Journées de méthodologie statistique de l'Insee 2022

IMPACT DU NETTOYAGE DES DONNÉES SUR LA QUALITÉ D'UN APPARIEMENT

Heidi KOUMARIANOS (*)

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

heidi.koumarianos@insee.fr

Mots-clés : appariement, nettoyage de données, qualité

Domaine concerné : Combinaison de sources, données administratives

Résumé

L'Insee souhaite favoriser l'utilisation massive et plus industrialisée des sources administratives et ce notamment dans la sphère des statistiques socio-démographiques. C'est dans ce cadre notamment que se place par exemple le programme de Répertoires Statistiques d'Individus et de Logements (RéSIL) qui vise à construire un système de répertoires d'individus, de ménages et de locaux d'habitation, durable et évolutif, mis à jour à partir de sources administratives diverses.
Aussi, les appariements entre sources vont être encore plus au cœur du système d'information de l'Insee et il convient donc de mettre en place les processus les plus automatisés et les plus robustes possibles afin d'obtenir les gains d'efficience et de qualité souhaités.

Un processus d'appariement se décompose en plusieurs phases :

- le nettoyage des données ;
- la réduction de la taille du problème (blocage ou indexation) ;
- la comparaison des unités au sein des paires ;
- la classification des paires retenues.

Cet article se concentre sur la première de ces phases, en cherchant à mesurer son impact sur les suivantes.

Lors d'un processus d'appariement, l'étape de nettoyage des données en entrée revêt une grande importance : elle s'assure du respect des formats de données, de la comparabilité des informations, et permet d'adapter les règles d'appariement aux caractéristiques des jeux de données.
Identifiée comme cruciale et comme la plus chronophage par les statisticiens en charge d'opérations d'appariement, elle fait rarement l'objet de mesures d'efficacité.

On présente d'abord les enjeux de l'étape de standardisation et de nettoyage de données individuelles. Cette étape s'appuie sur une analyse préalable de la qualité des données, notamment par le biais de contrôles formels, ou de conformité à un référentiel (géographique par exemple) et/ou d'appartenance à des nomenclatures.

Après ce rappel théorique, une application est proposée sur différents jeux de données individuelles (données fiscales et données de l'enquête annuelle de recensement notamment), pour essayer de différencier l'impact de la normalisation sur des données de qualité et d'origine (données administratives versus données d'enquêtes) différentes. Plusieurs scénarios sont comparés, en mobilisant différents niveaux de nettoyage : une première normalisation minimale pour rendre possible les comparaisons, une normalisation plus classique (suppression des caractères non attendus, conformité à un référentiel géographique), et un nettoyage de données plus poussé (exclusion d'enregistrements de trop faible qualité, modification plus importante des chaînes de caractères pour éliminer civilités, relations…).

Dans chacun des cas étudiés, on essaie de mesurer la charge de travail liée à la normalisation, et son impact sur les données (part de données modifiées, enregistrements exclus…).

On analyse ensuite son impact sur les phases ultérieures du processus d'appariement.

On s'attache enfin à essayer de quantifier l'impact de la normalisation et du nettoyage des données sur le résultat de l'appariement (nombre d'enregistrements appariés, de façon exacte ou floue, impact sur la déformation des distributions, estimations de faux positifs et faux négatifs), en lien avec la méthode d'appariement retenue (probabiliste versus déterministe).

L'Insee souhaite favoriser l'utilisation massive et plus industrialisée des sources administratives et ce, notamment dans la sphère des statistiques socio-démographiques. C'est dans ce cadre notamment que se place le programme de Répertoires Statistiques d'Individus et de Logements (RéSIL) qui vise à construire un système de répertoires d'individus, de ménages et de locaux d'habitation, durable et évolutif, mis à jour à partir de sources administratives diverses.
Les différentes sources portent sur des thématiques différentes, et les appariements entre sources vont être encore plus au cœur du système d'information de l'Insee, afin de produire des données riches pour les analyses statistiques et économiques.
Il convient donc de mettre en place les processus les plus automatisés et les plus robustes possibles afin d'obtenir les gains d'efficience et de qualité souhaités.

En matière de méthodologie d'appariement, l'accent est souvent mis sur les étapes de comparaison des paires d'enregistrements et de classification de ces paires (choix des mesures de similarité et de la pondération des différentes variables, calcul d'un score ou d'une distance, sélection déterministe ou probabiliste des paires).
Mais dans la pratique, les personnes expérimentées soulignent l'importance de l'étape de préparation des données, allant même jusqu'à indiquer que le succès de l'appariement repose très majoritairement sur cette première étape. Celle-ci est également perçue comme la plus chronophage. Ce constat a suscité la curiosité, et un premier pas pour mesurer l'impact de cette étape.

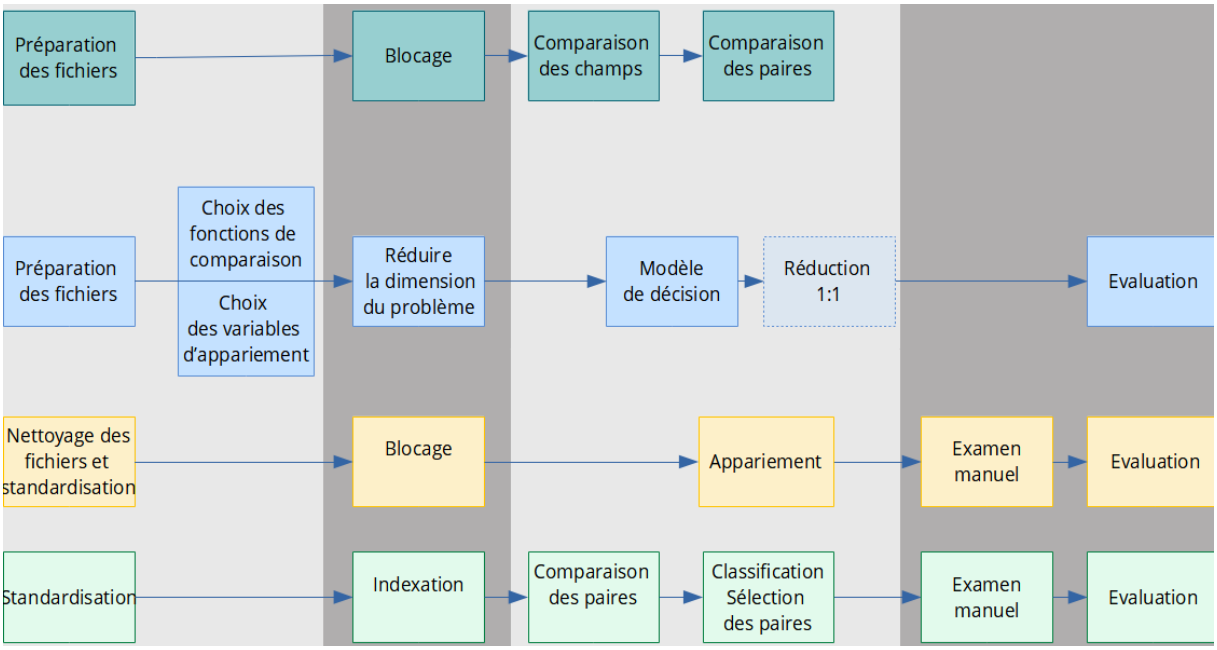
On s'intéressera ici à un appariement de données individuelles, issues de sources de nature et de qualité différentes.

1. Processus d'appariement, bref rappel sur les étapes et les mesures de qualité

1.1. Les étapes d'un processus d'appariement :

Les appariements font l'objet d'une littérature importante, et si la finalité est simple, la mise en œuvre peut être très variée.
Plusieurs approches systématiques ont été proposées, et convergent vers un découpage agrégé des différentes phases d'un processus d'appariement relativement cohérent.

Schéma 1 : Propositions de découpage en étapes d'un appariement de données



Sources :
[8](AUGUSTINE et alii, 2018), [6] (Tuoto et alii, 2010), [9] et [10] (ABS, 2018), [3] (Christen, 2012)

Si la formulation varie un peu, les étapes proposées par Christen en dernière ligne du schéma, se retrouvent de façon systématique :

- le nettoyage des données, qui comprend la standardisation, ou normalisation ;
- la réduction de la taille du problème (blocage ou indexation), i. e. la création d'un ensemble de paires potentielles, souvent plus petit que l'ensemble théorique des paires ;
- la comparaison des unités au sein des paires, qui comprend la comparaison des caractéristiques individuelles ;
- la classification des paires retenues, i.e. la sélection ou rejet des paires en fonction de critères de similarité, et parfois d'unicité.

Une étape postérieure d'évaluation de la qualité de l'appariement est souvent suggérée.

La mise en oeuvre de ces différentes étapes peut varier, mais le schéma mental qu'elles décrivent se retrouve dans tous les processus d'appariement.

Lors d'un processus d'appariement, il est également fréquent de décomposer la séquence d'appariement en deux étapes, la première étant un appariement exact sur l'ensemble des variables, après normalisation. Cette première étape est souvent perçue comme plus qualitative, en raison du caractère strict des comparaisons (elle n'est malgré tout pas une garantie absolue contre les faux positifs). Elle permet également de réduire la taille du problème.
En effet, un appariement est un problème de taille N^2 (théoriquement l'espace de comparaisons des paires est le produit cartésien des données), et très gourmand en ressources, ce qui conduit à proposer des solutions pour limiter la taille du problème ou le découper en problèmes plus petits.

Cet article se concentre sur les impacts de la première étape sur la suite du processus, et notamment la qualité de l'appariement.

1.2. Comment mesurer l'impact du nettoyage des données ?

Lors d'un processus d'appariement, des mesures classiques sont utilisées pour en évaluer la qualité : le taux d'appariement est la variable la plus immédiate à laquelle on pense, mais n'est cependant pas suffisante. En effet, on peut vouloir privilégier un « meilleur » appariement à un taux d'appariement plus élevé. Aussi, d'autres mesures caractérisent les paires potentielles (issues du produit cartésien des données) en quatre sous populations correspondant au statut des paires à l'issue de l'appariement (tableau 1).

