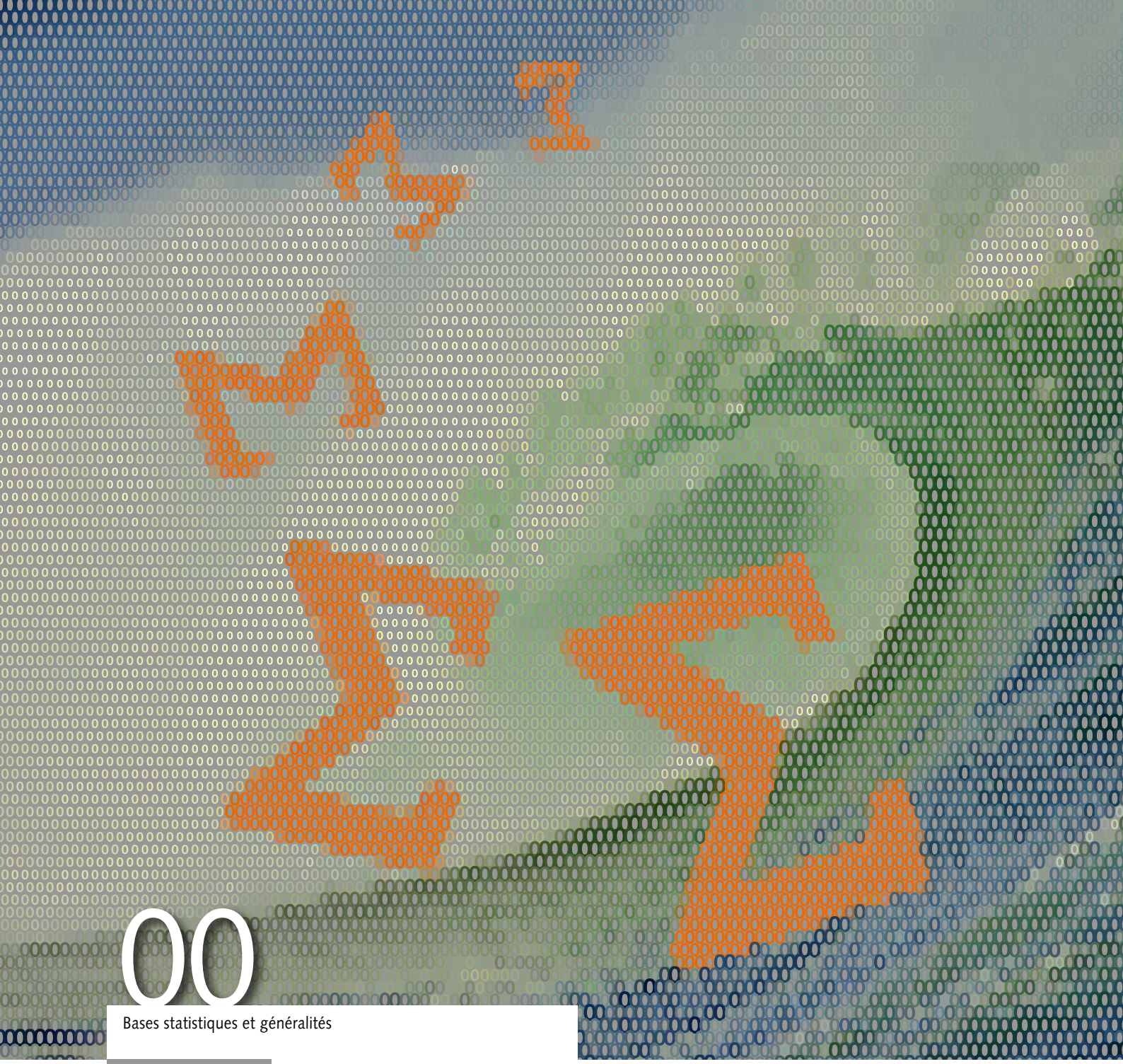


00

Bases statistiques et généralités

338-0071

Rapport de méthodes

Echantillonnage boule de neige

La méthode de sondage déterminé par les répondants

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Département fédéral de l'intérieur DFI
Office fédéral de la statistique OFS

Neuchâtel, 2014

La série «Statistique de la Suisse»
publiée par l'Office fédéral de la statistique (OFS)
couvre les domaines suivants:

0 Bases statistiques et généralités

1 Population

2 Espace et environnement

3 Vie active et rémunération du travail

4 Economie nationale

5 Prix

6 Industrie et services

7 Agriculture et sylviculture

8 Energie

9 Construction et logement

10 Tourisme

11 Mobilité et transports

12 Monnaie, banques, assurances

13 Protection sociale

14 Santé

15 Education et science

16 Culture, médias, société de l'information, sport

17 Politique

18 Administration et finances publiques

19 Criminalité et droit pénal

20 Situation économique et sociale de la population

21 Développement durable et disparités régionales et internationales

Rapport de méthodes

Echantillonnage boule de neige

La méthode de sondage déterminé par les répondants

Auteur Matthieu Wilhelm, Université de Neuchâtel

Editeur Office fédéral de la statistique (OFS)

Office fédéral de la statistique (OFS)
Neuchâtel, 2014

Editeur: Office fédéral de la statistique (OFS)

Complément d'information: Jean-Pierre Renfer, tél. 032 713 66 62, e-mail: Jean-Pierre.Renfer@bfs.admin.ch
Matthieu Wilhelm, Université de Neuchâtel, e-mail: Matthieu.Wilhelm@unine.ch

Réalisation: Section Méthodes statistiques, OFS

Diffusion: Office fédéral de la statistique, CH-2010 Neuchâtel
tél. 032 713 60 60 / fax 032 713 60 61 / e-mail: order@bfs.admin.ch

Internet: www.statistique.admin.ch

Numéro de commande: 338-0071

Prix: gratuit

Série: Statistique de la Suisse

Domaine: 0 Bases statistiques et généralités

Langue du texte original: Français

Page de couverture: OFS; concept: Netthoevel & Gaberthüel, Bienne; photo: © NorthShoreSurfPhotos – Fotolia.com

Graphisme/Layout: Section DIAM, Prepress/Print

ISBN: 978-3-303-00515-6

Table des matières

1 Introduction **7**
- 1.1 La méthode RDS et ses applications . . . 7
- 1.2 Bref historique de la méthode RDS . . . 8
- 1.3 Description de la méthode RDS . . . 8

2 Estimateurs **10**
- 2.1 Notations . . . 10
- 2.2 L'estimateur naïf . . . 11
- 2.3 Considérations mathématiques sur l'estimateur naïf . . . 12
- 2.4 L'estimateur de Salganik-Heckathorn . . . 14
- 2.5 L'estimateur de Salganik-Heckathorn, côté maths . . . 16
- 2.6 L'estimateur d'Heckathorn . . . 18
- 2.7 L'estimateur de Volz-Heckathorn . . . 19
- 2.8 L'estimateur de Volz-Heckathorn, côté maths . . . 20
- 2.9 Hypothèses mathématiques des estimateurs classiques . . . 22
- 2.10 Sensibilité des estimateurs . . . 27

3 Simulations **27**
- 3.1 L'homophilie . . . 28
  - 3.1.1 L'homophilie de processus . . . 29
  - 3.1.2 L'homophilie de réseau . . . 30
  - 3.1.3 Homophilie de réseau et homophilie de processus . . . 31
- 3.2 Dépendance aux germes . . . 32
- 3.3 Tirage avec ou sans remise . . . 33
- 3.4 Non-réponse . . . 36
- 3.5 Conclusion sur les simulations . . . 39

4 Conclusions **40**
- 4.1 A l'épreuve de la réalité . . . 40
- 4.2 Evaluation de la méthode RDS pour l'OFS . . . 40

Annexes **42**

A Eléments de théorie des processus stochastiques **42**

A Simulations supplémentaires **44**

Rapports de méthodes de la section méthodes statistiques **51**

Préambule

Dans le cadre du master en ingénierie mathématique de l'école polytechnique fédérale de Lausanne, le troisième des quatre semestres est consacré à un stage. Ayant pris une orientation plutôt statistique, j'ai souhaité postuler pour un stage dans la section de méthodologie de l'OFS. Le sujet de stage était déjà fixé. Suite à une demande externe concernant la méthode RDS, la section de méthodologie avait souhaité se renseigner sur la question et avait finalement considéré que ce sujet conviendrait parfaitement à un stagiaire. J'ai donc travaillé sur la méthode d'échantillonnage déterminée par les répondants (RDS) entre août et décembre 2012. Il en a résulté deux rapports distincts, un code informatique permettant de simuler des processus RDS ainsi qu'un guide de l'utilisateur pour ce code [1].
Je tiens à remercier chaleureusement Jean-Pierre Renfer, qui m'a suivi durant toute la période de mon stage. Son sérieux et son enthousiasme ont été contagieux. Je tiens aussi à remercier Philippe Eichenberger, chef de la section de méthodologie statistique de l'OFS. Son accueil au sein de son groupe et l'atmosphère qu'il y a imprégnée ont rendu cette période extrêmement agréable. Finalement, je souhaiterais remercier vivement tous les membres de la section de méthodologie statistique. Ils ont été extrêmement accueillants et chaleureux. Ils m'ont aussi permis de me sentir faire partie d'une équipe motivée, dynamique et très compétente. Ces mois furent tout simplement incroyables, je le leur dois en grande partie.
Je remercie Jean-Marc Nicoletti, Jean-Pierre Renfer et Marie Dupraz pour la relecture attentive de ce rapport, qui m'a permis de sensiblement l'améliorer.
Par ailleurs, ma vie à Neuchâtel n'aurait pas eu la même saveur sans mes grands-parents, Françoise et François Wilhelm qui m'ont régulièrement offert un asile et de bons petits plats.

Résumé

L'objectif de ce rapport est de présenter la méthode d'échantillonnage déterminée par les répondants, plus connue sous son acronyme anglais, RDS (Respondent-Driven Sampling). Après une très brève introduction à la théorie des sondages, la méthode RDS est introduite dans son contexte historique. On présente ensuite les différents estimateurs utilisés dans la plupart des études existantes. Le développement mathématique de ces estimateurs ainsi que les hypothèses nécessaires à leur dérivation sont passés en revue. On fait un survol théorique des estimateurs couramment utilisés, de leurs propriétés et de leurs faiblesses. Les estimateurs les plus récents font l'objet d'un autre rapport. On fait de plus des simulations afin d'illustrer certaines situations. On conclut à un préavis négatif pour l'utilisation de cette méthode, du moins dans l'état actuel de la recherche. On conseille donc d'y renoncer à ce stade dans les enquêtes devant satisfaire aux exigences de la statistique publique.

1. disponible sur http://www.unine.ch/members/matthieu.wilhelm

1 Introduction

Le contexte de la théorie des sondages

La théorie des sondages a commencé à se développer au siècle dernier. La raison de ce développement est apparue en même temps que le concept d'un état moderne dont les besoins en statistiques sont importants. Toutefois, pendant longtemps, la statistique se concevait comme quelque chose d'exhaustif, ne devant pas être le reflet d'une réalité, mais bel et bien la réalité elle-même, condensée en quelques chiffres. En effet, à l'époque le principe même d'échantillonnage a été totalement rejeté. Il était considéré comme peu sérieux, le recensement étant alors le seul moyen valable d'établir des statistiques. Dans le courant du XX$^{\text{ième}}$ siècle, et en particulier pendant sa seconde moitié, une base mathématique a permis à la théorie des sondages de devenir l'un des domaines importants des statistiques. De nombreuses recherches ont permis une amélioration notable des méthodes liées aux sondages (Tillé, 2001). C'est encore à l'heure actuelle un domaine en pleine expansion, sur lequel de nombreux groupes de recherche travaillent à travers le monde.
Ce développement mathématique de la théorie des sondages a surtout été possible grâce à l'apport de la théorie des probabilités. On distingue deux types de sondages : les sondages probabilistes et les sondages non-probabilistes (appelés aussi empiriques). Le sondage probabiliste attribue une probabilité non-nulle à tous les individus faisant partie de la population que l'on souhaite étudier, puis sélectionne un échantillon de cette population en fonction de ces probabilités. Le sondage non-probabiliste est caractérisé par une hypothèse supplémentaire au sujet de la population cible : l'homogénéité de la distribution du caractère observé. Ainsi, la sélection des individus participant à l'enquête n'est pas nécessairement purement aléatoire. C'est une hypothèse très forte, qui ne se vérifie que rarement.
On peut donner quelques exemples afin d'illustrer ces concepts. Premièrement, on définit un plan de sondage comme une loi de probabilité définie sur toute la population. Ainsi, tout sondage probabiliste est défini par son plan de sondage. Le sondage probabiliste le plus simple est lorsque l'on suppose que chaque individu a une même probabilité d'être choisi, c'est à dire que la loi de probabilité définie par le plan de sondage est uniforme. De fait, il est appelé *plan simple*. Le principal problème des sondages non-probabilistes est qu'il n'est pas possible d'estimer le biais ni la variance d'une estimation sur un tel échantillon. C'est pourquoi, sauf dans des cas bien particuliers, la statistique officielle n'utilise pas d'échantillon non-probabiliste.
Pour plus de détails au sujet de la théorie des sondages, on peut se référer à (Särndal *et al.*, 1992; Cochran, 1977; Tillé, 2001).

1.1 La méthode RDS et ses applications

Parmi les méthodes non-probabilistes, une méthode en particulier est l'objet de notre attention : la méthode d'échantillonnage dite « boule de neige ». Cette dernière a été développée par Leo A. Goodman (Goodman, 1961). Cette méthode non-probabiliste propose d'échantillonner une population de la manière suivante : dans un premier temps on fait un tirage aléatoire au sein de la population cible. Puis, on demande à chacun des individus ayant été sélectionnés dans ce premier tirage d'inclure k « ami(s) » dans l'enquête. Ces derniers sont admis dans l'enquête s'ils n'y sont pas déjà présents, c'est-à-dire qu'il ne font pas partie du tirage initial. Finalement, on peut itérer cette opération s fois. Une telle procédure d'échantillonnage est appelée « procédure d'échantillonnage boule de neige à s étapes et k noms. Cette dernière a avant tout pour but d'augmenter la taille d'un échantillon et de faire ensuite des estimations sur le nombre de liens bilatéraux ou de triangles au sein du réseau. Cette idée intéressante n'avait à l'origine pas pour but de permettre une inférence sur la population considérée mais plutôt de pouvoir faire une

inférence sur les caractéristiques du réseau social, telles que les liens. Considérée comme méthode d'échantillonnage à part entière, la procédure d'échantillonnage boule de neige est non-probabiliste puisqu'il est a priori impossible de quantifier les probabilités de sélection dans l'échantillon. On voit donc que les deux approches, celle qui vise à constituer un échantillon d'une population cible et celle qui visait à reconstruire un réseau, sont différentes. Il est donc important d'être attentif à l'usage de l'expression « échantillonnage boule de neige »(Gile et Handcock, 2012), celui-ci pouvant parfois prêter à confusion.
La méthode d'échantillonnage déterminée par les répondants, respondent-driven sampling en anglais, qui sera abrégée RDS ci-après, appartient à la famille des méthodes boule de neige. L'idée principale de l'échantillonnage boule de neige est d'augmenter la taille d'un échantillon en utilisant les réseaux sociaux des personnes recrutées. La technique RDS reprend cette idée et l'adapte de manière à être applicable. Cependant, le but visé par la méthode RDS est tout à fait différent puisqu'il s'agit d'estimer des proportions de sous-populations au sein de la population échantillonnée. En particulier, la méthode RDS a été développée par des chercheurs venus du domaine de la sociologie dans le but de faire des études sur des populations dites difficiles à atteindre. La nature même de ce type de population empêche toute procédure d'échantillonnage traditionnelle, qui nécessite des bases de données. Il est important de pouvoir comprendre les personnes qui vivent en marge de la société. Il est donc essentiel de pouvoir utiliser une méthode d'échantillonnage qui soit applicable, dont on peut tirer des estimations non-biaisées et dont les erreurs sont quantifiables. La nature non-probabiliste de ce type d'échantillonnage semble rendre impossible une quelconque inférence. Toutefois, en faisant des hypothèses supplémentaires sur la population cible, de récents travaux semblent indiquer qu'une inférence est peut-être envisageable dans certains cas.
Le but de ce rapport est de faire une synthèse des travaux accomplis dans ce domaine jusqu'ici. De plus, on s'efforce d'avoir un regard critique sur les différents aspects que l'on aborde et d'apporter, lorsque cela semble nécessaire et possible, certaines précisions.

1.2 Bref historique de la méthode RDS

La méthode d'échantillonnage RDS a été développée par Douglas D. Heckathorn, relativement récemment (Heckathorn, 1997, 2002). Elle avait avant tout pour but d'être appliquée dans le domaine de la recherche sociologique sur les populations difficiles à atteindre, telles que les personnes toxicomanes ou séropositives. Toutefois, les fondements mathématiques se sont révélés quelque peu ténus, les hypothèses permettant de faire des estimations étant souvent irréalistes. L'utilisation en constante augmentation de cette méthode, plus de 130 études en 2008 déjà (Johnston *et al.*, 2008), a poussé un certain nombre d'autres chercheurs, à s'y intéresser. Ces derniers ont permis un développement rapide de nouveaux estimateurs, moins sensibles lorsque les hypothèses ne sont que partiellement satisfaites et plus fondés mathématiquement.
Aujourd'hui, et bien que cette méthode n'ait pas toujours été utilisée avec toutes les précautions nécessaires (Gile et Handcock, 2010; Heimer, 2005; McCreesh *et al.*, 2012), ses récents développements sont prometteurs. On pourrait envisager son utilisation dans le cadre de la statistique officielle, pour autant que la méthodologie y relative soit conforme au code de bonnes pratiques.

1.3 Description de la méthode RDS

Concrètement, la méthode se présente de la manière suivante :

1. On choisit, théoriquement aléatoirement, un certain nombre de « germes » parmi la population cible. Ils constituent la vague 0. Ces individus sont ceux qui engendrent l'échan-

tillon, d'où leur nom.

2. On demande à ces germes de participer à l'enquête et de distribuer un nombre limité de coupons (de 1 à 3, selon les enquêtes), identifiables, à des personnes de leur entourage faisant partie de la population cible. Ainsi les recrues deviennent recruteurs. Les coupons sont en fait des « permis de participer » à l'enquête. De plus, ils permettent d'identifier le recruteur qui donne le coupon.
3. On répète le processus jusqu'à ce que la taille de l'échantillon désirée soit atteinte.

Grâce aux coupons, on est à la fois capable d'identifier les individus mais aussi de reconstruire partiellement le réseau social des personnes ayant participé à l'enquête. On verra plus tard qu'il s'agit d'un aspect essentiel. On peut voir une illustration d'un échantillonnage RDS sur un réseau sur la figure 1.

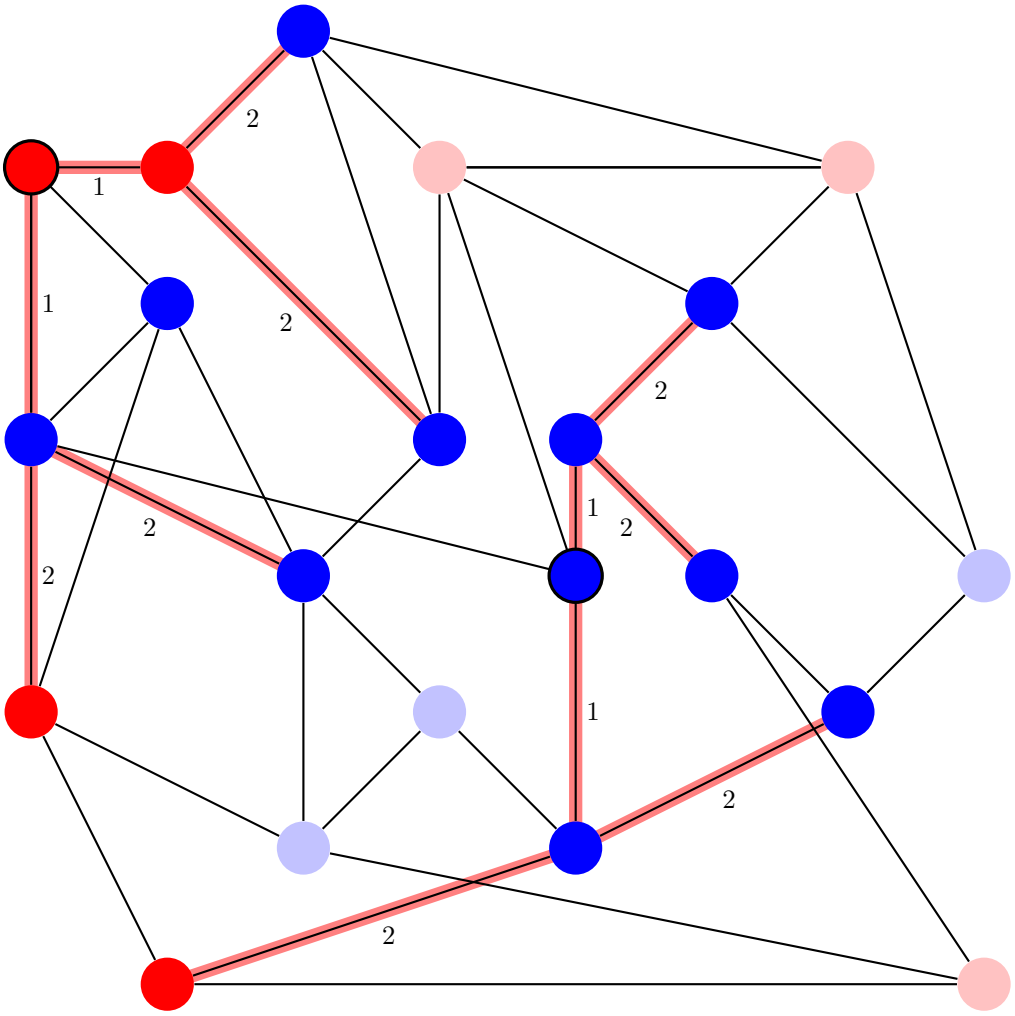


FIGURE 1 Exemple de réseau sur lequel a lieu un échantillonnage RDS. Dans cet exemple, on suppose qu'il y a deux catégories, infectés et sains. Les individus en rouge appartiennent à la catégorie « infectés » et les bleus à la catégories « sains ». Les points dont le contour est plus foncé sont les germes et les points en pastel sont ceux qui ne sont pas sélectionnés. Les arêtes qui sont surlignées sont celles utilisées dans le processus de recrutement. De plus, on peut voir le long des arêtes le numéro de la vague.

En général, on cherche à estimer des proportions au sein de populations difficiles à atteindre, et donc tous les estimateurs présentés dans ce qui suit tentent d'atteindre ce but. On peut citer

comme exemple, des études visant à estimer la proportion de personnes séropositives au sein de la population de personnes toxicomanes, dans une ville donnée.
En général, pour représenter un réseau social, on utilise les graphes. Sur ce graphe, on représente en général les individus par un sommet et les liens entre les individus sont représentés par une arête, appelée parfois lien par la suite, entre les sommets correspondants.
On remarque donc d'emblée que la méthode RDS est au carrefour de plusieurs domaines des mathématiques. Entre la théorie des sondages, la théorie des chaînes de Markov et la théorie des réseaux. Pour plus d'informations sur ces sujets, on peut se référer en ce qui concerne les chaînes de Markov, à (Karlin et Taylor, 1975) ou à (Durett, 2010). En ce qui concerne la théorie des réseaux et en particuliers sur la simulation de réseaux, on peut se référer à (Newman, 2003) et à (Newman *et al.*, 2001). Pour plus d'informations sur la théorie des graphes, on peut se référer à (Diestel, 2010).

