Statistics Canada Statistique Canada

Canada

Statistics Canada
www.statcan.gc.ca

# Qualité des couplages d'enregistrements: défis et solutions

*9e Colloque francophone sur les sondages*

**Université du Québec en Outaouais**

**11-14 Octobre 2016**

A. Dasylva

# Aperçu

1. Besoins
2. Méthodes de couplage
3. Erreurs de couplage
4. Évaluation avec des vérifications manuelles
5. Évaluation sans vérifications manuelles

# 1. Besoins

## Priorités de l'agence

- Usage accru de données qui ne proviennent pas d'enquêtes, ex. données administratives ou mégadonnées
- Accès public accru (au Canada et à l'international) à des données pertinentes avec plus de qualité et d'actualité

## Applications

- Informations sur les erreurs de couplage pour les données couplées dans les Centres de recherche sur les données (CRDs)
- Prolonger la pertinence de données d'enquêtes longitudinales

# 2. Méthodes de couplage

- 2.1 Généralités
- 2.2 Couplage probabiliste

# 2.1 Généralités

- **But**: lier des enregistrements, qui se rapportent à la même entité, ex. une personne.
- **Terminologie**
  - Paire appariée: ayant trait à une seule entité
  - Paire non appariée: contraire de paire appariée
  - Paire liée: *classée* comme ayant trait à une seule entité
  - Paire non liée: contraire de paire liée
  - Paire aléatoire: paire obtenue en tirant un enregistrement au hasard dans chaque table
- **Méthodes de couplage** utilisées dans l'agence
  - déterministe ou hiérarchique
  - probabiliste, voir Fellegi et Sunter (1969)

# 2.2 Couplage probabiliste

1. Sélectionner un sous-ensemble de paires potentielles avec des pochettes
2. Dans chaque paire, comparer les variables pour produire un vecteur de résultats
3. Assigner un poids de couplage $W$ à partir d'un rapport de cotes
   - ❖ Une somme de poids pour chaque variable sous l'hypothèse d'indépendance conditionnelle
4. Prendre une décision à partir de deux seuils $S_1 \leq S_2$
   a. Paires définitives: $W \geq S_2$; lier
   b. Paires rejetées: $W \leq S_1$; ne pas lier
   c. Paires possibles (zone grise): décision manuelle

# 3. Erreurs de couplage

- 3.1 Sources et types d’erreurs
- 3.2 Mesures d’erreurs
- 3.3 Impact des erreurs

# 3.1 Sources et types d'erreurs

- **Sources**
  - Lier sans clé unique, avec un pseudo-identifiant, ex. nom, date de naissance, sexe
  - Erreurs typographiques
  - Variations orthographiques, alias, ou formats différents
- **Types**
  - Faux positif**:** paire non appariée qui est liée
  - Faux négatif**:** paire appariée qui n'est pas liée; d'un grand intérêt dans les couplages, qui visent surtout à minimiser les faux positifs.

# 3.2 Mesures d'erreurs

VP: Vrai Positif

VN: Vrai Négatif

FP: Faux Positif

FN: Faux Négatif

**TFN**: Taux de FN

**TFP**: Taux de FP

**Pr**: Précision

| | **Liée (L)** | **Non liée (NL)** | |
|---|---|---|---|
| **Appariée** | VP | **FN** | $TFN = FN/(VP + FN)$ |
| **Non appariée** | **FP** | VN | $TFP = FP/(VN + FP)$ |
| | $Pr = VP/(VP + FP)$ | | |

# 3.3 Impact des erreurs

- En général, les erreurs produisent du biais et peuvent augmenter la variance des estimateurs.
- **Modèles de régression**
  - Lahiri and Larsen (2005)
  - Chambers et Kim (2015)
- **Tables de contingence**
  - Chipperfield et coll. (2011)

# 4. Évaluation avec des vérifications manuelles

- 4.1 Notations
- 4.2 Plan de sondage
- 4.3 Protocole de vérification
- 4.4 Calcul des taux d'erreurs
- 4.5 Estimation par calage
- 4.6 Erreurs de vérification
- 4.7 Exemple

