Journées de méthodologie statistique de l'Insee 2022

PROBABILISTES OU DÉTERMINISTES, DES MÉTHODES D'APPARIEMENTS AU BANC D'ESSAI DU PROGRAMME RÉSIL

Olivier Haag (*), Heïdi Koumarianos(**) , Lucas Malherbe(**)

(*) Insee, Direction des statistiques démographiques et sociales

(**) Insee, Direction de la méthodologie et de la coordination statistique et internationale

olivier.haag@insee.fr, heidi.koumarianos@insee.fr, lucas.malherbe@insee.fr

Mots-clés *(6 maximum)* : appariement, sources administratives, Fellegi et Sunter

Domaine concerné : Combinaison de sources, données administratives

Résumé

L'objet de cet article est de comparer les résultats de l'appariement des individus présents dans la source fiscale (Fichier d'Imposition des Personnes (FIP)) et l'Enquête Annuelle de Recensement (EAR) 2019, obtenus par différentes méthodes. L'objectif étant de voir si les méthodes probabilistes (non mise en production à l'Insee à ce jour) permettent d'obtenir des meilleurs résultats que des méthodes déterministes. Ceci afin de décider si le programme de Répertoires Statistiques d'Individus et de Logements (Résil) devra se doter d'un outil d'appariement probabiliste pour son moteur d'identification et d'appariement des sources administratives.

La première partie de l'article sera consacrée à une description succincte du programme qui vise à construire un système de répertoires statistiques d'individus, de ménages et de locaux d'habitation, durable et évolutif, mis à jour à partir de sources administratives diverses. Et elle décrira plus précisément pourquoi les appariements seront fondamentaux pour ce programme non seulement pour la constitution des répertoires mais aussi parce que le système de répertoires servira d'ossature au système d'information de la DSDS.

La deuxième partie de l'article s'attachera à présenter les deux sources utilisées pour cet appariement en pointant les défauts de qualité qui auront un impact sur la qualité finale des appariements. Ainsi, par exemple, on constate que sur les variables identifiantes utiles pour l'appariement (nom, prénom, dates et lieux de naissance, adresse), la source fiscale présente beaucoup moins d'anomalies (0,8 % des individus) que l'EAR (13,2 %). Pour cette dernière on note une moindre qualité pour les personnes répondant par questionnaire papier qui comportent en plus des erreurs liées à la saisie, du fait de la qualité de certaines écritures manuscrites.

La troisième partie présentera les différentes méthodes mises en œuvre :

- Rapsodie : Cet outil développé par le pôle « Revenus Fiscaux et Sociaux » à la DR Insee de Bretagne met en œuvre une méthode d'appariement déterministe. Un zoom sera fait sur l'intérêt d'un examen visuel d'un échantillon de paires issues du processus d'appariement. Il per-

met non seulement de mesurer la qualité de l'appariement (estimations des taux de faux positifs et faux négatifs) mais aussi de proposer des règles de gestion qui peuvent permettre d'améliorer l'appariement final lorsque des erreurs fréquentes sont identifiées (inversion des noms et prénoms dans l'EAR par exemple) ;

- Relais : Cet outil développé par Istat met en œuvre la méthode d'appariement probabiliste de Fellegi et Sunter. Cet outil dispose d'une IHM permettant de paramétrer les appariements (choix des distances et des seuils par exemple) et pourrait donc être utile dans le cadre de Résil. Il présente toutefois à ce jour des problèmes de performances qui empêchent l'appariement de fichiers volumineux. Le test de cet outil s'est donc fait uniquement sur les départements 48 et 35 ;
- Packages R et Python mettant en œuvre des méthodes probabilistes de type Fellegi et Sunter ainsi que des méthodes déterministes faisant intervenir du *machine learning*. L'objectif de ce test est plutôt de mesurer les performances que la méthodologie en tant que telle.

La dernière partie aura pour objectifs de comparer les résultats obtenus par ces différentes méthodes non seulement en termes de taux d'appariement mais aussi en termes de représentativité de la population des individus appariés. Ainsi, par exemple, l'appariement réalisé par Rapsodie permet de retrouver 92,7 % des individus de l'EAR dans FIP avec un taux de faux-positifs de l'ordre de 0,3 %. Toutefois, la population appariée présente un biais de couverture des moins de 20 ans et à un degré moindre des plus de 90 ans, ainsi qu'une sous-représentation des individus nés à l'étranger.

Abstract en anglais

The purpose of this article is to compare the results of the matching of individuals present in the French tax data and the Annual Census Survey (EAR) 2019, obtained by different methods. The objective is to see whether probabilistic methods (not yet in production at INSEE) provide better results than deterministic methods. This is in order to decide whether The French programme to build statistical registers of individuals and housing (Résil) should be equipped with a probabilistic matching tool for its administrative source identification and matching engine.

1 – Problématique (Résil)

Le programme de Répertoires Statistiques d'Individus et de Logements (Résil) vise à construire un système de répertoires statistiques d'individus, de ménages et de locaux d'habitation, durable et évolutif, mis à jour à partir de sources administratives diverses.

Dans ce contexte, les appariements seront fondamentaux non seulement pour la constitution des répertoires mais aussi parce que le système de répertoires servira d'ossature au système d'information de la DSDS. Il permettra en effet l'appariement avec d'autres sources : données d'enquêtes, données administratives, soit directement soit par le biais d'une identification préalable.

Ainsi, dans le but de définir l'offre d'identification proposée par Résil, il a été décidé de tester différentes méthodes d'appariement afin de choisir celle(s) qui semble(nt) la plus efficace non seulement en termes de qualité statistique mais aussi d'un point de vue de performance informatique (essentiel compte tenu des volumes à traiter).

Par ailleurs, la production de répertoires s'accompagnera d'une mesure de leur qualité, et en particulier de leur couverture afin de dépasser la situation actuelle où l'on constate des écarts entre les sources fiscales ou sociales et le recensement sans pouvoir les expliquer ni les imputer à l'une ou l'autre de ces sources. Elle sera mise en œuvre par comparaison entre la population définie par Résil et les données du recensement en utilisant la méthodologie capture-recapture et le modèle DSE (Dual System Estimation). Dans le cas de la France, l'existence de collectes annuelles de type recensement est un atout très important pour mesurer la montée en qualité d'un système de répertoires en construction. Tous les ans, les enquêtes annuelles de recensement couvrent 5 millions de logements et 9,3 millions d'habitants, ce qui fournit un échantillon de taille considérable pour servir de base à une telle opération.

