Composante du produit nº 11-522-X au catalogue de Statistique Canada
La série des symposiums internationaux de Statistique Canada : recueil

Article

Symposium 2008 :
Collecte des données : défis, réalisations et nouvelles orientations

Peut-on établir des statistiques officielles à partir d'enquêtes en ligne reposant sur le principe de l'autosélection?

par Jelke Bethlehem

2009

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2008 de Statistique Canada
Collecte des données : défis, réalisations et nouvelles orientations

Peut-on établir des statistiques officielles à partir d'enquêtes en ligne reposant sur le principe de l'autosélection?

Jelke Bethlehem[1]

Résumé

À première vue, les enquêtes en ligne semblent constituer une façon intéressante et attrayante de recueillir des données. Elles permettent d'avoir un accès simple, économique et rapide à un grand nombre de personnes. Il y a cependant un revers à cette médaille. Compte tenu des problèmes d'ordre méthodologique qu'elles posent, les enquêtes en ligne peuvent produire des résultats fortement biaisés, surtout si elles ont recours à la méthode d'autosélection des répondants plutôt qu'à l'échantillonnage probabiliste, comme cela devrait être le cas. Le sous-dénombrement constitue également un grave problème. On peut alors se demander si les enquêtes en ligne conviennent à la collecte des données pour les besoins de la statistique officielle. La présente communication porte sur les problèmes du sous-dénombrement et de l'autosélection dans les enquêtes en ligne et tente de montrer comment on peut intégrer la collecte des données par Internet aux pratiques courantes de collecte des données pour les besoins de la statistique officielle.

Mots clés : Enquête en ligne, autosélection, sous-dénombrement, échantillonnage probabiliste, enquête à mode de collecte mixte.

1. Introduction

Au cours des dernières décennies, le domaine de la recherche par enquête a subi des transformations radicales. Il y a eu d'abord le passage de la traditionnelle interview papier et crayon à l'interview assistée par ordinateur. Et aujourd'hui, notamment dans le secteur des études de marché, l'enquête en ligne remplace de plus en plus l'interview sur place et les enquêtes postales et téléphoniques. L'essor de la recherche en ligne n'a rien d'étonnant. Une enquête en ligne constitue une façon simple d'avoir accès à un grand nombre de personnes. On peut distribuer les questionnaires à très faible coût. On n'a pas besoin d'intervieweur et il n'y a pas de frais d'envoi postal ni d'impression. On peut lancer une enquête très rapidement. On perd peu de temps entre le moment où le questionnaire est prêt et le début du travail sur le terrain. Enfin, l'enquête en ligne offre de nouvelles possibilités attrayantes, comme l'utilisation du multimédia (son, images, animation et films).

À première vue, l'enquête en ligne semble avoir beaucoup en commun avec d'autres types d'enquête. Il s'agit simplement d'un autre mode de collecte des données : on ne pose pas de questions sur place ni par téléphone, mais par Internet. Ce qui diffère, toutefois, c'est que bon nombre d'enquêtes en ligne n'appliquent pas les principes de l'échantillonnage probabiliste. En prélevant des échantillons aléatoires, on peut appliquer la théorie des probabilités, ce qui permet de calculer des estimations sans biais et de déterminer l'exactitude des estimations. Depuis les années 1940, on applique avec succès le paradigme de l'échantillonnage probabiliste en statistique officielle et universitaire ainsi que dans le secteur des études de marché (mais dans une mesure nettement moindre). Les enquêtes en ligne ont souvent recours à la méthode d'autosélection des répondants plutôt qu'à l'échantillonnage probabiliste, ce qui risque de compromettre la qualité des résultats d'enquête, car on ne peut appliquer la théorie de l'échantillonnage probabiliste et les estimations sont souvent fortement biaisées.

Les organismes statistiques nationaux de nombreux pays doivent composer d'une part avec des contraintes budgétaires et, d'autre part, avec des demandes de plus en plus pressantes de renseignements détaillés. L'enquête en ligne peut-elle jouer un rôle dans la recherche d'une solution à ce dilemme? Dans la présente communication, nous soutenons que les enquêtes reposant sur le principe de l'autosélection ne peuvent jouer un tel rôle. Toutefois,

[1]Jelke Bethlehem, Statistics Netherlands, B.P. 24500, 2490 HA The Hague, The Netherlands (jbtm@cbs.nl)

lorsqu'on mène une enquête en ligne dans le cadre de l'échantillonnage probabiliste, on peut se servir de l'autosélection soit comme mode de collecte unique, soit comme un des modes utilisés dans une enquête à mode de collecte mixte.

