LE SONDAGE DETERMINE SELON LES REPONDANTS: UN SURVOL METHODOLOGIQUE

Pierre Lavallée

Statistique Canada, Ottawa (Ontario), K1A 0T6, pierre.lavallee@statcan.gc.ca

1. Introduction

Lorsque l'on doit effectuer un sondage, il arrive que la population d'intérêt soit une *population cachée ou difficile à joindre*. Cette dernière peut être difficile à joindre à cause de problèmes de rareté et d'absence de base de sondage qui sont des problèmes associés à l'échantillonnage. La population d'intérêt peut aussi être difficile à joindre à cause de problèmes de mesure, d'entrevue : les membres de la population sont difficiles à persuader, à interviewer ou à mesurer.

Heureusement pour le sondeur, les populations difficiles à joindre se retrouvent souvent sous forme de *réseaux*, c'est-à-dire que les membres sont regroupés en cercles de connaissances, en associations, ou par proximité géographique. C'est le cas, par exemple, des individus atteints du SIDA, des utilisateurs de drogues, des individus ayant des relations sexuelles avec des individus de même sexe, des individus appartenant à un groupe ethnique particulier, et même des musiciens de jazz dont l'exemple sera discuté plus loin.

On retrouve plusieurs méthodes d'enquête pour les populations cachées ou difficiles à joindre. Parmi celles-ci, on a le Sondage ciblé (*Targeted Sampling*), le Sondage par informateur-clé (*Key Informant Sampling*), ainsi que les méthodes par chaînage (*Chain-Referral methods*). Parmi ces dernières, on retrouve le Sondage par réseaux (*Network Sampling*), le Sondage en grappes adaptatif (*Adaptive Cluster Sampling*), le Sondage « boule-de-neige » (*Snowball Sampling*), et le **Sondage déterminé selon les répondants** (*Respondent-Driven Sampling*).

Comme le titre l'indique, le Sondage déterminé selon les répondants (SDR) fera l'objet d'un survol méthodologique dans le présent document. On parlera en premier lieu des méthodes de sondage par chaînage en général, pour se concentrer par la suite sur le SDR. On discutera ensuite des hypothèses sous-jacentes au SDR et des différentes méthodes d'estimation qui en découlent. Finalement, nous aborderons certains problèmes potentiels reliés au SDR et les effets de ces problèmes sur les estimations.

2. Les méthodes de sondage par chaînage

Les méthodes de sondage par chaînage suivent le processus d'enquête suivant. Au départ, on sélectionne un échantillon s **pas nécessairement probabiliste** de la population cible U (population cachée ou difficile à joindre). Cet échantillon sert de contacts initiaux. Chaque individu de s fournit alors les noms d'individus de U. Les enquêteurs contactent ensuite ces individus et leur demandent de participer à l'enquête. L'étape suivante consiste à demander à chaque répondant de fournir d'autres noms d'individus de U. Le processus d'enquête continue ainsi pour un certain nombre de vagues.

Le SDR est une version évoluée du Sondage « boule-de-neige ». Cette dernière méthode est certainement la méthode de chaînage la plus connue. À notre connaissance, Coleman (1958) a été le premier à mentionner cette méthode, mais sa popularité est certainement due à Goodman (1961). Par la suite, plusieurs auteurs ont utilisé cette méthode, notamment Frank (1977, 1978), ainsi que Frank et Snijders (1994).

Un Sondage « boule-de-neige » à τ vagues et κ noms se déroule selon les étapes suivantes :

1. On tire un échantillon aléatoire *s* de *n* individus d'une population cible *U*, ce qui constitue la vague 0.
2. On demande à chacun des *n* individus de *s* de nommer κ noms d'individus appartenant à la même population cible. Les individus nommés par les individus de *s* (et qui ne font pas partie de *s*) forment la 1re vague du sondage. On obtient alors des grappes de taille $\kappa+1$ qui peuvent être chevauchantes.
3. On demande à chaque individu de la 1re vague de nommer à son tour κ individus cibles. Les nouveaux individus nommés par les individus de la 1re vague (et qui ne font pas partie ni de la 1re vague, ni de *s*) forment la 2e vague du sondage.
4. On continue jusqu'à ce que τ vagues soient complétées.

Dans un article clé, Erickson (1979) a mentionné plusieurs problèmes reliés aux méthodes de sondage par chaînage. Premièrement, l'inférence doit être basée sur l'échantillon initial puisque les individus supplémentaires obtenus par chaînage ne sont pas choisis aléatoirement, ni sans biais connus. Deuxièmement, ces méthodes tendent à être biaisées envers les individus les plus coopératifs, ce qui est envenimé si l'échantillon initial est constitué de volontaires (*volontarisme*). Troisièmement, il peut y avoir un biais de *masquage* parce qu'on peut tenter de « protéger » des amis en ne les mentionnant pas. Finalement, le chaînage se faisant par liens de réseautage, les individus avec de grands réseaux de connaissances ont plus de chances d'être sélectionnés.

Selon Heckathorn et Jeffri (2001), « si les sources de biais sont bien comprises, on peut éliminer les biais étrangers à la méthode, et quantifier et contrôler ceux inhérents à la méthode. » C'est dans cette optique qu'a été développé le SDR en réponse aux problèmes soulevés par Erickson (1979) relatifs aux méthodes par chaînage.

3. Le Sondage déterminé selon les répondants

Le SDR a été proposé en premier lieu par Heckathorn (1997) pour étudier des populations cachées ou difficiles à joindre. Par la suite, plusieurs articles ont suivi, notamment Heckathorn et Jeffri (2001), Heckathorn (2002), Salganik et Heckathorn (2004), Volz et Heckathorn (2008), Gile et Handcock (2010), Seman (2010) et bien d'autres.