# 4.1 Notations

- $N$ **paires potentielles**
- Paire $i = 1, \ldots, N$
  - **Statut d'appariement**: $M_i$ égal à 1 si la paire est appariée et à 0 sinon
  - **Lien**: $L_i$ égal à 1 si la paire est liée et à 0 sinon
  - **Vecteur de résultats** $\boldsymbol{\gamma}_i$
  - **Poids de couplage** $W_i$
- **Proportion de mélange** $\lambda$
  - Probabilité qu'une paire potentielle soit appariée
  - $M_i \sim Bernoulli(\lambda)$

# 4.2 Plan de sondage

- Tirer un échantillon probabiliste de paires
- Probabilité d'inclusion ($\pi_i$) positive pour toute paire de poids $W_i \geq \theta$, où $\theta \ll S_1$
- Strates optimisées selon le poids de couplage $W_i$ (donné)
- Répartition optimale (ex. Neyman) en dérivant les variances de strates selon $M_i | W_i \sim Bernoulli(p_i)$ où
  - $\hat{p}_i = \left[1 + \left(1/\hat{\lambda} - 1\right)e^{-w_i}\right]^{-1}$
  - $\hat{\lambda}$ Maximise $\sum_{i=1}^{N} \log(\lambda w_i + 1 - \lambda)$

# 4.3 Protocole de vérification

- Faire vérifier chaque paire par un ou plusieurs vérificateurs indépendants n'ayant pas participé au développement du couplage.
- Ignorer le lien ($L_i$) le vecteur de résultats ($\boldsymbol{\gamma}_i$) et le poids de couplage ($W_i$)
- Contraindre chaque vérificateur à prendre une décision, c.-à-d. oui ou non
- Noter toute référence à une source externe, ex. internet

# 4.4 Calcul des taux d'erreurs

Soit $s$ l'échantillon de vérification manuelle

$$\widehat{TFN} = \left[\sum_{i \in s} \pi_i^{-1} M_i (1 - L_i)\right] \Big/ \left[\sum_{i \in s} \pi_i^{-1} M_i\right]$$

$$\widehat{Pr} = \left[\sum_{i \in s} \pi_i^{-1} L_i M_i\right] \Big/ \left[\sum_{i \in s} \pi_i^{-1} L_i\right]$$

# 4.5 Estimation par calage

- **Principe**: voir Dasylva (2015)
  - Pour chaque paire (au-dessus du seuil $\theta$) dériver la variable auxiliaire $\widehat{M}_i = P\left(M_i = 1 \middle| W_i\right)$ à partir d'un modèle
  - Vérifier un échantillon de paires pour déterminer $M_i$
  - Estimer les totaux impliquant les $M_i$ en utilisant les vérifications de l'échantillon et en calant par rapport aux $\widehat{M}_i$
  - Estimer les taux d'erreurs à partir des totaux estimés
- **Avantages**
  - Estimateurs convergents quelle que soit l'adéquation du modèle
  - Plus de précision pour une répartition sous-optimale

# 4.6 Erreurs de vérification

- 4.6.1 Défis de la vérification manuelle
- 4.6.2 Principe des vérifications répétées
- 4.6.3 Estimation des erreurs

# 4.6.1 Défis de la vérification manuelle

- **Coût élevé**
- **Problème de fiabilité**
  - Caractère subjectif des décisions manuelles
  - Les vérificateurs experts sont parfois incompétents
    - Dans la zone grise
    - Lorsque les variables de couplage ne donnent pas assez d'information, voir Newcombe et coll. (1983).
  - Un problème souvent ignoré
    - Fellegi et Sunter (1969), Guiver (2011), Heasman (2014)

# 4.6.2 Principe des vérifications répétées

- Adapter la méthode des interviews répétées, voir Biemer (2011)
- Faire des vérifications répétées par des vérificateurs indépendants, pour chaque paire
- Détecter les erreurs par les conflits

