Nº 11-522-x au catalogue
ISSN : 1709822-X

Recueil du Symposium de 2021 de Statistique Canada
Adopter la science des données en statistique officielle pour répondre aux besoins émergents de la société

Progrès dans l'utilisation de l'information auxiliaire pour l'estimation à partir d'échantillons non probabilistes

par Ramón Ferri García

Date de diffusion : le 15 octobre 2021

Statistique Canada Statistics Canada

Canada

Receuil du Symposium de 2021 de Statistique Canada
Adopter la science des données en statistique officielle pour répondre aux besoins émergents de la société

Progrès dans l'utilisation de l'information auxiliaire pour l'estimation à partir d'échantillons non probabilistes

Ramón Ferri García[1]

Résumé

De récents développements des modes d'administration de questionnaires et d'extraction de données ont favorisé l'utilisation d'échantillons non probabilistes, qui présentent souvent un biais de sélection découlant d'un manque de plan de sondage ou de l'autosélection des participants. Ce biais peut être traité par plusieurs ajustements, dont l'applicabilité dépend du type d'information auxiliaire disponible. Des poids de calage peuvent être utilisés lorsque seuls des totaux de population de variables auxiliaires sont disponibles. En cas de disponibilité d'une enquête de référence respectant un plan de sondage probabiliste, plusieurs méthodes peuvent être appliquées, comme l'ajustement sur le score de propension, l'appariement statistique ou l'imputation de masse, ainsi que des estimateurs doublement robustes. En cas de disponibilité d'un recensement complet de la population cible pour certaines covariables auxiliaires, des estimateurs fondés sur des modèles de superpopulation (souvent utilisés en échantillonnage probabiliste) peuvent être adaptés au cas d'échantillonnage non probabiliste. Nous avons étudié la combinaison de certaines de ces méthodes, afin de produire des estimations moins biaisées et plus efficaces, ainsi que l'utilisation de techniques de prédiction modernes (comme la classification par apprentissage automatique et des algorithmes de régression) dans les étapes de modélisation des ajustements décrits. Nous avons en outre étudié l'utilisation de techniques de sélection de variables avant l'étape de modélisation de l'ajustement sur le score de propension. Les résultats indiquent que les ajustements fondés sur la combinaison de plusieurs méthodes peuvent améliorer l'efficacité des estimations et que l'utilisation de l'apprentissage automatique et de techniques de sélection de variables peut contribuer à réduire le biais et la variance des estimateurs dans une plus grande mesure dans plusieurs situations.

Mots clés : échantillonnage non probabiliste; calage; ajustement sur le score de propension; appariement.

1. Introduction

Les méthodes d'enquête classiques présentent d'importants inconvénients, en ce qui concerne les taux de réponse, les coûts et les erreurs de couverture, et de nouvelles méthodes d'administration des questionnaires pourraient y remédier. Les enquêtes en ligne et par téléphone intelligent en font partie. Les sources de données passives souvent utilisées dans les analyses des sources dites de mégadonnées (GPS, moissonnage du Web, applications mobiles) peuvent également être considérées comme de nouvelles méthodes d'obtention d'échantillons d'une population.

Les nouvelles méthodes permettent d'obtenir des échantillons plus actuels à un moindre coût et elles offrent de nombreuses autres possibilités dans la conception des questionnaires ou même le ciblage de strates non démographiques accessibles au moyen d'outils en ligne. Toutefois, elles comportent des erreurs non dues à l'échantillonnage qui doivent être prises en compte. Par exemple, les problèmes de connexion Internet ou d'informatisation du questionnaire peuvent entraîner des erreurs de mesure, et l'absence d'intervieweur peut augmenter les comportements de satisfaction de l'enquête. La source d'erreur la plus pertinente dans les nouvelles méthodes d'enquête est le biais de sélection, attribuable à trois mécanismes différents : l'erreur de couverture, l'erreur de réponse et le biais d'autosélection.

