Recueil du Symposium 2014 de Statistique Canada
Au-delà des méthodes traditionnelles d'enquêtes : l'adaptation à un monde en évolution

Au sujet des corrections du biais pour les enquêtes en ligne

Lingling Fan, Wendy Lou et Victoria Landsman[1]

Résumé

Les enquêtes en ligne excluent l'entièreté de la population sans accès à Internet et ont souvent de faibles taux de réponse. Par conséquent, l'inférence statistique fondée sur des échantillons d'enquêtes en ligne requiert que soit disponible de l'information supplémentaire sur la population non couverte, que les méthodes d'enquête soient choisies avec précaution afin de tenir compte des biais possibles, et que l'interprétation et la généralisation des résultats à une population cible se fassent prudemment. Dans le présent article, nous nous concentrons sur le biais de non-couverture, et explorons l'utilisation d'estimateurs pondérés et d'estimateurs par imputation hot-deck pour corriger le biais sous le scénario idéal où l'information sur les covariables a été obtenue pour un échantillon aléatoire simple de personnes faisant partie de la population non couverte. Nous illustrons empiriquement les propriétés des estimateurs proposés sous ce scénario. Nous discutons d'extensions possibles de ces approches à des scénarios plus réalistes.

Mots clés : Enquêtes en ligne à participation volontaire; non-couverture; pondération; imputation.

1. Introduction

Dans les enquêtes en ligne fondées sur l'échantillonnage probabiliste, chaque unité de la population cible possède une probabilité positive connue d'être échantillonnée; donc, ces enquêtes peuvent être utilisées pour faire des inférences valides au sujet de la population. Pour les appliquer, nous devons d'abord définir la population et la base de sondage, puis générer un échantillon aléatoire. Cependant, dans plusieurs circonstances, il est difficile d'identifier les unités de la population cible et de contacter un échantillon probabiliste de la population (Alvarez et VanBeselaere, 2005). Par exemple, dans les enquêtes en ligne visant tous les électeurs admissibles, la base de sondage n'existe pas, puisque les répondants potentiels n'ont pas tous accès à Internet, ce qui empêche de réaliser des enquêtes fondées sur un échantillonnage probabiliste.

Dans les enquêtes en ligne non probabilistes, il est plus difficile de faire des inférences statistiques valides au sujet de la population. Ces enquêtes s'appuient souvent sur des panels volontaires, c'est-à-dire qu'un recrutement « ouvert à tous » est utilisé en premier lieu pour créer un panel volontaire à partir duquel seront ensuite sélectionnés aléatoirement les participants à l'enquête par échantillonnage probabiliste. Généraliser au sujet de la population en se fondant sur les résultats d'une enquête à participation volontaire peut poser problème, puisque le panel initial est un échantillon autosélectionné; les personnes qui n'ont pas accès à Internet ou celles qui ne se sont pas inscrites au panel volontaire ne seront jamais échantillonnées, ce qui cause un biais de non-couverture.

L'objectif de la présente étude est de se pencher sur la performance des techniques statistiques pour tenir compte du biais de non-couverture dans les enquêtes en ligne à participation volontaire. En particulier, nous nous concentrons sur les méthodes fondées sur le score de propension et les méthodes d'imputation, qui ont été employées à grande échelle pour traiter l'échantillonnage biaisé dans divers domaines de la statistique appliquée (p. ex. Valliant et Dever, 2011; Andridge et Little, 2010; Chen et Jun, 2000).

[1] 1er auteur : Lingling Fan, Université de Toronto, 100, St. George Street, Département de sciences statistiques, Toronto (Ontario) M5S 3G3; 2e et 3e auteurs : Wendy Lou et Victoria Landsman, Université de Toronto, École de santé publique Dalla Lana, Édifice des sciences de la santé, 155, College Street, Toronto (Ontario) M5T 3M7.

