Note méthodologique sur les appariements probabilistes

Ce document est issu de l'article [3] et cherche à se concentrer sur le modèle de Fellegi et Sunter [9] qui a inspiré l'ensemble des méthodes d'appariement et qui est le plus cité dans la littérature. Initialement prévu pour réaliser des appariements en routine et donc à vocation à produire de nouvelles bases de données, ce modèle est de plus en plus utilisé pour produire des appariements unique à des fins de recherche ou pour la réalisation d'étude statistique.

1 Les méthodes de couplage

Les méthodes de couplage indirect et plus particulièrement la méthode de Fellegi et Sunter sont utilisées lorsqu'il n'existe pas de moyen de retrouver de façon certaine des individus en utilisant un identifiant direct comme le NIR. L'usage de ce dernier étant réglementé pour son caractère discriminant, il est rare de le voir présent dans une base de données. Il est de même rare qu'un numéro de gestion utilisé dans une base apparaisse dans une autre. Sans identifiant direct, le couplage devra se faire sur des informations communes aux bases de données à apparier. Ainsi il faut d'abord détecter dans chacune des bases les variables indirectement identifiantes rendant possible la réalisation d'un couplage. On appellera identifiant indirect toute variable ou combinaison de variables qui permet de retrouver l'identité numérique d'un individu (dans une base de données), sans utiliser l'identité physique (dans le monde réel). Voici la liste des variables utilisées en pratique pour la ré-identification indirecte :

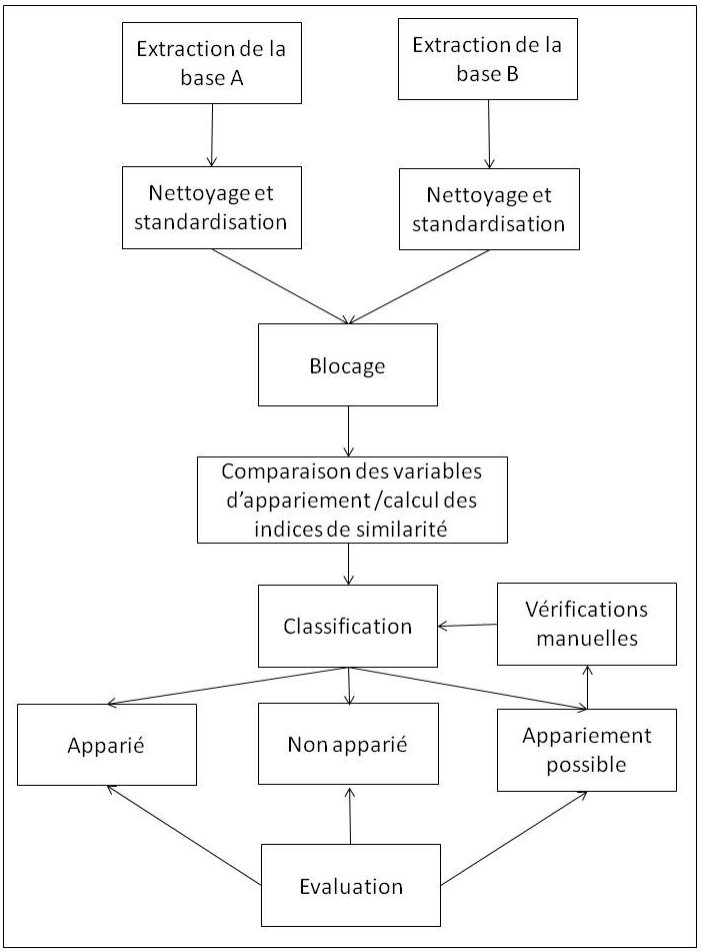
- Nom et prénom (en raison de leur non unicité, et des erreurs possibles d'écriture, ils ne sont pas considérés comme des identifiants certains et physiques des personnes)
- Sexe
- Date de naissance
- Date de décès
- Date de soin
- Adresse
- Département et commune de domicile
- Département et commune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

L'appariement est un processus long qui l'on peut synthétiser dans la figure 1. La standardisation et le nettoyage de données sont nécessaires pour rendre comparable, le plus possible, les bases à apparier sur les variables indirectement identifiantes. Le blocage est un moyen pour réduire le nombre de couple d'individu à évaluer, sans le blocage le nombre de couple à évaluer est de $n_{E_A} \times n_{E_B}$. Il existe plusieurs façon de comparer les variables de couplage, on peut utiliser pour cela des indices de similarité s'il on veut prendre en compte des erreurs typographiques par exemple [3]. On peut

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

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

FIGURE 1 – Schéma décrivant le processus complet d'un appariement



aussi transformer les variables en les simplifiant avant des de comparer les variables une à une. L'algorithme Soundex est un algorithme d'indexation phonétique qui peut être utiliser pour gommer ces erreurs typographiques. La classification consistera ici à appliquer la règle de décision de Fellegi et Sunter qui satisfaitun critère d'optimalité. Cette dernier étape peut être itérative en fonction de l'analyse des résultats obtenus.

1.1 Les modèles probabilistes

Newcombe et al. [16], en 1959, sont les premiers à avoir posé le problème des couplages de données en un problème d'inférence bayésienne. Les auteurs se posaient alors la question de la faisabilité de réaliser des couplages automatisés. En 1968, ce sont Fellegi et Sunter [9] qui formalisent mathématiquement l'approche de Newcombe [16] tout en démontrant une certaine forme d'optimalité dans la règle de classement. Cette formalisation se fait au travers d'une vraisemblance (modèle probabiliste) exprimant le processus par lequel les données ont été générées.