2 Estimateurs

Dans un premier temps, plusieurs estimateurs ont été développés (Heckathorn, 1997, 2002; Salganik et Heckathorn, 2004; Volz et Heckathorn, 2008). L'ensemble de ces estimateurs sont dits classiques car ils ont été largement utilisés (Johnston *et al.*, 2008) dans des études, et le sont encore à l'heure actuelle (McCreesh *et al.*, 2012). D'autres estimateurs ont été développés très récemment, principalement sous l'impulsion de Krysta J. Gile et de Mark. S Handcock (Gile, 2012; Gile et Handcock, 2011). Ces deux courants de recherches ont amené au sujet une dynamique qui a conduit à de rapides progrès dans ce domaine.La table 2 contient l'expression ainsi que l'idée principale qui a mené à l'élaboration de ces estimateurs.
Les premiers estimateurs ont été développés par Douglas D. Hekathorn et ses collègues (Heckathorn, 1997, 2002; Salganik et Heckathorn, 2004; Volz et Heckathorn, 2008). On présente ici ces quelques estimateurs, ainsi que les hypothèses mathématiques qui sont faites.
On cherche à échantillonner une population difficile d'accès, elle-même partitionnée en plusieurs sous-ensembles distincts, dont on veut estimer la proportion au sein de la population cible. Tous les estimateurs visent à atteindre ce but. La plupart du temps, on ne considère que deux sous-ensembles, infectés et bien-portants.

2.1 Notations

Avant de poursuivre, on va préciser quelques notations. De manière formelle, on suppose que l'on a une population cible S, de taille N, et une partition de l'ensemble $S = \{S_1, \ldots, S_K\}$ où les cardinalités des ensembles $S_1, \ldots, S_K$ sont données respectivement par $N_1, \ldots, N_K$. Supposons que l'on obtienne un échantillon s de taille n de la population S. Parmi les n individus appartenant à notre échantillon, n_1 appartiennent S_1, n_2 à S_2 et ainsi de suite. On pose encore n_{ij} le nombre d'individus appartenant à S_i ayant recruté un individu appartenant à S_j. Finalement, on pose $s_1 = S_1 \cap s$, $s_2 = S_2 \cap s$ et ainsi de suite. En général, les lettres en minuscules font référence à l'échantillon alors que les lettres en majuscules font référence à l'ensemble de la population cible. On remarque que, pour des questions de notations usuelles, n désigne aussi le nombre de pas effectués par la chaîne de Markov.
On note en général π_i la quantité $\frac{N_i}{N}$, que l'on veut estimer au sein de la population cible. On utilise aussi la notation $I_X(i)$ pour désigner la fonction indicatrice d'un ensemble X quelconque. Plus précisément, on a :

$$I_X(i) = \begin{cases} 1 & \text{si } i \in X, \\ 0 & \text{si } i \notin X. \end{cases}$$

La plupart des notations sont résumées dans la table 1.

Notation	Signification
S	ensemble désignant la population cible
s	échantillon de la population cible
N	nombre d'individus de la population cible
n	nombre d'individus de l'échantillon
M	nombre de sous-ensembles disjoints de S
$S_1, \ldots, S_M$	sous-ensembles disjoints de S
$s_1, \ldots, s_M$	sous-ensembles correspondants dans l'échantillon
$\pi_1, \ldots, \pi_M$	proportions des sous-ensembles au sein de la population cible S
P	matrice des probabilités de transition
$p_1, \ldots, p_N$	probabilités d'un individu d'appartenir à l'échantillon
$I_X(i)$	fonction indicatrice de l'ensemble X appliquée à l'individu i
d_i	nombre de degrés, i.e de connexions au sein du réseau, d'un individu i
$\overline{d}_S$	degré moyen de la population cible
$\overline{d}_s$	degré moyen de l'échantillon
R_i	somme des degrés d'un sous-ensemble i
T_{ij}	nombre de liens entre les sous-ensembles i et j
$\hat{x}$	estimation de la variable x
$\overline{x}$	valeur moyenne de la variable x

TABLE 1 Table récapitulative des notations.

2.2 L'estimateur naïf

Dans son article, Heckathorn (Heckathorn, 2002), introduit la notion de processus de Markov. Il résume cette notion en considérant qu'il s'agit d'un processus qui a deux principales caractéristiques. La première, c'est qu'il a un nombre fini d'états, et la deuxième est que la probabilité d'être recruté ne dépend que de la nature du recruteur, et non de la nature des deux (ou plus) recruteurs précédents (pour une définition plus mathématique, voir annexe A). Ces deux hypothèses semblent vérifiées dans de nombreux cas. On considère que le recrutement est le processus qui fait passer d'un état à un autre, où l'état représente le sous-ensemble auquel appartiennent les individus de l'échantillon (voir annexe A).
Il modélise donc le processus de recrutement de la manière suivante : il suppose qu'il s'agit d'une chaîne de Markov, dont les états sont les sous-ensembles de la population cible. Il inventorie, pour chaque élément de l'échantillon à quel sous-ensemble il appartient et à quel sous-ensemble son recruteur appartient. Ainsi, on peut estimer les probabilités de transition d'un état à un autre. Pour illustrer la situation, prenons l'exemple d'une étude qui aurait pour but d'étudier la distribution raciale des personnes toxicomanes (Heckathorn, 1997). Supposons que l'on ait une partition des personnes toxicomanes entre afro-américains, hispaniques, asiatiques, blancs, et autres. Ainsi, d'après notre échantillonnage, on peut estimer les probabilités de transition, c'est-à-dire que la probabilité qu'un afro-américain recrute un autre afro-américain, qu'un afro-américain recrute un blanc, qu'un hispanique recrute un afro-américain etc... Il existe un théorème qui stipule, sous certaines hypothèses, qu'au bout d'un certain temps, la chaîne commence à avoir un comportement stationnaire, c'est-à-dire que la probabilité d'être dans un état donné ne dépend plus de l'état initial (voir théorème 6, en annexe A). Il s'agit d'une sorte d'équilibre. C'est sur ce théorème que la démarche d'estimation d'Heckathorn (Heckathorn, 1997, 2002) se base. Son hypothèse est que la distribution stationnaire de la chaîne de Markov représente les différentes proportions des sous-ensembles dans la population totale.
Dans un premier temps, on estime les probabilités de transition. Par exemple, on estime $\widehat{P}_{ij} = \frac{n_{ij}}{n_i}$. On peut donc construire la matrice des probabilités de transition P. Finalement, on résout

le problème au valeurs propres :

$$\widehat{P}^T \hat{\boldsymbol{\pi}} = \hat{\boldsymbol{\pi}} \tag{1}$$

pour obtenir $\hat{\boldsymbol{\pi}}$ et donc $\hat{\pi}_i$, qui sera un estimateur de $\frac{N_i}{N}$. On a donc finalement :

$$\mu_n = \hat{\boldsymbol{\pi}},$$

où μ_n désigne simplement l'estimateur naïf, présenté dans (Heckathorn, 1997).
Dans le cas où les données sont ajustées, cela force la matrice P à être symétrique (voir section 2.3). De plus, si l'on suppose qu'il n'y a pas d'homophilie et que les tailles des sous-ensembles sont égales, alors on obtient finalement que l'estimateur naïf coïncide avec la proportion au sein de l'échantillon, c'est-à-dire :

$$\mu_n = \frac{1}{n}\boldsymbol{n}.$$

où $\boldsymbol{n}^T = (n_1, \ldots, n_m)$. Autrement dit, on a simplement

$$\hat{\pi}_i = \frac{n_i}{n}.$$

Cet estimateur est dit naïf en raison de sa simplicité. Dans ce cadre, et dans ce qui suit, l'homophilie est la propension que l'ont les individus à avoir des liens sociaux avec des individus en certains points semblables à eux. Cette définition concerne la traduction du terme anglophone « homophily » et ne désigne pas la tendance à soutenir la cause homosexuelle. L'homophilie peut aussi désigner la propension, lors du recrutement, à plutôt choisir une recrue parmi certaine catégorie. On reviendra plus tard sur ces concepts et ils seront explicités.
Les problèmes de cet estimateur sont multiples et ils proviennent des hypothèses fortes qu'il requiert. Ainsi, on peut conclure que, bien qu'il soit intéressant, cet estimateur ne semble pas pouvoir satisfaire aux exigences de fiabilité nécessaires.
Des biais peuvent apparaître, dus à la non stationnarité des probabilités de transition, à la sélection des germes, au modèle de réciprocité des relations ou simplement dans l'estimation des probabilités de transition. Finalement, le plus gros problème réside dans le fait que ce modèle est biaisé si les homophilies ne sont pas égales entre les différents sous-ensembles (Heckathorn, 1997, 2002), ce qui est est totalement utopique (pour une définition de l'homophilie, voir annexe A) et est aussi biaisé si les tailles des sous-ensembles ne sont pas égales.
Pour conclure, on peut dire qu'il a le mérite d'être fondateur. Toutefois, il est soumis à de fortes hypothèses, qui ne sont que très rarement vérifiées dans des cas concrets d'études. Lorsqu'il est apparu, il semblait séduisant car il prétendait offrir un cadre méthodologique sérieux en vue d'étudier des populations difficiles à atteindre. Mais ce n'est pas suffisant pour le considérer comme valide. En effet, dans l'article (Heckathorn, 1997), les estimations des erreurs sont faites via une méthode bootstrap non-précisée où il n' y a ni nombre de simulations, ni explication de la procédure de génération d'un réseau social ni paramètres de la simulation. De plus, aucune simulation en violant les hypothèses n'est faite non plus. Or, puisque ces dernières sont impossible à vérifier, cela aurait mérité que l'on y porte intérêt. De fait, de récentes études ont montré la grande sensibilité de cet estimateur à des violation des hypothèses (Gile et Tomas, 2011).

2.3 Considérations mathématiques sur l'estimateur naïf

Les hypothèses sous lesquelles une distribution stationnaire existe sont complexes (voir (Karlin et Taylor, 1975)) et nécessitent une attention particulière. Si elle existe, une distribution stationnaire est unique. De plus, la solution du problème (1) est bien le vecteur de probabilités stationnaires désiré. Pour plus de précisions, voir annexe A. Dans la pratique, il est courant

qu'une chaîne de Markov soit stationnaire. On se permet de citer le théorème dont s'inspire Heckathorn en annexe A. Les hypothèses nécessaires au théorème sont en général satisfaites mais il est important de les vérifier systématiquement.
Pour pouvoir tenir compte de l'homophilie, il suppose le modèle suivant (Fararo et Skvoretz, 1984) : supposons qu'un individu recruté, appartenant au groupe S_k devienne recruteur au cours de l'enquête. Il doit alors choisir ses recrues. On suppose que le processus suivant gouverne son choix : soit il fait un choix uniformément aléatoire dans le groupe S_k (c'est-à-dire que chaque individu du groupe S_k peut-être recruté avec probabilité $\frac{1}{N_k}$), avec probabilité H_k (c'est donc le fait de faire son choix au sein de son propre groupe qui a pour probabilité H_k, et non directement le choix de sa recrue), soit il fait un choix aléatoire au sein de la totalité de la population S (donc y compris dans son propre groupe), et ce, avec probabilité $1 - H_k$. Supposons que $x \in S_k$ représente le recruteur, $y \in S_k$ un individu quelconque de son propre groupe, et $z \in S_l, \quad l \neq k$ un individu quelconque n'appartenant pas à son groupe. Alors, on a, selon ce modèle :

$$\begin{aligned} P_{kk} = P(y \in s_k | x \in s_k) &= H_k + (1 - H_k)\frac{N_k}{N}, \\ P_{kl} = P(z \in s_l | x \in s_k) &= (1 - H_k)\frac{N - N_k}{N}, \forall l \text{ t.q } l \neq k. \end{aligned} \tag{2}$$

Heckathorn démontre donc finalement que si l'homophilie, désignée ici par H, est égale parmi tous les groupes, alors son estimation est non-biaisée par l'homophilie (Heckathorn, 1997, 2002).
Parmi les hypothèses, on suppose que les liens sont bilatéraux, c'est-à-dire que si un individu x recrute un autre individu y, inversement y aurait pu recruter x (Heckathorn, 1997, 2002). Mathématiquement, cette hypothèse suppose donc que le graphe du réseau social n'est pas dirigé. Afin d'inclure une estimation de l'homophilie, il utilise le modèle présenté ci-dessus (Fararo et Skvoretz, 1984).
Soit T_{ij} le nombre de liens qui lient le groupe S_i au groupe S_j. Puisque l'on a fait l'hypothèse de réciprocité, on a bien :

$$T_{ij} = T_{ji}.$$

De plus, si l'on pose que $\pi_i = \frac{N_i}{N}$ et que P_{ij} est donné comme dans (2), alors on peut poser :

$$T_{ij} = \pi_i N_i P_{ij}.$$

On obtient alors trivialement un système d'équations linéaires :

$$\begin{cases} \sum_{i=1}^{M} \pi_i = 1 \\ \pi_j N_j P_{ji} = \pi_i N_i P_{ij} \quad \forall i < j, \quad j = 2, \ldots, M. \end{cases} \tag{3}$$

ce système est donc évidemment surdéterminé, pour $M > 2$. Heckathorn (Heckathorn, 2002) propose d'utiliser l'estimateur des moindres carrés. Toutefois, cette dernière méthode, bien que puissante, ne garantit pas que la somme des probabilités demeure égale à 1. Donc, je suppose (il n'est pas fait mention de la procédure exacte) qu'il divise les proportions estimées $\hat{\pi}_1, \ldots, \hat{\pi}_M$ par la valeur $\hat{\pi} = \sum_{i=1}^{M} \hat{\pi}_i$, de manière à normaliser les proportions. Toutefois, cette manière est peu rigoureuse car elle ne garantit, a priori, pas que le résultat demeure optimal (c'est-à-dire que la solution ainsi trouvée soit celle qui minimise effectivement la somme des carrés). En lieu et place, il faudrait résoudre le problème suivant :

$$\min_{\hat{\boldsymbol{\pi}}:\ \sum_{i=1}^{M} \hat{\pi}_i = 1} \quad \sum_{\substack{j=2 \\ i<j}}^{M} (\hat{\pi}_j N_j \widehat{P}_{ji} - \hat{\pi}_i N_i \widehat{P}_{ij})^2 \tag{4}$$

Le système ci-dessus signifie que l'on souhaite minimiser la distance (au sens euclidien du terme) entre ces valeurs, sous la contrainte que la somme des probabilités soit égale à 1. Il s'agit donc d'un problème d'optimisation sous contraintes d'égalité. Le problème (4) peut se résoudre à l'aide de la méthode des multiplicateurs de Lagrange. De plus, il est possible d'utiliser une notion de distance différente, selon l'usage désiré. Ici, on utilise implicitement la distance euclidienne, mais on pourrait utiliser n'importe quelle norme[2]. Toujours dans (Heckathorn, 2002), une autre méthode est utilisée pour résoudre le système (1). Il s'agit de lissage de données. En fait, idéalement, les différents sous-ensembles devraient recruter de manière également efficaces, c'est-à-dire qu'il devrait y avoir autant de recrues que de recrutés dans un sous-ensemble donné. Cela reflète l'hypothèse de la réciprocité des liens. Donc, au lieu de compter exactement les recrues, on utilise une version modifiée du dénombrement de personnes ayant participé à l'enquête, appelé « comptage démographiquement ajusté » . On pose $n^*_{ij} = \widehat{P}_{ij}\hat{\pi}_i n$. Ainsi, on ajuste le nombre de recrues en fonction, non pas de la proportion effective au sein de l'échantillon mais en fonction de la proportion estimée au sein de la population totale. Deuxièmement, on pose que $\tilde{n}_{ij} = \frac{n^*_{ij}+n^*_{ji}}{2} = \tilde{n}_{ji}$ et on utilise ces valeurs pour résoudre le système surdéterminé ci-dessus.
Cette démarche peut paraître quelque peu artificielle, notamment dans la mesure où l'on suppose qu'utiliser la solution du système (3) en lieu et place de la proportion dans l'échantillon réduira le bais, alors que rien ne dit qu'il ne l'augmentera pas, même si cela peut paraître étonnant. De plus, le fait d'utiliser la moyenne arithmétique est quelque peu arbitraire et purement artificiel, car on veut symétriser les données. On aurait pu utiliser une moyenne géométrique ou harmonique. Il semble donc que la solution qui utilise le lagrangien est, a priori, un peu plus convaincante. Toutefois, ce n'est pas la solution qui est privilégiée dans (Heckathorn, 2002).
L'estimateur d'Heckathorn a la propriété d'être asymptotiquement non-biaisé. Toutefois, cette affirmation est à considérer avec précaution. Le terme « asymptotiquement » signifie ici qu'il faudrait pousser le processus (nombre d'individus échantillonnés) très loin et que la population cible devrait être infinie. Pour beaucoup, cela était un gage de validité. Toutefois, la présente situation (population finie, nombre de vagues restreint) est loin d'être asymptotique. On peut ajouter que, par construction, il s'agit d'un estimateur consistant, au sens de Cochran (Cochran, 1977).

2.4 L'estimateur de Salganik-Heckathorn

L'estimateur de Salganik-Heckathorn est présenté dans un article paru en 2004 (Salganik et Heckathorn, 2004). Actuellement, on y fait référence en parlant de l'estimateur RDS I. L'idée clé de ce nouvel estimateur est non pas d'utiliser l'échantillon pour en déduire directement une estimation sur la population totale mais plutôt de l'utiliser pour faire une estimation du réseau social puis d'en déduire une estimation sur la population totale après coup. Cela permet en effet d'utiliser pleinement la méthode RDS car elle donne une information sur la structure du réseau social en plus de simplement recruter des individus de la population cible.
On fait les hypothèses suivantes : premièrement, on suppose que le réseau social forme une unique composante connexe. Deuxièmement, on suppose que le processus d'échantillonnage est avec remise, c'est-à-dire que si un individu est sélectionné dans l'enquête, il peut à nouveau l'être, ce qui n'est évidemment pas le cas. La raison d'être de cette hypothèse est qu'avec remise, les probabilités de sélection d'un individu dans l'échantillon demeurent constantes au cours du temps, ce qui n'est pas le cas en pratique. Dans ce cas, le modèle présenté précédemment, qui suppose des probabilités de transition constantes au cours du temps, n'est

2. Une norme est forcément convexe. Si de plus, elle est strictement convexe, alors on peut montrer que la solution du système lagrangien est l'unique solution du problème (4). La plupart des normes usuelles sont, de fait, strictement convexes. A noter qu'en utilisant la norme euclidienne, le système à résoudre devient linéaire.

pas valable. En toute rigueur, elles devraient se modifier au cours du temps. On reviendra plus tard sur cette hypothèse ainsi que sur ses conséquences. On suppose encore que chaque participant reçoit et distribue un unique coupon, ce qui signifie que le processus d'échantillonnage forme une chaîne sur le réseau social que constitue la population cible. On fait encore l'hypothèse que le recruteur choisit de manière uniformément aléatoire les recrues parmi ses amis. Finalement, on suppose que tous les participants sont en mesure de donner une estimation précise de leur degré, c'est-à-dire qu'ils sont capables de dire avec précision combien de personnes ils connaissent au sein de la population cible.
On présente l'estimateur développé dans (Salganik et Heckathorn, 2004) en reprenant la même notation que dans la section précédente (pour plus de détails, voir section 2.1). On définit d'abord R_i comme la somme des degrés d'un groupe S_i donné, c'est-à-dire :

$$R_i = \sum_{j \in S_i} d_j.$$

On considère les probabilités de liens inter-groupes C_{ij}, définies comme :

$$C_{ij} = \frac{T_{ij}}{R_i},$$

où T_{ij} est simplement le nombre de liens entre un individu de S_i et individu de S_j. Ainsi, C_{ij} est simplement la proportion de liens de S_i vers S_j. Finalement, on définit le degré moyen d'un groupe comme :

$$\overline{d}_i = \frac{R_i}{N_i}.$$

Dans ce qui suit, on se limite au cas où l'on ne considère que deux groupes distincts, S_1 et S_2. On reviendra plus tard au cas où l'on considère plus de deux groupes. Puisque l'on a $T_{ij} = T_{ji}$ et que l'on peut écrire $T_{ij} = N_i \overline{d}_i C_{ij}$, on a le système suivant :

$$\begin{cases} \pi_1 + \pi_2 & = & 1, \\ \pi_1 \overline{d}_1 C_{12} - \pi_2 \overline{d}_2 C_{21} & = & 0. \end{cases} \tag{5}$$

En fait, on retrouve le problème vu dans la section précédente (voir équation (4)). Tout l'intérêt réside dans la méthode utilisée pour estimer les différents termes de l'équation nécessaires à l'estimation de π_1 et π_2.
En ce qui concerne l'estimation de C_{ij}, on utilise simplement l'estimation naïve, à savoir :

$$\widehat{C}_{ij} = \frac{r_{ij}}{\sum_{k=1}^{M} r_{ik}},$$

où r_{ij} désigne le nombre de recrutements par un individu de S_i d'un individu S_j dans l'échantillon. Il s'agit de l'équivalent de la quantité R_{ij} au sein de l'échantillon.
Quant à l'estimation du degré moyen par groupe, deux méthodes, menant au même estimateur sont proposées (Salganik et Heckathorn, 2004). Avant tout, rappelons que parmi les hypothèses que l'on a faites, on suppose que chaque recrue est à même de donner une estimation précise de son degré, c'est-à-dire du nombre de personnes appartenant à la population cible dans son entourage. Estimer le degré moyen d'un groupe en utilisant la moyenne empirique des degrés semble être une mauvaise idée. En effet, la méthode RDS a tendance à suréchantillonner des individus qui sont très sociaux, donc cet estimateur a tendance à surestimer le degré moyen et n'est donc pas adéquat. On doit développer un autre estimateur. Ce dernier peut-être construit de deux manière, la première en utilisant la distribution empirique du nombre de degrés, la seconde en utilisant un estimateur de type Hansen-Hurvitz. On ne présente pas

ici le détail des calculs menant à ces estimateurs et on se contente de donner le résultat. On estime donc $\overline{d}_i$ par :

$$\widehat{\overline{d}}_i = \frac{n_i}{\sum_{j=1}^{n_i} \frac{1}{d_j}}, \tag{6}$$

qui est la moyenne harmonique des degrés au sein du sous-ensemble considéré. Finalement, on peut résoudre facilement le système (5), en remplaçant les différentes valeurs inconnues par leurs estimations. Cet estimateur présente un certain nombre de problèmes. Outre le fait que les hypothèses sont difficiles à vérifier, elles sont rarement satisfaites dans la pratique. De plus, dans le cas où l'on considère plus de deux sous-ensembles, le système (5) devient :

$$\left\{ \begin{array}{lcl} \sum_{i=1}^{M} \pi_i & = & 1 \\ \pi_i \overline{d}_i C_{ij} - \pi_j \overline{d}_j C_{ji} & = & 0 \end{array} \right. , \quad i, j = 1, \ldots, M, \quad i \neq j.$$

qui est un système surdéterminé, si $M > 2$, c'est-à-dire s'il y a plus de deux sous-ensembles. On présente ici le cas où il y a deux sous-ensembles, et on peut se référer à la section 2.5 pour le cas où $M > 2$. Si $M = 2$, on a alors :

$$\mu_{\text{S-H}} = \hat{\pi}_1 = \frac{\widehat{\overline{d}}_2 \widehat{C}_{21}}{\widehat{\overline{d}}_1 \widehat{C}_{12} + \widehat{\overline{d}}_2 \widehat{C}_{21}} = \frac{\widehat{C}_{21}}{\frac{\widehat{\overline{d}}_1}{\widehat{\overline{d}}_2} \widehat{C}_{12} + \widehat{C}_{21}},$$

où $\mu_{\text{S-H}}$ désigne l'estimateur de Salganik-Heckathorn.
Les hypothèses sont, là aussi, très fortes. Il semble irréaliste qu'elles soient satisfaites. Une note en bas de page précise que le fait de considérer le tirage avec remise n'a qu'une influence négligeable (Salganik et Heckathorn, 2004). Toutefois, d'autres études postérieures sont venues contredire cette affirmation (Gile et Handcock, 2010). Deuxièmement, jamais une étude n'a été faite en ne distribuant qu'un seul coupon par personne, cela rendrait l'enquête trop lente. A nouveau, cette hypothèse n'est pas vérifiée. Finalement, il n'existe pas d'estimateur analytique de la variance et les estimations en utilisant une méthode bootstrap sont les seules possibles.