Dans ce contexte, cet article comparera les résultats de l'appariement des individus présents dans la source fiscale FIP (Fichier d'imposition des Personnes) et l'EAR 2019 obtenus par différentes méthodes.

La première partie de l'article s'attachera à présenter les deux sources utilisées pour cet appariement en pointant les défauts de qualité qui auront un impact sur la qualité finale des appariements.

La deuxième partie présentera les différentes méthodes mises en œuvre ([1] Lucas Malherbe 2022):

- Rapsodie : Cet outil développé par le pôle « Revenus Fiscaux et Sociaux » de Rennes met en œuvre une méthode d'appariement déterministe. Un zoom sera fait sur l'intérêt d'un examen visuel d'un échantillon de paires issues du processus d'appariement. Il permet non seulement de mesurer la qualité de l'appariement (estimations des taux de faux positifs et faux négatifs) mais aussi de proposer des règles de gestion qui peuvent permettre d'améliorer l'appariement final lorsque des erreurs fréquentes sont identifiées (inversion des noms et prénoms dans l'EAR par exemple) ;
- Relais : Cet outil développé par Istat met en œuvre la méthode d'appariement probabiliste de Fellegi et Sunter ([2] Fellegi, I. P. and Sunter, A. B. (1969)). Cet outil dispose d'une IHM permettant de paramétrer les appariements (choix des distances et des seuils par exemple) et pourrait donc être utile dans le cadre de Résil. Il présente toutefois à ce jour des problèmes de performances qui empêchent l'appariement de « gros fichiers ». Le test de cet outil s'est donc fait uniquement sur les départements 35, 48.

- Packages R et Python mettant en œuvre des méthodes probabilistes de type Fellegi et Sunter. En plus des départements ci-dessus, ces packages ont également été testés sur le 18e arrondissement parisien et le département 68.

La troisième partie aura pour objectifs de comparer les résultats obtenus par ces différentes méthodes non seulement en termes de taux d'appariement mais aussi en termes de représentativité de la population des individus appariés.

2- Présentation des données

Comme évoqué ci-dessus, la qualité des appariements dépend non seulement du processus d'appariement proprement dit mais aussi de la qualité des fichiers en entrée.
C'est pourquoi cette partie s'attache à présenter les deux sources utilisées mais aussi la qualité des variables utilisées pour l'appariement qu'elles contiennent. Ces variables sont les suivantes :

- noms (marital et de naissance) ;
- prénoms ;
- date et département de naissance ;
- adresse de résidence.

Cette phase a donc consisté à mesurer pour chaque variable des fichiers EAR et FIP utiles pour l'appariement les différents indicateurs suivants :

- taux de non-réponse partielle ;
- taux de valeurs erronées (lorsque les modalités de la variable appartiennent à une liste ou à un intervalle) ;
- taux de valeur douteuses[1]

Le nombre de doublons (modalités identiques sur chacune des variables) a également été calculé.

2.1 – Description des données du RP

Les données utilisées pour ce test sont celles de l'Enquête Annuelle de Recensement de 2019 ([3] Insee Méthode (2005)). Elles concernent 5 millions de logements et 9 millions de personnes. Pour notre étude seules les données sur les individus de plus de 15 ans vivant en logements ordinaires ont été mobilisées (afin d'être sur un champ comparable à celui de FIP) .

Seule une partie du territoire est donc recensée chaque année (1/5 des communes de moins de 10 000 habitants et environ 8 % des logements en communes de plus de 10 000 habitants).

Seules les informations collectées relatives à l'état civil et à l'adresse des individus ont été mobilisées pour ce test.

À noter que les indicateurs « qualité » présentés ci-dessus ne permettent pas de repérer tous les problèmes. Ils ne permettent par exemple pas de repérer les erreurs engendrées par la saisie optique sur les noms et prénoms dans l'EAR et qui empêchent bon nombre d'appariements exacts. C'est pourquoi il a été décidé de distinguer trois sous-populations de l'EAR dans les différentes analyses :

- les individus qui ont répondu par internet (non soumis aux problèmes de saisie) ;
- ceux de l'EDP ayant répondu par questionnaire papier (pour lesquels le niveau de qualité de saisie des variables « nom » et « prénom » demandé est plus important) ;
- les autres (répondant par questionnaire papier et hors EDP).

[1]ex. : Nom ou prénom contenant autre chose que des lettres, nom ou prénom de moins de 3 caractères, nom ou prénom contenant une succession d'au moins 3 lettres identiques, ou que des consonnes ou que des voyelles, etc.

2.2 – Description des données FIP

Le fichier d'imposition des personnes (FIP) a pour objectif de permettre l'identification des contribuables. Ces données sont issues des informations collectées par les services des impôts (par exemple suite à un questionnaire) ou des déclarations d'impôt sur le revenu et d'impôt sur la fortune immobilière souscrite par les contribuables. Il contient des renseignements sur l'état civil et l'adresse relatifs aux personnes qui entrent dans le champ d'application de l'impôt sur le revenu, de la taxe d'habitation, de la contribution à l'audiovisuel public, de la taxe annuelle sur les logements vacants ou de l'impôt sur la fortune immobilière. Pour les individus de moins de 15 ans, seule l'année de naissance est disponible c'est pourquoi ils ont été exclus du champ de cette étude. Au final le champ retenu représente plus de 80 % des individus présents dans FIP.
Ce fichier contient l'ensemble des membres des foyers fiscaux ayant payé l'un des impôts listés ci-dessus.
Ainsi, par exemple pour un couple avec 3 enfants dont un de plus de 15 ans et encore à charge, on disposera de l'état civil complet des parents et de l'enfant de plus de 15 ans et uniquement des années de naissance des deux plus jeunes. Ces 2 derniers ne font donc pas partie du champ de ce test d'appariement.

