

Recueil du Symposium de 2024 de Statistique Canada : Le futur des statistiques officielles

Sélection de seuils basée sur le modèle pour effectuer des couplages dans le domaine agricole

par Christian Arsenault

Date de diffusion : le 8 septembre 2025



Statistique
Canada

Statistics
Canada

Canada

Sélection de seuils basée sur le modèle pour effectuer des couplages dans le domaine agricole

Christian Arsenault¹

Résumé

La sélection optimale de seuils représente un défi crucial au chapitre du couplage probabiliste, ayant d'importantes répercussions sur l'exactitude et la fiabilité des ensembles de données couplés. Le présent document analyse le rendement du modèle de voisinage, un modèle d'erreur récemment proposé qui modélise des erreurs de couplage en fonction du nombre de couplages de chaque enregistrement. On a évalué trois algorithmes de sélection des seuils à l'aide du modèle de voisinage, mettant en évidence les formes et les limites de chacun. On a analysé leur rendement à l'aide d'études de simulation, qui ont montré que les méthodes utilisant le modèle de voisinage ont obtenu un biais relatif inférieur par rapport à deux méthodes établies pour la sélection des seuils. En outre, l'utilité pratique a été validée par des tests d'adéquation réalisés sur quatre ensembles de données agricoles, ce qui montre que le modèle peut être utilisé dans des applications de la vie réelle.

Mots-clés : Appariement statistique; modélisation probabiliste; erreurs de couplage.

1. Introduction

Le couplage d'enregistrement est une composante cruciale de l'infrastructure du Programme de la statistique agricole de Statistique Canada. Dans le cas du Recensement de l'agriculture et des enquêtes, il est devenu de plus en plus important d'utiliser des données administratives. De nombreuses tâches de couplage sont réalisées à l'aide G-Coup, le système généralisé de Statistique Canada pour le couplage probabiliste. Il a été mis en place en fonction de la méthodologie de Fellegi et Sunter (1969). Parmi les principaux défis associés aux projets de couplage, il y a l'évaluation de la qualité du couplage et l'établissement du seuil de similarité pour accepter les couplages. En l'absence de renseignements de base permettant de déceler les correspondances réelles, les seuils sont souvent établis manuellement et des revues par le commis sont utilisées pour évaluer l'exactitude du couplage. Puisque la revue par le commis est une initiative coûteuse et peu efficace, il est évident qu'il faut procéder à d'autres études dans ce domaine. Le présent article analyse l'approche décrite par Dasylyva et Goussanou (2022) pour établir des seuils de couplage sans devoir entraîner des données ou réaliser une revue manuelle, d'un point de vue théorique et pratique. Cette méthode sera mise à l'essai dans le cadre d'une étude par simulation où la correspondance réelle de chaque paire est connue. Elle sera également mise à l'essai avec quatre couplages administratifs agricoles afin d'évaluer sa viabilité en pratique.

2. Théorie

2.1 Renseignements généraux sur le couplage d'enregistrements

Le couplage d'enregistrements a pour objectif de déterminer quels enregistrements renvoient à la même entité dans deux fichiers distincts. Cette tâche est futile lorsqu'on dispose d'une clé primaire pour le couplage, comme un numéro d'entreprise ou un numéro d'assurance sociale. Toutefois, des difficultés surgissent lorsqu'on ne dispose que de variables qui peuvent ne pas être uniques à une entité donnée ou qui peuvent être enregistrées avec des erreurs, comme les noms d'entreprise ou les dates de naissance. Supposons qu'on a deux registres complets A et B qui renvoient à la

¹ Christian Arsenault, Statistique Canada, 150, promenade Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6, christian.arsenault@statcan.gc.ca

même population avec des variables de couplage K . Si l'on considère toutes les paires possibles dans le produit cartésien de $A \times B$, chaque paire est dite appariée si ses enregistrements renvoient à la même entité, et non appariée dans le cas contraire. Lors de l'échantillonnage d'une paire, on désigne l'événement montrant qu'une paire correspond par M , et l'événement montrant qu'une paire ne correspond pas par U . Fellegi et Sunter (1969) ont décrit la méthode probabiliste du couplage d'enregistrements, alors que deux enregistrements sont couplés si leur similarité dépasse un certain seuil. Cette similarité se fonde sur le vecteur de concordance d'une paire, $(\gamma^{(1)}, \dots, \gamma^{(K)})$, qui est modélisé avec un mélange linéaire logarithmique,

