HAL
open science

Evaluation et amélioration des méthodes de chaînage de données

Xinran Li

To cite this version:

Xinran Li. Evaluation et amélioration des méthodes de chaînage de données. Médecine humaine et pathologie. Université d'Auvergne - Clermont-Ferrand I, 2015. Français. NNT : 2015CLF1MM02 . tel-01244375

HAL Id: tel-01244375

https://theses.hal.science/tel-01244375

Submitted on 15 Dec 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UMR 6284 UdA – CNRS

Université d'Auvergne – Clermont I

ÉCOLE DOCTORALE DES SCIENCES DE LA VIE, SANTE, AGRONOMIE, ENVIRONNEMENT

Année 2015 N° d'ordre :

THÈSE

pour l'obtention du grade de

DOCTEUR D'UNIVERSITÉ

Spécialité : BIOSTATISTIQUE et INFORMATIQUE MEDICALE

Présentée et soutenue publiquement par

Xinran LI

le 29 janvier 2015

Évaluation et amélioration des méthodes de chaînage de données

---

Sous la direction de

M. Jean-Yves Boire

M. Lemlih Ouchchane

Jury :

M. Jean-Yves BOIRE	Directeur	PU-PH, Université d'Auvergne
M. Lemlih OUCHCHANE	Directeur	MCU-PH, Université d'Auvergne
M. Claude DUBRAY	Examinateur	PU-PH, Université d'Auvergne
M. Jacques DEMONGEOT	Examinateur	PU-PH, Université de Grenoble
M. Pascal STACCINI	Rapporteur	PU-PH, Université de Nice
M. Jean-Charles DUFOUR	Rapporteur	MCU-PH, Université d'Aix-Marseille

UMR 6284 UdA – CNRS

Université d'Auvergne – Clermont I

ÉCOLE DOCTORALE DES SCIENCES DE LA VIE, SANTE, AGRONOMIE, ENVIRONNEMENT

Année 2015 N° d'ordre :

THÈSE

pour l'obtention du grade de

DOCTEUR D'UNIVERSITÉ

Spécialité : BIOSTATISTIQUE et INFORMATIQUE MEDICALE

Présentée et soutenue publiquement par

Xinran LI

le 29 janvier 2015

Évaluation et amélioration des méthodes de chaînage de données

---

Sous la direction de

M. Jean-Yves Boire

M. Lemlih Ouchchane

Jury :

M. Jean-Yves BOIRE	Directeur	PU-PH, Université d'Auvergne
M. Lemlih OUCHCHANE	Directeur	MCU-PH, Université d'Auvergne
M. Claude DUBRAY	Examinateur	PU-PH, Université d'Auvergne
M. Jacques DEMONGEOT	Examinateur	PU-PH, Université de Grenoble
M. Pascal STACCINI	Rapporteur	PU-PH, Université de Nice
M. Jean-Charles DUFOUR	Rapporteur	MCU-PH, Université d'Aix-Marseille

Remerciements

Je tiens tout d'abord à remercier le Professeur Jean-Yves BOIRE pour avoir accepté de diriger cette thèse, pour m'avoir permis de l'effectuer dans les meilleures conditions possibles au sein de son laboratoire, et enfin pour m'avoir soutenu et encouragé tout au long de ces trois années de travail.

Je remercie très sincèrement le Docteur Lemlih OUCHCHANE pour m'avoir dirigé pendant ce travail de thèse et pour m'avoir accordé sa confiance dans le projet GINSENG. Je le remercie également pour m'avoir fait bénéficier de ses connaissances scientifiques et de ses méthodes de travail, et aussi pour sa disponibilité, son sens du dialogue et tout le temps qu'il a su me consacrer.

Je voudrais remercier le Professeur Pascal STACCINI et le Docteur Jean-Charles DUFOUR d'avoir accepté d'être rapporteurs de cette thèse. Je les remercie pour le temps qu'ils ont consacré et pour leurs remarques constructives qui ont permis d'améliorer ce manuscrit.

Je voudrais également exprimer mes sincères remerciements au Professeur Jacques DEMONGEOT pour sa participation active et constructive aux valorisations scientifiques de cette thèse. Je suis également honoré de sa présence au sein de ce jury.

Je tiens à remercier tout particulièrement le Professeur Claude DUBRAY pour me faire l'honneur de présider ce jury.

Je remercie le Docteur Aline GUTTMANN pour m'avoir fait partager ses connaissances, particulièrement en programmation R. J'ai eu plaisir à collaborer avec elle dans le cadre du projet GINSENG.

Je remercie particulièrement Madame Sylvie ROUX pour sa gentillesse, son efficacité et son soutien indéfectible, notamment dans les arcanes administratives, et aussi et surtout pour le temps qu'elle a consacré à la lecture et relecture de ce manuscrit pour corriger mes fautes de français.

Je remercie Madame Stéphanie LEGER, initialement responsable de mon master, que j'ai eu plaisir à retrouver dans nos collaborations autour du projet GINSENG. Je lui suis reconnaissant d'avoir consolidé mes bases théoriques en Statistique, mises à profit, je l'espère, au cours de cette thèse.

Je remercie également mes amis et collègues du service Biostatistique et du projet GINSENG, en particulier Audrey, Cécile, Emilie, Fatma, Jia, Jian, Johann, Juliette, Gaétan, Li, Libo, Sébastien, Wei, pour leurs idées, leurs aides et leurs encouragements.

Je tiens à exprimer ma gratitude et mon affection à mes parents qui m'ont soutenu durant toutes ces années. Je réserve des remerciements tout particuliers à Rui, mon épouse, pour sa patience, et le soutien qu'elle a su m'apporter depuis le début de bien des aventures dont celle-ci. Et enfin, je dédie ce travail à notre fille Zhixia, pour la joie qu'elle nous apporte.

Résumé

Le chaînage d'enregistrements est la tâche qui consiste à identifier parmi différentes sources de données les enregistrements qui concernent les mêmes entités. En l'absence de clé d'identification commune, cette tâche peut être réalisée à l'aide d'autres champs contenant des informations d'identifications, mais dont malheureusement la qualité n'est pas parfaite. Pour ce faire, de nombreuses méthodes dites « de chaînage de données » ont été proposées au cours des dernières décennies.

Afin d'assurer le chaînage valide et rapide des enregistrements des mêmes patients dans le cadre de GINSENG, projet qui visait à mettre en place une infrastructure de grille informatique pour le partage de données médicales distribuées, il a été nécessaire d'inventorier, d'étudier et parfois d'adapter certaines des diverses méthodes couramment utilisées pour le chaînage d'enregistrements. Citons notamment les méthodes de comparaison approximative des champs d'enregistrement selon leurs épellations et leurs prononciations, les chaînages déterministe et probabiliste d'enregistrements, ainsi que leurs extensions. Ces méthodes comptent des avantages et des inconvénients qui sont ici clairement exposés.

Dans la pratique, les champs à comparer étant souvent imparfaits du fait d'erreurs typographiques, notre intérêt porte particulièrement sur les méthodes probabilistes de chaînage d'enregistrements. L'implémentation de ces méthodes probabilistes proposées par Fellegi et Sunter (PRL-FS) et par Winkler (PRL-W) est précisément décrite, ainsi que leur évaluation et comparaison. La vérité des correspondances des enregistrements étant indispensable à l'évaluation de la validité des résultats de chaînages, des jeux de données synthétiques sont générés dans ce travail et des algorithmes paramétrables proposés et détaillés.

Bien qu'à notre connaissance, le PRL-W soit une des méthodes les plus performantes en termes de validité de chaînages d'enregistrements en présence d'erreurs typographiques dans les champs contenant les traits d'identification, il présente cependant quelques caractéristiques perfectibles. Le PRL-W ne permet par exemple pas de traiter de façon satisfaisante le problème de données manquantes. Notons également qu'il s'agit d'une méthode dont l'implémentation n'est pas simple et dont les temps de réponse sont difficilement compatibles avec certains usages de routine. Certaines solutions ont été proposées et évaluées pour pallier ces difficultés, notamment plusieurs approches permettant d'améliorer l'efficacité du PRL-W en présence de données manquantes et d'autres destinées à optimiser les temps de calculs de cette méthode en veillant à ce que cette réduction du temps de traitement n'entache pas la validité des décisions de chaînage issues de cette méthode.

Mots clés : chaînage d'enregistrements, comparaison approximative du champ d'enregistrement, méthodes probabilistes, évaluation

Abstract

Record linkage is the task of identifying which records from different data sources refer to the same entities. Without the common identification key among different databases, this task could be performed by comparison of corresponding fields (containing the information for identification) in records to link. To do this, many record linkage methods have been proposed in the last decades.

In order to ensure a valid and fast linkage of the same patients' records for GINSENG, a research project which aimed to implement a grid computing infrastructure for sharing medical data, we first studied various commonly used methods for record linkage. These are the methods of approximate comparison of fields in record according to their spellings and pronunciations; the deterministic and probabilistic record linkages and their extensions. The advantages and disadvantages of these methods are clearly demonstrated.

In practice, as fields to compare are sometimes subject to typographical errors, we focused on probabilistic record linkage. The implementation of these probabilistic methods proposed by Fellegi and Sunter (PRL-FS) and Winkler (PRL-W) is described in details, and also their evaluation and comparison. Synthetic data sets were used in this work for knowing the truth of matches to evaluate the linkage results. A configurable algorithm for generating synthetic data was therefore proposed.

To our knowledge, the PRL-W is one of the most effective methods in terms of validity of linkages in the presence of typographical errors in the field. However, the PRL-W does not satisfactorily treat the missing data problem in the fields, and the implementation of PRL-W is complex and has a computational time that impairs its opportunity in routine use. Solutions are proposed here with the objective of improving the effectiveness of PRL-W in the presence of missing data in the fields. Other solutions are tested to simplify the PRL-W algorithm and both reduce computational time and keep and optimal linkage accuracy.

Keywords: record linkage, approximate comparison of record field, probabilistic methods, evaluation

Table des matières

Remerciements 5
Résumé 6
Abstract 7
Table des matières 8
Table des tableaux 12
Table des figures 13
Introduction générale 14
Chapitre 1 Etat de l'art sur les techniques de chaînage d'enregistrements 17
1.1. Comparaison approximative des champs d'enregistrements 18
1.1.1. Méthode de Jaro-Winkler 18
1.1.1.1. Distance de Jaro 18
1.1.1.2. Distance de Jaro-Winkler 18
1.1.2. Méthode de Levenshtein 19
1.1.2.1. Distance de Levenshtein 19
1.1.2.2. Similarité de Levenshtein 20
1.1.3. Version française du Soundex 21
1.2. Comparaison des couples d'enregistrements 21
1.2.1. Chaînage déterministe des enregistrements 22
1.2.1.1. Comparaison de tous les champs disponibles 22
1.2.1.2. Comparaison d'une partie des champs disponibles 23
1.2.1.3. Comparaison approximative de tous les champs disponibles 23
1.2.2. Chaînage probabiliste des enregistrements 24
1.2.2.1. Méthode de Fellegi et Sunter 24
1.2.2.2. Méthode de Winkler 27
1.2.2.3. Méthode de DuVall 29
Chapitre 2 Implémentation et évaluation des méthodes de chaînage existantes 32
2.1. Création des données synthétiques servant à l'évaluation des méthodes de chaînage 34
2.1.1. Introduction 34
2.1.2. Matériel et méthodes 35
2.1.2.1. Approche globale de genèse des jeux de données synthétiques 35
2.1.2.2. Préparation des données source 36
2.1.2.3. Génération d'un jeu d'enregistrements fictifs 36

2.1.2.4. Création d'un couple de jeux de données synthétiques 37
2.1.2.5. Ajout d'erreurs typographiques aux jeux de données synthétiques 37
2.1.3. Résultats 39
Description des types et taux d'erreurs observés dans les données synthétiques 39
Exemple d'enregistrements synthétiques 40
Temps de calcul nécessaire au processus de génération des données 41
2.1.4. Discussion et Perspectives 41
2.1.5. Conclusion 42
2.2. Implémentation de la méthode chaînage probabiliste des enregistrements d'après Fellegi-Sunter 43
2.2.1. Introduction 43
2.2.2. Matériel et méthodes 44
2.2.2.1. Approche globale pour implémenter le PRL-FS 44
2.2.2.2. Construction des couples d'enregistrements 44
2.2.2.3. Calcul des poids de chaque couple d'enregistrements 44
2.2.2.4. Décision de chaînage des enregistrements 45
2.2.2.5. Estimation des paramètres m, u et p par l'algorithme EM 45
2.2.2.6. Evaluation de l'implémentation de la méthode PRL-FS 47
2.2.3. Résultats 47
2.2.3.1. Réalisation d'un chaînage par la méthode PRL-FS 47
2.2.3.2. Evaluation de l'exactitude de l'estimation des paramètres 48
2.2.3.3. Evaluation du résultat de chaînage par la méthode PRL-FS 49
2.2.4. Discussion et Perspective 51
2.2.5. Conclusion 52
2.3. Implémentation de la méthode chaînage probabiliste des enregistrements d'après Winkler 53
2.3.1. Introduction 53
2.3.2. Matériel et méthodes 54
2.3.2.1. Jeux de données 54
2.3.2.2. Calcul des poids de chaque couple d'enregistrements 54
2.3.2.3. Décision de chaînages des enregistrements 56
2.3.2.4. Estimation des paramètres par l'algorithme EM 56
2.3.2.5. Evaluation de l'exactitude de l'estimation des paramètres et de la performance de la méthode PRL-W 58
2.3.3. Résultats 58
2.3.3.1. Exactitude de l'estimation des paramètres requis par le PRL-W 58

2.3.3.2. Performance du PRL-W 61
2.3.4. Discussion et Perspective 63
2.3.5. Conclusion 64
Chapitre 3 Adaptation et amélioration des méthodes de chaînage existantes 65
3.1. Amélioration de la méthode probabiliste d'après Winkler face aux données manquantes dans les champs 66
3.1.1. Contexte 66
3.1.2. Objectifs 67
3.1.3. Matériel et méthodes 67
3.1.3.1. Approche globale 67
3.1.3.2. Genèse des jeux de données 67
3.1.3.3. Calcul des poids du couple d'enregistrements par le PRL-W original 68
3.1.3.4. Calcul des poids du couple d'enregistrements par le PRL-W avec le traitement des données manquantes 68
3.1.3.5. Décision de chaînages des enregistrements 70
3.1.4. Résultats 70
3.1.4.1. Comparaison des nombres de faux négatifs et de faux positifs 71
3.1.4.2. Comparaison des capacités de chaque solution pour réduire le nombre de mauvaises décisions 72
3.1.4.3. Comparaison des temps de calcul 72
3.1.5. Discussion 73
3.1.6. Conclusion 75
3.2. Méthode alternative pour le calcul du poids des enregistrements en considérant la similarité des champs 76
3.2.1. Contexte 76
3.2.2. Objectifs 76
3.2.3. Matériel et méthodes 76
3.2.3.1. Jeux de données 76
3.2.3.2. Définition de la fonction de calcul du poids du champ 77
3.2.3.3. Décision de chaînages des enregistrements 78
3.2.3.4. Evaluation et comparaison des méthodes PRL 78
3.2.4. Résultats 79
3.2.4.1. Comparaison des nombres de faux négatifs et de faux positifs 79
3.2.4.2. Comparaison des temps de calcul 80
3.2.5. Discussion 81

3.2.6. Conclusion 82
3.3. Proposition d'une méthode de chaînage par la combinaison linéaire des scores de similarité de chaque champ 83
3.3.1. Introduction 83
3.3.2. Matériel et méthodes 83
3.3.2.1. Jeux de données 83
3.3.2.2. Calcul des scores de similarité des champs 83
3.3.2.3. Poids de concordance des champs dans le PRL-FS 84
3.3.2.4. Calcul des poids du couple d'enregistrements 84
3.3.2.5. Décision de chaînages des enregistrements 84
3.3.2.6. Evaluation et comparaison des méthodes PRL 84
3.3.3. Résultats 85
3.3.4. Discussion et Conclusion 85
Conclusion générale 87
Bibliographie 89
Liste des publications et des communications 93

Table des tableaux

Tableau 1.1 Tableau de correspondance pour le codage Soundex 21
Tableau 1.2 Exemple des enregistrements provenant des bases de données A et B pour tester le DRL 23
Tableau 1.3 Exemple d'enregistrements provenant des 2 bases de données A et B pour tester le PRL 25
Tableau 1.4 Exemple des probabilités m_i et u_i estimées par l'algorithme EM pour le PRL-FS..... 26
Tableau 1.5 Extrait des probabilités mi, sk, i et ui, sk, iestimées par l'algorithme EM pour le PRL-W. La comparaison du champ *sexe* étant binaire, $m[0, 0.6]$ correspond à $m\{0\}$, et $m[0.98, 1]$ correspond à $m\{1\}$ 28
Tableau 1.6 Extrait des probabilités mi, d et ui, destimées par l'algorithme EM pour la méthode de DuVall, valeurs à titre d'exemple. La comparaison du champ *sexe* étant binaire, $m3,0$ correspond à l'identité des sexes, et $m3,1$ à leur différence..... 30

Tableau 2.1 Types d'erreurs introduites dans les jeux de données synthétiques 37
Tableau 2.2 Distribution des erreurs dans les champs 40
Tableau 2.3 Exemple des enregistrements générés 41
Tableau 2.4 Temps moyens pour 100 réalisations du processus de synthèse de données..... 41
Tableau 2.5 Paramètres estimés et poids de concordance et de discordance calculés en utilisant ces paramètres 47
Tableau 2.6 Tous les patterns possibles pour les résultats de comparaisons des champs et leurs fréquences (classés en ordre décroissant selon les poids des couples d'enregistrements)..... 48
Tableau 2.7 PRL-FS : différences entre paramètres estimés et observés sur 100 réalisations..... 49
Tableau 2.8 Nombres cumulés de 4 catégories de résultats (VP, FP, FN et VN) par les méthodes DRL et PRL-FS (100 réalisations)..... 50
Tableau 2.9 Sous-intervalles du JWSS pour chaque champ..... 56

Tableau 3.1 Les comparaisons appariées des cinq stratégies de chaînage d'enregistrements à l'égard des nombres de mauvaises décisions en utilisant les t-tests avec des corrections de Holm 72
Tableau 3.2 Fréquence cumulée multivariée des décisions de chaînage pour les mêmes couples d'enregistrements générées par le PRL-W par rapport au PRL-W avec des solutions traitant des données manquantes dans 100 processus (BD : bonnes décisions, MD : mauvaises décisions) 72
Tableau 3.3 Comparaisons appariées des cinq stratégies de chaînage d'enregistrements à l'égard des temps de calcul en utilisant les t-tests avec des corrections de Holm..... 73
Tableau 3.4 Les comparaisons appariées des 3 méthodes PRL à l'égard des nombres de mauvaises décisions en utilisant les t-tests avec des corrections de Holm..... 79
Tableau 3.5 Les comparaisons appariées des 3 méthodes PRL à l'égard du temps de calcul en utilisant les t-tests avec des corrections de Holm..... 81

Table des figures

Figure 2.1 Principes du processus de chaînage d'enregistrements 35
Figure 2.2 Processus de création d'un couple de jeux de données servant à l'évaluation des méthodes de chaînage d'enregistrements 36
Figure 2.3 Echantillons de noms, prénoms rattachés aux sexes, dates de naissance 37
Figure 2.4 Exemple de substitution d'une lettre par une des lettres voisines sur clavier informatique 39
Figure 2.5 Protocole d'évaluation de l'implémentation de la méthode de chaînage PRL-FS 44
Figure 2.6 Nombres de mauvaises décisions dans chacun des 100 processus de chaînage par les méthodes DRL et PRL-FS 50
Figure 2.7 Temps de calcul du DRL et du PRL-FS sur 100 réalisations 50
Figure 2.8 Paramètres estimés et observés de chaque sous-intervalle des champs nom, prénom, sexe et date de naissance dans un processus de chaînage d'enregistrements, voir le tableau 2.9 pour les valeurs des sous-intervalles 59
Figure 2.9 Paramètres *m* estimés et observés pour le champ « nom » ayant un JWSS dans [0.90, 0.92[en 100 estimations 59
Figure 2.10 Paramètres p estimés et observés dans 100 estimations 60
Figure 2.11 Relation entre l'inexactitude des estimations et les mauvaises décisions. Les lignes bleue et rouge sont respectivement les courbes de tendance polynomiale 2ème ordre pour les nombres de faux négatifs et de faux positifs dans les 100 processus 61
Figure 2.12 Nombres de faux positifs et de faux négatifs engendrés par les deux méthodes PRL dans 100 processus de chaînages 62
Figure 2.13 Temps de calcul des deux méthodes PRL dans 100 processus de chaînages 62

Figure 3.1 Les distributions des nombres de faux négatifs et de faux positifs générés par le PRL-W et les PRL-W avec les trois solutions pour traiter le problème de données manquantes dans 100 exécutions 71
Figure 3.2 Distributions des temps de calcul utilisant le PRL-W et le PRL-W avec trois solutions de traitement des données manquantes dans 100 processus 73
Figure 3.3 Distributions de nombres de faux négatifs et de faux positifs par les méthodes PRL-FS, PRL-W et PRL-I dans 100 processus 79
Figure 3.4 Nombres de faux négatifs et de faux positifs par les méthodes PRL-FS, PRL-W et PRL-I dans 100 processus 80
Figure 3.5 Distribution des temps de calcul des trois méthodes PRL dans 100 processus 80
Figure 3.6 Temps de calcul des trois méthodes PRL dans 100 processus 81
Figure 3.7 Nombres de mauvaises décisions par les méthodes PRL-FS, PRL-W et RL-CS dans 100 processus 85
Figure 3.8 Temps de calcul des méthodes PRL-FS, PRL-W et RL-CS dans 100 processus 85

Introduction générale

Le chaînage d'enregistrements (*record linkage*) est la tâche qui consiste à identifier parmi différentes sources de données les enregistrements qui concernent les mêmes entités. Ces entités peuvent être des individus, par exemple des patients, clients ou employés, mais elles peuvent aussi bien représenter des produits, des commandes ou des documents, et en fait cette tâche est utile dans de nombreux domaines. Le domaine médical est tout particulièrement intéressé par le chaînage d'enregistrements car la combinaison des informations d'un même patient provenant de différentes sources –afin par exemple de construire une image plus complète de son parcours de soins– est essentielle à la description fine de sa prise en charge ou, plus largement, aux études épidémiologiques [1].

Dans le cadre d'un projet de l'Agence Nationale de la Recherche (ANR-10-TECS-0008 GINSENG[1]), destiné à mettre en place une infrastructure de grille informatique pour le partage de données médicales et la réalisation d'études épidémiologiques dans la région Auvergne, nous avons été amenés à fédérer des bases de données médicales distribuées. Il s'agit notamment des bases de données des associations de dépistage des cancers ARDOC[2] et ABIDEC[3], des laboratoires d'anatomo-cytopathologie Unilabs-Sipath, et du réseau de santé périnatale d'Auvergne, pour ne citer que ceux à l'origine du projet. Le chaînage valide et rapide des enregistrements d'un même patient dispersés dans ces diverses bases de données constitue donc un verrou primordial.

Le défi majeur dans le chaînage d'enregistrements tient à l'absence de clé d'identification commune parmi différentes bases de données à chaîner [2], ce qui est précisément le cas dans le contexte de notre projet où aucune interopérabilité n'était préalablement planifiée entre les bases. Pour pouvoir réaliser un chaînage dans de telles conditions, il est nécessaire d'utiliser les champs (*fields*) d'enregistrements contenant les informations d'identifications, tels que nom, prénom, sexe et date de naissance. L'autorisation CNIL pour l'utilisation de ces informations d'identifications, avec préservation de la confidentialité, et à des fins de validation des identités et de chaînage des données dans le cadre de ce projet, a été obtenue en novembre 2013 (délibération n° 2013-364).

Plusieurs options méthodologiques sont disponibles pour assurer cette tâche de chaînage d'enregistrements de patients dans le contexte du projet GINSENG ou de façon plus générale.

Une des méthodes les plus simples est le chaînage déterministe d'enregistrements (*Deterministic Record Linkage, DRL*) [3]. Avec cette méthode, la comparaison de chaque paire de champs au sein d'un couple d'enregistrements (*record pair*), chaque enregistrement provenant d'une source de données spécifique, fonde le chaînage sur la base de la concordance de tous les champs comparés. Du fait que la qualité du contenu des champs à comparer n'est pas toujours parfaite, et afin de réduire l'éventualité d'absence de chaînage, les contraintes de chaînage peuvent être assouplies, par exemple, à partir d'une règle de concordance de (*n-1)* champs parmi *n* champs comparés. Toutefois, le crédit accordé à la concordance des champs pour chaîner des enregistrements est variable d'un champ à un autre [4]. Par exemple, la constatation d'une concordance sur le champ *nom* est plus

[1] Global Initiative for Sentinel E-health Network on Grid
[2] Association Régionale des Dépistages Organisés des Cancers
[3] Association Bourbonnaise Interdépartementale de Dépistage des Cancers

vraisemblablement associée au fait qu'il s'agit de deux enregistrements concernant la même personne que la constatation d'une concordance sur le champ *sexe*. Malheureusement, le chaînage déterministe d'enregistrements ne prend pas en compte cette différence de contribution entre les différents champs disponibles.

C'est en 1959 que Newcombe et *al.* ont proposé pour la première fois une approche probabiliste en utilisant des *odds ratios* pour quantifier les contributions respectives d'une concordance ou d'une discordance de chaque champ dans le contexte du chaînage d'enregistrements [5]. En 1969, Fellegi et Sunter ont formalisé sur de solides bases mathématiques l'idée de Newcombe, entérinant le concept de chaînage probabiliste d'enregistrements (*Probabilistic Record Linkage, PRL*) [6]. Dans cette approche PRL, un poids (*weight*) de concordance et un poids de discordance, fondés sur des rapports de vraisemblance, sont attribués à chaque champ. Un poids est également calculé pour un couple d'enregistrements en additionnant les poids de concordance ou de discordance de tous les champs de ce couple. Pour tout couple d'enregistrements, plus son poids est élevé, plus il est vraisemblable qu'il s'agisse du même patient, la décision de chaînage pouvant ainsi être fondée sur l'application d'une règle de seuillage. Avec leur papier princeps « *A Theory For Record Linkage* », Fellegi et Sunter ont constitué les fondements mathématiques de la plupart des méthodes probabilistes de chaînage d'enregistrements, et ce jusqu'à aujourd'hui. En pratique, la méthode originelle de Fellegi et Sunter (PRL-FS) est toujours une des méthodes les plus couramment utilisées, et a été implémentée dans de nombreux logiciels libres [7].

Toutefois, les champs à comparer sont parfois, sinon souvent, soumis à des dégradations à type d'erreurs typographiques telles que les fautes d'épellation [8]. Ces dégradations rendent la méthode PRL-FS moins efficace. Par exemple, pour le champ *prénom* dans deux couples d'enregistrements tels que ("*Philippe*", "*Philipe*") et ("*Philippe*", "*Nicolas*"), la même contribution à type de « pénalité » (poids de discordance) est attribuée pour ces deux couples de prénoms puisqu'ils sont tous deux jugés discordants. Le fait que la discordance dans le premier couple soit très vraisemblablement causée par une simple erreur d'épellation est donc totalement ignorée.

Pour améliorer le PRL-FS en palliant cet inconvénient, Winkler a proposé en 1990 une extension du PRL-FS en incorporant des mesures de similarité entre chaînes de caractères dans les champs à comparer (PRL-W), ajoutant la démonstration de son gain de performance par rapport au PRL-FS [9]. A notre connaissance, le PRL-W est une des méthodes les plus performantes en termes de validité des décisions de chaînage d'enregistrements en présence d'erreurs typographiques dans les champs. Cependant, on note que cette méthode a très rarement été implémentée et qu'elle majore les temps de calculs par rapport au PRL-FS.

L'étude des diverses méthodes de chaînage aboutit invariablement à l'inventaire de certains inconvénients, qu'ils soient d'ordre théorique ou pratique. On peut citer l'exemple précis du fait que le PRL-W ne permet pas de traiter de façon satisfaisante le problème des champs avec données manquantes, situation pourtant fréquente [10]. De façon plus générale, les temps de calculs des méthodes les plus performantes, type PRL, sont longs et sont une difficulté supplémentaire à leur utilisation pratique, notamment en temps-réel, leur faisant préférer des méthodes moins coûteuses mais moins performantes, comme le DRL (éventuellement assorties de règles de décision). Citons également le passage à la décision de chaînage de ces diverses méthodes qui est une étape soit très largement éludée soit, la plupart du temps, fondée de manière empirique et pragmatique.

Dans le contexte du projet GINSENG, la nécessité de fédérer des bases de données médicales distribuées à des fins épidémiologiques, voire d'aide à la décision, nous a donc contraints à mettre en œuvre de façon pratique le chaînage de nos données et à évaluer la validité de ce dernier. Ce travail d'évaluation a été l'occasion de proposer des adaptations des méthodes de chaînage en vue d'améliorer leurs performances en termes de validité et de rapidité. Cette démarche est l'objet de ce travail de thèse qui décrit les étapes d'implémentation et d'évaluation des méthodes existantes mais aussi et surtout de leurs adaptations originales.

Outre cette **introduction générale**, le présent manuscrit est organisé en trois chapitres.

Le **chapitre 1** présente l'état de l'art du chaînage d'enregistrements. Il résume les principaux concepts du domaine selon deux niveaux de comparaisons des informations des patients : les champs d'enregistrements et les enregistrements proprement dits. Dans une première partie, on présente deux algorithmes de mesure de la similarité entre deux chaînes de caractères selon leurs épellations, ainsi qu'un algorithme phonétique d'indexation selon la prononciation de ces chaînes de caractères, et ce afin de comparer leur similarité à l'échelle du champ d'enregistrement. Dans une seconde partie, deux types de chaînage d'enregistrements, l'un déterministe et l'autre probabiliste, ainsi que leurs extensions, sont présentés.

Le **chapitre 2** décrit tout d'abord une approche de genèse des données synthétiques servant à réaliser les études de performances des méthodes de chaînage. Ensuite, on décrit précisément l'implémentation de deux méthodes de chaînage probabiliste d'enregistrements, proposées par Fellegi et Sunter d'une part et par Winkler d'autre part, incorporant toutes deux l'algorithme espérance-maximisation (*expectation-maximisation algorithm, EM*), pour finir par leur évaluation et leur comparaison en termes de validité des décisions de chaînage et de temps de calcul.

Le **chapitre 3** propose des améliorations et des adaptations de méthodes probabilistes de chaînage d'enregistrements. Dans une première partie, différentes solutions sont proposées et implémentées dans l'objectif d'améliorer l'efficacité de la méthode PRL-W en présence de données manquantes dans les champs d'enregistrements. Dans une deuxième partie, une alternative dérivée de la méthode PRL-W est proposée pour le calcul du poids du couple d'enregistrements qui vise la double contrainte d'une exactitude de décision de chaînage très proche de celle du PRL-W, et d'un moindre temps de calcul. Dans une troisième partie, on propose une méthode originale de chaînage d'enregistrements qui combine linéairement les scores de similarités de chaque champ au sein d'un couple d'enregistrements, et ce afin de créer de nouveaux poids pour les couples d'enregistrements dont le gain de valeur informationnelle est ensuite évalué. Chacune des parties de ce chapitre fait l'objet d'une évaluation en termes de validité de décision et de temps de calcul des méthodes proposées et/ou d'une comparaison vis-à-vis des méthodes PRL originelles.

Enfin, ce manuscrit s'achève sur une **conclusion générale**, qui intègre une discussion, notamment des enjeux, des limites ou difficultés, et également des perspectives ouvertes par ce travail de thèse sur le chaînage d'enregistrements.

Chapitre 1

Etat de l’art sur les techniques de chaînage d’enregistrements

1.1. Comparaison approximative des champs d'enregistrements

Comme indiqué en introduction générale, la qualité des informations d'identification des patients contenues dans les champs d'enregistrements à comparer n'est pas toujours strictement valide ou parfaite. Ces champs peuvent contenir des erreurs typographiques, notamment dues à des erreurs d'épellation, avec un taux décrit comme fluctuant entre 8,5% et 36,5% selon la littérature [8,11,12]. Supposons deux enregistrements associés à un seul et même patient, et dont un ou plusieurs champs contiendraient des erreurs typographiques. Sur la base d'une comparaison stricte des champs, et ce même après une standardisation des données, la décision de chaînage pour ces deux enregistrements resterait délicate du fait qu'il peut subsister un ou des champs n'ayant pas une concordance exacte.

Pour réduire cet éventuel risque de défaut de chainage (absence de chainage à tort), plutôt que de comparer de façon stricte les champs correspondants dans un couple d'enregistrements en renvoyant un résultat binaire soit « identique » soit « différent », il s'avère utile d'appliquer une fonction de comparaison approximative des chaînes de caractères qui permet de quantifier la similarité entre les contenus des champs à comparer. Nous présentons ci-après les fonctions de comparaison des chaînes de caractères les plus couramment utilisées.

1.1.1. Méthode de Jaro-Winkler

La mesure de similarité de Jaro-Winkler entre deux chaînes de caractères est particulièrement adaptée à la comparaison de chaînes courtes telles que des noms [13]. Le résultat de cette mesure est représenté par un score normalisé entre 0 et 1, 1 signifiant que les deux chaînes sont identiques, et ce score s'éloignant d'autant de 1 vers 0 que la similarité entre les deux chaînes devient faible. Cette méthode a été introduite par Jaro en 1978 [14,15], puis une variante en a été proposée par Winkler en 1990 [9].

