Journal de la Société Française de Statistique
Vol. 159 No. 3 (2018)

Revue Bibliographique des Méthodes de Couplage des Bases de Données : Applications et Perspectives dans le Cas des Données de Santé Publique

Title: An Overview of Record Linkage Methods: Applications and Perspective on Health Data

Said Karim Bounebache[1], Catherine Quantin[2], Éric Benzenine[2], Guillaume Obozinski[3] et Grégoire Rey[1]

Résumé : Le couplage des bases de données est un enjeu important en santé publique, particulièrement en cette période de multiplication des bases de données administratives et de cohortes (Loth, 2015). Cette procédure consiste à faire correspondre des informations concernant un individu issues de base de données différentes sans pouvoir utiliser un identifiant unique. En France, dans le cas des données médicales et administratives, le Numéro d'Identification au Répertoire (NIR) est un exemple d'identifiant susceptible d'être utilisé pour servir de clé de couplage. Cependant ce dernier restera, en dépit de la loi du 26 janvier 2016 de modernisation de notre système de santé, difficile d'accès en raison de sa qualité d'identifiant direct commun à de nombreuses bases de données. Nous présentons les méthodes de chaînage susceptibles d'être utilisées par des chercheurs, en nous concentrant sur le modèle génératif de Fellegi et Sunter qui est une approche non supervisée, ainsi que sur quelques méthodes issues de l'apprentissage statistique. Enfin nous présentons rapidement différentes approches pour réaliser une analyse statistique sur des données appariées et comment répercuter l'incertitude de l'appariement dans l'analyse.

Abstract: Record linkage has become a powerful tool for public health, since the rise of medical and administrative database or cohort (Loth, 2015). This process allows matching individual's information obtained from different databases which don't have necessarily a common identifier. Furthermore, if such common identifier exists it could take a long time to obtain the necessary approval to use it. In France, the NIR is the identifier which is the most likely to be an identifier at the national level. However, in order to use the NIR, it is still compulsory to obtain the authorization from the CNIL even after the change of law concerning the modernization of the French Healthcare system. This paper presents a broad set of methods to perform record linkage, in particular the method proposed by Fellegi and Sunter and its extensions. The aim is to give some guidelines to researchers and to introduce some approaches to incorporate uncertainty associated with the linkage in their analysis.

Mots-clés : couplage/appariement indirect, bases de données médicales et administratives, réseau bayésien naïf, modèle mixte.

Keywords: record linkage, healthcare database, naive bayes network, mixed model

Classification AMS 2000 : 35L05, 35L70

[1] Inserm-CépiDc.
E-mail : said.bounebache@inserm.fr and E-mail : gregoire.rey@inserm.fr
[2] CHU de Dijon.
E-mail : catherine.quantin@chu-dijon.fr and E-mail : eric.benzenine@chu-dijon.fr
[3] École des Ponts-PariTech.
E-mail : guillaume.obozinski@imagine.enpc.fr

 ISSN: 2102-6238

1. Introduction

Le couplage des bases de données est un enjeu important en santé publique, particulièrement en cette période de multiplication des bases de données administratives et de cohortes (Loth, 2015). On utilisera parfois le terme d'appariement ou de chaînage dont l'emploi est très répandu en santé. L'intérêt des couplages est multiple, il peut aussi bien servir à enrichir une base de données pour la production d'études épidémiologiques qu'à l'amélioration de la qualité des informations contenues dans une base de données. Le besoin de méthodologies spécifiques provient du fait que les bases de données n'ont pas systématiquement un identifiant direct commun permettant de chaîner les informations concernant un individu donné. Par ailleurs, quand cet identifiant existe, les procédures permettant de l'utiliser ont jusqu'à présent été très longues, complexes et sans la garantie d'aboutir. En France, dans le cas des données médico-administratives, le Numéro d'Identification au Répertoire (NIR) est un exemple d'identifiant susceptible d'être utilisé pour servir de clé de couplage. Cependant, ce dernier était et restera sans doute, malgré les évolutions juridiques récentes (loi du 26 janvier 2016 de modernisation du système de santé), d'usage restreint justement en raison de sa qualité d'identifiant direct commun à de nombreuses bases de données. La multiplication des bases de données administratives et des cohortes mais aussi le volume d'informations qu'elles contiennent ont fait du couplage un réel défi sur le plan théorique ainsi que sur le plan opérationnel. En effet le volume des données et le désir de formaliser les règles de décision prises ont rendu obsolète la réalisation de couplages sans une automatisation importante. C'est dans cette optique que les méthodes de couplage de données se sont étoffées au fil des ans. L'objectif de ce document est de présenter les méthodes de couplage susceptibles d'être utilisées par des chercheurs, ainsi que les moyens de les mettre en œuvre, afin que toute équipe puisse concevoir et réaliser de la façon la plus pertinente un projet incluant le rapprochement de base de données. Nous tenons à préciser que beaucoup de ces méthodes sont à l'état de recherche et n'ont jamais été appliquées en taille réelle, pour la production de données, notamment dans le cas français. Ces méthodes étant issues de disciplines différentes, allant de la statistique à l'informatique, et d'origine plutôt anglo-saxonne, il n'existe pas à notre connaissance de document qui fait la synthèse de l'ensemble de ces méthodes et qui prennent en compte la particularité du système français des bases données médicales et administratives et de la législation qui les protège. Nous ne cherchons pas l'exhaustivité, mais nous décrivons les principaux outils pour réaliser et encadrer un couplage sur le plan théorique et pratique sans entrer dans la problématique juridique. Nous aborderons aussi la problématique de la prise en compte des erreurs qui peuvent être induites par le couplage.

2. Les méthodes de couplage

Dans une situation idéale, pour réaliser un couplage (les termes de chaînage ou d'appariement apparaissent souvent notamment dans le domaine de la santé), l'existence d'un identifiant direct entre les bases de données suffirait à permettre l'alignement direct des informations concernant un individu. En France, le NIR est aujourd'hui l'identifiant le plus généralisé pouvant permettre des alignements. Cependant c'est justement pour son caractère discriminant et généralisé que son usage est très réglementé. Par ailleurs beaucoup de bases de données existantes ou en cours de recueil n'intègrent pas le NIR, mais plutôt des informations d'état civil comme le domicile ou bien

 ISSN: 2102-6238

TABLEAU 1. *Un exemple de deux bases de données, E_A issue d'une population A et E_B issue d'une population B, à apparier pour aligner la variable X et Y*

Base E_A				Base E_B			
X	Variables de couplage			Variables de couplage			Y
x_1	v_{11}	$\dots$	v_{1k}	v_{11}	$\dots$	v_{1k}	y_1
$\vdots$							$\vdots$
x_i	v_{i1}	$\dots$	v_{ik}	v_{i1}	$\dots$	v_{ik}	y_i
$\vdots$							$\vdots$
$x_{n_{E_A}}$	$v_{n_{E_A}1}$	$\dots$	$v_{n_{E_A}k}$	$v_{n_{E_B}1}$	$\dots$	$v_{n_{E_B}k}$	$y_{n_{E_B}}$

le nom et le prénom. Sans identifiant direct, l'appariement devra se faire sur des informations communes aux bases de données. Il s'agira souvent de comparer sur ces variables communes chaque individu de la base E_A avec les individus de la base E_B. On appellera une paire (a,b) un élément de $E_A \times E_B$, une paire (a,b) est un couple si l'individu a est identique à l'individu b. Ainsi il faut d'abord détecter dans chacune des bases les champs (on utilisera indifféremment le terme champ ou variable ou clé de couplage) indirectement identifiants rendant possible la réalisation d'un couplage. On appellera identifiant indirect toute variable ou combinaison de variables qui permet de retrouver l'identité numérique d'un individu (qui permet de reconnaitre un individu dans une base de données), sans utiliser l'identité physique (dans le monde réel). Voici la liste des variables utilisées en pratique pour la ré-identification indirecte :

- Nom et prénom (en raison de leur non unicité, et des erreurs possibles d'écriture, ils ne sont pas considérés comme des identifiants certains et physiques des personnes)
- Sexe
- Date de naissance
- Date de décès
- Date de soin
- L'adresse postale
- Département et commune de domicile
- Département et commune de décès
- Département et commune de naissance
- Établissement de soin

L'utilisation de ces champs se heurte à quelques difficultés. D'abord il faut s'assurer que la combinaison de variables est suffisamment discriminante. Pour cela une condition nécessaire est (mais pas suffisante car les distributions sont rarement uniformes) que le nombre de modalités obtenu en croisant les variables doit être au moins supérieur au nombre obtenu en additionnant les nombres d'individus dans les deux bases. Une autre difficulté provient de tous les facteurs par lesquels les informations d'un même individu peuvent être renseignées différemment ou mal renseignées dans des bases différentes. En premier lieu, il se peut que la définition des variables ne corresponde pas exactement. Par exemple le lieu de résidence au moment du décès pourra être approximé par des informations concernant le dernier lieu de résidence connu. D'autre part la présence d'erreurs dans la base peut rendre l'identification incertaine, ces erreurs peuvent être dues au mode de collecte des données, ou bien encore à l'hétérogénéité des formats et des types utilisés. Pour les erreurs liées au format et au type il suffira de précéder l'appariement d'une

 ISSN: 2102-6238

étape de standardisation (par des méthodes issues de l'analyse syntaxique) du contenu des variables de couplage. Cela consiste principalement à mettre sous la même forme les variables en éliminant par exemple tous les caractères non signifiants, à l'utilisation d'un index pour filtrer les termes, à gérer des abréviations, la casse, etc. Ainsi, le couplage est souvent précédé par une étape de préparation des données et de pré-traitement afin de rendre les informations comparables. Malgré toutes ces précautions il reste difficile voire impossible d'identifier les couplages exacts dans tous les cas. Ainsi nous distinguerons quatre types de résultat. Le premier cas correspond à celui où deux individus sont appariés de façon unique, on dira alors que le couplage est univoque. Le second correspond à la situation ou un individu d'une base est apparié à plusieurs individus de l'autre base, on dira alors que le couplage est multivoque. Le troisième survient quand l'algorithme ne sait pas arbitrer le statut d'une paire (qui correspond au cas de figure où il faudrait réaliser un traitement supplémentaire). On dira alors que le couplage est équivoque. Enfin lorsque l'algorithme arbitrera que les individus sont différents on dira que la paire est non couplée (ou non appariée). On dira que le couplage entre deux bases de données est complet lorsque ces deux bases contiennent des informations sur la même sous population, dans le cas contraire on dira que le couplage est incomplet. La plupart du temps ce problème de couplage peut se réduire à un simple problème de discrimination (apparié ou non apparié) ou bien de classification (qui consiste à faire des regroupements d'objets ou d'individus en classes homogènes). Il existe plusieurs types, non exclusifs, de méthodes de classification utilisées pour le couplage et proposées par Christen (2012) :

- La discrimination fondée sur un seuil, qui consiste à utiliser un indice de similarité (cf. annexe) pour comparer deux enregistrements et à définir un seuil par rapport auquel l'indice permet d'arbitrer le statut de la paire. On peut utiliser aussi deux seuils si l'on veut constituer une classe intermédiaire de paire à vérifier manuellement afin de réduire le nombre de faux positifs et de faux négatifs. Les seuils sont choisis manuellement (à dire d'expert) ou bien par apprentissage.
- La discrimination fondée sur une règle de décision, elle intègre les indices de similarité au sein d'expressions booléennes sous forme normale conjonctive. Les termes de la forme sont des tests utilisant les indices de similarité que l'on compare à un seuil.
- La discrimination probabiliste, qui consiste à utiliser un modèle probabiliste de génération des données (souvent appelé modèle génératif).
- La discrimination fondée sur un coût, se base sur les modèles probabilistes et introduit une fonction de coût.
- La discrimination supervisée, consiste à utiliser des modèles statistiques pour la discrimination comme la régression, les réseaux de neurones ou encore les séparateurs à vaste marge. elle permet de construire un système de classification (fonction qui permet de prédire la classe en fonction des attributs) sur la base d'exemple.
- La discrimination par apprentissage actif, est un procédé itératif qui consiste à améliorer l'apprentissage en introduisant au fur et à mesure des nouveaux cas pour lesquels la classification a échoué. Cette méthode est utilisée quand la taille du jeu de données d'apprentissage initial est faible.
- La discrimination par clustering, est un ensemble de méthodes consistant a mettre en évidence des groupes pour lesquels les ressemblances au sein des groupes sont fortes et les ressemblances sont faibles entre les groupes. L'idée est d'utiliser ce type de regroupement

 ISSN: 2102-6238

pour faire de la discrimination. L'avantage de ces méthodes est qu'elles n'ont pas besoin de données pour lesquelles la classe est connue. Ceci est très pratique pour le couplage puisque le statut est souvent difficile à obtenir.

– La discrimination collective, est un ensemble de méthodes qui utilise la théorie des graphes afin de réaliser un couplage. L'information étant structurée en réseaux, le couplage d'une paire d'enregistrement se fait en prenant en compte l'information concernant les autres paires d'enregistrements. Plus généralement, ce mode d'appariement permet de coupler des bases de données contenant des entités de nature différente. Par exemple, on peut coupler une base de données contenant des auteurs avec une base de données contenant des articles scientifique (Christen, 2012). Pour coupler un auteur avec un article on pourra exploiter des informations sur les co-auteurs de l'article et éventuellement d'autres collaborations de l'auteur.

En pratique les méthodes utilisées peuvent se trouver à l'intersection de plusieurs de ces catégories. Nous distinguerons donc deux grandes classes de méthodes de couplage : les couplages déterministes, qui correspondent aux méthodes à dire d'expert, pouvant faire appel à l'apprentissage statistique (analyse discriminante) et les couplages probabilistes, qui correspondent essentiellement à la proposition de Fellegi et Sunter (1969) et Copas et Hilton (1990) et à leurs extensions. Nous essaierons dans la suite de faire une présentation des principes attachés aux différentes classes, puis nous présenterons deux exemples en santé publique. Certaines de ces méthodes font appel à l'apprentissage machine, nous rappelons rapidement quelques notions de base. On appelle apprentissage machine toutes les méthodes qui permettent de construire un modèle de la réalité à partir de données, que l'on appelle données d'apprentissage, afin d'automatiser des tâches. Cet apprentissage est supervisé si l'algorithme utilisé pour apprendre se base sur l'utilisation d'exemples pour lesquels la tâche a déjà été réalisée (plus ou moins correctement) que l'on appelle des données étiquetées. Dans le cas du couplage cela correspond à des paires d'individus dont on sait qu'elles correspondent ou non à un même individu. Dans le cas contraire, l'apprentissage est dit non supervisé. Il arrive que l'on soit amener à apprendre sur un ensemble de données où une partie seulement des données sont étiquetées, on parlera alors d'apprentissage semi-supervisé.

2.1. Le blocage

Les bases de données médicales et administratives ou statistiques, comme celles que nous pouvons trouver dans le SNIIRAM (Système National d'Information Inter-Régimes de l'Assurance Maladie, géré par la Caisse Nationale de l'Assurance Maladie des Travailleurs Salariés CNAMTS), dans l'EDP (Echantillon Démographique Permanent, géré par l'INSEE) ou la BCMD (Base des Causes Médicales de Décès, gérée par l'Inserm), sont volumineuses et peuvent rendre difficile l'application directe des méthodes de couplage. L'échantillon $E_A \times E_B$, contient un total de $n_{E_A} \times n_{E_B}$ individus (sachant que la BCMD cumule environ 550 000 entrées chaque année). Il est donc nécessaire de réduire le nombre de comparaisons. On appelle blocage tout processus permettant de réduire le nombre des paires d'individu à comparer. Si ce blocage est fait en préambule à l'apprentissage, cela peut avoir des conséquences diverses. Dans une approche non supervisée du couplage, le blocage peut rendre plus efficace les méthodes de classification car il peut permettre de rééquilibrer les classes, à savoir réduire les écarts de taille entre les classes. Cependant

 ISSN: 2102-6238

cette étape entrainera des biais pour la classification (clustering) mais aussi pour l'estimation des paramètres des modèles probabilistes (si cela n'est pris en compte dans les spécifications). Ces biais concernent principalement l'ensemble des paires qui ne sont pas des couples, augmentant le nombre de couples mal classés ainsi que la part des traitements manuels à réaliser. Si ce blocage est fait après l'estimation des paramètres ou l'apprentissage, l'effet d'une réduction des comparaisons est connu en théorie (Fellegi et Sunter, 1969; Christen, 2012). En effet si l'on ne s'intéresse qu'à un sous-ensemble de $E_A \times E_B$ en imposant certaines concordances sur un ou plusieurs champs, il en résultera une réduction du risque d'apparier des individus à tort mais à une augmentation du nombre des non appariés à tort. Par ailleurs cela diminuera le nombre de paires à vérifier manuellement.

Il existe plusieurs méthodes de réduction du nombre de paires et on pourra se référer à Elmagarmind *et al.* (2007), Steorts *et al.* (2014b) et Batxer *et al.* (2003) pour une synthèse exhaustive. Nous présentons rapidement quelques-unes de ces techniques les plus connues.

La première et la plus simple consiste à choisir parmi une liste de variables, des variables suffisamment discriminantes et dont la qualité semble bonne, et à ne comparer que les couples dont la valeur du champ de référence est identique. Un exemple de similarité qui peut être utilisée est le Soundex (Fellegi et Sunter, 1969; Jaro, 1989; Herzog *et al.*, 2007), ou des algorithmes analogues, car cette indice minimise les distorsions, en comparaison avec l'utilisation d'une autre similarité (surtout pour les champs contenant des chaînes de caractères de type nom, prénom et adresse). Une alternative pour réduire les biais et les risques de non comparaison est d'appliquer plusieurs étapes en choisissant à chaque fois un ou plusieurs champs différents. On pourra utiliser l'indice de similarité utilisant les bigrammes (cf. annexe) en ne comparant que les entrées pour lesquels, pour une variable donnée, la similarité dépasse un certain seuil. Plus grand est le seuil, plus la taille des blocs sera réduite.

La méthode de la fenêtre glissante (Sorted Neighbourhood) consiste à choisir ou à créer une variable et d'ordonner les individus selon cette variable. Il s'agira ensuite de faire glisser une fenêtre, dont la taille est choisie arbitrairement, sur cette variable et de comparer les entrées dont la valeur de la variable est contenue dans la fenêtre.

Enfin on pourra utiliser l'indice de similarité cosinus (cf. annexe) dans une approche de type "proche voisin". Il s'agit de créer un jeu de données dont les blocs sont construits de la façon suivante : on choisit aléatoirement un individu parmi la population et on construit un premier bloc en prenant tous les individus de la population suffisamment proche. Ce bloc est ensuite retiré de la population, avant la sélection d'un second individu suivi de la construction d'un second bloc. Les individus suffisamment proches, le sont au sens de l'indice de similarité cosinus.

Batxer *et al.* (2003) s'emploient à comparer les quatre propositions précédentes et montrent que la méthode des bigrammes et la méthode cosinus ont des meilleures performances [1] que le blocage simple et la fenêtre glissante. Les méthodes de blocage ne sont pas intéressantes que pour le gain de temps réalisé mais aussi pour permettre l'usage de méthode de clustering pour l'apprentissage non supervisé. Les méthodes de K-moyennes sont performantes quand les classes latentes sont bien séparées avec des tailles homogènes, or on sait qu'au sein du produit cartésien des bases on s'attend au mieux à retrouver un nombre de personnes égal à la taille de la plus petite des bases.

[1] définies dans Elfeky *et al.* (2002)

 ISSN: 2102-6238

2.2. *Les couplages déterministes*

Les méthodes que nous allons présentées succinctement ici utilisent souvent des indices de similarités. Elles peuvent être utilisées dans des règles de décision, servir pour construire des fonction de coût dans des algorithmes d'optimisation combinatoire pour donner des relations univoques, ou encore servir à construire des modèles pour apprentissage.

Les méthode les plus simples consistent à utiliser une similarité (à ne pas confondre avec les distances qui mesurent l'écart et qui ont des propriétés mathématiques spécifiques, les similarités évaluent la proximité et n'ont pas de propriété particulière) et à choisir un seuil adéquat. Tout couple d'entrées dont la similarité sera inférieure [2] à ce seuil sera considéré comme une paire d'entrées d'individus distincts. Une façon de procéder est de choisir une des similarités citées en annexe et de concaténer l'ensemble des champs concernant une entrée en un champ unique. Il y a une perte d'information concernant les différents types de champs, qui une fois concaténés ensembles perdent leur spécificité (structure des erreurs, pouvoir discriminant). Une autre méthode consiste à garder les différents champs mais à construire une distance égale à la combinaison des distances ad hoc et de choisir un seuil de test, on pourra se référer à Monge et Elkan (1996), Monge et Elkan (1997) et Cohen (2000). Le choix du seuil est un point important de ces méthodes et celui-ci nécessite souvent l'emploi d'un jeu de données d'apprentissage.

