Recueil du Symposium 2018 de Statistique Canada
Combiner pour vaincre : innovations dans l'utilisation de multiples sources de données

# Règles de décision et estimation du taux d'erreur pour le couplage d'enregistrements au moyen d'un modèle de probabilité

Clayton Block[1]

## Résumé

Depuis 1997, Élections Canada maintient le Registre national des électeurs, une base de données sur les Canadiens âgés de 18 ans et plus, utilisé pour administrer les élections fédérales. Cette base de données est mise à jour à partir de plusieurs sources administratives fédérales et provinciales, couplées aux électeurs figurant dans la base de données au moyen de renseignements personnels comme les noms, la date de naissance, le sexe et l'adresse. Au départ, un logiciel de couplage commercial fondé sur la théorie de Fellegi-Sunter était utilisé dans ces activités de couplage. La méthodologie et le logiciel utilisés ont progressivement été remplacés par des solutions personnalisées, offrant davantage de souplesse dans le traitement des paires potentielles et réduisant les taux d'erreur de classification associés au processus de couplage. Une amélioration clé de la méthodologie est une reformulation de la règle de décision bien connue de Fellegi-Sunter, maintenant exprimée sous forme de probabilité d'intérêt et comparée à une tolérance d'erreur. Pour l'appariement fondé sur les renseignements personnels, les probabilités nécessaires sont calculées à partir des paires observées à l'aide d'un modèle de probabilité simple de la concordance due au hasard pour la date de naissance. Les hypothèses du modèle doivent être assez réalistes. Les probabilités calculées pour chaque paire peuvent également être simplement additionnées pour produire des estimations de deux types d'erreur d'appariement, ce qui n'exige aucun logiciel spécialisé ni aucune procédure mathématique complexe. Les méthodes décrites seront utilisées dans différents processus de couplage à Élections Canada, chacun présentant des taux d'appariement attendus différents. La crédibilité des taux d'erreur qui en résultent sera par la suite évaluée. Dans l'avenir, ces résultats pourraient être comparés à ceux obtenus au moyen de méthodes d'estimation des taux d'erreur concurrentes et plus compliquées afin d'en relever les différences.

Mots clés : Couplage d'enregistrements probabiliste; théorie de Fellegi-Sunter; taux d'erreur de classification.

## 1. Introduction

### 1.1 Contexte

Depuis 1997, Élections Canada maintient le Registre national des électeurs, une base de données sur les Canadiens âgés de 18 ans et plus, utilisé pour administrer les élections fédérales et les référendums. En plus des renseignements fournis directement par les électeurs eux-mêmes, la base de données est mise à jour à partir de plusieurs sources administratives fédérales et provinciales. Pour pouvoir être utilisés aux fins des mises à jour, les enregistrements des sources de renseignements doivent d'abord être appariés aux enregistrements de la base de données du Registre, un processus appelé le couplage d'enregistrements. Les renseignements figurant dans le Registre comprennent des noms, des adresses, des dates de naissance et des sexes.

### 1.2 Description du problème de couplage d'enregistrements

Dans une paire d'enregistrements, nous voulons déterminer si les deux enregistrements se rapportent à la même entité ou à des entités différentes. Les paires d'enregistrements peuvent provenir de la même source, comme dans la détection d'enregistrements en double, de sources différentes, comme dans l'appariement de fichiers, ou des deux cas de figure.

Une fois les décisions prises concernant les paires d'enregistrements en question, il serait également utile de déterminer combien d'erreurs de classification ont été faites. Il existe deux types d'erreurs possibles : l'acceptation

[1]Clayton Block, Élections Canada, 30 rue Victoria, Gatineau (Québec), Canada, K1A 0T6 (clayton.block@elections.ca)

d'une paire qui se rapporte en réalité à des entités différentes, et le rejet d'une paire qui se rapporte en réalité à la même entité. Ces deux types d'erreurs seront appelés faux + et faux –, respectivement.