1.2 Une théorie générale du couplage selon Fellegi et Sunter

Soient deux bases de données notées E_A et E_B issues d'une population A et d'une population B respectivement, contenant des informations sur ces populations de type : nom, prénom, âge, date de naissance, date de décès, adresse, etc. Chaque entrée de la base concerne un unique individu de la population. L'objectif du couplage est de déterminer si les entrées dans chaque base se rapportent "probablement" à un même individu, compte tenu du fait que les informations des bases ont pu être altérées par la production d'erreurs ou que l'individu a pu changer de statut entre les deux entrées. Cette problématique se traduit d'un point de vue probabiliste en un problème de discrimination : pour chaque entrée de la base E_A on associe une entrée de la base E_B. L'ensemble des couples ainsi construit est noté $E_A \times E_B$ (le produit cartésien de E_A et E_B). Cet ensemble $E_A \times E_B$ est alors séparé en deux classes M et U qui sont respectivement l'ensemble des couples concernant un individu identique et l'ensemble des couples d'entrée concernant des individus différents. On notera dans la suite $\delta(a, b)$ le vecteur des concordances de la paire $(a, b) \in E_A \times E_B$, où $\delta_i(a, b) = 1$ si les champs i concordent et $\delta_i(a, b) = 0$ dans le cas contraire.

Tableau 2 – Un exemple de table de concordance

Vecteur δ					$E_A \times E_B$	
1	1	1	1	1	a_1	b_1
1	1	1	1	1	a_1	b_2
1	1	1	1	0	a_1	b_3
⋮	⋮	⋮	⋮	⋮	⋮	⋮
0	0	0	0	1	a_n	b_{m-1}
0	0	0	0	0	a_n	b_m

Cette classification se faisant en fonction des valeurs contenues dans les différents champs, la question du pouvoir informationnel (ou discriminant) des différents champs se pose. En effet le champ renseignant le sexe d'un individu contiendra moins d'information (au sens de Shannon) que le prénom ou encore l'adresse. Pour intégrer ce fait, [16] ont développé le concept du "poids" d'un champ basé sur la probabilité de voir les deux champs correspondant prendre la même valeur. Dans [9], les auteurs proposent la définition suivante : pour chaque champs i on définit les probabilités $m_i(\delta) = \mathbb{P}(\delta_i = 1|M)$ et $u_i(\delta) = \mathbb{P}(\delta_i = 1|U)$. Le poids w_i du champ i est calculé comme le logarithme du rapport entre les deux probabilités précédentes en fonction de la valeur de δ_i. Si $\delta_i(a, b) = 1$, alors $w_i(1) = \log\left(\frac{m_i}{u_i}\right)$, dans le cas contraire $w_i(0) = \log\left(\frac{1-m_i}{1-u_i}\right)$. Le poids global du couple noté w est la somme des poids des différents champs. Selon que ce poids global dépasse un seuil

donné, la décision est prise de classer le couple (a, b) comme M ou U. L'hypothèse sous-jacente est une hypothèse forte d'indépendance, conditionnellement à une variable latente, indiquant la concordance des champs. [21], [22] montrent que cette hypothèse est fausse en pratique et qu'elle réduit par ailleurs la qualité de la règle de décision. La log-vraisemblance du modèle de Fellegi-Sunter est donnée par :

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

On notera que cette vraisemblance fait l'hypothèse, généralement fausse, de l'indépendance entre les paires. Nous verrons dans la suite que cette hypothèse est responsable des appariements multivoques. La solution présentée dans [9] est une règle de décision plus subtile car elle propose de classer certains couples dans un troisième ensemble contenant les couples non classés. La règle de décision peut alors s'écrire :

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

Avec $w = \log\left(\dfrac{m(\delta)}{u(\delta)}\right)$. Cette règle de décision est optimale dans le sens où étant donné un pourcentage de faux positifs et un pourcentage de faux négatifs tolérés, la règle précédente minimise la taille de la troisième classe $\widetilde{C}$ (les seuils $T_{\widetilde{M}}$ et $T_{\widetilde{U}}$ sont des fonctions de ces pourcentages). Elle est basée sur un test UPP (Uniformément Plus Puissant) bien connu puisque c'est une ré-interprétation du test de Neyman et Pearson [17].

Pour calculer le seuil $T_{\widetilde{M}}$ il faudra additionner les probabilités u des scores les plus élevés vers les scores les plus faibles jusqu'à ce que la somme dépasse la probabilité de classer a tort un couple dans M. De même pour calculer le seuil $T_{\widetilde{U}}$, on additionnera les probabilités m jusqu'à ce que la somme dépasse la probabilité de classer a tort un couple dans U. Dans la table ci-dessus, si on tolère 0.01%, en additionnant les 4 premières probabilités u, on arrive à une probabilité de coupler à tort de 0.0095%. En additionnant la probabilité de la cinquième configuration on dépasse cette probabilité d'apparier à tort. Le score de cette cinquième configuration est le seuil $T_{\widetilde{M}}$.