2. Enquêtes en ligne

2.1 Échantillonnage sur le Web

Dans leur ouvrage-phare, Horvitz et Thompson (1952) ont jeté les bases de l'échantillonnage probabiliste tel qu'on l'applique aujourd'hui aux statistiques officielles. Selon eux, on peut toujours calculer des estimateurs sans biais des caractéristiques d'une population, pourvu qu'on prélève les échantillons au moyen de l'échantillonnage probabiliste et que chaque élément de la population possède une probabilité connue et strictement positive d'être sélectionné. Dans ces conditions, on peut en outre calculer les erreurs-types des estimations et, par conséquent, les intervalles de confiance. Il est donc possible d'établir l'exactitude des estimations. On peut aussi utiliser la méthode de Horvitz-Thompson dans les enquêtes dont le plan d'échantillonnage est complexe, comme l'échantillon aléatoire stratifié, l'échantillon en grappes ou l'échantillon à deux degrés.

Malheureusement, bon nombre d'enquêtes en ligne reposent en quelque sorte sur le principe de l'autosélection. On met simplement l'enquête sur le Web. La participation exige avant tout que les répondants connaissent l'existence de l'enquête. Il faut qu'ils visitent par hasard le site Web ou qu'ils donnent suite à un bandeau publicitaire, à un courriel ou à un autre message publicitaire. Deuxièmement, il faut qu'ils prennent la décision de remplir le questionnaire sur Internet. Le spécialiste de la recherche par enquête n'a aucun contrôle sur le processus de sélection.

Aux Pays-Bas, tous les grands sondages d'opinion sont menés auprès de panels en ligne établis selon le principe de l'autosélection. Les valeurs de certaines variables démographiques sont enregistrées à l'étape du recrutement. On peut donc comparer la répartition de ces variables dans le cadre d'une enquête à leur répartition au sein de la population. On peut employer des techniques de repondération pour tenter de corriger la surreprésentation ou la sous-représentation de certains groupes. À titre d'exemple d'une vaste enquête transversale en ligne menée aux Pays-Bas, citons *21minuten.nl*, enquête censée fournir des réponses à des questions concernant d'importants problèmes de la société néerlandaise. En 2006, en une période de six semaines, environ 170 000 personnes ont rempli les questionnaires en ligne. Une enquête semblable a été menée en Allemagne (*Perspektive Deutschland*). Une étude menée auprès de sociétés d'études de marché néerlandaises et portant sur 19 de leurs panels en ligne révèle que la plupart d'entre eux ont recours à l'autosélection; voir Vonk et coll. (2006).

On prétend parfois que les résultats d'enquêtes en ligne reposant sur le principe de l'autosélection sont « représentatifs » en raison du nombre élevé de répondants ou grâce aux méthodes avancées de repondération. Le terme « représentatif » est plutôt déroutant; voir Kruskall et Mosteller (1979a, 1979b et 1979c). Il peut revêtir plusieurs sens et on l'emploie souvent dans un sens très approximatif pour exprimer une vague notion de bonne qualité. On considère souvent un nombre élevé de répondants comme un gage de validité et de fiabilité. Toutefois, il y a nettement lieu de se demander si la grande taille de l'échantillon est aussi significative selon qu'elle a été obtenue par autosélection des répondants ou par échantillonnage probabiliste.

À l'heure actuelle, bon nombre d'enquêtes en ligne font face à deux problèmes méthodologiques fondamentaux. Nous avons déjà mentionné le premier : l'autosélection. Les chercheurs n'ont aucun contrôle sur le mécanisme de sélection. Les probabilités de sélection sont inconnues. On ne peut donc pas calculer d'estimations sans biais, ni établir l'exactitude des estimations. Le deuxième problème est celui du sous-dénombrement. Comme on recueille les données par Internet, les personnes sans accès à Internet ne peuvent jamais participer à une enquête en ligne. Par conséquent, les résultats de la recherche ne peuvent s'appliquer qu'à la « population Internet » et non à l'ensemble de la population. Dans les sections qui suivent, nous analysons ces deux problèmes de manière plus détaillée.

2.2 Sous-dénombrement

Les enquêtes en ligne sont confrontées au sous-dénombrement puisque la population cible est habituellement beaucoup plus vaste que la population Internet. En 2007, selon les données d'Eurostat (le bureau statistique de l'Union européenne), 54 % des ménages de l'UE avaient accès à Internet. On observe de grandes variations entre les pays. Les plus forts pourcentages d'accès à Internet ont été constatés aux Pays-Bas (83 %), en Suède (79 %) et au Danemark (78 %), et les taux les plus faibles, en Bulgarie (19 %), en Roumanie (22 %) et en Grèce (25 %).