Tableau 1 : les quatre statuts des paires d'enregistrements pour un appariement entre deux fichiers de taille N

	Vraies paires (mêmes individus)	**Fausses paires (individus différents)**
Paires acceptées dans le fichier résultat	Vrai positif (de l'ordre de N)	Faux positif (à minimiser)
Paires refusées dans le fichier résultat	Faux négatif (à minimiser)	Vrai négatif (de l'ordre de N^2-N)

De ces quatre sous populations, on déduit deux mesures principales de la qualité de l'appariement. La première, appelée précision, est la proportion de vraies paires parmi les paires retenues ; elle s'écrit de la manière suivante :

$$P = \frac{Nombre\ de\ vraies\ paires\ acceptées}{Nombre\ de\ paires\ acceptées} = \frac{Vrais\ positifs}{Vrais\ positifs + Faux\ positifs}$$

La seconde, appelée rappel, est la proportion de vraies paires retenues parmi l'ensemble des vraies paires ; elle s'écrit de la manière suivante :

$$R = \frac{Nombre\ de\ vraies\ paires\ acceptées}{Nombre\ de\ vraies\ paires} = \frac{Vrais\ positifs}{Vrais\ positifs + Faux\ négatifs}$$

La précision pourra être calculée pour les différents scénarios testés. Le rappel est moins évident à calculer, car il nécessite de connaître l'intégralité des vraies paires, ce qui n'est pas possible en général dans un cas réel.

On s'intéressera également au déroulement du processus, notamment l'impact de la normalisation sur les volumes d'appariement réalisés dans les étapes d'appariement exact, puis approché.

2. Processus d'appariement : description et enjeux de l'étape de nettoyage des données

Parmi les étapes d'un appariement, la première est celle de la standardisation ou nettoyage des données, appelée aussi parfois normalisation des données.

Cette étape est souvent jugée comme cruciale, importante, voire la plus importante en termes de temps de travail et de son poids sur la qualité finale de l'appariement.

Les termes nettoyage ou standardisation sont d'ailleurs un peu réducteurs, et l'on pourra y préférer la notion de préparation de données : on aura ainsi en tête, non seulement le nettoyage, mais également une dimension de qualification des données, et de préparation des étapes suivantes car les constats et les choix opérés à cette étape peuvent impacter les suivantes.

Le premier objectif de cette étape est de constituer un ensemble de variables comparables au sein des jeux de données que l'on souhaite apparier, en harmonisant les formats et en rapprochant les concepts lorsque cela est possible.

Le deuxième objectif est de nettoyer les données, en essayant d'ôter le bruit (erreurs de saisie, de codage) tout en conservant l'information la plus précise et discriminante possible.

Le troisième objectif est d'identifier les variables à utiliser lors du processus d'appariement, en sélectionnant celles dont la quantité d'informations est intéressante, et dont la qualité est bonne pour les deux fichiers.

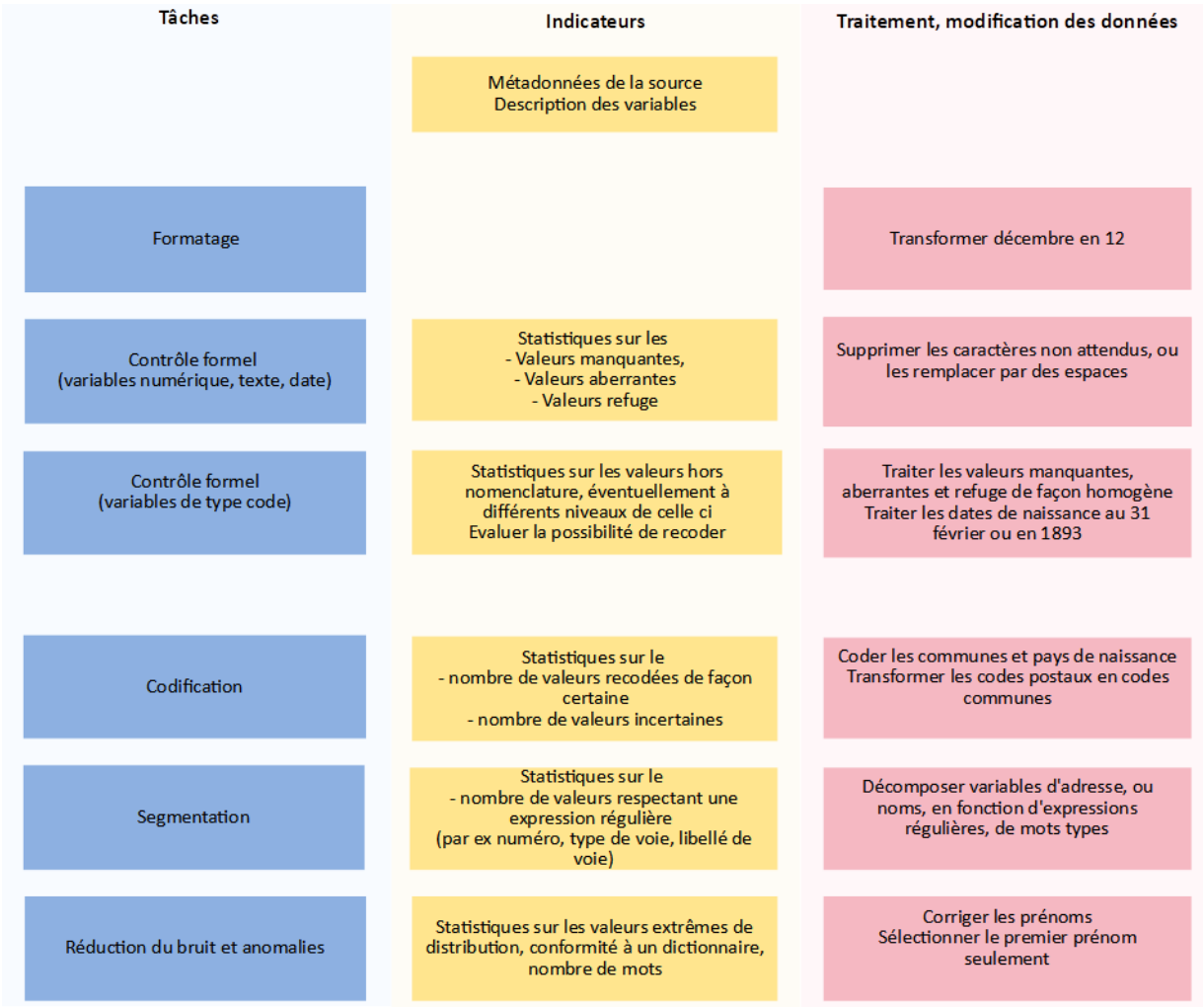
2.1. Les différentes tâches de l'étape de préparation des données

Il semble utile de proposer une décomposition des différentes tâches que recouvre cette étape de préparation des données.

Sanmartin (2017) propose de découper la tâche de préparation des données en plusieurs étapes : « la normalisation des variables de couplage, l'évaluation des variables de couplage, la détermination des enregistrements visés par le couplage (et notamment l'exclusion d'enregistrements), l'évaluation des résultats de la préparation de données, et le commencement du rapport sur le couplage d'enregistrements ».

Cette proposition souligne le fondement du processus complet d'appariement sur cette étape préalable qui est nécessaire pour « déterminer le degré d'exactitude et d'exhaustivité ainsi que pour accroître leur comparabilité et, finalement, pour veiller à ce qu'elles soient de qualité suffisamment élevée pour contribuer utilement au couplage d'enregistrements. La disponibilité et la qualité des variables de couplage détermineront les détails de la stratégie de couplage d'enregistrements. » [5] (Sanmartin et alii, 2017)

Schéma 2 : Proposition de plusieurs types de tâches de qualification et nettoyage des données



2.1.1. Les tâches de formatage et de contrôle formel

La première tâche, la plus évidente, consiste à mettre les variables d'appariements dans des formats similaires : format texte, numérique ou date, et homogénéisation de la casse et des signes diacritiques .Ainsi on ne comparera pas un libellé de commune à un code postal, ou un mois noté « 12 » à un mois renseigné « décembre ». Cette tâche est indispensable pour réaliser un appariement.

Pour deux variables représentant le même concept, on dispose à la fin de cette étape de valeurs similaires (12 et non pas décembre), dans le même format ('12' et non pas 12), et la même casse ('FRANCOIS' et non pas 'François').

Cela peut nécessiter un éventuel recodage, qui sera à ce stade déterministe et univoque, par exemple pour le sexe (1,2 versus m, f) ou les dates de naissance.

La deuxième tâche s'intéresse ensuite au respect des valeurs possibles d'une variable de format numérique, texte ou date, et notamment de l'appartenance à un ensemble de valeurs autorisées. Une date de naissance par exemple, respecte un format particulier : il n'existe pas de personne née un 30 février ou un 31 juin ; un prénom comporte habituellement des lettres, éventuellement un tiret, une apostrophe, des caractères accentués, mais ne comporte généralement pas de chiffre.

Une troisième tâche peut être identifiée, similaire dans l'esprit à la précédente : il s'agit ici de valider qu'une valeur appartient bien à une liste fermée. C'est souvent le cas pour les variables géographiques, qu'il s'agisse d'un code commune ou d'un code pays.

Lors de ces tâches de formatage et de contrôle formel , il est nécessaire de repérer comment sont encodées les valeurs manquantes (absence de valeur, ou valeur extrême, ou code spécifique), et de faire des choix pour les valeurs ne respectant pas le format attendu.

2.1.2. Les tâches de codification, ou de conformité à une nomenclature

Lors de l'étape de nettoyage des données, on peut également être amené à transformer les données existantes.

