JMS Journées de méthodologie statistique de l'Insee 2018

Appariement de l'enquête care par identification du plus proche echo

Patrick Jabot(*), Pierre-Eric Treyens(*)

(*)INSEE-Bretagne, Pôle Revenus Fiscaux et Sociaux

patrick.jabot@insee.fr pierre-eric.treyens@insee.fr

Mots-clés. Appariement, Indexation, Blocking, Enrichissement d'enquête.

Résumé en 350 mots

Une des missions du Pôle Revenus Fiscaux et Sociaux consiste en l'appariement de données administratives et d'enquête. Le processus utilisé actuellement se base sur un processus séquentiel de relâchement successif de clés de moins en moins robustes. Il nécessite ainsi une intervention humaine pour chaque séquence et une expertise de la source pour décider quelles clés d'appariements sont successivement les plus pertinentes. Décision qui rend par ailleurs le processus difficilement reproductible. Le processus présenté ici cherche à apparier l'enquête Capacité Aides et REssources des seniors (CARE) avec des données fiscales et s'appuie sur les quatre étapes fondamentales d'un appariement : normalisation des données identiques pour les bases d'enquête et fiscale, indexation des données afin de limiter les calculs nécessaires, choix d'une distance caractérisant l'écart entre un individu des deux bases et enfin le choix d'une règle qui permet de décider si un appariement est valide ou non.

Dans un premier temps, la normalisation des bases n'a été retouchée qu'à la marge par rapport au processus actuel. Pour l'indexation, la *clé de blocage* retenue est le code commune qui permet de réduire considérablement les comparaisons nécessaires (plus de 400 fois moins de comparaisons) et donc les temps de calculs. Non seulement grâce à l'indexation en elle-même mais aussi parce que certaines communes ne comptent aucun individu enquêté. La troisième étape a été la construction d'une distance non pondérée qui mesure l'écart entre deux individus à partir de huit variables : le nom, le prénom, le département, le jour, le mois et l'année de naissance, le ou les deux mots directeurs de l'adresse et bien entendu le code de la commune. Le cadre de travail retenu est celui d'une classification déterministe. Ainsi, une fois un seuil choisi, si le plus proche écho dans la base fiscale admet une distance inférieure à ce seuil, l'appariement est considéré comme réalisé. Une attention particulière a été portée au choix de ce seuil, afin notamment de minimiser les risques d'accepter un appariement faux et une méthode a été proposée afin de fournir un seuil d'appariementt conservateur.

Abstract en 5 à 10 lignes

For many years, the national Institutes of Statistics (Insee in France) need to link survey and administrative files to enrich data sources and improve the statistical studies. Data matching is this action consisting in identifying and merging records from distincts databases. But the matching process can be built in several ways and must make possible the link between very large files that contain millions of records. This work is a proposal to improve data matching done by the PRFS team in Rennes by simple organizational and technical tasks, in a deterministic approach.

Introduction

En 2015, le niveau de vie médian de la population s'élève à 20 300 euros annuels et le taux de pauvreté s'établit à 14.2 %. Ces chiffres proviennent de l'Enquête Revenus Fiscaux et Sociaux (ERFS) mais le terme d'enquête pourrait sembler légèrement abusif. En effet, si une part de l'information nécessaire à leur calcul provient effectivement de l'Enquête Emploi en Continu (EEC), ces calculs ne sont rendus possibles qu'en enrichissant les données de l'enquête à l'aide de sources fiscales et sociales. A l'INSEE, le Pôle Revenu Fiscaux et Sociaux (PRFS) de Rennes est le référent national en matière d'enrichissement de données. Et les appariements préalables à ces enrichissements ne concernent pas uniquement des enquêtes comme Budget de Famille (BDF), Histoire de Vie et Patrimoine (HVP) ou encore Prestation de compensation du Handicap : Exécution dans la Durée et Reste à charge (PHEDRE), ils concernent aussi d'autres fichiers administratifs comme l'Echantillon Interrégime de Retraités (EIR) ou encore celui des professionnels de santé (CNAMTS). Ainsi, la qualité de l'enrichissement dépend en grande partie de la qualité de l'appariement préalable et cette qualité se mesure à l'aune de deux indicateurs principaux, le taux d'appariement et la rapidité du processus d'appariement. Le but final de ces enrichissements est de mettre en relation bijective et avec la plus forte probabilité l'identifiant de l'enquête et celui de la base d'enrichissement.

Actuellement, les appariements au PRFS se basent sur une approche intuitive. On construit tout d'abord une clé qui agrège l'ensemble de l'information disponible sur chacun des individus. Une première séquence conduit alors à considérer comme appariés les individus pour lesquels la clé complète concorde entre le fichier d'enquête et le fichier administratif. Ensuite, des clés plus partielles sont constituées de manière à éluder les difficultés d'appariement liées à des erreurs de saisie ou de codage. Chaque tour de clé permet d'extraire des individus rapprochés, un reste à traiter étant alors constitué pour la vague d'appariement suivante. Au final, ce procédé conduit à des résultats d'appariement sérieux, au prix toutefois de plusieurs contraintes fortes (faible reproductibilité, nécessité d'une bonne connaissance du fichier d'enquête, durée de l'opération d'appariement relativement élevée).

A la différence de ce qui se fait actuellement, le processus développé ici utilise les quatre étapes habituelles d'un appariement : préparation et indexation des données, choix d'une distance entre individus et enfin prise de décision. Pour mettre au point cette nouvelle méthode, l'enquête Capacités, Aides et REssources des seniors (CARE) sera appariée avec un fichier permanent de la Direction Générale des Finances Publiques, le fichier FIP. Cette enquête individus présente le double avantage de renseigner le patronyme et d'être de taille raisonnable (15 000 individus). La première section de ce document présente le cadre actuel des appariements réalisés au PRFS. Ensuite, les quatre sections suivantes reprennent les quatres étapes ci-dessus. Ainsi, la deuxième section présente les données utilisées (enquête CARE et fichier fiscal FIP), les traitements de normalisation appliqués à ces données ainsi qu'un appariement exact permettant de limiter les futurs temps de calculs. La troisième décrit la clé de blocage utilisée (le code de la commune) qui permet de partitionner les bases de données en sous-blocs indépendants à comparer. Ensuite, la quatrième partie définira la distance retenue entre deux individus et la cinquième fournira un cadre de travail permettant d'objectiver le seuil de significativité retenu ainsi que les résultats obtenus. Enfin, une dernière partie sera consacrée à quelques remarques terminales.

