Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d’enquête

Classification par entropie maximale aux fins de couplage d’enregistrements

par Danhyang Lee, Li-Chun Zhang et Jae Kwang Kim

Date de diffusion : le 21 juin 2022

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Techniques d'enquête, juin 2022
Vol. 48, n° 1, pp. 1-24
Statistique Canada, n° 12-001-X au catalogue

Classification par entropie maximale aux fins de couplage d'enregistrements

Danhyang Lee, Li-Chun Zhang et Jae Kwang Kim[1]

Résumé

Dans le cadre d'un couplage d'enregistrements, on associe des enregistrements résidant dans des fichiers distincts que l'on pense être reliés à la même entité. Dans la présente étude, nous abordons le couplage d'enregistrements comme un problème de classification et adaptons la méthode de classification par entropie maximale de l'apprentissage automatique pour coupler des enregistrements, tant dans l'environnement d'apprentissage automatique supervisé que non supervisé. L'ensemble de couplages est choisi en fonction de l'incertitude connexe. D'une part, notre cadre de travail permet de surmonter certaines failles théoriques persistantes de l'approche classique dont les pionniers ont été Fellegi et Sunter (1969); d'autre part, l'algorithme proposé est entièrement automatique, contrairement à l'approche classique qui nécessite généralement un examen manuel afin de résoudre des cas indécis.

Mots-clés : Couplage probabiliste; ratio de densité; faux couplage; correspondance manquante; échantillonnage.

1. Introduction

La combinaison de renseignements provenant de plusieurs sources de données est un problème rencontré dans de nombreuses disciplines. Pour combiner des renseignements provenant de différentes sources, on suppose qu'il est possible de déterminer les enregistrements associés à la même entité, ce qui n'est pas toujours le cas en pratique. L'entité peut être une personne, une entreprise, la criminalité, etc. Si les données ne comprennent pas de numéro d'identification unique, la détermination des enregistrements provenant de la même entité devient un problème difficile. Le *couplage d'enregistrements* est le terme décrivant le processus de couplage des enregistrements que l'on pense être associés à la même entité. Alors que le couplage d'enregistrements peut entraîner le couplage d'enregistrements d'un seul fichier informatique pour déterminer les doublons (ce que l'on appelle une *déduplication*), nous nous concentrons plutôt sur le couplage d'enregistrements de fichiers distincts.

Le couplage d'enregistrements est employé depuis plusieurs décennies dans l'échantillonnage permettant de produire des statistiques officielles. En particulier, le couplage de fichiers administratifs avec des données de l'échantillon d'enquête peut considérablement améliorer la qualité et la résolution de statistiques officielles. À titre d'applications, Jaro (1989) ainsi que Winkler et Thibaudeau (1991) ont fusionné des données d'enquêtes postérieures au recensement et de recensement aux fins d'évaluation de la couverture du recensement. Zhang et Campbell (2012) ont couplé des fichiers de données du recensement de la population au fil du temps, alors qu'Owen, Jones et Ralphs (2015) ont couplé des registres administratifs pour créer un seul ensemble de données statistiques sur la population. L'approche classique dont Fellegi et Sunter (1969) ont été les pionniers (la méthode la plus populaire de couplage d'enregistrements en pratique) a servi avec succès dans ces applications.

1. Danhyang Lee, Department of Information Systems, Statistics and Management Science, University of Alabama, Tuscaloosa (Alabama), États-Unis; Li-Chun Zhang, Department of Social Statistics and Demography, University of Southampton, Southampton, Royaume-Uni, Bureau central des statistiques norvégien, Oslo, Norvège et faculté de mathématiques, université d'Oslo, Oslo, Norvège. Courriel : L.Zhang@soton.ac.uk; Jae Kwang Kim, Department of Statistics, Iowa State University, Ames (Iowa), États-Unis.

La règle de décision probabiliste de Fellegi et Sunter (1969) repose sur la notion de test du rapport de vraisemblance, qui permet de déterminer la probabilité qu'une paire d'enregistrements donnée soit une réelle correspondance. En appliquant la notion du test du rapport de vraisemblance, on doit estimer les paramètres du modèle sous-jacent et déterminer les seuils de la règle de décision. Winkler (1988) et Jaro (1989) traitent l'état de correspondance comme une variable non observée et proposent un algorithme espérance-maximisation (algorithme EM) pour l'estimation des paramètres, ce que nous appellerons la « procédure WJ ». Voir Herzog, Scheuren et Winkler (2007), Christen (2012) ainsi que Binette et Steorts (2020) pour obtenir une vue d'ensemble. Toutefois, comme nous l'expliquons à la section 2, justifier que la procédure WJ est un algorithme EM nécessite l'hypothèse cruciale que les mesures de correspondance entre les paires d'enregistrements (appelées *vecteurs de comparaison*) soient indépendantes d'une paire d'enregistrements à une autre, ce qui est impossible à vérifier en réalité. Newcombe, Kennedy, Axford et James (1959) abordent la dépendance entre des vecteurs de comparaison au moyen de l'application de données. Voir également, par exemple, Tancredi et Liseo (2011), Sadinle (2017) ainsi que Binette et Steorts (2020) pour des analyses de ce sujet. Des approches bayésiennes du couplage d'enregistrements sont également disponibles dans la littérature (Steorts, 2015; Sadinle, 2017; Stringham, 2021). Les approches bayésiennes pour aborder les problèmes du couplage d'enregistrements nous permettent de quantifier l'incertitude relative aux décisions de correspondance. Toutefois, la recherche stochastique fondée sur un algorithme MCMC (méthode de Monte Carlo par chaîne de Markov) dans l'approche bayésienne fait intervenir un fardeau de calcul supplémentaire.

Pour élaborer une autre approche, nous remarquons d'abord que le problème du couplage d'enregistrements est essentiellement un problème de classification, selon lequel chaque paire d'enregistrements est classée dans une catégorie « correspondance » ou « non-correspondance ». Diverses techniques de classification fondées sur des approches d'apprentissage automatique ont été utilisées en couplage d'enregistrements (Hand et Christen, 2018; Christen, 2012, 2008; Sarawagi et Bhamidipaty, 2002). Dans la présente étude, nous adaptons la méthode de classification par entropie maximale au couplage d'enregistrements. En particulier, nous pouvons considérer le rapport de vraisemblance de la méthode que proposent Fellegi et Sunter (1969) comme un cas particulier du ratio de densité et appliquer la méthode d'entropie maximale à l'estimation du ratio de densité. Nigam, Lafferty et McCallum (1999) utilisent, par exemple, l'entropie maximale pour la classification de texte et Nguyen, Wainwright et Jordan (2010) élaborent une théorie plus unifiée de la méthode d'entropie maximale pour l'estimation du ratio de densité. Il existe, cependant, une différence clé entre le couplage d'enregistrements et le cadre standard des problèmes de classification : les différentes paires d'enregistrements ne sont pas des « unités » distinctes, car le même enregistrement fait partie de nombreuses paires d'enregistrements.

Même si nous présentons notre algorithme de couplage d'enregistrements par entropie maximale tant pour des environnements supervisés que non supervisés, nos principales contributions concernent le cas non supervisé. Les approches supervisées nécessitent des données d'apprentissage, c'est-à-dire des paires d'enregistrements présentant des états connus de vraie correspondance et de vraie non-correspondance. De

telles données d'apprentissage ne sont souvent pas disponibles dans les situations réelles ou doivent être préparées manuellement, ce qui est coûteux et prend beaucoup de temps (Christen, 2007). Le cas non supervisé est donc, de loin, le plus courant en pratique. Dans le cas non supervisé, cependant, il n'est pas possible d'estimer directement le ratio de densité à partir des vraies correspondances et non-correspondances observées et il est difficile de créer conjointement des modèles de l'état de correspondance non observé et des résultats de comparaison observés pour toutes les paires d'enregistrements. Nous élaborons donc un nouvel algorithme itératif, afin d'estimer conjointement le ratio de densité ainsi que la classification par entropie maximale définie dans l'enregistrement non supervisé et prouvons sa convergence. Nous élaborons également les mesures associées à l'incertitude du couplage.

Nous montrons en outre que la procédure WJ peut être intégrée comme cas particulier de notre approche de l'estimation, mais sans la nécessité de l'hypothèse d'indépendance entre les paires d'enregistrements. Cela indique que la procédure WJ peut être justifiée sans l'hypothèse d'indépendance et explique la raison pour laquelle elle fournit des résultats raisonnables dans de nombreuses situations. Les mesures de l'incertitude élaborées dans la présente étude guident le choix de l'ensemble de couplages. Il s'agit d'une importante amélioration pratique par rapport à l'approche classique, qui ne fournit pas directement de mesure de l'incertitude pour l'ensemble final de couplages. Notre procédure est entièrement automatique, sans nécessiter de revue par le commis exigeante en ressources et nécessaire dans l'approche classique.

La présente étude est structurée de la manière suivante. La section 2 présente la configuration de base et l'approche classique. À la section 3, nous élaborons la méthode proposée dans le cadre d'un couplage d'enregistrements supervisé. À la section 4, nous étendons la méthode proposée à l'environnement plus difficile d'un couplage d'enregistrements non supervisé. Des discussions sur des approches d'estimation ainsi que des renseignements techniques connexes sont présentés à la section 5 et dans la documentation supplémentaire. Les résultats d'une étude de simulation complète sont présentés à la section 6. Nous présentons une conclusion et des commentaires sur des travaux à venir à la section 7.

2. Problèmes de l'approche classique

Supposons que nous disposons de deux fichiers de données, A et B; nous pensons qu'ils contiennent de nombreuses entités communes, mais aucun doublon au sein de chaque fichier. Tout enregistrement dans A et un autre dans B peut ou non faire référence à la même entité. Notre objectif est de trouver les vraies correspondances parmi toutes les paires possibles des deux fichiers de données. L'*espace de comparaison* bipartite $\Omega = A \times B = M \cup U$ consiste en des *correspondances* M et des *non-correspondances* U entre les enregistrements des fichiers A et B. Pour toute paire d'enregistrements $(a, b) \in \Omega$, $\boldsymbol{\gamma}_{ab}$ est le *vecteur de comparaison* entre un ensemble de variables *clés* respectivement associées à $a \in A$ et $b \in B$, comme le nom, le sexe, la date de naissance. Les variables clés et le vecteur de comparaison $\boldsymbol{\gamma}_{ab}$ sont entièrement observés dans Ω. Dans les cas où des erreurs peuvent influer sur les

variables clés, une correspondance (a,b) peut ne pas présenter de correspondance complète sur le plan de γ_{ab} et une non-correspondance (a,b) peut tout de même correspondre à certaines (parfois l'intégralité) des variables clés.