1.1.1.1. Distance de Jaro

La méthode de base de Jaro pour la distance entre chaînes s_1 et s_2 est définie par [16]:

$$d_J = \frac{1}{3}\left(\frac{c}{|s_1|} + \frac{c}{|s_2|} + \frac{c-t}{c}\right) \tag{1.1}$$

où :

$|s_i|$ est la longueur de la chaîne de caractères s_i
c est le nombre de caractères communs entre deux chaînes
t est le nombre de transpositions de caractères entre deux chaînes
S'ajoute une condition pour considérer deux caractères identiques de s_1 et de s_2 comme un caractère commun dans deux chaînes, à savoir que la différence entre les positions de s_1 et de s_2 dans leurs chaînes respectives ne doit pas dépasser la valeur suivante :

$$\frac{\max(|s_1|, |s_2|)}{2} - 1 \tag{1.2}$$

1.1.1.2. Distance de Jaro-Winkler

Pour favoriser les chaînes ayant un préfixe commun, Winkler a ajusté d_J en utilisant deux paramètres l et p, cette nouvelle distance étant définie comme suit [9]:

$$d_{JW} = d_J + (lp(1 - d_J)) \quad (1.3)$$

où :

d_J est la distance de Jaro entre s_1 et s_2
l est la longueur du préfixe commun (4 caractères maximum)
p est le paramètre qui ajuste l'importance du préfixe commun, qui doit être inférieur à 0,25, Winkler ayant utilisé $p = 0{,}1$ dans son travail princeps

Exemples

Exemple 1 :
Soient les 2 chaînes de caractères $s_1 =$ Catherine, $s_2 =$ Katehrina, dont la table de correspondance est construite ci-dessous :

	C	A	T	H	E	R	I	N	E
K	**0**	0	0	0	0	0	0	0	0
A	0	**1**	0	0	0	0	0	0	0
T	0	0	**1**	0	0	0	0	0	0
E	0	0	0	**0**	1	0	0	0	0
H	0	0	0	1	**0**	0	0	0	0
R	0	0	0	0	0	**1**	0	0	0
I	0	0	0	0	0	0	**1**	0	0
N	0	0	0	0	0	0	0	**1**	0
A	0	0	0	0	0	0	0	0	**0**

- $|s_1| = 9$
- $|s_2| = 9$
- $c = 7$ (nombre de 1 dans la zone verte de la table, la zone doit être inférieure à $\max(9{,}9)/2 - 1 = 3{,}5$)
- $t = 1$ (E/H)
- $l = 0$
- $p = 0{,}1$

→$d_{JW}(s_1, s_2)$ =0,8042

Exemple 2 :
Soient 2 chaînes de caractères $s_1 =$ Katrine, $s_2 =$ Katehrina, dont la table de correspondance est construite ci-dessous :

	K	A	T	R	I	N	E
K	**1**	0	0	0	0	0	0
A	0	**1**	0	0	0	0	0
T	0	0	**1**	0	0	0	0
E	0	0	0	**0**	0	0	1
H	0	0	0	0	**0**	0	0
R	0	0	0	1	0	**0**	0
I	0	0	0	0	1	0	**0**
N	0	0	0	0	0	1	0
A	0	0	0	0	0	0	0

- $|s_1| = 7$
- $|s_2| = 9$
- $c = 7$ (nombre de 1 dans la zone verte de la table, la zone doit être inférieure à $\max(7{,}9)/2 - 1 = 3{,}5$)
- $t = 0$
- $l = 3$
- $p = 0{,}1$

→$d_{JW}(s_1, s_2)$ =0,8815

1.1.2. Méthode de Levenshtein

1.1.2.1. Distance de Levenshtein

La distance de Levenshtein, autre mesure de similarité entre deux chaînes de caractères, compte le nombre minimum d'opérations nécessaires pour passer d'une chaîne de caractères à une autre, où une opération correspond à l'insertion, la suppression ou la substitution d'un caractère dans cette chaîne [17].

Pour calculer le nombre minimum d'opérations nécessaires, un algorithme de programmation dynamique (*dynamic programming*) peut être utilisé [18]. Après avoir construit une table de correspondance pour s_1 et s_2 sous forme d'une matrice vide contenant $|s_1| + 1$ lignes et $|s_2| + 1$ colonnes, cet algorithme commence par le remplissage de la première ligne et la première colonne de la matrice avec les numéros de ligne et de colonne correspondants. La cellule $d[0, j]$ (ligne 0 et

colonne j ($0 \leq j \leq |s_1|$)) contient la valeur j, et la cellule $d[i,0]$ (ligne i ($0 \leq i \leq |s_2|$) et colonne 0) contient la valeur i. Les cellules restantes de la matrice sont renseignées par l'approche récursive suivante [7]:

$$[i,j] = \begin{cases} d[i-1,j-1] & si\ s_1[i] = s_2[j] \\ min \begin{cases} d[i-1,j]+1 \\ d[i,j-1]+1 \\ d[i-1,j-1]+1 \end{cases} & si\ s_1[i] \neq s_2[j] \end{cases} \tag{1.4}$$

Exemples

Exemple 1 :
Soient les 2 chaînes de caractères s_1 = Catherine, s_2 = Katehrina. Le calcul de la distance de Levenshtein entre s_1 et s_2 par l'algorithme de programmation dynamique est effectué à partir de la matrice ci-dessous :

		0	1	2	3	4	5	6	7	8	9
			C	**A**	**T**	**H**	**E**	**R**	**I**	**N**	**E**
0		**0**	1	2	3	4	5	6	7	8	9
1	**K**	1	**1**	2	3	4	5	6	7	8	9
2	**A**	2	2	**1**	2	3	4	5	6	7	8
3	**T**	3	3	2	**1**	2	3	4	5	6	7
4	**E**	4	4	3	2	**2**	3	4	5	6	7
5	**H**	5	5	4	3	2	**3**	4	5	6	7
6	**R**	6	6	5	4	3	3	**3**	4	5	6
7	**I**	7	7	6	5	4	4	4	**3**	4	5
8	**N**	8	8	7	6	5	4	3	2	**3**	4
9	**A**	9	9	8	7	6	5	4	3	3	**4**

La cellule en bas à droite de la matrice correspond à la distance de Levenshtein entre s_1 et s_2.
→$d_{Levenshtein}$ = 4

Exemple 2 :
Soient les 2 chaînes de caractères s_1 = Katrine, s_2 = Katehrina. Le calcul de la distance de Levenshtein entre s_1 et s_2 par l'algorithme de programmation dynamique est effectué à partir de la matrice ci-dessous :

		0	1	2	3	4	5	6	7
			K	**A**	**T**	**R**	**I**	**N**	**E**
0		**0**	1	2	3	4	5	6	7
1	**K**	1	**0**	1	2	3	4	5	6
2	**A**	2	1	**0**	1	2	3	4	5
3	**T**	3	2	1	**0**	1	2	3	4
4	**E**	4	3	2	1	**1**	2	3	3
5	**H**	5	4	3	2	2	**2**	3	4
6	**R**	6	5	4	3	2	3	**3**	4
7	**I**	7	6	5	4	3	2	3	**4**
8	**N**	8	7	6	5	4	3	2	**3**
9	**A**	9	8	7	6	5	4	3	**3**

La cellule en bas à droite de la matrice correspond à la distance de Levenshtein entre s_1 et s_2.
→$d_{Levenshtein}$ = 3

1.1.2.2. Similarité de Levenshtein

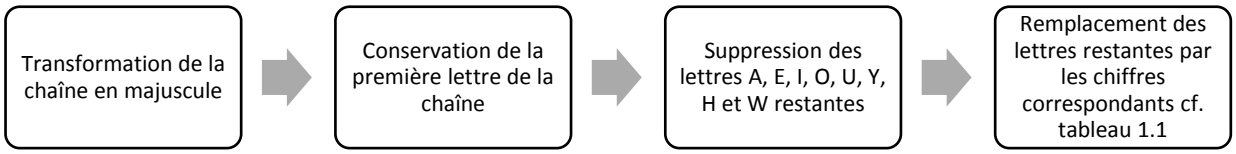
Ce principe du calcul de la similarité entre chaînes de caractères par la distance de Levenshtein, *i.e.* le minimum d'opérations permettant de passer d'une chaîne de caractères à une autre présente cependant une limite. Supposons par exemple deux couples de chaînes, avec dans le premier couple des d'environ 5 caractères, et dans le second couple des chaînes d'environ 10 caractères. Supposons également que pour ces 2 couples, on obtienne la même mesure de distance de Levenshtein, par exemple 4 (configuration parfaitement possible). Dans ce cas, on peut difficilement admettre que le degré de similarité pour ces deux couples de chaînes est semblable. Pour pallier cette limite, Levenshtein a introduit la notion de *similarité* par la prise en compte la longueur des chaînes à comparer, aboutissant à une mesure de similarité comprise entre 0 et 1, exprimée comme suit [7]:

$$sim_{Levenshtein}(s_1, s_2) = 1 - \frac{d_{Levenshtein}(s_1, s_2)}{\max(|s_1|, |s_2|)} \tag{1.5}$$

1.1.3. Version française du Soundex

Soundex est un algorithme servant à encoder phonétiquement des chaînes de caractères selon leur prononciation en anglais. Il a été conçu par Russell et Odell et breveté en 1918 [19]. Le principe de cet algorithme consiste à ce que les chaînes prononcées similairement soient encodés de la même façon par Soundex, ceci afin de pouvoir juger de la correspondance phonétique entre des chaînes malgré des différences d'épellations.

Un code Soundex est composé de 4 caractères, commençant par une lettre puis trois chiffres. La lettre correspond au premier caractère de la chaîne à coder, et les chiffres sont encodés selon les consonnes restantes dans cette chaîne. Un tableau de correspondance (version française) est utilisé pour le codage, les consonnes à prononciation similaire disposant du même code, les voyelles n'étant sont codées. Le processus de cet algorithme se déroule comme suit :



Si le code obtenu contient moins de 4 éléments, il est complété à droite par un 0 ; si le code obtenu contient plus de 4 éléments, seuls les 4 premiers éléments du code sont conservés.

Lettres	B, P	C, K, Q	D, T	L	M, N	R	G, J	X, Z, S	F, V
Code	1	2	3	4	5	6	7	8	9

Tableau 1.1 Tableau de correspondance pour le codage Soundex

Exemples

- $s_1 =$ Catherine

Catherine → CATHERINE → CTRN → C365, donc $soundex(s_1) = C365$

- $s_2 =$ Katehrina

Katehrina → KATEHRINA → KTRN → K365, donc $soundex(s_2) = K365$

- $s_3 =$ Katrine

Katrine → KATRINE → KTRN → K365, donc $soundex(s_3) = K365$

1.2. Comparaison des couples d'enregistrements

La décision de procéder ou non chaînage de deux enregistrements résulte de la prise en considération des résultats de comparaison de chaque couple de champs en correspondance. Pour dresser la liste de tous les couples d'enregistrements à considérer, il est nécessaire de procéder au produit cartésien des enregistrements de chacune des 2 sources de données. Supposons une procédure de chaînage d'enregistrements impliquant deux bases de données A et B, chacune de ces bases incluant des champs contenant les informations d'identifications des patients. La procédure de chaînage entre les bases A et B est fondée sur la liste de ces champs d'identification en commun

entre ces 2 bases. L'ensemble des couples d'enregistrements à comparer est donc le résultat du produit cartésien entre A et B, *i.e.* l'ensemble de tous les couples possibles entre ces 2 bases [6]:

$$A \times B = \{(a, b); a \in A, b \in B\} \tag{1.6}$$

où les enregistrements des A et B sont respectivement notés comme a et b.

Le principe de la procédure de chaînage des enregistrements consiste à affecter chaque couple provenant de l'ensemble $A \times B$ à l'un des deux sous-ensembles suivants [6]:

$$M = \{(a, b); a = b, a \in A, b \in B\} \tag{1.7}$$

et

$$U = \{(a, b); a \neq b, a \in A, b \in B\} \tag{1.8}$$

où M est l'ensemble des couples d'enregistrements concernant un même patient, et U est l'ensemble des couples d'enregistrements concernant des patients différents. La section suivante décrit les principales méthodes retrouvées dans la littérature.

1.2.1. Chaînage déterministe des enregistrements

1.2.1.1. Comparaison de tous les champs disponibles

Le chaînage déterministe d'enregistrements (*Deterministic Record Linkage, DRL*) est une des méthodes les plus simples pour effectuer un chaînage entre deux bases de données. Pour sa réalisation, on peut utiliser un vecteur de comparaison (*comparison vector*) γ^j, pour chaque couple j, et qui est défini comme suit [6]:

$$\gamma_i^j = \begin{cases} 1 \; si \; les \; champs \; i \; sont \; identiques \; au \; sein \; d'un \; couple \; d'enregistrements \; \mathrm{j} \\ 0 \; sinon \end{cases} \tag{1.9}$$

où
$i = 1, 2, \dots, n$ et $j = 1, 2, \dots, N$, avec n est le nombre de champs comparés, et N est le nombre de couples d'enregistrements possibles entre les deux bases à chaîner
j indique le $j^{ème}$ couple d'enregistrements dans le produit cartésien $A \times B$

La comparaison du $j^{ème}$ couple d'enregistrements consiste à comparer chaque champ correspondant au sein de ce couple, comme :

$$\gamma^j = [\gamma_1^j, \gamma_2^j, \dots, \gamma_n^j] \tag{1.10}$$

Une décision de chaînage peut donc être prise, fondée sur la concordance de tous les champs comparés, soit : si $\gamma^j = [1, 1, \dots, 1]$, alors le $j^{ème}$ couple d'enregistrements est considéré comme concernant un même patient, c'est-à-dire $\gamma^j \in M$; dans le cas contraire, $\gamma^j \in U$.

Exemple

Supposons que les trois enregistrements ci-dessous proviennent des bases de données A et B :

BDD	Identifiant du patient	Nom	Prénom	Sexe	Date de naissance
A	a1	Martin	Catherine	F	01/12/1980

BDD	Identifiant du patient	Nom	Prénom	Sexe	Date de naissance
B	b1	Martin	Katehrina	F	01/12/1980
B	b2	Martin	Catherine	F	01/12/1980

Tableau 1.2 Exemple des enregistrements provenant des bases de données A et B pour tester le DRL

Les champs d'enregistrements *nom*, *prénom*, *sexe* et *date de naissance* sont indicés respectivement comme champs 1, 2, 3 et 4. Soient γ^1 et γ^2, les couples d'enregistrements (a1, b1) et (a1, b2), la comparaison de chaque champ correspondant au sein de ces deux couples aboutit à :

$$\gamma^1 = [\gamma_1^1, \gamma_2^1, \gamma_3^1, \gamma_4^1] = [1, 0, 1, 1] \in U$$

$$\gamma^2 = [\gamma_1^2, \gamma_2^2, \gamma_3^2, \gamma_4^2] = [1, 1, 1, 1] \in M$$

1.2.1.2. Comparaison d'une partie des champs disponibles

La qualité de données pour les champs à comparer n'étant pas toujours parfaite, afin de réduire le nombre de défaut de chaînage, la règle de chaînage de la méthode *DRL* peut être élargie à la concordance de *n-1* champs parmi *n* champs comparés [11,20], c'est-à-dire que si le nombre d'éléments dans γ^j étant égal à 0 est inférieur ou égal à 1, alors on peut considérer que $\gamma^j \in M$.

Exemple

Reprenons le couple d'enregistrements γ^1 à comparer dans l'exemple précédent, sous cette nouvelle règle, assouplie, de décision de chaînage, γ^1 peut être classé comme un couple à chaîner, malgré une discordance entre les épellations du champ *prénom* (« *Catherine* » *versus* « *Katehrina* »).

1.2.1.3. Comparaison approximative de tous les champs disponibles

Cette méthode confronte le résultat de la comparaison des champs en assouplissant la règle de décision à partir d'un critère qui ne se contente pas des 2 seules modalités exclusives concernant le vecteur γ^j (qui renvoie à un résultat binaire soit de concordance exacte soit de discordance).

Les méthodes présentées dans la section 1.1, afin de comparer les champs au sein d'un couple d'enregistrements, peuvent être utilisées pour définir les critères de décision de champs à considérer comme étant « approximativement » concordants à partir des règles suivantes [13,21,22] :

- Pour la distance de Jaro-Winkler

  $$d_{JW}(r_i^j[A]; r_i^j[B]) \geq \theta_{JW,i}$$

- Pour la distance ou la similarité de Levenshtein

  $$d_{Levenshtein}(r_i^j[A]; r_i^j[B]) \leq \theta_{dL,i}$$

  $$sim_{Levenshtein}(r_i^j[A]; r_i^j[B]) \geq \theta_{sL,i}$$

- Pour la comparaison des codes Soundex

  $$soundex(r_i^j[A]) = soundex(r_i^j[B])$$

où :

$r_i^j[X]$ est la chaîne de caractères du champ i pour le couple d'enregistrements j de la base de données X

Il est bien entendu que le choix des valeurs de seuil $\theta_{JW,i}$, $\theta_{dL,i}$ et $\theta_{sL,i}$ est fixé en fonction du contexte d'application, par exemple, le type de valeur dans les champs à comparer (*e.g.* lettres, chiffres ou code), la tolérance du chaînage à tort (excès de chaînage) et de l'absence de chaînage à tort (défaut de chaînage), le type et le taux d'erreurs dans les champs, etc. Cependant, malgré tout le soin apporté à ce choix, il demeure empirique.

Exemple

A partir des mêmes exemples d'enregistrements du Tableau 1.2, supposons que l'on souhaite prendre une décision sur le chaînage ou non pour le couple d'enregistrements $r^1 = (a1, b1)$, et que l'on utilise la méthode DRL avec la comparaison approximative des champs par la distance de Jaro-Winkler.

Fixons les valeurs $\theta_{JW,1} = 0.8,\ \theta_{JW,2} = 0.8,\ \theta_{JW,3} = 0.9$ et $\theta_{JW,4} = 1$, définies empiriquement, puis calculons la distance de Jaro-Winkler pour chaque champ au sein du couple r^1 :

$$d_{JW}(r^1) = d_{JW}([r_1^1, r_2^1, r_3^1, r_4^1]) = [1, 0.8042, 1, 1]$$

Ainsi, la décision de chaînage de ces deux enregistrements peut être prise pour ce couple sur la base d'une constatation que les distances de tous les champs sont conformes à leurs critères respectifs, considérant ainsi ces champs comme « approximativement » concordants.

1.2.2. Chaînage probabiliste des enregistrements

Comme évoqué en introduction, la constatation d'une concordance sur un champ comme le *nom* est plus vraisemblablement susceptible de correspondre à deux enregistrements relatifs à une même personne que la constatation d'une concordance sur un champ comme le *sexe* ou le *prénom*. La contribution de la concordance pour chaîner des enregistrements semble donc variable d'un champ à un autre [4] [15]. Les méthodes déterministes de chaînage d'enregistrements ne sont malheureusement pas en mesure de prendre en compte cette différence de contribution, au contraire des méthodes probabilistes, dont l'idée a été introduite par Newcombe [5], puis formalisée par Fellegi et Sunter [6].

1.2.2.1. Méthode de Fellegi et Sunter

Dans le chaînage probabiliste des enregistrements par la méthode de Fellegi et Sunter (PRL-FS), à chaque couple d'enregistrements j est attribué un poids (*weight*) w fondé sur un ratio de vraisemblance [6]:

$$w^j = log_2\left(\frac{P(\gamma^j | r^j \in M)}{P(\gamma^j | r^j \in U)}\right) \tag{1.11}$$

Un seuil de décision (*decision threshold*) peut ensuite éventuellement être déterminé, les couples d'enregistrements avec un poids au-dessus de ce seuil étant considérés comme issus du même patient, et ceux dont le poids est en dessous de ce seuil étant considérés comme des enregistrements issus de patients différents.

Pour simplifier le calcul des deux probabilités dans l'équation (1.11), Fellegi et Sunter ont utilisé l'hypothèse d'indépendance conditionnelle suivante [6]:

$$P(\gamma^j | r^j \in M) = \prod_{i=1}^{n} m_i^{\gamma_i^j} (1 - m_i)^{1-\gamma_i^j} \tag{1.12}$$

$$P(\gamma^j | r^j \in U) = \prod_{i=1}^{n} u_i^{\gamma_i^j} (1 - u_i)^{1-\gamma_i^j} \tag{1.13}$$

où :

m_i est la probabilité de concordance du champ i sachant que le couple d'enregistrements j implique le même patient
u_i est la probabilité de concordance du champ i sachant que le couple d'enregistrements j implique des patients différents

Les probabilités m_i et u_i peuvent être exprimées comme suit :

$$m_i = P(\gamma_i^j = 1 | r_j \in M) \tag{1.14}$$
$$u_i = P(\gamma_i^j = 1 | r_j \in U) \tag{1.15}$$

Pour estimer les probabilités m_i et u_i, on pourrait utiliser un couple de jeux de données d'apprentissage pour lequel la vérité de chaînage serait connue (c'est-à-dire la connaissance de la réalité des couples impliquant un même patient et de ceux impliquant des patients différents) [7,23], mais ces jeux de données ne sont que très rarement disponibles.

L'estimation, par méthode du maximum de vraisemblance, des probabilités m_i et u_i est donc très largement réalisée par l'algorithme EM, qui est considéré comme une des méthodes les plus efficaces [15,24]. Mais à notre connaissance, le détail de son implémentation est très rarement décrit dans la littérature. La description de son implémentation sera donc détaillée au chapitre 2 de ce manuscrit.

Exemple

Supposons que les quatre enregistrements ci-dessous proviennent des bases de données A et B, et soumettons les décisions de chaînage pour les couples (a11, b11), (a11, b12) et (a11, b13) au PRL-FS :

BDD	Identifiant du patient	Nom	Prénom	Sexe	Date de naissance
A	a11	Bernard	Catherine	F	10/12/1980

BDD	Identifiant du patient	Nom	Prénom	Sexe	Date de naissance
B	b11	Bernard	Catherine	F	01/12/1980
B	B12	Bernard	Cahterine	F	01/12/1980
B	B13	Bernard	Anne	F	01/12/1980

Tableau 1.3 Exemple d'enregistrements provenant des 2 bases de données A et B pour tester le PRL

Tout d'abord, on procède à l'estimation des probabilités m_i et u_i avec l'algorithme EM, fondée sur l'ensemble des couples $A \times B$, et on obtient :

	Nom	Prénom	Sexe	Date de naissance
$\boldsymbol{m_i}$	0,81065	0,80769	0,98964	0,81361
$\boldsymbol{u_i}$	0,00012	0,00316	0,50118	0,00067

Tableau 1.4 Exemple des probabilités m_i et u_i estimées par l'algorithme EM pour le PRL-FS

En utilisant ces probabilités m_i et u_i estimées pour calculer les poids de chaque couple, et en confrontant ces poids avec un seuil de décision θ_{FS} étant ici égal à 12 (ce seuil n'est pas universel, voir la section 2.2 pour le choix du seuil), examinons les décisions de chaînage pour chacun de ces couples.

- Couple (a11, b11)

Notons ce couple comme r^1, et indiçons les champs d'enregistrements *nom*, *prénom*, *sexe* et *date de naissance* respectivement par 1, 2, 3 et 4.

$$\gamma^1 = [\gamma_1^1, \gamma_2^1, \gamma_3^1, \gamma_4^1] = [1, 1, 1, 0]$$

$$w^1 = log_2\left(\frac{P(\gamma^1|r^1 \in M)}{P(\gamma^1|r^1 \in U)}\right) = log_2\left(\frac{\prod_{i=1}^n m_i^{\gamma_i^1}(1-m_i)^{1-\gamma_i^1}}{\prod_{i=1}^n u_i^{\gamma_i^1}(1-u_i)^{1-\gamma_i^1}}\right) = log_2\left(\frac{m_1 \times m_2 \times m_3 \times (1-m_4)}{u_1 \times u_2 \times u_3 \times (1-u_4)}\right)$$
$$= 19.28$$

$w^1 > \theta_{FS}$ → a11 et b11 peuvent être considérés comme des enregistrements issus du même patient.

- Couple (a11, b12)

Notons ce couple comme r^2.

$$\gamma^2 = [\gamma_1^2, \gamma_2^2, \gamma_3^2, \gamma_4^2] = [1, 0, 1, 0]$$

$$w^2 = log_2\left(\frac{P(\gamma^2|r^2 \in M)}{P(\gamma^2|r^2 \in U)}\right) = log_2\left(\frac{\prod_{i=1}^n m_i^{\gamma_i^2}(1-m_i)^{1-\gamma_i^2}}{\prod_{i=1}^n u_i^{\gamma_i^2}(1-u_i)^{1-\gamma_i^2}}\right) = log_2\left(\frac{m_1 \times (1-m_2) \times m_3 \times (1-m_4)}{u_1 \times (1-u_2) \times u_3 \times (1-u_4)}\right)$$
$$= 8.9$$

$w^2 < \theta_{FS}$ → a11 et b12 ne peuvent être considérés comme des enregistrements issus du même patient.

- Couple (a11, b13)

Notons ce couple comme r^3.

$$\gamma^3 = [\gamma_1^3, \gamma_2^3, \gamma_3^3, \gamma_4^3] = [1, 0, 1, 0]$$

$$w^3 = log_2\left(\frac{P(\gamma^3|r^3 \in M)}{P(\gamma^3|r^3 \in U)}\right) = log_2\left(\frac{\prod_{i=1}^n m_i^{\gamma_i^3}(1-m_i)^{1-\gamma_i^3}}{\prod_{i=1}^n u_i^{\gamma_i^3}(1-u_i)^{1-\gamma_i^3}}\right) = log_2\left(\frac{m_1 \times (1-m_2) \times m_3 \times (1-m_4)}{u_1 \times (1-u_2) \times u_3 \times (1-u_4)}\right)$$
$$= 8.9$$

$w^3 < \theta_{FS}$ → a11 et b13 ne peuvent être considérés comme les enregistrements issus du même patient.

On peut constater que le même poids est attribué aux couples (a11, b12) et (a11, b13) puisqu'ils présentent tous 2 des discordances sur les champs *prénom* et *date de naissance*. Mais, en comparant les discordances de *prénom* dans ces deux couples, ("*Catherine*", "*Cahterine*") *versus* ("*Catherine*", "*Anne*"), la discordance dans le premier couple est probablement causée par une erreur d'épellation de type inversion de caractères. Il semble donc inapproprié d'attribuer systématiquement le même poids pour ces deux couples, et le poids attribué devrait être proportionnel à la mesure de similarité entre les chaînes dans le champ à comparer [15]. C'est sur ce principe que Winkler a proposé une méthode de chaînage probabiliste (PRL-W) qui tienne compte du degré de similarité entre chaînes de caractères dans le calcul des poids d'un couple d'enregistrements.

1.2.2.2. Méthode de Winkler

Le principe de la méthode probabiliste de Winkler (PRL-W) est très analogue à celui de la méthode de PRL-FS dont il s'inspire, leur différence tient à la définition des probabilités m et u. Dans cette méthode, Winkler a utilisé la distance de Jaro-Winkler (voir la section 1.1.1.2) pour quantifier la similarité entre les chaînes dans les champs à comparer. Cette distance est un score normalisé entre 0 et 1. L'intervalle du score [0,1] est divisé en une collection de l sous-intervalles s_k disjoints, pour chaque champ i et pour chaque sous-intervalle $s_{k,i}$, $m_{i,s_{k,i}}$ est la probabilité que le score de similarité du champ i se situe dans $s_{k,i}$ sachant que le couple d'enregistrements j concerne le même patient, et $u_{i,s_{k,i}}$ est la probabilité que le score de similarité du champ i se situe dans $s_{k,i}$ sachant que le couple d'enregistrements j concerne des patients différents [9]:

$$m_{i,s_{k,i}} = P(\varphi_i^j \in s_{k,i} | r_j \in M) \tag{1.16}$$

$$u_{i,s_{k,i}} = P(\varphi_i^j \in s_{k,i} | r_j \in U) \tag{1.17}$$

où :

φ_i^j est la distance Jaro-Winkler entre deux chaînes dans le champ i au sein du couple j. Pour simplifier la notation, nous avons abandonné l'expression $d_{JW}(r_i^j)$ vue en section 1.2.1.2

Sous l'hypothèse d'indépendance conditionnelle, on peut calculer deux probabilités :

$$P(\varphi^j | r^j \in M) = \prod_{i=1}^{n} \prod_{k=1}^{l} m_{s_{k,i}}^{I[\varphi_i^j \in s_{k,i}]} \tag{1.18}$$

$$P(\varphi^j | r^j \in U) = \prod_{i=1}^{n} \prod_{k=1}^{l} u_{s_{k,i}}^{I[\varphi_i^j \in s_{k,i}]} \tag{1.19}$$

où :

φ^j est le vecteur de comparaison de similarité par la distance Jaro-Winkler

En utilisant l'équation (1.11), le poids pour le couple d'enregistrements j est donc :

$$w^j = log_2 \left(\frac{\prod_{i=1}^{n} \prod_{k=1}^{l} m_{i,s_{k,i}}^{I[\varphi_i^j \in s_{k,i}]}}{\prod_{i=1}^{n} \prod_{k=1}^{l} u_{i,s_{k,i}}^{I[\varphi_i^j \in s_{k,i}]}} \right) \tag{1.20}$$

Comme pour la méthode PRL-FS, les probabilités m et u utilisées dans cette méthode peuvent être estimées par l'algorithme EM [25]. L'estimation des probabilités ainsi que le choix du seuil de décision de chaînage pour cette méthode seront détaillés dans la section 2.3.

Exemple

Reprenons les couples d'enregistrements r^1, r^2 et r^3 de l'exemple précédent (voir le Tableau 1.3 et le Tableau 1.4), et utilisons la méthode PRL-W afin de fonder les décisions de chaînage pour ces couples. L'estimation des probabilités $m_{i,s_{k,i}}$ et $u_{i,s_{k,i}}$ avec l'algorithme EM fondé sur l'ensemble des couples $A \times B$ est d'abord réalisée, et on obtient :

	Nom	Prénom	Sexe	Date de naissance
$\boldsymbol{m_{i,[0,0.60]}}$	0,00001	0,00001	0,00597	0,00001
…	…	…	/	…
$\boldsymbol{m_{i,[0.88,0.90]}}$	0,00895	0,01045	/	0,00001
…	…	…	/	…
$\boldsymbol{m_{i,[0.96,0.98]}}$	0,09104	0,08656	/	0,18059
$\boldsymbol{m_{i,[0.98,1]}}$	0,81343	0,79253	0,99402	0,81194
$\boldsymbol{u_{i,[0,0.60]}}$	0,86649	0,79943	0,50030	0,72429
…	…	…	/	…
$\boldsymbol{u_{i,[0.88,0.90]}}$	0,00026	0,00136	/	0,04082
…	…	…	/	…
$\boldsymbol{u_{i,[0.96,0.98]}}$	0,00005	0,00092	/	0,00556
$\boldsymbol{u_{i,[0.98,1]}}$	0,00013	0,00271	0,49969	0,00066

Tableau 1.5 Extrait des probabilités $m_{i,s_{k,i}}$ et $u_{i,s_{k,i}}$ estimées par l'algorithme EM pour le PRL-W. La comparaison du champ *sexe* étant binaire, $m_{[0,0.6]}$ correspond à $m_{\{0\}}$, et $m_{[0.98,1]}$ correspond à $m_{\{1\}}$

En utilisant ces probabilités $m_{i,s_{k,i}}$ et $u_{i,s_{k,i}}$ estimées pour calculer les poids de chaque couple, et en confrontant ces poids avec un seuil de décision θ_W étant ici égal à 6 (voir la section 2.3 pour le choix spécifique du seuil), les décisions de chaînage pour ces couples peuvent être prises (les probabilités estimées et le seuil choisi sont donnés ici à titre d'exemple) :

- r^1 : couple (a11, b11)

$$\varphi^1 = [\varphi_1^1, \varphi_2^1, \varphi_3^1, \varphi_4^1] = [1, 1, 1, 0.8963]$$

$$w^1 = log_2\left(\frac{\prod_{i=1}^{n}\prod_{k=1}^{l} m_{i,s_{k,i}}^{I[\varphi_i^1 \in s_{k,i}]}}{\prod_{i=1}^{n}\prod_{k=1}^{l} u_{i,s_{k,i}}^{I[\varphi_i^1 \in s_{k,i}]}}\right) = log_2\left(\frac{m_{1,[0.98,1]} \times m_{2,[0.98,1]} \times m_{3,[0.98,1]} \times m_{4,[0.88,0.90]}}{u_{1,[0.98,1]} \times u_{2,[0.98,1]} \times u_{3,[0.98,1]} \times u_{4,[0.88,0.90]}}\right) = 15.75$$

$w^1 > \theta_W$ → a11 et b11 peuvent être considérés comme des enregistrements issu du même patient.

- r^2 : couple (a11, b12)

$$\varphi^2 = [\varphi_1^2, \varphi_2^2, \varphi_3^2, \varphi_4^2] = [1, 0.9704, 1, 0.8963]$$

$$w^2 = log_2\left(\frac{\prod_{i=1}^{n}\prod_{k=1}^{l} m_{i,s_{k,i}}^{I[\varphi_i^2 \in s_{k,i}]}}{\prod_{i=1}^{n}\prod_{k=1}^{l} u_{i,s_{k,i}}^{I[\varphi_i^2 \in s_{k,i}]}}\right) = log_2\left(\frac{m_{1,[0.98,1]} \times m_{2,[0.96,0.98]} \times m_{3,[0.98,1]} \times m_{4,[0.88,0.90]}}{u_{1,[0.98,1]} \times u_{2,[0.96,0.98]} \times u_{3,[0.98,1]} \times u_{4,[0.88,0.90]}}\right) = 14.12$$

$w^2 > \theta_W$ → a11 et b12 peuvent être considérés comme les enregistrements issu du même patient.

- r^3 : couple (a11, b13)

$\varphi^3 = [\varphi_1^3, \varphi_2^3, \varphi_3^3, \varphi_4^3] = [1, 0.4537, 1, 0.8963]$

$$w^3 = log_2\left(\frac{\prod_{i=1}^{n}\prod_{k=1}^{l} m_{i,s_{k,i}}^{I[\varphi_i^3 \in s_{k,i}]}}{\prod_{i=1}^{n}\prod_{k=1}^{l} u_{i,s_{k,i}}^{I[\varphi_i^3 \in s_{k,i}]}}\right) = log_2\left(\frac{m_{1,[0.98,1]} \times m_{2,[0,0.60]} \times m_{3,[0.98,1]} \times m_{4,[0.88,0.90]}}{u_{1,[0.98,1]} \times u_{2,[0,0.60]} \times u_{3,[0.98,1]} \times u_{4,[0.88,0.90]}}\right) = -8.72$$

$w^3 < \theta_W$ → a11 et b13 ne peuvent pas être considérés comme les enregistrements issu d'un même patient.