Le problème se complique du fait de la répartition inégale de l'accès à Internet au sein de la population. Dans bien des pays, les personnes âgées, les personnes peu instruites et les minorités ethniques sont habituellement très sous-représentées parmi les personnes qui ont accès à Internet. Bethlehem (2007) décrit la situation observée aux Pays-Bas.

Pour mieux comprendre l'incidence du sous-dénombrement sur les estimations, supposons qu'on prélève un échantillon aléatoire pertinent dans la population Internet. Soit la population cible de l'enquête qui se compose de N personnes. À chaque personne k correspond une valeur Y_k de la variable cible Y. On suppose que l'enquête en ligne a pour but d'estimer la moyenne de la population $\bar{Y} = (Y_1 + Y_2 + ... + Y_N) / N$ de la variable cible Y.

On répartit la population U en une sous-population U_I de taille N_I de personnes ayant accès à Internet et une sous-population U_{NI} de taille N_{NI} de personnes sans accès à Internet. Appelons la sous-population U_I la population Internet. Supposons qu'un échantillon aléatoire simple est prélevé sans remplacement à partir de la population Internet. La moyenne de l'échantillon $\bar{y}_I$ est un estimateur sans biais de la moyenne $\bar{Y}_I$ de la population Internet, mais pas nécessairement de la moyenne de la population cible. Bethlehem (2007) montre que le biais de cet estimateur est égal à

$$B(\bar{y}_{HT}) = E(\bar{y}_{HT}) - \bar{Y} = \bar{Y}_I - \bar{Y} = \frac{N_{NI}}{N}(\bar{Y}_I - \bar{Y}_{NI}) . \qquad (2.2.1)$$

L'importance de ce biais est déterminée par deux facteurs. Le premier est la taille relative N_{NI} / N de la sous-population sans Internet. Le biais diminue donc à mesure que la couverture Internet augmente. Le deuxième facteur est le contraste $\bar{Y}_I - \bar{Y}_{NI}$ entre les moyennes de la population Internet et de la population sans Internet. Plus la moyenne de la variable cible diffère pour ces deux sous-populations, plus le biais est important.

Comme la couverture Internet augmente régulièrement, le facteur N_{NI} / N diminue, ce qui a pour effet de réduire le biais. Toutefois, on ignore si le contraste diminue aussi. Au contraire, il se pourrait que le groupe (restreint) de personnes sans Internet se démarque de plus en plus du reste de la population. Par conséquent, il peut encore subsister un biais substantiel.

2.3 Autosélection

La participation à une enquête en ligne reposant sur le principe de l'autosélection exige que les répondants connaissent l'existence de l'enquête et qu'ils décident de remplir le questionnaire sur Internet. Ainsi, chaque élément k de la population a la probabilité inconnue ρ_k de participer à l'enquête, pour k = 1, 2, ..., N. Bethlehem (2007) montre que la valeur prévue de la moyenne de l'échantillon est égale à

$$E(\bar{y}) \approx \bar{Y}^* = \frac{1}{N}\sum_{k=1}^{N}\frac{\rho_k}{\bar{\rho}}Y_k \qquad (2.3.1)$$

où $\bar{\rho}$ est la moyenne de toutes les propensions à répondre. Le biais de cet estimateur est égal à

$$B(\bar{y}) = E(\bar{y}) - \bar{Y} \approx \bar{Y}^* - \bar{Y} = \frac{C_{\rho Y}}{\bar{\rho}} = \frac{R_{\rho Y} S_\rho S_Y}{\bar{\rho}}, \qquad (2.3.2)$$

où $C_{\rho Y}$ est la covariance entre la variable cible et les probabilités de réponse, $R_{\rho Y}$ est le coefficient de corrélation, S_ρ est l'écart-type des probabilités de réponse et S_Y est l'écart-type de la variable cible. On peut montrer que dans le pire des cas (S_ρ prend sa valeur maximale et le coefficient de corrélation $R_{\rho Y}$ est égal à +1 ou à -1), la valeur absolue du biais est égale à

$$\left|B_{\max}(\bar{y})\right| = S_Y \sqrt{\frac{1}{\bar{\rho}} - 1}\,. \qquad (2.3.3)$$

