An extension of Fellegi-Sunter record linkage model for mixed-type data with application to SNDS

Thanh Huan VO*
Joint work with G. Chauvet, A. Happe, E. Oger, S. Paquelet and V. Garès

The 42nd Annual Conference of the International Society for Clinical Biostatistics

*INSA Rennes and IRT b<>com
ThanhHuan.Vo@b-com.com

Contents

Introduction

Introduction

- Motivation
  - SNDS (Système National des Données de Santé): is a national health information system of the French population
  - GETBO: venous thromboembolism (VTE) cases recorded between 2013 and 2015 in Brest

SNDS

	$\mathbf{X} \in \mathbb{R}^p$	$\mathbf{Y} \in \mathbb{R}^m$	$\mathbf{Z} \in \mathbb{R}^n$
1	Observed	Observed	Unobserved
...			
...			
n_A			

GETBO

	$\mathbf{X} \in \mathbb{R}^p$	$\mathbf{Y} \in \mathbb{R}^m$	$\mathbf{Z} \in \mathbb{R}^n$
1	Observed	Unobserved	Observed
...			
...			
n_B			

- Record linkage
  - process of combining data from different sources that refers to the same entity
  - no identifying information is available

Example

	Postal code	Cancer	Date of echodoppler
a_1	29001	1	10/03/2014
a_2	29002	0	17/05/2013
a_3	29003	0	19/11/2013
a_4	29002	0	01/03/2014

Database A

	Postal code	Cancer	Date of echodoppler
b_1	29001	1	12/03/2014
b_2	29002	0	17/05/2013

Database B

Table 1: Example of two databases with three matching variables: Postal code, cancer and date of echodoppler

Matching variables: are chosen among those in common between databases

- Categorical variables:
  - Binary data: sex, diagnosis code, ...
  - More than 3 categories: postal code, month of birth,...
- Continuous variables:
  - age, duration from an origin of dates (date of medical acts,...)

Record linkage model

Outline

Comparison step

Let K be the number of matching variables and

$$a_i = \left(a_i^1, \ldots, a_i^K\right), \quad i = 1, \ldots, n_A$$
$$b_j = \left(b_j^1, \ldots, b_j^K\right), \quad j = 1, \ldots, n_B$$

For each record pair (a_i, b_j), we define a comparison vector

$$\gamma_{ij} = (\gamma_{ij}^1, \ldots, \gamma_{ij}^K)$$

where

- $\gamma_{ij}^k = h^k(a_i^k, b_j^k)$
  - and $h^k(\cdot, \cdot)$ is a comparison function for the k^{th} matching variable.

Simple comparison approach

In a simple approach, we define for $k = 1, \ldots, K$

$$\gamma_{ij}^k = h^k(a_i^k, b_j^k) = \mathbb{1}_{a_i^k = b_j^k} \tag{1}$$

	Postal code	Cancer	Date of echodoppler
a_1	29001	1	10/03/2014
a_2	29002	0	17/05/2013
a_3	29003	0	19/11/2013
a_4	29002	0	01/03/2014

Database A

	Postal code	Cancer	Date of echodoppler
b_1	29001	1	12/03/2014
b_2	29002	0	17/05/2013

Database B

	γ^1	γ^2	γ^3
γ_{11}	1	1	0
γ_{12}	0	0	0
γ_{21}	0	0	0
γ_{22}	1	1	1
γ_{31}	0	0	0
γ_{32}	0	1	0
γ_{41}	0	0	0
γ_{42}	1	1	0

Table 2: Simple comparison matrix

Proposed comparison approach

- For categorical matching variables with L different categories

  $\longrightarrow$ L^2 configurations of possible pairs

  - Assign a number from 1 to L^2 **(no order meaning)** for each possible configuration

    Example: For a binary matching variable, we have

$$\begin{cases} h(0,0) = 1 \\ h(0,1) = 2 \\ h(1,0) = 3 \\ h(1,1) = 4 \end{cases} \tag{2}$$

    If we want to reduce the number of parameters

$$\begin{cases} h(0,0) = 1 \\ h(0,1) = h(1,0) = 2 \\ h(1,1) = 3 \end{cases} \tag{3}$$

- For continuous matching variables: Using distance (1-norm, 2-norm,...)