Ces mécanismes peuvent entraîner des biais importants si les caractéristiques des personnes participant aux échantillons diffèrent de celles des non-participants aux échantillons (Elliott et Valliant, 2017). La formule d'erreur élaborée par Meng (2018) montre que l'erreur d'estimation a trois sources : la quantité de données tirées de la population, la variabilité des données elles-mêmes et la corrélation entre le mécanisme de sélection et la variable d'intérêt. En cas de différences dans les valeurs de la variable d'intérêt entre les personnes échantillonnées et non

[1]Ramón Ferri García, Département de statistique et de recherche opérationnelle, Université de Grenade, Avenida de la Fuente Nueva S/N, Grenade, Espagne, 18071 (rferri@ugr.es)

échantillonnées, il y aura une forte corrélation qui entraînera une erreur d'estimation plus importante. Selon Meng (2018), le moyen le plus efficace de réduire l'erreur d'estimation n'est pas d'augmenter la taille de l'échantillon, ce qui entraînerait des améliorations très lentes, mais de réduire cette corrélation. C'est pourquoi certains ajustements ont été proposés dans la littérature pour réduire le biais de sélection. Ils sont décrits dans la section 2. À la section 3, nous décrivons certaines avancées récentes, puis nous examinons des pistes de recherche à la section 4.

2. Ajustements pour des échantillons non probabilistes

2.1 Cadre

Soit U une population cible de taille N, et $U_{pc} \subset U$ une population potentiellement couverte composée des unités pouvant appartenir à l'échantillon non probabiliste. Soit s_r un échantillon probabiliste de taille n_r tiré de U avec un plan de sondage donné et des poids de sondage d^r, et soit s_v un échantillon non probabiliste de taille n_v tiré de U_{pc} sans plan de sondage. Soit y une variable d'intérêt qui a été mesurée dans l'échantillon non probabiliste, mais non dans l'échantillon probabiliste, et $\boldsymbol{x}$ un ensemble de variables auxiliaires qui ont été mesurées dans les deux échantillons. Considérons une variable indicatrice R qui mesure l'inclusion d'une personne dans l'échantillon non probabiliste, de sorte que

$$R_i = \begin{cases} 1 & i \in s_v \\ 0 & i \notin s_v \end{cases}$$

2.2. Calage

Le calage est un ajustement qui ne nécessite pas d'échantillon probabiliste. Cette méthode a été élaborée à l'origine par Deville et Särndal (1992) pour obtenir des estimations plus efficaces, mais elle a ensuite été adaptée pour corriger la non-réponse aux enquêtes, et il lui suffit de connaître les totaux de population de $\boldsymbol{x}$. La méthode se fonde sur l'obtention d'un nouveau vecteur de poids, w, qui minimise une fonction de distance $g(.,.)$ par rapport aux poids de sondage d d'un échantillon donné s, tout en respectant les équations de calage

$$\sum_{i \in s} \boldsymbol{x}_i \, w_i = \sum_{i \in U} \boldsymbol{x}_i$$

selon lesquelles, lors de l'estimation des totaux de population des variables auxiliaires avec les nouveaux poids, ils devraient être exactement identiques aux totaux de population réels. Puisqu'il n'y a pas de plan de sondage dans les échantillons non probabilistes, il n'y a pas de poids de sondage pour lesquels la distance devrait être minimisée. Le choix habituel consiste à supposer des poids uniformes dans ce contexte, de sorte que $d = N/n_v$. Une étude par simulations réalisée par Bethléem (2010) a donné de bons résultats avec cette configuration.

2.3 Ajustement sur le score de propension

Un autre ajustement, fondé sur une méthode très populaire mise au point par Rosenbaum et Rubin (1983) pour effectuer une inférence sur des expériences non randomisées, est l'ajustement sur le score de propension (ASP). La conception théorique originale pour les enquêtes en ligne a été réalisée par Lee (2006) et elle continue d'être développée depuis. Dans l'ASP, nous supposons que les probabilités d'inclusion pour s_v, π, sont inconnues mais liées à un ensemble de covariables, de sorte que

$$\pi_i = P(R = 1 \mid \boldsymbol{x}_i), \; i \in U$$

Cette hypothèse nous permet d'estimer ces propensions au moyen d'un modèle prédictif M (habituellement une régression logistique) en l'ajustant pour prédire la variable indicatrice au moyen des deux échantillons regroupés.

$$\hat{\pi}_i^* = E_M[R^* = 1 \mid \boldsymbol{x}_i], i \in s_r \cup s_v,$$

où R^* est l'équivalent de la variable indicatrice R, mais en tenant compte seulement des échantillons regroupés de sorte que

$$R_i^* = \begin{cases} 1 & i \in s_v \\ 0 & i \in s_r \end{cases}$$

La variable R^* peut être considérée comme une mesure de R. On peut ensuite transformer les propensions estimées en poids, en utilisant la méthode de probabilité inverse $w_i = 1/\pi_i$ ou la version modifiée $w_i = (1 - \pi_i)/\pi_i$ étudiée dans Schonlau et Couper (2017), qui tient compte du fait que s_v n'appartient pas à la population cible de s_r. Les autres solutions possibles comprennent la stratification des propensions, qui permettent de classer les personnes ayant des propensions semblables dans une même strate. Certaines de ces solutions ont été présentées dans Lee et Valliant (2009) et Valliant et Dever (2011).

