Méthodologie de l'appariement probabiliste à l'Institut de la statistique du Québec

INSTITUT
DE LA
STATISTIQUE
DU QUÉBEC

Québec

Pour tout renseignement concernant l'ISQ
et les données statistiques dont il dispose,
s'adresser à :

Institut de la statistique du Québec
200, chemin Sainte-Foy
Québec (Québec) G1R 5T4

Téléphone :
418 691-2401
1 800 463-4090 (Canada et États-Unis)

Site Web : statistique.quebec.ca

Ce document est disponible seulement en version électronique.

Dépôt légal
Bibliothèque et Archives nationales du Québec
4e trimestre 2020
ISBN 978-2-550-87998-5 (en ligne)

Décembre 2020

Ce document a été réalisé à
l'Institut de la statistique du Québec par :

Gabriel Ouimet, statisticien
Direction de la méthodologie

Sous la coordination de :

Jimmy Baulne, coordonnateur
Direction de la méthodologie

Révision :

Julie Boudreault, révision linguistique
Direction de la diffusion et des communications

Pour tout renseignement concernant
le contenu de cette publication :

Direction de la méthodologie
Institut de la statistique du Québec
200, chemin Sainte-Foy, 3e étage
Québec (Québec) G1R 5T4

Téléphone :
418 691-2401
1 800 463-4090 (Canada et États-Unis)

Site Web : statistique.quebec.ca

Notice bibliographique suggérée

INSTITUT DE LA STATISTIQUE DU QUÉBEC (2020). *Méthodologie de l'appariement probabiliste à l'Institut de la statistique du Québec*, Québec, L'Institut, 2020, 10 p. [statistique.quebec.ca/fr/fichier/methodologie-appariement-probabiliste.pdf]

Table des matières

En quoi consiste l'appariement ?

L'appariement de deux fichiers de données consiste à identifier les individus qui sont présents dans les deux fichiers en créant des paires d'enregistrements leur correspondant. Lorsqu'un identificateur unique des individus est présent dans les deux fichiers, l'appariement est réduit à sa plus simple expression : on parle alors de jumelage. Par contre, lorsqu'il n'existe pas de tel identificateur unique, il est alors nécessaire de procéder à un appariement de nature beaucoup plus complexe. L'appariement, qu'on dit « exact[1] », peut être de nature déterministe ou probabiliste.

La méthode d'appariement déterministe est basée sur la comparaison exacte d'une combinaison de variables communes aux deux fichiers que l'on désire apparier. Les valeurs doivent concorder exactement pour qu'un lien entre deux individus soit identifié. Pour que l'appariement déterministe soit performant, les données doivent être complètes, dépourvues d'erreurs et présentes pour la presque totalité des enregistrements.

Cette méthode d'appariement probabiliste est basée sur la probabilité que deux enregistrements de deux fichiers puissent correspondre au même individu. Avec cette méthode, il est possible de maximiser l'utilisation des informations disponibles, puisqu'elle permet de moduler l'importance attribuée à certaines valeurs et de prendre en compte les données manquantes et les erreurs. En d'autres termes, cette méthode se base sur la probabilité que deux enregistrements forment une « bonne » paire, c'est-à-dire qu'ils correspondent au même individu. L'appariement probabiliste repose sur des règles de comparaison qui tirent parti du pouvoir discriminant des variables et permet d'envisager toute une gamme de concordances.

Pour la majorité des projets, les avantages de la méthode probabiliste (meilleur taux d'appariement, minimisation des erreurs d'appariement) surpassent largement ses désavantages (complexité supérieure et coûts plus élevés). Pour cette raison, l'ISQ recommande habituellement l'utilisation de la méthode probabiliste pour faire l'appariement des fichiers dans le cadre d'un projet qu'il réalise. Une description sommaire de cette méthode est présentée dans ce document. Plus de détails peuvent être fournis sur demande.

1. Un autre type d'appariement est celui dit « statistique ». Pour l'appariement statistique, il n'est pas nécessaire d'identifier le même individu dans les deux fichiers à apparier. Il suffit de trouver un individu qui possède les mêmes caractéristiques (âge, sexe, région, niveau de scolarité, revenu, etc.), peu importe qu'il s'agisse du même individu ou pas.

Méthodologie de l'appariement probabiliste

Dans tout processus d'appariement de fichiers, on doit d'abord choisir quelles variables seront utilisées. Celles-ci devraient être aussi discriminantes que possible et être présentes dans la majorité des enregistrements.