2.5 L'estimateur de Salganik-Heckathorn, côté maths

Tout d'abord, on introduit quelques notations :

$$NI(j)_{w=k} = \left\{ \begin{array}{ll} 1 & \text{si le nœud } j \text{ est sélectionné à la vague } k, \\ 0 & \text{sinon,} \end{array} \right.$$

et

$$EI(e_{j \to l})_{r=k} = \left\{ \begin{array}{ll} 1 & \text{si l'arête reliant } j \text{ et } l \text{ est sélectionnée à la vague } k, \\ 0 & \text{sinon,} \end{array} \right.$$

où l'on dit par convention qu'une arête $e_{j \to l}$ est sélectionnée pendant la vague k si l'extrémité de l'arête, le nœud l, appartient à la vague k. On remarque que l'on considère séparément les deux caractéristiques du réseau, nœuds et arêtes. Supposons de plus que l'on sélectionne les germes avec une probabilité proportionnelle au degré des individus de la population cible, c'est-à-dire que l'on ait :

$$P(NI(j)_{w=0} = 1) = \alpha \, d_j, \quad \forall j \in S, \text{pour } \alpha \in \mathbb{R}.$$

Tout d'abord, remarquons que, puisque l'on a $\sum_{j \in S} P(NI(j)_{w=0} = 1) = \alpha \sum_{j \in S} d_j = 1$, on a alors $\alpha = 1 / \sum_{i \in S} d_i$. Donc,

$$P(NI(j)_{w=0})) = \frac{d_j}{\sum_{i \in S} d_i}, \quad \forall j \in S.$$

On rappelle que l'on suppose que la recrue choisit de manière uniformément aléatoire parmi les connaissances qu'il a au sein de la population cible. On peut donc écrire, grâce à la formule de Bayes :

$$\begin{aligned} P(EI(e_{j\to l})_{r=1}=1) &= P\left(NI(j)_{w=0}=1\right) P\left(EI(e_{j\to l})_{r=1}=1 | NI(j)_{w=0}=1\right) \\ &= \frac{d_j}{\sum_{i\in S} d_i}\frac{1}{d_j} = \frac{1}{\sum_{i\in S} d_i}. \end{aligned}$$

Cela montre qu'en sélectionnant les germes avec une probabilité proportionnelle à leur degré, la sélection des arêtes est, elle, uniformément aléatoire[3]. Finalement, on peut calculer la probabilité qu'un individu soit sélectionné lors de la première vague (sachant que les germes constituent la vague 0) :

$$P\left(NI(j)_{w=1}=1\right) = d_j P\left(EI(e_{l\to j})_{r=1}=1\right) = \frac{d_j}{\sum_{i\in S} d_i} = P\left(NI(j)_{w=0}=1\right).$$

Ainsi, on voit que la probabilité de sélection demeure stable au cours du processus. Plus largement, on peut facilement démontrer l'assertion suivante :

Proposition 1
Supposons un processus d'échantillonnage RDS satisfaisant les hypothèses décrites dans la section 2.4. En particulier, on suppose que la probabilité de sélection d'un germe est proportionnelle à son degré et que le processus de sélection est avec remise. Alors, la probabilité de sélection ne change pas au cours du processus et est aussi proportionnelle au degré de l'individu.

Démonstration. On procède par récurrence sur k, le nombre de vague. On a déjà montré la pas de récurrence. En supposant que le résultat est vrai pour $k \in \mathbb{N}$, on montre de la même manière que le résultat demeure vrai pour $k+1$. □

Ainsi la probabilité qu'un individu soit choisi demeure stable au cours du temps. On remarque aussi que l'on a pas exclu du choix l'arête d'où l'on provient. Ceci illustre le fait que l'on considère un tirage avec remise.
En ce qui concerne l'estimateur du nombre de degrés, (voir équation (6)), on reconnaît ici la moyenne harmonique des degrés. Il n'est fait aucune mention de la pertinence de l'utilisation d'un tel estimateur. En règle générale, la moyenne harmonique s'utilise dans des contextes relativement particuliers et il serait intéressant de s'attarder sur la nature de l'utilisation de cette moyenne. Finalement, dans le cas où l'on a plus de deux sous-ensembles, on doit résoudre le système :

$$\begin{cases} \sum_{i=1}^{M} \pi_i &= 1 \\ \pi_i \overline{d}_i C_{ij} - \pi_j \overline{d}_j C_{ji} &= 0 \end{cases}, \quad i,j = 1,\dots,M, \quad i \neq j,$$

qui est surdéterminé, puisque $M > 2$. De manière similaire à la méthode présentée dans la section (2.2), on peut alors étudier la solution du problème suivant :

$$\min_{\hat{\boldsymbol{\pi}}:\ \sum_{i=1}^{M}\hat{\pi}_i=1} \quad \sum_{\substack{j=2 \\ i<j}}^{M} g\left(\hat{\pi}_i \widehat{\overline{d}}_i \widehat{C_{ij}} - \hat{\pi}_j \widehat{\overline{d}}_j \widehat{C}_{ji}\right), \tag{7}$$

3. Ici, on considère la probabilité $P(EI(e_{j\to l})_{r=1}=1)$. Il ne s'agit pas exactement de la probabilité que l'arête $e_{j\to l}$ fasse partie du réseau échantillonné car il faudrait aussi compter la probabilité que l'arête $e_{l\to j}$ soit échantillonnée, ce qui ferait simplement doubler la probabilité.

où $g(.)$ est une norme sur $\mathbb{R}$. On peut résoudre ce problème en utilisant la méthodes des multiplicateurs de Lagrange. Comme vu précédemment, ce problème admet une unique solution si la fonction $g(.)$ est strictement convexe.
Cet estimateur est, comme son prédécesseur, asymptotiquement non biaisé, sous les hypothèses très forte précisées précédemment. Il ne s'agit cependant pas d'une garantie de précision suffisante puisque les conditions asymptotiques ne sont jamais remplies. On a aussi montré que l'estimateur de Salganik-Heckathorn pouvait ne pas être consistant (voir chapitre 3).

2.6 L'estimateur d'Heckathorn

L'estimateur d'Heckathorn est une version légèrement modifiée de l'estimateur de Salganik-Heckathorn développée dans (Heckathorn, 2007). Actuellement, on y fait référence en parlant de l'estimateur RDS DS. Bien qu'en apparence très semblable, la différence se situe dans l'estimation du degré moyen par sous-ensemble. Le but est de pouvoir réduire le biais induit par l'activité différenciée (le fait que certains sous-ensembles ont plus de liens relativement à d'autres sous-ensembles) et le recrutement différencié (qui est la propension d'un groupe à recruter dans un groupe donné). Dans ce qui suit, on ne va présenter cet estimateur que brièvement. Toutefois, la modélisation est fondamentalement différente et l'estimateur d'Heckathorn est très utilisé, notamment en raison du fait qu'il est désormais implémenté dans le logiciel d'analyse RDS développé par Heckathorn et ses collègues (Volz *et al.*, 2007). Puisque l'estimateur d'Heckathorn est le même que l'estimateur de Salganik-Heckathorn à la différence près qu'il estime d'une manière différente le degré moyen par groupe, on ne présente que l'estimateur du degré moyen par sous-ensemble. On le note $\overline{ad}_i$, pour un sous-ensemble i donné (ad pour *adjusted degree*). On pose :

$$\overline{ad}_i = \frac{\sum_{j \in s_i} \text{RCD}_j}{\sum_{j \in s_i} \frac{1}{d_j} \text{RCD}_j},$$

où RCD_j est la composante du dégré liée au recrutement, *recruitment component of degree*, de l'individu j. On définit alors RCD_j comme :

$$\text{RCD}_j = \frac{\widehat{E}_i}{\frac{n_i}{n}}, \quad \forall j \in s_i,$$

où $\widehat{E}_i$ est la distribution stationnaire de la chaîne de Markov où l'on considère que les états sont les différents sous-ensembles et que les probabilités de transition sont ajustées en utilisant le lissage de données (voir section 2.3). Finalement $\widehat{\boldsymbol{E}}$ est la solution du problème aux valeurs propres :

$$\widehat{\boldsymbol{E}} = \widehat{\boldsymbol{E}} \widehat{P}^*,$$

où $\widehat{P}^*$ est la matrice des probabilités de transition ajustée (pour le détail, voir (Heckathorn, 2007, p.171-175) et 2.3) et $\widehat{E}_i$ est la $i^{\text{ème}}$ composante du vecteur $\widehat{\boldsymbol{E}}$. Pour plus de simplicité, on considère uniquement le cas où il y a deux sous-ensembles, indexés par 1 et 2. Alors, l'estimateur d'Heckathorn est donné par :

$$\mu_{\text{H}} = \hat{\pi}_1 = \frac{\overline{ad}_2 \widehat{C}_{21}}{\overline{ad}_1 \widehat{C}_{12} + \overline{ad}_2 \widehat{C}_{21}}.$$

A l'instar des autres estimateurs couramment utilisés, celui-ci est tout autant problématique. En effet, il requiert des hypothèses très fortes et rencontre les mêmes difficultés théoriques que l'estimateur de Volz-Heckathorn. Ces obstacles théoriques sont largement commentés dans la section qui suit.

2.7 L'estimateur de Volz-Heckathorn

L'estimateur de Volz-Heckathorn est présenté dans (Volz et Heckathorn, 2008). Actuellement, on y fait référence en parlant de l'estimateur RDS II. Cet estimateur a pour but de montrer que l'on peux estimer les probabilités de sélection des individus échantillonnés grâce à une méthode RDS. Ainsi, l'avancée qui est proposée est très importante. On essaie de montrer que l'on peut considérer l'échantillon engendré par un processus RDS comme un échantillon probabiliste. Ce changement de perspective est radical. Aussi, on prête une attention particulière aux arguments avancés.

On fait les hypothèses suivantes :

Echantillonnage avec remise On considère que l'échantillonnage est avec remise, c'est-à-dire qu'un individu déjà échantillonné peut l'être à nouveau. Sans cette hypothèse, on est obligé de considérer que les probabilités d'inclusion changent au cours du temps, ce qui rend l'estimation difficile. Dans les faits, cette hypothèse n'est jamais vérifiée puisque l'on empêche les participants de participer plus d'une fois à l'enquête. Comme il y a, en général, une compensation financière pour participer à l'enquête, certains individus sont tentés d'en faire une ressource financière.

Un seul coupon par recrue On considère que le processus d'échantillonnage est une chaîne, c'est-à-dire que chaque recrue ne donne qu'un seul coupon. Cette hypothèse n'est pas respectée en général car en donnant plusieurs coupons, on diminue exponentiellement le nombre de vagues nécessaires à atteindre une taille donnée. Par exemple, pour un échantillon de 500 individus, il faut 500 vagues avec un coupon par recrue, 9 vagues pour 2 coupons, 6 avec 3 coupons. Il est donc très utile de multiplier le nombre de coupons et cela se fait toujours en pratique.

Petite fraction de l'échantillon On suppose qu'une faible fraction de la population totale est inclue dans l'échantillon, de manière à ce que l'hypothèse de remise n'ait qu'un faible impact. Cette hypothèse découle directement de l'hypothèse d'un échantillonnage avec remise et la remplace, puisque qu'elle n'est jamais satisfaite.

Estimation précise du nombre de degré On suppose que chaque personne participant à l'enquête est capable de donner une estimation précise de son degré, c'est-à-dire du nombre de personne parmi ses connaissances appartenant à la population cible. Cette hypothèse est utile car elle permet d'avoir une meilleure connaissance du réseau. Dans le cas présent, on constate qu'elle est même essentielle au calcul de l'estimateur.

Choix aléatoire de la recrue On suppose que parmi toutes ses connaissances, la recrue choisit de manière uniformément aléatoire à qui elle va donner son coupon. Cette hypothèse n'est certes pas respectée, certains individus ayant plus de chance d'appartenir, *in fine*, à l'échantillon, en particulier ceux qui ont beaucoup de connaissances au sein de la population cible. Toutefois, certain auteurs, (Heckathorn, 2002), soutiennent que cette hypothèse est souvent satisfaite. Lorsque le recrutement n'est pas uniformément aléatoire, on parle de *recrutement différencié*. Une étude plus poussée à ce sujet montre que certains estimateurs sont sensibles à ce type de biais (Gile et Tomas, 2011). Ainsi, on peut d'emblée dire qu'il ne s'agit pas d'une hypothèse anodine.

Réciprocité Afin de simplifier la représentation du graphe du réseau social, on suppose que les liens sont bilatéraux et de même poids. Cette hypothèse est en fait double. Premièrement, on suppose que le graphe qui représente le réseau social n'est pas pondéré (*unweighted graph*, en anglais), c'est-à-dire que toutes les arêtes ont le même poids. A contrario, on pourrait imaginer qu'un réseau social soit plutôt pondéré. De plus, on suppose que les liens sont réciproques, ce qui signifie que les liens sociaux qui unissent un recruteur et sa recrue sont réciproques et que la recrue aurait pu recruter son recruteur.

Hypothèses sur la chaîne de Markov Pour pouvoir modéliser le processus de recrutement comme

une chaîne de Markov, et ensuite pouvoir conclure qu'il existe une distribution stationnaire, on suppose que le réseau social possède certaines propriétés. La première est le fait que le réseau est *irréductible*, c'est-à-dire que pour tout individu appartenant à la population cible, il existe une probabilité non nulle qu'il appartienne à l'échantillon, quelque soient les germes. Deuxièmement, on suppose que la chaîne est *apériodique* et *positive et récurrente*. Ces notions sont liées aux chaînes de Markov et ce n'est pas la vocation de ce travail que de traiter ce sujet en détail. On peut se référer à (Karlin et Taylor, 1975). Ces hypothèses permettent d'assurer qu'au bout d'un certain nombre de vagues, l'échantillon sera indépendant du choix des germes.

La différence principale entre cet estimateur et les estimateurs précédents est que l'on considère une chaîne de Markov dont l'espace des états n'est pas constitué simplement des différents sous-ensembles de la population cible mais de tous les individus.
Si l'on pose $\hat{d}_j$ comme le degré approximatif de l'individu j de l'échantillon, l'estimateur de π_i, la vraie proportion d'individus appartenant au sous-ensemble i de la population S est donné par :

$$\mu_{\text{V-H}} = \pi_i = \frac{\sum_{j \in s_i} \frac{1}{\hat{d}_j}}{\sum_{k \in s} \frac{1}{\hat{d}_k}},$$

où $\mu_{\text{V-H}}$ désigne l'estimateur de Volz-Heckathorn. Cet estimateur est asymptotiquement non biaisé, sous les hypothèses faites ci-dessus.

2.8 L'estimateur de Volz-Heckathorn, côté maths

Sous l'hypothèse que les germes sont choisis avec une probabilité proportionnelle à leur degré, on peut démontrer que la probabilité d'appartenir à l'échantillon est alors aussi proportionnelle au degré (voir 2.5 et (Salganik et Heckathorn, 2004)). Ainsi, on peut supposer que la matrice P des probabilités de transition est donnée par :

$$(P)_{ij} = \begin{cases} \frac{1}{d_i} & \text{si } i \text{ et } j \text{ sont voisins,} \\ 0 & \text{sinon.} \end{cases}$$

Finalement, on obtient la distribution stationnaire :

$$x_i = \frac{d_i}{\sum_{j \in S} d_j},$$

où x_i représente la distribution stationnaire la chaîne de Markov dont P est la matrice des probabilités de transition. La forme de cette distribution stationnaire est donc soumise à l'hypothèse que l'on a faite sur les probabilités de transition.
Si l'on utilise la condition de stationnarité ci-dessus, et en supposant que la probabilité de sélection p_i dans l'échantillon d'un individu i est proportionnelle au degré, on a :

$$p_i = \frac{d_i}{N\bar{d}_S},$$

où $\bar{d}_S$ et N désignent respectivement le degré moyen de la population et la taille de la population cible. On peut alors estimer cette quantité par :

$$\hat{p}_i = \frac{d_i}{N\widehat{\bar{d}}_S},$$

bien que N soit en général inconnu. Par ailleurs, de la même manière que dans (Salganik et Heckathorn, 2004), on estime $\overline{d}_S$ en utilisant l'estimateur de la moyenne harmonique :

$$\widehat{\overline{d}}_S = \frac{n}{\sum_{j \in s} \frac{1}{d_j}}. \tag{8}$$

Supposons que l'on s'intéresse au total sur l'échantillon d'une variable y. En utilisant l'estimateur de Hansen-Hurwitz [4](Hansen et Hurwitz, 1943) pour déterminer la taille, on trouve alors comme estimation du total sur la population de la variable y :

$$\widehat{T}_y = \frac{1}{n} \sum_{j \in s} \frac{y_j}{\hat{p}_j} = \frac{1}{n} \sum_{j \in s} \frac{\widehat{\overline{d}}_S N y_j}{d_j} = \frac{\widehat{\overline{d}}_S N}{n} \sum_{j \in s} \frac{y_j}{d_j}.$$

Puisque l'on s'intéresse à la moyenne de la valeur de la variable y, on a alors :

$$\widehat{\overline{y}} = \frac{1}{N} \widehat{T}_y = \frac{\widehat{\overline{d}}_S}{n} \sum_{j \in s} \frac{y_j}{d_j}.$$

En remplaçant $\widehat{\overline{d}}_S$ par sa valeur estimée en (8), on obtient finalement :

$$\hat{\overline{y}} = \frac{\sum_{j \in s} \frac{y_j}{d_j}}{\sum_{j \in s} \frac{1}{d_j}}.$$

Si l'on souhaite estimer la proportion du sous-ensemble k dans la population cible, il suffit de poser que y_i est la variable qui indique l'appartenance de l'individu i au sous-ensemble k. L'estimateur de Volz-Heckathorn s'écrit alors :

$$\mu_{\text{V-H}} = \hat{\pi}_k = \frac{\sum_{j \in s_k} \frac{1}{d_j}}{\sum_{j \in s} \frac{1}{d_j}}.$$

Cet estimateur peut être vu comme un estimateur de Hansen-Hurwitz généralisé de la quantité (Gile et Tomas, 2011) :

$$\frac{\sum_{j \in S_k} \frac{p_j}{d_j}}{\sum_{i \in S} \frac{p_i}{d_i}}.$$

Cet estimateur est asymptotiquement non-biaisé si les probabilités de sélections sont proportionnelles au degré.
Puisque l'on dispose d'estimations des probabilités de sélection, il est alors imaginable d'avoir une estimation de la variance de l'estimateur, afin de quantifier l'incertitude liée à $\mu_{\text{V-H}}$. Le plus évident est d'utiliser l'estimateur de la variance de l'estimateur de Hansen-Hurwitz. Toutefois, celui-ci est dérivé en faisant l'hypothèse que chaque individu de l'échantillon est sélectionné de manière indépendante, ce qui est loin d'être le cas d'un processus RDS. On ne présente que le résultat, pour l'estimateur de π_k, la proportion d'individus appartenant au sous-ensemble k de la population cible :

$$\widehat{\text{var}}(\hat{\pi}_k) = \widehat{V}_1 + \frac{\hat{\pi}_k^2}{n} \left((1-n) + \frac{2}{n_k} \sum_{i=2}^{n} \sum_{j=1}^{i-1} \widehat{P}_{kk}^{i-j} \right), \tag{9}$$

4. On fait implicitement l'hypothèse d'un échantillonnage avec remise.

où

$$\widehat{V}_1 = \frac{\widehat{V}(Z_i)}{n} = \frac{1}{n(n-1)} \sum_{j \in s} (Z_i - \hat{\pi}_k)^2,$$

et

$$Z_i = \frac{\hat{d}_S}{d_i} I_k(i),$$

où $I_k(i)$ est la fonction indicatrice du sous-ensemble k.
A l'aide de quelques manipulations algébriques, on peut écrire :

$$\sum_{i=2}^{n} \sum_{j=1}^{i-1} \widehat{P}_{kk}^{i-j} = \sum_{l=1}^{(n-1)} (n-l) P_{kk}^{l} = \sum_{m=1}^{n-1} m P_{kk}^{n-m} = P^n \sum_{m=1}^{n-1} m \left(\frac{1}{P}\right)^m.$$

Or, on a :

$$\sum_{n=1}^{N} n x^n = (N-1)\frac{x^N}{1-x} + \frac{1-x^{N+1}}{(1-x)^2} - \frac{1}{1-x}.$$

Il encore possible de simplifier l'équation (9) en utilisant cette dernière expression. Toutefois, lors des simulations, il est apparu que cet estimateur de variance est en général trop petit. Il est même régulièrement négatif, ce qui est tout de même problématique. Dans les faits, le processus d'échantillonnage est donc modélisé comme une marche aléatoire, à l'équilibre, c'est-à-dire dont la distribution est stationnaire. On suppose aussi que le processus est sans branchements, c'est-à-dire qu'il ne passe que d'un individu à un autre, et que l'on exclut le cas où plusieurs coupons sont distribués par un même recruteur.

2.9 Hypothèses mathématiques des estimateurs classiques

On peut voir que les hypothèses diffèrent naturellement pour chacun des estimateurs mais qu'un grand nombre d'entre elles se recoupent. De plus, l'estimateur de Volz-Heckathorn est le plus couramment utilisé. C'est pourquoi on se contente d'évoquer ici les hypothèses de ce dernier. Plusieurs articles se sont attardés à discuter ces différentes hypothèses (Gile et Handcock, 2010; Gile et Tomas, 2011; Lu *et al.*, 2012). Deux problèmes distincts se posent. Le premier est que les hypothèses ne sont que rarement satisfaites et il est donc nécessaire de pouvoir évaluer comment réagissent les estimateurs à la violation des hypothèses sur lesquelles ils sont basés. Deuxièmement, les hypothèses faites dans (Volz et Heckathorn, 2008) ne sont pas suffisantes pour assurer que l'estimateur est non-biaisé. On peut montrer (Gile et Handcock, 2010; Gile et Tomas, 2011; Lu *et al.*, 2012) que dans des cas souvent rencontrés dans la pratique, l'estimateur est fortement biaisé, notamment par l'activité différenciée ou par la non-réponse. Par ailleurs, on considère que les estimateurs qui ont précédé l'estimateur de Volz-Heckathorn sont désuets. En effet, il est montré que dans la très grande majorité des cas, l'estimateur $\mu_{\text{V-H}}$ supplante l'estimateur $\mu_{\text{S-H}}$, au sens du critère MSE (pour Mean Square Error)(Gile et Handcock, 2010). C'est pourquoi, on se concentre sur ces hypothèses qui sont en grande partie celles de l'estimateur de Volz-Heckathorn :

connexité On suppose que le réseau social dont est constitué la population cible est connexe, ou au moins que tous les germes appartiennent à la grande composante connexe (*giant component*, voir (Newman, 2003)). Au delà de cette simple hypothèse, on suppose que le réseau satisfait aux conditions nécessaires pour que si l'on considère le processus d'échantillonnage comme une chaîne de Markov, celle-ci admette une distribution stationnaire (voir annexe A).