1.2.1 Estimation des paramètres pour le modèle de Fellegi et Sunter

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

1.2.1.1 la méthode des fréquences d'erreur :

Cette méthode est inspirée par [16] qui propose d'automatiser le couplage en utilisant le pouvoir discriminant de l'information (mesuré par l'entropie de Shannon). Reprenons l'exemple de [9] sur

TABLEAU 3 – Calcul du seuil $T_{\widetilde{M}}$ de discrimination

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

TABLEAU 4 – Classification d'une table de concordance à l'issu du processus d'appariement décrit ci-dessus

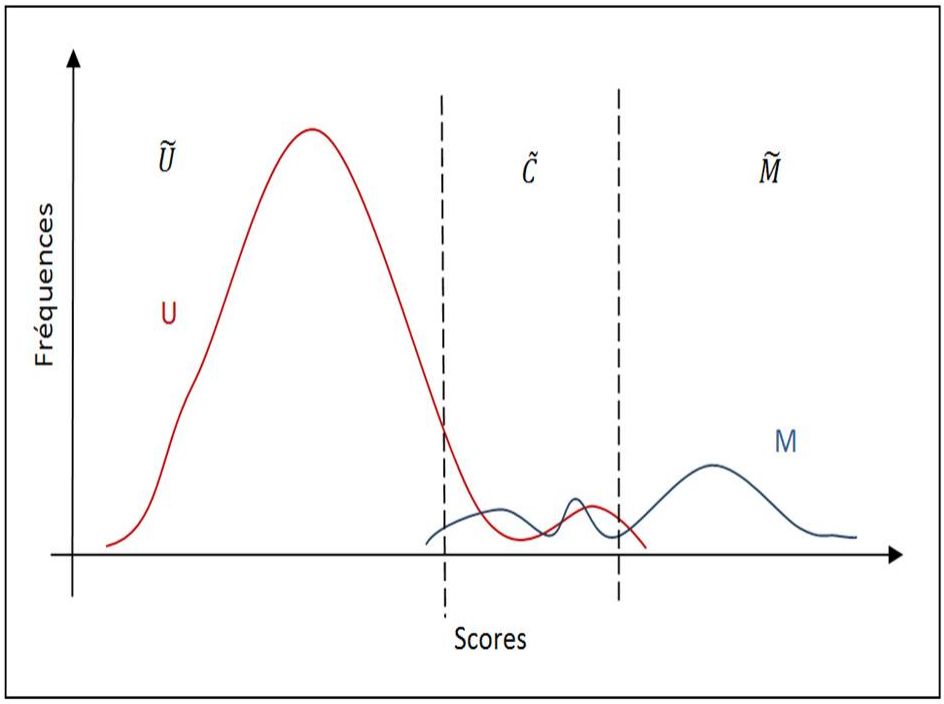
classification	Vecteur δ					M	U	Nb total
$\widetilde{M}$	1	1	1	1	1	511	4	515
	1	1	0	1	1	268	9	287
	1	1	1	0	1	115	10	125
	0	1	1	1	1	58	54	112
$\widetilde{C}$	1	0	1	1	1	49	10	59
	1	1	1	1	0	45	60	105
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\widetilde{U}$	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	0	0	0	0	1	3	1150	1153
	0	0	0	0	0	1	9579	9580

la variable Nom pour laquelle le nombre d'occurrence pour chaque modalité dans la population J ($J = E_A$ ou E_B) est donné par $(O_{J,t})_{1\leq t\leq n_J}$, où n_J est le nombre de modalité du nom pour la population J. Soient maintenant e_J la probabilité qu'un nom dans la population J soit mal reporté, e_{J0} la probabilité que le nom n'apparaisse pas dans la base générée à partir de J, et e_T la probabilité que le nom ait changé après son entrée de la base alors :

$$\begin{cases} m(\delta_{nom}=1) \approx 1 - e_{E_A} - e_{E_B} - e_T - e_{E_A0} - e_{E_B0} \\ u(\delta_{nom}=1) \approx (1 - e_{E_A} - e_{E_B} - e_T - e_{E_A0} - e_{E_B0}) \quad \times \sum_t \dfrac{O_{E_A,t} O_{E_B,t}}{N_{E_A} N_{E_B}} \end{cases} \tag{3}$$

Le calcul des probabilités de (3) a nécessité de faire des hypothèses supplémentaires ainsi que des approximations dans le calcul des probabilités m et u. On remarquera par ailleurs que les probabilités d'erreur, de non inclusion, ne dépendent pas de la modalité de la variable Nom. Cette

FIGURE 2 – Distribution des poids w à l'issu du processus d'appariement décrit ci-dessus



méthode demande des informations sur les populations (ici le nombre d'occurrence des noms) et la méthode de production des bases ainsi que sur les mécanismes de production des erreurs ou des divergences apparaissant dans les bases. Au final cette approche est utilisable dans une démarche supervisée, ou en tout cas nécessite des informations complémentaires obtenues par le biais de bases exhaustives. La méthode suivante est plus pertinente au regard de la situation dans laquelle sont réalisés les couplages des bases médicales et administratives.

1.2.1.2 Estimation non supervisée par l'algorithme d'espérance-maximisation ("EM") :

C'est Winkler [23] qui propose l'algorithme "EM" pour le modèle de Fellegi-Sunter. Jaro [13] le teste une première fois sur s données de recensement en 1989, les résultats présentés sont très prometteurs quant à l'utilisation de cet algorithme (dans un cadre orienté chaîne de caractères). L'algorithme "EM" a été développé pour la première fois dans [7], il permet de faire des estimations par maximum de vraisemblance dans un cadre non convexe en présence de variables latentes. Si δ est la réalisation d'une variable aléatoire Δ dont la loi est donnée par $\mathbb{P}_\theta(\delta)$ (aussi appelée vraisemblance $L(\theta)$), l'estimation par maximum de vraisemblance consiste alors à trouver la valeur θ qui rend le plus probable l'observation θ. Autrement dit il s'agit de déterminer la valeur du paramètre θ qui maximise la vraisemblance. Parfois l'expression de la loi de Δ rend difficile l'utilisation des méthodes classiques d'optimisation, l'idée principale de l'algorithme "EM" est que la loi de Δ n'est que la loi marginale ou la partie observable d'une variable non observée T. Dans le cas des couplages probabilistes, $T = (\Delta, S)$ où S peut être une donnée manquante ou bien une variable latente correspondant au vrai statut des couples à savoir s'ils correspondent ou non à un même individu. Il s'agira alors de se ramener à la maximisation d'une vraisemblance de la variable T. L'avantage étant que le calcul du θ qui maximise la vraisemblance de T est cette fois plus facile à effectuer (on