1 Un mode intuitif d'appariement : les poupées russes

Une première approche relativement naturelle de l'appariement consiste à se représenter l'opération comme une succession de rapprochements des observations grâce à des clés de moins en moins complètes. Ainsi, dans un premier temps, l'appariement repose sur l'identité entre les clés exhaustives agrégeant l'ensemble de l'information disponible dans les fichiers. Ce rapprochement est jugé le plus fort car il repose sur l'association stricte des données présentes et suggère un faible risque de doublons ou d'erreurs. Dans une deuxième séquence, des clés moins complètes sont constituées de manière à apparier des candidats pour lesquels des écarts plus ou moins importants sont attendus. Par exemple, le fait de constituer un clé sans prendre en compte le sexe de l'individu peut permettre de s'extraire d'éventuelles erreurs de saisie ou de codage de cette information. Les variables conservées sont alors jugées suffisamment discriminantes pour interdire de mauvais appariements : en l'occurrence, le prénom présente généralement un caractère sexué qui limite le risque d'erreur ou d'échos multiples.

Ce mode d'appariements successifs est aujourd'hui utilisé au PRFS pour l'enrichissement des enquêtes dites ponctuelles c'est-à-dire celles qui ne sont pas reconduites chaque année. Il produit des résultats consistants avec des taux d'appariement généralement supérieurs à 90 %, variables selon la qualité du fichier d'enquête à enrichir. Au-delà de sa relative rusticité, cette approche présente l'intérêt d'une certaine flexibilité. Il est notamment possible de multiplier les clés selon le taux d'appariement cible. Des erreurs de saisie des prénoms peuvent ainsi conduire à retenir les premiers caractères du prénom pour faciliter le rapprochement. Mais les inconvénients de cette méthode sont également nombreux. Les résultats de cette approche sont sensibles aux choix et à la séquence de passage des clés, la durée et la complexité de l'opération est liée au nombre de clés constituées, enfin elle implique une connaissance approfondie du fichier d'enquête à apparier, complétée idéalement par une expérience des opérations d'appariement de l'enquête précédentes. La fragilité de la reproductibilité, la caractère laborieux et aléatoire de la génération des clés et la durée de l'opération qui peut atteindre aujourd'hui plusieurs semaines plaident pour l'examen de solutions alternatives.

L'image des *poupées russes* illustre bien le procédé utilisé pour réaliser les appariements au PRFS aujourd'hui. Une succession de clés de moins en moins exhaustives sont produites pour rapprocher des observations. Ces clés doivent s'imbriquer pour constituer un tamis de plus en plus large de manière à retenir des appariements plus approximatifs. Aujourd'hui, le rapprochement sur la clé complète extrait entre 45 et 55 % en moyenne des individus. Le rendement des tours suivants est plus variable, de quelques unités jusqu'à des ensembles substantiels. Sur certaines enquêtes, le nombre de clés créées peut s'avérer très conséquent, de l'ordre de la centaine pour l'enquête Budget de Famille 2016. Toutefois, il est plus commun de créer quelques dizaines de clés pour obtenir un taux d'appariement sérieux. La combinaison d'une séquence de travail plus normalisée et d'une méthode plus générique d'appariement des individus non retrouvés par l'utilisation de la clé complète constitue la piste de travail de ce document.

2 Préparation des données et appariement exact

2.1 Données fiscales d'enrichissement et données d'enquête

Le fichier fiscal qui sert à l'enrichissement de l'enquête est le fichier FIP fourni par la DGFIP. Ce fichier recense des foyers fiscaux formés par une personne de référence et éventuellement son conjoint et/ou une ou plusieurs personnes à charge (jusqu'à 4). Dans le cadre d'une enquête individus à apparier, n'importe quelle personne du foyer peut être interrogée. Il est donc nécessaire d'obtenir une base individuelle à partir de la base FIP. La figure 1 détaille le processus

d'individualisation des données. Ainsi, chacun des foyers fiscaux (l'identifiant fiscal de chacun de ces foyers est le DIRINDIK) renverra dans la base individualisée un nombre de lignes égal au produit entre le nombre de personnes dans le foyer et le nombre d'adresses connues pour ce foyer. Cela conduit donc à une nouvelle base de données de niveau individuel contenant plus de 97 millions de lignes. Le fichier d'enquête, **CARE**, est quant à lui directement un fichier

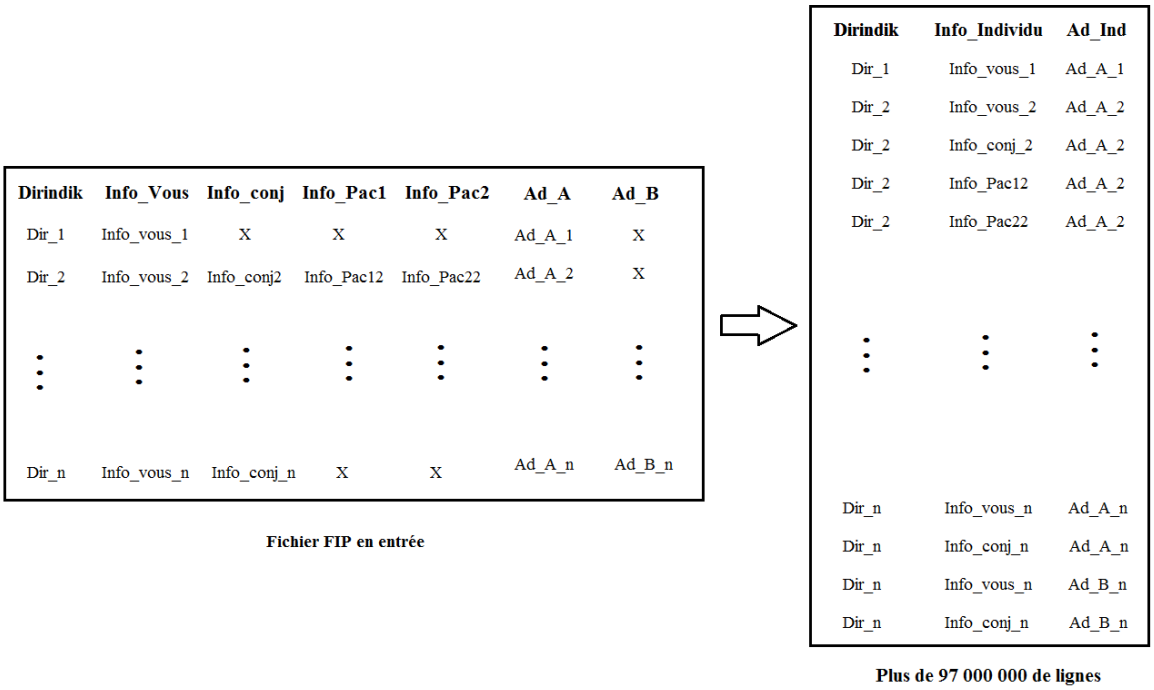

FIGURE 1 – Processus d'individualisation des données

individuel. Toutefois, pour quelques unes des personnes enquêtées (autour de 350), deux adresses sont disponibles. Ainsi, il sera possible d'indentifier deux échos pour un même individu. Quand le cas se produira, la première adresse de l'enquête sera retenue. Avant d'être soumis à la même transformation que le fichier FIP, le fichier d'enquête est préalablement nettoyé des observations pour lesquelles on peut anticiper qu'aucun appariement ne sera possible. Ainsi, dès lors que les seules informations disponibles seront le prénom, l'identifiant de la ville et les mots directeurs de l'adresse, les individus sont exclus du champ à apparier. Cela concerne 163 individus. Finalement, entre cette suppression et la déduplication de certains individus ayant deux adresses, le fichier à apparier compte 15 180 observations.

