Les appariements : finalités, pratiques et enjeux de qualité

Heidi Koumarianos[*], *Olivier Lefebvre*[**] *et Lucas Malherbe*[***]

Les appariements rapprochent les données individuelles issues de fichiers différents. Dans un contexte de disponibilité croissante de sources, notamment administratives, ils sont de plus en plus fréquemment utilisés dans la statistique publique, à des fins d'analyse, pour éclairer des questions nouvelles, ou pour améliorer des processus de production. Ces traitements statistiques posent souvent des difficultés, liées aux imperfections des données utilisées et à leur volume.

Un cadre juridique approprié est nécessaire pour les mettre en œuvre, compte tenu des enjeux de respect des secrets, en particulier la protection des données à caractère personnel. Une bonne connaissance des données utilisées et une attention particulière pour déterminer le paramétrage sont également indispensables pour assurer la meilleure qualité possible du résultat au regard de l'usage attendu, un appariement n'étant jamais fiable à 100 %.

La qualité de ces appariements est donc un enjeu majeur pour la statistique publique et passe par des évaluations directes du processus, nécessairement complétées par l'étude des impacts statistiques des appariements sur les données produites.

Record linkage reconciles individual data that are taken from various data files. It is used more and more frequently in official statistics, whether for analytical purposes, to investigate new subjects, or to improve production processes. Such statistical processing often raises issues relating to the imperfections of the data used and their volume.

Given the issues related to the compliance with data confidentiality, especially regarding personal data protection, an appropriate legal framework is required to implement them. A good knowledge of the data used and particular care in determining the parameters are also necessary to ensure the best possible quality of the results, since record linkage is never 100% reliable.

The quality of these matches is therefore a major issue for official statistics, and requires a direct assessment of the process, necessarily supplemented by a study of the statistical impact of the linkage on the data produced.

* Méthodologue, DMCSI, Insee.
heidi.koumarianos@insee.fr

** Maître d'ouvrage du programme Résil, DSDS, Insee.
olivier.lefebvre@insee.fr

*** À la date de rédaction, data scientist, DMCSI, Insee.

La mesure des revenus des ménages ou le suivi de l'insertion professionnelle des diplômés ont un point commun : ces deux opérations reposent sur l'utilisation de plusieurs sources de données, qu'il faut combiner entre elles. Au niveau le plus fin, il faut réunir, pour chaque individu ou ménage, les données qui le concernent dans chacune des sources. Cela permet de couvrir l'ensemble des sources de revenus des ménages, qu'ils soient imposables ou non imposables, ou encore de suivre le parcours professionnel des diplômés, et notamment leurs conditions d'entrée sur le marché du travail.

Combiner différentes sources permet une observation plus riche et plus efficace. Utilisable dans bien des domaines, l'appariement (l'opération qui permet d'associer et combiner les sources) constitue en quelque sorte une technique de collecte, avec ses contraintes techniques, ses défis méthodologiques, son encadrement juridique, ses enjeux déontologiques. La plupart des instituts de statistique utilisent cette technique pour la production de données statistiques, en lien avec l'utilisation croissante de données administratives, souvent très précises mais très ciblées quant à leur contenu, et qui demandent donc à être complétées.

L'Insee et plus généralement la statistique publique procèdent à des appariements depuis de longues années ; cette pratique s'est progressivement étendue grâce au développement de techniques performantes de traitement des données et à un accès facilité aux fichiers source. On peut ainsi citer Fidelimmo (André et Meslin, 2022), qui permet de mieux analyser le patrimoine immobilier des ménages et les effets redistributifs ou non de la taxe foncière, InserJeunes (Midy, 2021) pour la mesure de l'insertion professionnelle des apprentis, ou encore Sirus[1] (Hachid et Leclair, 2022), colonne vertébrale de la statistique d'entreprises.

▶ Apparier pour enrichir ou mieux comprendre des sources…

L'appariement, lorsqu'il est effectué pour un usage statistique, permet en première approche d'apporter un complément d'information à un fichier statistique existant. Plus généralement, on peut distinguer plusieurs usages de l'appariement :

- **Compléter le champ d'analyse** : le dispositif Filosofi[2] par exemple utilise différentes sources pour reconstituer les revenus des ménages, qu'il s'agisse de revenus du travail ou de prestations sociales. Il permet d'avoir une vue plus complète des revenus des ménages, à une échelle géographique fine pouvant aller jusqu'aux quartiers d'une ville, si ceux-ci sont d'une taille suffisante.
- **Éclairer certains phénomènes** : par exemple, apparier des fichiers de diplômés de l'enseignement supérieur avec des fichiers d'emploi permet de décrire l'insertion professionnelle des jeunes diplômés.

1 Sirus : Système d'immatriculation au répertoire des unités statistiques.
2 Filosofi : ensemble d'indicateurs sur les revenus localisés sociaux et fiscaux.

• **Connaître l'impact d'une aide sociale ou d'une aide à destination des entreprises** : apparier le fichier des bénéficiaires à un fichier décrivant la situation (emploi, réussite dans l'enseignement supérieur, résultats financiers) permet de savoir si l'aide a été suivie d'effets et mettre en place une évaluation de l'impact de l'aide.

• **Mieux comprendre le contenu des sources analysées** : par exemple, l'appariement du fichier des demandeurs d'emploi avec celui de l'enquête Emploi a permis de mieux comprendre les évolutions différentes du chômage au sens du Bureau International du Travail et du nombre de demandeurs d'emploi inscrits à France Travail[3].

▶ … ou encore améliorer des processus de production

Au-delà de la construction de données nouvelles, combiner les sources permet d'améliorer notablement des processus de production statistique. On modifie la phase de collecte des informations ou de contrôle ou d'évaluation de la qualité des sources (cohérence avec les concepts que l'on souhaite mesurer, mesure de la couverture). Plus précisément, un appariement conduit à :

• **Alléger les questionnaires d'enquête** : le principe est de ne pas demander à un ménage (ou à une entreprise) une information qu'il (ou elle) a déjà transmise à une administration, selon le principe « Dites-le nous une fois ». Par exemple, l'appariement de l'enquête Emploi avec les données fiscales permet de réduire le nombre de questions portant sur le revenu.

• **Mettre à jour un répertoire ou un référentiel** (RNIPP[4], Sirene[5], REU[6], Résil[7], Sirus, etc.) : on ajoute de nouvelles entités dans le répertoire (auquel cas il est essentiel de vérifier qu'elles n'y figurent pas déjà, pour éviter les doublons) ou on met à jour certaines caractéristiques (auquel cas il est essentiel de vérifier qu'on met à jour la bonne observation). La qualité des répertoires est indispensable à celle des processus statistiques (Espinasse et Roux, 2022 ; Demotes-Mainard, 2019).

• **Analyser la couverture d'une source en l'appariant à un référentiel** : c'est un progrès important dans l'analyse des évolutions ou dans le traitement des « trous de collecte », comme pour les non-répondants d'une enquête (ce qui est actuellement possible en statistique d'entreprises avec Sirus et qui pourra l'être en statistique démographique et sociale avec Résil).

▶ De quoi parle-t-on ?

En pratique, l'appariement désigne l'opération de rapprochement, au niveau de chacune des unités d'observation, de deux fichiers de données A et B, soit pour enrichir l'un des fichiers avec des variables supplémentaires ou mises à jour, soit pour créer un nouveau fichier contenant tout ou partie des variables de chacun des fichiers ***(Figure 1)***.

3 France Travail : depuis 2023, France Travail succède à Pôle emploi.
4 RNIPP : Répertoire national d'identification des personnes physiques.
5 Sirene : Système national d'identification et du répertoire des entreprises et de leurs établissements.
6 REU : Répertoire électoral unique.
7 Résil : Répertoire Statistique des Individus et des Logements. Voir l'article d'Olivier Lefebvre sur Résil dans ce même numéro.