1.2.2.3. Méthode de DuVall

Winkler a proposé la distance de Jaro-Winkler pour tenir compte de la similarité entre les chaînes de caractères dans le calcul des poids des couples et ainsi étendre la méthode PRL-FS. D'autres choix de mesure de similarité sont possibles. C'est ainsi, qu'en 2010, DuVall a proposé d'utiliser la distance de Levenshtein (voir la section 1.1.2.1) pour quantifier la similarité entre les chaînes [26]. Cette méthode est fondée sur le même principe que la méthode PRL-W, interviennent donc les probabilités m et u, où $m_{d,i}$ est la probabilité que la distance de Levenshtein du champ i soit égale à d sachant que le couple d'enregistrements j implique le même patient, et $u_{d,i}$ est la probabilité que la distance de Levenshtein du champ i soit égale à d sachant que le couple d'enregistrements j implique des patients différents [26]:

$$m_{i,d} = P(\delta_i^j = d | r_j \in M) \tag{1.21}$$
$$u_{i,d} = P(\delta_i^j = d | r_j \in U) \tag{1.22}$$

où :

δ_i^j est la distance Levenshtein entre deux chaînes de caractères du champ i au sein du couple j, pour simplifier la notation, on abandonne l'expression $d_{Levenshtein}(r_i^j)$ de la section 1.2.1.2.

Sous l'hypothèse d'indépendance conditionnelle, on peut calculer les deux probabilités suivantes :

$$P(\delta^j | r^j \in M) = \prod_{i=1}^{n}\prod_{d=0}^{o} m_{i,d}^{I[\delta_i^j = d]} \tag{1.23}$$
$$P(\delta^j | r^j \in U) = \prod_{i=1}^{n}\prod_{d=0}^{o} u_{i,d}^{I[\delta_i^j = d]} \tag{1.24}$$

où :

δ^j est le vecteur de comparaison de similarité par la distance de Levenshtein
o est la distance de Levenshtein maximale dans toutes les comparaisons

En utilisant l'équation (1.11), le poids pour le couple d'enregistrements j est donc :

$$w^j = log_2\left(\frac{\prod_{i=1}^{n}\prod_{d=0}^{o} m_{i,d}^{I[\delta_i^j = d]}}{\prod_{i=1}^{n}\prod_{d=0}^{o} u_{i,d}^{I[\delta_i^j = d]}}\right) \tag{1.25}$$

Comme pour les méthodes PRL-FS et PRL-W, les probabilités m et u utilisées dans cette méthode peuvent être estimées par l'algorithme EM [26].

Exemple

Utilisons la méthode de Duvall pour fonder les décisions de chaînage des couples d'enregistrements r^1, r^2 et r^3 dans l'exemple de la section 1.2.2.1 (voir le Tableau 1.3 et le Tableau 1.4). On commence par l'estimation des probabilités $m_{i,d}$ et $u_{i,d}$ avec l'algorithme EM fondé sur l'ensemble des couples $A \times B$, et on obtient :

	Nom	Prénom	Sexe	Date de naissance
$\boldsymbol{m_{i,0}}$	0,81763	0,81464	0,99252	0,81165
$\boldsymbol{m_{i,1}}$	0,07025	0,06576	0,00747	0,18684
$\boldsymbol{m_{i,2}}$	0,04334	0,05680	/	0,00001
…	…	…	/	…
$\boldsymbol{m_{i,7}}$	0,00001	0,00149	/	0,00001
$\boldsymbol{u_{i,0}}$	0,00015	0,00311	0,49960	0,00062
$\boldsymbol{u_{i,1}}$	0,00005	0,00120	0,50039	0,00554
$\boldsymbol{u_{i,2}}$	0,00007	0,00065	/	0,00403
…	…	…	/	…
$\boldsymbol{u_{i,7}}$	0,87317	0,78975	/	0,72359

Tableau 1.6 Extrait des probabilités $m_{i,d}$ et $u_{i,d}$ estimées par l'algorithme EM pour la méthode de DuVall, valeurs à titre d'exemple. La comparaison du champ *sexe* étant binaire, $m_{3,0}$ correspond à l'identité des sexes, et $m_{3,1}$ à leur différence

En utilisant les probabilités estimées ci-dessus pour le calcul des poids des couples d'enregistrements, et en confrontant ces poids avec un seuil de décision θ_D étant ici égal à 1, examinons les décisions de chaînage pour chacun des couples r^1, r^2 et r^3.

- r^1 : couple (a11, b11)

$$\delta^1 = [\delta_1^1, \delta_2^1, \delta_3^1, \delta_4^1] = [0, 0, 0, 2]$$

$$w^1 = log_2\left(\frac{\prod_{i=1}^{n}\prod_{d=0}^{o} m_{i,d}^{I[\delta_i^1=d]}}{\prod_{i=1}^{n}\prod_{d=0}^{o} u_{i,d}^{I[\delta_i^1=d]}}\right) = log_2\left(\frac{m_{1,0} \times m_{2,0} \times m_{3,0} \times m_{4,2}}{u_{1,0} \times u_{2,0} \times u_{3,0} \times u_{4,2}}\right) = 12.78$$

$w^1 > \theta_D$ → a11 et b11 peuvent être considérés comme des enregistrements issus du même patient.

- r^2 : couple (a11, b12)

$$\delta^2 = [\delta_1^2, \delta_2^2, \delta_3^2, \delta_4^2] = [0, 2, 0, 2]$$

$$w^2 = log_2\left(\frac{\prod_{i=1}^{n}\prod_{d=0}^{o} m_{i,d}^{I[\delta_i^2=d]}}{\prod_{i=1}^{n}\prod_{d=0}^{o} u_{i,d}^{I[\delta_i^2=d]}}\right) = log_2\left(\frac{m_{1,0} \times m_{2,2} \times m_{3,0} \times m_{4,2}}{u_{1,0} \times u_{2,2} \times u_{3,0} \times u_{4,2}}\right) = 11.19$$

$w^2 > \theta_D$ → a11 et b12 peuvent être considérés comme des enregistrements issus du même patient.

- r^3 : couple (a11, b13)

$$\delta^3 = [\delta_1^3, \delta_2^3, \delta_3^3, \delta_4^3] = [0, 7, 0, 2]$$

$$w^3 = log_2\left(\frac{\prod_{i=1}^{n}\prod_{d=0}^{o} m_{i,d}^{I[\delta_i^3=d]}}{\prod_{i=1}^{n}\prod_{d=0}^{o} u_{i,d}^{I[\delta_i^3=d]}}\right) = log_2\left(\frac{m_{1,0} \times m_{2,7} \times m_{3,0} \times m_{4,2}}{u_{1,0} \times u_{2,7} \times u_{3,0} \times u_{4,2}}\right) = -4.30$$

$w^3 < \theta_D$ → a11 et b13 ne peuvent pas être considérés comme les enregistrements issus d'un même patient.

Chapitre 2

Implémentation et évaluation des méthodes de chaînage existantes

Dans ce chapitre, nous présentons en détails l'implémentation de deux méthodes de chaînage probabiliste d'enregistrements mentionnées dans le chapitre précédent, l'une est la méthode de Fellegi-Sunter qui est une des méthodes probabilistes les plus utilisées, l'autre est la méthode de Winkler qui est l'une des méthodes probabilistes les plus performantes, si ce n'est, à notre connaissance, la plus efficace. Afin d'implémenter ces deux méthodes, il faut disposer de données permettant de mettre en œuvre cette procédure de chaînage d'enregistrements. Pour ce faire, on génère des couples de jeux de données synthétiques partageant des enregistrements concernant les mêmes personnes. De plus, et afin d'évaluer les performances de ces deux méthodes en termes de taux de mauvaises décisions et de temps de calcul, il nous faut disposer de données où la réalité du chaînage soit connue (*cf. infra*). Ces thématiques sont l'objet de ce chapitre qui est organisé en trois sections.

La première section de ce chapitre (2.1) décrit une proposition d'algorithme de genèse de jeux de données synthétiques servant à évaluer les méthodes de chaînage proposées. De tels jeux de données nous permettent de connaître la vérité quant à la correspondance des enregistrements, qui pourrait être assimilée à la disponibilité d'un « gold standard » indispensable à l'évaluation de la validité des décisions de chaînages des autres algorithmes. En outre, connaître et maîtriser la qualité des données (la proportion et le type d'erreurs dans chaque jeu de données) est très utile pour évaluer la capacité des méthodes de chaînage d'enregistrements à faire face aux différents types et taux d'erreurs. Cette section a fait l'objet de plusieurs valorisations intitulées « *Évaluation des algorithmes de rapprochement de patients par traits d'identification nominatifs* » [27] et « *Comparaison de performance des algorithmes de rapprochement de patients* » [28] présentées respectivement au Congrès ADELF-SFSP[4] et au Congrès ADELF-EMOIS[5]. Le contenu de cette section, constituant une synthèse et une extension de ces deux travaux, a été soumis à la « *Revue Epidémiologie et Santé Publique* ».

Dans la deuxième section (2.2), le processus de l'implémentation de la méthode de chaînage probabiliste proposée par Fellegi et Sunter est décrit [6]. Dans ce processus, l'estimation des probabilités m_i et u_i servant au calcul des poids des couples d'enregistrements est une étape cruciale rarement détaillée dans la littérature. Cette section propose donc une description de cette estimation par l'algorithme EM, ainsi que l'évaluation de la méthode implémentée. Le contenu de cette section est une extension d'une valorisation intitulée « *Utilisation de l'algorithme EM pour*

[4] Congrès Association des Epidémiologistes de Langue Française - Société Française de Santé Publique 2013 à Bordeaux, un prix de recherche a été accordé au travail de l'auteur [27] dans ce congrès

[5] Congrès Association des Epidémiologistes de Langue Française - Evaluation, Management, Organisation, Information, Santé 2014 à Paris

estimer les paramètres du chaînage probabiliste d'enregistrements » (*DOI: 10.1016/j.respe.2014.06.081*) [29] présenté au Congrès ADELF- EPITER[6], et qui a également été soumis à la « *Revue d'Epidémiologie et de Santé Publique* ».

La troisième section (2.3) présente l'implémentation de la méthode de chaînage probabiliste proposée par Winkler [9]. Cette section commence par décrire en détails comment utiliser l'algorithme EM –incluant la formalisation des équations utilisées dans les itérations d'espérance et de maximisation– pour estimer les paramètres nécessaires à la mise en œuvre de cette méthode. Ensuite, en utilisant les mêmes couples de jeux de données, cette méthode est comparée à celle de Fellegi-Sunter sur deux critères de jugement principaux, le taux de mauvaises décisions et le temps de calcul. Cette section correspond à une extension de la valorisation intitulée « *Implementation of an Extended Fellegi-Sunter Probabilistic Record Linkage Method Using the Jaro-Winkler string Comparator* » (*DOI : 10.1109/BHI.2014.6864381*) [25].

[6] Congrès de l'Association des Epidémiologistes de Langue Française - Association pour le développement de l'Epidémiologie de Terrain, 2014 à Nice

2.1. Création des données synthétiques servant à l'évaluation des méthodes de chaînage

2.1.1. Introduction

Dans le cadre de la réalisation d'un projet régional visant à mettre en place un réseau sentinelle sur grille informatique pour l'e-santé et l'épidémiologie en Auvergne [30], un des verrous techniques consistait à chaîner les enregistrements d'un même patient dispersés dans différentes bases de données. Ces bases concernaient originellement les données des associations de dépistage des cancers du sein, du colon et du col utérin en Auvergne (ARDOC et ABIDEC), des laboratoires d'anatomo-cytopathologie (Unilabs-Sipath), et du réseau de santé périnatale d'Auvergne. En l'absence de clé d'identification commune, la réalisation de cette tâche de chaînage des enregistrements dans différentes bases nécessite de s'appuyer sur les champs d'enregistrements contenant les informations d'identification, tels que le *nom*, le *prénom*, le *sexe* et la *date de naissance*, etc. Il existe plusieurs méthodes alternatives permettant d'aboutir à la décision de chaînage fondée sur les comparaisons de champs d'enregistrements. L'évaluation et la comparaison de ces alternatives nécessitent du matériel (des données) et un protocole expérimental (méthodes et critères de jugement) à mêmes de dégager laquelle de ces méthodes serait la plus performante et adaptée à cette tâche de chaînage pour notre projet.

En général, un processus de chaînage d'enregistrements comprend trois grandes étapes qui sont schématisées à la Figure 2.1. La **première étape** consiste à construire des couples d'enregistrements dont chacun provient d'une des bases de données à chaîner, et à effectuer un éventuel prétraitement pour que les champs à comparer entre deux bases se présentent au même format. Dans la **deuxième étape**, la mise en œuvre d'une méthode de chaînage d'enregistrements, par comparaison des champs au sein des couples d'enregistrements, aboutit au calcul et à l'affectation d'un poids pour chaque couple qui est le reflet de l'éventualité qu'il implique une seule et même personne. Enfin, la **troisième étape** est celle de la décision de chaînage qui est fondée sur la confrontation entre le poids du couple attribué et une règle de seuillage simple ou double. Avec le seuil unique, les couples sont classés en deux catégories : *à chaîner* et *à ne pas chaîner* ; avec le double seuil, les couples sont classés en trois catégories, incluant la catégorie *à décider manuellement* qui est ajoutée aux 2 précédentes.

On peut proposer deux grands types d'évaluation de la performance d'une méthode de chaînage. La première consiste à évaluer globalement, et éventuellement par un indicateur unique tel qu'une aire sous la courbe ROC, la capacité de la méthode de chaînage à discriminer les couples d'enregistrements impliquant la même personne de ceux impliquant des personnes différentes. La deuxième consiste à dénombrer les défauts de chaînage (absence de chaînages à tort ou séparation à tort) et les excès de chaînages (chaînages à tort). Dans les deux types d'évaluations, il est indispensable de disposer de l'information sur la véritable correspondance des enregistrements (*i.e.* la réalité du statut identiques ou différents des enregistrements d'un couple, et ce pour tous les couples possibles d'enregistrements), comme l'offrirait un *gold standard*, pour établir la validité des chaînages. Une telle information est bien trop difficile à obtenir sur données réelles, car il serait nécessaire d'effectuer des vérifications pour tous les couples d'enregistrements possibles entre deux bases à chaîner, ce qui demanderait un travail considérable, long et coûteux. Ce d'autant qu'il est difficile d'identifier si oui ou non deux enregistrements impliquent la même personne notamment

lorsque des fautes d'épellation ou des erreurs typographiques sont présentes. Par exemple, pour deux enregistrements (*Martin, Philippe, M, 19730512*) et (*Martin, Philipe, M, 19730512*), on peut constater qu'il existe une discordance pour le prénom, mais on ne sait pas si cette discordance est causée par une faute d'épellation ou si les deux enregistrements impliquent deux personnes différentes portant le même nom, ayant le même sexe et étant nées à la même date.

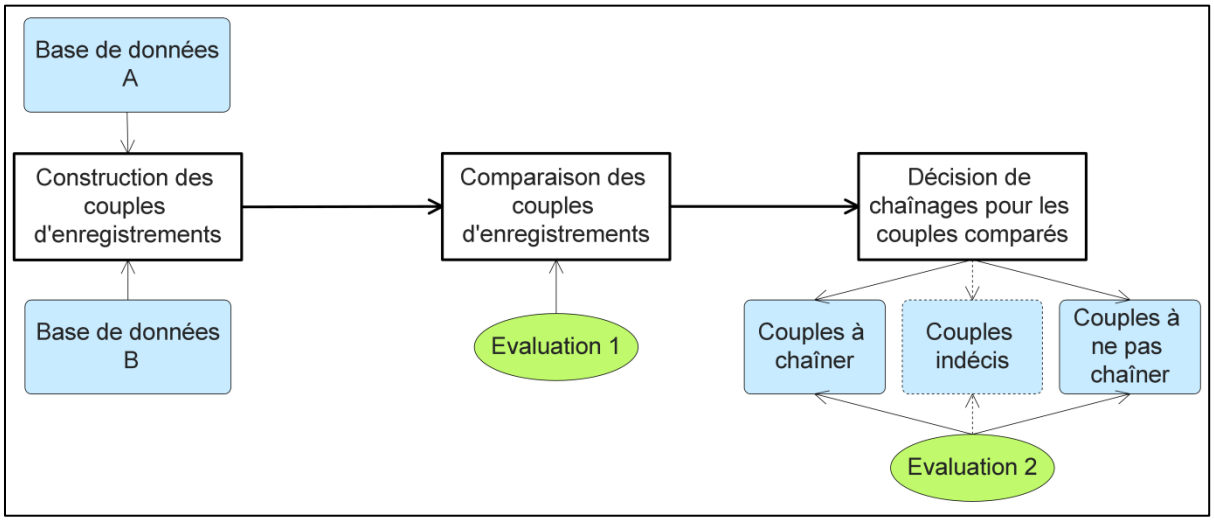


Figure 2.1 Principes du processus de chaînage d'enregistrements

Les contraintes susmentionnées nous obligent à chercher une alternative de l'utilisation des données réelles afin d'évaluer les méthodes de chaînage. Dans cette étude, on propose d'utiliser les données synthétiques ayant des caractéristiques aussi réalistes que possible. Les données synthétiques présentent donc de multiples avantages, dont le premier est leur disponibilité en taille théoriquement illimitée, mais également de respecter des caractéristiques prédéfinies, notamment de structure, de permettre l'introduction de dégradations maîtrisées en proportion et type d'erreurs dans les enregistrements dans chaque jeu de données à chaîner, et enfin de conserver intacte et valide la réalité du chaînage des données qui constitue le résultat idéal [31].

La section qui suit décrit les principes d'un algorithme modulable de génération de couples de jeux de données synthétiques servant au chaînage et à son évaluation, et permettant l'adjonction aléatoire de dégradations des champs d'enregistrements (omission, insertion, substitution ou inversion d'un caractère dans une chaîne au sein des champs) contrôlées selon des types et des taux d'erreurs prédéterminés.

2.1.2. Matériel et méthodes

2.1.2.1. Approche globale de genèse des jeux de données synthétiques

La Figure 2 résume les trois étapes principales pour la création d'un couple de jeux de données synthétiques partageant des enregistrements communs, matériel indispensable pour implémenter et évaluer des méthodes de chaînage d'enregistrements. Au préalable, on prépare des données source (*seed data*) servant à construire des enregistrements fictifs (2.1.2.2). Avec ces données source, N_S enregistrements fictifs sont générés dans la **première étape** (2.1.2.3), chacun de ces enregistrements se compose de cinq champs, *nom*, *prénom*, *sexe*, *date de naissance*, et une clé d'identification unique et permanente. Dans la **deuxième étape** (2.1.2.4), les jeux de données A et B sont générés par tirage

aléatoire sans remise de N_A et N_B enregistrements à partir du jeu des N_S enregistrements générés précédemment. Dans la **troisième étape** (2.1.2.5), des erreurs typographiques sont introduites selon une typologie déterminée et pour une certaine proportion contrôlée d'enregistrements sélectionnés de manière aléatoire dans chacun des jeux de données A et B (l'introduction d'erreurs ne s'applique évidemment pas au champ contenant la clé d'identification). La réalisation de ces trois étapes est présentée en détails dans les sections suivantes.

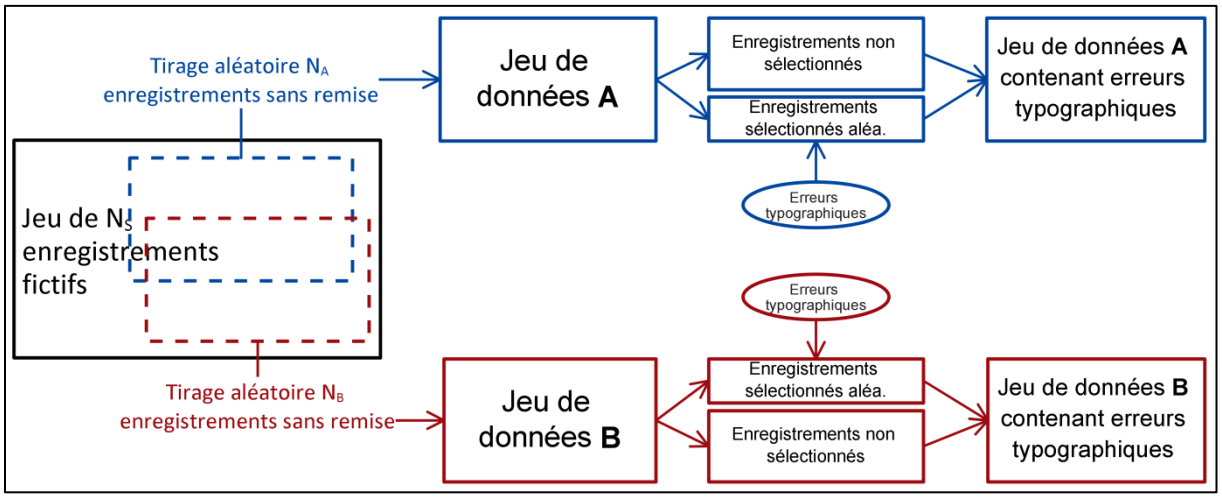


Figure 2.2 Processus de création d'un couple de jeux de données servant à l'évaluation des méthodes de chaînage d'enregistrements

2.1.2.2. Préparation des données source

Pour générer les enregistrements fictifs de patients, il faut disposer des données source (*seed data*) contenant les champs *nom*, *prénom*, *sexe* et *date de naissance*. Ces données peuvent provenir d'une liste d'enregistrements librement accessibles au public, ou d'enregistrements dont l'exploitation est autorisée. Dans le cadre de la réalisation de notre projet, une autorisation CNIL est obtenue pour « *l'utilisation des données relatives à l'identification du patient (nom, prénom, date de naissance, sexe, code postal de résidence) à des fins de validation des identités et de chaînage des données dans le cadre de la construction de l'outil* ». Dans cette étude, on a donc utilisé les champs *nom*, *prénom*, *sexe* et *date de naissance* des enregistrements d'une base de données de patients comme structure de nos données source pour générer les enregistrements fictifs.

2.1.2.3. Génération d'un jeu d'enregistrements fictifs

Dans notre étude, nous avons séparé les contenus des 4 champs de tous les enregistrements réels pour créer 3 colonnes dont l'une correspondait au « prénom sexué » composé par le *prénom* et le *sexe* correspondant, les champs nom et date de naissance constituant une colonne chacun à eux seuls. (Figure 2.3).

Nom	Prénom	Sexe	Date de naiss.
Martin	Jean	M	19821208
Bernard	Philippe	M	19850507
Thomas	Nathalie	F	19901211
Petit	Isabelle	F	19790512
Robert	Michel	M	19851010
Richard	Sylvie	F	19750302
Durand	Catherine	F	19730417
Dubois	Patrick	M	19820519
...	...	...	...
...	...	...	...

Figure 2.3 Echantillons de noms, prénoms rattachés aux sexes, dates de naissance

A cette étape, chaque enregistrement fictif est construit en concaténant aléatoirement les 3 éléments tirés respectivement des 3 colonnes indiqué à la Figure 2.3. Par exemple, l'enregistrement fictif (*Martin, Philippe, M, 19730417*) est créé en concaténant les champs *Martin*, *Philippe M* et *19730417*, ces derniers étant tirés au sort. On attribue ensuite une clé d'identification à cet enregistrement fictif permettant de l'identifier de manière unique et définitive (cette clé n'étant pas concernée par le processus d'ajouts d'erreurs). En répétant cette procédure N_S fois, un jeu contenant N_S enregistrements fictifs est ainsi généré.

2.1.2.4. Création d'un couple de jeux de données synthétiques

A partir du jeu de données contenant N_S enregistrements fictifs précédemment générés, N_A et N_B enregistrements sont respectivement tirés sans remise pour construire les jeux de données synthétiques A et B. Les nombres N_A et N_B sont choisis en respectant la contrainte $N_A + N_B = \alpha \times N_S$ avec $1 < \alpha < 2$, afin que ces deux jeux de données aient donc N_C enregistrements communs, N_C variant de $(\alpha - 1) \times N_S$ à $min(N_A, N_B)$.

2.1.2.5. Ajout d'erreurs typographiques aux jeux de données synthétiques

Dans la réalité, les champs relatifs à l'identification d'un patient contiennent en général des erreurs typographiques. A cette étape, on introduit donc des erreurs pour une certaine proportion d'enregistrements sélectionnés aléatoirement dans chacun des jeux de données synthétiques A et B. Les types d'erreurs introduites dans les jeux de données synthétiques sont listés au Tableau 2.1.

Types d'erreurs	Description	Exemple
Omission	Supprimer un caractère dans une chaîne	*Michael* → *Michel*
Insertion	Insérer un caractère dans une chaîne	*Michael* → *Micharel*
Substitution	Remplacer un caractère dans une chaîne par un caractère différent	*Michael* → *Mickael*
Transposition	Echanger deux caractères adjacents dans une chaîne	*Michael* → *Micheal*

Tableau 2.1 Types d'erreurs introduites dans les jeux de données synthétiques

Selon une étude de validation de données patients, il s'agit des fautes d'épellation les plus courantes dans les traits d'identification nominatifs [32]. Dans un jeu de données, ces erreurs se produisent dans une certaine proportion d'enregistrements. Dans la littérature, cette proportion peut atteindre 36,5% [32]. Dans notre algorithme, le type et la proportion des erreurs sont modulables afin de construire des jeux de données réalistes quant aux tâches de chaînages sur données réelles.

L'ajout de ces 4 types d'erreurs typographiques dans les champs d'enregistrement est décrit dans les sections suivantes.

2.1.2.5.1. Omission d'un caractère

Dans notre algorithme, l'omission d'un caractère dans un champ n'est pas totalement aléatoire pour que les erreurs introduites soient les plus proches possibles de la réalité. Le processus de l'omission commence par la recherche de caractères adjacents identiques dans une chaîne au sein d'un champ, et si tel est le cas, l'un des caractères est supprimé, car une telle suppression n'engendre pas forcément de modification de la prononciation de cette chaîne de caractères, par exemple, de *Yannick* à *Yanick*. En l'absence de lettres adjacentes identiques dans la chaîne, le processus recherche la lettre *h* qui est susceptible de ne pas se prononcer (lettre muette). Si la lettre *h* existe dans une chaîne, elle est supprimée, par exemple, de *Jonathan* à *Jonatan*. Si une chaîne ne comporte ni lettres adjacentes identiques, ni la lettre *h*, alors la suppression d'une lettre est effectuée aléatoirement. Ce type d'erreur est seulement appliqué aux champs *nom* et *prénom*.

2.1.2.5.2. Insertion d'un caractère

Comme pour l'omission d'un caractère, la saisie par erreur d'un caractère en plus dans le *nom* ou le *prénom* pourrait aussi être due à la ressemblance phonétique entre deux chaînes de caractères (la chaîne originale et la chaîne mal saisie), par exemple, de *Philipe* à *Phillipe*. On définit d'abord une liste de lettres susceptibles d'être répétées par erreur, telles que : *c, f, l, m, n, p, r, s, t*. Si une de ces lettres existe dans une chaîne à perturber (et si elle n'est pas la première lettre de cette chaîne), on en choisit aléatoirement une et on la répète une fois. Si aucune de ces lettres n'existe, on répète une consonne qui est choisie aléatoirement dans cette chaîne.

2.1.2.5.3. Substitution d'un caractère

Dans la saisie d'un *nom* ou d'un *prénom*, on suppose qu'une substitution de caractère est susceptible de se produire à cause de la ressemblance phonétique entre certaines lettres, ou d'une faute de frappe sur une lettre voisine sur clavier informatique (AZERTY par exemple). Dans le processus de genèse de l'erreur de substitution d'un caractère, on détermine donc deux règles de remplacements de lettres selon les deux causes susmentionnées. La première règle est de remplacer une lettre par une autre lettre ayant le même code Soundex (voir la version française de la table de correspondance Soundex, Tableau 1.1), par exemple, la saisie de *Catherine* au lieu de *Katherine*. La deuxième règle est de remplacer une lettre par une des lettres voisines sur le clavier informatique, par exemple, *h* est remplacée par une des lettres *g, y, u, j, n* et *b* (voir Figure 2.4). Dans ce processus de substitution, on utilise d'abord la première règle, et si elle n'est pas applicable (*i.e.* aucune des lettres dans une chaîne ne partageant le même code Soundex avec une autre lettre, par exemple, les lettres *l* et *r*), alors c'est la deuxième règle qui est utilisée. Pour la perturbation d'une *date*, la deuxième règle est appliquée mais en respectant le format correct de la date. Quant au *sexe*, le processus de substitution remplace *M* par *F* et *F* par *M*.

Figure 2.4 Exemple de substitution d'une lettre par une des lettres voisines sur clavier informatique

2.1.2.5.4. Transposition de deux caractères adjacents

Dans le processus d'introduction des erreurs de transposition de deux caractères adjacents, on choisit d'abord aléatoirement un caractère dans une chaîne à perturber. Considérons ce $k^{ième}$ caractère de la chaîne, si ce caractère n'est pas le dernier caractère de la chaîne, on échange ce $k^{ième}$ caractère avec le $(k+1)^{ième}$ de cette chaîne ; sinon, on échange le $k^{ième}$ caractère avec le $(k-1)^{ième}$ caractère de cette chaîne. Ce type d'erreurs est seulement appliqué au *nom* et au *prénom*.

A l'issue de cette étape d'ajout aléatoire des erreurs, un de ces quatre types d'erreurs typographiques est introduit dans les champs pour une certaine proportion des enregistrements générés dans l'étape précédente. Le couple de jeux de données servant à l'évaluation des méthodes de chaînage d'enregistrements est ainsi créé et disponible pour analyse.

2.1.3. Résultats

L'algorithme de génération du couple de jeux de données synthétiques est implémenté en R (version 2.15.1) [33] sur une station de travail avec un CPU de 2,0 GHz (Intel® Xeon® E5-2620) et 16 Go de RAM.

Pour évaluer nos algorithmes de chaînage, la génération de couple de jeux de données a été réalisée avec les paramètres suivants :

- chaque jeu de données contient 2 000 enregistrements,
- le taux moyen/attendu d'enregistrements perturbés est fixée à environ 30%,
- la répartition des types d'erreurs au sein de ces enregistrements perturbés est de 10%, 10%, 1% et 10% pour les champs *nom*, *prénom*, *sexe* et *date de naissance*, respectivement,
- *nom* et/ou *prénom* sont éventuellement soumis aux 4 types d'erreurs,
- *date de naissance* et *sexe* ne sont soumis qu'aux erreurs de substitution de caractère,
- enfin, ce processus de génération est en général répété 100 fois.

Description des types et taux d'erreurs observés dans les données synthétiques

Le Tableau 2 décrit la distribution des erreurs observées dans les jeux de données synthétiques. On y lit la fréquence moyenne (en pourcentage) sur 100 réalisations. Un pattern se code en 4 caractères qui représentent respectivement l'existence (code 1) ou non (code 0) d'une erreur dans les champs *nom*, *prénom*, *sexe* et *date de naissance*. Par exemple, un pattern 0110 signifie que des erreurs sont présentes pour les champs *prénom* et *sexe*. On peut observer dans chaque jeu qu'environ 72% des

enregistrements ne sont pas sujets à des erreurs, soit environ 28% des enregistrements perturbés, ce qui correspond approximativement au taux d'erreurs attendu prédéfini (environ 30%). Pour la distribution des erreurs au sein des enregistrements perturbés, les fréquences des erreurs dans le *nom*, le *prénom*, le *sexe* et la *date de naissance* sont calculées en additionnant respectivement les patterns qui sont sous la forme de 1xxx, x1xx, xx1x et xxx1. On peut constater que les fréquences observées sont respectivement 10%, 10%, 1% et 10%, ce qui correspond exactement aux taux attendus prédéfinis.

Pattern	Fréquence en %	
	Jeu de données A	Jeu de données B
0000	72.093%	72.253%
0001	8.087%	7.993%
0010	0.753%	0.703%
0011	0.087%	0.060%
0100	8.037%	7.930%
0101	0.873%	0.947%
0110	0.063%	0.107%
0111	0.007%	0.007%
1000	8.073%	7.977%
1001	0.827%	0.910%
1010	0.073%	0.097%
1011	0.003%	0.003%
1100	0.893%	0.910%
1101	0.120%	0.080%
1110	0.007%	0.020%
1111	0.003%	0.003%
Total	100.000%	100.000%

Tableau 2.2 Distribution des erreurs dans les champs

Exemple d'enregistrements synthétiques

Dans notre processus, chaque enregistrement généré comporte une clé d'identification unique qui permet d'identifier si deux enregistrements impliquent ou non une même personne. Dans le Tableau 2.3, l'enregistrement du jeu de données A est soumis respectivement à des erreurs de transposition et de substitution dans les champs *nom* et *prénom*. Avec l'*identifiant unique*, malgré ces variations des épellations du *nom* et du *prénom*, on peut établir la vérité de correspondance qui permettra d'évaluer les méthodes de chaînage des enregistrements. Par exemple, en utilisant le chaînage probabiliste des enregistrements de Fellegi-Sunter [6] (1.2.2.1), la décision sera un non-chaînage de ces deux enregistrements, alors qu'avec le chaînage probabiliste de Winkler [9] (1.2.2.2), la décision sera un chaînage. Ce n'est que grâce au champ *identifiant unique* disponible dans les données synthétiques, que l'on peut évaluer ces deux méthodes de chaînage, et conclure que la méthode de Winkler est meilleure que la méthode de Fellegi-Sunter face aux erreurs typographiques dans les champs d'enregistrements.

Jeu de données	Identifiant unique	Nom	Prénom	Sexe	Date de naissance	Erreur
A	ID0512	Matrin	Catherina	F	01/12/1980	1100

Jeu de données	Identifiant unique	Nom	Prénom	Sexe	Date de naissance	Erreur
B	ID0512	Martin	Catherine	F	01/12/1980	0000

Tableau 2.3 Exemple des enregistrements générés

Temps de calcul nécessaire au processus de génération des données

Enfin, la consommation en temps de calcul est notée avec, reportés au Tableau 2.4, les temps moyens de 100 exécutions pour générer différentes tailles de jeux de données, le taux attendu d'enregistrements à perturber étant fixé à environ 30%.