Il existe des tentatives plus sophistiquées qui essaient de s'affranchir du choix arbitraire d'un seuil. Par exemple Guha *et al.* (2004) cherchent à apparier une requête à une base relationnelle. Pour cela on classera la base de données de l'élément le plus proche à l'élément le moins proche à la requête autant de fois qu'il y a de variable d'appariement. Chaque classement se fait en comparant les éléments de la base avec la requête en calculant l'écart via un indice de similarité sur une des variables d'appariement. Enfin, après avoir construit une distance qui mesure l'écart entre les classements, les auteurs proposent alors un algorithme d'optimisation combinatoire qui permet de construire un classement médian. L'algorithme n'a pas besoin de seuil en revanche il faut paramétrer le nombre d'éléments dans le classement. Issue du traitement automatique du langage, Ananthakrishna *et al.* (2002) cherchent à détecter des doublons dans un entrepôt de bases de données en exploitant la structure hiérarchique existante (sur des données géographiques) entre les différentes bases. Pour chaque entité de la base un score est calculé sur la base d'un indice de similarité. Plus ce dernier est élevé plus il y a de chance que cette entité soit un doublon. Enfin une technique de détection de valeurs aberrantes est utilisée pour détecter les doublons de façon floue, à savoir les entités avec des score extrêmement grand, pour construire un seuil. Chanduri *et al.* (2005) proposent un algorithme de type proche voisin pour détecter des doublons. Comme précédemment et sous certaines hypothèses il est possible de construire un score sous la forme d'une fonction de répartition dont les variations permettront de détecter le seuil. Si les hypothèses sont bonnes alors la densité de probabilité, dérivée du score, doit avoir deux profils en cloche une représentant le pattern des doublons tandis que la seconde représentante celui des entités uniques. Notons que Domingo-Ferrer et Torra (2002) et Torra et Domingo-Ferrer (2003) montrent que l'algorithme simple, qui consiste à choisir pour un individu d'une base l'individu le plus proche de l'autre base, présente des performances satisfaisantes. Jaro (1989) propose

[2] La plupart des indices de similarité calculent le nombre de points communs et non le nombre de dissemblances. Dans une situation où l'on mesure effectivement la dissemblance, c'est lorsque l'indice est supérieure à un certain seuil que nous déciderons de ne pas coupler une paire.

 ISSN: 2102-6238

(dans le cas des modèles probabilistes mais qui reste valable pour les indices de similarité) de calculer les distances entre les individus de chacune des paires et de trouver à posteriori la relation univoque entre les deux bases qui maximise un score sous l'hypothèse que l'une des bases est incluse dans l'autre. L'algorithme utilisé est un algorithme bien connu en recherche opérationnelle qui résout les problèmes d'allocation linéaire. Goldstein *et al.* (2017) proposent une approche non supervisée utilisant l'analyse des correspondances pour construire un système de scores attachés aux modalités prises par les indices de similarité (quitte à discrétiser l'indice). Cette proposition se ramène à résoudre un simple système linéaire, dont le passage à l'échelle fait apparaître des problèmes de conditionnement. Les auteurs ne proposent pas pour l'instant de solution non supervisée pour le choix d'un seuil optimal permettant de discriminer les paires. Le système de score construit dans cette formulation semble plus juste que celui fournit par le modèle probabiliste de Fellegi et Sunter (1969). Même si le classement des paires reste relativement identique notamment à cause de l'hypothèse d'indépendance conditionnelle utilisée dans ce dernier et ce même si la contribution de chaque variable de concordance dans cette nouvelle formulation reste additive.

Les couplages utilisant des règles de décision peuvent être vus comme une extension des méthodes de comparaison à un seuil, puisqu'elle mélange l'expertise sur les bases de données à coupler et l'utilisation d'indice de proximité. Les règles sont souvent spécifiées par un expert, qui a une connaissance de la base et de la façon dont celle-ci a été produite. Ainsi c'est l'expert qui, au travers de règles qu'il a mises en place, et qu'il a formalisé dans un algorithme (sous forme normale conjonctive),attribue le caractère probable d'une association entre deux enregistrements. Lim *et al.* (1996) et Herandez et Stolfo (1998) sont deux exemples simples de ce type d'appariement. Enfin nous verrons dans les applications un exemple d'appariement exploitant la connaissance de la base avec la production de la base Amphi (Lamarche-Vadel *et al.*, 2013). Il est possible d'apprendre sur un jeu de données et de proposer des règles de décision pour réaliser un couplage. L'évaluation de la performance d'une règle de décision se fait par des indicateurs usuels comme la précision, le rappel, les taux de faux et de vrais positifs (négatifs) ou encore l'analyse de la courbe ROC (Cornuéjols et Miclet, 2010; Izenman, 2008; Hastie *et al.*, 2001). Un premier exemple est l'algorithme de boosting qui permet sur la base d'un ensemble de systèmes simples de classification de construire un système plus complexe. La discrimination supervisée utilise souvent un ou plusieurs indices de similarité qu'il est possible de combiner en pénalisant les coefficients (en les contraignant à être positifs par exemple) pour prendre en compte des colinéarités entre les différentes similarités. Quand les deux classes ne sont pas linéairement séparables, on utilisera des méthodes de linéarisation (utilisant des noyaux c'est le cas des séparateurs à vaste marge par exemple). Cochinwala *et al.* (2001) réalise une analyse détaillée d'un couplage utilisant des arbres de décisions dans les problématiques de couplage. Dans le cas où le NIR n'est pas disponible, et plus généralement quand il n'est pas possible d'avoir accès à des données étiquetées, on pourra appliquer les méthodes ci-dessus sur des données qu'on aura étiquetées par l'intermédiaire d'un algorithme automatique de classification (Christen, 2008a, 2007). Verykios *et al.* (2000) poursuivent les développements de Cochinwala *et al.* (2001)en étiquetant les données de façon floue. Cette étiquetage se fait par l'utilisation d'une méthode de classification bayesienne (Cheeseman et Stutz, 1997) sur l'espace des comparaisons [3]. Une fois

[3] L'espace Γ contenant les vecteurs dont chaque coordonnée est le résultat de la comparaison de deux individus ou deux entités sur une variable d'appariement, éventuellement en utilisant un indice de similarité

 ISSN: 2102-6238

mis en évidence une structure latente des données, on fait correspondre à chaque classe obtenue un label (apparié, non apparié, appariement possible). Enfin l'apprentissage de l'arbre de décisions se fait sur cette catégorisation, en utilisant pour chaque nœud de l'arbre une des coordonnées du vecteur des comparaisons. Elmagarmind *et al.* (2007) proposent une bibliographie riche sur les méthodes d'apprentissage des similarités.

2.3. Les modèles probabilistes

Newcombe et Kennedy (1959) sont les premiers à avoir posé le problème du couplage de données en un problème d'inférence bayésienne. L'objectif était d'étudier la faisabilité d'automatisation de cette tâche. En 1968, Fellegi et Sunter (1969) formalisent mathématiquement l'approche de Newcombe et Kennedy (1959) tout en démontrant une certaine forme d'optimalité dans la règle de classement. Cette formalisation se fait par le biais d'un modèle probabiliste exprimant le processus par lequel les données ont été générées.

2.3.1. Une théorie générale du couplage selon Fellegi et Sunter

Soient deux bases de données notées E_A et E_B issues d'une population A et d'une population B respectivement, contenant des informations sur ces populations de type : nom, prénom, âge, date de naissance, date de décès, adresse, etc. On notera $v_i(d)$ la i-ième variable commune entre ces deux bases pour l'individu d, $1 \leq i \leq K$. Chaque entrée de la base concerne un individu de la population. L'objectif du couplage est de déterminer si chaque entrée dans une base se rapporte "probablement" à une entrée dans une autre base, compte tenu du fait que les informations des bases ont pu être altérées par la production d'erreurs ou que l'individu a pu changer de statut entre les deux entrées. Cette problématique se traduit d'un point de vue probabiliste en un problème de discrimination : pour chaque entrée de la base E_A on associe une entrée de la base E_B. L'ensemble des paires ainsi construit est noté $E_A \times E_B$ (le produit cartésien de E_A et E_B). Cet ensemble $E_A \times E_B$ est alors séparé en deux classes : l'ensemble des couples noté M et l'ensemble des non couples noté U. On définit par commodité par la variable $S_{a,b}$ la fonction indicatrice d'être un couple, ainsi $M = \{S = 1\}$ et $U = \{S = 0\}$. Le couplage consistera donc à estimer la séparation entre ces deux classes. Fellegi et Sunter (1969) proposent de réaliser cette séparation dans un autre ensemble que $E_A \times E_B$. Ils proposent en effet de réaliser cette séparation dans l'espace des comparaisons Γ contenant l'image de $E_A \times E_B$ par une fonction δ qui indique variable par variable s'il y a concordance entre deux individus ou bien deux entités. On notera dans la suite $\delta(a,b)$ le vecteur des concordances de la paire $(a,b) \in E_A \times E_B$, où $\delta_i(a,b) = 1$ si $v_i(a) = v_i(b)$ et $\delta_i(a,b) = 0$ dans le cas contraire.

Puisque cette séparation se fait en fonction des variables communes se pose alors la question de leur pouvoir informationnel (ou discriminant) au sens de Shannon. En effet le champ renseignant le sexe d'un individu contiendra moins d'information que le prénom ou encore l'adresse. Pour intégrer ce fait, Newcombe et Kennedy (1959) ont développé le concept du "poids" d'un champ basé sur la probabilité de voir deux variables concorder sur une valeur. Dans Fellegi et Sunter (1969), les auteurs proposent la définition suivante : pour chaque champs i on définit les probabilités $m_i(\delta) = \mathbb{P}(\delta_i = 1|M)$ et $u_i(\delta) = \mathbb{P}(\delta_i = 1|U)$. Le poids w_i du champ i est calculé

 ISSN: 2102-6238

TABLEAU 2. *Un exemple de table de concordance*

Vecteur δ					$E_A \times E_B$	
1	1	1	0	1	a_1	b_1
1	0	1	0	0	a_1	b_2
0	1	1	1	0	a_1	b_3
⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	0	0	0	1	$a_{n_{E_A}}$	$b_{n_{E_B}-1}$
0	0	1	1	1	$a_{n_{E_A}}$	$b_{n_{E_B}}$

comme le logarithme du rapport entre les deux probabilités précédentes en fonction de la valeur de δ_i. Si $\delta_i(a,b) = 1$, alors $w_i(1) = \log\left(\frac{m_i}{u_i}\right)$, dans le cas contraire $w_i(0) = \log\left(\frac{1-m_i}{1-u_i}\right)$. Le poids global, ou score, du couple noté w est la somme des poids des différents champs. Selon que ce score global dépasse un seuil donné, la décision est prise de classer le couple (a,b) comme M ou U. Enfin on supposera que conditionnellement à $S_{a,b}$ les variables $(\delta_i(a,b))_{1\leq i\leq K}$ sont indépendantes. Ce qui implique par exemple que l'on considère le jour, le mois et l'année de naissance comme trois variables séparées alors pour un individu la discordance sur le jour de naissance est indépendante de discordance sur le mois ou l'année. C'est évidemment une hypothèse irréaliste, Thibaudeau (1993), Tromp *et al.* (2008) montrent que cette hypothèse est fausse en pratique (sur des exemples moins triviaux) et qu'elle réduit par ailleurs la qualité de la règle de décision. Nous verrons dans la suite que cette hypothèse est responsable des appariements multivoques. Nous verrons dans la suite que cette hypothèse est responsable en grande partie des appariements multivoques. Cependant cette hypothèse d'indépendance permet de simplifier grandement le problème de l'appariement. La log-vraisemblance du modèle de Fellegi-Sunter est donnée par :

$$\begin{aligned} l(\theta) &= \sum_{(a,b)\in E_A\times E_B} \log\left(\pi \times m[\delta(a,b)] + (1-\pi)\times u[\delta(a,b)]\right) \\ m[\delta(a,b)] &= \prod_{1\leq i\leq K} m_i^{\delta_i(a,b)} \times (1-m_i)^{(1-\delta_i(a,b))} \\ u[\delta(a,b)] &= \prod_{1\leq i\leq K} u_i^{\delta_i(a,b)} \times (1-u_i)^{(1-\delta_i(a,b))} \end{aligned} \tag{1}$$

La solution présentée dans Fellegi et Sunter (1969) est plus complexe car elle propose de classer certaines paires dans un troisième ensemble contenant les paires non classées. La règle de décision peut alors s'écrire :

$$(a,b) \in \begin{cases} \widetilde{M} \text{ si } w > T_{\widetilde{M}} \\ \widetilde{C} \text{ si } T_{\widetilde{U}} \leq w \leq T_{\widetilde{M}} \\ \widetilde{U} \text{ sinon} \end{cases} \tag{2}$$

Avec le score $w = \log\left(\frac{m(\delta)}{u(\delta)}\right)$. Cette règle de décision est optimale dans le sens où étant donné un pourcentage de faux positifs et un pourcentage de faux négatifs tolérés, la règle pré-

 ISSN: 2102-6238

cédente minimise la taille de la troisième classe $\widetilde{C}$ (les seuils $T_{\widetilde{M}}$ et $T_{\widetilde{U}}$ sont des fonctions de ces pourcentages). Finalement le test (2) est un test UPP (Uniformément Plus Puissant) puisque c'est une ré-interprétation du test de Neyman et Pearson (1933).

Pour calculer le seuil $T_{\widetilde{M}}$ il faudra additionner les probabilités u des scores les plus élevés vers les scores les plus faibles jusqu'à ce que la somme dépasse la probabilité de classer a tort un couple dans M. De même pour calculer le seuil $T_{\widetilde{U}}$, on additionnera les probabilités m jusqu'à ce que la somme dépasse la probabilité de classer a tort un couple dans U. Dans la table 3, si on tolère 0.01%, en additionnant les 4 premières probabilités u, on arrive à une probabilité de coupler à tort de 0.0095%. En additionnant la probabilité de la cinquième configuration on dépasse cette probabilité d'apparier à tort. En pratique on choisira une probabilité d'apparier à tort de $2,5\%$ et une probabilité de ne pas apparier à tort de $2,5\%$ ce qui correspond aux pratiques courantes lorsqu'on réalise un test d'hypothèse à 5% (Fournel *et al.*, 2009; Quantin *et al.*, 2009). Le score de cette cinquième configuration est le seuil $T_{\widetilde{M}}$.

TABLEAU 3. *Calcul du seuil $T_{\widetilde{M}}$ de discrimination : si on tolère* 0.01%*, en additionnant les 4 premières probabilités u, on arrive à une probabilité de coupler à tort de* 0.0095%*. En additionnant la probabilité de la cinquième configuration on dépasse cette probabilité d'apparier à tort. Le score de cette cinquième configuration est le seuil $T_{\widetilde{M}}$*

classification	Γ					m	u	Score
$\widetilde{M}$	1	1	1	1	1	64%	0.0004%	5.2
	1	1	0	1	1	2%	0.0005%	3.6
	1	1	1	0	1	16%	0.0063%	3.4
	0	1	1	1	1	1%	0.0023%	2.75
$\widetilde{C}$	1	0	1	1	1	0.6%	0.0016%	2.6
	1	1	1	1	0	11%	0.08%	2.15
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\widetilde{U}$	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	0	0	0	0	1	0.0001%	0.176%	−3.24
	0	0	0	0	0	0.00002%	35%	−6.3

TABLEAU 4. *Classification d'une table de concordance à l'issue du processus d'appariement*

classification	Γ					M	U	Nb total
$\widetilde{M}$	1	1	1	1	1	511	4	515
	1	1	0	1	1	268	9	287
	1	1	1	0	1	115	10	125
	0	1	1	1	1	58	54	112
$\widetilde{C}$	1	0	1	1	1	49	10	59
	1	1	1	1	0	45	60	105
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\widetilde{U}$	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	0	0	0	0	1	3	1150	1153
	0	0	0	0	0	1	9579	9580

 ISSN: 2102-6238

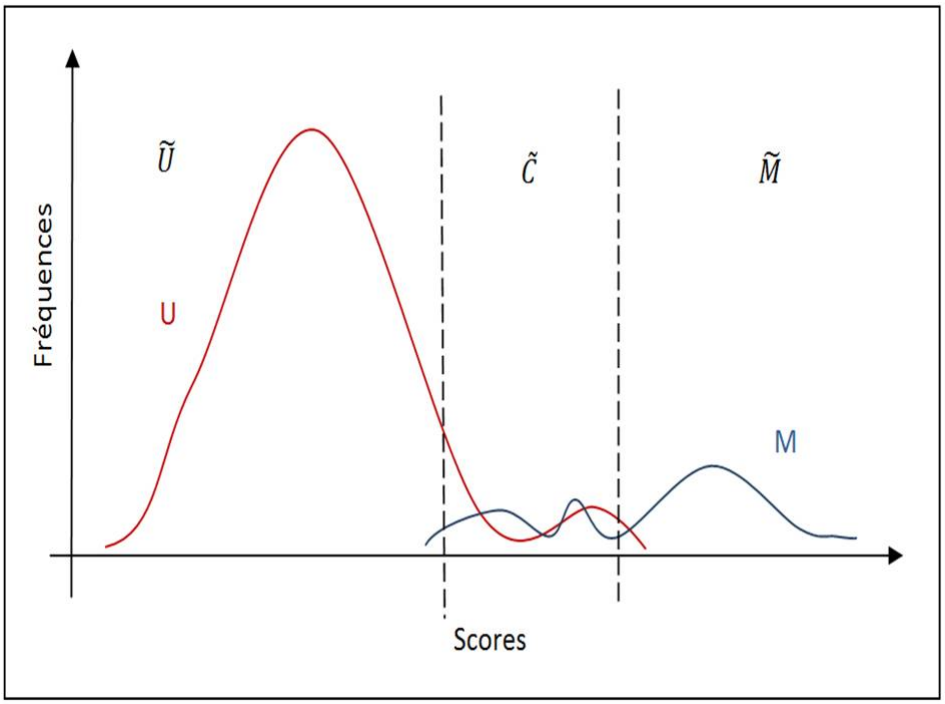

FIGURE 1: Distribution des scores w à l'issue du processus d'appariement décrit ci-dessus

2.3.2. *Prise en compte d'un indice de similarité en présence d'un nombre important d'erreurs*

Il est parfois nécessaire d'appliquer un traitement orthographique avant le couplage. Jaro, et plus tard Winkler (1990), ont proposé une correction des scores w pour des bases comportant des taux d'erreurs typographiques trop importants pour que le couplage précédent soit suffisamment performant. Cette méthode consiste à calculer des poids intermédiaires se situant entre le poids de concordance 1 et le poids de concordance 0. Il s'agit d'utiliser la similarité de Jaro (ou celle de Winkler-Jaro) comme coefficient de relaxation entre les deux poids précédents. Si $\tau_{Jaro}(a_i, b_i)$ est l'indice de Jaro (voir en annexe) appliqué aux champs i :

$$\begin{cases} \log\left(\dfrac{m_i}{u_i}\right), \text{si } \tau_{Jaro}(a_i,b_i) = 1 \\ \max\left(\log\left(\dfrac{m_i}{u_i}\right) - \left(\log\left(\dfrac{m_i}{u_i}\right) - \log\left(\dfrac{1-m_i}{1-u_i}\right)\right)(1-\tau_{Jaro}(a_i,b_i)), \log\left(\dfrac{1-m_i}{1-u_i}\right)\right), \text{sinon} \end{cases}$$

De cette manière la présence d'erreur ne se traduit plus systématiquement par l'addition dans le score global du poids associé à la concordance 0. On pourra remarquer que cette méthode peut aussi réduire le nombre des couplages multivoques résultant de la qualité relative des bases de données à coupler. On peut remarquer, que dans dans le modèle de Fellegi et Sunter (1969) (le modèle proposé n'imposait pas que les concordances prennent que deux modalités), il est possible de remplacer les vecteurs de concordances par des vecteurs d'indice de similarité directement à condition éventuellement de discrétiser ces indices.

