

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

Utilisation de l'algorithme Relief pour la vérification et l'imputation dans le cadre du Recensement de la population du Canada

par Irwin Khuu

Date de diffusion : le 8 septembre 2025



Statistique
Canada

Statistics
Canada

Canada

Utilisation de l'algorithme Relief pour la vérification et l'imputation dans le cadre du Recensement de la population du Canada

Irwin Khuu¹

Résumé

Par le passé, le processus de vérification et d'imputation du Recensement de la population au Canada se déroulait à l'aide d'une méthodologie d'imputation par le plus proche voisin, selon laquelle la distance entre une unité rejetée et un donneur potentiel est obtenue au moyen d'une combinaison pondérée de variables auxiliaires. La révision du modèle entre les cycles peut être un processus compliqué et long, étant donné qu'il n'existe pas d'approche normalisée pour la sélection des variables et la pondération entre les sujets. La présente étude illustre la possibilité pour l'algorithme de sélection des variables Relief de créer une approche de sélection et de pondération des variables axée sur l'apprentissage automatique qui est normalisée et comparable entre les cycles de recensement et parmi les nombreux sujets du recensement. Un aperçu de la façon dont ce processus peut être appliqué en pratique sera présenté, suivi des résultats sur plusieurs sujets qui indiquent une amélioration générale par rapport aux méthodes précédentes.

Mots clés : Vérification et imputation; Recensement canadien; sélection de variables; Relief; apprentissage automatique.

1. Introduction

Le processus de vérification et d'imputation est un processus complexe utilisé pour le Recensement de la population du Canada, qui demande beaucoup de temps et comporte des centaines de règles de vérification pour un sujet donné du questionnaire. Dans le cadre du recensement, l'imputation est effectuée à l'aide du système généralisé de Statistique Canada connu sous le nom de Système canadien de vérification et d'imputation du recensement (SCANVIR), qui fonctionne à l'aide d'une méthodologie d'imputation par plus proche voisin. Cette méthodologie permet de déterminer efficacement les mesures minimales d'imputation de changements pour une imputation simultanée et multivariée par donneur pour une combinaison de variables quantitatives et qualitatives.

Les réponses sont réparties entre les enregistrements qui ne satisfont pas à une ou plusieurs règles de vérification et nécessitent par conséquent une imputation (enregistrements rejetés) et ceux qui satisfont aux règles de vérification (enregistrements acceptés). Pour chaque enregistrement rejeté, on trouve les enregistrements les plus semblables (donneurs potentiels). Les sous-ensembles de données des donneurs potentiels sont déterminés de sorte que lorsqu'ils sont utilisés pour imputer l'enregistrement rejeté, ils n'entraînent plus d'échecs aux règles de vérification (mesures d'imputation).

Par conséquent, la qualité de l'imputation dépend de la capacité du modèle d'imputation à mesurer avec succès la similitude entre les voisins. Dans SCANVIR, la distance entre une paire donneur et enregistrement échoué est définie par une combinaison pondérée de fonctions de distance :

$$D_{fp} = \sum_i w_i D_i(V_{fi}, V_{pi}) \quad (1)$$

où pour chaque variable du modèle i , $w_i \geq 0$ est le poids attribué et $D_i(V_{fi}, V_{pi})$ est une fonction sur $[0,1]$ représentant la distance entre la valeur de l'enregistrement rejeté V_{fi} et la valeur du donneur potentiel V_{pi} . Les facteurs de pondération permettent de préciser l'importance de la similitude entre les valeurs rejetées et de donneur potentiel pour chaque variable.

¹Irwin Khuu, Statistique Canada, 150, promenade Tunney's Pasture, Canada, K1A 0T6 (irwin.khuu@statcan.gc.ca)

Par conséquent, le problème de l'élaboration d'un modèle de SCANVIR de grande qualité se concentre sur la sélection et la pondération des variables auxiliaires. Dans le cadre du recensement canadien, les stratégies décisionnelles pour ce problème sont élaborées par sujet grâce à la collaboration des spécialistes du sujet et de la méthodologie. Il s'agit généralement d'un processus complexe et chronophage, d'autant plus compliqué par l'ajout potentiel de questions, de variables ou de changements entre les cycles du recensement. Par conséquent, l'élaboration d'un processus systématique et généralement applicable pouvant être normalisé et comparé entre les nombreux sujets du recensement mérite d'être approfondie.