Comme mentionné plus tôt, le SDR est une version évoluée du Sondage « boule-de-neige ». Comme pour toutes les méthodes par chaînage, on suppose que les meilleurs individus pour pénétrer au sein d'une population cachée sont leurs propres pairs. Le SDR consiste donc à enquêter au sein de réseaux d'individus cibles en sélectionnant au départ un ou plusieurs des éléments des réseaux appelés « germes ». Notons que dans la population cible *U*, on suppose généralement un seul réseau, ce qui implique que deux individus cibles peuvent toujours être reliés entre eux par chaînage au sein du réseau.

Voici les étapes du SDR à κ noms pour la sélection de *n* individus :

1. Tirer un échantillon *s* **pas nécessairement aléatoire** d'individus cibles, ce qui constitue la vague 0. Ces individus sont appelés « germes ».

2. Donner κ coupons à chaque germe et leur demander de donner les coupons à d'autres individus cibles parmi leurs pairs. Parce que chaque coupon est identifié, on peut suivre les patrons de recrutement au sein de la population cible. Les individus recrutés — qui sont aussi des individus cibles — par les individus de la vague 0 forment la vague 1.
3. Donner κ coupons à chaque individu de la vague 1 et leur demander de donner les coupons à d'autres individus cibles parmi leurs pairs. Les individus recrutés par les individus de la vague 1 forment la vague 2.
4. On répète le processus jusqu'à l'obtention d'un échantillon de *n* individus.

À la fin du processus, l'échantillon contient les germes (vague 0) et les individus cibles recrutés. La figure suivante donne le patron de recrutement de personnes pour l'étude ECHO (*Eastern Connecticut Health Outreach*) destinée à la prévention du SIDA (Heckathorn, 1997). À la Figure 1, on peut voir que le germe est un homme de race noire et que la composition de l'échantillon se diversifie rapidement.

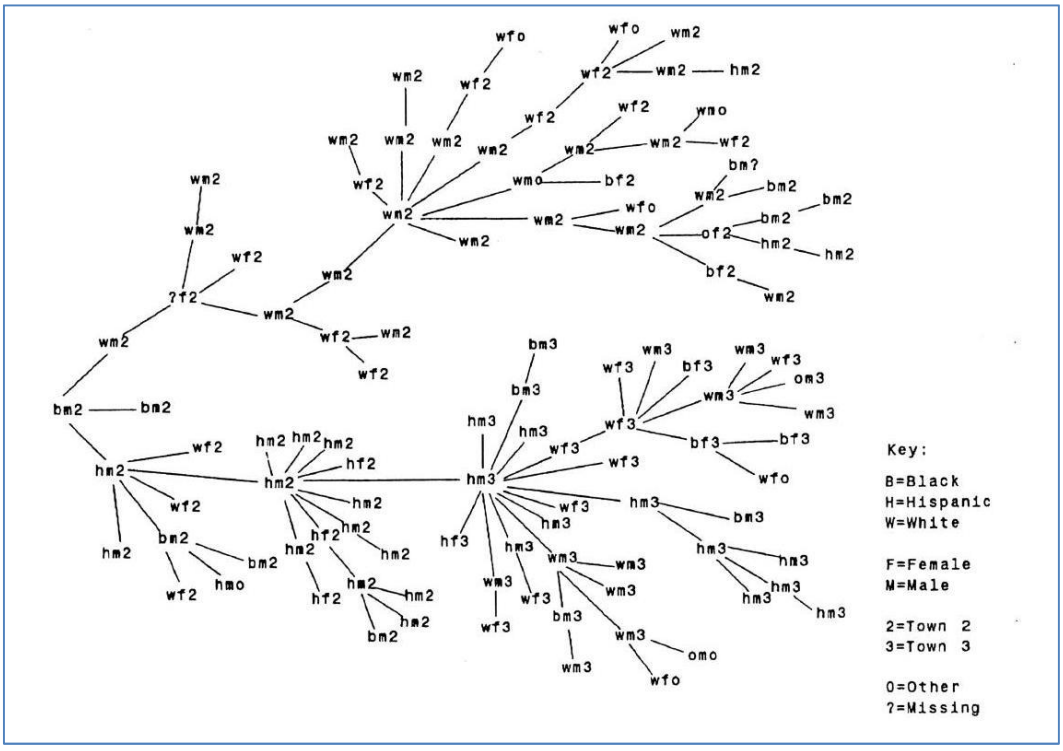

Figure 1. Patron de recrutement du SDR en partant d'un seul germe.

Un autre exemple classique de l'utilisation du SDR est celui décrit par Heckathorn et Jeffri (2001) qui présente une étude sur des musiciens de jazz dans quatre villes américaines. Cette population est considérée comme cachée à cause du côté marginal que les musiciens de jazz aiment se donner. Pour cette enquête, on a tiré un échantillon de musiciens de jazz en visitant certains bars connus. Cet échantillon n'est donc pas aléatoire. Ces premiers musiciens de jazz sont les germes. On a donné $\kappa = 4$ coupons à chaque musicien de jazz et on leur a demandé de donner les coupons à des musiciens de jazz parmi leurs pairs. Les individus recrutés forment la vague 1. On a ensuite donné $\kappa = 4$ coupons à chaque individu de la vague 1 et on leur a demandé de donner les coupons à d'autres musiciens de jazz parmi leurs pairs. Les individus recrutés ici forment la vague 2. On a répété le processus jusqu'à l'obtention d'un nombre donné de musiciens de jazz dans chaque ville.

Pour le SDR, le paramètre κ doit être suffisamment grand pour que le recrutement continue, même si certains recruteurs arrêtent. Par contre, il doit être suffisamment petit pour qu'il y ait un maximum de vagues avant l'obtention de la taille d'échantillon *n*. En pratique, on choisit généralement $\kappa = 3$ ou 4. Notons que plus il y a de vagues, plus on peut espérer couvrir les réseaux en entier. De plus, le processus d'enquête convergera alors vers un équilibre, quel que soit le choix des germes de la vague 0.