échantillonnage avec remise On suppose que l'échantillonnage est fait avec remise. Cette hypothèse en entraîne une autre, celle que la fraction échantillonnée soit faible, afin que la différence entre un échantillonnage avec ou sans remise soit négligeable. Cette dernière est, elle, nécessaire pour utiliser des estimateurs de Hansen-Hurwitz.

un seul coupon par recrue On suppose qu'un seul coupon est donné à chaque recrue et que cette dernière l'utilise. A nouveau, cette hypothèse est double puisqu'elle exclut implicitement la non-réponse. Le fait que le coupon soit unique est nécessaire afin de modéliser l'échantillonnage comme une marche aléatoire et non pas comme un processus de branchement.

choix aléatoire de la recrue On suppose qu'au moment du recrutement, le choix parmi les différents individus connectés est uniformément aléatoire. Cette hypothèse est nécessaire pour pouvoir estimer les probabilités d'inclusion et de rendre « probabiliste » la méthode RDS.

choix aléatoire des germes On suppose que les germes sont choisis de manière aléatoire parmi la population cible, avec une probabilité proportionnelle au degré. Cette hypothèse permet de supposer que le choix des germes est une réalisation de la distribution stationnaire de la chaîne de Markov.

réciprocité des liens On suppose que les liens sont réciproques, c'est-à-dire qu'une recrue aurait aussi pu recruter son recruteur. Cette hypothèse est double car on suppose aussi que les liens ne sont pas pondérés (certains liens pourraient avoir plus de poids que d'autres). En terme de théorie des graphes, on suppose donc que le réseau est à la fois non pondéré et non orienté (*unweighted* et *undirected*, en anglais). Cette hypothèse simplificatrice rend possible une estimation.

estimation précise du nombre de degré Chaque individu appartenant à l'échantillon est supposé être en mesure de donner une estimation précise de son degré, c'est-à-dire du nombre de personne qu'il connaît au sein de la population cible. Pour le calcul de certains estimateurs, cette information est nécessaire. De manière heuristique, elle permet d'obtenir plus d'informations sur le réseau sans que cela ne demande de ressources supplémentaires.

Ces conditions sont nécessaires pour utiliser les modles développés dans ce qui précède.
De plus, d'autres aspects peuvent influencer les estimateurs. Ces aspects ne sont pas d'emble abordés comme des hypothèses nécessaires (Volz et Heckathorn, 2008) mais sont tout de même à considérer (Gile et Handcock, 2010; Gile et Tomas, 2011). On abordera ici essentiellement ce qui a trait à la structure du réseau, au taux de sondage ainsi qu'à la non-réponse.

La dépendance aux germes Une des principales propriétés que l'on prête à un échantillon engendré par un processus RDS est qu'au bout d'un certain nombre de vagues, la dépendance aux germes doit disparaître. Cette hypothèse est soutenue par le fait que l'on modélise le processus de recrutement comme une chaîne de Markov. Toutefois, dans les faits, les hypothèses sur lesquelles se fondent ce processus ne sont pas scrupuleusement satisfaites, puisque le tirage est sans remise et que le processus d'échantillonnage n'est pas une marche aléatoire sur le réseau. De fait, la dépendance aux germes ne disparaît pas complètement et il faut un certain nombre de vagues pour que le biais soit considérablement amenuisé (8 semble être un nombre de vagues suffisant (Gile et Handcock, 2010)), surtout si les germes sont choisis de manière peu aléatoire. On trouve une illustration de ce phénomène dans (Gile et Handcock, 2010). On voit que dans certains cas, le biais introduit est important. En dépit des remarques rassurantes à ce sujet dans (Lu *et al.*, 2012), il semble qu'il s'agisse là d'un point à traiter avec précaution. Toutefois, les simulations mettant en évidence un tel biais dans (Gile et Handcock, 2010) sont extrêmes et n'ont qu'une faible probabilité de survenir. Mais il est clair que ce phénomène, bien que, rare a priori, reste envisageable. Dans tous les cas, il semble que les études RDS menées ont tendance à ne pas utiliser assez de vagues.
On peut imaginer que pour diminuer le biais lié à la sélection non-probabiliste des germes, on utilise pour l'estimation uniquement les individus à partir de la deuxième ou la troisième

vague. Cette idée est communément appliquée dans les procédures MCMC, où l'on fait un *burn in*, c'est-à-dire que l'on lance une chaîne de Markov et que, pour estimer sa distribution, on n'utilise que les valeurs simulées au-delà d'un certain seuil. On peut imaginer appliquer idée similaire. Toutefois, les simulations montrent que l'option optimale semble être de ne supprimer que les germes de l'échantillon et de la première vague. Si l'on enlève les vagues suivantes, dans certaines situations, un biais apparaît, parfois relativement important (Gile et Handcock, 2010). Pour comprendre ce phénomène contre-intuitif, il faut se rappeler que le processus est sans remise. Pour illustrer ceci, on peut considérer la situation suivante : supposons que l'on considère une variable indicatrice (infecté ou pas) et que l'on ait 20 germes, tous infectés. La population totale est constituée de 1000 individus, dont 200 infectés et on suppose que l'homophilie de processus est très forte, c'est-à-dire qu'il y a un recrutement différencié. Alors, les premières vagues, qui comprennent 100 individus disons, sont très majoritairement constituées d'individus infectés. Alors quelle que soit l'estimation, la probabilité d'inclusion est au moins diminuée de moitié pour des individus infectés. Ainsi, l'estimation s'en voit très biaisée. Le fait que la population est finie et que le tirage est sans remise rend possible ce type de biais (Gile et Handcock, 2010). Il est donc préconisé de supprimer les germes et la première vague de l'échantillon en vue de l'estimation (Gile et Handcock, 2010).

L'homophilie dans le processus de recrutement Un autre grand problème de la méthode RDS est le processus de recrutement. Sa force est d'utiliser les réseaux sociaux pour échantillonner des populations difficiles à atteindre. Toutefois, c'est aussi une grande faiblesse. L'enquêteur n'a aucune prise sur la manière dont est mené le recrutement par les individus. Plusieurs biais peuvent alors apparaître dont l'homophilie dite de processus, qui est la propension à recruter un individu parmi les gens de son propre milieu. Cette problématique est plus complexe qu'il n'y paraît. Dans le cadre de l'estimation d'une proportion au sein d'une population difficile à atteindre, le cas le plus évident est lorsque l'homophilie est corrélée avec la variable observée (exemple : on cherche à estimer la proportion de personnes de couleur noir au sein la communauté homosexuelle et on suppose qu'il y a de l'homophilie entre les différentes catégories ethniques). Toutefois, cela peut-être plus complexe lorsque l'homophilie a plutôt lieu dans un domaine différent que celui que l'on vise (exemple : on reprend l'exemple précédent en considérant que l'homophilie est liée au sexe, c'est-à-dire femme-femme, homme-homme). Finalement, on peut aussi manquer un facteur d'homophilie. Par exemple, lorsque qu'une partition au sein de la population cible n'est pas remarquée par les enquêteurs (Heimer, 2005). Par exemple, prenons le cas d'une étude sur le milieu gay latino de Los Angeles (Ramirez-Valles *et al.*, 2005). Les enquêteurs ont pu manquer le fait que des personnes nées sur sol américain et des personnes nées en Amérique du sud avaient tendance à se comporter différemment, y compris dans le recrutement. A contrario, on pourrait imaginer que les enquêteurs supposent de l'homophilie où il n'y a pas de raison d'en avoir. Par exemple, lorsque l'on suppose que les personnes séropositives recrutent plus au sein de leur milieu. Ce n'est pas évident, puisque le fait d'être séropositif ne l'est pas non plus de prime abord, que de l'homophilie existe dans un tel milieu (Heimer, 2005). De manière générale, on peut encore considérer la situation où les individus issus d'un sous-ensemble donné ont tendance à privilégier le recrutement parmi un sous-ensemble précis (celui auquel ils appartienne ou, justement, un autre). C'est une généralisation de la notion d'homophilie de processus. On parle alors de recrutement différencié, ce qui comprend donc l'homophilie. Cette problématique est donc très complexe et échappe à une modélisation mathématique simple. Il est important de bien étudier de manière qualitative la population cible avant le début d'une enquête RDS (Salganik, 2012). Dans tous les cas, même sous sa forme la plus simple, l'homophilie biaise tout de même les estimations.

L'échantillonnage avec remise Une des hypothèses les plus fortes semble être le fait que l'on considère le tirage avec remise. Cette hypothèse est faite car afin de contourner le fait que les probabilités d'inclusion changent au cours du processus. De manière plus fondamentale, cela permet de modéliser le processus d'échantillonnage comme un processus de Markov homogène en temps. Si ce n'était pas le cas, la modélisation s'en verrait terriblement complexifiée. Par ailleurs, elle implique directement l'hypothèse que la fraction échantillonnée doit être relativement faible, puisque, dans ce cas, la différence entre un échantillonnage avec ou sans remise est négligeable. Sachant que toute enquête RDS la viole sciemment, il est important de pouvoir comprendre à quel point les estimations sont sensibles à une telle violation. Il y a une apparente contradiction entre deux hypothèses : celle d'un échantillonnage avec remise, qui implique qu'une petite fraction de la population cible est échantillonnée, et celle de la nécessité d'un grand nombre de vagues pour faire disparaître la dépendance de l'échantillon aux germes. Une des manière de régler ce problème est de limiter le nombre de coupons, mais cela ralonge le temps nécessaire à l'enquête. Le fait de permettre une remise implique une vitesse de convergence vers la distribution stationnaire beaucoup plus lente, puisque l'individu peut choisir celui qui l'a choisi, par exemple, ou encore lui-même. De plus, puisque le nombre de tirages possibles est beaucoup plus important, N^n au lieu de $\binom{N}{n}$, la variance s'en voit accrue. D'après les simulations, le fait d'ajouter une remise peut parfois péjorer l'estimation, ce qui est étonnant mais pas inexplicable. En effet, ne pas faire de remise accélère le mélange des sous-ensembles au sein de l'échantillon, d'où un biais important lorsque les germes ne sont pas choisis de manière aléatoire (Gile et Handcock, 2010). Par contre, la sensibilité de l'estimateur à la fraction de l'échantillon est plus problématique. Lorsque l'on fait varier l'activité relative parmi les sous-ensembles, on remarque que l'estimation est problématique dans certaines situations (Gile et Handcock, 2010).

La connexité La connexité du réseau peut parfois être une hypothèse difficile à vérifier. Il est donc important de s'assurer qualitativement de l'existence d'une forte probabilité de connexité du réseau qui constitue la population cible. Si ce n'est pas le cas, il est nécessaire d'utiliser la méthode RDS séparément sur les différentes composantes connexes du réseau. On peut rencontrer le problème que l'estimation finale perde tout sens. En effet, une étude RDS n'a pour but que d'estimer des populations au sein d'un même réseau social. Si l'on explore deux réseaux sociaux distincts, donc pas connexes, il est possible que ceux-ci ne soient pas comparables, avec à la clé des résultats dénués de sens. On peut illustrer cette situation avec l'exemple suivant : supposons que l'on cherche à étudier les personnes toxicomanes séropositives en Suisse et que la ville de Zürich ait une politique très active dans ce domaine et que des campagnes de prévention aient déjà eu lieu, avec un grand succès. La ville de Genève aurait quant elle une politique très répressive, sans aucune prévention. On pourrait supposer que dans une telle situation, 25% des personnes toxicomanes soient séropositives à Zürich, tandis qu'à Genève, ce taux passerait à 40%. Une étude RDS menée entre Genève et Zürich pourrait estimer le pourcentage de personnes séropositives parmi les personnes toxicomanes de 35%. Ainsi, ce résultat est dénué de sens, puisque les deux populations, de Zürich et de Genève, sont totalement différentes [5].
Si l'on étudie sciemment au cours d'une même enquête, plusieurs composantes qui ne sont pas connectées, on fait implicitement l'hypothèse, très forte, que la répartition du caractère observé est homogène au sein des deux populations. Dans ce genre de cas, il est recommandé de faire plusieurs enquêtes différentes sur chaque composante connexe ou alors d'admettre que l'enquête n'est pas faisable d'un point de vue méthodologique et d'y renoncer (Gile et Handcock, 2010).

5. Ce cas est bien évidemment fictif et ne sert qu'à illustrer un propos.

Structure du réseau Un des aspect lié à la structure du réseau est l'activité relative d'un groupe i par rapport à un groupe j. Elle est simplement définie comme :

$$w_{ij} = \frac{\overline{d}_i}{\overline{d}_j}, \tag{10}$$

où $\overline{d}_i$ représente le degré moyen du sous-ensemble i. C'est un indicateur de la capacité d'un sous-ensemble à recruter et à être recruté, relativement à un autre groupe. Puisque dans la plupart des cas, on ne considère que deux groupes, on suppose que i représente le groupe des infectés et j le groupe des bien portants et on omet de préciser l'indice. De fait, si la quantité w est proche de 1, il s'agit d'un cas idéal. Lorsque le rapport w est différent de 1, cela peut biaiser dramatiquement l'estimation (Gile et Handcock, 2010; Gile et Tomas, 2011). Dans ce cas, on dit qu'il y a une activité différenciée (*differential activity*, c'est ainsi que l'on y fait référence dans la littérature).

Par ailleurs, on peut aussi considérer une homophilie de réseau. En effet, il est raisonnable de supposer que certains individus ont plus de liens avec des individus qui leur sont semblables. Dans la société, on parle de communautarisme dans certains cas pour illustrer ce genre de situation.

Non-réponse En ce qui concerne la non-réponse, cet aspect est peu évoqué dans la littérature, à l'exception notable de (Gile et Handcock, 2010; Gile et Tomas, 2011). En général, lorsque le niveau de non-réponse augmente, cela augmente inévitablement la variance de l'estimation. Il est aussi raisonnable de considérer une situation où la non-réponse varie en fonction du sous-ensemble considéré ou en fonction du degré moyen du groupe considéré. Ceci n'est pas forcément un artifice mathématique et cela peut très bien représenter une situation réelle. On peut imaginer une étude sur des personnes qui se prostituent et pour laquelle la proportion de non-réponse est plus élevée parmi celles qui n'ont pas de permis de séjour. La non-réponse est un problème inhérent à tous les types de sondages et la procédure RDS n'échappe pas à cette règle. En particulier, lorsqu'elle est associée à d'autres types de perturbations (recrutement différencié, activité différenciée), des biais très importants peuvent apparaître.

Conclusion En règle générale, lorsqu'une seule des hypothèses est violée, les simulations montrent que l'estimation n'est pas nécessairement biaisée. Toutefois, lorsqu'elles s'accumulent, des biais, parfois importants peuvent apparaître (Gile et Handcock, 2010; Gile et Tomas, 2011). De plus, on remarque que certaines d'entre elles sont impossibles à vérifier dans un cas concret. La seule manière de certifier la méthode RDS est donc l'usage de la simulation, pour tester les hypothèses les plus extrêmes.

Très peu de données fiables existent au sujet des personnes difficiles à atteindre. Il est donc très difficile de pouvoir comparer les résultats obtenus via la méthode RDS à des données existantes. Par ailleurs, les données récoltées à l'aide de la méthode RDS sont peu nombreuses à être disponible librement (Salganik, 2012). On peut quand même noter une étude de (McCreesh *et al.*, 2012), qui est une référence dans sa méthodologie (Salganik, 2012). On y étudie une population dont les caractéristiques sont connues, en vue d'évaluer la méthode RDS. L'estimateur de Volz-Heckathorn y est utilisé. Les réultats y sont sévères pour la méthode RDS et montrent que les réserves mentionnées jusqu'ici sont fondées. Nous reviendrons plus tard sur un bref compte-rendu de cette étude (voir 4.1)

2.10 Sensibilité des estimateurs

Pour évaluer définitivement la méthode RDS et ses estimateurs, il est essentiel, ainsi que le préconise (Salganik, 2012), de la soumettre à deux types d'épreuves. D'une part, des simulations, d'autres part des échantillonnages via la méthode RDS menés sur des populations pour lesquelles des données fiables sont déjà disponibles, en vue d'une comparaison des deux résultats. Rares sont les études qui permettent une telle comparaison et elles sont appelées à se développer. Cependant, il faut admettre qu'elles sont complexes à mettre en place et nécessitent une solide infrastructure. A cet égard, l'étude ménée par McCreesh *et al.* est exemplaire (McCreesh *et al.*, 2012).
Il est essentiel de garder à l'esprit que la méthode RDS est avant tout une méthode qui vise à être appliquée. Les besoins pratiques doivent demeurer primordiaux. Malheureusement, seules des simulations sont en mesures de certifier pleinement cette méthode car il est impossible d'effectuer un nombre suffisant d'enquêtes, avec des paramètres connus, pour valider la méthode.
On a déjà mentionné le fait que si une seule des hypothèses est violée, cela ne conduit pas nécessairement à un biais, alors que dans le cas où plusieurs le sont, les estimateurs y deviennent nettement plus sensibles. Les simulations doivent donc être employées de manière soumettre les différents estimateurs à des situations, peut-être peu probables mais qui ont une probabilité non négligeable de survenir.
Finalement, un dernier point mérite d'tre mentionné. Les estimateurs de Salganik-Heckathorn et de Volz-Heckathorn ne sont pas consistants (Cochran, 1977, p.21), c'est-à-dire que l'on a pas forcément :

$$\lim_{n \to N} \mu_{\text{V-H/S-H}}(S_i) = \frac{N_i}{N} = \pi_i,$$

où $\mu_{\text{V-H/S-H}}(S_i)$ désigne l'estimateur de Volz-Heckathorn (ou de Salganik-Heckathorn) de la proportion d'un groupe S_i au sein de la population, c'est-à-dire la proportion au sein de l'échantillon. En effet, puisqu'ils ne se basent pas directement sur la population échantillonnée mais plutôt sur la structure du réseau, un biais n'est pas impossible. Cette justification est essentiellement heuristique mais elle est vérifiée dans les simulations (Gile et Handcock, 2010).

3 Simulations

La problématique des populations difficiles à atteindre est telle qu'il est difficile de tester des estimateurs sur des données réelles. En effet, pour ce faire, il faut utiliser une procédure d'échantillonnage RDS sur une population dont on connaît a priori les véritables caractéristiques. Or, de telles études sont rares car très difficiles à mettre en œuvre. A notre connaissance, il n'existe à l'heure actuelle qu'une seule étude de ce genre (McCreesh *et al.*, 2012). La modélisation mathématique des réseaux sociaux que constituent les populations difficiles à atteindre semble constituter une difficulté pour l'instant infranchissable. Par ailleurs, le processus d'échantillonnage est lui aussi d'une complexité telle qu'il n'existe pas de formulation mathématique de ce processus qui soit pleinement satisfaisante.
Dans ce contexte, la simulation semble être le moyen le plus approprié de mener à bien une étude empirique des différents estimateurs existants. Toutefois, une grande précaution est de mise. Il est en effet nécessaire de garder à l'esprit que ce type de procédure ne peut pas rendre compte de la complexité d'une situation réelle mais seulement en être une approximation.
Le premier ingrédient nécessaire à une simulation d'une procédure RDS est la génération d'un graphe, qui doit être une approximation d'un réseau social. Cette étape est extrêmement complexe. Dans notre cas, on a choisi de se limiter à un certain nombre de propriétés que l'on peut imposer au réseau. Par ailleurs, la génération d'un tel réseau est aléatoire. Pour plus de

précision quant aux détails d'implémentation de la génération du réseau, on peut se référer au guide d'utilisation qui accompagne le code (Wilhelm, 2012).
Le deuxième ingrédient est le processus d'échantillonnage lui-même. C'est-à-dire qu'à partir du réseau, on doit être capable de lancer le processus qui permet d'obtenir l'échantillon. Toutefois, pour rendre compte de la spécificité du processus d'échantillonnage RDS, l'objet ainsi obtenu doit rendre possible la construction de la totalité du processus. Concrètement, il ne s'agit pas seulement de savoir qui a été échantillonné mais aussi de savoir qui a recruté qui et de pouvoir ainsi reconstituer le cheminement du recrutement au sein du réseau. En outre, il est nécessaire de pouvoir faire varier certains paramètres qui caractérisent le processus de recrutement.
Finalement, après avoir obtenu un échantillon, il suffit de calculer la valeur des estimateurs. Ainsi, on peut facilement identifier 3 étapes distinctes pour le processus : La génération du réseau, l'échantillonnage et le calcul des estimateurs. On ne va pas s'attarder sur les possibilités offertes par le code mais ce dernier permet de reproduire la plupart des simulations existantes dans la littérature. Pour tous les détails, on peut toujours se référer au guide de l'utilisateur. Toute l'implémentation est faite grâce au logiciel $\mathbf{R}$ (R Core Team, 2012) et en particulier grâce au package `statnet` (Handcock *et al.*, 2003).
Dans toutes les simulations qui apparaissent, si rien n'est précisé, les valeurs par défaut sont les suivantes :

- *nombre de simulations* 1000 simulations
- *taille du réseau* $N = 1000$
- *taille de l'échantillon* $n = 500$
- *proportion de la population infectée* 20 %
- *homophilie de réseau* pas d'homophilie de réseau
- *homophilie de processus* pas d'homophilie de processus
- *activité relative* $w = 1$, c'est-à-dire pas d'activité différenciée
- *degré moyen du réseau* $\overline{d} = 7$
- *nombre de germes* 10 germes
- *nombre de coupons* 3 coupons
- *estimation du degré* exacte
- *non réponse* pas de non-réponse
- *remise* avec remise
- *choix des germes* uniformément aléatoireme parmi la population

3.1 L'homophilie

Dans un premier temps, on a soumis les différents estimateurs à une variation de l'homophilie. Il s'agit de différencier deux types d'homophilie : l'homophilie de réseau ainsi que l'homophilie du processus d'échantillonnage. S'il y a de l'homophilie de réseau, c'est que la probabilité pour un lien donné de connecter deux individus de même type est plus grande que la probabilité de connecter deux individus de type différent. Cette homophilie est donc à situer au niveau du réseau. L'homophilie du processus d'échantillonnage est quant à elle due au fait que lors du processus lui-même, certains individus ne choisissent pas leurs recrues de manière uniformément aléatoire parmi les individus avec lesquels ils sont connectés mais ont une plus forte probabilité de choisir quelqu'un qui partage avec eux une certaine caractéristique. En général, cette caractéristique est indépendante (au sens mathématique du terme) de la variable que l'on étudie, alors il est facile de démontrer que l'homophilie n'influence pas l'estimation de la proportion au sein de la population totale. Lorsqu'au contraire l'homophilie est justement la variable observée (par exemple, si l'on cherche à estimer la prévalence du SIDA et si les personnes saines ont tendance à plutôt recruter des personnes saines), alors l'échantillonnage sera forcément biaisé, c'est-à-dire que la prévalence du SIDA est significativement différente

au sein de la population totale qu'au sein de l'échantillon.

3.1.1 L'homophilie de processus

En modifiant graduellement l'homophilie de réseau, on peut observer le comportement des estimateurs lorsque l'homophilie varie. On simule 1000 réseaux qui possèdent les mêmes caractéristiques et, sur chacun des réseaux, on effectue une procédure d'échantillonnage RDS. Pour chaque échantillon, on calcule la valeur des 4 estimateurs abordés au cours de ce travail.