Des études antérieures (Stelmack, 2019) ont déterminé que la famille d'algorithmes Relief était une stratégie prometteuse pour l'imputation du sujet de la catégorie d'admission dans le Recensement du Canada de 2016, surpassant d'autres méthodes de sélection d'entités comme le rapport d'accroissement de l'information et la forêt aléatoire. La présente étude étend l'utilisation des algorithmes Relief à d'autres sujets présents dans le Recensement de 2021 et compare leur rendement à celui des modèles du Recensement de 2021.

2. Aperçu de Relief

La famille d'algorithmes Relief est un ensemble d'algorithmes de sélection de caractéristiques qui, comme SCANVIR, utilisent l'approche du plus proche voisin pour déterminer l'importance d'une variable. Autrement dit, les variables auxiliaires utiles pour la prédiction d'une variable d'intérêt devraient avoir des valeurs semblables parmi les enregistrements ayant la même valeur dans la variable d'intérêt. L'algorithme Relief original (Kira et Rendell, 1992a) a généralement été supplanté par ReliefF (Kononenko, 1994), extension qui permet à l'algorithme de prendre en charge les données manquantes et les variables d'intérêt à catégories multiples, ainsi que de fournir généralement des estimations de poids plus fiables.

En général, pour une variable cible donnée, l'algorithme calcule les poids des variables auxiliaires en procédant à une itération au moyen d'une liste d'enregistrements et dans chaque cas (l'enregistrement cible) en trouvant les correspondances les plus proches (voisins ayant la même valeur de variable cible) et les non-correspondances (voisins ayant une valeur de variable cible différente). Pour chaque voisin, il met ensuite à jour le poids des variables auxiliaires, soit positivement si la variable auxiliaire diffère entre l'enregistrement de la cible et la non-correspondance la plus proche, soit négativement si la variable auxiliaire diffère entre l'enregistrement de la cible et la correspondance la plus proche. Le résultat est un vecteur de poids dans $[-1,1]$, où chaque élément représente la pertinence d'une variable auxiliaire, c.-à-d. sa capacité à prédire la variable cible. Les poids proches de zéro, ou négatifs, indiquent des prédicteurs non pertinents, qui peuvent être retirés d'un modèle de prédiction sans compromettre la qualité de la prédiction. Des explications plus détaillées figurent dans Kononenko (1994), Kira et Rendell (1992b) et Urbanowicz (2018).

Pour la mise en œuvre dans SCANVIR, les poids de Relief peuvent servir de mesure pertinente et peuvent être utilisés à la fois pour éliminer les variables qui semblent improductives pour prédire une variable cible donnée, ainsi qu'en étant transformés et appliqués comme poids d'imputation de SCANVIR.

3. Méthodologie

Pour chaque sujet examiné, des réponses de grande qualité ont été relevées afin de les utiliser pour l'entraînement et la mise à l'essai d'un modèle d'imputation fondé sur Relief. Dans chaque ensemble de données, les enregistrements sont divisés en un ensemble d'entraînement, pour entraîner le modèle Relief, et un ensemble d'essai, afin d'évaluer le rendement du modèle et de le comparer au modèle du Recensement de 2021.

La création d'un modèle Relief comprend à la fois la sélection de variables auxiliaires et la pondération. Pour la sélection des variables, l'utilisation de l'évaluation automatique de l'importance des variables permet l'introduction de nouvelles variables auxiliaires potentielles au modèle sans nécessiter de lourds efforts de la part des spécialistes du sujet et de la méthodologie pour déterminer la meilleure façon de les inclure dans le modèle d'imputation.

L'algorithme Relief lui-même est mis en œuvre dans R à l'aide du progiciel **CORElearn** (Robnik-Šikonja et Savicky, 2022).

La mise en œuvre de l'algorithme Relief ne tient compte que d'une seule variable cible dans son entraînement. Dans le cas d'une imputation multivariée, un modèle Relief distinct est généré pour chaque variable cible. Les poids sont normalisés et la moyenne est établie pour tous les modèles, afin d'obtenir un ensemble général de poids, puis ils sont remis à l'échelle en vue de leur mise en œuvre dans SCANVIR.