2.2 Normalisation des données et création des sous-clés

Les informations utilisées pour l'appariement et donc présentes conjointement dans les deux fichiers sont le nom, le prénom, le département de naissance, le jour, le mois et l'année de naissance, l'identifiant de la commune (ou de l'arrondissement, variable CODEGEO), le premier mot directeur de l'adresse et éventuellement le second s'il y a lieu ainsi que le sexe. Dans la base d'enquête, un seul nom est donné alors que le fichier d'enrichissement contient aussi le nom de naissance. Cette information sera aussi utilisée dans le calcul de la distance entre le nom de l'enquête et celui présent dans le fichier FIP (cf. infra). Par ailleurs, le programme de normalisation utilisé afin de normaliser les différentes variables est identique pour les deux bases de données. Plus que la qualité de cette normalisation [1], il est surtout indispensable que le processus

1. Ce processus de normalisation est commun à l'ensemble des appariement réalisés au PRFS. Ainsi, il n'impactera pas les différences observées entre le processus d'appariement utilisé actuellement et celui

de normalisation des informations contenues dans les deux bases de données soit identique afin de s'assurer de la comparabilité de l'information disponible, pour au pire travailler à erreurs constantes. Que ce soit pour les noms, les prénoms et les adresses, les caractères spéciaux, les accents, les chiffres... sont exclus. L'adresse subit les mêmes transformations et ne sont finalement retenus que le ou les deux mots directeurs de l'adresse. Ces mots directeurs sont le ou les deux plus longs mots de l'adresse et ne doivent pas faire moins de quatre caractères.

Finalement, chaque information disponible conjointement dans les deux bases peut se voir comme une sous-clé. A l'aide de ces sous-clés, une clé totale est créée en les concaténant à la fois dans le fichier FIP et dans le fichier CARE. Il faut noter que si cette clé totale concaténée avec l'identifiant fiscal (DIRINDIK) fournit un identifiant unique, ce n'est pas le cas de la clé totale prise isolément. Cela arrive après un déménagement en dehors du département d'origine qui implique un changement de DIRINDIK. Comme l'appariement se fera au niveau de l'identifiant de la commune ou de l'arrondissement afin d'éviter des calculs trop lourds, conserver comme identifiant unique le couple concaténé DIRINDIK et cle_totale permettra de retrouver l'individu dans le département qui correspondra à celui dans lequel il a été enquêté. A ce stade, le fichier FIP individualisé et normalisé ne compte désormais plus que 84.5 millions de lignes.

2.3 Construction d'une clé totale et appariement exact

La concaténation de toutes les sous-clés utilisées pour l'appariement permet de créer une clé totale. Quand la valeur prise par cette clé est exactement la même dans le fichier d'enquête et dans la base fiscale d'enrichissement, on a un appariement que l'on qualifiera d'exact. Cela concerne 56,0 % des individus enquêtés. Il est à noter que certains individus peuvent être retrouvés plusieurs fois dans FIP puisque certains d'entre eux ont deux adresses dans CARE (changement d'adresse ou décès du conjoint ce qui modifie le DIRINDIK de la femme quand elle survit à son conjoint). Pour le premier cas, la première adresse donnée dans le fichier d'enquête est l'adresse privilégiée pour l'appariement. Dans le second, on conserve le DIRINDIK de la personne survivante.

On obtient alors deux premiers reliquats d'individus pour lesquels l'appariement exact n'a pas fonctionné, un pour l'enquête mais aussi pour le fichier FIP dans lequel les individus appariés sont aussi retirés. Le principal intérêt est de limiter les temps de calculs auxquels nous devrons faire face dans l'étape suivante de l'appariement, celle par recherche du plus proche écho. Plus précisément, cet appariement exact est un cas particulier de la méthode générale qui consiste à apparier les individus de l'enquête ayant une distance nulle avec les individus de la base fiscale. On peut donc, sans perte de généralité, considérer que cette approche en deux temps est parfaitement équivalente à l'approche en une étape. La distance utilisée sera décrite dans la partie suivante mais on comprend aisément qu'une distance nulle est équivalente à l'égalité stricte pour toutes les sous clés entre deux individus que l'on cherche à comparer.

3 Indexation et clé de blocage

Initialement, la base de données individualisée de FIP comptait 84.5 millions d'observations et celle de CARE plus de 15 000. La comparaison de chaque observation de la base d'enquête à chaque observation de la base FIP aurait donc conduit à calculer un peu moins de 1 300 milliards de fois la distance entre deux observations puis à trier plus de 15 000 fois une base comptant 84.5 millions d'observations afin de récupérer l'écho le plus proche au sens de cette distance. Quand bien même, un de ces calculs de distance ne prendrait qu'un cent millième de seconde, il faudrait

présenté dans ce papier.

près de 150 jours de traitement sans compter les phases de triages. Ce problème calculatoire porte le nom de **complexité quadratique**. Supposons maintenant qu'une variable permette de séparer les deux bases de données en deux parties de même taille (par exemple le sexe). Le nombre de comparaisons ne serait alors plus que de deux fois 7500 (la moitié de 15 000) multiplié par 42.25 millions. C'est-à-dire que le nombre de calculs à réaliser aurait été divisé par deux, tout comme les temps de traitements nécessaires à l'appariement des données. Même si l'étape d'appariement exact permet pour un temps de calcul très court (autour de 30 minutes) de limiter les comparaisons futures de 56,0% et donc de devoir réaliser autour de 550 milliards de comparaisons. Cela reste insuffisant pour espérer comparer les deux reliquats entre eux naïvement.

Afin de contourner ce problème de complexité quadratique, la méthode privilégiée dont l'intuition a été donnée ci-dessus s'appelle l'**indexation** (*indexing* en anglais). Elle consiste à utiliser une variable appelée **clé de blocage** (*blocking key, sorting key etc.*) qui sert à partitionner les deux bases de données en sous-bases de données, ou sous-blocs, définies selon les modalités de cette clé de blocage. Il reste alors à chercher le plus proche écho d'un individu apaprtenant à un sous-bloc dans le sous-bloc correspondant de la base fiscale. Afin d'être efficace, cette clé de blocage doit présenter deux qualités essentielles, un fort pouvoir discriminant afin de limiter le plus possible la complexité quadratique et être renseignée le mieux possible pour que le véritable écho se trouve dans le bon sous-bloc.