Typiquement, on recueille l'information suivante durant le processus d'enquête du SDR : (i) les variables d'intérêt *y* pour l'enquête; (ii) le *degré* ϑ_k de chaque individu *k* enquêté; (iii) les numéros d'identification des coupons afin d'établir les patrons de recrutement. Le degré ϑ_k de l'individu *k* représente le nombre de connaissances que possède cet individu au sein de la population cible. Le degré correspond donc au nombre d'individus qui pourraient potentiellement être recrutés par cet individu. On doit aussi s'assurer de la non-duplication des individus enquêtés, bien que le sondage soit fait avec remise et qu'ainsi, un individu peut être enquêté deux fois ou plus.

Le SDR a été développé autour de l'utilisation d'un *système dual d'incitatifs* (récompenses monétaires). On retrouve ainsi des incitatifs primaires qui sont des récompenses pour participer à l'enquête, ce qui est relativement courant dans les enquêtes. On retrouve aussi des incitatifs secondaires qui sont des récompenses pour **recruter** des individus cibles visés par l'enquête. Les incitatifs secondaires peuvent s'avérer plus efficaces que les incitatifs primaires (Heckathorn, 1990). Étant donné que l'on demande aux individus enquêtés à une vague donnée de recruter eux-mêmes d'autres individus parmi leurs pairs, en plus de la récompense s'ajoute une certaine pression par les recruteurs. À la différence du SDR, le Sondage « boule-de-neige » n'utilise que des incitatifs primaires (participation à l'enquête). Une autre différence entre le SDR et le Sondage « boule-de-neige » est qu'avec le SDR, les enquêteurs ne demandent pas aux individus de leur mentionner leurs pairs, mais plutôt de les recruter directement dans l'enquête, ce qui s'avère très efficace.

4. Hypothèses et estimation

Avec le SDR, on s'intéresse en général à l'estimation de proportions au sein de la population cible *U*. On désire donc estimer P_i, la proportion d'individus faisant partie du groupe *i* au sein de *U,* ou encore la proportion d'individus ayant la caractéristique *i* au sein de *U*. Par exemple, P_i peut être la proportion de femmes parmi les musiciens de jazz. Un autre exemple est celui où on s'intéresse à l'ethnicité des consommateurs de drogues par injection.

On dénombre jusqu'à présent trois grandes approches pour estimer les proportions P_i : (1) l'estimation à l'aide de chaînes de Markov; (2) l'estimation à l'aide du modèle de réciprocité; (3) l'estimation à partir d'une approche probabiliste. Nous allons maintenant décrire chacune des trois approches.

4.1 Estimation à l'aide de chaînes de Markov

Selon Heckathorn (1997), le processus de recrutement du SDR peut se voir comme une chaîne de Markov régulière (ergodique et non cyclique) d'ordre un. En effet, (i) il y a un nombre limité d'états (groupes ethniques, sexes); (ii) le recrutement est sans lien avec les

recruteurs passés (processus sans mémoire); (iii) deux individus cibles peuvent toujours être reliés entre eux par chaînage au sein du réseau (*ergodicité*); et (iv) tout individu peut être rejoint à n'importe quelle vague (*non-cyclicité*). Voir la Figure 2 tirée de Heckathorn (1997).

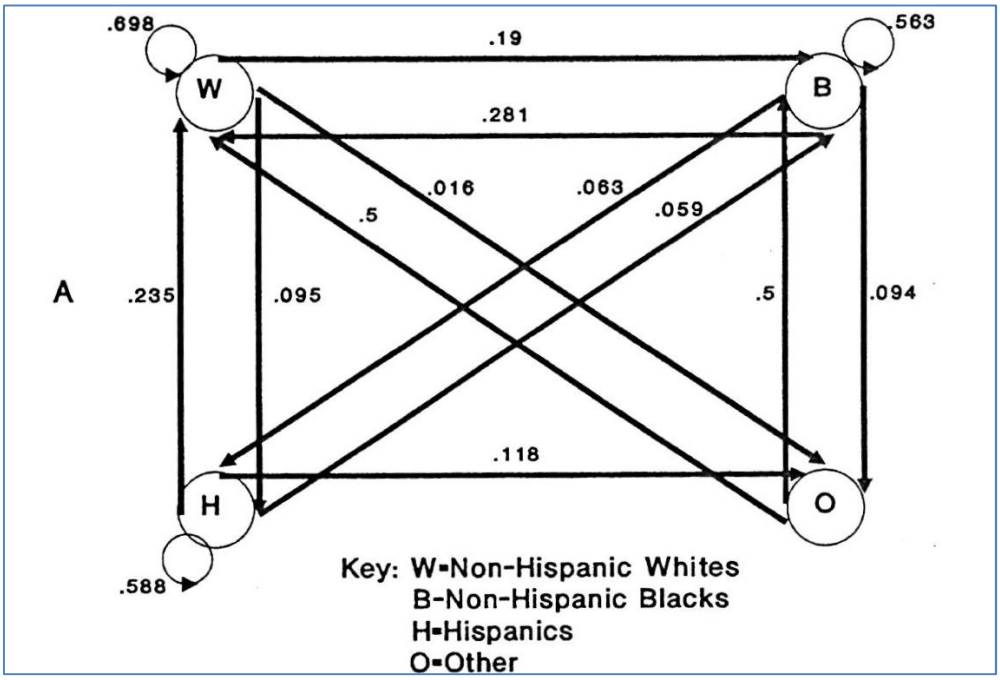

Figure 2. Représentation du réseau selon une chaîne de Markov où les noeuds représentent les caractéristiques des individus et les flèches, les probabilités de recrutement empiriques.