Il peut s'agir notamment d'une tâche de codification, pour les variables géographiques comme le pays ou la commune de naissance, afin de se conformer à une liste de codes. Il arrive en effet que l'on ne dispose que de libellés pour ces variables.

On peut alors choisir soit d'utiliser ce libellé comme variable d'appariement, soit de réaliser une première étape de codification de ces libellés géographiques, dans une nomenclature géographique commune aux deux fichiers.

C'est aussi un choix possible lorsque les codes présents dans les données ne sont pas concordants avec la géographie de référence : il peut alors y avoir recodage dans la géographie souhaitée, si on dispose toutefois des informations nécessaires.

Cela peut aussi consister à réaliser un transcodage, pour mettre dans la même nomenclature deux variables codées dans des listes différentes (code commune *versus* code postal par exemple).

Ce transcodage peut comporter des choix non univoques, entraînant alors des erreurs potentielles.

2.1.3. Les tâches de segmentation

On peut souhaiter découper certains champs textuels, pour en extraire une information importante, ou se ramener à des informations comparables au sein des jeux de données. C'est le cas de l'adresse, qui peut faire l'objet d'une seule variable, ou de plusieurs décrivant différents attributs (nom, type de voie, libellé de voie, complément d'adresse).

C'est le cas également du nom, qui peut être simple, ou contenir nom de naissance et nom d'usage au sein d'une même variable.

Ce type de tâches devient plus complexe, et entraîne potentiellement plus de risques de perte d'information.

2.1.4. Les tâches visant à réduire le bruit

Enfin, des opérations plus poussées peuvent être envisagées, afin de réduire l'impact notamment de l'orthographe sur la correspondance des champs texte. C'est le cas notamment de la phonétisation, que l'on n'a pas testée ici, ou de la comparaison à un dictionnaire de noms ou de prénoms.

On peut également lors de cette étape tenter de supprimer les valeurs n'apportant pas de réelle information, comme le « SP » (sans prénom) des fichiers fiscaux, ou autres motifs de gestion (« A SUPPRIMER »).

Pour tous ces traitements, plusieurs choix sont possibles lorsque l'on constate une anomalie pour une valeur :

- supprimer la valeur en anomalie ;
- tenter de la « corriger », totalement ou partiellement (supprimer les caractères non attendus d'un champ nom par exemple) ;
- la conserver telle quelle ;

- conserver une partie seulement de l'information (département de naissance et non commune par exemple, ou seulement le premier prénom) ;
- indexer plusieurs types d'informations (par exemple l'ensemble des prénoms, mais également le premier) et les utiliser conjointement dans le processus de comparaison.

Tous ces choix peuvent être faits pour un petit nombre d'individus, ou pour l'ensemble de la base.
Le non-respect d'un format (code commune) peut conduire à privilégier l'usage du département de naissance par exemple, plutôt que la commune si la qualité globale de la variable est sujette à caution.

2.2. Les conséquences du nettoyage de données :

L'objectif premier de l'étape de nettoyage des données est de les rendre le plus comparable possible. On peut également souhaiter nettoyer les données pour en supprimer les anomalies, ou utiliser une information moins fine, comme le premier prénom au lieu de l'ensemble des prénoms.

Cette étape veut donc déterminer quelles sont les bonnes informations à utiliser. Mais en normalisant de façon importante, on peut diminuer le caractère discriminant des informations dont on dispose, et entraîner une dégradation des résultats.
Le challenge est donc de distinguer les variations légitimes (des orthographes d'un prénom par exemple) des erreurs qui peuvent subvenir lors du processus de collecte des données [3].

Ainsi, l'absence de normalisation entraînera des éventuelles non-concordances, et donc des faux négatifs. À l'inverse, la normalisation excessive pourrait entraîner l'apparition de faux positifs.
Des procédés comme la phonétisation peuvent ainsi créer des faux positifs, en réduisant le caractère discriminant des variables.

Une tâche de normalisation peut entraîner :
- l'apparition d'erreurs (par exemple MONIQUE à la place de MOJIBUR, ou erreur de codification d'un pays de naissance CONGO),
- la diminution de la quantité d'information (sans introduire d'erreur toutefois) comme l'utilisation d'un seul prénom seulement.

2.3. Les enseignements que l'on peut tirer de l'étape du nettoyage des données :

Cette première étape est l'occasion également de faire un premier bilan sur la qualité des variables, et d'identifier celles qui pourront être utilisées *in fine* comme variables d'appariement, ainsi que les méthodes de comparaison à utiliser.

Il est intéressant de produire lors de cette étape un ensemble de statistiques visant à décrire la qualité des informations. [5](Christen 2019)

3. Description des données et scénarios de nettoyage des données :

3.1. Données utilisées :

Les données utilisées sont celles de l'enquête annuelle de recensement, du fichier d'imposition sur les personnes et du fichier tous salariés pour l'année 2019.
Elles ont été mobilisées dans le contexte des travaux autour de RéSIL [1] et du code statistique non signifiant (CSNS)[2], en collaboration avec les équipes de ces projets.

La description ci-dessous concerne les données en entrée du processus d'appariement, elles ne sont pas une description de la source elle-même.

Afin de réduire le volume des données et les temps de traitement impartis, les travaux ont porté sur le département de la Lozère et sur celui des Hauts de Seine.
La Lozère a fait l'objet de plusieurs tests dans le cadre du projet RéSIL, et avait été retenue en raison de sa petite taille et des temps d'exécution raisonnables qui en découlent.
Le département des Hauts-de-Seine a été choisi en raison de la diversité de population qui y réside, entraînant une plus grande variété de traits d'identité.
Les données sur le plus petit département ont fait l'objet de plus de tests, pour des raisons de faisabilité étant donné leur volume.

L'enquête annuelle de recensement (EAR) comprend :

- une variable pour le ou les prénoms ;
- une variable pour le ou les noms ;
- plusieurs variables concernant le lieu de naissance (libellé de la commune, code commune et code département pour les naissances en France, le libellé du pays pour les naissances à l'étranger) ;
- une variable pour la date de naissance.

.

L'enquête de recensement est bimodale (papier et internet). Les données identifiantes ne font pas partie des variables d'intérêt et ne font pas l'objet de traitement.
Les informations identifiantes des bulletins papier sont toutefois saisies par un prestataire de saisie, au moyen d'outils de saisie optique. Cela entraîne des erreurs en particulier sur les variables de type texte.
Ces erreurs sont de l'ordre de 12 % des bulletins papier dans les EAR récentes, et la valeur à l'issue de la saisie peut être très différente de la vraie valeur. (KONSTANTOPOULOS au lieu de TRISTANT EPOUSE LEROY, ANNICK au lieu de CHRISTINE, KONAIN au lieu de ROMAIN, ou encore PASCALINE LOIS au lieu de ABRAHAM).
La collecte internet représente environ l60 % des individus en 2019, on peut donc s'attendre à environ 5 % de noms et prénoms entachés d'erreurs dues à la saisie optique.

[1] Le projet RéSIL vise à construire un répertoire d'individus et de logements à partir de plusieurs sources administratives.
[2] Le projet CSNS vise à délivrer un code statistique non signifiant, attribué à partir du NIR ou de traits d'identité appariés au préalable au répertoire national des personnes physiques.

Le fichier d'imposition sur les personnes (FIP) comprend :

- une variable pour le ou les prénoms ;
- une variable pour le nom de naissance ;
- une variable pour le nom marital ;
- une variable de type code concernant le lieu de naissance ;
- trois variables pour les jour, mois et année de naissance.

Le fichier tous salariés comprend :

- une variable pour le ou les prénoms ;
- une variable pour le nom de naissance ;
- une variable pour le nom marital ;
- une variable de type code concernant le lieu de naissance ;
-
- trois variables pour les jour, mois et année de naissance.

Le tableau 2 compare les informations disponibles dans chacune de ces sources.

Tableau 2 : Exemple fictif des informations dont on dispose

	EAR	FIP	Tous salariés
Exemple 1			
Nom	MARTIN EPOUSE DELAHAYE		
Nom de naissance		MARTIN	MARTIN
Nom marital		DELAHAYE	DELAHAYE
Prénoms	MARIE	MARIE CAMILLE	MARIE CAMILLE
Code commune ou pays de naissance		99352	99352
Nom de la commune de naissance	ORANIE		
Nom du pays de naissance	ALGERIE		
Date de naissance	19790927	19790927	19790927
Exemple 2			
Nom	BERNARD		
Nom de naissance		BERNARD	BERNARD
Nom marital			
Prénoms	PHILIPPE	PHILIPPE SERGE	PHILIPPE SERGE
Code commune ou pays de naissance	93031	93031	93031
Nom de la commune de naissance			
Nom du pays de naissance			
Date de naissance	19860810	Trois variables 1986, 08, et 10	Trois variables 1986, 08, et 10

NB : tous les exemples d'informations d'état civil sont inspirés d'exemples réels, mais ne correspondent pas à de vraies personnes (mélange d'informations provenant de plusieurs individus).

3.2. Informations qualitatives et descriptives sur les données

Les tableaux ci-dessous décrivent certaines caractéristiques des fichiers utilisés.
La qualification des variables est une partie importante de l'étape de préparation des données.