Une faible proportion des données de l'ensemble d'essais, selon le sujet, comporte une non-réponse artificielle induite dans la ou les variables d'intérêt. La non-réponse a été déterminée au moyen d'un mécanisme de propension à répondre par données manquantes entièrement au hasard (MCAR). Les non-réponses sont imputées dans SCANVIR à l'aide du modèle Relief avec sélection de variables et du modèle du Recensement de 2021. Les résultats sont vérifiés par rapport à la valeur réelle, afin d'obtenir des estimations de l'exactitude globale, du macrorappel et de la macroprécision. De plus, l'adéquation de la distribution a été vérifiée à l'aide d'un test du khi carré et d'une inspection de la statistique V de Cramér.

4. Ensembles de données

Les sujets du recensement examinés sont la langue officielle et la langue maternelle. Seules les réponses qui ne nécessitaient pas de vérifications déterministes ou d'imputation par donneur ont été utilisées pour l'analyse, afin d'estimer de façon fiable la qualité de notre modèle d'imputation par donneur sans être confondues par d'autres étapes de vérification et d'imputation. De plus, de nombreux modules d'imputation par donneur de SCANVIR comprennent leur propre stratification de la population étudiée. Ils sont également adoptés dans les critères de sélection du modèle Relief. Autrement dit, des modèles Relief distincts sont créés pour chaque strate unique.

4.1 Langue officielle

Le sujet des langues officielles impute une seule variable d'intérêt : les langues officielles connues de la personne. La réponse peut être l'anglais, le français, bilingue ou ni l'un ni l'autre. Quatre modèles Relief correspondant aux collectivités autochtones, aux logements collectifs, aux ménages comptant une seule personne et au reste de la population ont été créés. En raison de considérations relatives à la durée de l'entraînement, même si environ 5,65 millions de répondants étaient admissibles à l'entraînement du modèle Relief, un ensemble d'entraînement réduit et équilibré en catégorie contenant 1,42 million de réponses a été utilisé pour générer les modèles Relief. Environ 870 000 réponses ont été utilisées dans l'ensemble d'essais, y compris les collectivités autochtones, les logements collectifs et les membres de la population générale. Les données ont été imputées dans trois strates distinctes : collectivités autochtones, logements collectifs et population générale. Le taux de non-réponse dans l'ensemble d'essais a été établi à 2 % pour la population générale, tandis qu'en raison du petit nombre de cas, la non-réponse pour les collectivités autochtones et les logements collectifs a été établie à 10 %.

Le modèle du Recensement de 2021 utilisait sept variables auxiliaires, soit des données démographiques, administratives et des réponses sur la langue déjà traitées. En utilisant Relief, trois à cinq variables auxiliaires supplémentaires se sont révélées convenables, intégrant des renseignements supplémentaires au niveau de la famille de recensement et du ménage, des données géographiques, ainsi que des renseignements sur le statut d'Autochtone pour les répondants des collectivités autochtones.

4.2 Langue maternelle

Le sujet de la langue maternelle impute trois variables (deux variables binaires indiquant si l'anglais ou le français est la langue maternelle), ainsi qu'une variable écrite à cardinalité élevée pour les langues non officielles, avec des centaines de valeurs potentiellement imputables. Cinq modèles Relief correspondant aux collectivités autochtones dans les réserves, aux logements collectifs, aux familles de recensement comptant une seule personne, aux ménages comptant une seule personne et aux familles de recensement comptant deux enfants et deux adultes ont été créés. Les modèles Relief ont été entraînés sur un total de 2,03 millions d'enregistrements. Environ 1,33 million de réponses ont

été utilisées dans l'ensemble d'essais, y compris les collectivités autochtones dans les réserves, les logements collectifs et la population générale. Les données ont été imputées dans trois strates distinctes, soit les logements collectifs, les familles de recensement comptant une seule personne et les familles de recensement comptant deux adultes et deux enfants. Le taux de non-réponse dans l'ensemble d'essais a été établi à 7 % pour les logements privés, tandis qu'en raison du petit nombre de cas, la non-réponse pour les logements collectifs a été établie à 15 %.

Le modèle du Recensement de 2021 utilisait dix variables auxiliaires liées aux données administratives, à la réponse au questionnaire sur la langue à la maison et à l'information au niveau du ménage pour tous les répondants, et trois variables auxiliaires supplémentaires concernant le statut d'Autochtone pour les personnes vivant dans une réserve. Les modèles Relief comprenaient six autres variables, y compris des renseignements démographiques et géographiques ainsi que la réponse au questionnaire sur les langues officielles.

