Nº 11-522-X au catalogue
ISSN 1709-822X

Recueil du Symposium de 2022 de Statistique Canada : Désagrégation des données : dresser un portrait de données plus représentatif de la société

Étude de cas de l'utilisation de Splink : couplage du recensement pour trouver des doublons

par Mary Cleaton, Johanna Hall, Rachel Shipsey, Zoe White et Kristina Xhaferaj

Date de diffusion : le 25 mars 2024

Statistique Canada Statistics Canada

Canada

Recueil du Symposium de 2022 de Statistique Canada
Désagrégation des données : Dresser un portrait de données plus représentatif de la société

Étude de cas de l'utilisation de Splink : couplage du recensement pour trouver des doublons

Mary Cleaton, Johanna Hall, Rachel Shipsey, Zoe White et Kristina Xhaferaj[1]

Résumé

Les auteures ont utilisé le progiciel de couplage probabiliste Splink mis au point par le ministère de la Justice du Royaume-Uni pour relier les données du recensement de l'Angleterre et du pays de Galles à elles-mêmes afin de trouver des réponses en double au recensement. Un grand ensemble étalon-or des doublons confirmés du recensement était disponible, ce qui signifiait que la qualité des résultats de la mise en œuvre de Splink pouvait être assurée. Le présent article décrit la mise en œuvre et les fonctionnalités de Splink, donne des détails sur les configurations et les paramètres que nous avons utilisés pour ajuster Splink à notre projet en particulier, et donne les résultats que nous avons obtenus.

Mots clés : Splink, recensement, couplage probabiliste, Fellegi-Sunter, espérance-maximisation

1. Introduction

1.1 Introduction à Splink

Splink est un progiciel Python à code source ouvert qui a été conçu en 2020 par le ministère de la Justice du Royaume-Uni (MdJ) pour faciliter les projets de couplage de données à grande échelle. Il est accessible au public par GitHub (Linacre, et al., 2022) et est compatible avec AWS Athena et Spark, ce qui permet le couplage des mégadonnées. Splink met en œuvre l'algorithme d'espérance-maximisation (EM) (Dempster, et al., 1977) et la méthode de Fellegi-Sunter de couplage probabiliste des données (Fellegi & Sunter, 1969).

1.1 Doublons du recensement

Dans le Recensement de 2021 d'Angleterre et du pays de Galles (ci-après le « Recensement de 2021 »), on a estimé qu'il y avait environ 420 000 enregistrements en double, ce qui contribuait à un surdénombrement global de 0,96 % (ONS, 2022). Le projet de couplage vise à trouver les enregistrements en double à différents emplacements géographiques. Ce type de duplication dans un recensement se produit pour diverses raisons, notamment quand des personnes :

- emménagent dans un nouveau logement pendant la période du recensement et remplissent un questionnaire de recensement pour leurs deux adresses;
- ont plusieurs résidences et remplissent un questionnaire de recensement pour au moins deux d'entre elles;
- les étudiants ayant une résidence pendant leur semestre et une adresse personnelle qui remplissent un questionnaire de recensement pour les deux;
- des enfants qui passent du temps avec chacun de leurs parents séparés.

1.2 Étalon-or

À la suite du Recensement de 2021, l'ONS a utilisé une stratégie de déclaration simple fondée sur la date de naissance, le prénom et le nom de famille pour repérer les doublons potentiels dans des échantillons des données du recensement. Ces doublons candidats ont été épurés au moyen d'un nouvel algorithme de classification appelé algorithme de vérification automatique (ACA pour *Automatic checking algorithm*) (Shipsey & White, 2020). L'algorithme de vérification automatique divise les doublons candidats en catégories de statut « accepté automatiquement », « rejeté automatiquement » ou « incertain ». Les candidats classés comme incertains sont

[1] Toutes les auteures : Direction de la méthodologie et de la qualité, Office for National Statistics, Angleterre et pays de Galles, Royaume-Uni, NP10 8XG (datalinkage@ons.gov.uk).

examinés manuellement. Cet exercice a produit un ensemble de 3 219 099 enregistrements du recensement échantillonnés, qui étaient classés soit comme étant en double (23 434 enregistrements) soit comme n'étant pas en double (3 195 665 enregistrements). On a rassemblé les enregistrements et leurs doublons (le cas échéant) pour créer un ensemble de données étalon-or.

1.3 Buts et objectifs

Les objectifs du projet étaient doubles. Premièrement, il s'agissait de mettre à l'essai une autre méthode de détection des doublons dans un ensemble de données, en utilisant le Recensement de 2021 comme étude de cas. Deuxièmement, il visait à illustrer Splink, en fournissant une étude de cas de l'utilisation de Splink aux fins de couplage de mégadonnées réelles. L'article décrit notre utilisation de Splink, mais ne mentionne pas davantage la méthode de surdénombrement qui avait été proposée.