Nombre d'enregistrements	1000	2000	4000	6000	8000	10000
Temps (en secondes)	0.180	0.368	1.010	2.313	4.250	7.678

Tableau 2.4 Temps moyens pour 100 réalisations du processus de synthèse de données

2.1.4. Discussion et Perspectives

L'utilisation de données synthétiques pour évaluer les méthodes de chaînage apparaît donc comme une stratégie incontournable. Ce travail propose un algorithme modulable utile à la synthèse des jeux de données dans le but d'évaluer la validité des méthodes de chaînage des enregistrements, *i.e.* la conformité à la réalité des décisions de chaînage issues de ces méthodes. Dans un processus de génération de données, 3 paramètres pour chaque jeu de données sont modulables : le nombre d'enregistrements, le taux d'enregistrements soumis à des erreurs typographiques et les types d'erreurs pour chaque champ d'enregistrements.

Les résultats démontrent que notre algorithme permet de générer les jeux de données conformes aux attentes prédéterminées dans le processus de génération, et que ce processus consomme relativement peu de temps de calcul. Avec notre algorithme, un identifiant unique est attribué pour chaque enregistrement généré permettant d'établir la vérité des correspondances des enregistrements afin d'évaluer les résultats de chaînages. Une telle information sur les correspondances des enregistrements est très difficile à obtenir sur les bases de données réelles ne partageant pas de clé d'identification commune, et demanderait en pratique la vérification extrêmement coûteuse de chaque couple d'enregistrements sans être certain d'identifier si deux enregistrements impliquent ou non les mêmes personnes, notamment en présence d'erreurs typographiques dans les champs.

Dans cette étude, pour que les erreurs ajoutées aux données synthétiques soient proches de la réalité, on a défini les règles d'omission, d'insertion, de substitution et de transposition de caractères. Mais, malgré ces règles, un ajout automatique des erreurs sans dictionnaire peut parfois manquer de réalisme, notamment dans une langue donnée. Ainsi, on a pu observer dans nos données synthétiques une transformation du prénom *Christophe* en *Christope* par l'omission de lettre *h*, ou de *Patrick* devenu *Patricc* par la substitution d'un caractère selon la règle fondée sur la table de correspondance de Soundex (remplacement de *k* par *c*, ou l'inverse). Toutefois, de telles erreurs dans des prénoms aussi courants devraient être rares en pratique. Dans ce contexte de génération des données, le but de l'adjonction des erreurs dans les enregistrements est destiné à évaluer les méthodes de chaînage et, à notre connaissance, aucune des méthodes existantes n'est en

mesure de nuancer la similarité entre les couples (*Patrick*, *Patricc*) et (*Patrick*, *Patrikk*) ou entre les couples (*Christophe*, *Christope*) et (*Christophe*, *Christoph*). Ce petit défaut de réalisme ne gêne donc en rien dans l'évaluation des méthodes actuelles de chaînage.

Les jeux de données générées dans cette étude sont issus des données source (*seed data*). Pour généraliser ce travail, par exemple dans le contexte de l'absence d'autorisation d'utilisation des données personnelles, les données source pourraient être constituées par des enregistrements librement accessibles au public [34]. Dans ce cas, les données générées pourraient être publiées en ligne comme les jeux de données challenge à destination des utilisateurs ou des auteurs des méthodes de chaînage pour tester ou évaluer leurs méthodes. De plus, l'algorithme modulable peut être fourni aux utilisateurs pour générer eux-mêmes les jeux de données synthétiques en fonction de leurs besoins (nombre d'enregistrements, le taux et le type d'erreurs dans chaque jeu), mais également pour l'étendre avec de nouvelles fonctionnalités comme, par exemple, l'ajout de nouvelles erreurs dans les enregistrements telles que des données manquantes dans les champs, l'inversion des nom et prénom, etc.

2.1.5. Conclusion

On a proposé dans cette section du manuscrit, un algorithme de génération des données, ensuite, on a décrit la structure globale et les principaux éléments de l'algorithme, et on a illustré le choix des paramètres. Les résultats démontrent que notre algorithme permet de générer efficacement les jeux de données conformes à nos attentes. Cet algorithme de génération reste perfectible, mais la version actuelle pourrait satisfaire nombre d'utilisateurs pour générer les jeux de données servant à évaluer les méthodes existantes de chaînage des enregistrements, ou d'autres études éventuelles.

2.2. Implémentation de la méthode chaînage probabiliste des enregistrements d'après Fellegi-Sunter

2.2.1. Introduction

Comme dit précédemment, pour créer un réseau sentinelle sur grille informatique pour l'e-santé et l'épidémiologie en Auvergne [30], on a besoin de chaîner les enregistrements des mêmes patients dispersés dans différentes bases de données qui ne partagent pas de clé d'identification commune. Dans un tel cas, la réalisation de la tâche de chaînage des enregistrements nécessite une comparaison des champs d'enregistrements contenant les informations d'identification, tels que nom, prénom, sexe et date de naissance, etc. Pour identifier si deux enregistrements concernent une même personne *via* la comparaison de leurs champs, une des méthodes les plus simples est le chaînage déterministe d'enregistrements (*Deterministic Record Linkage, DRL*) [3]. Avec cette méthode, après une comparaison de chaque champ correspondant au sein d'un couple d'enregistrements (*record pair*) dont chaque enregistrement provient d'une source de données, une décision de chaînage pourrait être prise, fondée sur une concordance de tous les champs comparés. Cependant, la qualité des données pour les champs à comparer n'est pas toujours parfaite, les champs peuvent être soumis à des dégradations telles que des fautes d'épellation et des erreurs typographiques [8]. Ces erreurs dans les champs rendent la méthode DRL moins efficace, par exemple, supposons que deux enregistrements concernent une même personne, ici (*Martin, Michel, M, 19821212*) et (*Martin, Michael, M, 19821212*), une erreur de saisie étant présente dans le champ prénom d'un enregistrement, la méthode DRL serait mise en défaut quant à la décision de chaînage de ces deux enregistrements.

Pour pouvoir trouver la correspondance des enregistrements malgré la variation de saisie dans leurs champs à comparer, la méthode de chaînage probabiliste des enregistrements (*Probabilistic Record Linkage, PRL*) formalisée par Fellegi et Sunter (PRL-FS, voir 1.2.2.1) [6] est une des techniques de choix dans ce contexte. Avec cette méthode, après avoir construit tous les couples d'enregistrements, chaque enregistrement provenant d'une des bases de données à chaîner, un poids (*weight*) de concordance et un poids de discordance fondés sur les ratios de log vraisemblance sont estimés pour chaque champ d'enregistrement. Pour un couple d'enregistrements, son poids est calculé en additionnant le poids de concordance ou le poids de discordance de chaque champ. Dans un couple d'enregistrements, si les contenus pour un champ sont les mêmes, le poids de concordance de ce champ est utilisé pour le calcul du poids du couple d'enregistrements ; sinon c'est le poids de discordance de ce champ est utilisé. Pour la décision de chaînage, si le poids du couple d'enregistrements est supérieur à un seuil de décision (*decision threshold*) déterminé, ce couple d'enregistrements est considéré comme un couple à chaîner [35].

La méthode de chaînage PRL-FS est une des méthodes de chaînage les plus utilisées. L'estimation des paramètres requis pour le calcul des poids de couples d'enregistrements et le choix du seuil de décision est une étape cruciale du processus de la méthode PRL-FS [11,29]. L'utilisation de l'algorithme EM (espérance-maximisation) pour estimer les paramètres est reconnue comme une des méthodes les plus efficaces [15,23]. Dans la littérature, de nombreuses études décrivent le chaînage d'enregistrements par le PRL-FS [11,22,24,36–38], mais à notre connaissance, la mise en œuvre de l'algorithme EM pour estimer des paramètres requis ainsi que le choix du seuil de décision à l'aide d'un des paramètres estimés est rarement décrite en détails.

L'objectif de cette étude est donc de proposer cette description détaillée de l'implémentation de la méthode de chaînage PRL-FS ainsi que son évaluation. Dans cette étude, on a formalisé les paramètres requis dans l'implémentation du PRL-FS, puis on a détaillé l'estimation des paramètres par l'algorithme EM. Avec les jeux de données synthétiques sachant quels couples d'enregistrements concernent les mêmes personnes, la comparaison des paramètres estimés avec les paramètres observés est effectuée. La méthode PRL-FS est implémentée en utilisant les paramètres estimés, puis est comparée avec la méthode DRL vis-à-vis du taux de mauvaises décisions et du temps de calcul.

2.2.2. Matériel et méthodes

2.2.2.1. Approche globale pour implémenter le PRL-FS

Notre étude de l'implémentation de la méthode de chaînage PRL-FS se décompose en quatre étapes successives (Figure 2.5). Après avoir construit des couples d'enregistrements dont chaque enregistrement provient d'une des bases de données, un poids est calculé pour chaque couple d'enregistrements. Un seuil de décision est ensuite déterminé, les couples d'enregistrements dont le poids est supérieur à ce seuil seront considérés comme impliquant la même personne. Enfin, une évaluation du résultat de chaînages des enregistrements est réalisée.

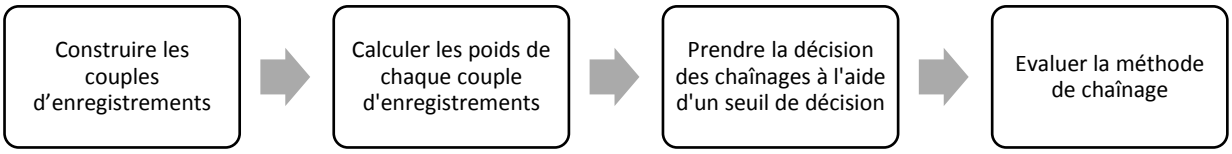


Figure 2.5 Protocole d'évaluation de l'implémentation de la méthode de chaînage PRL-FS

2.2.2.2. Construction des couples d'enregistrements

Dans cette étude, le chaînage est effectué avec les jeux de données synthétiques, car, pour évaluer la méthode de chaînage des enregistrements, on a besoin de l'information sur la vérité de correspondances des enregistrements comme le ferait un *gold standard* pour établir la validité des chaînages. Une telle information est très difficile à obtenir sur les données réelles qui ne partagent pas de clé d'identification commune, car elle nécessite une vérification de tous les enregistrements à comparer entre deux sources de données, ce qui serait un travail considérable.

Le couple des jeux de données (A et B) est créé par un algorithme de genèse des données synthétiques (2.1). Chaque jeu de données générées contient 1 000 enregistrements, et chaque enregistrement contient des champs *nom*, *prénom*, *sexe*, *date de naissance* et une *clé d'identification unique*. Les erreurs typographiques sont introduites dans 10%, 10%, 1% et 10% des champs *nom*, *prénom*, *sexe*, *date de naissance*, respectivement (soit dans environ 30% des enregistrements).

Pour rappel, les couples d'enregistrements qui sont l'objet du processus de chaînage des jeux de données A et B sont tous les couples d'enregistrements possibles résultant du produit cartésien entre ces 2 jeux de données.

2.2.2.3. Calcul des poids de chaque couple d'enregistrements

Les poids des champs dans la méthode PRL-FS sont définis comme les rapports de log vraisemblance fondés sur les paramètres m et u, où m est la probabilité que les contenus pour un champ soient identiques étant donné que le couple d'enregistrements concerne à la même personne, et u est la

probabilité que les contenus pour un champ soient identiques étant donné que le couple d'enregistrements correspond à des personnes différentes [39]. Pour un couple d'enregistrements donné, si les contenus pour le champ i sont identiques, alors le poids de ce champ est :

$$w_i^+ = log_2(m_i/u_i) \tag{2.1}$$

si les contenus pour le champ i sont différents, alors le poids de ce champ est :

$$w_i^- = log_2((1 - m_i)/(1 - u_i)) \tag{2.2}$$

En général, w_i^+ est une valeur positive et w_i^- est une valeur négative.

Pour un couple d'enregistrements j, son poids est calculé comme la somme des poids de tous les champs au sein de ce couple d'enregistrements :

$$w^j = \sum_{i=1}^{n}\left(w_i^+ \times I\left[\gamma_i^j = 1\right] + w_i^- \times I\left[\gamma_i^j = 0\right]\right) \tag{2.3}$$

où :

γ_i^j est une indicatrice (binaire), elle est égale à 1 si les contenus du champ i sont identiques au sein du couple d'enregistrements j, sinon elle est égale à 0.

2.2.2.4. Décision de chaînage des enregistrements

Pour prendre la décision de chaînage, *i.e.* choisir entre chaîner ou ne pas chaîner les 2 enregistrements composant un couple, le poids de chaque couple d'enregistrements peut, dans le cas le plus simple, être confronté à seuil de décision déterminé. Les couples d'enregistrements ayant un poids supérieur au seuil sont considérés comme des couples à chaîner, sinon ils sont considérés comme des couples à ne pas chaîner. Le choix du seuil est une étape importante du processus de chaînage car il aura une incidence fondamentale sur le résultat de chaînages.

En utilisant l'information sur la vérité des correspondances des enregistrements, le seuil de décision optimal est facile à trouver à l'aide d'une ROC-analyse [40]. Mais cette information étant inconnue dans la pratique, nous proposons d'utiliser le paramètre p, *i.e.* la proportion de couples d'enregistrements impliquant une seule et même personne parmi tous les couples d'enregistrements possibles entre deux jeux de données. Supposons que les jeux de données A et B contiennent respectivement N_A et N_B enregistrements, le nombre de couples d'enregistrements concernant les même personnes peut donc être calculé par $p \times N_A \times N_B$. En triant les couples d'enregistrements selon leurs poids en ordre décroissant, il paraît raisonnable de choisir une valeur de seuil de décision à proximité du poids du $(p \times N_A \times N_B)^{ème}$ couple d'enregistrements, notamment si l'on accorde du crédit à la valeur informationnelle de ces poids.

2.2.2.5. Estimation des paramètres m, u et p par l'algorithme EM

Pour calculer les poids de concordance et de discordance d'un champ i, on a besoin des valeurs des paramètres m_i et u_i. La valeur du paramètre p est également nécessaire et ce en dehors du choix du seuil de décision. Afin d'estimer ces paramètres, l'algorithme EM –méthode d'estimation du maximum de vraisemblance impliquant des variables non observées– peut donc être utilisé. Cette méthode commence par une initialisation des paramètres avec des valeurs raisonnables. L'étape E calcule l'espérance de la fonction de vraisemblance en utilisant les valeurs initiales des paramètres. L'étape M maximise la fonction de vraisemblance en utilisant l'espérance calculée à l'étape E afin

d'obtenir de nouveaux paramètres. On itère les étapes E et M jusqu'à ce que les estimations de tous les paramètres convergent [41–43].

Formulation

Dans le contexte du chaînage d'enregistrements, γ_i^j est une information observée qui indique la concordance ou la discordance du champ i du couple d'enregistrements j (voir 2.2.2.3), et g^j est une information non observée et définie comme :

$$g^j = \begin{cases} 1 \;\; si\ le\ couple\ d'enregistrements\ j\ implique\ une\ même\ personne \\ 0 \;\; si\ le\ couple\ d'enregistrements\ j\ implique\ des\ personne\ différentes \end{cases}$$

La fonction de log vraisemblance des données complètes est exprimée comme suit [15] :

$$\begin{aligned} \log f(m,u,p|\boldsymbol{g},\boldsymbol{\gamma}) = \textstyle\sum_{j=1}^{N} g^j \log\left(p \prod_{i=1}^{n} m_i^{\gamma_i^j}(1-m_i)^{\gamma_i^j}\right) \\ + \textstyle\sum_{j=1}^{N}(1-g^j)\log\left((1-p)\prod_{i=1}^{n} u_i^{\gamma_i^j}(1-u_i)^{\gamma_i^j}\right) \end{aligned} \tag{2.4}$$

où :

N est le nombre total de couples d'enregistrements (par exemple, pour les jeux de données A et B qui contiennent respectivement N_A et N_B enregistrements, ce nombre est $N_A \times N_B$).

n est le nombre de champs à comparer.

$\boldsymbol{g} = [g^1,\ g^2, \dots, g^N]^T$

$\boldsymbol{\gamma} = \begin{bmatrix} \gamma_1^1 & \cdots & \gamma_n^1 \\ \vdots & \ddots & \vdots \\ \gamma_1^N & \cdots & \gamma_n^N \end{bmatrix}$

A l'étape E de l'algorithme EM, l'espérance de g^j est calculée comme suit [15] :

$$E[g^j|\boldsymbol{\gamma}] = \frac{p\prod_{i=1}^{n} m_i^{\gamma_i^j}(1-m_i)^{\gamma_i^j}}{p\prod_{i=1}^{n} m_i^{\gamma_i^j}(1-m_i)^{\gamma_i^j} + (1-p)\prod_{i=1}^{n} u_i^{\gamma_i^j}(1-u_i)^{\gamma_i^j}} \tag{2.5}$$

A l'étape M de l'algorithme EM, en annulant les dérivées partielles pour chacun des paramètres m_i, u_i et p de la fonction (2.4), les paramètres estimés sont obtenus comme suit [15] :

$$\hat{m}_i = \frac{\sum_{j=1}^{N} g^j \gamma_i^j}{\sum_{j=1}^{N} g^j} \tag{2.6}$$

$$\hat{u}_i = \frac{\sum_{j=1}^{N}(1-g^j)\gamma_i^j}{\sum_{j=1}^{N}(1-g^j)} \tag{2.7}$$

$$\hat{p} = \frac{\sum_{j=1}^{N} g^j}{N} \tag{2.8}$$

Implémentation

Dans la mise en œuvre de l'algorithme EM pour estimer les paramètres pour la méthode PRL-FS, on commence par choisir les valeurs de départ des paramètres m_i, u_i et p, respectivement fixées à 0.9,

0.1 et 0.0001. Pour la valeur de γ_i^j, elle est obtenue par la comparaison des contenus du champ i au sein du couple d'enregistrements j. En utilisant ces valeurs, l'espérance de g^j est calculée dans la première étape E avec l'équation (2.5). Ensuite, l'étape M estime les paramètres m_i, u_i et p en utilisant l'espérance de g^j calculée à l'étape E avec les équations (2.6) à (2.8). Dans la deuxième étape E, l'espérance de g^j est calculée en utilisant les nouveaux paramètres m_i, u_i et p estimés à la première étape M. On itère les étapes E et M jusqu'à ce que la différence entre deux valeurs d'estimation consécutives des paramètres (surtout p qui nécessite une estimation précise) soit inférieure à 10^{-8}. Les m_i, u_i et p obtenus à la dernière itération seront utilisés pour le calcul des poids de couples d'enregistrements et le choix du seuil de décision.

2.2.2.6. Evaluation de l'implémentation de la méthode PRL-FS

Pour évaluer la méthode de chaînage PRL-FS, on a utilisé 100 couples de jeux de données A et B créés par notre processus de génération des données synthétiques. Avec chaque couple de jeux de données A et B, on a effectué le chaînage en utilisant respectivement la méthode PRL-FS et la méthode DRL. La performance de l'algorithme EM pour estimer les paramètres requis dans le PRL-FS est d'abord évaluée en comparant les paramètres estimés avec les paramètres observés. Ensuite, la méthode PRL-FS est évaluée en elle-même puis comparée avec la méthode DRL vis-à-vis du nombre d'erreurs de chaînage (absence ou excès de chaînage), ainsi que du temps de calcul.

2.2.3. Résultats

2.2.3.1. Réalisation d'un chaînage par la méthode PRL-FS

Plaçons-nous dans un processus de chaînage d'enregistrements provenant des 2 jeux de données A et B, chacun contenant 1 000 enregistrements, et pour lesquels des erreurs typographiques sont introduites dans environ 30% des enregistrements. Après avoir construit le produit cartésien des couples d'enregistrements, la méthode PRL-FS consiste tout d'abord à estimer les paramètres m, u et p par l'algorithme EM. Le Tableau 2.5 présente les paramètres estimés et les poids de concordance et de discordance qui en résultent (à l'aide des équations 2.1 et 2.2), avec une simplification de notation pour les champs nom, prénom, sexe et date de naissance qui sont respectivement notés C1, C2, C3 et C4.

Champ	$\hat{m}_i$	$\hat{u}_i$	Poids de concordance	Poids de discordance	$\hat{p}$
C1	0.81676	0.00016	12.342	-2.448	0.000663
C2	0.81156	0.00305	8.053	-2.403	
C3	0.98869	0.49952	0.985	-5.467	
C4	0.80696	0.00066	10.267	-2.372	

Tableau 2.5 Paramètres estimés et poids de concordance et de discordance calculés en utilisant ces paramètres

Ensuite, au sein du chaque couple d'enregistrements, les champs sont comparés, le résultat des comparaisons est représenté sous la forme d'un pattern se composant de 4 codes binaires. Les 4 codes signifient respectivement la concordance (représentée par 1) ou la discordance (représentée par 0) des champs C1, C2, C3 et C4. Le poids du couple d'enregistrements est calculé à l'aide des équations 2.3 en utilisant l'information fournie par le pattern correspondant au couple. Par exemple, pour un couple d'enregistrements dont le résultat de comparaison des champs est 1110 (concordances des champs C1, C2 et C3 et discordance du champ C4), son poids est 12.342 + 8.053 + 0.985 – 2.372 = 19.008.

Le Tableau 2.6 résume tous les patterns possibles, ainsi que la fréquence de chaque pattern (la fréquence d'un pattern est le nombre de couples d'enregistrements dont les résultats de comparaisons de champs correspondent à ce pattern). Le poids des couples d'enregistrements ayant ces 16 combinaisons de concordance et de discordance des champs sont aussi listés.

Pattern (C1 C2 C3 C4)	Nombre de coulpes d'enregistrements	Poids	Nombre cumulé
1111	353	31.65	353
1101	3	25.19	356
1011	80	21.19	436
1110	81	19.01	517
0111	81	16.86	598
1001	1	14.74	599
1100	3	12.56	602
0101	1	10.40	603
Seuil de décision			
1010	108	8.55	711
0011	334	6.40	1045
0110	1516	4.22	2561
1000	70	2.09	2631
0001	337	-0.05	2968
0100	1553	-2.23	4521
0010	497293	-6.23	501814
0000	498186	-12.69	1000000

Tableau 2.6 Tous les patterns possibles pour les résultats de comparaisons des champs et leurs fréquences (classés en ordre décroissant selon les poids des couples d'enregistrements)

Pour choisir le seuil de décision de chaînage, le paramètre $\hat{p}$ est utilisé. Grâce à ce paramètre, le nombre estimé de couples d'enregistrements impliquant une seule et même personne est estimé par $\hat{p} \times N_A \times N_B$, soit 0.000663 x 1000 x 1000 = 663. Ensuite, en triant les couples d'enregistrements par ordre décroissant de poids (voir le Tableau 2.6), la valeur du seuil de décision à choisir doit être à proximité du poids du 663ème couple d'enregistrements, soit entre 8.55 et 10.40. Par conséquent, les couples d'enregistrements ayant un poids supérieur à 8.55 sont considérés comme des couples à chaîner, au contraire des autres couples d'enregistrements qui sont considérés comme des couples à ne pas chaîner.

2.2.3.2. Evaluation de l'exactitude de l'estimation des paramètres

Le processus de chaînage a été répété 100 fois avec différents jeux de données A et B, au sein de chaque processus, et une estimation des paramètres par l'algorithme EM est effectuée. Sur 100 réalisations, les valeurs des paramètres ont en moyenne convergé après 21 itérations des étapes E et M. Pour évaluer l'exactitude de l'estimation, on a comparé les 9 paramètres estimés (les paramètres m et u pour les 4 champs *nom*, *prénom*, *sexe* et *date de naissance*, respectivement, et le paramètre p) avec les paramètres observés correspondants.

Le Tableau 2.7 résume les différences entre ces deux types de paramètres m et u sur les 100 réalisations, les paramètres estimés et observés pour le champ i sont respectivement notés $\hat{m}_i$, $\hat{u}_i$ et m_i, u_i.

	$min(\lvert\hat{m}_i - m_i\rvert)$	$max(\lvert\hat{m}_i - m_i\rvert)$	$mean(\lvert\hat{m}_i - m_i\rvert)$
C1	0.00044	0.03047	0.00879
C2	0.0001	0.03219	0.00953
C3	2.62×10^{-5}	0.00915	0.00238
C4	7.27×10^{-5}	0.03047	0.00908
	$\mathbf{min}(\lvert\hat{u}_i - u_i\rvert)$	$\mathbf{max}(\lvert\hat{u}_i - u_i\rvert)$	$\mathbf{mean}(\lvert\hat{u}_i - u_i\rvert)$
C1	5.43×10^{-9}	1.83×10^{-5}	5.08×10^{-6}
C2	1.74×10^{-9}	1.73×10^{-5}	4.19×10^{-6}
C3	1.18×10^{-7}	1.57×10^{-5}	4.85×10^{-6}
C4	2.19×10^{-7}	1.67×10^{-5}	5.55×10^{-6}

Tableau 2.7 PRL-FS : différences entre paramètres estimés et observés sur 100 réalisations

Pour le paramètre p, sur 100 réalisations, la différence moyenne entre ses valeurs estimées et observées ($mean(|\hat{p} - p|)$) est 3.44×10^{-05}, la différence minimale ($min(|\hat{p} - p|)$) et la différence maximale ($max(|\hat{p} - p|)$) sont respectivement 2.49×10^{-08} et 1.00×10^{-05}.

On peut observer que les différences entre ces 9 paramètres estimés et observés sont toutes très faibles. Pour évaluer l'impact éventuel de ces légères différences sur la décision de chaînage, on a utilisé respectivement ces deux types de paramètres pour le calcul des poids de couples d'enregistrements et pour le choix du seuil de décision. L'utilisation des paramètres estimés ou observés mène exactement aux mêmes résultats de chaînage.

Quant au temps de calcul d'une telle estimation, sur une station de travail avec un CPU de 2,0 GHz (Intel (R) Xeon (R) E5-2620) et 16 Go de RAM, l'algorithme EM étant implémenté en R (version 2.15.1) [33], il était en moyenne de 30,29 secondes.

2.2.3.3. Evaluation du résultat de chaînage par la méthode PRL-FS

Afin de démontrer la meilleure performance de la méthode PRL-FS par rapport à la méthode DRL, le processus de chaînage est effectué avec le même couple de jeux de données par chacune de ces deux méthodes. Ce processus a été répété 100 fois avec, à chaque fois, un nouveau couple de jeux de données utilisé. La validité des chaînages est évaluée par rapport à l'information sur la vérité de correspondances des enregistrements, utilisée comme *gold standard*. Vis-à-vis d'un résultat de décision de chaînage, on peut donc parler de vrai positif (VP, chaînage de deux enregistrements impliquant la même personne), faux positif (FP, chaînage de deux enregistrements impliquant des personnes différentes), faux négatif (FN, non chaînage de deux enregistrements impliquant la même personne) ou vrai négatif (VN, non chaînage de deux enregistrements impliquant des personnes différentes). Le Tableau 2.8 montre les nombres cumulés de ces 4 catégories de résultats menés par les méthodes DRL et PRL-FS sur 100 réalisations.

Dans ce tableau, on peut constater que la méthode DRL ne génère aucun faux négatif, car la décision de chaînage de cette méthode est fondée sur une concordance exacte de tous les champs comparés. Cependant, elle génère un nombre considérable de faux négatifs, environ 47% des couples d'enregistrements impliquant la même personne ne peuvent pas être chaînés avec cette méthode. En revanche, la méthode PRL-FS permet de chaîner environ 95% des couples d'enregistrements concernant la même personne, aux dépens de la génération de certains faux positifs.

		Vérité de correspondances	
		Même personne	Personnes différentes
DRL	à chaîner	35582	0
	à ne pas chaîner	31140	99933278
PRL-FS	à chaîner	60605	413
	à ne pas chaîner	6127	99932855

Tableau 2.8 Nombres cumulés de 4 catégories de résultats (VP, FP, FN et VN) par les méthodes DRL et PRL-FS (100 réalisations)

Outre la comparaison des résultats menés par les deux méthodes de chaînage présentés en nombres cumulés sur 100 processus, on compare, dans la suite, la performance de ces deux méthodes sur chacun de ces 100 processus. La Figure 2.6 présente le nombre de faux négatifs générés par le DRL (courbe en vert) et les nombres de faux négatifs et de faux positifs générés par le PRL-FS (diagramme empilé en rouge et bleu). On peut constater que l'utilisation du PRL-FS au lieu du DRL permet une réduction moyenne de 246 mauvaises décisions sur 10^6 couples d'enregistrements.

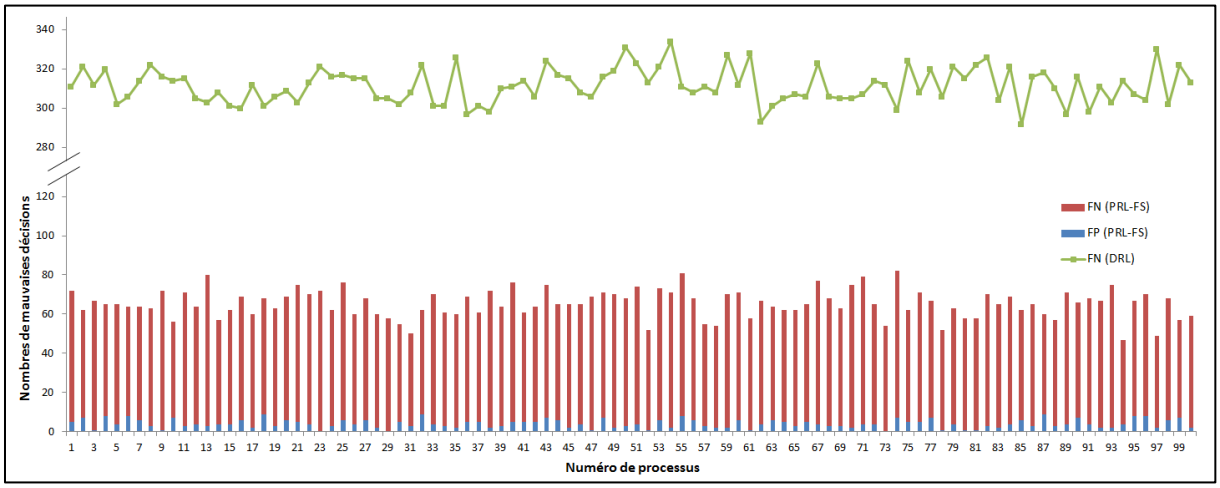


Figure 2.6 Nombres de mauvaises décisions dans chacun des 100 processus de chaînage par les méthodes DRL et PRL-FS

Un autre critère pour évaluer la performance de la méthode de chaînage est le temps nécessaire pour achever un processus de chaînage. Les temps de calcul moyens sur 100 processus pour le DRL et le PRL-FS sont respectivement 10.78 et 44.48 secondes, ces temps de calcul apparaissant relativement stables entre les processus (Figure 2.7).

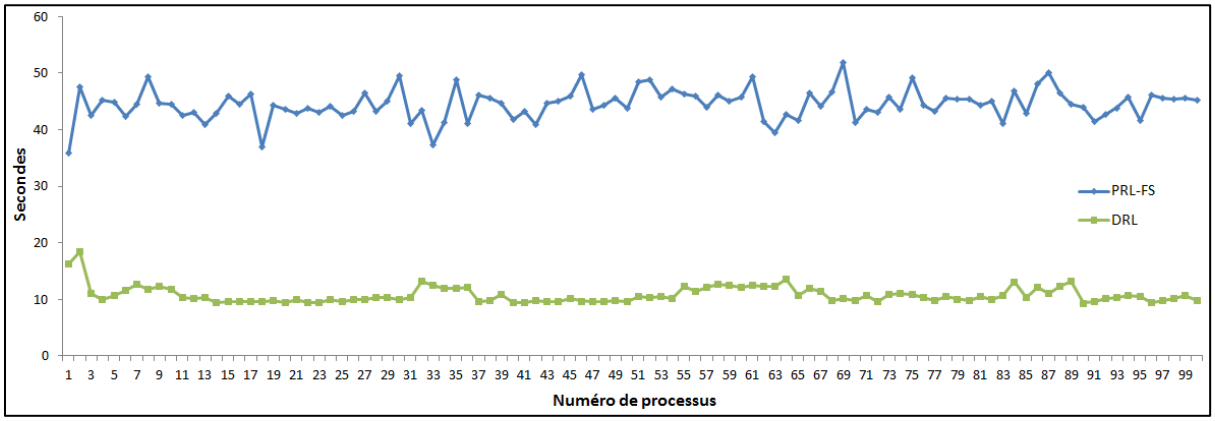


Figure 2.7 Temps de calcul du DRL et du PRL-FS sur 100 réalisations

2.2.4. Discussion et Perspective

L'estimation des paramètres requis pour le calcul des poids de couples d'enregistrements et le choix du seuil de décision est un aspect important de la méthode de chaînage PRL-FS, mais à notre connaissance, la description en détail de la mise en œuvre de l'estimation ainsi que l'utilisation d'un des paramètres estimés pour le choix du seuil de décision sont rarement présentées dans la littérature. Ce travail a donc détaillé l'estimation des paramètres requis et la réalisation du processus de chaînage PRL-FS à l'aide des paramètres estimés, afin de fournir une méthodologie à suivre pour les personnes souhaitant implémenter le PRL-FS voire ses variantes dans leurs travaux.

Dans cette étude, pour évaluer l'estimation des paramètres, chaque processus de chaînage est réalisé en utilisant respectivement les paramètres estimés et observés. On a observé que l'utilisation de ces deux types de paramètres mène exactement aux mêmes résultats de chaînages dans les 100 processus, ce qui démontre que l'algorithme EM est une méthode efficace pour estimer les paramètres requis dans le PRL-FS.