La clé de blocage choisie est l'identifiant de la commune car il présente un double avantage. Non seulement il est bien renseigné, que ce soit dans le fichier d'enquête ou dans le fichier FIP, mais en plus, après l'appariement exact, il ne compte plus que 1 895 modalités sur plus de 35 000 modalités présentes dans le fichier FIP. Cela permet donc de réduire considérablement les comparaisons nécessaires afin de réaliser l'appariement. Finalement, dans de rares cas (comparativement aux 84.5 millions d'observations que compte le fichier FIP individualisé), l'identifiant de la commune pour une observation du fichier FIP est la concaténation entre le numéro du département et la chaîne de caractères '000' et plus rarement encore (moins de 1 000 cas) il est de la forme '000' ou '00000'. Dans le premier cas, les observations dont l'identifiant de la commune commence par le numéro du département sont incluses dans le sous-bloc défini par l'identifiant de la commune et dans le second, ils sont inclus dans tous les sous-blocs issus de FIP. Par exemple, quand l'identifiant de la commune dans la base CARE est '01034', le sous-bloc dans CARE est constitué des individus de CARE dont l'identifiant commune est "01034". Par contre, le sous-bloc provenant de FIP dans lequel on recherchera les échos les plus proches sera constitué non seulement des individus dont l'identifiant de commune est "01034" mais aussi de ceux dont cet identifiant vaut "01000", "000" et "00000". Si cela rallonge inévitablement les temps de calcul, cela n'est pas rédhibitoire comme pouvait l'être la comparaison systématique des deux bases prises intégralement. En procédant de la sorte, le nombre final de comparaisons dépasse très légèrement 1.3 milliard, soit 1 000 fois moins que la comparaison de la base intégrale avec le fichier FIP et 440 fois moins que la comparaison des reliquats obtenus après l'appariement exact.

A ce stade, le reliquat issu de FIP et obtenu après l'appariement exact est départementalisé. En effet, il est plus rapide d'un point de vue calculatoire de lire une centaine de fois la table FIP pour isoler des tables départementales et ensuite de lire chacune de ces tables autant de fois qu'une commune du département compte des individus enquêtés que de lire près de 2 000 fois la table FIP pour obtenir le même résultat. Maintenant que l'étape de création des sous-blocs à partir de l'identifiant communal est terminé, il faut comparer chaque individu d'un sous-bloc de l'enquête avec chaque individu du sous-bloc correspondant dans le fichier FIP.

4 Distance retenue

4.1 Présentation générale

En mathématiques, une distance d doit vérifier trois propriétés.

$$\begin{array}{lll} \forall\,(x,\,y,\,z) & d(x,y) = d(y,x) & \text{Propriété de symétrie} \\ & d(x,y) = 0 \iff x = y & \text{Propriété de séparation} \\ & d(x,z) \leqslant d(x,y) + d(y,z) & \text{Inégalité triangualaire} \end{array}$$

Dans le cadre de ce papier, nous allons définir la distance d entre un élément de CARE et un élément de FIP comme une somme pondérée de distances d_i (à définir) entre les différentes sous clés i présentes dans les deux bases. Pour rappel, ces sous-clés sont le nom, le prénom, l'identifiant de la commune, le sexe, le département de naissance, le jour, le mois et l'année de naissance ainsi que le ou les deux premiers mots directeurs de l'adresse. On a alors

$$d\left(\text{Ind}_{CARE}, \text{Ind}_{FIP}\right) = \sum_{i=1}^{9} \alpha_i d_i \left(\text{Ind}_{CARE}, \text{Ind}_{FIP}\right) \tag{1}$$

$$\begin{array}{lll} \text{avec} & i & \in (\text{Nom, Prénom, Sexe, Identifiant commune,} \\ & & \quad \text{Département, Jour Mois et Année de naissance et Adresse}) \end{array}$$

Toutefois, ce que nous appellerons distance par la suite ne vérifiera que partiellement la propriété de symétrie (cf. partie 4.2.1) et une partie de la propriété de séparation, $d(x,x) = 0$. Ainsi, l'appellation distance pourrait paraître abusive, mais comme l'intuition portée par le terme semble suffisamment pertinente, nous le conserverons. Cependant, cette distance doit posséder un pouvoir discriminant suffisant pour se rapprocher le plus possible de la propriété de séparation (dans la partique une distance nulle conduit à identifier certainement le bon individu). L'idée sera donc d'identifier un **seuil** S tel que pour un individu a de la base d'enquête et un individu b de la base d'enrichissement, on ait

$$\begin{array}{lll} d(a,b) & > & S \Rightarrow a \neq b \\ d(a,b) & < & S \Rightarrow a = b \end{array} \tag{2}$$

En effet, un même individu dans les deux sources peut avoir des informations différentes pour plusieurs raisons (mauvaise saisie de la part de l'enquêteur, erreur dans les fichiers d'enrichissement, orthographes différentes sur des patronymes ou des prénoms, des libellés d'adresse différents avec des abréviations dans l'une et pas dans l'autre etc.). Ainsi, la distance d'un individu à lui-même peut ne pas être nulle et l'enjeu est donc de l'identifier avec suffisamment de certitudes en dépit de l'écart d'information observé dans les deux sources. Bien entendu, on ne pourra jamais être totalement certain de la justesse d'un appariement et on se retrouve finalement dans une situation relativement semblable à celle du statisticien face à un test. En rejetant les appariements au-delà d'une certaine valeur de S, nous limiterons le risque d'accepter à tort un appariement, i.e. d'associer à un individu de l'enquête un écho qui ne serait pas lui-même dans la base d'enrichissement.

Pour l'instant, aucune hypothèse n'est faite sur les pondérations à apporter, elles seront donc égales à 1 et la formule est alors simplement celle de l'équation (1) non pondérée

$$d\left(\text{Ind}_{CARE}, \text{Ind}_{FIP}\right) = \sum_{i=1}^{9} d_i \left(\text{Ind}_{CARE}, \text{Ind}_{FIP}\right) \tag{3}$$

Dans la suite de cette partie, nous présenterons les différentes distances retenues dans les formules 2 et 3. Toutefois, le choix de ces distances est nécessairement subjectif. En fait, il existe énormément de distances pour mesurer l'écart entre deux chaînes de caractères. Par exemple,

la distance de Levenshtein compte le nombre minimal de caractères qu'il faut supprimer, insérer ou remplacer pour passer d'une chaîne de caractères à une autre. Ainsi, passer du prénom "KARYNE" au prénom "KRYSTEL" demande 4 opérations (suppression du A dans KARYNE, suppression du S ou du T dans KRYSTEL, remplacement du T ou du S par un N dans KRYSTEL et suppression du L dans KRYSTEL) alors que passer de "KRYSTEL" à "CHRISTELLE" en demande 5. Pourtant les deux prénoms "KRYSTEL" et "CHRISTELLE" semblent plus proches et ils le seraient si la distance utilisée était phonétique. Le choix de la distance ne se limite pas à ces deux exemples, il existe en effet des dizaines de distances entre deux chaînes de caractères comme celle de Jaro-Winckler utilisée par Statistiques Canada. Ce choix pléthorique montre bien à quel point le choix de la distance ne peut être que subjectif. Mais cette subjectivité n'est pas problématique si la distance choisie remplit correctement son rôle : discriminer parmi plusieurs échos celui qui est le plus vraisemblablement le bon. Le choix retenu ici est de se servir de la distance de Levenshtein qui, selon les sous clés, sera bornée afin de limiter les écarts possibles dans les différents calculs (cf. ci-dessous). Par ailleurs, chacune des distances entre les sous-clés sera comprise entre 0 et 1.