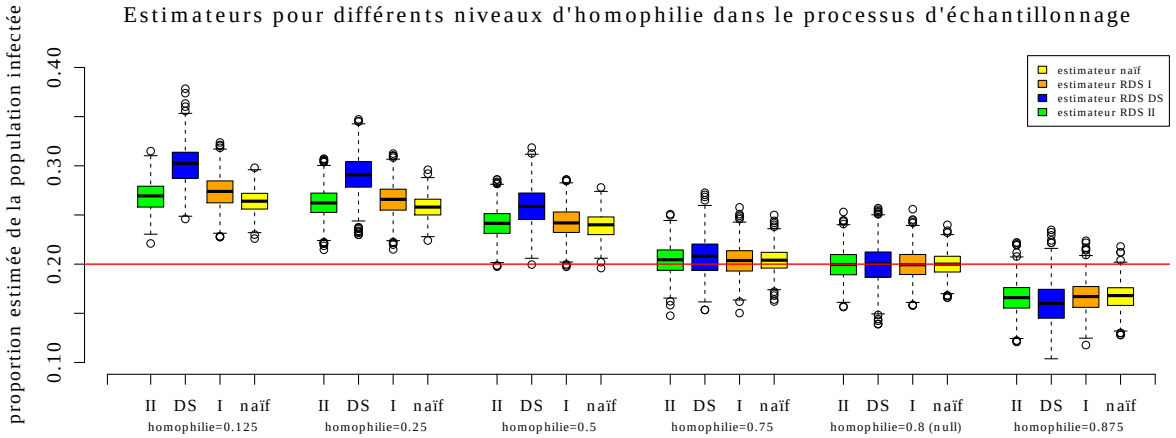


FIGURE 2 Boîtes à moustache des estimateurs dans plusieurs configurations d'homophilie de processus. Précisément, les probabilités de transitions infecté/sain et et infecté/infecté restent inchangées et valent $1/2$. La probabilité de transition sain/sain prend les valeurs de $1/8$, $1/4$, $1/2$, $3/4$ et $7/8$ et la valeur `null`. La valeur `null` signifie que les individus sont recrutés uniformément aléatoirement parmi toutes les connections du recruteur. Ce niveau se situe en moyenne à 0.8, ce qui correspond à la proportion d'individus sains au sein de la population. Les autres valeurs sont celles par défaut.

On remarque un certain nombre de choses sur la figure 2. Tout d'abord, aucun des estimateurs ne semble être en mesure de corriger le biais dû à l'homophilie. Tous échouent à estimer de manière non biaisée la véritable proportion de la population infectée. Plus la probabilité de transition est proche de 0.8, moins le biais des estimateurs semble important. Le biais semble donc clairement sensible à une variation de l'homophilie. De plus, le signe du biais semble dépendre de l'homophilie, ce qui est raisonnable. Lorsque la valeur est supérieure à 0.8, on remarque que, même avec une faible différence de niveau d'homophilie, le biais est immédiatement très important. Ceci est principalement lié à la valeur de niveau d'homophilie relativement élevée. Finalement, le biais de tous les estimateurs semble suivre une même tendance lorsqu'ils sont soumis à un même niveau d'homophilie.
Les variances quant à elles semblent peu ou pas affectées. Pour chacun des estimateurs, elles semblent demeurer à des niveaux semblables quelque soit le niveau d'homophilie de processus. De fait, l'écart entre les valeurs minimale et maximale de l'écart-type est inférieur à 0.02, quelque soit l'estimateur.
On remarque de sensibles différences entre les estimateurs. Premièrement, si aucun ne parvient à corriger le biais induit par le recrutement différencié, l'estimateur RDS DS y semble plus sensible que les autres. De plus, souvent l'estimateur le moins biaisé semble être l'estimateur naïf, et lorsque ce n'est pas le cas, il en est très proche. Ce fait est particulièrement étonnant.

Les variances des estimateurs RDS I et RDS II sont très proches, dans toutes les situations et semblent raisonnables. La variance de l'estimateur naïf est la plus faible en général et les estimateurs RDS I et RDS II ont une variance qui est très proche l'une de l'autre. La variance de l'estimateur RDS DS est dans tous les cas plus grande que celle des autres estimateurs.
On remarque donc qu'aucun de ces estimateurs ne parvient à corriger le biais induit par l'homophilie liée au processus de recrutement. Le biais induit peut-être d'important suivant le niveau d'homophilie. Toutefois, de tels niveaux d'homophilie ne semblent pas pouvoir être totalement exclus de situations expérimentales réelles. En effet, la stigmatisation de certaines populations font qu'elles sont systématiquement rejetées par le reste de la population, à l'instar des intouchables dans la société indienne par exemple.
Finalement, il est aussi important de préciser que, puisque c'est un mécanisme lié au processus de recrutement, l'enquêteur potentiel n'est pas totalement impuissant face à une telle situation. En effet, on peut imaginer la procédure suivante, pour éviter effectivement qu'un tel phénomène se produise :

- Demander à la recrue de compter le nombre de personne qu'il connaît au sein de la population cible. Pour chacune de ces personnes, lui donner un surnom et ainsi établir une liste de ses connaissances au sein du réseau qui soit anonyme pour l'enquêteur.
- Attribuer un entier pour chacune de ces personnes et faire un tirage aléatoire parmi ces personnes. Le choix de la distribution des coupons n'est donc pas totalement du ressort de la recrue.
- Vérifier lorsque les nouvelles recrues participent à l'enquête si elles se reconnaissent dans le surnom qui les désigne.

De cette manière, on peut éviter, dans une certaine mesure, le recrutement différencié.

3.1.2 L'homophilie de réseau

L'homophilie de réseau est une caractéristique propre au réseau. Il n'est pas évident que la structure même du réseau permette qu'un échantillonnage RDS produise un échantillon non-biaisé, quand bien même ce dernier est sans homophilie. En fait, certaines caractéristiques du réseau semblent influencer le processus. Plus précisément, en admettant un échantillonnage idéal, c'est-à-dire qui satisfait toutes les hypothèses désirées, peut-on être sûr que les estimateurs se comportent de manière fiable' Dans ce qui suit, on fait varier deux paramètres différents, l'homophilie de réseau et le niveau d'activité relatif. Ce sont deux paramètres propres au réseau.
On peut voir sur la figure 3 comment se comportent les estimateurs lorsque l'on fait varier le niveau d'homophilie de réseau et le niveau d'activité relatif. Dans ces différents cas, le processus d'échantillonnage est idéal, dans le sens où l'on n'y inclut aucun biais. Tout d'abord, on voit que selon la situation, le fait de varier ces paramètres influence significativement les estimateurs. Il est très étonnant de voir que les deux paramètres testés simultanément semblent influencer les estimations de manière presque indépendante. En effet, il semble que l'homophilie de réseau n'a une influence que sur la variance des estimateurs tandis que l'activité relative ait surtout une influence sur les biais des estimateurs ainsi qu'une légère influence sur la variance, la faisant baisser. Dans toutes les situations, on peut observer que l'estimateur naïf est celui qui se comporte le moins bien. Ensuite, on note que le biais augmente lorsque l'activité relative augmente. En raison de la configuration du réseau, le seuil minimal de w est de 0.8 (voir (Wilhelm, 2012)). Finalement, lorsque que l'activité relative augmente, alors un biais apparaît. De plus, pour un même niveau d'homophilie de réseau, la variance diminue aussi lorsque l'activité relative augmente. Par contre, à activité relative égale, la variance augmente sensiblement avec le niveau d'homophilie. Toutes ces remarques sont valables pour tous les estimateurs excepté l'estimateur naïf.

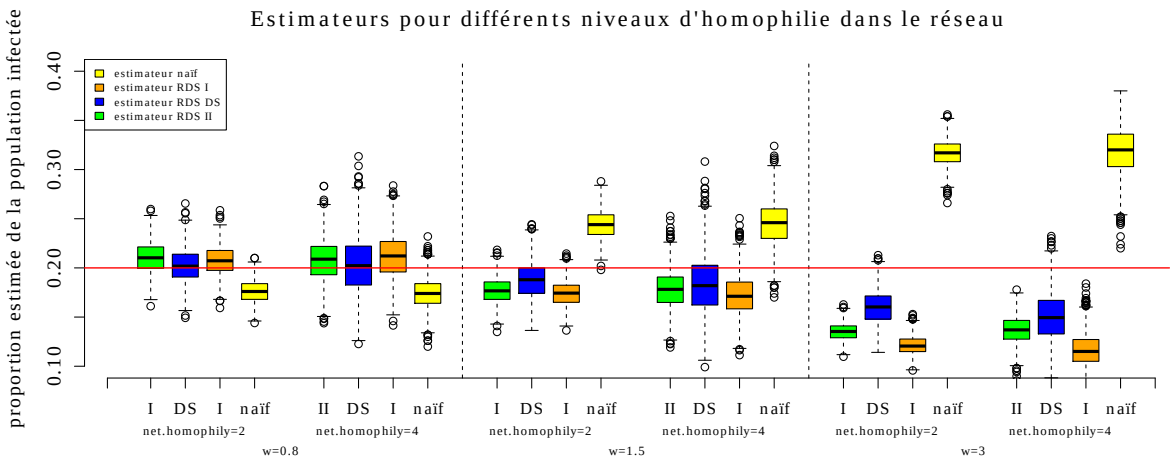


FIGURE 3 Boîtes à moustache des estimateurs dans plusieurs configurations d'homophilie de réseau et de niveau d'homophilie de processus. Ici, le paramètre `net.homophily` désigne le rapport entre le nombre d'arêtes internes aux groupes, c'est-à-dire les arêtes infecté/infecté ou sain/sain et le nombre d'arêtes inter-groupe. Si l'on distribue uniformément les arêtes sur le réseau (en excluant les liens d'un individu avec lui-même), ce rapport est de 2.12 dans la configuration par défaut de notre réseau. Le niveau d'activité w est le rapport entre le degré moyen des individus sains et le degré moyen des individus infectés (voir équation (10)). Tous les autres paramètres sont choisis par défaut.

Dans toutes les situations, hormis la situation idéale, c'est-à-dire lorsque le paramètre `group.homophily` $= 2$ et $w = 1$, l'estimateur naïf semble être biaisé, voire très biaisé. Si l'on examine le comportement des estimateurs un à un, on remarque que l'estimateur RDS DS a en général une variance plus grande que les deux autres. Toutefois, son biais est moindre que pour tous les autres estimateurs lorsqu'il y a une forte activité relative. On a vu à la section 2.6 que l'estimateur RDS DS est sensé être plus performant en présence d'activité différenciée et de recrutement différencié. Or, s'il parvient à être moins biaisé en présence d'activité relative importante, comme on peut le voir sur la figure 3, il ne parvient pas à corriger le biais dû à l'homophilie de processus, comme on peut le voir sur la figure 2. Quant aux estimateurs RDS I et RDS II, ils semblent se comporter de manière très similaire, tant du point de vue de la variance que du point de vue du biais.

3.1.3 Homophilie de réseau et homophilie de processus

Les deux simulations qui précèdent nous permettent de mettre en lumière que les deux types d'homophilie, de processus d'échantillonnage et de réseau, ont une grande influence sur le biais et la variance des estimateurs. De plus l'activité relative est elle aussi un facteur qui compte beaucoup dans l'estimation.
On peut voir que l'homophilie de réseau et l'homophilie de processus influent toutes les deux sur l'estimation. Il est donc naturel de vouloir combiner ces deux facteurs de manière à voir quels sont les effets conjugués sur les estimateurs. Ainsi, on peut éventuellement identifier les facteurs les plus influents.
On voit sur la figure 4 que le biais semble plutôt être lié à l'homophilie de processus qu'à l'homophilie de réseau. Dans ce cas précis, le biais est très fort lorsque la probabilité de transition sain/sain est très faible. En effet, puisque la proportion d'individu infecté est très faible, il peut résulter d'une telle probabilité de transition une probabilité de sélection d'un individu très forte. La figure 4 montre aussi que l'homophilie de réseau a tendance, comme on l'avait déjà re-

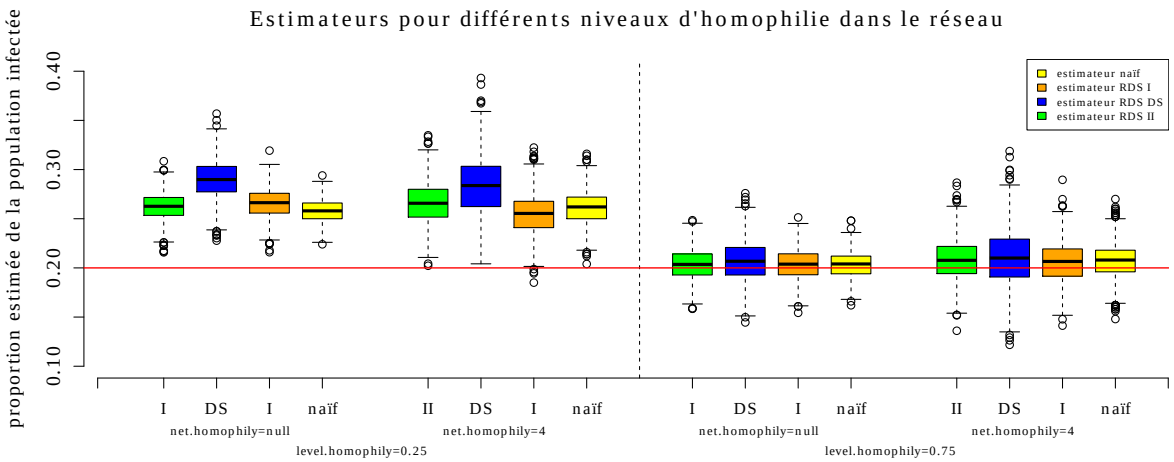


FIGURE 4 Boîtes à moustache des estimateurs dans plusieurs configurations d'homophilie de réseau et de niveau d'activité relatif. Ici, le paramètre `net.homophily` désigne le rapport entre le nombre d'arêtes internes aux groupes, c'est-à-dire les arêtes infecté/infecté ou sain/sain et le nombre d'arêtes inter-groupe. Si l'on distribue uniformément les arêtes sur le réseau (en excluant les liens d'un individu avec lui-même), ce rapport est de 2.12 dans la configuration par défaut de notre réseau. Par ailleurs, le paramètre `level.homophily` dénote la probabilité de transition sain/sain au cours de l'échantillonnage. Tous les autres paramètres sont choisis par défaut.

marqué précédemment, à augmenter la variance des estimateurs. Ici, le cas où la probabilité de transition est de 0.75 semble, en général, être proche de la probabilité qu'aurait un individu d'être sélectionné sans homophilie, dont le niveau correspond en moyenne à 0.8.
D'après les figures 2, 3 et 4, on peut voir que les effets qui sont liés au réseau semblent s'additionner à ceux qui sont liés au processus, tant dans le biais que dans la variance. Il ne semble pas que l'une ou l'autre variable propre au processus ait une influence sur l'effet qu'a une variable propre au réseau, et réciproquement. Cela laisse penser qu'il existe une certaine indépendance entre les deux types de variables, celle propres au réseau et celles propres au processus d'échantillonnage. En particulier, il semble que les estimateurs se comportent en général relativement bien en présence d'homophilie de processus, excepté lorsque cette dernière est extrême, mais par contre ils sont sensibles a l'activité relative. Ce dernier facteur semble être celui qui biaise le plus les estimations.

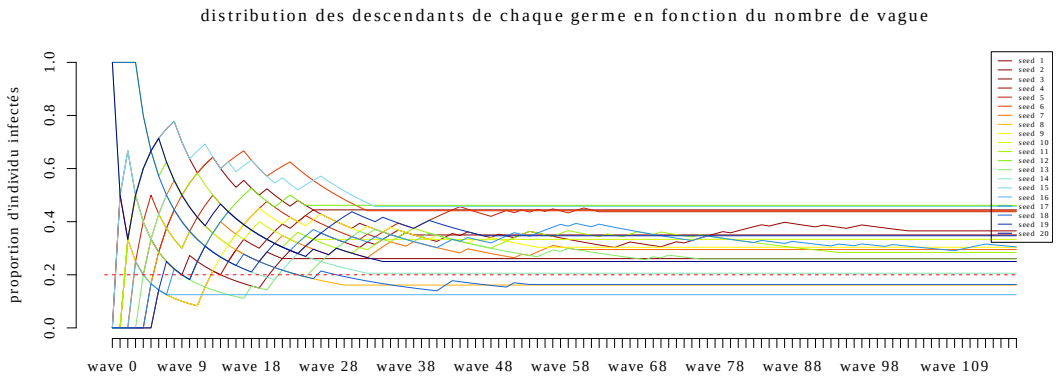
3.2 Dépendance aux germes

L'un des principaux reproches que l'on peut faire à la méthode RDS est le fait que les germes ne sont que très rarement tirés aléatoirement. Dans ce cas, il est donc intéressant de tester la dépendance de l'échantillon final aux germes. Il est dit que la dépendance aux germes devrait disparaître au fil des vagues, en utilisant un argument propre aux chaînes de Markov (voir annexe A)(Heckathorn, 1997, 2002). Il est clair que choisir une distribution des germes (par rapport à la variable d'intérêt) la plus proche possible de la véritable distribution serait idéal. En particulier, si l'homophilie est raisonnable ou faible, alors la distribution stationnaire de la chaîne de Markov (voir section 2.2) doit être la véritable distribution du caractère infectieux au sein du réseau. Toutefois, comme cela a été mentionné précédemment, il n'est pas clair que ces distributions coïncident, précisément à cause de facteurs tels que l'homophilie de processus ou l'activité relative. On peut voir dans les figures 5 et 6 comment évolue la dépendance aux germes au fil du temps. Sur la figure 5, on peut voir la proportion d'individus infectés parmi

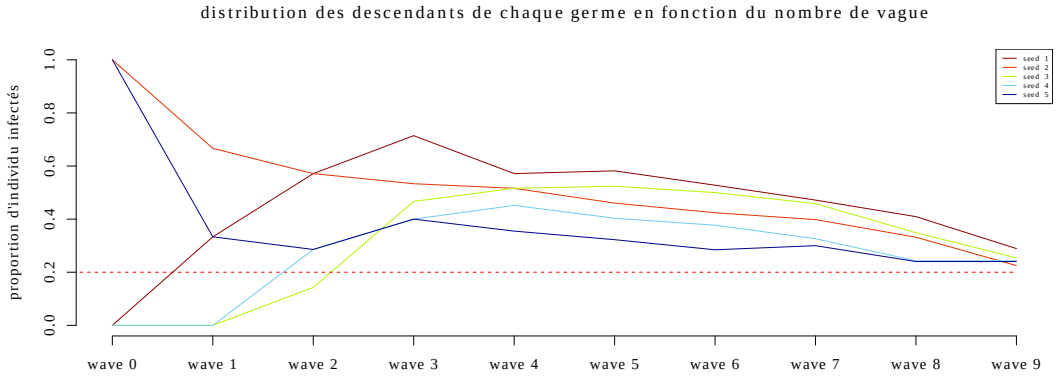
les descendants des différents germes au fil des vagues. Théoriquement, on doit arriver rapidement à un équilibre puisque l'homophilie de processus est nulle (Heckathorn, 1997). Toutefois, en raison de la configuration particulière de notre réseau, c'est-à-dire une forte homophilie de réseau et une forte activité différentielle, l'équilibre atteint n'est pas toujours la véritable proportion. Ainsi, on démontre que même en s'affiranchissant des problèmes liés au processus, la nature du réseau suffit à biaiser fortement le processus. D'autre part, sur la figure 6 on peut voir que s'il est vrai que la distribution semble atteindre un équilibre (qui correspond à l'estimateur trivial), cet équilibre, même après un nombre de vagues très important, ne correspond pas à la véritable proportion. Cette simulation a été répétée un certain nombre de fois et on obtient des résultats similaires. C'est une limitation sérieuse qui est mise en évidence ici. De plus, sur la figure 6, on peut voir que non seulement la proportion est fausse et ne varie que très peu mais les estimateurs ont un comportement semblable, à savoir qu'ils sont dans un état « stationnnaire » sans pour autant être proche de la véritable proportion.
Finalement, on peut conclure de ces deux figures que, contrairement à ce qui est communément admis, l'équilibre atteint ne correspond pas forcément à la véritable proportion d'individus infectés. Le fait donc que le processus atteigne un certain équilibre n'est pas forcément bon signe.
Dans ces conditions idéales, c'est-à-dire lorsque le processus et que le réseau lui-même sont idéaux, c'est-à-dire lorsque l'on n'a ni activité différenciée ni homophilie de processus, les estimateurs convergent en très peu de vagues vers la véritable proportion. On peut l'observer en annexe, sur les figures 14 et 15. De plus, ni le nombre de germes, ni le nombre de coupons ne semblent véritablement influer sur le processus. Dans ces conditions, tous les estimateurs se révèlent très performants. De plus, la moyenne de l'échantillon semble elle aussi très rapidement proche de la véritable proportion.
On étudie le comportement des estimateurs lorsque les germes ne sont pas choisis de manière aléatoire. On peut en voir le résultat sur la figure 7. On fait varier le niveau d'homophilie afin de voir si cela n'agrave pas l'estimation. Au vu de la figure 7, on remarque que la proportion d'individus infectés parmi les germes ne fait pas sensiblement varier les estimateurs. Ceci confirme que la dépendance aux germes à tendance à disparaître assez rapidement lorsque le réseau est idéal, comme c'est le cas sur figure 7.
Les simulations ont montré deux aspects différents. Le premier, c'est que sauf cas extrême, il suffit de peu de vagues pour que l'équilibre soit atteint, c'est-à-dire que ni la proportion au sein de l'échantillon ni les estimateurs eux-mêmes ne varient beaucoup au fil des vagues. Par contre, on a montré que dans certaines conditions, l'équilibre atteint ne correspond pas à la proportion réelle. Il est donc illusoire de croire qu'augmenter le nombre de vagues permettrait de corriger le biais induit par la structure du réseau. Il est donc recommandé de procéder à un nombre suffisant de vagues, typiquement 5 ou 6 mais aller nettement au delà ne saurait corriger le biais introduit par la structure du réseau, s'il existe.

