Insee
Mesurer pour comprendre

L'application RAPSODIE

Pôle Revenus Fiscaux et Sociaux

Jean-François Portier & Pierre-Eric Treyens

12 avril 2021

Plan de la présentation

I. Historique des appariements au PRFS

Premier appariement au PRFS en 2000 (ERF)

- Chaîne figée avec des programmes à soumettre en batch
- Beaucoup de phases manuelles
- Autour de 80 jours pour ERF jusqu'en 2015
- Autour de 40 jours avec le tirage dans la TH

En 2015, des enquêtes individus nécessitent de nouveaux programmes

- Basés sur le relâchement très progressif de clés successives
- Abandon des reprises manuelles

Dans les deux cas, de bons résultats mais fastidieux à obtenir, peu reproductibles et demandant une forte expertise des fichiers

I. Origine de Rapsodie

En 2018, une nouvelle méthode d'appariement est testée sur l'enquête CARE

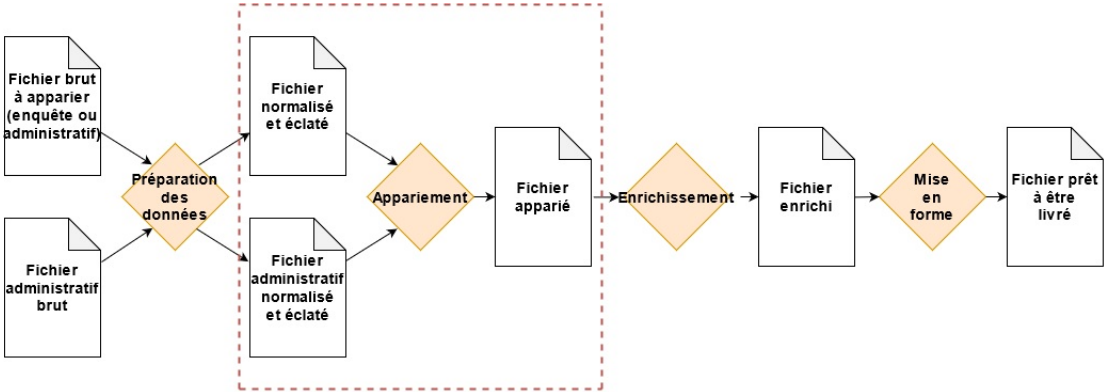
- Méthode déterministe par recherche du plus proche écho (plus faible distance)
- utilisant la commune comme clé de blocage
- et qui a donné de très bons taux d'appariement (trop ?)

Pensée pour répondre à plusieurs besoins

- Réduire les délais de production
- Générique
- Simple d'utilisation
- Sans reprise manuelle
- Reproductible

I. Schéma simplifié de l'application RAPSODIE

L'appariement, un rouage de l'application Rapsodie



Plan de la présentation

II. Les données mobilisées

Les sources utilisées pour l'appariement ...

- Le Fichier d'Imposition des Personnes (FIP)
- Les fichiers sociaux (CNAF, CNAV, MSAV et MSAF)
- Les fichiers d'enquête ou administratifs à enrichir

... et pour l'enrichissement

- Le Permanent des Occurrences de Traitement des Émissions (POTE)
- Les fichiers sociaux (CNAF, CNAV, MSAV et MSAF)

II. Déduplication des fichiers sociaux

Certains individus perçoivent des prestations sur plus de deux identifiants

- Identification des doublons sur la base d'une clé Nom-Prénom-Date de naissance-Sexe et d'une autre information parmi
  - l'adresse, le nom de naissance (si différent du nom d'usage),
  - la date de naissance des enfants ou du conjoint,
  - des informations de revenus, un tuilage dans les prestations.
- Agrégation des prestations perçues
- Récupération des différentes adresses disponibles pour chaque individu

II. Préparation des données

Une étape très importante !

- Recodification
  - Renommage/formatage des variables
  - Recodification des modalités
  - Pour FIP, ajout de variables fiscales
- Traitement des adresses : Récupération d'au plus deux mots directeurs par adresse disponible
- Éclatement : Duplication des lignes du fichier selon les différentes informations disponibles sur l'adresse, le nom ou le prénom

Un exemple : Le schéma simplifié de la table FIP en sortie

Dirindik	IdFiscal	Prénom	Nom	Commune	Adresse	Anrev	Id_gestion	Anrev_id
35001095177893	100028919306	JEANNE	DUPONT	RENNES	NOEL FAIL	2017	100028919306	2017
35001095177893	100028919306	JEANNE	MARTIN	RENNES	NOEL FAIL	2017	100028919306	2017
35001095177893	100028919306	JEANNE	DUPONT	ST-MALO	TREHOUART	2017	100028919306	2017
35001095177893	100028919306	JEANNE	MARTIN	ST-MALO	TREHOUART	2017	100028919306	2017
75060758094593	100028919306	JEANNE	DUPONT	PARIS	LECLERC		100028919306	2017
75060758094593	100028919306	JEANNE	MARTIN	PARIS	LECLERC		100028919306	2017
75568487411494	100028919306	JEANNE	MARTIN	PARIS	PAUL FORT		100028919306	2017
83542314012493	200126256781	JEANNE	MARTIN	TOULON	COL CANTO		200126256781	