$$P(\gamma) = P(M)m(\gamma) + P(U)u(\gamma) \quad (2.1)$$

où $m(\gamma) = P(\gamma|M)$ et $u(\gamma) = P(\gamma|U)$ sont désignés par le nom probabilités conditionnelles m et u respectivement. Selon l'hypothèse d'indépendance conditionnelle, chacune de ces probabilités conditionnelles est le produit des probabilités marginales correspondantes de $\gamma^{(1)}, \dots, \gamma^{(K)}$, et le poids de couplage est la somme des poids associés aux différentes variables. Le poids de couplage, $w(\gamma) = \log(m(\gamma)/u(\gamma))$, est ensuite utilisé pour déterminer la décision de couplage optimale. Fellegi et Sunter ont suggéré que la règle de couplage soit déterminée par un seuil supérieur θ et un seuil inférieur θ' avec des probabilités de couplage τ et τ' respectives, dont le calcul fait l'objet d'études en cours, en plus d'être le sujet principal du présent rapport. Cependant, on va mettre l'accent sur le problème associé au fait d'attribuer un seuil simple θ , au-dessus duquel on considère que les paires correspondent, comme c'est habituellement fait en pratique.

2.2 Travaux antérieurs

On discute depuis quelque temps des calculs de seuil et de l'estimation des erreurs dans le cadre du couplage d'enregistrements probabiliste. Les contributions dans ce domaine ont été nombreuses, chacune ayant ses limites en termes d'utilisation pratique. Dans le présent rapport, on discute de deux de ces méthodes, soit la méthode établie par Belin et Rubin (1995), et celle produite par Sariyar, Borg et Pommerening (2011), qui se fonde sur la théorie de valeur extrême. La méthode élaborée par Belin et Rubin définit le seuil en modélisant les poids de couplage comme mélange de deux répartitions, alors que la première répartition représente les faux positifs, et la deuxième, les vrais positifs. Même si cette méthode évite de s'appuyer sur l'hypothèse d'indépendance conditionnelle, des données d'entraînement sont nécessaires dans le cadre d'un processus de calage. En pratique, ces données d'entraînement peuvent se fonder sur des revues par le commis, un processus coûteux. La méthode élaborée par Sariyar et coll. (2011) est conçue pour tenir compte de certaines limites de la méthode décrite ci-dessus, y compris la nécessité d'avoir des données d'entraînement. Cette méthode exige des données spécialisées, afin de déterminer l'intervalle de fonction d'excès moyen lorsqu'il est linéaire, ce qui peut être une tâche difficile en fonction de la répartition du poids.

2.3 Modèle de voisinage

Comme le décrivent Dasylyva et Goussanou (2022), on a suggéré une méthode pour calculer le seuil de couplage et l'erreur de couplage en découlant, sans devoir utiliser de données de l'utilisateur ou de données d'entraînement. Elle estime l'exactitude du couplage en modélisant le nombre de couplages d'un enregistrement au moyen d'un modèle dit de voisinage. Des notes sont nécessaires pour décrire ce modèle. Tout d'abord, on suppose une population de N entités avec deux fichiers complets pour la population, ne renfermant aucun doublon. Le modèle s'applique également lorsque les fichiers ne sont pas des recensements. Cependant, il est plus pratique de décrire le modèle avec cet exemple simplifié. Chaque population a donc N enregistrements, alors qu'on désigne les valeurs du premier fichier comme v_i , où i correspond à l'entité i , et prend sa valeur de l'ensemble V fini. On considère ensuite que le deuxième fichier est une permutation aléatoire inconnue de i , désigné par $\pi(\cdot)$, alors que les valeurs de l'enregistrement pour la même entité dans les deux fichiers sont indépendantes et identiquement distribuées. Le *voisinage* d'un enregistrement v_i , désigné par $B(v_i)$, renvoie au sous-ensemble de V , de manière à ce que v_i et v'_j soient couplés. Selon ces définitions, on peut définir différentes mesures d'exactitude du couplage. Un faux négatif se produit lorsque $v'_{\pi(i)} \notin B(v_i)$, et un faux positif se produit lorsque $v'_{\pi(i')} \in B(v_i)$ pour i et i' distincts. Il existe de nombreuses mesures d'exactitude qui peuvent être définies en fonction de cela. Cependant, on va mettre l'accent sur la précision et le rappel. La précision représente la proportion des enregistrements couplés qui sont de véritables correspondances, et le rappel est la proportion des véritables correspondances qui ont été couplées. On peut utiliser les termes taux de faux négatif (FNR) et taux de faux positif (FPR) du couplage pour les décrire. Plus précisément,