Figure 1 - Un exemple d'appariement pour étudier le lien entre profession et niveau de diplôme

Table A

Identifiant | Traits d'identité (variables d'appariement) | Variable statistique d'intérêt

ID_A	Nom	Prénom	Date de naissance	Profession
A1	RAFIK	Kélian	03/02/2001	Enseignant
A2	GIROUT	Claudia	05/08/1993	Esthéticienne
A3	REGARD	Esteban	23/11/1997	Cariste
A4	MOREAU	Lya	12/11/2002	Agent de police
A5	PEREZ	Maria	27/06/1987	Cadre commercial
A6	LEFEBVRE	Olivier	27/06/1999	Statisticien

?

Table B

Identifiant | Traits d'identité (variables d'appariement) | Variable statistique d'intérêt

ID_B	Nom	Prénom	Date de naissance	Dipôme le plus élevé
B1	LEFEVRE	Oscar	02/04/1975	DUT
B2	RAFIK	Kélian	03/02/2001	Master
B3	REGAR	Esteban	23/11/1997	Bac
B4	GIROUT	Claudia	05/08/1993	Brevet d'études professionnelles
B5	PEREZ	Marina	27/06/1987	Doctorat

Table de résultats

Id	Date de naissance	Profession	Dipôme le plus élevé
R1	03/02/2001	Enseignant	Master
R2	05/08/1993	Esthéticienne	BEP
R3	23/11/1997	Cariste	Bac
R4	27/06/1987	Cadre commercial	Doctorat

Note de lecture : À partir des deux fichiers A et B, contenant respectivement la profession et le diplôme, on veut construire un fichier de résultat contenant, pour les individus figurant dans les deux fichiers, leur profession et leur diplôme. Le processus de construction de ce fichier est détaillé dans la suite de l'article.

> ***Si les fichiers ont un identifiant commun de bonne qualité, l'opération technique est triviale… Elle se résume alors à une « jointure » sur cet identifiant.***

Un appariement peut être destiné à un usage administratif (par exemple une vérification de droits), opérationnel (par exemple fusionner des fichiers de clients) ou à un usage statistique (***encadré 1***).

La suite de cet article porte sur la manière de réaliser les appariements et sur leurs usages statistiques.

Pour que ce rapprochement soit pertinent, il faut savoir à quelle ligne[8] du fichier B correspond chaque ligne du fichier A, en s'assurant donc qu'elles portent toutes deux sur la même unité d'observation. Si les fichiers ont un identifiant commun de

8 Une ligne correspond à un enregistrement représentant un individu.

bonne qualité, l'opération technique est triviale : deux lignes ayant le même identifiant portent naturellement sur la même unité d'observation. Elle se résume alors à une « jointure » sur cet identifiant (il convient néanmoins d'en traiter tous les autres aspects : encadrement juridique et déontologique, analyse de la qualité statistique du résultat, etc.). Dans le cas contraire, deux approches sont possibles :

- l'identification individuelle : on trouve un identifiant commun en comparant ces deux fichiers A et B à un troisième fichier, C, souvent de taille plus importante, qui joue le rôle de référentiel. Cette étape d'identification vise donc à chercher successivement, pour les observations des fichiers A et B, à quelle ligne du fichier C elles correspondent, et à faire la « jointure » sur l'identifiant correspondant[9] ;

- la confrontation de paires : on confronte directement les deux fichiers, en cherchant parmi toutes les paires d'observations possibles, lesquelles correspondent à la même entité ; on utilise pour cela des techniques spécifiques, fondées soit sur l'application de règles de décisions successives, soit sur des modèles probabilistes.

Il s'agit d'appariements de micro-données. Il existe également des appariements statistiques, fondés sur l'appartenance à des strates, que l'on n'évoquera pas ici. Les anglophones désignent les premiers par « *record linkage* » et les seconds par « *propensity score matching* » (Rosenbaum et Rubin, 1983).

▶ Encadré 1. Appariement, enrichissement, interconnexion, couplage : tous synonymes ?

Un **enrichissement** de données d'un fichier par un autre fichier consiste à trouver, pour un individu donné, les informations qui le concernent dans deux fichiers différents, puis à constituer un troisième fichier avec les données ainsi rassemblées.

Même si les appariements ne désignent que la première phase de cette opération (celle qui consiste à relier entre elles deux observations relatives à la même entité), les statisticiens désignent souvent cette technique sous le nom d'« **appariement** » de fichiers (ou de données). C'est ce terme qui figure dans la loi de 1951 et le décret d'application de la loi pour une République numérique*.

Les rédacteurs de la loi relative à l'Informatique, aux fichiers et aux libertés de 1978 et du Règlement général pour la protection des données ont employé les termes « **interconnexion** », « **rapprochement** », ou « **mise en relation** » de fichiers. L'interconnexion de fichiers et leur rapprochement constituent deux formes de mises en relation de fichiers ; le terme d'interconnexion est plus souvent employé pour des mises en relation présentant un fort degré d'automatisation, voire totalement automatisées.

D'autres termes peuvent être utilisés par les statisticiens pour désigner les appariements : **combinaison ou couplage** de fichiers (le dernier terme étant celui utilisé au Canada à la fois par Statistique Canada et dans la directive** qui régit ces opérations).

Pour clarifier le sujet et mieux asseoir la notion d'appariement en droit français, le décret fondant Résil propose une définition de l'appariement*** :

« Ces appariements constituent des mises en relation, au sens du 3° du I de l'article 33 de la loi [Informatique et Libertés], entre les données à caractère personnel enregistrées sur le « répertoire statistique des individus et des logements » et des sources de données statistiques tierces. Ils donnent lieu à la création de nouveaux fichiers, lesquels constituent des traitements de données à caractère personnel au sens du [RGPD]. »

* *Décret n° 2016-1930 du 28 décembre 2016 portant simplification des formalités préalables relatives à des traitements à finalité statistique ou de recherche (voir fondements juridiques).*

** *Statistique Canada. 2017. Directive sur le couplage de microdonnées. https://www.statcan.gc.ca/fr/enregistrement/politique4-1.*

*** *Décret n° 2024-12 du 5 janvier 2024 portant création d'un traitement automatisé de données à caractère personnel dénommé « répertoire statistique des individus et des logements » (Résil) (voir fondements juridiques).*

9 Si l'un des deux fichiers est lui-même un référentiel, par exemple si on cherche à mettre à jour ce dernier, on ne réalise l'opération d'identification que sur l'autre fichier.

▶ Un traitement de données nécessite un cadre juridique adapté

L'appariement de données personnelles est un traitement de données au sens juridique qu'il faut traiter comme tel. Un responsable du traitement doit être identifié, charge à lui de s'acquitter des obligations imposées par le Règlement général pour la protection des données (RGPD) et la loi Informatique et Libertés[10] : vérification du respect des principes de nécessité, minimisation et proportionnalité[11], inscription au registre des traitements de son administration, réalisation d'une étude d'impact si ce traitement présente certaines caractéristiques (par exemple s'il porte sur une population très nombreuse ou mobilise des variables sensibles, etc.). Quand l'appariement porte sur des données statistiques ou est réalisé à des fins statistiques, les données sont placées sous la protection de la loi de 1951 sur l'obligation, la coordination et le secret en matière de statistiques[12] (*Redor, 2023*).

▶ Une pratique souvent délicate

Quand on ne dispose pas d'identifiant commun aux deux fichiers, l'appariement doit être effectué sur des variables permettant d'identifier sans ambiguïté les personnes.