2.3 – Comparaison de la qualité des deux sources

Tableau1 : Résumé des indicateurs sur les fichiers en entrée FIP et EAR

		FIP	EAR
Nb individus retenus		55 102 356	7 243 345
Individus ayant au moins une anomalie	Une seule anomalie	0,5 %	12,3 %
	Plus d'une anomalie	0,3 %	0,9 %
	Part des personnes nées à l'étranger	48,2 %	10,8 %
Individus avec département de naissance en erreur [2]		0,2 %	6,5 %
Individus avec prénoms douteux[3]		0 %	5,2 %

La majorité des cas d'individus ayant au moins une anomalie sont des individus dont seul le département de naissance est manquant. En revanche lorsqu'un individu cumule plus d'erreurs, c'est souvent des valeurs manquantes combinées sur le jour et mois de naissance.

D'autres résultats aboutissent au diagnostic suivant :

- Pour les personnes nées à l'étranger la date de naissance du 1er janvier est surreprésentée (3 fois plus de personnes nées le 1er janvier que les autres jours de l'année dans FIP par exemple) (sans doute lié à des défauts d'état civil dans certains pays) dans FIP comme dans l'EAR ;
- si on ventile par mode de collecte pour l'EAR, comme attendu, les réponses par internet sont légèrement meilleures (12 % des individus ayant au moins une erreur contre 15 % pour le papier).

[2] Y compris pour les personnes nées à l'étranger pour lesquels on attend un département 99.

[3] *Sont considérés comme douteux, les valeurs manquantes ou les prénoms contenant au moins un caractère spécial ou un chiffre, ou les prénoms de moins de 3 caractères, ou pour les prénoms d'au moins 3 caractères ceux qui contiennent uniquement voyelles ou consonnes ou au moins 3 lettres identiques qui se suivent.*

En conclusion, on peut dire que sur les variables utiles pour l'appariement, le fichier FIP est d'excellente qualité a priori. Celui de l'EAR l'est un peu moins mais comme peu d'individus présentent plusieurs problèmes les conséquences devraient être limitées sur l'efficacité du processus d'appariement. Rappelons toutefois que ces tableaux ne prennent pas en compte les potentielles erreurs de saisie sur noms et prénoms. Sur ce point, un test a été mené pour mesurer le taux d'appartenance des prénoms de l'EAR à la liste des prénoms présente sur le site Insee.fr. On constate que plus de 10 % des prénoms de l'EAR ne se retrouvent pas dans la liste !

3 – Présentation des méthodes

3.1 – Présentation de Rapsodie

Rapsodie est un outil d'appariement et d'enrichissement développé par le pôle Revenus Fiscaux et Sociaux de la DR de Rennes ([4] P. Jabot and alii (JMS2010))**.**

Il repose sur le principe de tours d'appariement successifs afin d'optimiser les temps de traitements. Au départ il y avait en théorie plus de 750 tera de paires à étudier.

1. Le 1er permet d'apparier les individus des deux fichiers qui ont exactement les mêmes noms, prénoms, sexe, date de naissance et commune de résidence. Cette étape permet d'apparier près de 70 % des individus de l'EAR et donc de diviser par 3 le nombre de paires à étudier ;
2. Le 2e permet d'apparier parmi les individus non appariés au tour précédent, selon une méthode de type « plus proche écho » avec comme clé de blocage la commune de résidence. Ainsi sont appariés deux individus qui résident dans la même commune dans les deux fichiers et dont la somme pondérée[4] des distances[5] des écarts entre chaque variable (nom, prénom, date et département de naissance et 2 premiers mots significatifs de l'adresse) est inférieure à un seuil défini *a priori*. L'ajout de règles de gestion[6] permet de considérer certains individus comme appariés même si la somme de leurs distances dépasse le seuil. À noter que si pour un individu donné il existe plusieurs paires dont la distance est inférieure au seuil, seule la paire ayant la distance minimale est retenue. L'utilisation de « clés de blocage » permet de limiter le nombre de paires à analyser en le divisant par plus de 5 000.
3. Le 3e permet d'apparier parmi les individus non appariés aux tours précédents ceux qui ont exactement les mêmes noms, prénoms, dates de naissance et sexes, et des communes de résidence différentes.

Pour mémoire, la qualité du processus d'appariement ne se mesure pas uniquement par le biais du taux d'appariement : en effet on peut avoir apparié 100 % d'un fichier mais en ayant commis 30 % d'erreurs. Il convient également de prendre en compte les populations suivantes ([5] J DOIDGE and alii)

Tableau 2 : Types de paires rencontrés dans un appariement imparfait

	Vraies paires (mêmes individus)	**Fausses paires (individus différents)**
Acceptées dans le fichier résultat	**Vrai positif** (de l'ordre de 7 millions)	**Faux-positif** (à minimiser)
Refusées dans le fichier résultat	**Faux-négatif** (à minimiser)	**Vrai négatif** (de l'ordre de 400 millions de millions)

[4] Les distances des noms et prénoms ont été pondérées par 1,5 et les autres distances par 1.

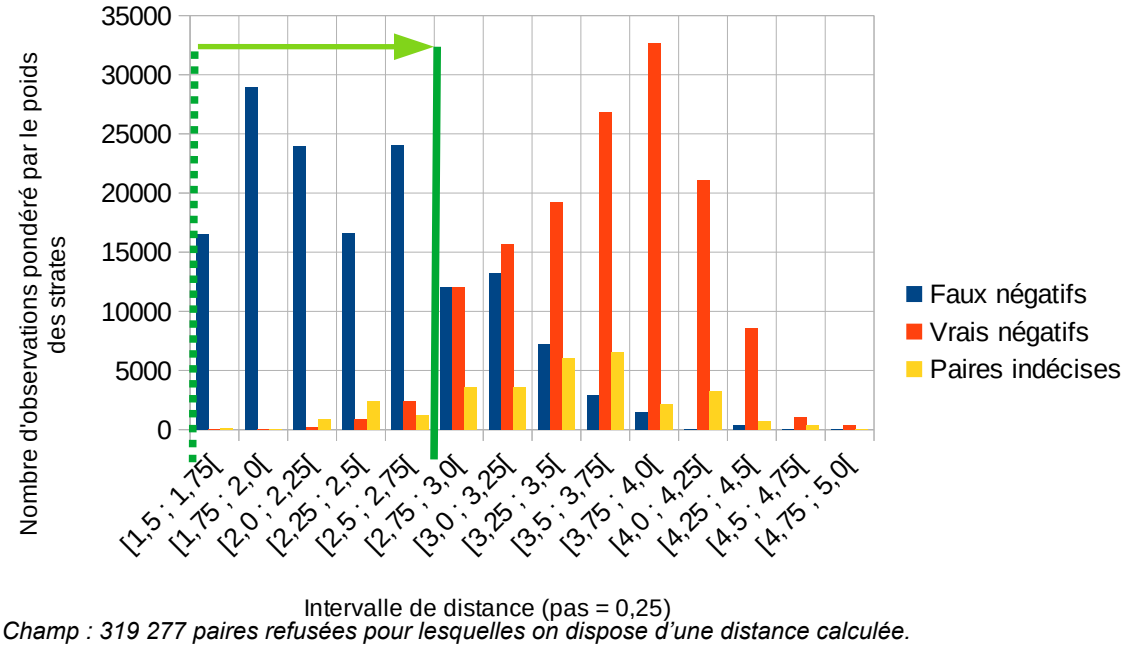
[5]La distance de Levenshtein est utilisée pour les libellés (nom, prénom, adresse), et pour les autres variables (éléments de date de naissance et code géographique) la distance vaut 0 s'ils sont égaux et 1 sinon.

