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

Utilisation de panels en ligne pour les statistiques officielles

Jelke Bethlehem[1]

Résumé

Les progrès en informatique, ainsi que les nouveaux défis sociétaux, tels que la hausse des taux de non-réponse et la diminution des budgets, peuvent entraîner des changements de méthodes d'enquête pour la production des statistiques officielles. De nos jours, l'usage des panels en ligne est devenu très répandu dans le domaine des études de marché. La question qui se pose est celle de savoir si ces panels conviennent pour les statistiques officielles. Permettent-ils de produire des statistiques de haute qualité au sujet de la population générale? Le présent article a pour objet de répondre à cette question en explorant divers aspects méthodologiques, dont le sous-dénombrement, la sélection de l'échantillon et la non-réponse. Statistics Netherlands a procédé à un essai au moyen d'un panel en ligne. Certains résultats sont décrits.

Mots clés : panel en ligne, représentativité, sous-dénombrement, non-réponse, échantillonnage.

1. Introduction

1.1 Collecte de données pour les statistiques officielles

Les instituts nationaux de statistique (INS) ont pour tâche de produire des statistiques précises. Traditionnellement, les données nécessaires à la production de ces statistiques sont recueillies en réalisant des enquêtes sur place ou par téléphone. Cette méthode de collecte des données est onéreuse, mais l'expérience a montré qu'il s'agissait du prix à payer pour obtenir des données de haute qualité. De nos jours, dans de nombreux pays, les contraintes budgétaires obligent les INS à rechercher des moyens moins coûteux de recueillir les données tout en maintenant un haut niveau de qualité.

À première vue, un panel en ligne semblerait être une solution de rechange prometteuse. La collecte de données en ligne remporte de plus en plus de succès, particulièrement dans le domaine des études de marché. Rien d'étonnant à cela, car il s'agit d'un moyen simple, rapide et bon marché de recueillir de grandes quantités de données. Une fois qu'un panel en ligne a été mis en place, il est simple de réaliser une enquête. Aucune procédure complexe et coûteuse de sélection d'échantillons n'est nécessaire. Il suffit d'envoyer un courriel aux membres du panel (ou à un échantillon de ceux-ci). Aucun intervieweur n'est nécessaire, et il n'y a pas de frais d'envoi des questionnaires imprimés par la poste. Il suffit de placer un questionnaire électronique sur Internet.

La vitesse est un autre avantage de la collecte des données en ligne. Une nouvelle enquête peut être lancée rapidement. Il existe des exemples d'enquête en ligne dont la conception du questionnaire, ainsi que la collecte, l'analyse et la publication des données n'ont pas pris plus d'un jour. En raison de ces avantages et de leur nature longitudinale, les panels en ligne sont devenus un outil puissant pour la réalisation de sondages d'opinion. Ainsi, durant les dernières semaines de la campagne pour les élections parlementaires de 2012 aux Pays-Bas, quatre grands sondages d'opinion nationaux distincts ont été effectués quotidiennement, et ils étaient tous fondés sur des panels en ligne. Voir Bethlehem (2013).

Les panels en ligne peuvent être utilisés de diverses façons en statistique officielle :

- pour des études longitudinales, lorsque le même jeu de variables est mesuré pour le même groupe de personnes à différents points dans le temps. L'accent est mis sur la mesure du changement;

[1]Jelke Bethlehem, Université de Leiden, Faculté des sciences sociales et du comportement, Wassenaarseweg 52, 2333 AK Leiden, Pays-Bas (j.g.bethlehem@fsw.leidenuniv.nl).

- pour des études transversales, où le panel sert de base de sondage pour des enquêtes particulières qui peuvent porter sur différents sujets. Des échantillons peuvent être sélectionnés pour différents groupes (personnes âgées, personnes ayant un niveau élevé de scolarité, etc.);
- Pour des enquêtes ponctuelles (non planifiées) rapides qu'un INS peut réaliser à la demande de tierces parties.

Le présent article est axé sur l'utilisation de panels en ligne pour réaliser des enquêtes transversales auprès de la population générale. Afin de pouvoir faire des inférences statistiques appropriées au sujet d'une population, le recrutement des membres du panel et la sélection de l'échantillon doivent se fonder sur l'échantillonnage probabiliste. Le panel LISS aux Pays-Bas (Scherpenzeel, 2008), le KnowledgePanel aux États-Unis (Knowledge Networks, 2012) et le panel ELIPSS en France (Lynn, 2013) en sont des exemples.