Un fichier comporte plusieurs types d'informations jouant un rôle différent dans un processus d'appariement. On peut classer ces informations en trois catégories :

- les informations identifiantes primaires : il s'agit de traits d'identité associés à un individu de manière unique et très stable dans le temps. Pour une personne, il s'agit de ses nom(s), prénom(s), lieu et date de naissance[13] ;
- les informations identifiantes secondaires : ce sont des informations qui ne sont pas associées de manière unique et permanente à un individu, mais peuvent permettre d'améliorer le processus d'appariement. Pour une personne, il pourra s'agir par exemple de sa commune et de son adresse de résidence ;
- les autres informations ne sont généralement pas utilisées dans un processus d'appariement, mais sont des variables d'intérêt pour le fichier statistique produit. Elles peuvent toutefois être mobilisées lors de l'évaluation de la qualité, en détectant par exemple des incohérences dans les enregistrements appariés.

Cela pose plusieurs difficultés : ces informations ne sont pas toujours présentes dans les fichiers et peuvent être entachées d'erreurs ou d'omissions. Leur comparaison est très coûteuse en ressources informatiques, et ce coût croît rapidement avec la taille des données.

10 Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés (voir les références juridiques en fin d'article).
11 Selon le RGPD, « les données à caractère personnel doivent être [...] adéquates, pertinentes et limitées à ce qui est nécessaire au regard des finalités pour lesquelles elles sont traitées (minimisation des données) ».
12 Loi n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques (voir les références juridiques en fin d'article).
13 Pour un établissement, la raison sociale et l'adresse constituent des informations primaires.

▶ Quels critères utiliser quand on n'a pas d'identifiant commun ?

Une première condition est bien sûr de disposer des informations identifiantes dans les deux fichiers, et que ces informations soient comparables, c'est-à-dire que leur contenu sémantique et leur représentation soient identiques. Par exemple, on dispose de la commune de résidence dans les deux fichiers et cette dernière figure dans les deux cas sous forme de libellé, ou de code selon une nomenclature identique.

Outre ce critère nécessaire de comparabilité des données, trois conditions principales sont requises pour réaliser avec succès un appariement :

- la richesse des informations ;
- la qualité des informations ;
- un processus efficace pour traiter un important volume de données.

Les informations dont on dispose doivent être suffisamment nombreuses, discriminantes et précises pour permettre de distinguer les individus les uns des autres. Connaître le mois de naissance par exemple est beaucoup moins informatif que connaître le prénom (un douzième de la population partage la même valeur pour le mois de naissance). Plus l'information est précise, plus elle permettra de distinguer un individu d'un autre. Il sera ainsi plus intéressant d'utiliser une date de naissance complète plutôt qu'une année, une commune de naissance plutôt qu'un département, etc.

On peut aussi chercher à mobiliser un plus grand nombre d'informations, en utilisant une adresse de résidence, en plus des traits d'identité. Il peut toutefois subsister malgré tout, des cas non univoques (en cas d'homonymie par exemple), qui sont d'autant plus nombreux que les données sont peu précises ou comportent des erreurs.

▶ Les données identifiantes ne sont jamais parfaites : comment s'en accommoder ?

La deuxième difficulté est liée à la qualité des données. Toute information manquante, incomplète ou erronée est susceptible de nuire à la qualité de l'appariement, en entraînant des appariements à tort, ou à l'inverse, en « ratant » de nombreuses paires. Lors d'une enquête statistique, les traits d'identité des personnes interrogées ne font pas partie des variables d'intérêt, et peuvent alors souffrir de défauts de qualité. Or, ces informations sont indispensables, dès lors que l'on envisage un processus d'appariement.

Pour pallier ces difficultés, le statisticien dispose d'outils permettant de préparer les données utilisées pour l'identification en vue d'améliorer le résultat de son appariement.

Ces outils ne sont pas magiques et ne peuvent créer une bonne information là où elle manque ou est erronée. On peut toutefois recourir à un processus de normalisation des données, notamment pour améliorer la comparabilité :

- on norme les données dans un format identique de part et d'autre : transformer

décembre en 12, ou un nom de commune en son code commune ou encore utiliser une casse similaire (suppression des caractères accentués, passage en minuscules, etc. (Cotton et Haag, 2023)) ;

- on structure ensuite les informations : normaliser un libellé d'adresse (type de voie, nom de voie, numéro dans la voie, indice de répétition, commune) ou encore identifier un nom dans un champ comportant aussi la civilité ; ce second point est toutefois moins évident, car il nécessite d'analyser les libellés.

Parfois le sujet est plus délicat, lorsqu'un fichier administratif comporte dans un même champ le nom marital et le nom de naissance ou encore le nom de plusieurs titulaires d'une carte grise par exemple. Ce processus est efficace dans la mesure où les traitements entrepris sont assez déterministes (comme rassembler les Bd, boul, Boulevard sous une seule dénomination). Mais attention à ne pas aller trop loin : la volonté de supprimer des données erronées peut conduire à supprimer de l'information, et finalement dégrader le processus (Koumarianos, 2022). Ce reproche est parfois adressé aux algorithmes phonétiques : ils ont pour objectif de neutraliser les différences orthographiques, mais ils peuvent alors considérer à tort que Lefebvre et Lefèvre, ou Schmidt et Schmitt sont deux modalités identiques (Randall et alii, 2013).

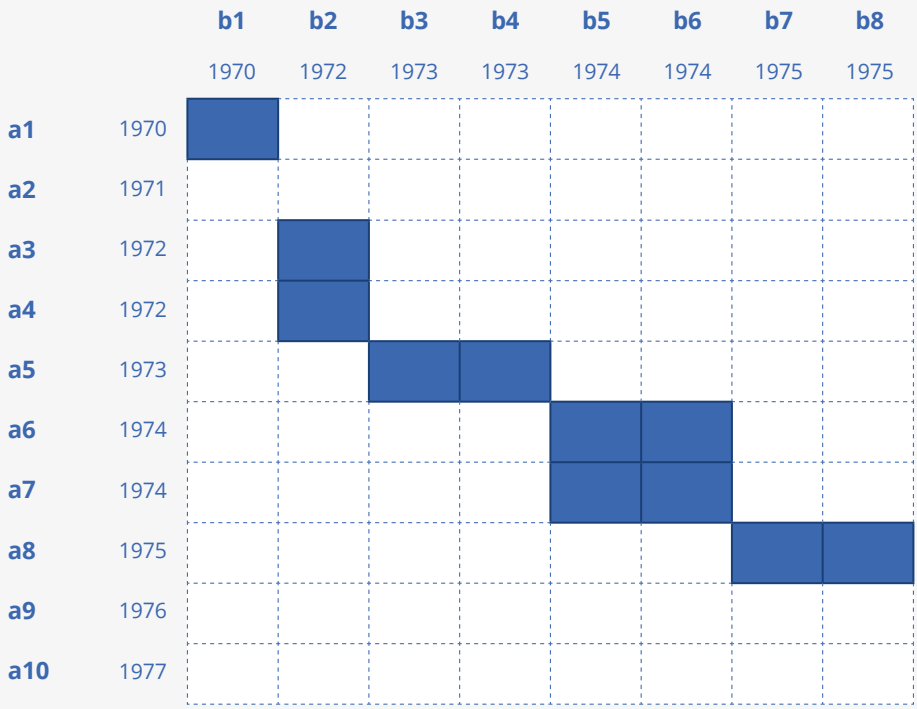
Lorsqu'il subsiste des erreurs dans les informations, de type faute de frappe ou faute d'orthographe, il sera souvent plus efficace de gérer ces problèmes ultérieurement dans le processus d'appariement, en mobilisant des mesures de similarité tenant compte de ces potentielles coquilles, plutôt que d'établir une comparaison s'appuyant sur une égalité stricte.