Bethlehem (1988) montre que la formule (2.3.2) s'applique également dans la situation où l'on a prélevé un échantillon probabiliste et où l'on constate ensuite la non-réponse à l'étape du travail sur le terrain. Par conséquent, l'expression (2.3.3) offre un moyen de comparer les biais éventuels de divers plans d'enquête. Par exemple, toutes les enquêtes régulières de Statistics Netherlands reposent sur l'échantillonnage probabiliste. Leurs taux de réponse oscillent autour de 70 %, ce qui signifie que le biais absolu maximal est égal à $0{,}65 \times S_y$. Parmi les enquêtes menées aux Pays-Bas, l'une des plus vastes enquêtes en ligne reposant sur le principe de l'autosélection est l'enquête *21minuten.nl*. En 2006, en une période de six semaines, environ 170 000 personnes ont rempli le questionnaire en ligne. La population cible de cette enquête n'était pas définie, car tout le monde pouvait y participer. Si l'on suppose que la population cible englobe tous les Néerlandais de 18 ans et plus, la propension à répondre moyenne est égale à 170 000 / 12 800 000 = 0,0133. Le biais absolu maximal est donc égal à $8{,}61 \times S_y$. On peut conclure que le biais de la vaste enquête en ligne peut être 13 fois supérieur à celui de l'enquête probabiliste de moindre envergure.

On peut aussi illustrer l'incidence de l'autosélection en prenant pour exemple les élections générales tenues aux Pays-Bas en 2006. Pour prévoir le résultat de ces élections, divers organismes d'enquête ont mené des sondages d'opinion dont les résultats sont résumés dans le tableau 2.3-1. *Politieke Barometer*, *Peil.nl* et *De Stemming* sont des sondages d'opinion menés par des sociétés d'études de marché. Ils reposent tous sur des échantillons tirés de panels en ligne. Pour réduire le risque de biais, on a procédé à la repondération. Les sondages ont été menés la veille des élections. L'écart absolu moyen indique l'importance (moyenne) des écarts entre les résultats des sondages et celui des élections. Les écarts sont particulièrement importants dans le cas de partis volatils comme les sociaux-démocrates (PvdA), les socialistes (SP) et les populistes (PVV). Par exemple, l'un des sondages avait prévu que le parti socialiste (SP) obtiendrait 32 sièges au parlement, alors qu'il en a obtenu seulement 25.

Dans le cadre de l'Enquête sur les élections au parlement néerlandais (DPES), Statistics Netherlands a effectué le travail sur le terrain au cours des semaines qui ont précédé les élections. On a suivi le principe de l'échantillonnage probabiliste en prélevant un échantillon probabiliste véritable (à deux degrés) dans le registre de la population. On a interviewé les répondants sur place (au moyen de l'IPAO). Les prévisions de cette enquête sont nettement supérieures à celles qui reposent sur les sondages d'opinion en ligne. Les prévisions et les résultats des élections ne diffèrent que pour quatre partis, et les écarts sont d'au plus un siège.

Tableau 2.3-1.
Élections au parlement néerlandais, 2006. Résultats des élections et de divers sondages d'opinion

	Résultat des élections	*Politieke Barometer*	*Peil.nl*	*De Stemming*	Enquête 2006 (DPES)
Taille de l'échantillon	…	1 000	2 500	2 000	2 600
Sièges au parlement :					
CDA (chrétiens-démocrates)	41	41	42	41	41
PvdA (sociaux-démocrates)	33	37	38	31	32
VVD (libéraux)	22	23	22	21	22
SP (socialistes)	25	23	23	32	26
GL (parti vert)	7	7	8	5	7
D66 (libéraux-démocrates)	3	3	2	1	3
ChristenUnie (parti chrétien)	6	6	6	8	6
SGP (parti chrétien)	2	2	2	1	2
PvdD (parti animalier)	2	2	1	2	2
PVV (populistes)	9	4	5	6	8
Autres partis	0	2	1	2	1
Écart absolu moyen	…	1,27	1,45	2,00	0,36

… n'ayant pas lieu de figurer

L'échantillonnage probabiliste offre en outre l'avantage d'assurer une protection contre certains groupes de la population qui tentent de manipuler les résultats, notamment dans les sondages d'opinion. L'autosélection n'offre pas cette protection. On en a observé un exemple lors de l'attribution du Prix du livre de l'année 2005 (*NS Publieksprijs*), un prestigieux prix littéraire. Le choix du gagnant était déterminé au moyen d'un sondage mené sur un site Web. Le public pouvait voter pour l'un des livres finalistes ou pour un autre livre, au choix. Plus de 90 000 personnes ont participé à l'enquête. Le gagnant a été la nouvelle traduction de la Bible lancée par les Sociétés bibliques des Pays-Bas et des Flandres. Sans être finaliste, cet ouvrage a néanmoins recueilli une majorité écrasante (72 %) grâce à une campagne lancée (entre autres) par des sociétés bibliques, un radiodiffuseur chrétien et un journal chrétien. Bien que le concours se soit déroulé de façon tout à fait dans les règles, on ne peut vraiment pas considérer le groupe de votants comme représentatif de la population néerlandaise.