2. Splink

Splink est une mise en œuvre de l'algorithme de Fellegi-Sunter (Fellegi & Sunter, 1969), qui est la méthode standard de couplage probabiliste de données utilisée dans le monde entier. Nous avons utilisé la version Splink v2.1.4 et avons fourni des commentaires à l'équipe de développement du ministère de la Justice. Par la suite, des améliorations ont été apportées au progiciel, et une nouvelle version a été publiée : Splink v3.

2.1 Valeurs m et u

La méthode de Fellegi-Sunter est équivalente à un algorithme bayésien naïf. Il faut entrer des valeurs m et u pour chaque variable utilisée pour l'appariement. La valeur m est la probabilité que les valeurs d'une variable de couplage donnée correspondent quand la paire candidate est un appariement vrai. Les valeurs m peuvent être calculées à partir des données d'entraînement, comme un ensemble de données étalon-or, ou à l'aide de l'algorithme EM. Dans Splink, les modèles locaux peuvent servir à calculer les valeurs m si aucun ensemble de données étalon-or n'est disponible.

La valeur u est la probabilité que les valeurs d'une variable de couplage donnée correspondent quand la paire candidate est un non-appariement vrai (c.-.à-.d. une concordance uniquement due au hasard). Les valeurs u peuvent être facilement calculées au moyen de la fonction Splink « estimate_u_values ».

2.2 Instructions de cas

Les instructions de cas SQL sont une série de clauses « if-else » (sinon-si) qui décrivent les états de la concordance pour une variable donnée. Par exemple, les états de la concordance peuvent être une concordance exacte, l'absence de concordance ou une concordance partielle, selon ce que détermine un comparateur de chaînes comme la distance d'édition de Levenshtein ou le score de similarité de Jaro-Winkler. Toutes les instructions de cas SQL utilisées dans cette étude de cas sont présentées dans (Xhaferaj, 2022).

Par exemple, l'instruction de cas pour la variable « prénom » utilise la logique suivante :

- s'il manque un prénom, le niveau de concordance est de -1;
- sinon, si les prénoms concordent exactement, le niveau de concordance est de 3;
- sinon, si le score de Jaro-Winkler entre les prénoms est ≥ 0,88 ou si les prénoms et les deuxièmes prénoms sont transposés, ou si les prénoms et les noms de famille sont transposés, alors le niveau de concordance est de 2;
- sinon, si la distance d'édition de Levenshtein normalisée entre les prénoms est ≥ 0,401, alors le niveau de concordance est de 1;
- sinon, le niveau de concordance est de 0.

2.3 Poids de concordance

Au moyen des valeurs m et u, un poids de concordance est calculé pour chaque niveau de concordance de l'instruction de cas comme étant $log_2(m/u)$. Quand une clause d'instruction de cas est associée à une non-

concordance, le poids de concordance est négatif. Quand une clause d'instruction de cas indique une concordance partielle, le poids de concordance peut être positif ou négatif, selon le poids des autres clauses de cette instruction de cas. Le Tableau 2.3-1 affiche les valeurs m et u et les poids de concordance associés pour l'instruction de cas de la variable « prénom » décrite à la section 2.2.

Tableau 2.3-1
Valeurs *m* et *u* et poids de concordance pour les cinq niveaux de concordance de la variable « prénom »

Niveau de concordance	Valeur m	Valeur u	Poids de concordance
3	0,7798	0,001490522	9,0311
2	0,1393	0,001386703	6,6502
1	0,0780	0,024442757	1,6736
0	0,0030	0,972680018	-8,3524
-1	s.o.	s.o.	0

2.4 Groupage

En théorie, les modèles de couplage probabiliste comparent chaque combinaison potentielle d'enregistrements au moyen d'une jointure cartésienne. La majorité de ces comparaisons sont redondantes, car les enregistrements ont peu en commun. Le groupage est une technique servant à réunir des paires candidates qui ont en commun des caractéristiques, réduisant ainsi l'espace de recherche et le nombre de comparaisons nécessaires. Une série de règles de groupage sert à sélectionner les comparaisons les plus prometteuses. Par exemple, la règle de groupage 1 nécessite un appariement exact entre « prénom » et « nom de famille », plus « deuxième prénom » pour former un appariement ou pour qu'il soit manquant; la règle de groupage 5 nécessite un appariement exact entre région, « initiales du prénom », « initiales du nom de famille », « date de naissance » et « sexe ». Les règles de groupage doivent être suffisamment souples pour saisir toutes les comparaisons d'appariements potentiels tout en étant suffisamment strictes pour produire un nombre de comparaisons dont le calcul est réalisable. Le groupage est utilisé à la fois pour le modèle global et les modèles locaux dans Splink.