Dans l'approche classique de Fellegi et Sunter (1969), la nature probabiliste de γ_{ab} est reconnue, du fait des perturbations entraînant des erreurs dans les variables clés. Les méthodes connexes sont appelées *couplage d'enregistrements probabiliste*. Pour expliquer la méthode de couplage d'enregistrements probabiliste de Fellegi et Sunter (1969), disons que $m(\gamma_{ab}) = f\left(\gamma_{ab} \middle| (a,b) \in M\right)$ la fonction de masse de probabilité que les valeurs discrètes γ_{ab} peuvent prendre avec $(a,b) \in M$. De la même manière, nous pouvons définir $u(\gamma_{ab}) = f\left(\gamma_{ab} \middle| (a,b) \in U\right)$. Le ratio :

$$r_{ab} = \frac{m(\gamma_{ab})}{u(\gamma_{ab})}$$

est alors la base du test du rapport de vraisemblance (TRV) pour $H_0 : (a,b) \in M$ par rapport à $H_1 : (a,b) \in U$. Soit $M^* = \{(a,b) : r_{ab} > c_M\}$, les paires classées comme correspondances et $U^* = \{(a,b) : r_{ab} < c_U\}$ les non-correspondances; le reste des paires est classé par revue par le commis, où (c_M, c_U) sont respectivement les seuils associés aux probabilités de faux couplages (de paires dans U) et de faux non-couplages (de paires dans M), définis sous la forme :

$$\mu = \sum_{\gamma} u(\gamma)\, \delta(M^*; \gamma) \quad \text{et} \quad \lambda = \sum_{\gamma} m(\gamma)\, \delta(U^*; \gamma), \tag{2.1}$$

où $\delta(M^*; \gamma) = 1$, si $\gamma_{ab} = \gamma$ signifie $(a,b) \in M^*$ et 0 autrement; de même pour $\delta(U^*; \gamma)$.

En pratique, les probabilités $m(\gamma)$ et $u(\gamma)$ sont inconnues; tout comme la *prévalence* de vraies correspondances, définie par $\pi = |M| / |\Omega| := n_M / n$. Soit $\boldsymbol{\eta}$, l'ensemble contenant π et les paramètres inconnus de $m(\gamma)$ et $u(\gamma)$. Soit $g_{ab} = 1$, si $(a,b) \in M$ et 0 si $(a,b) \in U$. Avec les données complètes $\{(g_{ab}, \gamma_{ab}) : (a,b) \in \Omega\}$, Winkler (1988) et Jaro (1989) supposent que la probabilité est :

$$h(\boldsymbol{\eta}) = \sum_{(a,b) \in \Omega} g_{ab} \log\left(\pi m(\gamma_{ab})\right) + \sum_{(a,b) \in \Omega} (1 - g_{ab}) \log\left((1-\pi)\, u(\gamma_{ab})\right). \tag{2.2}$$

Un algorithme EM s'ensuit en traitant $g_\Omega = \{g_{ab} : (a,b) \in \Omega\}$ comme les données manquantes.

L'approche classique présente deux problèmes fondamentaux.

[Problème I] Le couplage d'enregistrements n'est pas une application directe du test du rapport de vraisemblance, car *toutes* les paires dans Ω doivent être évaluées plutôt que toute paire *donnée*. La classification de Ω en M^* et U^* est généralement incohérente, puisqu'un enregistrement donné peut appartenir à plusieurs paires dans M^*. Une déduplication de M^* après classification serait alors nécessaire, ce qui ne fait *pas* partie de la formulation théorique ci-dessus. En particulier, il manque une méthodologie connexe d'estimation de l'incertitude relative à l'ensemble couplé final, comme le nombre de faux couplages qu'il comprend ou les correspondances restantes en dehors de cet ensemble.

[Problème II] En réalité, les vecteurs de comparaison entre deux paires ne sont pas indépendants, tant qu'ils partagent un enregistrement. Par exemple, avec $(a,b) \in M$ et $\boldsymbol{\gamma}_{ab}$ ne faisant pas l'objet d'erreurs, alors $g_{ab'}$ doit être 0, pour $b' \neq b$ et $b' \in B$, tant qu'il n'existe pas d'enregistrement en double dans A ni dans B et que $\boldsymbol{\gamma}_{ab'}$ dépend uniquement des erreurs de variables clés de b'. Alors que, marginalement, $g_{ab'} = 1$ avec une probabilité π et $\boldsymbol{\gamma}_{ab'}$ dépend également des erreurs de variables clés de a. Il s'ensuit que $h(\boldsymbol{\eta})$ dans (2.2) ne correspond pas à la répartition réelle de données couplées de $\boldsymbol{\gamma}_\Omega = \{\boldsymbol{\gamma}_{ab} : (a,b) \in \Omega\}$, même lorsque les probabilités marginales m et u sont correctement définies. De la même manière, même si l'on peut *marginalement* définir $\pi = \Pr[(a,b) \in M \,|\, (a,b) \in \Omega]$ pour une paire d'enregistrements sélectionnée *aléatoirement* dans Ω, il ne s'ensuit pas que $\log f(g_\Omega) = n_M \log \pi + (n - n_M)\log(1-\pi)$ *conjointement* comme dans (2.2). Pour ces deux raisons, $h(\boldsymbol{\eta})$ selon (2.2) ne peut pas être le logarithme du rapport de vraisemblance des données complètes.

Dans les deux prochaines sections, nous élaborons une classification par entropie maximale du couplage d'enregistrements, afin d'éviter les problèmes mentionnés ci-dessus; d'autres discussions de l'approche classique sont présentées ensuite.

3. Classification par entropie maximale : environnement supervisé

Comme nous l'avons mentionné à la section 1, le problème de couplage d'enregistrements est un problème de classification. La classification par entropie maximale est utilisée en restauration d'image ou en analyse de texte (Gull et Daniell, 1984; Berger, Della Pietra et Della Pietra, 1996). La *classification par entropie maximale (CEM)* a été proposée pour l'apprentissage supervisé dans le cadre de problèmes de classification standard, lorsque les unités sont connues, mais que les catégories réelles des unités sont inconnues à part un échantillon d'*unités étiquetées*. Soit $Y \in \{1, 0\}$ la catégorie réelle et $\mathbf{X}$, le vecteur aléatoire de caractéristiques. Le ratio de densité est :

$$r(\mathbf{x}; \boldsymbol{\eta}) = \frac{f(\mathbf{x} \,|\, Y = 1; \boldsymbol{\eta})}{f(\mathbf{x} \,|\, Y = 0; \boldsymbol{\eta})} := \frac{f_1(\mathbf{x}; \boldsymbol{\eta})}{f_0(\mathbf{x}; \boldsymbol{\eta})},$$

où f_1 et f_0 sont respectivement des fonctions de densité conditionnelles données par $Y = 1$ ou 0, et $\boldsymbol{\eta}$ contient les paramètres inconnus. Pour la CEM fondée sur $r(\mathbf{x})$, nous constatons $\hat{\boldsymbol{\eta}}$ qui maximise la divergence de Kullback-Leibler (KL) de f_0 à f_1 soumise à une contrainte, c'est-à-dire :

$$D = \int_{S_1} f_1(\mathbf{x}; \boldsymbol{\eta}) \log r(\mathbf{x}; \boldsymbol{\eta})\, d\mathbf{x} \quad \text{soumise à} \quad \int_{S_1} f_0(\mathbf{x}; \hat{\boldsymbol{\eta}})\, r(\mathbf{x}; \hat{\boldsymbol{\eta}})\, d\mathbf{x} = 1,$$

où S_1 est le soutien de $\mathbf{X}$ étant donné $Y = 1$; la contrainte de normalisation survient, puisque $r(\mathbf{x}; \hat{\boldsymbol{\eta}})\, f_0(\mathbf{x}; \hat{\boldsymbol{\eta}})$ est une estimation de $f_1(\mathbf{x})$. Sous réserve d'un soutien commun $S_1 = S_0$, où S_0 est le

soutien de $\mathbf{X}$ étant donné $Y = 0$, nous pouvons utiliser la fonction de répartition empirique de X sur $\{\mathbf{x}_i : y_i = 1\}$ à la place de f_1 pour D, et celle sur $\{\mathbf{x}_i : y_i = 0\}$ à la place de f_0 pour la contrainte. En ayant obtenu $\hat{r}_{\mathbf{x}} = r(\mathbf{x}; \hat{\boldsymbol{\eta}})$, nous pouvons classer toute unité en fonction du vecteur de caractéristiques connexe $\mathbf{x}$ en fonction de $\Pr(Y = 1 \mid \mathbf{x}; \hat{p}, \hat{r}_{\mathbf{x}})$, où $\hat{p}$ est une estimation de la prévalence $p = \Pr(Y = 1)$.

Nous décrivons la façon dont la notion de CEM pour l'apprentissage supervisé peut être adaptée au problème de couplage d'enregistrements dans les sous-sections suivantes.

3.1 Ratio de probabilité du couplage d'enregistrements

Pour une CEM basée sur l'apprentissage supervisé aux fins de couplage d'enregistrements, supposons que M est observé pour Ω donné et que le classificateur entraîné doit être appliqué aux paires d'enregistrements hors de Ω. Pour bien comprendre le concept, supposons que B est un échantillon non probabiliste chevauchant la population P et A est un échantillon probabiliste de P présentant des probabilités d'inclusion connues. Alors que $\boldsymbol{\gamma}_M = \{\boldsymbol{\gamma}_{ab} : (a, b) \in M\}$ peut être considéré comme un échantillon à unités indépendantes et identiquement distribuées (IID), puisque chaque (a, b) dans M fait référence à une entité distincte, cela n'est pas le cas pour $\{\boldsymbol{\gamma}_{ab} : (a, b) \notin M\}$, dont la répartition *conjointe* interfère avec le modèle.

Ratio de probabilité (I)

Soit $r_q(\boldsymbol{\gamma})$ le *ratio de probabilité* déterminé par :

$$r_q(\boldsymbol{\gamma}) = \frac{m(\boldsymbol{\gamma})}{q(\boldsymbol{\gamma})},$$

où $m(\boldsymbol{\gamma})$ est la fonction de probabilité de masse de $\boldsymbol{\gamma}_{ab} = \boldsymbol{\gamma}$ étant donné $g_{ab} = 1$, et $q(\boldsymbol{\gamma})$ celle sur $\boldsymbol{\gamma}_\Omega = \{\boldsymbol{\gamma}_{ab} : (a, b) \in \Omega\}$. La mesure de la divergence de KL de $q(\boldsymbol{\gamma})$ à $m(\boldsymbol{\gamma})$ et la contrainte de normalisation sont :

$$D_f = \sum_{\boldsymbol{\gamma} \in S(M)} m(\boldsymbol{\gamma}) \log r_q(\boldsymbol{\gamma}) \quad \text{et} \quad \sum_{\boldsymbol{\gamma} \in S(M)} \hat{q}(\boldsymbol{\gamma}) \hat{r}_q(\boldsymbol{\gamma}) = 1,$$

où $S(M)$ est le soutien de $\boldsymbol{\gamma}_{ab}$ étant donné $g_{ab} = 1$. Cette configuration permet à $S(M)$ d'être un sous-ensemble de S, où S est le soutien de tous les $\boldsymbol{\gamma}_{ab}$ possibles. Il s'ensuit que, en fonction de l'échantillon à unités IID $\boldsymbol{\gamma}_M$ de taille $n_M = |M|$, la fonction objective à *minimiser* pour r_q peut s'exprimer sous la forme :

$$Q_f = \sum_{(a,b) \in M} \frac{f(\boldsymbol{\gamma}_{ab})}{n_M(\boldsymbol{\gamma}_{ab})} r_q(\boldsymbol{\gamma}_{ab}) - \frac{1}{n_M} \sum_{(a,b) \in M} \log r_q(\boldsymbol{\gamma}_{ab}), \tag{3.1}$$

où $n_M(\boldsymbol{\gamma}_{ab}) = \sum_{(i,j) \in M} \mathbb{I}(\boldsymbol{\gamma}_{ij} = \boldsymbol{\gamma}_{ab})$ en fonction du soutien observé $S(M)$.

