Proposition d'un nouveau processus d'appariement au Pôle Revenus Fiscaux et Sociaux (RFS)

Une application à l'enquête CARE

P. Jabot[1] P.E. Treyens[1]

[1]Pôle Revenus Fiscaux et Sociaux
Direction régionale de Bretagne
Insee

JMS 2018

Plan de la présentation

I. Introduction

- La mission du Pôle RFS, la valorisation des données fiscales et sociales

- Un rouage décisif pour l'enrichissement, l'appariement
- Pour des raisons juridiques notamment, un rouage délicat

- Relier sources d'informations sans identifiant unique et évident
- Construire identifiant par combinaison de variables

- Comment améliorer un appariement gourmand en ressources ?
    - *Muscler le process de production*
    - *Ciseler le rouage technique*

I. Contexte du test

- Un test centré sur le rouage technique de l'appariement
- Le fichier d'enquête Capacité Aides et REssources des seniors (CARE)
  - *Une base de 15 000 individus*
  - *Un fichier de bonne qualité*
- Une base fiscale : Le fichier FIP
  - *Une base de 50 000 000 de foyers fiscaux*
  - *Un fichier qui comprend des variables personnelles et d'adresse*
- Les variables communes
  - *Département, jour, mois et année de naissance*
  - *Prénom, nom (et le nom de naissance de FIP)* et sexe
  - *L'identifiant de la commune et des mots directeurs de l'adresse*

II. Le processus actuel

Un processus séquentiel

- Construit graduellement à partir des expériences précédentes.
- Processus séquentiel qui tente d'épouser les différences de qualité des fichiers.
- Fondé sur la construction d'un jeu de clés quelquefois très fourni qui s'adapte aux caractéristiques des variables.

- Au final, ce processus présente différentes caractéristiques :
  - *Peu industrialisé, gourmand en interventions manuelles*
  - *L'appariement est basé sur une expertise présente et passée des fichiers*

III. Un nouveau processus d'appariement

Un appariement *déterministe* non supervisé

- Les différentes étapes du nouveau processus
  - *Préparation des données et appariement exact*
  - *Indexation*
  - *Choix d'une distance entre deux individus*
  - *Choix d'un seuil pour l'appariement*

III. Individualisation, normalisation et appariement exact

- Le fichier FIP est individualisé $\Rightarrow$ Plus de 84 000 000 de lignes
- Exclusion de 163 individus de CARE ayant trop de données manquantes
- Le reliquat du fichier CARE (14 837 individus) est aussi individualisé $\Rightarrow$ 15 180 lignes
- Le même processus de normalisation des adresses est appliqué aux deux bases
- Création d'une clé totale dans FIP et CARE
- Fusion sur cet identifiant unique $\Rightarrow$ 8 306 appariements exacts (56 %)
- Récupération d'un reliquat CARE et d'un reliquat FIP en sortie de cette étape

III. Indexation

Position du problème

- Calculer la distance de chaque individu CARE à chaque individu FIP
  - *Sans appariement exact ⇒ 1 300 milliards de comparaisons*
  - *Avec appariement exact ⇒ 550 milliards de comparaisons*
  - *Trop long dans les deux cas ! (Complexité quadratique)*
- L'indexation (indexing ou blocking) : Le moyen de contourner cet écueil
  - *Idée : Partitionner les bases CARE et FIP en sous blocs-disjoints*
  - *Comparer systématiquement les individus dans les sous-blocs correspondants*
  - *Quelle clé de blocage choisir ? L'identifiant de la commune*
  - *Permet de passer à 1.3 milliards de comparaisons*

III. Choix de la distance

Distance entre deux individus (1)

- Soit a un individu de CARE et b un individu de FIP.
- On définit la distance entre deux individus comme la somme non pondérée des *sous-distances* entre les variables caractérisant un individu

$$\begin{aligned} d(a,b) &= d_{Nom}(Nom_a, Nom_b) + \dots \\ &\quad \dots + d_{Adresse}(Adresse_a, Adresse_b) \end{aligned}$$

  - $\neq$ Distance mathématique
  - *Chaque sous-distance est comprise entre 0 et 1*
  - *On utilise 8 variables : Nom (+ nom de naissance dans FIP), prénom, sexe, département, année, mois et jour de naissance et le ou les deux mots directeurs de l'adresse*

III. Choix des distances

Basé principalement sur la **distance de Levenshtein** $\Rightarrow$ Nombre minimal de caractères à supprimer, insérer ou remplacer pour passer d'une chaîne de caractères à une autre

- *Par exemple, la distance entre "MARTHE" et "MARIE" est de 2. Suppression du H et transformation du T en I (ou inversement)*
- D'autres distances existent comme la distance phonétique, la distance de Jaro-Winkler (utilisée par StatCan) non implémentées sous SAS etc.
- Les distances sont bornées et divisées par cette borne pour appartenir à $[0, 1]$

III. Critère de sélection du seuil

La question centrale

- Un appariement est considéré comme réalisé sous 3 conditions
  - *L'individu FIP apparié est celui dont la distance est minimale*
  - *Sous réserve que le distance soit inférieure à un seuil donné*
  - *En cas de double appariement (sur les 2 adresses care), la première est privilégiée*
- Si le choix de la distance est subjectif, le seuil doit l'être le moins possible
- **Question : Comment fixer ce seuil le plus objectivement possible ?**