Pour les résultats de chaînage obtenus dans les 100 processus, par rapport à la décision fondée sur une concordance exacte de tous les champs comparés (la méthode DRL) qui permet de chaîner seulement environ 53% des couples d'enregistrements impliquant les mêmes personnes, la méthode PRL-FS permet de chaîner environ 95% des couples impliquant les mêmes personnes, au dépend de générer environ 4 faux positifs sur 10^6 couples d'enregistrements et de consommer 33.7 secondes en plus en moyenne. Cette méthode devrait être particulièrement intéressante pour combiner les données dispersées des mêmes patients –contenant éventuellement des erreurs typographiques– servant à la recherche épidémiologique. En revanche, dans la production de soins, le chaînage à tort des enregistrements peut entraîner des conséquences graves, ce qui doit être absolument évité. Pour éliminer les faux positifs générés avec cette méthode, une vérification manuelle des couples d'enregistrements ayant des poids proches de la valeur du seuil de décision doit donc être effectuée.

Le PRL-FS est une méthode couramment utilisée grâce à son efficacité et à sa simplicité ; cependant, cette méthode présente certaines limites. On peut constater que les poids des couples d'enregistrements calculés dans cette méthode sont les valeurs discrètes avec peu de modalités. La capacité de discriminer les couples d'enregistrements entre eux avec ces poids est donc restreinte. Par exemple, dans la section 2.2.3.1, à l'aide du paramètre p estimé, on devrait prendre la décision de chaînages pour les 663 premiers couples d'enregistrements (étant classés en ordre décroissant selon leurs poids), mais on a décidé de chaîner que les 603 premiers couples d'enregistrements, car les $603^{\text{ème}}$ au $711^{\text{ème}}$ couples d'enregistrements ont le même poids (voir le Tableau 2.6). Par conséquent, les éventuels faux négatifs sont générés.

Par ailleurs, on a constaté que, par exemple, pour le champ prénom dans deux couples d'enregistrements tels que ("*Philippe*", "*Philipe*") et ("*Philippe*", "*Nicolas*"), le même poids de discordance est attribué pour ces deux couples de prénoms, alors que la discordance dans le premier couple est possiblement et seulement causée par une erreur d'épellation.

Pour élargir les modalités des poids des couples d'enregistrements, et aussi pour prendre en compte les similarités des chaînes de caractères dans les champs à comparer, une extension de la méthode PRL-FS est proposée par Winkler [9]. Avec la méthode de Winkler, après avoir calculé le score de similarité des chaînes de caractères dans les champs, une série de poids pour chaque champ est

estimée en fonction du score de similarité du champ. L'implémentation et l'application de cette méthode ont été rarement présentées, on décrira donc dans la section 2.3 l'implémentation de la méthode de Winkler qui est fondée sur la méthodologie présentée dans cette section.

2.2.5. Conclusion

Dans cette étude, on a décrit en détail le processus de l'implémentation de la méthode de chaînage PRL-FS ainsi que son évaluation. Les paramètres requis dans l'implémentation du PRL-FS sont formalisés, et l'estimation de ces paramètres par l'algorithme EM est détaillée. Le PRL-FS est implémenté à l'aide des paramètres estimés, puis il est comparé avec le DRL à l'égard du taux de mauvaises décisions et du temps de calcul. Les résultats sont : (1) la réalisation d'un processus de chaînage par la méthode PRL-FS, (2) un algorithme EM efficace pour estimer les paramètres, (3) la démonstration que le PRL-FS est meilleur que le DRL. En outre, notre étude permet de faciliter la mise en œuvre de la méthode PRL-FS.

2.3. Implémentation de la méthode chaînage probabiliste des enregistrements d'après Winkler

2.3.1. Introduction

Chaîner correctement et efficacement les enregistrements des mêmes patients distribués dans différentes sources de données est essentiel pour la prestation des soins et la recherche épidémiologique. En raison de la législation de protection de la vie privée, l'utilisation d'un identifiant unique du patient –tel que le numéro de sécurité sociale– pour chaîner les enregistrements des patients n'est pas autorisée dans de nombreux pays [11]. Pour rendre le chaînage possible, on a comparé une série de champs d'identification disponibles (par exemple, nom, prénom, sexe et date de naissance, etc.) dans les couples d'enregistrements dont chaque enregistrement provient d'une source de données, afin de prendre une décision de chaînage selon une considération globale des résultats des comparaisons de chaque champ. Malheureusement, ces champs sont parfois soumis à des erreurs typographiques [26,32,44]. Par conséquent, on a besoin d'une méthode de chaînage efficace avec une base théorique solide pour chaîner les enregistrements des patients.

De nombreuses méthodes de chaînage des enregistrements ont été introduites et utilisées au cours des dernières décennies. Une stratégie de chaînage couramment utilisée est le chaînage probabiliste formalisée par Fellegi et Sunter (PRL-FS) [6]. Dans cette méthode, à chaque champ d'enregistrement est attribué un poids de concordance et un poids de discordance fondés sur les ratios de log vraisemblance. Pour un couple d'enregistrements, son poids est calculé en additionnant le poids de concordance ou de discordance de chaque champ au sein du ce couple. Si les contenus pour un champ sont identiques, le poids de concordance de ce champ est utilisé pour le calcul du poids du couple d'enregistrements ; sinon le poids de discordance de ce champ est utilisé. Pour la décision de chaînage, le couple d'enregistrements peut être considéré comme un couple à chaîner si son poids est supérieur à un seuil de décision déterminé [35].

La méthode PRL-FS est relativement facile à mettre en œuvre, mais sa limite tient au fait que le résultat de comparaisons des champs est binaire (concordance/discordance). Par exemple, pour le champ nom dans deux couples d'enregistrement tels que ("*Durand*", "*Durant*") et ("*Durand*", "*Nicolas*"), le même poids est attribué pour ces deux couples de noms, comme ils sont tous discordants. Jaro a proposé que le poids attribué pour le champ doit être proportionnel à la mesure de similarité entre les chaînes de caractères dans ce champ [15]. Winkler a ensuite présenté une méthode de chaînage probabiliste des enregistrements (PRL-W) qui tient compte de la similarité entre les chaînes de caractères dans le calcul du poids des champs [9].

Dans la méthode de PRL-W, la similarité entre les chaînes de caractères dans chaque champ à comparer est quantifiée par un score normalisé entre 0 et 1. Plus le score est proche de 1, plus les chaînes sont similaires. L'intervalle du score [0,1] est divisé en une collection de sous-intervalles disjoints, pour chaque champ et pour chaque sous-intervalle, un poids est attribué, contrairement au PRL-FS où chaque champ a seulement deux poids possibles.

Winkler a formalisé le calcul de ces nouveaux poids pour les champs quand la vérité de correspondances des enregistrements (c'est-à-dire la connaissance de quels couples d'enregistrements concernant les mêmes personnes et de quels couples concernant les personnes différents) est préalablement connue [9]. A notre connaissance, l'application et l'implémentation de

la méthode PRL-W dans la pratique (en l'absence de la vérité de correspondances des enregistrements) n'ont jamais été détaillées dans la littérature. Dans cette section du manuscrit, on décrit en détails comment utiliser l'algorithme EM pour estimer les paramètres nécessaires à la mise en œuvre du PRL-W, et on formalise les équations utilisées dans les étapes d'espérance et de maximisation. Pour évaluer l'exactitude de l'estimation, on compare chaque paramètre estimé avec le paramètre observé correspondant. Enfin, on compare la performance des méthodes PRL-FS et PRL-W en utilisant les jeux de données synthétiques.

2.3.2. Matériel et méthodes

2.3.2.1. Jeux de données

Pour évaluer la performance des méthodes de chaînage des enregistrements, on a besoin de la vérité des correspondances des enregistrements avec lesquelles on peut confronter notre résultat des chaînages. Un tel travail utilisant des données réelles nécessitera des vérifications extrêmement coûteuses sans être sûr de trouver tous les faux positifs (chaînage de deux enregistrements concernant les personnes différentes) et les faux négatifs (non chaînage de deux enregistrements concernant la même personne) dans le résultat de chaînage. On a donc choisi d'effectuer notre étude en utilisant les jeux de données synthétiques.

Dans cette étude, le couple des jeux de données est créé par un algorithme de genèse des données synthétiques (2.1). On a choisi chaque jeu de données contenant 1 000 enregistrements. Les types d'erreurs sont omission, insertion, substitution ou inversion d'un ou plusieurs caractères dans les champs nom, prénom, sexe et date de naissance. Ces quatre types d'erreurs ont été choisis d'après une étude de validation des données de patients [32], dans laquelle ce sont des erreurs typographiques les plus communes dans les champs d'identification. Ces erreurs ont été appliquées à une certaine proportion d'enregistrements dans chaque jeu de données générées. Dans la littérature, la proportion d'erreurs typographiques varie de 8,5% à 36,5% dans différentes bases de données étudiées [11,12,32]. Par conséquent, dans cette étude, on a choisi que les erreurs étaient appliquées à approximativement 30% des enregistrements sélectionnés aléatoirement dans chaque jeu de données. Pour la distribution des erreurs parmi les enregistrements perturbés, il y a 32%, 32%, 32% et 4% d'erreurs dans les champs nom, prénom, sexe et date de naissance, respectivement. Un enregistrement peut avoir plusieurs champs avec des erreurs. Dans nos jeux de données générées, environ 90% des enregistrements a des erreurs dans un seul champ, 9% des enregistrements a des erreurs dans deux champs, 0,9% des enregistrements a des erreurs dans trois champs et 0,1% des enregistrements a des erreurs dans tous les champs. Le processus de genèse des jeux de données synthétiques a été répété 100 fois, donc l'évaluation des méthodes de chaînage des enregistrements reflète le résultat dans 100 exécutions.

2.3.2.2. Calcul des poids de chaque couple d'enregistrements

Par rapport à la méthode PRL-FS, Winkler a amélioré le calcul du poids des champs, la similarité entre les chaînes de caractères dans les champs à comparer est prise en compte. Dans la méthode PRL-W, la similarité entre les chaînes est quantifiée par la distance Jaro-Winkler (voir la section 1.1.1.2). Cette distance prend simultanément en compte la longueur de chaque chaîne de caractères, le nombre de caractères communs et le nombre de transpositions de caractères entre les deux chaînes. Cette distance est représentée sous la forme d'un score normalisé entre 0 et 1, qui est proportionnel au degré de similitude entre les chaînes.

Dans notre étude, on a utilisé la fonction de comparaison des chaînes de caractères « *jarowinkler* » dans le package de R « *RecordLinkage* » pour calculer le score de similarité Jaro-Winkler (JWSS) [45]. Par exemple, en utilisant cette fonction avec ses arguments par défaut, les JWSS des couples de noms ("*Durand*", "*Durant*") et ("*Durand*", "*Nicolas*") sont 0,9333 et 0,4365, respectivement. Pour le champ date de naissance, on a considéré ses valeurs au format de date comme des chaînes de caractères dans le calcul du score de similarité, le JWSS pour les dates de naissance peut donc être calculé comme pour les noms. Par exemple, le JWSS pour ("*08/12/1982*", "*12/08/1982*") est de 0,95. Pour le champ sexe, ses valeurs sont normalisées en « *M* » ou « *F* » par un prétraitement de données, donc le JWSS pour le champ sexe ne peut qu'être 0 ou 1.

Comme le PRL-FS, les poids des champs dans le PRL-W sont aussi les rapports de log vraisemblance, mais les probabilités m et u sont définies différemment. Après avoir calculé le score de similarité pour chaque champ dans chaque couple d'enregistrements, l'intervalle du score [0,1] est divisé en une collection de l sous-intervalles disjoints, pour chaque champ i et pour chaque sous-intervalle $s_{k,i}$, $m_{i,s_{k,i}}$ est la probabilité que le score de similarité du champ i se situe dans $s_{k,i}$ sachant que le couple d'enregistrements concerne la même personne, et le $u_{i,s_{k,i}}$ est la probabilité que le score de similarité du champ i se situe dans $s_{k,i}$ sachant que le couple d'enregistrements concerne les personnes différentes. Si le champ i a un score de similarité dans $s_{k,i}$, alors le poids de ce champ est :

$$w_{i,s_{k,i}} = log_2(m_{i,s_{k,i}}/u_{i,s_{k,i}}) \tag{2.9}$$

Pour un couple d'enregistrements j, son poids est calculé comme la somme des poids de chaque champ :

$$w^j = \sum_{i=1}^{n}\left(w_{i,s_{k,i}} \times I\left[JWSS_i^j \in s_{k,i}\right]\right) \tag{2.10}$$

où :
n est le nombre de champs à comparer.
$JWSS_i^j$ est le score de similarité Jaro-Winkler du champ i au sein du couple d'enregistrements j.
$s_{k,i}$ est la $k^{ième}$ sous-intervalle pour le champ i.

La définition des sous-intervalles disjoints $s_{k,i}$ pour le champ i est présentée dans le Tableau 2.9. La largeur des sous-intervalles est généralement de 0,02, sauf les cas suivants :

- Pour les champs nom et prénom, le [0, 0.6[est considéré comme le premier sous-intervalle, car des couples de noms et de prénoms ayant un JWSS inférieur à 0,60 sont associés –dans la majorité des cas– à des couples d'enregistrements concernant des personnes différentes [9].
- Pour le champ date de naissance, le [0, 0.8[est considéré comme le premier sous-intervalle, car des couples de dates ayant un JWSS inférieur à 0,80 sont généralement associés à des couples d'enregistrements concernant des personnes différentes selon notre observation.
- Pour le champ sexe, sa valeur ne peut qu'être « *M* » ou « *F* », donc le JWSS pour ce champ peut seulement être 0 ou 1.

Le JWSS égale à 1 est traité séparément, car cela signifie que les contenus du champ sont identiques.

sous-intervalle du JWSS	Champ (i)			
	Nom $i = 1$	Prénom $i = 2$	Sexe $i = 3$	Date de naissance $i = 4$
[0, 0.60[	s_0	s0	s0 = [0, 1[	s0= [0, 0.80[
[0.60, 0.62[	s1	s1		
[0.62, 0.64[	s2	s2		
[0.64, 0.66[	s3	s3		
[0.66, 0.68[	s4	s4		
[0.68, 0.70[	s5	s5		
[0.70, 0.72[	s6	s6		
[0.72, 0.74[	s7	s7		
[0.74, 0.76[	s8	s8		
[0.76, 0.78[	s9	s9		
[0.78, 0.80[	s10	s10		
[0.80, 0.82[	s11	s11		s1
[0.82, 0.84[	s12	s12		s2
[0.84, 0.86[	s13	s13		s3
[0.86, 0.88[	s14	s14		s4
[0.88, 0.90[	s15	s15		s5
[0.90, 0.92[	s16	s16		s6
[0.92, 0.94[	s17	s17		s7
[0.94, 0.96[	s18	s18		s8
[0.96, 0.98[	s19	s19		s9
[0.98, 1[	s20	s20		s10
{1}	s21	s21	s1	s11

Tableau 2.9 Sous-intervalles du JWSS pour chaque champ

2.3.2.3. Décision de chaînages des enregistrements

Comme la méthode PRL-FS, pour prendre la décision de chaînages, le poids de chaque couple d'enregistrements est confronté avec un seuil de décision déterminé. Ce seuil peut être choisi à l'aide du paramètre p, qui est la proportion de couples d'enregistrements concernant les mêmes personnes parmi tous les couples d'enregistrements possibles entre deux jeux de données. D'après la définition du PRL-W, plus le poids de couple d'enregistrements est élevé, plus la possibilité que le couple d'enregistrements concerne la même personne est grande. En triant les couples d'enregistrements selon leurs poids en ordre décroissant, il est donc raisonnable de choisir une valeur de seuil de décision à proximité du poids du $(p \times N_A \times N_B)^{\text{ème}}$ couple d'enregistrements (un chaînage implique que deux jeux de données contiennent respectivement N_A et N_B enregistrements).

2.3.2.4. Estimation des paramètres par l'algorithme EM

En utilisant un couple de jeux de données synthétiques où la vérité de correspondances des enregistrements est connue, les paramètres $m_{i,s_{k,i}}$, $u_{i,s_{k,i}}$ et p pourraient être calculés comme :

$$m_{i,s_{k,i}} = \frac{\#JWSS_i^j \in s_{k,i} \text{ et le couple d'enregistrements } j \text{ concerne la même personne}}{\#\text{couples d'enregistrements concerne la même personne}}$$

$$u_{i,s_{k,i}} = \frac{\#JWSS_i^j \in s_{k,i} \text{ et le couple d'enregistrements } j \text{ concerne les personnes différentes}}{\#\text{couples d'enregistrements concerne les personnes différentes}}$$

$$p = \frac{\#\text{couples d'enregistrements concerne la même personne}}{\#\text{couples d'enregistrements}}$$

Cependant, dans la pratique, la vérité de correspondances des enregistrements est inconnue. L'algorithme EM –une méthode pour trouver des estimations du maximum de vraisemblance des paramètres dans les modèles probabilistes qui dépendent des variables non observées– peut donc être utilisé pour estimer ces paramètres.

La fonction de log vraisemblance des données complètes est exprimée comme :

$$\begin{aligned}\log f(m,u,p|\boldsymbol{g},\boldsymbol{\gamma}) &= \sum_{j=1}^{N} g^j \log\left(p \prod_{i=1}^{n} \prod_{k=0}^{l_i} m_{i,s_{k,i}}^{I\left[JWSS_i^j \in s_{k,i}\right]}\right) \\ &+ \sum_{j=1}^{N} (1-g^j) \log\left((1-p) \prod_{i=1}^{n} \prod_{k=0}^{l_i} u_{i,s_{k,i}}^{I[JWSS_i^j \in s_{k,i}]}\right)\end{aligned} \tag{2.11}$$

où :
g^j est une valeur non observée indiquant si le couple d'enregistrements j concerne la même personne (cette valeur doit être 1 si le couple d'enregistrements j concerne la même personne; 0 sinon)
N est le nombre total de couples d'enregistrements.
n est le nombre de champs à comparer.
l_i est le nombre de sous-intervalles pour le champ i.
$\boldsymbol{g} = [g^1, g^2, \dots, g^N]^T$
$\boldsymbol{\gamma} = \begin{bmatrix} \gamma_1^1 & \cdots & \gamma_n^1 \\ \vdots & \ddots & \vdots \\ \gamma_1^N & \cdots & \gamma_n^N \end{bmatrix}$

Par itération E, l'espérance de g^j est calculée comme suit :

$$E\left[g^j|\boldsymbol{\gamma}\right] = \frac{p \prod_{i=1}^{n} \prod_{k=0}^{l_i} m_{i,s_{k,i}}^{I\left[JWSS_i^j \in s_{k,i}\right]}}{p \prod_{i=1}^{n} \prod_{k=0}^{l_i} m_{i,s_{k,i}}^{I\left[JWSS_i^j \in s_{k,i}\right]} + (1-p) \prod_{i=1}^{n} \prod_{k=0}^{l_i} u_{i,s_{k,i}}^{I[JWSS_i^j \in s_k,i]}} \tag{2.12}$$

Par itération M, mettant les dérivées partielles pour chacun des paramètres m_i, u_i et p de la fonction (2.11) égale à zéro, et en respectant les contraintes suivantes :

$$\sum_{k=0}^{l_i} m_{i,s_{k,i}}^{I\left[JWSS_i^j \in s_{k,i}\right]} = 1 \quad \text{et} \quad \sum_{k=0}^{l_i} u_{i,s_{k,i}}^{I\left[JWSS_i^j \in s_{k,i}\right]} = 1$$

Les paramètres inconnus peuvent donc être estimés, ils sont analogues aux paramètres estimés dans le PRL-FS :

$$\widehat{m}_{i,s_{k,i}} = \frac{\sum_{j=1}^{N} g^j I\left[JWSS_i^j \in s_{k,i}\right]}{\sum_{j=1}^{N} g^j} \tag{2.13}$$

$$\hat{u}_{i,s_{k,i}} = \frac{\sum_{j=1}^{N} (1-g^j) I\left[JWSS_i^j \in s_{k,i}\right]}{\sum_{j=1}^{N} (1-g^j)} \tag{2.14}$$

$$\hat{p} = \frac{\sum_{j=1}^{N} g^j}{N} \tag{2.15}$$

Dans la mise en œuvre de l'algorithme EM pour estimer les paramètres, après avoir initialisé les paramètres $m_{i,s_{k,i}}$, $u_{i,s_{k,i}}$ et p, au cours des itérations E et M sont obtenues lorsque le critère de convergence est satisfait. Dans notre étude, on a défini la convergence comme la différence entre deux valeurs du paramètre p estimé consécutivement inférieur à 10^{-8}. Les $m_{i,s_{k,i}}$, $u_{i,s_{k,i}}$ et p obtenus dans la dernière itération sont utilisés pour le calcul des poids de couples d'enregistrements et le choix du seuil de décision.

2.3.2.5. Evaluation de l'exactitude de l'estimation des paramètres et de la performance de la méthode PRL-W

On a comparé des valeurs estimées et observées de chaque paramètre pour évaluer l'exactitude de l'algorithme EM dans l'estimation des paramètres requis par le PRL-W. Les résultats de chaînages menés par le PRL-W en utilisant les paramètres estimés et observés ont également été comparés afin de vérifier si l'inexactitude dans l'estimation peut influer sur le résultat de chaînages. Enfin, on a comparé le PRL-W avec le PRL-FS en ce qui concerne la capacité à réduire le nombre de faux positifs et de faux négatifs, et aussi le temps de calcul.

2.3.3. Résultats

2.3.3.1. Exactitude de l'estimation des paramètres requis par le PRL-W

Dans notre étude, la genèse du couple de jeux de données A et B, et le processus de chaînage d'enregistrements par la méthode PRL-W ont été répétés 100 fois. Dans l'estimation des paramètres, les paramètres ont convergé à la 12ième itération en moyenne. On a comparé les 117 paramètres estimés (44, 44, 24, et 4 paramètres m et u pour les champs nom, prénom, sexe et date de naissance, respectivement, et le paramètre p) avec les paramètres observés correspondants dans chacune des 100 exécutions, leurs différences étaient très faibles, avec des différences minimales et maximales de $1{,}38 \times 10^{-11}$ et 0,0083, respectivement.

Paramètre m et u

- Comparaison au sein d'un processus de chaînage d'enregistrements

La figure 2.8 montre une comparaison de tous les paramètres $m_{i,s_{k,i}}$ et $u_{i,s_{k,i}}$ estimés et observés de chaque champ obtenu dans un processus de chaînage comme un exemple, qui est aléatoirement choisi parmi nos 100 processus. On peut observer que chaque paramètre estimé et le paramètre observé correspondant ont été presque parfaitement superposés.

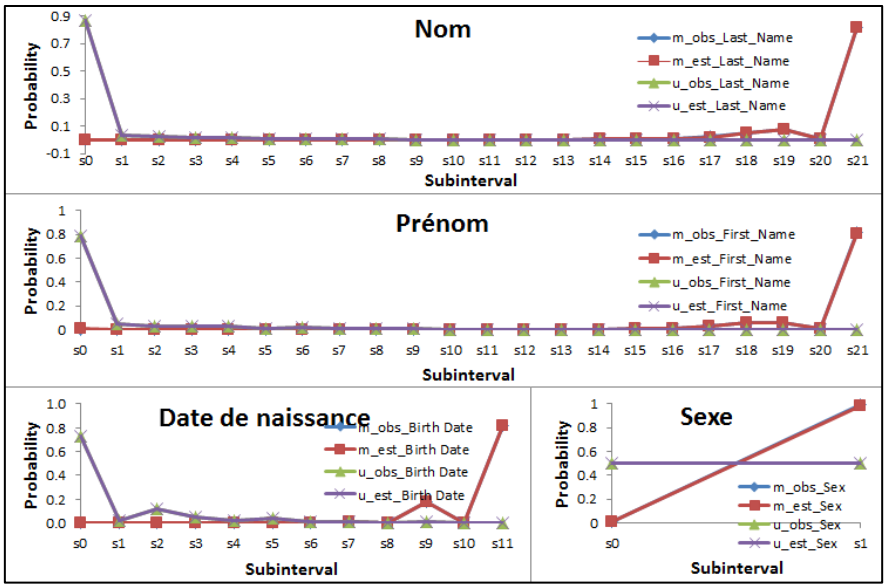


Figure 2.8 Paramètres estimés et observés de chaque sous-intervalle des champs nom, prénom, sexe et date de naissance dans un processus de chaînage d'enregistrements, voir le tableau 2.9 pour les valeurs des sous-intervalles

- Comparaison entre les processus de chaînage d'enregistrements

Pour démontrer si l'exactitude de l'estimation dépend des jeux de données, on a comparé chaque paramètre estimé et observé correspondant dans les 100 processus de chaînage d'enregistrements. Les différences entre ces deux types de paramètres ont toujours été relativement faibles, et l'exactitude de l'estimation est stable, indépendamment des jeux de données.

La figure 2.9 affiche des comparaisons des $m_{1,[0.90,0.92[}$ observés et estimés dans 100 estimations (en utilisant 100 couples de jeux de données) choisi à titre d'exemple parmi 116 paramètres m et u. On peut constater que les paramètres estimés et observés correspondants ont été presque entièrement superposés, avec les différences absolues entre 1,006 × 10^{-8} et 0,0015.

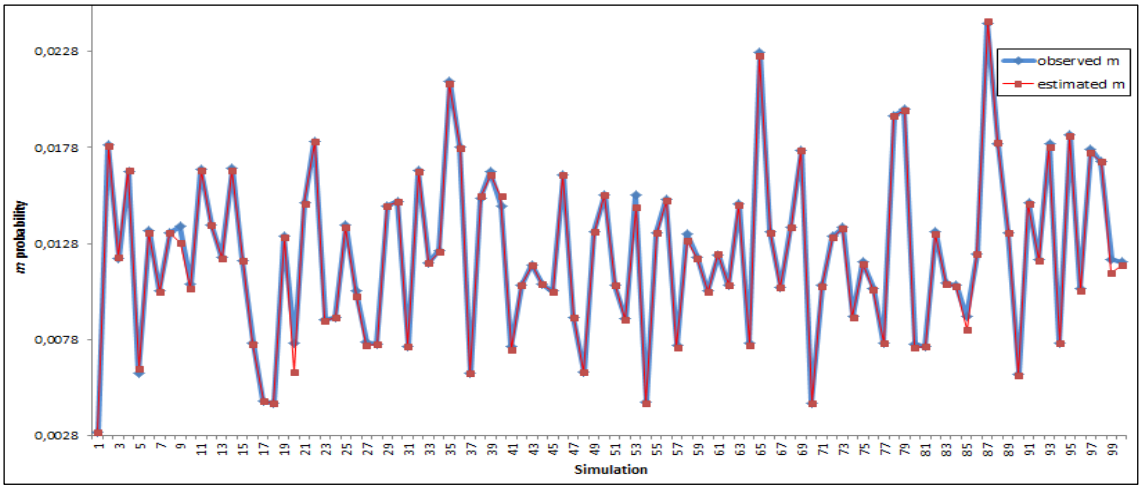


Figure 2.9 Paramètres *m* estimés et observés pour le champ « nom » ayant un JWSS dans [0.90, 0.92[en 100 estimations

Pour évaluer l'impact possible de ces légères différences entre les paramètres $m_{i,s_{k,i}}$ et $u_{i,s_{k,i}}$ estimés et observés sur le résultat de chaînages, on a utilisé ces deux types de paramètres pour calculer le poids des couples d'enregistrements. Dans chaque processus de chaînage, l'utilisation des poids

calculés par les paramètres $m_{i,s_{k,i}}$ et $u_{i,s_{k,i}}$ estimés et observés conduit exactement aux mêmes résultats de chaînages.

Paramètre p

Outre les paramètres $m_{i,s_{k,i}}$ et $u_{i,s_{k,i}}$, on a comparé la différence entre le paramètre p estimé et observé, et on a mesuré l'impact de cette différence sur le résultat de chaînages. La figure 2.10 montre une comparaison entre $\hat{p}$ et p dans 100 estimations. Les valeurs de $\hat{p}$ et p correspondant dans chaque estimation étaient toujours proches, avec des différences $(\hat{p} - p)$ allant de -4,37 × 10^{-6} à 5,96 × 10^{-6}.

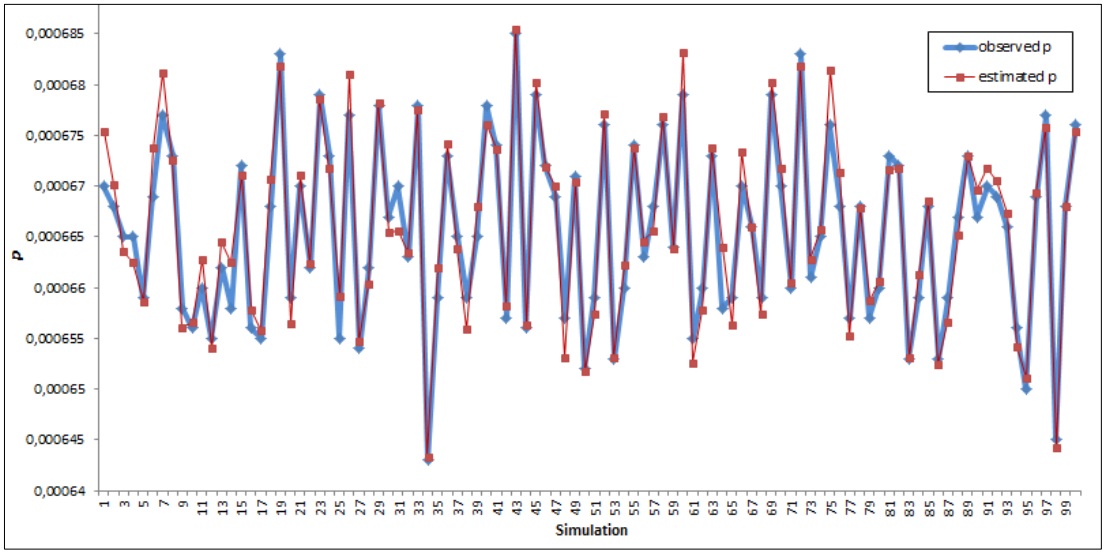


Figure 2.10 Paramètres p estimés et observés dans 100 estimations

L'impact possible de ces différences sur le résultat de chaînage a été évalué comme suit: dans chacun de nos 100 processus de chaînage d'enregistrements, on a utilisé respectivement $\hat{p}$ et p pour définir le seuil de décision, et on a compté les nombres de mauvaises décisions générées à l'issue de chaque processus. On a tracé, dans la figure 2.11, les nombres de faux positifs et de faux négatifs en fonction des valeurs de $\hat{p} - p$. La ligne verticale dans cette figure représente $\hat{p} = p$, à gauche de cette ligne, $\hat{p} < p$, et à droite de cette ligne, $\hat{p} < p$. On peut constater que plus la valeur $\hat{p} - p$ est grande, plus le nombre de faux positifs est élevé ; plus la valeur $\hat{p} - p$ est faible, plus le nombre de faux négatif est élevé, ce qui signifie qu'une sous-estimation ou une surestimation de p pourrait conduire respectivement à une augmentation éventuelle des faux négatifs ou faux positifs.

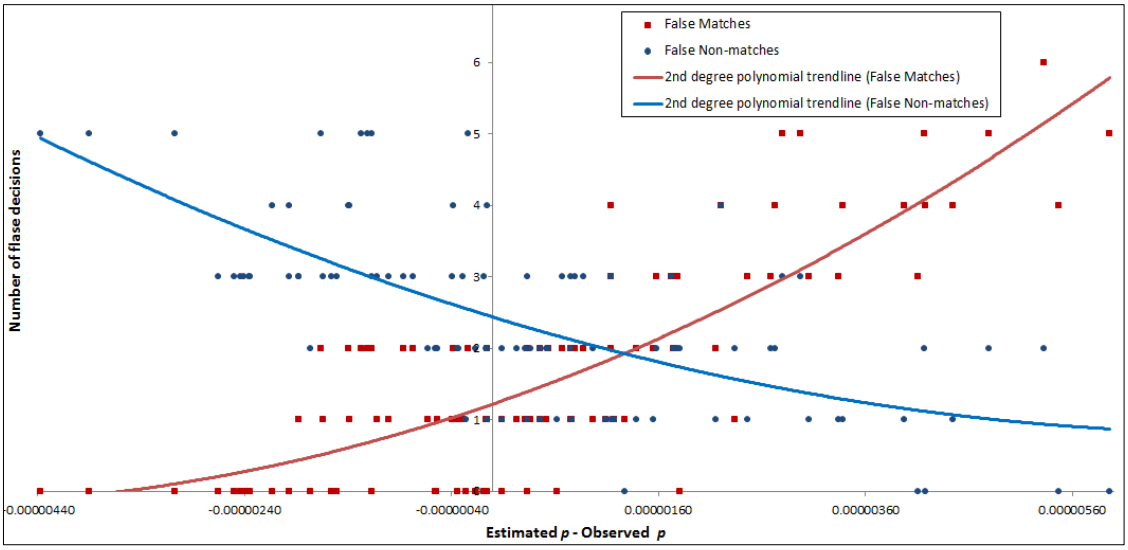


Figure 2.11 Relation entre l'inexactitude des estimations et les mauvaises décisions. Les lignes bleue et rouge sont respectivement les courbes de tendance polynomiale 2ème ordre pour les nombres de faux négatifs et de faux positifs dans les 100 processus

2.3.3.2. Performance du PRL-W

Réduction des mauvaises décisions par rapport au PRL-FS

En utilisant les mêmes jeux de données, on a implémenté respectivement les méthodes de chaînage PRL-FS et PRL-W. Les nombres de mauvaises décisions engendrées par ces deux méthodes ont été comparés. Comme le montre la figure 2.12, on peut observer que le PRL-W est plus performant que le PRL-FS sur à la fois la réduction de faux positifs et de faux négatifs dans chacun des 100 processus de chaînages. Sur 10^6 couples d'enregistrements, le nombre moyen de faux positifs et de faux négatifs engendrés par les PRL-FS étaient respectivement de 5,01 et 61,63 en moyenne; ces nombres engendrés par le PRL-W étaient respectivement de 1,64 et 2,38 en moyenne.

