Recueil du Symposium 2016 de Statistique Canada
Croissance de l'information statistique : défis et bénéfices

Aperçu du couplage d'enregistrements de données d'entreprises à Statistique Canada : Comment coupler les enregistrements « non couplables »

Javier Oyarzun et Laura Wile[1]

Résumé

Le mandat de Statistique Canada comprend la production de données statistiques en vue de faire la lumière sur les questions d'actualité touchant les entreprises. Le couplage des enregistrements de données d'entreprises est un aspect important de l'élaboration, de la production, de l'évaluation et de l'analyse de ces données statistiques. Comme le couplage d'enregistrements peut faire intrusion dans la vie privée, Statistique Canada n'y recourt que si l'intérêt public est manifeste et l'emporte sur les inconvénients de l'intrusion. Le couplage d'enregistrements connaît un renouveau déclenché par un usage plus important de données administratives par un grand nombre de programmes statistiques. Le couplage d'enregistrements de données d'entreprises pose de nombreux défis. Par exemple, plusieurs fichiers administratifs ne contiennent pas d'identificateurs communs, les données sont consignées dans des formats non normalisés, certaines données contiennent des erreurs typographiques, les fichiers de données administratives sont habituellement de grande taille, et enfin, l'évaluation de multiples paires d'enregistrements rend les comparaisons absolues difficiles, voire parfois impossibles.

Étant donné l'importance et les défis du couplage d'enregistrements, Statistique Canada a élaboré une norme en vue d'aider les utilisateurs à optimiser leur processus de couplage d'enregistrements de données d'entreprises. Ainsi, ce processus comprend l'exploitation d'une stratégie de groupement des enregistrements qui réduit le nombre de paires d'enregistrements à comparer et à apparier, l'utilisation d'un logiciel interne de Statistique Canada pour procéder à des couplages déterministes et probabilistes, et la création de champs standardisés pour le nom et l'adresse des entreprises dans le Registre des entreprises de Statistique Canada. Le présent article donne un aperçu de la méthode de couplage d'enregistrements de données d'entreprises et examine divers projets économiques qui font appel au couplage d'enregistrements à Statistique Canada, notamment dans les domaines des Comptes nationaux, du commerce international, de l'agriculture et du Registre des entreprises.

Mots-clés : Couplage d'enregistrements, appariement déterministe, Registre des entreprises, partenariat, standardisation.

1. Introduction

Le mandat de Statistique Canada comprend la production de données statistiques pour faire la lumière sur les questions d'actualité touchant les entreprises. Le couplage d'enregistrements de données d'entreprises est un aspect important de l'élaboration, de la production, de l'évaluation et de l'analyse de ces données statistiques. Statistique Canada procède au couplage d'enregistrements de données d'entreprises depuis de nombreuses années, souvent par appariement direct en utilisant le numéro d'entreprise (NE). Statistique Canada effectue aussi de très nombreux couplages d'enregistrements portant sur des données recueillies auprès des particuliers. Ces dernières données peuvent être plus stables que les données d'entreprises, vu que les particuliers gardent généralement le même nom et le même sexe, et la même date de naissance. La nature évolutive des entreprises complexifie l'exécution du couplage d'enregistrements, car les entreprises peuvent changer de nom, fermer leurs portes, les rouvrir, et procéder à des fusions et à des acquisitions. Le nombre de projets de couplage d'enregistrements de données d'entreprises a augmenté récemment à Statistique Canada, en partie parce qu'un plus grand nombre de programmes statistiques utilisent des sources de données administratives (qui peuvent ne pas contenir de NE) et qu'un nombre croissant de programmes existants de Statistique Canada sont tenus d'établir un lien avec le Registre des entreprises (RE) centralisé de l'organisme. En raison de l'importance du couplage d'enregistrements de données d'entreprises et des défis qu'il pose,

[1] Javier Oyarzun, Statistique Canada, Division des méthodes d'enquêtes auprès des entreprises, Immeuble R.-H.-Coats, 100 promenade Tunney's Pasture, Ottawa (Ontario) Canada, K1A 0T6 (javier.oyarzun@canada.ca);
Laura Wile, Statistique Canada, Division des méthodes d'enquêtes auprès des entreprises, Immeuble R.-H.-Coats, 100 promenade Tunney's Pasture, Ottawa (Ontario) Canada, K1A 0T6 (laura.wile@canada.ca).