2. Cadre des enquêtes en ligne à participation volontaire

Soit une population finie U de N individus, et soit y_k une mesure du résultat et $X_k = (X_{k1}, X_{k2}, \dots, X_{kp})'$, un vecteur de dimension p de variables explicatives (covariables) pour l'unité k, $k = 1, \dots, N$. Deux variables indicatrices, W_k et V_k, définissent respectivement l'accès à Internet et la volonté de participer à un panel volontaire : $W_k = 1$ si l'unité k a accès à Internet et $W_k = 0$ autrement; $V_k = 1$ si l'unité k choisit de participer à un panel volontaire et $V_k = 0$ autrement. Il est commode de présenter la population cible comme l'union de deux ensembles disjoints : un panel volontaire, V, c'est-à-dire un ensemble de tous les individus k tels que $V_k = 1$ de taille N_1, et un ensemble complémentaire, V^c de taille N_0. Notons que les individus dans V^c pourraient ou non avoir accès à Internet, tandis que chaque individu dans le panel volontaire V a accès à Internet. Soit S_V le sous-échantillon tiré d'un panel volontaire, avec s_k la probabilité que le volontaire k soit sélectionné dans S_V, c'est-à-dire $s_k = P(k \in S_V | V_k = 1, W_k = 1)$; soit π_k la probabilité que l'unité k participe à l'enquête en ligne, qui reflète la décision d'un participant k de participer à une enquête. Dans la présente étude, nous supposons que les probabilités s_k sont connues, qu'il n'y a pas de non-réponse et que chaque unité ayant accès à Internet s'est inscrite au panel volontaire (c'est-à-dire que le panel volontaire et l'ensemble d'unités ayant accès à Internet sont identiques). Sous ces trois hypothèses, les probabilités π_k peuvent être calculées selon (Valliant et Dever, 2011)

$$\pi_k = P(V_k = 1 | X_k) s_k$$

où les probabilités $P(V_k = 1 | X_k)$ sont inconnues et doivent être estimées.

Supposons que S_V est un échantillon aléatoire simple de participants volontaires de taille n_1 tiré d'un panel volontaire V. Notre objectif est d'estimer la moyenne de population $\bar{Y}$ pour l'ensemble de la population $U = V \cup V^c$ d'après cet échantillon. Pour estimer le paramètre cible, de l'information supplémentaire est nécessaire sur la population non couverte V^c. Différents scénarios peuvent être envisagés. Dans la présente étude, nous supposons que nous sélectionnons un échantillon aléatoire simple S_R de taille n_0 à partir de V^c contenant l'information sur les covariables uniquement. Cet échantillon est appelé échantillon de référence (figure 1 à droite).