L'article explore l'utilisation possible de panels en ligne pour les statistiques officielles. Les panels en ligne permettent-ils de produire des statistiques aussi précises que celles obtenues au moyen d'enquêtes par IPAO? Pour essayer de répondre à cette question, nous abordons un certain nombre de problèmes, dont le sous-dénombrement, le recrutement et la non-réponse.

2. Panels en ligne

2.1 Sous-dénombrement

Une enquête réalisée en partant d'un panel en ligne peut présenter un certain sous-dénombrement, parce que la population cible d'une enquête est habituellement beaucoup plus vaste qu'uniquement les personnes ayant une connexion à Internet. Ainsi, selon Eurostat, l'organisme statistique de l'Union européenne, 79 % des ménages de l'UE avaient accès à Internet en 2013. La proportion variait considérablement d'un pays à l'autre. Les pays où la pénétration d'Internet était la plus élevée étaient l'Islande (96 %), les Pays-Bas (95 %), la Norvège (94 %) et le Luxembourg (94 %). L'accès à Internet était le plus faible en Turquie (49 %), en Bulgarie (54 %), en Grèce (56 %) et en Roumanie (58 %). Un problème encore plus épineux tient au fait que l'accès à Internet n'est pas réparti uniformément dans la population. On constate habituellement dans de nombreux pays que les personnes âgées, les personnes peu instruites et les minorités ethniques sont gravement sous-représentées parmi le groupe ayant accès à Internet.

Bethlehem et Biffignandi (2012, chapitre 8) montrent que le biais, lié au sous-dénombrement, de la moyenne d'échantillon, en tant qu'estimateur de la moyenne de population d'une variable d'intérêt, dépend de deux facteurs :

- le pourcentage de personnes qui ont accès à Internet, le biais étant d'autant plus faible que la pénétration d'Internet est élevée;
- le contraste entre les personnes qui ont et n'ont pas accès à Internet. Il s'agit de la différence entre les moyennes de population des personnes ayant accès et de celles n'ayant pas accès à Internet. Le biais est d'autant plus important que la moyenne de la variable d'intérêt diffère entre les deux groupes.

Dans de nombreux pays, le pourcentage de personnes qui n'ont pas accès à Internet n'est pas négligeable. En outre, il existe des différences importantes entre les personnes qui ont accès et celles qui n'ont pas accès à Internet. Des groupes particuliers sont sous-représentés dans la population ayant accès à Internet. Cela porte à conclure qu'en général, un échantillon aléatoire tiré de la population ayant une connexion à Internet produira des estimations biaisées des paramètres de la population cible.

Il faut s'attendre à ce que la pénétration d'Internet augmente avec le temps. Donc, le pourcentage de personnes ayant accès à Internet augmentera, ce qui réduira le biais. Cependant, il n'est pas certain que le contraste entre les personnes ayant et celles n'ayant pas Internet diminuera aussi avec le temps. Il est même possible qu'il augmente, car le groupe n'ayant pas accès à Internet pourrait différer de plus en plus de celui ayant accès à Internet. Donc, l'effet combiné d'un plus grand nombre de personnes ayant accès à Internet et d'un plus grand contraste ne se traduira pas nécessairement par une réduction du biais.

Un moyen de résoudre le problème du sous-dénombrement consiste à offrir l'accès à Internet aux personnes sans connexion à Internet qui sont sélectionnées dans l'échantillon. Cette approche a été mise en œuvre pour le

KnowledgePanel aux États-Unis, et pour le panel LISS aux Pays-Bas. Leenheer et Scherpenzeel (2013) montrent que cette mesure a amélioré la représentativité du panel LISS. La population enquêtée se rapprochait davantage de la population générale. Lynn (2013) mentionne le panel français ELIPSS à chaque membre duquel ont été fournis une tablette et un abonnement 3G. L'avantage de cette approche est que tous les membres du panel utilisent le même appareil pour remplir le questionnaire, ce qui évite les effets d'appareil.

Néanmoins, fournir un dispositif Internet aux membres du panel soulève aussi de nouvelles questions. Cette approche est-elle encore possible si le sous-dénombrement est important? Et qu'en est-il des personnes qui n'ont aucune expérience d'Internet? Cela donnera-t-il lieu à des interviews incomplètes ou à des erreurs de mesure? Une autre solution au problème du sous-dénombrement consiste à étendre le panel en ligne pour en faire un panel à mode de collecte mixte. Les personnes échantillonnées qui n'ont pas accès à Internet seront alors approchées selon un autre mode que le questionnaire en ligne (envoi par la poste, ITAO ou IPAO). Quelle que soit l'approche utilisée pour réduire le sous-dénombrement (offre de l'accès gratuit à Internet ou l'établissement d'un panel à mode de collecte mixte), le coût du panel augmentera.