Statistique Canada s'emploie à élaborer une stratégie *généralisée* de couplage d'enregistrements destinée à faciliter l'exécution de tels couplages. Le présent article expose en détail une proposition de méthodologie de couplage d'enregistrements de données d'entreprises. La section 2 donne des exemples de projets de couplage d'enregistrements de données d'entreprises menés à Statistique Canada. La section 3 présente certaines difficultés que l'on peut rencontrer lorsqu'on essaie de coupler des enregistrements de données d'entreprises. La section 4 décrit la méthodologie proposée de couplage d'enregistrements de données d'entreprises. Enfin, la dernière section expose les plans concernant les travaux à venir.

2. Le couplage d'enregistrements et ses applications à Statistique Canada

À la présente section, nous passons brièvement en revue les projets faisant appel au couplage d'enregistrements de données d'entreprises menés à Statistique Canada et dont sont tirées les idées en vue d'établir une stratégie généralisée de couplage d'enregistrements.

Plusieurs divisions de Statistique Canada apparient des fichiers de données administratives ne contenant pas de numéro d'entreprise (NE) au Registre des entreprises (RE).

- Par exemple, les données sur les biens exportés sont fournies par l'Agence des services frontaliers du Canada (ASFC) et le United States Bureau of Census, ainsi que par d'autres sources. Les déclarations sont utilisées dans le couplage d'enregistrements pour créer et tenir à jour le Registre des exportateurs de Statistique Canada. Étant donné l'absence de NE dans le Registre des exportateurs, la Division du commerce et des comptes internationaux apparie ce registre au RE en utilisant les noms et les adresses des entreprises pour obtenir des renseignements démographiques (voir Auger (2015) ou Byrd (2016) pour des renseignements détaillés). En outre, les fichiers de données sur les brevets obtenus auprès de l'Office de la propriété intellectuelle du Canada, du United States Patent and Trademark Office et de l'Office européen des brevets sont également appariés au RE.

- En 2012, en vue de réduire le nombre de bases de sondage tenues à jour par Statistique Canada, le Registre des fermes (RF) a été transféré dans le RE. Comme le RF ne contenait pas de NE pour toutes les fermes, des travaux exhaustifs de couplage d'enregistrements ont été menés afin de rapprocher les deux registres (Dongmo Jiongo *et coll.*, 2013). Ce rapprochement a révélé une discordance entre le nombre d'entreprises désignées comme appartenant au secteur agricole dans le RE et le nombre de fermes incluses dans le RF. Cette discordance est expliquée principalement par les partenariats non identifiés (voir la section 2.2 pour des renseignements plus détaillés sur les partenariats commerciaux) ou les fermes classées incorrectement. Des couplages d'enregistrements ont été effectués en continu afin de repérer les duplicatas en utilisant le nom d'entreprise, le numéro de téléphone, l'adresse de l'entreprise et des données administratives (fiscales) (Gutoskie, 2016).

Afin d'être constituée en société, une entreprise doit obtenir un numéro d'entreprise auprès de l'Agence du revenu du Canada (ARC). Il existe plusieurs raisons pour lesquelles une « même » entreprise pourrait posséder plusieurs NE dans le RE. Par exemple, une entreprise peut posséder un certain nombre de NE différents pour les besoins de la déclaration de ses données financières, ou l'ARC peut attribuer un nouveau NE à une entreprise si celle-ci subit un changement, tel un changement de propriété (Rollin, 2013). En outre, dans le cas d'une entreprise individuelle particulière qui ne nécessite pas de NE, de multiples entrées pour la « même » entreprise peuvent figurer dans le RE si l'entreprise n'a pas été identifiée correctement dans une structure de partenariat.

