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

Le sondage indirect appliqué aux modèles de Capture-Recapture avec dépendance entre les sources.

Herménégilde Nkurunziza et Ayi Ajavon[1]

Résumé

Capture-Recapture est une méthode largement utilisée pour estimer la taille inconnue d'une population. La méthode consiste à tirer, de la population d'intérêt, deux échantillons indépendants. L'estimateur de Petersen de la taille de la population, souvent utilisé, est fonction de la taille et du chevauchement entre les deux échantillons. Lavallée et Rivest (2012) se sont intéressés au cas où les échantillons sont issus d'un *sondage indirect* et ont introduit une généralisation de l'estimateur de Petersen basée sur *la méthode généralisée de partage des poids*. En pratique, l'hypothèse d'indépendance sur laquelle repose l'estimateur n'est pas souvent vérifiée (Brenner(1995)). Dans le présent article, nous nous intéressons aux modèles de capture-recapture avec dépendance entre les sources et proposons une extension de l'estimateur de Lavallée et Rivest (2012). Nous analysons les propriétés de l'estimateur obtenu et présentons une illustration de la méthode à l'aide de données simulées.

Mots-clés : Échantillonnage indirect; méthode généralisée de partage de poids; base de sondage; estimateur; dépendance.

1. Contexte et objectifs

1.1 Le sondage indirect

Dans les conditions normales de sondage, on a une base de sondage pour la population d'intérêt. De cette base de sondage, on tire un échantillon pour faire des estimations. Cependant, il arrive des situations où on n'a pas de base de sondage de la population d'intérêt, mais on a celle d'une autre population liée d'une certaine manière à celle-ci. Le sondage indirect est utilisé pour enquêter les populations difficiles à sonder, à trouver ou à atteindre, parce que les individus de la population cible sont peu nombreux ou parce qu'il n'existe pas de base de sondage fiable.

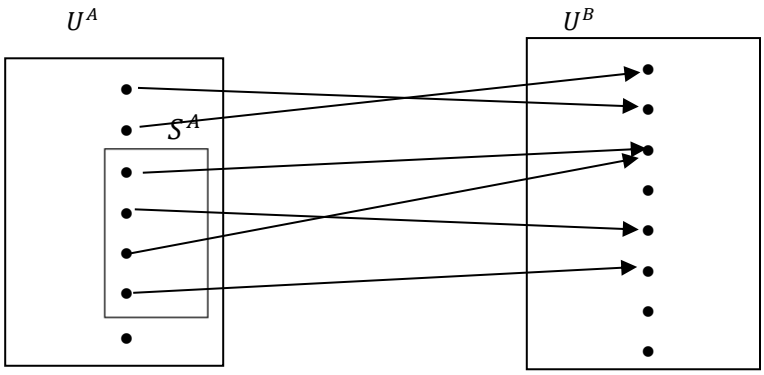
Il est aussi utilisé dans le cas de populations pour lesquelles il est difficile d'obtenir une mesure, une entrevue (par exemple si la loi ne le permet pas) (Lavallée (2016), Kiesl (2016)). On utilise alors une base de sondage différente de la population cible, mais reliée d'une certaine façon à la population difficile à échantillonner (Lavallée, 2016).

Par exemple, en ayant la liste des organisations qui offrent des services (ex. repas, logis) aux personnes « sans domicile fixe », on pourrait estimer le nombre total de ces derniers dans une ville donnée en utilisant le sondage indirect. Ou encore on pourrait faire une enquête sur des enfants si on a seulement la liste des parents.

Formellement, on a deux populations U^A et U^B (pour garder les mêmes notations que dans Lavallée (2002)) reliées entre elles. On désire produire une estimation pour U^B, alors qu'une base de sondage est disponible pour U^A seulement. On tire alors un échantillon de U^A qui est alors utilisé pour faire l'estimation pour U^B.