À partir de cette constatation, Heckathorn (1997) pose le Théorème 1 suivant:

> *Avec la progression du recrutement d'une vague à l'autre, on converge vers un mélange équilibré d'individus recrutés qui est indépendant de l'ensemble des individus de départ (germes).*

Tel que l'on peut le voir dans les Figures 3a et 3b suivantes tirées de Heckathorn (1997), on constate, en effet, empiriquement qu'en partant de deux germes différents (respectivement, un individu noir et un individu blanc), on obtient rapidement les mêmes proportions de personnes au sein des individus recrutés.

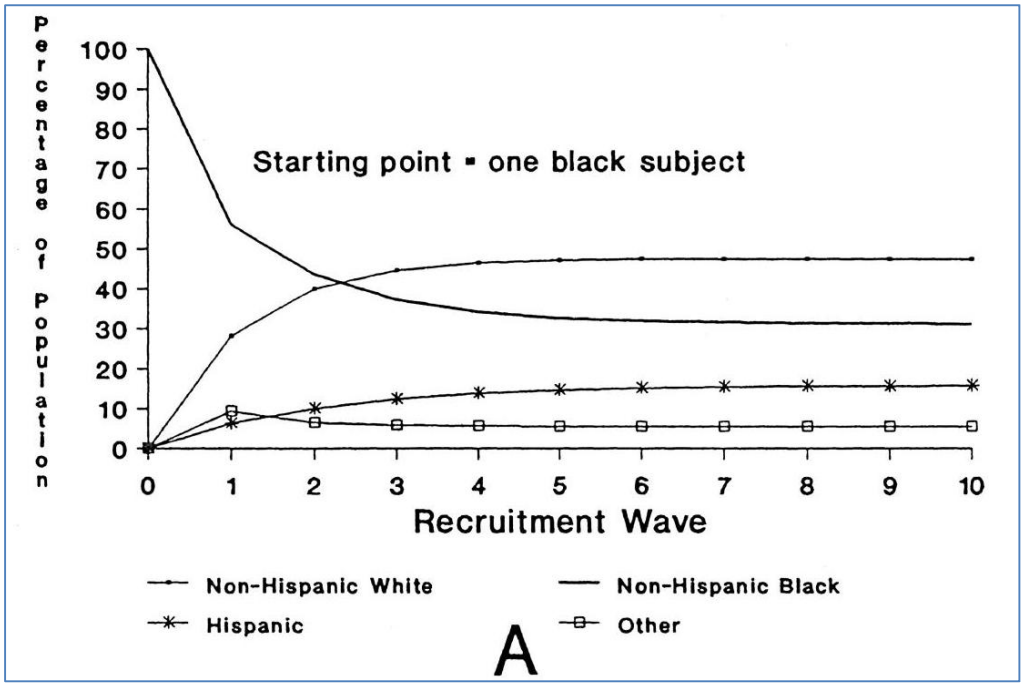

Figure 3a. Proportions d'individus recrutés à partir d'un seul germe : homme noir.

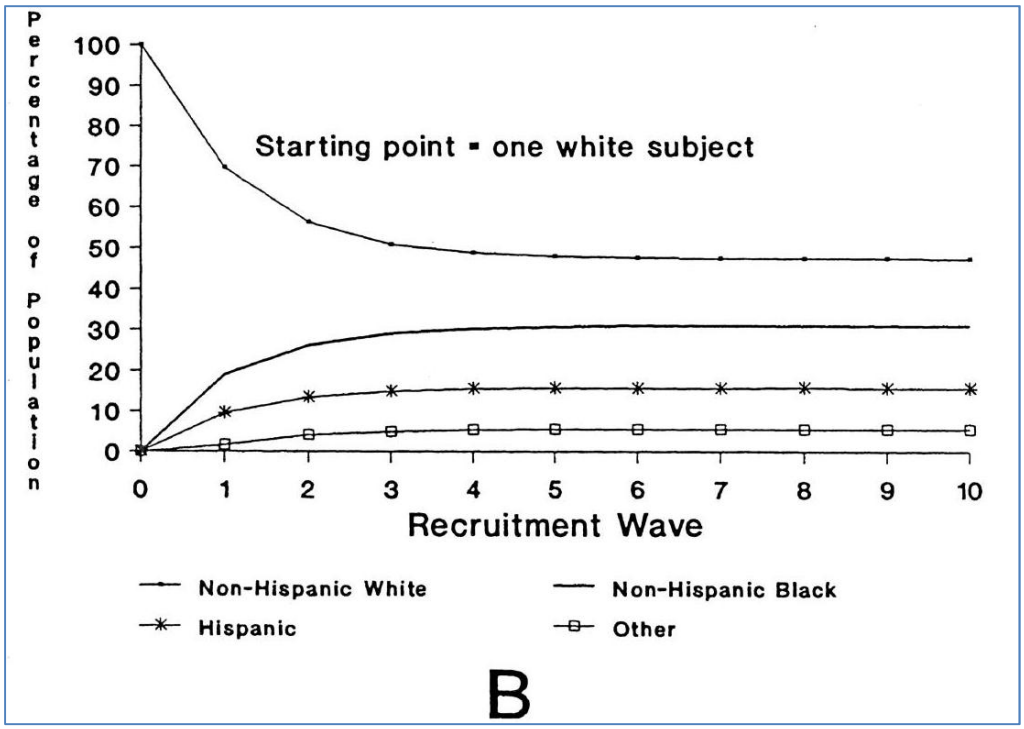

Figure 3a. Proportions d'individus recrutés à partir d'un seul germe : homme blanc.