obtient même une expression explicite dans beaucoup de cas notamment celui du mélange Gaussien). En pratique, il est souvent plus facile de manipuler la log-vraisemblance, et dans la littérature on maximisera plutôt $\log(\mathbb{P}_\theta(\delta))$ ce que nous ferons dans la suite. Plus précisément, cette estimation se fait en deux étapes : la première appelée étape "espérance" consiste à calculer la vraisemblance conditionnelle $V(\theta,\tilde{\theta}) = E\left(\log(\mathbb{P}_\theta(T))\,|\Delta=\delta,\tilde{\theta}\right)$, celle-ci est suivi de l'étape "Maximisation" qui consiste à trouver la valeur θ qui maximise V. Pour calculer V il faut d'abord choisir une valeur initiale $\tilde{\theta}$, l'algorithme alterne alors les étapes "E" et "M" successivement jusqu'à obtenir une stabilisation de la vraisemblance. En effet, on peut démontrer que la suite $(\tilde{\theta}_n)_n$ augmente la valeur de la vraisemblance. Cette monotonie fait de l'algorithme EM un algorithme plus stable que d'autres algorithmes très utilisé comme celui de Newton-Raphson. Par ailleurs ee théorème de Wu [27] garantit la convergence de la séquence des paramètres pour le modèle de Fellegi et Sunter. En théorie l'algorithme converge d'autant plus vite que Δ a de l'information sur S. En appliquant l'algorithme sur la log-vraisemblance on obtient l'estimation suivante à chaque itération. Le paramètre θ se décompose en trois paramètres $\theta = (m,u,\pi)$

Étape "E" :

$$\begin{aligned}
\mathbb{P}_{(n-1)}(M|\delta(a,b)) &= \frac{\pi^{(n-1)}\prod_{1\le i\le K}(m_i^{(n-1)})^{\delta_i(a,b)}(1-m_i^{(n-1)})^{1-\delta_i(a,b)}}{\mathbb{P}_{(n-1)(\delta(a,b))}} \\
\mathbb{P}_{(n-1)}(U|\delta(a,b)) &= \frac{(1-\pi)^{(n-1)}\prod_{1\le i\le K}(u_i^{(n-1)})^{\delta_i(a,b)}(1-u_i^{(n-1)})^{1-\delta_i(a,b)}}{\mathbb{P}_{(n-1)(\delta(a,b))}}
\end{aligned} \tag{4}$$

avec

$$\mathbb{P}_{(n-1)}(\delta(a,b)) = \pi^{(n-1)}\prod_{1\le i\le K}(m_i^{(n-1)})^{\delta_i(a,b)}(1-m_i^{(n-1)})^{1-\delta_i(a,b)} \tag{5}$$

$$+ \quad (1-\pi)^{(n-1)}\prod_{1\le i\le K}(u_i^{(n-1)})^{\delta_i(a,b)}(1-u_i^{(n-1)})^{1-\delta_i(a,b)} \tag{6}$$

Étape "M" :

Pour $i\in\{1,...,K\}$

$$m_i^{(n)} = \frac{\sum_{(a,b)\in E_A\times E_B}\mathbb{P}_{(n-1)}(M|\delta(a,b))\delta_i(a,b)}{\sum_{(a,b)\in E_A\times E_B}\mathbb{P}_{(n-1)}(M|\delta(a,b))} \tag{7}$$

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

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

Le critère d'arrêt de cet algorithme en pratique consiste souvent à comparer la vraisemblance obtenue en utilisant les derniers paramètres estimés avec la vraisemblance obtenue lors de l'itération précédente (on pourra aussi faire une comparaison classique de l'écart entre une estimation et la

précédente). En plus de permettre un traitement non supervisé des appariements probabilistes, le calcul précédent montre que l'algorithme EM a pour avantage d'être assez simple d'utilisation. D'un point de vue pratique l'algorithme se comporte assez bien en présence d'erreurs typographiques (en nombre relatif, mais peut devenir instable si ce nombre est trop important et/ou converger vers de mauvaises solutions), il peut être utilisé comme un outil exploratoire pour donner de bonnes valeurs initiales des probabilités m et u (dans le cas où d'autres modèles mieux ajustés sont utilisés) et enfin il attribue en général aux champs les plus discriminants les meilleurs scores. Cependant, contrairement à une approche supervisée, l'algorithme EM a tendance à donner un taux réel de faux positifs supérieur au taux de faux positifs autorisé. [2] montre sur un exemple qu'une probabilité de faux positifs tolérée de 0.1% donne un taux réel de faux positifs de 1.5%. Dans ce même article il montre que, pour un taux d'erreur de 0.5%, il doit autoriser au modèle un taux d'erreur théorique de 10^{-7}. Par ailleurs si l'hypothèse d'indépendance est violée, les estimations sont alors très biaisées et nécessitent une correction manuelle, a posteriori la plupart du temps [21, 25]. L'algorithme EM classique souffre d'une dépendance aux valeurs initiales et les estimations successives peuvent tendre vers des points qui ne sont au mieux que des points stationnaires de la vraisemblance (c'est-à-dire vers des extrema locaux ou des points selles). L'algorithme peut donc converger vers des points qui posent des problèmes d'interprétation puisqu'il y a de multiples résultats possibles qui n'ont pas forcément les bonnes propriétés. Enfin l'algorithme EM pourrait souffrir d'une vitesse de convergence relativement lente (sur un volume de données important). L'expérience accumulée dans [13] et [24] montre qu'en pratique et sous certaines conditions l'algorithme EM est robuste aux conditions initiales et la convergence peut être très rapide.