[6]Exemple de règle de gestion, : prénom EAR = nom FIP, prénom FIP = nom EAR et les autres variables exactes (date de naissance, adresse)

En outre, il est également très important d'analyser la représentativité de population résultant de l'appariement afin d'en mesurer les biais éventuels. On pourra ainsi les corriger avant de commencer à l'utiliser pour produire des statistiques.
La mesure de la qualité de l'appariement entre FIP et l'EAR et l'analyse des résultats ont été réalisées par Maxime Huguet lors de son stage de fin d'étude du Céfil. Il a notamment analysé visuellement 3 000 paires (1 000 acceptées et 2 000 refusées), afin non seulement d'estimer la qualité des appariements (estimation des faux positifs et faux négatifs) mais aussi pour l'optimiser en proposant un seuil d'acceptation des paires plus efficace et de nouvelles règles de gestion.

La sélection des paires à contrôler visuellement s'est faite de façon aléatoire par un sondage stratifié selon la distance des paires, en sur-représentant les distances qui sont au voisinage des seuils de décision. En plus des caractéristiques des individus dans les deux sources utilisées pour l'appariement, afin de faciliter le diagnostic, le fichier de contrôle contenait en plus la composition du ménage des individus des paires dans les deux sources (liste des individus habitant dans le même logement de l'EAR et liste des individus du même foyer fiscal de FIP pour les individus de la paire).
Cette stratégie permettait de disposer de plus de paires proches du seuil et ainsi d'optimiser ce dernier à partir d'un nombre suffisant de paires analysées.
Les résultats présentés ci-dessous sont pondérés. Le taux de faux positifs est nul au seuil de 1,6 et très faible (0,3 %) au seuil de 2,7. Pour les faux négatifs, les résultats (pondérés) de ces contrôles figurent dans le graphique ci-dessous :

Graphique 1 : Distribution du type de paires selon leur distance[7]

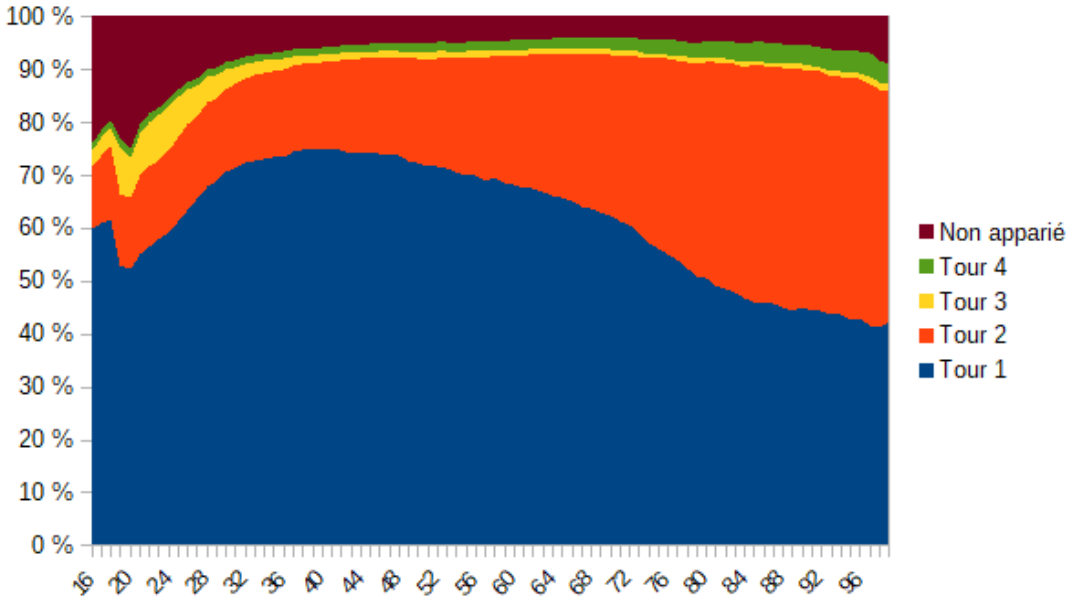


Champ : 319 277 paires refusées pour lesquelles on dispose d'une distance calculée.

Ce graphique permet d'identifier le seuil optimal pour l'acceptation des paires. En effet, avant 2,75 on se rend compte que le taux de faux négatifs est vraiment supérieur au taux de faux positifs et qu'au-delà l'éventuel gain en appariement s'accompagne d'un nombre de faux positifs très important. Cette analyse a donc permis de fixer ce seuil à 2,75 (alors qu'il avait été fixé à 1,6 à l'origine) ; cela ajoute un « quatrième tour » aux trois tours définis précédemment. Cette décision conduit à assumer un taux de faux-positifs de l'ordre de 0,3 %, mais permet par ailleurs de gagner 2 points sur le taux d'appariement. Au final avec ces paramètres, 92,7 % des individus de l'EAR (restreints aux individus de plus de 15 ans en logement ordinaire) sont appariés avec un individu de FIP.

[7]Les paires indécises représentent des cas où l'examen visuel ne permettait pas de trancher de façon certaine entre faux positif ou négatif et ce malgré l'information complémentaire.

Graphique 2 : les taux d'appariement entre EAR et FIP par tour et par âge.