Sous l'hypothèse d'une chaîne de Markov régulière (ergodique et non cyclique) d'ordre un, on a alors le résultat asymptotique suivant (Kemeny et Snell, 1960). Supposons qu'il n'y a que deux groupes (états) possible *a* et *b*. Soient P_a^E et P_b^E, les proportions à l'équilibre pour les groupes *a* et *b*. On a alors :

$$1 = P_a^E + P_b^E$$

$$P_a^E = P_a^E R_{aa} + P_b^E R_{ba} \qquad (1)$$

$$P_b^E = P_a^E R_{ab} + P_b^E R_{bb}$$

où R_{ij} sont les proportions de recrutement en partant du groupe *i* au groupe *j* ($i = a, b$ et $j = a, b$). Par exemple, R_{ab} est la proportion d'individus appartenant au groupe *a* qui recrute des individus appartenant au groupe *b*. Si on se réfère à la Figure 2, on peut voir que la proportion d'individus hispaniques recrutant des individus non hispaniques blancs est de 0,235. À partir de (1), on obtient

$$P_a^E = \frac{R_{ba}}{1 - R_{aa} + R_{ba}} \tag{2}$$

et

$$P_b^E = \frac{1 - R_{aa}}{1 - R_{aa} + R_{ba}} \tag{3}$$

En estimant les proportions R_{ij} à partir des données mesurées au cours du processus d'enquête du SDR, on obtient $\hat{P}_a^E$ et $\hat{P}_b^E$, des estimateurs des proportions P_a et P_b où

$$\hat{P}_a^E = \frac{\hat{R}_{ba}}{1 - \hat{R}_{aa} + \hat{R}_{ba}} \tag{4}$$

et

$$\hat{P}_b^E = \frac{1 - \hat{R}_{aa}}{1 - \hat{R}_{aa} + \hat{R}_{ba}} \tag{5}$$

4.2 Estimation à l'aide du modèle de réciprocité

Avec le SDR, on constate une abondance de *liens réciproques*, c'est-à-dire que s'il y a un lien de l'individu k à l'individu k', il y a aussi un lien de l'individu k' à l'individu k. Soit T_{ij}, le nombre de liens réciproques entre les groupes *i* et *j*. Puisque les liens sont réciproques, par définition, $T_{ij} = T_{ji}$. Selon Heckathorn (1999), on a

$$T_{ij} = P_i \bar{\vartheta}_i R_{ij} \tag{6}$$

où $\bar{\vartheta}_i$ est le *degré moyen* des individus au sein du groupe *i*. Rappelons que le degré ϑ_k d'un individu *k* est le nombre de connaissances qu'il possède au sein de la population cible. La quantité $\bar{\vartheta}_i$ est donc le nombre moyen de connaissances des individus appartenant au groupe *i*. Puisque $T_{ij} = T_{ji}$, on a donc pour le cas des deux groupes *a* et *b* :

$$P_a \bar{\vartheta}_a R_{ab} = P_b \bar{\vartheta}_b R_{ba} \tag{7}$$

Puisque $P_b = 1 - P_a$, on obtient

$$P_a = \frac{\bar{\vartheta}_b R_{ba}}{\bar{\vartheta}_b R_{ba} + \bar{\vartheta}_a R_{ab}} \tag{8}$$

Comme pour l'estimation à partir des chaînes de Markov, on estime les proportions R_{ij} à partir des données mesurées au cours du processus d'enquête du SDR. Il reste alors à estimer le degré moyen $\overline{\vartheta}_i$. En supposant que chaque individu *k* est recruté au hasard avec probabilité proportionnelle au degré ϑ_k, Salganik et Heckathorn (2004) ont proposé l'estimateur de type Hansen-Hurwitz (1943) suivant :

$$\hat{\overline{\vartheta}}_i = \frac{n_i}{\sum_{k=1}^{n_i} \frac{1}{\vartheta_k}} \tag{9}$$

où n_i est le nombre d'individus enquêtés dans le groupe *i*. Donc, pour les deux groupes *a* et *b*, on a

$$\hat{P}_a^R = \frac{\hat{\overline{\vartheta}}_b \hat{R}_{ba}}{\hat{\overline{\vartheta}}_b \hat{R}_{ba} + \hat{\overline{\vartheta}}_a \hat{R}_{ab}} \tag{10}$$

On note qu'en général, $\hat{P}_a^R \neq \hat{P}_a^E$.

4.3 Estimation à partir d'une approche probabiliste

Pour estimer les proportions P_i selon une approche probabiliste, Volz et Heckathorn (2008) ont posé les hypothèses de base suivantes : (i) le sondage est effectué avec remise; (ii) nous avons une petite fraction de sondage; (iii) chaque individu ne recrute qu'un seul autre individu (c'est-à-dire, $\kappa = 1$); (iv) on mesure le degré ϑ_k de chaque individu enquêté *k*; (v) la mesure du degré ϑ_k de chaque individu est exacte; (vi) le recrutement se fait de façon aléatoire; (vii) les liens sont réciproques; et (viii) le recrutement peut se modéliser selon un processus de Markov. On peut aisément voir que plusieurs de ces hypothèses peuvent être violées. Par exemple, comme le degré ϑ_k de l'individu *k* représente le nombre de connaissances que possède cet individu au sein de la population cible, on peut très bien voir que ce dernier peut être très difficile à mesurer avec exactitude. L'aspect aléatoire du recrutement est aussi très discutable.