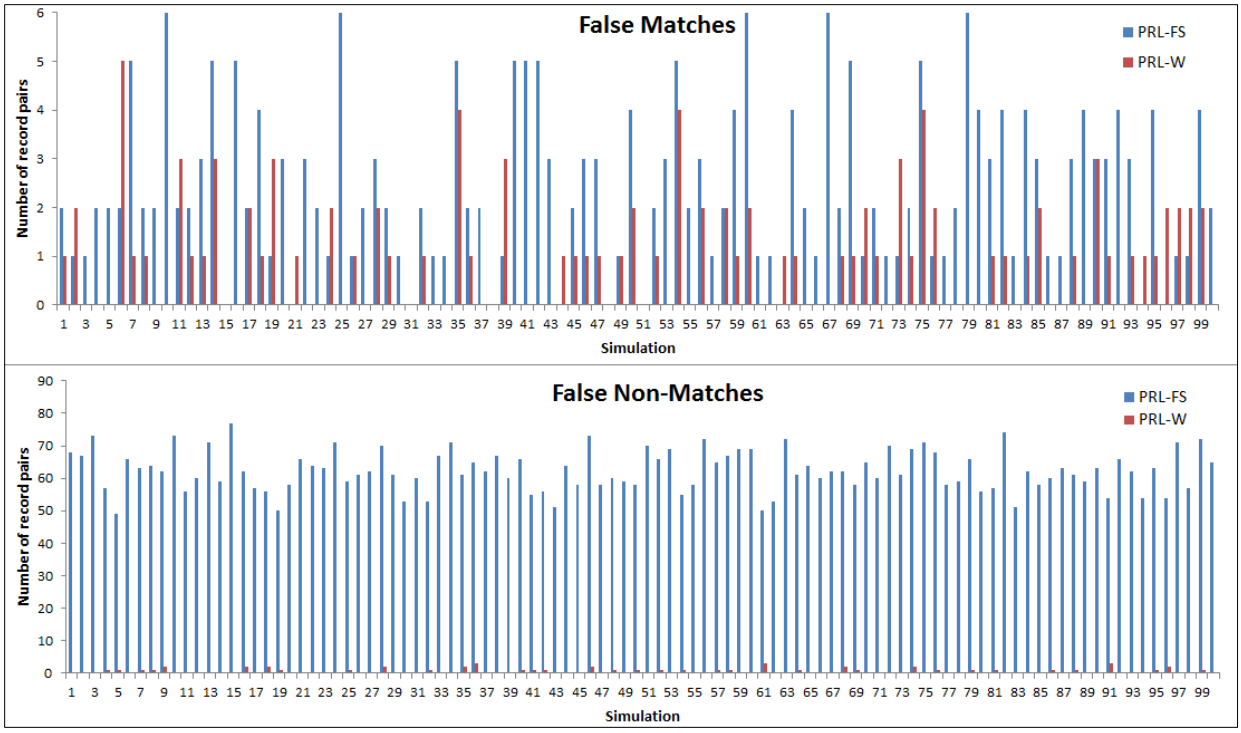


Figure 2.12 Nombres de faux positifs et de faux négatifs engendrés par les deux méthodes PRL dans 100 processus de chaînages

Comparaison des temps de calcul

En plus du nombre de mauvaises décisions, on a comparé le temps de calcul pour effectuer les deux méthodes (figure 2.13). On a utilisé une station de travail avec un CPU de 2,0 GHz (Intel (R) Xeon (R) E5-2620) et un RAM de 16 Go, les deux méthodes étaient implémentées en R (version 2.15.1) [33]. Les temps de calcul du PRL-FS et du PRL-W étaient respectivement 128,01 et 155,37 secondes en moyenne dans les 100 processus.

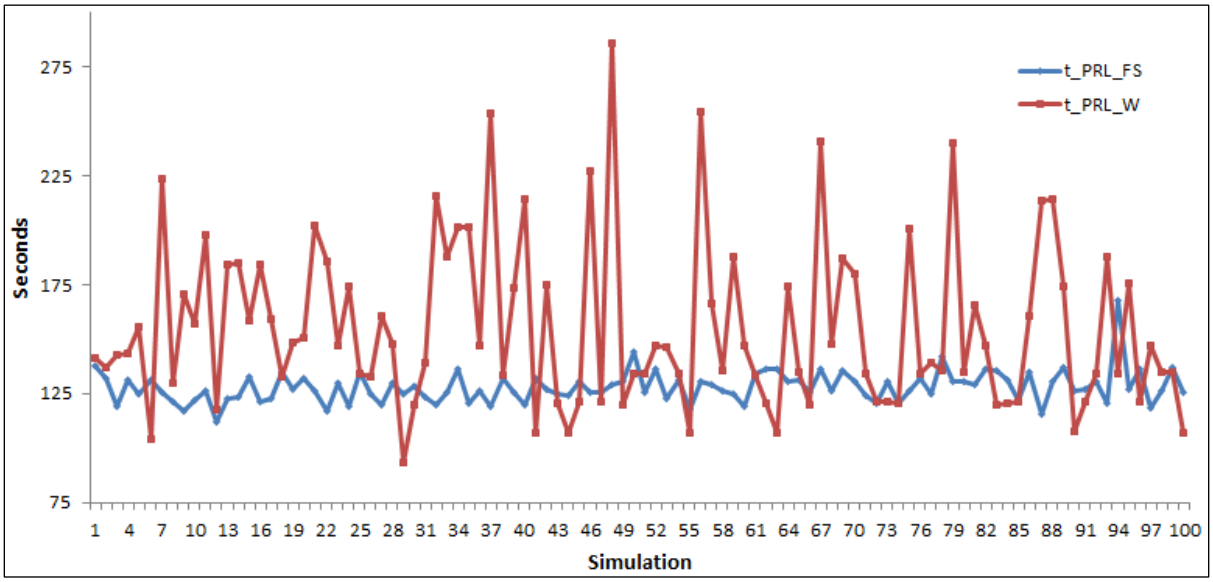


Figure 2.13 Temps de calcul des deux méthodes PRL dans 100 processus de chaînages

Afin d'évaluer l'impact de la taille des jeux de données sur les temps de calcul, on a effectué 10 processus de chaînage PRL-W en utilisant respectivement les jeux de données contenant $1000 \times k$ $(k = 1, 2, 3, 4\ et\ 5)$ enregistrements. Leurs temps de calcul étaient respectivement environ 155, 671, 1 422, 2 665 et 3 992 secondes en moyenne dans 10 processus. On peut constater que le temps de calcul augmente rapidement avec le nombre d'enregistrements dans les jeux de données à chaîner.

2.3.4. Discussion et Perspective

Dans cette étude, on a utilisé les jeux de données synthétiques dont une proportion des enregistrements contient des erreurs typographiques. La proportion et les types des erreurs configurés dans les jeux de données sont choisis selon la littérature [32], il s'agit des erreurs typographiques les plus courantes dans les champs d'identification. Ce travail démontre que le PRL-W est une méthode efficace pour chaîner les jeux de données contenant ces types d'erreurs.

Mais, outre les erreurs typographiques, le problème de données manquantes dans les champs se produit aussi fréquemment avec un taux médian de 10,9% [10,46]. On a généré les jeux de données supplémentaires contenant les valeurs manquantes dans les champs (données non présentées), afin de tester le PRL-W, et on a constaté que cette méthode présente une limite. Dans le calcul de JWSS pour les champs, la valeur manquante est considérée comme une chaîne vide, le JWSS entre une chaîne vide et une chaîne non vide est 0 parce que les contenus dans les deux chaînes sont totalement différents. Le poids du champ correspond au JWSS égale à 0 est donc attribué à un champ contenant la chaîne vide comme pour les champs contenant deux chaînes non vides avec des valeurs totalement différentes (aucune similarité). Il semble inapproprié d'attribuer systématiquement le même poids au champ ayant un JWSS égale à 0 sans considérer le facteur de données manquantes. Par conséquent, dans notre futur travail, on va proposer plusieurs méthodes pour traiter le problème des données manquantes dans les champs à comparer.

Une autre limite de la présente implémentation est le temps de calcul pour les jeux de données contenant un grand nombre d'enregistrements. Avec le PRL-W, un temps de calcul d'environ 155 secondes pour chaîner les jeux de données dont chacun contient 1 000 enregistrements devrait être acceptable pour la plupart des situations. Mais, dans la pratique, la taille des jeux de données serait beaucoup plus grande. On a observé que quand la taille des jeux de données est multipliée par k, l'exécution du processus de chaînage PRL-W nécessite environ 2^k fois plus de temps de calcul. Avec une telle relation entre la taille des jeux de données et le temps de calcul. La réalisation d'un chaînage de deux jeux de données dont chacun contient 100 000 enregistrements (1 000$\times k$, $k =$ 100), le temps de calcul devrait théoriquement être d'environ 1 550 000 secondes (155$\times 2^k$, $k =$ 100), soit d'environ 17 jours et 22 heures. On doit donc chercher une méthode pour réduire le temps de calcul du PRL-W dans le chaînage des jeux de données de grande taille sans altérer la qualité du chaînage. Pour ce faire, on a fait deux propositions.

La première proposition est d'appliquer le procédé de *blocking* pour réduire le nombre de couples d'enregistrements à comparer. Ce procédé divise un jeu de données en plusieurs blocs selon une clé de *blocking* prédéfinie. Les enregistrements ayant la même valeur de clé de *blocking* sont classés dans le même bloc [7]. Par exemple, choisir le mois de naissance comme la clé de *blocking* et définir ("01", "02", ..., "12") comme les valeurs de clé, un jeu de données peut donc être divisé en 12 blocs. Dans un chaînage de deux jeux de données dont chacun contient 12 000 enregistrements, sans le

procédé de *blocking*, 1,44 × 10^8 couples d'enregistrements doivent être comparés ; avec le procédé de *blocking*, 12 blocs dont chacun contient environ 1 000 enregistrements peuvent être créés (supposant que le mois de naissance est uniformément réparti), donc seulement environ 12 × 10^6 couples d'enregistrements doivent être comparés, soit une réduction de 1,32 x 10^8 couples à comparer. La limite de ce procédé est la qualité de données pour la clé de *blocking* serait exigée. Par exemple, des erreurs dans la date de naissance telles que l'inversion du jour et du mois pourraient séparer les enregistrements concernant les mêmes personnes en différents blocs, cela génère donc les faux négatifs dans le résultat de chaînages.

La deuxième proposition est d'utiliser les échantillons des enregistrements dans chacun des deux jeux de données à chaîner pour effectuer l'estimation des paramètres par l'algorithme EM, au lieu d'utiliser tous les enregistrements. Avec cette méthode, on a besoin de déterminer la proportion et/ou le nombre minimaux des enregistrements nécessaires pour l'estimation des paramètres, avec lesquels l'exactitude de l'estimation reste satisfaisante.

2.3.5. Conclusion

On a décrit avec précision comment utiliser l'algorithme EM pour estimer les paramètres nécessaires à l'implémentation du PRL-W, et on a formalisé les équations utilisées dans les étapes d'espérance et de maximisation. La meilleure performance du PRL-W par rapport au PRL-FS a été démontrée en utilisant les jeux de données synthétiques. Les limites du PRL-W ont été discutées, et quelques améliorations possibles pour le PRL-W ont été proposées.

Chapitre 3

Adaptation et amélioration des méthodes de chaînage existantes

Ce chapitre se compose de trois sections :

La **section 3.1** est tirée d'une *soumission en cours* :

« ***Improving Probabilistic Record Linkage Method with Field Similarity Consideration Faced with Missing Data*** »

pour le "JAMIA".

La **section 3.2** est tirée d'une *soumission en cours* :

« ***An Alternative Method for Weight Computation of Probabilistic Record Linkage with Field Similarity Consideration*** »

pour le "JBHI".

La **section 3.3** est traduite (avec une adaptation et modification pour ce manuscrit) de la publication :

« ***An empiric weight computation for record linkage using linearly combined fields' similarity scores*** »

Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE, pages 1346-1349

Xinran Li, Aline Guttmann, Jacques Demongeot, Jean-Yves Boire, Lemlih Ouchchane

DOI : 10.1109/EMBC.2014.6943848

3.1. Amélioration de la méthode probabiliste d'après Winkler face aux données manquantes dans les champs

3.1.1. Contexte

Aujourd'hui, de plus en plus d'enregistrements médicaux électroniques de patients sont distribués dans les systèmes d'information hétérogènes [2], et la combinaison des enregistrements des mêmes patients est souvent nécessaire que ce soit pour la pratique clinique ou la recherche. Pour savoir quels enregistrements concernent les mêmes patients sans l'identifiant unique à travers différents systèmes, on peut comparer les champs d'identification communs (par exemple nom, prénom, sexe et date de naissance) entre deux enregistrements. Pour réaliser une telle tâche, de nombreuses méthodes de chaînage d'enregistrements ont été proposées et appliquées au cours des dernières décennies. A notre connaissance, la méthode étendue du chaînage probabiliste des enregistrements Fellegi-Sunter en tenant compte de la similarité des champs proposée par Winkler (PRL-W) [9], est l'une des méthodes les plus efficaces pour effectuer le chaînage d'enregistrements lorsque les champs d'enregistrement ont des erreurs typographiques. On a implémenté le PRL-W et on a mesuré sa performance dans une étude précédente [25].

Dans la méthode PRL-W, à chaque champ est attribuée une série de poids en fonction de la similarité de champ. Pour quantifier la similarité entre deux chaînes pour un champ, le score de similarité Jaro-Winkler (JWSS) est utilisé [9]. Ce score est normalisé entre 0 (aucune similarité) et 1 (concordance exacte). L'intervalle du score [0,1] est divisé en une collection de sous-intervalles s_k disjoints. Le poids du champ i est le rapport de log-vraisemblance entre les probabilités $m_{i,s_{k,i}}$ et $u_{i,s_{k,i}}$, où $m_{i,s_{k,i}}$ est la probabilité que le score de similarité du champ i se situe dans $s_{k,i}$ sachant que le couple d'enregistrements concerne le même patient, et le $u_{i,s_{k,i}}$ est la probabilité que le score de similarité du champ i se situe dans $s_{k,i}$ sachant que le couple d'enregistrements concerne les patients différents [9]. Pour un couple d'enregistrements, son poids est la somme des poids de chaque champ au sein de ce couple d'enregistrements. Pour prendre une décision de chaînage, on peut comparer le poids du couple d'enregistrements avec un certain seuil, au-dessus duquel le couple d'enregistrements peut être considéré comme couple à chaîner, et en dessous duquel le couple d'enregistrements peut être considéré comme couple à ne pas chaîner [7].

Le PRL-W possède une solide base théorique et est une méthode efficace pour chaîner les jeux de données contenant des erreurs typographiques courantes, telles que l'omission, l'insertion, la substitution ou la transposition d'un ou de plusieurs caractères dans les champs d'identification. Mais cette méthode a une limite, elle ne peut pas traiter efficacement le problème de données manquantes dans les champs, qui s'est souvent produite avec un taux médian de données manquantes de 10,9% selon une étude de Forster [10]. La méthode PRL-W traite passivement les données manquantes comme suit : une donnée manquante dans le champ est considérée comme une chaîne vide, et le JWSS pour les chaînes « vide, non vide » ou les chaînes « vide, vide » est 0 car il n'y a pas d'informations valables fournies par le champ pour calculer le JWSS, et le poids du champ correspondant à JWSS étant 0 est donc attribué. Il semble inapproprié d'attribuer le même poids pour les champs contenant des données manquantes comme pour les champs non vides contenant les données totalement différentes (aucune similarité).

Par conséquent, on a étudié les solutions existantes pour traiter les données manquantes dans les champs d'identification dans la littérature, et on a sélectionné certaines solutions pertinentes parmi elles pour être adaptées pour la méthode PRL-W afin d'améliorer sa performance face à des données manquantes, ils sont : attribuer un poids de zéro pour le champ contenant des données manquantes [47] ; estimer un poids pour le champ contenant les données manquantes comme il le fait pour estimer les poids de concordance et de discordance du champ fondés sur le rapport de log-vraisemblance [23] ; et utiliser un algorithme de l'imputation de distance (*Distance Imputation Algorithm*) qui estime le poids du champ contenant des données manquantes fondé sur le résultat de comparaisons des autres champs au sein du même couple d'enregistrements [46]. Ces solutions ont été initialement conçues pour la méthode PRL-FS et ont certaines limites. On a donc proposé de les adapter et de les améliorer afin de pouvoir être appliquée pour la méthode PRL-W.

3.1.2. Objectifs

Notre objectif est de proposer différentes solutions de traitement des données manquantes pour la méthode PRL-W fondées sur les approches existantes. Lorsque les données manquantes sont introduites dans les champs d'identification, les solutions que l'on a proposées devraient améliorer l'exactitude des décisions de chaînage sans augmenter les temps de calcul. On a formalisé les solutions proposées, puis on les a mises en œuvre et on les a évaluées.

3.1.3. Matériel et méthodes

3.1.3.1. Approche globale

Pour effectuer cette étude, on a choisi d'utiliser –comme dans nos autres études– des jeux de données synthétiques afin de connaître la vérité de correspondances des enregistrements avec laquelle on peut confronter notre résultat des chaînages. Les jeux de données sont générés par notre algorithme de genèse des données synthétiques, dans chaque jeu, les erreurs typographiques et les valeurs manquantes sont présentes dans une proportion des enregistrements. On a ensuite effectué les processus de chaînage d'enregistrements en utilisant respectivement le PRL-W original et le PRL-W avec différentes solutions de traitement des données manquantes. Enfin, on a comparé les nombres de mauvaises décisions générées par chaque stratégie de chaînage d'enregistrements, ainsi que leurs temps de calcul.

3.1.3.2. Genèse des jeux de données

En utilisant l'algorithme de genèse des données synthétiques que l'on a proposé (2.1), les jeux de données A et B dont chacun contient 1 000 enregistrements sont créés. Chaque enregistrement comprend des champs nom, prénom, sexe, date de naissance et une clé d'identification unique. Les erreurs typographiques sont présentes dans 10%, 10%, 1% et 10% des champs nom, prénom, sexe et date de naissance, respectivement. A partir de ces jeux de données générés, 10% des enregistrements ont été aléatoirement choisis, puis le contenu d'un ou plusieurs champs d'identification dans ces enregistrements ont été supprimés afin de produire les données manquantes. Ce taux de données manquantes a été choisi selon l'étude de Forster [10]. Ce processus de genèse des jeux de données a été répété 100 fois.

3.1.3.3. Calcul des poids du couple d'enregistrements par le PRL-W original

Comme présenté précédemment (2.3.2.2), avec la méthode originale de PRL-W, le poids du champ i ayant un JWSS se situe dans $s_{k,i}$ est :

$$w_{i,s_{k,i}} = log_2(m_{i,s_{k,i}}/u_{i,s_{k,i}}) \tag{3.1}$$

Pour le couple d'enregistrements j, son poids est calculé comme la somme des poids de chaque champ :

$$w^j = \sum_{i=1}^{n}\left(w_{i,s_{k,i}} \times I\left[JWSS_i^j \in s_{k,i}\right]\right) \tag{3.2}$$

Pour mettre en œuvre le PRL-W, les JWSS pour les champs ont été calculés par la fonction « *jarowinkler* » inclus dans le package de R « *RecordLinkage* » [45]; et les paramètres requis $m_{i,s_{k,i}}$ et $u_{i,s_{k,i}}$ ont été estimées en utilisant l'algorithme d'espérance-maximisation (EM) [41], qui sont détaillés dans notre étude précédente [25].

3.1.3.4. Calcul des poids du couple d'enregistrements par le PRL-W avec le traitement des données manquantes

Afin de gérer le problème de données manquantes dans le chaînage d'enregistrements, on a proposé trois solutions fondées sur les méthodes existantes pour traiter les données manquantes. Avant l'application de ces solutions, on a examiné chaque champ au sein de chaque couple d'enregistrements afin de trouver si les données manquantes se sont produites. Les résultats de l'examen ont été notés comme suit : $M_i^j = 1$ si les données manquantes se sont produites dans le champ i du couple d'enregistrements j ; $M_i^j = 0$ sinon. On ne distingue pas si l'une ou les deux chaînes dans le champ à comparer sont manquantes, car ni le couple de chaînes « vide, non vide », ni le couple de chaînes « vide, vide » ne fournissent aucune information à propos de la concordance, de la discordance ou de la similarité du champ à comparer, c'est-à-dire que ces deux situations donne la même valeur informationnelle pour le résultat de comparaison du champ.

Solution 1 : Attribution du poids zéro pour les champs contenant des données manquantes

Autrement que le PRL-W original où le poids le plus faible du champ (c'est-à-dire le poids pour le champ n'ayant pas de similarité) est attribué à un champ contenant des données manquantes, cette solution propose d'attribuer un poids de zéro [47]. Car aucune information valable ne peut être obtenue dans la comparaison du champ quand les données sont manquantes, il semble raisonnable d'attribuer un poids « nul » plutôt que d'attribuer un poids pour « pénaliser » ou « récompenser » la présence de données manquantes. Avec cette solution, le poids pour le couple d'enregistrements j est calculé comme suit :

$$w'^j = \sum_{i=1}^{n}\left(w_{i,s_{k,i}} \times I\left[JWSS_i^j \in s_{k,i}\right] \times (1 - M_i^j)\right) \tag{3.3}$$

Solution 2 : Estimation du poids pour le champ contenant les données manquantes fondées sur le rapport de log-vraisemblance

Dans cette solution, on a étendu l'estimation des poids des champs dans la méthode originale de PRL-W. En plus de l'estimation des poids du champ pour chaque sous-intervalle de JWSS, on a proposé d'estimer un poids pour le champ contenant des valeurs manquantes comme pour les sous-

intervalles de JWSS. Pour ce faire, les probabilités $m_{M,i}$ et $u_{M,i}$ sont introduites, où $m_{M,i}$ est la probabilité que le champ i ait des données manquantes sachant que le couple d'enregistrements concerne la même personne, et le $u_{M,i}$ est la probabilité que le champ i ait des données manquantes sachant que le couple d'enregistrements concerne les personnes différentes. De manière analogue, le poids de champ i ayant des données manquantes est le rapport de log-vraisemblance entre $m_{M,i}$ et $u_{M,i}$:

$$w_{M,i} = log_2(m_{M,i}/u_{Mi}) \tag{3.4}$$

Pour un couple d'enregistrements j, son poids est calculé comme la somme des poids de chaque champ au sein de ce couple d'enregistrements :

$$w''^{j} = \sum_{i=1}^{n}\left(w_{i,s_{k,i}} \times I\left[JWSS_i^j \in s_{k,i}\right] \times (1 - M_i^j) + w_{M,i} \times M_i^j\right) \tag{3.5}$$

Les probabilités $m_{M,i}$ et $u_{M,i}$ requises dans l'équation 3.4 peuvent être estimées en utilisant l'algorithme EM que pour $m_{i,s_{k,i}}$ et $u_{i,s_{k,i}}$.

Solution 3 : Calcul du poids pour le champ contenant des données manquantes fondées sur les autres champs au sein du même couple d'enregistrements

Les solutions précédentes attribuent un poids pour le champ contenant des données manquantes indépendamment du résultat de comparaisons des autres champs au sein du même couple d'enregistrements. Par exemple, au sein d'un couple d'enregistrements, en supposant que le champ i soit manquant, que tous les autres champs soient concordants, ou les autres champs soient discordants, le même poids pour le champ i est attribué. Mais il semble plus raisonnable d'attribuer un poids plus élevé (récompense) pour le champ i dans le premier cas (tous les autres champs sont concordants) et un poids plus faible (pénalité) pour le champ i dans le second cas (tous les autres champs sont discordants). Par conséquent, sur la base de l'algorithme de calculs de distance [46], on a proposé une méthode pour obtenir le poids du champ contenant les données manquantes en tenant compte du résultat des comparaisons des autres champs au sein du même couple d'enregistrements.

Cette méthode comprend deux étapes. La première étape consiste à calculer la probabilité conditionnelle du champ i étant approximativement concordant sachant que tous les autres champs étant approximativement concordants. La deuxième étape consiste à calculer le poids du champ i quand il contient des données manquantes en utilisant la probabilité conditionnelle du champ i calculée précédemment.

Etape 1: Calcul de probabilité conditionnelle

Pour jeux de données A et B dont chacun contient n champs, on a noté $C_A = \{C_{A1}, C_{A2}, C_{A3}, \ldots, C_{An}\}$ comme l'ensemble de champs pour le jeu de données A, et $C_B = \{C_{B1}, C_{B2}, C_{B3}, \ldots, C_{Bn}\}$ comme l'ensemble de champs pour le jeu de données B.

La probabilité conditionnelle du champ i est calculée en utilisant tous les couples d'enregistrements possibles (entre les jeux de données A et B) qui ne contiennent pas données manquantes :

$$P(C_{Ai} \sim C_{Bi} | {C_{Ai}}^c \sim {C_{Bi}}^c) = \frac{P(C_{Ak} \sim C_{Bk}; k=1 \text{ à } n)}{P(C_{Ak} \sim C_{Bk}; k=1 \text{ à } n \; sauf \; i)} \tag{3.6}$$

où :

${C_{Ai}}^c$ et ${C_{Bi}}^c$ sont respectivement les compléments de C_{Ai} dans l'ensemble C_A et les compléments de C_{Bi} dans l'ensemble C_B.
$C_{Ax} \sim C_{Bx}$ est l'événement où le contenu dans le champ C_{Ax} et le contenu dans le champ C_{Bx} sont approximativement concordants.

Dans cette solution, les contenus du champ dans un couple d'enregistrements sont considérés comme approximativement concordants si le JWSS du champ est supérieure ou égale à un seuil ∂ donné, c'est-à-dire si $JWSS(C_{Ax}, C_{Bx}) \geq \partial$, alors $C_{Ax} \sim C_{Bx}$.

Étape 2 : Calcul du poids du champ contenant des données manquantes

Dans cette étape, au sein de chaque couple d'enregistrements j, si les données manquantes se sont produites dans le champ i et les JWSS de tous les autres champs sont égaux ou supérieurs à ∂, alors la probabilité conditionnelle $P(C_{Ai} \sim C_{Bi} | {C_{Ai}}^c \sim {C_{Bi}}^c)$ est utilisée pour ajuster le poids de concordance du champ i qui est calculé par la méthode originale de PRL-W, afin d'obtenir un poids pour le champ i contenant des données manquantes :

$$w_{M',i} = w_{i,\{1\}} \times P(C_{Ai} \sim C_{Bi} | {C_{Ai}}^c \sim {C_{Bi}}^c) \tag{3.7}$$

où :
$w_{i,\{1\}}$ est le poids du champ i (calculé par la méthode originale de PRL-W) lorsque le JWSS de ce champ est égal à 1.

Pour un couple d'enregistrements j, son poids est la somme des poids de chaque champ au sein de ce couple d'enregistrements :

$$w'''^j = \sum_{i=1}^{n} \left(w_{i,s_{k,i}} \times I\left[JWSS_i^j \in s_{k,i}\right] \times (1 - M_i^j) + w_{M',i} \times M_i^j \right) \tag{3.8}$$

Cette solution est uniquement appliquée aux couples d'enregistrements n'ayant qu'un champ avec des données manquantes. Pour les couples d'enregistrements ayant plus d'un champ contenant des données manquantes, la solution 1 est utilisée.

3.1.3.5. Décision de chaînages des enregistrements

Comme les méthodes PRL-FS et PRL-W, la décision de chaînages des enregistrements est prise à l'aide du paramètre p, qui est la proportion de couples d'enregistrements concernant les mêmes personnes parmi tous les couples d'enregistrements possibles entre deux jeux de données. Ce paramètre peut être estimé par l'algorithme EM. Supposons que deux jeux de données à chaîner contiennent respectivement N_A et N_B enregistrements, il semble raisonnable de prendre la décision de chaînages pour les $p \times N_A \times N_B$ premiers couples d'enregistrements qui sont classés en ordre décroissant selon leurs poids.

3.1.4. Résultats

Dans notre étude, la genèse des jeux de données a été répétée 100 fois. Avec chaque couple de jeux données générés A et B, on a effectué les processus de chaînage d'enregistrements en utilisant respectivement le PRL-W original et les PRL-W avec les trois solutions de traitement des données manquantes. Dans les sections suivantes, on a noté les PRL-W avec les solutions 1 et 2 comme PRL-W_S1 et PRL-W_S2, respectivement ; la solution 3 se décompose en PRL-W_S3_0.8 et PRL-W_S3_0.9, respectivement, selon les seuils $\partial = 0.8$ et $\partial = 0.9$ choisis pour considérer les contenus d'un champ approximativement concordants (suivant la valeur du seuil).

3.1.4.1. Comparaison des nombres de faux négatifs et de faux positifs

La Figure 3.1 montre les distributions des nombres de faux négatifs et de faux positifs (sur 10^6 couples d'enregistrements entre les jeux de données A et B) générés par le PRL-W et les PRL-W avec les trois solutions pour traiter le problème de données manquantes dans 100 exécutions. On peut constater que toutes les solutions peuvent améliorer l'exactitude de chaînage pour le PRL-W face aux données manquantes dans les champs. Le PRL-W_S2 a généré le moins de faux négatifs (12,97 en moyenne) ; les PRL-W_S1, PRL-W_S3_0.8 et PRL-W_S3_0.9 avaient des performances similaires pour réduire le nombre de faux négatifs (31,22, 29,46 et 29,55 en moyenne, respectivement) par rapport au PRL-W original (41,64 en moyenne). Pour les nombres moyens de faux positifs générés par le PRL-W, PRL-W_S1, PRL-W_S2, PRL-W_S3_0.8 et PRL-W_S3_0.9, leurs différences étaient moins importantes, ils étaient 3,16, 1,56, 2,45, 1,11 et 1,5, respectivement.

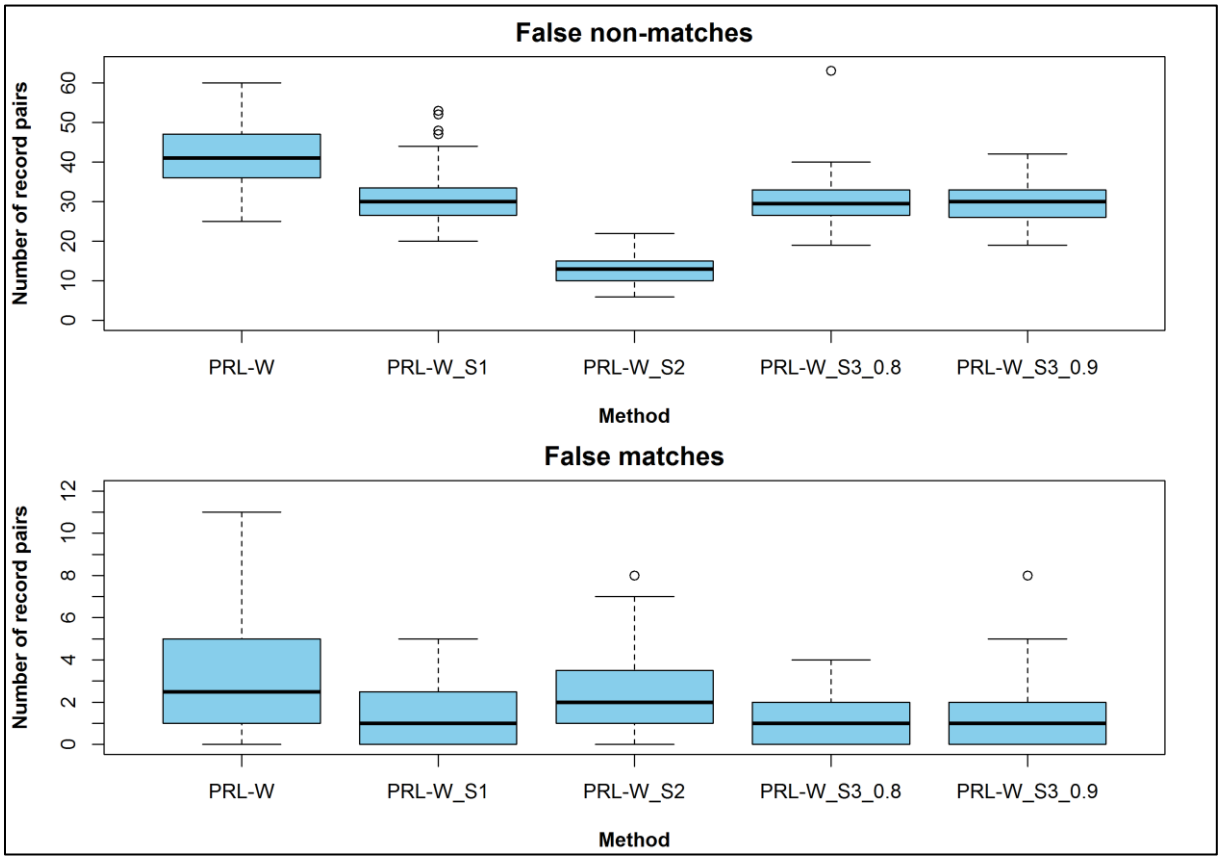


Figure 3.1 Les distributions des nombres de faux négatifs et de faux positifs générés par le PRL-W et les PRL-W avec les trois solutions pour traiter le problème de données manquantes dans 100 exécutions

Pour vérifier si ces différences de nombres de mauvaises décisions générées par chaque stratégie de chaînage d'enregistrements étaient significatives, les *t*-tests appariés (*pairwise t-test*) avec des corrections par la technique de Holm ont été réalisés (Tableau 3.1). On peut observer que les nombres de mauvaises décisions générées par le PRL-W original et par le PRL-W avec des solutions de traitement des données manquantes étaient significativement différentes. En comparant les nombres de mauvaises décisions générées respectivement par les trois solutions, les solutions 1 et 3 n'ont pas eu de différences significatives, et toutes les deux ont significativement différé de la solution 2. Pour la solution 3, l'utilisation du seuil $\partial = 0.8$ ou $\partial = 0.9$, les nombres de mauvaises décisions générées n'étaient pas significativement différents.

