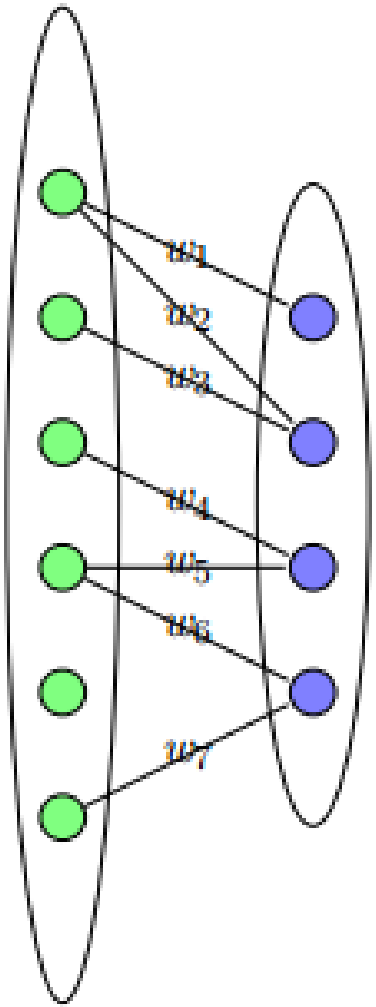
Modélisation

Introduction à la fonction de coût

- Fonction de coût $c(\gamma)$ = **risque de passage** d'une échelle à l'autre.
- Fonction définie à partir d'une distance $c(cov(x_i), cov(y_j))$ associées aux **distributions des covariables.**

$$c(\gamma) = \sum_{i,j} \gamma_{i,j} c(cov(x_i), cov(y_j))$$

Fonction de risque associée au transport



Définition du risque

- Plus la distribution des covariables sur la base A sera éloignée de la distribution des covariables sur la base B plus grand sera le risque.
- Le risque est défini à partir des **écarts d'entropie des distributions de certaines co-variables d'intérêt**.

Il s'agira de minimiser ce risque.

Modélisation

Fonction de coût

Soit K le nombre de covariables associées à la variable d'intérêt.
Soit S le nombre de modalités prises par chaque covariable.
On définit la fonction de coût par :

$$c(\gamma) = \sum_{k=0}^{K} \sum_{i} \sum_{j} \sum_{s=0}^{S} \gamma_{i,j} \left| p_{i,s}^{k} \ln p_{i,s}^{k} - q_{j,s}^{k} \ln q_{j,s}^{k} \right|$$

Où $p_{i,s}^{k} = \mathbb{P}(\mathrm{Cov}_k X = a_s | x_i)$ et $q_{j,s}^{k} = \mathbb{P}(\mathrm{Cov}_k Y = b_s | y_j)$
avec la convention $p \ln(p) = 0$ si $p = 0$.

Modélisation

Transport optimal

Soit $i \in [1, n]$. Soit $c^A(i)$ la classe de l'élément i sur la base A.
Etant donnée une classe c_1 de la base A et d_1 une classe de la base B, $N(c_1, d_1)$ donne le **nombre de transitions possibles** de la classe c_1 vers la classe d_1 par le transport optimal.

$$N(c_1, d_1) = \sum_{i=1}^{n} \mathbb{1}_{\left[c^A(i)=c_1,\, T_{\text{opt}}(c^A(i))=d_1\right]}$$

Modélisation

Règle d'allocation individuelle

Pour tout i fixé, $c(\hat{i})$ est la classe affectée par estimation à l'individu i. On définit :

$$V^k(i) = \left\{ d_1 \mid \sum_{i=1}^{k} \mathbb{1}_{\left[c^A(i)=c_1, c(\hat{i})=d_1\right]} \leq N(c_1, d_1) \right\}$$

On a finalement :

$$\text{Ind} = \underset{j \mid c^B(j) \in V^k(i)}{\arg\min} \; d(\text{Cov}(j), \text{Cov}(i))$$

Et

$$c(\hat{i}) = c^B(\text{Ind})$$

Plan

Plan

Définition des variables

- On se donne deux bases B_1 et B_2.
- **X** est codée de 1 à 4 sur B_1 et **Y** de 1 à 3 sur B_2.
- On se donne deux covariables $(\mathbf{A}_i)_i$ et $(\mathbf{B}_i)_i$ liées à **X** et **Y** et codées de 1 à 5.