 ISSN: 2102-6238

2.3.3. le couplage univoque (1-1) :

La règle de décision de Fellegi et Sunter (1969) consiste à calculer un score d'un couple d'entrée et de comparer ce score à des seuils pour savoir si ces entrées concernent un même individu. Comme chaque couple est testé indépendamment des autres, il peut arriver qu'une entrée de E_A soit associée à plusieurs entrées de la base E_B de même qu'une entrée de E_B peut être associée à plusieurs entrées de la base E_A. Cela peut arriver aussi quand il y a des doublons c'est à dire lorsque des individus paraissent identiques au regard des variables utilisées lors du couplage. Par ailleurs le seuil peut être dépassé sans qu'il y ait concordance sur l'intégralité des champs, une entrée peut être couplée avec plusieurs entrée avec lesquels il concorde sur des combinaisons de variables différentes. Dans le modèle présenté ci-dessous, ce genre de situation vient du fait que le modèle de Fellegi et Sunter n'est pas une vraie vraisemblance. La question que l'on se pose alors est de savoir, dans le cas où l'on cherche à fournir comme correspondance pour chaque individu d'une base au plus un individu d'une autre base, comment choisir les bons couplages. Jaro interprète cette question du choix comme un problème de transport dégénéré connu sous le nom d'affectation linéaire, qui est formulé comme suit :

$$\begin{cases} \text{Maximiser} \sum_a \sum_b C_{a,b} x_{a,b} \\ \sum_a x_{a,b} \leq 1 \\ \sum_b x_{a,b} \leq 1 \end{cases} \tag{3}$$

Où $C_{a,b}$ est un coût qui est une fonction de $W(\delta(a,b))$ (cela peut être le poids lui-même) et $x_{a,b}$ est une indicatrice qui prend la valeur 0 ou 1 selon que la paire est gardée ou non. L'idée n'est donc pas de choisir pour un individu d'une base sa contrepartie dans l'autre avec lequel il a le score le plus élevé, mais il s'agira de maximiser le score global de toutes les paires restantes mises en relation par une bijection. Enfin une fois que les couplages multivoques auront été supprimés on pourra alors comparer les paires restantes avec le seuil de décision. Ce problème d'optimisation peut être résolu par une succession d'opérations simples, et peut donc être implémenté facilement. Par ailleurs ces algorithmes sont bien connus et il existe souvent des instructions préexistantes, voire des packages d'optimisation, qui peuvent être utilisés : la procédure OPTNET (Wright, 2011) en SAS par exemple résout des problèmes d'allocation linéaire.

2.3.4. Estimation des paramètres pour le modèle de Fellegi et Sunter

Les auteurs proposent deux méthodes pour l'estimation des paramètres $m = (m_i)_{1\leq i\leq K}$ et $u = (u_i)_{1\leq i\leq K}$, une première nécessitant une connaissance précise du mécanisme de production d'erreurs ou de divergences dans les bases.

2.3.4.1. la méthode des fréquences d'erreur : Inspiré par Newcombe et Kennedy (1959) qui propose d'automatiser le couplage en utilisant le pouvoir discriminant de l'information (mesuré par l'entropie de Shannon). Reprenons l'exemple de Fellegi et Sunter (1969) sur la variable Nom pour laquelle le nombre d'occurrence pour chaque modalité dans la population J ($J = E_A$ ou E_B) est donné par $(O_{J,t})_{1\leq t\leq n_J}$, où n_J est le nombre de modalité du nom pour la population J. Soient

 ISSN: 2102-6238

maintenant e_J la probabilité qu'un nom dans la population J soit mal reporté, e_{J0} la probabilité que le nom n'apparaisse pas dans la base générée à partir de J, et e_T la probabilité que le nom ait changé après son entrée de la base alors :

$$\begin{cases} m(\delta_{nom}=1) \approx 1-e_{E_A}-e_{E_B}-e_T-e_{E_A0}-e_{E_B0} \\ u(\delta_{nom}=1) \approx (1-e_{E_A}-e_{E_B}-e_T-e_{E_A0}-e_{E_B0}) \quad \times \sum_t \frac{O_{E_A,t}O_{E_B,t}}{N_{E_A}N_{E_B}} \end{cases} \tag{4}$$

Le calcul des probabilités de (4) a nécessité de faire des hypothèses supplémentaires ainsi que des choix dans les approximations de calcul pour les probabilités m et u. On remarquera par ailleurs que les probabilités d'erreur, de non inclusion, ne dépendent de la modalité de la variable Nom. Cette méthode demande des informations sur les populations (ici le nombre d'occurrence des noms) et la méthode de production des bases ainsi que sur les mécanismes de production des erreurs ou des divergences apparaissant dans les bases. Au final cette approche est utilisable dans une démarche supervisée. La méthode suivante est plus pertinente au regard de la situation dans laquelle sont réalisés les couplages des bases médicales et administratives.

2.3.4.2. Estimation non supervisée par l'algorithme d'espérance-maximisation (EM) : C'est Winkler (1988) qui propose l'algorithme EM pour le modèle de Fellegi-Sunter. Jaro (1989) le teste une première fois sur des données de recensement en 1989, les résultats présentés sont très prometteurs quant à l'utilisation de cet algorithme (dans un cadre orienté chaîne de caractères). L'algorithme EM a été développé pour la première fois dans Dempster *et al.* (1977), il permet de faire des estimations par maximum de vraisemblance dans un cadre non convexe en présence de variables latentes. Si δ est la réalisation d'une variable aléatoire Δ dont la loi est donnée par $\mathbb{P}_\theta(\delta)$ (aussi appelée vraisemblance $L(\theta)$), l'estimation par maximum de vraisemblance consiste alors à trouver la valeur θ qui rend le plus probable l'observation θ. Autrement dit il s'agit de déterminer la valeur du paramètre θ qui maximise la vraisemblance. Parfois l'expression de la loi de Δ rend difficile l'utilisation des méthodes classiques d'optimisation, l'idée principale de l'algorithme EM est que la loi de Δ n'est que la loi marginale ou la partie observable d'une variable non observée T. Dans le cas des couplages probabilistes, $T=(\Delta,S)$ où S peut être une donnée manquante ou bien une variable latente correspondant au vrai statut des couples à savoir s'ils correspondent ou non à un même individu. Il s'agira alors de se ramener à la maximisation d'une vraisemblance de la variable T. L'avantage étant que le calcul du θ qui maximise la vraisemblance de T est cette fois plus facile à effectuer (on obtient même une expression explicite dans beaucoup de cas notamment celui du mélange Gaussien). En pratique, il est souvent plus facile de manipuler la log-vraisemblance, et dans la littérature on maximisera plutôt $\log(\mathbb{P}_\theta(\delta))$ ce que nous feront dans la suite. Plus précisément, cette estimation se fait en deux étapes : la première appelée étape "espérance" consiste à calculer la vraisemblance conditionnelle $V(\theta,\tilde{\theta}) = E\left(\log(\mathbb{P}_\theta(T))\,|\Delta=\delta,\tilde{\theta}\right)$, celle-ci est suivi de l'étape "Maximisation" qui consiste à trouver la valeur θ qui maximise V. Pour calculer V il faut d'abord choisir une valeur initiale $\tilde{\theta}$, l'algorithme alterne alors les étapes "E" et "M" successivement jusqu'à obtenir une stabilisation de la vraisemblance. En effet, on peut démontrer que la suite $(\tilde{\theta}_n)_n$ augmente la valeur de la vraisemblance. Cette monotonie fait de l'algorithme EM un algorithme plus stable que

 ISSN: 2102-6238

d'autres algorithmes très utilisés comme celui de Newton-Raphson[4]. Par ailleurs le théorème de Wu et Jeff (1983)[5] garantit la convergence de la séquence des paramètres pour le modèle de Fellegi et Sunter. En théorie l'algorithme converge d'autant plus vite que Δ a de l'information sur S. En appliquant l'algorithme sur la log-vraisemblance on obtient l'estimation suivante à chaque itération. Le paramètre $\theta = (m, u, \pi)$

Étape E :

$$\begin{aligned} \mathbb{P}_{(n-1)}(M|\delta(a,b)) &= \frac{\pi^{(n-1)} \prod_{1\leq i\leq K} (m_i^{(n-1)})^{\delta_i(a,b)} (1-m_i^{(n-1)})^{1-\delta_i(a,b)}}{\mathbb{P}_{(n-1)(\delta(a,b))}} \\ \mathbb{P}_{(n-1)}(U|\delta(a,b)) &= \frac{(1-\pi)^{(n-1)} \prod_{1\leq i\leq K} (u_i^{(n-1)})^{\delta_i(a,b)} (1-u_i^{(n-1)})^{1-\delta_i(a,b)}}{\mathbb{P}_{(n-1)(\delta(a,b))}} \end{aligned} \tag{5}$$

avec

$$\mathbb{P}_{(n-1)}(\delta(a,b)) = \pi^{(n-1)} \prod_{1\leq i\leq K} (m_i^{(n-1)})^{\delta_i(a,b)} (1-m_i^{(n-1)})^{1-\delta_i(a,b)} \tag{6}$$

$$+ \quad (1-\pi)^{(n-1)} \prod_{1\leq i\leq K} (u_i^{(n-1)})^{\delta_i(a,b)} (1-u_i^{(n-1)})^{1-\delta_i(a,b)} \tag{7}$$

Étape M : Pour $i \in \{1, \ldots, K\}$

$$m_i^{(n)} = \frac{\sum_{(a,b)\in E_A\times E_B} \mathbb{P}_{(n-1)}(M|\delta(a,b))\delta_i(a,b)}{\sum_{(a,b)\in E_A\times E_B} \mathbb{P}_{(n-1)}(M|\delta(a,b))} \tag{8}$$

$$u_i^{(n)} = \frac{\sum_{(a,b)\in E_A\times E_B} \mathbb{P}_{(n-1)}(U|\delta(a,b))\delta_i(a,b)}{\sum_{(a,b)\in E_A\times E_B} \mathbb{P}_{(n-1)}(U|\delta(a,b))} \tag{9}$$

$$\pi_i^{(n)} = \frac{\sum_{(a,b)\in E_A\times E_B} \mathbb{P}_{(n-1)}(M|\delta(a,b))}{n_{E_A} \times n_{E_B}} \tag{10}$$

Le critère d'arrêt de cet algorithme consiste souvent à comparer la vraisemblance obtenue en utilisant les derniers paramètres estimés avec la vraisemblance obtenue lors de l'itération précédente (on pourra aussi mesurer l'écart entre deux estimations consécutives). En plus de permettre

[4] Nous avons choisi de porter notre intérêt sur l'algorithme EM pour sa simplicité d'utilisation et les propriétés de stabilité de l'algorithme, cependant il est possible d'utiliser d'autres algorithmes comme les méthodes de Newton et de quasi Newton et les méthodes classiques de descente du gradient. Ces dernières sont par ailleurs plus efficace si les fonctions de vraisemblances sont concaves et unimodales ce qui n'est pas le cas en générale avec les modèles de mélanges dont le paramètre de mélange est la loi de probabilité d'une variable latente. Dans ce cas ces algorithmes sont très sensibles au choix du paramètre d'initialisation, l'algorithme est à la fois simple à implémenter dans le cadre de l'appariement probabiliste (forme fermée la plupart du temps ce qui implique des opérations algébriques qui ne nécessitent pas, contrairement à la méthode de Newton, l'inversion de la matrice d'information observée). Par ailleurs sa convergence est très rapide en pratique, et son comportement asymptotique est stable pour peu que l'on choisisse une base d'apprentissage appropriée. On trouvera par ailleurs une comparaison entre les méthodes de Newton et l'algorithme EM dans McLachlan et Krishnan (2008). Il est possible d'utiliser des méthodes de descente adaptées à certaines fonctions non convexe (on dira α-convexe, voir annexe de Cornuéjols et Miclet (2010).

[5] En effet la vraisemblance possède de bonne propriété de régularité par ailleurs, les paramètres que nous cherchons à estimer sont essentiellement des probabilités et donc l'espace des paramètres a aussi de bonnes propriétés. (Wu et Jeff, 1983)

 ISSN: 2102-6238

un traitement non supervisé des appariements probabilistes, le calcul précédent montre que l'algorithme EM a pour avantage d'être assez simple d'utilisation. D'un point de vue pratique l'algorithme se comporte assez bien en présence d'erreurs typographiques (en nombre relatif, mais peut devenir instable si ce nombre est trop important et/ou converger vers de mauvaises solutions), il peut être utilisé comme un outil exploratoire pour donner de bonnes valeurs initiales des probabilités m et u (dans le cas où d'autres modèles mieux ajustés sont utilisés) et enfin il attribue en général aux champs les plus discriminants les meilleurs scores. Cependant, contrairement à une approche supervisée, l'algorithme EM a tendance à donner un taux réel de faux positifs supérieur au taux de faux positifs autorisé. Belin (1990) montre sur un exemple qu'une probabilité de faux positifs tolérée de 0.1% donne un taux réel de faux positifs de 1.5%. Dans ce même article il montre que, pour un taux d'erreur de 0.5%, il doit autoriser au modèle un taux d'erreur théorique de 10^{-7}. Par ailleurs si l'hypothèse d'indépendance est violée, les estimations sont alors très biaisées et nécessitent une correction manuelle, a posteriori la plupart du temps (Thibaudeau, 1993; Winkler, 1993). L'algorithme EM classique souffre d'une dépendance aux valeurs initiales et les estimations successives peuvent tendre vers des points qui ne sont au mieux que des points stationnaires de la vraisemblance qui éventuellement des extremums locaux ou des points selles (en théorie seulement en pratique il est difficile de converger vers ces points). L'algorithme peut donc converger vers des points qui posent des problèmes d'interprétation puisqu'il y a de multiples résultats possibles qui n'ont pas forcément les bonnes propriétés. Enfin l'algorithme EM pourrait souffrir d'une vitesse de convergence relativement lente (sur un volume de données important). L'expérience accumulée dans Jaro (1989) et Winkler (1990) montre qu'en pratique et sous certaines conditions l'algorithme EM est robuste aux conditions initiales et la convergence peut être très rapide. L'algorithme EM a connu plusieurs modifications au cours du temps, qui ont permis d'améliorer la convergence et la précision des résultats, mais aussi qui en ont fait un outil facilement utilisable dans les extensions du modèle de Fellegi et Sunter présentées ci-dessous. En effet lorsque la vraisemblance est très compliquée et/ou le volume de données important, l'algorithme doit faire face à des problèmes computationnels liés à l'usage d'algorithme d'optimisation lors de la phase de maximisation. Certaines évolutions répondent à ces problèmes d'estimation comme l'algorithme ECM (Meng et Rubin, 1993) qui propose de remplacer l'unique étape "M" de l'algorithme EM par une succession d'étape "M" techniquement plus simple à traiter. Ce dernier est très utile lorsque l'on veut manipuler des modèles latents sans hypothèse d'indépendance conditionnelle de type log-linéaire (Winkler, 1993). Deux revues complètes sur les algorithmes EM présentent les points forts de chacun d'entre eux : McLachlan et Krishnan (2008), Meng et Van Dyk (1997) et Foulley (2002).

2.3.4.3. Estimation sans biais des paramètres de (1) : Jaro (1989, 1995) propose plusieurs développements pratiques pour réaliser le couplage, en commençant par expliciter l'estimation des paramètres. En effet nous avons vu que les méthodes de blocage biaisent l'estimation des paramètres $(u_i)_i$, Jaro remarque que la proportion de paires correspondant à un unique individu est très faible en comparaison des autres paires, autrement dit $U \approx E_A \times E_B$. Ainsi il est possible d'approximer les paramètres $(u_i)_i$ directement en comparant aléatoirement les lignes de E_A et de E_B. L'algorithme EM ne sert alors qu'à estimer les paramètres $(m_i)_i$ et π.

 ISSN: 2102-6238

2.3.4.4. Calcul non supervisé des occurrences et du pouvoir discriminant d'une modalité : Lorsque Fellegi et Sunter (1969) propose la formule (4), ils utilisent en fait des probabilités intermédiaires $m(\delta_i = 1, v_i(a) = t_i)$, $1 \leq i \leq K$. La somme sur t de ces probabilités donnant $m(\delta_i = 1) = m_i$, de même pour le calcul de $m(\delta_i = 1) = m_i$. Il est possible de calculer le pouvoir discriminant des modalités de chacune des variables en calculant les occurrences $O(i)_{E_A \cap E_B, t}$, $O(i)_{E_A, t}$ et $O(i)_{E_B, t_i}$. $O(i)_{E_A \cap E_B, t_i}$ est le nombre de couples pour lesquels $\delta_i = 1$ et $v_i(a) = t$. Alors, si l'on connait la classe M sur un exemple, $\hat{m}(\delta_i = 1, v_i(a) = t_i) = \dfrac{O(i)_{E_A \cap E_B, t_i}}{\#M}$. Dans le cas de l'appariement non supervisé, Jaro (1995) propose de réaliser un premier appariement, après avoir initialisé les paramètres m et u du modèle de Fellegi et Sunter (1969). Puis sur les classes $\widetilde{M}$ et $\widetilde{U}$, on estime les probabilités $m(\delta_i = 1, v_i(a) = t_i)$ et $u(\delta_i = 1, v_i(a) = t_i)$ pour tout $1 \leq i \leq K$. Enfin on recalcule les probabilités m et u, puis on réalise un nouveau couplage et ainsi de suite jusqu'à la convergence des paramètres m et u. Cet algorithme est utilisé par Fournel *et al.* (2009) et Quantin *et al.* (2009).

2.3.5. Extension du modèle Fellegi et Sunter

Le modèle de Fellegi et Sunter, avec l'hypothèse d'indépendance conditionnelle, s'apparente à une classe bien connue de modèles en apprentissage les modèles de Bayes "simple" ou "naïf", c'est-à-dire que la loi conditionnelle du vecteur de concordance sachant la classe à laquelle appartient le couple (a,b) d'entrée correspond à la loi d'un vecteur aléatoire dont les coordonnées sont indépendantes. Cette hypothèse sous-entend par exemple pour un modèle contenant la variable sexe et la variable prénom que la concordance des sexes serait conditionnellement indépendante de la concordance des prénoms. Cette hypothèse n'est pas réaliste en pratique surtout quand des méthodes de réduction des comparaisons sont utilisées comme le blocage, qui crée des dépendances entre les variables de couplage, en particulier dans le groupe des paires d'individus différents (Thibaudeau, 1993). Si l'hypothèse d'indépendance conditionnelle est violée, cela peut entrainer un biais dans l'estimation des probabilités d'erreur. La proportion des erreurs est même une fonction croissante de la dépendance entre les variables (Kelley, 1986; Jaro, 1989; Thibaudeau, 1993). Thibaudeau (1993) propose de revenir sur l'hypothèse d'indépendance conditionnelle en utilisant un modèle log-linéaire pour la loi des u_i. La loi du couple (Δ, S) s'écrit alors :

$$\log\left(\mathbb{P}_\theta(\delta, s)\right) = \mu + \alpha_s + \sum_{i=1}^{K} \rho^i_{\delta_i} + \sum_{i=1}^{K} \varphi^i_{s,\delta_i} + (1-s)\left(\sum_{k=2}^{\sharp D} \sum_{i_1 < \ldots < i_k \in D} h^{i_1,\ldots,i_k}_{\delta_{i_1},\ldots,\delta_{i_k}}\right) \tag{11}$$

L'ensemble D correspond à l'ensemble des variables de concordances pour lesquelles une relation de dépendance a été détectée. Les paramètres ci-dessus doivent vérifier une liste de contraintes :

$$\begin{cases} \alpha_1 = -\alpha_0 \\ \rho^i_1 = -\rho^i_0 \\ \varphi^i_{1,\delta_i} = -\varphi^i_{0,\delta_i}, \varphi^i_{s,1} = -\varphi^i_{s,0} \\ h^{i_1,\ldots,i_k}_{\delta_{i_1},\ldots,0,\ldots,\delta_{i_k}} = -h^{i_1,\ldots,i_k}_{\delta_{i_1},\ldots,0,\ldots,\delta_{i_k}} \end{cases} \tag{12}$$

L'estimation des paramètres se fait en utilisant un algorithme de type EM proposé par Haberman dans Haberman (1979). Winkler (1993) propose un algorithme EM modifié pour des

 ISSN: 2102-6238

modèles log-linéaires, qui permet de prendre en compte des contraintes convexes (c'est à dire la restriction de l'espace des paramètres du modèle dans un domaine convexe) sans que celui-ci nécessite un effort algorithmique important. Larsen et Rubin (2001) introduisent un modèle général de mélange de lois de probabilité qui inclut les extensions précédentes. Ils proposent une procédure itérative de sélection de modèle notamment sur le nombre de composantes du mélange. Cette procédure est basée sur une vérification manuelle de certaines paires à "risque" qui va permettre de calibrer au mieux les composantes du mélange. On trouvera en annexe une estimation du taux d'appariés à tort. Winkler (2002) donne les conditions nécessaires pour l'utilisation efficace d'un algorithme EM dans le cadre non supervisé :

- L'ensemble M doit être suffisamment grand (il doit représenter au moins 5% de l'ensemble $E_A \times E_B$). Ceci est faux en général si l'on regarde l'ensemble des comparaisons possibles, sauf si l'estimation est réalisée sur un sous-ensemble de paires qui contient un nombre suffisant de paires dont le score de concordances sera très élevé.
- La classe des couples "identiques" doit être suffisamment séparée de la classe des couples d'entrées "différents", ce qui implique que le taux d'erreurs typographiques doit être relativement bas ou que ces erreurs doivent être compensées par la présence de variables redondantes.
- Dans le cas où l'on fait l'hypothèse d'indépendance conditionnelle, les catégories de variable de mélange doivent être bien choisies.