3. Repondération

3.1 Repondération traditionnelle

La repondération englobe une famille de techniques visant à améliorer la qualité des estimations d'enquête à l'aide de données auxiliaires. On entend par *données auxiliaires* un ensemble de variables mesurées dans le cadre de l'enquête et pour lesquelles on dispose de renseignements sur la répartition au sein de la population (ou de l'ensemble de l'échantillon). En comparant la répartition d'une variable auxiliaire au sein de la population avec sa répartition parmi les réponses, on peut évaluer si l'échantillon est représentatif ou non de la population (à l'égard de cette variable). Si ces répartitions diffèrent considérablement, on doit conclure que l'échantillon est sélectif. Pour corriger cette situation, on peut calculer des coefficients de pondération qu'on attribue à tous les enregistrements d'éléments observés. On peut maintenant obtenir des estimations des caractéristiques d'une population à l'aide des valeurs pondérées au lieu des valeurs non pondérées. On a souvent recours à la repondération pour corriger des résultats d'enquêtes entachées de non-réponse; voir, par exemple, Bethlehem (2002).

La *stratification a posteriori* est une méthode de pondération bien connue et souvent utilisée. Pour effectuer la stratification a posteriori, on a besoin d'une ou de plusieurs variables auxiliaires qualitatives qui, ensemble, répartissent la population cible en un certain nombre de strates (ou sous-populations). On attribue des coefficients de pondération identiques à tous les éléments de la même strate. Le biais de l'estimation fondée sur des données pondérées est infime s'il n'y a (en moyenne) aucun écart entre les participants et les non-participants. C'est le cas

s'il existe un lien étroit entre la variable cible et les variables de stratification. Dans les textes publiés, on parle alors d'une réponse « manquant au hasard » (MAR ou *Missing at Random*). La variation des valeurs de la variable cible se manifeste entre les strates, mais pas à l'intérieur des strates. En d'autres termes, les strates sont homogènes à l'égard de la variable cible. Malheureusement, on ne dispose pas très souvent de telles variables auxiliaires, ou encore la corrélation est faible.

3.2 Pondération fondée sur les scores de propension

Plusieurs sociétés d'études de marché utilisent la *pondération fondée sur les scores de propension* pour corriger le risque de biais dans leurs enquêtes en ligne; voir, par exemple, Börsch-Supan et coll. (2004) et Duffy et coll. (2005). On obtient des *scores de propension* en modélisant une variable qui indique si une personne participe ou non à l'enquête. On utilise habituellement un modèle de régression logistique où la variable indicatrice est la variable dépendante et les variables attitudinales sont les variables explicatives. On suppose que ces variables attitudinales expliquent pourquoi une personne participe ou non. L'ajustement du modèle de régression logistique se ramène à estimer la probabilité de participer (score de propension) compte tenu des valeurs des variables explicatives.

On utilise les scores de propension estimatifs pour stratifier la population. Chaque strate comporte des éléments présentant (à peu près) les mêmes scores de propension. Si, effectivement, tous les éléments d'une strate ont la même propension à répondre, il n'y a pas de biais lorsque l'on utilise uniquement les éléments de la population Internet aux fins de l'estimation. Selon Cochran (1968), cinq strates sont habituellement suffisantes pour éliminer une grande partie du biais. La société d'études de marché Harris Interactive a été l'une des premières à appliquer la pondération fondée sur les scores de propension; voir Terhanian et coll. (2001).

Deux conditions sont nécessaires à l'application de la pondération fondée sur les scores de propension. Premièrement, on doit disposer de variables auxiliaires pertinentes, soit des variables pouvant expliquer si une personne est disposée ou non à participer à l'enquête en ligne. On utilise souvent des variables qui mesurent les attitudes et le comportement en général. Deuxièmement, on doit disposer des valeurs de ces variables pour les participants et les non-participants.

3.3 Repondération fondée sur une enquête de référence

En l'absence de variables auxiliaires pertinentes, on peut envisager de mener une *enquête de référence* reposant sur un petit échantillon probabiliste. On recueille les données selon un mode différent de l'enquête en ligne, par exemple l'IPAO (interview sur place assistée par ordinateur, avec ordinateurs portatifs) ou l'ITAO (interview téléphonique assistée par ordinateur). Dans l'hypothèse d'une non-réponse nulle ou ignorable, cette enquête de référence produit des estimations sans biais de la répartition des variables auxiliaires au sein de la population. Plusieurs sociétés d'études de marché ont appliqué la méthode de l'enquête de référence; voir, par exemple, Börsch-Supan et coll. (2004) et Duffy et coll. (2005).