Modèle de simulation

$$X_i = f_1(A_i, B_i) + \gamma_1 \mathcal{N}(0,1) \text{ et } Y_i = f_2(A_i, B_i) + \gamma_2 \mathcal{N}(0,1)$$

où :

$$f_1(A_i, B_i) = \alpha_1 A_i + \beta_1 B_i \qquad sh$$
$$f_2(A_i, B_i) = \alpha_2 A_i + \beta_2 B_i \qquad sh$$

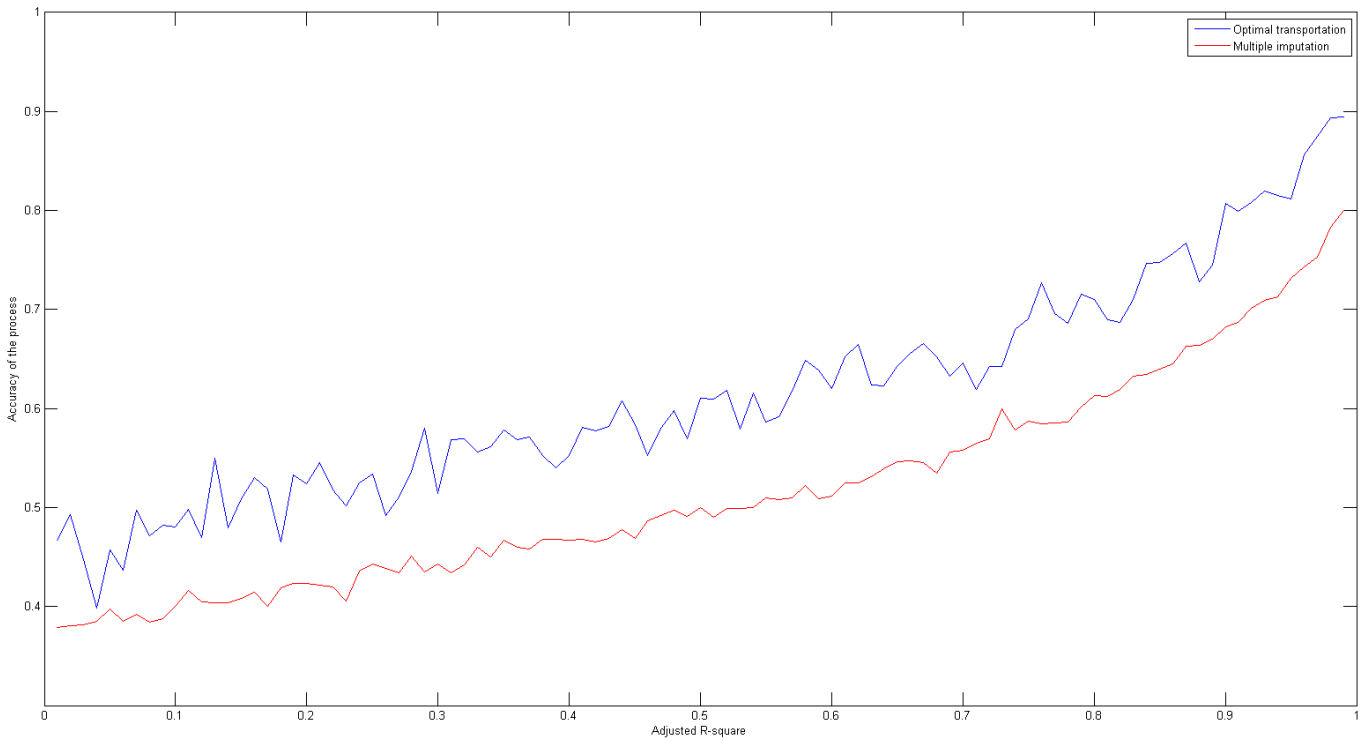
On sélectionne 500 parmi les 1000 simulés –> Base B_1.
Les 500 restants –> Base B_2.
$r = 100$