Tableau 3 : Répartition des enregistrements des trois fichiers selon le nombre de mots pour les variables nom et prénom

	EAR	FIP	Tous salariés
Nombre de mots de la variable nom ou nom de naissance			
0	0,68 %	0 %	0,00 %
1	80,26 %	90,05 %	88,54 %
2	10,91 %	7,93 %	9,49 %
3	7,02 %	1,55 %	1,61 %
4 ou plus	1,13 %	0,48 %	0,36 %
Nombre de mots du prénom			
0	1,41 %	0,04 %	0,03 %
1	89,86 %	73,3 %	89,94 %
2	7,92 %	17,49 %	9,03 %
3	0,72 %	8,15 %	0,90 %
4 ou plus	0,09 %	1,02 %	0,10 %
Nom ne comportant pas de voyelles	0,99 %	0,01 %	0,01 %
Nom ne comportant que des initiales	1,17 %	0 %	0,01 %
Proportion de nom marital non vide*		50,9 %	23,60 %
Prénom ne comportant pas de voyelles	1,41 %	0,04 %	0,03 %
Prénom ne comportant que des initiales	1,41 %	0,04 %	0,03 %

* Les données portent sur l'ensemble des sources pour les deux départements sauf pour le nom marital. On a réduit la population aux femmes majeures de moins de 65 ans (champ comparable pour les trois sources)

Tableau 4 : Proportion de codes communes de naissance manquants ou non conformes au COG

	EAR 92	FIP 92	Tous salariés 92
Valeurs manquantes	33,04 %	0,00 %	2,16 %
Valeurs non conformes au COG	0,37 %	3,19 %	7,33 %

Tableau 5 : Proportion de valeurs manquantes ou non valides pour les jour et mois de naissance

Valeur	EAR	FIP	Tous salariés
Date incomplète ou non valide			
0000		0,22 %	
Mois+00		Quelques valeurs	
Mois+99			Quelques valeurs
WWWW			0,06 %
..			0,3 %
Mois sans jour	Quelques valeurs		
Jour sans mois	Quelques valeurs		
Pics de distribution des dates			
1er janvier	1,09 %	1,47 %	0,88 %
31 décembre	0,45 %	0,45 %	0,4 %
jours « ordinaires »	De 0,22 % à 0,32 %	De 0,24 % à 0,32 %	De 0,23 % à 0,32 %

L'analyse de ces informations nous conduit à en tirer des directives pour le processus d'appariement :

- les informations partielles ou manquantes sont représentées de plusieurs manières différentes : cela nous conduit à normaliser l'encodage des valeurs manquantes pour les variables de date ;
- on note également les valeurs refuge classiques du 1er janvier et dans une moindre mesure du 31 décembre ;
- un quart des femmes ont renseigné un nom marital dans le fichier tous salariés, et la moitié dans le FIP. Dans l'EAR, on ne dispose pas de la distinction des noms : cela nous conduit à utiliser plutôt le nom de naissance que le nom marital, en essayant toutefois de le repérer lorsque c'est possible dans l'EAR ;
- 25 % des prénoms comportent 2 ou 3 mots dans le fichier FIP, contre environ 10 % dans les autres fichiers : cela nous conduit à utiliser plutôt le premier prénom seulement ;
- 7 % des lieux de naissance du fichier tous salariés n'appartiennent pas au code officiel géographique de l'année en cours ou à un code pays : en effet, cette information n'est pas enregistrée en tant que telle, mais extraite à partir ddu NIR ; cela conduit à privilégier la concordance sur le lieu de naissance sur deux positions (le département ou naissance à l'étranger) plutôt que le code commune de naissance ou pays, après avoir codé dans l'EAR les codes pays à partir des libellés fournis. Il serait toutefois possible d'essayer d'améliorer le code issu du NIR en utilisant l'historique de géographie.

3.3. Description des scénarios de normalisation :

Trois niveaux de formatage et nettoyage des données ont été réalisés :
Le premier niveau est le plus fruste et il a intentionnellement été choisi comme plus faible que ce qui est fait le plus souvent.
Les variables de noms et prénoms ont été conservées telles quelles, dès lors qu'elles contenaient au moins un caractère alphabétique (cela signifie que les valeurs ne contenant que des chiffres ou des caractères spéciaux ont été supprimées). Le pays de naissance de l'EAR n'a pas été codifié ce qui induit une absence de code commune pour les personnes nées à l'étranger dans l'EAR.
La date de naissance a été mise dans des formats similaires (agrégation, ou désagrégation selon les cas).

Le deuxième niveau de nettoyage correspond ce qui pourrait être fait assez classiquement dans un processus d'appariement. Certains de ces traitements sont notamment faits par l'outil Rapsodie utilisé au sein de l'Insee pour apparier les données fiscales avec d'autres sources de données. On est allé toutefois un peu au-delà notamment pour essayer d'identifier les équivalents mal orthographiés du mot EPOUSE.
Les caractères numériques et spéciaux sont supprimés des variables noms et prénoms, et remplacés par des espaces.
Si le prénom comporte des mots d'une seule lettre, ils sont supprimés.
Les prénoms composés classiques pour lesquels le tiret est absent sont modifiés (MARIE PIERRE devient MARIE-PIERRE) afin d'identifier le mieux possible le premier prénom.
On essaie également de repérer les civilités (premier mot du nom correspondant à 'MME', 'M', 'ME','MNE','MR','MLLE','MMM') et les mentions de relation (EPOUSE, VEUVE, NEE, EP) sont repérées pour en déduire les noms de naissance et noms maritaux.
On cherche également les chaînes de texte approchantes (distance de Levensthein[3] de 1 pour les mots EPOUX et VEUVE et distance de Levenshtein de 1 ou 2 pour les mots EPOUSE et DIVORCE).

Les lieux de naissance non codés dans l'EAR ont été codés dans la mesure du possible, avec Sicore[4]. Les résultats ont été enrichis par une comparaison des libellés de pays et communes des bases d'apprentissage Sicore avec ceux des bases de données et les distances de Levensthein faibles ont été conservées.
Les lieux de naissance des fichiers administratifs n'ont pas été corrigés. En effet, on ne disposait pas de libellé de lieu de naissance, et l'on ne disposait donc pas d'un moyen de correction, les anomalies pouvant provenir d'un code commune ayant changé, d'un code postal au lieu d'un code commune, ou d'une erreur de frappe.
Les informations de dates de naissance non conformes ont été supprimées (jour et mois à 00, 99, XX, WW, ou année inférieure à 1905).

Le troisième niveau va au-delà de ce qui est fait habituellement, en utilisant les listes de noms et prénoms disponibles sur insee.fr[5] afin de vérifier la présence des noms et prénoms dans le dictionnaire.

[3]La distance de Levensthein mesure le nombre d'éditions élémentaires (suppression, insertion ou substitution d'un caractère) pour transfomer une chaîne de caractères en une autre.
[4]Outil interne de codification des variables de libellés de communes, libellés de pays, activités, professions et diplômes.
[5]Le fichier des prénoms contient des données sur les prénoms attribués aux enfants nés en France entre 1900 et 2020 (plusieurs conditions d'effectifs sont appliquées, dont celle-ci : pour une année de naissance donnée, le prénom a été attribué au moins 3 fois à des personnes de sexe féminin ou de sexe masculin).
Le fichier des noms contient les noms attribués au moins 30 fois de 1891 à 2000 en France (métropole et Dom) et dénombrent 218 982 noms différents au niveau national.

Un fichier de travail a été constitué avec les données de l'ensemble de l'EAR, de trois départements pour les données fiscales et d'emploi. Les noms et prénoms retrouvés plus de 20 fois dans les fichiers administratifs ont été ajoutés au dictionnaire pour prendre en compte une partie des prénoms rares.

Les noms et prénoms n'appartenant pas au dictionnaire ont ensuite été appariés et lorsqu'il existait un écho à une distance de Levensthein plus petite que trois, les noms et les prénoms ont alors été modifiés (les échos étaient triés par distance de Levensthein croissante et nombre de personnes portant ce nom ou ce prénom dans le dictionnaire décroissant).

Cette conformité à un dictionnaire a été mise en place pour détecter les erreurs orthographiques, notamment celles occasionnées par la saisie optique des bulletins papier de l'EAR.

Tableau 6 : Différentes orthographes appariées avec CHRISTOPHE et MONIQUE dans l'EAR

CHRISTAPHE	240	MONIGUE	282
CARISTOPHE	78	MONQUE	110
CHRISTEPHE	39	MORIQUE	103
CHRITOPHE	35	MONIQUES	95
CHRIOTOPHE	34	MUNIQUE	78
CHAISTOPHE	33	MANQUE	75
CHRISTOPTE	31	MANIGUE	67
CHRSTOPHE	31	MOIQUE	62
CHRISTOGHE	25	MNIQUE	58
CHRISTOPE	25	MONIGNE	52
		MONIQNE	30
		LONIQUE	28
		MONIAUE	27
		MONIQE	27
		HONIQUE	25
CHISTAPHE, HRISTOPHE, CHNSTOPHE, CHRISTOHE, CHRISTPHE, CHUSTOPHE, CHASTOPHE, CHRISTOPRE, CHRESTOPHE, CHRISTOPLE, CHRISTOPME, SHRISTOPHE, CHRISTOPNE, CHISTOGHE, CHRISOPHE, CHRIGTOPHE, CHRISIOPHE, CHRISTOPPE, GHRISTOPHE, CHNISTOPHE		MONIQU, MOUIQUE, JOSIQUE, MONIQUEE, MONIQUEN, MONNQUE, MONIUE, MORIGUE, MGNIQUE, MONIGUES, MOURQUE, YONIQUE, MANIQUES, MONEQUE, MONIQUEL, MONIQHE, MONIQIE, MONIQUUE, TONIQUE	
concernent entre 10 et 25 personnes	298	concernent entre 10 et 25 personnes	258
545 orthographes supplémentaires	823	672 orthographes supplémentaires	1027

Source : ensemble de l'EAR 2019

Dans les fichiers administratifs, ces types d'erreur sont beaucoup moins fréquents.