En calculant une répartition estimative de la population dans le cadre de la stratification a posteriori, on obtient la même valeur prévue de l'estimateur. Les conditions dans lesquelles on réduit le biais sont donc les mêmes que pour l'estimateur normal de stratification a posteriori. La méthode de l'enquête de référence a ceci d'intéressant qu'on peut utiliser n'importe quelle variable aux fins de la repondération, du moment qu'on la mesure dans les deux enquêtes. Par exemple, certaines sociétés d'études de marché utilisent des variables « webographiques » ou « psychographiques » pour répartir la population en « groupes de mentalité ». On suppose que les personnes appartenant à un même groupe ont plus ou moins le même niveau de motivation et d'intérêt à participer à ce genre d'enquête. Si tel est le cas, on peut utiliser efficacement ces variables aux fins de la repondération. Pour ce faire, on doit naturellement disposer, à l'égard de la population, de renseignements pertinents sur les variables psychographiques, fondés sur un taux de réponse élevé dans des échantillons aléatoires.

Schonlau et coll. (2004) décrivent l'enquête de référence de Harris Interactive. Il s'agit d'une enquête ITAO qui utilise la composition aléatoire. Cette enquête de référence sert à ajuster plusieurs enquêtes en ligne. Schonlau et coll. (2003) soulignent que le succès de cette méthode repose sur deux hypothèses : 1) les variables webographiques peuvent expliquer l'écart entre les répondants à l'enquête en ligne et les autres membres de la population cible;

2) l'enquête de référence n'est pas entachée d'une non-réponse non ignorable. En pratique, il n'est pas facile de satisfaire à ces conditions. Schonlau et coll. (2007) montrent que l'utilisation de variables webographiques aux fins de la repondération peut s'avérer efficace, mais pas toujours.

La méthode de l'enquête de référence présente aussi un inconvénient. Bethlehem (2007) montre que si une enquête de référence sert à estimer la répartition des variables auxiliaires au sein de la population, la variance de l'estimateur de stratification a posteriori dépend pour une large part de la taille de l'enquête de référence, de moindre envergure. Le nombre élevé d'observations de l'enquête en ligne ne contribue donc pas à produire des estimations exactes. La méthode de l'enquête de référence réduit le biais des estimations au prix d'une variance plus élevée. La taille effective de l'échantillon de l'enquête en ligne est du même ordre de grandeur que celle de l'enquête de référence.

Il convient de mentionner que les questions attitudinales sont beaucoup moins fiables que les questions factuelles. Les répondants n'ont peut-être jamais pensé aux sujets abordés dans les questions attitudinales. Ils doivent se faire une idée au moment même où la question est posée. Leurs réponses peuvent dépendre de leur situation actuelle et peuvent varier avec le temps. Une question attitudinale peut donc donner lieu à des erreurs de mesure importantes.

4. Des enquêtes en ligne pour les statistiques officielles ?

Peut-on utiliser les enquêtes en ligne reposant sur le principe de l'autosélection pour recueillir les données pour les besoins des statistiques officielles ? L'examen des graves problèmes méthodologiques abordés à la section 2 nous amène à conclure qu'il est très difficile, sinon impossible, de faire une inférence valide au sujet de la population étudiée. L'autosélection risque d'entraîner un biais dans les estimations des caractéristiques de la population. Ce risque ressemble à l'effet de la non-réponse dans les enquêtes reposant sur l'échantillonnage probabiliste traditionnel, mais nous avons montré que dans les enquêtes reposant sur le principe de l'autosélection, le biais pouvait s'avérer beaucoup plus important.

La repondération permet de réduire le biais, mais uniquement si le mécanisme d'échantillonnage satisfait à la condition d'une réponse manquant au hasard (*Missing at Random* ou MAR). Les variables de pondération doivent alors être en corrélation étroite avec les variables cibles de l'enquête et les probabilités de réponse. Souvent, on ne dispose pas de ces variables. On pourrait résoudre ce problème en menant une enquête de référence. Selon certains chercheurs, les variables webographiques semblent pouvoir expliquer le comportement de réponse. Il convient de souligner que les variables webographiques sont des variables attitudinales. Elles sont beaucoup plus difficiles à mesurer que des variables factuelles et, par conséquent, peuvent donner lieu à des erreurs de mesure importantes.