2.2 Recrutement du panel

Créer un panel en ligne qui permet de faire des inférences statistiques valides au sujet d'une population générale requiert la sélection d'un échantillon probabiliste. Il ne s'agit pas d'une tâche facile, parce qu'habituellement, aucune base de sondage appropriée n'est disponible. Par conséquent, de nombreux panels en ligne font appel à une certaine forme d'autosélection. L'autosélection (également appelée adhésion volontaire, ou *opt-in* en anglais) signifie qu'il appartient entièrement aux gens de choisir de participer au panel. Les membres du panel sont les personnes qui ont Internet, qui voient une invitation, qui visitent le site approprié et qui décident de participer.

Dans le cas de l'autosélection, le chercheur n'exerce aucun contrôle sur le processus de sélection. La probabilité de participation de toute personne est inconnue, de sorte qu'il est impossible de construire des estimateurs sans biais. Un autre problème tient au fait que des personnes ne faisant pas partie de la population cible peuvent devenir membres du panel. En outre, une personne pourrait s'inscrire au panel plusieurs fois (éventuellement sous différentes identités). Un autre problème est dû au fait que certaines personnes peuvent essayer de manipuler les résultats de l'enquête. Bronzwaer (2012) donne un exemple. Durant la campagne pour les élections parlementaires aux Pays-Bas en 2012, un groupe de 2 500 personnes ont demandé à être membres d'un panel en ligne (Peil.nl) utilisé pour réaliser des sondages d'opinion politique. Leur idée était de se comporter d'abord comme des électeurs en faveur du parti des personnes âgées (50PLUS), puis de changer progressivement en faveur des chrétiens démocrates. Malheureusement pour eux, leur complot a été déjoué quand l'organisme de sondage a constaté une forte augmentation soudaine du nombre de personnes se portant volontaires pour le panel. Néanmoins, cette anecdote montre qu'il est techniquement possible de manipuler un panel en ligne recruté par autosélection.

Pour montrer les effets de l'autosélection sur les estimateurs, supposons que chaque membre de la population ayant accès à Internet possède une probabilité inconnue de participer à une enquête. Un chercheur naïf émettant l'hypothèse d'un échantillonnage aléatoire simple utilisera la moyenne d'échantillon comme estimateur. Bethlehem et Biffignandi (2012, chapitre 9) montrent que cet estimateur est biaisé et que le biais est déterminé par trois facteurs :

- la probabilité moyenne de participation, le biais étant d'autant plus petit que la probabilité moyenne de participation est faible;
- la variation des probabilités de participation, le biais étant d'autant plus grand que la variation des probabilités est importante;
- la corrélation entre la variable d'intérêt et les probabilités de participation, le biais étant d'autant plus grand que la corrélation est forte.

Il convient de souligner que la probabilité de participation moyenne peut être très faible, et donc le biais, très grand. Par exemple, aux Pays-Bas, il existe des panels en ligne comptant 100 000 membres. La population cible (tous les Néerlandais de 18 ans et plus) compte 12,8 millions de personnes. Donc, la probabilité moyenne de participation est égale à 100 000 / 12 800 000 = 0,008.

Le biais disparaît si toutes les probabilités de participation sont égales. Le mécanisme d'autosélection est alors comparable à un échantillonnage aléatoire simple. Le biais disparaît également si la participation ne dépend pas de la valeur de la variable d'intérêt. L'emploi d'un panel autosélectionné est hors de question pour produire des statistiques précises au sujet de la population générale. En effet, un groupe de travail spécial de l'American Association for Public Opinion Research (AAPOR) a conclu que les chercheurs doivent éviter les panels en ligne non probabilistes lorsqu'un de leurs objectifs est d'estimer avec précision les valeurs de population (voir Baker et coll., 2010).

Idéalement, la base de sondage pour un panel en ligne est une liste des adresses de courrier électronique de toutes les personnes faisant partie de la population cible. Ce genre de liste pourrait exister, par exemple pour tous les étudiants d'une université, ou pour tous les employés d'une grande entreprise, mais malheureusement, une telle liste n'existe pas pour la population générale. Il faut donc recourir à un autre mode de recrutement. Voici quelques exemples.