Ratio de probabilité (II)

Sous réserve que $S(M) \subseteq S(U)$, où $S(U)$ est le soutien de $\boldsymbol{\gamma}_{ab}$ sur U, il est possible de considérer le ratio de probabilité comme étant :

$$r(\boldsymbol{\gamma}) = \frac{m(\boldsymbol{\gamma})}{u(\boldsymbol{\gamma})}$$

où $u(\boldsymbol{\gamma})$ est la probabilité de $\boldsymbol{\gamma}_{ab} = \boldsymbol{\gamma}$ étant donné $g_{ab} = 0$. Cela donne :

$$r_q(\boldsymbol{\gamma}) = \frac{m(\boldsymbol{\gamma})}{q(\boldsymbol{\gamma})} = \frac{m(\boldsymbol{\gamma})}{\pi m(\boldsymbol{\gamma}) + (1-\pi)u(\boldsymbol{\gamma})} = \frac{r(\boldsymbol{\gamma})}{\pi(r(\boldsymbol{\gamma})-1)+1}$$

où $q(\boldsymbol{\gamma}) = \pi m(\boldsymbol{\gamma}) + (1-\pi)u(\boldsymbol{\gamma})$, de sorte que $r_q(\boldsymbol{\gamma})$ et $r(\boldsymbol{\gamma})$ sont univoques. Parallèlement, la mesure de divergence de KL de $u(\boldsymbol{\gamma})$ à $m(\boldsymbol{\gamma})$ est fournie par :

$$D = \sum_{\boldsymbol{\gamma} \in S(M)} m(\boldsymbol{\gamma}) \log r(\boldsymbol{\gamma})$$

et la fonction objective à *minimiser* pour r peut désormais s'exprimer sous la forme :

$$Q = \sum_{(a,b) \in M} \frac{u(\boldsymbol{\gamma}_{ab})}{n_M(\boldsymbol{\gamma}_{ab})} r(\boldsymbol{\gamma}_{ab}) - \frac{1}{n_M} \sum_{(a,b) \in M} \log r(\boldsymbol{\gamma}_{ab}). \tag{3.2}$$

Modèle de $\boldsymbol{\gamma}$: selon le modèle multinomial, il est possible de simplement utiliser la fonction de répartition empirique de $\boldsymbol{\gamma}$ sur $\boldsymbol{\gamma}_\Omega$ sous forme $f(\boldsymbol{\gamma})$, pour chaque niveau distinct de $\boldsymbol{\gamma}$, tant que $|\Omega|$ est grand par rapport à $|S|$. Il en va de même pour $m(\boldsymbol{\gamma})$ sur $\boldsymbol{\gamma}_M$ et $u(\boldsymbol{\gamma})$ sur U. Pour le couplage hors de Ω, la valeur estimée de $m(\boldsymbol{\gamma})$ tirée de $M(\Omega)$ s'applique, si la sélection de A tirée de P n'est pas informative.

Pour $\boldsymbol{\gamma}$ constitué de K indicateurs binaires de correspondance, $\gamma_k = 0, 1$ pour $k = 1, \ldots, K$, il existe jusqu'à 2^K niveaux distincts de $\boldsymbol{\gamma}$ qui peuvent parfois être relativement grands par rapport à $|M|$. Un modèle plus parcimonieux de $m(\boldsymbol{\gamma}; \boldsymbol{\theta})$ couramment utilisé est défini par :

$$m(\boldsymbol{\gamma}; \boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{\gamma_k} (1-\theta_k)^{1-\gamma_k} \tag{3.3}$$

où $\theta_k = \Pr(\gamma_{ab,k} = 1 \mid g_{ab} = 1)$ et $\gamma_{ab,k}$ est la k^{e} composante de $\boldsymbol{\gamma}_{ab}$. Il est possible de modéliser θ_k en fonction des répartitions des variables clés donnant lieu à $\boldsymbol{\gamma}$, ce qui est fondé sur les fréquences différentielles de leurs valeurs, comme le fait que certains noms sont plus courants que d'autres. De façon similaire, $u(\boldsymbol{\gamma}; \boldsymbol{\xi})$ peut être modélisé comme dans (3.3) en ayant recours aux paramètres ξ_k au lieu de θ_k, où $\xi_k = \Pr(\gamma_{ab,k} = 1 \mid g_{ab} = 0)$.

Il convient de mentionner que (3.3) sous-entend une indépendance conditionnelle parmi les indicateurs de correspondance. Winkler (1993) et Winkler (1994) ont démontré que même lorsque l'hypothèse d'indépendance conditionnelle n'est pas vérifiée, les résultats fondés sur une hypothèse d'indépendance conditionnelle sont relativement robustes. Des modèles plus complexes permettant des valeurs corrélées de γ_k peuvent également être envisagés. Voir Armstrong et Mayda (1993) ainsi que Larsen et Rubin

(2001) pour obtenir des analyses de ces modèles. Voir Xu, Li, Shen, Hui et Grannis (2019) pour consulter une étude visant à comparer des modèles avec ou sans γ_k corrélés.

3.2 Ensembles de classification par entropie maximale aux fins de couplage d'enregistrements

Sous réserve qu'il n'existe aucun enregistrement en double dans A ni B, un *ensemble de classification* aux fins de couplage d'enregistrements, appelé $\hat{M}$, consiste en des paires d'enregistrements de Ω, dans le cadre desquelles tout enregistrement dans A ou B figure, tout au plus, dans une paire d'enregistrements dans $\hat{M}$. Soit l'*entropie* d'un ensemble de classification $\hat{M}$ définie par :

$$D_{\hat{M}} = \frac{1}{\left|\hat{M}\right|} \sum_{(a,b)\in\hat{M}} \log r\left(\boldsymbol{\gamma}_{ab}\right). \tag{3.4}$$

Un ensemble de CEM d'une taille donnée $n^* = \left|\hat{M}\right|$ est le premier ensemble de classification de taille n^*, obtenu par déduplication d'ordre décroissant de $r\left(\boldsymbol{\gamma}_{ab}\right)$ sur Ω. Il est possible que $(a,b') \notin \hat{M}$ et $r\left(\boldsymbol{\gamma}_{ab'}\right) > r\left(\boldsymbol{\gamma}_{a',b'}\right)$ pour $(a',b') \in \hat{M}$, s'il existe $(a,b) \in \hat{M}$ avec $r\left(\boldsymbol{\gamma}_{ab}\right) > r\left(\boldsymbol{\gamma}_{ab'}\right)$.

Un ensemble de CEM de taille n^* n'est pas nécessairement l'ensemble de classification le plus grand possible présentant l'entropie maximale, qu'il faudra appeller ensemble *maximal* de CEM et qui est l'ensemble de classification le plus grand de sorte que $r\left(\boldsymbol{\gamma}_{ab}\right) = \max_{\boldsymbol{\gamma}} r\left(\boldsymbol{\gamma}\right)$ pour chaque (a,b) qu'il contient. En pratique, un ensemble maximal de CEM est donné par le premier passage d'un *couplage déterministe*, qui consiste uniquement en des paires d'enregistrements présentant une correspondance parfaite *et* unique de toutes les variables clés.

La méthodologie de couplage probabiliste pour un ensemble de CEM est utile si nous souhaitons permettre des couplages supplémentaires, même si leurs variables clés ne correspondent pas parfaitement. Pour l'incertitude associée à un ensemble de CEM donné $\hat{M}$, nous considérons deux types d'erreurs. Tout d'abord, nous définissons le *taux de faux couplages (TFC)* parmi les couplages dans $\hat{M}$ comme étant :

$$\psi = \frac{1}{\left|\hat{M}\right|} \sum_{(a,b)\in\hat{M}} \left(1 - g_{ab}\right) \tag{3.5}$$

qui est différent de μ selon (2.1) où le dénominateur est $|U|$. Ensuite, le *taux de correspondances manquantes (TCM)* de $\hat{M}$, associé à la probabilité de faux non-couplages λ dans (2.1), est fourni par :

$$\tau = 1 - \frac{1}{n_M} \sum_{(a,b)\in\hat{M}} g_{ab}. \tag{3.6}$$

Alors que μ et λ dans (2.1) sont des probabilités théoriques, les taux TFC et TCM sont des erreurs réelles.

Il est instructif d'envisager la situation selon laquelle on nous demanderait de former des ensembles de CEM dans Ω avec toutes les estimations nécessaires associées au ratio de probabilité $r\left(\boldsymbol{\gamma}\right)$, qui peut être

obtenu dans l'environnement d'apprentissage supervisé, sans que n_M, g_Ω ou M ne soient donnés directement.

Tout d'abord, l'ensemble de CEM parfait devrait avoir la taille n_M. Soit $n(\boldsymbol{\gamma}) = \sum_{(a,b)\in\Omega} \mathbb{I}(\boldsymbol{\gamma}_{ab} = \boldsymbol{\gamma})$. Nous pouvons obtenir n_M d'après la résolution de l'équation fixe suivante :

$$n_M = \sum_{(a,b)\in\Omega} \hat{g}(\boldsymbol{\gamma}_{ab}) = \sum_{\gamma\in S} n(\boldsymbol{\gamma})\,\hat{g}(\boldsymbol{\gamma}) \tag{3.7}$$

où

$$\hat{g}(\boldsymbol{\gamma}) := \Pr\left(g_{ab} = 1 \,\middle|\, \boldsymbol{\gamma}_{ab} = \boldsymbol{\gamma}\right) = \frac{\pi r(\boldsymbol{\gamma})}{\pi\left(r(\boldsymbol{\gamma})-1\right)+1} = \frac{n_M r(\boldsymbol{\gamma})}{n_M\left(r(\boldsymbol{\gamma})-1\right)+n} \tag{3.8}$$

et la probabilité est définie relativement à un échantillonnage entièrement aléatoire d'une seule paire d'enregistrements de Ω. Pour constater que $\hat{g}(\boldsymbol{\gamma})$ selon (3.8) satisfait à (3.7), prenons note que $\hat{g}(\boldsymbol{\gamma}) = n_M m(\boldsymbol{\gamma}) / n(\boldsymbol{\gamma})$ satisfait à (3.7) pour toute valeur de $m(\boldsymbol{\gamma})$ bien définie et $n(\boldsymbol{\gamma})/n = \pi m(\boldsymbol{\gamma}) + (1-\pi)\,u(\boldsymbol{\gamma})$ par définition.