Volz et Heckathorn (2008) ont proposé l'estimateur de type Hansen-Hurwitz suivant pour l'estimation d'une moyenne $\overline{Y}$ au sein de la population cible:

$$\hat{\overline{Y}}^{VH} = \frac{\sum_{k=1}^{n} y_k / \vartheta_k}{\sum_{k=1}^{n} 1/\vartheta_k} \tag{11}$$

Pour le cas des deux groupes *a* et *b*, pour l'estimation de proportions, on pose $y_k = 1$ si l'unité *k* appartient au groupe *a*, et 0 sinon. L'estimateur (11) devient ainsi

$$\hat{P}_a^{VH} = \frac{\sum_{k=1}^{n_a} 1/\vartheta_k}{\sum_{k=1}^{n} 1/\vartheta_k} \tag{12}$$

où n_a est le nombre d'individus enquêtés appartenant au groupe *a*. Notons que l'on peut réécrire (12) selon

$$\hat{P}_a^{VH} = \frac{\sum_{k=1}^{n_a} 1/\vartheta_k}{\sum_{k=1}^{n} 1/\vartheta_k} = \left(\frac{n_a}{n}\right)\left(\frac{\hat{\bar{\vartheta}}_U}{\hat{\bar{\vartheta}}_a}\right) \qquad (13)$$

où $\hat{\bar{\vartheta}}_a = n_a / \sum_{k=1}^{n_a} \frac{1}{\vartheta_k}$ et $\hat{\bar{\vartheta}}_U = n / \sum_{k=1}^{n} \frac{1}{\vartheta_k}$. Le facteur $\hat{\bar{\vartheta}}_U / \hat{\bar{\vartheta}}_a$ est l'*effet de réseautage*. Il est intéressant de noter que si $n_a \hat{R}_{ab} = n_b \hat{R}_{ba}$ (c'est-à-dire que le nombre d'individus recrutés du groupe *a* au groupe *b* est égal à celui du groupe *b* au groupe *a*), alors $\hat{P}_a^{VH} = \hat{P}_a^{R}$ (Volz et Heckathorn, 2008).

4.4 Estimation de la variance

Pour estimer la variance de l'estimateur $\hat{P}_i^R$ donné par (10), Heckathorn (2002) a proposé une méthode similaire au bootstrap. Les étapes sont les suivantes :

1. On sélectionne au hasard un des germes k_0 de la vague 0. Supposons que $k_0 \in i$, c'est-à-dire que l'individu k_0 appartient au groupe *i*.
2. On sélectionne au hasard k_1, un des individus recrutés à la vague 1 par le germe k_0 en se basant sur les proportions $\hat{R}_{ij}$. Supposons que $k_1 \in i'$, c'est-à-dire que l'individu k_0 appartient au groupe i'.
3. On sélectionne au hasard k_2, un des individus recrutés par k_1 à la vague 2 en se basant sur les proportions $\hat{R}_{i'j}$.
4. On continue jusqu'à ce qu'on dispose de *n* individus, ce qui constitue une réplique *g*.
5. On calcule l'estimations $\hat{P}_i^{R,g}$ pour chaque groupe *i*.
6. On recommence le processus *G*=10 000 fois, disons.

La variance de $\hat{P}_i^R$ se calcule alors selon

$$\widehat{Var}(\hat{P}_i^R) = \frac{1}{G}\sum_{g=1}^{G}\left(\hat{P}_i^{R,g} - \hat{\bar{P}}_i^R\right)^2 \qquad (14)$$

où $\hat{\bar{P}}_i^R = \sum_{g=1}^{G} \hat{P}_i^{R,g} / G$ est la moyenne des estimations pour le groupe *i*.

Notons que cette approche bootstrap peut aussi s'appliquer aux estimateurs $\hat{P}_i^E$ et $\hat{P}_i^{VH}$ donnés respectivement par (4) et (12). Il est aussi possible de procéder par sélection avec remise de germes de la vague 0. On note cependant peu de différence dans les résultats (Heckathorn, 2002).

4.5 Estimations pour la population entière

Rappelons que les proportions P_i se rapportent à la population cible *U* seulement. Par exemple, P_a peut représenter la proportion de femmes **parmi** les musiciens de jazz. On peut cependant aussi s'intéresser à l'estimation des effectifs N_i et des totaux $Y_i = \sum_{k=1}^{N_i} y_k$ où N_i

est le nombre d'individus appartenant au groupe *i*. De plus, on peut vouloir produire des estimations de proportions P_i^* par rapport à la population **entière** U^* de taille N^*, où $U \subseteq U^*$ et $N \leq N^*$. Pour revenir à l'exemple précédent, N_i serait le nombre de femmes musiciennes de jazz et P_i^*, la proportion de femmes musiciennes de jazz parmi **toute** la population.

Pour produire des estimations de quantités telles que N_i, Y_i ou encore P_i^*, Heckathorn et Jeffri (2001) ont proposé d'utiliser l'estimation par capture-recapture (Seber, 1982, et Thompson, 2002). Pour ce faire, on utilise, par exemple, une liste externe des individus de la population cible *U*. Notons que cette liste est nécessairement incomplète puisque la population cible est une population difficile à joindre pour laquelle, par définition, nous n'avons pas de base de sondage adéquate. Par exemple, dans le cas des musiciens de jazz, on pourrait utiliser la liste des musiciens de jazz **inscrits** comme musiciens professionnels.

Soit n_i^L, le nombre d'individus de la liste externe appartenant au groupe *i*; n_i^{SDR}, le nombre d'individus échantillonnés par SDR au sein du groupe *i*; et $n_i^{L\&SDR}$, le nombre d'individus appartenant aux deux. On peut alors estimer N_i, Y_i et P_i^* en utilisant respectivement

$$\hat{N}_i = \frac{n_i^{SDR} \times n_i^L}{n_i^{L\&SDR}} \tag{15}$$

$$\hat{Y}_i = \frac{\left(\sum_{k=1}^{n_i^{SDR}} y_k\right)\left(\sum_{k=1}^{n_i^L} y_k\right)}{\sum_{k=1}^{n_i^{L\&SDR}} y_k} \tag{16}$$

et

$$\hat{P}_i^* = \frac{\hat{N}_i}{N^*} \tag{17}$$

On note que l'estimation (17) requiert de connaître la taille N^* de la population **entière** U^*, ce qui n'est souvent pas disponible.

5. Problèmes potentiels et effets sur les estimations

Le SDR possède plusieurs problèmes potentiels reliés à la violation des hypothèses de base utilisées dans la construction des estimateurs. Quelques-uns de ces problèmes sont abordés dans la présente section.