- La Division de l'analyse économique (DAE) de Statistique Canada tient à jour une base de données longitudinales sur les entreprises canadiennes appelée Fichier de microdonnées longitudinales des Comptes nationaux. Afin de suivre longitudinalement une entreprise en vue d'analyser la main-d'œuvre, il est impossible de se fier uniquement aux NE, parce qu'une entreprise peut changer de NE sans que son effectif subisse des changements, situation qui est considérée comme une fausse création d'entreprise ou une fausse disparition d'entreprise. Pour obtenir des renseignements sur l'entrée ou la sortie d'une entreprise, on procède à l'heure actuelle à un suivi de la main-d'œuvre (Rollin, 2013). D'autres variables de couplage possibles pourraient être étudiées pour d'autres types d'analyses, en s'appuyant sur le nom de l'entreprise, l'adresse de l'entreprise et la classification industrielle (Système de classification des industries de l'Amérique du Nord). Les travaux réalisés dans le cadre de ce projet faciliteront le processus d'identification des entreprises prédécesseurs et successeurs dans le RE.

- Les données administratives du RE sont mises à jour régulièrement par couplage d'enregistrements en utilisant des fichiers administratifs fournis par l'ARC. En outre, le couplage d'enregistrements peut être utilisé pour évaluer la qualité de la base de sondage afin de déceler et de résoudre les cas de duplication, de surdénombrement ou de sous-dénombrement (Oyarzun et Wile, 2016). Récemment, des études ont été lancées en vue d'améliorer le dépistage des partenariats dans le RE. Un partenariat existe quand deux ou plusieurs particuliers sont partenaires dans la même entreprise individuelle. Aux fins de l'impôt, chaque partenaire doit déclarer la même information financière, ainsi que la part du partenariat qui lui revient. Comme il est expliqué plus haut, les partenariats sous forme d'entreprises individuelles sans NE peuvent être difficiles à repérer. Ne pas réussir à détecter les partenariats peut entraîner un surdénombrement qui donne lieu à une surestimation pour des variables telles que le revenu. Par conséquent, les partenariats doivent être repérés et le couplage d'enregistrements peut aider à compléter cette tâche. Un exemple de cette application du couplage d'enregistrements sera examiné plus en détail à la section 4.

3. Défis posés par le couplage d'enregistrements de données d'entreprises

Quand Fellegi et Sunter (1969) ont formalisé le couplage probabiliste d'enregistrements, la plupart des chercheurs n'avaient pas accès aux données administratives et informatisées qui sont disponibles aujourd'hui. La tendance récente à utiliser des données administratives a entraîné un renouveau du couplage d'enregistrements et de nombreux nouveaux défis. Les défis abordés dans le présent article se répartissent en quatre catégories, à savoir i) le traitement des données administratives, ii) les caractéristiques des noms et adresses des entreprises, iii) la stratégie de standardisation et iv) la détermination des vrais appariements.

3.1 Traitement des données administratives

L'utilisation de fichiers de données administratives à Statistique Canada a augmenté considérablement ces dernières années. Les données administratives peuvent avoir plusieurs usages – par exemple, elles peuvent être utilisées pour la création de bases de sondage, l'échantillonnage, l'imputation, l'estimation et l'analyse. Toutefois, ces données sont souvent dépourvues d'identificateurs communs ou uniques et, s'ils existent, ces identificateurs sont parfois inexacts. En outre, les fichiers de données administratives peuvent contenir des renseignements consignés dans des formats non standardisés ou présenter des erreurs typographiques. En outre, les ensembles de données administratives sont parfois très grands, ce qui rend souvent impossible l'évaluation de chacune des paires potentielles lorsqu'on procède au couplage d'enregistrements. Par conséquent, il est parfois nécessaire d'effectuer un prétraitement supplémentaire des fichiers de données administratives.

