Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d'enquête

Suivi de la non-réponse aux enquêtes auprès des entreprises

par Elisabeth Neusy, Jean-François Beaumont, Wesley Yung,
Mike Hidiroglou et David Haziza

Date de diffusion : le 21 juin 2022

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Techniques d'enquête, juin 2022
Vol. 48, n° 1, pp. 103-128
Statistique Canada, n° 12-001-X au catalogue

Suivi de la non-réponse aux enquêtes auprès des entreprises

Elisabeth Neusy, Jean-François Beaumont, Wesley Yung, Mike Hidiroglou et David Haziza[1]

Résumé

Au cours des deux dernières décennies, les taux de réponse aux enquêtes ont régulièrement diminué. Dans ce contexte, il est devenu de plus en plus important pour les organismes statistiques d'élaborer et d'utiliser des méthodes permettant de réduire les effets négatifs de la non-réponse sur l'exactitude des estimations découlant d'enquêtes. Le suivi des cas de non-réponse peut être un remède efficace, même s'il exige du temps et des ressources, pour pallier le biais de non-réponse. Nous avons mené une étude par simulations à l'aide de données réelles d'enquêtes-entreprises, afin de tenter de répondre à plusieurs questions relatives au suivi de la non-réponse. Par exemple, en supposant un budget fixe de suivi de la non-réponse, quelle est la meilleure façon de sélectionner les unités non répondantes auprès desquelles effectuer un suivi ? Quel effort devons-nous consacrer à un suivi répété des non-répondants jusqu'à la réception d'une réponse ? Les non-répondants devraient-ils tous faire l'objet d'un suivi ou seulement un échantillon d'entre eux ? Dans le cas d'un suivi d'un échantillon seulement, comment sélectionner ce dernier ? Nous avons comparé les biais relatifs Monte Carlo et les racines de l'erreur quadratique moyenne relative Monte Carlo pour différents plans de sondage du suivi, tailles d'échantillon et scénarios de non-réponse. Nous avons également déterminé une expression de la taille de l'échantillon de suivi minimale nécessaire pour dépenser le budget, en moyenne, et montré que cela maximise le taux de réponse espéré. Une principale conclusion de notre expérience de simulation est que cette taille d'échantillon semble également réduire approximativement le biais et l'erreur quadratique moyenne des estimations.

Mots-clés : Non-réponse; suivi; enquêtes auprès des entreprises.

1. Introduction

La recherche en matière de collecte de données est un sujet d'intérêt au sein des organismes statistiques nationaux souhaitant accroître les taux de réponse ou réduire les coûts de la collecte de données. Compte tenu des coûts élevés que représente la collecte de données d'enquête, même une petite augmentation de l'efficacité des procédures de collecte de données peut se traduire par des économies monétaires appréciables. Étant donné que les taux de réponse ont diminué au cours des 20 dernières années, tant pour les enquêtes sociales que pour les enquêtes économiques, le biais de non-réponse a également suscité des préoccupations croissantes.

Dans l'une des premières études traitant de la non-réponse, Hansen et Hurwitz (1946) ont proposé de sélectionner un sous-échantillon de non-répondants, également appelé « un échantillon de suivi de la non-réponse », afin d'éliminer le biais de non-réponse. Cette procédure était la suivante : des questionnaires étaient envoyés par la poste et, après un certain temps, des intervieweurs procédaient à un suivi personnel auprès d'un échantillon de non-répondants, afin d'obtenir leurs réponses. Ils ont montré la façon dont les réponses à l'envoi par la poste initial pouvaient être combinées à celles de l'échantillon de suivi de non-réponse pour obtenir un estimateur sans biais d'un total ou d'une moyenne de population. Ils ont postulé une hypothèse forte voulant que chaque unité de l'échantillon de suivi réponde. Toutefois, dans

1. Elisabeth Neusy, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6. Courriel : elisabeth.neusy@statcan.gc.ca. Jean-François Beaumont, Wesley Yung et Mike Hidiroglou, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6. David Haziza, University of Ottawa, 150, Louis-Pasteur Private, Ottawa (Ontario), Canada, K1N 6N5.

l'environnement d'aujourd'hui, cette hypothèse n'est pas réaliste, car les entreprises et les particuliers sont de plus en plus réticents à répondre aux enquêtes.