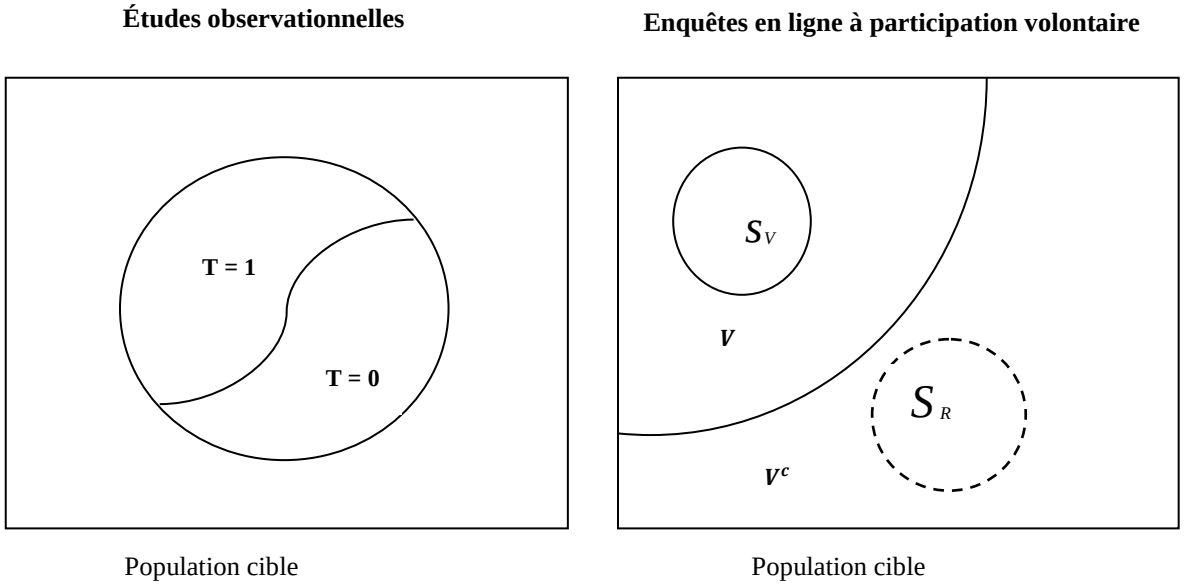

Figure 1 : Etudes observationnelles (à gauche) et enquêtes en ligne à participation volontaire sous le scénario idéal (à droite)

Nos simulations sont motivées par les données de sondages électoraux simulées de Bethlehem (Bethlehem, 2010), où la population cible U comprend $N = 30\,000$ unités âgées de 18 à 80 ans avec 19 058 (63,5 %) volontaires. Nous avons constaté que 35,2 % des unités votaient pour le Nouveau Parti Internet (NPI) dans l'ensemble de la population, que 45,6 % votaient pour le NPI parmi les volontaires, et que les jeunes natifs étaient plus susceptibles de participer à des panels volontaires et de voter pour le NPI.

3. Méthodes

3.1 Estimateur pondéré

Un estimateur pondéré (de type Horvitz-Thompson) est fréquemment utilisé par les statisticiens d'enquête pour estimer les quantités dans une population finie à partir d'un échantillon donné (p. ex., Valliant et Dever, 2011). Dans le scénario idéal mentionné plus haut, l'estimateur pondéré de la moyenne de population $\bar{Y}$ peut s'écrire en utilisant l'information résultante disponible pour l'échantillon enquêté S_V uniquement, sous la forme

$$\hat{\bar{Y}}_{HT} = \frac{\sum_{k \in S_V} y_k \hat{w}_k}{\sum_{k \in S_V} \hat{w}_k}, \ldots\ldots(3,1)$$

où les poids sont définis comme étant $\hat{w}_k = (\hat{\pi}_k s_k)^{-1}$.

Les inverses des s_k sont souvent appelés poids de base dans la littérature sur les sondages (Valliant et Dever, 2011). Pour le scénario utilisé dans la présente étude, les poids de base sont égaux à $\frac{N_1}{n_1}$ et $\frac{N_0}{n_0}$ pour chaque individu dans l'échantillon enquêté et dans l'échantillon de référence, respectivement.

Les probabilités π_k reflètent le mécanisme d'autosélection d'une personne en vue de participer au panel volontaire. Puisque l'échantillon enquêté S_V et l'échantillon de référence S_R sont disjoints sous le scénario utilisé, les probabilités π_k peuvent être considérées comme les scores de propension où S_V sert de branche de traitement et S_R de branche de contrôle (Rosenbaum et Rubin, 1983) dans les études observationnelles (voir la figure 1 pour la comparaison entre l'étude observationnelle et l'étude en ligne à participation volontaire). Ces probabilités sont inconnues et doivent être estimées. En utilisant l'information sur les covariables disponibles pour les individus de l'échantillon de référence, les probabilités π_k peuvent être estimées en ajustant des modèles de régression à l'échantillon joint $S_V \cup S_R$. Le plus souvent, la régression logistique est le modèle privilégié, et nous avons choisit de l'utiliser également.

Il est important de souligner que le plan d'échantillonnage utilisé pour obtenir l'échantillon joint $S_V \cup S_R$ est informatif, puisque les individus enquêtés et l'échantillon de référence ont été sélectionnés avec des probabilités s_k différentes. Dans de tels cas, un modèle de régression logistique pondéré dans lequel sont intégrés les inverses des s_k (poids de base) est requis pour obtenir des estimations valides des π_k (Valliant et Dever, 2011).