2.5 Le modèle global de Splink

Le modèle global de Splink est une mise en œuvre de la méthode de Fellegi-Sunter. L'utilisateur entre une valeur de probabilité a priori, qui est une estimation de la probabilité qu'une paire d'enregistrements sélectionnée au hasard soit une paire appariée. Le modèle global prend la probabilité a priori, ainsi que les valeurs m et u et utilise une série d'instructions de cas SQL définis par l'utilisateur pour filtrer les paires candidates et attribuer le poids approprié pour chaque variable à la paire candidate. Les poids sont additionnés pour calculer un score d'appariement final. Habituellement, un seuil est choisi par l'utilisateur, de telle sorte que les paires candidates dont la note est supérieure au seuil sont acceptées comme étant des appariements (c.-à-d. des doublons, si l'on apparie un ensemble de données à lui-même). Le modèle global de Splink peut être entièrement personnalisé par l'utilisateur en fonction de ses besoins, au moyen des fonctionnalités décrites dans la section 3.

Dans le modèle global, il est généralement préférable de réaliser le groupage au moyen de règles nombreuses plutôt que strictes qui permettent tous les différents types d'erreurs dans les données. Cela permet de s'assurer que tous les appariements possibles sont saisis tout en réduisant le nombre de paires candidates qui doivent être notées. Toutes les règles de groupage du modèle global utilisées dans la présente étude de cas sont indiquées dans (Xhaferaj, 2022).

2.6 Le modèle local de Splink

Le modèle local de Splink est une mise en œuvre de l'algorithme EM. Une série de modèles locaux sert à générer des valeurs m pour chacune des variables d'appariement et pour chaque clause des instructions de cas. Pour chaque modèle local, un sous-ensemble des variables d'appariement sert au groupage, et des valeurs m sont attribuées aux autres variables d'appariement. Par exemple, dans le modèle local 1, « prénom » et « nom de famille » peuvent servir au groupage, tandis que les valeurs m sont calculées pour « deuxième prénom », « date de naissance », « code postal » et « sexe ». Dans le modèle local 2, la « date de naissance » et le « code postal » peuvent servir au groupage, tandis que les valeurs m sont calculées pour le « prénom », le « nom de famille » et le « sexe ». Certaines des variables utilisées pour l'appariement auront plusieurs ensembles de valeurs m calculées, p. ex. « sexe » dans l'exemple donné ci-dessus. Par défaut, Splink utilise la moyenne harmonique des

différentes valeurs m. Cependant, l'utilisateur peut choisir d'utiliser un ensemble spécifique de valeurs m si cela semble plus approprié (voir l'exemple à la section 4.6).

Pour chaque modèle local, l'utilisateur doit entrer une valeur de probabilité a priori, qui est la proportion estimée d'appariements parmi les enregistrements groupés dans ce modèle. Il est extrêmement important que la probabilité a priori du modèle local soit appropriée. Dans le cas contraire, elle peut exécuter une itération vers un maximum local plutôt que le maximum global, ce qui donne des valeurs m absurdes. Les utilisateurs doivent toujours vérifier que les valeurs m produites par l'algorithme sont sensées, et modifier la probabilité a priori ou les instructions de cas jusqu'à ce qu'ils considèrent que les sorties sont satisfaisantes.

3. Fonctionnalités de Splink

3.1 Splink permet plusieurs états de concordance

La méthode originale de Fellegi-Sunter ne permettait que des états binaires de concordance (c.-à-d. concordance totale ou non-concordance totale), bien que les modifications apportées à la méthode permettent l'utilisation d'états de concordance partielle (Winkler, 1990). Splink, quant à lui, permet d'utiliser par défaut plusieurs états de concordance, avec des états de concordance partielle intégrés de façon transparente au processus de couplage au moyen d'instructions de cas SQL. Le fait de permettre de multiples niveaux de concordance de cette façon permet d'adapter et d'épurer les couplages sans besoin de codage supplémentaire.

Les instructions de cas SQL de Splink sont un moyen particulièrement utile d'intégrer les états de concordance partielle, car elles permettent des options nuancées, comme des instructions logiques à plusieurs variables. Par exemple, le « numéro de référence de propriété unique » ne peut être utilisé que s'il s'agit d'une correspondance exacte, la concordance dérivant sinon de la similarité du « code postal ». Ce genre d'instruction de cas complexe à plusieurs variables permet l'utilisation de combinaisons de variables qui, sinon, iraient à l'encontre de l'hypothèse d'indépendance conditionnelle de la méthode de Fellegi-Sunter.