3.2 Caractéristiques des noms et des adresses des entreprises

Les noms et les adresses d'entreprises sont souvent disponibles quand on réalise un couplage d'enregistrements. Statistique Canada a effectué des travaux de grande envergure sur les enquêtes-ménages en vue d'établir des règles d'appariement pour les noms et les adresses des particuliers et des ménages. Toutefois, ces règles pourraient ne pas convenir pour les entreprises. Les noms d'entreprises peuvent être moins distincts que les noms de personnes et leur longueur peut être plus variable (par exemple, un nom d'entreprise peut être un acronyme, un simple mot ou contenir 10 mots ou plus). Les noms d'entreprises ont également tendance à contenir des mots communs tels que « ferme » ou « société de portefeuille ». Les adresses posent de nombreuses difficultés également – par exemple, une même entreprise peut posséder diverses adresses susceptibles de donner lieu à plusieurs appariements, ou de nombreuses entreprises peuvent avoir la même adresse, ce qui peut créer de faux appariements (par exemple, des entreprises situées dans la même tour de bureaux). Comme les noms d'entreprises, à l'heure actuelle, les adresses incluses dans le RE ne sont pas standardisées.

3.3 Stratégie de standardisation

La standardisation des noms (parsage) est un processus crucial destiné à rendre les variables plus comparables. À Statistique Canada différents groupes possèdent leurs propres stratégies de standardisation des noms et des adresses

d'entreprises, chacune de ces stratégies ayant des objectifs différents et une qualité variable. Certaines règles fondamentales de la standardisation consistent à convertir les lettres en majuscules, à classer les mots par ordre alphabétique, à éliminer les accents français, à laisser tomber les mots triviaux (par exemple, laisser tomber « compagnie » et « limitée »), à convertir les mots à une orthographe standardisée (par exemple, les noms de province pourraient être standardisés en vue d'utiliser une représentation à deux caractères). Des règles plus complexes que cela sont nécessaires pour standardiser les noms d'entreprises et les adresses d'entreprises. Toutefois, il faut déterminer jusqu'où il faut pousser la standardisation, car celle-ci peut avoir une incidence sur la détermination des appariements résultants.

3.4 Détermination des vrais appariements

Lorsqu'on essaie de coupler des enregistrements, l'étape finale consiste à utiliser une méthode qui créera des paires d'enregistrements selon une approche déterministe[2] ou probabiliste[3]. À Statistique Canada, un logiciel appelé MixMatch (Lachance, 2014) est utilisé pour effectuer les couplages déterministes d'enregistrements et un système généralisé appelé G-Link sert à effectuer les couplages probabilistes d'enregistrements. Pour l'une et l'autre approches, l'utilisateur doit examiner attentivement quelle stratégie de couplage convient. Ainsi, dans le cas de MixMatch, l'utilisateur doit déterminer quels comparateurs de chaînes de caractères et quels paramètres doivent être utilisés. Dans le cas de G-Link, l'utilisateur doit déterminer les poids de couplage qui seront utilisés pour chaque champ de couplage afin de décider si deux entités concordent. Pour les deux outils, l'utilisateur doit envisager l'utilisation du groupement ou mise en bloc (*blocking*)[4] pour réduire le temps de traitement. Enfin, l'utilisateur doit créer une méthode pour attribuer un score d'évaluation de la qualité aux paires qui sont décelées. Parfois, un examen manuel pour évaluer les appariements potentiels est nécessaire. Cette dernière étape peut prendre beaucoup de temps et requiert l'expertise des divisions spécialisées.

4. Méthodologie proposée de couplage des enregistrements de données d'entreprises

En s'appuyant sur des idées dégagées des projets de couplage d'enregistrements de données d'entreprises réalisés à Statistique Canada, la Division des méthodes d'enquêtes auprès des entreprises a proposé une méthodologie de couplage d'enregistrements de données d'entreprises. En théorie, le couplage probabiliste d'enregistrements est la méthode appropriée pour apparier les entreprises. En pratique, cette méthode est souvent gourmande en ressources informatiques pour les grands ensembles de données (en ce qui concerne tant les observations que les variables) tels que le RE. Le produit cartésien qui est nécessaire lorsque l'on adopte une approche probabiliste ne peut tout simplement pas être traité par les ordinateurs actuels. Par contre, la méthode de couplage déterministe d'enregistrements peut être plus praticable et convenir aux besoins de l'utilisateur.