Ainsi on retrouve 17 personnes prénommées CHRISTAPOR, CHRISTOPE, HRISTOPHE, CHRISTOHE, CHRITOPHE, CHRSITOPHE, CHRSTOPHE ou CHRISOPHE dans les fichiers FIP et Tous salariés du département des Hauts-de-Seine ou encore 25 personnes prénommées CENIQUE, COPIHUE,

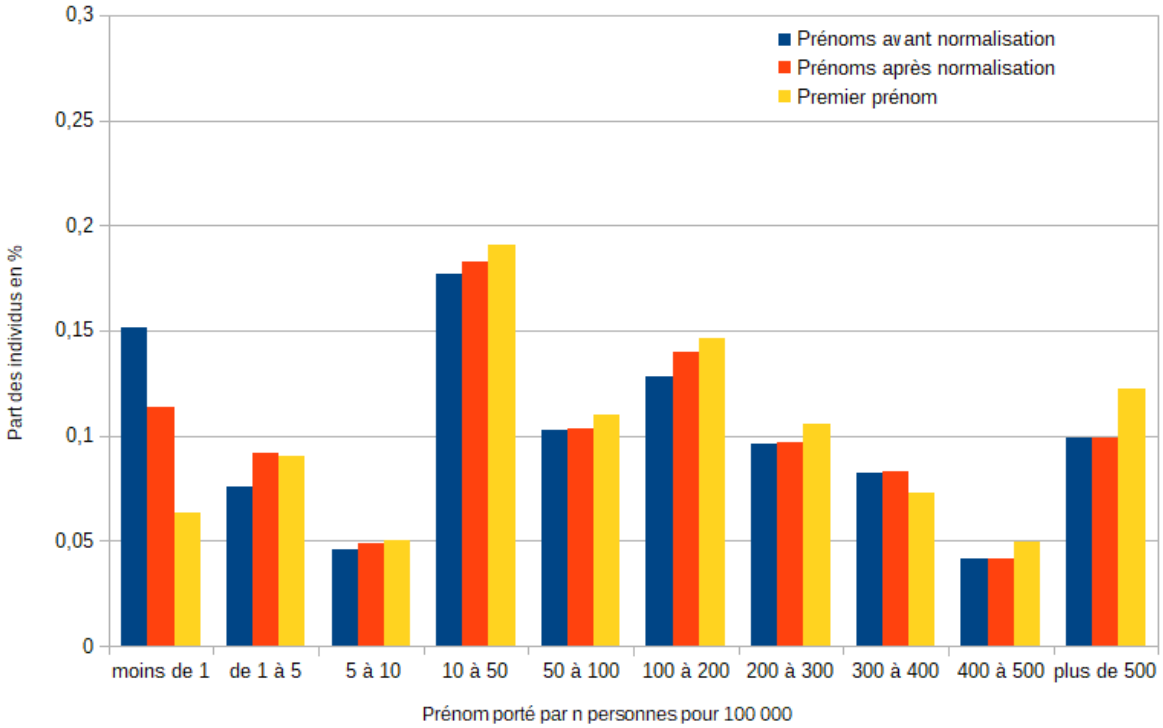
LENIQUE, MENGXUE, MINGYUE, MOJIBUR, MOZIBUR, ORNIQUE, ULNIQUE, VENIQUE, YONGXUE, HONGYUE, LINDIQUE, MENGYUE, RONIQUE, ZE MONIQUE.
La correction ici est beaucoup plus questionnable, et l'absence des prénoms rares dans le dictionnaire construit est préjudiciable.
On peut imaginer que MOJIBUR est un vrai prénom par exemple.

Des améliorations sont possibles pour ces tables de noms et prénoms, probablement en adaptant le seuil d'acceptation d'un prénom corrigé, en termes de distance absolue, ou relative à la taille du nom, ainsi que la prise en compte du lieu de naissance et de la date.

Tableau 7 : Nombre moyen de personnes partageant le même nom avant et après normalisation

	EAR	FIP	Tous salariés
Nom avant normalisation	2	3,3	3,7
Nom après normalisation	2,5	4,9	5,8
Prénoms avant normalisation	4,8	4,7	8,8
Prénoms après normalisation	6,7	5,3	12,6
Premier prénom après normalisation	9,4	29,6	27,6

Graphique 1 : Distribution du nombre de personnes selon la fréquence de leur prénom, avant et après normalisation – Données EAR



Graphique 2 : Distribution du nombre de personnes selon la fréquence de leur prénom, avant et après normalisation – Données FIP

Graphique 3 : Distribution du nombre de personnes selon la fréquence de leur prénom, avant et après normalisation – Données Tous salariés

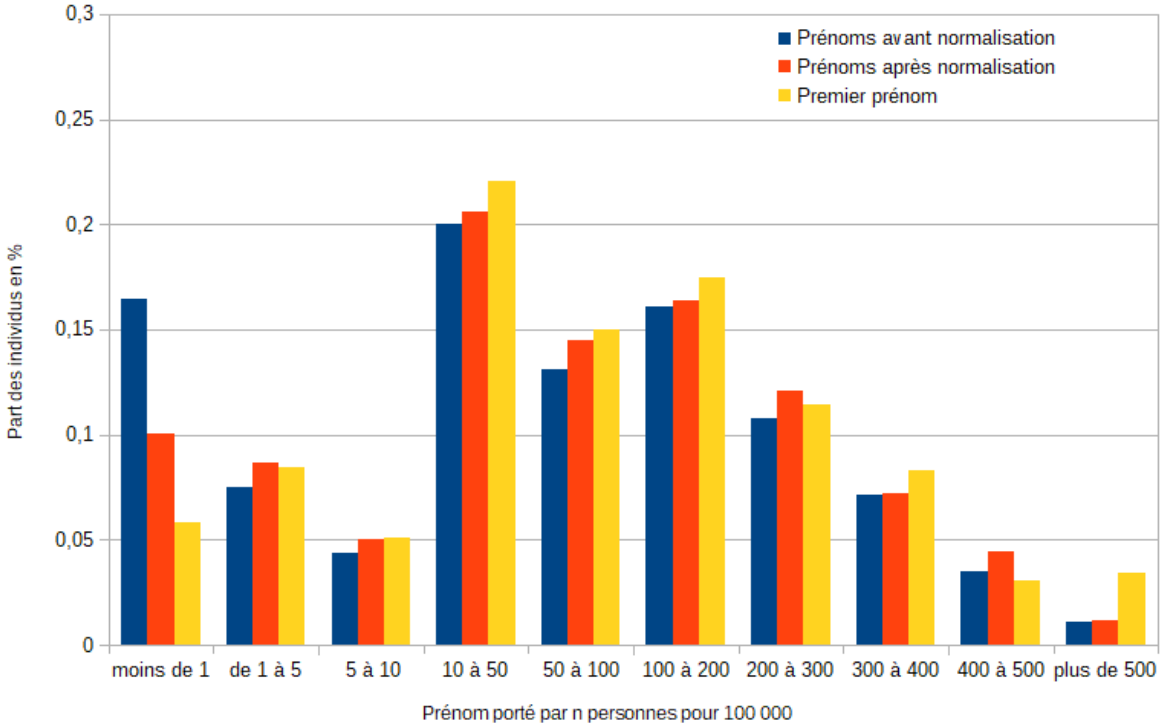


Tableau 8 : Entropie de Shannon des variables de noms et prénoms

	EAR	FIP	Tous salariés
Nom avant normalisation	15,12	16,47	16,41
Nom après normalisation	14,74	15,88	15,85
Prénom avant normalisation	11,69	13,11	11,65
Prénom après normalisation	11,31	12,89	11,38
Premier prénom	10,98	10,72	10,85
Lieu de naissance sur 5 positions avant codification	5,93		
Lieu de naissance sur 5 positions après codification	6,58*	7,68	6.69
Lieu de naissance sur 2 positions	3,56	3,733	4,06

* L'entropie augmente, car le code du lieu de naissance était absent pour les naissances à l'étranger. La quantité d'information apportée par le libellé du lieu de naissance n'a pas été mesurée.

Note : l'entropie de Shannon est une mesure de la quantité d'information d'une source donnée. Elle se calcule par la formule suivante :

$$E = -\sum_{\text{ensemble des valeurs distinctes } i} p_i \times \log(p_i)$$

où p_i est la proportion d'individus ayant la valeur i
Cela nous permet de mesurer la perte d'information liée à la normalisation mais également à la sélection d'informations plus agrégées.

Il reste toutefois des anomalies que l'on ne peut pas corriger :

- présence de nom et prénom dans un même champ ;
- présence de plusieurs noms et/ou prénoms sans espace : MAZETTIFANNY ;
- présence d'éléments de civilité ou relation sans espace (NB on a essayé pour EPOUSE correctement orthographié, c'est plus difficile pour CREMIEREROUSE MENUD ou bien encore AMIIHEPMIIM UMRINE) ;
- présence de relations avec beaucoup de fautes CAUBET EGASSE PARIS ;
- interversion nom marital et nom de naissance ;
- interversion nom et prénom ;
- interversion mois et jour si inférieur à 12 ;
- erreur sur le siècle pour les annnées entre 1910 et 1922 au lieu de 2010 à 2022.

3.4. Enseignements tirés de l'étape de préparation des données

Le tableau ci-dessous présente la répartition des modifications apportées dans les différents fichiers.
Le premier constat est que les fichiers administratifs ont été peu corrigés.
Les principales corrections dans l'EAR ont porté sur la codification du pays de naissance, et la tentative de repérage des noms maritaux. Elles sont plus fréquentes pour les réponses papier que les réponses internet.

Dans la deuxième étape de normalisation, la proportion de noms et prénoms modifiés suite à la comparaison avec un dictionnaire est plus importante pour l'enquête.