$$\begin{aligned}
FNR &= \frac{1}{N} \sum_{i=1}^N I(v'_{\pi(i)} \notin B(v_i)) \\
FPR &= \frac{1}{N(N-1)} \sum_{i=1}^N \sum_{i' \neq i}^N I(v'_{\pi(i')} \in B(v_i)) \\
Précision &= \frac{1 - FNR}{1 - FNR + (N-1)FPR} \quad (2.2)
\end{aligned}$$

Maintenant qu'on a défini le concept de voisinage, on peut montrer comment il peut servir à estimer les erreurs de couplage et à établir le seuil. En supposant qu'on dispose de deux fichiers complets ne comprenant aucun doublon, on désigne le nombre de voisins d'un enregistrement v_i par n_i , qui sert à modéliser les erreurs de couplage. Les deux équations suivantes montrent la relation entre n_i et les faux négatifs (FN) et les faux positifs (FP) comprenant v_i :

$$FN = \begin{cases} 1 & \text{if } n_i = 0, \\ 0 \text{ or } 1 & \text{if } 1 \leq n_i \leq N-1, \\ 0 & \text{if } n_i = N \end{cases}, \quad FP = \begin{cases} 0 & \text{if } n_i = 0, \\ n_i - 1 \text{ or } n_i & \text{if } 1 \leq n_i \leq N-1, \\ N-1 & \text{if } n_i = N \end{cases} \quad (2.3)$$

Avec ces renseignements, on peut modéliser les erreurs de couplage. On peut considérer n_i comme la somme des vrais positifs et des faux positifs. Le premier se fonde sur le couplage ou non de v_i à l'enregistrement correspondant dans l'autre fichier. Il s'agit simplement d'un essai de Bernoulli avec une probabilité $p_N(v_i)$ conditionnelle à v_i . En ce qui concerne les faux positifs, ils suivent une loi binomiale se fondant sur $N-1$ essais (un pour chaque enregistrement non correspondant dans l'autre fichier) et une probabilité de réussite $\lambda_N(v_i)$ conditionnelle à v_i . Si N est grand, la deuxième distribution converge vers une distribution de Poisson avec le paramètre λ . Par conséquent, le nombre de voisins d'un enregistrement peut être modélisé par le mélange fini suivant avec les G catégories :

$$n_i \sim \sum_{g=1}^G \alpha_g \left(\text{Bernoulli}(p_g) * \text{Poisson}(\lambda_g) \right) \quad (2.4)$$

Les paramètres du mélange peuvent être estimés à l'aide du maximum de vraisemblance, et un estimateur pour la précision du couplage découle de la loi des grands nombres.

$$\text{Precision} \rightarrow \left(1 + \frac{\sum_{g=1}^G \alpha_g \lambda_g}{\sum_{g=1}^G \alpha_g p_g} \right)^{-1} \text{ as } N \rightarrow \infty \quad (2.5)$$

Le fait d'avoir un estimateur de la précision du couplage nous donne non seulement une indication de la qualité, mais fournit également les données qui nous permettent d'établir le seuil.

2.4 Algorithmes pour choisir le seuil à l'aide du modèle de voisinage

Maintenant qu'on dispose d'un modèle d'erreur, on peut l'utiliser pour définir le seuil de couplage. Le seuil est établi en fournissant une précision cible si l'objectif est de maximiser le rappel, ou en précisant un rappel cible si l'objectif est de maximiser la précision. Pour simplifier le tout, on va supposer qu'une précision cible a été définie. On a envisagé trois algorithmes pour établir le seuil. Ils sont décrits dans les trois sections qui suivent.

2.4.1 Recherche exhaustive

Dans le cadre de cette méthode, on évalue chaque poids de paire candidate parmi les poids de paires observées dans le couplage, afin de déterminer s'il correspond à la précision cible fournie. L'algorithme est le suivant.

1. Trier les paires en fonction du poids descendant.
2. Pour chaque poids de paire unique, calculer la précision estimative.
 - Si la précision estimative est égale ou inférieure à la précision cible, il faut réutiliser le poids de la paire actuelle.