4.2 Les distances entre les sous clés

4.2.1 Distance entre les prénoms et entre les noms

Dans le cas des prénoms (respectivement des noms), la distance de Levenshtein est utilisée. Comme la distance entre deux sous clés doit être comprise en 0 et 1, il faudra donc la normaliser pour qu'elle reste bornée quand on comparera un individu CARE à chaque individu FIP du même sous-bloc. Le choix de diviser la distance obtenue par le maximum observé des distances entre les prénoms (respectivement les noms) aurait souvent conduit à obtenir des distances trop faibles pour certaines comparaisons alors que les prénoms (respectivement les noms) auraient été fortement différents. Aussi, le choix a été fait de borner la distance de Levenshtein par la valeur 5 puis de la diviser par 5. La distance finale est donc bien toujours comprise entre 0 et 1. Plus précisément, dès lors que les prénoms ou les noms sont identiques, la distance est nulle. Le cas des prénoms et des noms composés ou des deuxième voire des troisièmes prénoms souvent renseignés dans la base FIP nous a conduit à tester si le prénom (respectivement le nom) dans une des deux bases était inclus dans le prénom ou le nom de l'autre base. Quand cela se produisait, la distance entre les deux prénoms a été fixée à 0,1 de manière arbitraire. Par ailleurs, dans le cas du nom, la base FIP renseigne aussi le nom de naissance. Quand le nom de l'enquête était le nom de naissance, la distance renvoyée a été fixée à 0. Sinon, la distance est simplement la plus faible entre deux distances de Levenshtein bornée par 5 et divisée par 5, celle au nom de famille et celle au nom de naissance [2] [3].

Par exemple, la distance de Levenshtein entre les prénoms "MARIE" et "MARIA" est de 1, il suffit en effet de changer la lettre E en la lettre A. Ainsi, la distance normalisée entre ces deux prénoms est de 0,2. Par contre, la distance retenue entre "MARIE" et "MARIE THERESE" ne sera que de 0,1. Le choix a donc été fait de privilégier une graphie exacte en terme de proximité entre deux chaînes de caractères plutôt qu'une graphie ne différant que d'une lettre. D'autre part, si une femme a pour nom "DUPOND" dans l'enquête et "MARTIN" comme nom dans FIP et "DUPOND" comme nom de naissance dans FIP, la distance est nulle. Si par contre, son de naissance dans FIP avait été "DUPONT", la distance aurait été de 0,2.

2. Il est à noter qu'en procédant ainsi, on trouve 17,5% des individus de l'enquête dont la distance à un individu FIP est nulle. Ces 17,5 % s'ajoutent aux 56 % obtenus lors de l'appariement exact.

3. C'est l'utilisation du nom de naissance présent dans FIP et absent de CARE qui empêche d'avoir la propriété de symétrie des distances.

4.2.2 Distance entre les jours de naissance, le mois de naissance et le département de naissance

Que ce soit le jour, le mois ou le département de naissance, ces variables sont codées sous la forme de chaînes de deux caractères. Par ailleurs, dans FIP comme dans CARE, ces variables ne sont parfois pas renseignées. Dans FIP, elles peuvent même être imputées au 1er janvier ou au 31 décembre, essentiellement pour des personnes nées à l'étranger. Ainsi, dès lors que l'information est absente, la distance a été fixée à 1. Cela implique donc que même si l'information est absente pour l'individu de l'enquête et pour l'individu FIP et que l'on a bien une égalité entre deux chaînes de caractères vides, la distance est fixée arbitrairement à 1. L'autre question qui se posait était de choisir si, par exemple, un mois de naissance valant "12" dans CARE était plus proche ou moins proche d'un mois de naissance valant "10" dans FIP que d'un mois de naissance valant "09". En d'autres termes, est-ce-qu'une différence d'un chiffre entre les deux chaînes doit être autant pénalisée qu'une différence de deux ? Le choix réalisé a été de considérer que deux nombres ayant un chiffre commun au même emplacement étaient plus proches que deux nombres sans chiffre en commun. Ainsi, la distance entre "12" et "11" est de 0,5 après normalisation alors que celle entre "12" et 09" est de 1 et celle entre "12" et "21" est aussi de 1.

4.2.3 Distance entre les années de naissance

Contrairement aux cas prédécents, les années de naissance sont codées sur quatre caractères. Là encore, si une information est manquante dans CARE ou dans FIP la distance est égale à 1 et si les deux informations sont identiques dans les deux bases, elle est nulle. De plus, comme les deux premiers chiffres de l'année de naissance sont moins significatifs que les deux derniers mais qu'on a tout de même voulu conserver une part de leur pertinence dans le calcul de la distance, le choix a été fait d'utiliser la distance de Levenshtein entre les deux chaînes mais bornée à deux puis divisée par deux.

4.2.4 Distance entre les libellés d'adresse

Comme nous l'avons déjà évoqué dans la partie concernant la normalisation, un ou deux mots directeurs sont récupérés à partir du libellé de l'adresse, le premier des deux étant le plus long des deux. Toutefois, le processus de normalisation peut conduire à deux mots de même taille et dans ce cas, ils peuvent s'inverser d'une base de données à l'autre. De même, si un des mots est mal orthographié (lettre manquante), la position des deux peut s'inverser. Il a donc fallu prendre en considération ces différents cas dans le calcul de la distance entre les libellés des adresses.

Dès lors qu'aucun mot n'est récupéré dans une des deux bases, soit parce que l'adresse n'est pas renseignée soit parce que le processus de normalisation y conduit, la distance est égale à 1. Si les observations de chaque base ne comporte qu'un seul mot directeur pour l'adresse, la distance entre les deux informations est la distance de Levenshtein bornée à 6 puis divisée par 6. Quand un libellé ne comporte qu'un mot directeur et que l'autre en comporte deux, la distance retenue est le minimum entre deux distances, la distance de Levenshtein bornée à 6 puis divisée par 6 au premier mot du libellé et la même distance mais au second mot du libellé. Enfin, quand les deux adresses comportent deux mots, la distance est le minimum entre deux sommes. La première est celle entre la distance de Levenshtein bornée par 3 du premier mot directeur de l'adresse dans CARE au premier mot directeur de l'adresse dans FIP et la même distance entre le second mot directeur de l'adresse dans CARE et le second mot directeur dans FIP. La seconde est l'inverse de la première, i.e. on compare le premier mot directeur de CARE au second de FIP et le seconde CARE au premier de FIP.