Pour prendre des décisions sensées et éclairées, il est essentiel de bien comprendre les conséquences de chaque type d'erreur et, dans la mesure du possible, d'en tenir compte dans la prise de décisions. Dans le contexte de la tenue à jour du Registre à Élections Canada, les paires faux + peuvent faire en sorte que les enregistrements du Registre soient corrompus et comportent des renseignements erronés qui deviennent difficiles à corriger, et que des électeurs légitimes n'apparaissent pas sur une liste électorale, ou qu'ils y figurent à une adresse incorrecte. Cela incommode les électeurs et peut entraîner une perception publique négative d'Élections Canada, et pourrait même donner lieu à une couverture médiatique négative de l'organisation. Par ailleurs, les paires faux – peuvent faire en sorte que des mises à jour légitimes soient manquées ou que des non-électeurs ou des enregistrements en double soient conservés sur une liste électorale. Cela pourrait également entraîner une perception publique négative d'Élections Canada et peut-être même une couverture médiatique négative.

Il convient de souligner que les deux types d'erreurs peuvent entraîner des conséquences négatives. Compte tenu de ces conséquences, Élections Canada estime qu'un faux + est plus grave qu'un faux –, non seulement parce que les conséquences sont plus graves, mais également parce qu'ils sont plus difficiles à corriger après coup. En réalité, de nombreux cas de faux – sont facilement corrigés par la suite, lorsque davantage de renseignements deviennent disponibles à propos des paires, comme des mises à jour d'adresse.

En plus de simplement accepter ou rejeter les paires d'enregistrements, une troisième possibilité pourrait toujours être considérée. Si l'on ne sait pas si l'on doit accepter ou rejeter une paire compte tenu des renseignements disponibles, la décision elle-même pourrait être reportée jusqu'à ce que suffisamment de renseignements puissent être obtenus pour résoudre le cas, par exemple en communiquant avec les entités concernées. Malheureusement, compte tenu du temps, des renseignements et des ressources limités disponibles, et du fardeau supplémentaire que cela imposerait aux électeurs, il ne s'agit pas d'une possibilité réaliste pour Élections Canada.

## 2. Règles de décision pour le couplage des enregistrements

### 2.1 Le couplage d'enregistrements est un problème de probabilité

Prenons l'exemple fictif mais réaliste suivant d'une paire d'enregistrements :

| Nom | Date de naissance | Adresse |
|---|---|---|
| Robert J. Smith | 9 juillet 1963 | 123, rue Principale, K1L 5T4 |
| Bob Smith | 9 juillet 1963 | 246, promenade Elm, R1M 4T9 |

En l'absence de tout autre renseignement, rien ne permet de savoir avec certitude s'il s'agit d'une personne qui a déménagé à une nouvelle adresse ou de deux personnes différentes qui ont simplement des noms semblables et la même date de naissance. Devant cette incertitude, l'outil qu'il convient d'utiliser est la probabilité.

Si *M* désigne un appariement vrai, *U* désigne un non-appariement vrai et *Résultats* est la synthèse de tout ce que nous pouvons observer de pertinent à propos d'une paire donnée, une bonne règle de décision serait

| | |
|---|---|
| Rejeter la paire si | $P_M < Tolérance\ au\ faux +$ |
| Accepter la paire si | $P_U < Tolérance\ au\ faux -$ |
| Reporter la décision | Sinon |

où

$$P_M = \Pr(M|Résultats) \quad (1)$$

$$P_U = \Pr(U|Résultats) = 1 - P_M \quad (2)$$

Si le report de la décision n'est pas une possibilité, une des tolérances est simplement écartée. La décision devient celle d'accepter ou de rejeter la paire d'après la tolérance conservée et un type d'erreur doit être laissé sans contrôle.

L'espace d'échantillon utilisé pour définir les probabilités d'intérêt est simplement l'ensemble de toutes les paires d'enregistrements possibles pouvant être formées à partir des enregistrements disponibles et de tous les renseignements qui sont disponibles à propos de ces enregistrements et paires d'enregistrements.