▶ L'appariement est une opération gourmande, comment la rendre frugale ?

Enfin un troisième enjeu est celui de la performance informatique du processus. Un processus d'appariement revient à sélectionner, au sein de deux fichiers de tailles N_1 et N_2, les paires d'individus identiques au sein de l'ensemble des paires possibles. Cet ensemble est très grand, de taille $N_1 \times N_2$, alors que le nombre de paires d'individus identiques est inférieur ou égal au minimum de (N_1,N_2). En dépit des progrès techniques, un appariement reste un problème de grande taille. Apparier deux fichiers de 60 000 lignes signifie raisonner dans un ensemble de 3,6 milliards de paires potentielles. Parmi l'ensemble des paires, un grand nombre rassemble deux individus qui ne se ressemblent pas du tout. Il n'est donc pas efficace de constituer l'ensemble théorique de toutes les paires. On cherche souvent à le réduire à un sous-ensemble de paires plus vraisemblables. C'est ce qu'on appelle le blocage : réduire la dimension du problème en ne comparant, par exemple, que des individus nés la même année (***figure 2***). Les informations utilisées pour le blocage doivent être d'excellente qualité. Dans le cas contraire, cela entraînerait de nombreux appariements manqués.

> ***C'est ce qu'on appelle le blocage : réduire la dimension du problème.***

▶ Figure 2 - Exemple de blocage sur l'année de naissance



Lecture : Dans cet exemple fictif, on cherche à apparier le fichier A qui comporte 10 individus et le fichier B qui en comporte 8.
On bloque sur l'année de naissance, ce qui permet de travailler sur les paires d'individus nées la même année.
Au lieu de constituer 80 paires à des fins de comparaison (l'ensemble des carrés de la matrice), seules 11 paires sont étudiées (les carrés bleus).

▶ Réussir son appariement : un équilibre délicat entre théorie et technicité, connaissance des données et empirisme

Pour réaliser un appariement de données, il est fréquent de recourir à un outil dédié. On peut mobiliser des packages d'outils statistiques « au cas par cas », mais dans un contexte de production répétée, il est courant de disposer d'outils dédiés à cette opération. Ces outils peuvent être génériques et permettre l'appariement de n'importe quel jeu de données : ils comportent alors généralement un ensemble de paramètres à choisir avec soin pour disposer d'un résultat de bonne qualité.

Au sein des Instituts nationaux statistique (INS), il existe des outils généralistes d'appariement, développés en interne ou plus largement par une autre administration et qui proposent un ensemble d'outils (fonctions de comparaison, de classification, choix

de méthodes de blocage) comme l'outil Relais développé par Istat[14] (Cibella et alii, 2012), G-link développé par StatCan[15] (Chevrette, 2011) ou SPLink utilisé par l'ONS[16] britannique (Cleaton et alii, 2022).

D'autres outils sont plus spécifiques et sont conçus pour répondre à un besoin plus ciblé : par exemple, les outils récemment développés au sein du service statistique public français comme l'outil Rapsodie[17], spécialisé dans l'appariement avec les données fiscales (Jabot et Treyens, 2018) ou l'outil Inserjeunes (Midy, 2021).

Les méthodes utilisées ne sont pas spécifiques, mais l'utilisation régulière sur un certain type de données conduit à des sélections de règles et de paramètres particulièrement adaptés à un jeu de données spécifique : bien coder des informations manquantes au sein d'un fichier, par exemple la modalité SNP (Sans Nom Prénom) dans un fichier administratif, ou prendre en compte les valeurs refuge pour les dates de naissance au 01/01 ou tous les 15 chaque mois[18].

▶ Encadré 2. Les principales étapes d'un processus d'appariement

De nombreux auteurs s'accordent à formaliser les différentes étapes d'un processus d'appariement (*Christen, 2012*), en distinguant :

– une phase de préparation des données (qui comprend l'analyse de la qualité, la normalisation des variables) ;

– une phase de constitution des paires qui tient compte des problématiques de volume et optimise le sous-ensemble constitué ;

– une phase de comparaison des individus au sein des paires qui utilise des fonctions plus ou moins complexes de calcul de similarité ou de distance ;

– une phase de classification qui sélectionne les paires retenues et écarte les paires rejetées ;

– une phase d'évaluation toujours nécessaire qui conduit parfois à modifier les étapes précédentes.

À chaque étape d'un appariement (***encadré 2***), l'expertise et la connaissance des données sont nécessaires et permettent souvent d'améliorer les résultats du processus. Il n'existe pas de solution toute prête, adaptée à tous les appariements. Il est nécessaire de tenir compte de la qualité des données, de leurs caractéristiques pour sélectionner la méthode pertinente et choisir les bons paramètres pour les fonctions de comparaison et de classification.

> ***L'appariement est un processus comportant une dimension de réglages fins d'un ensemble de paramètres.***

Quel que soit l'outil choisi, l'appariement est un processus comportant une dimension de réglages fins d'un ensemble de paramètres, souvent approchés de manière itérative et empirique.

14 Relais : *Record Linkage At Istat* ; Istat : Institut national de statistique italien.
15 G-Link : *Generalized system for record linkage* ; StatCan : Institut national de statistique canadien.
16 SPLink : *probabilistic record linkage at scale* ; ONS : Office for National Statistics, Institut national de statistique du Royaume-Uni.
17 Rapprochement des données sociales, des enquêtes et des impôts.
18 Il s'agit de valeurs du domaine de définition de la variable concernée, attribuées parfois en cas de non-réponse. Ainsi la date du 01/01 est très souvent attribuée lorsqu'un jour de naissance est inconnu.

Comparer les informations, puis sélectionner les paires : le cœur d'un processus d'appariement

La suite de l'article n'est pas une description détaillée des étapes de comparaison et de classification mais donne les grandes lignes des étapes centrales d'un processus d'appariement (Christen, 2012, et Malherbe, 2023 pour une présentation plus complète).

Après avoir identifié un ensemble de paires potentielles, par exemple après une étape de blocage, on procède à la classification de ces paires. Pour chacune d'elles, on compare les deux enregistrements liés. Ceci permet de déterminer s'il s'agit d'une paire d'individus vraiment identiques, vraiment différents ou si un doute subsiste. Il existe une grande variété de méthodes de classification. Elles diffèrent sur un ensemble d'éléments, en particulier le caractère plus ou moins automatique de la définition des paramètres et le recours ou non à un ensemble de paires annotées[19]. L'approche dite probabiliste se caractérise par un degré d'automatisation relativement élevé tout comme l'utilisation du machine learning[20], là où les autres approches nécessitent de définir certains paramètres clés de façon manuelle, grâce à l'expertise et la connaissance des données du statisticien.

Les méthodes déterministes : système de règles…

Cette méthode consiste à apparier les deux fichiers au cours de plusieurs étapes en commençant par des règles strictes et en relâchant progressivement les contraintes. Les individus appariés à une étape ne sont plus considérés pour les étapes suivantes.

> ***Plusieurs étapes en commençant par des règles strictes et en relâchant progressivement les contraintes.***

La première étape est en général un appariement exact : si toutes les variables d'appariement d'une ligne du fichier A sont identiques à celles d'une ligne du fichier B (par exemple, mêmes nom,prénom, date et lieu de naissance), on apparie les 2 lignes. Les étapes suivantes autorisent des légères différences et deviennent de moins en moins strictes. Les autoriser peut se faire soit en excluant simplement un champ de la comparaison, soit en imposant une contrainte plus souple qu'une correspondance exacte. L'idée étant de relâcher progressivement les contraintes, ce sont plutôt les champs les moins discriminants ou ceux contenant le plus d'erreurs qui sont relâchés en premier, par exemple le jour de naissance plutôt que le nom de famille : on perd moins d'information et on retire des données erronées pouvant conduire à un appariement à tort.