Copas et Hilton (1990) proposent une approche supervisée, en formulant le couplage en un test de Neyman-Pearson de rapport de vraisemblance à l'aide de modèles prenant en compte les modalités (et non les concordances) des variables indirectement identifiantes sur lesquelles une erreur à pu être introduite (connu sous le nom de "hit miss model"). Pour finir, Sadinle et Fienberg (2013) étendent le modèle Fellegi et Sunter (1969) pour les couplages multiples et Schürle (2005) démontre que l'algorithme EM appliqué au modèle saturé (avec dépendance conditionnelle), pour l'appariement entre deux bases de données, converge en une itération.

2.3.6. Calibration des poids et correction des seuils

Nous avons présenté ci-dessus quelques méthodes qui tentent d'améliorer le modèle de Fellegi et Sunter en introduisant directement des ajustements dans le modèle qui nécessite par ailleurs des algorithme EM permettant d'en estimer les paramètres. Nous proposons maintenant deux stratégies pour améliorer l'appariement. Contrairement aux précédentes elles n'essaient d'améliorer directement un modèle en prenant en compte des dépendances entre les erreurs mais tentent de corriger les outputs de cette dernière en : corrigeant les seuils de discrimination dans le modèle de Fellegi et Sunter, ou bien en calibrant les scores à des probabilités d'appartenance à une classe ou une autre correspondant plus au calcul de la valeur prédictive positive et la valeur prédictive négative. Les deux méthodes présentées ci-dessous sont supervisées, elles analysent les poids produits par un modèle de couplage qui peut être déterministe ou bien probabiliste. La première méthode (Rogot *et al.*, 1986) consiste à corriger les seuils de discrimination des poids. Les auteurs illustrent sur un exemple que les seuils choisis a priori vont provoquer dans un cas une surabondance de paires à vérifier manuellement du côté des paire de la classe U, et dans l'autre un excès de faux positif. Leur méthode de correction se fait en deux temps (Figure 2) : dans un premier temps un couplage est fait avec deux bases de données qui vont servir de bases d'appren-

 ISSN: 2102-6238

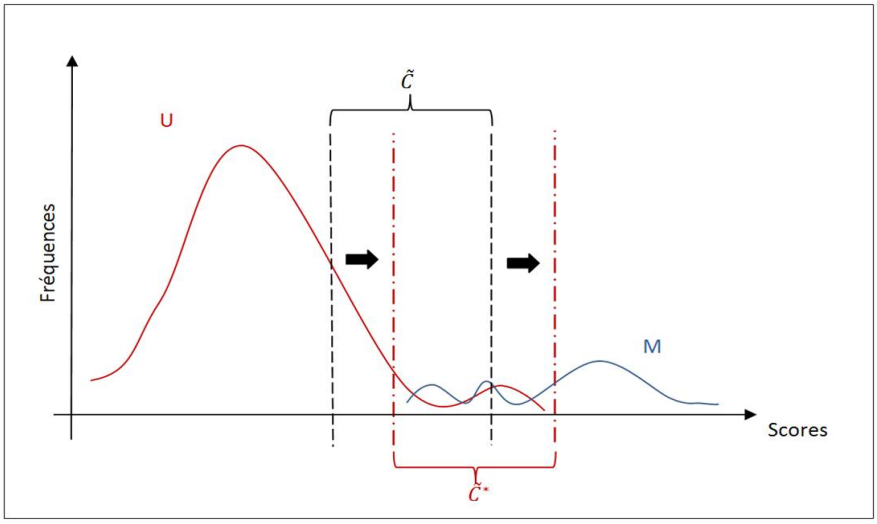

FIGURE 2: Schéma de la correction des seuils à l'issue du processus décrit dans Rogot *et al.* (1986)

tissage. Il s'agira alors de détecter, sur le résultat, les seuils qui encadrent la classe $\tilde{C}$, à savoir l'intervalle sur lequel se superposent les scores des deux classes. Enfin on utilisera un second jeu de données pour élargir cet intervalle afin d'être robuste aux fluctuations d'échantillonnage.

Une version non supervisée de cette méthode à été proposé dans Sariyar *et al.* (2011) utilisant la théorie des valeurs extrêmes. Elle consiste à estimer les seuils de la classe $\tilde{C}$, à savoir le plus haut score obtenu par une paire de la classe U pour la borne droite et le plus bas score de la classe M pour la borne de gauche. Cette dernière n'a pas fait l'objet d'une évaluation poussée sur des données réelle. L'autre méthode (Rubin et Belin, 1991; Belin et Rubin, 1995) propose de calibrer les scores avec des probabilités d'appartenance à l'une des deux classes ainsi que les seuils discriminants à des probabilités d'erreur. Plus précisément, cette approche consiste à modéliser par un mélange Gaussien la répartition des poids des paires préalablement normalisés par une transformation de Box-Cox (Box et Cox, 1964), puis de déterminer pour chaque seuil le taux de faux positifs où de faux négatifs. Le couplage pilote permet de choisir le bon paramétrage de la transformation de Box-Cox pour chaque classe M et U. L'algorithme EM estime ensuite les paramètres du mélange, sous la contrainte que le ratio des variances des deux composantes du mélange reste constant. On peut alors déterminer la probabilité d'être un faux positif ou faux négatif pour chaque paire. Plus précisément on considère la fonction T_{BC} suivante :

$$T_{BC}(w,\iota,\kappa) = \begin{cases} \dfrac{w^\kappa - 1}{\iota^{\kappa-1}\kappa}, \kappa \neq 0 \\ \iota \log(w), \kappa = 0 \end{cases} \tag{13}$$

Cette transformation, appelée transformation de Box-Cox, va permettre la normalisation des scores. Pour chaque classe M et U il faudra estimer les paramètres κ_M et κ_U respectivement, de même pour ι_M et ι_U qui correspondent respectivement aux moyennes géométriques des poids

 ISSN: 2102-6238

de la classe M et U. Une fois estimées les paramètres de la transformation [6] , on applique un algorithme EM pour estimer les paramètres du mélange sous la contrainte que le ratio des variances des deux composantes du mélange reste constant. Enfin en appliquant le théorème de Bayes on trouve l'estimation du taux de faux positifs pour un seuil τ. En posant $R(\tau,l) = \frac{T_{BC}(\tau,\hat{\iota}_l,\hat{\kappa}_l)-\hat{m}_l}{\hat{\sigma}_l}$:

$$\hat{\varepsilon}(\tau) = \frac{\hat{p}\left[1-F_{N(0,1)}\left(R(\tau,U)\right)\right]}{\hat{p}\left[1-F_{N(0,1)}\left(R(\tau,U)\right)\right]+(1-\hat{p})\left[1-F_{N(0,1)}\left(R(\tau,M)\right)\right]}$$

La fonction $F_{N(0,1)}$ est la fonction de répartition d'une loi normale centrée réduite. Les valeurs $(\hat{p},\hat{m}_U,\hat{\sigma}_U,\hat{m}_M,\hat{\sigma}_M)$ correspondent aux estimations des paramètres du mélange Gaussien. La formule ci-dessus donne la probabilité des appariements à tort pour chaque seuil choisi. Ainsi si l'on veut un appariement avec une erreur de 5%, on choisira comme seuil l'antécédent de cette erreur par la fonction $\hat{\varepsilon}$. Enfin la probabilité sachant le score d'être dans la classe U est donnée par :

$$\mathbb{P}(U|w) = \frac{\hat{p}\left[\frac{w}{\hat{\iota}_U}\right]^{\hat{\kappa}_U-1} e^{\frac{1}{2}(R(\tau,U))^2}}{\hat{p}\left[\frac{w}{\hat{\iota}_U}\right]^{\hat{\kappa}_U-1} e^{\frac{1}{2}(R(\tau,U))^2}+(1-\hat{p})\left[\frac{w}{\hat{\iota}_M}\right]^{\hat{\kappa}_M-1} e^{\frac{1}{2}(R(\tau,M))^2}}$$

2.3.7. L'analyse bayésienne du couplage

Il existe une approche purement bayésienne du couplage de données. En effet, la plupart des méthodes précédentes privilégient l'usage des modèles de mélange avec l'emploi de variables latentes et de la règle de Bayes pour la règle de décision sans proposer de loi à priori sur les paramètres. Dans Fortini *et al.* (2001), les auteurs proposent une analyse bayésienne du modèle de Fellegi et Sunter avec un a priori de Dirichlet pour les paramètre m et u, où cette fois le statut des paires est devenu un paramètre du modèle à estimer. L'objectif est d'obtenir des estimations de la probabilité de concordance sachant les observations plus précises et d'inclure de façon endogène à la vraisemblance l'unicité des couplages (celle-ci étant traitée à posteriori dans le modèle de Fellegi et Sunter par des algorithmes d'allocation linéaire). Les auteurs proposent trois approches dont la principale utilise des algorithmes d'Hasting-Metropolis et du recuit simulé (Duflo, 1997) pour déterminer les modes de la loi a posteriori de la matrice des configurations des couplages (matrice dont les lignes représentent les individus de la base E_A et les colonnes les individus de la base E_B, cette matrice contient des 1 et des 0, les 1 indiquant que l'individu de la base E_A correspond à l'individu de la base E_B) entre les bases de données (ce n'est plus l'algorithme EM qui est utilisé). On notera dans cette approche que l'hypothèse d'indépendance conditionnelle de la concordance des champs n'est pas nécessaire, de même pour l'hypothèse sur l'indépendance des paires. Dans Larsen (2005), un modèle de Bayes hiérarchique est utilisé afin de produire des

[6] La définition de transformation de Box-Cox donnée ci-dessus fait l'hypothèse que les scores sont positifs. Ce qui n'est pas le cas en général avec le modèle de Fellegi-Sunter. Dans sa forme la plus générale la transformation de Cox-Box propose un décalage pour rendre les scores positifs.

 ISSN: 2102-6238

estimations par bloc de données et prendre en compte les variations issues de la position géographique. Enfin McGlincy (2004) propose une version bayésienne de la méthode fréquentielle de Fellegi et Sunter (équation (4)), par imputation multiple des probabilités d'erreur. Winkler (2002, 2000) s'inspire des travaux de Nigam *et al.* (2000). Ces derniers proposent une approche semi-supervisée de la discrimination textuelle dans le cas où il est difficile d'obtenir des données étiquetées. Cette méthode de discrimination incluant à la fois des données étiquetées et non étiquetées démontrent qu'un apprentissage réalisé de cette façon donne de meilleurs résultats que les approches plus classiques de l'apprentissage supervisé avec peu de données étiquetées ou bien qu'un apprentissage totalement non supervisé. En effet, dans le cas non supervisé, lors de la phase d'estimation, la présence de données étiquetées limite l'éventail des valeurs que peuvent prendre les paramètres estimés. Comme pour Fellegi et Sunter, les auteurs emploient un modèle de réseau bayésien, un mélange de lois multinomiales, et appliquent sur ce modèle un algorithme EM. L'algorithme produit dans un premier temps un système de discrimination supervisé sur les données ad hoc qui vont servir à initialiser l'algorithme EM. Les auteurs apportent deux innovations : l'ajout dans la vraisemblance d'un paramètre λ qui module l'importance à donner aux informations non étiquetées et la possibilité de ne pas faire correspondre les classes d'intérêt avec les composantes du mélanges ce qui entraîne un ajustement de l'étape "E" dans l'algorithme EM. Le choix du paramètre λ et le choix des composantes du mélange qui devront constituer les ensembles M et U se fait en utilisant une méthode de validation croisée. L'avantage principal de ce modèle provient du fait qu'il capture un peu mieux les dépendances existantes entre les variables. Il est possible d'inclure, comme dans le cas non supervisé, des structures de dépendance via des modèles log-linéaires. Finalement Winkler (2007) adapte cette procédure au cas non supervisé en étiquetant de façon floue un jeu de données étiquetées en assignant à certaines paires le statut de "pseudo-apparié" et de "pseudo-non apparié". Plus précisément on désigne une paire comme un couple selon que le score, qu'on aura choisi au préalable est au-dessus d'un seuil élevé, dans le cas contraire on ne couplera pas la paire.

$$\begin{aligned}\log\left(L(\delta,s,\theta)\right) = \log\left(\mathbb{P}_{Dir}(\theta)\right) &+ \lambda \sum_{i\in D_l}\sum_{j=1}^{J} s_{i,j}\log\left(\mathbb{P}_\theta(\delta_i|C_j)\mathbb{P}_\theta(C_j)\right) \\ &+ (1-\lambda)\sum_{i\in D_l}\sum_{j=1}^{J}\log\left(\mathbb{P}_\theta(\delta_i|C_j)\mathbb{P}_\theta(C_j)\right)\end{aligned} \tag{14}$$

La loi $\mathbb{P}_{Dir}(\theta)$ est une loi à priori de Dirichlet, l'ensemble D_l correspond aux paires étiquetées et l'ensemble D_u correspond aux paires sans étiquette, le paramètre $\lambda \in [0,1]$ sert à moduler entre une approche complètement supervisée ($\lambda = 1$) et une approche complètement non supervisée ($\lambda = 0$). Les classes $(C_j)_{1\leq j\leq J}$ ne correspondent pas aux classes M et U, ce qui permet de compenser l'hypothèse d'indépendance conditionnelle. Ce n'est qu'a posteriori qu'il sera décidé quelles classes correspondent à M et quelles classes correspondent à U. Le choix des composantes du mélange qui devront constituer l'ensemble M et U se fait par la validation croisée. Comme dans le modèle de Fellegi et Sunter (1969) les concordances sont indépendantes conditionnellement à la classe C_j.

Les approches les plus récentes sur le couplage bayesien, inspirées de Copas et Hilton (1990) et de la discrimination collectiviste, utilisent un modèle bayésien hiérarchique qui permet de

 ISSN: 2102-6238

gérer à la fois le couplage et la détection des doublons en appariant en même temps sur plusieurs base de données (Steorts *et al.*, 2013, 2014a). Celle-ci est permise par l'algorithme SMERED (Split and MErge REcord linkage and De-duplication) qui est l'interprétation d'un algorithme Split-Merge, un algorithme mixte de chaîne de Markov adapté au modèle de mélange dont l'a priori est un processus de Dirichlet (Jain et Neal, 2004). Enfin ce modèle est amélioré par des méthodes bayésiennes empiriques pour éviter de spécifier les lois a priori et pour étendre le modèle aux variables dont les modalités sont des chaînes de caractères (Steorts, 2015).

$$
\begin{aligned}
v_{ijl}|\phi_{ij}, \overline{v}_{\varphi_{ij}l}, k_{ijl}, \theta_l &\sim \begin{cases} 1_{\overline{v}_{\phi_{ij}l}} & , \text{if } k_{ijl} = 0 \\ \text{Multinomiale}(1, \theta_l) & , \text{if } k_{ijl} = 1 \end{cases} \\
k_{ijl} &\sim \text{Bernoulli}(\lambda_l) \\
\overline{v}_{hl}|\theta_{hl} &\sim \text{Multinomiale}(1, \theta_l) \\
\theta_l &\sim \text{Dirichlet}(\vartheta_l) \\
\lambda_l &\sim \text{Beta}(o_l, p_l) \\
\pi\left[(\phi_{ij})_{ij}\right] &\propto 1
\end{aligned}
\tag{15}
$$

La variable v_{ijl} correspond à la l-ième variable d'appariement reportée dans la base i pour le j-ième enregistrement. k_{ijl} est la variable binaire qui indique si v_{ijl} a été reportée avec une erreur ou non. $\overline{v}_{hl}$ est la vraie valeur de la variable l pour l'individu h et enfin $(\phi_{ij})_{ij}$ est la matrice qui indique l'identité de l'individu qui est associé à l'enregistrement j dans la base i. A la différence des modèles probabilistes présentés précédemment, les appariements se font entre des enregistrements et un individu et non entre des enregistrements. Cela permet de gérer à la fois appariement et recherche des doublons. Cependant, le nombre de paramètres du modèle augmente avec la taille de la population latente, ce qui limite l'efficacité d'un échantillonneur de Gibbs et justifie l'usage du SMERED.

3. Exploitation des données appariées

Nous avons présenté un ensemble de méthodes et de modèles pour l'appariement de base de données, mais aucune de ces méthodes ne peut produire un résultat parfait. Leur performance dépend de plusieurs facteurs, dont les principaux sont la disponibilité de variables de couplage, ainsi que la qualité de ces variables. Pour éviter d'avoir un nombre de faux positifs important, certains statisticiens préfèrent des stratégies conservatrices qui réduisent effectivement le nombre de faux positifs mais en dégradant la sensibilité, pouvant accroître les biais de sélection. D'autres au contraire conserveront le plus de couples possibles au prix d'une augmentation des faux positifs et donc une augmentation des biais de classement.

3.1. Le bruit lié aux couplages

Neter *et al.* (1965) montre sur un exemple que des erreurs de couplage peuvent provoquer des biais très importants sur les analyses statistiques. Dans le cas de la régression linéaire simple et

 ISSN: 2102-6238

quand l'appariement est complet, il est facile de calculer le biais induit par les erreurs de couplage. Considérons par exemple le modèle $Y = X\beta + \varepsilon$ où $\varepsilon \sim N(0, \sigma^2)$ où la variable dépendante est incluse dans une base de données et les variables explicatives dans une autre. Lorsque nous apparions ces bases, nous observons le couple $((X_i, Z_i))_i$ où Z_i est la variable :

$$Z_i = \begin{cases} Y_i, \text{ avec une probabilité } p_{ii} \\ Y_k, \text{ avec une probabilité } p_{ik} \end{cases} \tag{16}$$

Sans prise en compte de l'erreur de couplage l'estimation de β donne $\hat{\beta} = (X^tX)^{-1}X^tZ$ dont l'espérance donne :

$$\begin{aligned} E(\hat{\beta}|X) &= E((X^tX)^{-1}X^tZ|X) && (17)\\ &= E((X^tX)^{-1}X^tY|X) + E((X^tX)^{-1}X^t(Z-Y)|X) && (18)\\ &= \beta + (X^tX)^{-1}X^t(PX\beta - X\beta) && (19)\\ &= \beta + \text{biais} && (20) \end{aligned}$$

Où la matrice P est la matrice contenant les probabilités de permutation entre l'allocation issue de l'appariement et la vraie allocation, induit par le processus d'appariement. On remarquera que le biais dont il est question dans le calcul ci-dessus, est un biais généré par la permutation d'au moins deux individus. Ce qui correspond presque à une généralisation du biais de classement aux données continues, a ceci près que le mode de production de la donnée n'empêche pas l'exploitation de la bonne donnée. Elle pourrait même permettre l'exploitation simultanée de plusieurs candidats possibles, ce qui correspond au contenu des méthodes présentées ci-dessous.