La méthodologie proposée présentée ici a la capacité de générer des appariements reproductibles et tient compte des variables de couplage d'enregistrements indépendamment les unes des autres. Bon nombre de techniques de couplage déterministe d'enregistrements s'appuient sur une série hiérarchique de critères d'appariement et s'arrêtent lorsqu'un appariement a été décelé – par exemple, critère 1 : les unités doivent concorder en ce qui concerne le nom d'entreprise et l'adresse d'entreprise; sinon, critère 2 : les unités doivent concorder en ce qui concerne le nom d'entreprise; sinon, critère 3 : les unités doivent concorder en ce qui concerne l'adresse d'entreprise.
Lorsque la nouvelle méthodologie proposée est appliquée indépendamment aux a) noms d'entreprise, b) adresses d'entreprise et c) données administratives, des étapes similaires mais indépendantes sont réalisées (figure 4-1).

[2] Le *couplage déterministe d'enregistrement* produit des appariements basés sur un ou plusieurs identificateurs qui concordent dans les ensembles de données disponibles.
[3] Le *couplage probabiliste d'enregistrement* s'appuie sur une plus vaste gamme d'identificateurs possibles, et comprend le calcul d'un poids pour chaque identificateur en se basant sur sa capacité à déterminer correctement un appariement ou un non-appariement, puis à utiliser ces poids pour calculer la probabilité que deux enregistrements donnés se rapportent à une même entité.
[4] Le *groupement*, comme la stratification, sert à définir fonctionnellement dans un grand ensemble de données un plus petit ensemble de données sur des individus ayant au moins une caractéristique commune.

Figure 4-1
Étapes de la méthodologie proposée de couplage d'enregistrements

4.1 Standardisation

La standardisation des noms d'entreprises est effectuée à l'aide d'un système généralisé de Statistique Canada appelé G-Code. Une stratégie de parsage, qui résulte de la combinaison de multiples stratégies utilisées par divers programmes à Statistique Canada, est fournie à G-Code. La standardisation des adresses est réalisée en utilisant un système généralisé appelé PCODE. Ce dernier est un autre système généralisé de Statistique Canada qui prend une adresse canadienne, la normalise conformément aux lignes directrices de la Société canadienne des postes et génère un code postal pour l'adresse. À Statistique Canada, le prétraitement des données administratives est effectué par la Division des données administratives (DDA). Après réception des données fournies par l'ARC, la DDA procède à la vérification, à l'imputation, à la détection des valeurs aberrantes et à la détermination des partenariats.

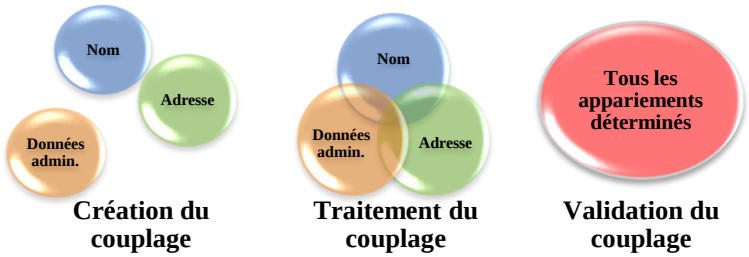
4.2 Appariement

Les appariements d'enregistrements de données d'entreprises sont déterminés en exécutant le couplage d'enregistrements sur le nom d'entreprise, l'adresse d'entreprise, une combinaison des deux ou toute autre variable appropriée. Pour les besoins du présent article, les auteurs apparient les enregistrements en utilisant le nom d'entreprise, l'adresse d'entreprise et les données administratives dans trois processus indépendants. Premièrement, pour déterminer les appariements en utilisant les noms d'entreprises, un couplage déterministe d'enregistrements est exécuté à l'aide de MixMatch (Lachance, 2014). Deuxièmement, des appariements fondés sur la proximité géographique sont créés en utilisant les adresses d'entreprises. Enfin, les appariements en se servant de données administratives sont déterminés au moyen d'une mesure de différence entre des variables communes. Dans l'exemple de la détection des partenariats, les enregistrements possédant de nombreux champs de revenu en commun sont plus susceptibles de représenter un partenariat que ceux ne possédant qu'un seul champ en commun.