Comme mentionné précédemment, l'appariement probabiliste est une méthode d'appariement basée sur la vraisemblance que deux enregistrements de deux fichiers différents puissent correspondre au même individu. Cette méthode repose sur deux hypothèses statistiques importantes. Premièrement, l'erreur qui entache une variable doit être indépendante de celle qui entache d'autres variables. Deuxièmement, la concordance accidentelle d'une variable doit être indépendante de la concordance accidentelle d'une autre. En d'autres termes, si les champs sont liés (corrélation entre les données), alors la mesure de la ressemblance entre les deux enregistrements d'une paire est faussée. Dans ce cas, cette mesure de la ressemblance, qui permet d'évaluer s'il s'agit du même individu, peut avoir l'air mieux (ou pire) qu'elle ne l'est en réalité.

L'appariement probabiliste s'effectue à l'aide d'un logiciel spécialisé. Dans le cadre des projets réalisés par l'ISQ, c'est le logiciel G-Coup qui est utilisé. G-Coup est un système probabiliste de couplage d'enregistrements conçu par Statistique Canada.

Création des paires d'enregistrements à l'aide de pochettes

Le processus d'appariement de deux fichiers avec la méthode probabiliste se déroule de la façon suivante : tout d'abord, on doit constituer des paires d'enregistrements où chacune des paires contient un enregistrement provenant du premier fichier et un enregistrement provenant du second fichier. Pour des raisons d'efficacité, il est préférable d'éviter de comparer chacun des enregistrements du premier fichier avec tous les enregistrements du second fichier. Par conséquent, on établit des critères, qu'on appelle « pochettes », afin de limiter le nombre de paires d'enregistrements possibles. On n'évaluera que les paires dont les deux membres appartiennent à la même pochette. Ces paires sont appelées les « paires potentielles ».

Certaines variables peuvent être utilisées soit seules, soit en combinaison pour définir les pochettes et ainsi former des paires d'enregistrements qui ont minimalement ces éléments en commun.

Une fois les paires créées, on fait une première évaluation de la possibilité que les deux enregistrements d'une paire correspondent au même individu. Cette évaluation se fait sur la base de règles qui comparent des informations « comparables » entre les deux fichiers, tel le nom de famille. C'est la personne responsable de l'appariement qui doit définir ces règles et attribuer un poids indiquant l'importance de chacune des règles et de chacun des niveaux de concordance d'une règle. Ces poids sont en lien direct avec la probabilité qu'une paire d'enregistrements soit une bonne paire, c'est-à-dire qu'elle corresponde au même individu. On fait ensuite la somme des informations obtenues pour l'ensemble des règles, ce qui nous donne un poids global pour chaque paire d'enregistrements.

Une fois les poids des règles déterminés et le poids global calculé, on doit décider des seuils inférieur et supérieur qui nous permettront de classer chaque paire d'enregistrements dans l'une des catégories suivantes : définitive (poids global excédant le seuil supérieur), possible (poids global entre le seuil supérieur et le seuil inférieur) et rejetée (poids global sous le seuil inférieur). La première catégorie contient les paires pour lesquelles on considère que les enregistrements correspondent au même individu, alors que la dernière catégorie contient les paires pour lesquelles on estime que les enregistrements ne correspondent pas au même individu. Il reste alors la catégorie d'incertitude qui, elle, contient les paires d'enregistrements qui devront être examinées manuellement pour déterminer leur classification. Il est donc important que le nombre de paires de cette catégorie soit relativement petit.

Le logiciel G-Coup permet également de raffiner les poids attribués aux règles, pour tenir compte de la fréquence des valeurs. Par exemple, un nom de famille rare augmentera les chances que la paire d'enregistrements corresponde au même individu. Cette pratique permet de mieux discriminer les paires sur l'échelle de vraisemblance en donnant plus d'importance aux concordances de valeurs peu fréquentes.

Certains enregistrements d'un fichier peuvent être jumelés avec plusieurs enregistrements d'un autre fichier. Il est donc nécessaire de résoudre les conflits ainsi créés, soit manuellement ou à l'aide du logiciel, de façon à ce que chaque individu du fichier de départ soit associé à au plus un individu de l'autre fichier.

En résumé, une paire d'enregistrements sera évaluée, et classée dans l'une des catégories « définitive », « possible » ou « rejetée », seulement si les deux enregistrements de cette paire concordent pour les critères d'au moins une des pochettes. Les autres paires d'enregistrements seront rejetées sans être évaluées. Cela implique que, moins les pochettes sont restrictives, plus le nombre de paires d'enregistrements comparées est élevé et plus cela nécessite des ressources informatiques. À l'opposé, plus elles sont restrictives, plus il y a de risques de manquer de bonnes paires d'enregistrements qui correspondraient au même individu. Il est donc essentiel de maintenir un juste équilibre entre ces deux besoins.