Figure 1.1-1
Représentation graphique du sondage indirect

[1] Herménégilde Nkurunziza, Statistique Canada, promenade Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6 (hermenegilde.nkurunziza@canada.ca); Ayi Ajavon, Statistique Canada, promenade Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6 (Ayi.Ajavon@canada.ca)



Le défi majeur du sondage indirect est de définir le lien entre U^A et U^B (Kiesl, 2016). Une fois que ce lien est établi, l'autre défi est de pouvoir associer une probabilité de sélection, ou un poids de sondage, aux unités enquêtées de la population cible U^B(Deville et Lavallée, 2006). Pour résoudre ce dernier problème, Lavallée (1995) a proposé la Méthode généralisée de partage de poids (MGPP) qui permet d'associer des poids aux unités *k* de U^B ayant un lien avec les unités échantillonnées *i* de U^A, comme suit :

-Un échantillon S^A est sélectionné de U^A, $\pi_i > 0$ est la probabilité de sélection de l'unité i. Donc le poids de sondage, s'il n'y a pas d'autres ajustements, est $\frac{1}{\pi_i}$. Le poids de sondage d'une unité k de U^B est donnée par (Lavallée, 2002):

$\widehat{w}_k = \sum_{i \in S^A} \frac{1}{\pi_i} \frac{l_{i,k}}{L_k^B}$ (1)

où, $l_{i,k} = 1$ si l'unité *i de* U^A est liée à l'unité k de U^B ou $l_{i,k} = 0$ sinon.

$L_k^B = \sum_{i \in U^A} l_{i,k}$ (2)

Ce poids pourrait être considéré comme une moyenne des poids de sondage des unités de la population U^A liées à k (Lavallée, 2002). Une fois que les poids des unités liées à S^A sont établis, l'estimation pour U^B se fait comme dans le cas classique. La non-réponse des unités de S^A est traité comme dans le cas classique d'enquête.

La difficulté majeure de la MGPP est liée, comme déjà mentionnée ci-haut, à celle d'établir si un élément i de U^A est liée à l'unité k de U^B. Ceci peut créer des biais dans les estimations. Dans la pratique, on peut obtenir les liens entre les éléments i et k, par exemple, lors de l'entrevue des unités sélectionnées dans l'échantillon S^A (Deville et Lavallée, 2006) ou par appariement.

1.2 Capture-recapture

La méthode de capture-recapture, quant à elle, a été longtemps utilisée pour estimer la taille d'une population inconnue. Un exemple classique est l'estimation du nombre de poisson dans un lac (Lavallée et Rivest, 2012). Elle consiste à tirer, de la population à estimer, un échantillon de taille, disons n_1, à étiqueter les unités échantillonnées et à les retourner dans la population. On tire ensuite un deuxième échantillon de la population, disons de taille n_2, et on estime la taille totale de la population en fonction de la taille de ces deux échantillons et du nombre n_{12} d'unités communes aux deux échantillons. L'estimateur de Peterson est alors utilisé.

$\widehat{N} = \frac{n_1 n_2}{n_{12}}$ (3)

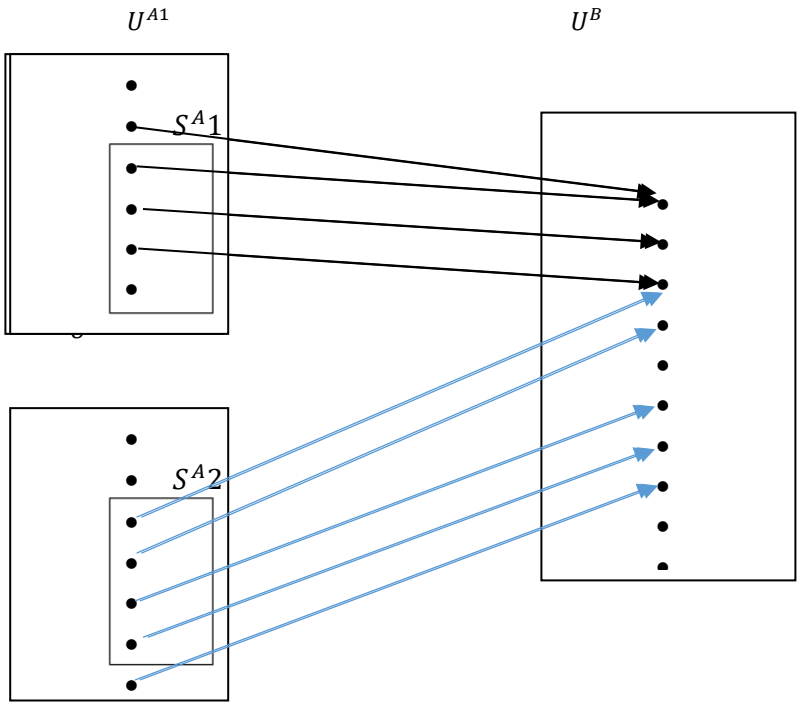
La méthode est actuellement utilisée davantage en biologie et en épidémiologie, pour estimer la taille des populations difficiles à atteindre ou à dénombrer, par exemple le nombre d'individus atteints d'une certaine maladie (Lavallée et Rivest (2012), Tilling (1999)), mais aussi dans d'autres domaines (Corrao et al, 2000, Kiesl, 2016). L'estimateur de Peterson est aussi utilisé pour estimer la taille d'une population partiellement couverte par deux fichiers. Dans ce dernier cas, l'estimateur de Peterson est donné par :

$\widehat{N}_{pet} = \frac{N_1 N_2}{N_{12}}$ (4)

où N_1 et N_2 sont les nombres d'enregistrements des deux fichiers respectivement, et N_{12} est le nombre d'enregistrements communs aux deux fichiers.