19 Ce sont des paires pour lesquelles, après observation humaine le plus souvent, on apporte et parfois commente l'information suivante : individus identiques, individus différents, impossible de décider.

20 L'apprentissage automatique (*machine learning* en anglais) est un champ d'étude de l'intelligence artificielle qui vise à donner aux machines la capacité d'« apprendre » à partir de données, via des modèles mathématiques. Cette approche n'a pas fait ses preuves dans le cadre des appariements en raison du caractère asymétrique du problème de classification d'une part (on cherche n paires parmi un ensemble de taille n^2), et du faible nombre de variables d'autre part (Malherbe, 2023).

▶ ... ou somme pondérée de mesures de similarité...

Une autre approche consiste à calculer des mesures de similarité pour chaque champ identifiant puis à les agréger pour obtenir une mesure globale de similarité pour chaque paire. Une mesure de similarité textuelle est un nombre qui représente la proximité de deux mots ou textes. Les mesures de similarité classiques pour les appariements reposent sur la distance de Levenshtein[21] ou celle de Jaro-Winkler[22] (Herzog et alii, 2007). La similarité globale au niveau de la paire s'obtient ensuite par une somme pondérée des similarités de chaque champ. Les poids associés aux différentes variables sont définis de façon empirique par le statisticien, sur la base de leur caractère plus ou moins discriminant ainsi que de leur qualité.

Cette méthode s'accompagne presque systématiquement de la sélection d'un seuil. Dans ce cas, une paire n'est liée que si sa similarité globale dépasse le seuil.

▶ ... ou appariement par moteur de recherche : un outil efficace pour gérer d'importants volumes de données

Une troisième approche, beaucoup moins conventionnelle, consiste à utiliser un moteur de recherche textuelle, comme Elasticsearch[23] ou Solr. Ce type d'outils est plutôt conçu initialement pour rechercher efficacement de l'information dans un ensemble de textes très vaste, comme l'ensemble des produits sur un site d'e-commerce par exemple. Il peut pourtant se révéler très utile pour réaliser des appariements, en particulier lorsque les fichiers sont très volumineux.

D'un point de vue pratique, un appariement avec un moteur de recherche se déroule de façon assez différente des approches précédentes. La première étape consiste à indexer l'un des deux fichiers, généralement le plus volumineux. Cette opération consiste à stocker les données de ce fichier de façon à ce que ce soit très efficace d'y rechercher de l'information. La structure de données utilisée dans ce cadre s'appelle un index inversé[24]. La seconde étape consiste à effectuer des requêtes, c'est-à-dire rechercher dans cet index une correspondance pour chacun des individus de l'autre fichier. L'outil fournit en sortie l'ensemble des individus correspondant à la requête et les classe grâce à un score de pertinence. Les moteurs de recherche sont très flexibles dans la définition des requêtes, laissant à l'utilisateur une grande marge de manœuvre sur les filtres et les éléments qui entrent dans le score de pertinence. Cette approche peut intervenir en complément d'une première phase d'appariement exact, ce qui permet de réduire la taille des fichiers à traiter.

21 La distance de Levenshtein est une distance, au sens mathématique du terme, donnant une mesure de la différence entre deux chaînes de caractères. Elle est égale au nombre minimal de caractères qu'il faut supprimer, insérer ou remplacer pour passer d'une chaîne à l'autre.

22 La distance de Jaro-Winkler mesure la similarité entre deux chaînes de caractères. Il s'agit d'une variante proposée en 1999 par William E. Winkler, découlant de la distance de Jaro qui est principalement utilisée dans la détection de doublons.

23 Pour des précisions sur Elasticsearch, voir l'article « Le code statistique non signifiant (CSNS) : un service pour faciliter les appariements de fichiers », (Bénichou et alii, 2023).

24 Structure qui donne, pour chaque mot trouvé dans un corpus, la liste des documents où il se trouve.

Cette approche a fait ses preuves puisqu'elle est employée notamment dans le cadre du code statistique non signifiant (Bénichou et alii, 2023). Elle le sera également pour Résil. Il est important de noter qu'elle est plus adaptée à des tâches d'identification, c'est-à-dire lorsqu'on recherche des individus dans un répertoire ou dans un fichier quasi exhaustif sur la population d'intérêt. Il s'agit alors plutôt d'effectuer un grand nombre de recherches « individuelles », indépendantes les unes des autres, lorsqu'un processus d'appariement plus classique raisonnera sur deux ensembles de données pris dans leur totalité.

▶ L'approche probabiliste

L'approche probabiliste provient d'un cadre mathématique établi (Fellegi et Sunter, 1969). Le principe est d'attribuer à chaque paire une probabilité de correspondre à un seul et même individu, calculée grâce à un ensemble de paramètres. Ces paramètres ne sont pas choisis manuellement, mais estimés directement à partir des paires d'individus à apparier (Winkler, 2000). Ils représentent la capacité des différentes variables identifiantes à discriminer les paires. Une correspondance sur le nom de famille constitue ainsi un meilleur indice pour lier une paire d'individus qu'une correspondance sur le genre. Les paramètres du modèle d'appariement probabiliste captent cette information via une probabilité conditionnelle notée u. Ce paramètre représente la probabilité d'observer la même valeur sur un champ donné, sachant que les deux individus de la paire sont différents. Par exemple, la valeur de cette probabilité pour le mois de naissance serait approximativement 1/12.

Par ailleurs, la qualité des différentes variables identifiantes est prise en compte dans les paramètres via la probabilité m. Celle-ci est définie comme la probabilité d'observer la même valeur sur un champ donné au sein d'une paire d'individus identiques. Si les données étaient de parfaite qualité, cette valeur vaudrait toujours 1, mais c'est rarement le cas. On peut alors interpréter la grandeur 1-m comme le taux d'erreur sur un champ donné.

Une fois estimés les paramètres u et m pour chaque variable, il est possible d'obtenir pour chaque paire une probabilité de correspondre au même individu. La règle de décision consiste alors à comparer cette probabilité à un seuil défini par le statisticien pour choisir de lier ou non la paire. La valeur du seuil est à fixer en fonction de l'objectif poursuivi et du type d'erreur acceptable. Si le seuil est élevé, on prend peu de risques sur les paires qui sont liées, mais on risque d'en manquer. À l'inverse, si le seuil est bas, le taux d'appariement est important, mais on a un risque de lier des paires à tort (voir plus bas le paragraphe sur le statut des paires).

La méthode d'appariement probabiliste s'appuie sur les données elles-mêmes pour l'estimation des paramètres u et m. Cela permet de tenir compte du caractère informatif de chaque variable, sans nécessiter une connaissance fine des données. Cependant, cette méthode est très coûteuse en ressources de calcul, ce qui la rend difficile à mettre en œuvre sur les volumes de données habituels dans les processus d'appariement. En matière de qualité, la méthode probabiliste fait jeu égal avec des outils déterministes adaptés aux données (Haag et alii, 2022), mais ne peut rivaliser, du point de vue des ressources informatiques et du temps de traitement sur des données individuelles volumineuses.

Quelle que soit la méthode, un appariement est un processus imparfait. Il convient d'évaluer sa qualité, tant lors de l'élaboration du processus que lors de sa mise en œuvre.

▶ Qualité des appariements et enjeux statistiques

Dans un premier temps, évaluer la qualité permet d'identifier d'éventuelles pistes d'amélioration pour l'appariement. Par exemple, si une sous-population spécifique est mal appariée, il convient d'y porter une attention particulière, par exemple en améliorant le nettoyage et la normalisation des données sur cette sous-population. L'examen manuel de paires permet également de repérer des erreurs fréquentes (telles que des interversions de noms et de prénoms, ou l'utilisation d'anciens noms de communes) et d'adapter l'appariement pour les éviter.