3.2 Splink permet des ajustements de fréquence de terme

Les valeurs communes dans les variables peuvent être prises en compte au moyen d'ajustements de la fréquence de terme (Winkler, 2000), que Splink inclut par défaut. Cela peut être particulièrement utile pour les variables de chaîne avec asymétrie de la cardinalité (p. ex. les valeurs communes comme « Smith » dans la variable « nom de famille » d'un ensemble de données du Royaume-Uni). On utilise une sous-pondération pour tenir compte de leur fréquence élevée dans le ou les ensembles de données.

Une nouvelle fonctionnalité de Splink v3 est que l'ajustement de la fréquence de terme peut également servir à ajuster automatiquement les poids de concordance pour les variables catégoriques. Par exemple, en Angleterre et au pays de Galles, la concordance sur la variable « pays de naissance » n'augmente pas la confiance dans le fait que les paires candidates constituent un appariement si le pays de naissance est l'Angleterre ou le pays de Galles. Toutefois, si le « pays de naissance » concorde et que la valeur est le Mexique, cela accroîtrait la confiance dans le fait que les paires candidates constituent un appariement. Cette fonctionnalité n'était pas disponible quand nous avons effectué notre couplage. C'est pourquoi nous avons tenu compte des différences de cardinalité en utilisant différents niveaux de concordance dans les instructions de cas.

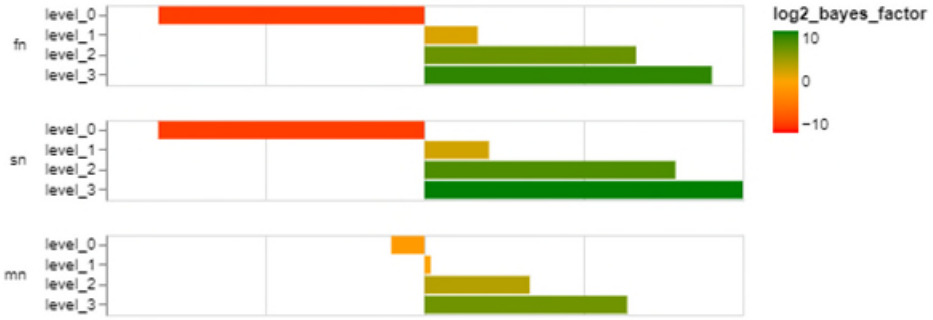
3.3 Splink produit des sorties diverses, y compris des visualisations

Splink peut produire des sorties diverses permettant d'étudier les données, d'interroger les décisions de groupage, de vérifier le bien-fondé des résultats, d'afficher des résumés de résultats, d'adapter les résultats et de fournir une plus grande transparence quant à la façon dont les résultats ont été obtenus. En voici deux exemples. D'autres exemples et explications sont disponibles sur le site Splink GitHub (Linacre, et al., 2022).

Exemple 1 : La Figure 3.3-1 indique les poids de concordance pour trois variables différentes. Elle montre que la non-concordance (niveau 0) sur le deuxième prénom a un poids beaucoup moins négatif que la non-concordance sur le prénom ou le nom de famille. On peut s'y attendre puisque le deuxième prénom est généralement moins bien enregistré dans les ensembles de données que le prénom ou le nom de famille. Cela illustre la possibilité de s'appuyer sur ce type de graphique de Splink pour détecter et vérifier le bien-fondé des poids de concordance

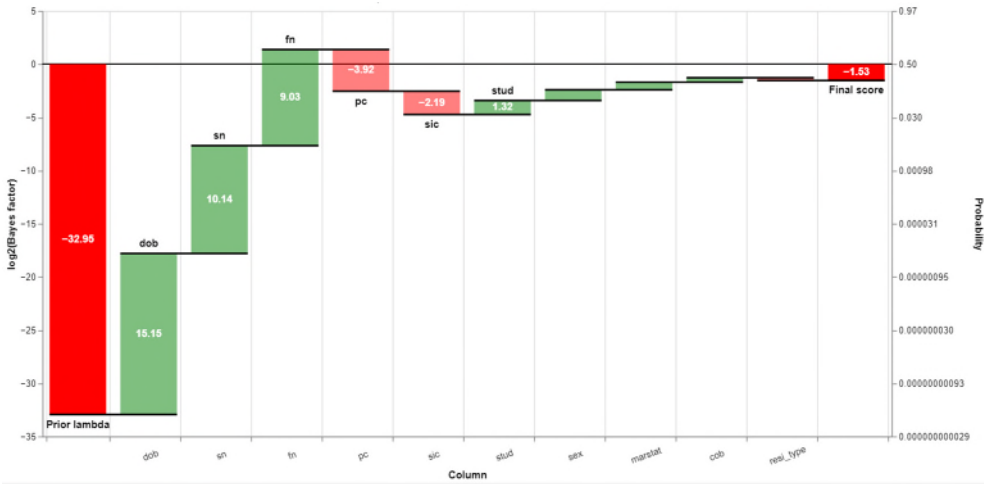
comparativement les uns aux autres. Un tableau complet montrant les poids de concordance pour toutes les variables de l'étude de cas se trouve dans (Xhaferaj, 2022).