5. Résultats

Lors de la vérification des résultats du modèle, divers types de mesures de la qualité liées à l'imputation sont utilisés : exactitude au niveau de l'enregistrement, précision et rappel au niveau de la catégorie et adéquation de la distribution. Au niveau de l'enregistrement, l'exactitude est simplement déterminée en vérifiant si l'enregistrement avec réponse manquante a la même valeur imputée qu'à l'origine. Toutefois, dans les ensembles de données asymétriques selon les catégories, l'exactitude peut être une piètre mesure de la qualité de l'imputation. On se penche plutôt sur la précision au niveau de la catégorie suivante : la proportion d'enregistrements imputés correctement sur tous les enregistrements imputés; et pour le rappel, la proportion d'enregistrements imputés correctement sur tous les enregistrements de la catégorie. Comme l'analyse de catégories individuelles peut être exigeante pour les variables imputées à grande cardinalité, les macro-moyennes des valeurs de précision et de rappel parmi toutes les catégories sont fournies.

L'adéquation de la distribution est une mesure encore plus importante de la qualité de l'imputation en ce qui concerne les répercussions sur une analyse plus poussée. Il faut s'attendre à ce qu'en vertu du modèle de non-réponse MCAR utilisé ici, la distribution des données imputées corresponde de près à la distribution des données non imputées. L'adéquation de la distribution est mesurée en examinant les résultats d'un essai d'adéquation du khi carré sur les données imputées par rapport aux fréquences relatives dans les données non imputées. Le test du khi carré peut s'avérer peu fiable dans les cas de grande cardinalité avec un nombre de cellules faible ou nul, de sorte que des réalisations éparpillées de catégorie sont regroupées en une seule catégorie « Autre », le cas échéant. Dans le cadre d'une imputation bien adaptée, les statistiques obtenues lors du test du khi carré devraient être faibles; indiquant une correspondance étroite entre les données imputées et non imputées.

La qualité approximative de la distribution est examinée plus en détail à l'aide de la statistique V de Cramér. Bien qu'elle soit généralement utilisée pour mesurer l'association entre deux variables catégoriques, elle peut être adaptée pour jouer un rôle d'adéquation en comparant les fréquences observées aux chiffres attendus. Aucun test d'hypothèse particulier n'est associé à la mesure, mais comme pour le test du khi carré, on s'attend à ce que la statistique V de Cramér soit faible; les différences entre le modèle du Recensement de 2021 et le modèle Relief peuvent être évaluées de façon heuristique.

5.1 Langue officielle

Comme mentionné précédemment, le sujet des langues officielles impute les langues officielles connues de l'enregistrement, avec quatre réponses possibles : anglais, français, bilingue ou ni l'un ni l'autre. Les données ont été séparées en trois strates : collectivités autochtones, logements collectifs et population générale.

Le tableau 5.1-1 présente les résultats en matière d'exactitude, de macroprécision et de macrorappel entre les modèles de production de 2021 et Relief pour chaque strate. Dans toutes les strates, le modèle Relief dépasse le modèle du Recensement de 2021 pour toutes les mesures. L'inspection des résultats de l'imputation montre que, malgré la présence d'améliorations parmi les quatre catégories, elles étaient particulièrement répandues parmi les catégories « bilingues » et « ni l'une ni l'autre », qui présentaient des résultats de précision et de rappel particulièrement bas.

Tableau 5.1-1**Imputation de la langue officielle : exactitude, précision et rappel**

Strate	Exactitude		Macroprécision		Macrorappel	
	Production 2021	Relief	Production 2021	Relief	Production 2021	Relief
Collectivités autochtones	90,27 %	94,80 %	57,53 %	75,26 %	55,08 %	70,62 %
Logements collectifs	89,07 %	94,26 %	57,21 %	76,04 %	57,64 %	77,04 %
Population générale	82,87 %	89,88 %	70,48 %	82,43 %	68,60 %	81,77 %

Le tableau 5.1-2 fournit des mesures de l'adéquation de la distribution entre les modèles de production de 2021 et Relief pour chaque strate. Dans toutes les strates, tant pour le test du khi carré que pour la statistique *V* de Cramér, le modèle Relief dépasse le modèle du Recensement de 2021 en matière d'adéquation de la distribution. Plus particulièrement, pour la population générale, l'inspection des résultats d'imputation montre que le modèle du recensement surestime le nombre d'enregistrements de la catégorie « anglais », tout en sous-estimant les catégories « bilingue » et « ni l'un ni l'autre ».