> ***Il est important de s'assurer de la qualité des données à l'issue du processus d'appariement et d'évaluer l'impact de l'appariement sur les résultats de l'étude.***

De nombreux appariements servent à des études statistiques. Il est important de s'assurer de la qualité des données à l'issue du processus d'appariement et d'évaluer l'impact de l'appariement sur les résultats de l'étude. En effet, un défaut de qualité de l'appariement entraîne des incohérences sur les données individuelles (appariements erronés) ou un défaut de représentativité (lorsque les appariements manqués portent plus spécifiquement sur certaines populations).

La qualité des données appariées peut être évaluée de différentes manières, complémentaires. Il est parfois possible de mesurer la qualité du processus lui-même, en s'appuyant sur la proportion de la population appariée ou l'étude des paires retenues ou rejetées. Il est également souhaitable de compléter l'analyse dans une perspective plus statistique, en comparant les populations étudiées avant et après appariement. A-t-on, par exemple, la même structure par âge, la même répartition sur le territoire ?

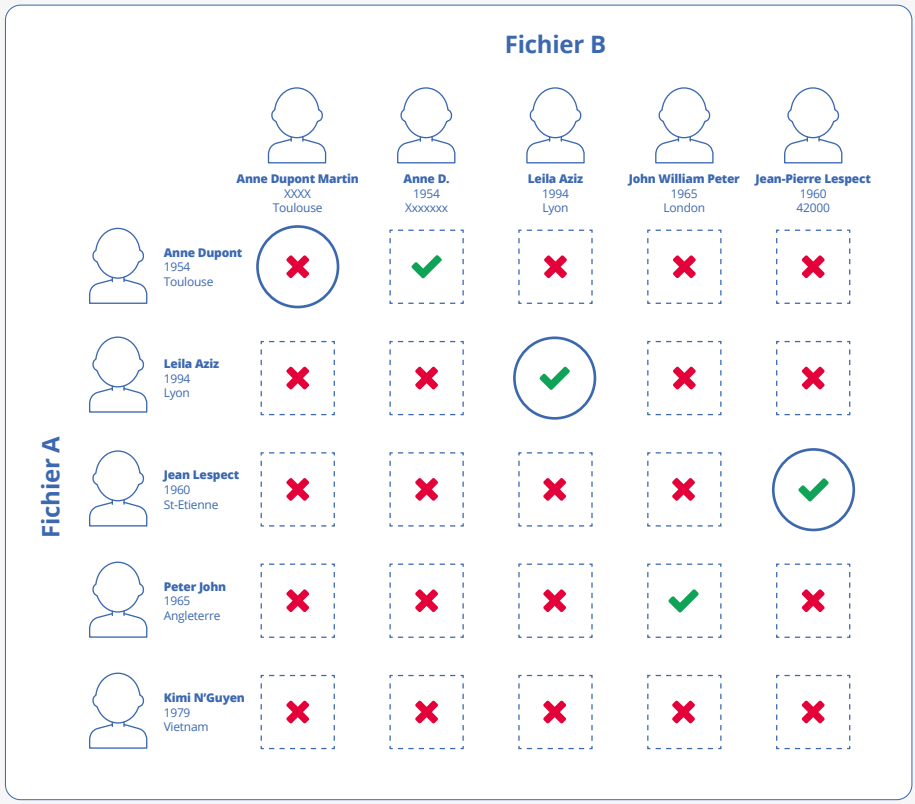
▶ Des mesures d'évaluation fondée sur le statut des paires

Lorsqu'on réalise un appariement, on peut se tromper de deux manières : lier à tort deux enregistrements, c'est-à-dire considérer deux enregistrements comme représentant la même personne à tort (les faux positifs), ou ne pas lier deux enregistrements représentant la même personne (les faux négatifs).

Lorsqu'on dispose du « vrai » résultat, on peut alors caractériser les paires en quatre groupes, en fonction de leur statut réel et du statut prédit à l'issue du processus d'appariement (***figure 3***), tel que :

- Les « bonnes décisions » :
  - les vrais positifs (VP) sont les paires d'individus identiques (concordantes) qui ont été liés par le processus ;

▶ Figure 3 - Statut des paires



Statut prédit	Statut réel		
		Individus identiques (paires concordantes) ✔	Individus différents (paires non concordantes) ✖
	Paire liée ○	Vrais positifs ✔	Faux positifs ✖
	Paire non liée ▯	Faux négatifs ✔	Vrais négatifs ✖

Lecture : 3 paires sont liées, dont 2 sont des vraies positives et 1 est une fausse positive. 22 paires sont rejetées, dont 2 sont des fausses négatives et 20 des vraies négatives.

• les vrais négatifs (VN) sont les paires d'individus différents (non concordantes) qui n'ont pas été liés par le processus ;

- Les « mauvaises décisions » :
  • les faux positifs (FP) sont les paires d'individus différents (non concordantes) qui ont été liés à tort par le processus ;
  • les faux négatifs (FN) sont les paires d'individus identiques (concordantes) qui n'ont pas été liés à tort (en quelque sorte oubliées ou non trouvées) par le processus.

Cette caractérisation est un outil intéressant, mais elle ne constitue pas une évaluation quantitative de la performance. Il est cependant possible de définir de telles mesures à partir des effectifs de chacune de ces catégories.

▶ Deux indicateurs usuels pour décrire la qualité d'un appariement

Lors d'un appariement, les effectifs des classes sont extrêmement déséquilibrés : pour deux fichiers de taille n, le nombre de paires d'individus identiques est proche de n tandis que le nombre de paires d'individus différents est approximativement de n^2. On choisit généralement des mesures comme la précision et le rappel, qui ne font pas appel au nombre de paires négatives (***figure 4***).

La précision se définit de la façon suivante :

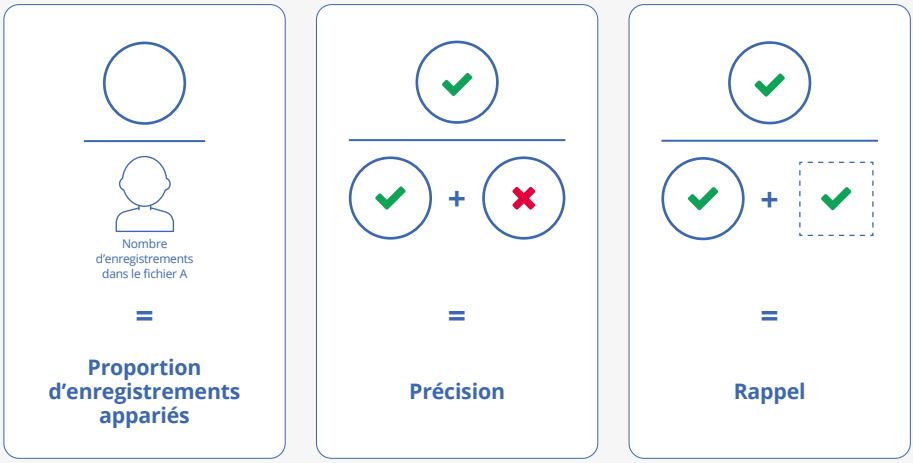
$$\text{Précision} \begin{cases} = \dfrac{\text{Vrais positifs}}{\text{Vrais positifs + Faux positifs}} \\ = \dfrac{\text{Nombre de paires liées ET concordantes}}{\text{Nombre de paires liées}} \\ = \text{Taux de réussite sur les paires liées} \end{cases}$$

Une précision élevée signifie que l'erreur sur ce modèle, lorsqu'il lie une paire, est rare. Cependant, cela ne donne pas d'information sur sa capacité à identifier un grand nombre de paires. Dans le cas extrême, un modèle liant une seule paire à juste titre aurait une précision parfaite de 1. Un tel modèle n'est pourtant pas satisfaisant. C'est pourquoi le rappel intervient souvent en complément de la précision. Le rappel, aussi appelé sensibilité, correspond à la proportion de cas positifs identifiés comme tels par le modèle.