La mise en application de cette méthode est simple et facile. Cependant, elle peut être lente, en fonction du nombre de poids de la paire unique.

2.4.2 Recherche binaire

La recherche binaire a un temps d'exécution logarithmique, car l'espace de recherche est divisé par deux à chaque itération. Cependant, elle présente un inconvénient, celui de nécessiter un paramètre de tolérance. L'algorithme est le suivant :

1. Il faut définir une tolérance quant à l'optimalité pour laquelle la précision cible peut différer de la précision estimative.
2. Il faut établir le poids de la paire médiane pour l'espace de recherche actuel comme seuil avant de calculer la précision cible.
 - Si la précision estimative correspond à la tolérance quant à l'optimalité, il faut l'établir comme seuil définitif.
 - Si la précision estimative est inférieure à la cible, la récursivité vise les valeurs supérieures du poids des paires.
 - Sinon, la récursivité vise les valeurs inférieures du poids des paires.

2.4.3 Partitionnement récursif

Le troisième algorithme comprend l'utilisation du partitionnement récursif pour établir un arbre décisionnel à l'aide de l'exactitude du couplage obtenue à l'aide du modèle de voisinage, afin d'estimer l'hétérogénéité de chaque nœud. Si l'on tient compte de notre vecteur de résultats de comparaison, $\gamma = (\gamma^{(1)}, \dots, \gamma^{(K)})$, on peut créer un arbre décisionnel qui vise à créer des nœuds homogènes en termes de probabilité d'appariement conditionnelle (c.-à-d. la probabilité conditionnelle qu'une paire corresponde en raison des résultats de comparaison) en assurant un fractionnement des variables de couplage $\gamma^{(k)}$ possibles. Si le fractionnement de la variable de couplage $\gamma^{(k)}$ entraîne la création de deux nœuds R_1, R_2 correspondant à la concordance et à la non-concordance de la variable de couplage, la forme de la fonction coût du fractionnement t est :

$$\hat{J} = \left(\sum_i I(X_i \in R_1(t)) \right) Var(Y|X \in R_1(t)) + \left(\sum_i I(X_i \in R_2(t)) \right) Var(Y|X \in R_2(t)) \quad (2.6)$$

Les deux variances dans cette somme pondérée peuvent être écrites en termes de précision pour établir le lien avec le modèle de voisinage :

$$Var(Y|X \in R(t)) = E[Y|X \in R(t)](1 - E[Y|X \in R(t)]) \quad (2.7)$$

$$E[Y|X \in R] = \text{Précision} \rightarrow \left(1 + \frac{\sum_{g=1}^G \alpha_g \lambda_g}{\sum_{g=1}^G \alpha_g p_g} \right)^{-1} \text{ lorsque } N \rightarrow \infty \quad (2.8)$$

On peut maintenant minimiser la fonction coût à l'aide d'une précision estimative du modèle de voisinage, afin de déterminer les fractions optimales. Voici comment on peut résumer l'algorithme :

1. Il faut assurer le partitionnement récursif des données à l'aide de l'erreur estimée du modèle de voisinage comme critère de fractionnement.
2. Trier les nœuds terminaux de l'arbre décisionnel obtenu en augmentant les poids de couplage. Pour chaque nœud terminal :
 - Si la précision cumulative actuelle du couplage est supérieure à la cible, il faut classer les paires dans la catégorie des paires correspondantes et passer au nœud suivant.
 - Si la précision cumulative actuelle du couplage est égale ou inférieure à la cible et si la précision du nœud précédent est supérieure à la cible, il faut classer les enregistrements de ce nœud dans la

- catégorie des paires correspondantes, avec une probabilité telle que la précision cible est atteinte. Il faut établir le seuil du couplage comme le poids de paire le plus élevé de ce nœud.
- Sinon, il faut classer les paires dans la catégorie des paires non correspondantes pour le nœud.

Cette méthode fournit des renseignements sur la qualité du couplage au niveau de la paire, puisque chaque paire sera affectée à un nœud avec une précision estimative. En outre, les variables de couplage utilisées lors du fractionnement donnent un aperçu des variables les plus efficaces pour diviser les paires en sous-groupes homogènes, en montrant essentiellement quelles variables étaient importantes lors du couplage. Cependant, l'algorithme est lent lorsque les variables de couplage et les poids de paires uniques sont nombreux, compliquant l'application en pratique.