Règles de comparaison utilisées pour l'appariement

La méthode d'appariement probabiliste repose sur des règles de comparaison qui tirent parti du pouvoir discriminant des variables et a l'avantage de permettre l'utilisation de divers niveaux de concordance. Une règle de comparaison est un ensemble de spécifications définissant les comparaisons à effectuer sur les enregistrements d'une paire. Le résultat d'une règle de comparaison peut être une concordance complète (par exemple, Baulne c. Baulne), une concordance partielle (par exemple, Baulne c. Beaulne), un désaccord (par exemple, Baulne c. Tremblay) ou un résultat manquant (par exemple, lorsque l'une des valeurs est manquante).

Outre la concordance complète, pour laquelle les données doivent être parfaitement identiques, une concordance partielle de certaines informations sert généralement à évaluer la vraisemblance des paires d'enregistrements. À titre d'exemple, supposons que le nom d'une personne dans un des fichiers contienne une faute d'orthographe. Lorsque comparés à l'aide d'une règle, les deux champs de nom vont correspondre en tout point, sauf sur la position de la faute d'orthographe. On parle alors de concordance partielle. Plusieurs niveaux de concordance partielle peuvent être définis. De plus, chacune des règles utilisées pour discriminer les paires d'enregistrements inclut un niveau de désaccord. Dans le cadre de la méthode probabiliste, un désaccord entre deux informations d'une paire peut avoir une incidence négative sur la vraisemblance de cette paire.

Chacune des variables communes aux deux fichiers à apparier, dont la comparaison est pertinente, fera généralement l'objet d'une règle. À titre d'exemple, une règle pourrait comparer la date de naissance de l'individu d'un enregistrement du premier fichier à celle d'un enregistrement du second fichier. Si la date concorde parfaitement, c'est-à-dire jour/mois/année, alors la paire se voit attribuer un certain poids positif. Si la date concorde partiellement, par exemple sur l'année et le mois de naissance, mais que les chiffres du jour sont inversés, alors la paire se voit attribuer un poids positif, qui est inférieur à celui d'une concordance complète. Finalement, si la date ne concorde ni complètement ni partiellement, alors le poids attribué est négatif.

La vraisemblance d'une paire d'enregistrements est fonction du poids de l'ensemble des règles de comparaison utilisées. Cela veut dire que, bien que la date de naissance concorde parfaitement, si le prénom de l'individu et le prénom de son père ne concordent pas, alors la paire aura un faible poids global, donc une faible vraisemblance d'être une bonne paire, c'est-à-dire de correspondre au même individu. Dans le cas contraire, si toutes ces informations concordent, alors le poids global sera plus élevé, tout comme la vraisemblance que ce soit une bonne paire.

De plus, il existe plusieurs types de règles : des règles numériques (par exemple, pour la comparaison de l'âge des individus), des règles de caractères (par exemple,

pour la comparaison du nom des individus) et des règles de dates. Il est également possible de créer des règles matricielles (par exemple, pour la comparaison de noms ou de prénoms composés). Ces dernières sont particulièrement utiles lorsqu'une seule variable d'un fichier peut être comparée à plusieurs variables d'un second fichier. La présence de plusieurs adresses ou noms de famille pour un même individu est plus facile à considérer sous un angle matriciel.

En somme, lors d'un appariement, un ensemble de règles est construit et appliqué à chacune des paires dans le but d'évaluer leur vraisemblance. Ces règles présentent de nombreuses particularités qui nous permettent de compenser les erreurs et les réponses manquantes. Ainsi, on augmente les chances d'identifier toutes les bonnes paires, en minimisant du même coup les risques d'accepter de mauvaises paires d'enregistrements.

Utilisation de poids de fréquence

Le poids évoqué dans les paragraphes précédents est aussi appelé le poids « de niveau ». Il correspond au poids de base appliqué aux différents niveaux de résultats d'une règle. Un des avantages de la méthode probabiliste est qu'il est possible de moduler l'importance donnée à certaines concordances. L'utilisation d'un poids de fréquence est un élément qui permet une telle modulation. Le poids de fréquence, lorsqu'il est utilisé pour une règle, remplace la valeur du poids de niveau attribué à cette règle.

Les chances que deux enregistrements correspondent au même individu sont beaucoup plus élevées lorsqu'on obtient une concordance sur un nom de famille rare (dans la population en général ou dans une sous-population particulière, telle une région) que lorsque l'on obtient une concordance sur un nom de famille plus commun. Ainsi, l'ajustement du poids en fonction de la fréquence des valeurs permet d'augmenter la certitude que deux enregistrements d'une paire correspondent au même individu.