Champ : Personne de plus de 15 ans de l'EAR 2019

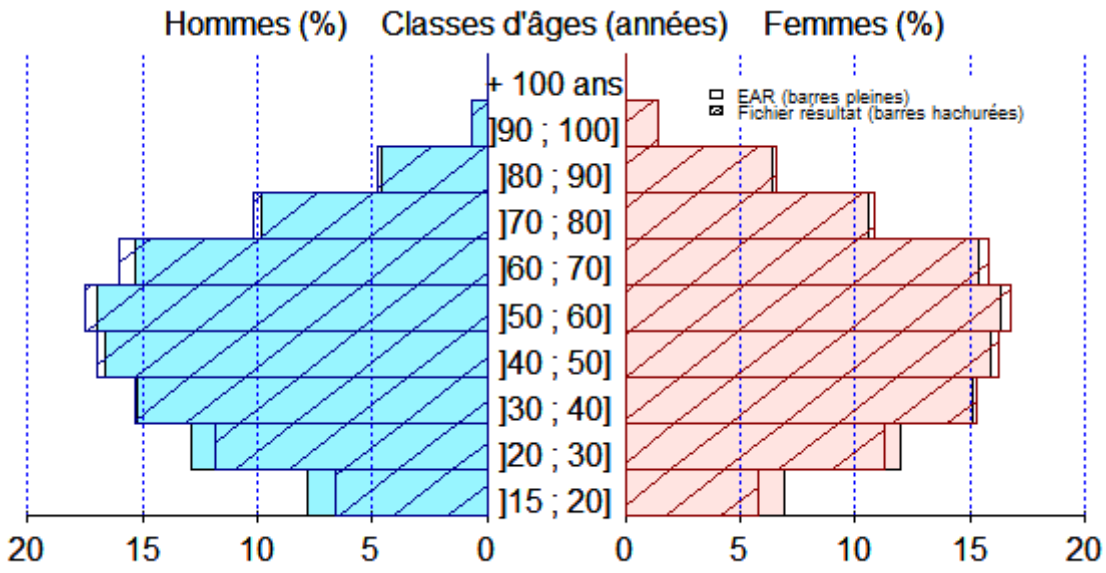
La première conclusion est que la population appariée est biaisée, car elle sous-représente les jeunes individus de moins de 30 ans et notamment ceux entre 18 et 22 ans. Si on souhaite utiliser cette population pour produire des statistiques il conviendra par exemple de la caler sur la structure par âge de la population française des plus de 15 ans.

On constate également que l'appariement exact (tour 1) permet d'atteindre un taux d'appariement de 65 % ce qui est déjà un bon résultat. En revanche on note que les plus jeunes et les plus âgés ont des taux d'appariement exact moins bons. Les raisons de ces moindres performances ne sont a priori pas exactement les mêmes.

- Pour les plus âgés, cette baisse peut principalement provenir de leur plus forte propension à répondre par papier qui se traduit par plus de fautes dans les patronymes qui empêchent l'appariement exact. Ce problème est en partie corrigé par les tours 2 et 4 (tours d'appariement approchés, avec des seuils respectifs d'acceptation des paires de 1,6 et 2,75) qui permettent d'apparier des individus ayant des caractéristiques proches mais pas obligatoirement identiques.
- Pour les plus jeunes, les raisons sont de deux ordres :
  - leur localisation peut différer d'une source à l'autre[8] ce qui empêche un appariement exact. Ainsi, on voit que l'appariement exact France entière qui supprime la contrainte sur la commune de résidence (tour 3 dans ce graphique) permet de récupérer davantage de jeunes.
  - le biais de couverture de la base FIP sur les 18/20 ans qui se retrouve donc dans la population appariée.

[8]Les jeunes étudiants peuvent être recensés dans leur logement étudiant mais sont rattachés au foyer fiscal de leur parent et donc localisés à l'adresse des parents dans FIP.

Le premier constat de biais est également visible sur la pyramide des âges ci-dessous.

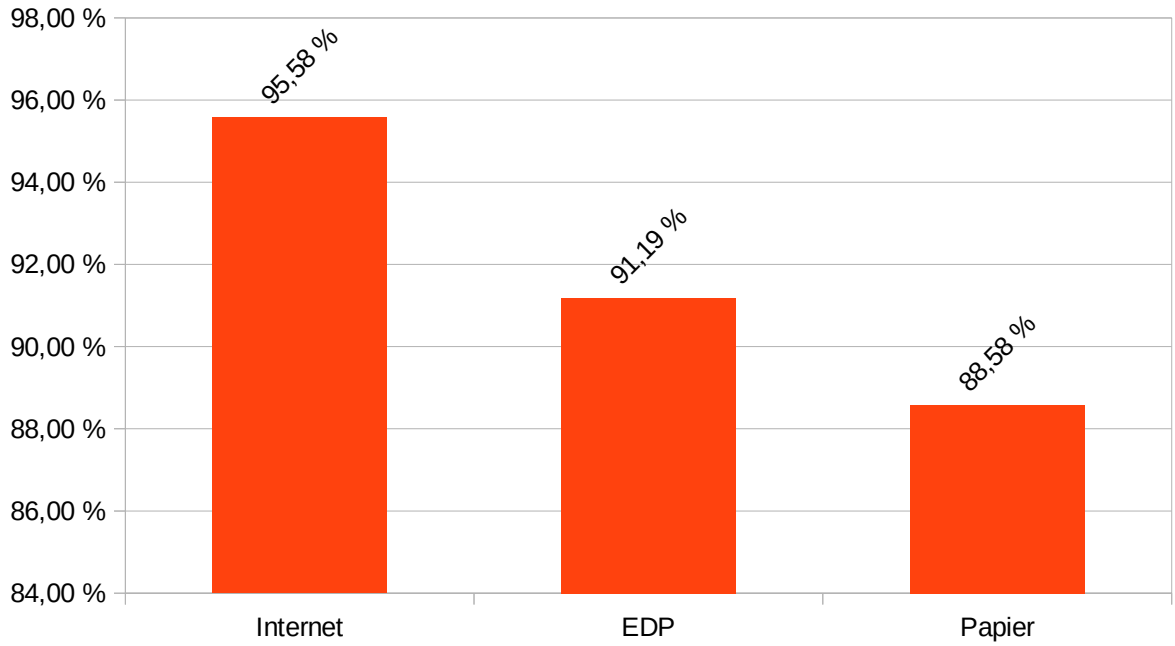


Graphique 3 : Pyramides des âges (en %) de l'EAR et de la population appariée
Champ : 7 243 345 individus de l'EAR