Tableau 9 : Pourcentage d'enregistrements en fonction du nombre de modifications - pour les variables de nom de naissance, nom marital, prénom, lieu de naissance, date de naissance

	EAR	FIP	Tous salariés
Modifications lors de l'étape 1 de normalisation			
Aucune modification	74,65 %	93,36 %	96,89 %
1 modification – Date de naissance	0,04 %	0,03 %	0,48 %
1 modification – Lieu de naissance	10,56 %		
1 modification – Prénom	2,78 %	6,48 %	2,47 %
1 modification nom marital		0,03 %	0,03 %
1 modification – nom de naissance	0,53 %	0,09 %	0,10 %
Modification sur nom de naissance ET nom marital	9,22 %	0,01 %	
2 modifications	0,42 %	0,00 %	0,03 %
3 modifications	1,74 %		
4 modifications	0,06 %		
Modifications lors de l'étape 2 de normalisation			
	EAR	**FIP**	**Tous salariés**
Aucune modification	86,06 %	93,50 %	94,09 %
1 modification – Prénom	5,59 %	0,51 %	0,73 %
1 modification - nom marital	0,56 %	1,19 %	0,48 %
1 modification – nom de naissance	6,66 %	4,26 %	4,36 %
Modification sur nom de naissance ET nom marital	0,10 %	0,27 %	0,11 %
2 modifications	1,04 %	0,25 %	0,21 %
3 modifications		0,04 %	0,01 %

3.5. Scénarios d'appariements testés

Plusieurs scénarios d'appariements ont été testés ici, en faisant varier les données (pour avoir des qualités relatives de fichiers différentes), les choix d'appariement (exact, déterministe flou), ainsi que l'identification au répertoire national d'identification des personnes physiques (RNIPP) par l'algorithme d'identification du CSNS.

Chaque scénario a été décliné pour les trois niveaux de normalisation sur les variables nom (ou nom de naissance), prénoms, date de naissance et commune ou pays de naissance.
Pour les appariements exacts et déterministes flous, on a dans un second temps, remplacé les prénoms par le premier prénom seulement, la commune de naissance en France par le département de naissance, et le pays de naissance par un indicateur de naissance à l'étranger, afin de mesurer les impacts respectifs du nettoyage des données, et des choix de variables mobilisées dans l'appariement.

Les variables d'adresse et de commune de résidence ont été exclues à dessein. En effet, le temps manquait pour réaliser l'analyse manuelles des paires. Afin d'évaluer la précision, on a utilisé les résultats d'un appariement précédent, ainsi que la concordance de l'adresse et/ou de la commune de résidence.

3.5.1. Appariement exact entre deux sources

Le premier type d'appariement testé est un appariement exact.

Il a été mis en œuvre sur les données de l'EAR et FIP pour le département de la Lozère, pour les individus de 15 ans et plus.

L'appariement exact a été réalisé sur les variables de nom (ou nom de naissance), de date de naissance, de commune de naissance et de l'ensemble des prénoms.

Les paires trouvées par Rapsodie ou concordantes sur l'adresse ont été d'emblée considérées comme vraies paires.
Les doutes restent limités sur les autres paires, dans un contexte d'appariement exact et localisé.

3.5.2. Appariement déterministe flou entre deux sources

On a également effectué un appariement déterministe par valeurs approchées (les conditions d'acceptation des paires autorisant une correspondance non exacte), afin d'essayer d'évaluer dans quelle mesure le passage d'une correspondance exacte (sur tous les prénoms) à une correspondance partielle (inclusion d'une variable dans l'autre par exemple) peut compenser l'absence de normalisation.

On a conservé une correspondance exacte pour la date et le lieu de naissance (en faisant varier toutefois le niveau de précision de l'information, en passant de 5 positions ou 2 positions), et on a introduit une relâche sur les variables de nom et prénom (inclusion de 3-grams).

3.5.3. Identification au RNIPP

Le troisième scénario testé est l'identification des individus au RNIPP, via l'algorithme d'identification du CSNS, pour les fichiers de l'EAR, FIP et tous salariés des Hauts-de-Seine.

Dans ce contexte, chaque individu est recherché dans une base de données à portée nationale et comportant 130 millions de lignes.

Il convient de souligner des différences très importantes entre ces tests :
- les premiers, bien que comparant des données d'enquête à une source administrative, mettent en vis-à-vis des fichiers portant sur des champs proches et de petite taille. Les fichiers ont pu être normalisés de façon identique au préalable ;
- le troisième présente une forte dissymétrie des données, de par leur volume et par le champ couvert. De plus, les données du RNIPP ne subissent pas en entrée la même normalisation.

L'algorithme est par ailleurs unique, quel que soit le fichier.
Il se décompose en une étape d'appariement exacte, plusieurs étapes avec relâche simple (premier prénom seulement, et relâches successives sur différentes variables). Ces étapes ne renvoient que des échos uniques. La dernière étape fonctionne par valeur approchée, qui autorise des correspondances partielles sur plusieurs variables simultanément. Cette dernière étape sélectionne le meilleur écho et délivre donc un score.

Les variantes avec premier prénom et département de naissance n'ont pas été soumises à ce scénario : en effet, la comparaison avec le premier prénom seulement est effectuée lors de la deuxième étape du processus CSNS, et la troisième étape consiste en une relâche de la contrainte sur le lieu de naissance.
Enfin, ce troisième scénario permet de connaître facilement le statut des paires, puisque le fichier tous salariés comporte le NIR. On peut estimer très rapidement les taux de faux positifs à l'issue du processus (on fait toutefois l'hypothèse que les NIR présents dans le fichier tous salariés sont exempts d'erreurs).

NB : l'ensemble des tests présentés ici avec le CSNS ont été menés avec une version provisoire de l'algorithme à l'hiver 2022.

4. Résultats

4.1. Mesures de l'impact du nettoyage des données, mesures de qualité d'un appariement

Afin d'essayer d'évaluer l'impact de l'étape de préparation des données, on s'intéressera au nombre d'enregistrements appariés (taux d'appariement global, qui mesure en creux de façon très imparfaite les faux négatifs), ainsi qu'à une évaluation des paires appariées à tort.

Dans le troisième scénario, on s'intéressera également à l'étape d'identification des enregistrements.

4.2. Résultats chiffrés*

4.2.1. Résultats des appariements exacts

Tableau 10 : Résultats de l'appariement exact EAR FIP sur la Lozère

	Nombre d'enregistrements		Nombre de paires		
	EAR	FIP	norme 0	norme 1	norme 2
sur code commune et tous les prénoms	10 127	58 516	3 001	3 956 (+31,8 %)	4 165 (+38,8 %)
sur code département et tous les prénoms			3 210 (+7 %)	4 307 (+43,5 %)	4 534 (+51,0 %)
sur code département et premier prénom			3 932 (+31,0 %)	5 201 (+73,3 %)	5 475 (+82,4 %)

Le taux d'appariement passe de 29,6 (avec les variables les plus fines, et non normalisées) à 54 % des enregistrements de l'EAR (à titre de comparaison, il est de 94,5 % à l'issue de Rapsodie, qui teste également la concordance des noms maritaux entre eux, et du nom de naissance d'une source avec le nom marital de l'autre source, qui mobilise également les informations d'adresse).

Dans ce premier test, aucune paire identifiée lors de la première étape n'est modifiée lorsqu'on augmente le degré de normalisation (lorsqu'on se déplace vers la droite dans ce tableau).

La réduction de l'information au département de naissance et au premier prénom n'a pas non plus modifié ou supprimé de paires trouvées au cours des étapes moins normalisées.
Autrement dit, les améliorations apportées n'ont pas modifié la composition des paires, dans ce contexte.

Toutes les paires appariées ici l'ont été par Rapsodie, sauf 26 (des EPASSE EPAUSE etc. au lieu de EPOUSE, mais aussi des JEAR LUCP ou des SBASTIGN).
95,9 % des paires présentent la même commune de résidence (on rappelle qu'on est ici sur une correspondance exacte des noms, prénoms, date, lieu de naissance).

Tableau 11 : Résultats de l'appariement exact Tous salariés FIP sur la Lozère

	Nombre d'enregistrements		Nombre de paires		
	Tous salariés	FIP	norme 0	norme 1	norme 2
sur code commune et tous les prenoms	31 604	58 516	17 741	18 240 (+2,8 %)	18 296 (+3,1 %)
sur code departement et tous les prénoms			18 322 (+3,3 %)	18 831 (+6,1 %)	18 889 (+6,5 %)
sur code dep et premier prenom			22 556 (+27,1 %)	22 976 (+29,5 %)	23 019 (+29,8 %)

Le taux d'appariement passe de 56,1 % à 72,8 % des enregistrements du fichier tous salariés.

Dans ce second test, une centaine de paires est perdue pour les cases jaunes (en bas à droite du tableau). Cela est dû notamment à une différence de conformité des prénoms au dictionnaire (erreur dans le traitement des prénoms multiples – à corriger)
En revanche, la réduction de l'information utilisée (sur le premier prénom et le département) n'a pas d'impact négatif. Une seule paire est modifiée et correspond probablement à une même personne, présente dans deux foyers fiscaux au sein du FIP. Une des deux lignes comprend seulement le premier prénom.

À ce stade, le constat semble donc très en faveur du nettoyage des données, qui augmente énormément le taux d'appariement, notamment pour les données de l'EAR (+38 %), et ce sans aggraver le nombre de faux positifs.
Le constat est moins flagrant pour les sources administratives (+3 % seulement), mais qui sont de meilleure qualité au départ.
On note également que le bon choix des variables d'appariement (premier prénom et lieu de naissance sur deux positions) a également un fort effet, et ce, même sans nettoyage des données (+31 % pour EAR FIP, + 27 % pour tous salariés - FIP).

4.2.2. Résultats des appariements déterministes flous

Tableau 12 : Résultats de l'appariement déterministe flou EAR FIP sur la Lozère

	Nombre d'enregistrements		Nombre de paires		
	EAR	FIP	norme 0	norme 1	norme 2
sur code commune et tous les prenoms	10 127	58 516	4 253	4 775 (+ 12,27 %%)	4 914 (+15,54 %)
sur code dep et premier prenom			4 950 (+16,39 %)	5 629 (+32,35 %)	5 794 (+36,23 %)

Le taux d'appariement passe de 42 % à 57,2 %. L'apport du nettoyage des données ou de la simplification de l'information utilisée est moins fort lorsqu'on utilise des correspondances floues.