- Sélectionner un échantillon aléatoire à partir d'un registre de population, et envoyer aux personnes sélectionnées une lettre contenant l'adresse Internet de l'enquête et un code d'ouverture de séance unique. Cette approche est utilisée à l'heure actuelle par Statistics Netherlands pour réaliser des enquêtes en ligne et des enquêtes à mode mixte;
- Sélectionner un échantillon aléatoire à partir d'un registre de population, utiliser des annuaires téléphoniques pour relier autant que possible des numéros de téléphone aux personnes sélectionnées, appeler ces personnes et les inviter à devenir membres du panel. Si aucun numéro de téléphone ne peut être lié à une personne, essayer d'obtenir la coopération de cette personne au moyen d'une interview par IPAO. Une approche de ce type a été utilisée pour le panel LISS (voir Leenheer et Scherpenzeel, 2013);
- Sélectionner un échantillon aléatoire de numéros de téléphone (par exemple, par la méthode de composition aléatoire), appeler chaque numéro sélectionné et inviter un membre du ménage choisi au hasard à devenir membre du panel.

Les deux premières méthodes de recrutement ne peuvent être utilisées que s'il existe un registre de population, comme dans les pays scandinaves et aux Pays-Bas. Sinon, une méthode similaire pourrait consister à utiliser une liste d'adresses, par exemple un fichier d'adresses postales (FAP).

Les grands organismes de sondage qui réalisent de nombreux sondages pourraient envisager de recruter un panel parmi les participants à l'une de leurs enquêtes par IPAO ou par ITAO. Par exemple, Statistics Netherlands demande toujours aux participants aux enquêtes par IPAO/ITAO s'ils accepteraient de participer de nouveau à certaines enquêtes dans l'avenir. Les personnes coopératives pourraient aussi être d'accord de devenir membres d'un panel. En outre, il arrive parfois que leur adresse de courrier électronique soit déjà disponible, ce qui permet d'éviter un mode de recrutement différent. Cependant, un inconvénient important de cette approche est que les participants aux enquêtes par IPAO/ITAO ne constituent pas un échantillon aléatoire de la population générale. Le panel pourrait donc manquer de représentativité et donner des estimations biaisées.

2.3 Non-réponse

La non-réponse est un problème important dans le cas des panels en ligne. Elle a lieu à deux étapes, à savoir 1) durant la phase de recrutement et 2) dans des sondages particuliers effectués à partir du panel. La non-réponse au recrutement peut être élevée parce que la participation à un panel en ligne requiert un engagement et un effort importants de la part des répondants. Les taux de réponse aux enquêtes en ligne basées sur l'échantillonnage probabiliste sont souvent inférieurs à 40 %. Voir, par exemple, Cook, Heath et Thompson (2000), Kaplowitz, Hadlock et Levine (2004), ainsi que Lozar Manfreda et coll. (2008). Par contre, le taux de réponse à une enquête particulière tirée du panel est souvent élevé, car les membres du panel ont tous accepté de participer à une enquête régulièrement.

La non-réponse peut avoir diverses causes. Celles-ci dépendent de l'approche de recrutement adoptée pour créer le panel. Si une base de sondage avec adresses de courrier électronique est disponible, la non-réponse peut être le résultat d'une impossibilité à prendre contact (adresse électronique incorrecte, message intercepté par un filtre antipourriel, message non lu), d'un refus (après avoir lu le message du courriel), ou d'une incapacité à répondre (par

exemple navigateur ne fonctionnant pas correctement). Si le recrutement a lieu par la poste ou par téléphone, les causes de non-réponse sont similaires à celle des enquêtes par la poste ou par téléphone.

L'attrition est un type particulier de non-réponse qui peut se produire dans le cas d'enquêtes réalisées à partir d'un panel. Les membres du panel peuvent se fatiguer de répondre à des questionnaires d'enquête particuliers et, par conséquent, décider d'arrêter de participer. Une fois qu'ils s'arrêtent, ils ne recommencent jamais à participer.

Le problème de la non-réponse tient au fait qu'elle peut être sélective. Pour montrer ce que peuvent être les effets de la non-réponse, on suppose habituellement que chaque membre de la population possède une probabilité de réponse inconnue. Alors, selon Bethlehem, Cobben et Schouten (2011), le biais de la moyenne des réponses (en tant qu'estimateur de la moyenne de population) est déterminé par trois facteurs, à savoir :

- la probabilité moyenne de réponse, le biais étant d'autant plus petit que la probabilité moyenne de réponse est élevée;
- la variation des probabilités de réponse, le biais étant d'autant plus grand que la variation des probabilités est importante;
- la corrélation entre la variable d'intérêt et les probabilités de réponse, le biais étant d'autant plus grand que la corrélation est forte.

Le biais disparaît si toutes les probabilités de réponse sont égales. Alors, la réponse peut être considérée comme un échantillon aléatoire simple. Le biais disparaît également si les probabilités de réponse ne dépendent pas de la valeur de la variable d'intérêt.