La plupart des études publiées ces 15 dernières années ont porté sur des plans de collecte adaptatifs, également appelés « plans d'enquête adaptatifs », « plans de sondage adaptatifs », « plans de collecte dynamiques », « conception d'enquête adaptative » ou simplement « plans adaptatifs ». Groves et Heeringa (2006) ont défini un plan de sondage adaptatif comme un plan qui utilise des paradonnées, ou des données du processus de collecte, pour apporter des modifications aux procédures de la collecte de données afin d'obtenir des estimations de meilleure qualité par coût unitaire. Beaumont, Bocci et Haziza (2014) ont fait remarquer que la littérature relative aux plans de collecte adaptative était axée principalement sur l'élaboration de procédures visant à réduire le biais de non-réponse d'un estimateur non ajusté pour la non-réponse (voir, par exemple, Schouten, Cobben et Bethlehem, 2009; Peytchev, Riley, Rosen, Murphy et Lindblad, 2010). Beaumont et coll. (2014) ont fait valoir que toute information (par exemple les données auxiliaires, les paradonnées) pouvant être utilisée au cours de la collecte de données pour réduire le biais de non-réponse peut être également utilisée à l'étape de l'estimation. En d'autres termes, le biais dû à la non-réponse pouvant être supprimé à l'étape de la collecte par une procédure de collecte adaptative peut également être supprimé à l'étape de l'estimation par des ajustements appropriés des poids pour la non-réponse. Ils ont suggéré que les procédures de collecte adaptatives, comme la priorisation des appels, ne peuvent pas réduire le biais de non-réponse dans une plus grande mesure qu'un ajustement adéquat des poids pour la non-réponse. Tourangeau, Brick, Lohr et Li (2017) ont également souligné, dans leur article de synthèse, les limites des procédures de collecte adaptatives pour réduire le biais de non-réponse et les coûts.

Jusqu'à maintenant, les ouvrages publiés portant sur des travaux de recherche relatifs à la collecte ont principalement ciblé les enquêtes auprès des ménages, et il existe peu d'études sur ce sujet pour les enquêtes auprès des entreprises, à deux exceptions près : Bosa, Godbout, Mills et Picard (2018) et Thompson, Kaputa et Bechtel (2018). Bosa et coll. (2018) ont développé un score propre à chaque variable reflétant l'importance de suivre une unité d'échantillonnage donnée et ont suggéré une procédure de collecte adaptative reposant sur ce score. Les unités exhibant un score élevé contribuent le plus à réduire la variance des estimateurs ponctuels. La priorité est donnée à ces unités dans un contexte d'opérations de collecte coûteuses, comme le suivi téléphonique. Thompson et coll. (2018) se sont penchés sur le sous-échantillonnage des non-répondants et ont étudié le problème de la répartition du sous-échantillon soumise à certaines contraintes appliquées au taux de réponse et à la taille de l'échantillon dans des domaines d'intérêt prédéterminés.

Même si les enquêtes-entreprises reposent généralement sur des plans de sondage simples, comme des plans de sondage aléatoire simple stratifié ou des plans d'échantillonnage de Bernoulli, elles présentent certaines caractéristiques qui posent des défis en matière de collecte. Une caractéristique particulière est que les populations des entreprises sont grandement asymétriques, un faible pourcentage d'entreprises représentant la majeure partie des activités économiques. Par conséquent, les enquêtes-entreprises

comportent généralement une strate à tirage complet, au sein de laquelle toutes les unités sont sélectionnées avec certitude, et une strate à tirage partiel, au sein de laquelle les unités sont généralement sélectionnées à l'aide d'un échantillonnage aléatoire simple sans remise ou d'un échantillonnage de Bernoulli. Les unités dans la strate à tirage complet correspondent aux grandes entreprises. Ne pas obtenir de réponse de ces grandes entreprises pourrait aboutir à des estimations présentant un biais important. Par conséquent, toutes ces unités font généralement l'objet d'un suivi, et des efforts sont déployés pour assurer la réception de leurs réponses. Les grandes entreprises disposent généralement d'un personnel (par exemple des comptables) capable de répondre aux variables du questionnaire. En revanche, les petites entreprises peuvent devoir payer un comptable externe afin d'obtenir les renseignements demandés; cela pourrait être un facteur contribuant à la non-réponse pour ces entreprises. Une autre caractéristique des enquêtes-entreprises est que la collecte est généralement effectuée en deux étapes. Tous d'abord, des lettres sont envoyées aux unités d'échantillonnage par la poste ou par courriel, les invitant à remplir un questionnaire électronique en ligne. Après un certain temps, on entreprend un suivi des unités non répondantes par interview téléphonique assistée par ordinateur.