Une enquête de référence n'est efficace que si elle porte réellement sur un échantillon probabiliste sans non-réponse, ou avec non-réponse ignorable, c'est-à-dire avec réponse manquant entièrement au hasard (MCAR ou *Missing Completely at Random*). En pratique, il peut être difficile de satisfaire à cette condition. Si les estimations d'une enquête de référence sont biaisées à cause de la non-réponse non ignorable, on ne fait que remplacer le biais de l'enquête en ligne par le biais de l'enquête de référence, ce qui n'aide pas vraiment à résoudre le problème.

L'autosélection constitue un grave problème, qu'on peut néanmoins résoudre en appliquant l'échantillonnage probabiliste. On peut prélever un échantillon aléatoire (d'adresses, par exemple) dans une base de sondage et envoyer à chaque adresse sélectionnée une lettre demandant de remplir un questionnaire sur Internet. Des codes d'identification uniques garantissent que les bonnes personnes répondent aux questions. En fait, la seule différence avec une enquête postale est que sur Internet, le questionnaire papier est remplacé par un questionnaire électronique.

Il faut également résoudre le problème du sous-dénombrement dans les enquêtes en ligne. Il est intéressant de signaler qu'entre 60 % et 70 % seulement des ménages néerlandais ont encore un téléphone fixe dont le numéro est inscrit à l'annuaire. Si l'on sélectionne un échantillon d'après l'annuaire téléphonique, il manque donc un ménage sur trois. Les enquêtes téléphoniques traditionnelles sont donc entachées, elles aussi, d'un sous-dénombrement. Ce dernier devrait diminuer avec le temps. Du point de vue de la couverture, une enquête en ligne peut s'avérer plus efficace qu'une enquête téléphonique, du moins aux Pays-Bas.

Si le sous-dénombrement dans les enquêtes en ligne constitue vraiment un problème, on pourrait le résoudre en offrant simplement l'accès à Internet aux personnes qui en sont dépourvues. On peut citer comme exemple le panel LISS; voir Scherpenzeel (2008). Établi en 2007, ce panel néerlandais se compose d'environ 5 000 ménages. Pour le constituer, Statistics Netherlands a prélevé un échantillon probabiliste dans la population néerlandaise. On a invité des ménages à devenir membres du panel au moyen d'une enquête IPAO ou ITAO. Aux ménages sans accès à Internet, on a offert une connexion Internet gratuite et un ordinateur personnel d'utilisation simple, mis au point pour les personnes qui n'avaient jamais utilisé Internet auparavant.

Une enquête en ligne peut-elle remplacer une enquête IPAO ou ITAO? En ce qui concerne la collecte des données, il existe une différence importante entre les enquêtes IPAO et ITAO, d'une part, et les enquêtes en ligne, d'autre part. Dans les enquêtes IPAO et ITAO, les intervieweurs effectuent le travail sur le terrain. Ils jouent un rôle important en persuadant les gens de participer à l'enquête et peuvent aussi les aider à remplir le questionnaire. Dans une enquête en ligne, il n'y a pas d'intervieweur. Il s'agit d'une auto-interview. La qualité des données recueillies peut donc être inférieure à cause des taux de non-réponse élevés et du plus grand nombre d'erreurs dans les réponses aux questions. Selon De Leeuw et Collins (1997), les taux de réponse ont tendance à être plus élevés quand un intervieweur pose les questions. Toutefois, la réponse aux questions délicates est plus élevée sans intervieweur.

Une enquête en ligne peut constituer l'un des modes retenus dans le cadre d'une méthode de collecte des données à mode de collecte mixte. Chaque mode de collecte (interview sur place, enquête téléphonique, enquête postale, enquête en ligne, etc.) présente des avantages et des inconvénients. Le mélange des modes de collecte offre la possibilité de compenser la faiblesse de chaque mode. On peut ainsi réduire les coûts tout en améliorant les taux de réponse et la qualité des données. Il existe plusieurs stratégies de collecte des données à mode de collecte mixte. L'une est la méthode simultanée : on répartit l'échantillon en groupes que l'on approche en même temps par des modes différents. Une deuxième stratégie à mode de collecte mixte est la méthode séquentielle : on approche d'abord tous les membres de l'échantillon par le mode le moins coûteux (Internet). On procède ensuite au suivi des non-répondants par l'autre mode le plus économique (ITAO) et, enfin, au suivi des derniers non-répondants par le mode le plus coûteux (IPAO). Une troisième stratégie pourrait consister à laisser les répondants choisir le mode de collecte des données qu'ils préfèrent. Elle est souple et conviviale, mais certains travaux de recherche semblent indiquer que les taux de réponse baissent dès qu'on offre aux répondants le choix du mode de collecte; voir Dillman (2008).