L'exemple suivant a été observé dans le processus d'appariement. Les deux mots sont "HENRI"

et "DUNAN" dans CARE et "DUNANT" et "HENRI" dans FIP, la distance de Levenshtein entre "HENRI" et "DUNANT" est de 5 et celle entre "DUNAN" et "HENRI" est de 4. En bornant les deux distances à 3, on retrouve finalement une distance de 1 entre les deux sous clés. Si on inverse les observations, on se retrouve à comparer "HENRI" avec "HENRI" et "DUNANT" avec "DUNAN" et la distance n'est plus que de 0.16667 à cause de la mauvaise orthographe du nom du fondateur de la Croix Rouge.

4.2.5 Distance entre les identifiants de communes

Il peut sembler curieux de calculer une distance suivant l'identifiant de la commune qui sert de variables d'indexation. Toutefois, on a inclus dans le sous-bloc FIP associé à un identifiant commune l'ensemble des adresses de FIP s'écrivant comme la concaténation du numéro de département et de la chaîne de caractère '000'(cela arrive lors d'un déménagement dans un autre département)et celles dont l'identifiant commune est '000' ou '00000'. Pour cette sous clé, la distance sera nulle si l'identifiant de la commune est le même, elle sera de 0,5 si l'identifiant de la commune dans FIP s'écrit comme la concaténation ci-dessus et elle sera de 1 si l'identifiant de la commune dans FIP est '000' ou '00000'. Cela permet de ne pas écarter définitivement les individus de FIP dont l'identifiant de commune n'est pas exactement celui de CARE tout en les pénalisant, surtout si l'information sur le département ne coïncide pas.

4.2.6 Distance entre les sexes

Finalement, pour le sexe, la distance est de 1 dès lors que cette information n'est pas présente dans une des deux observations à comparer ou qu'elle diffère. Elle est de 0 si l'information est identique.

4.3 Schéma général du calcul des distances

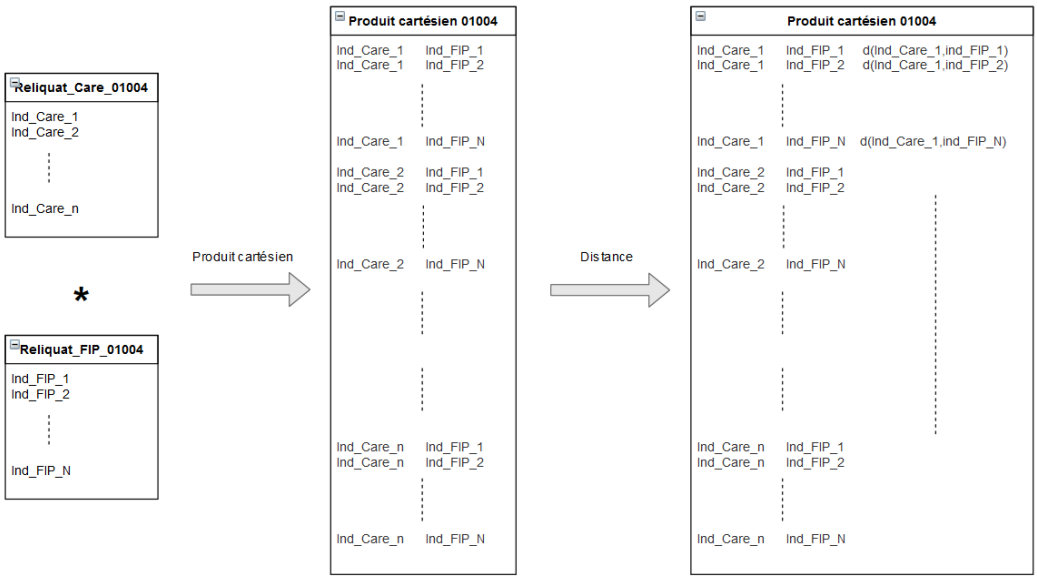

FIGURE 2 – Schéma du produit cartésien et du calcul des distances pour la commune 01004

Pour chaque sous-bloc, le produit cartésien est réalisé entre les individus issus des bases d'enquête et fiscale. Ensuite, la distance entre chaque individu est calculée pour chaque ligne de ce produit cartésien tel que schématisé dans la figure 2 pour la commune 01004. Par la suite, les

distances sont ordonnées et la question du seuil en-deçà duquel un couple est considéré comme réellement apparié sera discuté dans la partie suivante.

5 Seuil retenu pour l'appariement

5.1 Un cadre de travail similaire à celui de la théorie des tests

Après que les distances d'un individu de la base d'enquête à ceux de la base fiscale ont été calculées, il faut fixer un seuil en-dessous duquel l'appariement est validé et au-dessus duquel il ne l'est pas. Dans ce cadre, deux erreurs peuvent se produire ; accepter un appariement à tort ou en rejeter un à tort. De notre point de vue, l'erreur la plus pénalisante est de valider à tort un appariement. Ainsi, le risque associé à cette erreur sera le risque de première espèce et le seuil sera choisi afin de le contrôler [4].

Soit A la base d'enquête, B la base fiscale et K le nombre de modalités de la clé de blocage, ici le nombre de codes commune/arrondissement présents dans la base d'enquête. On appelle A_k et B_k les sous-blocs des bases A et B dont chaque individu possède la modalité k de la clé de blocage avec $k \in [\![1, K]\!]$. Soit a_k^i un individu i choisi au hasard dans A_k, on a donc deux hypothèses concurrentes

$$H_0 : \text{Rejeter l'appariement de } a_k^i \qquad \text{vs.} \qquad H_1 : \text{Accepter l'apparriement de } a_k^i \tag{4}$$

qui seront notées par la suite

$$H_0 : \text{Accepter } a_k^i \notin B_k \qquad \text{vs.} \qquad H_1 : \text{Accepter } a_k^i \in B_k \tag{5}$$

Comme on cherche à minimiser les chances d'accepter $a_k^i \in B_k \mid a_k^i \notin B_k$. On définit alors le risque de première espèce associé au sous-bloc k comme

$$\alpha(k) = \Pr\left(\text{Accepter } a_k^i \in B_k \mid a_k^i \notin B_k\right) \tag{6}$$

Or, dans notre méthode, accepter qu'un individu de la base d'enquête ait au moins un écho dans la base fiscale revient à trouver un ou des individus dans la base fiscale pour lesquels la distance à l'individu de l'enquête soit inférieure à un seuil S. Il vient alors, avec $b_{j,k}$ un individu j du sous-bloc B_k,

$$\alpha(k) = \Pr\left(\underset{j \in [\![1, \text{Card}(B_k)]\!]}{min} d(a_k^i, b_{j,k}) < S \mid a_k^i \notin B_k\right) \tag{7}$$