# 4.6.3 Estimation des erreurs de vérification

- **Hypothèses**
  - Vérificateurs échangeables et supposés conditionnellement indépendants étant donné le statut d'une paire et d'autres variables explicatives, ex. expérience du vérificateur
- **Solution de la décision à la majorité**
  - Négliger les erreurs de la décision à la majorité
  - Besoin d'une information abondante et de qualité
  - Estimation simple, mais approximative
- **Analyse de classes latentes**: voir Biemer (2011)
  - Estimation basée sur la maximisation d'une vraisemblance
  - Estimation complexe, mais plus précise

# 4.7 Exemple

- 4.7.1 Méthodes
- 4.7.2 Résultats

# 4.7.1 Méthodes

- **Fichiers intrants**
  - *Enquête sur la santé dans les collectivités canadiennes (ESCC)*: 2,3M d'enregistrements. provenant de périodes de collecte commençant en 2000, 2003, 2005, et de 2007 à 2011
  - *Base canadienne de données de mortalité (BCDM)*: 3,6M d'enregistrements.
- **Variables**
  - Date de naissance
  - Sexe
  - Nom de famille
  - Prénom
  - Code postal

# 4.7.1 Méthodes

- **Distribution des paires**
  - Potentielles : 418M
  - Définitives (D) : 114K
  - Possibles (P) : 22K
  - Rejetées (R) : 418M
- **Vérifications manuelles**
  - Un échantillon de 1000 paires stratifiées ayant un poids au-dessus de 1
  - Trois vérificateurs indépendants pour chaque paire

Voir Sanmartin et coll. (2015) pour plus de détails

# 4.7.2 Résultats

## Erreurs de couplage pour les seuils S1=S2=92

| | | Liée (L) | Non liée (NL) | |
|---|---|---|---|---|
| **Vérification manuelle** | **Appariée** | VP<br>34 298 | **FN**<br>**855,09** | 35 153,1<br>**TFN=2,43%** |
| | **Non appariée** | **FP**<br>**473,57** | VN<br>1 161 015 | 1 161 489<br>**TFP=0,04%** |
| | | 34 771,6<br>**Pr=98,64%** | 1 161 870 | 1 196 642 |

# 4.7.2 Résultats

## Taux d'erreurs de vérification globaux et par strate

| Strate | Intervalle de poids | TFP (%) | TFN (%) |
|---|---|---|---|
| 1 | 1,51 – 23,51 | 0,00 | 0,00 |
| 2 | 23,52 – 49,51 | 0,54 | 33,33 |
| 3 | 49,52 – 73,51 | 4,01 | 5,88 |
| 4 | 73,52 – 97,51 | 6,86 | 5,49 |
| 5 | 97,52 – 123,51 | 11,11 | 1,09 |
| 6 | 123,52 – 149,51 | 0,00 | 0,27 |
| 7 | 149,52 – 163,51 | 0,00 | 0,53 |
| 8 | 163,51 – 194,52 | 0,00 | 0,27 |

| | | Vérification manuelle | |
|---|---|---|---|
| | | classée appariée | classée non appariée |
| Vérité | Appariée | TVP | **TFN=2,97%** |
| | Non appariée | **TFP=0,15%** | TVN |

# 5. Évaluation sans vérifications manuelles

- 5.1 Limites des vérifications manuelles
- 5.2 Méthode des pochettes conditionnellement indépendantes (CI)
- 5.3 Autres méthodes

# 5.1 Limites des vérifications manuelles

- **Coût élevé**
- **Erreurs à la vérification**
- **Situations particulières**
  - Couplage de données anonymes, ex. en santé
  - Variables de couplage apportant peu d'information, ex. couplage de données sociales sans les noms

# 5.2 Méthode des pochettes CI

- **Principe**: voir Dasylva et Sinha (2014)
  a. Pochettes fondées sur des variables, qui sont indépendantes des variables de couplages dans chaque table
  b. Estimer la distribution des paires non appariées dans les pochettes par des paires aléatoires
  c. Déduire la distribution des paires appariées
- **Forces**
  - Prise en compte des interactions
  - Méthode s'appliquant aussi aux couplages non probabilistes
- **Faiblesses**
  - Disponibilité de variables ayant les propriétés requises