Détermination des seuils

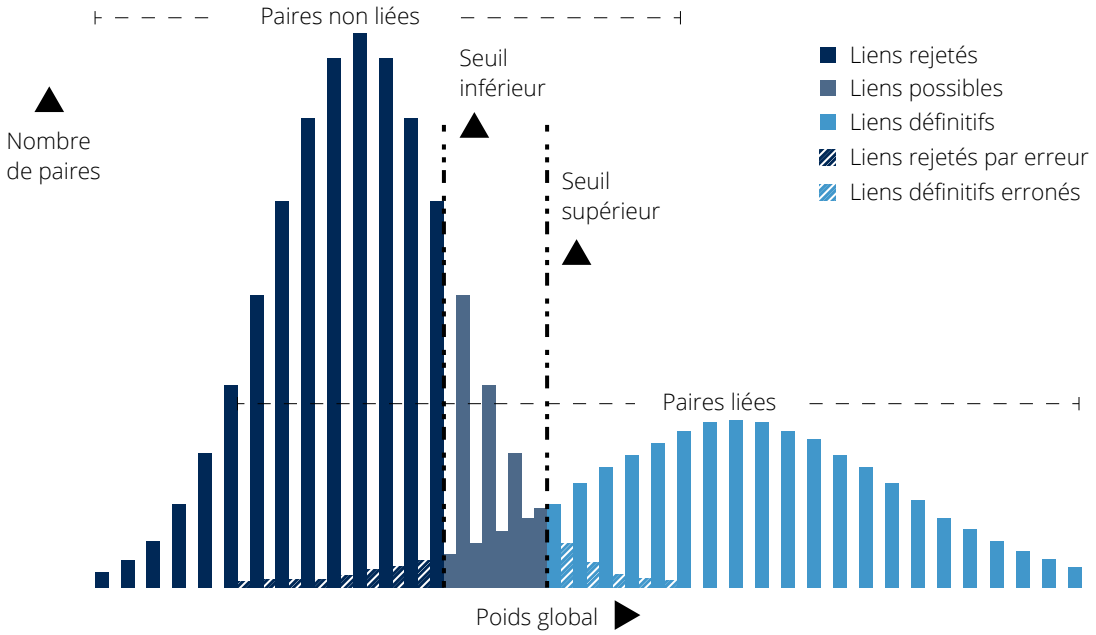
La détermination des seuils est une phase importante de l'appariement probabiliste, car elle permet de rejeter certaines paires d'enregistrements, d'en confirmer d'autres comme étant de bonnes paires, mais surtout d'établir le nombre de paires incertaines qui devront être vérifiées manuellement. À cette étape, on doit déterminer, sur la base de la distribution des poids globaux, un seuil inférieur en deçà duquel les paires seront rejetées, et un seuil supérieur au-delà duquel les paires seront considérées comme définitives. Entre ces deux seuils se trouvent les paires possibles qui devront être confirmées ou infirmées par une vérification manuelle.

Le succès de l'appariement réside dans la capacité à bien séparer les deux ensembles de paires, soit celui des paires liées (dans lequel on trouve les bonnes paires, c'est-à-dire celles dont les enregistrements correspondent au même individu) et celui des paires non liées (dans lequel on trouve les paires d'enregistrements qui ne correspondent pas au même individu). Plus il est difficile de séparer les deux ensembles de paires, plus la zone de chevauchement est importante, et plus le nombre de paires à réviser manuellement est élevé.

Afin de faciliter la compréhension des notions comme le poids global d'une paire d'enregistrements et les seuils inférieur et supérieur, la figure 1 montre la distribution théorique des ensembles des paires liées et non liées en fonction du poids global des paires. Cette figure montre les deux distributions en forme de « cloche », l'une étant généralement beaucoup plus petite que l'autre en raison du nombre peu élevé de bonnes paires comparativement au nombre très élevé de mauvaises paires. Cette image permet de distinguer clairement les deux ensembles de paires, soit celui des paires non liées à gauche, et celui des paires liées à droite. Il est vrai que cette séparation théorique des ensembles est rarement aussi évidente en pratique.

Figure 1

Distribution théorique des ensembles des paires liées et non liées[1]



1. Reproduction du schéma de la page 15 du Manuel de concepts de la documentation de G-Coup. STATISTIQUE CANADA (2010). G-Coup : Manuel de concepts. Solutions généralisées – Division de développement de systèmes, Statistique Canada, 20 p.

« L'Institut de la statistique du Québec est l'organisme gouvernemental responsable de produire, d'analyser et de diffuser des informations statistiques officielles, objectives et de qualité pour le Québec. Celles-ci enrichissent les connaissances, éclairent les débats et appuient la prise de décision des différents acteurs de la société québécoise. »

« La statistique au service de la société : la référence au Québec »

statistique.quebec.ca

Institut de la statistique Québec