Figure 3.3-1
Pondération de concordance attribuée pour les niveaux de concordance de trois variables : prénom (fn), nom de famille (sn) et deuxième prénom (mn)



Exemple 2 : La Figure 3.3-2 fournit un exemple de graphique en cascade Splink pour une paire candidate choisie. La probabilité a priori est appliquée en premier (probabilité a priori lambda). Dans l'exemple présenté, la paire candidate concorde sur « date de naissance » (dob), « nom de famille » (sn) et « prénom » (fn), « statut d'étudiant » (stud), « sexe », « état matrimonial » (marstat) et « pays de naissance » (cob), mais « deuxième prénom » était manquant (et n'est donc pas présent sur l'axe des x) et il y avait une non-concordance sur « code postal » (pc) et « code d'industrie » (sic). De plus, la concordance sur l'« état matrimonial » et le « pays de naissance » portait sur des réponses courantes et n'ajoutait donc pas beaucoup de poids au score d'appariement (score final). Pour le « type de résidence » (resi_type), les deux enregistrements comportaient « ménage ». Bien qu'il s'agisse d'une correspondance, un faible poids négatif a été attribué, car la concordance sur cette valeur de cette variable n'ajoute pas à la probabilité que cette paire candidate soit formée d'enregistrements en double. Dans l'ensemble, les graphiques en cascade de Splink montrent la façon dont le score de la paire candidate a été calculé, ce qui permet la traçabilité des décisions de Splink.

Figure 3.3-2
Graphique en cascade montrant les poids de concordance attribués à une paire d'enregistrements candidate donnée



4. Le couplage recensement-recensement de 2021 comme étude de cas

4.1 Variables

Comme l'étude de cas utilisait les données du recensement, nous disposions d'une grande variété de variables de couplage de grande qualité. Les données utilisées pour le couplage étaient « prénom », « deuxième prénom », « nom de famille », « date de naissance », « sexe », « pays de naissance », « type de résidence » (établissement communautaire ou ménage), « état matrimonial », « statut d'étudiant », codes postaux (« code postal du

domicile », « autre code postal », « code postal du domicile d'il y a un an », « code postal du lieu de travail ») et « groupe d'industries ».

De plus, des variables ont été calculées pour être utilisées dans le groupage et les instructions de cas. Les variables de code postal comprenaient la zone de code postal et le secteur de code postal (p. ex. pour le code postal PO15 5RR, la zone est PO et le secteur est PO15 5). Les variables de nom comprenaient les première et deuxième parties du double métaphone (Phillips, 2000), l'initiale du nom, le deuxième caractère du nom et le dernier caractère du nom.

4.2 Instructions de cas

Des instructions de cas distinctes ont été écrites pour « prénom », « deuxième prénom », « nom de famille », « sexe », « date de naissance », « pays de naissance », « type de résidence », « état matrimonial », « statut d'étudiant », les codes postaux et « groupe d'industries ». En procédant par essais et erreurs, nous avons déterminé les combinaisons et les clauses les plus appropriées. Les visualisations de Splink ont été inestimables dans ce processus, car nous avons examiné les scores obtenus par des paires ayant des caractéristiques et des erreurs différentes, c.-à-d. de « bons » appariements, des appariements avec erreurs et des non-appariements. La liste complète des instructions de cas finalement utilisées dans l'étude de cas est fournie dans (Xhaferaj, 2022).

Dans ce projet en particulier, nous avons utilisé des instructions de cas distinctes pour « prénom », « deuxième prénom » et « nom de famille ». Cependant, dans les couplages ultérieurs de données de moindre qualité, la création d'une seule instruction de cas combinant toutes les variables de nom s'est montrée plus efficace. Les utilisateurs doivent essayer plusieurs solutions pour trouver celle qui convient le mieux à leur(s) ensemble(s) de données.

4.3 Valeurs m : à partir de l'étalon de référence

Les 23 434 doublons confirmés avec l'étalon-or ont été filtrés au moyen des instructions de cas. On a calculé les valeurs m en divisant le nombre d'enregistrements satisfaisant à chaque clause par le nombre d'enregistrements ayant des valeurs non nulles pour cette variable. Les valeurs u correspondantes ont été estimées au moyen de la fonction « estimate_u_values » de Splink.