Remarque : Si l'IDFiscal est manquant, on crée un IDFiscal de gestion

Plan de la présentation

III. L'appariement exact

Une étape permettant uniquement de réduire les délais

- Création d'une clé en concaténant les informations identifiantes disponibles dans les deux fichiers
- Fusion des deux fichiers *via* cette clé
- Suppression des individus retrouvés dans les deux bases et constitution de deux restes à apparier
- Permet d'apparier entre 50% et 70% des individus d'une enquête

III. L'appariement par plus proche écho

Choix d'une clé de blocage

- Modulable dans RAPSODIE (simple pour les enquêtes, double pour les fichiers administratifs entre eux)
- Choix usuels
  - Commune de résidence puis année de naissance pour les enquêtes (appariement séquentiel avec suppression des individus retrouvés)
  - Commune de résidence **&** année de naissance pour les fichiers administratifs (ou trop volumineux)
- Variables bien renseignées avec de nombreuses modalités

$\Longrightarrow$ Forte diminution de la **complexité quadratique**

III. L'appariement par plus proche écho

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

$$d(a,b) = d_{Nom}\left(Nom_a, Nom_b\right) + \ldots$$
$$\ldots + d_{Adresse}\left(Adresse_a, Adresse_b\right)$$

- Chaque sous-distance est comprise entre 0 et 1
- Les variables utilisées peuvent être le nom, le prénom, la date de naissance, le département de naissance, le sexe, la commune et les mots directeurs de l'adresse.
- Basé principalement sur la distance de Levenshtein

III. L'appariement par plus proche écho

Sélection des appariements acceptés

- Pour chaque individu, on récupère l'écho pour lequel la distance est la plus faible
- Sélection des appariements valides
  - Choix empirique d'un seuil (94.8% dans ERFS)
  - complété par des règles de gestion spécifiques (96.5% dans ERFS)
  - voire par des reprises manuelles opportunistes
- Quelques spécificités selon les sources
- Prise en compte de la dimension ménage de la source

III. Quelques résultats obtenus

	CNAM-TS (2017)	CUI (2016)	ERFS (2019)
Informations générales			
Nombre d'individus	≃310 000	≃46 000	≃88 000
Appariement exact			
Effectifs et proportion	≃262 000 (84,5%)	≃32 000 (69,6%)	≃72 100 (81,9%)
Temps de calculs	20 minutes	14 minutes	20 minutes
Appariement par plus proche écho (Commune)			
Effectifs et proportion	≃7 000 (2,3%)	≃8 000 (17,4%)	≃12 250 (13,9%)
Temps de calculs	≃26 heures	≃11 heures	≃10 heures
Appariement par plus proche écho (Année de naissance)			
Effectifs et proportion	≃27 000 (8,7%)	≃4 000 (8,7%)	NA
Temps de calculs	≃73 heures	≃20 heures	≃24 heures
Ensemble			
Effectifs et proportion	≃296 000 (95,5%)	≃44 000 (95,7%)	≃84 000 (95,8%)
Temps de calculs	≃100 heures	≃31 heures	≃34 heures

III. Un peu de probabiliste dans le déterministe ?

Une méthode présentée lors des JMS

- Récupérer des appariements dont on est certain (appariement exact ou distance très faible)
- Retirer ces individus de la base administrative (et les personnes du même foyer pour FIP)
- Apparier ces individus avec le reliquat de la base administrative
- Le taux d'appariement obtenu permet d'estimer le risque d'accepter un appariement à tort

Plan de la présentation

III. Les fichiers utilisés lors de l'enrichissement

Le fichier issu de l'appariement

- Identifiant de l'enquête / identifiant du fichier d'enrichissement
- Pour un enrichissement fiscal : la colonne de déclarant

Le fichier de tous les individus de l'enquête

- Les individus appariés, non appariés et les mineurs (non soumis à l'appariement)

Les fichiers utilisés pour l'enrichissement

- Enrichissement fiscal : POTE, PLFC
- Enrichissement social : CNAF, CNAV, MSAF et MSAV

III. La fusion avec le fichier d'enrichissement

Fusion réalisée sur l'identifiant du fichier d'enrichissement

- Spécificités pour un enrichissement fiscal
  - Récupération de toutes les données du foyer fiscal pour chaque individu apparié
  - uniquement lorsque des revenus sont présents pour le millésime
  - uniquement la dernière déclaration
  - deux déclarations en cas de décès
- Fusion simple pour un enrichissement social

III. Les traitements complémentaires

Pour l'enrichissement fiscal

- Calcul des agrégats individuels et de niveau foyer
- Attribution des revenus individuels
- Création des individus de la déclaration non enquêtés
- Traitement des décès (cumul ou imputation)
- Calcul de l'impôt et récupération de la TH
- Ajout des non appariés et des non soumis à l'appariement
- Calcul d'indicatrices (complétude et provenance)

Pour l'enrichissement social

- Calcul des prestations sociales de l'allocataire
- Imputation des prestations non présentes (complément RSA)

III. Création des livrables

- **Sélection et ordonnancement des variables**
- **Création des labels**
- **Liste des tables livrées pour l'enrichissement fiscal**
  - Table ménages (Revenus, Impôts, TH)
  - Table foyers (impôts, revenus)
  - Table individus (revenus individualisables)
- **Liste des tables pour l'enrichissement social**
  - Une table avec le montant des prestations selon la source

Merci de votre attention !