Ensuite, à part l'ensemble maximal de CEM, on devrait accepter les paires discordantes. Dans l'environnement d'apprentissage supervisé, nous observons la fonction de répartition empirique de $\boldsymbol{\gamma}$ sur M, donnant lieu à $\hat{\theta}_k = n_M(1;k)/n_M$, où $n_M(1;k)$ est le nombre de correspondances pour la k^{e} variable clé sur M. L'ensemble de CEM parfait $\hat{M}$ devrait présenter ces taux de correspondance. Nous obtenons alors, pour $k = 1,\ldots,K$,

$$\hat{\theta}_k = \frac{1}{\left|\hat{M}\right|} \sum_{(a,b)\in\hat{M}} \mathbb{I}(\gamma_{ab,k} = 1) \quad \text{pour} \quad \left|\hat{M}\right| = n_M. \tag{3.9}$$

Ainsi, quelle que soit la modélisation de $m(\boldsymbol{\gamma})$, l'ensemble de CEM parfait devrait satisfaire conjointement aux équations de $K+1$ définies dans (3.7) et (3.9), compte tenu de la connaissance de $r(\boldsymbol{\gamma})$.

4. Classification par entropie maximale aux fins de couplage d'enregistrements non supervisé

Soit $\mathbf{z}$ le vecteur K des variables clés, qui peut être imparfait pour deux raisons : il n'est pas suffisamment riche si les vraies valeurs $\mathbf{z}$ ne sont pas uniques pour chaque entité distincte sous-jacente aux deux fichiers à coupler ou il peut faire l'objet d'erreurs si le $\mathbf{z}$ observé n'est pas égal à sa vraie valeur. Supposons que A contienne uniquement les vecteurs $\mathbf{z}$ distincts du premier fichier, après suppression de tout autre enregistrement présentant un vecteur $\mathbf{z}$ en double par rapport à un enregistrement conservé dans A. En d'autres termes, si le premier fichier contient initialement deux enregistrements ou plus ayant exactement la même valeur de clé combinée, alors seulement l'un d'entre eux sera conservé dans A pour le couplage d'enregistrements avec le deuxième fichier. De la même façon, supposons que B contient les

enregistrements uniques du deuxième fichier. La raison d'une *déduplication séparée des clés* est qu'aucune comparaison des deux fichiers ne peut distinguer les $\mathbf{z}$ en double dans l'un ou l'autre fichier, ce qui est un problème à résoudre autrement.

Si A et B sont prétraités comme cela a été décrit ci-dessus, l'ensemble maximal de CEM M_1 comprend uniquement les paires d'enregistrements présentant une correspondance parfaite de toutes les variables clés. Pour un couplage probabiliste au-delà de M_1, nous pouvons suivre le même processus de CEM que pour l'apprentissage supervisé, tant qu'il est possible d'obtenir une estimation du ratio de probabilité, selon celle pouvant former l'ensemble de CEM de la taille choisie. Néanmoins, pour estimer les taux TFC (3.5) et TCM (3.6) connexes, une estimation de n_M est également nécessaire.

4.1 Algorithme d'une classification par entropie maximale non supervisée

L'idée maintenant est d'appliquer (3.7) et (3.9) conjointement. Puisque définir $\hat{n}_M = |M_1|$ et $\hat{\theta}_k \equiv 1$ associés à l'ensemble maximal de CEM satisfait automatiquement à (3.7) et à (3.9), pour un couplage probabiliste, il faut supposer $n_M > |M_1|$ et $\theta_k < 1$ pour au moins certains $k = 1, \ldots, K$. De plus, à moins d'indications externes contraires, nous pouvons seulement supposer un soutien commun $S(M) = S(U)$ dans l'environnement non supervisé. Soit :

$$r(\boldsymbol{\gamma}) = m(\boldsymbol{\gamma}; \boldsymbol{\theta}) / u(\boldsymbol{\gamma}; \boldsymbol{\xi}) \tag{4.1}$$

où la probabilité d'observer $\boldsymbol{\gamma}$ est $m(\boldsymbol{\gamma}; \boldsymbol{\theta})$ selon (3.3), étant donné qu'une paire d'enregistrements sélectionnée aléatoirement dans Ω appartient à M; sinon à $u(\boldsymbol{\gamma}; \boldsymbol{\xi})$, donné de la même manière par (3.3) en ayant recours aux paramètres ξ_k au lieu de θ_k. Un algorithme itératif d'une CEM non supervisée est fourni ci-dessous.

I. Soit $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_K^{(0)})$ et $n_M^{(0)} = |M_1|$, où M_1 est l'ensemble maximal de CEM.

II. Pour la t^{e} itération, supposons que $g_{ab}^{(t)} = 1$ si $(a, b) \in M^{(t)}$ et 0 autrement.

   i. Actualisons $u(\boldsymbol{\gamma}; \boldsymbol{\xi}^{(t)})$ à l'aide de (4.4), dont nous traiterons ci-dessous, étant donné $\mathbf{g}^{(t)} = \{g_{ab}^{(t)} : (a, b) \in \Omega\}$, et calculons :

$$\theta_k^{(t)} = \frac{1}{|M^{(t)}|} \sum_{(a,b) \in \Omega} g_{ab}^{(t)} \, \mathbb{I}(\gamma_{ab,k} = 1), \tag{4.2}$$

   qui maximise D_M dans (3.4) pour $u(\boldsymbol{\gamma}; \boldsymbol{\xi}^{(t)})$, $M^{(t)} = \{(a, b) \in \Omega : g_{ab}^{(t)} = 1\}$ et $|M^{(t)}| = \sum_{(a,b) \in \Omega} g_{ab}^{(t)}$ donnés. Une fois $\boldsymbol{\theta}^{(t)}$ et $\boldsymbol{\xi}^{(t)}$ obtenus, on peut mettre $n_M^{(t)} = \sum_{\gamma} n(\boldsymbol{\gamma}) \, \hat{g}^{(t)}(\boldsymbol{\gamma})$ à jour, où

$$\hat{g}^{(t)}(\boldsymbol{\gamma}) \equiv \hat{g}(\boldsymbol{\gamma}; \boldsymbol{\theta}^{(t)}, \boldsymbol{\xi}^{(t)}) = \min \left\{ \frac{|M^{(t)}| \, r^{(t)}(\boldsymbol{\gamma})}{|M^{(t)}| (r^{(t)}(\boldsymbol{\gamma}) - 1) + n}, 1 \right\}$$

$$r^{(t)}(\boldsymbol{\gamma}) \equiv r(\boldsymbol{\gamma}; \boldsymbol{\theta}^{(t)}, \boldsymbol{\xi}^{(t)}) = \frac{m(\boldsymbol{\gamma}; \boldsymbol{\theta}^{(t)})}{u(\boldsymbol{\gamma}; \boldsymbol{\xi}^{(t)})}.$$

ii. Pour $\boldsymbol{\theta}^{(t)}, \boldsymbol{\xi}^{(t)}$ et $n_M^{(t)}$ donnés, nous constatons l'ensemble de CEM $M^{(t+1)} = \left\{(a,b) \in \Omega : g_{ab}^{(t+1)} = 1\right\}$ de telle sorte que $\left|M^{(t+1)}\right| = n_M^{(t)}$ par déduplication d'ordre décroissant de $r^{(t)}(\boldsymbol{\gamma}_{ab})$ sur Ω. Il maximise l'entropie $Q^{(t)}(\mathbf{g})$:

$$Q^{(t)}(\mathbf{g}) \equiv Q\left(\mathbf{g} \mid \boldsymbol{\psi}^{(t)}\right) = \frac{1}{n_M^{(t)}} \sum_{(a,b)\in\Omega} g_{ab} \log r^{(t)}(\boldsymbol{\gamma}_{ab}), \tag{4.3}$$

en fonction de $\mathbf{g}$.

III. Répétons jusqu'à $n_M^{(t)} = n_M^{(t+1)}$ ou $\left\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t+1)}\right\| < \epsilon$, où ϵ est une petite valeur positive.

Une propriété de convergence théorique de l'algorithme proposé et sa validation sont présentées dans les documents supplémentaires.

Il convient de mentionner que, dans la mesure où $\Omega = M \cup U$ est hautement déséquilibrée, où la prévalence de $g_{ab} = 1$ est très proche de 0, nous pourrions simplement ne pas tenir compte des contributions de M et utiliser :

$$\hat{\xi}_k = \frac{1}{n} \sum_{(a,b)\in\Omega} \mathbb{I}\left(\gamma_{(ab,k)} = 1\right) \tag{4.4}$$

dans le cadre du modèle (3.3) de $u(\boldsymbol{\gamma}; \boldsymbol{\xi})$, auquel cas il n'y aurait pas de mise à jour de $u(\boldsymbol{\gamma}; \boldsymbol{\xi}^{(t)})$. D'autres possibilités d'estimer $u(\boldsymbol{\gamma}; \boldsymbol{\xi})$ feront l'objet d'une discussion à la section 5.2.

Le tableau 4.1 présente un aperçu de la CEM aux fins de couplage d'enregistrements dans l'environnement supervisé ou non supervisé. Dans l'environnement supervisé, nous observons $\boldsymbol{\gamma}$ pour les paires d'enregistrements correspondantes dans M, de telle sorte que la probabilité $m(\boldsymbol{\gamma})$ puisse être estimée directement à partir d'eux. Alors que dans le cas d'une CEM dans l'environnement non supervisé, nous ne pouvons pas séparer l'estimation de $m(\boldsymbol{\gamma})$ et n_M.

Tableau 4.1
Classification par entropie maximale pour un couplage d'enregistrements dans un environnement supervisé ou non supervisé

	environnement supervisé	**Non supervisé**
$\boldsymbol{\Omega = M \cup U}$	**Observé**	**Non observé**
Ratio de probabilité	$r_q(\boldsymbol{\gamma})$ généralement applicable $r(\boldsymbol{\gamma})$ étant donné $S(M) \subseteq S(U)$	$r(\boldsymbol{\gamma})$ généralement en supposant $S(M) = S(U)$
Modèle de $\boldsymbol{\gamma}$	Multinomial dans le cas de résultats de comparaison discrets uniquement Directement ou au moyen de variables clés et d'erreurs de mesure	
Ensemble de CEM	Guidé par TFC et TCM Nécessite en plus l'estimation de n_M	
Estimation	$m(\boldsymbol{\gamma}; \boldsymbol{\theta})$ de $\boldsymbol{\gamma}_M$ dans Ω n_M selon (3.7) hors de Ω	$m(\boldsymbol{\gamma}; \boldsymbol{\theta})$ et n_M selon (3.7) et (3.9) conjointement

4.2 Taux d'erreur

Une CEM aux fins de couplage d'enregistrements devrait généralement être guidée par les taux d'erreur, le TFC et le TCM, sans être limitée à l'estimation de n_M.

Il convient de mentionner que $\{\hat{g}_{ab}: (a,b) \in \hat{M}\}$ de tout ensemble de CEM $\hat{M}$ font partie des plus grands dans Ω, car la CEM suit l'ordre décroissant de $\hat{r}_{ab}$, à l'exception de la déduplication nécessaire lorsqu'il existe plusieurs paires pour un enregistrement donné. Pour exercer un plus grand contrôle sur le TFC, supposons que ψ est le TFC le plus grand et considérons la procédure de bissection suivante.

i. Choisissons une valeur de seuil c_ψ et formons l'ensemble de CEM correspondant $\hat{M}(c_\psi)$, où $\hat{r}_{ab} \geq c_\psi$ pour tout $(a,b) \in \hat{M}(c_\psi)$.
ii. Calculons le TFC estimé de l'ensemble de CEM obtenu $\hat{M}$ sous la forme :

$$\hat{\psi} = \frac{1}{|\hat{M}|} \sum_{(a,b) \in \hat{M}} (1 - \hat{g}_{ab}). \tag{4.5}$$

Si $\hat{\psi} > \psi$, alors on augmente c_ψ; si $\hat{\psi} < \psi$, alors on réduit c_ψ.