Dans la présente étude, nous nous concentrons sur la strate à tirage partiel et tentons de répondre aux questions suivantes : i) Pour un budget de suivi fixe, quel effort devrions-nous consacrer à un suivi répété des non-répondants jusqu'à l'obtention d'une réponse ? ii) Devrions-nous effectuer un suivi auprès de tous les non-répondants ou en sélectionner un échantillon ? iii) Dans le cas de la sélection d'un échantillon de non-répondants, quels plans de sondage mèneraient à des estimateurs plus efficaces ? À notre connaissance, la détermination d'une taille d'échantillon et d'un plan de sondage du suivi appropriés n'a pas été étudiée dans la littérature.

Dans le reste de l'article, nous présentons nos analyses sur le suivi de la non-réponse dans le contexte des enquêtes-entreprises. La stratégie de suivi proposée, qui consiste en un plan de sondage du suivi, une procédure de collecte des données et un estimateur, est présentée à la section 2. À la section 3, nous indiquons quelques propriétés théoriques de la stratégie de suivi proposée. La section 4 rend compte d'une étude par simulations menée pour étudier les propriétés de l'estimateur de Hansen-Hurwitz, ajusté pour la non-réponse, d'un total de population selon différents plans de sondage du suivi et scénarios de réponse. Enfin, la section 5 présente un résumé de nos principales conclusions. Même si nous focalisons sur les enquêtes-entreprises, nous estimons que la plupart de nos conclusions s'appliquent également aux enquêtes sociales.

2. Stratégie de suivi proposée

Supposons une population finie U de N unités, réparties en L strates, $U_1,\ldots,U_h,\ldots,U_L$, respectivement de tailles $N_1,\ldots,N_h,\ldots,N_L$, tel que $U=\bigcup_{h=1}^{L}U_h$ et $N=\sum_{h=1}^{L}N_h$. Nous souhaitons estimer le total de population $Y=\sum_{h=1}^{L}\sum_{i\in U_h}y_{hi}$, où y_{hi} est la valeur de la variable d'intérêt y pour $i\in U_h$. Dans chaque strate U_h, un échantillon s_{1h} de taille n_{1h} est sélectionné selon un échantillonnage

aléatoire simple sans remise. L'échantillon total obtenu, $s_1 = \bigcup_{h=1}^{L} s_{1h}$, est de taille n_1. Nous désignons par $\pi_{1hi} = n_{1h}/N_h$ la probabilité que l'unité $i \in U_h$ soit sélectionnée dans s_{1h}. Les n_{1h} unités échantillonnées dans la strate h sont invitées, par la poste ou par courriel, à remplir un questionnaire électronique en ligne. Nous appelons cela « l'envoi par la poste ». Si toutes les unités échantillonnées répondent à l'envoi par la poste, il est possible d'utiliser l'estimateur par dilatation sans biais de Y, également appelé « l'estimateur d'échantillon complet » :

$$\hat{Y}_{\text{COMPLET}} = \sum_{h=1}^{L} \sum_{i \in s_{1h}} w_{1hi} y_{hi}, \tag{2.1}$$

où $w_{1hi} = 1/\pi_{1hi}$ désigne les poids de sondage associés à $i \in s_{1h}$.

En pratique, les unités échantillonnées ne répondent pas toutes à l'envoi par la poste. Supposons qu'après un certain temps, n_{1hr} des n_{1h} unités échantillonnées répondent dans la strate h. Nous désignons l'ensemble de répondants de la strate h par s_{1hr} et la probabilité de réponse pour l'unité $i \in s_{1h}$ par p_{1hi}. Un échantillon de n_2 unités, s_2, est alors sélectionné dans l'ensemble de tous les non-répondants à l'envoi postal, $s_{1,\text{nr}}$. Nous désignons par s_{2h} l'ensemble de n_{2h} unités sélectionnées pour un suivi dans la strate h au sein de l'ensemble de non-répondants à l'envoi postal de la strate h, $s_{1h,\text{nr}}$. Nous désignons la probabilité que le non-répondant de l'envoi par la poste $i \in s_{1h,nr}$ soit sélectionné dans l'échantillon de suivi s_2 par π_{2hi}. Nous supposons que cette probabilité puisse s'écrire sous la forme $\pi_{2hi} = n_2 \pi^*_{2hi}$, où π^*_{2hi} ne dépend pas de la taille de l'échantillon de suivi n_2 et satisfasse à la condition suivante :