Le mode de recrutement a une incidence sur le taux de réponse au recrutement. Habituellement, les enquêtes assistées par intervieweur (IPAO, ITAO) sont caractérisées par des taux de réponse plus élevés que les enquêtes autoadministrées (par la poste, en ligne). Du point de vue de la représentativité, le recrutement assisté par intervieweur devrait donc être privilégié. Cependant, cela accroît considérablement le coût du recrutement.

Il est important de s'en tenir au paradigme de l'échantillonnage probabiliste : les échantillons doivent être sélectionnés par échantillonnage probabiliste, et les probabilités de sélection doivent être connues. L'autosélection (adhésion volontaire) a été rejetée en tant que technique d'échantillonnage scientifique valable, parce que les probabilités de sélection sont inconnues. En outre, ces probabilités peuvent être nulles pour certains groupes. Mais qu'en est-il d'un échantillon probabiliste approprié qui fait l'objet d'une quantité importante de non-réponse? Cette situation ne ressemble-t-elle pas quasiment à un sondage avec autosélection? Bethlehem (2010) montre que les expressions pour le biais de non-réponse et le biais d'autosélection sont similaires, l'expression pour le biais de non-réponse contenant les probabilités de réponse et celle pour le biais de sélection contenant les probabilités de participation. Les probabilités de réponse sont en général nettement plus élevées que les probabilités de participation. Cela signifie que le biais d'autosélection peut être beaucoup plus grand que le biais de non-réponse. À titre d'exemple, le taux de réponse au recrutement du panel LISS basé sur l'échantillonnage probabiliste était de 54 %, ce qui correspond à une probabilité de réponse moyenne de 0,54. Le taux de participation à un panel en ligne autosélectionné aux Pays-Bas était de 0,008. Cela implique que, dans le pire cas, le biais d'autosélection peut être plus de 12 fois plus grand que le biais de non-réponse.

Le taux de réponse n'est pas le seul facteur qui détermine le biais de non-réponse. Un autre facteur important est la variation des probabilités de réponse. Celle-ci est mesurée par l'écart-type. Si toutes les probabilités de réponse sont égales, l'écart-type est nul, et il n'y a pas de biais. Plus les probabilités de réponse varient, plus le biais est grand. Schouten, Cobben et Bethlehem (2009) proposent l'indicateur R comme mesure de la représentativité. Cet indicateur est défini comme étant $R = 1 - 2S_\rho$, où S_ρ est l'écart-type des probabilités de réponse. R est égal à 1 si toutes les probabilités de réponse sont les mêmes. Il en est ainsi en cas de représentativité complète. Le manque de représentativité est d'autant plus important que la valeur de R est proche de 0.

Le tableau 2.3.1 donne un exemple de non-réponse dans un panel en ligne. Les données proviennent du panel LISS et sont tirées de Scherpenzeel et Schouten (2011). Le tableau montre les taux de réponse aux diverses phases du processus. Les pourcentages se rapportent à l'échantillon initial. Un contact a pu être établi avec 91 % des ménages échantillonnés. Dans 75 % des cas, les ménages ont consenti à participer à une brève interview de recrutement.

Cinquante-quatre pour cent des ménages ont décidé de participer au panel, mais seulement 48 % sont devenus actifs dans le panel. Donc, le taux de réponse à la phase de recrutement était de 54 %. Le tableau montre aussi l'effet de l'attrition. Le taux de réponse diminue au cours du temps. Après quatre ans, seulement un ménage échantillonné sur trois dans l'échantillon original est encore actif dans le panel.

Tableau 2.3.1

Taux de réponse et indicateur *R* pour le panel LISS

Phase	Réponse	Indicateur *R*
Contact de recrutement	91 %	0,85
Interview de recrutement	75 %	0,80
Accepte de participer au panel	54 %	0,71
Actif dans le panel en 2007	48 %	0,67
Actif dans le panel en 2008	41 %	0,70
Actif dans le panel en 2009	36 %	0,75
Actif dans le panel en 2010	33 %	0,78

L'indicateur *R* est encore élevé à la phase de prise de contact (0,85), mais il diminue au cours du processus de recrutement. L'indicateur *R* n'est que de 0,67 quand le panel débute en 2007, ce qui indique que ce dernier n'est pas tellement représentatif. Étonnamment, l'indicateur *R* augmente de nouveau au fil des ans. Apparemment, l'érosion rend la composition du panel plus équilibrée.

2.4 Correction du biais

Au moins trois phénomènes peuvent avoir une incidence sur la représentativité d'un panel en ligne, à savoir le sous-dénombrement, l'autosélection et la non-réponse. Le manque de représentativité peut aboutir à des estimations biaisées. Le redressement par pondération englobe une famille de techniques de correction visant à réduire le biais en se servant de variables auxiliaires. Ces variables auxiliaires sont définies ici comme des variables qui sont mesurées dans une enquête et dont la distribution dans la population est connue.