2.4 Estimateurs par modélisation de superpopulation

Les deux ajustements décrits dans les sections précédentes sont des ajustements fondés sur le plan, axés sur la modélisation des propensions à la participation, mais on trouve aussi dans la littérature des méthodes fondées sur un modèle. Dans ces méthodes, nous supposons que la population U est la réalisation d'une variable aléatoire de superpopulation qui est liée à certaines covariables au moyen d'une fonction donnée avec un modèle m et un terme d'erreur :

$$y_i = m(\boldsymbol{x}_i) + e_i, i \in U,\ e \sim N(0,\ \sigma^2)$$

Pour la classe d'estimateurs décrite dans cette sous-section, nous supposons qu'un recensement complet de U est disponible pour $\boldsymbol{x}$, ce qui signifie que nous pouvons considérer que l'échantillon probabiliste s_r observé dans notre étude est un recensement. Dans cette situation, nous pouvons utiliser les données de l'échantillon non probabiliste s_v pour ajuster un modèle prédictif SP à la variable d'intérêt y, puis l'utiliser afin de prédire ses valeurs pour l'ensemble de la population, de façon à obtenir le vecteur de nouvelles valeurs prédites, $\hat{y}$:

$$\hat{y}_j = E_{SP}\left[y_j \middle| \boldsymbol{x}_j\right], j \in U$$

On peut utiliser les valeurs prédites pour obtenir des estimations du total de la population de y, Y, en utilisant plusieurs formules, en remplaçant les échantillons probabilistes originaux, pour lesquels ces estimateurs ont été élaborés, par l'échantillon non probabiliste :

- l'estimateur fondé sur un modèle (Royall, 1970) :

$$\hat{Y}_{MB} = \sum_{i \in s_v} y_i + \sum_{i \in U - s_v} \hat{y}_i$$

- l'estimateur assisté par un modèle (Cassel et coll., 1976) :

$$\hat{Y}_{MA} = \sum_{i \in U} \hat{y}_i + \sum_{i \in s_v} w_i\,(y_i - \hat{y}_i)$$

- l'estimateur calé par un modèle (Wu et Sitter, 2001) :

$$\hat{Y}_{MC} = \sum_{i \in s_v} w^{MC}\, y_i,$$

où w^{MC} sont les poids de calage qui minimisent la distance avec les poids de sondage de s_v tout en respectant les équations de calage

$$\frac{1}{N}\sum_{i\in s_v} w^{MC} = 1, \sum_{i\in s_v} w^{MC}\,\hat{y}_i = \sum_{i\in U}\hat{y}_i$$

La restriction $\frac{1}{N}\sum_{i\in s_v} w^{MC} = 1$ peut être éliminée, ce qui donne un autre estimateur $\hat{Y}^*{}_{MC}$.

2.5 Appariement d'échantillons (imputation massive)

Dans une situation où l'échantillon probabiliste s_r n'est pas le recensement complet, mais seulement un sous-ensemble de la population (tiré selon un plan d'échantillonnage probabiliste), nous pouvons appliquer la même méthode fondée sur la prédiction des valeurs de y dans s_r, qui est connue dans la littérature sous le nom d'appariement statistique, d'appariement d'échantillons ou d'imputation massive. Cette méthode a d'abord été introduite par Rivers (2007) et Vavreck et Rivers (2008). Elle se fonde sur l'ajustement d'un modèle prédictif SM au moyen des données de l'échantillon non probabiliste s_v qu'elle applique dans l'échantillon probabiliste s_r pour prédire les valeurs non observées de la variable d'intérêt y :

$$\hat{y}_j = E_{SM}\left[y_j \middle| \boldsymbol{x}_j, R_j\right], j \in s_r$$

Les valeurs prédites peuvent être traitées comme des valeurs réelles dans les estimateurs habituels de la moyenne de la population $\bar{Y}$ et du total de la population avec les poids de sondage de l'échantillon probabiliste :

$$\hat{Y}_{SM} = \sum_{i\in s_r} d_i^r \hat{y}_i\,, \hat{\bar{Y}}_{HT}^{SM} = \frac{1}{N}\sum_{i\in s_r} d_i^r \hat{y}_i\,, \hat{\bar{Y}}_{H}^{SM} = \frac{\sum_{i\in s_r} d_i^r \hat{y}_i}{\sum_{i\in s_r} d_i^r}$$