$$\sum_{h=1}^{L} \sum_{i \in s_{1h,\text{nr}}} \pi^*_{2hi} = 1. \tag{2.2}$$

Cette condition est satisfaite pour l'échantillonnage aléatoire simple, l'échantillonnage aléatoire simple stratifié, avec répartition proportionnelle ou répartition de Neyman, et l'échantillonnage avec probabilité proportionnelle à la taille.

Les unités de l'échantillon s_2 font l'objet d'un suivi téléphonique. Si toutes les n_{2h} unités répondent au suivi, $h = 1, \ldots, L$, l'estimateur de Hansen et Hurwitz sans biais (1946) du total Y peut être utilisé :

$$\hat{Y}_{\text{HH}} = \sum_{h=1}^{L} \sum_{i \in s_{1hr}} w_{1hi} y_{hi} + \sum_{h=1}^{L} \sum_{i \in s_{2h}} w_{1hi} w_{2hi} y_{hi}, \tag{2.3}$$

où $w_{2hi} = 1/\pi_{2hi}$ est le poids de sondage du suivi pour l'unité $i \in s_{2h}$. L'objectif de l'échantillon s_2 est d'estimer le total inconnu $\sum_{h=1}^{L} \sum_{i \in s_{1h,nr}} w_{1hi} y_{hi}$. Si une variable x fortement associée à la variable d'intérêt y est disponible avant la sélection de l'échantillon de tous les non-répondants à l'envoi postal, il semble naturel d'utiliser $w_{1hi} x_{hi}$ comme variable auxiliaire pour la stratification ou comme mesure de la taille pour l'échantillonnage avec probabilité proportionnelle à la taille.

Comme l'a signalé un arbitre, il est important d'attendre la fin de la collecte de données de l'envoi par la poste avant de sélectionner l'échantillon de suivi. Si des unités répondent à l'envoi par la poste après la sélection de l'échantillon de suivi, il est nécessaire de prendre des décisions quant à la gestion de ces

répondants tardifs. S'ils ne sont pas éliminés, il peut être difficile d'obtenir un estimateur sans biais comme (2.3) sans introduire des hypothèses de modèle (voir Beaumont, Bocci et Hidiroglou, 2014). Ce problème peut également avoir une incidence sur la longueur de la période de collecte.

Comme cela est mentionné dans l'introduction, il est peu probable que toutes les unités de l'échantillonnage de suivi répondent. Supposons qu'après la fin de la période de collecte de données, n_{2hr} unités ont répondu au suivi dans la strate h. Nous désignons par s_{2hr} l'ensemble des n_{2hr} répondants de la strate h. Nous utilisons la version de l'estimateur de Hansen et Hurwitz (1946) ajusté pour la non-réponse :

$$\hat{Y}_{\text{HH}-\text{NA}} = \sum_{h=1}^{L} \sum_{i \in s_{1hr}} w_{1hi} y_{hi} + \sum_{h=1}^{L} \sum_{i \in s_{2hr}} w_{1hi} w_{2hi} a_{2hi} y_{hi}, \tag{2.4}$$

où a_{2hi} est un ajustement de poids pour la non-réponse. Dans le cas d'une non-réponse uniforme, un ajustement de poids approprié est l'inverse du taux de réponse pondéré global :

$$a_{2hi} = a_2 = \frac{\sum_{h=1}^{L} \sum_{j \in s_{2h}} w_{1hj} w_{2hj}}{\sum_{h=1}^{L} \sum_{j \in s_{2hr}} w_{1hj} w_{2hj}}, \quad i \in s_{2hr}, h = 1, \ldots, L. \tag{2.5}$$

Une hypothèse moins restrictive est une non-réponse uniforme au sein d'une strate. Dans ce cas, un ajustement de poids approprié serait l'inverse du taux de réponse pondéré dans la strate :

$$a_{2hi} = a_{2h} = \frac{\sum_{j \in s_{2h}} w_{2hj}}{\sum_{j \in s_{2hr}} w_{2hj}}, \quad i \in s_{2hr}, h = 1, \ldots, L. \tag{2.6}$$

Il convient de mentionner que l'ajustement des poids pour la non-réponse (2.6) peut uniquement être calculé si $n_{2hr} > 0$ pour toutes les strates. Sinon, des versions non pondérées de (2.5) et (2.6) peuvent également être envisagées.