3.2 Estimateur imputé

L'imputation est une méthode fréquemment utilisée pour traiter la non-réponse partielle. Dans cette méthode, on tente de créer un ensemble complet de données comblant les valeurs manquantes. Diverses méthodes d'imputation, y compris des méthodes d'imputation statistiques et d'imputation sous apprentissage machine ont été élaborées. L'imputation hot-deck est amplement appliquée en pratique; par exemple, elle est très souvent utilisée par les organismes statistiques publics et les organismes spécialisés dans les sondages. Dans l'imputation hot-deck, les valeurs manquantes pour un non-répondant (un receveur) sont remplacées par les valeurs observées d'un répondant (un donneur) dont les caractéristiques sont similaires à celles du non-répondant (Andridge et Little, 2010).

Diverses approches peuvent être utilisées pour définir des groupes d'unités similaires, dont la création de classes d'imputation et des méthodes de mesure de distance. Dans la création de classes d'imputation, les répondants et les non-répondants sont affectés à des classes en se fondant sur les variables auxiliaires. Dans les méthodes de mesure de distance, une mesure de distance sert à évaluer le degré de proximité des donneurs potentiels par rapport aux receveurs. La mesure de distance est choisie en fonction de la nature des variables auxiliaires utilisées dans l'imputation.

Il existe plusieurs méthodes d'imputation hot-deck, et nous en avons choisies deux pour la présente étude : l'imputation par la méthode du plus proche voisin (IPPV) et l'imputation hot-deck aléatoire (HDA) (pondérée). Dans la méthode HDA, pour chaque receveur, un donneur est sélectionné aléatoirement dans chaque classe d'imputation, et la valeur du donneur est attribuée au receveur. La seule différence entre l'imputation HDA et l'imputation HDA pondérée est que les poids de base pour les unités figurant dans l'ensemble de donneurs sont

intégrés dans l'imputation. Dans le cas de la méthode IPPV, la valeur manquante y_j associée à chaque receveur j est imputée par une valeur y_i provenant du donneur, où le donneur i est le plus proche voisin de j mesuré par les variables X en utilisant la mesure de distance de Gower (puisque nous avons une imputation des variables continues ainsi que catégoriques). Ces deux méthodes d'imputation hot-deck d'usage répandu sont choisies pour plusieurs raisons. Premièrement, les deux méthodes possèdent plusieurs bonnes propriétés : les valeurs imputées sont des valeurs réelles; les variables auxiliaires qui interviennent dans les variables d'appariement sont utilisées dans l'imputation; aucun modèle explicite reliant Y et X n'est utilisé. Deuxièmement, la théorie des estimateurs par la méthode IPPV est bien établie (Chen et Jun, 2000) : Chen et Jun ont prouvé que, sous certaines conditions de régularité sur la distribution des covariables et le mécanisme de réponse, les moyennes d'échantillon IPPV et toute fonction lisse des moyennes d'échantillon sont asymptotiquement sans biais; la distribution empirique et les estimateurs de quantiles IPPV sont asymptotiquement sans biais; Chen et Jun ont également obtenu la variance approximative des estimateurs sous IPPV. Troisièmement, l'imputation hot-deck aléatoire possède certaines propriétés bien connues (p. ex., Rubin, 1987); notamment, elle peut fournir une distribution et des estimateurs de quantiles valides, mais peut ne pas être aussi efficiente que l'imputation IPPV (Chen et Jun, 2000).

Dans la méthode d'imputation, nous utilisons les méthodes d'imputation hot-deck décrites ci-haut pour imputer la variable cible y_j, pour $j \in S_R$ sous le scénario idéal (figure 1), puis nous estimons la moyenne de la population en utilisant les données imputées et les données de l'échantillon en ligne à participation volontaire en tant que données finales :

$$\hat{\bar{Y}}_{IMP} = \frac{N_1}{N}\bar{y}_1 + \frac{N_0}{N}\bar{y}_0, \ldots\ldots (3,2)$$

où $\bar{y}_1 = \frac{1}{n_1}\sum_{i \in S_V} y_i$, $\bar{y}_0 = \frac{1}{n_0}\sum_{j \in S_R} \hat{y}_j$, avec $\hat{y}_j, j \in S_R$, les valeurs imputées.