III. Critère de sélection du seuil

Un cadre de travail similaire à celui des test

- Soit k une modalité définissant un sous-bloc A_k dans CARE et B_k dans FIP
- Soit un individu a choisi au hasard dans A_k, on peut commettre deux types d'erreur

$$\text{Accepter l'appariement sachant que } a \notin B_k \quad (1)$$

$$\text{Rejeter l'appariement sachant que } a \in B_k \quad (2)$$

C'est-à-dire

$$\text{Accepter que } a \in B_k \quad \text{sachant que} \quad a \notin B_k$$

$$\text{Accepter que } a \notin B_k \quad \text{sachant que} \quad a \in B_k$$

- Comme le risque (1) semble plus grave. On a alors

$$H_0 \ : \ a \notin B_k \qquad \text{vs.} \qquad H_1 \ : \ a \in B_k$$

III. Critère de sélection du seuil

Hypothèse simplificatrice

- Ainsi, le risque de première espèce associé au bloc k et au seuil S est

$$\begin{aligned} \alpha(k,S) &= \mathbb{P}_k\left(\text{Accepter } a \in B_k \mid a \notin B_k\right) \\ &= \mathbb{P}_k\left(\min_{b_{j,k}\in B_k} d(a,b_{j,k}) < \mathrm{S} \mid a \notin B_k\right) \end{aligned}$$

- Idéalement, il faudrait calculer $\alpha(k,\mathrm{S})$, i.e. donner la distribution de $\min d(a,b_{j,k})$ sachant que $a \notin B_k$ pour chaque $k \Rightarrow$ Trop long !

Hypothèse de travail retenue

On suppose que pour tout S, $\alpha(k,\mathrm{S})$ croît avec la taille de B_k. I.e., plus il y a d'individus dans un sous-bloc, plus il y a de chances de trouver un écho proche à tort.

III. Critère de sélection du seuil

- Soit $B_{\bar{k}}$ le sous-bloc ayant le plus grand cardinal
- Sous l'hypothèse retenue, le niveau du test (pour un seuil S) est donc

$$\alpha = \mathbb{P}\left(\min_{b_{j,\bar{k}} \in B_{\bar{k}}} d(a, b_{j,\bar{k}}) < \mathrm{S} \mid a \notin B_{\bar{k}} \right) \tag{3}$$

- Or, pour que l'appariement ait du sens, l'individu a tiré au hasard devrait souvent se retrouver dans $B_{\bar{k}}$
  $\Rightarrow$ Pour s'assurer que $a \notin B_{\bar{k}}$, on va tirer l'individu au hasard dans $\bar{A}_{\bar{k}}$

III. Critère de sélection du seuil

- Finalement, on va utiliser comme estimateur théorique du niveau

$$\alpha_T = \mathbb{P}\left(\min_{b_{j,\bar{k}} \in B_{\bar{k}}} d(a, b_{j,\bar{k}}) < \mathrm{S} \mid a \in \bar{A}_{\bar{k}} \right) \tag{4}$$

- et comme estimation de cet estimateur théorique (pour un seuil S donné)

$$\widehat{\alpha_T} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{d(a,B_{\bar{k}}) < \mathrm{S} \text{ tel que } a \in \bar{A}_{\bar{k}}\}} \tag{5}$$

III. Critère de sélection du seuil

Dans la pratique ?

- On choisit $\bar{k} = 75000 \cup 75115 \Rightarrow$ plus d'un million d'individus
- Pour $\bar{A}_{\bar{k}}$, on exclut tous les codes commune commençant par "75"
- On réalise un tirage par SAS de 5 000 individus dans $\bar{A}_{\bar{k}}$
- On corrige *à la main* les cas des déménagements restants
  - *8 cas pour lesquels la distance était inférieure ou égale à 1*
- On obtient la distribution suivante

Quantile	0.001	0.005	0.01	0.025	0.05
Seuil	1.62	1.80	1.93	2.1	2.23

$\Rightarrow$ Si la distance est inférieure à 1.93, on a moins de 1% de chances d'accepter un appariement à tort

IV. Bilan de l'appariement

Seuil d'appariement retenu	Nombre et taux d'appariement du reliquat 6 531 individus enquêtés non appariés exactement	Appariement global sur les 14 837 individus appariables 8306 appariements exacts
Distance $= 0$	2 601	10 907 (soit 73.51%)
Distance < 1.4	6 480	14 786 (soit 99.66%)
Distance < 1.62	6 488	14 794 (soit 99.71%)
Distance < 1.8	6 488	14 794 (soit 99.71%)
Distance < 1.93	6 488	14 794 (soit 99.71%)
Distance < 2.1	6 504	14 810 (soit 99.82%)
Distance < 2.23	6 507	14 813 (soit 99.84%)

Remarque : Taux calculés pour 14 837 individus *appariables* sur 15 000

IV. Bilan de l'appariement

- En termes de temps de calculs
  - *Phase de normalisation et d'individualisation : Moins de 12 h*
  - *Appariement exact : Moins de 30 minutes*
  - *Appariement du plus proche écho au niveau communal (et création d'une base récupérant tous les échos dont la distance est inférieure à une valeur donnée, ici 3) : Moins de 14h*
  - *Calibrage du seuil : Moins de 9h pour 1 000 individus*
- ... à ne pas confondre avec les temps de traitement
- Au moins 93.7 % d'enrichissements contre 90 % pour la méthode actuelle