Résultats sur simulations

En fonction du R^2 ajusté

(γ_1, γ_2) est tel que $R^2 = \{0.1, \ldots, 0.99\}$

Comparaison avec l'imputation multiple



Plan

Données MCAR

Modèle de simulation

$$X_i = f_1(A_i, B_i) \text{ et } Y_i = f_2(A_i, B_i)$$

où :

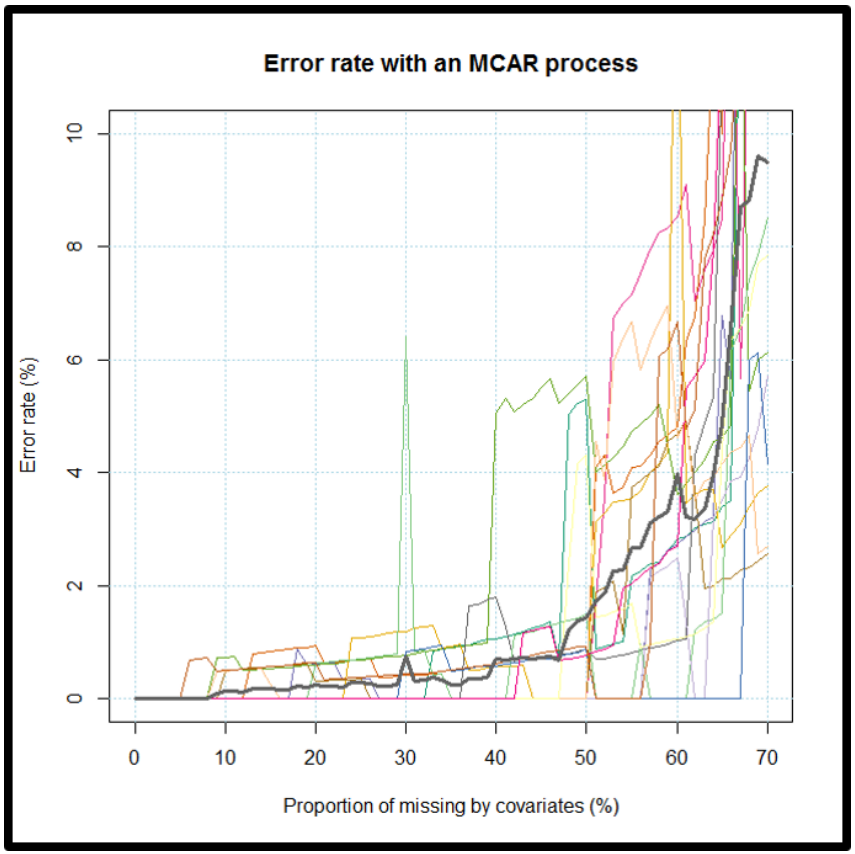
$$f_1(A_i, B_i) = \alpha_1 A_i + \beta_1 B_i \qquad sh$$
$$f_2(A_i, B_i) = \alpha_2 A_i + \beta_2 B_i \qquad sh$$

Pour chacune des covariables $(\mathbf{A}_i)_i$ et $(\mathbf{B}_i)_i$, on associe un processus de non-réponse
Pour $i = \{1, \ldots, 1000\}$,
$R_i^A = 1$ si $\mathbf{A}$ est présent pour i et 0 sinon.
On créé respectivement R_i^B