2.5 Test d'adéquation

Le dernier élément théorique qu'on va utiliser dans les résultats est le test d'adéquation. Cette méthode a été décrite par Dasylyva et Goussanou (2024). Il s'agit d'un test de khi carré pour évaluer la validité du modèle de voisinage. Puisque le modèle de voisinage indique que les voisins de chaque enregistrement devraient suivre un mélange dans lequel chaque composante est la convolution d'une distribution de Bernoulli avec une distribution de Poisson, cela signifie qu'on a des fréquences attendues de voisins qu'on devrait observer pour un couplage donné. Dans ce cas, il est tentant d'utiliser un test de khi carré. Cependant, un tel test exige un niveau d'indépendance, ce qui n'est pas le cas en raison des observations corrélées. Dasylyva et Goussanou (2024) suggéraient, comme solution, d'utiliser un sous-ensemble de valeurs m_N et n_i , afin que les observations soient approximativement indépendantes, alors que le choix de $m_N \propto N^{2/5}$ est optimal. Puis, pour un nombre donné de catégories, on peut vérifier si le modèle s'applique à un ensemble de données précis. Le test se fonde sur le nombre d'enregistrements qui n'ont pas de couplage, qui ont un couplage, deux couplages, ..., $d - 1$ couplages, et d couplages ou plus. Notre statistique du test en découlant a une distribution au khi carré approximative avec $d - 2G$ degrés de liberté, où G désigne le nombre de catégories du mélange fini. Dans le cadre de ce test d'adéquation, notre hypothèse nulle est ensuite rejetée si la statistique du test est supérieure au niveau critique $\chi^2_{d-2G, 1-\alpha}$. Cependant, lors de nos tests, on utilisera une procédure bootstrap paramétrique pour évaluer le niveau critique, qui est robuste aux petites valeurs de m_N .

3. Résultats

Cette section présente les résultats de l'utilisation du modèle de voisinage dans le cadre d'une simulation et dans le monde réel. La section 3.1 présentera une étude par simulation réalisée pour comparer les méthodes du voisinage, de la TVE et de BR pour la sélection du seuil. On comparera la précision estimative de chaque méthode à la vérification sur le terrain des simulations, afin de déterminer l'efficacité de chaque méthode. Dans la section 3.2, le test de khi carré modifié sera appliqué à différents ensembles de données agricoles, afin d'évaluer l'applicabilité du modèle de voisinage en pratique.

3.1 Simulations

Dans le cadre de notre étude de simulation, 100 populations de $N = 10000$ unités ont été produites, chaque unité ayant des caractéristiques catégorielles $K = 10$ tirées uniformément de $\{1, 2, 3\}$, ainsi qu'une variable de blocage $B \in \{1, \dots, 5000\}$. À partir de chaque population simulée, deux recensements ont été produits. Des coquilles ont été saisies pour chaque variable de couplage catégorique. Des coquilles ont été insérées avec une probabilité proportionnelle à k , plus précisément, $e_k = 0.05 + \frac{k}{10} (0.2 - 0.05)$. Ainsi, chaque variable de couplage a été saisie sans erreur avec une probabilité $1 - e_k$. En présence d'une coquille, une valeur a été choisie aléatoirement parmi les deux autres valeurs possibles pour cette variable. Pour simuler l'effet de l'établissement d'un bloc, des paires potentielles ont été créées en comparant uniquement les enregistrements correspondant aux critères de la méthode des blocs. Dans le cas de ceux ne correspondant pas aux critères de la méthode des blocs, on a comparé les variables de couplage en créant un vecteur de résultat γ_k , et un poids de couplage a été attribué à la forme $w = 10 \log_2 B + w_1 + \dots + w_K$, où

$$w_k = \gamma_k \log_2 \left(\frac{m_k}{u_k} \right) + (1 - \gamma_k) 10 \log_2 \left(\frac{1 - m_k}{1 - u_k} \right), \quad (3.1)$$