Comme cela a été mentionné plus tôt, le suivi des non-répondants sélectionnés dans s_2 s'effectue par téléphone. Dans la procédure de collecte des données proposée, une liste d'appels est d'abord créée en classant aléatoirement les unités de s_2. On appelle ensuite séquentiellement ces unités jusqu'à ce que la liste d'appels soit vide ou que le budget de suivi soit entièrement épuisé, selon ce qui se produit en premier. Chaque tentative d'appel effectuée auprès des unités de s_2 aboutit à l'un des trois résultats suivants :

1. Réponse : une réponse est obtenue de l'unité. L'unité est retirée de la liste d'appels, afin de ne pas la rappeler.
2. Non-réponse finale : l'unité est classée comme étant un « non-répondant »; elle ne devrait pas être rappelée et est retirée de la liste d'appels. L'exemple le plus courant de ce résultat est le refus de répondre à l'enquête.

3. Toujours en cours : le traitement de l'unité n'est pas terminé et l'unité doit être rappelée; elle est donc replacée à la fin de la liste d'appels. Un exemple de ce résultat est une tentative n'ayant pas permis de prise de contact ou une prise de contact menant à un rendez-vous de rappel.

Les issues de « réponse » et de « non-réponse finale » sont tous deux des issues définitives, au sens où l'unité est retirée de la liste d'appels et du processus de collecte, contrairement à l'une « toujours en cours », pour lequel l'unité est replacée dans la liste d'appels afin d'être rappelée. Une unité pour laquelle le processus de collecte des données dont l'issue est « réponse » ou « non-réponse finale » est désignée comme étant finalisée ou résolue; dans le cas contraire, le cas n'est pas résolu. Il existe deux types de non-répondants après la collecte de données : i) des unités dont le traitement est finalisé et ayant un résultat de « non-réponse finale »; et ii) des unités non résolues. Ces deux types de non-répondants sont pris en compte dans l'estimation fondée sur l'estimateur ajusté pour la non-réponse (2.4).

Nous supposons que, pour une unité échantillonnée donnée, les résultats des tentatives d'appels sont indépendants et que la probabilité associée à chacun des trois résultats possibles demeure constante tout au long de la période de collecte de données complète. Pour une unité échantillonnée donnée $i \in s_{2h}$, $h = 1, \ldots, L$, la probabilité d'une « réponse » est désignée par $P_{2hi}^{(1)}$; la probabilité d'une « non-réponse finale », par $P_{2hi}^{(2)}$; et la probabilité d'un résultat « toujours en cours », par $P_{2hi}^{(3)}$. En pratique, l'hypothèse d'indépendance et l'hypothèse de probabilité constante pourraient ne pas être entièrement vérifiées. On s'attend à ce que l'hypothèse d'indépendance soit plus plausible, si les probabilités sont conditionnelles à des prédicteurs puissants et si le délai entre deux tentatives d'appel successives pour la même unité n'est pas trop court. L'hypothèse de probabilité constante ne se vérifie pas lorsque les probabilités dépendent de prédicteurs pouvant varier au cours de la collecte de données, comme l'heure de la journée ou le jour de la semaine où a lieu la tentative d'appel. Même s'il peut être possible d'étendre notre modèle à des prédicteurs variant dans le temps, cela compliquerait nos développement théoriques et l'étude par simulations. Ces hypothèses sont retenues tout au long de l'article dans le but de simplifier nos analyses. Il s'agit d'une limite de nos analyses dont il faudra tenir compte lors de l'interprétation des résultats.