  Example: $a_1^3 = 10/03/2014, b_1^3 = 12/03/2014$

  $\longrightarrow \gamma_{11}^3 = |a_1^3 - b_1^3| = 2$

Proposed comparison approach

	Postal code	Cancer	Date of echodoppler
a_1	29001	1	10/03/2014
a_2	29002	0	17/05/2013
a_3	29003	0	19/11/2013
a_4	29002	0	01/03/2014

Database A

	Postal code	Cancer	Date of echodoppler
b_1	29001	1	12/03/2014
b_2	29002	0	17/05/2013

Database B

	γ^1	γ^2	γ^3
γ_{11}	1	1	0
γ_{12}	0	0	0
γ_{21}	0	0	0
γ_{22}	1	1	1
γ_{31}	0	0	0
γ_{32}	0	1	0
γ_{41}	0	0	0
γ_{42}	1	1	0

(a) Simple approach

	γ^1	γ^2	γ^3
γ_{11}	1	3	2
γ_{12}	0	2	297
γ_{21}	0	2	299
γ_{22}	1	1	0
γ_{31}	0	2	113
γ_{32}	0	1	186
γ_{41}	0	2	9
γ_{42}	1	1	288

(b) Proposed approach

Table 4: Two different comparison approaches

Outline

Modeling

Comparison vector γ_{ij} of record pairs (a_i, b_j) is a mixed-type vector that includes

- K_1 categorical values
- K_2 continuous values

$$\gamma_{ij} \equiv \left(\gamma_{ij}^{1}, \ldots, \gamma_{ij}^{K_1}, \gamma_{ij}^{K_1+1}, \ldots, \gamma_{ij}^{K_1+K_2}\right)$$

Mixture model (P. Fellegi and B. Sunter, 1969)

$$\mathbb{P}(\gamma) = \mathbb{P}(\gamma|M)\mathbb{P}(M) + \mathbb{P}(\gamma|U)\left[1 - \mathbb{P}(M)\right]$$

Classification

- Once all parameters are estimated

  $\longrightarrow$ Estimate probability of matching for all record pairs using Bayes formula

$$q_{ij} = \mathbb{P}\left((a_i, b_j) \in M | \gamma_{ij}\right) = \frac{\mathbb{P}\left(\gamma_{ij} | M\right) \mathbb{P}(M)}{\mathbb{P}(\gamma_{ij})} \tag{4}$$

- Classify the set of all record pairs into
  - Matched set:
$$M = \{(a_i, b_j) | q_{ij} \geq \tau\}$$
  - Unmatched set:
$$U = \{(a_i, b_j) | q_{ij} < \tau\}$$

  where τ is a predefined threshold (e.g. 0.5)

Application

Context

- Databases:
  - SNDS: 48 102 medical acts corresponding to 32 382 patients
  - GETBO: 1919 medical acts corresponding to 1332 patients

    $\longrightarrow$ Firstly, we find matched pairs of medical acts

    $\longrightarrow$ then, deduce pairs of patients

- Blocking variables:
  - Month of birth
  - Type of medical acts (echodoppler, scintigraphy, angiography, ...)

  $\longrightarrow$ Reduce $48\,102 \times 1\,919 = 92\,307\,738$ to $4\,308\,847$ possible pairs

- Four matching variables:
  - Year of birth
  - Residency code
  - Gender
  - Date of medical acts

Methods

- **FS-ext**: Our proposed model for mixed-type data
  - Binary comparison for year of birth and residency code
  - Three categorical comparison (3) for gender
  - Absolute distance for date of medical acts

- **FS**: Traditional model
  - Binary comparison for all matching variables

- **Deterministic method**: a pair of medical acts is classified as a match if
  - the same type of medical act, month, year of birth, gender, residency code, and,
  - the difference between date of medical acts is less than or equal to 3 days

Comparison of results

	Classified as a match by					
	FS-ext	**FS**	**Deterministic method**	Number of pairs of patients	$\overline{\hat{q}}_{\text{FS-ext}}(sd)$	$\overline{\hat{q}}_{\text{FS}}(sd)$
	X	X	X	867	0.993 (0.003)	0.996 (0)
	X	X		245	0.900 (0.045)	0.911 (0)
	X			34	0.868 (0.136)	
		X		2		0.911 (0)
Total	1146 (86%)	1114 (83.6%)	867 (65%)			

Table 5: Comparison of three different record linkage methods with the number of pairs, the average of estimated posterior probability of matching mean($\hat{q}$) and the standard deviation (in parentheses)

Concluding remarks

- In the Monte Carlo simulation, our proposed approaches improve the performance of Fellegi-Sunter model in both scenarios
  - Low prevalence binary matching variables
  - Continuous matching variables
- In application, our extension model predicts more matching patients in SNDS for patients registered in GETBO with high probability

Concluding remarks

- In the Monte Carlo simulation, our proposed approaches improve the performance of Fellegi-Sunter model in both scenarios
  - Low prevalence binary matching variables
  - Continuous matching variables
- In application, our extension model predicts more matching patients in SNDS for patients registered in GETBO with high probability

Thank you for your attention!

References

Ivan P. Fellegi and Alan B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64:1183–1210, 12 1969. doi: 10.1080/01621459.1969.10501049.