2.6 Estimation doublement robuste

La méthode proposée récemment par Chen et coll. (2020) considère la combinaison de l'approche fondée sur le plan et de l'approche fondée sur un modèle dans leur estimateur doublement robuste. Dans cette estimation, on prédit les valeurs de la variable cible dans l'échantillon probabiliste et on les utilise comme dans l'appariement d'échantillons, mais on ajoute un terme qui tient compte des erreurs de prédiction, telles qu'elles sont observées dans l'échantillon non probabiliste, et on élève chaque erreur de prédiction en la pondérant par la propension inverse estimée avec l'ASP. L'estimateur doublement robuste de la moyenne de la population peut être défini comme suit :

$$\hat{\bar{Y}}^{DR} = \frac{1}{N}\sum_{i\in s_v} \frac{y_i - \hat{y}_i}{\hat{\pi}_i} + \frac{1}{N}\sum_{i\in s_r} d_i^r \hat{y}_i,$$

où on peut remplacer N par $\hat{N} = \sum_{i\in s_r} d_i^r$ pour obtenir l'estimateur de Hájek. Cette méthode est doublement robuste parce qu'elle est robuste en cas de spécifications erronées dans la prédiction de $\hat{\pi}$ faite dans l'ASP ou dans la prédiction de $\hat{y}$ faite dans l'appariement d'échantillons.

3. Progrès récents

3.1 Combinaison d'ajustement sur le score de propension (ASP) et de calage

Compte tenu de notre cadre théorique, certaines contributions ont été faites à partir de plusieurs méthodologies. La première (Ferri-García et Rueda, 2018) étudie la combinaison de l'ASP et du calage, en substituant les poids de sondage dans le calage par les poids de propension obtenus au moyen de l'ASP. Cette combinaison avait été étudiée dans la littérature (Lee et Valliant, 2009), mais dans cette contribution, nous avons examiné l'utilisation des totaux de population estimés dans le calage, au lieu des totaux réels, de sorte que les équations de calage sont

$$\sum_{i \in s_v} \boldsymbol{x}_i w_i = \sum_{i \in s_r} \boldsymbol{x}_i$$

Nous avons également étudié deux procédures différentes pour transformer les propensions en poids : la pondération de probabilité inverse $d_i = (1 - \hat{\pi}_i^*)/\hat{\pi}_i^*, i \in s_v$ et la stratification de la propension proposée dans Lee et Valliant (2009). La simulation comportait une population fictive avec quatre covariables et une variable d'intérêt représentant le vote pour trois partis politiques fictifs. L'estimation du pourcentage de vote pour chacun d'eux était assujettie aux mécanismes de valeurs manquantes entièrement au hasard (MCAR pour *Missing Completely At Random*), de valeurs manquantes au hasard (MAR pour *Missing At Random*) et de valeurs manquantes non dues au hasard (MNAR pour *Missing Not At Random*). Plusieurs configurations ont également été prises en compte pour les covariables.

Les résultats de la racine de l'erreur quadratique moyenne (REQM) observés dans la simulation sont indiqués dans le tableau 3.1-1. La REQM de la combinaison de méthodes pour les données MAR était légèrement inférieure à celle obtenue si l'on utilise seulement l'ASP, bien que l'erreur ait varié significativement entre les scénarios pris en compte dans la simulation. Les résultats semblent indiquer que l'utilisation de totaux estimés peut fonctionner aussi bien que l'utilisation de totaux de population réels, et que la combinaison est utile si les bonnes covariables sont utilisées.