NB. : Les résultats sont non pondérés et ne concernent donc que les individus répondants de l'EAR)

Comme déjà évoqué par ailleurs, le mode de collecte a un impact sur le taux de l'appariement, comme le montre le graphique suivant :

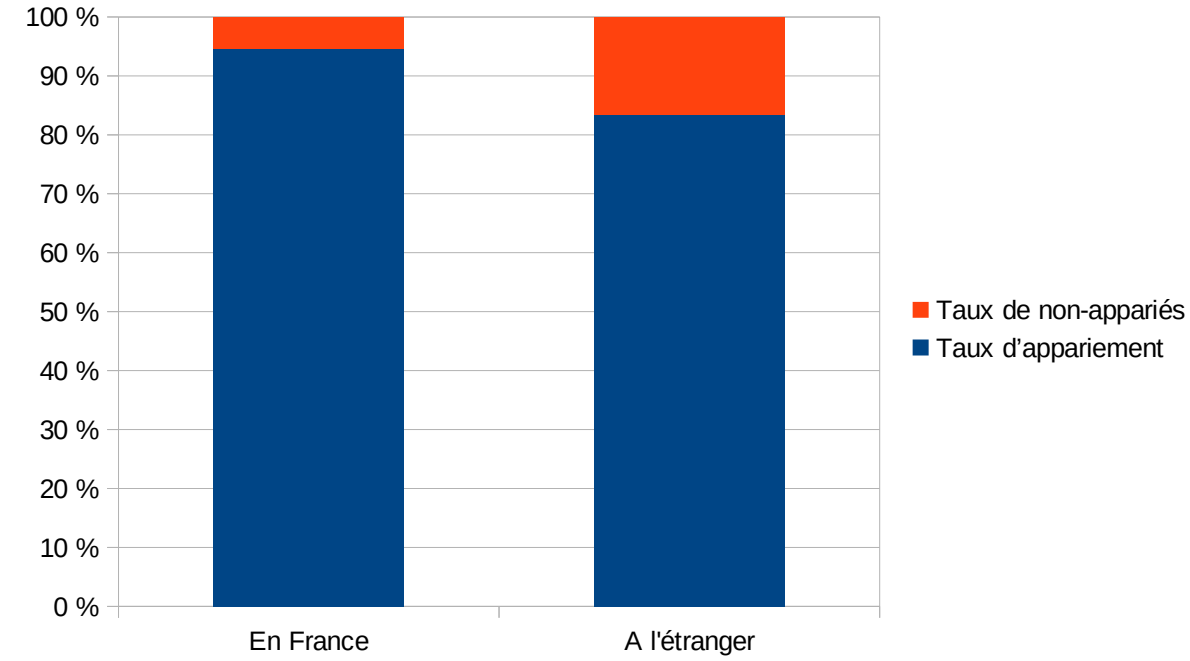
Graphique 4 : Taux d'appariement des plus de 15 ans de l'EAR avec FIP selon le mode de collecte de l'EAR



Champ : 7 243 345 individus de l'EAR

On constate par ailleurs un biais non négligeable lié au pays de naissance puisque que 95 % des individus de l'EAR nés en France sont appariés contre 83 % pour ceux nés à l'étranger. Ceci s'explique en partie par les imprécisions sur la date et lieu de naissance et potentiellement également sur les fautes plus fréquentes sur les patronymes.

Graphique 5 : Taux d'appariement des plus de 15 ans de l'EAR avec FIP selon le lieu de naissance



Champ : 7 243 345 individus de l'EAR

3.2 – Présentation de Relais

Relais est un outil d'appariement développé par l'institut national de statistiques italien Istat.
L'outil est doté d'une interface utilisateur graphique qui permet à un utilisateur non expert de mettre en œuvre un appariement, selon une méthode déterministe ou probabiliste. ([6] (Cibella et alii, 2010).
Ainsi Relais permet de télécharger ses données au sein de l'outil, de sélectionner ses variables d'appariement, de choisir une méthode de réduction du problème (blocage par exemple) ainsi que les fonctions de comparaison des enregistrements. La méthode probabiliste permet de déterminer les seuils de conservation/rejet des paires, et propose une estimation d'indicateurs « qualité » tels que le rappel[9] et la précision[10].

La méthode retenue pour cette étude est la méthode probabiliste, suivant la théorie de Fellegi-Sunter.
Les données utilisées sont celles de l'EAR et de FIP, déjà normalisées par l'outil Rapsodie, pour les départements de la Lozère et de l'Ille-et-Vilaine.

Relais est fortement limité sur les volumes de données mobilisés. Les données sur le département de la Lozère ont pu être traitées au cours d'un même processus d'appariement.
En revanche, les contraintes de taille ont pesé plus fortement pour réaliser l'appariement de données pour l'Ille-et-Vilaine, et ont conduit à découper le fichier en 7 : 5 fichiers découpés selon un critère communal, et 2 autres découpés selon un critère communal et d'année de naissance.
Ce découpage induit une sous optimalité pour le processus : d'une part l'estimation des probabilités dans le cadre de la théorie de Fellegi-Sunter est réalisée plusieurs fois, et peut donc différer d'un

[9]Le rappel représente la proportion de vraies paires correctement identifiées parmi l'ensemble des vraies paires.
[10]La précision représente la proportion de vraies **paires** parmi les paires retenues

fichier à l'autre, d'autre part le découpage sur l'année de naissance constitue alors de fait un blocage partiel supplémentaire.

Pour chacun de ces fichiers, un blocage a été effectué sur la commune de résidence.
Les critères de comparaison retenus sont les suivants :
- distance de Levensthein (normée) supérieure à 0,9 pour le nom de naissance ;
- distance de Levenshtein (normée) supérieure à 0,8 pour le prénom ;
- égalité pour les variables d'année de naissance, de mois de naissance, de jour de naissance et de département de naissance ;
- distance de Levenshtein normée supérieure à 0,8 pour les deux premiers mots directeurs de l'adresse.
NB : au sein de Relais, on ne peut pas préciser d'union de conditions (comme par exemple égalité des noms de naissance ou égalité des noms d'usage).
Le seuil d'acceptation des paires est une probabilité de 0,9.