4. Résultats des simulations

Afin d'évaluer la performance des estimateurs pondérés et imputés pour l'estimation de la moyenne de population, nous nous sommes servis du biais relatif moyen (BRM) donné par $BRM = \frac{\bar{\theta}-\theta}{\theta} * 100\ \%,\ \bar{\theta} = \frac{1}{M}\sum_{m=1}^{M}\theta_m$ où M est le nombre de simulations, θ_m est la moyenne de la population estimée, calculée d'après les $m-th$ simulations, et θ est la moyenne réelle de la population. Comme le biais des estimateurs dépend des tailles de l'échantillon de volontaires et de l'échantillon de référence, diverses combinaisons de tailles d'échantillon de volontaires et d'échantillon de référence ont été utilisées dans la simulation. Pour chaque combinaison, le nombre de simulations est fixé à 1 000 ($M = 1\ 000$), et le BRM empirique est calculé sur l'ensemble des 1 000 échantillons. Les valeurs du BRM calculées en se fondant sur les estimateurs pondérés ainsi que les estimateurs imputés ont été comparées aux valeurs du BRM calculées en se fondant sur les estimateurs d'échantillonnage aléatoire simple (EAS) provenant de l'ensemble de la population et du panel volontaire. Selon les résultats théoriques existants, la moyenne pour l'EAS provenant de l'ensemble de la population et l'estimateur pondéré (3,1) avec score de propension estimé par la régression logistique pondérée devraient être approximativement sans biais, tandis que la moyenne pour l'EAS provenant du panel volontaire et l'estimateur pondéré (3,1) avec score de propension estimé par régression logistique non pondérée devraient être biaisés.

Le tableau 1 et la figure 2 donnent les résultats des simulations en utilisant l'estimateur pondéré (3,1), indiquant que la moyenne pour l'EAS sur l'ensemble de la population est sans biais. La moyenne pour l'EAS sur le panel volontaire et l'estimateur pondéré avec estimation non pondérée des paramètres sont biaisés; l'estimateur pondéré avec estimation pondérée des paramètres est approximativement sans biais comme nous nous y attendions. Le tableau 2 et la figure 3 montrent les résultats des simulations en utilisateur l'estimateur imputé (3,2), indiquant que les estimateurs imputés par les méthodes IPPV, HDA et HDA pondérée (où les poids de base de l'échantillon de volontaires sont intégrés dans l'imputation HDA) sont approximativement sans biais. Qui plus est, les deux estimateurs ont de très petites variances comparativement à l'estimateur de la moyenne pour l'EAS sur l'ensemble de la population (voir les diagrammes à moustaches aux figures 2 et 3), et par comparaison, on peut voir que l'estimateur imputé (3,2) présente une performance légèrement supérieure à l'estimateur pondéré (3.1) en ce qui concerne la réduction du biais. Donc, en nous fondant sur les études par simulation, nous pouvons estimer empiriquement le biais et la variance des deux estimateurs, car nous savons qu'il est difficile de calculer les variances analytiquement.

Tableau 1 : Biais relatifs en pourcentage de la proportion estimée de personnes votant pour le NPI sur 1 000 échantillons en utilisant l'estimateur pondéré pour différentes tailles d'échantillons de volontaires et de référence

		EAS	Web	SCORE DE PROPENSION P	SCORE DE PROPENSION P-PB
n_1	n_0				
1 000	1 000	-0,02	29,61	-8,69	1,78
2 000	1 000	-0,01	29,68	4,32	1,89
3 000	1 000	0,00	29,72	10,69	1,80
4 000	1 000	0,05	29,62	14,45	1,71
2 000	2 000	-0,04	29,62	-8,71	1,77

EAS - EAS de taille n_1 de U; Web - EAS de taille n_1 du panel volontaire;
SCORE DE PROPENSION P - $\hat{\bar{Y}}_{HT}$ avec régression logistique non pondérée; SCORE DE PROPENSION P-PB -$\hat{\bar{Y}}_{HT}$ avec régression logistique pondérée.