Tableau 3.1-1
REQM moyenne des estimations dans les scénarios pris en compte dans Ferri-García et Rueda (2018)

REQM des estimations	Sans pondération	ASP	ASP + calage (totaux réels)	ASP + calage (totaux estimés)
Variable MAR	0,0367	0,0226	0,0224	0,0225
Variable MNAR	0,1171	0,1013	0,1023	0,1023

3.2 Combinaison d'ajustement sur le score de propension (ASP) et d'appariement

La combinaison de l'ASP et de l'appariement d'échantillons a également été envisagée (Castro-Martín et coll., 2021) d'une façon légèrement différente de l'estimateur doublement robuste, par l'utilisation des propensions estimées pour construire des modèles pondérés (où les poids d'entrée sont $w_i = \frac{1}{\hat{\pi}_i^*}, i \in s_v$), qui pourraient ensuite servir à prédire les valeurs de la variable cible. Cela signifie que l'estimateur est le même que dans l'appariement d'échantillons, mais que le modèle utilisé aux fins de prédiction est entraîné au moyen des poids déjà obtenus dans l'ASP. Notre étude par simulations a porté sur trois pseudo-populations ayant deux plans de sondage différents pour obtenir s_v dans chacune d'elles. Pour en savoir plus, voir Castro-Martín et coll. (2021). Les résultats concernant l'erreur quadratique moyenne (EQM) observée dans les simulations sont résumés dans le tableau 3.2-1. Ils semblent montrer que la combinaison décrite d'ASP et d'appariement pourrait atteindre les mêmes niveaux d'efficacité que l'estimateur doublement robuste, ou des niveaux quelque peu supérieurs.

Tableau 3.2-1
Efficacité moyenne et médiane (en pourcentage) de chaque méthode et nombre de fois qu'elle a été parmi les meilleures méthodes (EQM supérieure de moins de 1 % à l'EQM minimale) dans les simulations réalisées par Castro-Martín et coll. (2021)

Méthode	Moyenne	Médiane	Meilleure
ASP + appariement	65,8	66,4	18
Doublement robuste	64	65,2	18
Appariement	61,8	64,2	14
ASP	46,6	53,9	6

3.3 L'apprentissage automatique dans l'ASP

Pour ce qui est de l'ASP, nous avons étudié l'utilisation d'algorithmes de classification de l'apprentissage automatique (AA) dans l'estimation des propensions comme solution de substitution à la régression logistique (Ferri-García et

Rueda, 2020). À cette fin, deux études par simulations ont été réalisées au moyen de la même population fictive que dans Ferri-García et Rueda (2018) et d'une pseudo-population basée sur des données réelles de l'enquête espagnole sur les conditions de vie. Dans les deux études, les propensions ont été estimées au moyen d'un large éventail d'algorithmes prédictifs.

Les résultats de l'EQM obtenus dans la première étude par simulations se trouvent dans le tableau 3.3-1. Quand le mécanisme de sélection est MCAR, les arbres de décision donnent de meilleurs résultats pour ce qui est de l'erreur quadratique moyenne, en particulier l'algorithme des arbres de classification et de régression (CART pour *Classification And Regression Trees*). Quand le mécanisme de sélection est MAR, on peut observer que, bien que la régression logistique donne de bons résultats, la méthode des k-plus proches voisins (K-NN) et surtout l'algorithme Gradient Boosting Machine (GBM) pourraient être utiles dans des situations diverses. Enfin, quand le mécanisme de sélection est MNAR, la forêt aléatoire est le meilleur choix pour l'estimation des propensions.

Tableau 3.3-1
EQM moyenne fournie par chaque algorithme et nombre de fois qu'elle a été parmi les meilleures (EQM inférieure à 1 % de plus que l'EQM minimale) dans la simulation sur la population fictive de Ferri-García et Rueda (2020)

Variable d'intérêt	Mesure	Régression logistique	C4.5	C5.0	CART	K-NN (k-plus proches voisins)	Bayésien naïf	Forêt aléatoire	GBM
Variable MCAR	EQM moyenne	1,3	1,2	1,0	0,7	1,5	3,1	11,7	1,0
	Meilleure	0	4	3	11	0	0	0	5
Variable MAR	EQM moyenne	3,7	38,0	42,8	61,3	4,5	16,5	67,5	26,6
	Meilleure	12	0	0	0	1	0	0	10
Variable MNAR	EQM moyenne	102,9	166,9	175,7	201,3	90,5	73,2	75,3	144,9
	Meilleure	0	0	0	0	2	1	19	0

3.4 L'apprentissage automatique dans la modélisation de superpopulation

L'utilisation de l'apprentissage automatique dans la modélisation de superpopulation est examinée par Ferri-García et coll. (2021) dans une situation où un recensement complet de la population est disponible. Ces estimateurs dépendent fortement de la spécification du modèle. Et, de fait, on observe que le choix du modèle est le facteur le plus important, indépendamment d'autres facteurs comme la taille de l'échantillon ou la formule utilisée dans l'estimation. Dans une étude par simulations utilisant les mêmes pseudo-populations que dans Castro-Martín et coll. (2021), nous avons constaté que les modèles de régression pénalisée comme Ridge ou elastic-net (Glmnet) étaient meilleurs que d'autres solutions, bien que certains algorithmes comme K-NN donnaient de bons résultats pour certaines situations.