1.2.1.3 Calcul non supervisé des occurrences et du pouvoir discriminant d'une modalité :

Lorsque [9] propose la formule (3), les auteurs utilisent en fait des probabilités intermédiaires $m(\delta_i = 1, v_i(a) = t_i)$, $1 \leq i \leq K$. La somme sur t de ces probabilités donnant $m(\delta_i = 1) = m_i$, de même pour le calcul de $m(\delta_i = 1) = m_i$. Il est possible de calculer le pouvoir discriminant des modalités de chacune des variables en calculant les occurrences $O(i)_{E_A \cap E_B, t}$, $O(i)_{E_A, t}$ et $O(i)_{E_B, t_i}$. $O(i)_{E_A \cap E_B, t_i}$ est le nombre de couples pour lesquels $\delta_i = 1$ et $v_i(a) = t$. Alors, si l'on connait la classe M sur un exemple, $\hat{m}(\delta_i = 1, v_i(a) = t_i) = \dfrac{O(i)_{E_A \cap E_B, t_i}}{\#M}$. Dans le cas de l'appariement non supervisé, [12] propose de réaliser un premier appariement, après avoir initialisé les paramètres m et u du modèle de [9]. Puis sur les classes $\widetilde{M}$ et $\widetilde{U}$, on estime les probabilités $m(\delta_i = 1, v_i(a) = t_i)$ et $u(\delta_i = 1, v_i(a) = t_i)$ pour tout $1 \leq i \leq K$. Enfin on recalcule les probabilités m et u, puis on réalise un nouveau couplage et ainsi de suite jusqu'à la convergence des paramètres m et u. Cet algorithme est utilisé par [10] et [18].

1.3 Le blocage et estimation sans biais

Les bases de données médicales et administratives ou statistiques, comme celles que nous pouvons trouver dans le SNDS (Système National des Données de Santé, géré par la Caisse Nationale de l'Assurance Maladie CNAM), dans l'EDP (Echantillon Démographique Permanent, géré par l'INSEE) ou la BCMD (Base des Causes Médicales de Décès, gérée par l'Inserm), sont volumineuses et peuvent rendre difficile l'application directe des méthodes de couplage. L'échantillon $E_A \times E_B$, contient un total de $n_A \times n_B$ individus (sachant que la BCMD cumule environ 550 000 entrées

chaque année). Il est donc nécessaire de réduire le nombre de comparaisons. Si ce blocage est fait avant l'analyse de données, cela peut avoir des conséquences diverses. Dans une approche de l'apprentissage machine non supervisés le blocage peut permettre de rendre plus efficace les méthodes de clustering. En revanche dans le cas du modèle probabiliste cela peut entrainer un biais dans l'estimation des paramètres. Si ce blocage est fait après l'estimation des paramètres ou l'apprentissage des règles de décision, l'effet d'une réduction des comparaisons est connu en théorie [9, 6]. En effet si l'on ne s'intéresse qu'à un sous-ensemble de l'espace des comparaisons il en résultera une réduction du risque d'apparier des individus à tort et une augmentation du nombre des non appariés à tort. De même la classe intermédiaire sur laquelle l'algorithme n'a pas pu statuer est aussi réduite. Si ce blocage est réalisé au moment de la phase d'estimation des paramètres, les estimations des paramètres $(u_i)_i$ seront biaisées.

Il existe plusieurs méthodes de réduction du nombre de paires et on pourra regarder dans [8, 20, 1] pour une synthèse exhaustive. La méthode la plus simple consiste à choisir parmi les champs les plus discriminants et dont la "qualité" est à priori bonne, et à ne comparer que les couples dont la valeur de ces champs de référence est identique. Les algorithmes qui sont souvent utilisés pour construire des champs de blocage, sont du type Soundex, (surtout pour les champs contenant des chaînes de caractères de type nom, prénom et adresse).

Les méthodes de blocage ne sont pas intéressantes que pour le gain de temps réalisé mais aussi pour permettre l'usage de méthode de clustering pour l'apprentissage non supervisé de règles de décision. Les méthodes de K-moyennes sont performantes quand les classes latentes sont bien séparées avec des tailles homogènes, or on sait qu'au sein du produit cartésien des bases on s'attend au mieux à retrouver un nombre de personnes égal à la taille de la plus petite des bases.

Si le blocage est utilisé lors de l'estimation des paramètres alors les paramètres estimés seront biaisés, principalement les paramètres $(u_i)_i$. Jaro [13, 12] propose un ajustement dans l'estimation en remarquant que la proportion de paires correspondant à un unique individu est très faible en comparaison des autres paires, autrement dit $U \approx E_A \times E_B$. Ainsi il est possible d'approximer les paramètres $(u_i)_i$ directement en comparant aléatoirement les lignes de E_A et de E_B. L'algorithme EM ne sert alors qu'à estimer les paramètres $(m_i)_i$ et π.

2 Application du couplage probabiliste :

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

Ce couplage a été réalisé pour démontrer l'applicabilité des méthodes de Fellegi et Sunter et réduire le suivi statistique de patients atteints d'un cancer aux personnes encore vivantes. En l'absence d'un identifiant unique, la détermination du statut vital des patients a été réalisé par le croisement de données hospitalières de l'institut Gustave Roussy (IGR) et des données de la base nationale de mortalité de l'INSEE, après avoir rendu ces informations anonymes [10]. La base de données de l'IGR ne renseigne le statut vital des patients que dans 55% des cas. En revanche en France tous les décès sont enregistrés dans la base nationale de mortalité de l'INSEE. L'ensemble des patients hospitalisés à l'IGR (10 489), domiciliés en France métropolitaine ou dans les départements d'outre-mer, hospitalisés pour la première fois pour une tumeur maligne entre 1998 et 2000 à l'institut Gustave-Roussy ont été inclus. Du côté des données de mortalité de l'INSEE sont incluses les données des années 1998 à 2004 (environ 3,5 millions). Après anonymisation par hachage

(avec l'algorithme SHA), les fichiers de mortalité et de morbidité hospitalière ont été chaînés sur le nom, le premier prénom, la date de naissance et le code de la commune de naissance au Service de Biostatistique et Information Médicale du CHU de Dijon.