Une description plus détaillée de l'estimateur de Peterson est donnée dans Lavallée et Rivest (2012). Ils se sont intéressés à la méthode de capture-recapture dans un contexte où les échantillons sont issus d'un *sondage indirect*, c'est-à-dire le cas où les 2 fichiers (A1 et A2) ne représentent pas la population d'intérêt, mais d'autres populations liées d'une certaine manière à la population d'intérêt. Graphiquement, la situation se présente comme suit.

Figure 1.2-1
Capture-recapture et sondage indirect



Ils ont introduit une généralisation de l'estimateur de Petersen basée sur la MGPP, appelée estimateur par capture–recapture généralisé (CReG) , sous la forme suivante :

$\widehat{N}^{B}_{CReG} = \frac{\widehat{N}^{B}_{A1}\widehat{N}^{B}_{A2}}{\widehat{N}^{B}_{A1,A2}}$ (5)

où

$\widehat{N}^{B}_{A1}$ (respectivement $\widehat{N}^{B}_{A2}$) est l'estimation à partir des unités liées à celles de A1 (respectivement A2), avec des poids de sondage donnés par (1).

$\widehat{N}^{B}_{A1,A2}$ est l'estimation à partir des unités liées à des unités de A1 et A2, et dont les poids sont donnés par

$w_k^{A1,A2} = \left(\frac{1}{L_k^{A1}}\sum_{i\in S^{A1}}\frac{l_{ik}}{\pi_i^{A1}}\right)\left(\frac{1}{L_k^{A2}}\sum_{i\in S^{A2}}\frac{l_{ik}}{\pi_i^{A2}}\right)$ (6).

Ils ont ensuite montré comment la méthode pourrait être appliquée à l'estimation du total de n'importe quelle variable d'intérêt Y.

2. Le sondage indirect appliqué aux modèles de capture-recapture avec dépendance entre les sources

Comme déjà mentionné ci-haut, l'estimateur (5) repose sur l'hypothèse d'indépendance entre les sources. En pratique, cette hypothèse n'est pas toujours vérifiée (Brenner, 1995). La dépendance entre les sources conduit à un biais pour les estimateurs basés sur l'hypothèse d'indépendance (Chao (2001), Corrao et al (2000)). La dépendance peut provenir des situations suivantes: dans le cas de capture-recapture des animaux, la première capture pourrait par exemple créer un sentiment de peur/panique chez les animaux, créant ainsi une corrélation négative entre les captures. En épidémiologie, les listes utilisées peuvent être dépendantes (Chao, 2001).

Lorsqu'il y a dépendance positive entre les sources, la probabilité de trouver des cas sur un fichier augmente la probabilité pour ces cas de se trouver sur l'autre. Lorsqu'il y a dépendance négative entre les sources, la probabilité de trouver des cas sur un fichier diminue la probabilité pour ces cas de se trouver sur l'autre (Brenner, 1995).
Très peu d'études se sont intéressées au cas où les sources sont dépendantes. Dans le présent article, nous nous intéressons justement au sondage indirect appliqué aux modèles de capture-recapture avec dépendance entre les sources et proposons une extension de l'estimateur (5).

Dans le cas de dépendance négative entre les 2 sources, nous proposons une extension du CReG comme suit :

$$\widehat{N}^B_{\text{CReGd}} = \frac{\widehat{N}^B_{A1}\widehat{N}^B_{A2}}{\widehat{N}^B_{A1,A2}+\frac{|cov(X_{A1},X_{A2})|}{\min(p_1-p_{12},\ p_2-p_{12})}(\min(\widehat{N}^B_{A2},\widehat{N}^B_{A1})-\widehat{N}^B_{A1,A2})} \quad (7)$$

où :