3.3 Tirage avec ou sans remise

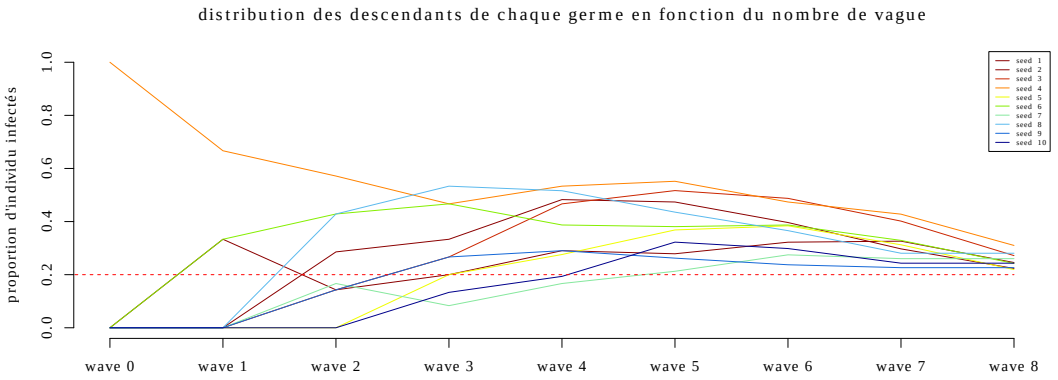
On a vu que pour certains estimateurs, on fait l'hypothèse que le tirage est avec remise, ceci pour utiliser l'estimateur de Horwitz-Thomson. On considère donc des situations ou les estimateurs semble ne pas parvenir à estimer correctement la véritable proportion et on simule un même processus avec remise pour observer les différences potentielles. Dans ce cas, on remarque sur la figure 8 que faire un tirage avec ou sans remise a un effet étonnant. Premièrement, dans tous les cas, le processus avec remise augmente la variance de tous les estimateurs. Ceci est explicable car, dans le cas avec remplacement, le nombre d'échantillons possibles est sensiblement plus grand. Par contre, parfois, le processus avec remise diminue le biais tandis que parfois il l'augmente, et ce, pour tous les estimateurs. En fait, il semble ajouter



(a) $n = 1000$ avec 1 coupons et 20 germes

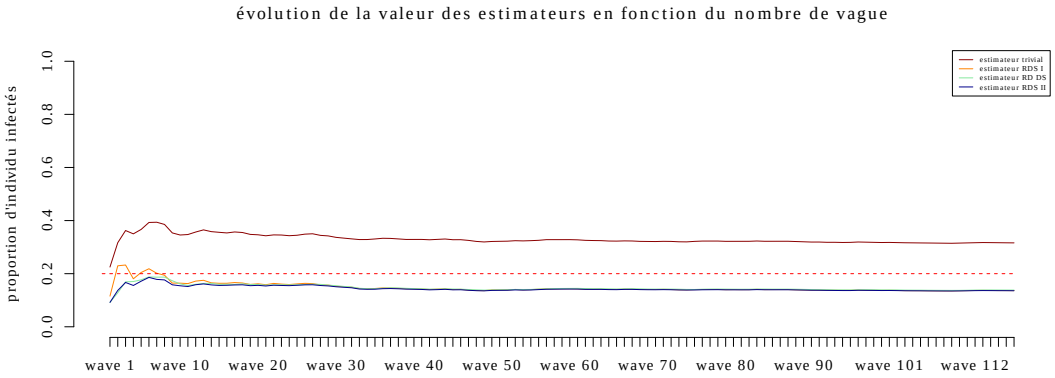


(b) $n = 3000$ avec 2 coupons et 5 germes

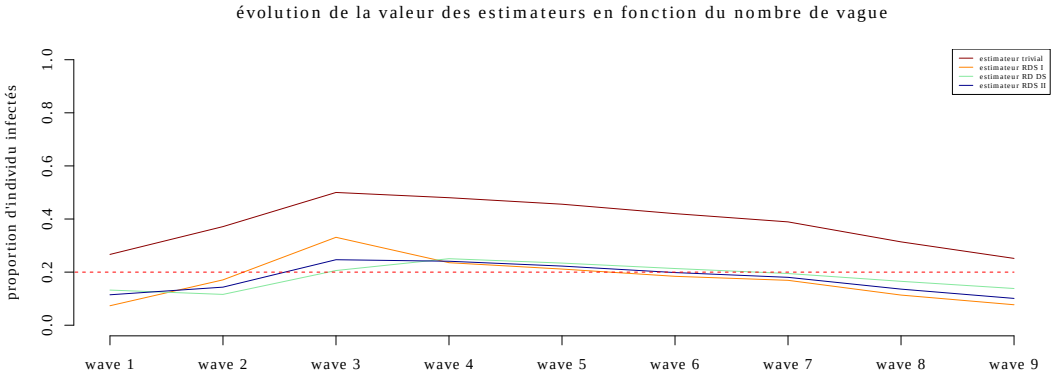


(c) $n = 3000$ avec 2 coupons et 10 germes

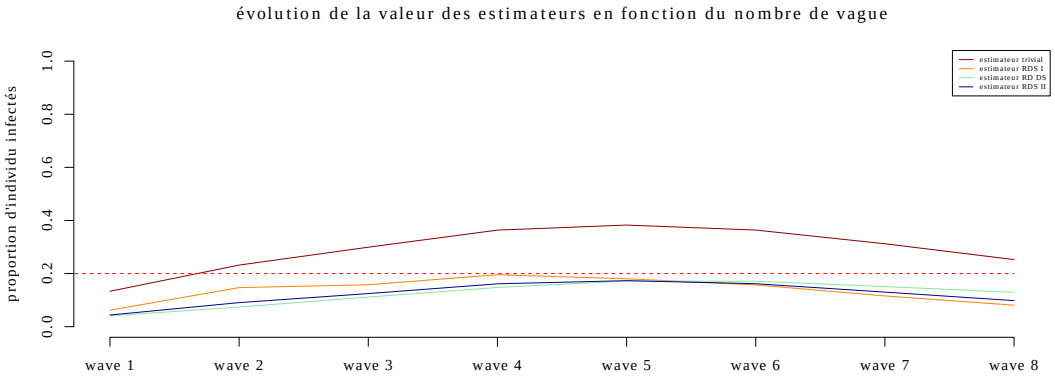
FIGURE 5 Distributions des descendants des germes en fonction du nombre de vagues. Le réseau simulé a toujours les mêmes caractéristiques : le paramètre `group.homophily` est égal à 4, l'activité relative w est égale à 3 et il y a 4000 individus dans la population totale. Tous les autres paramètres du processus et du réseau sont ceux par défaut.



(a) $n = 1000$ avec 1 coupons et 20 germes



(b) $n = 3000$ avec 2 coupons et 5 germes



(c) $n = 3000$ avec 2 coupons et 10 germes

FIGURE 6 Evolution des estimateurs au cours du temps. Le réseau simulé ainsi que les échantillons sont les mêmes que ceux de la figure 5.

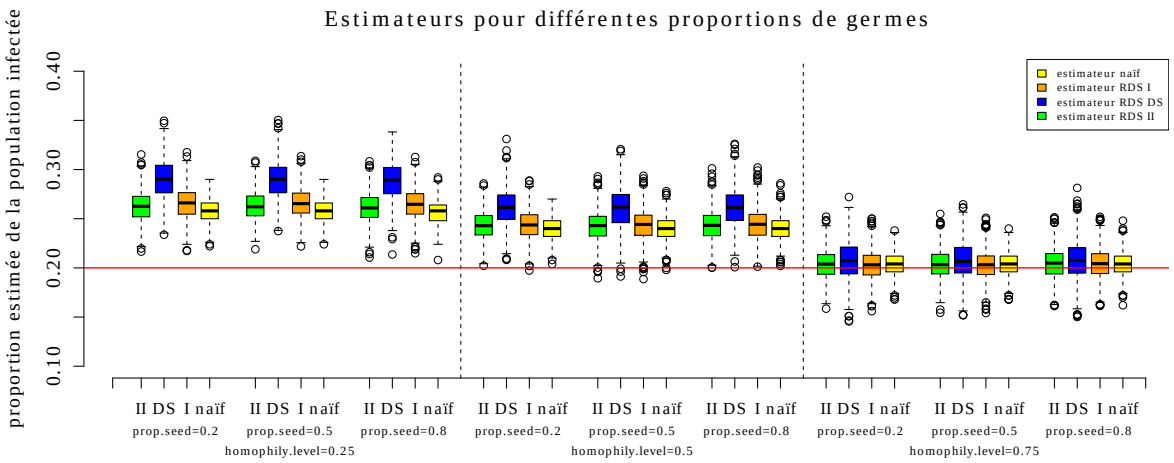


FIGURE 7 Boîtes à moustache des estimateurs dans plusieurs configurations d'homophilie de processus et de proportions de germes sains. Le nombre de vagues est de 4. Tous les autres paramètres sont choisis par défaut.

un biais qui varie en fonction de l'homophilie de processus. Dans les deux situations, l'homophilie de processus est biaisée par rapport à l'équilibre (l'équilibre correspond en moyenne à 0.8). Donc, un processus avec remplacement a tendance à augmenter la proportion de l'estimation de la taille du groupe qui est favorisé par l'homophilie de processus. A noter que sur la figure 8, la situation de gauche est en fait très particulière. La valeur de l'activité différentielle $w = 3$ a tendance à introduire un grand biais négatif. Toutefois, l'homophilie de processus qui vaut 0.25 aura tendance à introduire un grand biais positif. Dans ce cas, les deux biais se compensent, raison pour laquelle l'estimation sans remplacement semble très bonne. Cependant, puisque l'homophilie de processus est très forte, l'introduction d'un tirage avec remplacement biaise beaucoup les estimations.
Pour conclure, il semblerait que le fait que le processus soit fait avec remise introduit un biais qui va dans le même sens que l'homophilie de processus. Il ne semble donc pas qu'il s'agisse d'une hypothèse cruciale dans la pratique. Elle est plutôt à considérer en parallèle avec l'homophilie de processus.
On a vu dans la section 2.9 que l'hypothèse de non-remise peut entraîner une autre hypothèse. En effet, si le taux d'échantillonnage est faible alors la différence entre un échantillon avec ou sans remise est négligeable. Ainsi, on peut remplacer l'hypothèse selon laquelle l'échantillonnage est fait avec remise par une hypothèse selon laquelle le taux d'échantillonnage est faible. Dans ce qui suit, on utilise volontairement des taux d'échantillonnage importants pour étudier l'impact de la violation d'une telle hypothèse.
La figure 9 illustre un aspect essentiel des estimateurs. En effet, on voit, qu'en cas d'activité différenciée, l'estimateur RDS II, n'est pas consistant. Ceci est un problème théorique très important. De fait, utiliser un estimateur qui n'est pas consistant est problématique et constitue une très sérieuse objection à l'utilisation des estimateurs RDS. En effet, tous sont inconsistants, à l'exception de l'estimateur naïf, ainsi que l'on peut aussi le voir sur les figures 9 et 16.

3.4 Non-réponse

La non-réponse est l'un des principaux problèmes auquel on est confronté lors d'enquêtes. Elle peut être de nombreux types mais dans le cas d'un échantillonnage de type RDS, on différencie deux types de non-réponse. Le premier est lorsqu'un coupon n'est pas distribué. Dans ce cas,

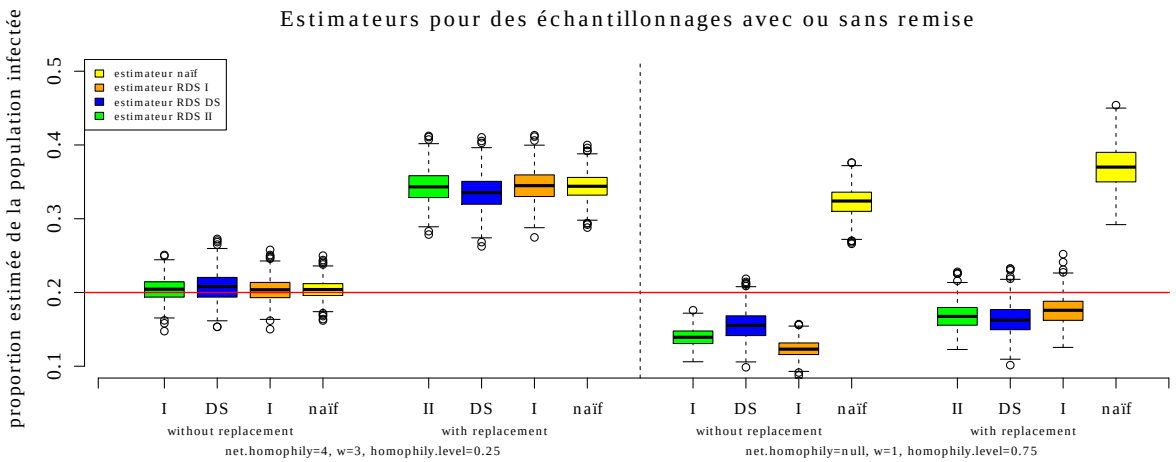
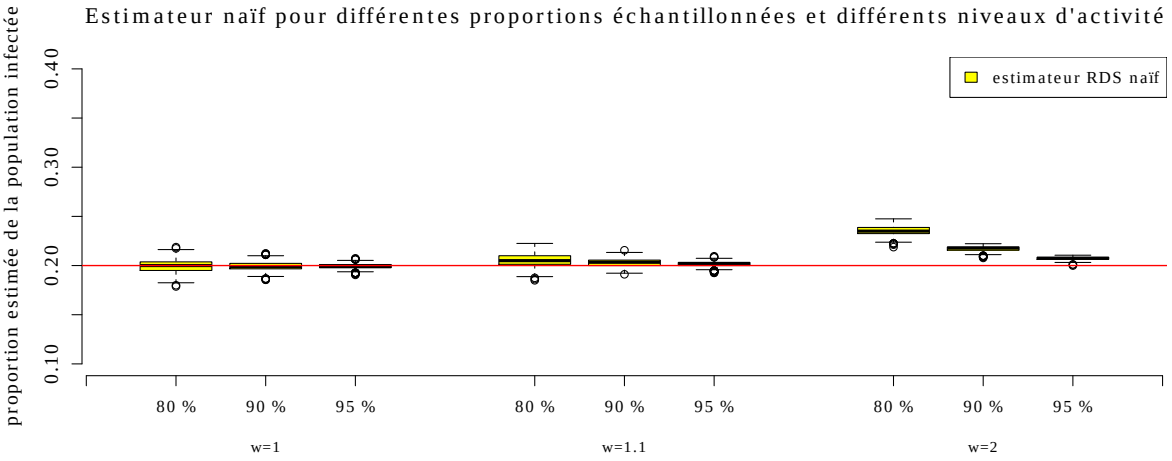


FIGURE 8 Boîtes à moustache des estimateurs avec un échantillonnage avec ou sans remise. Les autres paramètres sont indiqués en légende. Il y a 4 vagues. Tous les autres paramètres sont choisis par défaut.

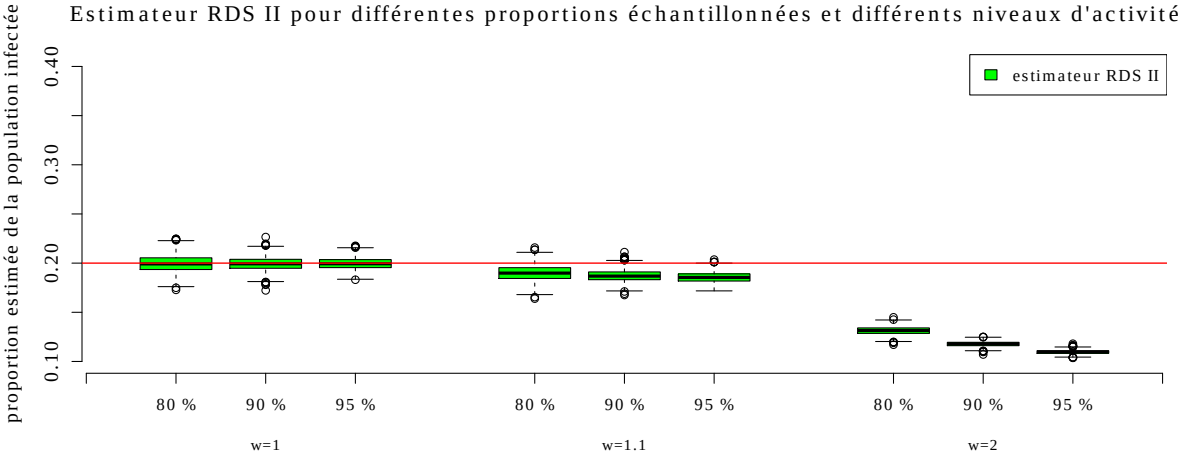
le problème se situe au niveau du recruteur. Le second est lorsqu'un coupon est transmis mais que la recrue renonce à participer à l'étude. La différence entre ces deux types est subtile. Dans le cas où le coupon est transmis, cela signifie que la recrue qui refuse de répondre ne peut pas être recrutée (même au cours d'une vague ultérieure). Ici, nous nous concentrons sur la non-réponse de second type, c'est-à-dire celle qui a lieu lorsque qu'une recrue refuse de participer à l'étude. On dénote dans la suite par « niveau de non-réponse » la probabilité d'un individu de ne pas répondre lorsqu'il est recruté. Lorsque ce dernier est dénoté par un vecteur à deux composantes, il s'agit du niveau de non-réponse de chacun des deux statuts infectieux.

Premièrement, on a fait varier le niveau de non-réponse, comme on peut le voir sur la figure 10. Dans ce cas, le résultat est conforme à ce à quoi on peut s'attendre. En effet, une augmentation de la non-réponse pour le groupe des infectés induit un biais négatif tandis qu'une augmentation de la non-réponse dans le groupe des individus sains induit un biais positif. Il semble que la magnitude du biais soit comparable, si l'on fait varier le niveau de la non-réponse d'un seul des deux groupes et que l'autre demeure nul. Par contre, pour un même niveau de non-réponse, le biais est plutôt négatif. Ceci est étonnant et est probablement dû à la différence de proportion entre les individus sains et ceux qui sont infectés au sein de la population totale. Il semble toutefois que la non-réponse a, dans ce cas, un effet relativement prévisible.

Sur la figure 11, on peut voir quels sont les effets de la non-réponse lorsqu'elle est couplée à de l'homophilie de processus. Dans ce cas, l'homophilie est une homophilie relative, c'est-à-dire que lorsqu'un individu est en passe de recruter un autre individu, il a 1.5 ou 2 fois plus de chance de recruter un individu infecté qu'un individu sain. C'est donc une homophilie de processus relative dans la mesure où les poids de sélection de chacun des individus connectés au recruteur varient selon la distribution d'individus sains parmi les individus connectés au recruteur. Concrètement, le poids de sélection d'un individu infecté est `mult.homophily` fois plus élevé que le poids d'un individu sain. On donc fait varier ce paramètre pour un niveau de non-réponse fixe. On remarque que les effets des deux facteurs semblent s'additionner. Il en résulte une estimation qui est proche de la réalité lorsque le taux de non-réponse chez les personnes infectées augmente et que le facteur `mult.homophily` augmente aussi. Comme dans la plupart des simulations précédentes, on voit que l'estimateur d'Heckathorn (RDS DS) a tendance à avoir une plus grande variance à être plus biaisé que les autres. De plus, à nouveau l'estimateur naïf semble le meilleur dans toutes les situations, tant du point de vue du biais que



(a) Estimateur naïf



(b) Estimateur RDS II

FIGURE 9 Estimateurs naïf et RDS II (Volz-Heckathorn) en fonction de l'activité relative et du taux de sondage.

du point de vue de la variance.

On voit sur la figure 12 que la non-réponse semble être un effet qui se superpose aux effets du processus d'échantillonnage et à la structure du réseau. Elle semble en effet induire un biais négatif qui s'ajoute aux autres biais. Il ne semble pas que l'effet varie en importance selon les conditions d'échantillonnage et de processus.

En conclusion, on peut remarquer que la non-réponse est un phénomène qui peut induire un biais. On suppose que ce qui motive la non-réponse est le statut infectieux, ce qui est réaliste dans bon nombre de situation réelle d'enquêtes RDS. Pourtant, il se pourrait qu'elle soit motivée par d'autres facteurs, peu ou pas corrélés avec le statut infectieux, auquel cas elle aurait un effet relativement faible. Toutefois, l'observation la plus intéressante est peut-être celle qui indique que pour un même niveau de non-réponse, l'effet n'est pas nul sur les estimations, ce qui aurait pourtant pu être supposé. Donc, la non-réponse biaise de toute façon les estimations.

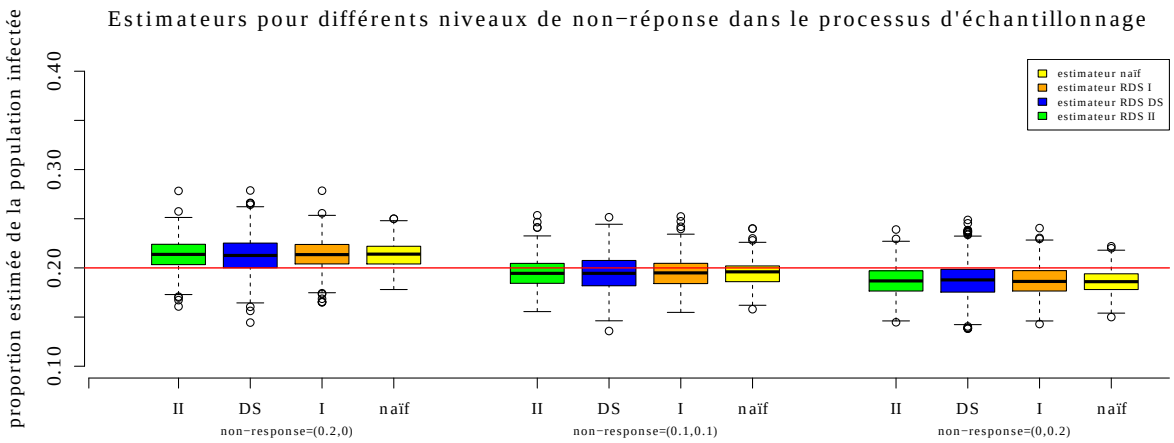


FIGURE 10 Boîtes à moustache des estimateurs pour plusieurs niveaux de non-réponse. Tous les autres paramètres sont choisis par défaut, ce qui signifie dans ce cas que le réseau et le processus sont considérés comme idéaux.

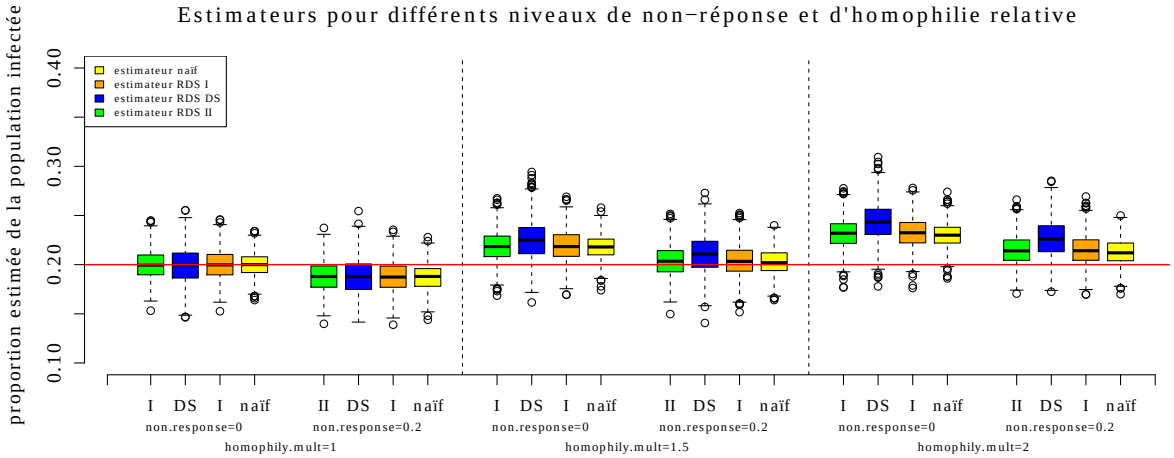


FIGURE 11 Boîtes a moustache des estimateurs pour plusieurs niveaux de non-réponse. Tous les autres paramètres sont choisis par défaut, ce qui signifie dans ce cas que le réseau et le processus sont considérés comme idéaux.

3.5 Conclusion sur les simulations

Les simulations ont permis d'illustrer les propos du chapitre 2. Ainsi, on a pu voir différentes situations où la violation des hypothèses nécessaires à l'utilisation des estimateurs RDS est problématique. Il semble aussi qu'ils sont souvent moins performants que l'estimateur naïf, tant d'un point de vue de la variance que d'un point de vue du biais. L'estimateur RDS DS est en général caractérisé par une nettement plus grande variance. Toutefois, il se comporte mieux que les autres estimateurs lorsque le réseau a une activité différenciée. En général, les estimateurs RDS I et RDS II se comportent de manière semblable, même si, d'un point de vue de l'erreur carrée moyenne, l'estimateur RDS II semble être légèrement meilleur (Gile et Handcock, 2010).

Finalement, le fait que les estimateurs RDS ne soient pas consistant est très problématique, quand bien même cela se produit dans des cas assez particuliers.