5.1 Convergence asymptotique

Quel doit être le nombre de vagues requises pour que les résultats asymptotiques précédents soient utilisables? En réponse à cette question, Heckathorn (1997) a donné le Théorème 2 suivant :

L'échantillon d'individus cibles obtenu par SDR converge rapidement vers l'équilibre selon un taux de convergence géométrique.

Donc, selon ce théorème, peu de vagues sont nécessaires pour que les résultats soient utilisables. C'est d'ailleurs ce que l'on constate à la Figure 4 suivante où la composition de l'échantillon semble s'équilibrer après seulement cinq vagues.

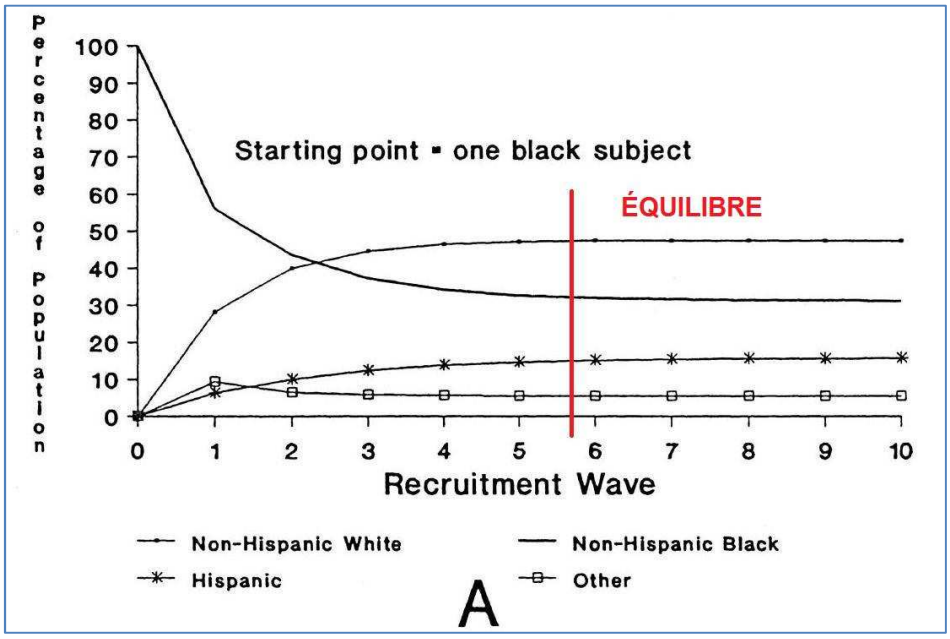

Figure 3a. Proportions d'individus recrutés à l'équilibre

5.2 Chaîne de Markov

Avec une chaîne de Markov régulière d'ordre un, on suppose que l'on passe d'un état à **un seul** autre état. On suppose donc qu'un individu ne recrute qu'un seul autre individu, c'est-à-dire que κ=1. Cependant, en général, $\kappa \geq 1$ et on choisit souvent en pratique $\kappa = 3$ ou 4, ce qui produit un recrutement arborescent comme le montre la Figure 1.

Tel que le mentionne Heckathorn (2002) cependant, une structure arborescente peut s'analyser comme un ensemble de structures linéaires : chaque individu enquêté a été recruté par un seul recruteur, qui a été recruté par un seul recruteur, et ainsi de suite. Ce dernier ajoute que l'on doit vérifier empiriquement que le modèle markovien s'applique en comparant la composition de l'échantillon avec celle espérée théoriquement s'il agissait bien d'une chaîne de Markov, c'est-à-dire les données à l'équilibre.

5.3 Biais d'homophilie

Les individus ne recrutent généralement pas selon les mêmes proportions entre les groupes. La tendance est de recruter davantage dans son propre groupe (*homophilie*), c'est-à-dire que $R_{ii} > R_{ij}$ pour $i \neq j$.

Soit π_i^H, la probabilité qu'il y ait homophilie dans le recrutement pour le groupe *i*. On suppose ainsi que

$$R_{ii} = \pi_i^H + (1-\pi_i^H)P_i \qquad (18)$$

où P_i est la proportion de membres du groupe *i* au sein de *U*. Pour le cas des deux groupes *a* et *b*, on obtient à partir de (2):

$$P_a^E = \frac{P_a(1-\pi_b^H)}{1-\pi_a^H + P_a(\pi_a^H - \pi_b^H)} \tag{19}$$

Pour que $\hat{P}_a^E$ donné par (4) estime sans biais P_a, il faut que $P_a^E = P_a$. Ceci est le cas si : (i) $\pi_i^H = 0$ et, dans ce cas, il n'y a pas d'homophilie; ou (ii) $\pi_i^H = \pi_j^H$, c'est-à-dire que la probabilité d'homophilie du recrutement est la même entre les groupes. Ce dernier résultat a mené au Théorème 3 de Heckathorn (1997) :

> *Un échantillon obtenu par SDR est sans biais si la probabilité d'homophilie dans le recrutement est la même entre les groupes*.

En pratique, on peut s'attendre à ce que la probabilité d'homophilie π_i^H soit au moins positivement corrélée entre les groupes. En effet, selon le Principe de Simmel (1955), *des ennemis communs augmentent la solidarité au sein de groupes.* À partir de simulations, Heckathorn (1997) a montré que la violation de l'hypothèse d'égalité des π_i^H doit être grande pour biaiser substantiellement l'échantillon.

5.4 Biais de recrutement différentiel

Il y a *recrutement différentiel* lorsqu'un groupe donné a un recrutement particulièrement efficace. Ce dernier tend à surreprésenter certains groupes au sein de l'échantillon. Cependant, l'utilisation d'un nombre limité de noms κ tend à réduire ce biais. Par ailleurs, l'estimation à partir du modèle de réciprocité corrige aussi en partie ce biais (Heckathorn et Jeffri, 2001). Heckathorn (2002) souligne que l'équilibre de la chaîne de Markov n'est pas influencé par ce biais puisque l'estimateur $\hat{P}_i^E$ donné par (4) dépend des $\hat{R}_{ij}$ qui sont des **proportions** de recrutement pour un groupe donné *i*. Donc, si un groupe donné *i* recrute avec plus d'efficacité, $\hat{R}_{ij}$ reste inchangé.