4.3 Attribution de scores de similitude

Après l'achèvement de chaque étape de couplage, les trois ensembles d'appariements (c.-à-d. appariements fondés sur les noms d'entreprises, appariements fondés sur les adresses d'entreprises et appariements fondés sur les données administratives) sont fusionnés (voir la figure 4.3-1). Trois scores de similitude sont attribués à chaque appariement, un pour chacune des trois variables utilisées pour déterminer les appariements.

Figure 4.3-1
Combinaison des appariements fondés sur les noms d'entreprises, des appariements fondés sur les adresses et des appariements fondés sur les données administratives



4.3.1 Score sur le nom d'entreprise

Pour déceler les enregistrements en double, le score sur le nom d'entreprise est dérivé de la distance d'édition généralisée (DEG) qui résume le degré de différence entre deux chaînes de caractères. La DEG est une généralisation de la distance d'édition de Levenshtein, qui est une mesure de similarité entre deux chaînes. Cette distance d'édition mesure le nombre de suppressions, d'insertions ou de remplacements de caractères uniques qui sont nécessaires pour transformer une chaîne *a* en une autre chaîne *b*. Par exemple, la chaîne « ballons » comparée à la chaîne « baolln » recevrait une pénalité, traduite par un score plus élevé (S_{NOM}), à cause de la permutation du « o » et du « s » manquant. Mathématiquement, la distance de Levenshtein entre deux chaînes *a* et *b* de longueur $|a|$ et $|b|$, respectivement, peut être calculée comme il suit :

$$S_{NOM} = lev_{a,b}(|a|, |b|)$$

où

$$lev_{a,b}(i,j) = \begin{cases} \max(i,j) & si\ \min(i,j) = 0 \\ min\begin{cases} lev_{a,b}(i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+1_{a_i \neq b_j} \end{cases} & autrement \end{cases}$$

où $i = 1, \ldots, |a|$ et $j = 1, \ldots, |b|$, et $1_{a_i \neq b_j}$ est la fonction indicatrice égale à 0 quand $a_i = b_j$ et égale à 1 autrement, et $lev_{a,b}(i,j)$ est la distance entre les *i* premiers caractères de *a* et les *j* premiers caractères de *b*. Notons que le premier élément de l'argument *min* correspond à une suppression, le second, à une insertion et le troisième représente une concordance ou une discordance. Par conséquent, un score plus faible signifie que les deux chaînes sont davantage similaires. Dans cet exemple, « ballons » et « baolln » obtiendraient un score DEG (S_{NOM}) de 120 (20 pour la permutation du « o » et 100 pour l'insertion de « s »).

4.3.2 Score sur l'adresse d'entreprise

Pour le couplage des adresses d'entreprises, la mesure de distance peut être fondée sur la distance physique entre les deux enregistrements appariés. Cette mesure peut être obtenue en évaluant les coordonnées d'après le système mondial de localisation (GPS). La formule pour mesurer la distance ($S_{ADRESSE}$) entre deux coordonnées GPS (L1 et L2) est la suivante :

$$S_{ADRESSE} = R * C = 6\,371 * C$$

où *R* est le rayon de la Terre (6 371 kilomètres en moyenne), *C* est une fonction de la latitude (ϕ) et de la longitude (λ) de L1 et L2 :

$$C = 2 * atan2(\sqrt{a}, \sqrt{1-a})$$

a est défini comme étant

$$a = \sin^2\left(\frac{L_{1,\phi} - L_{2,\phi}}{2}\right) + cos(L_{1,\phi}) * cos(L_{2,\phi}) * \sin^2\left(\frac{L_{1,\lambda} - L_{2,\lambda}}{2}\right)$$

Par exemple, si deux adresses sont situées dans le même immeuble, la différence entre leurs latitudes et entre leurs longitudes serait égale à 0, rendant les paramètres *a* et *C* égaux à 0. Par conséquent, le score de distance ($S_{ADRESSE}$) entre les deux adresses serait 0.