$\widehat{N}^B_{A1}$ et $\widehat{N}^B_{A2}$ sont définies comme dans (5) ci-haut. $\widehat{N}^B_{A1,A2}$ est l'estimation à partir des unités liées à celles de A1 et à celles de A2, en utilisant des poids définis comme dans (1) par

$$w_k^{A1,A2} = \frac{1}{(L_k^{A1}+L_k^{A2})}\left(\sum_{i\in S^{A1}}\frac{l_{ik}}{\pi_i^{A1}}+\sum_{i\in S^{A2}}\frac{l_{ik}}{\pi_i^{A2}}\right) \quad (8)$$

$cov(X_{A1},\ X_{A2})$ est la covariance entre le fait d'être sur le fichier A1 et celui d'être sur le fichier A2, et on a $-1 \le cov(X_{A1},\ X_{A2}) \le 1$ (Brenner, 1995)

$X_{A1i} = 1$ si l'élément *i* est sur le fichier A1 et $X_{A1i} = 0$ sinon

$X_{A2i} = 1$ si l'élément *i* est sur le fichier A2 et $X_{A2i} = 0$ sinon

p_1 est la probabilité d'être dans le domaine A_1, p_2 la probabilité d'être dans le domaine A_2, et p_{12} la probabilité d'être dans le domaine $A_1 \cap A_2$. La quantité $p_1 - p_{12}$ est égale à l'espérance de la partie liée à A1 moins l'intersection et $p_2 - p_{12}$ est égale à l'espérance de la partie liée à A2 moins l'intersection

Le terme $\frac{|cov(X_{A1},X_{A2})|}{\min(p_1-p_{12},\ p_2-p_{12})}\left(\min(\widehat{N}^B_{A2},\widehat{N}^B_{A1})-\widehat{N}^B_{A1,A2}\right)$ peut être vu comme terme de correction du biais due à la dépendance entre les sources.

Note : Dans l'égalité (7),

1) si les deux sources sont indépendantes $cov(X_{A1},\ X_{A2}) = 0$ et nous retrouvons l'expression (5) de $\widehat{N}^B_{CReG}$.
2) si les deux sources sont très dépendantes, avec corrélation négative $cov(X_{A1},\ X_{A2}) \cong \min((1-p_1)(1-p_2)\ ,\ p_1p_2\ )$.
3) Si $p_1 + p_2 > 1$, alors $(1-p_1)(1-p_2) <\ p_1p_2$ (Brenner, 1995) . On aura
4) $$\widehat{N}^B_{\text{CReGd}} \cong \frac{\max(\widehat{N}^B_{A1},\widehat{N}^B_{A2})\min(p_1-p_{12},\ p_2-p_{12})}{(1-p_1)(1-p_2)} \cong \frac{\max(\widehat{N}^B_{A1},\widehat{N}^B_{A2})}{\max((1-p_1),(1-p_2))}\ .$$

Dans le cas de dépendance positive, on obtient un estimateur similaire à (7) par symétrie.

2.1 Propriété de l'estimateur proposé

Lemme : convergence faible.

Considérons l'estimateur $\widehat{N}^B_{\mathrm{CReGd}} = \frac{\widehat{N}^B_{A1}\widehat{N}^B_{A2}}{\widehat{N}^B_{A1,A2}+\frac{|cov(X_{A1},X_{A2})|}{\min(p_1-p_{12},\ p_2-p_{12})}(\min(\widehat{N}^B_{A2},\widehat{N}^B_{A1})-\widehat{N}^B_{A1,A2})}$.

La dépendance négative entre les sources implique qu'il existe $\theta \geq 0$ telle que $p_1p_2 = (1-\theta)p_{12} + \theta \min(p_1, p_2)$ (Kimeldorf and Sampson, 1989).

Alors l'estimateur $\widehat{N}^B_{\mathrm{CReGd}}$ converge en probabilité vers la taille de la population.

Preuve : La preuve est identique au cas indépendant (Lavallée et Rivest, 2012), on montre simplement que $\widehat{N}^B_{\mathrm{CReGd}}/N^B$ converge faiblement vers 1 du fait que le numérateur $\widehat{N}^B_{A1}\widehat{N}^B_{A2}$ et le dénominateur convergent vers la même quantité.

3. Simulation

On cherche à estimer le nombre d'utilisateurs de téléphones cellulaires dans une ville, à partir des fichiers A1 et A2 fournis par les deux seuls fournisseurs. A1 contient 1000 numéros et A2 contient 800 numéros. On sélectionne par EAS un échantillon de 500 numéros de chaque fichier (les probabilités d'inclusion sont p_1=1/2 et p_2=5/8). On appelle chaque numéro sélectionné et on prend les informations du propriétaire, ceci crée le lien entre les deux listes et le fichier des personnes.