Le biais présenté ci-dessus n'inclus pas les biais de sélection induits par l'algorithme d'appariement, notamment les biais dûs au choix des variables d'appariement sélectionnées ainsi que de leur proxy. Par exemple si l'on cherche à mettre en évidence un lien entre un événement de santé et un décès, en couplant par exemple la base des causes médicales de décès avec les données du SNIIRAM (Lamarche-Vadel *et al.*, 2013), il est tentant, pour récupérer les dates de décès dans le SNIIRAM, d'utiliser la date de sortie de l'hôpital avec le mode de sortie "Par décès". Ce qui implique que les moyens pour retrouver la date de décès sont plus fiables pour les individus décédés à l'hôpital (par ailleurs il est possible dans ce cas de retrouver le département de décès ce qui n'est pas le cas des décès en dehors de ces établissements). Le risque est alors d'avoir dans les appariés une sur-représentation de cette population. Réaliser un appariement notamment en choisissant mal le seuil d'acceptation pourrait renforcer artificiellement le lien entre le décès et un évènement de santé. Dans l'exemple de la base AMPHI qui sera présenté plus loin, lorsqu'on cherche à coupler les causes de décès avec les données du SNIIRAM, on constate que les populations vivant à l'étranger ou ayant une origine étrangère sont les plus difficiles à apparier. Donc si l'on cherche à étudier les inégalités sociales de santé il faut avoir à l'esprit qu'il y aura un biais pour cette population. On trouvera dans Ford *et al.* (2006) un exemple d'appariement dans l'état de la Nouvelle-Galles du Sud, Australie, dans lequel la base des résumés de sortie (Inpatient Statistics Collection, ISC) est couplée à la base périnatale (Midwives Data Collection, MDC) couvrant toutes les naissances après 20 semaines de gestation. Ces bases contiennent des informations relatives à l'enfant et à la mère. L'appariement se fait entre mères d'un côté et entre enfants de l'autre et concerne des naissances allant du 1 Janvier 2000 au 31 Décembre 2002.

 ISSN: 2102-6238

Les appariés correspondent à 98.8% de la base MDC et 99% de la base ISC. On constate que dans la population des non appariés du côté des mères on retrouvera une sur-représentation des femmes susceptibles d'avoir ou d'avoir eu des grossesses à risque, on trouvera aussi une sur-représentation des naissances par césarienne, de morts nés etc. De même du côté des nouveaux nés, on trouvera une sur-représentations de prématurés et des décès hospitaliers. Ainsi un taux d'appariement très élevé n'élude pas la question de la présence de biais dans l'analyse statistique. Bohensky *et al.* (2010) réalisent une analyse bibliographie sur les appariements et mettent en avant un certain nombre d'articles pour lesquels il a été constaté un lien existant entre le résultat de l'appariement et les caractéristiques des individus (sexe, âge, groupe, situation économique et sociale, situation géographique, santé). Il est important de noter que même un appariement avec le NIR n'évite pas ce genre de situation. Legleye *et al.* (2017) montrent que la récupération du NIR lors d'enquête n'est pas systématique. En fait ils démontrent au travers d'une enquête que les répondants sont plus favorables à fournir leurs coordonnées de naissance (le nom, le prénom et la date de naissance) que leur NIR avec des disparités en fonction du sexe ; les hommes sont plus favorables à fournir des informations directement identifiantes que les femmes. Que le taux d'acceptation pour fournir des informations directement identifiantes varie significativement en fonction de l'âge. Enfin les personnes les plus diplômés et que les cadres sont plus favorables à fournir leur NIR. Au final, les méthodes présentées ci-dessous ne permettent pas de faire l'économie d'une description des bases à apparier, du processus d'appariement (règles de décisions, choix des variables, etc...) afin de prendre en compte ces informations dans l'analyse statistique.

3.2. L'analyse statistique sur des données appariées

En épidémiologie et en santé publique où l'on cherche souvent à mesurer un risque dans le but de mettre en évidence une association, il existe 3 pratiques assez répandues pour prendre en compte l'incertitude dans un couplage. Idéalement on pourrait disposer d'un jeu de données constitué du couplage de deux bases de données représentatives en utilisant un identifiant direct comme le NIR par exemple (ce type de jeux données est souvent appelé un gold standard), ce qui permettrait de calculer des valeurs prédictives, ou bien de calibrer des scores et de corriger les biais (Lash *et al.*, 2009). Ce point a été traité plus haut, cependant nous tenons à rappeler qu'il est difficile d'avoir accès à ce type de données pour des raisons de coût et pour des raisons réglementaires. Une seconde approche va consister à observer la distribution des individus appariés et de la comparer à la distribution des individus non appariés, cette méthode permet de voir si le fait d'être apparié ou non est lié au facteur d'exposition ou bien à la variable réponse. Cependant pour détecter les non appariés il est nécessaire qu'au moins une des deux bases contienne l'autre. Enfin la troisième approche, la plus connue, consiste à réaliser une analyse de sensibilité afin de déterminer le sens des biais lorsqu'on décide d'inclure dans la classe $\widetilde{M}$ des cas très ambigus en faisant varier le seuil par exemple. Pour réaliser une analyse de sensibilité il est nécessaire d'avoir une information sur l'amplitude du couplage (comme le score dans la proposition de Fellegi et Sunter (1969)). En effet le résultat d'un appariement n'est pas nécessairement binaire, avec des individus appariés correctement et des non trouvés (ce qui correspond à une situation de données manquantes). En faisant varier des seuils d'appariement, il est possible de mesurer comment les individus couplés à tort impactent les résultats. On trouvera dans Harron *et al.* (2017) une analyse comparative de ces trois approches.

 ISSN: 2102-6238

Dans la continuité de Neter *et al.* (1965), d'autres travaux se sont intéressés à l'utilisation des données chaînées dans la production d'études scientifiques et l'estimation des biais dus aux erreurs de couplage. Scheuren et Winkler (1993) proposent une analyse de l'effet de la présence des faux positifs dans le cadre d'une analyse linéaire ordinaire. Les auteurs proposent un ajustement dans le cas de la régression linéaire à condition de connaitre les probabilités de permutation entre deux individus donnés, ainsi qu'une évaluation de cet ajustement. En réalité il est peu réaliste d'estimer la matrice des permutations, les auteurs proposent de conserver les couplages pour lesquels la probabilité d'être issu du même individu est la plus élevée. En pratique il est difficile, et parfois impossible, d'estimer ces probabilités de permutation. On utilisera plutôt $p_{ik} = P(M|\delta(a_i, b_k))$ et notamment celles fournies par l'estimation dans Belin et Rubin (1995). Dans la continuité, Lahiri et Larsen (2005) proposent une méthode applicable à des modèles linéaires pour analyser des données appariées selon la méthode proposée dans Larsen et Rubin (2001). Ils obtiennent un ajustement non biaisé et plus performant que la correction proposée par Scheuren et Winkler (1993). Ils proposent deux méthodes pour calculer la variance de cet estimateur. La première suppose que les paramètres du modèle de mélange utilisés pour le couplage sont connus. Dans ce cas l'estimation de la variance est explicite. Dans le cas contraire, on peut estimer cette variance en utilisant un bootstrap paramétrique. Chambers (2009), Kim et Chambers (2012a,b) présentent une extension du modèle de Lahiri et Larsen aux équations d'estimation (Estimating Equation) incluant les modèles linéaires et la régression logistique. Cette formulation permet de prendre en compte le cas des couplages multiples (pour la production de données longitudinales par exemple), et le cas des appariements incomplets.

Par ailleurs, Hof et Zwinderman (2012) proposent plusieurs améliorations à l'estimation proposée par Lahiri et Larsen (2005). Premièrement ils en proposent une extension directe dans le cas où les covariables se trouvent dans des bases différentes et à la condition que la base qui contient la variable à expliquer soit une base exhaustive pour les autres ou, autrement dit, que toutes les bases contenant uniquement des covariables soient un échantillon de la base qui contient la variable expliquée. Ensuite, les auteurs reformulent la régression linéaire sur des données appariées comme une simple régression linéaire pondérée sans contrainte préalable comme l'inclusion d'une des bases dans l'autre. Ce résultat fournit une estimation toujours biaisée dans la plupart des couplages incomplets (situation qui est en dehors du spectre de Lahiri et Larsen (2005)). Cependant, ce cadre théorique permet des manipulations simples pour calibrer l'analyse statistique afin d'en améliorer la performance sur la base de logiciels existants et donc n'impliquant pas de développements spécifiques. Cette formulation s'adapte très facilement aux régressions logistiques. On trouvera une analyse des performances dans Hof et Zwinderman (2012) de ces ajustements en fonction de différents scénarios. On notera que dans l'analyse comparative, les méthodes incluant une phase d'imputation ont les biais les plus faibles. En effet les analyses faites sur des données simulées montrent qu'il vaut mieux ne pas forcer les appariements et retirer les appariés à tort. Ceci est très difficile sans hypothèses complémentaires et donc sans connaissance sur la distribution des données.

Enfin Hof et Zwinderman (2015) proposent une méthode d'estimation dans sa formulation la plus générale (il n'inclut pas les données longitudinales produites par des appariements multiples ou successifs et les données avec une structure hiérarchique) dans le cadre du maximum de vraisemblance. Ils simplifient le modèle en justifiant la validité de la pseudo vraisemblance retenue, à savoir que l'estimation par l'algorithme EM des coordonnées de θ qui jouent un rôle

 ISSN: 2102-6238

dans le terme d'intérêt $L(y_b|\theta,x_a,S_{ab}=1)$ est identique pour la vraisemblance et la pseudo vraisemblance. Il s'agira alors de maximiser la pseudo vraisemblance :

$$\begin{aligned} l(Y,X,[\Delta]|\theta) &= \sum_{a\in E_A}\sum_{b\in E_B}\{\log[L(y_b|\theta,x_a,S_{ab}=1)]+\log[L(x_a|\theta,S_{ab}=1)] \\ &\quad +\log[\varpi_{ab1}]\}\,\mathbb{P}(S_{ab}=1|y_b,\theta,x_a,\delta(a,b))+\{\log[L(y_b|\theta,S_{ab}=0)] \\ &\quad +\log[L(x_a|\theta,S_{ab}=0)]+\log[\varpi_{ab0}]\}\,\mathbb{P}(S_{ab}=0|y_b,\theta,x_a,\delta(a,b)) \end{aligned} \quad (21)$$

Où Y est la variable à expliquer, X est l'ensemble des covariables, $[\Delta]=(\delta(a,b))_{ab}$ la matrice des vecteurs de comparaisons entre les individus des deux bases, S la matrice qui contient le vrai statut des paires. La simplification du modèle se fait sous les hypothèses suivantes :

- Y et X sont indépendantes conditionnellement à l'évènement $\{S=0\}$. Autrement dit, il n'existe aucun lien entre ces deux variables pour les paires de l'ensemble U.
- Le couple (Y,X) est indépendant des variables de couplages conditionnellement à S. Autrement dit, connaissant le statut de la paire, la concordance ou la discordance des variables de couplage ne contient pas d'information sur le couple (Y,X).
- $\varpi_{abk}=L(\delta(a,b)|\theta,S_{ab}=k)\mathbb{P}(S_{ab}=k)$ pour $k=0,1$.

Les auteurs montrent que la maximisation de cette pseudo vraisemblance peut encore être simplifiée en réalisant l'estimation des paramètres en deux étapes : en appliquant dans un premier temps le modèle de Fellegi et Sunter pour estimer les paramètres ϖ_{abk}, puis en les intégrant dans la pseudo vraisemblance ci-dessus pour estimer les autres paramètres en utilisant un algorithme EM sur l'équation (21).

Utiliser le bootstrap pour estimer la variance est souvent difficile sur des modèles aussi complexes et des bases de données aussi volumineuses. On utilisera plutôt l'inverse de la matrice Hessienne de la pseudo log-vraisemblance. Cette approche est intéressante dans le sens où elle se décline pour beaucoup de modèles, par ailleurs elle permet également de traiter le cas où les covariables sont réparties sur plusieurs bases. Cependant les tests réaliser sur données simulées avec différents scénarios montrent qu'il n'est pas efficace de considérer l'ensemble des paires.

Le principe proposé par Goldstein *et al.* (2012) consiste à faire l'appariement en deux temps. D'abord réaliser un appariement du type de Fellegi et Sunter afin de déterminer les appariés les plus sûrs. Il ne reste alors que les non trouvés et les appariements multivoques. Les auteurs proposent de considérer ces derniers comme des données manquantes. Plus précisément, si le couplage sert à enrichir la base E_A avec une ou plusieurs variables de la base E_B, une fois que l'on a ajouté dans la base les informations trouvées issues des appariements dit sans équivoque ou sûrs, nous pouvons appliquer à la base E_A les méthodes d'imputations multiples. Il s'agira d'appliquer la méthode classique pour les non trouvés et d'appliquer une méthode d'imputation multiple pour laquelle l'imputation dans la base E_A se fait en utilisant une loi a priori qui est dépendante des poids du couplage pour les appariements multivoques. On suppose dans cette approche que la base E_B, est exhaustive et contient l'ensemble des individus de la base E_A que nous cherchons à apparier. Les non trouvés de la base E_A sont les individus pour lesquels aucune paire constituée avec la base E_B n'a obtenu de score suffisamment élevé. La méthode d'imputation fait l'hypothèse de normalité (au sens de suivre la loi normale) sur l'ensemble des variables. Goldstein *et al.* (2009) proposent d'étendre la méthode de Schafer (1997), en appliquant une transformation de Box-Cox, ainsi qu'une transformation analogue pour les données catégorielles ou discrètes dans

 ISSN: 2102-6238

le cas où les données ne sont clairement pas gaussiennes. Ils intègrent par ailleurs l'information contenue dans les poids du couplage probabiliste, en les convertissant en loi a priori. Une proposition de loi a priori pour le couplage d'un individu $a \in E_A$ avec un individu $b \in E_B$ peut avoir l'expression suivante :

$$p_{a,b} = \frac{\mathbb{P}(M|w(\delta(a,b))}{\sum_{a\in E_A, b\in E_B} \mathbb{P}(M|w(\delta(a,b))} \tag{22}$$

Une autre proposition consiste à utiliser directement les poids une fois que ces derniers auront été rendus positifs. La loi utilisée pour l'imputation multiple des couplages est alors

$$v_{a,b}(y_{E_B}) \propto f(y_{E_B}|y_{E_A})p_{a,b} \tag{23}$$

Harron *et al.* (2014) utilisent cette méthode et mettent en évidence son efficacité à condition d'estimer la loi a priori en utilisant un gold standard et de bien choisir un seuil d'acception des paires à prendre en compte dans la formule (22) et (23). Il reste toutefois à étudier la sensibilité de l'imputation à la loi a priori et au biais d'estimation de ses paramètres. Enfin, comme nous l'avons vu plus haut, il existe une approche bayésienne du couplage (Fortini *et al.*, 2001). Il est par ailleurs possible de propager l'incertitude liée au couplage à l'analyse statistique. Tancredi et Liseo (2011b) proposent un modèle bayésien intégrant à la fois le couplage et l'inférence statistique de telle sorte que l'inférence prenne en compte l'incertitude sur le couplage. Ils proposent également une ré-interprétation du problème de régression sur des données appariées en un problème de sélection de modèle. Dans le même esprit, Tancredi et Liseo (2011a) proposent un modèle hiérarchique pour l'estimation de la taille d'une population donnée en exploitant les informations observées sur les variables de couplage et intégrant les modèles de capture-recapture.

4. Applications

Nous présenterons ci-dessous trois applications d'appariement, la première, réalisée au sein de la CNAMTS est un couplage déterministe basé sur des règles de décision simples. Les deux autres utilisent un appariement probabiliste qui est une application du modèle de Fellegi et Sunter pour laquelle des aménagements ont été faits afin de préserver l'anonymat des individus des bases concernées. Ces trois exemples ont été réalisés pour démontrer l'applicabilité de ces méthodes, il n'y aura pas d'analyses comparatives pour démontrer l'efficacité de l'appariement probabilistes par rapport à une règle déterministe. Par ailleurs ces études ayant été réalisées indépendamment et dans des conditions différentes, les mesures de performances ne sont pas les mêmes selon les exemples. À notre connaissance ces exemples sont représentatifs de la pratique des appariements indirects réalisés en France. Par ailleurs ces exemples impliquent des bases de données qui sont souvent sollicitées pour compléter des informations individuelles, comme le statut vital, les causes multiples de décès, la prise d'un médicament ou encore un acte médical. La réglementation française et les institutions comme la CNIL encouragent l'utilisation d'un tiers de confiance pour la réalisation de l'appariement afin de séparer les variables identifiantes des variables d'intérêt contenant des informations sensibles. Ce tiers de confiance peut être un organisme au sein duquel seulement quelques personnes sont habilitées à avoir accès aux données de couplage et peuvent donc réaliser l'appariement. Ce genre de montage complique l'étape de vérification des résultats comme le montrent les exemples ci-dessous.

 ISSN: 2102-6238

Suite au processus de couplage, certaines paires sont appariés de manière certaine, d'autre sont " à valider " en consolidant grâce à des informations annexes (provenant de variables qui ne sont pas utilisées pour le chainage). Selon le contexte de travail, on est amené à se poser les questions suivantes :

- Souhaite-t-on avoir un couplage complètement exhaustif ? En effet un défaut minime d'exhaustivité peut être tolérable à des fins descriptives, mais sera non tolérable dans d'autres contextes.
- Quelles informations annexes a-t-on à disposition ? Sous quelle forme ? informatiques ou papier ?
- Quels moyens sont disponibles pour le processus de vérification ? En effet il faut mettre en vis-à-vis du nombre de situations qui nécessitent une vérification.

En aval du couplage on peut se demander s'il est nécessaire d'appliquer un traitement orthographique. Comme nous l'avons dit plus haut, cela va dépendre de la qualité des données (saisie manuelle, lecture automatique à partir d'une carte vitale). Sur des données de moindre qualité, cela permet de rattraper des erreurs typographiques et donc faciliter l'appariement. Sur des données " parfaites " il n'y aura pas de gain, au contraire un risque de générer des faux positifs.

Ce qui soulève un autre problème en rapport par le contexte de travail. Si l'on est autorisé à travailler sur des données en clair on peut appliquer différents calculs d'indice de similarité entre les chaînes de caractères comparées. Si l'on confie à un tiers les deux bases anonymisées au préalable, on ne peut plus faire ces calculs. On pourrait décomposer des chaînes de caractères en N-uplets avant anonymisation , permettant ainsi de connaitre le nombre de N-uplets communs lors de la phase de comparaison.

Au final, cela reste difficile de fournir une recette clef en main car certains éléments vont dépendre du contexte de l'étude. Le processus d'appariement qui a été suivi dans les trois exemples est synthétisé dans la figure 3. Les deux exemples ci-dessous utilisant le même modèle probabiliste ont utilisé le même algorithme algorithme d'apprentissage et le même algorithme de classification ont été décrits par Jaro (1995) et ont été rappelé précédemment. L'algorithme d'apprentissage itératif est celui présenté au paragraphe 2.3.4.4. Après l'estimation des paramètres, les scores sont calculés selon Fellegi et Sunter (1969) et puis arrive l'étape de résolution du problème d'allocation linéaire formuler en (3) pour éliminer les appariements multivoques. En revanche, comme nous l'avons déjà dit, l'étape de validation est très dépendante des conditions du couplages.

En revanche ce qui est dépendant du contexte sont le nettoyage et la standardisation des données, les étapes de blocage, et enfin les vérifications itératifs à l'issue des étapes de classifications. L'un des objectif de cette partie est d'illustrer ces différentes étapes afin de bien se rendre compte des moyens à mobiliser pour réaliser un appariement.