Il se définit comme suit :

$$\text{Rappel} \begin{cases} = \dfrac{\text{Vrais positifs}}{\text{Vrais positifs + Faux négatifs}} \\ = \dfrac{\text{Nombre de paires liées ET concordantes}}{\text{Nombre de paires concordantes}} \\ = \text{Taux d'identification des paires concordantes} \end{cases}$$

▶ Figure 4 - Indicateurs usuels s'appuyant sur le statut des paires



Lecture : Dans la figure 3, la proportion d'enregistrements appariés est de 60 % (3/5), la précision de 66 % (2/3) et le rappel de 50 % (2/4).

Un rappel élevé signifie que les paires sont facilement identifiées par le modèle. Là aussi, dans le cas extrême, toutes les paires possibles seraient sélectionnées, le rappel vaudrait 1, mais le résultat comporterait beaucoup de paires appariées à tort.

▶ L'usage final des données appariées conduit à arbitrer entre précision et rappel

Pour évaluer la qualité d'un appariement, il est nécessaire de définir des objectifs et de réaliser un arbitrage entre les faux négatifs (ou paires concordantes oubliées) et les faux positifs (paires non concordantes acceptées). La notion de qualité est toujours liée à l'usage envisagé. Ainsi, lorsque les techniques d'appariement sont utilisées à des fins opérationnelles (dans le cadre d'opérations de gestion administrative par exemple), on porte une grande attention à chaque résultat individuel et on cherche le plus souvent à éviter les faux positifs (meilleure précision possible). Dans un contexte statistique, la précision est souhaitable, mais on souhaite également éviter un biais de représentativité induit par un défaut de rappel. Il est impossible d'être parfait sur les deux aspects : si on augmente le rappel et le taux d'appariement, alors la précision se dégrade et inversement. Selon l'usage envisagé des données appariées (enrichissement de données, évaluation de la couverture), on effectue un choix sur le niveau de précision attendue.

Les outils nécessaires pour évaluer la qualité d'un appariement par la qualité des paires

Si la proportion d'enregistrements appariés se calcule très facilement pour n'importe quel appariement, ce n'est pas le cas de la plupart des autres mesures de qualité. Celles-ci nécessitent en effet des informations supplémentaires sur les fichiers appariés, le plus souvent un échantillon de paires annotées. Dans le meilleur des cas, on dispose d'un *gold standard* ou fichier étalon. Il s'agit d'un échantillon de paires, représentatif des fichiers à apparier et dont le vrai statut est connu. L'obtention d'un tel échantillon est différent selon chaque situation.

Cependant, dans la majorité des cas, il n'existe pas de *gold standard* et il faut donc ajouter une étape d'annotation manuelle pour qualifier un ensemble de paires, en faisant intervenir un observateur humain : la paire proposée par le processus est-elle une paire « valide » ?, ou le processus a-t-il apparié à tort ?, ou bien encore est-il impossible de trancher ? Cette étape d'annotation manuelle prend du temps et peut donc s'avérer très coûteuse.

D'autres moyens de mesurer au moins partiellement ces grandeurs existent, en utilisant par exemple une sous-population dont on connaît a priori le statut d'appariement attendu, ou en observant la cohérence des données appariées (Doidge et alii, 2020). Lors d'un processus d'appariement, il est fréquent d'annoter manuellement un échantillon de paires, généralement choisies dans un sous ensemble de paires au statut « incertain » ; c'est le cas des paires pour lesquelles la similarité est proche du seuil de rejet, afin d'évaluer le taux de vrais et faux positifs de part et d'autre de ce seuil. Si les paires rejetées sont très majoritairement des faux négatifs, on modifie le seuil afin d'accepter ces paires rejetées à tort, au détriment d'un petit nombre de faux positifs supplémentaires.

Évaluer la qualité d'un appariement par son impact sur les données

Si les indicateurs précédents permettent d'évaluer le niveau de qualité d'un appariement, ils ne sont pas toujours simples à mesurer par l'utilisateur final, d'autant que le processus est parfois exécuté par un service tiers. C'est souvent le cas, à des fins de protection des données individuelles notamment ou en raison de la technicité de certaines opérations (préparation des données, paramétrage de l'outil d'appariement).

Lorsque le service tiers effectue l'appariement, il a connaissance de l'ensemble des variables d'appariement et il est donc en mesure d'évaluer la qualité du processus grâce aux méthodes mentionnées précédemment. Ce n'est en général pas le cas de l'utilisateur final, qui ne dispose pas des variables d'appariement (notamment les traits d'identité). Aussi, il est souhaitable que le service tiers réalisant l'appariement, produise et lui transmette des évaluations de son processus et des indicateurs de qualité associés au résultat de l'appariement. Ces mesures sont importantes, mais elles ne sont pas suffisantes pour évaluer l'impact statistique des erreurs d'appariement.

> ***Évaluer la qualité d'un appariement ne relève donc pas de la seule responsabilité de l'entité qui réalise l'appariement, mais s'appuie sur les travaux complémentaires du ou des services qui exploitent les données appariées.***

En effet, l'utilisateur dispose d'un plus grand nombre de variables, les variables d'intérêt, qu'il utilise pour produire des statistiques (par exemple, diplôme, profession, niveau de revenu, etc.). Ces informations permettent d'évaluer l'impact de l'appariement de façon statistique, par son impact sur la population d'intérêt notamment via des distributions ou statistiques des variables d'intérêt. L'utilisateur peut ainsi vérifier la représentativité de la population appariée par rapport au fichier source : par exemple, a-t-on déformé la structure par âge de la population ?

Ce deuxième niveau d'analyse, plus statistique et tourné vers l'usage est indispensable pour évaluer les éventuels biais induits par le processus d'appariement. Dès lors, si le statisticien a connaissance d'un défaut de représentativité dans la population appariée, il peut mobiliser des traitements statistiques adéquats, comme c'est le cas lors du traitement de toute source statistique.

Évaluer la qualité d'un appariement ne relève donc pas de la seule responsabilité de l'entité qui réalise l'appariement, mais s'appuie sur les travaux complémentaires du ou des services qui exploitent les données appariées.

Conclusion

Les appariements de données se développent ces dernières années au sein de la statistique publique, portés à la fois par une demande croissante de données enrichies et par la disponibilité croissante des données administratives ainsi que l'augmentation des ressources computationnelles.

Ils sont essentiels à différents processus statistiques et seront au cœur du programme Résil[25].

La qualité de ceux-ci est un enjeu important à évaluer tant lors de la réalisation de ces appariements que lors des utilisations faites ultérieurement des données appariées.

S'il existe des outils et des méthodes identifiées pour réaliser des appariements, ils doivent nécessairement être accompagnés de travaux d'analyse des données et d'expertise du statisticien pour sélectionner les paramètres les plus adéquats pour les jeux de données concernés.

25 Voir l'article d'Olivier Lefebvre sur Résil dans ce même numéro.

▶ Fondements juridiques

- Loi n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques. In : *Site de Légifrance*. [en ligne]. Mise à jour le 25 mars 2019. [Consulté le 22 février 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000888573*.

- Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés. In : *Site de Légifrance*. [en ligne]. Mise à jour le 21 février 2024. [Consulté le 22 février 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000886460*.