$$m_k = (1 - \varepsilon_k)^2 + \varepsilon_k^2 / (d - 1), \quad (3.2)$$

$$u_k = 1/d \quad (3.3)$$

Il convient de noter que, puisque les résultats sont simulés, w est le poids de couplage probabiliste réel, qui peut être calculé directement à partir des probabilités m - et u -. Pour chaque simulation, on a calculé le seuil de couplage selon deux des méthodes de voisinage candidat, ainsi que les méthodes BR et de la TVE. La méthode BR nécessite des données d'entraînement pour un processus de calage, et la méthode de la TVE nécessite la sélection d'un paramètre critique, tous deux obtenus par échantillonnage aléatoire de 10 % des paires, ce qui est le cas idéal pour les deux méthodes si la revue par le commis ne comporte aucune erreur, comme c'est le cas ici. Les deux autres méthodes se fondent sur le modèle de voisinage, y compris une recherche exhaustive (voisin-1 dans les figures) et une recherche binaire (voisin-2 dans les figures). Le partitionnement récursif a été omis de l'étude par simulation du fait des longues durées d'exécution. Les deux figures suivantes montrent le rendement de chaque algorithme en termes de différence relative entre la précision obtenue en fonction du seuil de pondération sélectionné et la précision cible pour une précision cible de 0,9 et de 0,8 respectivement.

Figure 3.1-1

Différence relative entre la précision atteinte et la précision cible - précision cible = 0,9

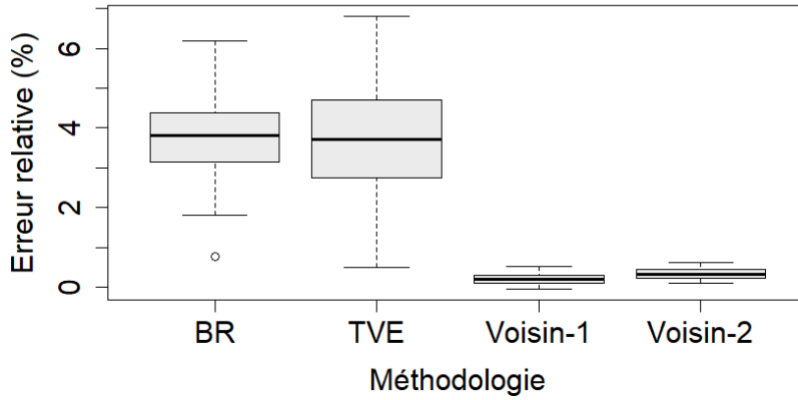
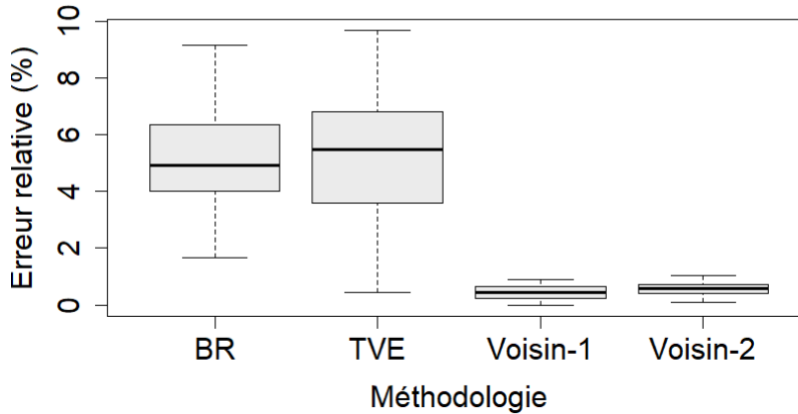


Figure 3.1-2

Différence relative entre la précision atteinte et la précision cible - précision cible = 0,8



Les résultats pour une précision cible de 0,9 et de 0,8 étaient similaires, ayant dans les deux cas un biais relatif de moins de 10 % pour toutes les méthodes dans le cadre de toutes les simulations. Cependant, par rapport à la méthode BR et à la TVE, les deux méthodes utilisant le modèle de voisinage ont constaté un biais et une variance inférieurs. Cela signifie qu'elles offrent un meilleur rendement lors des simulations sans exiger d'intervention humaine ou de revue par le commis. Tandis que la recherche exhaustive et la recherche binaire ont affiché un rendement similaire,

on privilégie la recherche binaire en pratique, car son exécution est plus rapide, en faisant un choix plus pratique lorsqu'il y a de nombreux poids de paires uniques dans le cadre du problème de couplage.