TABLEAU 5 – Étape additionnelle de validation automatique et manuelle de couplage des enregistrements sur le nom (N), le prénom (P), la date de naissance (DN) et le code commune de naissance (CN). La colonne Niveau indiquant le niveau de concordance.

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

Pour éviter de comparer trop brutalement les noms et prénoms, un algorithme phonétique d'indexation adapté à la langue française a été utilisé juste avant la fonction de hachage. Ceci a permis de gommer certaines erreurs dues à des fautes de saisie. Une étape supplémentaire de vérification automatique et manuelle a permis de corriger les erreurs que l'algorithme phonétique ne sait pas gérer. Cette étape est décrite dans la table 5 dont les conditions sont les suivantes :

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

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

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

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

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

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

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

La performance de cette appariement est estimée sur sa capacité à prédire le statut vital. Comme nous l'avons dit la base de l'IGR connait le statut vital pour 55% des individus. Le 45% restant ont été complétés par une demande de statut vital au Répertoire National d'Identification des Personnes Physiques (RNIPP). Les résultats du couplage montrent l'intérêt de l'utilisation du couplage

probabiliste pour obtenir des informations sur le statut vital d'un nombre important de patients à un moindre coût, puisque la proportion de bien classés était de $97,2\%$, la sensibilité de $94,8\%$ et la spécificité de $99,5\%$. Les valeurs prédictives négative (VPN) et positive (VPP) sont respectivement de $95,3\%$ et de $99,4\%$. Si l'on ajoute l'étape de vérification manuelle, la proportion de bien classés passe à $98,4\%$ avec une spécificité à $99,4\%$ et une sensibilité à $97,2\%$, la VPP est toujours de $99,4\%$ et la VPN est de $97,4\%$. Ces résultats étaient meilleurs pour les patients nés en France, avec un taux de bien classés de $98,3\%$ (respectivement $99,2\%$ avec l'étape supplémentaire) une sensibilité de $96,8\%$ (respectivement $98,5\%$), une spécificité de $99,8\%$ (identique avec l'étape supplémentaire), une VPP à $99,8\%$ (identique avec l'étape supplémentaire) et une VPN à 97% (respectivement $98,6\%$). Les résultats étaient moins bons pour les patients nés à l'étranger avec un taux de bien classés à $90,7\%$, une sensibilité à $82,8\%$ et spécificité à $97,7\%$ (respectivement un taux de $93,7\%$, une sensibilité de $89,8$, une spécifié à 97.2%), une VPP à $97,0\%$ (respectivement $96,6\%$) et une VPN à $86,5\%$ (respectivement $91,5\%$). L'ajout d'information complémentaire comme le second prénom ou le lieu de naissance (plus précis que le pays) a permis d'améliorer un peu les résultats. Enfin les performances sont dépendantes du sexe, la spécificité et la sensibilité sont meilleures chez les hommes (resp. $99,9\%$ et $97,9\%$) que chez les femmes (respectivement $99,8\%$ et 95%). Cela est probablement dû au fait que le nom de jeune fille n'est pas toujours présent.

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

Le Réseau Périnatal de Bourgogne (RPB) inclut tous les établissements publics et privés de la région prenant en charge les femmes enceintes et les nouveau-nés (environ 18 000 naissances annuelles réparties sur 18 établissements). Un recueil continu de 42 indicateurs a été mis en place en 1998 (25 indicateurs pour la mère et 17 pour le nouveau né). Les informations sont extraites des résumés du Programme de Médicalisation de Systèmes d'Information (PMSI) recueillis pour toutes les hospitalisations. En effet, tout séjour hospitalier, effectué dans un établissement de santé public ou privé, fait l'objet d'un résumé dans l'objectif d'établir le budget des hôpitaux en fonction de leur activité. Les indicateurs n'existant pas dans le PMSI, tels que les facteurs de risques psychosociaux, font l'objet d'un recueil sur une fiche adjointe au résumé PMSI, constituant un "résumé élargi". Pour le traitement des données médicales, le chaînage des "résumés élargis" est réalisé à deux niveaux différents. D'une part, les "résumés élargis" d'une même personne, mère ou nouveau-né, doivent pouvoir être reliés lorsqu'il y a hospitalisations successives (plusieurs unités ou établissements différents). D'autre part, les "résumés élargis" de la mère doivent être reliés à ceux de l'enfant afin d'évaluer l'impact postnatal des facteurs de risques et des pathologies maternelles. Le couplage de données anonymes a alors été rendu possible par l'utilisation du logiciel "Anonymat" à partir de six variables, saisies chez la mère et son bébé : le nom de jeune fille de la mère, son prénom et sa date de naissance, le prénom de l'enfant et sa date de naissance, le code postal de résidence de la mère. Avant transmission, les fichiers sont validés au sein de chaque établissement. De plus, l'exhaustivité et la qualité du recueil des données de chaînage sont systématiquement contrôlées par l'équipe coordinatrice du RPB, qui assure le chaînage mère-enfant (pour 99.9% des nouveau-nés) selon la méthode de [12] et [18]. Dans cet appariement plusieurs hypothèses sont faites, d'abord il est impossible de coupler une mère avec le mauvais nouveau né si l'ensemble des variables concordent. De plus chaque mère doit correspondre à un nouveau né et chaque nouveau né doit avoir une mère. Le cas contraire impliquerait la non exhaustivité des bases ou bien la présence d'erreur. Ainsi plusieurs

appariements probabilistes suivant [12] ont été fait, le premier permet de coupler les cas non problématiques. Les non trouvés sont ensuite appariés en retirant une voire deux variables afin de savoir d'où peuvent provenir les écarts. Les résultats sont ensuite renvoyés aux établissement du réseau afin de faire des vérifications (les premiers appariements ayant permis de limiter les vérifications aux cas les plus problématiques). À la suite de cette étape de vérifications un dernier appariement est réalisé sur l'ensemble des variables. La méthode est évalué ensuite sur un ensemble de test validé manuellement, dont on trouvera les résultats dans la table 6.