Parmi les paires retenues, seules 49 paires n'ont pas été trouvées par Rapsodie. Une partie de ces paires divergentes avait fait l'objet d'annotations dans le cadre de travaux Résil : 2 sont des paires identifiées comme faux positifs, 3 sont indécises, et 4 paires n'appartenaient pas au fichier d'annotation. Pour les 42 autres, elles ont été retenues comme vraies positives.

Ici encore, le nettoyage des données semble apporter un meilleur taux d'appariement, au prix d'une modification du taux de faux positif très faible.
Ici encore, la prise en compte du bon niveau d'information semble plus importante que le nettoyage des valeurs erronées.

Enfin, contrairement à l'appariement exact, les différents niveaux de normalisation ont conduit à la perte d'un certain nombre de paires (de l'ordre de 10 à 60 paires selon les cas).

4.2.3. Résultats de l'identification au CSNS

Tableau 13 : Résultats de l'identification du fichier Tous salariés des Hauts-de-Seine par l'algorithme du CSNS

	Nombre d'enregistrements			Estimation de faux positifs		
	Norme 0	**Norme 1**	**Norme 2**	Norme 0	Norme 1	Norme 2
Enregistrements identifiés	888 382 (99,02 %)	888 244 (99,00 %)	880 706 (98,16 %)	0,43 %	0,42 %	1,43 %
Enregistrements non identifiés	8 811	8 949	16 487			
	Ventilation des enregistrements identifiés par étape			**Estimation de faux positifs par étape**		
Étape d'appariement exact	32,06 %	32,12 %	22,91 %	0,16 %	0,16 %	0,14 %
Étapes par relâche simple	64,28 %	64,22 %	69,28 %	0,27 %	0,27 %	0,37 %
Étape par valeur approchée	2,68 %	2,67 %	5,97 %	13,57 %	13,48 %	24,36 %
Non identifiés	0,98 %	1,00 %	1,84 %			

On constate ici que l'impact de la première normalisation est quasi nul.

Celui de la deuxième normalisation entraîne en revanche une dégradation, tant dans le déroulement du processus (il y a moins d'enregistrements identifiés lors de l'étape exacte), que dans les résultats : le taux d'identification a baissé et le taux de faux positifs a augmenté.

Le phénomène est toutefois assez faible.

Avec le premier niveau de normalisation 98 % des enregistrements ayant subi une modification n'ont pas changé d'identification.

Les enregistrements modifiés par le deuxième niveau de normalisation ont conservé également la même identification dans 91 % des cas.

Tableau 14 : Résultats de l'identification du fichier EAR des Hauts-de-Seine par l'algorithme du CSNS

	Norme 0	Norme 1	Norme 2
Identifiés	118 630 (94,16%)	118 994 (94,45%)	118 103 (93,74%)
Non identifiés	7 360	6 996	7 887
Étape d'appariement exact	10,50 %	18,29 %	13,74 %
Étapes par relâche simple	66,90 %	63,27 %	65,29 %
Étape par valeur approchée	16,76 %	12,89 %	14,71 %
Non identifiés	5,84 %	5,55 %	6,26 %

On note une légère amélioration du taux d'identification pour le niveau de normalisation 1, avec un report des enregistrements identifiés lors de la dernière étape vers les étapes précédentes, plus strictes.

Le niveau de normalisation 2 semble plutôt dégrader la qualité de l'appariement, avec une bascule dans le sens inverse vers l'étape la moins stricte, et une baisse du taux d'identification.

Pour les données de l'EAR et de FIP, on ne dispose pas de moyen d'évaluer rapidement les faux positifs. Une première analyse simple est de dénombrer les enregistrements pour lesquels l'identification a changé.

Tableau 15 : changement d'identification pour les enregistrements de l'EAR ayant été modifiés lors de la normalisation 1

	A été identifié lors d'une étape différente	A été identifié lors de la même étape	Total
Résultat différent	1 316	1 330	2 646 (6,69%)
Résultat identique	20 562	16 324	36 686 (93,31%)
Total	21 878 (55,34%)	17 654 (44,66%)	39 532

Tableau 16 : changement d'identification pour les enregistrements de l'EAR ayant été modifiés lors de la normalisation 2

	A été identifié lors d'une étape différente	A été identifié lors de la même étape	Total
Résultat différent	2 104	3 753	5 857 (19,21%)
Résultat identique	16 691	7 943	24 634 (80,79%)
Total	18 795 (61,64%)	11 696 (38,36%)	30 491

On peut émettre de sérieux doutes sur la qualité des individus identifiés lors de la même étape au sein du processus, mais pour lesquelles l'identification a changé (les 1 330). Ces cas se produisent lors de l'étape par valeur approchée, qui est celle entraînant le plus grand nombre de faux positifs, notamment parce que tous les enregistrements de bonne qualité sont généralement identifiés au préalable.

En revanche on peut supposer que ceux ayant changé d'étape, pour une étape plus stricte (les 1 316) sont vraisemblablement mieux identifiés. Il s'agit pour moitié environ des personnes nées à l'étranger pour lesquelles le pays de naissance a pu être codé.
Le reste des échos devrait être analysé mais cela n'a pas pu être fait par manque de temps.
On note toutefois une forte stabilité de l'identification puisque 93 % des enregistrements sont toujours identifiés de la même façon.

Le constat est moins bon pour le niveau 2 de normalisation, avec un changement d'identification plus fréquent. Cette fois-ci les enregistrements n'ayant pas changé d'étape sont plus nombreux, et ceux qui changent se retrouvent souvent dans l'étape la moins stricte.

On peut donc supposer une dégradation des taux de faux positifs avec ce deuxième niveau de normalisation.

Le même exercice sur les données fiscales confirme cet enseignement. Les évolutions sont moins marquées, puisque le volume de données modifiées par la normalisation est plus faible, mais la dégradation, notamment sur les étapes d'identification se produit de façon légère dès l'étape exacte, le déplacement au sein des étapes allant plutôt dans le sens le moins strict.

Tableau 17 : Résultats de l'identification du fichier FIP des Hauts-de-deSeine par l'algorithme du CSNS

	Norme 0	Norme 1	Norme 2
Identifiés	1 1846 33 (99,40%)	1 1836 54 (99,00%)	1 173 402 (98,46%)
Non identifiés	7130	8109	18361
Etape d'appariement exact	42,32 %	41,97 %	31,71 %
Etapes par relâche simple	54,10 %	54,27 %	60,26 %
Etape par valeur approchée	2,99 %	3,09 %	6,49 %
Non identifiés	0,60 %	0,68 %	1,54 %

Tableau 18 : changement d'identification pour les enregistrements de FIP ayant été modifiés lors de la normalisation 1

	A été identifié lors d'une étape différente	A été identifié lors de la même étape	Total
Résultat différent	998	831	1 829 (4,55 %)
Résultat identique	13 899	24 479	38 378 (95,45 %)
Total	14 897 (37,05 %)	25 310 (62,95 %)	40 207

Tableau 19 : changement d'identification pour les enregistrements de FIP ayant été modifiés lors de la normalisation 2

	A été identifié lors d'une étape différente	A été identifié lors de la même étape	Total
Résultat différent	19 043	9 085	28 128 (10,86 %)
Résultat identique	168 713	62 268	230 981 (89,14 %)
Total	187 756 (72,46 %)	71 353 (27,54 %)	259 109

4.3. Enseignements que l'on peut en tirer et limites

Le nettoyage des données : 80 % de la réussite de l'appariement ?

Les tableaux ci-dessous synthétisent les résultats de taux d'appariement et de précision pour les différents scénarios.

Pour les scénarios 4 et 5, la précision a été évaluée de façon grossière à partir des changements d'identification lors de l'identification au RNIPP par le CSNS. On peut supposer que le taux réel est supérieur, mais que les écarts entre les différents scénarios sont une bonne approche de l'évolution réelle.

Tableau 20 : Comparatif des taux d'appariement des différents scénarios pour l'appariement entre l'EAR et FIP

	Norme 0	Norme 1	Norme 2	**Norme 0**	**Norme 1**	**Norme 2**
	Tous les prénoms et lieux de naissance sur 5 positions			Premier prénom et lieu de naissance sur 2 positions		
exact ear fip	29,6 %	39,1 %	41,1 %	38,8 %	51,4 %	54,1 %
flou ear fip	42,0 %	47,2 %	48,5 %	48,9 %	55,6 %	57,2 %

Le tableau ci-dessus évalue l'impact des différents choix que l'on peut faire lors d'un processus d'appariement :
- le nettoyage des données ;
- la sélection des variables et niveaux des variables d'appariement ;
- le choix des fonctions de comparaison.

Ces choix ne sont pas indépendants ni exclusifs.
Le choix d'une fonction de comparaison plus lâche peut compenser certaines absences de nettoyage.
Chacun des choix que l'on peut faire influe sur le taux d'appariement final.

On constate d'abord que le premier niveau de normalisation a un effet plus important (5 à 12 points) que le second (1 à 3 points).
Ces effets ont des ordres de grandeur similaires (en nombre de points supplémentaires) que l'on utilise les variables les plus précises (tous les prénoms et lieux de naissance détaillés) ou les variables les plus agrégées.
En revanche, l'effet de la normalisation est plus faible lorsqu'on utilise des comparaisons floues plutôt que des comparaisons exactes.

Lorsqu'on fixe le niveau de normalisation et le choix des variables, l'utilisation de la comparaison floue permet de rapporter entre 3 et 12 points.

Lorsqu'on fixe le niveau de normalisation et les fonctions de comparaison, l'utilisation de variables moins précises apporte de 7 à 12 points.

On se situe ici dans un contexte où les données norminatives de l'EAR comportent des erreurs.
L'absence totale de nettoyage des données aboutit dans tous les cas à une perte sur le taux d'appariement.

Le tableau 21 concerne lui un appariement entre deux sources administratives, de meilleure qualité pour les informations nominatives notamment.