Résultats

On se donne un taux t de non-réponse par covariable tel que $t_A = t_B = t$

$$R_i^A \sim \textbf{Bern}(t) \text{ et } R_i^B \sim \textbf{Bern}(t)$$



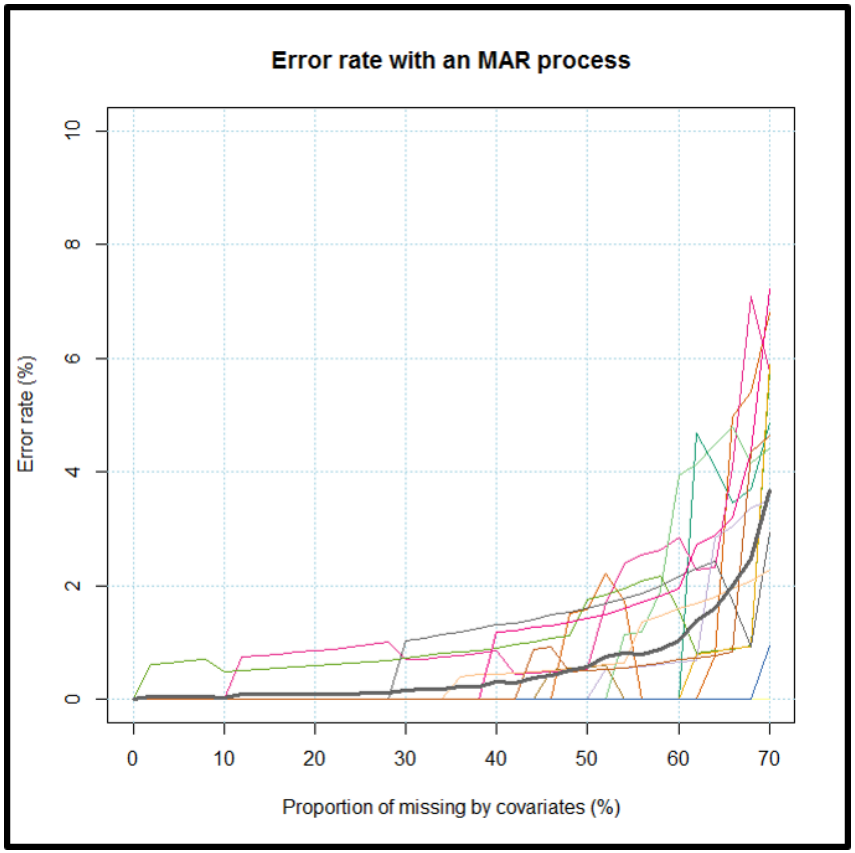
Données MAR

Modèle de simulation

Simulation d'un processus MAR sur les covariables **A** et **B**.

- On dispose de 5 pré-covariables communes aux covariables : le sexe, une variable binaire, une variable qualitative, une variable ordonnée et l'âge (3 classes).
- La variable Age influe sur R^B, la variable Sexe influe sur R^A.
- On fait varier le taux d'erreur par covariable de 0 à 70%.

Résultats


Error rate with an MAR process
Error rate (%)
0
2
4
6
8
10
0
10
20
30
40
50
60
70
Proportion of missing by covariates (%)

Outline

Conclusions

- Comparativement au processus d'imputation multiple, notre méthode par transport de mesures est **plus efficace** mais **moins stable** :
  - De 55% à 10% d'erreur selon la valeur du R^2 ajusté pour le processus de fusion de bases par transport optimal.
  - De 60% à 20% d'erreur selon la valeur du R^2 ajusté pour la méthode d'imputation multiple.
- Avec présence de données manquantes en MCAR et MAR, on ne dégrade pas trop l'efficacité de notre méthode de fusion de bases :
  - Pour des variables d'intérêt entièrement déterminées par des covariables, on a au plus 10% d'erreur en moyenne pour une part de données manquantes MCAR allant jusqu'à 70%.
  - Pour des variables d'intérêt entièrement déterminées par des covariables, on a au plus 4% d'erreur en moyenne pour une part de données manquantes MAR allant jusqu'à 70%.

Cadre d'application

Hypothèse forte

Fusionner les bases suppose que les **distributions** de la variable d'intérêt sur les deux bases soient **semblables**

Bases d'application

- Données longitudinales pour reconstitution de cohortes
- Données transversales sur populations comparables

Perspectives

- Tester la validité de l'approche sur données réelles
- Tester la validité de l'approche lorsqu'on introduit des données manquantes de type MNAR dans la base
- Tester la validité de l'approche en combinant alea et données manquantes
- Définir un "poids" pour pondérer le processus de fusion lorsque les populations ont des distributions différentes sur la variable d'intérêt
- Introduire des données en langage naturel (cf partenariat avec l'IRIT)

merci

Lot 1 : Alignements de variables.