La comparaison de la distribution des réponses pour une variable auxiliaire à la distribution de cette variable dans la population montre clairement si la réponse est représentative ou non de la population (en ce qui concerne la variable en question). Si les deux distributions diffèrent considérablement, on est forcé de conclure que la réponse est sélective. Pour corriger une telle situation, des poids d'ajustement sont calculés pour les participants à l'enquête. Ceux qui appartiennent à un groupes sous-représentés reçoivent un poids supérieur à 1 et ceux qui appartiennent à un groupe surreprésenté, un poids plus petit que 1. Les estimations des caractéristiques de la population sont ensuite calculées en utilisant les données pondérées au lieu des données non pondérées.

La poststratification est une méthode de pondération simple et souvent utilisée. L'estimation par la régression généralisée et l'estimation par le raking ratio sont des méthodes de pondération plus avancées. Il existe aussi des méthodes de pondération fondées sur les probabilités de réponse estimées. On parle alors de pondération par la propension à répondre. Un traitement plus approfondi des méthodes de pondération figure, par exemple, dans Bethlehem et Biffignandi (2012) et dans Särndal et Lundström (2005). Ces méthodes de pondération ne permettent de réduire le biais de sélection que si les variables auxiliaires utilisées satisfont deux conditions :

- les variables auxiliaires doivent être fortement corrélées avec les probabilités de réponse ou de participation. Elles doivent pouvoir expliquer le mécanisme de sélection;
- les variables auxiliaires doivent être fortement corrélées aux variables d'intérêt de l'enquête ou du panel. Elles doivent être suffisamment explicatives pour bien prédire les valeurs des variables d'intérêt.

Le biais disparaîtra complètement si l'on utilise toutes les variables auxiliaires qui sont nécessaires pour expliquer complètement le mécanisme de sélection et les variables d'intérêt. Si l'on utilise seulement un sous-ensemble, le biais sera réduit mais non éliminé.

Des problèmes de sélection ont lieu aux deux phases du panel en ligne. Durant le recrutement, il peut y avoir sous-dénombrement, autosélection et non-réponse. Il peut aussi y avoir non-réponse aux enquêtes réalisées à partir du

panel en ligne. Idéalement, la correction du biais doit se faire en deux étapes. La sélectivité lors du recrutement est un phénomène différent de la non-réponse à une enquête. Par conséquent, ces deux phénomènes peuvent nécessiter des modèles distincts contenant des variables auxiliaires différentes. En outre, un beaucoup plus grand nombre de variables auxiliaires sont disponibles pour corriger le biais dû à la non-réponse à l'enquête. Pour de nombreux panels en ligne, les nouveaux membres répondent à un sondage en vue d'établir leur profil. Il s'agit d'un questionnaire initial destiné à recueillir des renseignements démographiques de base. Toutes ces variables peuvent être utilisées pour pondérer les données recueillies en réponse à l'enquête. Souvent, un moins grand nombre de variables auxiliaires sont disponibles pour le redressement par pondération à la phase de recrutement. En résumé, dans un panel en ligne, le redressement par pondération est un processus en deux étapes :

1) calculer les poids pour tous les membres du panel de telle façon que le panel devienne représentatif de la population cible;
2) pour chaque enquête, calculer les poids de telle façon que l'enquête devienne représentative du panel.

Les poids finaux sont obtenus en multipliant les poids au recrutement par les poids de l'enquête.

3. Un panel en ligne pilote

3.1 Recrutement

Statistics Netherlands a réalisé un essai pilote de panel en ligne en 2012. L'objectif principal était d'acquérir une première expérience de la création de panels en ligne. Comme l'organisme disposait de peu de temps et de ressources limitées, la décision a été prise de recruter les membres du panel parmi les répondants à une enquête existante, OViN (*Onderzoek Verplaatsingen in Nederland*), qui est une enquête sur la mobilité. L'échantillon de cette enquête a été tiré aléatoirement du registre de la population. À la fin du questionnaire, on a demandé à tous les répondants s'ils étaient disposés à participer à d'autres enquêtes de Statistics Netherlands. Ceux dont la réponse était positive (et qui avaient accès à Internet) ont été invités par courriel à devenir membres du panel en ligne pour une période d'un an et à remplir un questionnaire chaque mois.