4.4 Probabilité a priori dans le modèle global

On a estimé la probabilité a priori pour le modèle global en divisant la proportion de doublons dans l'ensemble étalon-or par le nombre total de paires candidates produites par une jointure cartésienne, $N(N-1)/2$, où N est le nombre d'enregistrements du recensement.

4.5 Groupage dans le modèle global

La stratégie de groupage visait à capturer tous les vrais appariements (doublons) tout en minimisant la charge de calcul. Deux facteurs influent sur la charge de calcul : le nombre total de comparaisons créées par une règle de groupage et le nombre maximal de comparaisons dans un bloc individuel. Par exemple, une règle utilisant « l'initiale du prénom » et le « nom de famille » pourrait produire un nombre total de comparaisons dont le calcul est faisable. Cependant, chaque paire ayant une combinaison donnée de prénom et de nom de famille serait placée dans un seul bloc. Si 10 000 personnes s'appelaient « J. Smith », ce seul bloc contiendrait 49 995 000 comparaisons, ce qui entraînerait une asymétrie et rendrait le calcul inefficace. La fonction de Splink « get_total_comparisons » a servi à vérifier le nombre total de comparaisons créées par une règle. La fonction de Splink « column_combination_frequencies_chart » a servi à vérifier la taille des blocs individuels créés par des règles. Les règles qui créaient plus de 10 millions de comparaisons et/ou au moins un bloc de plus de 100 personnes ont été réécrites.

Pour la plupart des variables, si la valeur de la variable était manquante pour un ou les deux enregistrements d'une paire, il ne faut pas les sélectionner lors du groupage réalisé au moyen de cette variable. Cependant, comme il est significatif de ne pas avoir de deuxième prénom, les auteures ont écrit une fonction pour traiter les deuxièmes prénoms manquants comme des appariements. Cela permettait de ne pas exclure les personnes dont le deuxième prénom est manquant dans les blocs créés au moyen de règles qui incluaient le deuxième prénom.

On ne s'attend pas nécessairement à ce que les codes postaux du domicile soient appariés quand on cherche des réponses en double du recensement. C'est pourquoi il a fallu comparer les différentes variables du code postal pour saisir tous les doublons possibles. Au total, 12 règles contenant un code postal permuté, p. ex. « code postal du domicile » correspond exactement à « autre code postal », ont été incluses pour en tenir compte. Bien qu'il soit intuitif d'essayer d'inclure les différentes combinaisons sous forme d'énoncés OU, p. ex. « code postal du domicile » correspond exactement à « autre code postal » OU « autre code postal » correspond exactement au « code postal du domicile », nous avons constaté qu'il était plus efficace sur le plan du calcul d'écrire des règles de groupage distinctes.

Au total, 42 règles de groupage ont été utilisées, ce qui a généré environ 3 milliards de paires candidates. Il a fallu environ quatre heures pour exécuter le modèle global de Splink avec ce nombre de candidats. Les règles de groupage du modèle global se trouvent dans (Xhaferaj, 2022).

4.6 Valeurs m : à partir des modèles locaux de Splink

Bien que les valeurs m créées au moyen de l'étalon-or aient été utilisées dans le modèle global final, il était également important de démontrer que des valeurs semblables pouvaient être générées au moyen des modèles locaux de Splink. Cette fonctionnalité sera nécessaire dans les projets de couplage pour lesquels il n'existe pas d'étalon-or. Il a fallu une certaine expérimentation pour concevoir des valeurs de probabilité a priori et des groupages appropriés pour les modèles locaux afin que des valeurs m raisonnables soient calculées. Le Tableau 4.6-1 montre les valeurs de probabilité a priori et les conditions de groupage employées pour chacun des modèles locaux utilisés dans notre dernière version.

Tableau 4.6-1
Description des trois modèles locaux utilisés pour générer des valeurs m dans Splink

Modèle	Probabilité a priori	Conditions de groupage
1	0,50	Appariement exact pour : prénom, deuxième prénom, nom de famille, date de naissance, sexe
2	0,25	Appariement exact pour : date de naissance, sexe, tout code postal (domicile, autre, domicile d'il y a un an, lieu de travail)
3	0,50	Appariement exact pour : prénom, deuxième prénom, nom de famille, tout code postal (comme dans le modèle 2)

Les valeurs m générées par chacun des trois modèles locaux sont disponibles dans (Xhaferaj, 2022). En utilisant les mêmes valeurs u que précédemment, nous avons calculé les poids de concordance. Les visualisations de Splink ont été utilisées pour soutenir les décisions concernant l'utilisation des valeurs m de moyenne harmonique ou le résultat d'un modèle local unique. Par exemple, pour l'état matrimonial, on a utilisé seulement les valeurs m du modèle 2, puisque le modèle 1, le modèle 3 et la moyenne harmonique ont donné un poids négatif pour la concordance sur l'« état matrimonial » des mariages entre personnes de sexe opposé. Bien qu'il s'agisse d'un statut de concordance commun et que, par conséquent, une certaine sous-pondération puisse être attendue, cette sous-pondération semblait intuitivement incorrecte. Les utilisateurs doivent toujours vérifier que les valeurs m produites par les modèles locaux sont sensées avant de les utiliser dans le modèle global.