## 2.2 Considérations générales relatives aux règles de décision

Les probabilités nécessaires pour prendre des décisions judicieuses dans le couplage des enregistrements ne sont pas connues et doivent donc en quelque sorte être estimées. Les estimations qui en résulteront ne seront pas parfaites et dépendront fortement des renseignements utilisés. Comme dans toute décision, la subjectivité doit pouvoir jouer un rôle au besoin. Pour assurer la cohérence logique, en plus des probabilités calculées, il y a lieu de considérer les principes directeurs généraux suivants dans les décisions relatives au couplage des enregistrements :

1. Tous les renseignements disponibles à propos des enregistrements d'une paire qui sont pertinents à la décision doivent être utilisés.
2. Les aspects subjectifs du processus doivent être établis à l'avance dans la mesure du possible et doivent être intégrés logiquement aux considérations plus objectives.
3. Les paires ayant des *Résultats* communs doivent faire l'objet de la même décision.
4. Les paires ayant des *Résultats* « meilleurs » qu'une autre paire qui a été acceptée doivent être acceptées.
5. Les paires ayant des *Résultats* « pires » qu'une autre paire qui a été rejetée doivent être rejetées.

La détermination des paires qui sont « meilleures » ou « pires » que les autres constitue le point où les aspects objectifs et subjectifs du problème peuvent entrer en conflit et peut-être ne jamais être résolus parfaitement.

## 2.3 Couplage d'enregistrements probabiliste

Une approche couramment utilisée pour résoudre ce problème est appelée le couplage d'enregistrements probabiliste (Fellegi et Sunter, 1969). Plutôt que d'utiliser la probabilité d'intérêt $P_M$ directement, la règle de décision envisagée est

| | |
|---|---|
| Rejeter la paire si | $R < Seuil\ inférieur$ |
| Accepter la paire si | $R > Seuil\ supérieur$ |
| Reporter la décision | Sinon |

où $$R = \Pr(Résultats|M)/\Pr(Résultats|U) \quad (3)$$

Si toutes les probabilités sont correctement spécifiées, on peut facilement démontrer à l'aide de la règle de Bayes que cette règle est équivalente du point de vue mathématique à la règle indiquée à la section 2.1 ci-dessus.

Pour calculer *R*, l'approche envisagée consiste à limiter les *Résultats* aux résultats de chaque comparaison des champs pertinents dans les enregistrements eux-mêmes. Une hypothèse de simplification clé est que les résultats de la comparaison pour chaque champ inclus sont tous indépendants les uns des autres.

Les probabilités sont estimées de manière itérative à partir d'un sous-ensemble de toutes les paires possibles, obtenu en imposant la concordance stricte pour plusieurs combinaisons différentes de champs ou de composantes de champs. Plutôt que d'utiliser l'estimation de *R* directement, cette dernière est transformée en un poids pour chaque paire, lequel est comparé à des seuils de poids. Les poids et les seuils sont généralement rajustés après l'examen des résultats d'après des échantillons d'enregistrements.

## 2.4 Inconvénients du couplage d'enregistrements probabiliste

Lorsqu'il a été établi en 1997, le programme de mise à jour du Registre d'Élections Canada utilisait un logiciel commercial pour effectuer le couplage d'enregistrements probabiliste à l'aide de l'approche décrite ci-dessus. Au fil

du temps, on a observé plusieurs inconvénients dans la méthode, les plus importants d'entre eux violant certains des principes directeurs que nous souhaitons observer :

1. Les seuils de tolérance ne sont pas spécifiés directement, mais sont plutôt vaguement contrôlés en rajustant des valeurs seuils n'ayant aucune signification en-dehors du processus de couplage. Cela les rend très subjectifs, alors qu'ils devraient idéalement être complètement objectifs.
2. Les décisions fondées seulement sur le poids calculé font invariablement en sorte que de nombreuses paires sont acceptées tout en étant manifestement pires que certaines paires rejetées et que d'autres paires sont rejetées tout en étant manifestement meilleures que certaines paires rejetées. Une grande intervention est nécessaire pour rétablir la cohérence logique.
3. En raison de la complexité de la variation des noms en usage, il est utile d'examiner manuellement les cas présentant certains degrés de concordance partielle. Les poids calculés sont à eux seuls peu utiles pour déterminer les paires qui exigent un tel examen. Une grande intervention est nécessaire pour éviter d'ajouter une subjectivité inutile au processus.
4. L'inconvénient le plus important est la très grande sous-utilisation des renseignements disponibles qui sont pertinents pour prendre des décisions relatives au couplage, examinée en détail à la section 3.