Comme $\alpha(k)$ dépend du sous-bloc B_k choisi initialement, il faudrait idéalement donner la distribution de $min\ d(a_k^i, b_{j,k})$ sachant que $a_k^i \notin B_k$ pour chaque k ou à défaut, calculer le niveau du test comme le sup $\alpha(k)$ sur k pour chaque seuil S. Ce qui pratiquement s'avère trop lourd en temps de calculs. Toutefois, on peut supposer que pour un S donné, $\alpha(k)$ croit avec le cardinal du sous-bloc B_k. En effet, plus la taille d'un sous-bloc augmente, plus il devient probable d'y trouver des individus possédant des caractéristiques proches de a_k^i. Sous cette hypothèse, le choix le plus raisonnable est donc d'estimer la distribution empirique de $min\ d(a_{\bar{k}}^i, b_{j,\bar{k}})$ sachant que $a_{\bar{k}}^i \notin B_{\bar{k}}$ avec $B_{\bar{k}}$ le sous-bloc de B ayant le plus grand cardinal. Utiliser cette distribution devrait alors conduire à un test plus conservateur puisque le risque de première espèce dans chacun des sous-blocs B_k y est plus faible (sous l'hypothèse retenue). Or, pour que l'appariement ait du

4. Dans la littérature [1], il est possible de considérer deux seuils avec entre les deux des appariements potentiels. Ce n'est pas le cas dans notre processus d'appariement pour lequel les appariements potentiels seront considérés comme non réalisés. La qualité des résultats obtenus permettant de s'affranchir de cette subtilité.

sens, il faudrait que l'individu a_k^i puisse se retrouver dans le sous-bloc B_k, ce qui semble contradictoire avec la nécessité de donner la loi de $min\, d(a_{\bar{k}}^i, b_{j,\bar{k}})$ sachant que $a_{\bar{k}}^i \notin B_{\bar{k}}$. Finalement, la solution retenue afin de contourner cet écueil est d'approximer la distribution étudiée à l'aide de celle de $min\, d(c_i, b_{j,\bar{k}})$ sachant que $c_i \notin B_{\bar{k}}$, avec c_i un élément tiré aléatoirement dans $\bar{A}_{\bar{k}}$, le complémentaire de $A_{\bar{k}}$ dans A. Le but étant naturellement de se placer autant que possible sous l'hypothèse que c_i n'appartient pas à $B_{\bar{k}}$. Finalement, pour un seuil donné, l'estimateur théorique retenu du niveau est la probabilité pour un individu c_i tiré au hasard dans $\bar{A}_{\bar{k}}$ d'avoir un écho dont la distance à B_k est inférieure à un seuil S fixé

$$\alpha_T = \mathbb{P}\left(\min_{b_{j,\bar{k}} \in B_{\bar{k}}} d(c_i, b_{j,\bar{k}}) < \mathrm{S} \mid c_i \in \bar{A}_{\bar{k}} \right) \tag{8}$$

Bien entendu, le code commune n'est alors pas retenu dans le calcul de la distance afin de ne pas faire augmenter artificiellement le seuil d'acceptation de H_1 sous H_0. Finalement, en tirant aléatoirement N individus c_i dans $\bar{A}_{\bar{k}}$, on estimera α_T (pour un seuil donné) par

$$\widehat{\alpha_T} = \frac{1}{N} \sum_{i=1}^{N} 1_{\{d(c_i, B_{\bar{k}}) < \mathrm{S} \text{ tel que } c_i \in \bar{A}_{\bar{k}}\}} \tag{9}$$

5.2 Estimation de la distribution et résultats des simulations

Le code commune ayant le plus d'observations dans la base fiscale est 75000, cela renvoie donc à des individus ayant eu une adresse en région parisienne. Dans le processus de calcul des échos au sein d'un sous-bloc, le 15e arrondissement de la ville de Paris, dont le code commune est 75115, s'y retrouve donc associé comme cela a été décrit dans la section 3. Finalement, le sous-bloc $B_{\bar{k}}$ choisi est la réunion disjointe des sous-blocs B_{75000} et B_{75115} et son cardinal dépasse le million d'observations. Ensuite, afin d'estimer empiriquement la distribution de $min\, d(c_i, b_{j,\bar{k}})$ sachant que $c_i \notin B_{\bar{k}}$, 5000 individus c_i sont tirés aléatoirement, sur la base d'un sondage aléatoire simple, parmi les individus dont le code commune ne commence pas par 75. En effet, un individu ayant par exemple déménagé depuis le 14 e arrondissement parisien peut apparaître dans FIP avec un code commune valant 75000 et donc décaler la distribution recherchée vers la gauche.

On obtient alors le graphique 3 où il apparaît que la probabilité d'obtenir une distance inférieure à 1.66, dès lors que l'individu n'appartient pas à la commune, est quasiment nulle.

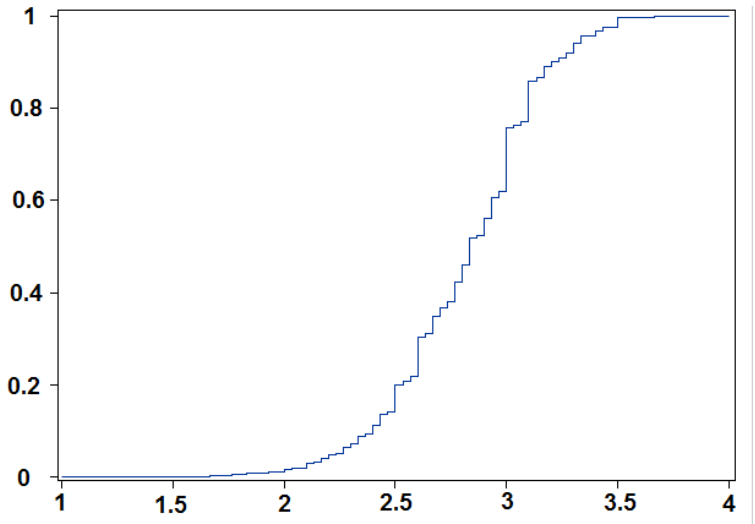

FIGURE 3 – Fonction de répartition estimée du minimum des distances des individus

Plus précisément, sur les 5 000 individus sélectionnés, la distance minimale est inférieure ou égale à 1 pour 8 individus. Or ce cas là ne se produit que s'il agit du même individu qui, suite à un déménagement, apparaît dans les deux sous-ensembles. Pour ces derniers, on remplace la distance obtenue par la seconde distance la plus faible. Par ailleurs, une seule personne obtient une distance de 1.4, a priori il s'agit de deux jumelles dont le prénom et le libellé de l'adresse sont suffisamment proches pour obtenir une distance très faible. Finalement, les premiers quantiles sont les suivants