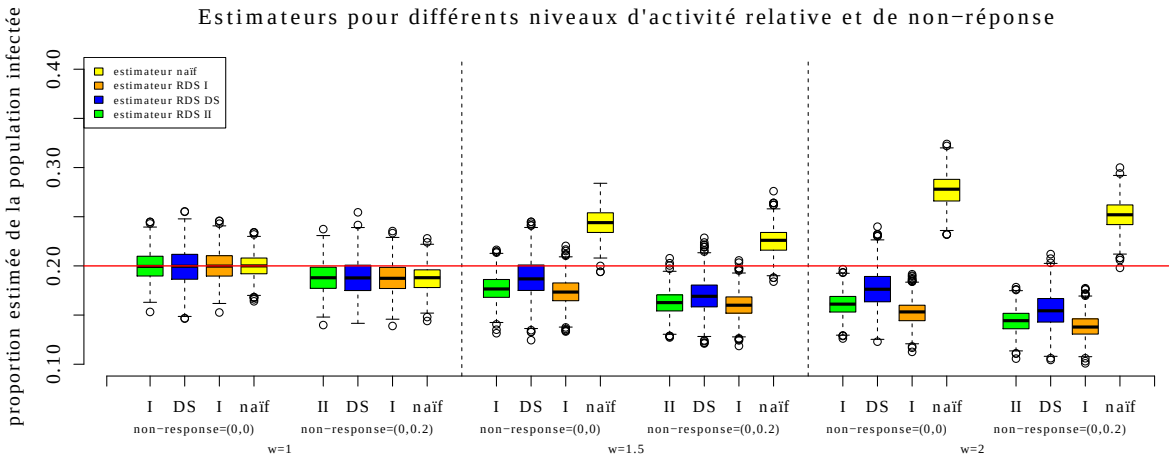


FIGURE 12 Boîtes à moustache des estimateurs pour plusieurs niveaux de non-réponse. Tous les autres paramètres sont choisis par défaut.

Le code développé et qui a permis de faire ces simulations offre de nombreuses autres possibilités. En effet, il est possible de faire varier presque tous les paramètres mentionnés dans les simulations. Celles qui sont montrées ici constituent un choix subjectif de ce que l'on veut illustrer. Toutefois, de nombreuses autres configurations sont possibles.

4 Conclusions

4.1 A l'épreuve de la réalité

L'étude ménée par (McCreesh *et al.*, 2012) a permis de tester la méthode RDS sur une population dont on connaissait les caractéristiques précises. Comme ce fait est relativement exceptionnel, on en fait un bref compte-rendu. En effet, à notre connaissance, il s'agit de la seule étude du genre. Elle a été menée sur une communauté de villages ruraux en Ouganda. Pour chaque individu, on a reporté les caractéristiques suivantes : âge, appartenance tribale, village de provenance, statut socio-économique et religion. Pour chacune de ces caractéristiques, on a estimé les proportions au sein de la population en utilisant les estimateurs $\mu_{\text{S-H}}$ et $\mu_{\text{V-H}}$ et on a calculé les intervalles de confiance à 95% en utilisant la méthode bootstrap de Salganik (Salganik, 2006). Le but de l'enquête est d'évaluer les estimateurs couramment utilisés. En fait, les deux principaux estimateurs utilisés, l'estimateur de Volz-Heckathorn et l'estimateur d'Heckathorn sont moins performants que la moyenne empirique dans plus de la moitié des cas (63% des cas pour l'estimateur $\mu_{\text{S-H}}$ et 67% pour l'estimateur $\mu_{\text{V-H}}$) et ne parviennent donc pas à corriger les différents biais inhérents à cette méthode d'échantillonage. Les intervalles de confiance proposés (Salganik, 2006) ont un taux de couverture nettement plus faible que prévu, entre 50 % et 74%, selon le type d'estimation, contre 95% théoriquement.

4.2 Evaluation de la méthode RDS pour l'OFS

Sachant que ni les simulations, ni même des conditions réelles ne donnent des résultats satisfaisants, comme le montre l'étude de McCreesh *et. al* (McCreesh *et al.*, 2012), on peut affirmer que les estimateurs couramment utilisés aujourd'hui ne sont pas satisfaisants du point de vue de la rigueur scientifique. En outre, comme cela a déjà été mentionné, ces estimateurs ne sont

Estimateur	Expression	Idée principale
estimateur naïf	$\mu_n = \frac{n_1}{n}$	Moyenne empirique de l'échantillon
estimateur de Salganik-Heckathorn	$\mu_{\text{S-H}} = \frac{\widehat{\widehat{D}}_1 \widehat{C}_{21}}{\widehat{\widehat{D}}_1 \widehat{C}_{21} + \widehat{\widehat{D}}_2 \widehat{C}_{12}}$	Utilisation des proportions de liens inter-groupes et des degrés des individus échantillonnés
estimateur d'Heckathorn	$\mu_{\text{H}} = \frac{\widehat{\widehat{AD}}_1 \widehat{C}_{21}}{\widehat{\widehat{AD}}_1 \widehat{C}_{21} + \widehat{\widehat{AD}}_2 \widehat{C}_{12}}$	Estimateur de Salganik-Heckathorn avec une modification de l'estimation du degré moyen
estimateur de Volz-Heckathorn	$\mu_{\text{V-H}} = \frac{\sum_{i \in s_1} 1/\hat{d}_i}{\sum_{i \in s} 1/\hat{d}_i}$	Probabilité d'inclusion d'individu proportionnelle au nombre de degrés
estimateur de Gile (ou successive sampling)	$\mu_{\text{S-S}} = \frac{\sum_{i \in s_1} 1/\hat{\pi}_i}{\sum_{i \in s} 1/\hat{\pi}_i}$	Modification de l'estimateur $\mu_{\text{V-H}}$, où la probabilité d'inclusion est estimée à l'aide d'une procédure itérative qui permet de s'affranchir de l'hypothèse de non-remise

TABLE 2 Résumé des principaux estimateurs RDS, dans l'ordre chronologique. Pour un échantillon obtenu via une procédure RDS, tous les estimateurs visent à estimer la proportion d'individus appartenant au sous-ensemble S_1 au sein d'une population S.

pas toujours consistants et non-biaisés, qui sont des propriétés essentielles pour un usage dans la statistique publique. Cette étude douche les espoirs des chercheurs qui croyaient détenir une méthode permettant enfin des inférences sur ces populations difficiles à atteindre. On peut encore ajouter que cette étude a été menée sur une population qui n'est ni difficile à atteindre ni sujette à stigmatisation. Or, on peut supposer que ces deux facteurs auraontient plutôt tendance à augmenter l'incertitude des estimations. On peut conclure que la méthode RDS, dans l'état actuel, est très utile pour pouvoir échantillonner des populations difficiles à atteindre mais ne parvient pas à fournir des estimateurs qui permettent de corriger le biais intrinsèque à la méthode d'échantillonnage. Il est donc honnête d'admettre que cette méthode ne satisfait pas aux critères de rigueur scientifique. Cependant, dans le cadre de la statistique officielle, les populations ne sont pas toujours difficiles à atteindre et on ne peut exclure que dans certains cas, les hypothèses soient satisfaites. Dès lors, le principal problème est l'estimation de la variance.

Annexes

A Eléments de théorie des processus stochastiques

La plupart de ce qui suit a pour base l'ouvrage (Karlin et Taylor, 1975) ou des notes de cours diverses [6]. On suppose que les notions de base de la théorie de la probabilité sont connues et on suppose que l'on se donne un espace de probabilité $(\Omega, \mathcal{F}, P)$.

Définition 2 (Processus stochastique)
Soit S un espace métrique muni de la tribu borélienne $\mathcal{B}(S)$ (c'est-à-dire la σ-algèbre engendrée par les ouverts de S, au sens de la métrique de S). Soit T un ensemble arbitraire. On appelle processus stochastique une famille de variables aléatoires $\{X_t\}_{t\in T}$ définies sur un même espace de probabilité $(\Omega, \mathcal{F}, P)$, indexée par $t \in T$ et à valeurs dans S. Dans ce qui suit, on va toujours utiliser $S \subset \mathbb{R}$ et $T \subset \mathbb{N}$. De plus, on va supposer que S est dénombrable. Pour plus de précisions, voir (Karlin et Taylor, 1975, p. 21-22).

Définition 3 (Chaîne de Markov)
Soient $X_1, X_2, \ldots,$ une suite de variables aléatoires définies sur $(\mathcal{F}, \Omega, \mathbb{P})$ et à valeurs dans $\mathbb{R}$. On dit que $\{X_n\}_{n\in\mathbb{N}}$ est une chaîne de Markov du premier ordre si la propriété suivante

$$P(X_{n+1} = x|X_0, \ldots, X_n) = P(X_{n+1} = x|X_n) \tag{11}$$

est satisfaite. On ne considère donc que les chaînes de Markov à temps discret. De plus le nombre de valeurs que peuvent prendre les variables aléatoires $X_1, X_2, \ldots$ est supposé fini, de cardinalité M et sera noté S dans ce qui suit. On appelle S l'espace des états, et un élément $i \in S$ un état. Plus généralement, on appelle *chaîne de Markov* un processus de Markov ayant un espace d'état fini ou au moins dénombrable. Pour plus de précisions, voir (Karlin et Taylor, 1975, p. 28).

Définition 4 (Probabilité de transition d'une chaîne de Markov)
On appellera probabilité de transition à un pas la valeur $P(n, i, j) = P(X_{n+1} = i|X_n = j)$. Si cette valeur ne dépend pas de n, ce qui correspond au cas que nous traiterons, alors on l'appellera probabilité de transition et on dira que les probabilités de transition sont stationnaires. Ce type de chaîne de Markov est dite homogène en temps. Si c'est le cas, on peut former la matrice des probabilités de transition P, de taille $M \times M$ et dont les coefficients sont donnés par $P_{ij} = P(X_1 = i|X_0 = j)$. On pose encore $P^n_{ij} = P(X_0 = i|X_n = j)$. On peut alors (pour les détails, voir (Karlin et Taylor, 1975, p.58-59)) écrire que :

$$P^n_{ij} = (P^n)_{ij},$$

où P^n est la matrice P élevée à la puissance n. On remarque donc que la chaîne de Markov peut être étudiée en utilisant uniquement la distribution initiale ainsi que les probabilités de transition.

Définition 5 (Distribution stationnaire d'une chaîne de Markov)
La distribution stationnaire d'une chaîne de Markov est la distribution qui satisfait l'équation :

$$\pi_i = \sum_{j\in S} \pi_j P_{ji}.$$

6. En particulier les notes du cours « Applied Stochastic Processes » donné par le Professeur Mountford, durant l'année 2010 à l'EPFL et dactylographié par les bons soins de Yoav Zemel, que je remercie au passage.

On remarque que cette équation peut-être écrite de manière plus compacte sous forme matricielle :

$$\boldsymbol{\pi} = \boldsymbol{\pi} P. \tag{12}$$

Il s'agit en fait d'un vecteur propre « à gauche » pour la valeur propre 1. On peut réécrire cette équation sous la forme :

$$P^T \boldsymbol{\pi}^T = \boldsymbol{\pi}^T,$$

afin de voir apparaître le problème sous la forme d'un système aux valeurs propres traditionnel. De plus, on sait que

$$\pi_i = \lim_{n\to\infty} P_{ii}^n = \lim_{n\to\infty} P_{ji}^n,$$

où $P_{ij}^n = P(X_0 = i, X_n = j)$. Donc, on peut interpréter π_i comme la probabilité que la chaîne se trouve dans l'état i, lorsque qu'un grand nombre de transitions se sont passées. On remarque donc que lorsque n devient grand, la dépendance en X_0 disparait, comme on peut le voir dans la seconde égalité ci-dessus. On peut dire que la chaîne de Markov a tendance à être indépendante de son point de départ, et donc de la distribution de X_0. Finalement, on peut se référer à (Karlin et Taylor, 1975, thm 1.2 et 1.3, p.83-84) en ce qui concerne ces résultats. De plus, cet ouvrage peut servir de référence dans le domaine des processus stochastiques, tout du moins pour une introduction.

Théorème 6 (voir (Karlin et Taylor, 1975))
Si une chaîne de Markov est irréductible, positive récurrente et apériodique, alors pour tout $i \in S$ (S désigne l'espace des états), on a :

$$\lim_{n\to\infty} P(X_n = i) = \lim_{n\to\infty} P_{ji}^n = \pi_i, \forall j \in S,$$

où π_i est l'unique distribution stationnaire de la chaîne de Markov.

A Simulations supplémentaires

On présente les simulations supplémentaires. Premièrement, les mêmes simulations que celles représentées sur les figures 5 et 6 sur des réseaux différents. Concrètement, on a utilisé les mêmes paramètres de processus et on a utilisé un réseau avec une homophilie de groupe nulle et aucune activité relative, c'est-à-dire $w = 1$. On peut remarquer que, dans ces conditions idéales, c'est-à-dire celles qui sont faites dans (Heckathorn, 1997), les estimateurs se comportent extrêmement bien, tout comme l'estimateur naïf.
Ensuite les figures 16 et 13 représentent les estimateurs en fonction de l'activité relative et de la proportion échantillonnée. La figure 16 montre les estimateurs qui ne figurent pas sur la figure 9. Quant à la figure 13, il s'agit d'un résumé des figures 9 et 16. Toutefois, on l'a mis en annexe car elle est peu lisible. Elle a toutefois l'avantage de montrer la totalité des données sur un seul graphique.

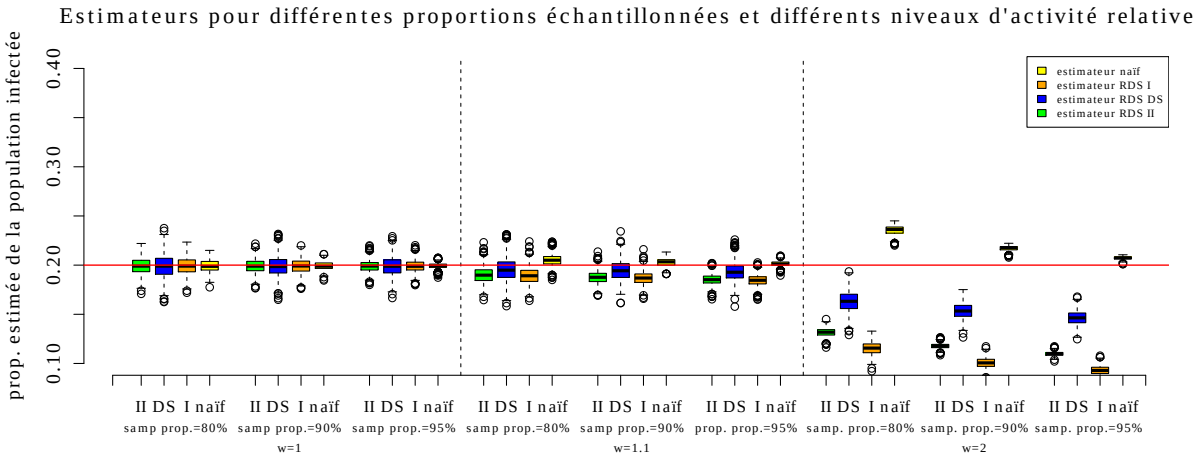
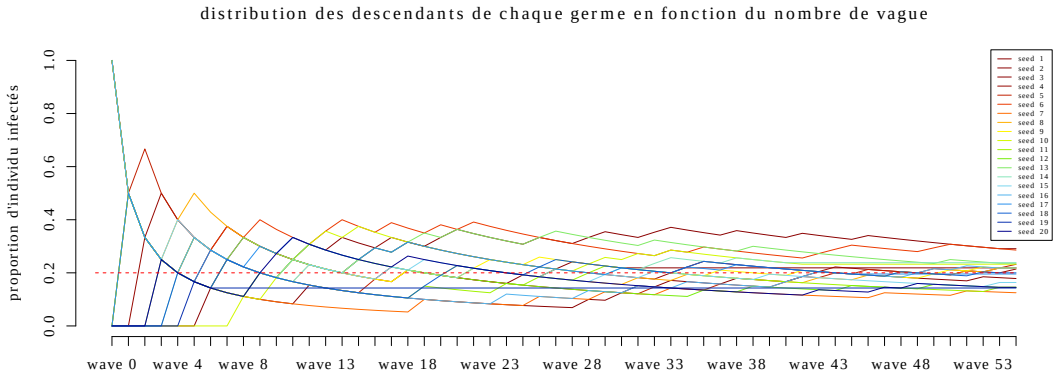
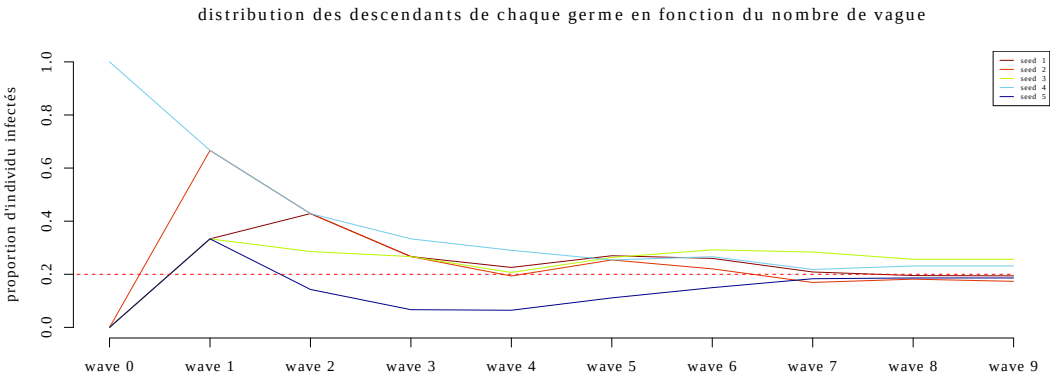


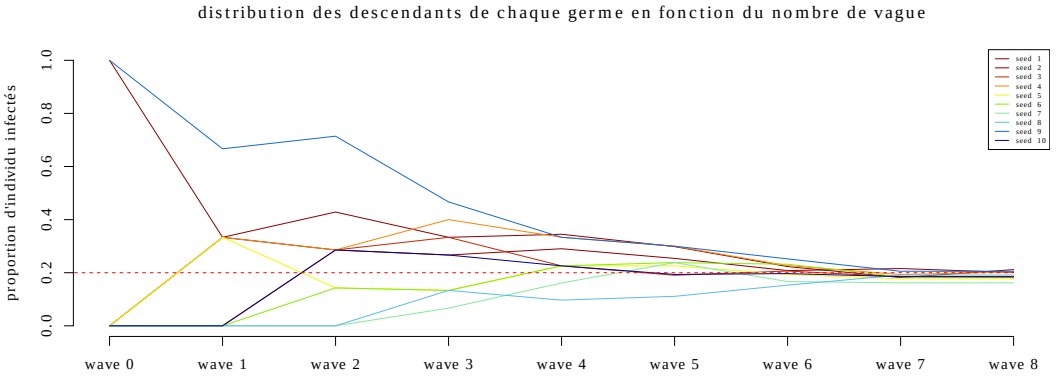
FIGURE 13 Estimateurs en fonction de l'activité relative et du taux de sondage.



(a) $n = 1000$ avec 1 coupons et 20 germes

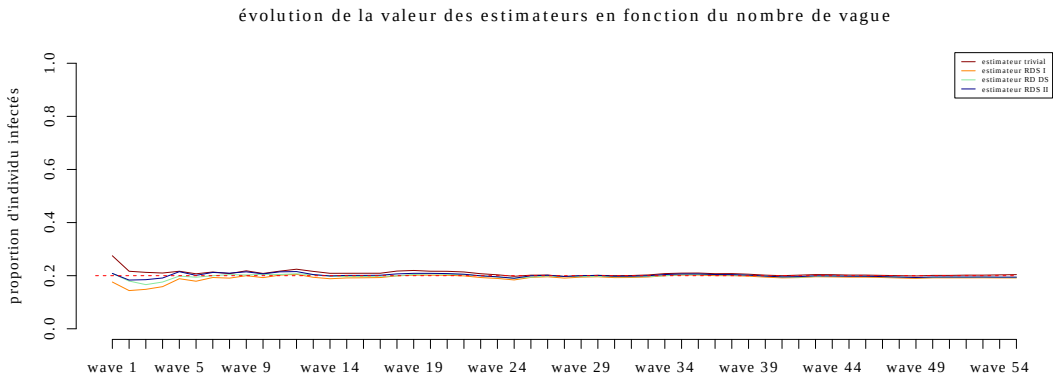


(b) $n = 3000$ avec 2 coupons et 5 germes

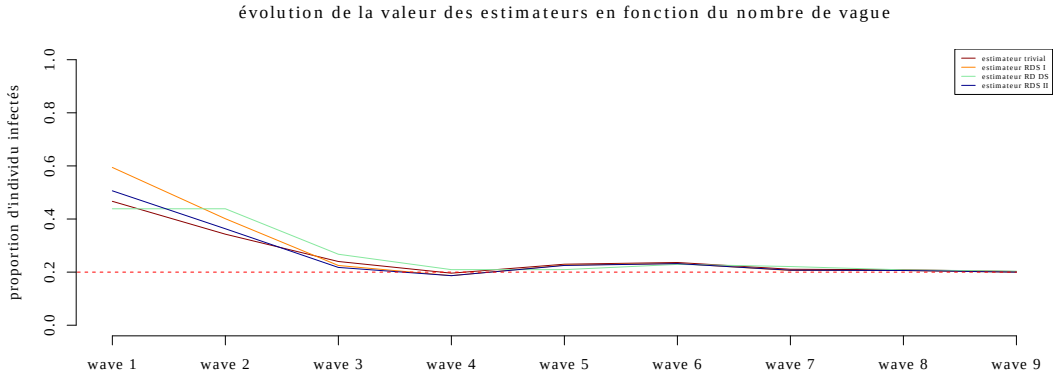


(c) $n = 3000$ avec 2 coupons et 10 germes

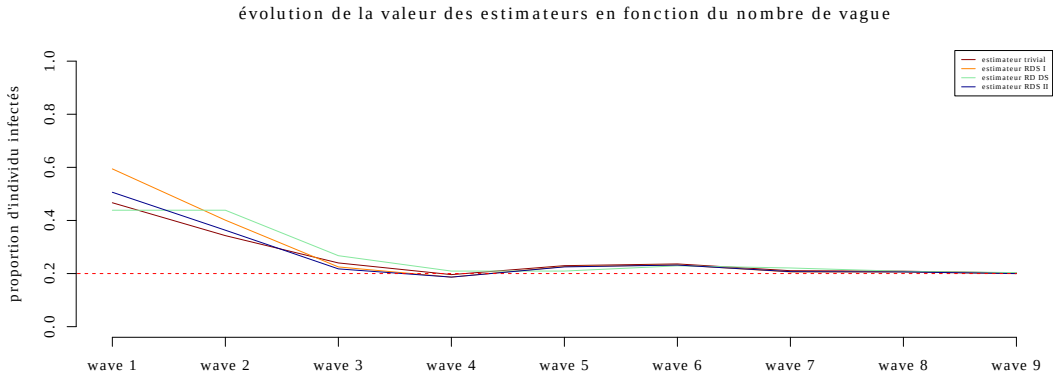
FIGURE 14 Distributions des descendants des germes en fonction du nombre de vagues. Le réseau simulé avait toujours les mêmes caractéristiques : le paramètre `group.homophily` était tel que l'homophilie de réseau était nulle, l'activité relative w était égale à 1 et il y avait 4000 individus dans la population totale. Il n' y a pas d'homophilie. Toutes les autres valeurs du processus et du réseau étaient celles par défaut, sauf précision contraire.