Plusieurs tentatives d'appels téléphoniques peuvent être nécessaires pour entrer en contact avec une unité et résoudre le cas. Les gestionnaires de la collecte de données peuvent souhaiter imposer une limite supérieure au nombre de tentatives d'appels à effectuer pour toute unité d'échantillonnage de suivi. Si une unité est toujours en cours malgré l'atteinte de cette limite, elle est retirée de la liste d'appels et le cas demeure non résolu à la fin de la collecte de données. Soit K la limite supérieure du nombre de tentatives d'appels. En supposant que chaque unité non résolue à la fin de la collecte de données atteigne toujours le nombre maximal de tentatives K, la probabilité que l'unité $i \in s_{1h,\mathrm{nr}}$ réponde lorsqu'elle est sélectionnée dans l'échantillon s_2 peut s'écrire $p_{2hi}(K) = \sum_{k=1}^{K} p_{2hik}$, où p_{2hik} est la probabilité que l'unité $i \in s_{1h,\mathrm{nr}}$ réponde à la k^{e} tentative lors de sa sélection dans s_2. Selon ces hypothèses, il est aisé de constater que $p_{2hik} = \left(P_{2hi}^{(3)}\right)^{k-1} P_{2hi}^{(1)}$. Nous obtenons donc :

$$\begin{aligned} p_{2hi}(K) &= \sum_{k=1}^{K} p_{2hik} \\ &= P_{2hi}^{(1)} \sum_{k=0}^{K-1} \left(P_{2hi}^{(3)}\right)^k \\ &= P_{2hi}^{(1)} \frac{1-\left(P_{2hi}^{(3)}\right)^K}{1-P_{2hi}^{(3)}}. \end{aligned} \tag{2.7}$$

Dans la section suivante, l'équation (2.7) sera utilisée pour déterminer une taille d'échantillon de suivi appropriée.

3. Quelques propriétés théoriques de la stratégie de suivi proposée

Soit C le budget total consacré au suivi de la non-réponse, pouvant être défini en termes d'unités monétaires ou temporelles. Un coût est enregistré pour chaque tentative d'appel et dépend du résultat de l'appel. Nous désignons respectivement par $c^{(1)}$, $c^{(2)}$ et $c^{(3)}$, le coût par tentative d'appel pour une issue de « réponse », de « non-réponse finale » et « toujours en cours ». Pour simplifier notre développement, nous supposons que ces coûts sont les mêmes pour chaque unité d'échantillonnage et ne varient pas au cours de la collecte de données. Supposons que $c_{hi} = \sum_{k=1}^{K} c_{hik}$ soit le coût d'une unité résolue $i \in s_{2h}$ ou atteignant le nombre maximal de tentatives d'appels pour cette unité, où c_{hik} est le coût de la k^{e} tentative d'appel pour l'unité $i \in s_{2h}$. Si une unité $i \in s_{2h}$ est résolue à la l^{e} tentative, c_{hik} est définie comme zéro pour toutes les valeurs $k > l$. Par conséquent, le coût c_{hik} est soit nul, si l'unité $i \in s_{2h}$ a été résolue avant la k^{e} tentative, soit $c^{(1)}$, $c^{(2)}$ ou $c^{(3)}$, selon le résultat de l'appel. Pour une taille d'échantillon donnée n_2 et une valeur fixe de K, le coût total du suivi, $\sum_{h=1}^{L} \sum_{i \in s_{2h}} c_{hi}$, est une variable aléatoire lorsque chaque unité d'échantillonnage fait l'objet d'un suivi jusqu'à sa résolution ou jusqu'au nombre maximal de tentatives d'appel. En tenant compte des attentes quant au coût total en fonction du plan de sondage du suivi et du mécanisme de non-réponse, conditionnel à $s_{1,\text{nr}}$, nous obtenons le coût de suivi espéré :

$$\tilde{C}(n_2, K) = \sum_{h=1}^{L} \sum_{i \in s_{1h,\text{nr}}} \pi_{2hi} \tilde{c}_{hi}(K), \tag{3.1}$$

où $\tilde{c}_{hi}(K) = \sum_{k=1}^{K} \tilde{c}_{hik}$ est le coût espéré d'une unité résolue $i \in s_{1h,\text{nr}}$ ou atteignant le nombre maximal de tentatives d'appels lorsque cette unité est sélectionnée dans s_2 et $\tilde{c}_{hik}$ est le coût espéré de la k^{e} tentative d'appel, $k \le K$, pour cette unité. En considérant que $c_{hik} \neq 0$ uniquement si l'unité i n'a pas été résolue avant la k^{e} tentative, il est aisé de constater que le coût espéré $\tilde{c}_{hik}$ est :

$$\tilde{c}_{hik} = \left(P_{2hi}^{(3)}\right)^{k-1} \left(c^{(1)} P_{2hi}^{(1)} + c^{(2)} P_{2hi}^{(2)} + c^{(3)} P_{2hi}^{(3)}\right).$$