L'itération entre les deux étapes mènerait finalement à une valeur de c_ψ donnant un $\hat{\psi}$ se rapprochant autant que possible de ψ, pour un ratio de probabilité donné $\hat{r}(\gamma)$.

L'ensemble de CEM final $\hat{M}$ peut être choisi en fonction de l'estimation du TFC correspondante $\hat{\psi}$. Il est également possible de prendre en compte l'estimation du TCM fournie par :

$$\hat{\tau} = 1 - \sum_{(a,b) \in \hat{M}} \hat{g}_{ab} / \hat{n}_M \tag{4.6}$$

où $\hat{n}_M$ est donné par un algorithme de CEM non supervisée. Il convient de souligner que si $|\hat{M}| = \hat{n}_M$, alors $\hat{\psi} = \hat{\tau}$, mais pas si $\hat{M}$ est guidé par une valeur cible donnée du TFC ou du TCM.

À la section 6.2, nous étudierons la performance des ensembles de CEM guidés par les taux d'erreur au moyen de simulations.

5. Discussion

Nous présentons et comparons ci-après deux autres approches dans l'environnement non supervisé, notamment les façons dont certains de leurs éléments peuvent être intégrés à l'approche de CEM. D'autres approches moins pratiques sont présentées dans la documentation supplémentaire.

5.1 Approche classique

Nous rappelons les problèmes I et II de l'approche classique mentionnée à la section 2.

D'un point de vue pratique, il est possible de traiter le problème I par une méthode de déduplication de l'ensemble M^* des paires d'enregistrements classées, où $\hat{r}(\gamma_{ab})$ est supérieur à la valeur de seuil pour toutes les valeurs $(a,b) \in M^*$. À titre de « progrès par rapport à des méthodes antérieures d'attribution ponctuelle », Jaro (1989) choisit l'ensemble couplé $\hat{M}^* \subseteq M^*$, qui maximise la somme des $\log \hat{r}(\gamma_{ab})$ soumise à la contrainte d'un couplage univoque. Puisque $\hat{g}_{ab}$ est une fonction monotone de $\hat{r}(\gamma_{ab})$, cela

revient à choisir $\hat{M}^*$ qui maximise le nombre attendu de correspondances qu'il contient, défini comme suit :

$$n_M^* = \sum_{(a,b)\in \hat{M}^*} \hat{g}_{ab}.$$

Toutefois, n_M^* n'est toujours pas associé aux probabilités de faux couplages et de non-couplages définis par (2.1). Comme nous l'illustrons ci-dessous, cela ne permet pas non plus de contrôler directement les erreurs des $\hat{M}^*$ couplés.

Prenons le couplage de deux fichiers contenant 100 enregistrements chacun. Supposons que la méthode d'attribution de Jaro donne $\left|\hat{M}^*\right| = 100$ à une occasion, où 80 couplages correspond à $\hat{g}_{ab} \approx 1$ et 20 couplages correspond à $\hat{g}_{ab} \approx 0{,}75,$ de telle sorte que $n_M^* \approx 95.$ Supposons que cette méthode donne 90 couplages avec $\hat{g}_{ab} \approx 1$ et 10 couplages avec $\hat{g}_{ab} \approx 0{,}5$ à une autre occasion, où $n_M^* \approx 95.$ Clairement, n_M^* ne contrôle pas directement les erreurs de couplage dans $\hat{M}^*.$ De plus, il n'existe aucune raison convaincante d'accepter 100 couplages dans ces deux occasions, simplement parce que 100 couplages univoques sont possibles.

En formant l'ensemble de CEM, on traite directement le problème I, en fonction du concept d'entropie maximale, qui est pertinent dans de nombreux domaines de l'étude scientifique. Sa mise en œuvre est simple et rapide pour des ensembles de données volumineux. Les taux d'erreur estimés TFC (4.5) et TCM (4.6) sont directement définis pour un ensemble de CEM donné.

Le problème II concerne l'estimation de paramètres. Comme nous l'avons expliqué auparavant, appliquer l'algorithme EM fondé sur la fonction objective (2.2), comme le proposent Winkler (1988) et Jaro (1989) *n'est pas* une approche valide d'estimation du maximum de vraisemblance (EMV). Cette procédure WJ peut facilement se comparer à celle fournie à la section 4.1, où les deux adoptent le même modèle (3.3) et le même estimateur de $u(\boldsymbol{\gamma};\boldsymbol{\xi})$ au moyen de $\hat{\xi}_k$ fournie par (4.4). Il devient alors clair que la même formule est utilisée pour mettre à jour $n_M^{(t)}$ dans chaque itération, mais qu'une formule différente est utilisée pour :

$$\theta_k^{(t)} = \frac{1}{n_M^{(t)}} \sum_{(a,b)\in\Omega} \hat{g}_{ab}^{(t)} \gamma_{ab,k} \tag{5.1}$$

où le numérateur est dérivé de *toutes* les paires dans $\Omega,$ alors que $\theta_k^{(t)}$ selon (4.2) repose uniquement sur les paires de l'ensemble de CEM $M^{(t)}.$ Nous remarquons que ces deux méthodes diffèrent uniquement dans l'environnement non supervisé, mais qu'elles deviendraient identiques dans l'environnement supervisé, où il est possible d'utiliser la valeur binaire observée g_{ab} plutôt que la valeur fractionnelle estimée $\hat{g}_{ab}.$

Ainsi, nous pouvons intégrer la procédure WJ en tant que variation de l'algorithme d'une CEM non supervisée, où les formules (5.1) et (4.4) sont spécifiquement choisies. C'est pourquoi cette méthode peut fournir des estimations paramétriques raisonnables dans de nombreuses situations, malgré l'idée erronée qu'il s'agit de l'EMV. Des simulations serviront plus tard à comparer empiriquement les deux formules (4.2) et (5.1) pour $\theta_k^{(t)}.$

5.2 Approche de l'estimation du maximum de vraisemblance

Nous dérivons, ci-dessous, un autre estimateur de ξ_k selon l'approche de maximum de vraisemblance, qui peut être intégrée à l'algorithme de CEM proposé, au lieu de (4.4). Cela nécessite un modèle des variables clés qui explique les hypothèses des erreurs de variables clés. Soit z_k la k^{e} variable clé prenant la valeur de $1,\ldots,D_k$. Copas et Hilton (1990) envisagent un processus aléatoire non informatif de génération dans le cadre duquel le z_k observé peut prendre la vraie valeur malgré la perturbation. Copas et Hilton (1990) démontrent que le modèle aléatoire est plausible dans l'environnement d'apprentissage supervisé fondé sur des ensembles de données étiquetés.

Nous adaptons ce modèle aléatoire à l'environnement non supervisé comme suit. Tout d'abord, pour tout $(a,b)\in M$, soit $\alpha_k = \Pr\left(e_{ab,k}=1\right)$, où $e_{ab,k}=1$ si la paire associée de variables clés est soumise à *toute forme de perturbation* pouvant éventuellement entraîner une non-correspondance de la k^{e} variable clé, et $e_{ab,k}=0$ autrement. Soit :

$$\theta_k = \left(1-\alpha_k\right) + \alpha_k \sum_{d=1}^{D_k} m_{kd}^2 = 1-\alpha_k\left(1-\sum_{d=1}^{D_k} m_{kd}^2\right)$$

où nous supposons que α_k doit être positif pour certaines valeurs de $k=1,\ldots,K$, et

$$m_{kd} = \Pr\left(z_{ik}=d \mid g_{ab}=1, e_{ab,k}=1\right) = \Pr\left(z_{ik}=d \mid g_{ab}=1, e_{ab,k}=0\right)$$

pour $i=a$ ou b. Ensuite, pour tout enregistrement i dans A ou B, supposons que $\delta_i=1$, s'il a une correspondance dans l'autre fichier et que $\delta_i=0$ autrement. Étant donné $\delta_i=0$, avec ou sans perturbation, supposons $\Pr\left(z_{ik}=d \mid \delta_i=0\right)=u_{kd}$. Nous obtenons $\beta_{kd} := m_{kd}\equiv u_{kd}$ si δ_i *n'est pas informatif*. Une hypothèse légèrement moins stricte est que δ_i soit uniquement non informatif dans l'un des deux fichiers. Pour faire preuve de plus de résilience par rapport à un échec éventuel, nous pouvons supposer que m_{kd} se vérifie pour tous les enregistrements dans *le plus petit* fichier et permettre à u_{kd} de différer pour les enregistrements où $\delta_i=0$ dans *le plus grand* fichier. Supposons que $n_A < n_B$. Soit :

$$p = \Pr\left(\delta_b=1\right) = E\left(n_M\right)/n_B = n_A\pi$$

la probabilité qu'un enregistrement dans B ait une correspondance dans A. Nous pouvons supposer que $\mathbf{z}_A=\left\{\mathbf{z}_a : a\in A\right\}$ est indépendant dans A, ce qui donne :

$$\ell_A = \sum_{a\in A}\sum_{k=1}^{K} \log m_{ak}$$

où $m_{ak}=\sum_{d=1}^{D_k} m_{kd}\,\mathbb{I}(z_{ak}=d)$. Le logarithme du rapport de vraisemblance des données complètes en fonction de $\left(\delta_B,\mathbf{z}_B\right)$ est :

$$\ell_B = \sum_{b\in B}\delta_b \log\left(p\prod_{k=1}^{K} m_{bk}\right) + \sum_{b\in B}\left(1-\delta_b\right)\log\left(\left(1-p\right)\prod_{k=1}^{K} u_{bk}\right) \tag{5.2}$$

où $m_{bk}=\sum_{d=1}^{D_k} m_{kd}\,\mathbb{I}(z_{bk}=d)$ et $u_{bk}=\sum_{d=1}^{D_k} u_{kd}\,\mathbb{I}(z_{bk}=d)$, selon une hypothèse de valeurs $\left(\delta_b,\mathbf{z}_b\right)$ indépendantes pour toutes les entités dans B.

Selon une modélisation distincte de $\mathbf{z}_A$ et $(\mathbf{z}_B, \delta_B)$, supposons que $\hat{m}_{kd}$ est l'EMV fondée sur ℓ_A, à partir d'où un algorithme EM d'estimation de p et que u_{kd} découle de (5.2) lorsque nous traitons δ_B comme données manquantes. L'estimation est réalisable uniquement si $\{u_{kd}\}$ et $\{m_{kd}\}$ ne sont pas exactement identiques, alors que l'EMV de n_M présente une importante variance, où $\{m_{kd}\}$ et $\{u_{kd}\}$ sont proches, même si ces valeurs ne sont pas exactement égales.