TABLEAU 6 – Évaluation de la procédure : sensibilité, spécificité, taux de vrais positifs (VP), taux de vrais négatifs (VN), taux de faux positifs (FP), taux de faux négatifs (FN)

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

La cause principale de la présence des faux positives est le nombre important de données manquantes en effet parmi les faux positifs dans 85,11% des cas il y a concordance sur toutes les variables sauf celles qui sont manquantes.

3 Analyse statistique sur des données appariées

En épidémiologie et en santé publique où l'on cherche souvent à mesurer un risque dans le but de mettre en évidence une association, il existe 3 pratiques assez répandues pour prendre en compte l'incertitude dans un couplage. Idéalement on pourrait disposer d'un jeu de données constitué du couplage de deux bases de données représentatives en utilisant un identifiant direct comme le NIR par exemple (ce type de jeux de données est souvent appelé un gold standard), ce qui permettrait de calculer des valeurs prédictives, ou bien de calibrer des scores et de corriger les biais [15]. Ce point a été traité plus haut, cependant nous tenons à rappeler qu'il est difficile d'avoir accès à ce type de données pour des raisons de coût et pour des raisons réglementaires. Une seconde approche va consister à observer la distribution des individus appariés et de la comparer à la distribution des individus non appariés, cette méthode permet de voir si le fait d'être apparié ou non est lié au facteur d'exposition ou bien à la variable réponse. Cependant pour détecter les non appariés il est nécessaire qu'au moins une des deux bases contienne l'autre. Enfin la troisième approche, la plus connue, consiste à réaliser une analyse de sensibilité afin de déterminer le sens des biais lorsqu'on décide d'inclure dans la classe $\widetilde{M}$ des cas très ambigus en faisant varier le seuil par exemple. Pour réaliser une analyse de sensibilité il est nécessaire d'avoir une information sur l'amplitude du couplage (comme le score dans la proposition de [9]). En effet le résultat d'un appariement n'est pas nécessairement binaire, avec des individus appariés correctement et des non trouvés (ce qui

correspond à une situation de données manquantes). En faisant varier des seuils d'appariement, il est possible de mesurer comment les individus couplés à tort impactent les résultats. On trouvera dans [11] une analyse comparative de ces trois approches. Il existe évidemment des méthodes plus sophistiquées pour prendre en compte les incertitudes liées à l'appariement que l'on trouvera dans [3], elles sont supervisées dans la plupart des cas et sont donc rarement applicables en pratique.

4 Logiciels de couplage

Les principaux logiciels libres pour le couplage probabiliste sont R (au travers du package Record Linkage), FRIL et Febrl [5]. Febrl est une plateforme en open source basée sur Python qui propose une interface graphique pour réaliser son appariement. Il contient des fonctionnalités avancées de nettoyage et de standardisation des données (notamment des méthodes de Markov cachées). Il contient aussi les similarités [3] : indice de Jaro, q-grams, phonétique (il y en a 26 en tout [4]). Il propose les méthodes classiques de couplage probabiliste comme Fellegi et Sunter, à savoir le calcul du score et la discrimination mais pas l'estimation des paramètres du modèle. Il propose également d'autres méthodes utilisant les séparateurs à vaste marge applicable au cadre non supervisé. Par exemple, il permet de développer son propre algorithme de couplage. Febrl est un outil très utile pour débuter et comprendre le fonctionnement des algorithmes de couplage et pour faire des comparaisons entre les différents algorithmes et méthodes, il peut créer des bases synthétiques pour tester des méthodes d'appariement. Cependant, son modèle de gestion de données, chargées en mémoire, ne permet pas de traiter nativement des bases de données très volumineuses. FRIL est une plateforme basée sur Java qui supporte moins d'indices de similarité que Febrl (Levenshtein, q-gram, Jaro-Winkler, Soundex). La règle de décision appliquée dans le processus d'appariement est celle de Fellegi et Sunter (l'estimation des paramètres se faisant par l'algorithme EM). On pourra également trouver dans [26] une présentation sur l'utilisation de la procédure SQL avec la clause JOIN, qui est un outil puissant pour réaliser le produit cartésien des bases de données en incluant du blocage simple uniquement. SAS propose aussi des similarités comme la distance de Levenshtein via les fonctions COMPGED, COMPLEV et SPEDIS ou le Soundex via la fonction SOUNDEX. Enfin il existe un package R pour le couplage probabiliste, Record Linkage, qui contient un ensemble basique de fonctions de similarités (phonétiques, Jaro-Winkler, Levenshtein). Le package permet l'appariement non supervisé via le modèle probabiliste de Fellegi et Sunter ainsi que l'estimation des paramètres via l'algorithme EM, des méthodes de classifications non supervisées (K-mean clustering, bagged clustering). Les arbres de décision, le séparateur à vaste marge (SVM) et les réseaux de neurones sont aussi implémentés pour l'apprentissage supervisé. Pour l'expérimentation le module emprunte le générateur de données synthétique de Febrl. Ce module est aussi l'un des rares à contenir une fonctionnalité de calibration non supervisée des scores [19]. Au final la plateforme Febrl et le package R Record Linkage sont de bons outils pour réaliser des appariements de tout type et pour faire l'expérimentation au regard de la richesse des fonctions qu'ils proposent comme la possibilité de simuler des données ainsi que l'éventail des outils présentés dans ce document. FRIL et SAS seront plus puissant pour des gros volume de données mais sont plus limités en fonctionnalité et nécessiteront plus souvent l'implémentation de nouvelles fonctionnalités (pour le cas de SAS). La réglementation impose l'utilisation d'un serveur/espace sécurisé pour réaliser les traitements statistiques sur des données de santé. Or ces espaces contraignent souvent les outils mis à disposition pour des raisons de sécurité. Même s'il n'y a pas vraiment d'outil spécifique pour l'appariement en santé nous recommandons toutefois les développements en SAS, R et Python que