Nombres de faux négatifs				
	PRL-W	**PRL-W_S1**	**PRL-W_S2**	**PRL-W_S3_0.8**
PRL-W_S1	$<2\times10^{-16}$	-	-	-
PRL-W_S2	$<2\times10^{-16}$	$<2\times10^{-16}$	-	-
PRL-W_S3_0.8	$<2\times10^{-16}$	0.092	$<2\times10^{-16}$	-
PRL-W_S3_0.9	$<2\times10^{-16}$	0.092	$<2\times10^{-16}$	0.912
Nombres de faux positifs				
	PRL-W	**PRL-W_S1**	**PRL-W_S2**	**PRL-W_S3_0.8**
PRL-W_S1	2.70×10^{-10}	-	-	-
PRL-W_S2	0.011	0.00091	-	-
PRL-W_S3_0.8	5.40×10^{-16}	0.17115	1.60×10^{-7}	-
PRL-W_S3_0.9	$6.00E\times10^{-11}$	0.79936	0.00039	0.19791

Tableau 3.1 Les comparaisons appariées des cinq stratégies de chaînage d'enregistrements à l'égard des nombres de mauvaises décisions en utilisant les t-tests avec des corrections de Holm

3.1.4.2. Comparaison des capacités de chaque solution pour réduire le nombre de mauvaises décisions

En plus de la comparaison des nombres totaux de mauvaises décisions générées par chaque stratégie de chaînage d'enregistrements, on a comparé les décisions pour les mêmes les couples d'enregistrements générés par le PRL-W et le PRL-W avec des solutions de traitement des données manquantes. La répartition de fréquence multivariée (le PRL-W contre le PRL-W avec la solution x) pour les décisions des mêmes les couples d'enregistrements a été montré dans le Tableau 3.2. Dans ce tableau de contingence, les valeurs impliquant les mêmes décisions par des stratégies différentes pour les mêmes couples d'enregistrements ont été indiquées en texte brut, les valeurs concernant les décisions différentes par des stratégies différentes pour les mêmes couples d'enregistrements ont été indiquées en italique. Ces valeurs représentent les nombres cumulés dans 100 processus de chaînage. On peut observer que la solution 2 est plus performante que toutes les autres solutions. Dans un processus de chaînage, par rapport au PRL-W original, l'utilisation de la solution 2 permet de réduire en moyenne 29,97 couples d'enregistrements mal classés au dépend de générer en moyenne 0,59 mauvaises décisions supplémentaires.

		PRL-W_S1		**PRL-W_S2**		**PRL-W_S3_0.8**		**PRL-W_S3_0.9**	
		BD	**MD**	**BD**	**MD**	**BD**	**MD**	**BD**	**MD**
PRL-W	**BD**	99994972	*548*	99995461	*59*	99994482	*1038*	99993803	*1717*
	MD	*1750*	2730	*2997*	1483	*2461*	2019	*3092*	1388

Tableau 3.2 Fréquence cumulée multivariée des décisions de chaînage pour les mêmes couples d'enregistrements générées par le PRL-W par rapport au PRL-W avec des solutions traitant des données manquantes dans 100 processus (BD : bonnes décisions, MD : mauvaises décisions)

3.1.4.3. Comparaison des temps de calcul

On a utilisé une station de travail avec un CPU de 2,0 GHz (Intel (R) Xeon (R) E5-2620) et un RAM de 16 Go, les stratégies PRL-W, PRL-W_S1, PRL-W_S2, PRL-W_S3_0.8 et PRL-W_S3_0.9 étaient implémentées en R (version 2.15.1) [33], leurs temps de calcul étaient 145,8, 146,4, 122,2, 156,2 et

155,9 secondes en moyenne dans les 100 processus, respectivement. On peut observer que le PRL-W_S2 nécessite moins de temps de calcul (Figure 3.2).

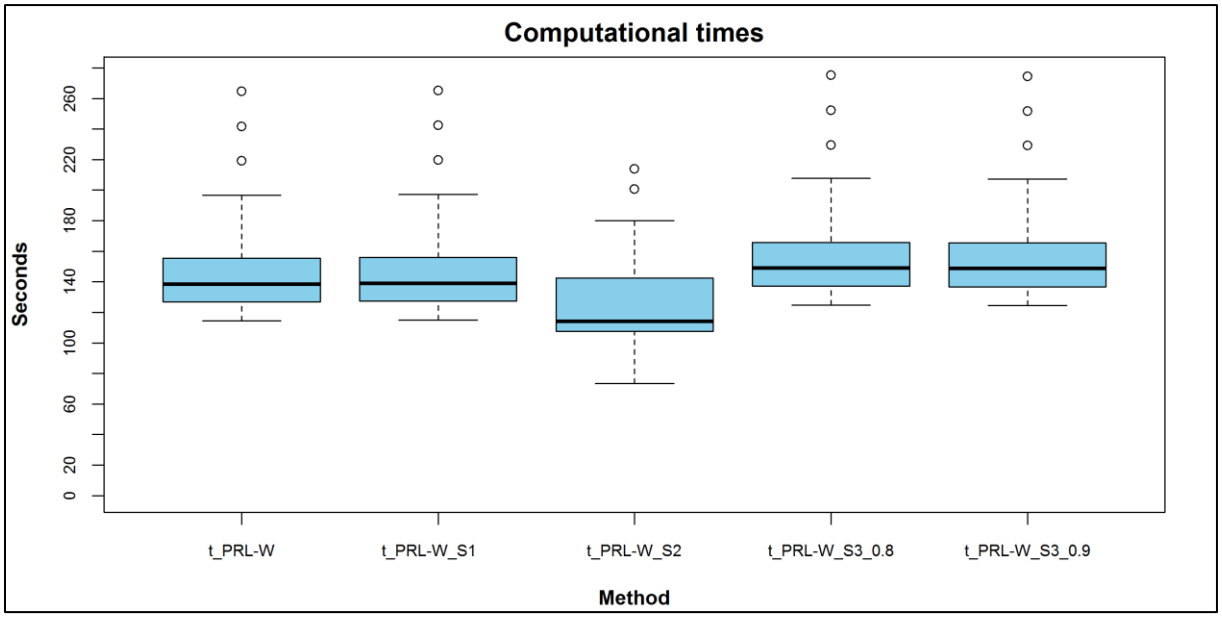


Figure 3.2 Distributions des temps de calcul utilisant le PRL-W et le PRL-W avec trois solutions de traitement des données manquantes dans 100 processus

A travers des *t*-tests appariés pour comparer les temps de calcul des 5 stratégies de chaînage (Tableau 3.3), la différence entre les temps de calcul pour PRL-W_S2 et pour les autres stratégies est statistiquement significative.

	Temps de calcul			
	PRL-W	**PRL-W_S1**	**PRL-W_S2**	**PRL-W_S3_0.8**
PRL-W_S1	1.000	-	-	-
PRL-W_S2	1.5×10^{-9}	6.9×10^{-10}	-	-
PRL-W_S3_0.8	0.028	0.030	$<2\times10^{-16}$	-
PRL-W_S3_0.9	0.030	0.030	$<2\times10^{-16}$	1.000

Tableau 3.3 Comparaisons appariées des cinq stratégies de chaînage d'enregistrements à l'égard des temps de calcul en utilisant les t-tests avec des corrections de Holm

3.1.5. Discussion

Pour les jeux de données contenant des erreurs typographiques courantes et sans données manquantes dans les champs d'identification, le PRL-W est l'une des méthodes les plus efficaces pour effectuer la tâche de chaînage d'enregistrements. Malheureusement, la méthode PRL-W devient moins efficace quand les données manquantes se sont produites dans les champs d'identification. On a donc proposé trois solutions pour améliorer le résultat de chaînages du PRL-W dans le cas de la présence de données manquantes dans les champs à comparer. Par rapport au PRL-W, toutes les solutions de traitement des données manquantes proposées dans cette étude peuvent réduire considérablement le nombre de mauvaises décisions. Quant au temps de calcul par rapport au PRL-W, l'utilisation de la solution 3 l'augmente légèrement (environ 10 secondes en plus pour chaîner deux jeux de données contenant respectivement 1000 enregistrements) ; il n'y avait pas

d'augmentation significative du temps de calcul en utilisant la solution 1 ; et avec la solution 2, le temps de calcul a été significativement réduit.

La solution 1 est l'une des solutions les plus simples pour gérer le problème des données manquantes, avec laquelle le champ contenant des données manquantes est simplement attribué un poids de zéro, parce que la comparaison des valeurs manquante est non informative. Par rapport au traitement par défaut du PRL-W que le champ contenant des données manquantes est considérée comme une discordance de champ et le poids le plus faible du champ est attribué, la solution 1 permet d'attribuer un poids « neutre », ce qui diminue l'éventualité d'absences de chaînage du couple d'enregistrements. Mais, cette solution présente l'inconvénient que le poids zéro est systématiquement attribué au champ contenant des données manquantes indépendamment du statut des autres champs au sein du même couple d'enregistrements.

La solution 2 a proposé d'étendre le résultat catégoriel de comparaison du champ dans le PRL-W original en ajoutant un nouveau niveau au résultat de comparaison qui est l'événement de données manquantes, le poids pour le champ contenant des données manquantes est estimé sur la même base théorique que l'estimation des poids pour les sous-intervalles de JWSS. Par rapport aux autres solutions qui ajustent le poids du champ estimé par le PRL-W original (traitement des données manquantes *post*-PRL-W), cette solution estime directement le poids pour le champ contenant des données manquantes par le PRL-W étendu (traitement des données manquantes *intra*-PRL-W), qui présente les avantages suivants : (1) les poids estimés pour le champ sont plus pertinents, car ils sont estimés par l'algorithme EM avec une considération globale de tous les niveaux du résultat de la comparaison du champ, y compris le niveau pour l'événement de données manquantes, contrairement au PRL-W original où les données manquantes sont inappropriées et classées au niveau du champ n'ayant pas de similarité ; (2) le temps de calcul par rapport au PRL-W originale est réduit, parce que : dans l'estimation des paramètres de PRL-W par l'algorithme EM, les étapes d'espérance et de maximisation itèrent jusqu'à ce que les critères de convergence soient satisfaits. On a observé que le temps de l'estimation est proportionnel au nombre d'itérations pour la convergence des paramètres, et l'introduction de ce nouveau « données manquantes » a conduit à une réduction du nombre d'itérations, de sorte que le temps de calcul est réduit.

La solution 3 consiste à attribuer un poids pour le champ contenant les données manquantes sur la base de règles d'association pour les relations entre les champs. Les règles d'association de chaque champ ont été créées fondées sur tous les couples d'enregistrements ne contenant pas de données manquantes. Dans un processus de chaînage, lorsque la valeur manquante s'est produite dans un champ, le poids pour ce champ est ajusté en utilisant la règle d'association de ce champ. Le défaut de cette méthode est que les contenus d'un champ sont approximativement concordants à partir d'une valeur empirique d'un seuil. Dans cette étude, on a fixé le seuil à 0,8 et 0,9. Par exemple, avec le seuil de 0,8, les couples de chaînes ("*Yohann*", "*Johan*") et ("*Yohann*", "*Yoan*") ont tous été considéré comme approximativement concordants puisque leurs JWSS étaient 0,8222 et 0,9111, respectivement ; mais avec le seuil de 0,9, seul le couple de chaîne ("*Yohann*", "*Yoan*") a été considéré comme approximativement concordant.

Avec chacune de ces trois solutions, on a observé que la réduction des faux positifs est moins évidente que celle des faux négatifs. Le PRL-W original attribue le poids du champ le plus faible pour les champs contenant les données manquantes, de sorte que les couples d'enregistrements ayant

des données manquantes ont moins de possibilité d'être classés comme couples à chaîner. En conséquence, le nombre de faux positifs générés par le PRL-W est relativement faible, et, sur ce point, l'amélioration potentielle apportée par des solutions annexes est donc limitée. Parmi toutes les solutions proposées, la solution 2 présente la meilleure performance à la fois dans la réduction des mauvaises décisions (considération globale des nombres de faux positifs et de faux négatifs) et sur l'amélioration du temps de calcul. Cependant, il est moins efficace pour réduire les faux positifs par rapport aux autres solutions. Pour gérer cette difficulté, un examen manuel pourrait être réalisé pour les couples d'enregistrements contenant des données manquantes étant considérés comme les couples à chaîner. Un tel examen manuel peut également toujours être proposé pour les autres solutions afin de réduire le risque de faux positifs.

3.1.6. Conclusion

Dans cette étude, pour améliorer la performance du PRL-W en présence de données manquantes, on a proposé trois solutions fondées sur des approches de traitement des données manquantes. Lorsque les données manquantes se sont produites dans les champs d'identification, toutes les solutions proposées permettent d'améliorer l'exactitude des décisions de chaînage du PRL-W sans augmenter (voire réduire) le temps de calcul.

3.2. Méthode alternative pour le calcul du poids des enregistrements en considérant la similarité des champs

3.2.1. Contexte

Le PRL-W est une extension du PRL-FS en prenant en compte les mesures de similarité des chaînes de caractères dans les champs à comparer, cette méthode a une solide base théorique et elle est une des méthodes les plus performantes pour le chaînage d'enregistrements contenant des erreurs typographiques courantes, telles que l'omission, l'insertion, la substitution ou la transposition d'un ou de plusieurs caractères dans les champs d'identification.

Dans la méthode de PRL-W, la similarité entre les chaînes de caractères dans chaque champ à comparer est quantifiée par un JWSS compris entre 0 et 1. L'intervalle du score [0,1] est divisé en une collection de sous-intervalles disjoints, pour chaque champ et pour chaque sous-intervalle, un poids est attribué. Cependant, l'utilisation du PRL-W au lieu de PRL-FS entraîne l'estimation des paramètres en plus. Par exemple, dans l'étude de Winkler [9], l'implémentation du PRL-W nécessite 42 paramètres à estimer pour chacun des champs nom et prénom, contrairement au PRL-FS où seulement 2 paramètres sont à estimer par champ. Cette augmentation du nombre de paramètres à estimer complique non seulement la mise en œuvre du processus de chaînage, mais aussi augmente le temps de calcul.

Par conséquent, on cherche à développer une méthode alternative pour calculer les poids du champ qui permet également de tenir compte de la similarité du champ, mais avec moins de paramètres à estimer. Pour ce faire, on propose une fonction de calcul des poids du champ contenant les arguments (de fonction) suivants : (1) le poids de concordance (w_i^+), et (2) le poids de discordance (w_i^-) du champ i estimés par la méthode PRL-FS, afin de quantifier l'importance de l'information fournie champ i ; (3) le JWSS du champ i au sein du couple d'enregistrements j ($JWSS_i^j$) utilisé dans la méthode PRL-W, afin de quantifier la similarité du champ comparé. On émet l'hypothèse que la performance du chaînage d'enregistrements en utilisant des poids calculés par notre fonction pourrait combiner à la fois l'avantage du PRL-FS (temps de calcul court) et l'avantage du PRL-W (haute exactitude de chaînage).

3.2.2. Objectifs

L'objectif de ce travail est de proposer une fonction appropriée des w_i^+, w_i^- et $JWSS_i^j$ pour calculer les poids du champ i au sein du couple d'enregistrements j, qui peut être exprimée en $w_i^j = f(w_i^+, w_i^-, JWSS_i^j)$. On cherche à assurer que le poids w_i^j est proportionnel à la similarité du champ i, de sorte que plus les contenus du champ sont similaires, plus le poids de champ est élevé. On évalue la fonction susmentionnée en comparant les décisions de chaînage subséquentes avec celles générées par le PRL-FS et le PRL-W.

3.2.3. Matériel et méthodes

3.2.3.1. Jeux de données

Dans cette étude, le processus de chaînage est effectué en utilisant les jeux de données synthétiques, avec lesquels on peut connaître la vérité de correspondances des enregistrements qui

sert à évaluer notre résultat des chaînages. Grâce à l'algorithme de genèse des données synthétiques (2.1), les jeux de données A et B comprenant chacun 1 000 enregistrements sont créés. Chaque enregistrement contient des champs nom, prénom, sexe, date de naissance et d'une clé d'identification unique. Les erreurs typographiques sont présentes dans 10%, 10%, 1% et 10% des champs nom, prénom, sexe et date de naissance, respectivement. Ce processus de genèse des jeux de données a été répété 100 fois.

3.2.3.2. Définition de la fonction de calcul du poids du champ

On cherche à définir une fonction qui permet de calculer les poids du champ proportionnels aux similarités du champ comme les poids calculés avec la méthode PRL-W, mais sans l'estimation de beaucoup plus de paramètres. On propose de créer une série de poids pour chaque champ i compris entre w_i^- et w_i^+, qui sont estimés par la méthode PRL-FS possédant une solide base théorique.

La fonction doit satisfaire les conditions suivantes : lorsque le $JWSS_i^j$ est inférieur à un certain seuil que les contenus dans le champ i sont considérés comme pour les personnes différentes, le poids w_i^- est attribué ; lorsque le $JWSS_i^j$ est égal à 1, le poids w_i^+ est attribué ; lorsque le $JWSS_i^j$ est entre le seuil susmentionné et 1, le poids entre w_i^- et w_i^+ est attribué par une interpolation linéaire.

Pour créer une telle fonction pour chaque champ, on doit (1) estimer les poids w_i^- et w_i^+, et (2) calculer le score de similarité $JWSS_i^j$.

Estimation des poids de discordance et de concordance avec le PRL-FS

On rappelle ici que le poids du champ dans la méthode PRL-FS est défini dans la section 2.2.2.3. Notons que m_i est la probabilité que les contenus du champ i sont identiques sachant que le couple d'enregistrements appartient à la même personne, et que u_i est la probabilité que les contenus pour le champ i sont identiques sachant que le couple d'enregistrements appartient à des personnes différentes. Pour un couple d'enregistrements, si les contenus pour le champ i sont identiques, alors le poids de ce champ est :

$$w_i^+ = log_2(m_i/u_i) \tag{3.9}$$

si les contenus pour le champ i sont différents, alors le poids de ce champ est :

$$w_i^- = log_2((1 - m_i)/(1 - u_i)) \tag{3.10}$$

En général, w_i^+ est une valeur positive et w_i^- est une valeur négative.

Les probabilités m_i et u_i utilisées dans les équations 3.9 et 3.10 peuvent être estimées en utilisant l'algorithme EM.

Mesure la similarité des contenus du champ

La similarité entre les contenus du champ est quantifiée par le JWSS comme c'est le cas dans le PRL-W. Les JWSS de chaque champ au sein de chaque couple d'enregistrements ont été calculés par la fonction « *jarowinkler* » inclus dans le package de R « *RecordLinkage* » [45]. En général, une telle mesure de similarité est appliquée pour comparer un couple de chaînes de lettres. Néanmoins, on a également utilisé le JWSS pour quantifier la similarité des dates pour le champ date de naissance, puisque l'on a considéré les valeurs du champ comme une chaîne sous le format *aaaammjj*. Par

exemple, le JWSS pour ("*08/12/1969*", "*12/08/1969*") est de 0,96. Pour le champ sexe, ses valeurs sont normalisées en « *M* » ou « *F* », donc le JWSS pour ce champ ne peut qu'être 0 ou 1.

Calcul des poids intermédiaires entre les poids de discordance et de concordance du champ

En utilisant le poids de concordance (w_i^+) et le poids de discordance (w_i^-) pour le champ i estimés, et le JWSS pour le champ i au sein du couple d'enregistrements j, on a proposé la fonction suivante :

$$w_i^j = \begin{cases} w^- & \forall\, JWSS_i^j \in [0, T_i[\\ \dfrac{w^+ - w^-}{1 - T_i}\left(JWSS_i^j - T_i\right) + w^- & \forall\, JWSS_i^j \in [T_i, 1[\\ w^+ & \forall\, JWSS_i^j = 1 \end{cases} \tag{3.11}$$

où :
T_i est le seuil de JWSS pour le champ i, en dessous duquel les couples d'enregistrements sont considérés comme pour les personnes différentes. On a fixé $T_i = 0.6$ pour les champs nom et prénom selon Winkler [9]. Par exemple, le JWSS pour les prénoms "Robert" et "Richard" est 0,5857, avec le seuil égal à 0,6, ces deux prénoms sont considérés comme les prénoms impliquant les personnes différentes. On a fixé $T_i = 0.8$ pour le champ date de naissance sur la base des observations empiriques. Par exemple, le JWSS pour les dates de naissance "19820327" et "19670815" est 0,7422, ces deux dates de naissance sont très peu probables de correspondre à la même personne. Enfin, on a fixé $T_i = 1$ pour le champ sexe, comme le JWSS pour ce champ est binaire.

Pour le couple d'enregistrements j, son poids correspond à la somme des poids de chaque champ :

$$w^j = \sum_{i=1}^{n} w_i^j \tag{3.12}$$

Le processus de chaînage d'enregistrements utilisant ce poids est noté PRL-I.

3.2.3.3. Décision de chaînages des enregistrements

Dans un processus de chaînage par une méthode PRL, un poids est attribué à chaque couple d'enregistrements, plus ce poids est élevé, plus il est vraisemblable qu'il s'agisse d'une même personne. Pour évaluer le pouvoir discriminant des poids calculés par chaque méthode PRL (PRL-FS, PRL-W et PRL-I) dans la décision de chaînages, on a utilisé la vérité de correspondances des enregistrements comme un *gold standard* pour définir le seuil de décision optimal de chaque méthode PRL par une ROC-analyse [40]. Pour chaque méthode, les couples d'enregistrements ayant un poids supérieur au seuil sont considérés comme les couples à chaîner, sinon ils sont considérés comme les couples à ne pas chaîner.

3.2.3.4. Evaluation et comparaison des méthodes PRL

Avec chaque couple de jeux de données A et B, on a effectué un processus de chaînage d'enregistrements en utilisant respectivement les méthodes PRL-FS, PRL-W et PRL-I. Pour leur implémentation, la réalisation du PRL-FS est présentée dans la section 2.2 ; on a détaillé comment mettre en œuvre le PRL-W dans la section 2.3 [25] ; le PRL-I peut être implémenté en utilisant les méthodes décrites dans cette présente étude. La comparaison de ces trois méthodes PRL était fondée sur les mesures de faux positifs (chaînages de deux enregistrements concernant les

personnes différentes), de faux négatifs (non chaînages des enregistrements concernant les mêmes personnes), ainsi que sur les temps de calcul.

3.2.4. Résultats

3.2.4.1. Comparaison des nombres de faux négatifs et de faux positifs

Sur 10^6 couples d'enregistrements, le nombre moyen de faux négatifs/faux positifs générés par les méthodes PRL-FS, PRL-W et PRL-I étaient respectivement 61,89 / 2,62, 1,44 / 1,06 et 1,62 / 1,1. Dans la Figure 3.1, on a les distributions de nombres de faux négatifs et de faux positifs générés par les méthodes PRL-FS, PRL-W et PRL-I dans 100 processus. On peut constater que le PRL-W et le PRL-I ont des performances similaires, et que les deux méthodes sont largement meilleures que les PRL-FS.

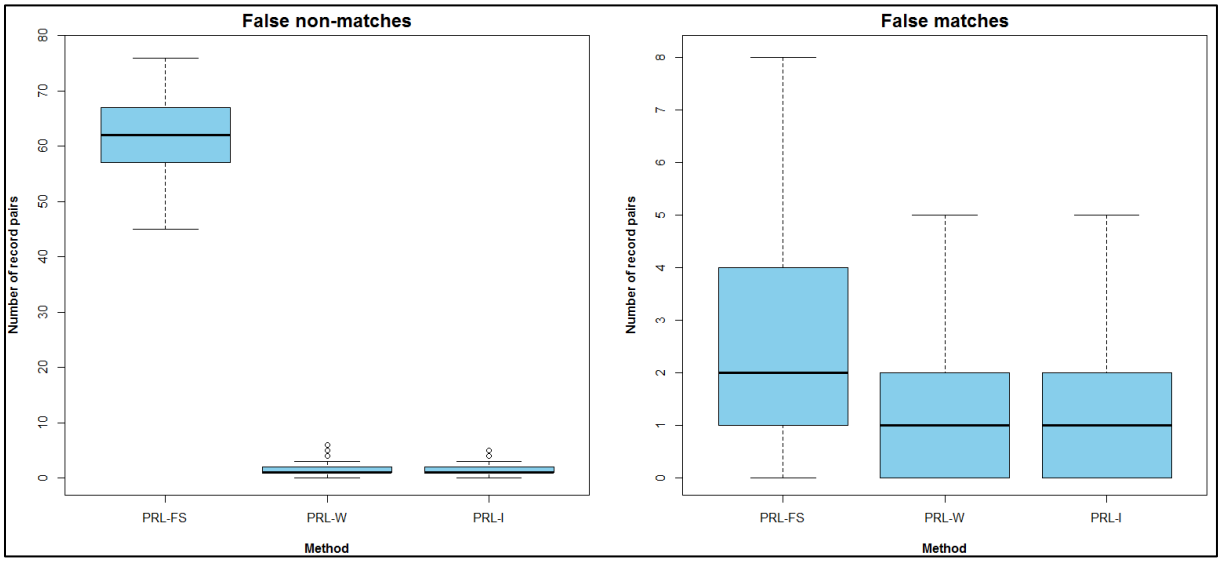


Figure 3.3 Distributions de nombres de faux négatifs et de faux positifs par les méthodes PRL-FS, PRL-W et PRL-I dans 100 processus

Les *t*-tests appariés (avec des corrections par la technique de Holm) montrent que la différence entre les nombres de mauvaises décisions générées par le PRL-W et le PRL-I n'étaient pas significatives (Tableau 3.4).

Nombres de faux négatifs		
	PRL-FS	**PRL-W**
PRL-W	$<2\times10^{-16}$	-
PRL-I	$<2\times10^{-16}$	0.75

Nombres de faux positifs		
	PRL-FS	**PRL-W**
PRL-W	$<1.2\times10^{-14}$	-
PRL-I	$<3.4\times10^{-14}$	0.83

Tableau 3.4 Les comparaisons appariées des 3 méthodes PRL à l'égard des nombres de mauvaises décisions en utilisant les t-tests avec des corrections de Holm

De plus, on a comparé le PRL-W et le PRL-I en ce qui concerne les nombres de faux négatifs et de faux positifs dans chacun des 100 processus de chaînage. Face au même couple de jeux de données, on peut observer que le PRL-W et le PRL-I ont des performances similaires dans les décisions de chaînage (Figure 3.4).

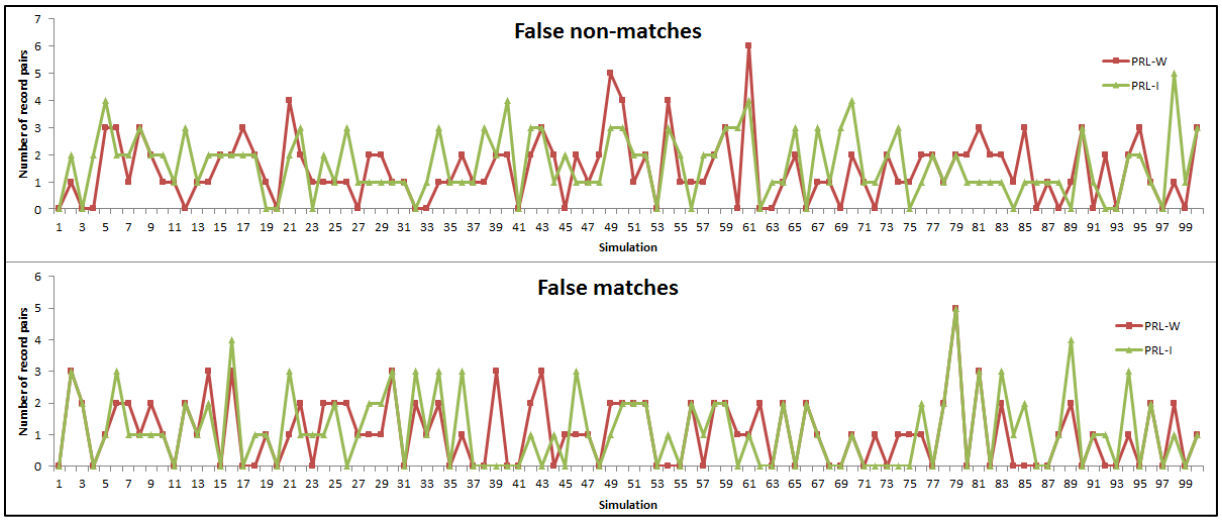


Figure 3.4 Nombres de faux négatifs et de faux positifs par les méthodes PRL-FS, PRL-W et PRL-I dans 100 processus

3.2.4.2. Comparaison des temps de calcul

En utilisant R (version 2.1.51) [33] sur une station de travail avec un CPU de 2,0 GHz (Intel (R) Xeon (R) E5-2620) et un RAM de 16 Go, les temps de calcul de PRL-FS, PRL-W et PRL-I étaient 123,74, 163,18 et 128,11 secondes en moyenne sur 100 processus, respectivement. On peut observer que les méthodes PRL-FS et PRL-I avaient besoin de moins de temps de calcul par rapport à la méthode PRL-W (Figure 3.5).

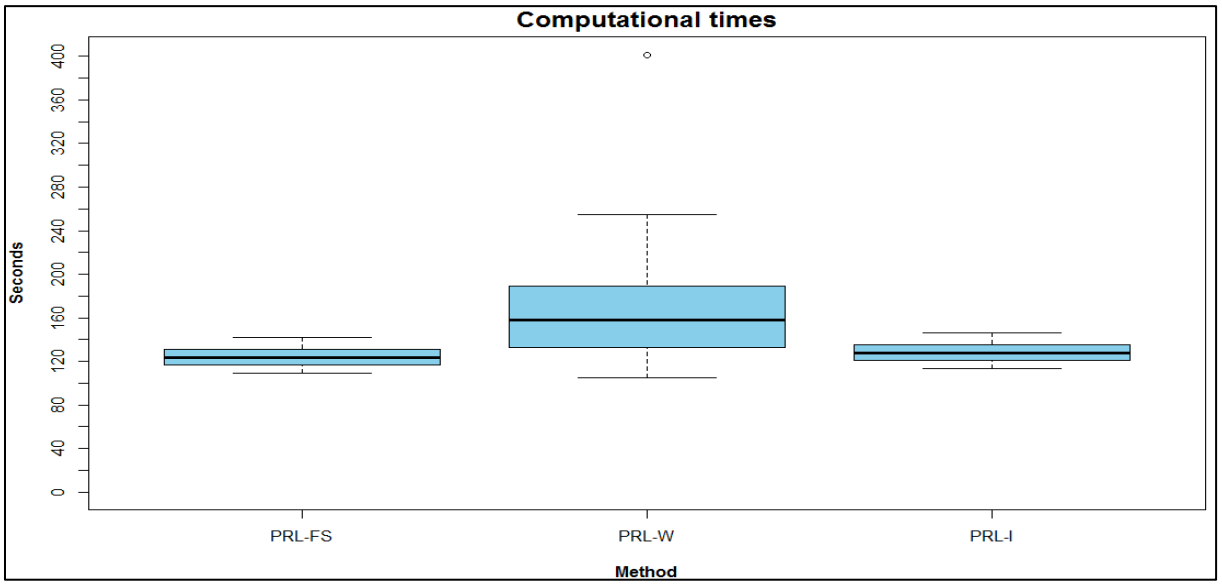


Figure 3.5 Distribution des temps de calcul des trois méthodes PRL dans 100 processus

En comparant les temps de calcul de ces trois méthodes PRL (Tableau 3.5), la réduction du temps de calcul pour PRL-I comparée avec le PRL-W est statistiquement significative.

Temps de calcul		
	PRL-FS	**PRL-W**
PRL-W	$<2\times10^{-16}$	-
PRL-I	0.23	$<2\times10^{-16}$

Tableau 3.5 Les comparaisons appariées des 3 méthodes PRL à l'égard du temps de calcul en utilisant les t-tests avec des corrections de Holm

Comme le montre la Figure 3.6, on peut observer que les temps de calcul du PRL -W étaient beaucoup plus sensibles que ceux de la méthode PRL-I. En comparant les variations de temps de calcul entre les séries appariées de PRL-W et de PRL-I (par le test de corrélation de Pearson entre la somme et la différence des temps de calcul), la variabilité du PRL-W s'est avérée être significativement plus élevée que PRL-I (le coefficient de corrélation des moments de Pearson est de 0,9269 et la p-value $<2,2 \times 10^{-16}$). Les variations des temps de calcul de PRL-I et de PRL-FS sont presque les mêmes car leurs différences résident dans le calcul supplémentaire du JWSS dans le PRL-I.

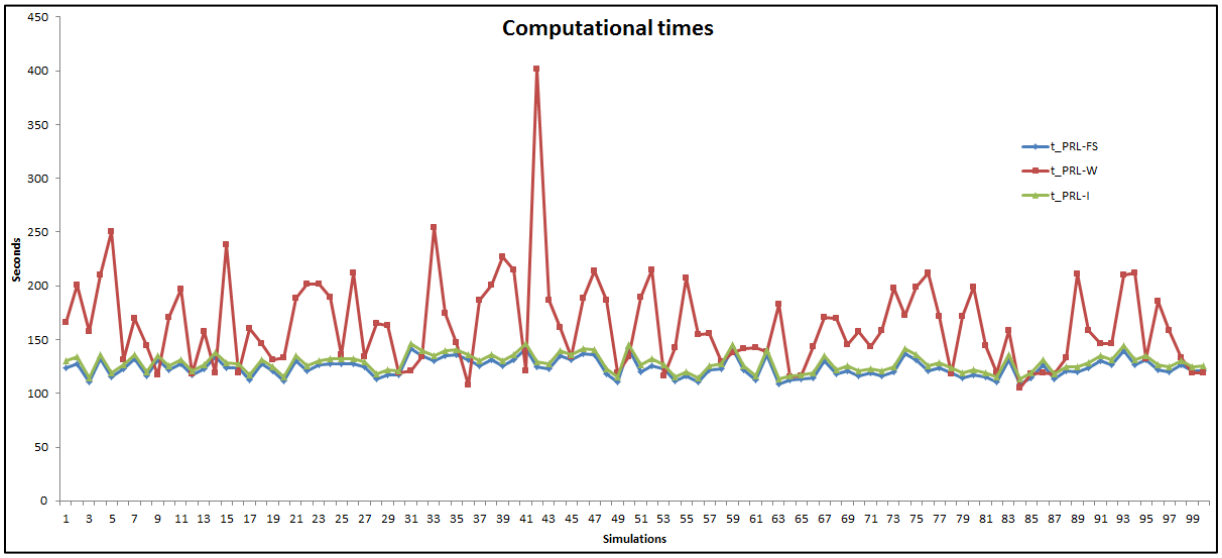


Figure 3.6 Temps de calcul des trois méthodes PRL dans 100 processus

3.2.5. Discussion

Ce travail a proposé une méthode alternative au PRL-W pour simplifier l'estimation des paramètres et pour réduire le temps de calcul. Le PRL-I a un pouvoir discriminant similaire au PRL-W, et le PRL-I et nécessite moins de temps de calcul.