Parallèlement, la proximité de $\{m_{kd}\}$ et $\{u_{kd}\}$ n'influe pas sur l'approche de la CEM, selon laquelle $\hat{n}_M$ est obtenue en résolvant (3.7) avec $\hat{r}(\boldsymbol{\gamma}) = \hat{m}(\boldsymbol{\gamma}) / \hat{u}(\boldsymbol{\gamma})$, auquel cas l'estimation de $\hat{u}(\boldsymbol{\gamma})$ est en effet la plus faible lorsque $\{m_{kd}\} = \{u_{kd}\}$. De plus, il est possible d'intégrer un *algorithme EM de profil*, fondé sur (5.2) étant donné $n_M^{(t)}$, pour mettre $u(\boldsymbol{\gamma}; \boldsymbol{\xi}^{(t)})$ à jour dans l'algorithme de CEM non supervisé de la section 4.1. À la t^{e} itération, lorsque $t \geq 1$, étant donné que $p^{(t)} = n_M^{(t)} / \max(n_A, n_B)$ et que $\hat{m}_{kd}$ est estimé à partir du fichier le plus petit A, nous obtenons $u_{kd}^{(t)}$ grâce à :

$$\xi_k^{(t)} = \left((1 - p^{(t)}) \sum_{d=1}^{D_k} u_{kd}^{(t)} \hat{m}_{kd} + p^{(t)} \left(1 - \frac{1}{n_A} \right) \sum_{d=1}^{D_k} \hat{m}_{kd}^2 \right) \Big/ (1 - p^{(t)} / n_A). \tag{5.3}$$

6. Étude par simulations

6.1 Configuration

Pour étudier la faisabilité pratique de l'algorithme d'une CEM non supervisée aux fins de couplage d'enregistrements, nous menons une étude par simulations fondée sur les ensembles de données énumérés au tableau 6.1, qui sont diffusés par ESSnet-DI (McLeod, Heasman et Forbes, 2011) et offerts gratuitement en ligne. Chaque enregistrement dans un ensemble de données comprend des variables clés synthétiques connexes, qui peuvent être déformées par des valeurs manquantes et des coquilles lors de leur création, de manière à imiter des erreurs réelles (McLeod et coll., 2011).

Tableau 6.1
Description de l'ensemble de données (taille entre parenthèses)

Ensemble de données	Description
Recensement (25 343)	Ensemble de données fictif représentant certaines observations d'un recensement décennal.
SIC (24 613)	Observations fictives d'un système d'information sur la clientèle (SIC), données administratives combinées provenant des systèmes sur les taxes et les avantages.
DRP (24 750)	Observations fictives provenant de données du registre des patients (DRP) des services de santé nationaux.

Nous tenons compte des clés de couplage du nom de famille (PERNAME1), du prénom (PERNAME2), du sexe (SEX) et de la date de naissance (DOB). Pour modéliser les variables clés, nous divisons la date de naissance en trois variables clés (le jour [DAY], le mois [MON], l'année [YEAR]). Pour les variables de texte, comme le nom de famille et le prénom, nous les divisons en quatre variables clés à l'aide de l'algorithme de codage Soundex (Copas et Hilton, 1990, page 290), qui réduit un nom à un

code comprenant la première lettre suivie de trois chiffres, par exemple Copas ≡ C120, Hilton ≡ H435. Les douze variables clés du couplage d'enregistrements sont présentées au tableau 6.2.

Tableau 6.2
Douze variables clés disponibles dans les trois ensembles de données

Variable		Description	Nombre de catégories
PERNAME1	1	Première lettre du nom de famille	26
	2	Premier chiffre du code Soundex de nom de famille	7
	3	Deuxième chiffre du code Soundex de nom de famille	7
	4	Troisième chiffre du code Soundex de nom de famille	7
PERNAME2	1	Première lettre du prénom	26
	2	Premier chiffre du code Soundex de prénom	7
	3	Deuxième chiffre du code Soundex de prénom	7
	4	Troisième chiffre du code Soundex de prénom	7
SEX		Homme/Femme	2
DOB	DAY	Jour de naissance	31
	MON	Mois de naissance	12
	YEAR	Année de naissance (1910 à 2012)	103

Nous définissons deux scénarios afin de générer des fichiers de couplage. Nous utilisons la variable d'identification unique (PERSON-ID) pour l'échantillonnage, disponible dans les trois ensembles de données. Nous échantillons respectivement $n_A = 500$ et $n_B = 1\,000$ personnes à partir des DRP et du SIC. Soit p_A la proportion des enregistrements dans le fichier le plus petit (DRP) également sélectionnés dans le plus gros fichier (SIC), avec laquelle nous pouvons faire varier le degré de chevauchement, c'est-à-dire l'ensemble de personnes qui correspondent AB, entre A et B. Nous utilisons $p_A = 0{,}8; 0{,}5$ ou $0{,}3$ selon le scénario.

Scénario I (non informatif)

- Échantillon aléatoire $n_0 = n_B / p_A$ personnes issues du recensement.
- Échantillon aléatoire n_A parmi ces n_0 comme les personnes des DRP, désigné par A.
- Échantillon aléatoire n_B parmi ces n_0 comme les personnes du SIC, désigné par B.

Selon ce scénario, δ_a et δ_b ne sont pas informatifs pour la répartition des variables clés. Pour toute p_A donnée, nous avons $E(n_M) = n_A p_A$ et $\pi = E(n_M)/n_0$, où n_M est le nombre aléatoire de personnes qui correspondent entre les fichiers A et B simulés.

Scénario II (informatif)

- Échantillon aléatoire n_A issu du Recensement $\cap$ DRP $\cap$ SIC, désigné par A dans le DRP.
- Échantillon aléatoire $n_M = n_A p_A$ issu de A comme étant les personnes appariées, désigné par AB.
- Échantillon aléatoire $n_B - n_M$ issu de SIC \ A avec SEX = F, YEAR ≤ 1970 et MON impair, désigné par B_0. Soit $B = AB \cup B_0$ les personnes échantillonnées de SIC.

Dans ce scénario, la répartition des variables clés est la même dans A, que $\delta_a = 1$ ou non, mais est différente pour les enregistrements $b \in B_0$, ou $\delta_b = 0$. Par conséquent, le scénario II est informatif. Pour toute p_A donnée, $n_M = n_A p_A$ et $\pi = p_A / n_B$ sont fixes.

6.2 Résultats : estimation

Pour l'algorithme de CEM non supervisée fourni à la section 4.1, on peut choisir (4.2) ou (5.1) pour mettre à jour $\theta_k^{(t)}$. De plus, nous pouvons utiliser directement (4.4) pour $\hat{\xi}_k$ ou (5.3) pour mettre à jour $\xi_k^{(t)}$ de façon itérative. En particulier, choisir (5.1) et (4.4) permet d'intégrer en fait la procédure de Winkler (1988) et Jaro (1989) d'estimation des paramètres. Il convient de mentionner que l'approche de CEM diffère toujours de celle de Jaro (1989), relativement à la formation de l'ensemble couplé $\hat{M}$.

Le tableau 6.3 affiche une comparaison de la performance de l'algorithme de CEM non supervisée, à l'aide de différentes formules pour $\theta_k^{(t)}$ et $\xi_k^{(t)}$, pour lesquelles la taille de $\hat{M}$ est égale à l'estimation correspondante $\hat{n}_M$. Nous incluons en outre $\hat{\theta}_k = n_M(1;k)/n_M$ estimé directement à partir des paires appariées dans M, comme si M était disponible pour l'apprentissage supervisé, ainsi que (4.4) pour $\hat{\xi}_k$. Les paramètres et les taux d'erreur réels sont fournis en plus de leurs estimations.

Tableau 6.3
Paramètres et moyennes de leurs estimations, moyennes des taux d'erreur et leurs estimations, pour 200 simulations. Médiane de l'estimation de n_M donnée par $\tilde{n}_M$

Scénario I											Scénario II									
Paramètre		Formule		Estimation							Paramètre		Formule		Estimation					
π	$E(n_M)$	$\theta_k^{(i)}$	$\xi_k^{(i)}$	$\hat{\pi}$	$\hat{n}_M$	$\tilde{n}_M$	TFC	TCM	$\widehat{\text{TFC}}$	$\widehat{\text{TCM}}$	π	n_M	$\theta_k^{(i)}$	$\xi_k^{(i)}$	$\hat{\pi}$	$\hat{n}_M$	$\tilde{n}_M$	TFC	TCM	$\widehat{\text{TFC}}$	$\widehat{\text{TCM}}$
0,0008	400	$\hat{\theta}_k$	(4.4)	0,00080	400,0	397	0,0264	0,0266	0,0357	0,0357	0,0008	400	$\hat{\theta}_k$	(4.4)	0,00080	398,3	400	0,0230	0,0273	0,0326	0,0326
		(4.2)	(5.3)	0,00082	407,9	405	0,0425	0,0257	0,0509	0,0509			(4.2)	(5.3)	0,00080	401,4	401	0,0305	0,0277	0,0403	0,0403
		(4.2)	(4.4)	0,00083	414,7	407	0,0549	0,0244	0,0620	0,0620			(4.2)	(4.4)	0,00081	405,2	404	0,0379	0,0262	0,0467	0,0467
		(5.1)	(4.4)	0,00081	406,0	405	0,0399	0,0269	0,0503	0,0503			(5.1)	(4.4)	0,00080	401,4	401	0,0316	0,0286	0,0438	0,0438
0,0005	250	$\hat{\theta}_k$	(4.4)	0,00050	251,6	249	0,0340	0,0301	0,0370	0,0370	0,0005	250	$\hat{\theta}_k$	(4.4)	0,00050	249,6	250	0,0284	0,0302	0,0334	0,0334
		(4.2)	(5.3)	0,00052	258,3	255	0,0559	0,0296	0,0533	0,0533			(4.2)	(5.3)	0,00050	251,8	251	0,0383	0,0320	0,0410	0,0410
		(4.2)	(4.4)	0,00053	266,9	256,5	0,0742	0,0277	0,0680	0,0680			(4.2)	(4.4)	0,00052	257,7	253	0,0513	0,0295	0,0516	0,0516
		(5.1)	(4.4)	0,00052	261,7	259	0,0676	0,0305	0,0636	0,0636			(5.1)	(4.4)	0,00051	255,4	253,5	0,0510	0,0336	0,0520	0,0520
0,0003	150	$\hat{\theta}_k$	(4.4)	0,00030	152,3	151	0,0439	0,0356	0,0381	0,0381	0,0003	150	$\hat{\theta}_k$	(4.4)	0,00030	150,5	150	0,0382	0,0355	0,0350	0,0350
		(4.2)	(5.3)	0,00033	165,9	156,5	0,0873	0,0244	0,0620	0,0620			(4.2)	(5.3)	0,00031	153,0	153	0,0559	0,0377	0,0452	0,0452
		(4.2)	(4.4)	0,00041	205,4	161	0,1632	0,0308	0,1251	0,1251			(4.2)	(4.4)	0,00032	158,5	155	0,0708	0,0342	0,0558	0,0558
		(5.1)	(4.4)	0,00054	271,4	169	0,3015	0,0785	0,1639	0,1639			(5.1)	(4.4)	0,00038	189,3	156	0,1414	0,0524	0,0903	0,0903