4.3.3 Score sur les données administratives

Pour attribuer un score à des appariements déterminés en utilisant des données administratives, dans le présent exemple des données financières, on attribue à chaque appariement une mesure de distance qui tient compte de la grandeur des valeurs des données fiscales et du nombre de champs (représenté par *i* dans l'équation qui suit) qui sont communs aux deux unités (x_1 et x_2) :

$$S_{ADMIN} = \frac{1}{Log\left(\sum_i (x_{1,i} + x_{2,i})\right)} * \log(1 + \sum_i \left(x_{1,i} - x_{2,i}\right)^2)$$

Par exemple, un mari et sa femme produisant la même déclaration financière obtiendraient un score de différence (S_{ADMIN}) de 0 parce que la somme de $\left(x_{1,i} - x_{2,i}\right)$ serait nulle.

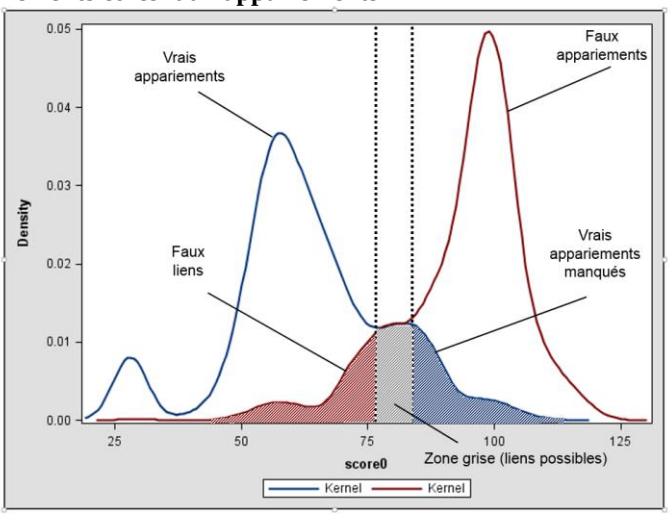
4.4 Production d'un score final

Un score global (ou final) est attribué à chaque appariement en se basant sur les trois scores indépendants (nom, adresse et données administratives) pour faire la distinction entre les appariements forts et les appariements plus faibles. Par exemple, un appariement dont le score pour le nom d'entreprise, l'adresse d'entreprise et les données financières (données administratives) est faible est plus susceptible d'être un vrai appariement qu'un appariement dont le score est élevé. Un score global peut alors être établi en se basant sur les trois scores présentés à la section 4.3 et calculé comme il suit :

$$\mathrm{G_L} = \mathrm{W_1} S_{NOM,L} + \mathrm{W_2} S_{ADRESSE,L} + \mathrm{W_3} S_{ADMIN.L}$$

où S_{NOM} est le score de similitude des noms d'entreprises, $S_{ADRESSE}$ est le score de similitude des adresses d'entreprises et S_{ADMIN} est le score de similitude des données administratives. Chacun de ces scores (nom, adresse et données administratives) doit être pondéré par les valeurs ($\mathrm{W_1}, \mathrm{W_2}, \mathrm{W_3}$) pour que leur contribution au score global, $\mathrm{G_L}$, où *L* représente un appariement (*link*), soit du même ordre de grandeur. Les appariements dont les scores globaux sont les plus faibles devraient être acceptés automatiquement (sous un seuil prédéterminé), ou classés par ordre de priorité (triés), analysés par l'analyste de la division spécialisée et rejetés au besoin. Les méthodologistes, les spécialistes du domaine, ou un système informatique peuvent déterminer les bornes/seuils qu'il convient d'appliquer afin de faire la distinction entre les « vrais appariements » et les « faux appariements ». Cependant, une zone grise entre le seuil supérieur pour la détermination des « faux appariements » et le seuil inférieur pour la détermination des « vrais appariements » doit être utilisée pour déterminer les cas qui pourraient être des « vrais appariements », mais qui pourraient également être des « faux appariements ». Les cas compris dans cette zone grise doivent alors être examinés par les spécialistes du domaine pour déterminer s'ils sont correctement classés comme des « vrais appariements » ou des « faux appariements ». Comme l'illustre la figure 4.4-1, la méthodologie offre à l'utilisateur l'option de modifier le seuil afin de a) obtenir un plus grand nombre de « vrais appariements » (et conséquemment un plus grand nombre de « faux appariements »), ou b) obtenir moins de « vrais appariements » (et conséquemment moins de « faux appariements »).