Tableau 5.1-2**Imputation de la langue officielle : Test du khi carré et *V* de Cramér**

Strate	Modèle du Recensement de 2021				Modèle Relief			
	Test du khi carré			V de Cramér	Test du khi carré			V de Cramér
	Statistique de l'essai	dl	Valeur de p	Statistique	Statistique de l'essai	dl	Valeur de p	Statistique
Collectivités autochtones	5,869	3	0,12	0,0465	4,78	3	0,18	0,0420
Logements collectifs	4,953	3	0,18	0,0672	0,747	3	0,86	0,0260
Population générale	22,961	3	$4,12 \times 10^{-5}$	0,0212	2,216	3	0,53	0,0066

dl = degrés de liberté

5.2 Langue maternelle

Comme mentionné précédemment, le sujet de la langue maternelle impute trois variables : deux variables binaires pour les langues maternelles (anglais et français), ainsi qu'une variable écrite pour les langues non officielles. Les données ont été séparées en trois strates, soit les logements collectifs, les familles de recensement comptant une seule personne et les familles de recensement comprenant deux adultes et deux enfants. Bien que des résultats aient été obtenus pour toutes les variables imputables dans chaque strate, par souci de brièveté, seuls les résultats de l'imputation de la réponse écrite sont présentés ici.

Le tableau 5.2-1 présente les résultats en matière d'exactitude, de macroprécision et de macrorappel entre les modèles de production de 2021 et Relief pour chaque strate lors de l'imputation des réponses écrites dans la langue maternelle. Dans toutes les strates, le modèle Relief dépasse le modèle de 2021 pour toutes les mesures, surtout en matière de précision et de rappel, pour lesquels les résultats du modèle de 2021 étaient relativement médiocres.

Tableau 5.2-1**Imputation des réponses écrites sur la langue maternelle : exactitude, précision et rappel**

Strate	Exactitude		Macroprécision		Macrorappel	
	Production 2021	Relief	Production 2021	Relief	Production 2021	Relief
Logements collectifs	90,38 %	94,18 %	38,51 %	48,55 %	31,14 %	34,21 %
Familles de recensement comptant une seule personne	89,92 %	93,13 %	35,83 %	57,55 %	27,35 %	46,05 %
Familles de recensement comptant deux adultes et deux enfants	89,94 %	94,44 %	36,51 %	74,73 %	19,72 %	46,47 %

Le tableau 5.2-2 fournit des mesures de l'adéquation de la distribution entre les modèles de production de 2021 et Relief pour chaque strate lors de l'imputation des réponses écrites en matière de langue maternelle. En général, le modèle Relief présentait de légères améliorations pour les réponses écrites par rapport au modèle du recensement pour les logements privés, mais cette relation s'inversait pour les logements collectifs. Cela dit, un examen heuristique des mesures utilisées suggère qu'aucun des deux modèles n'a particulièrement bien réussi à maintenir les similitudes entre les données imputées et non imputées.

Tableau 5.2-2**Imputation des réponses écrites sur la langue maternelle : Test du khi carré et V de Cramér**

Strate	Modèle du Recensement de 2021			Modèle Relief			
	Test du khi carré			Test du khi carré			V de Cramér
	Statistique de l'essai	dl	Valeur de p	Statistique	Statistique de l'essai	dl	Valeur de p
Logements collectifs	30,881	25	0,193	0,0382	43,82	26	0,016
Familles de recensement comptant une seule personne	482,10	248	≈ 0	0,0066	261,71	248	0,149
Familles de recensement comptant deux adultes et deux enfants	740,13	213	≈ 0	0,0085	411,28	213	≈ 0

dl = degrés de liberté

Bien qu'elles ne soient pas présentées, des mesures de la qualité de l'imputation ont été obtenues pour l'imputation de la langue officielle et de la langue maternelle. En général, les mesures de l'exactitude, de la précision et du rappel étaient très élevées pour les deux modèles, même si des améliorations constantes étaient constatées avec le modèle Relief par rapport au modèle du Recensement de 2021. Toutefois, au niveau de la distribution, les résultats étaient généralement mitigés et aucun des deux modèles ne présentait d'avantage constant. De plus, selon les mesures utilisées, la capacité des modèles à maintenir l'adéquation de la distribution était en général quelque peu douteuse.