# 3. Une solution de rechange au couplage d'enregistrements probabiliste

## 3.1 Renoncer à l'entière généralité

Il convient de souligner que l'approche probabiliste de couplage d'enregistrements décrite ci-dessus est entièrement générale du fait qu'elle n'exige pas de connaître les types de champs de données utilisés. Cette généralité a un coût très élevé, car il est très utile de connaître quelque chose à propos des champs utilisés aux fins du couplage.

Lorsqu'il est question de couplage d'enregistrements, l'entière généralité n'est pas non plus réellement très utile. La vaste majorité des projets de couplage entrent dans deux grandes catégories : les projets liés aux enregistrements comprenant des renseignements sur des entités personnelles, comme des noms, adresses et dates de naissance, et ceux liés aux enregistrements comprenant des renseignements sur des entités commerciales. Les approches de couplage pour ces grandes catégories, et toute autre catégorie d'intérêt plus restreinte, peuvent avoir beaucoup de choses en commun, mais elles n'ont certainement pas à être identiques.

## 3.2 Intégration des connaissances du sens commun

Toutes les activités de couplage d'enregistrements d'Élections Canada concernent des enregistrements de renseignements personnels, et les décisions à propos des paires d'enregistrements reposent dans une large mesure sur le degré de concordance de ces renseignements. Par conséquent, il pourrait s'avérer très profitable d'intégrer à ces décisions les raisons pour lesquelles les champs pourraient ne pas concorder pour les appariements vrais et les raisons pour lesquelles ils pourraient concorder pour les non-appariements vrais.

La méthode décrite à la section 2 utilise ce qui est observé à propos de la paire, les *Résultats*, pour calculer une valeur synthèse, le poids total, lequel est utilisé pour prendre une décision à propos de la paire. Les raisons de la concordance imparfaite des renseignements personnels sont complexes. Une mesure synthèse unique écarte beaucoup de renseignements pertinents. On propose plutôt, dans la mesure du possible, que les *Résultats* soient eux-mêmes utilisés pour décider directement si une paire doit être acceptée, rejetée ou gardée pour un examen plus approfondi, comme l'illustre le tableau simplifié présenté ci-dessous.

**Tableau 3.2-1**
**Règle de décision fondée directement sur les *Résultats* observés**

| Degré de concordance | | | Décision |
|---|---|---|---|
| Nom/sexe | Date de naissance | Adresse | |
| Élevé | Élevé | Élevé | Accepter |
| Élevé | Élevé | Faible | Examiner davantage |
| Élevé | Faible | Élevé | Examiner davantage |
| Élevé | Faible | Faible | Rejeter |
| Faible | Élevé | Élevé | Examiner davantage |
| Faible | Élevé | Faible | Rejeter |
| Faible | Faible | Élevé | Rejeter |
| Faible | Faible | Faible | Rejeter |
| Non considéré sérieusement en raison d'une concordance insuffisante | | | Rejeter |

Même avec plusieurs degrés de concordance partielle pour chaque champ, le grand nombre de combinaisons possibles en comprendra relativement peu qui seraient « acceptables » dans au moins certaines circonstances. Le reste, qui comprendrait la vaste majorité des paires possibles, pourrait être rejeté sans risque. Autrement dit, on peut supposer sans risque que la probabilité que ces paires soient des appariements vrais compte tenu des *Résultats* observés est de zéro.

Évidemment, au bout du compte, les combinaisons de *Résultats* jugées inacceptables sont subjectives. Cependant, on peut néanmoins s'appuyer sur des critères objectifs et appliquer la méthode de manière automatisée pour assurer la cohérence. Ces critères peuvent reposer en partie sur des exigences cocmmerciales. Par exemple, la nécessité que les électeurs fournissent une preuve d'identité au bureau de scrutin pourrait limiter la mesure dans laquelle la non-concordance du nom est autorisée pour les appariements acceptés dans certaines applications de couplage.

## 3.3 Une approche probabiliste alternative

Les plus grands inconvénients de l'approche probabiliste de couplage d'enregistrements décrite à la section 2 sont que les seuils utilisés sont subjectifs et, plus important encore, que les renseignements disponibles, pertinents aux décisions, ne sont pas facilement tenus en compte. L'approche suivante permet d'éviter ces deux inconvénients.