Figure 4.4-1
Score pour les vrais appariements et les faux appariements



5. Travaux à venir

La création d'une stratégie généralisée de couplage d'enregistrements de données d'entreprises est encore en évolution. Des progrès seront réalisés en consultation avec les comités techniques et directeurs de Statistique Canada, ainsi que les groupes qui procèdent à l'heure actuelle à des couplages d'enregistrements de données d'entreprises. Dans le court terme, des travaux seront menés en vue d'ajouter des noms et des adresses d'entreprises standardisés au RE pour faciliter les travaux de couplage des utilisateurs. En outre, les processus utilisés pour standardiser les noms et les adresses d'entreprises devront être communiqués aux utilisateurs afin de pouvoir standardiser d'autres ensembles de données que le RE. Les travaux se poursuivront en vue d'établir une stratégie générale de couplage d'enregistrements et d'attribution de scores pour déterminer les appariements. De surcroît, une stratégie sera élaborée pour évaluer la qualité des couplages. La stratégie généralisée sera appliquée, par exemple 1) pour créer pour les analyses longitudinales un tableau des entreprises plus complet caractérisé par une plus grande continuité entre les entreprises (prédécesseur/successeur), 2) pour améliorer le processus de détection des partenariats dans le RE, et 3) pour aider les équipes des projets procédant à l'heure actuelle à des couplages d'enregistrements de données d'entreprises à Statistique Canada. L'objectif ultime est d'offrir aux analystes des entreprises un modèle leur permettant de déterminer les appariements avec un indicateur de qualité associé.

Remerciements

Nous tenons à remercier pour leur contribution les personnes suivantes qui ont rendu ce projet possible : Sylvie Auger, Susie Fortier, Kimberly Fyfe, Josh Gutoskie, Paul Hunsberger, Leon Jang, Rob Kozak, Martin Lachance, Pierre Lavallée, Danielle Lebrasseur, William Liu, Chris Mohl, Martin Montreuil, Sylvain Poirier, Alexander Reicker, Anne-Marie Rollin, Abdelnasser Saïdi, Ling Su, Anthony Yeung et Wesley Yung.

Bibliographie

Auger, S. (2015), "Exporter Register", unpublished document, Ottawa, Canada: Statistics Canada.

Byrd, C. (2016), "Exporter Register Process Overview", unpublished document, Ottawa, Canada: Statistics Canada.

Dongmo Jiongo, V., Émond, N. and J. Lynch, "The Migration of Agricultural Surveys to Statistics Canada's Business Register", *Proceedings of Statistics Canada Symposium 2013*, pp. 294 – 298.

Fellegi, I.P., and A.B. Sunter (1969), "A Theory of Record Linkage", *Journal of the American Statistical Association*, 64, pp. 1183-1210.

Gutoskie, J. (2016), "Use of G-CODE and MixMatch for 2016 CEAG Frame", unpublished document, Ottawa, Canada: Statistics Canada.

Lachance, M. (2014), "Useful functionalities for record linkage", *Proceedings of Statistics Canada Symposium 2014*.

Mayda, M. (2015). "Address Matching by the Address Register Team", unpublished document, Ottawa, Canada: Statistics Canada.

Oyarzun, J., Su, L. and D. Lebrasseur (2015). "An Overview of Record Linkage Applications in BSMD", Presented at Statistics Canada's Business Survey Methods Division Technical Committee, February 27, 2015.

Oyarzun, J. and L. Wile, (2016), "Business Register: Agriculture Project", internal document, Ottawa, Canada: Statistics Canada.

Rollin, A.-M. (2013), "Developing a Longitudinal Structure for the National Accounts Longitudinal Microdata File (NALMF)", *Proceedings of Statistics Canada Symposium 2013*, pp. 306-311.

Saïdi, A. (2014). “Overview of Record Linkage at Statistics Canada”, Technical Reported Presented at Statistics Canada’s Advisory Committee on Statistical Methods, May 5-6, 2014.

Statistics Canada (2014). “User Guide for G-Link”, Ottawa: Canada: Statistics Canada.