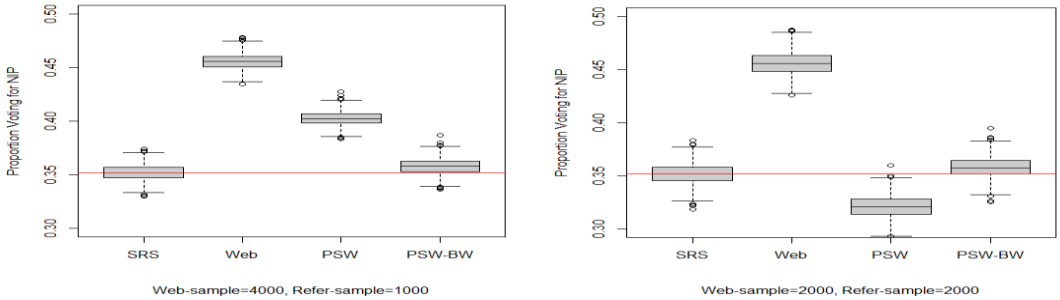

Figure 2 : Résultats de simulations fondés sur l'estimateur pondéré

Tableau 2 : Biais relatifs en pourcentage de la proportion estimée de personnes votant pour le NPI dans 1 000 échantillons en utilisant l'estimateur imputé pour différentes tailles d'échantillons de volontaires et de référence

		EAS	Web	IPPV	HDA	HDA-PB
n_1	n_0					
1 000	1 000	-0,35	30,40	0,30	5,04	4,67
2 000	1 000	-0,15	29,56	-0,30	0,79	0,77
3 000	1 000	-0,26	29,98	0,24	0,65	0,39
4 000	1 000	0,28	30,19	0,16	0,41	0,38
2 000	2 000	0,15	29,95	0,23	1,03	1,10

IPPV – $\hat{\bar{Y}}_{IMP}$ avec IPPV; HDA - $\hat{\bar{Y}}_{IMP}$ avec HDA; HDA-PB- $\hat{\bar{Y}}_{IMP}$ avec HDA pondéré.