Dans le couplage d'enregistrements comprenant des renseignements personnels, les noms et les dates de naissance sont les principaux champs qui permettent d'identifier les personnes. Pour les appariements vrais, la date de naissance est le seul champ disponible qui ne peut pas ne pas concorder pour des raisons légitimes. Comparativement aux noms, la date de naissance présente aussi un nombre relativement restreint de degrés de concordance partielle utiles.

Soit k qui représente les différents degrés de concordance partielle de la date de naissance. Par exemple, nous pourrions autoriser trois degrés, la concordance, la concordance partielle ou la non-concordance. Évidemment, il faudrait définir clairement les différents degrés. Si nous faisons abstraction de tout ce que nous savons à propos de la paire d'enregistrements, nous obtenons

$$Résultats = Résultats\ k_{DOB} \cap Résultats_{Autres} \qquad (4)$$

Supposons également qu'il est possible d'observer toutes les paires ayant des $Résultats_{autres}$. Autrement dit, nous ne les avons pas écartées même si elles ont déjà été rejetées. Nous pourrions alors simplement compter le nombre de paires présentant chaque résultat lié à la date de naissance pour obtenir

$$t_k = Nombre\ de\ paires\ présentant\ Résultats\ k_{DOB} \cap Résultats_{Autres} \qquad (5)$$
$$r_k = t_k / \sum_k t_k \qquad (6)$$

Si nous pouvions également connaître d'une certaine façon

$$x_k = Nombre\ d'appariements\ vrais\ présentant\ Résultats\ k_{DOB} \cap Résultats_{Autres} \qquad (7)$$

la probabilité d'intérêt pourrait alors être calculée par définition.

Autrement dit :

$$P_M = \Pr(M|Résultats\ k_{DOB} \cap Résultats_{Autres}) \equiv {}^{x_k}/_{t_k} \tag{8}$$

$$P_U = \Pr(U|Résultats\ k_{DOB} \cap Résultats_{Autres}) \equiv {}^{(t_k - x_k)}/_{t_k} \tag{9}$$

Supposons maintenant que, sans connaître $x_k$, nous connaissons au moins

$$p_k = \Pr(Résultats\ k_{DOB}|M \cap Résultats_{Autres}) \equiv {}^{x_k}/_{\sum_k x_k} \tag{10}$$

$$q_k = \Pr(Résultats\ k_{DOB}|U \cap Résultats_{Autres}) \equiv {}^{(t_k - x_k)}/_{\sum_k (t_k - x_k)} \tag{11}$$

L'intégration des équations (6) et (10) à l'équation (11) pour résoudre $x_k$ et l'intégration de cette résolution à l'équation (8) permet d'obtenir

$$P_M = \Pr(M|Résultats) = \frac{\frac{q_k - r_k}{r_k}}{\frac{q_k - p_k}{p_k}} = \frac{Distance\ relative\ de\ q_k\ à\ r_k}{Distance\ relative\ de\ q_k\ à\ p_k} \tag{12}$$

Pour obtenir des probabilités valides, il est entendu que $r_k$ doit toujours se situer entre $p_k$ et $q_k$. Étant donné que $r_k$ est observé pour toutes les paires, soit une combinaison d'appariements vrais et de non-appariements vrais, cela devrait toujours s'avérer vrai si $p_k$ et $q_k$ sont connus avec raisonnablement de précision.

## 3.4 Spécifier les probabilités nécessaires

Le calcul des probabilités dans l'équation (12) nécessite les valeurs de $p_k$ et de $q_k$ spécifiées dans les équations (10) et (11), respectivement.

Dans le cas des appariements vrais, la date de naissance ne concorderait pas seulement en raison d'inexactitudes dans ce champ. Si l'exactitude des dates de naissance dans le Registre était mesurée, cela fournirait les estimations des valeurs de $p_k$ nécessaires. En fait, ces estimations ont été produites en 2014, d'après un petit échantillon de 49 000 enregistrements, et sont présentées dans le tableau ci-dessous.