5.5 Biais dû à la différence des réseaux de connaissances

Dans le cas des deux groupes *a* et *b*, pour que $\hat{P}_a^R$ donné par (10) estime sans biais P_a, il faut que $P_a^R = P_a$. Par ailleurs, si les degrés ϑ_k des répondants *k* sont égaux, les $\bar{\vartheta}_i$ sont alors égaux entre les groupes. Dans le cas des deux groupes, on a alors $\bar{\vartheta}_a = \bar{\vartheta}_b$ et $R_{ab} = 1 - R_{aa}$. On peut ainsi déduire que

$$P_a^R = \frac{\bar{\vartheta}_b R_{ba}}{\bar{\vartheta}_b R_{ba} + \bar{\vartheta}_a R_{ab}} = \frac{\bar{\vartheta}_b R_{ba}}{\bar{\vartheta}_b R_{ba} + \bar{\vartheta}_b (1 - R_{aa})} = \frac{R_{ba}}{1 - R_{aa} + R_{ba}} = P_a^E \tag{20}$$

On obtient donc $P_a^R = P_a^E$, les proportions à l'équilibre. Heckathorn (2002) a déduit de ce résultat le Théorème 4 suivant :

Un échantillon obtenu par SDR est sans biais dans le sens où $P_a^R = P_a^E$ *si et seulement si les degrés moyens des groupes sont égaux.*

Donc, l'estimateur (10) corrige le biais dû à la différence de réseaux de connaissances en tenant compte des degrés moyens $\overline{\vartheta}_i$.

6. Conclusion

Le SDR est très utilisé en pratique pour enquêter auprès de populations cachées, difficiles à joindre. Comme on l'a vu cependant, le SDR n'est pas en soi une méthode de sondage probabiliste et l'estimation repose sur plusieurs hypothèses plus ou moins vérifiées en pratique. Il est donc important d'étudier la robustesse du SDR par rapport à la violation des différentes hypothèses. Mentionnons, notamment, les articles de Gile et Handcock (2010) et de Lu et coll. (2012) qui attaquent ce genre d'étude.

À la lumière de ce qui se trouve dans la littérature, il semble en conclusion que le SDR appartienne à cet ensemble de méthodes non probabilistes (par exemple, le Sondage par quotas) qui, malgré plusieurs problèmes potentiels théoriques et pratiques, fournissent à la fin des estimations raisonnables.

Bibliographie

Coleman, J.S. (1958). Relational analysis: The study of social organization with survey methods. *Human Organization*, Vol. 17, pp. 28-36.

Erickson, B.H. (1979). Some problems of inference from chain data. *Sociological Methodology*, Vol. 10, pp. 276-302.

Frank, O. (1977). Survey sampling in graphs. *Journal of Statistical Planning and Inference*, Vol. 1, pp. 23-5264.

Frank, O. (1978). Sampling and estimation in large social networks. *Social Networks,* Vol. 1, pp. 91-101.

Frank, O., Snijders, T. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, Vol. 10, pp. 53-67.

Gile, K.J., Handcock, M.S. (2010). Respondent-driven sampling : An assessment of current methodology. *Sociological Methodology,* Vol. 40, pp. 285-327.

Goodman, L.A. (1961). Snowball sampling. *Annals of Mathematical Statistics*, Vol. 32, No. 1, pp 148-170.

Hansen, MH., Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, Vol. 14, No. 4, pp. 333-362.

Heckathorn, D.D. (1990). Collective sanctions and compliance norms : A formal theory of group-medicated social control. *American Sociology Review*, Vol. 55, pp. 366-384.

Heckathorn, D.D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems,* Vol. 44, pp. 174-199.

Heckathorn, D.D. (1999). Respondent-driven sampling: Applications and extensions. Article présenté aux *Public Health Colloquium Series*, School of Public Health, University of Illinois at Chicago.

Heckathorn, D.D. (2002). Respondent-driven sampling II : Deriving valid population estimates from chainreferral samples of hidden populations. *Social Problems*, Vol. 49, pp. 11-34.

Heckathorn, D. D., Jeffri, J. (2001). Finding the beat: Using respondent-driven sampling to study jazz musicians. *Poetics*, Vol. 28, pp.307–329.

Kemeny, J.G., Snell, J.L. (1960). *Finite Markov Chains*. Van Nostrand, Princetown, New Jersey.

Lu, X., Bengtsson, L., Britton, T., Camitz, M., Kim, B.J., Thorson, A., Lijeros, F. (2012). The sensitivity of respondent-driven sampling. *Journal of the Royal Statistical Society*, A, Vol. 175, Part 1, pp. 191-216.

Salganik, M.J., Heckathorn, D.D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. *Social Methodology*, Vol. 34, pp. 193-239.

Seman, S. (2010). Échantillonnage espace-temps et échantillonnage déterminé selon les répondants des populations difficiles à joindre. *Methodological Innovations Online*, Vol. 5, No. 2, pp. 60-75.

Seber, G.A.F. (1982). The Estimation of Animal Abundance and Related Parameters, Second Edition. Griffin, Londres.

Simmel, G. (1955). *Conflict: The web of group affiliations*. Trans. K.H. Wolff and R. Bendix, Free Press, New York.

Thompson, S.K. (2002). *Sampling, 2nd Edition.* John Wiley and Sons, New York, 400 pages.

Volz, E., Heckathorn, D.D. (2008). Probability based estimation theory for respondent-driven sampling. *Journal of Official Statistics*, Vol. 24, No. 1, pp. 79-97.