Comme nous nous y attendions, les meilleurs résultats sont obtenus lorsque le paramètre θ_k est estimé directement à partir des paires correspondantes dans M, c'est-à-dire $\hat{\theta}_k = n_M(1;k)/n_M$, en conjonction avec (4.4) pour $\hat{\xi}_k$, même si $\hat{\xi}_k$ selon (4.4) n'est pas exactement sans biais. Néanmoins, l'estimateur approximatif $\hat{\xi}_k$ peut être amélioré, puisque nous observons que l'estimateur EM de profil fourni par (5.3) est plus efficace pour toutes les configurations, où les deux sont combinés avec (4.2) pour $\theta_k^{(t)}$. Pour ce qui est des deux formules de $\theta_k^{(t)}$ selon (4.2) et (5.1), et les estimateurs n_M et les aux d'erreur TFC et TCM obtenus, nous remarquons ce qui suit :

- Scénario I : Lorsque la taille de l'ensemble apparié M est relativement grande celle-ci correspondant à $p_A = 0{,}8$, il existe seulement de faibles différences en termes de moyenne et de médiane des deux estimateurs de n_M et la différence est seulement de quelques faux couplages en termes d'erreurs de couplage. La figure 6.1 montre que (4.2) donne quelques erreurs plus importantes de $\hat{n}_M$ que (5.1) pour les 200 simulations, lorsque $p_A = 0{,}8$ ou $\pi = 0{,}0008$. À mesure que la taille de l'ensemble apparié M diminue, les moyennes et médianes des estimateurs de n_M découlant de (4.2) et (5.3) sont plus proches des valeurs réelles que celles des autres estimateurs. En particulier, lorsque l'ensemble apparié M est relativement petit, où $\pi = 0{,}0003$, la formule (5.1) donne une estimation nettement moins bonne de n_M en tous points. Alors que cela est partiellement attribuable à l'utilisation de (4.4) plutôt que de (5.3), la majeure partie de la différence se résume au choix de $\theta_k^{(t)}$, ce qui peut être observé dans des comparaisons intermédiaires des résultats en fonction de (4.2) et de (4.4).
- Scénario II : L'utilisation de (4.2) et de (5.3) pour l'algorithme de CEM non supervisée est plus efficace que l'emploi des autres formules en termes à la fois d'estimation de n_M et des taux d'erreur pour les trois tailles de l'ensemble apparié (figure 6.2). Une amélioration relativement plus importante est obtenue en utilisant (4.2) et (5.3) pour des ensembles appariés plus petits.

Les résultats donnent à penser que la taille de l'ensemble apparié tend à influer davantage sur l'algorithme de CEM non supervisée dans le scénario I que dans le scénario II. Choisir (4.2) et (5.3) semble cependant fournir l'estimation la plus robuste de n_M et des taux d'erreur pour la petite taille de l'ensemble apparié M, quel que soit le caractère informatif des erreurs de variables clés. Cela doit être attribuable au fait que le numérateur de $\theta_k^{(t)}$ est calculé dans (5.1) pour toutes les paires dans Ω plutôt que par rapport à l'ensemble de CEM $M^{(t)}$, qui semble plus sensible lorsque le déséquilibre entre M et U est aggravé, alors que les tailles de A et B demeurent fixes.

Figure 6.1 Diagrammes à surfaces de $\hat{n}_M - n_M$ fondés sur 200 échantillons Monte Carlo dans le scénario I.

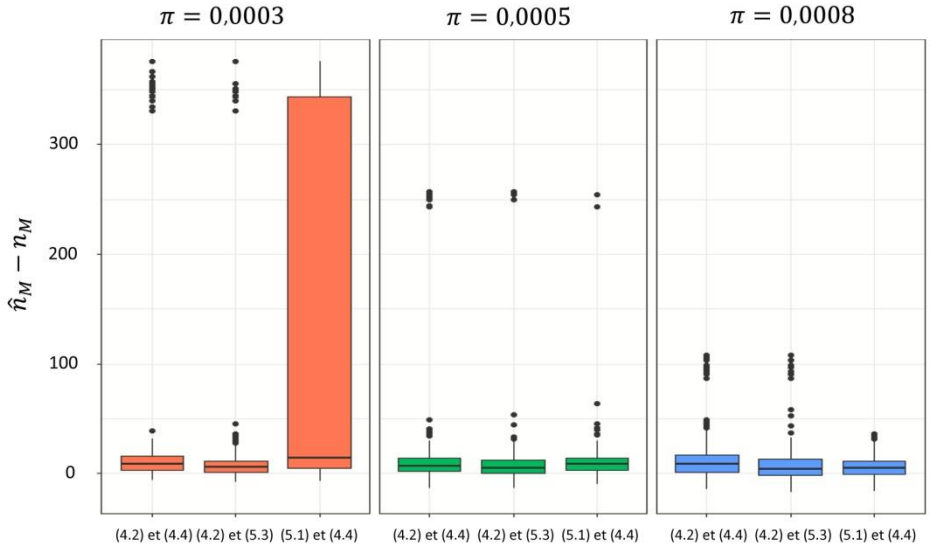
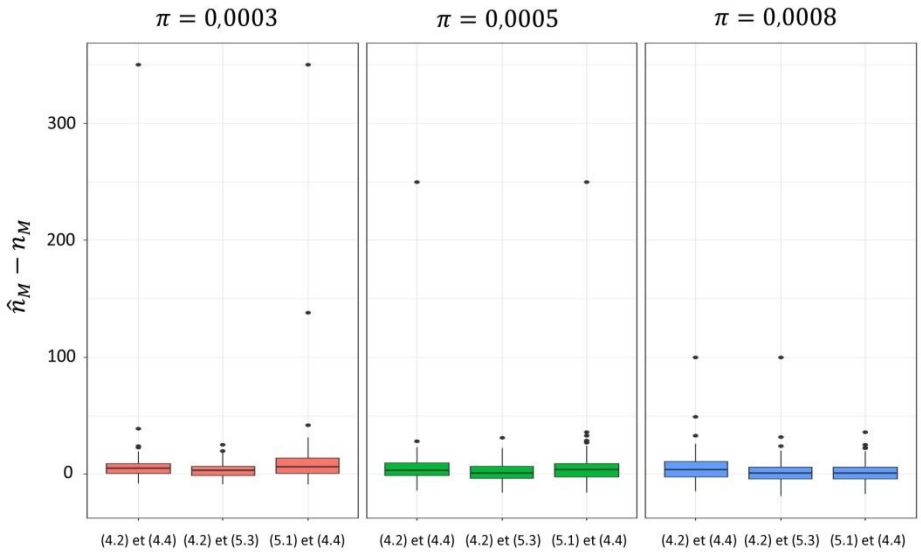


Figure 6.2 Diagrammes à surfaces de $\hat{n}_M - n_M$ fondés sur 200 échantillons Monte Carlo dans le scénario II.



Nous incluons également les autres résultats obtenus pour $p_A = 0{,}2;\ 0{,}15$ et $0{,}1$ dans la documentation supplémentaire. L'estimation $\hat{n}_M$ (ou $\hat{\pi}$) est de moins en moins bonne à mesure que p_A (ou π) diminue. Cela reflète les constats antérieurs; par exemple, Enamorado, Fifield et Imai (2019) ont montré qu'un plus grand degré de chevauchement entre les ensembles de données entraîne de meilleurs résultats de fusion en termes de taux d'erreurs ainsi que d'exactitude de leurs estimations. Sadinle (2017) met également ce problème en évidence. Le couplage d'enregistrements en cas de prévalence extrêmement faible de vraies correspondances est un problème qui doit être étudié plus attentivement de manière indépendante.

6.3 Résultats : ensemble de classification par entropie maximale

Faire en sorte que l'ensemble de CEM $\hat{M}$ corresponde à la taille estimée $\hat{n}_M$ n'est généralement pas une approche raisonnable pour le couplage d'enregistrements. Le couplage d'enregistrements devrait être guidé directement par l'incertitude connexe, c'est-à-dire les taux d'erreur TFC et TCM, selon leurs estimations (4.5) et (4.6), comme le décrit la section 4.2. Il convient de mentionner que cela ne nécessite pas l'estimation de n_M en plus de $r(\boldsymbol{\gamma})$.

Nous avons $\widehat{\text{TFC}} = \widehat{\text{TCM}}$ au tableau 6.3, car $|\hat{M}| = \hat{n}_M$ dans ce cas. Nous pouvons observer qu'ils reflètent le TFC plus étroitement que le TCM, en particulier lorsque $\hat{n}_M$ est estimé à l'aide des formules (4.2) et (5.3). Cela n'est pas très étonnant. Prenons par exemple l'ensemble maximal de CEM M_1 comprenant les paires dont les variables clés correspondent de manière complète et unique. Sous réserve de variables clés assez riches, comme dans la configuration proposée ici, nous pouvons nous attendre à ce que le TFC de M_1 soit faible, de telle sorte que même une estimation naïve $\widehat{\text{TFC}} = 0$ n'est probablement pas trop fautive. Parallèlement, le TCM réel présente une fourchette bien plus large d'une application à une autre, car la différence entre n_M et $|M_1|$ est déterminée par l'étendue des erreurs de variables clés, de telle sorte que l'estimation du TCM dépend de manière plus importante de celle de n_M. La situation est

similaire pour tout ensemble de CEM au-delà de M_1, tant que $\hat{g}_{ab}$ demeure très élevé pour tout $(a,b) \in \hat{M}$.

Le tableau 6.4 présente la performance de l'ensemble de CEM à l'aide de la procédure de bissection décrite à la section 4.2, pour les mêmes configurations qu'au tableau 6.3. Nous utilisons uniquement (4.2) pour $\theta_k^{(t)}$ et (5.3) pour $\xi_k^{(t)}$, afin d'obtenir la valeur $\hat{n}_M$ correspondante. Nous laissons le TFC cible être $\psi = 0{,}05$ ou $0{,}03$, lorsque ce dernier est clairement inférieur au TFC réel de $\hat{M}$ de taille $\hat{n}_M$ (tableau 6.3), en particulier lorsque la prévalence est relativement faible (à $\pi = 0{,}0003$) dans l'un ou l'autre des scénarios. Le tableau 6.4 présente les taux réels obtenus (TFC et TCM) et leurs estimations.