Tableau 3.4-1
Efficacité moyenne et médiane (en pourcentage) de chaque méthode avec chaque algorithme et nombre de fois qu'elle a été parmi les meilleures méthodes (EQM supérieure de moins de 1 % à l'EQM minimale) dans les simulations réalisées par Ferri-García et coll. (2021). AM = Assistée par un modèle; FM = Fondée sur un modèle; CM = Calée par un modèle.

Méthode	Algorithme	Moyenne	Médiane	Meilleure	Méthode	Algorithme	Moyenne	Médiane	Meilleure
AM	Régression Ridge	62,2	64,3	13	CM	Glmnet	61,5	63	12
FM	Régression Ridge	61,9	64,1	12	FM	Glmnet	61,3	63	9

AM	GLM	61,7	64,3	12	CM	K-NN (k-plus proches voisins)	59,1	53,1	7
FM	GLM	61,7	64,1	12	AM	K-NN (k-plus proches voisins)	58,5	52,7	7
CM	GLM	61,7	64,3	12	CM	LASSO bayésien	58,5	61,3	10
CM	Régression Ridge	61,6	62,8	11	AM	LASSO bayésien	58,2	61,2	11
AM	Glmnet	61,6	62,8	11	FM	Réseaux neuronaux régularisés bayésiens	57,9	61,8	8

3.5 Sélection automatisée des variables avant l'ASP

La littérature sur la pondération des scores de propension (Hirano et Imbens, 2001; Brookhart et coll., 2006) souligne l'importance du choix des covariables et conclut que le meilleur choix consiste à inclure des covariables liées à la variable d'intérêt. Cependant, on ne connaît pas toujours les associations ou les relations de causalité avant la procédure d'estimation, de sorte qu'on doit apprendre des données afin de sélectionner certaines variables et en éliminer d'autres. C'est pourquoi nous avons mené une étude sur l'utilisation d'algorithmes automatisés de sélection des variables afin d'améliorer les résultats de l'ASP (Ferri-García et Rueda, 2021). Nous avons utilisé certains sélecteurs pour la régression linéaire (stepwise, LASSO), les filtres (khi carré, CFS, OneR) et les mesures d'importance (forêt aléatoire, Boruta), que nous avons appliqués dans deux études par simulations, en utilisant des données fictives et des données réelles d'une enquête menée par le Centre espagnol de recherche sociologique.

Le tableau 3.5-1 présente certains résultats d'efficacité de la deuxième étude par simulations. Pour ce qui est de l'efficacité, certains algorithmes de sélection de variables, en particulier OneR ou CFS, sont associés à des estimations plus efficaces. Cet avantage peut s'expliquer par la plus faible variance des estimateurs, qui sont encore moins biaisés qu'une situation où toutes les variables sont utilisées quand l'ASP est conjugué au calage.

Tableau 3.5-1
Efficacité moyenne et médiane des estimations (EQM de l'algorithme/EQM en utilisant toutes les variables) et nombre de fois que son efficacité a été inférieure à 1 ou 0,9 dans l'étude par simulations utilisant des données réelles réalisée par Ferri-García et Rueda (2021). Calage du ratissage appliqué après l'ASP.

Algorithme	Moyenne	Médiane	Efficacité < 1	Efficacité < 0,9
Boruta	1,028	1,004	22	1
CFS	0,950	0,943	38	11
Khi carré	0,968	0,942	36	11
Ratio de gain	0,983	0,955	37	9
LASSO	0,976	0,956	34	7
Stepwise	1,015	1,010	21	2
OneR	0,965	0,943	42	14
Importance de forêt aléatoire	0,991	0,973	34	4