3.2 Données réelles

Dans les simulations, la méthode de recherche binaire est prometteuse, car elle est rapide et établit le seuil avec exactitude en termes de différence entre la précision cible et la précision obtenue. Cependant, il est tout aussi important que le modèle de voisinage corresponde aux données avec lesquelles il sera utilisé en pratique, car aucune revue par le commis ne sera réalisée. Pour vérifier cette exigence, on a déterminé quatre ensembles de données agricoles pour mettre à l'essai le modèle de voisinage afin de choisir le seuil.

1. Assurance-récolte : liste des exploitations agricoles participant à un programme fédéral d'assurance-récolte
2. Agri-stabilité : liste des exploitations agricoles qui participent au programme fédéral de stabilisation du revenu pour les agriculteurs
3. Producteurs acéricoles : liste des producteurs acéricoles au Québec
4. Producteurs biologiques : liste des fermes biologiques au Québec

Chaque liste a été couplée au Registre des entreprises de Statistique Canada, qui a produit des paires avec leurs poids probabilistes pour calculer un seuil. En outre, on a été en mesure d'exécuter les tests d'adéquation de khi carré modifié décrits dans la section 2.5. Le tableau suivant illustre les résultats obtenus.

Tableau 3.2-1

Résultats du test d'adéquation pour les ensembles de données en agriculture

Ensemble de données	Statistique du test	Niveau critique	Rejet de H_0
Producteurs acéricoles	0,558	3,194	Faux
Agri-stabilité fédérale	2,757	3,337	Faux
Producteurs acéricoles	1,659	3,558	Faux
Producteurs biologiques	0,073	5,425	Faux

On a mis à l'essai chaque ensemble de données pour une catégorie et des sous-ensembles d'enregistrements $m_N = 50$, qui allaient de pair avec la taille recommandée de $N^{2/5}$ pour chaque ensemble de données, tout en fournissant suffisamment de données pour estimer de manière fiable les paramètres du modèle. Dans chaque cas, la statistique du test était inférieure au niveau critique évalué lors de la procédure bootstrap. Il a donc été impossible de rejeter l'hypothèse nulle selon laquelle les fréquences observées suivaient celles précisées par le modèle de voisinage.

4. Discussion et travaux futurs

On a constaté que le modèle de voisinage est un outil efficace dans le cadre de notre étude par simulation et à l'aide de données réelles. Lorsqu'on utilise des données simulées, le modèle de voisinage a fourni un seuil de couplage avec une précision estimative se rapprochant de la vérification sur le terrain par rapport à la méthode Belin-Rubin ou à la théorie de valeur extrême lorsqu'elles sont entraînées à l'aide d'un échantillon composé de données simulées. En termes d'utilisation pratique, lorsqu'il est appliqué avec un algorithme de recherche binaire, il est exécuté rapidement dans l'étude par simulation ainsi qu'avec quatre ensembles de données agricoles, et il réussit également le test d'adéquation du khi carré modifié dans chaque cas. Bien que l'utilisation du partitionnement récursif a été difficile pour les grands ensembles de données ayant des distributions de poids de paires plus variées, cette méthode permet d'évaluer la qualité des couplages au niveau des paires. En raison du potentiel que montre le modèle de voisinage, on envisagera de l'utiliser dans le cadre de notre processus de couplage agricole. On recommande de valider le respect des hypothèses du modèle, et d'appliquer le test de khi carré modifié pour valider l'applicabilité du modèle au cas d'utilisation particulier.

Bibliographie

- Belin, T., et Rubin, D. (1995), “A Method for Calibrating False-Match Rates in Record Linkage”, *Journal of the American Statistical Association*, 90, p. 694-707.
- Dasylda, A., et Goussanou, A. (2022), “On the consistent estimation of linkage errors without training data”, *Japanese Journal of Statistics and Data Science*, 5, p. 181-216. Article accessible à l’adresse <https://doi.org/10.1007/s42081-022-00153-3>.
- Dasylda, A., et Goussanou, A. (2024), “Making statistical inferences about linkage errors”, *Japanese Journal of Statistics and Data Science*, 7, p. 17-56. Article disponible à l’adresse <https://doi.org/10.1007/s42081-023-00228-9>.
- Fellegi, I., et Sunter, A. (1969), “A theory of record linkage”, *Journal of the American Statistical Association*, 64, p. 1183-1210.
- Sariyar, M., Borg, A., et Pommerening, K. (2011), “Controlling false match rates in record linkage using extreme value theory”, *Journal of Biomedical Informatics*, 44, p. 648-654.