Tableau 6.4
Paramètres et moyennes de leurs estimations, moyennes des taux d'erreur et leurs estimations, pour 200 simulations, $n = |\Omega| = n_A n_B$

Scénario I									Scénario II								
Paramètre		Cible TFC	Estimation						Paramètre		Cible TFC	Estimation					
π	$E(n_M)$		$\hat{n}_M$	$\lvert\hat{M}\rvert/n$	$\lvert\hat{M}\rvert$	TFC	TCM	$\widehat{\text{TFC}}$	$\widehat{\text{TCM}}$	π	n_M		$\hat{n}_M$	$\lvert\hat{M}\rvert/n$	$\lvert\hat{M}\rvert$	TFC	TCM	$\widehat{\text{TFC}}$	$\widehat{\text{TCM}}$
0,0008	400	0,05	407,9	0,00080	401,9	0,0313	0,0280	0,0393	0,0527	0,0008	400	0,05	401,4	0,00080	397,8	0,0239	0,0294	0,0337	0,0418
		0,03		0,00079	395,0	0,0196	0,0328	0,0271	0,0568			0,03		0,00079	393,1	0,0164	0,0334	0,0256	0,0451
0,0005	250	0,05	258,3	0,00050	251,9	0,0396	0,0326	0,0385	0,0576	0,0005	250	0,05	251,8	0,00050	248,6	0,0305	0,0361	0,0328	0,0447
		0,03		0,00049	246,7	0,0246	0,0374	0,0264	0,0650			0,03		0,00049	245,2	0,0226	0,0416	0,0245	0,0497
0,0003	150	0,05	165,9	0,00031	153,4	0,0533	0,0403	0,0389	0,0783	0,0003	150	0,05	153,0	0,00030	150,1	0,0445	0,0443	0,0333	0,0514
		0,03		0,00030	149,3	0,0355	0,0483	0,0256	0,0905			0,03		0,00029	147,4	0,0322	0,0489	0,0238	0,0588

Nous pouvons constater que l'algorithme de CEM guidé par le TFC donne l'ensemble de CEM $\hat{M}$, dont la taille $|\hat{M}|$ est proche du n_M réel pour toutes les configurations. En effet, dans le scénario I, la moyenne de $|\hat{M}|$ est plus proche de n_M que la moyenne (ou médiane) de $\hat{n}_M$ pour toutes les simulations, ce qui découle directement de l'estimation des paramètres, en particulier lorsque l'ensemble de correspondances est relativement petit (à $\pi = 0{,}0003$) et que la performance de $\hat{n}_M$ est la plus sensible. En d'autres termes, le fait que $|\hat{M}|$ diffère de l'estimation $\hat{n}_M$ n'est pas nécessairement préoccupant pour l'algorithme de CEM guidé en ciblant le TFC.

Pour estimer le TCM avec (4.6), on peut soit utiliser $|\hat{M}|$ comme estimation de n_M, soit $\hat{n}_M$ à partir de l'estimation de paramètres basée sur (4.2) et (5.3). Dans le premier cas, nous obtiendrions $\widehat{\text{TCM}} = \widehat{\text{TFC}}$. Ce $\widehat{\text{TCM}}$ n'est pas déraisonnable en termes absolus puisque $|\hat{M}|$ est proche de n_M dans ce cas, comme nous pouvons le voir en comparant la moyenne de $\widehat{\text{TFC}}$ avec celle du TCM réel du tableau 6.4. Toutefois, il présente un inconvénient *a priori*, en ce sens qu'il diminue à mesure que le TFC cible diminue, même s'il est probable que nous omettions plus de correspondances réelles lorsque davantage de couplages sont exclus de l'ensemble de CEM $\hat{M}$. Utiliser $\hat{n}_M$ directement à partir de l'estimation des paramètres est logique de ce point de vue, puisque le n_M réel doit demeurer le même, quel que soit le TFC cible. Toutefois, l'estimateur $\widehat{\text{TCM}}$ pourrait alors devenir moins fiable du fait de la prévalence relativement faible π, où $\hat{n}_M$ pourrait être sensible dans de telles situations.

En résumé, l'estimation du TFC tend à être plus fiable que celle du TCM, en particulier si la prévalence π est relativement faible dans sa fourchette théorique $0 < \pi \leq \min(n_A, n_B)/n$. Les recommandations suivantes pour un couplage d'enregistrements non supervisé sont donc justifiées.

- Lors de la formation de l'ensemble de CEM $\hat{M}$ en fonction de l'incertitude du couplage, l'option de se fier au TFC estimé par (4.5) est plus robuste.
- L'estimation du TCM donnée par (4.6), dérivée de l'estimation des paramètres $\hat{n}_M$ basée sur (4.2) et (5.3), fournit une mesure supplémentaire de l'incertitude. Cependant, il convient de prendre en compte que cette mesure peut être sensible lorsque la prévalence π est relativement faible.
- Entre deux valeurs cibles du TFC, $\psi < \psi'$, nous pouvons prêter davantage attention à l'estimation de correspondances manquantes supplémentaires dans $\hat{M}(\psi)$ par rapport à $\hat{M}(\psi')$, fournie par

$$\sum_{(a,b) \in \hat{M}(\psi')} \hat{g}_{ab} - \sum_{(a,b) \in \hat{M}(\psi)} \hat{g}_{ab} = \sum_{(a,b) \in \hat{M}(\psi') \setminus \hat{M}(\psi)} \hat{g}_{ab} .$$

7. Observations finales

Nous avons élaboré une approche de classification par entropie maximale aux fins de couplage d'enregistrements. Cette approche fournit un cadre de travail de couplage d'enregistrements probabiliste unifié à la fois dans des environnements supervisés et non supervisés, au sein desquels un ensemble de classification cohérent de couplages est explicitement choisi relativement à l'incertitude connexe. La formulation théorique permet d'atténuer certains défauts persistants des approches classiques. De plus, l'algorithme de CEM proposé est entièrement automatique, contrairement à l'approche classique qui nécessite généralement un examen manuel afin de résoudre les cas indécis.

Un problème important nécessitant des recherches supplémentaires concerne l'estimation de paramètres pertinents dans le modèle d'erreurs de variables clés qui cause des difficultés en matière de couplage d'enregistrements. Tout d'abord, comme cela a été souligné plus tôt, traiter le couplage d'enregistrements comme un problème de classification permet d'étudier de nombreuses techniques modernes d'apprentissage automatique. Un défi clé à ce sujet est le fait que les différentes paires d'enregistrements ne sont pas des « unités » distinctes, de sorte que toute technique puissante d'apprentissage supervisé doit être adaptée à l'environnement non supervisé, au sein duquel il est impossible d'estimer les paramètres pertinents en fonction des correspondances et des non-correspondances réelles, y compris le nombre d'entités appariées. Ensuite, le modèle d'erreurs de variables clés ou de résultats de comparaison peut être amélioré. Une fois ces difficultés résolues ensemble, d'autres améliorations de l'estimation des paramètres peuvent être apportées si tout va bien, ce qui serait bénéfique à la fois à la classification de l'ensemble de couplages et à l'évaluation de l'incertitude connexe.

Les diverses formes possibles d'erreurs de variables clés informatives sont un autre aspect intéressant à étudier en pratique, dans la mesure où le modèle relatif aux entités appariées d'une façon ou d'une autre diffère de celui des entités non appariées. Des variations pertinentes de l'approche de CEM peuvent devoir être configurées dans différentes situations.

Remerciements

Les auteurs souhaitent remercier le rédacteur associé et les réviseurs pour leurs commentaires constructifs. Les travaux de Jae Kwang Kim sont en partie financés par la subvention MMS n° 1733572 de la Fondation nationale des sciences.

Documentation supplémentaire

Dans les documents supplémentaires (arXiv:2009.14797), nous présentons la propriété de convergence théorique de l'algorithme proposé et certains cas spéciaux des ensembles de CEM aux fins de couplage d'enregistrements et traitons de deux approches moins pratiques pouvant être intégrées à l'algorithme de CEM. Une étude par simulations supplémentaire comprenant de faibles niveaux de chevauchement des fichiers est également présentée.

Bibliographie

Armstrong, J.B., et Mayda, J.E. (1993). Estimation modéliste des taux d'erreur liés au couplage d'enregistrements. *Techniques d'*enquête, 19, 2, 147-158. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1993002/article/14459-fra.pdf.

Berger, A.L., Della Pietra, S.A. et Della Pietra, V.J. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, 22, 39-71.

Binette, O., et Steorts, R.C. (2020). (almost) all of entity resolution. *arXiv preprint arXiv:2008.04443*.

Christen, P. (2007). A two-step classification approach to unsupervised record linkage. Dans *Proceedings of the Sixth Australasian Conference on Data Mining and Analytics*, Citeseer, 70, 111-119.

Christen, P. (2008). Automatic record linkage using seeded nearest neighbour and support vector machine classification. Dans *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 151-159.

Christen, P. (2012). A survey of indexing techniques for scalable record linkage and deduplication. *IEEE Transactions on Knowledge and Data Engineering*, 24(9), 1537-1555.

Copas, J., et Hilton, F. (1990). Record linkage: Statistical models for matching computer records. *Journal of the Royal Statistical Society, Series A, (Statistics in Society)*, 153(3), 287-312.

Enamorado, T., Fifield, B. et Imai, K. (2019). Using a probabilistic model to assist merging of large-scale administrative records. *American Political Science Review*, 113(2), 353-371.

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

Gull, S.F., et Daniell, G.J. (1984). Maximum entropy method in image processing. *IEE Proceedings 131F*, 646-659.

Hand, D., et Christen, P. (2018). A note on using the f-measure for evaluating record linkage algorithms. *Statistics and Computing*, 28(3), 539-547.

Herzog, T.N., Scheuren, F.J. et Winkler, W.E. (2007). *Data Quality and Record Linkage Techniques*. Springer Science & Business Media.

Jaro, M.A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Floride. *Journal of the American Statistical Association*, 84(406), 414-420.

Larsen, M.D., et Rubin, D.B. (2001). Iterative automated record linkage using mixture models. *Journal of the American Statistical Association*, 96(453), 32-41.

McLeod, P., Heasman, D. et Forbes, I. (2011). Simulated data for the on the job training. *Essnet DI*, 70. Accessible à l'adresse http://www.crosportal.eu/content/job-training.

Newcombe, H.B., Kennedy, J.M., Axford, S. et James, A.P. (1959). Automatic linkage of vital records. *Science*, 130(3381), 954-959.

Nguyen, X., Wainwright, M.J. et Jordan, M.I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11), 5847-5861.

Nigam, K., Lafferty, J. et McCallum, A. (1999). Using maximum entropy for text classification. Dans *IJCAI-99 Workshop on Machine Learning for Information Filtering*, Stockholm, Suède, 1, 61-67.

Owen, A., Jones, P. et Ralphs, M. (2015). Large-scale linkage for total populations in official statistics. *Methodological Developments in Data Linkage*, 170-200.

Sadinle, M. (2017). Bayesian estimation of bipartite matchings for record linkage. *Journal of the American Statistical Association*, 112(518), 600-612.

Sarawagi, S., et Bhamidipaty, A. (2002). Interactive deduplication using active learning. Dans *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 269-278.

Steorts, R.C. (2015). Entity resolution with empirically motivated priors. *Bayesian Analysis*, 10, 849-875.

Stringham, T. (2021). Fast Bayesian record linkage with record-specific disagreement parameters. *Journal of Business & Economic Statistics*, 0(0), 1-14.

Tancredi, A., et Liseo, B. (2011). A hierarchical Bayesian approach to record linkage and population size problems. *The Annals of Applied Statistics*, 5(2B), 1553-1585.

Winkler, W.E. (1988). Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage. Dans *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 667-671.

Winkler, W.E. (1993). Improved decision rules in the Fellegi-Sunter model of record linkage. Dans *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 274-270.

Winkler, W.E. (1994). Advanced methods for record linkage. Dans *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 467-472.

Winkler, W.E., et Thibaudeau, Y. (1991). *An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 US Decennial Census*. Citeseer.

Xu, H., Li, X., Shen, C., Hui, S.L. et Grannis, S. (2019). Incorporating conditional dependence in latent class models for probabilistic record linkage: Does it matter? *Annals of Applied Statistics*, 13(3), 1753-1790.

Zhang, G., et Campbell, P. (2012). Data survey: Developing the statistical longitudinal census dataset and identifying its potential uses. *Australian Economic Review*, 45(1), 125-133.