6. Discussion

Les comparaisons entre les modèles Relief et du Recensement de 2021 ont indiqué des améliorations générales au niveau de l'unité et par catégorie pour l'imputation univariée et multivariée. Au niveau de la distribution, le modèle Relief présentait également une amélioration par rapport au modèle de 2021 dans le cas univarié, tandis que les résultats n'étaient pas concluants dans les cas d'imputation multivariée complexes de grande cardinalité, comme celui portant sur la langue maternelle. Les résultats relatifs à l'adéquation de la distribution dans le cas d'une imputation multivariée doivent être pris avec une certaine prudence, car les mesures utilisées ne peuvent pas évaluer de façon fiable de multiples changements liés à la distribution de variables multiples simultanément et peuvent donc être conservatrices.

Il existe de nombreuses possibilités pour poursuivre d'autres recherches. Il convient de souligner que ces résultats ont été obtenus pour les cas de non-réponse par mécanisme MCAR. La qualité du modèle Relief, comparativement aux modèles de recensement, en ce qui concerne les mécanismes de non-réponse par données manquantes de façon aléatoire (MAR) ou manquante de façon non aléatoire (MNAR), est un sujet pour d'autres recherches, même si des recherches (Stelmack, 2019) suggèrent que ces résultats peuvent toujours être valides dans de tels cas. L'imputation de variables métriques, comme le revenu, n'est pas non plus présente dans cette analyse; l'algorithme ReliefF ne conviendrait pas dans ce cas, mais des algorithmes Relief fondés sur la régression comme RReliefF (Robnik-Šikonja et Kononenko, 2003) sont également disponibles dans le progiciel CORElearn et pourraient être envisagés. En ce qui concerne la sélection des variables, seules celles ayant des poids Relief négatifs ont été retirées du modèle d'imputation Relief, mais d'autres critères de décision peuvent être envisagés. Par exemple, Kira et Rendell (1992b) suggèrent un seuil d'inclusion fondé sur l'inégalité de Chebyshev. Enfin, l'analyse montre que l'algorithme Relief est encore imparfait pour maintenir l'adéquation de la distribution dans l'imputation multivariée. La question de savoir si cette situation peut être améliorée de façon fiable demeure non réglée.

Compte tenu de toutes ces mises en garde, les résultats montrent que l'inclusion de Relief dans la création de modèles d'imputation a un grand potentiel pour améliorer la qualité de l'imputation dans les recensements canadiens à venir. Dans le cas de problèmes d'imputation plus complexes, les produits du modèle Relief devraient servir principalement de données de référence pour conseiller les spécialistes du domaine et la méthodologie sur les décisions possibles concernant l'inclusion de variables, plutôt que d'être utilisés en bloc.

7. Remerciements

L'auteur tient à remercier les autres membres de l'équipe responsable de la méthodologie de vérification et d'imputation du recensement, qui ont fourni des commentaires précieux et leur intuition pour l'élaboration de cette recherche.

Bibliographie

- Kira, K. et L.A. Rendell (1992a), « A Practical Approach to Feature Selection », *Machine Learning Proceedings 1992*, p. 249-256.
- Kira, K. et L.A. Rendell (1992b), « The Feature Selection Problem: Traditional Methods and a New Algorithm », *Proceedings of the AAAI Conference on Artificial Intelligence*, 10, p. 129-134.
- Kononenko, I. (1994), « Estimating Attributes: Analysis and Extensions of Relief », *Machine Learning: ECML-94*, p. 171-182.
- Kononenko, I., E. Šimec et M. Robnik-Šikonja. (1997), « Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF », *Applied Intelligence*, 7, p. 39-55.

- Robnik-Šikonja, M. et I. Kononenko. (2003), « Theoretical and Empirical Analysis of ReliefF and RReliefF », *Machine Learning*, 53, p. 23-69.
- Robnik-Šikonja, M. et P. Savicky. (2022), CORElearn: Classification, Regression and Feature Evaluation. Version du progiciel R 1.57.3.1. <https://CRAN.R-project.org/package=CORElearn>.
- Stelmack, A. (2019), « Improving Edit and Imputation Strategies Through Feature Selection », *2019 Joint Statistical Meetings – Government Statistics Section*, p. 3083-3095.
- Urbanowicz R.J., M. Meeker, W. La Cava, R.S. Olson et J.H. Moore. (2018), « Relief-based feature selection: Introduction and review », *Journal of Biomedical Informatics*, 85, p. 189-203.