On a évalué cette nouvelle méthode et on l'a comparée avec les méthodes PRL-FS et PRL-W. Afin d'évaluer correctement la valeur informationnelle des poids des couples d'enregistrements, on a utilisé les paramètres estimés pour calculer les poids de ces couples d'enregistrements et on a utilisé le seuil de décision optimal choisi par une ROC-analyse (seuil défini en utilisant la vérité de correspondances des enregistrements). Les nombres de mauvaises décisions ont été utilisés comme indicateur de performance parce qu'ils reflètent les pouvoirs discriminants des poids de couples d'enregistrements calculés par chaque méthode PRL.

Dans notre étude, chaque jeux de données contient 1 000 enregistrements, l'amélioration des temps de calcul peut ne pas sembler énorme (128,11 et 163,18 secondes en moyenne pour les méthodes

PRL-I et PRL-W, respectivement), mais le nombre de couples d'enregistrements augmente de façon quadratique avec la taille des jeux de données [48], et aussi le temps de calcul est proportionnel au nombre de couples d'enregistrements.

Avec une telle relation entre la taille des jeux de données et le temps de calcul pour un processus de chaînage, les temps nécessaires entre le PRL-I et le PRL-W pour chaîner deux jeux de données dont chacun contient 100 000 enregistrements, aurait pu être environ une centaine d'heures (à l'avantage du PRL-I).

Ce travail repose sur l'hypothèse que plus les contenus du champ sont similaires, plus le poids de champ est élevé. Cette approche permet d'attribuer des poids pertinents pour la plupart des couples d'enregistrements. Mais dans la pratique, on peut rencontrer des erreurs d'épellation comme la substitution de caractères d'une chaîne, par exemple, "*Christine*" devient "*Christina*" ou "*Kristine*". Au couple ("*Christine*", "*Christina*") est attribué un poids du champ supérieur au couple ("*Christine*", "*Kristine*"), avec leur JWSS de respectivement 0,9556 et 0,8843. Cependant, on ne peut pas conclure que le premier couple est plus susceptible de correspondre à la même personne que le second couple. La poursuite des travaux devrait évaluer la relation entre la propension de couples d'enregistrements à être la même personne et les JWSS de leurs champs, ce qui pourrait ne pas toujours être monotone croissante au sein du sous-intervalle de JWSS [0.6, 1[.

En outre, on peut également mettre en œuvre cette méthode en utilisant différents algorithmes de mesure de la similarité [21] au lieu d'utiliser le JWSS.

3.2.6. Conclusion

Dans cette étude, on a proposé une méthode alternative au PRL-W pour le calcul du poids des champs. Notre méthode est fondée sur l'interpolation des poids des champs estimés par une méthode de chaînage d'enregistrements couramment utilisé (PRL-FS). On a démontré que notre méthode peut atteindre une exactitude de décision de chaînages très similaire que le PRL-W en consommant moins de temps de calcul.

3.3. Proposition d'une méthode de chaînage par la combinaison linéaire des scores de similarité de chaque champ

3.3.1. Introduction

Comme présenté précédemment, le PRL-W est une des méthodes les plus performantes pour le chaînage probabiliste d'enregistrements contenant des erreurs typographiques courantes. Mais, l'utilisation du PRL-W au lieu de PRL-FS augmente le nombre de paramètres à estimer. En utilisant la largeur des sous-intervalles choisie par Winkler [9], l'implémentation du PRL-W nécessite 42 paramètres à estimer pour chacun des champs nom et prénom, contrairement au PRL-FS où on a seulement 2 paramètres à estimer par champ. Cette augmentation du nombre de paramètres à estimer complique non seulement la mise en œuvre du processus de chaînage, mais augmente aussi le temps de calcul.

Par conséquent, on se propose de développer une méthode de chaînage d'enregistrements où le poids du couple d'enregistrements est calculé par une combinaison linéaire des scores de similarité (JWSS) de chaque champ au sein de ce couple d'enregistrements. Pour les coefficients de cette combinaison linéaire, on choisit empiriquement les poids de concordance des champs utilisés dans la méthode PRL-FS, qui reflètent la contribution de chaque champ dans le calcul du poids du couple d'enregistrements (Pour rappel, plus le poids du couple d'enregistrements est élevé, plus il est vraisemblable qu'il s'agisse de la même personne). On émet l'hypothèse que la méthode de chaînage d'enregistrements utilisant les scores de similarité des champs linéairement combinés (RL-CS) est plus performant que le PRL-FS, et qu'elle peut atteindre une performance similaire du PRL-W.

En utilisant des jeux de données synthétiques, on a implémenté les méthodes RL-CS, PRL-FS et PRL-W. Puis, on a évalué et comparé ces trois méthodes en termes de mauvaises décisions (faux positif : chaînage de deux enregistrements concernant les personnes différentes ; faux négatif : absence de chaînage de deux enregistrements concernant la même personne) de chaînage et de temps de calcul.

3.3.2. Matériel et méthodes

3.3.2.1. Jeux de données

Comme dans les autres études, les couples de jeux de données synthétiques sont créés par notre algorithme de genèse des données synthétiques (2.1). Chaque jeu de données créé contient 1 000 enregistrements, et chaque enregistrement contient des champs nom, prénom, sexe, date de naissance et d'une clé d'identification unique. Les erreurs typographiques sont présentes dans 10%, 10%, 1% et 10% des champs nom, prénom, sexe et date de naissance, respectivement. Ce processus de genèse des jeux de données a été répété 100 fois.

3.3.2.2. Calcul des scores de similarité des champs

Pour quantifier la similarité entre les contenus du champ, le JWSS est utilisé. Dans cette étude, les JWSS ont été calculés en utilisant la fonction « *jarowinkler* » du package de R « *RecordLinkage* » [45]. Les JWSS de chaque champ seront linéairement combinés afin de calculer les poids des couples d'enregistrements dans la méthode RL-CS. Pour les coefficients de cette combinaison linéaire, on a proposé d'utiliser les poids de concordance des champs utilisés dans la méthode PRL-FS, qui seront présentés dans la section suivante.

3.3.2.3. Poids de concordance des champs dans le PRL-FS

Les poids de concordance des champs utilisés dans la méthode PRL-FS représentent l'importance des informations apportées par chaque champ. Par exemple, deux enregistrements dont le champ *nom* est concordant sont beaucoup plus susceptibles de concerner la même personne que deux enregistrements dont le champ *sexe* est concordant. Par conséquent, le poids de concordance du champ *nom* devrait être beaucoup plus élevé que le poids de concordance du champ *sexe*.

Ces poids sont définis comme les rapports de log vraisemblance fondés sur les paramètres m_i et u_i, où m_i est la probabilité que les contenus pour le champ i sont identiques sachant que le couple d'enregistrements concerne la même personne, et u_i est la probabilité que les contenus pour le champ i sont identiques sachant que le couple d'enregistrements concerne les personnes différentes [39]. Pour un couple d'enregistrements, si les contenus pour le champ i sont identiques, alors le poids de ce champ est :

$$w_i = log_2(m_i/u_i) \tag{3.13}$$

L'estimation des probabilités m_i et u_i peut être effectuée avec l'algorithme espérance-maximisation (EM) [41].

3.3.2.4. Calcul des poids du couple d'enregistrements

On a proposé de calculer le poids du couple d'enregistrements j comme suit [49] :

$$w^j = \sum_{i=1}^{n} w_i \times JWSS_i^j \tag{3.14}$$

Dans ce calcul du poids, l'importance des informations fournies par chaque champ (w^j) est constante, et est déterminée par (3.13). Pour le couple d'enregistrements j, le $JWSS_i^j$ est le JWSS de son champ i. Le poids du couple d'enregistrements j représente une combinaison linéaire des $JWSS_i^j$.

3.3.2.5. Décision de chaînages des enregistrements

Comme les poids calculés par les autres méthodes de chaînage d'enregistrements, plus le poids du couple d'enregistrements est élevé, plus il est vraisemblable qu'il s'agisse de la même personne. Prenons le paramètre p comme la proportion de couples d'enregistrements concernant les mêmes personnes parmi tous les couples d'enregistrements possibles entre deux jeux de données, pour que deux jeux de données à chaîner contiennent respectivement N_A et N_B enregistrements, il est raisonnable de prendre la décision de chaînages pour les $p \times N_A \times N_B$ premiers couples d'enregistrements étant classés en ordre décroissant selon leurs poids. Ce paramètre p peut être estimé par l'algorithme EM en même temps que l'estimation des paramètres m_i et u_i.

3.3.2.6. Evaluation et comparaison des méthodes PRL

En utilisant notre algorithme de genèse des données synthétiques, 100 couples de jeux de données étaient créés. Avec chaque couple de jeux de données, on a effectué un processus de chaînage d'enregistrements en utilisant respectivement les méthodes PRL-FS, PRL-W et RL-CS. La comparaison de ces trois méthodes étaient fondées sur le nombre de faux positifs et de faux négatifs, ainsi que sur les temps de calcul.

3.3.3. Résultats

La Figure 3.7 montre les nombres de mauvaises décisions sur 10^6 couples d'enregistrements menées par les méthodes PRL-FS, PRL-W et RL-CS dans 100 processus. On peut observer que le RL-CS et le PRL-W avaient une performance relativement similaire en termes des décisions de chaînage, et tous les deux sont plus performants que le PRL-FS. L'utilisation des méthodes PRL-FS, PRL-W et RL-CS ont respectivement généré 66,84, 4 et 12,51 mauvaises décisions en moyenne sur 100 processus de chaînages.

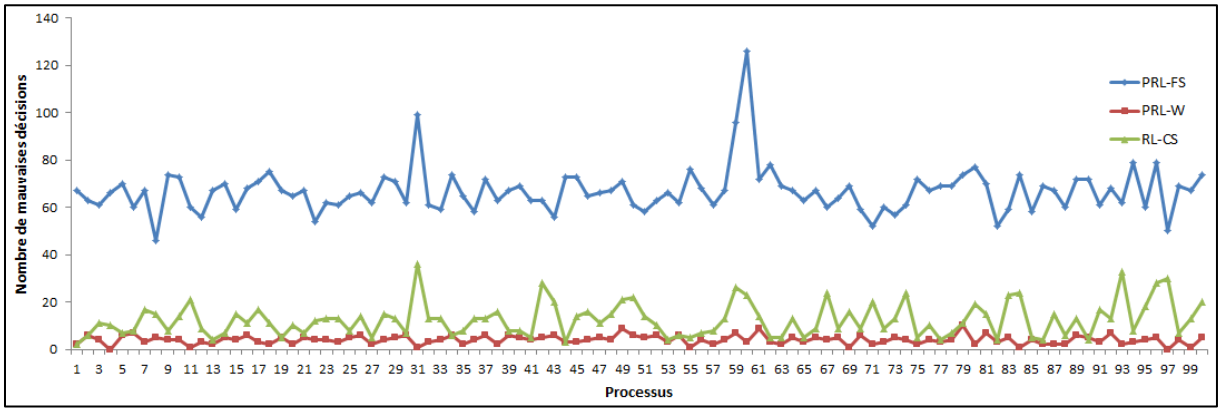


Figure 3.7 Nombres de mauvaises décisions par les méthodes PRL-FS, PRL-W et RL-CS dans 100 processus

De plus, on a comparé les temps de calcul de ces trois méthodes (Figure 3.8). On a utilisé une station de travail avec un CPU de 2,0 GHz (Intel (R) Xeon (R) E5-2620) et un RAM de 16 Go, les temps de calcul de PRL-FS, PRL-W et PRL-I étaient 123,74, 163,18 et 126,72 secondes en moyenne sur 100 processus, respectivement.

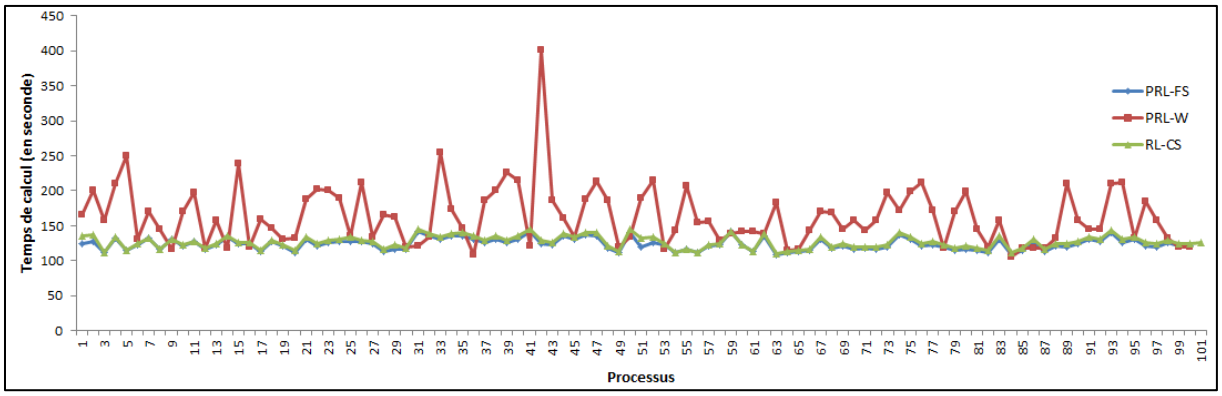


Figure 3.8 Temps de calcul des méthodes PRL-FS, PRL-W et RL-CS dans 100 processus

Parmi ces trois méthodes, le PL-CS est la méthode la plus efficace en termes de la considération globale de validité et de rapidité des chaînages.

3.3.4. Discussion et Conclusion

En utilisant des jeux de données synthétiques, on a démontré que le RL-CS est meilleur que le PRL-FS et possède une performance comparable au PRL-W en termes de décision de chaînages. On a donc préféré le RL-CS en raison de la simplicité de sa mise en œuvre (beaucoup moins de paramètres à estimer), ainsi que des temps de calcul réduits. Ce RL-CS garde les avantages du PRL-W en ce qui

concerne la prise en compte des informations fournies par chaque champ et de la similarité des contenus dans chaque champ à comparer.

Dans notre étude, les paramètres (m , u et p) requis par chaque méthode de chaînage d'enregistrements ont été obtenus par estimation. Les performances de ces méthodes pourraient dépendre de la performance de l'estimateur. On a donc comparé les résultats des chaînages issus des processus utilisant respectivement des paramètres estimés et observés. Dans chaque processus, l'utilisation des paramètres m et u estimés et observés (pour calculer les poids des couples d'enregistrements) conduit aux mêmes résultats de chaînages ; l'utilisation de p observé (pour choisir le seuil de décision) au lieu de p estimé a conduit –dans la plupart des cas– à un meilleur résultat de chaînages.

Cette étude illustre l'intérêt de la combinaison linéaire des JWSS des champs pour créer un poids du couple d'enregistrements, et démontre la meilleure performance de la méthode RL-CS en utilisant des jeux de données synthétiques. La genèse des données est effectuée en utilisant les types d'erreurs typographiques présentés dans la littérature, mais ces types d'erreurs restent loin d'être exhaustifs. Par exemple, l'inversion entre le nom et le prénom ou entre le nom de jeune fille et le nom marital pourrait être des erreurs possibles que l'on devrait pouvoir intégrer dans nos futurs travaux. En outre, comme on ne peut prévoir tous les scénarios possibles dans un travail réel de chaînage d'enregistrements, une méthode d'échantillonnage définissant les types et les taux d'erreurs dans les bases de données réelles serait très souhaitable afin d'adapter et d'améliorer le RL-CS dans un contexte spécifique.

Conclusion générale

Dans cette thèse, nous avons donc étudié différentes méthodes de chaînage de données/ d'enregistrements, *i.e.* des méthodes permettant de chaîner les informations d'une même personne dispersées dans différentes bases de données. Ce travail décrit l'implémentation et l'évaluation des méthodes existantes ainsi que quelques contributions originales proposées et justifiées par la nécessité d'améliorations et d'adaptations de ces méthodes à des contextes et situations particulières et fréquentes. Les travaux effectués lors de cette thèse ont été initialement motivés dans la recherche d'une solution de chaînage efficace de données dans le cadre du projet GINSENG. Ce dernier visait à mettre en place une infrastructure de grille informatique dédiée au partage de données médicales. La nécessité d'approfondir l'étude de ces méthodes de chaînages tenait au fait qu'elles obligent à des adaptations spécifiques au contexte dans lequel on souhaite les appliquer. Néanmoins, le caractère universel du problème qu'abordent ces méthodes, nous a conduits à envisager leur utilisation dans d'autres contextes, comme celui de l'épidémiologie ou de la veille sanitaire. Par exemple, les spécificités de l'identitovigilance, qui surveille et gère les risques et erreurs liés à l'identification des patients, nous ont amenés à proposer et évaluer des améliorations de ces méthodes pour une utilisation en routine.

Ce chaînage d'enregistrements ne pose bien sûr pas le moindre problème lorsqu'il s'adresse à des bases de données dont l'interopérabilité a été planifiée. Cette tâche de chaînage ne constitue un défi qu'en l'absence de clé d'identification commune parmi les différentes bases de données à chaîner, ce qui est souvent le cas dans le contexte du projet GINSENG et ce qui quasiment la règle lorsque l'on cherche à exploiter ou réutiliser des données observationnelles disponibles et dispersées. Pour réaliser un chaînage dans de telles conditions, la comparaison de champs d'enregistrement contenant les traits ou informations d'identifications devient incontournable. La plupart des systèmes de gestion de base de données (SGBD) permet d'effectuer ce type de tâche mais de manière déterministe, *i.e.* avec un chaînage fondé sur la concordance exacte de tous les champs comparés. Dans la pratique, la qualité de données des champs à comparer n'est cependant pas toujours parfaite. Ce manuscrit présente donc tout d'abord un état de l'art des techniques couramment utilisées pour la comparaison approximative des champs d'enregistrements, ainsi que les méthodes probabilistes permettant de chaîner efficacement les enregistrements contenant des erreurs typographiques courantes.

Dans la deuxième partie, nous avons souhaité présenté dans le détail l'implémentation de la méthode de chaînage PRL-FS, et ce pour deux raisons essentielles. Bien que cette méthode soit disponible dans de plusieurs logiciels, dont certains sont libres [7], le détail de son implémentation n'est que très rarement fourni. De plus, l'utilisation de ces logiciels peut être limitée par la dépendance à un système d'exploitation ou à d'autres logiciels, à des contraintes de taille limite de fichiers à traiter, et surtout et enfin à l'impossibilité d'une intégration facile à un SGBD. La présentation du PRL-FS est également très utile pour introduire la méthode de chaînage PRL-W, proposée par William Winkler, et qui dérive de cette méthode en y ajoutant la prise en compte du degré de similarité entre les contenus des champs à comparer. Le gain de performance du PRL-W est clairement démontrée [7,9,25], elle est sans aucun doute la méthode de chaînage la plus performante à ce jour. Cependant, cette méthode est rarement appliquée depuis sa publication en

1990, la plupart des études récentes s'appuyant sur le PRL-FS [36,46,50,51], et son implémentation détaillée est tout bonnement absente de la littérature à ce jour. Bien qu'il soit difficile d'attribuer cette absence à la complexité de sa mise en œuvre pratique sur données réelles, et également du fait qu'elle est d'un intérêt tout particulier dans notre travail, nous avons souhaité dédier une section entière à la méthode de chaînage PRL-W et aux détails précis de son implémentation. Notons enfin que notre proposition d'évaluation des méthodes probabilistes de chaînage de données (PRL-FS et PRL-W) s'entend en termes de validité, *i.e.* de conformité à la réalité des décisions issue des PRL. Il a donc été nécessaire de disposer de jeux de données où la vérité du chaînage soit connue, nous avons opté pour des données synthétiques et développé un algorithme paramétrable de genèse de ces données.

La troisième partie répond à divers questionnements par des propositions d'améliorations ou d'adaptations originales de méthodes probabilistes de chaînage d'enregistrements. Une première contribution consiste à proposer des solutions permettant d'améliorer la prise en compte des données manquantes lors de la mise en œuvre du PRL-W. Ces solutions sont l'objet d'une évaluation veillant notamment à ce que ces améliorations ne fassent pas aux dépens du temps de calcul. Une seconde contribution consiste à explorer des voies d'amélioration de la valeur informationnelle des poids de chaînages, i.e. maximisant la probabilité que pour deux couples d'enregistrements choisis aléatoirement, le poids de celui qui implique une seule et même personne soit supérieur à celui impliquant deux personnes authentiquement différentes. Parmi les approches proposées, l'une dérive du PRL-W, l'autre est une méthode simple de mise en œuvre et totalement originale fondée sur la combinaison linéaire des scores de similarité entre chaînes de caractères. Enfin, une troisième contribution focalise sur la décision de chaînage en elle-même. Une approche par seuillage fondée sur un des paramètres estimés par l'algorithme EM permet un choix optimal quant aux erreurs de chaînages et constitue selon nous une approche satisfaisante pour l'évaluation en épidémiologie par exemple.

En revanche, ce questionnement ouvre une des grandes perspectives de notre travail au sens où, pour une utilisation en production de soins, la sécurité de la prise en charge des patients obligent la mise en œuvre d'une approche par double seuillage. La constitution d'une liste critique à vérifier, en dehors de la liste des couples à chaîner ou à séparer de façon sûre, est la solution de choix, mais elle nécessite encore une évaluation intensive et systématique avant d'entériner son intégration en routine, *e.g.* en identitovigilance. D'autres perspectives sont envisagées sur des voies telles que l'amélioration de la valeur informationnelle des poids de chaînage en affranchissant ces méthodes de l'hypothèse d'indépendance conditionnelle ou en intégrant la distribution des occurrences d'erreurs typologiques. La voie de l'amélioration des temps de calculs est aussi envisagée, en optimisant les divers algorithmes impliqués dans ces méthodes PRL (réduction du nombre d'itérations de l'algorithme EM, partition des tâches d'estimation, hybridation DRL-PRL). Enfin, la portée de nos travaux tient essentiellement à notre capacité à proposer l'intégration opérationnelle de ces méthodes et de leurs adaptations aux systèmes auxquels elles pourraient être couplées, comme par exemple un système d'information hospitalier.

Bibliographie

1 Safran C, Bloomrosen M, Hammond We, *et al.* Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper. *J Am Med Inform Assoc* 2007;**14**:1–9. doi:10.1197/jamia.M2273

2 Haggerty JL, Reid RJ, Freeman GK, *et al.* Continuity of care: a multidisciplinary review. *BMJ* 2003;**327**:1219–21. doi:10.1136/bmj.327.7425.1219

3 Roos LL, Wajda A. Record linkage strategies. Part I: Estimating information and evaluating approaches. *Methods Inf Med* 1991;**30**:117–23.

4 Jaro MA. Probabilistic linkage of large public health data files. *Stat Med* 1995;**14**:491–8. doi:10.1002/sim.4780140510

5 Newcombe HB, Kennedy JM, Axford SJ, *et al.* Automatic Linkage of Vital Records Computers can be used to extract 'follow-up' statistics of families from files of routine records. *Science* 1959;**130**:954–9. doi:10.1126/science.130.3381.954

6 Fellegi IP, Sunter AB. A theory for record linkage. *J Am Stat Assoc* 1969;**64**:1183–210.

7 Christen P. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer 2012.

8 Friedman C, Sideli R. Tolerating spelling errors during patient validation. *Comput Biomed Res* 1992;**25**:486–509. doi:10.1016/0010-4809(92)90005-U

9 Winkler WE. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. Published Online First: 1990.http://www.eric.ed.gov/ERICWebPortal/recordDetail?accno=ED325505 (accessed 6 Dec2013).

10 Forster M, Bailey C, Brinkhof MWG, *et al.* Electronic medical record systems, data quality and loss to follow-up: survey of antiretroviral therapy programmes in resource-limited settings. *Bull World Health Organ* 2008;**86**:939–47.

11 Tromp M, Ravelli AC, Bonsel GJ, *et al.* Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage. *J Clin Epidemiol* 2011;**64**:565–72. doi:10.1016/j.jclinepi.2010.05.008

12 Winkler WE. Approximate string comparator search strategies for very large administrative lists. *Statistics* 2005;:02.

13 Cohen W, Ravikumar P, Fienberg S. A comparison of string metrics for matching names and records. In: *KDD Workshop on Data Cleaning and Object Consolidation*. 2003. 73–8.https://www.cs.cmu.edu/afs/cs/Web/People/wcohen/postscript/kdd-2003-match-ws.pdf (accessed 30 Jul2014).

14 Jaro MA. *Unimatch: A record linkage system: Users manual*. Bureau of the Census 1978. http://books.google.fr/books?hl=fr&lr=&id=was9AAAAIAAJ&oi=fnd&pg=PA7&dq=unimatch&ots=p2LRwh98uN&sig=weWe0mi3zli70EYCS-JUyxlf37A (accessed 6 Dec2013).

15 Jaro MA. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *J Am Stat Assoc* 1989;**84**:414–20. doi:10.1080/01621459.1989.10478785

16 Winkler WE. Overview of record linkage and current research directions. BUREAU OF THE CENSUS 2006.

17 Navarro G. A Guided Tour to Approximate String Matching. *ACM Comput Surv* 2001;**33**:31–88. doi:10.1145/375360.375365

18 Jokinen P, Tarhio J, Ukkonen E. A Comparison of Approximate String Matching Algorithms. *Softw Pr Exper* 1996;**26**:1439–58. doi:10.1002/(SICI)1097-024X(199612)26:12<1439::AID-SPE71>3.0.CO;2-1

19 Russell RC. *The soundex coding system*. US Patent 1,261,167 1918.

20 Li B, Quan H, Fong A, *et al.* Assessing record linkage between health care and Vital Statistics databases using deterministic methods. *BMC Health Serv Res* 2006;**6**:48. doi:10.1186/1472-6963-6-48

21 Cohen WW, Ravikumar PD, Fienberg SE. A Comparison of String Distance Metrics for Name-Matching Tasks. In: *IIWeb*. 2003. 73–8.http://dc-pubs.dbs.uni-leipzig.de/files/Cohen2003Acomparisonofstringdistance.pdf (accessed 7 Dec2013).

22 Quantin C, Binquet C, Bourquard K, *et al.* [Assessment of the discriminating power of identifiers for record linkage]. *Rev Dépidémiologie Santé Publique* 2004;**52**:431–40.

23 Samuels C. Using the EM Algorithm to Estimate the Parameters of the Fellegi-Sunter Model for Data Linking. 2012.

24 Winkler WE. Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage. In: *Proceedings of the Section on Survey Research Methods, American Statistical Association*. 2000. 667–71.

25 Li X, Guttmann A, Cipiere S, *et al.* Implementation of an extended Fellegi-Sunter probabilistic record linkage method using the Jaro-Winkler string comparator. In: *2014 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. 2014. 375–9. doi:10.1109/BHI.2014.6864381

26 DuVall SL, Kerber RA, Thomas A. Extending the Fellegi-Sunter probabilistic record linkage method for approximate field comparators. *J Biomed Inform* 2010;**43**:24–30. doi:10.1016/j.jbi.2009.08.004

27 Li X, Guttmann A, Cipière S, *et al.* Évaluation des algorithmes de rapprochement de patients par traits d'identification nominatifs, Clermont-Ferrand, France. *Rev DÉpidémiologie Santé Publique* 2013;**61**:S322.

28 Li X, Guttman A, Cipière S, *et al.* Comparaison de performance des algorithmes de rapprochement de patients. *Rev DÉpidémiologie Santé Publique* 2014;**62**:S76–7.

29 Li X, Guttmann A, Cipière S, *et al.* Utilisation de l'algorithme EM pour estimer les paramètres du chaînage probabiliste d'enregistrements. *Rev DÉpidémiologie Santé Publique* 2014;**62, Supplement 5**:S196. doi:10.1016/j.respe.2014.06.081

30 Projet ANR | ANR - Agence Nationale de la Recherche. http://www.agence-nationale-recherche.fr/projet-anr/?tx_lwmsuivibilan_pi2%5BCODE%5D=ANR-10-TECS-0008 (accessed 28 Oct2014).

31 Christen P, Vatsalan D. Flexible and extensible generation and corruption of personal data. In: *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM 2013. 1165–8.http://dl.acm.org/citation.cfm?id=2507815 (accessed 6 Nov2014).

32 Friedman C, Sideli R. Tolerating spelling errors during patient validation. *Comput Biomed Res* 1992;**25**:486–509. doi:10.1016/0010-4809(92)90005-U

33 R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: : R Foundation for Statistical Computing 2014. http://www.R-project.org

34 Talburt JR, Zhou Y, Shivaiah SY. SOG: A Synthetic Occupancy Generator to Support Entity Resolution Instruction and Research. *ICIQ* 2009;**9**:91–105.

35 Zhu VJ, Overhage MJ, Egg J, *et al.* An empiric modification to the probabilistic record linkage algorithm using frequency-based weight scaling. *J Am Med Inform Assoc JAMIA* 2009;**16**:738–45. doi:10.1197/jamia.M3186

36 Lebreton E, Vincelet C, Chatignoux E, *et al.* Chaînage d'enregistrements de séjours PMSI aux premiers certificats de santé : un test dans le Val d'Oise. *Rev DÉpidémiologie Santé Publique* 2014;**62**:257–66. doi:10.1016/j.respe.2014.04.006

37 Grannis SJ, Overhage JM, Hui S, *et al.* Analysis of a Probabilistic Record Linkage Technique without Human Review. *AMIA Annu Symp Proc* 2003;**2003**:259.

38 Méray N, Reitsma JB, Ravelli ACJ, *et al.* Probabilistic record linkage is a valid and transparent tool to combine databases without a patient identification number. *J Clin Epidemiol* 2007;**60**:883.e1–883.e11. doi:10.1016/j.jclinepi.2006.11.021

39 Blakely T, Salmond C. Probabilistic record linkage and a method to calculate the positive predictive value. *Int J Epidemiol* 2002;**31**:1246–52.

40 Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer 2009.

41 Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Ser B Methodol* 1977;:1–38.

42 Moon TK. The expectation-maximization algorithm. *Signal Process Mag IEEE* 1996;**13**:47–60.

43 Do CB, Batzoglou S. What is the expectation maximization algorithm? *Nat Biotechnol* 2008;**26**:897–9. doi:10.1038/nbt1406

44 Porter EH, Winkler WE. Approximate string comparison and its effect on an advanced record linkage system. In: *Advanced Record Linkage System. US Bureau of the Census, Research Report*. 1997. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.46.7347 (accessed 11 Dec2013).

45 Sariyar M, Borg A. The RecordLinkage Package: Detecting Errors in Data. http://journal.r-project.org/archive/2010-2/RJournal_2010-2_Sariyar+Borg.pdf (accessed 7 Dec2013).

46 Ong TC, Mannino MV, Schilling LM, *et al.* Improving record linkage performance in the presence of missing linkage data. *J Biomed Inform* doi:10.1016/j.jbi.2014.01.016

47 Herzog TN, Scheuren FJ, Winkler WE. *Data Quality and Record Linkage Techniques*. Springer 2007.

48 Bilenko M, Kamath B, Mooney RJ. Adaptive Blocking: Learning to Scale Up Record Linkage. In: *Sixth International Conference on Data Mining, 2006. ICDM '06*. 2006. 87–96. doi:10.1109/ICDM.2006.13

49 Li X, Guttmann A, Demongeot J, *et al.* An empiric weight computation for record linkage using linearly combined fields' similarity scores. In: *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2014. 1346–9. doi:10.1109/EMBC.2014.6943848

50 Shlomo N. Probabilistic Record Linkage for Disclosure Risk Assessment. In: Domingo-Ferrer J, ed. *Privacy in Statistical Databases*. Springer International Publishing 2014. 269–82.http://link.springer.com.sicd.clermont-universite.fr/chapter/10.1007/978-3-319-11257-2_21 (accessed 2 Dec2014).

51 Capuani L, Bierrenbach AL, Abreu F, *et al.* Accuracy of a probabilistic record-linkage methodology used to track blood donors in the Mortality Information System database. *Cad Saúde Pública* 2014;**30**:1623–32. doi:10.1590/0102-311X00024914

Liste des publications et des communications

2013 **Li, X**; Guttmann, A; Cipière, S; Maigne, L; Boire, J-Y; Ouchchane, L; ,"*Évaluation des algorithmes de rapprochement de patients par traits d'identification nominatifs*", Revue d'Épidémiologie et de Santé Publique,61,S322,2013,Elsevier

2014 **Li, X**; Guttman, A; Cipière, S; Maigne, L; Boire, J-Y; Ouchchane, L, "*Comparaison de performance des algorithmes de rapprochement de patients*", Revue d'Épidémiologie et de Santé Publique,62,,S76-S77,2014,Elsevier

2014 **Li, Xinran**; Guttmann, Aline; Cipiere, Sebastien; Maigne, Lydia; Demongeot, Jacques; Boire, Jean-Yves; Ouchchane, Lemlih, *"Implementation of an extended Fellegi-Sunter probabilistic record linkage method using the Jaro-Winkler string comparator"*, Biomedical and Health Informatics (BHI), 2014 IEEE-EMBS International Conference on",375-379,2014,IEEE

2014 **Li, X**; Guttmann, A; Cipière, S; Demongeot, J; Boire, J-Y; Ouchchane, L*,"Utilisation de l'algorithme EM pour estimer les paramètres du chaînage probabiliste d'enregistrements*", Revue d'Épidémiologie et de Santé Publique,62,,S196,2014,Elsevier Masson

2014 **Li, Xinran**; Guttmann, Aline; Demongeot, Jacques; Boire, Jean-Yves; Ouchchane, Lemlih, "An Empiric Weight Computation for Record Linkage Using Linearly Combined Fields' Similarity Scores", 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'14),1346-1349,2014,IEEE