Dans le cas des non-appariements vrais, on supposera que le degré de concordance pour la date de naissance est obtenu par pur hasard, indépendamment de toute autre considération. Pour une date de naissance donnée, on peut compter le nombre de personnes dans le Registre des électeurs ayant une date de naissance qui concorde parfaitement, qui concorde partiellement ou qui ne concorde pas et l'exprimer sous forme de fréquence relative. Cette dernière permet d'obtenir la probabilité qu'une nouvelle personne présente un degré de concordance précis pour la date de naissance avec une personne sélectionnée au hasard dans le Registre. Les valeurs pour un exemple type sont présentées dans le tableau ci-dessous.

**Tableau 3.4-1**
**Estimations de $p_k$ et de $q_k$ pour une date de naissance type (9 juillet 1963)**

| Degré de concordance | Appariement vrai ($p_k$) | Non-appariement vrai ($q_k$) |
|---|---:|---:|
| Concordance | 98,77 % | 0,01 % |
| Concordance partielle | 1,16 % | 0,29 % |
| Non-concordance | 0,07 % | 99,70 % |

## 3.5 Règles de décision finales

Une décision préliminaire pour les paires qui ne sont pas encore rejetées doit être fondée sur une combinaison de *Résultats* observés et sur la probabilité calculée résultante $P_M$. Certaines paires peuvent présenter des valeurs de *Résultats* qui justifient un examen manuel avant la prise de cette décision préliminaire. D'autres peuvent devoir être

acceptées pour des raisons opérationnelles, malgré une probabilité qui laisserait entendre le contraire. Enfin, la cohérence logique des valeurs de *Résultats* et des décisions préliminaires prises doit être vérifiée pour en arriver à une décision finale pour chaque paire.

## 3.6 Inclusion de tous les renseignements pertinents

Il convient de souligner que les $Résultats_{autres}$ ont jusqu'à maintenant été vaguement décrits comme étant tout ce qui est connu à propos de la paire à l'exception du degré de concordance pour la date de naissance. Dans le couplage probabiliste traditionnel, les probabilités doivent être estimées pour chaque champ inclus, les résultats pour ces champs étant présumés indépendants. Le retrait de ces deux exigences permet d'intégrer d'autres renseignements pertinents à propos des paires à ces règles de décision.

Par exemple, le degré de concordance pour l'adresse peut être spécifié de façon plus précise en intégrant tous les champs d'adresse pertinents, sans se préoccuper des violations de l'indépendance. Il n'est pas requis de limiter le nombre de degrés de concordance pour chaque champ et ces derniers doivent inclure les fréquences pertinentes pour minimiser la concordance due au hasard lorsque cela est indiqué. D'autres champs, qu'on juge peut-être de moindre importance aux fins du couplage, comme le statut (p. ex. actif ou décédé), peuvent aussi être inclus sans effort supplémentaire.

Plus important encore, les renseignements à propos d'autres paires peuvent aussi maintenant être intégrés facilement. Pour déterminer si une paire donnée doit être acceptée, il serait certainement pertinent de savoir que les enregistrements concernés étaient également en jeu dans d'autres paires « meilleures ». Tous les faits pertinents de ce genre peuvent simplement être ajoutés à la définition des $Résultats_{autres}$.

## 3.7 Estimation du taux d'erreur

Une fois qu'une paire est acceptée ou rejetée, seule une des deux erreurs de classification est possible. Une façon simple d'estimer le nombre d'erreurs de classification faites consiste à additionner simplement la probabilité des probabilités pertinentes pour toutes les paires.

Autrement dit :

$$Nombre\ de\ faux+ \cong \sum_{Acceptée} P_U = \sum_{Acceptée} (1 - P_M) \tag{13}$$

$$Nombre\ de\ faux- \cong \sum_{Rejetée} P_M \tag{14}$$

Il est à espérer que les probabilités pertinentes puissent être suffisamment bien estimées au moyen de cette méthode pour permettre de produire des estimations crédibles des taux d'erreur de classification dans le couplage pour un large éventail de projets de couplage d'enregistrements. Lorsque cette méthode d'estimation des taux d'erreur sera opérationnelle, il est à espérer que les résultats pourront être comparés aux méthodes d'estimation des taux d'erreur concurrentes afin d'en relever les différences.

## Bibliographie

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