Statistics Netherlands pouvait relier l'échantillon de l'enquête OViN à plusieurs registres. Donc, il existait un grand ensemble de variables auxiliaires. Les valeurs de ces variables devenaient disponibles pour les membres du panel ainsi que pour les personnes ne figurant pas dans le panel. Par conséquent, ces variables auxiliaires ont permis d'analyser la représentativité du panel.

L'échantillon pour le panel en ligne a été sélectionné parmi les participants à l'enquête OViN. Par conséquent, la population cible du panel en ligne était les répondants à l'enquête OViN et non la population générale. Cet aspect a été pris en compte lors de l'étude de la représentativité du panel en ligne.

Le processus de recrutement pour le panel en ligne est résumé au tableau 3.1.1. L'échantillon original de l'enquête OViN comprenait 12 406 personnes sélectionnées au hasard dans le registre de population. Le taux de réponse à l'enquête OViN était de 57,5 %, ce qui correspond à 6 928 personnes. De ces personnes, 4 251 (35,3 % de l'échantillon original) ont accepté de participer au panel. De ces personnes prêtes à participer, seulement 1 231 se sont réellement inscrites au panel, et seulement 1 134 ont réellement rempli le questionnaire de la première enquête réalisée à partir du panel. Donc, en fin de compte, seulement 9,4 % de l'échantillon initial de l'enquête OViN sont devenus actifs dans le panel. Il s'agit d'un résultat faible qui a suscité des préoccupations au sujet de la représentativité du panel.

Le tableau 3.1.1 contient l'indicateur R pour plusieurs étapes du processus de recrutement. La réponse à l'enquête OViN présentait un défaut de représentativité, car l'indicateur R était de 0,78. Étonnamment, l'indicateur R a augmenté pour atteindre 0,88 durant les étapes subséquentes. Le groupe restant est devenu plus représentatif à mesure que d'autres personnes abandonnaient le panel. Donc, la composition finale du panel était plus équilibrée que le groupe initial de répondants à l'enquête OViN.

Tableau 3.1.1

Processus de recrutement

Phase	Taille	Taux de réponse	Indicateur R
Échantillon OViN	12 406		
Réponse OViN	6 928	57,5 %	0,78
Disposé à entrer dans le panel	4 251	35,3 %	0,84
Sélectionné pour le panel	4 227	35,1 %	0,84
Inscrit au panel	1 231	10,2 %	0,88
Participant au panel	1 134	9,4 %	

Pour évaluer l'utilité du panel en ligne, nous avons analysé deux variables d'intérêt : le *niveau de scolarité* et l'*activité principale*. Toutes les valeurs pour ces deux variables étaient disponibles, parce qu'il était possible de relier l'échantillon de l'enquête OViN à des sources administratives. Le tableau 3.1.2 compare la distribution du niveau de scolarité dans le panel à la distribution correspondante dans l'échantillon OViN (la population cible). Le pourcentage de personnes ayant un faible niveau de scolarité dans le panel est de 2,6 % seulement, alors qu'il devrait être beaucoup plus élevé (5,5 %). Les personnes ayant un niveau élevé de scolarité semblent être surreprésentées dans le panel. Le pourcentage de personnes ayant un haut niveau de scolarité est de 45,5 % dans le panel, alors qu'il ne devrait être que de 33,6 %.

Tableau 3.1.2

Estimations (non pondérées et pondérées) pour le niveau de scolarité (%)

Niveau de scolarité	Panel	Pondéré	OViN
Primaire	2,6	4,3	5,5
Secondaire, premier cycle	15,2	16,5	21,0
Secondaire, deuxième cycle	34,4	35,8	37,6
Baccalauréat/maîtrise	45,5	40,6	33,6

Afin de déterminer si la pondération pouvait améliorer les estimations, nous avons construit un modèle de pondération contenant les variables *âge*, *revenu du ménage* et *statut socioéconomique*. Ces variables auxiliaires ont été choisies en raison de leur corrélation avec les variables d'intérêt et les probabilités de réponse. Le tableau 3.1.2 montre que la pondération a amélioré quelque peu les estimations. Elle s'approchait davantage des vraies valeurs de l'enquête OViN. Néanmoins, des biais importants persistaient.

Tableau 3.1.3

Estimations (non pondérées et pondérées) pour l'activité principale (%)

Activité principale	Panel	Pondérée	OViN
Femme/homme au foyer	11,9	12,2	12,5
Retraité(e)	16,8	17,8	14,7
À l'école/étudiant(e)	6,1	10,6	9,8
Handicapé(e)	2,4	2,8	2,8
Chômeur(se)	1,9	2,1	2,3
Occupé(e)	59,2	52,4	56,1

Le tableau 3.1.3 donne les résultats pour la variable d'intérêt *activité principale*. Les retraité(e)s et les personnes occupées sont sur-représentés, et les étudiants sont sous-représentés. Les effets de la pondération sont variables. Les

estimations pour certaines catégories (femme/homme au foyer, à l'école/étudiant(e), handicapé(e), chômeur(se)) s'améliorent. Les estimations pondérées sont plus proches de la vraie valeur. L'effet opposé est observé pour les retraité(e)s et les personnes occupées : les estimations pondérées s'écartent davantage de la vraie valeur.