4.1. Le projet AMPHI pour le détermination d'une mesure de la qualité des soins

Ce couplage a été fait dans le cadre d'un projet sur la mise en place d'indicateurs de mortalité pour les établissements de santé. Le projet AMPHI (Lamarche-Vadel *et al.*, 2013) a exigé le couplage de la BCMD avec une extraction du SNIIRAM contenant les séjours du Programme de Médicalisation des Systèmes d'Information en Médecine, Chirurgie obstétrique et Odontologie (PMSI-MCO) de 2008-2009 des bénéficiaires du Régime Général (RG) hors Section Locale Mu-

 ISSN: 2102-6238

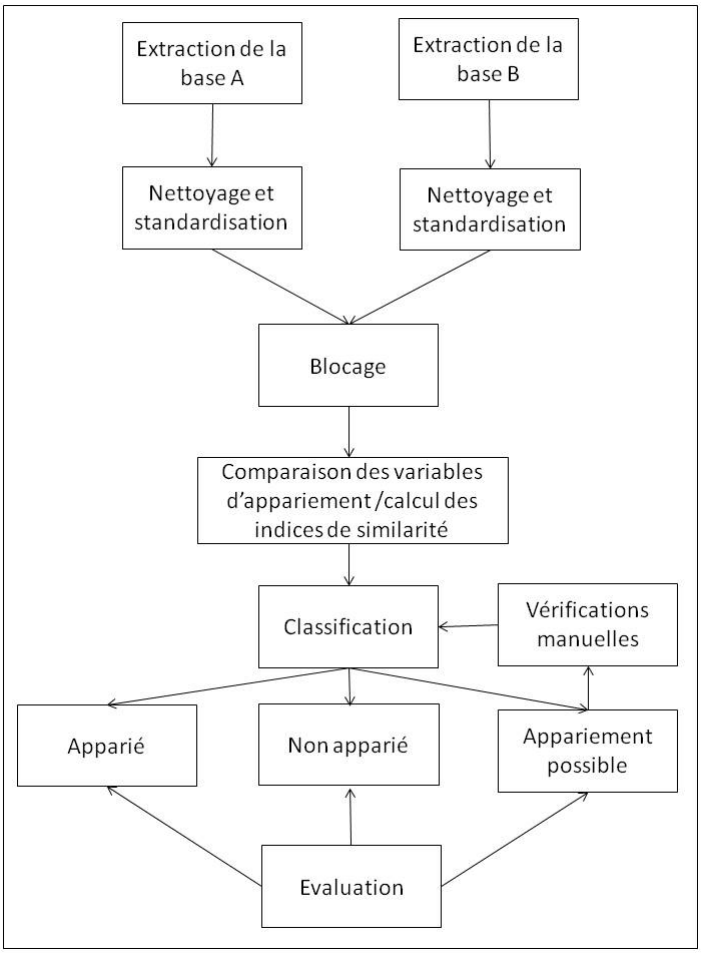

FIGURE 3: Schéma décrivant le processus complet d'un appariement

tualiste (SLM). La principale difficulté lors de la procédure de couplage a porté sur la commune de domicile du SNIIRAM. La variable "commune" était la moins fiable, celle-ci est souvent manquante sur la période 2008-2009 (16% en 2008, 4% en 2009), de plus elle mélange des codes postaux et des codes INSEE (le code INSEE étant la référence dans la BCMD). Par ailleurs, la commune était la variable la plus divergente même quand toutes les autres variables concordaient entre deux entrées. Une stratégie de couplage en deux temps, avec un traitement spécifique de la commune de domicile (comdom), a été appliquée aux sujets décédés en 2008-2010 (figure 4). Cette variable a été utilisée pour départager les cas de couplage multivoques entre une observation du SNIIRAM et plusieurs observations de la BCMD. Ce choix a été fait pour prendre en compte la dépendance entre la divergence des communes et des départements. Dans un premier temps, un couplage est réalisé entre les décédés du SNIIRAM et ceux de la BCMD sur les variables indirectement identifiantes disponibles à l'exception de la commune de domicile, à savoir, le sexe, le mois et l'année de naissance, le jour, le mois et l'année de décès, le département de domicile (depdom). L'année de décès étant la variable la plus fiable cette dernière a

 ISSN: 2102-6238

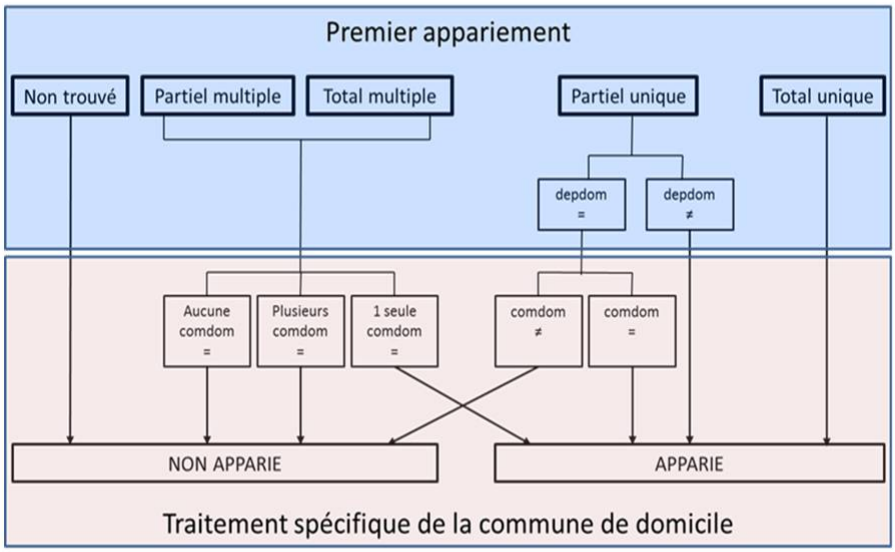

FIGURE 4: Méthode de couplage utilisée pour la production de la base AMPHI

été utilisée comme la variable de blocage. Ainsi nous ne comparons que les personnes décédées la même année. Dans un deuxième temps, la commune de domicile est utilisée pour départager les couplages multivoques et pour trancher les partiels uniques (ces derniers correspondant aux paires d'entrée dont les champs sont similaires sauf pour l'un d'entre eux).

Il permet d'obtenir de "bons résultats" car plus de 94% des individus de la base extraite du SNIIRAM ont trouvé une contrepartie unique dans la BCMD. Le taux de recouvrement, à savoir le nombre d'individus de l'extraction SNIIRAM couplés uniquement avec un décès rapporté au nombre d'individus dans l'extraction SNIIRAM) est supérieur à 91% pour toutes les classes d'âge sauf pour les moins de 20 ans : 92,2% pour les enfants décédés avant un an, 93% pour les sujets décédés entre un et vingt ans. Les différences se font plutôt au niveau géographique, en effet le taux de recouvrement est supérieur à 93% pour tous les départements de domicile à l'exception des Alpes-Maritimes et des départements de la région parisienne. Paris a le taux de recouvrement le plus faible (86%), puis viennent ensuite la Seine-Saint-Denis, les Hauts-de-Seine et le Val-de-Marne avec des taux de recouvrement compris entre 89% et 90% et enfin les Yvelines, l'Essonne et le Val d'Oise avec des taux aux alentours de 92%. Le taux de recouvrement n'est que de 27% pour les personnes domiciliées à l'étranger. Plus généralement, le taux par département pourrait être inversement corrélé avec le pourcentage de personnes nées à l'étranger. De façon générale et sur l'exemple de la base AMPHI on voit que ce type de méthode de couplage peut requérir un effort humain important et une expertise sur les données pour mettre en place des règles pertinentes. Quand il s'agit d'apparier des bases de données dont l'une est

 ISSN: 2102-6238

exhaustive, un couplage déterministe souvent utilisé consiste à épuiser les combinaisons de variables jusqu'à isolement d'une paire unique. Ce qui n'était pas le cas du projet AMPHI, puisque la BMCD contient tous les décès ayant eu lieu en France tandis que l'extraction du PMSI contient des personnes décédées éventuellement à l'étranger.

TABLEAU 5. *Taux de recouvrement de l'extraction du SNIIRAM pour le projet AMPHI par âge et sexe.*

Âge au décès	Sexe masculin		Sexe féminin		Total	
	Nb	%	Nb	%	Nb	%
Moins de 1 an	599	92.0%	433	91.9%	1032	92.2%
1 à 2 ans	415	94.3%	333	91.7%	748	93.2%
3 à 9 ans	784	92.5%	631	93.1%	1415	92.7%
10 à 19 ans	1789	93.6%	799	91.9%	2588	93.1%
20 à 29 ans	5038	92.1%	1702	91.3%	6740	91.9%
30 à 39 ans	7562	92.2%	3839	94.6%	11401	93.0%
40 à 49 ans	19146	94.7%	10786	96.3%	29932	95.3%
50 à 59 ans	43548	95.5%	21683	96.4%	65231	95.8%
60 à 69 ans	57309	94.9%	28000	96.1%	85309	95.3%
70 à 79 ans	84635	94.9%	58711	96.0%	143346	95.4%
80 à 89 ans	101089	95.0%	128697	94.6%	229786	94.8%
9 ans et plus	26531	95.0%	82145	94.2%	108676	94.4%

4.2. Le couplage probabiliste pour le suivi des patients atteints de tumeur maligne

Ce couplage a été réalisé pour démontrer l'applicabilité des méthodes de Fellegi et Sunter et réduire le suivi statistique de patients atteints d'un cancer aux personnes encore vivantes. En l'absence d'un identifiant unique, la détermination du statut vital des patients a été réalisé par le croisement de données hospitalières de l'institut Gustave Roussy (IGR) et des données de la base nationale de mortalité de l'INSEE, après avoir rendu ces informations anonymes (Fournel *et al.*, 2009). La base de données de l'IGR renseigne le statut vital des patients que dans 55% des cas. En revanche en France tous les décès sont enregistrés dans la base nationale de mortalité de l'INSEE. L'ensemble des patients hospitalisés à l'IGR (10 489), domiciliés en France métropolitaine ou dans les départements d'outre-mer, hospitalisés pour la première fois pour une tumeur maligne entre 1998 et 2000 à l'institut Gustave-Roussy ont été inclus. Les données de mortalité de l'INSEE des années 1998 à 2004 (environ 3,5 millions). Après anonymisation par hachage (avec l'algorithme SHA), les fichiers de mortalité et de morbidité hospitalière ont été chaînés sur le nom, le premier prénom, la date de naissance et le code de la commune de naissance au Service de Biostatistique et Information Médicale du CHU de Dijon.

Pour éviter de comparer trop brutalement les noms et prénoms, un algorithme phonétique d'indexation adapté à la langue française a été utilisé juste avant la fonction de hachage. Ceci a permis de gommer certaines erreurs dues à des fautes de saisie. Une étape supplémentaire de vérification automatique et manuelle a permis de corriger les erreurs que l'algorithme phonétique ne sait pas gérer. Cette étape est décrite dans la table 6 dont les conditions sont les suivantes :

Traitement 1 est une étape automatique qui consiste à intervertir le nom avec le nom de jeune fille.

 ISSN: 2102-6238

TABLEAU 6. *Étape additionnelle de validation automatique et manuelle de couplage des enregistrements sur le nom (N), le prénom (P), la date de naissance (DN) et le code commune de naissance (CN). La colonne Niveau indiquant le niveau de concordance.*

Pays de naissance	N	P	DN	CN	vérification	traitement
France	0	1	1	1	automatique	traitement 1
	0	1	1	1	manuelle	traitement 2
	1	1	0	1	automatique	traitement 3
	1	0	1	0	manuelle	traitement 4
	0	0	1	1	automatique	traitement 1
	1	0	0	1	automatique	traitement 5
À l'étranger	0	1	1	.	automatique	traitement 1
	0	1	1	.	manuelle	traitement 2
	1	1	0	.	automatique	traitement 6

Traitement 2 est une vérification manuelle sur le nom de famille, pour faire correspondre par exemple Von Schneider avec De Schneider.

Traitement 3 est une étape automatique qui vérifie, lorsque les dates de naissance divergent, s'il y a concordance de deux informations concernant le jour, le mois et l'année. S'il n'y a qu'une seule concordance on vérifiera la concordance sur le deuxième ou le troisième prénom.

Traitement 4 est une vérification manuelle que la divergence du prénom soit due à un prénom composé par exemple Jean et Jean-Paul.

Traitement 5 est une vérification automatique qu'il y a deux informations concordantes pour la date de naissance et la concordance entre les seconds ou troisièmes prénoms.

Traitement 6 est une vérification automatique qu'il y a deux informations concordantes sur la date de naissance.

Les étapes manuelles ont été réalisées à l'IGR où était stockée la base de correspondance entre les identifiants anonymisés et les identifiants. Le lieu de naissance a été choisi comme la variable bloquante, ce qui implique que seuls les individus avec un lieu de naissance parfaitement renseigné ont été sélectionnés.

La performance de cette appariement est estimée sur sa capacité à prédire le statut vital. Comme nous l'avons dit la base de l'IGR connait le statut vital pour 55% des individus. Le 45% restant ont été complétés par une demande de statut vital au Répertoire National d'Identification des Personnes Physiques (RNIPP). Les résultats du couplage montrent l'intérêt de l'utilisation du couplage probabiliste pour obtenir des informations sur le statut vital d'un nombre important de patients à un moindre coût, puisque la proportion de bien classés était de $97,2\%$, la sensibilité de $94,8\%$ et la spécificité de $99,5\%$. Les valeurs prédictives négative (VPN) et positive (VPP) sont respectivement de $95,3\%$ et de $99,4\%$. Si l'on ajoute l'étape de vérification manuelle, la proportion de bien classés passe à $98,4\%$ avec une spécificité à $99,4\%$ et une sensibilité à $97,2\%$, la VPP est toujours de $99,4\%$ et la VPN est de $97,4\%$. Ces résultats étaient meilleurs pour les patients nés en France, avec un taux de bien classés de $98,3\%$ (respectivement $99,2\%$ avec l'étape supplémentaire) une sensibilité de $96,8\%$ (respectivement $98,5\%$), une spécificité de $99,8\%$ (identique avec l'étape supplémentaire), une VPP à $99,8\%$ (identique avec l'étape supplémentaire) et une VPN à 97% (respectivement $98,6\%$). Les résultats étaient moins bons pour les patients nés à l'étranger avec un taux de bien classés à $90,7\%$, une sensibilité à

 ISSN: 2102-6238

82,8% et spécificité à 97,7% (respectivement un taux de 93,7%, une sensibilité de 89,8, une spécifié à 97.2%), une VPP à 97,0% (respectivement 96,6%) et une VPN à 86,5% (respectivement 91,5%). L'ajout d'information complémentaire comme le second prénom ou le lieu de naissance (plus précis que le pays) a permis d'améliorer un peu les résultats. Enfin les performances sont dépendantes du sexe, la spécificité et la sensibilité sont meilleures chez les hommes (resp. 99,9% et 97,9%) que chez les femmes (respectivement 99,8% et 95%). Cela est probablement dû au fait que le nom de jeune fille n'est pas toujours présent.

4.3. Le couplage probabiliste pour l'évaluation d'un réseau périnatal régional

Le Réseau Périnatal de Bourgogne (RPB) inclut tous les établissements publics et privés de la région prenant en charge les femmes enceintes et les nouveau-nés (environ 18 000 naissances annuelles réparties sur 18 établissements). Un recueil continu de 42 indicateurs a été mis en place en 1998 (25 indicateurs pour la mère et 17 pour le nouveau né). Les informations sont extraites des résumés du Programme de Médicalisation de Systèmes d'Information (PMSI) recueillis pour toutes les hospitalisations. En effet, tout séjour hospitalier, effectué dans un établissement de santé public ou privé, fait l'objet d'un résumé dans l'objectif d'établir le budget des hôpitaux en fonction de leur activité. Les indicateurs n'existant pas dans le PMSI, tels que les facteurs de risques psychosociaux, font l'objet d'un recueil sur une fiche adjointe au résumé PMSI, constituant un "résumé élargi". Pour le traitement des données médicales, le chaînage des "résumés élargis" est réalisé à deux niveaux différents. D'une part, les "résumés élargis" d'une même personne, mère ou nouveau-né, doivent pouvoir être reliés lorsqu'il y a hospitalisations successives (plusieurs unités ou établissements différents). D'autre part, les "résumés élargis" de la mère doivent être reliés à ceux de l'enfant afin d'évaluer l'impact postnatal des facteurs de risques et des pathologies maternelles. Le couplage de données anonymes a alors été rendu possible par l'utilisation du logiciel "Anonymat" à partir de six variables, saisies chez la mère et son bébé : le nom de jeune fille de la mère, son prénom et sa date de naissance, le prénom de l'enfant et sa date de naissance, le code postal de résidence de la mère. Avant transmission, les fichiers sont validés au sein de chaque établissement. De plus, l'exhaustivité et la qualité du recueil des données de chaînage sont systématiquement contrôlées par l'équipe coordinatrice du RPB, qui assure le chaînage mère-enfant (pour 99.9% des nouveau-nés) selon la méthode de Jaro (1995) et Quantin *et al.* (2009). Dans cet appariement plusieurs hypothèses sont faites, d'abord il est impossible de coupler une mère avec le mauvais nouveau né si l'ensemble des variables concordent. De plus chaque mère doit correspondre à un nouveau né et chaque nouveau né doit avoir une mère. Le cas contraire impliquerait la non exhaustivité des bases ou bien la présence d'erreur. Ainsi plusieurs appariements probabilistes suivant Jaro (1995) ont été fait, le premier permet de coupler les cas non problématiques. Les non trouvés sont ensuite appariés en retirant une voire deux variables afin de savoir d'où peuvent provenir les écarts. Les résultats sont ensuite renvoyés aux établissement du réseau afin de faire des vérifications (les premiers appariements ayant permis de limiter les vérifications aux cas les plus problématiques). À la suite de cette étape de vérifications un dernier appariement est réalisé sur l'ensemble des variables. La méthode est évalué ensuite sur un ensemble de test validé manuellement, dont on trouvera les résultats dans la table 7.

La cause principale de la présence des faux positives est le nombre important de données

 ISSN: 2102-6238

TABLEAU 7. *Évaluation de la procédure : sensibilité, spécificité, taux de vrais positifs (VP), taux de vrais négatifs (VN), taux de faux positifs (FP), taux de faux négatifs (FN)*

Année	Sensibilité	Spécificité	VP	VN	FN	FP
1998	89,04%	95,25%	85,68%	3,58%	10,55%	0,18%
1999	97,53%	97,82%	89,04%	8,52%	2,25%	0,19%
2000	97,24%	97,77%	86,35%	10,50%	2,45%	0,70%
2001	93,03%	87,47%	78,99%	13,20%	5,92%	1,89%
2002	94,17%	89,74%	79,72%	13,76%	4,94%	1,57%
2003	93,82%	86,84%	78,64%	14,05%	5,18%	2,13%
2004	88,26%	87,31%	69,14%	18,91%	9,20%	2,75%
2005	93,13%	86,02%	70,57%	20,84%	5,20%	3,39%
2006	97,09%	77,74%	73,97%	18,51%	2,21%	5,30%

manquantes en effet parmi les faux positifs dans 85,11% des cas il y a concordance sur toutes les variables sauf celles qui sont manquantes.

5. Logiciels de couplage

Les principaux logiciels libres pour le couplage probabiliste sont R, au travers du package Record Linkage (Borg et Sariyar, 2016), FRIL (Jurczyk, 2009) et Febrl (Christen, 2008b). Febrl est une plateforme en open source basée sur Python qui propose une interface graphique pour réaliser son appariement. Il contient des fonctionnalités avancées de nettoyage et de standardisation des données (notamment des méthodes de Markov cachées). Il contient aussi les similarités présentées dans ce document : indice de Jaro, q-grams, phonétique (il y en a 26 en tout (Christen, 2006)). Il propose les méthodes classiques de couplage probabiliste comme Fellegi et Sunter, à savoir le calcul du score et la discrimination mais pas l'estimation des paramètres du modèle. Il propose également d'autres méthodes utilisant les séparateurs à vaste marge applicable au cadre non supervisé. Par exemple, il permet de développer son propre algorithme de couplage. Febrl est un outil très utile pour débuter et comprendre le fonctionnement des algorithmes de couplage et pour faire des comparaisons entre les différents algorithmes et méthodes, il peut créer des bases synthétiques pour tester des méthodes d'appariement. Cependant, son modèle de gestion de données, chargées en mémoire, ne permet pas de traiter nativement des bases de données très volumineuses. FRIL est une plateforme basée sur Java qui supporte moins d'indice de similarité que Febrl (Levenshtein, q-gram, Jaro-Winkler, Soundex). La règle de décision appliquée dans le processus d'appariement est celle de Fellegi et Sunter (l'estimation des paramètres se faisant par l'algorithme EM). On pourra également trouver dans Wright (2011) une présentation sur l'utilisation de la procédure SQL avec la clause JOIN, qui est un outil puissant pour réaliser le produit cartésien des bases de données en incluant du blocage simple uniquement. SAS propose aussi des similarités comme la distance de Levenshtein via les fonctions COMPGED, COMPLEV et SPEDIS ou le Soundex via la fonction SOUNDEX. Enfin il existe un package R pour le couplage probabiliste, Record Linkage, qui contient un ensemble basique de fonctions de similarités (phonétiques, Jaro-Winkler, Levenshtein). Le package permet l'appariement non supervisé via le modèle probabiliste de Fellegi et Sunter ainsi que l'estimation des paramètres via l'algorithme EM, des méthodes de classifications non supervisées (K-mean clustering, bagged clustering). Les arbres de décision, le séparateur à vaste marge (SVM) et les réseaux de neurones sont aussi

 ISSN: 2102-6238

implémentés pour l'apprentissage supervisé. Pour l'expérimentation le module emprunte le générateur de données synthétique de Febrl. Ce module est aussi l'un des rares à contenir une fonctionnalité de calibration non supervisée des scores (Sariyar *et al.*, 2011). Au final la plateforme Febrl et le package R Record Linkage sont de bons outils pour réaliser des appariements de tout type et pour faire l'expérimentation au regard de la richesse des fonctions qu'ils proposent comme la possibilité de simuler des données ainsi que l'éventail des outils présentés dans ce document. FRIL et SAS seront plus puissant pour des gros volume de données mais sont plus limités en fonctionnalité et nécessiteront plus souvent l'implémentation de nouvelles fonctionnalités (pour le cas de SAS). La réglementation impose l'utilisation d'un serveur/espace sécurisé pour réaliser les traitements statistiques sur des données de santé individuelles. Or ces espaces contraignent souvent les outils mis à disposition pour des raisons de sécurité. Même s'il n'y a pas vraiment d'outil spécifique pour l'appariement en santé nous recommandons toutefois les développements en SAS, R et Python que l'on retrouvera le plus souvent disponible dans ces espaces sécurisés.