l'on retrouvera le plus souvent disponible dans ces espaces sécurisés.

5 Conclusion

Nous avons présenté le modèle de Fellegi et Sunter ainsi que les moyens pour estimer les paramètres du modèle de façon non supervisée. Nous avons démontré l'applicabilité de ces méthodes au travers de deux exemples. Il existe un large ensemble de méthodes qui ont été proposées depuis [9] que nous avons présenté dans [3]. Cependant ces généralisations n'ont pu faire leur preuve sur des données françaises. Ce fait est principalement dû à la réglementation qui protège les données individuelles et rend difficile la réalisation d'évaluation. Par ailleurs beaucoup de ces méthodes sont supervisées et sont donc sans intérêt dans le cadre d'un appariement occasionnelle. Enfin il est bien sûr possible d'utiliser des méthodes déterministes basées sur des règles de décision à dire d'expert comme celle proposé par [14]. Il n'est pas garanti qu'un appariement probabiliste puisse donner de meilleurs résultats qu'un appariement déterministe.

Références

[1] R. Batxer, P. Christen, and T. Churches. A comparison fast blocking methods for record linkage, 2003.

[2] T. Belin. A proposed improvement in computer matching. *Statistics of Income and Related Administrative Record Resarch*, pages 167–172, 1990.

[3] S. Bounebache, C. Quantin, E. Benzenine, G. Obozinski, and G. Rey. An Overview of Record Linkage Methods : Applications and Perspective on Health Data. *Journal de la Societe Française de Statistique*, 159(3) :79–123, Dec. 2018.

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

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

[6] P. Christen. *Data Matching : Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer-Verlag Berlin and Heidelberg GmbH & Co. K, 2012 ed. edition, 2012.

[7] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1) :1–38, 1977.

[8] A. Elmagarmind, P. Ipeirotis, and V. Verykios. Duplicate record detection : A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(1), 2007.

[9] I. Fellegi and A. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328) :1183–1210, 1969.

[10] I. Fournel, M. Schwarzinger, C. Binquet, E. Benzenine, C. Hill, and C. Quantin. Contribution of record linkage to vital status determination in cancer patients. *Studies in Health Technology and Informatics*, 150 :91–95, 2009.

[11] K. Harron, J. Doidge, H. Knight, R. Gilbert, H. Goldstein, D. Cromwell, V. Meulen, and H. Jan. A guide to evaluating linkage quality for the analysis of linked data. *International Journal of Epidemiology*, 46(5) :1699–1710, 2017.

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

[13] M. A. Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406), 1989.

[14] A. Lamarche-Vadel, E. Jougla, and G. Rey. *Base AMPHI : Base de données pour l'Analyse de la Mortalité Post-Hospitalisation en France en 2008-2010*. PhD thesis, Université Paris-Sud / Inserm-CépiDC, 2013.

[15] T. L. Lash, M. P. Fox, and A. K. Fink. *Applying Quantitative Bias Analysis to Epidemiologic Data*. Springer Publishing Company, Incorporated, 1st edition, 2009.

[16] H. Newcombe and J. Kennedy. Automatic linkage of vital records. *Science*, 130(3381) :954–959, 1959.

[17] J. Neyman and E. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231 :289–337, 1933.

[18] C. Quantin, B. Gouyon, P. Avillach, C. Ferdynus, P. Sagot, and J.-B. Gouyon. Using discharge abstracts to evaluate a regional perinatal network : Assessment of the linkage procedure of anonymous data. *International Journal of Telemedicine and Applications*, 2009, 2009.

[19] M. Sariyar, A. Borg, and K. Pommerening. Controlling false match rates in record linkage using extreme value theory. *Journal of Biomedical Informatics*, 44(4) :648–654, 2011.

[20] R. Steorts, S. Ventura, M. Sadinle, and S. Fienberg. A comparison of blocking methods for record linkage. In *Privacy in Statistical Databases*, pages 253–268. Springer, Cham, 2014.

[21] Y. Thibaudeau. The discriminant power of dependency structures in record linkage, 1992.

[22] M. Tromp, N. Méray, A. Ravelli, J. Reitsma, and G. Bonsel. Ignoring dependency between linking variables and its impact on the outcome of probabilistic record linkage studies. *Journal of the American Medical Informatics Association : JAMIA*, 15(5) :654–660, 2008.

[23] W. Winkler. Using the EM algorithm for weight computation in the fellegi-sunter model of record linkage, 1988.

[24] W. Winkler. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage, 1990.

[25] W. Winkler. Improved decision rules in the felligi sunter model of record linkage, 1993.

[26] G. Wright. Probabilistic record linkage in SAS. *Proceedings of Western Users of SAS Software*, 2011.

[27] C. Wu and F. Jeff. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1) :95–103, 1983.