# 5.3 Autres méthodes

## Mélange de multinomiales

- Modéliser la distribution du vecteur de résultats ($\boldsymbol{\gamma}_i$); voir Fellegi et Sunter (1969), Jaro (1989), Armstrong et Mayda (1993), Larsen et Rubin (2001), et Winkler (2006)
- Modèle inadéquat avec l'hypothèse d'indépendance conditionnelle (Belin et Rubin, 1995) et possiblement surparamètré en présence d' interactions (Kim, 1984)

## Mélange de normales transformées selon Box-Cox

- Modéliser la distribution du poids de couplage ($W_i$); voir Belin et Rubin (1995)
- Solution nécessitant un échantillon d'apprentissage et une bonne séparation des paires appariées et non appariées

# Questions?

Abel.dasylva@Canada.ca

# Bibliographie


- **Armstrong, J., and Mayda, J. (1993)**. “Model-based estimation of record linkage error rates”, *Survey Methodology*, 19, pp. 137-147, 1993.
- **Belin, T.R. and Rubin, D.B.. (1995)**. “A Method for calibrating false-match rates in record
- **Biemer, P. (2011)**. Latent Class Analysis of Survey Error, New Jersey:John Wiley.
- **Chambers W.E., and G., Kim (2015)**. “Secondary analysis of linked data”, in Methodological Developments in Data Linkage, pp.83-108, UK:John Wiley.
- **Dasylva, A., and S. , Sinha (2014)**. “Reducing the Structure of Statistical Models for Probabilistic Record Linkage”, poster presented at the *Joint Statistical Meetings*.
- **Dasylva, A (2015)**. “Design-based Estimation with Record-Linked Administrative Files”, in *Proceedings of the International Symposium on Statistical Methodology*.
- **Fellegi, I.P., and Sunter, A.B. (1969)**. “A Theory of Record Linkage”, *JASA*, 64, pp. 1183-1210.
- **Guiver, T. (2011)**. “Sampling-Based Clerical Review Methods in Probabilistic Linking”, unpublished report, Australia: Australia Bureau of Statistics.
- **Heasman, D. (2014)**. “Sampling a matching project to establish the linking quality”, *Survey Methodology Bulletin*, Office of National Statistics, 72, pp. 1-16.
- **Jaro, M. A. (1989)**. “Advances in record linkage methodology to matching the 1985 census of Tampa, Florida”, *Journal of the American Statistical Association*, 84, pp. 414-420.

# Bibliographie


- **Kim, B.S. (1984)**. “Studies of multinomial mixture models”, PhD thesis, University of North Carolina , Chapel Hill.
- **Lahiri, P. and Larsen, D. (2005)**. “Regression analysis with linked data”, JASA, 100, pp. 222-227.
- **Larsen, M., and Rubin, D. (2001)**. “Iterated automated record linkage using mixture models”, *JASA*, 96, pp. 32-41.
- **Newcombe, H.B., Smith, M.E., and Howe, G.R. (1983)**. “Reliability of computerized versus manual death searches in a study of the health of eldorado uranium workers”, *Computers in Biology and Medecine*, 13, pp. 157-169.
- **Sanmartin, C., Y., Decady, R.,Trudeau, A., Dasylva, M., Tjepkema, P., Fines, R., Burnett, N., Ross, and D., Manuel (2015)**.”Linking the Canadian Community Health Survey to the Canadian Mortality Database: A national resource to study mortality in Canada”, submitted to Health Reports.
- **Winkler, W.E., and Yancey, W.E. (2006)**. “Record-linkage error-rate estimation without training data”, in *Proceedings of the Section on Survey Research Methods*, ASA.
- **Winkler W.E. (2015)**. “Probabilistic linkage” in Methodological Developments in Data Linkage, pp.8-35, UK:John Wiley.

# Abbréviations

| | |
|---|---|
| **BCDM** | Base canadienne de données de mortalité |
| **CI** | Conditionnellement indépendants |
| **ESCC** | Enquête sur la santé dans les collectivités canadiennes |
| **TFN** | Taux de faux négatifs |
| **TFP** | Taux de faux positifs |
| **TVN** | Taux de vrais négatifs |
| **TVP** | Taux de vrais positifs |