4. Pistes de recherche

Il faudrait envisager d'approfondir les recherches sur l'estimation à partir d'échantillons non probabilistes utilisant de l'information auxiliaire. L'inclusion de poids de sondage dans les modèles prédictifs pour l'estimation de la propension devrait être étudiée. Bien qu'un estimateur convergent comportant des poids de sondage ait été élaboré par Chen et coll. (2020) pour la régression logistique, il se peut que d'autres stratégies de pondération conviennent mieux à d'autres modèles paramétriques et non paramétriques. L'autre question importante sur laquelle il faudrait se

pencher est l'atténuation du biais produit par le mécanisme MNAR. En effet, le biais qu'il produit est le plus difficile à traiter, d'après les connaissances actuelles. Les autres pistes de recherche comprennent la définition de propriétés théoriques et l'inclusion d'autres stratégies de prétraitement des données, comme l'équilibrage des classes ou le réglage des hyperparamètres, qui sont courantes en science des données et pourraient être utiles dans l'estimation à partir d'échantillons non probabilistes.

Bibliographie

Bethlehem, J. (2010), « Selection bias in web surveys », *International Statistical Review*, 78(2), p. 161-188.

Brookhart, M. A., S. Schneeweiss, K. J. Rothman, R. J. Glynn, J. Avorn, et T. Stürmer (2006), « Variable selection for propensity score models », *American journal of epidemiology*, 163(12), p. 1149-1156.

Cassel, C. M., C. E. Särndal, et J. H. Wretman (1976), « Some results on generalized difference estimation and generalized regression estimation for finite populations », *Journal of the American Statistical Association*, 63(3), p. 615-620.

Castro-Martín, L., M. Rueda, et R. Ferri-García (2021), « Combining Statistical Matching and Propensity Score Adjustment for inference from non-probability surveys », *Journal of Computational and Applied Mathematics* (sous presse).

Chen, Y., P. Li, et C. Wu (2020), « Doubly robust inference with nonprobability survey samples », *Journal of the American Statistical Association*, 115(532), p. 2011-2021.

Deville, J.-C., et C.-E. Särndal (1992), « Calibration Estimators in Survey Sampling », *Journal of the American Statistical Association*, 87(418), p. 376-382.

Elliott, M. R., et R. Valliant (2017), « Inference for nonprobability samples », *Statistical Science*, 32(2), p. 249-264.

Ferri-García, R., et M. Rueda (2018), « Efficiency of propensity score adjustment and calibration on the estimation from non-probabilistic online surveys », *SORT*, 42(2), p. 159-182.

Ferri-García, R., et M. Rueda (2020), « Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys », *Plos one*, 15(4), e0231500.

Ferri-García, R., et M. Rueda (2021), « Variable selection in Propensity Score Adjustment to mitigate selection bias in online surveys », *Statistical Papers* (en cours de révision).

Ferri-García, R., L. Castro-Martín, et M. Rueda (2021), « Evaluating Machine Learning methods for estimation in online surveys with superpopulation modeling », *Mathematics and Computers in Simulation*, 186, p. 19-28.

Hirano, K., et G. W. Imbens (2001), « Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization », *Health Services and Outcomes research methodology*, 2(3), p. 259-278.

Lee, S. (2006), « Propensity score adjustment as a weighting scheme for volunteer panel web surveys », *Journal of official statistics*, 22(2), p. 329-349.

Lee, S. et R. Valliant (2009), « Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment », *Sociological Methods & Research*, 37(3), p. 319-343.

Meng, X. L. (2018), « Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election », *The Annals of Applied Statistics*, 12(2), p. 685-726.

Rivers, D. (2007), « Sampling for web surveys », présentation aux Joint Statistical Meetings, Salt Lake City (Utah).

Rosenbaum, P. R., et D. B. Rubin (1983), « The central role of the propensity score in observational studies for causal effects », *Biometrika*, 70(1), p. 41-55.

Royall, R. M. (1970), « On finite population sampling theory under certain linear regression models », *Biometrika*, 57(2), p. 377-387.

Schonlau, M., et M. P. Couper (2017), « Options for conducting web surveys », *Statistical Science*, 32(2), p. 279-292.

Valliant, R., et J. A. Dever (2011), « Estimating propensity adjustments for volunteer web surveys », *Sociological Methods & Research*, 40(1), p. 105-137.

Vavreck, L., et D. Rivers (2008), « The 2006 cooperative congressional election study », *Journal of Elections, Public Opinion and Partis*, 18(4), p. 355-366.

Wu, C., et R. R. Sitter (2001), « A Model-Calibration Approach to Using Complete Auxiliary Information from Survey Data », *Journal of the American Statistical Association*, 96, p. 185-193.