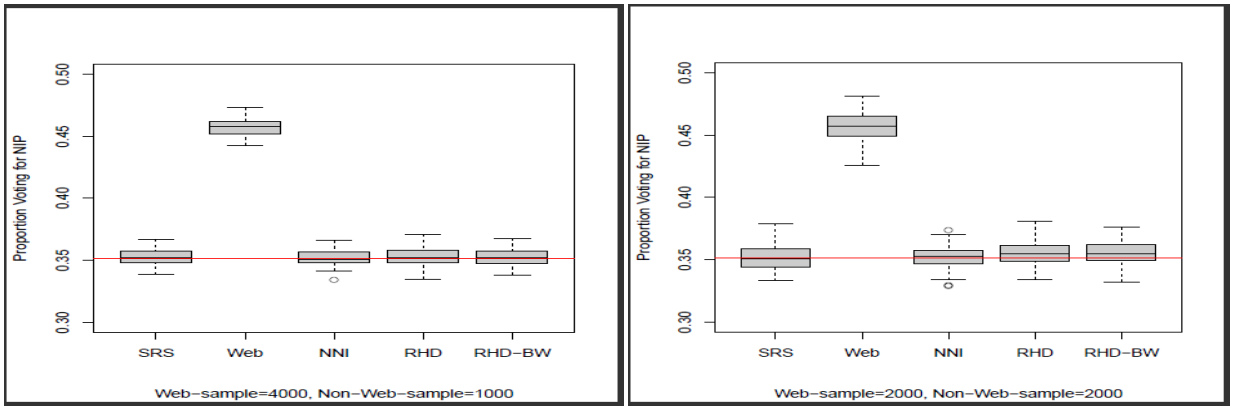

Figure 3 : Résultats de simulations fondés sur l'estimateur imputé

5. Discussion et conclusion

Il est bien connu que le biais d'autosélection dans les enquêtes en ligne à participation volontaire est plus inquiétant que dans tout autre mode d'enquête. Dans la présente étude, nous utilisons la pondération par les scores de propension et des méthodes d'imputation pour résoudre la question du biais de non-couverture dans les enquêtes en ligne à participation volontaire dans le cas du scénario idéal. En nous fondant sur notre ensemble de données de simulation, nous constatons que les estimateurs pondérés et imputés donnent de très bons résultats en ce qui concerne la réduction du biais de non-couverture, et que les estimateurs imputés pourraient performer un peu plus que les estimateurs pondérés. Dans la méthode de l'estimateur pondéré, nous notons que la détermination de la probabilité que l'unité k réponde à une enquête peut être difficile, de sorte que nous formulons trois hypothèses : i) la probabilité que l'unité k soit sous-échantillonnée est connue; ii) il n'y a pas de non-réponse; iii) toutes les unités ayant accès à Internet ont décidé de participer au panel volontaire. Sous ces hypothèses, nous nous concentrons sur l'estimation de la probabilité d'être un volontaire, laquelle pose aussi un défi, puisque les poids de base doivent être intégrés dans le modèle de régression afin d'obtenir des estimations valides. Cependant, comme la détermination des poids de base n'est pas facile à faire, nous avons choisi un cas relativement simple dans la présente étude. Des scénarios plus complexes seront examinés à l'avenir, selon la disponibilité d'échantillons de référence. Dans le scénario réaliste illustré à la figure 4, pour utiliser la méthode du score de propension, nous devons supposer que l'échantillon de volontaires et l'échantillon de référence sont disjoints, ce qui est une condition nécessaire, mais inexprimée dans la littérature sur les enquêtes avec échantillon de volontaires et échantillon de référence quand le modèle avec score de propension est utilisé (Valliant et Dever, 2011). Qui plus est, pour obtenir des estimations sans biais, nous devons déterminer les poids de base utilisés dans la régression logistique, mais l'exercice est difficile puisque l'échantillon de référence est constitué d'unités provenant de V ainsi que de V^c. En revanche, les imputations IPPV et HDA sont plus faciles à mettre en œuvre sous ce scénario réaliste, parce qu'elles ne requièrent pas l'hypothèse d'échantillons disjoints ni les poids de base. En outre, un plus grand nombre d'approches devraient être explorées sous le scénario réaliste, comme la stratification selon le score de propension, l'appariement selon le score de propension, et les estimateurs sous un modèle doublement robuste.

La présence de données manquantes est particulièrement fréquente dans les enquêtes en ligne à participation volontaire (en raison de la non-couverture et de la non-réponse). De nombreuses méthodes d'imputation ont été proposées, y compris des techniques d'imputation statistiques (telles que l'imputation hot-deck et l'imputation par la régression) et des techniques d'imputation fondées sur l'apprentissage machine (par exemple, CART, RF, et SVM, voir Hastie, 2009), pour traiter les données manquantes. Cependant, les méthodes d'imputation, surtout l'imputation sous apprentissage machine, ne sont habituellement pas employées pour les enquêtes en ligne. Les chercheurs évitent souvent d'utiliser des méthodes d'imputation fondées sur des arbres de classification, puisqu'elles peuvent être difficiles à interpréter. Dans de nombreux cas, cependant, nous ne cherchons pas à interpréter les arbres; nous nous soucions seulement de leur capacité à fournir des imputations raisonnables. Donc, l'application des méthodes d'imputation sous apprentissage machine, dont il a été prouvé que la performance est excellente, aux enquêtes en ligne devrait être étudiée plus en profondeur.