La collecte des données à mode de collecte mixte soulève une préoccupation importante : la qualité des données risque d'être compromise par les effets de mode. Selon ce phénomène, lorsqu'on pose à une personne la même question selon différents modes de collecte des données, on obtient des réponses différentes. Par exemple, supposons qu'on pose une question fermée comportant plusieurs options de réponse. Dans le cas d'une enquête sur place, on présente au répondant une affiche cartonnée montrant toutes les réponses possibles. Dans le cas d'une enquête téléphonique, l'intervieweur lit au répondant toutes les possibilités de réponse. Selon certains chercheurs, les répondants semblent préférer les dernières options de la liste. Quant aux répondants à une enquête en ligne qui se voient proposer une liste déroulante, ils ont tendance à opter pour les premières réponses de la liste.

On peut conclure qu'il y a parfois lieu d'utiliser les enquêtes en ligne comme mode de collecte des données pour les besoins des statistiques officielles. Il faudrait approfondir la recherche pour établir comment on peut procéder sans altérer la qualité des données recueillies.

Bibliographie

Bethlehem, J.G. (1988). Reduction of the nonresponse bias through regression estimation, *Journal of Official Statistics,* 4, 251-260.

Bethlehem, J.G. (2002). Weighting Nonresponse Adjustments Based on Auxiliary Information. *Survey Nonresponse* (Éds. R.M. Groves et coll.). New York: Wiley.

Bethlehem, J.G. (2007). Reducing the bias of web survey based estimates, Document de travail 07001, Voorburg, Pays-Bas: Statistics Netherlands.

Cochran, W.G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies, *Biometrics*, 24, 205-213.

Couper, M.P., Baker, R.P., Bethlehem, J.G., Clark, C.Z.F., Martin, J., Nicholls II, W.L. et O'Reilly, J.M. (Éds.) (1998). *Computer Assisted Survey Information Collection*, New York: Wiley.

De Leeuw, E. et Collins M. (1997). Data collection methods and survey quality. *Survey Measurement and Process Control* (Éds. L. Lyberg et coll.). New York: Wiley, pp. 199-220.

Dillman, D. (2008). How can we convince the general public to respond to Internet surveys, *MESS Workshop*. Zeist, Pays-Bas.

Duffy, B., Terhanian, G., Bremer, J. et Smith, K. (2005). Comparing data from online and face-to-face surveys, *International Journal of Market Research*, 47, 615-639.

Horvitz, D.G. et Thompson D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47, 663-685.

Kruskal, W. et Mosteller, F. (1979a). Representative Sampling, I: Non-scientific Literature, *International Statistical Review*, 47, 13-24.

Kruskal, W. et Mosteller, F. (1979b). Representative Sampling, II: Scientific Literature, Excluding Statistics, *International Statistical Review*, 47, 111-127.

Kruskal, W. et Mosteller, F. (1979c). Representative Sampling, III: the Current Statistical Literature, *International Statistical Review*, 47, 245-265.

Scherpenzeel, A. (2008). An online panel as a platform for multi-disciplinary research, *Access panels and online research, panacea or pitfall?* (Éds. I. Stoop et M. Wittenberg), Amsterdam: Aksant, 101-106.

Schonlau, M., Fricker, R.D. et Elliott, M.N. (2003). Conducting Research Surveys via E-mail and the Web, Rand Corporation, Santa Monica, CA.

Schonlau, M., Van Soest, A. et Kapteyn, A. (2007). Are 'webographic' or attitudinal questions useful for adjusting estimates from web surveys using propensity scoring?, *Survey Research Methods*, 1, 155-163.

Schonlau, M., Zapert, K., Payne Simon, L., Sanstad, K., Marcus, S., Adams, J., Spranca, M., Kan, H., Turner, R. et Berry, S. (2004). A Comparison between responses from propensity-weighted web survey and an identical RDD survey, *Social Science Computer Review*, 22, 128-138.

Terhanian, G., R. Smith, J. Bremer, et R. K. Thomas (2001). Exploiting Analytical Advances: Minimizing the Biases Associated with Internet-Based Surveys of Non-Random Samples, *ARF/ESOMAR: Worldwide Online Measurement,* ESOMAR Publication Services, 248, 247–272.

Vonk, T., Van Ossenbruggen, R. et Willems, P. (2006). The effects of panel recruitment and management on research results, a study among 19 online panels, *Panel Research 2006, ESOMAR World Research*, ESOMAR Publication Services, 317, 79-99.