L'effet de la normalisation est dans ce cas beaucoup plus faible, alors que le choix de variables d'appariement moins précises continue d'apporter un gain important pour le taux d'appariement.

Tableau 21 : Comparatif des taux d'appariement des différents scénarios pour l'appariement entre l'EAR et FIP

	Norme 0	Norme 1	Norme 2	**Norme 0**	**Norme 1**	**Norme 2**
	Tous les prénoms et lieux de naissance sur 5 positions			Premier prénom et lieu de naissance sur 2 positions		
exact fip ts	56,1 %	57,7 %	57,9 %	71,4 %	72,7 %	72,8 %

Dans le meilleur des cas, le nettoyage des données a permis d'augmenter de 40 % le taux d'appariement.

Ces premiers scénarios nous indiquent donc :
- que le nettoyage des données n'est pas suffisant pour obtenir un taux d'appariement satisfaisant : on se situe en effet loin des taux obtenus par un processus élaboré ;
- que le bon choix des variables a un effet égal ou supérieur en terme d'augmentation du taux d'appariement ;
- que la qualité des données en entrée est un élément important pour déterminer la stratégie.

L'analyse des résultats obtenus avec l'algorithme du CSNS nous permet de mettre en balance le taux d'appariement, et une estimation des taux de faux positifs.
Dans ce second scénario, le champ du fichier de référence est beaucoup plus grand (130 millions d'individus) et conduit à un risque de faux positifs plus important.

Le processus très abouti met en œuvre des fonctions de comparaison élaborées, ainsi que des comparaisons sur des informations partielles (un seul des prénoms par exemple).
Les autres leviers d'action (comparaison floue, restriction de l'information utilisée) sont donc déjà mis en œuvre au sein du processus lui-même et ne font pas l'objet de variantes.
La seule caractéristique qui varie ici est donc le nettoyage des données.
Celui ci a un impact très faible, et parfois négatif sur le taux d'appariement.
Enfin, l'estimation des taux de faux positifs met en évidence une augmentation de celui-ci, notamment pour le nettoyage le plus important.

Dans ce contexte de processus élaboré, on en conclue donc que le nettoyage des données en entrée n'est pas efficace et peut même dégrader la qualité de l'appariement.

Tableau 22 :Comparatif des taux d'identification par l'algorithme du CSNS pour les 3 fichiers et les différents niveaux de normalisation

	Norme 0	Norme 1	Norme 2
CSNS EAR	94,2 %	94,5 %	93,7 %
CSNS FIP	99,4 %	99,3 %	98,5 %
CSNS TS	99,0 %	99,0 %	98,2 %

Tableau 23: Comparatif des taux de faux positifs à l'issue de l'algorithme du CSNS pour les 3 fichiers et les différents niveaux de normalisation

	Norme 0	Norme 1	Norme 2
CSNS EAR	Inconnu >1,12 %	Inconnu >2,22 %	Inconnu >4,96 %
CSNS FIP	Inconnu >0,08 %	Inconnu >0,15 %	Inconnu >2,40 %
CSNS TS	0,43 %	0,42 %	1,43 %

Le nettoyage des données : 80 % du temps de travail pour réaliser un appariement ?

En terme de temps de travail, les travaux se sont étalés sur environ 6 semaines, comportant les tâches de nettoyage des données, les appariements, l'analyse des résultats et la rédaction.
La phase de nettoyage des données a pris entre deux et trois semaines, notamment pour la constitution du dictionnaire des prénoms qui a mobilisé des appariements volumineux (sans blocage de façon intentionnelle).
Mais par ailleurs, le réglage fin des règles de comparaison et le poids relatif accordé aux différentes variables peut également demander un effort important pour bien calibrer un algorithme d'appariement. Ce travail n'a pas été mené ici : les mesures de similarité ont été choisies rapidement, et sans tests préalables.

L'estimation du temps qu'il est nécessaire de consacrer au nettoyage des données est donc probablement sur évaluée ici.

Le processus CSNS a lui consacré un temps assez important à l'élaboration des différentes étapes d'appariement, et notamment aux poids accordés à chaque correspondance dans la requête par valeur approchée.
Le constat est sans appel : le temps passé à régler finement les règles de comparaison est plus efficace que le nettoyage des données.

Enfin, dans un processus d'appariement standard, il peut y avoir un temps d'évaluation de la qualité des appariements plus conséquent.
Cette évaluation peut conduire à revenir sur les choix effectués pendant ou après la phase de nettoyage, et à modifier le processus, et n'a pas été mesurée ici.

La préparation des données, étape cruciale du processus d'appariement ?

Plus que le nettoyage des données lui-même, ce qui semble le plus important au cours de la préparation des données est la qualification de celles-ci, et les conclusions qu'on en tire.

Les opérations de nettoyage ayant apporté le plus au processus d'appariement ici semblent être :
- la bonne distinction des informations dans les champs nominatifs (enlever les mots liés aux relations, les civilités) ;
- le codage du lieu de naissance lorsque celui-ci était absent.

La tentative de correction des erreurs de frappe ou de saisie permet certes de corriger certains noms et prénoms, mais introduit également de nouvelles erreurs d'une part, et peut être compensée par une relâche de la fonction de comparaison d'autre part.

Une approche systématique de la phase de qualification peut être intéressante, afin de ne pas s'attarder sur des cas particuliers, mais plutôt d'essayer d'identifier les problèmes les plus importants, ainsi que d'évaluer la possibilité d'y remédier.

La connaissance de la qualité des variables de couplage est indispensable pour déterminer ensuite la stratégie de couplage, à savoir le choix des variables, le séquencement en étapes du processus, et le choix des fonctions de comparaison.

4.4. Limites et travaux potentiels d'amélioration

Le nettoyage des données (comme la conformité des prénoms à un dictionnaire) a été appliqué sans distinction aux valeurs erronées (erreurs de saisie par exemple) et aux prénoms trop rares pour être présents dans le dictionnaire. Il pourrait être intéressant toutefois de développer un dictionnaire de synonymie pour les erreurs de saisie, en limitant sa portée (notamment aux personnes nées en France avant le changement de loi sur les prénoms de 1993), en limitant la distance acceptée pour la correction, voire en introduisant une validation manuelle du dictionnaire de synonymes.
Il est difficile cependant de distinguer des prénoms bien orthographiés mais peu fréquents d'erreurs de frappe ou de saisie portant sur un prénom courant.
De plus, certaines erreurs de frappe ou de saisie peuvent tout à fait correspondre à un prénom réel : Yvon à la place de Yvette, ou plus surprenant, Benoit à la place de Robert par exemple (cas réel d'erreur).

Une autre piste non explorée serait de s'intéresser à la corrélation des erreurs sur différentes variables, pour distinguer des enregistrements de moindre qualité et leur appliquer un traitement différent.

On a choisi de ne pas mener de travaux sur les adresses, afin de disposer d'éléments d'appréciation de la qualité des paires appariées. Seule l'information sur le département de résidence était prise en compte, puisque l'on a réduit le champ géogaphique pour les appariements directs (hors CSNS).
Mais les informations localisantes ont une grande importance dans les appariements, et la qualité et la comparabilité des adresses est un enjeu majeur de la mobilisation de cette information.

Enfin, ces analyses ont été menées sur des fichiers de plutôt bonne qualité, et sur les informations nominatives et de naissance des personnes. Les conclusions seraient certainement différentes pour des champs textuels plus riches et complexes, comme des libellés de produits ou des actes médicaux.

Conclusion

« La normalisation, c'est 80 % du succès d'un appariement », c'est une des phrases qui a provoqué la réflexion, et ces travaux.

A l'issue de ceux-ci, on dira plutôt que « la phase de préparation des données pose les fondations d'un processus d'appariement ». Si elle est correctement menée, en comportant notamment une étape préalable de qualification des données, elle conduit au choix des bonnes variables et permet d'affiner la stratégie d'appariement.
Le nettoyage des données, s'il est utile dans sa forme la plus simple, peut entraîner une dégradation des résultats de l'appariement, en particulier si le champ couvert par les données est vaste, et sans information localisante.

Bibliographie

[1] RANDALL S. M., FERRANTE A. M., BOYD J. H., SEMMENS J. B, « The effect of data cleaning on record linkage quality », *BMC Medical Informatics and Decision Making,* Vol. 13, n° 1, décembre 2013

[2] DOIDGE J., CHRISTEN P. et HARRON K., « Quality assessment in data linkage », published by the UK Government Analysis Function and Office for National Statistics as part of the National Statistician's Quality Review : Joined up data in government: the future of data linking methods, dernière mise à jour juin 2021

[3] CHRISTEN P., « Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection », Springer, 2012

[4] CHRISTEN P., CHURCHES T. et XI ZHU J., 2002. « Probabilistic Name and Address Cleaning and Standardisation », 2002

[5] SANMARTIN C., « *Modèle du processus d'un projet de couplage d'enregistrements* », Statistique Canada, 2017

[6] TUOTO T., CIBELLA N., FORTINI M. et SCANNAPIECO M., « From theory to practice: the software RELAIS as a solution for record linkage », 2010

[7] GILL L., « Methods for Automatic Record Matching and Linkage and their Use in National Statistics », National Statistics Methodological Series no. 25, 2001

[8] AUGUSTINE E., REDDY V., ROTHSTEIN J., « Linking Administrative Data: Strategies and Methods », California Policy Lab , 2018

[9] Data Linking Information Series, Sheet 4: Probabilistic linking, Australian Bureau of Statistics, 2018 https://toolkit.data.gov.au/Data_Linking_Information_Series_Sheet_4:_Probabilistic_linking.html

10] Data Linking Information Series, Sheet 5: Linked data quality, Australian Bureau of Statistics, 2018 (https://toolkit.data.gov.au/Data_Linking_Information_Series_Sheet_5:_Linked_data_quality.html)