- Décret n° 2016-1930 du 28 décembre 2016 portant simplification des formalités préalables relatives à des traitements à finalité statistique ou de recherche. In : *Site de Légifrance*. [en ligne]. Version initiale. [Consulté le 22 février 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000033735139/*.

- Décret n° 2024-12 du 5 janvier 2024 portant création d'un traitement automatisé de données à caractère personnel dénommé « répertoire statistique des individus et des logements » (Résil). In : *Site de Légifrance*. [en ligne]. Version initiale. [Consulté le 22 février 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000048866207*.

▶ Bibliographie

- ANDRÉ, Mathias et MESLIN, Olivier, 2022. Patrimoine immobilier des ménages : enseignements d'une exploitation de sources administratives exhaustives. In : *Courrier des statistiques*. [en ligne]. 20 janvier 2022. Insee. N° N7, pp. 107-125. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/6035944?sommaire=6035950*.

- BÉNICHOU, Yves-Laurent, ESPINASSE, Lionel et GILLES, Séverine, 2023. Le code statistique non signifiant (CSNS) : un service pour faciliter les appariements de fichiers. In : *Courrier des statistiques*. [en ligne]. 30 juin 2023. Insee. N° N9, pp 64-85. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/7635825?sommaire=7635842*.

- CHEVRETTE, Antoine, 2011. G-Link: A Probabilistic Record Linkage System. In : *NORC Conference Proceedings*, [en ligne]. Mai 2011. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://www.norc.org/content/dam/norc-org/pdfs/G-Link_Probabilistic%20Record%20Linkage%20paper_PVERConf_May2011.pdf*.

- CHRISTEN, Peter, 2012. *Data Matching–Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. 5 juillet 2012. Springer. ISBN 978-3642311635.

- CIBELLA, Nicoletta, SCANNAPIECO, Monica, TOSCO, Laura, TUOTO, Tiziana et VALENTINO, Luca, 2012. Record Linkage with RELAIS: Experiences and Challenges. In : *Site de Istat*. [en ligne]. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://www.istat.it/en/methods-and-tools/methods-and-it-tools/process/processing-tools/relais*.

- CLEATON, Mary, HALL, Johanna, SHIPSEY, Rachel, WHITE, Zoe et XHAFERAJ, Kristina, 2022. A case study of using Splink: Census duplicate matching. Proceedings of Statistics Canada Symposium 2022. In : *Plateforme open source GitHub de l'Office for National Statistics*. [en ligne]. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://github.com/Data-Linkage/Splink-census-linkage/blob/main/SplinkCaseStudy.pdf*.

- COTTON, Franck et HAAG, Olivier, 2023. L'intégration des données administratives dans un processus statistique – Industrialiser une phase essentielle. In : *Courrier des statistiques*. [en ligne]. 30 juin 2023. Insee. N° N9, pp 104-125. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/7635829?sommaire=7635842*.

- DEMOTES-MAINARD, Magali, 2019. Élire, un projet ambitieux au service du Répertoire électoral unique. In : *Courrier des statistiques*. [en ligne]. 27 juin 2019. Insee. N° N2, pp. 58-71. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4168399?sommaire=4168411*.

- DOIDGE, James, CHRISTEN, Peter et HARRON, Katie, 2020. Quality assessment in data linkage. National Statistician's Quality Review. In : Site de *UK government*. Mis à jour le 16 juillet 2021. [en ligne]. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://www.gov.uk/government/publications/joined-up-data-in-government-the-future-of-data-linking-methods/quality-assessment-in-data-linkage*.

- ESPINASSE, Lionel et ROUX, Valérie, 2022. Le Répertoire national d'identification des personnes physiques (RNIPP) au cœur de la vie administrative française. In : *Courrier des statistiques*. [en ligne]. 29 novembre 2022. Insee. N° N8, pp. 72-92. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/6665188?sommaire=6665196*.

- FELLEGI, Ivan P. et SUNTER, Alan B., 1969. A theory for record linkage. In : *Journal of the American Statistical Association*. Vol. 64, No 328, pp. 1183-1210. [en ligne]. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://courses.cs.washington.edu/courses/cse590q/04au/papers/Felligi69.pdf*.

- HAAG, Olivier, KOUMARIANOS, Heidi et MALHERBE, Lucas, 2022. Probabilistes ou déterministes, des méthodes d'appariements au banc d'essai du programme Résil. In : *Site des JMS de l'Insee*. [en ligne]. JMS 2022. [Consulté le 20 février 2024]. Disponible à l'adresse : *http://jms-insee.fr/jms2022s07_3/*.

- HACHID, Ali et LECLAIR, Marie, 2022. Sirus, le répertoire d'entreprises au service du statisticien. In : *Courrier des statistiques*. [en ligne]. 29 novembre 2022. Insee. N° N8, pp. 115-130. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/6665192?sommaire=6665196*.

- HERZOG, Thomas N., SCHEUREN, Fritz J. et WINKLER, William E., 2007. Data Quality and Record Linkage. In : *Researchgate*. [en ligne]. Janvier 2007. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://www.researchgate.net/publication/220695391_Data_Quality_and_Record_Linkage*.

- JABOT, Patrick et TREYENS, Pierre-Eric, 2018. Proposition d'un nouveau processus d'appariement au Pôle Revenus Fiscaux et Sociaux (RFS). Une application à l'enquête CARE. In : *Actes des journées de méthodologie statistique 2018*. [en ligne]. [Consulté le 20 février 2024]. Disponible à l'adresse : *http://www.jms-insee.fr/2018/S20_1_PPT_TREYENS_JMS2018.pdf*.

- KOUMARIANOS, Heidi, 2022. Impact du nettoyage des données sur la qualité d'un appariement. In : *Site des JMS de l'Insee*. [en ligne]. JMS 2022. [Consulté le 20 février 2024]. Disponible à l'adresse : *http://jms-insee.fr/jms2022s07_2/*.

- MALHERBE, Lucas, 2023. Appariements de données individuelles : concepts, méthodes, conseils. In : *Documents de travail n° M2023/03*. [en ligne]. 3 juillet 2023. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/7644535*.

- MIDY, Loïc, 2021. Un outil d'appariement sur identifiants indirects : l'exemple du système d'information des jeunes. In : *Courrier des statistiques*. [en ligne]. 8 juillet 2021. Insee. N° N6, pp. 82-99. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/5398689?sommaire=5398695*.

- RANDALL, Sean M., FERRANTE, Anna M., BOYD, James H. et SEMMENS, James B., 2013. The effect of data cleaning on record linkage quality. In : *BMC Medical Informatics and Decision Making*. Vol. 13, n° 1, pp. 64. [en ligne]. 5 juin 2013. [Consulté le 20 février 2024]. Disponible à l'adresse : *DOI 10.1186/1472-6947-13-64*.

- REDOR, Patrick, 2023. Confidentialité des données statistiques : un enjeu majeur pour le service statistique public. In : *Courrier des statistiques*. [en ligne]. 30 juin 2023. Insee. N° N9, pp. 46-63. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/7635823?sommaire=7635842*.

- ROSENBAUM, Paul R. et RUBIN, Donald B., 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. In : *Biometrika*. Vol. 70, No. 1, pp. 41-55. [en ligne]. Avril 1983. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://www.jstor.org/stable/2335942*.

- Statistique Canada. 2017. Directive sur le couplage de microdonnées. In : *site de Statistique Canada*. [en ligne]. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://www.statcan.gc.ca/fr/enregistrement/politique4-1*.

- WINKLER, William E., 2000. Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage. In : *Statistical Research Report Series*. RR2000/05, US Bureau of the Census. [en ligne]. 4 octobre 2000. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://courses.cs.washington.edu/courses/cse590q/04au/papers/WinklerEM.pdf*.