Un tableau montrant les valeurs m et les poids de concordance correspondants pour la méthode de l'étalon-or et la méthode des modèles locaux pour toutes les instructions de cas se trouve dans (Xhaferaj, 2022). La distribution de la fréquence des scores d'appariement obtenus au moyen des valeurs m générées par Splink était comparable à celle obtenue au moyen des valeurs m générées par la méthode étalon-or.

5. Résultats

Une évaluation visuelle de la distribution des scores d'appariement produite par Splink a démontré qu'elle avait réussi à distinguer les appariements (c.-à-d. les doublons) des non-appariements. La distribution était bimodale avec un pic important à gauche, représentant les non-appariements, et un pic beaucoup plus petit à droite, représentant les appariements. Ces pics étaient clairement séparables, mais, comme c'est habituellement le cas pour les couplages probabilistes, il y avait une zone de chevauchement. Ce phénomène est illustré dans la Figure

5-1 : le grand pic de gauche est clairement visible dans le graphique principal tandis que le petit pic de droite est visible dans l'encadré. Dans la Figure 5-2, la distribution des scores pour les appariements (doublons) et les non-appariements est indiquée. Les appariements se situent principalement dans le pic de droite, ce qui démontre que Splink a correctement détecté la majorité des vrais doublons.

Figure 5-1
Distribution des scores d'appariement générés par Splink au moyen des valeurs m générées à partir de l'ensemble de données étalon-or. L'encadré montre le même graphique avec un axe des y tronqué

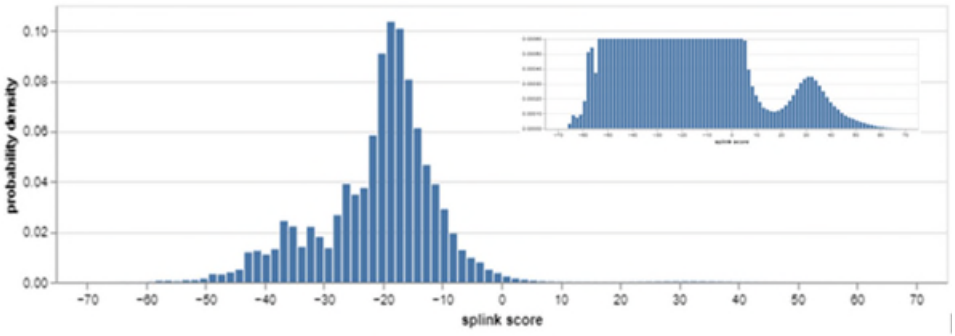
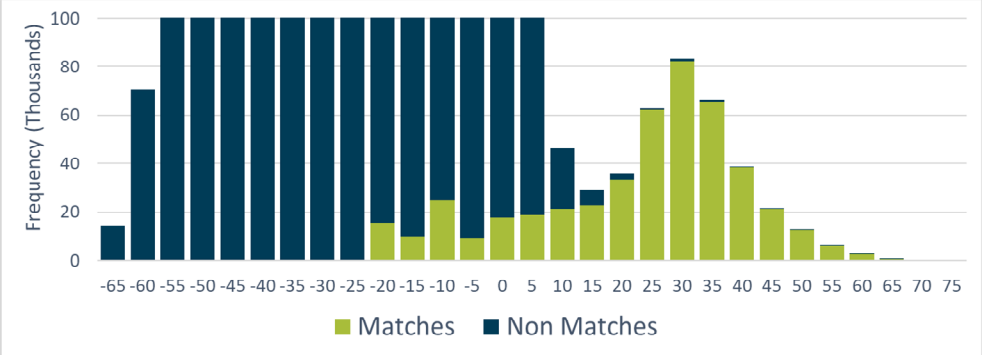


Figure 5-2
Distribution des scores d'appariement générés par Splink avec la proportion d'appariements et de non-appariements pour chaque fourchette de scores indiquée. Remarque : l'axe des y est tronqué