Relais détermine ensuite les motifs de concordance possibles, correspondant aux critères statistiques définis. Ainsi, la validation de tous les critères de comparaison n'est pas indispensable pour être une paire retenue.
Relais retient bien évidemment les paires remplissant tous les critères simultanément, mais également certaines ne remplissant que 4, 5 ou 6 de ces critères.

Pour le département de la Lozère, 8 575 paires sont retenues à l'issue d'un processus d'appariement probabiliste.
Relais fournit des indicateurs, fondés sur les probabilités de Fellegi-Sunter :

- la fréquence estimée d'un match est de 0,00107,
- la précision $P=\frac{Nombre\ de\ vraies\ paires\ acceptées}{Nombre\ de\ paires\ acceptées}=\frac{Vrais\ positifs}{Vrais\ positifs+Faux\ positifs}$ est de 0,99809 (soit une estimation des vrais positifs à 99,8 %, et de faux positifs à 0,2 % parmi les paires retenues),
- et le rappel $R=\frac{Nombre\ de\ vraies\ paires\ acceptées}{Nombre\ de\ vraies\ paires}=\frac{Vrais\ positifs}{Vrais\ positifs+Faux\ négatifs}$ est de 0,92656. Parmi l'ensemble inconnu de vraies paires, 92,6 % ont été retrouvées.

3.3 – Présentation de la librairie Python *recordLinkage (appelée « python » dans le chapitre 4)*

recordLinkage est une librarie Python, c'est donc un outil *open source.* Il permet de mettre en œuvre des appariements déterministes comme probabilistes. Dans le cadre de cet article, c'est la méthode probabiliste, issue du cadre de Fellegi-Sunter, qui a été testée.

La librairie propose un ensemble de fonctions permettant de mettre en œuvre les différentes étapes d'un appariement, comme le blocage, le calcul de distances et bien sûr la classification des paires.

Comme pour Relais, ce sont les données normalisées par Rapsodie qui ont été utilisées.

Pour la Lozère, la stratégie d'appariement retenue fait intervenir trois étapes. Les individus appariés au cours d'une étape ne sont plus considérés pour les suivantes.

1. La première étape consiste en un appariement exact sur le nom, le prénom, le sexe, la date de naissance et la commune de résidence.
2. La deuxième étape est un appariement probabiliste avec blocage sur la commune de résidence.
3. La troisième étape est un appariement probabiliste avec blocage sur l'année de naissance.

Pour l'Ille-et-Villaine, appliquer la même stratégie que pour la Lozère conduit à une consommation excessive de mémoire vive, dépassant 250 Go. Une étape a donc été ajoutée après l'appariement exact afin de réduire le nombre total de paires comparées. La stratégie est la suivante :

1. appariement exact ;
2. appariement probabiliste avec blocage sur la commune de résidence et l'année de naissance, où ne sont donc comparés que les individus qui résident dans la même commune et qui sont nés la même année ;
3. appariement probabiliste avec blocage sur la commune de résidence uniquement ;
4. appariement probabiliste avec blocage sur l'année de naissance uniquement.

Même en appliquant cette stratégie, la consommation de mémoire vive tutoie tout de même les 200 Go lors de l'étape la plus consommatrice, la 3^{e}, lors de laquelle plus de 400 milliards de paires potentielles sont traitées. Ce volume est majoritairement dû à la présence de grandes communes dans le département, particulièrement Rennes.

Comme pour Relais, l'algorithme probabiliste de classification implémenté dans la librairie *recordLinkage* fonctionne avec des variables binaires. Cela ne contraint pas à faire uniquement des comparaisons exactes, il reste possible d'utiliser des comparaisons floues, condition de décider d'un seuil pour les « binariser » par la suite. Les règles de comparaison retenues sont les suivantes :

- nom, prénom et les 2 mots directeurs de l'adresse : similarité de Jaro-Winkler avec un seuil à 0,92 ;
- commune de résidence et jour, mois, année et commune de naissance : comparaison exacte

L'algorithme fournit en sortie une probabilité pour chaque paire issue de l'étape de blocage, c'est-à-dire chaque paire pour laquelle on a effectué des comparaisons et calculé des distances. Le seuil d'acceptation des paires a été fixé ici à 0,5, mais il est possible d'ajuster ce seuil suite à l'analyse visuelle d'un échantillon de paires.

4– Comparaison des résultats

4.1 – Méthodologie de comparaison

La comparaison entre les 3 méthodes a été faite de façon détaillée sur 2 départements (48 et 35).

À noter que pour des raisons de performance, les appariements n'ont pu être menés de la même façon dans les 2 départements comme évoqué dans les points 3.2 et 3.3 ci-dessus.

Le principe de comparaison a été le suivant :

- Comparaison des taux d'appariement des 3 méthodes
- Identification des appariements différents
- Sélection d'un échantillon de 1000 paires appariées de façons différentes
- Mesure des faux positifs à partir de l'analyse visuelle de l'échantillon précédent.

4.2 – Principaux résultats

	Département 48	Département 35
Population de plus de 15 ans EAR	10 127	130 950
Population de plus de 15 ans FIP	62 823	874 304
Appariés de façon exacte	6 239	96 697
Appariés de façon exacte hors du département[11]	622	3 087
Taux d'appariement « Rapsodie » (en %)	91,3	94,8
Taux appariement « Relais » (en %)	92,1	93,7
Taux d'appariement « Python » (en %)	92,4	95,1
Faux positif « Rapsodie » (en %)	0,02	0,05
Faux positif « Relais » (en %)	0,3	0,03
Faux positif « Python » (en %)	0,3	0,7
Taux d'appariement « Rapsodie » corrigé (en %)[12]	91,3	94,7
Taux appariement « Relais » corrigé (en %)	91,9	93,6
Taux d'appariement « Python » corrigé (en %)	92,2	94,4

Ces premiers résultats montrent que les trois méthodes donnent des résultats très proches avec un très bon niveau de qualité (un taux de faux positifs faibles).

Les méthodes probabilistes donnent a priori de meilleurs résultats (cas du département 48 sur lequel ces méthodes ont pu être mises en œuvre sans contrainte). Toutefois, lorsqu'il est nécessaire de les contraindre (cf. point 3.2 et 3.3 ci-dessus) à cause de la taille des fichiers en entrée (cas du département 35) on constate alors que les résultats sont au même niveau voire un peu moins bons qu'une méthode déterministe dont les paramètres ont été optimisés à la suite de contrôles visuels.