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

TABLE 1 – Quantiles estimés de la distribution du minimum des distances entre les individus de $\bar{A}_{\bar{k}}$ et ceux de $B_{\bar{k}}$

Finalement, dès lors que la distance entre deux individus est inférieure à 1.4, on aura pour ainsi dire 100 % de chances de ne pas apparier à tort l'individu de l'enquête. Cela passera à 99.9 % de chances pour une distance inférieure à 1.62 etc. Par ailleurs, plus des trois quart des individus de l'enquête qui n'ont pas été appariés exactement vivent dans des communes de moins de 50 000 habitants, ce qui permet de penser que ces probabilités sont certainement plus fortes pour ces individus si l'on se place sous l'hypothèse de travail faite précédemment et selon laquelle, pour un seuil donné, le risque de première espèce était supérieur quand la taille de la ville augmentait. En terme d'appariements, on obtient les résultats suivants

Seuil d'appariement retenu	Nombre et taux d'appariement du reliquat 6 531 individus enquêtés non appariés exactement	Appariement global sur les 14 837 individus appariables 8306 appariements exacts	Nombre additionnel d'individus en augmentant le seuil *et* Probabilité de mal apparier ces individus additionnels	
Distance = 0	2 601	10 907 (soit 73.51%)	10 907	Risque de 0%
Distance < 1.4	6 480	14 786 (soit 99.66%)	3 969	Risque de 0%
Distance < 1.62	6 488	14 794 (soit 99.71%)	8	Risque de 0.1%
Distance < 1.8	6 488	14 794 (soit 99.71%)	0	Risque de 0.5%
Distance < 1.93	6 488	14 794 (soit 99.71%)	0	Risque de 1%
Distance < 2.1	6 504	14 810 (soit 99.82%)	16	Risque de 2.5%
Distance < 2.23	6 507	14 814 (soit 99.84%)	3	Risque de 5%

TABLE 2 – Taux d'appariement en fonction du seuil choisi
Note de lecture : 3 969 individus sont appariés avec une distance superieure à 0 et inférieure à 1.4. Ces individus ont 100% de chances d'être bien appariés.

On peut déduire du tableau 2 que pour les individus dont la distance est inférieure ou égale à 1.62, l'écho retrouvé est très certainement le bon. Dans la pratique, il se peut que l'on retrouve plusieurs échos dont la distance est inférieur à 1.62 dans la phase d'appariement approximatif, toutefois il s'agit quasiment à chaque fois de la même personne et donc l'appariement en tant que tel est bel et bien réalisé.

6 Conclusion

Dans ce document, l'accent a uniquement été mis sur l'appariement. Toutefois, le travail du Pôle Revenus Fiscaux et Sociaux n'est pas tant d'apparier un fichier d'enquête que de l'enrichir. Un retour en arrière sur le fichier fiscal s'impose. Ce fichier permanent renvoie des informations parfois datées sur des individus puisqu'il recense les différentes adresses d'un même individu, voire dans plusieurs départements et, dans ce cas, sous des identifiants fiscaux différents. Si d'un point de vue appariement les résultats sont très satisfaisants, tant d'un point de vue qualité que d'un point de vue durée, simplicité et reproductibilité. Cela ne suffit pas à enrichir l'enquête en fournissant par exemple des informations sur le revenu. Pour cela, un autre fichier est nécessaire : le Permanent des Occurences de Traitement des Emissions (POTE) qui regroupe les déclarations de revenus. Dans ce fichier, l'identifiant fiscal est bien présent et permet une fusion immédiate avec le fichier FIP pour enrichir l'enquête, sous réserve bien entendu que l'identifiant fiscal associé au meilleur écho de l'individu apparaisse dans le POTE.

La méthode présentée ici a été pensée et testée initialement pour l'appariement et il faudrait la modifier à la marge pour l'enrichissement. En effet, certains individus qui ont été appariés exactement ne se retrouvent pas dans le POTE et il faudrait les reverser dans le reliquat de l'enquête non enrichi pour identifier des échos approximatifs ayant une faible distance (et donc correspondant à un appariement valide) et pour lesquels l'information nécessaire à l'enrichissement est présente dans POTE. Nonobstant cette remarque, l'identifiant fiscal de 7 821 individus parmi les 8 306 appariés exactement est retrouvé dans le POTE et celui de 6 193 l'est sur les 6 531 individus appariés par leur plus proche écho. Finalement, sans reverser les individus appariés exactement dans le processus d'appariement approximatif, 14 014 individus sont enrichis sur les 15 000 de l'enquête [5], soit 93.4 % d'enrichissement.

Par ailleurs, même quand le meilleur écho n'appartient pas au fichier POTE, les échos suivants peuvent tout de même être reliés à l'individu de l'enquête et apparaître dans POTE. En se servant des quantiles obtenus dans le tableau 1 et en considérant qu'une distance inférieure à 1.4 renvoie un bon écho, on retrouve 6 202 échos enrichissables. Pour 12 individus, le plus proche écho enrichissable a une distance comprise entre 1.4 et 1.62. Une reprise manuelle permet alors de valider 8 nouveaux enrichissements. Si la distance du plus proche écho enrichissable est comprise entre 1.62 et 2.23, il est possible d'en identifier 23 autres sur les 35 dans ce cas. Si au-delà, le risque semble trop important, une reprise manuelle rapide permet de récupérer 31 enrichissements auquel s'ajoute les 9 dont l'écho enrichissable a une distance inférieure à 4. Enfin, on peut obtenir un taux d'enrichissement de 93.7 % mais le principal gain envisageable reste bien entendu l'identification du plus proche écho enrichissable que l'on obtiendrait sur les 485 individus appariés exactement et non enrichis.

A titre de comparaison, la méthode actuelle a permis d'enrichir 90 % des données au cours d'un processus nettement plus long que le processus proposé ici. La phase de normalisation des données a duré autour de 12h (normalement mutualisable en début d'année dans le cadre d'une mise en production) et la recherche des plus proches échos a nécessité un peu moins de 14h de calculs. Les autres étapes de fusions ou de tri sur la FIP individualisée n'ont quant à elles duré au plus que 30 minutes chacune. Par ailleurs, au-delà du temps machine requis qui pourrait augmenter fortement selon la taille de la base d'enquête, une fois le processus lancé l'intervention humaine n'est plus nécessaire hormis éventuellement une phase de reprise manuelle qu'il ne faut jamais totalement exclure dans un processus d'appariement.

5. Pour rappel, 163 individus ont été exclus du processus d'appariement à cause d'un trop grand nombre de données manquantes.

Références

[1] Christen P., *Data Matching*, Data-Centric Systems and Applications, DOI : 10.1007/978-3-642-31164-2_4, ©Springer-Verlag Berlin Heidelberg 2012.

[2] Fellegi I.P. and Sunter A.B., *A theory for record linkage*. Journal of the American Statistical Association **64**(328), 1183-1210 (1969).