6. Conclusion

Nous avons présenté un vaste panorama de méthodes de couplage ainsi que les moyens pour intégrer les résultats de ce couplage dans une analyse statistique afin de la rendre plus exacte. Nous avons également proposé quelques logiciels ou fonctions permettant de réaliser des appariements. L'objectif n'était pas de proposer une étude comparative qu'il est possible de trouver dans les sources que nous avons citées, mais plutôt de présenter l'état de l'art pour accompagner un projet de recherche nécessitant la réalisation d'un appariement, ou susciter des initiatives de développements spécifiques, car beaucoup de ces méthodes n'ont pas pu être testées et évaluées sur des données réelles. A ce titre nous n'abordons pas ici un sujet pourtant important concernant la mise en œuvre de ces méthodes sur le plan juridique. En effet, appliquer ces méthodes dans le respect des contraintes réglementaires (protection du NIR, espace confiné, tiers de confiance, circulation sécurisée, habilitation) peut nécessiter la mise en place d'une organisation impliquant différents acteurs. Nous avons mis en évidence la pluralité de ces méthodes en ayant à l'esprit les contraintes légales et opérationnelles des utilisateurs afin d'informer au mieux ces derniers avant d'initier un projet de recherche nécessitant un appariement. C'est pour cela que nous avons accentué cette revue sur les modèles probabilistes qui semblent respecter une grande partie de ces contraintes en offrant une boite à outils permettant l'exploitation des données appariées. Les évolutions récentes de la loi de modernisation de notre système de santé et de la loi pour une république numérique et la mise en place par l'Inserm d'une infrastructure de services chargée notamment d'effectuer des appariements et la mise à disposition des données de santé et administratives pourraient faciliter la réalisation d'études comparatives sur les méthodes d'appariement.

Références

AGRAWAL, R. et SRIKANT, R. (2002). Searching with numbers. *In Proceedings of the 11th International Conference on World Wide Web*, WWW '02, pages 420–431. ACM.

ANANTHAKRISHNA, R., CHANDHURI, S. et GANTI, V. (2002). Eliminating fuzzy duplicates in data warehouses. VLDB '02 Proceedings of the 28th international conference on Very Large Data Bases.

 ISSN: 2102-6238

BATXER, R., CHRISTEN, P. et CHURCHES, T. (2003). A comparison of fast blocking methods for record linkage. ACM SIGKDD '03 Workshop on Data Cleaning, Record Linkage and Object Consolidation.

BELIN, T. (1990). A proposed improvement in computer matching. *Statistics of Income and Related Administrative Record Resarch*, pages 167–172.

BELIN, T. et RUBIN, D. (1995). A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association*, 90(430):694–707.

BILENKO, M., COHEN, W., FEINBERG, S., MOONEY, R. et RAVIKUMAR, P. (2003). Adaptive name-matching in information integration. *IEEE Intelligent System*, 18:16–23.

BOHENSKY, M., JOLLEY, D., SUNDARARAJAN, V., EVANS, S., PILCHER, D., SCOTT, I. et BRAND, C. (2010). Data linkage : A powerful research tool with potential problems. *BMC Health Services Research*, 10:346.

BORG, A. et SARIYAR, M. (2016). *RecordLinkage : Record Linkage in R*. R package version 0.4-10, https ://CRAN.R-project.org/package=RecordLinkage.

BOX, G. et COX, D. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252.

CHAMBERS, R. (2009). Regression analysis of probability-linked data. *Official Statistics Research Series*, 4.

CHANDURI, S., GANTI, V. et MOTWANI, R. (2005). Robust identification of fuzzy duplicates. pages 865–876, Washington, DC, USA. IEEE Computer Society.

CHEESEMAN, P. et STUTZ, J. (1997). Bayesian classification (AutoClass) :theory and results. *Advances in Knowledge Discovery and Data Mining*.

CHRISTEN, P. (2006). A comparison of personal name matching : Techniques and practical issues. *In Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06)*, pages 290–294.

CHRISTEN, P. (2007). A two-step classification approach to unsupervised record linkage. *In Proceedings of the Sixth Australasian Conference on Data Mining and Analytics - Volume 70*, AusDM '07, pages 111–119. Australian Computer Society, Inc.

CHRISTEN, P. (2008a). Automatic training example selection for scalable unsupervised record linkage. *In Proceedings of the 12th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, PAKDD'08, pages 511–518. Springer-Verlag.

CHRISTEN, P. (2008b). Febrl - : An open source data cleaning, deduplication and record linkage system with a graphical user interface. *In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1065–1068. ACM.

CHRISTEN, P. (2012). *Data Matching : Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer-Verlag Berlin and Heidelberg GmbH & Co. K, 2012 ed. édition.

COCHINWALA, M., KURIEN, V., LALK, G. et SASHA, D. (2001). Efficient data reconciliation. *Information Science*, 137(1):1–15.

COHEN, W. (2000). Data integration using similarity joins and world-based information representation language. *ACM Transactions on Information Systems*, 18(3).

COHEN, W. W. (1998). Integration of heterogeneous databases without common domains using queries based on textual similarity. pages 201–212.

COPAS, J. et HILTON, F. (1990). Record linkage : Statistical models for matching computer records. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 153(3):287–320.

CORNUÉJOLS, A. et MICLET, L. (2010). *Apprentissage Artificiel : Concept et Algorithmes*. Algorithme. Eyrolles.

DEMPSTER, A., LAIRD, N. et RUBIN, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

DOMINGO-FERRER, J. et TORRA, V. (2002). Validating distance-based record linkage with probabilistic record linkage. *In Topics in Artificial Intelligence*, pages 207–215. Springer.

DUA, S. et CHOWRIAPPA, P. (2012). *Data Mining for Bioinformatics*. CRC Press.

DUFLO, M. (1997). *Algorithmes stochastiques*. Springer.

ELFEKY, M., VERYKIOS, V. et ELMAGARMID, A. (2002). TAILOR : a record linkage toolbox. *In Proceedings 18th International Conference on Data Engineering*, pages 17–28.

ELMAGARMIND, A., IPEIROTIS, P. et VERYKIOS, V. (2007). Duplicate record detection : A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(1).

FELLEGI, I. et SUNTER, A. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210.

FORD, J., ROBERTS, C. et TAYLOR, L. (2006). Characteristics of unmatched maternal and baby records in linked

 ISSN: 2102-6238

birth records and hospital discharge data. *Paediatric and Perinatal Epidemiology*, 20(4):329–337.

FORTINI, M., LISEO, B., NUCCITELLI, A. et SCANU, M. (2001). On Bayesian record linkage. *Research in Official Statistics*, 4(1).

FOULLEY, J.-L. (2002). Algorithme "EM" : Théorie et application au modèle mixte. *Journal de la Société Française de Statistique*, 18(3-4):57–109.

FOURNEL, I., SCHWARZINGER, M., BINQUET, C., BENZENINE, E., HILL, C. et QUANTIN, C. (2009). Contribution of record linkage to vital status determination in cancer patients. *Studies in Health Technology and Informatics*, 150:91–95.

GILL, L. (1999). Ox-link : The oxford medical record linkage system. *Record Linkage Techniques*.

GOLDSTEIN, H., CARPENTER, J., KENWARD, M. et LEVIN, K. (2009). Multilevel models with multivariate mixed response types. *Statistical Modelling*, 9(3):173–197.

GOLDSTEIN, H., HARRON, K. et CORTINA-BORJA, M. (2017). A scaling approach to record linkage. *Statistics in Medicine*, 36(16):2514–2521.

GOLDSTEIN, H., HARRON, K. et WADE, A. (2012). The analysis of record-linked data using multiple imputation with data value priors. *Statistics in Medicine*, 31(28):3481–3493.

GUHA, S., KOUDAS, N., MARATHE, A. et SRIVASTAVA, D. (2004). Merging the results of approximate match operations. *In Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30*, pages 636–647. VLDB Endowment.

HABERMAN, S. (1979). *Analysis of qualitative data. vol. 2, new developments*. Academic Press Inc.

HARRON, K., DOIDGE, J., KNIGHT, H., GILBERT, R., GOLDSTEIN, H., CROMWELL, D., MEULEN, V. et JAN, H. (2017). A guide to evaluating linkage quality for the analysis of linked data. *International Journal of Epidemiology*, 46(5):1699–1710.

HARRON, K., WADE, A., GILBERT, R., MULLER-PEBODY, B. et GOLDSTEIN, H. (2014). Evaluating bias due to data linkage error in electronic healthcare records. *BMC Medical Research Methodology*, 14:36.

HASTIE, T., TIBSHIRANI, R. et FRIEDMAN, J. (2001). *The Elements of Statistical Learning*. Springer.

HERANDEZ, M. et STOLFO, S. (1998). Real-world data is dirty : Data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery*, 2(1):9–37.

HERZOG, T. N., SCHEUREN, F. J. et WINKLER, W. E. (2007). *Data Quality and Record Linkage Techniques*. Springer Publishing Company, Incorporated, 1st édition.

HOF, M. et ZWINDERMAN, A. (2012). Methods for analyzing data from probabilistic linkage strategies based on partially identifying variables. *Statistics in Medicine*, 31(30):4231–4242.

HOF, M. et ZWINDERMAN, A. (2015). A mixture model for the analysis of data derived from record linkage. *Statistics in Medicine*, 34(1):74–92.

IZENMAN, A. (2008). *Modern Multivariate Statistical Techniques : Regression, Classification and Manifold Learning*. Springer Texts in Statistics. Springer.

JAIN, S. et NEAL, R. (2004). A split-merge Markov chain monte carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13(1):158–182.

JARO, M. (1995). Probabilistic linkage of large public health data files. *Statistics in Medicine*, 14(5):491–498.

JARO, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association*, 84(406).

JARO, M. A. et STATES., U. (1978). *UNIMATCH : a record linkage system : users manual*. Bureau of the Census, Washington.

JURCZYK, P. (2009). *FRIL : Fined-grained Record Integration and Linkage Tool Tutorial*. http ://fril.sourceforge.net/FRIL-Tutorial-3.2.pdf.

KELLEY, P. (1986). Robustness of the census bureau's record linkage system. *In Proceedings of the Section on Survey Research Methods, American Statistical Association*, pages 620–624.

KIM, G. et CHAMBERS, R. (2012a). Regression analysis under incomplete linkage. *Computational Statistics & Data Analysis*, 56(9):2756–2770.

KIM, G. et CHAMBERS, R. (2012b). Regression analysis under probabilistic multi-linkage. *Statistica Neerlandica*, 66(1):64–79.

LAHIRI, P. et LARSEN, M. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, 100(469):222–230.

LAMARCHE-VADEL, A., JOUGLA, E. et REY, G. (2013). *Base AMPHI : Base de données pour l'Analyse de la Mortalité Post-Hospitalisation en France en 2008-2010*. Thèse de doctorat, Université Paris-Sud / Inserm-CépiDC.

 ISSN: 2102-6238

LARSEN, M. (2005). Advances in record linkage theory : Hierarchical Bayesian record linkage theory. ASA Section on Survey Research Methods.

LARSEN, M. et RUBIN, D. (2001). Iterative automated record linkage using mixture models. *Journal of the American Statistical Association*, 96(453):32–41.

LASH, T. L., FOX, M. P. et FINK, A. K. (2009). *Applying Quantitative Bias Analysis to Epidemiologic Data*. Springer Publishing Company, Incorporated, 1st édition.

LEGLEYE, S., RICHARD, J.-B., REY, G., BECK, F. et GRIEVE, M. (2017). Testing the Acceptability of Asking Respondents for Identifying Information in a Cross-Sectional Survey of the General Population. *Population, English edition*, 72(4):697–713.

LIM, E., SRIVASTAVA, J., PRABHAKAR, S. et RICHARDSON, J. (1996). Entity identification in database integration. *Informatics and computer science*, 89(1).

LOTH, A. (2015). Données de santé : Anonymat et risque de ré-identification. Dossiers Solidarité Santé 64, Drees.

MCGLINCY, M. (2004). A Bayesian record linkage methodology for multiple imputation of missing links.

MCLACHLAN, G. et KRISHNAN, T. (2008). *The EM Algorithm and Extensions*. Wiley-Blackwell, 2nd edition édition.

MENG, X. et RUBIN, D. (1993). Maximum likelihood estimation via the ECM algorithm : A general framework. *Biometrika*, 80(2):267–278.

MENG, X. et VAN DYK, D. (1997). The EM algorithm-an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 59(3):511–567.

MONGE, A. et ELKAN, C. (1996). The field matching problem : Algorithms and applications. *In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 267–270. AAAI Press.

MONGE, A. et ELKAN, C. (1997). An efficient domain-independent algorithm for detecting approximately duplicate database record.

NETER, J., MAYNES, E. et RAMANATHAN, R. (1965). The effect of mismatching on the measurement of response error. *Journal of the American Statistical Association*, 60(312):1005–1027.

NEWCOMBE, H. et KENNEDY, J. (1959). Automatic linkage of vital records. *Science*, 130(3381):954–959.

NEYMAN, J. et PEARSON, E. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337.

NIGAM, K., MCCALLUM, A., THRUN, S. et MITCHELL, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2):103–134.

PHILIPS, L. (1990). Hanging on the metaphone. *Computer Language*, 7(12):39–44.

QUANTIN, C., GOUYON, B., AVILLACH, P., FERDYNUS, C., SAGOT, P. et GOUYON, J.-B. (2009). Using discharge abstracts to evaluate a regional perinatal network : Assessment of the linkage procedure of anonymous data. *International Journal of Telemedicine and Applications*, 2009.

ROGOT, E., SORLIE, P. et JOHNSON, N. (1986). Probabilistic methods in matching census samples to the national death index. *Journal of Chronic Diseases*, 39(9):719–734.

RUBIN, D. et BELIN, T. (1991). Recent developments in calibrating error rates for computer matching. Conference Paper 1991 Annual Research Conference :Proceedings : Bureau of the Census.

SADINLE, M. et FIENBERG, S. E. (2013). A generalized Fellegi-Sunter framework for multiple record linkage with application to homicide record systems. *Journal of the American Statistical Association*, 108:385–397.

SARIYAR, M., BORG, A. et POMMERENING, K. (2011). Controlling false match rates in record linkage using extreme value theory. *Journal of Biomedical Informatics*, 44(4):648–654.

SCHAFER, J. (1997). *Analysis of Incomplete Multivariate Data*. Chapman and Hall/CRC.

SCHEUREN, F. et WINKLER, W. (1993). Regression analysis of data files that are computer matched. Survey Methodology,.

SCHÜRLE, J. (2005). A method for consideration of conditional dependencies in the Fellegi and Sunter model of record linkage. *Statistical Papers*, 46(3):433–449.

SHAWE-TAYLOR, J. et CRISTIANINI, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA.

STEORTS, R. (2015). Entity resolution with empirically motivated priors. *Bayesian Analysis*, 10(4):849–875.

STEORTS, R., HALL, R. et FIENBERG, S. (2013). A Bayesian approach to graphical record linkage and de-duplication. *arXiv :1312.4645 [stat]*.

STEORTS, R., HALL, R. et FIENBERG, S. (2014a). SMERED : A Bayesian approach to graphical record linkage and de-duplication. *arXiv :1403.0211 [stat]*.

 ISSN: 2102-6238

STEORTS, R., VENTURA, S., SADINLE, M. et FIENBERG, S. (2014b). A comparison of blocking methods for record linkage. *In Privacy in Statistical Databases*, pages 253–268. Springer, Cham.

TAFT, R. (1970). *Name Search Techniques.* Technical Report Special no.1 New York State Identification and Intelligence System.

TANCREDI, A. et LISEO, B. (2011a). A hierarchical Bayesian approach to record linkage and population size problems. *The Annals of Applied Statistics*, 5(2):1553–1585.

TANCREDI, A. et LISEO, B. (2011b). Some advances on Bayesian record linkage and inference for linked data. *In Proceedings of the ESSnet Data Integration Workshop*. http ://www.ine.es/e/essnetdi_ws2011/ppts/Liseo_Tancredi.pdf.

THIBAUDEAU, Y. (1993). The discrimination power of dependency structures in record linkage. *SURVEY METHODOLOGY*, 19(1).

TORRA, V. et DOMINGO-FERRER, J. (2003). Record linkage methods for multidatabase data mining. *In* TORRA, P. V., éditeur : *Information Fusion in Data Mining*, numéro 123 de Studies in Fuzziness and Soft Computing, pages 101–132. Springer Berlin Heidelberg.

TROMP, M., MÉRAY, N., RAVELLI, A., REITSMA, J. et BONSEL, G. (2008). Ignoring dependency between linking variables and its impact on the outcome of probabilistic record linkage studies. *Journal of the American Medical Informatics Association : JAMIA*, 15(5):654–660.

VERYKIOS, V., ELMAGARMID, A. et HOUSTIS, E. (2000). Automating the approximate record-matching process. *information Science*, 126(1):83–98.

WINKLER, W. (1988). Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage. Bureau of the Census Statistical Research Report Series.

WINKLER, W. (1990). String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage.

WINKLER, W. (1993). Improved decision rules in the Fellegi-Sunter model of record linkage. Rapport technique, Bureau of the Census Statistical Research Report Series.

WINKLER, W. (1999). The state of record linkage and current research problem. Rapport technique, Statistical Research Division, U.S. Census Bureau.

WINKLER, W. (2000). Machine learning, information retrieval and record linkage. Rapport technique, American Statistical Association, Proceeding of the Section of Survey Research Methods.

WINKLER, W. (2002). Methods for record linkage and Bayesian networks. Rapport technique, Statistical Research Report Series RRS2002/05, US Bureau of the Census.

WINKLER, W. (2007). Automatically estimating record linkage false match rates. Rapport technique, Statistical Research Division, U.S. Census Bureau.

WRIGHT, G. (2011). Probabilistic record linkage in SAS. *Proceedings of Western Users of SAS Software*.

WU, C. et JEFF, F. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103.

Annexe A : Les indices de similarité

Un couplage trivial consiste à comparer les champs des deux bases que nous cherchons à relier, et de faire correspondre les entrées dont tous les champs renseignent des informations identiques. Dans le cas où la mise en correspondance échoue, on considère que les entrées se rapportent à des individus différents. Cette méthode est trop restrictive car elle ne prend pas en compte les difficultés relevées ci-dessus comme la qualité des bases de données, la production d'erreur lors de l'entrée, le changement de statut d'un individu et les modifications que cela peut entrainer (le changement de nom après un mariage par exemple). Pour pallier à ces situations, de nombreuses mesures de similarité ont été développées, permettant d'apprécier la proximité entre deux champs non strictement identiques et donc de définir une version relaxée de la concordance des champs. Chaque mesure, présentée ci-dessous, est plus adaptée à un type d'erreur. On trouvera pour l'ensemble de ces similarités une présentation détaillée dans Elmagarmind *et al.* (2007),Torra et Domingo-Ferrer (2003) et Dua et Chowriappa (2012).