Supposons que chaque compagnie ne donne pas plus d'un numéro à la même personne, mais une personne peut avoir plus d'un numéro provenant de fournisseurs différents.

Après les appels on complète le tableau comme dans l'exemple suivant (données fictives).

Numéro	X_{A1} (Numéro A1)	X_{A2} (Numéro A2)	Propriétaire
613 000 6644	1	0	Jean
819 333 9999	0	1	Alice
613 777 0000	1	0	Peter
613 777 8888	1	0	Alice
613 999 0000	0	1	Jean
613 000 2222	1	0	
819 000 5555	0	1	Smith

Dans cet exemple Jean et Alice ont chacun deux numéros, un pour chacun des deux fournisseurs.

Dans la situation décrite dans cette simulation, on voit qu'il y a une dépendance négative forte, donc nous pouvons prendre $cov(X_{A1},\ X_{A2}) \cong (1-p_1)(1-p_2)$.

Supposons de plus que :

- 400 des 500 numéros de l'échantillon S^{A1} sont associés à des propriétaires.
- 450 des 500 numéros de l'échantillon S^{A2} sont associés à des propriétaires.
- Dans les deux échantillons, 30 personnes ont un numéro pour chacun des deux fournisseurs.

En utilisant les expressions (1) et (6), on a les estimations suivantes

$\widehat{N}_{A1}^{B}$	$\widehat{N}_{A2}^{B}$	$\widehat{N}_{A1,A2}^{B}$
800	810	54

Comme la dépendance est négative et forte, et que $p_1 + p_2 > 1$, alors $\left|\mathrm{cov}(X_{A1},\ X_{A2})\right| \cong (1-p_1)(1-p_2)$ (Brenner,1995).

On aura $\widehat{N}^{B}_{CReGd} \cong \frac{\max(\widehat{N}^{B}_{A1},\widehat{N}^{B}_{A2})}{\max((1-p_1),(1-p_2))} = \frac{810}{1/2}$=1620.

Si on ignore la dépendance $\widehat{N}^{B}_{CReG} = \frac{800\times810}{96} = 6750$.

Ceci surestimerait le total de la population possédant un téléphone cellulaire.

Conclusion

Nous avons proposé un estimateur du total d'une population, par sondage indirect appliqué à la méthode de capture-recapture, en présence de dépendance entre les sources. L'estimateur présenté ici est seulement valide pour la dépendance négative. Pour la dépendance positive, un estimateur similaire est obtenu par symétrie. On suppose que les liens entre les unités sont correctement établis.

Bibliographie

Brenner, H. (1995), « Use and Limitations of the Capture-Recapture Method in Disease Monitoring with Two Dependent Sources », *Epidemiology*, 6(1), p. 42-48.

Chao, A. (2001), « An Overview of Closed Capture–Recapture Models », *Journal of Agricultural, Biological, and Environmental Statistics*, 6(2), p. 158–175.

Deville, J.-C., et P. Lavallée (2006), « Indirect Sampling: the Foundations of the Generalised Weight Share Method », *Survey Methodology*, 32, p. 165–176.

Giovanni, C. G. et al. (2000), « Capture-recapture methods to size alcohol related problems in a population », *J Epidemiol Community Health*, 54, p. 603–610.

Kiesl, H. (2016), « Indirect Sampling: A Review of Theory and Recent Applications », *German Statistical Society*, 10(4), p. 289-303.

Kimeldorf, G., et A. R. Sampson (1989), « A framework for positive dependence », *Ann. Inst. Statist. Math*, 41(1), p. 31-45.

Lavallée, P. (1995), « Cross-sectional Weighting of Longitudinal Surveys of Individuals and Households using the Weight Share Method », *Survey Methodology*, 21, p. 25–32.

Lavallée, P. (2002), *Le sondage indirect ou la méthode généralisée du partage de poids*, Bruxelles: Éditions de l'Université de Bruxelles.

Lavallée, P. (2016), « Le sondage indirect pour les populations difficiles à joindre », Cours offert à Statistique Canada, Canada.

Lavallée, P., et L. P. Rivest (2012), « Capture-recapture sampling and indirect sampling », *Journal of Official Statistics*, 28(1), p. 1-27.

Tilling, K., et J. A. C. Sterne (1999), « Capture-Recapture Models Including Covariate Effects », *American Journal of Epidemiology*, 149(4), p. 392-400.