(a) $n = 1000$ avec 1 coupons et 20 germes

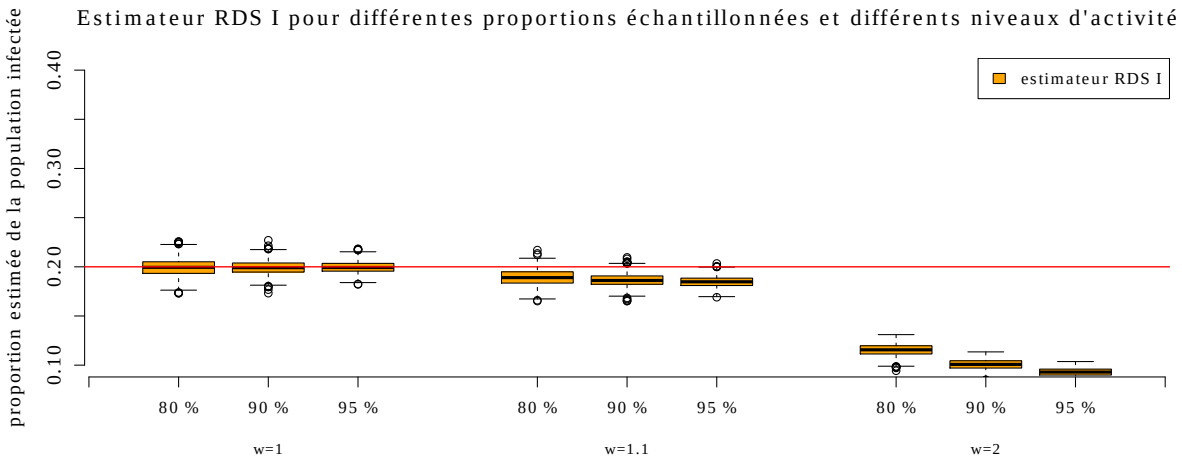


(b) $n = 3000$ avec 2 coupons et 5 germes

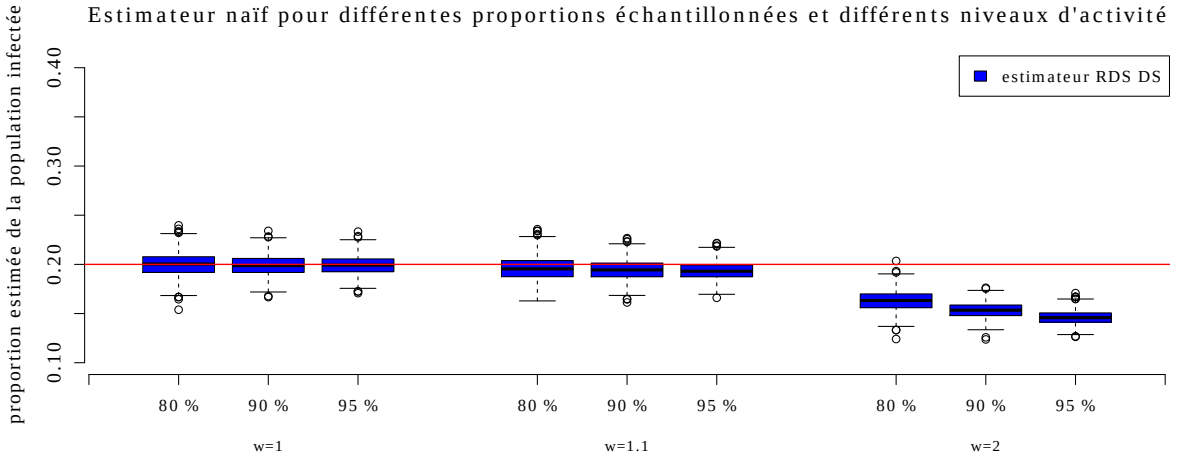


(c) $n = 3000$ avec 2 coupons et 10 germes

FIGURE 15 Evolution des estimateurs au cours du temps. Le réseau simulé ainsi que les échantillons sont les mêmes que ceux de la figure 14.



(a) Estimateur RDS I



(b) Estimateur RDS DS

FIGURE 16 Estimateurs RDS I et RDS DS (Salganik-Heckathorn et Heckathorn) en fonction de l'activité relative et du taux de sondage.

Références

William G. COCHRAN : *Sampling processes*. John Wiley & Sons, 1977.

Reinhard DIESTEL : *Graph Theory*. Springer, fourth édition, 2010.

Rick DURETT : *Probability : Theory and Examples*. Cambridge University Press, fourth édition, 2010.

Thomas J. FARARO et John SKVORETZ : Biaised networks and social structure theorems : Part II. *Social Networks*, 6, 1984.

Krista J. GILE : Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. *Journal of the American Statistical Association*, 106, 2012.

Krista J. GILE et Mark S. HANDCOCK : Respondent-driven sampling : An assesment of current methodology. *Sociological Methodology*, 40, 2010.

Krista J. GILE et Mark S. HANDCOCK : Network model-assisted inference from respondent-driven sampling data. 2011.

Krista J. GILE et Mark S. HANDCOCK : On the concept of snowball sampling. *Sociological Methodology*, 2012.

Krista J. GILE et Amber TOMAS : The effect of non-response and non-recruitment on estimators for respondent-driven sampling. *Electronic Journal of Statistics*, 5, 2011.

Leo A. GOODMAN : Snowball sampling. *Annals of Mathematical Statistics*, 32, 1961.

Mark S. HANDCOCK, David R. HUNTER, Carter T. BUTTS, Steven M. GOODREAU, Pavel N. KRIVITSKY et Martina MORRIS : *statnet : Software Tools for the Statistical Modeling of Network Data*. Seattle, WA, 2003. URL http://CRAN.R-project.org/package=statnet. Version 3.0-1 . Project home page at http ://statnet.org.

Morris H. HANSEN et William N. HURWITZ : On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 1943.

Douglas D. HECKATHORN : Respondent-driven sampling : A new approach to the study of hidden populations. *Social Problems*, 44, 1997.

Douglas D. HECKATHORN : Respondent-driven sampling II : Deriving valid population estimates from chain-refferal samples of hidden populations. *Social Problems*, 49, 2002.

Douglas D. HECKATHORN : Extension of respondent-driven sampling : Analyzing continuous variables and controlling for differential recruitment. *Sociological Methodology*, 37, 2007.

Robert HEIMER : Critical issues and further question about respondent-driven sampling : Comment on Ramirez-Valles, *et al.* (2005). *AIDS and Behavior*, 9, 2005.

Lisa JOHNSTON, Mohsen MALEKINEJAD, Carl KENDALL, Irene IUPPA et George RUTHERFORD : Implementation challenges to using respondent-driven sampling methodology for HIV biological and behavioral surveillance : Field experiences in international settings. *AIDS and Behavior*, 12, 2008.

Samuel KARLIN et Howard M. TAYLOR : *A First Course in Stochastic Processes*. Academic Press, second édition, 1975.

Xin LU, Linus BENGTSSON, Tom BRITTON, Martin CAMITZ, Beom Jun KIM, Anna THORSON et Fredrik LIJEROS : The sensitivity of respondent-driven sampling method. *Journal of the Royal Statistical Society : Series A*, 175, 2012.

Nicky MCCREESH, Simon D.W. FROST, Janet SEELEY, Joseph KATONGOLE, Matilda N. TARSH, Richard NDUNGUSE, Fatima JICHI, Natasha L. LUNEL, Dermot MAHER, Lisa G. JOHNSTON, Pam SONNENBERG, Andrew J. COPAS, Richard J. HAYES et Richard G. WHITE : Evaluation of respondent-driven sampling. *Epidemiology*, 23, 2012.

Mark J.E NEWMAN : The structure and function of complex networks. *SIAM Review*, 45, 2003.

Mark J.E NEWMAN, Steven H. STROGATZ et Duncan J. WATTS : Random graphs with arbitrary degree distribution and their applications. *Physical Review E*, 6402, 2001.

R CORE TEAM : *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL http://www.R-project.org/. ISBN 3-900051-07-0.

Jesus RAMIREZ-VALLES, Douglas D. HECKATHORN, Raquel V ?SQUEZ, Rafael M. DIAZ et Richard T. CAMPBELL : From networks to populations : The development and application of respondent-driven sampling among IDU's and latino gay men. *AIDS and Behavior*, 9, 2005.

Matthew J. SALGANIK : Variance estimation, design effects, and sample size calculations for respondent-driven sampling. *Journal of Urban Health*, 83, 2006.

Matthew J. SALGANIK : Respondent-driven sampling in the real world. *Epidemiology*, 23, 2012.

Matthew J. SALGANIK et Douglas D. HECKATHORN : Sampling and estimations in hidden population using respondent-driven sampling. *Sociological Methodology*, 34, 2004.

Carl-Erik SÄRNDAL, Bengt SWENSSON et Jan WRETMAN : *Model Assited Survey Sampling*. Springer, 1992.

Yves TILLÉ : *Théorie des sondages, échantillonage et estimation en populations finies*. Dunod, 2001.

Erik VOLZ et Douglas D. HECKATHORN : Probability based estimation theory for respondent-driven sampling. *Journal of Official Statistics*, 24, 2008.

Erik VOLZ, Cyprian WEJNERT, I. DEGANI et Douglas D. HECKATHORN : *Respondent-Driven Sampling Analysis Tool, version 6.0.1*. Cornell University, Ithaca, NY, 2007.

Matthieu WILHELM : *Codes pour la simulation de processus RDS : user's guide*. OFS, Neuchâtel, 2012.

Methodenberichte der Sektion Statistische Methoden des BFS
Rapports de méthodes de la section méthodes statistiques de l'OFS
Methodology reports published by the FSO's Statistical Methods Section

Wilhelm, M. (2014). Echantillonnage boule de neige. La méthode de sondage déterminé par les répondants. Numéro de commande : 338-0071

Assoulin, D. (2013). Wertschöpfungsstatistik. Revision 2009 : Stichprobenrahmen und Stichprobenplan. Bestellnummer : 338-0070

Ferster, M. (2013). EVS I - Energieverbrauchsstatistik 2002 bis 2007 : Stichprobenplan und Hochrechnung. Bestellnummer : 338-0069

Assoulin, D. (2013). Zusatzerhebung für die landwirtschaftliche Betriebszählung 2010 : Stichprobenplan und Hochrechnung. Bestellnummer : 338-0068

Potterat, J. (2012). Use of conversion keys for NOGA2002-2008. Order number : 338-0067-05

Andrade, B., Salamin P.-A. (2012). Enquête sur la situation sociale et économique des étudiant-e-s des hautes écoles suisses 2009. Cadre de sondage, plan d'échantillonnage et méthodes d'estimation. Numéro de commande : 338-0066

Potterat, J., Panchard, C., Kilchmann, D. (2012). Umweltschutzausgaben der Unternehmen 2009 (UWSA2009). Stichprobenplan, Einsetzungen, Gewichtung und Schätzverfahren. Bestellnummer : 338-0065

Potterat, J. (2012). Benutzung der Umsteigeschlüssel NOGA 2002-2008. Bestellnummer : 338-0064

Potterat, J. (2011). Kosten und Nutzen der Berufsbildung aus Sicht der Betriebe im Jahr 2009 (KNBB09). Stichprobenplan, Gewichtung und Schätzverfahren. Bestellnummer : 338-0063

Kilchmann, D., Potterat, J., Genoud, S. (2011). Gütertransporterhebung 2008. Stichprobenplan, Datenaufbereitung, Gewichtung und Schätzverfahren. Bestellnummer : 338-0062

Graf, E. (2010). Enquête suisse sur la santé 2007. Plan d'échantillonnage, pondérations et analyses pondérées des données. Numéro de commande : 338-0061

Eichenberger P., Hulliger B., Potterat J. (2010). Describing the Anticipated Accuracy of the Swiss Population Survey. Order number : 338-0060

Graf, E. (2010). Étude empirique de l'attrition du Panel Suisse de Ménages : vers une caractérisation du profil du non-répondant. Numéro de commande : 338-0059

Graf, E. (2009). Weightings of the Swiss Household Panel : SHP_I wave 9, SHP_II wave 4, SHP_I et SHP_II combined. Order number : 338-0058

Graf, E. (2009). Pondérations du Panel Suisse de Ménages : PSM_I vague 9, PSM_II vague 4, PSM_I et PSM_II combinés. Numéro de commande : 338-0057-05

Qualité, L., Tillé, Y. (2009). Estimation de la précision d'évolutions dans l'enquête sur la valeur ajoutée. Numéro de commande : 338-0056

Renaud, A., Panchard, C. et Potterat, J. (2008). Statistique de l'emploi. Révision 2007 : méthodes d'estimation. Numéro de commande : 338-0055

Graf, E. (2008). Pondérations du PSM. PSM_I vague 8, PSM_II vague 3, PSM_I et PSM_II combinés. Numéro de commande : 338-0054

Andrade, B., Graf, M. (2008). Enquête suisse sur la structure des salaires 2006. Aspects méthodologiques du modèle des salaires "Salarium". Numéro de commande : 338-0053

Renaud, A. (2008). Statistique de l'emploi. Révision 2007 : cadre de sondage et échantillonage. Numéro de commande : 338-0052

Graf, E. (2008). Pondérations du SILC pilote. SILC_I vague 2, SILC_II vague 1, SILC_I et SILC_II combinés. Numéro de commande : 338-0051

Kilchmann, D. (2008). Statistik der sozialmedizinischen Institutionen 1999-2004 und Krankenhausstatistik 1999-2002. Einsetzungen für fehlende Daten. Bestellnummer : 338-0050

Renaud, A. (2008). Technologies de l'information et de la communication. Estimations sur la base de la statistique de la valeur ajoutée. Numéro de commande : 338-0049

Assoulin, D. (2007). Wertschöpfungsstatistik. Einsetzungsversuche für fehlende Antworten grosser Unternehmen. Bestellnummer : 338-0048

Kilchmann, D. (2007). Beherbergungsstatistik Campingplätze. Stichprobenrahmen und Schätzverfahren 2005/06. Bestellnummer : 338-0047

Gabler, S., Häder, S. (2007). Haushalts- und Personenerhebungen. Machbarkeit von Random Digit Dialing in der Schweiz. Bestellnummer : 338-0046

Ferrez, J., Graf, M. (2007). Enquête suisse sur la structure des salaires. Programmes R pour l'intervalle de confiance de la médiane. Numéro de commande : 338-0045

Renaud, A. (2007). Harmonisation de la scolarité obligatoire en Suisse (HarmoS). Design général de l'enquête et échantillon des écoles. Numéro de commande : 338-0044

Potterat, J. (2007). Betriebszählung 2005. Statistische Methoden zur Schätzung der provisorischen Ergebnisse. Bestellnummer : 338-0043

Hulliger, B. (2006). Umweltschutzausgaben der Unternehmen 2003, Stichprobenplan, Datenaufbereitung und Schätzverfahren. Bestellnummer : 338-0042

Renfer, J.-P. (2006). Enquête sur les chiffres d'affaires du commerce de détail. Plan d'échantillonnage et méthodes d'estimation. Numéro de commande : 338-0041

Salamin, P.-A. (2006). Statistique de l'aide sociale dans le domaine de l'asile. Plan de sondage et extrapolations pour l'enquête pilote 2005. Numéro de commande : 338-0040

Renaud, A. (2006). Statistique suisse des bénéficiaires de l'aide sociale. Pondération des communes 2004. Numéro de commande : 338-0039

Graf, M. (2006). Swiss Earnings Structure Survey 2002-2004. Compositional data in a stratified two-stage sample : Analysis and precision assessment of wage components. Order number : 338-0038

Potterat, J. (2006). Pensionskassenstatistik 2004. Statistische Methoden zur Schätzung der provisorischen Ergebnisse. Bestellnummer : 338-0037

Potterat, J. (2006). Kosten und Nutzen der Berufsbildung aus Sicht der Betriebe im Jahr 2004. Stichprobenplan, Gewichtung und Schätzverfahren. Bestellnummer : 338-0036

Kilchmann, D. (2006). Vierteljährliche Wohnbaustatistik. Stichprobenplan, statistische Datenaufarbeitung und Schätzverfahren 2005. Bestellnummer : 338-0035

Kilchmann, D. (2006). Erhebung über Forschung und Entwicklung in der schweizerischen Privatwirtschaft 2004. Bereinigung der Stichprobe, Ersatz fehlender Werte und Schätzverfahren. Bestellnummer : 338-0034

Kilchmann, D., Eichenberger, P., Potterat, J. (2005). Volkszählung 2000. Statistische Einsetzungsverfahren Band 2. Bestellnummer : 338-0033

Kilchmann, D., Eichenberger, P., Potterat, J. (2005). Volkszählung 2000. Statistische Einsetzungsverfahren Band 1. Bestellnummer : 338-0032

Graf, M., Matei, A. (2005). Enquête suisse sur la structure des salaires 2002. La précision du salaire brut standardisé médian. Numéro de commande : 338-0031

Graf, E., Renfer, J.-P. (2005). Enquête suisse sur la santé 2002. Plan d'échantillonnage, pondération et estimation de la précision. Numéro de commande : 338-0030

Potterat, J. (2005). Mietpreis-Strukturerhebung 2003. Gewichtung und Schätzverfahren. Bestellnummer : 338-0029

Potterat, J. (2005). Landwirtschaftliche Betriebszählung 2003. Schätzverfahren für die Zusatzerhebung. Bestellnummer : 338-0028

Renaud, A. (2004). Coverage estimation for the Swiss population census 2000. Estimation methodology and results. Order number : 338-0027

Kilchmann, D. (2004). Revision des Schweizerischen Lohnindex. Schätzmethoden der Lohnindices und deren Varianzschätzer. Bestellnummer : 338-0026

Graf, M. (2004). Enquête suisse sur la structure des salaires 2002. Plan d'échantillonnage et extrapolation pour le secteur privé. Numéro de commande : 338-0025

Renaud, A. (2004). Analyse de données d'enquêtes. Quelques méthodes et illustration avec des données de l'OFS. Numéro de commande 338-0024

Renaud, A., Potterat, J. (2004). Estimation de la couverture du recensement de la population de l'an 2000. Echantillon pour l'estimation de la sous-couverture (P-sample) et qualité du cadre de sondage des bâtiments. Numéro de commande : 338-0023

Graf, M. (2004). Fusion de données. Etude de faisabilité. Numéro de commande : 338-0022

Potterat, J. (2003). Mietpreis-Strukturerhebung 2003. Entwicklung des Stichprobenplans und Ziehung der Stichprobe. Bestellnummer : 338-0021

Potterat, J. (2003). Landwirtschaftliche Betriebszählung 2003. Stichprobenplan der Zusatzerhebung. Bestellnummer : 338-0020.

Renaud, A. (2003). Estimation de la couverture du recensement de la population de l'an 2000. Echantillon pour l'estimation de la sur-couverture (E-sample). Numéro de commande : 338-0019

Hulliger, B. (2003). Bereinigung der Stichprobe, Ersatz fehlender Werte und Schätzverfahren. Erhebung über F+E in der schweizerischen Privatwirtschaft 2000. Bestellnummer : 338-0018

Renfer, J.-P. (2003). Enquête 2000 sur la recherche et le développement dans l'économie privée en Suisse. Plan d'échantillonnage. Numéro de commande : 338-0017

Potterat, J. (2003). Kosten und Nutzen der Berufsbildung aus Sicht der Betriebe. Schätzverfahren. Bestellnummer : 338-0016

Graf, M., Matei, A. (2003). Stratégie de choix des modèles de désaisonnalisation. Application aux séries de l'emploi total. Numéro de commande : 338-0015

Potterat, J., Salamin, P.A. (2002). Betriebszählung 2001. Methoden für die Datenbereinigung. Bestellnummer : 338-0014

Renaud, A. (2002). Programme international pour le suivi des acquis des é lèves (PISA). Plans d'échantillonnage pour PISA 2000 en Suisse. Numéro de commande : 338-0013

Renfer, J.-P. (2002). Enquête 2001 sur les coûts et l'utilité de la formation des apprentis du point de vue des établissements. Plan d' échantillonnage. Numéro de commande : 338-0012

Potterat, J., Salamin, P.A. (2002). Betriebszählung 2001. Stichprobenplan und Schätzverfahren für die provisorischen Ergebnisse. Bestellnummer : 338-0011

Graf, M. (2002). Enquête suisse sur la structure des salaires 2000. Plan d'échantillonnage, pondération et méthode d'estimation pour le secteur privé. Numéro de commande : 338-0010

Renaud, A., Eichenberger P. (2002). Estimation de la couverture du recensement de la population de l'an 2000. Procédure d'enquête et plan d'échantillonnage de l'enquête de couverture. Numéro de commande : 338-0009

Kilchmann, D., Hulliger, B. (2002). Stichprobenplan für die Obstbaumzählung 2001. Bestellnummer : 338-0008

Graf, M. (2002). Passage du concept établissement au concept entreprise. Numéro de commande : 338-0007

Salamin, P.A. (2001). La technique de la double enquête pour la statistique du transport routier de marchandise. Numéro de commande : 338-0006

Peters, R., Renfer, J.-P. et Hulliger, B. (2001). Statistique de la valeur ajoutée 1997-1998. Procédure d'extrapolation des données. Numéro de commande : 338-0005

Potterat, J., Hulliger, B. (2001). Schätzung der Sägereiproduktion mit der Sägerei-Erhebung PAUL. Bestellnummer : 338-0004

Graf, M. (2001). Désaisonnalisation. Aspects méthodologiques et application à la statistique de l'emploi. Numéro de commande : 338-0003

Hüsler, J., Müller, S. (2001). Schlussbericht Betriebszählung 1995 (BZ 95), Mehrfach imputierte Umsatzzahlen. Bestellnummer : 338-0002

Renaud, A. (2001). Statistique suisse des bénéficiaires de l'aide sociale. Plan d'échantillonnage des communes. Numéro de commande : 338-0001

Hulliger, B., Eichenberger, P. (2000). Stichprobenregister für Haushalterhebungen : Umstellung auf Telefonnummern ohne Namen und Adressen, Abläufe für Erstellung und Stichprobenziehung. Bestellnummer : 338-0000

Programme des publications de l'OFS

En sa qualité de service central de statistique de la Confédération, l'Office fédéral de la statistique (OFS) a pour tâche de rendre les informations statistiques accessibles à un large public.

L'information statistique est diffusée par domaine (cf. verso de la première page de couverture); elle emprunte diverses voies:

Moyen de diffusion	*Contact*
Service de renseignements individuels	032 713 60 11 info@bfs.admin.ch
L'OFS sur Internet	www.statistique.admin.ch
Communiqués de presse: information rapide concernant les résultats les plus récents	www.news-stat.admin.ch
Publications: information approfondie	032 713 60 60 order@bfs.admin.ch
Données interactives (banques de données, accessibles en ligne)	www.stattab.bfs.admin.ch

Informations sur les divers moyens de diffusion sur Internet à l'adresse www.statistique.admin.ch → Services → Les publications de Statistique suisse

Rapports de méthodes de la section méthodes statistiques

Les rapports de méthodes décrivent les méthodes mathématiques et statistiques à la base des résultats et des analyses de la statistique publique. Ils présentent également l'évaluation et le développement de nouvelles méthodes en vue d'une application future. Ces publications visent d'une part à documenter les méthodes utilisées ou envisagées dans un souci de transparence et de rigueur scientifique, et d'autre part à favoriser la collaboration avec le monde scientifique et universitaire.

Les résultats numériques présentés dans les rapports de méthodes illustrent les concepts mathématiques décrits, mais ne sont pas des résultats officiels des enquêtes concernées. De même, les méthodes réellement appliquées peuvent différer légèrement de celles décrites dans ces rapports.

Les rapports de méthodes sont disponibles sous forme électronique sur le site internet de l'OFS.

L'objectif de ce rapport est de présenter la méthode d'échantillonnage déterminée par les répondants, plus connue sous son acronyme anglais, RDS (Respondent-Driven Sampling). Après une très brève introduction à la théorie des sondages, la méthode RDS est introduite dans son contexte historique. On présente ensuite les différents estimateurs utilisés dans la plupart des études existantes. Le développement mathématique de ces estimateurs ainsi que les hypothèses nécessaires à leur dérivation sont passés en revue. On fait un survol théorique des estimateurs couramment utilisés, de leurs propriétés et de leurs faiblesses. On fait de plus des simulations afin d'illustrer certaines situations. On conclut à un préavis négatif pour l'utilisation de cette méthode, du moins dans l'état actuel de la recherche. On conseille donc d'y renoncer à ce stade dans les enquêtes devant satisfaire aux exigences de la statistique publique.

N° de commande
338-0071

Commandes
Tél.: 032 713 60 60
Fax: 032 713 60 61
E-mail: order@bfs.admin.ch

Prix
gratuit

ISBN 978-3-303-00515-6