3.2 Conclusion

Le panel en ligne pilote a montré que ce mode de collecte des données ne peut être utilisé que pour un nombre limité d'enquêtes de Statistics Netherlands. Les enquêtes par IPAO ne peuvent pas toutes être converties en enquêtes en ligne. Par exemple, certains questionnaires d'enquête sont simplement trop longs. En outre, le tirage d'échantillons parmi les répondants à des enquêtes par IPAO ou par ITAO antérieures donne des panels manquant sérieusement de représentativité. Qui plus est, les coûts sont importants. Les coûts de recrutement sont particulièrement élevés. Il en existe d'autres, comme les coûts annuels de maintien du panel (par exemple, pour qu'il demeure représentatif) et les coûts par enquête. La décision a donc été prise de ne pas adopter un panel en ligne pour le moment.

Bibliographie

Baker, R., Blumberg, S.J., Brick, J.M., Couper, M.P., Courtright, M., Dennis, J.M., Dillman, D., Frankel,, M.R., Garland, P.,Groves, R.M., Kennedy, C., Krosnick, J., Lavrakas, P.J., Lee, S., Link, M., Piekarski, L., Rao, K., Thomas, R.K. et Zahs, D. (2010), « Research Synthesis: AAPOR Report on online panels », *Public Opinion Quarterly*, 74, p. 711 à 781.

Bethlehem, J.G. (2010), « Selection Bias in Web Surveys », *International Statistical Review*, 78, p. 161 à 188.

Bethlehem, J.G. (2013), *De kwaliteit van internetpeilingen*, discours inaugural, Université de Leiden : Leiden.

Bethlehem, J.G. et Biffignandi, S. (2012), *Handbook of web surveys*, Hoboken: John Wiley & Sons.

Bethlehem, J.G., Cobben, F. et Schouten, B. (2011), *Handbook of nonresponse in household surveys*, Hoboken: John Wiley & Sons.

Bronzwaer, S. (2012), « Infiltranten probeerden de peilingen van Maurice de Hond te manipuleren », *NRC*, 13 septembre 2012.

Cook, C., Heath, F. et Thompson, R.L. (2000), « A meta-analysis of response rates in web- or internet-based surveys », *Education and Psychological Measurement*, 60, p. 821 à 836.

Kaplowitz, M.D., Hadlock, T.D. et Levine, R. (2004), « A comparison of web and mail survey response rates », *Public Opinion Quarterly*, 68, p. 94 à 101.

Knowledge Networks (2012), *KnowledgePanel design summary*, www.knowledgenetworks.com.

Leenheer, J. et Scherpenzeel, A.C. (2013), « Does it pay off to include non-internet households in an internet panel? », *International Journal of Internet Science*, 8, p. 17 à 29.

Lozar Manfreda, K., Bosnjak, M., Berzelak, J., Haas, I. et Vehovar, V. (2008), « Web surveys versus other survey modes – A meta-analysis comparing response rates », *International Journal of Market Research*, 50, p. 79 à 104.

Lynn, P. (2013), « Issues of coverage and sampling in web surveys for the general population: an overview ». Communication présentée à la conférence d'ouverture du NCRM Network of Methodological Innovation, Web Surveys for the general population: how, why and when?

Särndal, C.E. et Lundström, S. (2005), *Estimation in surveys with nonresponse*, Chichester: John Wiley & Sons.

Scherpenzeel, A. (2008), « An Online Panel as a Platform for Multi-Disciplinary Research ». Dans I. Stoop et M. Wittenberg (sous la dir. de), *Access panels and online research, panacea or pitfall?*, Aksant: Amsterdam, p. 101 à 106.

Schouten, B., Cobben, F. et Bethlehem, J.G. (2009), « Indicateurs de la représentativité de la réponse aux enquêtes », *Techniques d'enquête*, 35, vol. 1, p. 107 à 120.

Scherpenzeel, A. et Schouten, B. (2011), « LISS Panel R-indicator: representativity in different stages of recruitment and participation of an internet panel ». Communication présentée au XX^{e} atelier international sur la non-réponse aux enquêtes auprès des ménages, Bilbao, Espagne.