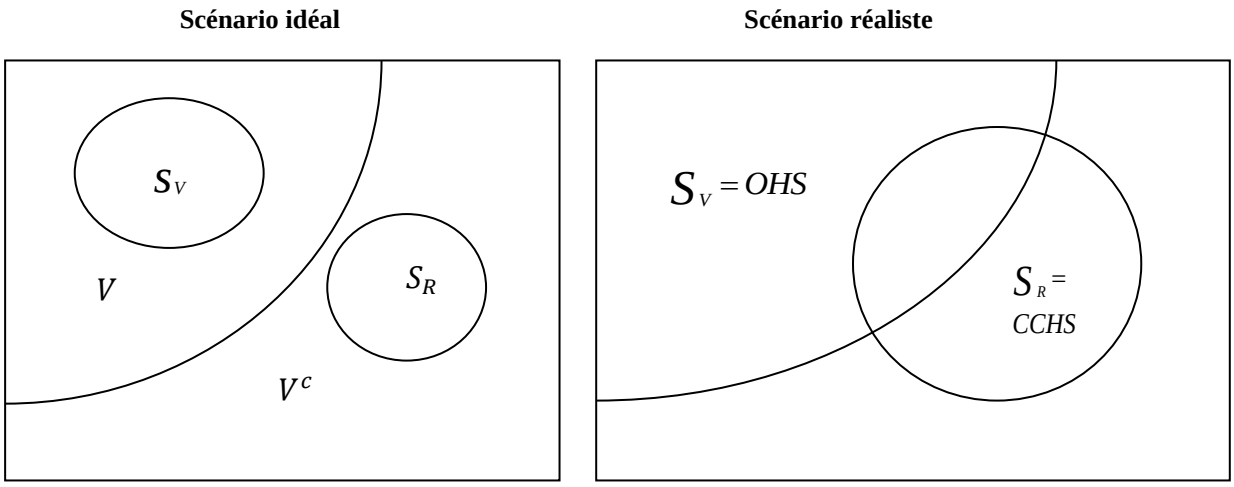

Figure 4 : Du scénario idéal (à gauche) au scénario réaliste (à droite)

Dans la présente étude, nous utilisons l'imputation hot-deck pour remplacer les valeurs manquantes dues à la non-couverture dans les enquêtes en ligne. La performance des estimateurs imputés fondés sur ces méthodes d'imputation hot-deck semble très prometteuse. Cependant, la plupart des méthodes d'imputation existantes, y compris les méthodes d'imputation statistiques et celles sous apprentissage machine, ne permettent pas d'intégrer dans l'imputation l'information sur le plan de sondage, comme les grappes, les strates et les poids, et elles doivent donc être modifiées afin de pouvoir mieux intégrer cette information dans l'avenir.

Bibliographie

ALAVAREZ, R. Michael et Carla VANBESELAERE. « Web-Based Surveys », *Encyclopedia of Social Measurement*, vol. 3, p. 955 à 962.

ANDRIDGE, Rebecca R. et Roderick J.A. LITTLE. 2010. « A Review of Hot Deck Imputation for Survey Non-response », *International Statistical Review*, vol. 78, nº 1, p. 40 à 64.

BETHLEHEM, Jelke. 2010, « Selection Bias in Web Surveys », *International Statistical Review*, vol. 78, nº 2, p. 161 à 188.

CHEN, Jiahua et Shao JUN. 2000. « Nearest Neighbor Imputation for Survey Data », *Journal of Official Statistics*, vol. 16, nº 2, p. 113 à 131.

COUPER, Mick P. 2000. « Web Surveys: A Review of Issues and Approaches », *Public Opinion Quarterly*, vol. 64, p. 464 à 494.

HASTIE, Trevor, Robert TIBSHIRANI et Jerome FRIEDMAN. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York, NY, Springer.

ROSEBAUM, Paul R. et Donald B. RUBIN. (1983), « The Central Role of the Propensity Score in Observational Studies for Causal Effects », *Biometrika*, vol. 70, nº 1, p. 41 à 55.

RUBIN, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

VALLIANT, Richard et Jill A. DEVER. 2011. « Estimating Propensity Adjustments for Volunteer Web Surveys », *Sociological Methods & Research*, vol. 40, nº 1, p. 105 à 137.