À noter qu'une telle optimisation des seuils des méthodes probabilistes n'a pas été mise en œuvre, on peut donc raisonnablement penser que le taux de faux positifs de ces méthodes pourrait être réduit mais au détriment du taux d'appariement global.

Un premier examen des faux positifs de la méthode « Python » avait quand même permis d'en réduire le nombre en augmentant le seuil d'acceptation des paires.

[11]Ces individus ont été retirés du calcul des taux d'appariement, car ils ne pouvaient pas être trouvés par les relais et la méthode python qui n'ont recherché des individus que dans les fichiers du 48 ou 35. Ce qui pose problème quand on voit que 10 % des appariés de Lozère l'ont été hors du département

[12]Le taux corrigé est obtenu en supprimant les faux positifs.

4.3 – Analyse des faux positifs

Paire retenue par			Paires correctes (en %)		Paires incorrectes (en %)		Paires indécises (en %)		Total	
Rapsodie	Relais	Python	Dép. 48	Dép .35	Dép. 48	Dép. 35	Dép. 48	Dep. 35	Dép. 48	Dep. 35
Oui	Oui	Non	80.0	90.0	20.0	3.0	0.0	7.0	5	227
Oui	Non	Oui	100.0	98.7	0.0	0.5	0.0	0.8	44	1155
Oui	Non	Non	97.7	84.0	2.3	6.5	0.0	9.5	43	851
Non	Oui	Oui	76,2	93.0	11,9	5.0	11,9	2.0	143	424
Non	Non	Oui	52.6	26.3	31.6	61.0	15.8	12.7	19	1465
Non	Oui	Non	72.0	82.3	20.0	9.4	8.0	8.3	25	113
Oui	Oui	Oui	100.0	100.0	0.0	0.0	0.0	0.0	8591	126 151

Champ : analyse des échantillons de paires différentes

Comme on pouvait s'y attendre, les taux de faux positifs sont les plus importants pour les paires qui ne sont identifiées que par une seule méthode d'appariement.

Une analyse des faux positifs de la méthode « Python » (61 % de faux positif pour les paires identifiées uniquement par cette méthode) montre qu'ils pourraient être en partie supprimés en augmentant le seuil d'acceptabilité des paires, ceci aura bien entendu en parallèle un impact sur le taux d'appariement final. En outre, on constate aussi qu'en général les faux positifs ont une année de naissance erronée. Ce constat pourrait également être pris en compte dans la spécification du modèle d'appariement probabiliste.

4.4 – Analyse des faux négatifs

Parmi les individus appariés selon Python mais pas par Rapsodie, dans 2/3 des cas pour le département 48, il s'agit de personnes résidant dans 2 villes différentes entre l'EAR et le fichier Fip. Ceci vient du fait que la méthode Rapsodie mise en œuvre utilisait la commune comme clé de blocage ! En outre même si in fine Rapsodie réalisait une dernière étape d'appariement exact France entière, elle ne permettait pas l'appariement de deux personnes vivant dans deux communes différentes dans les 2 sources et ayant des caractéristiques proches (faute de frappe dans le nom par exemple). Pour régler ce problème, il suffirait de faire un 2^{e} tour de Rapsodie en prenant par exemple l'année de naissance comme clé de blocage. Cette amélioration a été simulée sur le département 48 et aurait permis d'atteindre un taux d'appariement de l'ordre de 92.2 %. L'utilisation de cette seconde clé de blocage aurait donc permis de se rapprocher des résultats des méthodes probabilistes.

5– Conclusion

Ces travaux ont confirmé plusieurs éléments déjà observés par d'autres études du même type sur d'autres jeux de données :

- l'intérêt de l'analyse visuelle d'échantillons de paires acceptées ou refusées afin de mesurer la qualité de l'appariement et d'en améliorer sa qualité en optimisant ses paramètres.
- les difficultés de mise en œuvre des méthodes probabilistes sur gros fichiers. Ces problèmes sont non seulement dus au volume total d'enregistrements du fichier mais aussi à la taille maximale des sous-populations issues du blocage qui définit le plus gros produit cartésien à mettre en œuvre.
- Un algorithme de type déterministe bien paramétré permet d'avoir des résultats proches de ceux d'un algorithme probabiliste sans contrainte de taille de fichier. En outre, son fonctionnement est plus facile à vulgariser.

Cette expérimentation montre aussi le fait que des méthodes différentes peuvent être complémentaires et utiles pour juger de leur qualité respective. Une paire trouvée par différentes méthodes a moins de chance d'être un faux positif qu'une paire trouvée par une seule méthode.
Ce constat nous conduit à envisager dans le cadre de Résil de prévoir de mettre en place en plus de l'outil d'appariement déterministe en production, un outil d'appariement probabiliste qui pourrait fonctionner sur un échantillon et permettre de s'assurer de la qualité des résultats obtenus en production.
Enfin, il a également été décidé de mettre en place dans le cadre de Résil, une interface de mesure de la qualité des appariements par annotation de paires. Une telle interface aura en effet plusieurs intérêts :

- mesurer la qualité de l'identification (estimation de la précision et du rappel évoqué ci-dessus)
- optimiser le processus (en améliorant les paramètres ou en proposant de nouvelles règles de gestions au vu des résultats des contrôles visuels.
- disposer de paires étiquetées qui pourraient servir à l'apprentissage d'un modèle de type machine learning supervisé qui pourrait être envisagé pour la classification des paires ([7] L. Midy 2021).

Bibliographie

[1] Lucas Malherbe (, JMS 2022) « Méthodologie des appariements de données individuelles
[2] Fellegi, I. P. and Sunter, A. B. (1969), A theory for record linkage. Journal of the American Statistical Association 64, 1183–1210.
[3] Insee méthode (2005) "Pour comprendre le recensement de la population"
[4]Patrick JABOT, Pierre-Éric TREYENS (JMS 2010) "Appariement d'enquêtes avec des données administratives sociales ou fiscales "
[5] James Doidge, Peter Christen, Katie Harron, 2020. *Quality assessment in data linkage*. 2020. ONS.
[6] Cibela and alii (2010) "From theory to practice: the software RELAIS as a solution for record linkage" 2010
[7] Loïc Midy (courrier des statistique n°6 2021) "Un outil d'appariement sur identifiants indirects : l'exemple du système d'information sur l'insertion des jeunes "