- Objectif :
  - Fusion de bases de données
  - Problème : alignement de variables
    variables concernant le même objet soient codées de façon différente dans les 2 bases
  - Lien avec la question du transport de mesures

- Partenaires :
  - IMT (Nicolas SAVY - Sébastien DEJEAN - Laurent RISSER - X X)
  - INSERM Unité 1027 (Chloé DIMEGLIO)
  - IRIT (Mohand BENGHANEM - Mathieu SERRURIER - Nathalie AUSSENAC-GILLES)

Lot 2 : Gestion des données manquantes.

- Objectif :
  - Parfaire les connaissances sur l'Imputation multiple
  - Identifier des méthodes alternatives
    notamment dans le cadre complexe des données MNAR
  - Quantifier l'impact des données manquantes sur une réponse

- Partenaires :
  - IMT (Nicolas SAVY - Sébastien DEJEAN - Laurent RISSER - Cécile CHOUQUET - X X)
  - INSERM Unité 1027 (Chloé DIMEGLIO - Benoit LEPAGE)
  - IRIT (Mohand BENGHANEM - Mathieu SERRURIER - Nathalie AUSSENAC-GILLES)

Lot 3 : Intégration de données en langage naturel.

- Objectif :
  - Intégrer l'information contenue dans les messages textuels
    - expériences individuelles
    - étude de cas
    - entretiens
  - traitement du langage naturel
  - codage de cette information

- Partenaires :
  - IMT (Nicolas SAVY - Sébastien DEJEAN - Laurent RISSER - X X)
  - INSERM Unité 1027 (Chloé DIMEGLIO - Benoit LEPAGE)
  - IRIT (Mohand BENGHANEM - Mathieu SERRURIER - Nathalie AUSSENAC-GILLES)

Lot 4 : Sensibilité d'une base de données.

- Objectif :
  - Validitation des résultats issus de
    - fusion de bases
    - codage de l'information textuelle
    - gestion des données manquantes
  - Approche bayésienne :
    - Simuler des base de données en générant les variables selon leurs lois
    - Analyser la situation
    - Regarder la distribution des résultats

- Partenaires :
  - IMT (Nicolas SAVY - Laurent RISSER - X X)
  - INSERM Unité 1027 (Chloé DIMEGLIO - Benoit LEPAGE)

Lot 5 : Introduction de variables latentes.

- Objectif :
  - Un état de santé n'est pas forcément observé directement mais reflété par différentes mesures
  - Meilleure gestion de l'erreur de déclaration (voulue ou pas)
  - Meilleure gestion de l'enchainement des variables observée en passant par la "chaînes" des états latents
  - raisonnement poussé jusqu'à des variables latente comme l'état de santé vrai ou la qualité de vie vraie (si tant est que ces variables aient un sens)
  - Statistique : approche de l'estimation du style EM (markov caché)

- Partenaires :
  - IMT (Nicolas SAVY - Jérome DUPUIS - Laurent RISSER - X X)
  - INSERM Unité 1027 (Chloé DIMEGLIO - Benoit LEPAGE)

Lot 6 : Réflexion autour de la temporalité.

- Objectif :
  - Reconstitution de cohorte biographiques à partir de différentes sources de données
  - Les cohortes de naissance sont rares

- Partenaires :
  - IMT (Nicolas SAVY)
  - INSERM Unité 1027 (Thierry LANG, Cyrille DELPIERRE, Michelle KELLY-IRVING, Sébastien LAMY - Chloé DIMEGLIO - Benoit LEPAGE)
  - Autres equipes de l'INSERM
  - CHU

Lot 7 : Aspects éthique, juridiques et sociaux.

- Objectif :
  - Confidentialité des données
  - Anonymisation des données
  - Conséquences sociales et sanitaires de ces développements du croisement de Bases de Données

- Partenaires :
  - INSERM Unité 1027 (Thierry LANG, Emmanuelle RIAL-SEBAG)
  - IFERISS (laboratoire de Droit d'UT1, Sociologie,..)