Pour quantifier la capacité de Splink à trouver des doublons dans l'ensemble du recensement (c.-à-d. pas seulement avec l'étalon-or), des échantillons de 1 000 enregistrements et leurs paires candidates ont été sélectionnés de 12 fourchettes de scores et examinés manuellement. Les résultats montrent que, comme on pouvait s'y attendre, les paires d'enregistrements ayant un score d'appariement élevé avaient une grande probabilité d'être des doublons et que la probabilité d'être un doublon diminuait à mesure que le score diminuait (Tableau 5-1). Une légère exception a été observée pour les scores (x) dans la fourchette $25 \leq x < 30$. Aucune méthode de couplage automatique ne permet d'attribuer correctement toutes les paires; cependant, le petit nombre de faux positifs se produisant aux scores les plus élevés serait probablement acceptable pour la plupart des projets de couplage.

Tableau 5-1
Pourcentage d'enregistrements échantillonnés confirmés comme étant des enregistrements en double au moyen d'un examen manuel

Score d'appariement (x)	Doublons confirmés (%)	Score d'appariement (x)	Doublons confirmés (%)
$40 \leq x$	99,45	$5 \leq x < 10$	14,90
$30 \leq x < 40$	98,80	$0 \leq x < 5$	3,47
$25 \leq x < 30$	98,90	$-5 \leq x < 0$	0,57
$20 \leq x < 25$	92,81	$-10 \leq x < -5$	0,44
$15 \leq x < 20$	77,46	$-15 \leq x < -10$	0,07
$10 \leq x < 15$	45,90	$x < -15$	0,00

6. Conclusions

Splink a permis de réussir à coupler le recensement de 2021 d'Angleterre et du pays de Galles à lui-même afin de détecter les réponses en double du recensement. L'ensemble de données contenait environ 58 millions d'enregistrements. Quand ils ont été joints au moyen de règles de groupage optimisées, environ 3 milliards de paires candidates ont été générées, que le modèle global Splink a traitées en environ 4 heures.

Un ensemble de données étalon-or de doublons et de non-doublons avait déjà été créé à partir des données du recensement et pouvait donc être utilisé comme données d'entraînement aux fins du calcul de la valeur m. Cependant, la mise en œuvre du modèle local de Splink de l'algorithme EM a permis d'estimer avec succès des valeurs m comparables à celles obtenues à partir de l'étalon-or.

Les résultats de l'étude de cas étayent le fait que Splink est rapide sur le plan du calcul et exact sur le plan méthodologique. Les visualisations fournies dans Splink facilitent considérablement le réglage des algorithmes, y compris l'écriture de règles de groupage et d'instructions de cas, ainsi que la vérification du bien-fondé des paramètres. De plus, les utilisateurs peuvent afficher des résumés de leurs résultats et utiliser les graphiques en cascade pour assurer la transparence algorithmique au niveau des enregistrements individuels.

L'expérience de Splink acquise dans le cadre de cette étude de cas a montré que pour être le plus efficace possible, Splink nécessite des utilisateurs qui connaissent bien le couplage des données et comprennent le mode de fonctionnement du couplage probabiliste. Splink est un excellent outil qui guide l'utilisateur tout au long du processus de couplage probabiliste des données, du groupage et de la configuration des paramètres jusqu'à la production des résultats. Splink ouvre des perspectives fort encourageantes en matière de normalisation du couplage des données entre ministères et au sein des ministères, et au-delà. Nous recommandons son utilisation chaque fois qu'un couplage probabiliste est envisagé.

Bibliographie

Dempster, A. P., Laird, N. M. & Rubin, D. B., 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 39(1), p. 1-22.

Fellegi, I. P. & Sunter, A. B., 1969. A Theory for Record Linkage. *Journal of the American Statistical Association*, 64(328), p. 1183-1210.

Linacre, R. et al., 2022. *Splink.* [En ligne, Disponible à l'adresse : https://github.com/moj-analytical-services/splink, [Site consulté le 21 09 2022].

ONS, 2022. *Coverage estimation for Census 2021 in England and Wales.* [En ligne], Disponible à l'adresse : https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/coverageestimationforcensus2021inenglandandwales, [Site consulté le 11 11 2022].

Phillips, L., 2000. The Double Metaphone Search Algorithm. *C/C++ Users Journal*, 18(6), p. 38-43.

Shipsey, R. & White, Z., 2020. *Census to Census matching strategy 2021*, London: UK Statistics Authority.

Winkler, W. E., 1990. *String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage*, Washington DC: U.S. Bureau of the Census.

Xhaferaj, K., 2022. *ONS case-study of Splink use.* [En ligne] , Disponible à l'adresse : https://github.com/Data-Linkage/Splink-census-linkage, [Site consulté le 02 11 2022].