 ISSN: 2102-6238

– La similarité de Hamming est une distance utilisée en théorie des codes correcteurs d'erreur. Elle consiste à comparer deux mots issus d'un alphabet binaire de même longueur et de compter le nombre de lettres distinctes.
– La similarité de Levenshtein (Edit distance) : très utilisée dans l'alignement de séquences, cette similarité est une distance entre deux chaînes de caractères qui mesure le nombre minimal d'opérations nécessaires pour transformer une chaîne de caractères en une autre chaîne. Les opérations autorisées étant l'insertion, la suppression et l'échange de caractère. Chacune des opérations citée ci-dessus se voit généralement attribuer un coût différent, on pourrait considérer par exemple que l'opération de substitution coûte moins cher que celle d'ajout de caractère. La distance de Levenshtein est moins efficace quand les erreurs sont dues à des troncatures ou des omissions volontaires (par exemple quand le deuxième prénom est tronqué). Pour remédier à ce problème, la similarité "affine gap" autorise deux opérations supplémentaires : l'ajout d'un espace et l'extension d'un espace. L'extension étant une opération à laquelle on attribue un coût inférieur aux autres opérations (substitution, suppression...). La similarité de Smith-Waterman diminue le coût des dissemblances présentes en début et en fin de chaîne de caractères.
– La similarité de Jaro : Jaro et States. (1978) présente une définition séquentielle de cet indice. Adaptée aux chaînes de caractères, cet indice comptabilise les caractères communs entre deux chaînes de caractères de la façon suivante : Soient les chaînes de caractères C_a de longueur L_a et C_b de longueur L_b, autrement dit $C_a = a_1a_2...a_{L_a}$ et $C_b = b_1b_2...b_{L_b}$. Soient maintenant SC_a la sous chaîne de longueur SL de C_a qui contient les caractères en commun avec C_b et SC_b la sous chaîne de C_b, de même longueur SL que SC_a, qui contient des caractères qui sont contenus dans C_a. Dans cette définition un caractère a_k est commun à un caractère b_j si $a_k = b_j$ et si $k - f \leq j \leq k + f$ avec $f = \frac{min(L_a, L_b)}{2}$. Enfin on note $T_{a,b}$ le nombre de transpositions, c'est à dire le nombre de caractères de la sous chaîne SC_a qui n'est pas au même rang dans la sous chaîne SC_b. On a :

$$Sim_{Jaro}(C_a, C_b) = W_a \frac{SL}{L_a} + W_b \frac{SL}{L_b} + W_{a,b} \frac{SL - \frac{T_{a,b}}{2}}{SL} \tag{24}$$

Où W_l ($l = a, b$ ou a, b) est le poids mesurant la contribution de chaque proportion souvent choisi à 1/3. Cet indice est plutôt adapté aux chaînes de caractères courtes. Ce n'est pas une distance puisque la mesure est maximale quand cette mesure est égale à 1 et que les chaînes de caractères sont identiques, par ailleurs elle ne vérifie pas l'inégalité triangulaire. Winkler (1999) propose la variante suivante : si L est la longueur du préfixe en commun entre les sous-chaînes SC_a et SC_b alors la mesure de Jaro-Winkler est :

$$Sim_{Jaro-Winkler}(C_a, C_b) = Sim_{Jaro}(C_a, C_b) + p \frac{min(L, 4)}{10} (1 - Sim_{Jaro}(C_a, C_b)) \tag{25}$$

Où p est un poids attribué à l'importance que l'on porte au préfixe. Winkler motive cette extension par le fait que les préfixes sont souvent plus fiables que le reste de la chaîne. Pour les données numériques et/ou continues, Jaro propose la similarité suivante pour relaxer les poids :

$$\tau_{Jaro}(a_i, b_i) = 1 - min\left(k \frac{|a_i - b_i|}{min(a_i, b_i)}, 1\right) \tag{26}$$

 ISSN: 2102-6238

Où k est coefficient de relaxation. Cette similarité est adaptée pour l'âge par exemple ou un écart faible entre des individus âgés peut plus s'apparenter à une erreur tandis qu'un écart d'un an pour des personnes plus jeune (autour d'un an par exemple) signifiera plus probablement le fait que ces personnes ne sont pas les mêmes.

– La similarité Atomic strings : cet indice est plutôt adapté pour faire correspondre une chaîne de caractères avec son abréviation. Une chaîne atomique est une suite de caractères délimitée pas des caractères de ponctuation. Deux chaînes atomiques sont identiques si elles correspondent intégralement ou si l'une est un préfixe de l'autre. L'indice se calcule comme la proportion moyenne de chaînes atomiques qui sont identiques. Autrement dit soient Ca et Cb deux chaînes de caractères, soient $|C_a|$ et $|C_b|$ les nombres d'éléments atomiques respectifs. Soit k le nombre d'éléments atomiques qui concordent entre C_a et C_b alors :

$$Sim_{Atom}(C_a, C_b) = 2\frac{k}{|C_a| + |C_b|} \tag{27}$$

Cet indice de similarité est plus adapté aux champs contenant une adresse ou une référence bibliographique car elle permet la proximité d'une chaîne de caractères avec son diminutif.

– La similarité n-grammes : l'ensemble des n-grammes d'une chaîne de caractères est l'ensemble de ses sous chaînes de caractères de longueur égale à n. Pour deux chaînes de caractères, la similarité n-grammes est le ratio entre le nombre de n-grammes communs entre ces deux chaînes et la moyenne arithmétique du nombre de n-grammes de chacune des chaînes. Soient Ng_a et Ng_b les ensembles des n-grammes contenus dans la chaîne de caractères C_a et C_b respectivement.

$$Sim_{n.gram}(C_a, C_b) = 2\frac{\sharp(Ng_a \cap Ng_b)}{\sharp Ng_a + \sharp Ng_b} \tag{28}$$

Comme on le verra ensuite cet indice de similarité est pratique pour mettre en place un protocole de couplage sur des données cryptées. Elle a cependant le défaut d'accroître la quantité d'information à traiter, ce qui sera dommageable lors de couplage de grandes bases de données.

– La similarité cosinus : issue de la recherche d'information sur internet et du traitement automatique de texte, cette mesure présentée dans Cohen (1998) consiste à représenter les chaînes de caractères comme des vecteurs dans un espace euclidien, et de mesurer la similarité existante entre les représentations des deux chaînes. La représentation d'un document dans un espace vectoriel peut se faire en utilisant le codage $tf.idf$. Chaque chaîne est décrite vectoriellement au travers d'un vocabulaire de référence (contenant des lettres, des mots ou des morceaux de phrase), que l'on appelle un corpus V dont la taille est notée $|V|$. La chaîne de caractères sera projetée dans un espace vectoriel dont la dimension est exactement la taille du corpus. Pour une chaîne de caractères donnée on calcule les coordonnées de son vecteur représentatif de la façon suivante : chaque terme m du corpus est comparé aux différents mots de la chaîne, afin de déterminer leur fréquence d'apparition dans la chaîne, cette fréquence sera notée tfm. Ensuite pour chaque mot m, on regarde son pouvoir discriminant au travers du nombre, noté n_m, d'entrées dans la base qui contiennent le mot m, cette quantité est notée $idfm = \frac{N}{n_m}$. Soit une chaîne de caractères C_a, sa coordonnée m

 ISSN: 2102-6238

(on indexe par les mots du corpus) est donnée par pour tout $m \in V$,

$$C_a(m) = \log(1 + tf_m).\log(idf_m) \tag{29}$$

Une fois construit le vecteur représentatif des entrées, sur la base des deux quantités précédentes (fréquence et pouvoir discriminant), on compare les entrées des deux bases en mesurant l'angle formé par les vecteurs représentatifs en utilisant le produit scalaire. Pour deux chaînes C_a et C_b la similarité entre les deux chaînes est :

$$Sim_{cosine}(C_a, C_b) = \sum_{(m \in V} \frac{C_a(m)}{\sqrt{\sum_{m \in V}(C_a(m))^2}} \frac{C_b(m)}{\sqrt{\sum_{m \in V}(C_b(m))^2}} \tag{30}$$

Cette méthode est connue pour fonctionner dans des situations diverses et être insensible à l'ordre des mots dans une chaîne de caractères. Par ailleurs cet indice va avoir tendance à associer des chaînes de caractères contenant simultanément des mots dont l'occurrence est rare. Cependant elle ne prend pas en considération les erreurs de caractère. En effet selon la structure du corpus, si l'on retire une lettre à une chaîne de caractères la nouvelle chaîne pourra avoir une similarité très faible par rapport à la chaîne de départ. Bilenko *et al.* (2003) propose une version plus souple de la similarité cosinus :

$$Sim_{soft-TF.IDF}(C_a, C_b) = \sum_{(m \in V(\theta, C_a, C_b)} \frac{C_a(m)}{\sqrt{\sum_{m \in V}(C_a(m))^2}} \frac{C_b(m)}{\sqrt{\sum_{m \in V}(C_b(m))^2}} \max_{v \in C_b} sim(v, m) \tag{31}$$

L'ensemble $V(\theta, C_a, C_b)$ est constitué des mots m du corpus contenu dans C_a et pour lesquels il existe au moins un mot v dans C_b vérifiant que $sim(m, v) \geq \theta$. La similarité Sim utilisée est souvent celle de Jaro ou celle de Jaro-Winkler.

– Les similarités phonétiques : ces similarités sont différentes de celles présentées précédemment, elles ne correspondent pas à la mesure d'un écart, mais plutôt au codage d'une chaîne de caractères en fonction des phonèmes qui le composent. Ainsi deux chaînes de caractères dont l'orthographe est différente mais pas la prononciation phonétique vont pouvoir correspondre. L'outil le plus connu de l'analyse phonétique est l'algorithme d'indexation Soundex : il code les chaînes de caractères ne contenant que des lettres de l'alphabet de la façon suivante (pour la version française) : la première lettre de la chaîne est conservée, la chaîne est mise en majuscule et les voyelles sont supprimées ainsi que les lettres W et H. Pour la version française, les consonnes sont rangées parmi 9 classes notées de 1 à 9 (hormis la première lettre qui n'est pas codée). Après suppression des occurrences consécutives de chiffres, l'algorithme conserve la première lettre puis les trois premiers chiffres du codage (complété par des zéros si nécessaire). Ce type d'algorithme est très efficace pour corriger des erreurs apparaissant sur des chaînes de caractères correspondant à des noms et des prénoms. Le Soundex n'est cependant pas exempt de défauts, et il existe d'autres algorithmes (voir par exemple Taft (1970), Gill (1999) et Philips (1990).

Pour les bases de données administratives utilisées en Santé publique, il arrive souvent que les clés de couplage soient les noms et prénoms, l'adresse, les dates (naissances, décès, soins), les départements et les communes (naissances, décès, domiciles, soins). Or les indices ci-dessus seront surtout utiles pour des classes de champs cités à savoir les nom et prénom d'un côté

 ISSN: 2102-6238

et l'adresse de l'autre. Les similarités pour des champs numériques sont peu traitées dans la littérature, dans la plupart des cas les méthodes précédemment citées sont utilisées. Agrawal et Srikant (2002) préconisent d'adapter la similarité WHIRL présentée ci-dessus, pour exploiter les valeurs numériques autrement que comme des chaînes de caractères. Jaro propose aussi un indice Jaro (1989) pour le cas précis de l'âge.

Les mesures de similarité de type Jaro, Gram et WHIRL font partie des objets plus connus sous le nom de noyaux. En statistique et en apprentissage statistique, les noyaux sont des objets permettant de linéariser des problèmes d'estimation et de classification non linéaires. On trouvera dans Shawe-Taylor et Cristianini (2004) une présentation complète des méthodes de noyaux.

Annexe B : Estimation du taux de faux positif

Voici deux approches pour estimer le taux des appariés à tort (il est possible d'avoir les mêmes résultats pour les taux de non appariés à tort), les deux propositions sont données dans Larsen et Rubin (2001). La première est tout simplement l'application de la théorème de Bayes :

$$\mathbb{P}(U|\widetilde{M}) = \frac{\mathbb{P}(\widetilde{M}|U)\mathbb{P}(U)}{\mathbb{P}(\widetilde{M})} \tag{32}$$

Ce qui sous entend que le modèle utilisé est bien ajusté aux données (variable latente avec plus de deux modalités, dépendance conditionnelle). La seconde est une application de la loi des grands nombres :

$$\mathbb{P}(U|\widetilde{M}) = \frac{\sharp U \cap \widetilde{M}}{\sharp \widetilde{M}} = \frac{\sum_{\delta(a,b)\in\widetilde{M}} \left(1 - \mathbb{P}_{\hat{\theta}}(G=1|\delta(a,b)\right)}{\widetilde{n}} \tag{33}$$

La seconde proposition est la probabilité qu'un élément de $\widetilde{M}$ soit une paire d'individus différent.

Annexe C : Ajustement et correction des modèles pour prendre en compte l'incertitude de l'appariement

.1. *La méthode de Scheuren et Winkler (1993)*

L'estimation des paramètres (β, σ^2) du modèle linéaire simple est :

$$\hat{\beta}_{SW} = (X^tX)^{-1}X^tZ - (X^tX)^{-1}X^t\widehat{Biais} \tag{34}$$

$$\widehat{\sigma^2}_{SW} = \frac{\left(Z - \widehat{Biais} - X\hat{\beta}_{SW}\right)^t \left(Z - \widehat{Biais} - X\hat{\beta}_{SW}\right)}{(n-p-1)} \tag{35}$$

$$\widehat{Var}\left(\hat{\beta}_{SW}\right) = \widehat{\sigma^2}_{SW}(X^tX)^{-1} \tag{36}$$

Où $\widehat{Biais}$ est une estimation du biais que l'on calcule en choisissant pour chaque ligne de la matrice P les deux plus grandes probabilités d'association que l'on va noter p_{ij_1} et p_{ij_2} (respectivement la plus grande et la seconde plus grande).

$$\widehat{Biais}_i = (p_{ij_1} - 1)\, z_{j_1} + p_{ij_2} \tag{37}$$

 ISSN: 2102-6238

.2. *La méthode de Lahiri et Larsen (2005)*

L'estimation des paramètres est donnée par :

$$V = P^t X \tag{38}$$
$$\tilde{\beta}_{LL} = (V^t V)^{-1} V^t Z \tag{39}$$

Il y a deux cas de figure. Le premier suppose que les paramètres du modèle de mélange utilisés pour le couplage sont connus. Dans ce cas l'estimation de la variance est obtenue par :

$$var(\tilde{\beta}_{LL}) = (V^t V)^{-1} V^t \Sigma(Z) V (V^t V)^{-1} \tag{40}$$

$\Sigma(Z)$ est la matrice de variance-covariance du vecteur aléatoire Z dont l'expression est la suivante :

$$Var(Z_i) = \sigma^2 + \beta^t A_i \beta \text{ où } A_i = \sum_{j=1}^{n} p_{ij}(X_j - V_i)(X_j - V_i)^t \tag{41}$$

$$Cov(Z_i, Z_j) = \beta^t A_{ij} \beta \text{ où } A_{ij} = \sum_{l=1}^{n} \sum_{k=1, k\neq l}^{n} p_{iu} p_{jk}(X_l - V_i)(X_k - V_j)^t \tag{42}$$

Pour estimer ces quantités il suffira donc de remplacer (β, σ^2) par leur estimation. L'estimation de σ^2 est :

$$\widehat{\sigma^2}_{LL} = \max\left(0, \frac{S^2 - trace\left(\widehat{H} - V(V^t V)^{-1} V^t \widehat{H}\right)}{n - p}\right) \tag{43}$$
$$S^2 = Z^t \left(Id - V(V^t V)^{-1} V^t\right) \tag{44}$$
$$\widehat{H}_{ii} = \hat{\beta}_{LL}^t A_i \hat{\beta}_{LL} \tag{45}$$
$$\widehat{H}_{ij} = \hat{\beta}_{LL}^t A_{ij} \hat{\beta}_{LL} \tag{46}$$

Enfin dans le cas où les paramètres du mélange sont inconnus, on peut remplacer les paramètres du mélange par une estimation de ces paramètres. Afin de prendre en compte l'incertitude liée au couplage, les auteurs proposent un bootstrap paramétrique comme suit :

1. Estimer les paramètres du mélange que l'on va noter $\hat{\theta}$.
2. Produire B échantillons des variables du mélange en remplaçant θ par $\hat{\theta}$.
3. Pour chaque échantillon b, estimer un nouveau θ_b.
4. Réaliser le couplage et estimer pour chaque échantillon b le coefficient $\hat{\beta}_{LL}(\hat{\theta}_b)$.

L'estimation de la variance est :

$$var(\beta)_{boot} \approx \frac{1}{B}\sum_{b=1}^{B} \widehat{Var}\left(\hat{\beta}_{LL}(\hat{\theta}_b)\right) + \frac{1}{B}\sum_{b=1}^{B} \left(\hat{\beta}_{LL}\left(\hat{\theta}_b\right) - \hat{\beta}_{LL}\left(\hat{\theta}\right)\right)\left(\hat{\beta}_{LL}\left(\hat{\theta}_b\right) - \hat{\beta}_{LL}\left(\hat{\theta}\right)\right)^t \tag{47}$$

 ISSN: 2102-6238

Enfin dans le cadre de la régression logistique l'estimation itérative du paramètre β devient :

$$\left\{\begin{array}{rcl} \beta_{LL}^{(n)} & = & \beta_{LL}^{(n-1)} + (X^t P^t DPX)^{-1} X^t P^t \left[Z - \rho\left(X, \beta_{LL}^{(n-1)}\right)\right] \\ \rho\left(X, \beta_{LL}^{(n-1)}\right) & = & \left[f\left(PX_1\beta_{LL}^{(n-1)}\right) ... f\left(PX_{n_A}\beta_{LL}^{(n-1)}\right)\right]^t \\ D & = & \text{diag}\left(f\left(PX_i\beta_{LL}^{(n-1)}\right)\left(1 - f\left(PX_i\beta_{LL}^{(n-1)}\right)\right) | i = 1, ..., n_A\right) \end{array}\right. \tag{48}$$

Où f est la fonction logistique.

.3. *L'algorithme EM pour le modèle de Hof et Zwinderman (2015)*

On se place dans le cadre du modèle (21). on supposera que $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$.

Étape "E"

$$\begin{array}{rcl} logit\left(\mathbb{P}\left(S_{ab} = 1 | y_b, x_a, \delta(a,b); \theta^{(n-1)}\right)\right) & = & \log\left[\dfrac{L(y_b|\theta_1^{(n-1)}, x_a, S_{ab} = 1)}{L(y_b|\theta_2^{(n-1)}, x_a, S_{ab} = 0)}\right] \\ & + & \log\left[\dfrac{L(x_a|\theta_3^{(n-1)}, S_{ab} = 1)}{L(x_a|\theta_4^{(n-1)}, S_{ab} = 0)}\right] \\ & + & \log\left[\dfrac{\varpi_{ab1}}{\varpi_{ab0}}\right] \end{array} \tag{49}$$

Étape "M"

$$\max_{\theta_1} \sum_{a \in E_A} \sum_{b \in E_B} \{\log[L(y_b|\theta_1, x_a, S_{ab} = 1)]\} \mathbb{P}\left(S_{ab} = 1 | y_b, x_a, \delta(a,b); \theta^{(n-1)}\right) \tag{50}$$

$$\max_{\theta_2} \sum_{a \in E_A} \sum_{b \in E_B} \{\log[L(y_b|\theta_2, S_{ab} = 0)]\} \mathbb{P}\left(S_{ab} = 0 | y_b, x_a, \delta(a,b); \theta^{(n-1)}\right) \tag{51}$$

$$\max_{\theta_3} \sum_{a \in E_A} \sum_{b \in E_B} \{\log[L(x_a|\theta_3, S_{ab} = 1)]\} \mathbb{P}\left(S_{ab} = 1 | y_b, x_a, \delta(a,b); \theta^{(n-1)}\right) \tag{52}$$

$$\max_{\theta_4} \sum_{a \in E_A} \sum_{b \in E_B} \{\log[L(x_a|\theta_4, S_{ab} = 0)]\} \mathbb{P}\left(S_{ab} = 0 | y_b, x_a, \delta(a,b); \theta^{(n-1)}\right) \tag{53}$$

 ISSN: 2102-6238