Le coût espéré $\tilde{c}_{hi}(K)$ se réduit à :

$$\begin{aligned} \tilde{c}_{hi}(K) &= \sum_{k=1}^{K} \tilde{c}_{hik} \\ &= \left(c^{(1)} P_{2hi}^{(1)} + c^{(2)} P_{2hi}^{(2)} + c^{(3)} P_{2hi}^{(3)}\right) \sum_{k=0}^{K-1} \left(P_{2hi}^{(3)}\right)^k \\ &= \left(c^{(1)} P_{2hi}^{(1)} + c^{(2)} P_{2hi}^{(2)} + c^{(3)} P_{2hi}^{(3)}\right) \frac{1-\left(P_{2hi}^{(3)}\right)^K}{1-P_{2hi}^{(3)}}. \end{aligned} \tag{3.2}$$

En utilisant $\pi_{2hi} = n_2 \pi^*_{2hi}$ ainsi que la condition (2.2), nous pouvons déterminer la taille de l'échantillon de suivi nécessaire pour dépenser le budget C, en moyenne, tout en veillant à la résolution de chaque unité ou à l'atteinte du nombre maximal de tentatives, K. Nous pouvons donc déterminer la taille de l'échantillon de suivi de telle sorte que le coût espéré du suivi (3.1) soit exactement égal au budget C. Cette taille d'échantillon est :

$$n_2(C,K) = \frac{C}{\sum_{h=1}^{L} \sum_{i \in s_{1h,\mathrm{nr}}} \pi^*_{2hi} \tilde{c}_{hi}(K)}, \tag{3.3}$$

où $\tilde{c}_{hi}(K)$ est obtenu par la formule (3.2). Pour un budget fixe C, la taille de l'échantillon $n_2(C,K)$ est inversement liée à K et atteint son minimum lorsque $K = \infty$; c'est-à-dire lorsqu'il n'existe pas de limite supérieure au nombre d'appels. Cela signifie que pour un coût C, choisir une taille d'échantillon supérieure à $n_2(C,\infty)$ a un effet similaire à réduire la valeur de K, augmentant ainsi le nombre espéré d'unités non résolues. De plus, si l'on choisit une taille d'échantillon inférieure à $n_2(C,\infty)$, le coût espéré (3.1) est inférieur au budget C; c'est-à-dire qu'en moyenne, le budget n'est pas entièrement dépensé. La taille d'échantillon $n_2(C,\infty)$ est par conséquent la taille d'échantillon minimale permettant de dépenser le budget C, en moyenne.

Pour la taille d'échantillon $n_2(C,K)$ dans (3.3), le nombre espéré de répondants à l'enquête de suivi est :

$$\begin{aligned} \tilde{n}_{2r}(C,K) &= \sum_{h=1}^{L} \sum_{i \in s_{1h,\mathrm{nr}}} \pi_{2hi} p_{2hi}(K) \\ &= C \frac{\sum_{h=1}^{L} \sum_{i \in s_{1h,\mathrm{nr}}} \pi^*_{2hi} p_{2hi}(K)}{\sum_{h=1}^{L} \sum_{i \in s_{1h,\mathrm{nr}}} \pi^*_{2hi} \tilde{c}_{hi}(K)}, \end{aligned} \tag{3.4}$$

où $p_{2hi}(K)$ est obtenu dans (2.7) et le taux de réponse espéré est :

$$\frac{\tilde{n}_{2r}(C,K)}{n_2(C,K)} = \sum_{h=1}^{L} \sum_{i \in s_{1h,\mathrm{nr}}} \pi^*_{2hi} p_{2hi}(K). \tag{3.5}$$

Selon (2.7) et (3.5), nous observons que le taux de réponse espéré ne dépend pas du budget C et diminue à mesure que K diminue. Nous avons mentionné précédemment que choisir une taille d'échantillon supérieure à la taille d'échantillon minimale $n_2(C,\infty)$, pour un coût fixe C, avait un effet similaire à réduire la valeur de K. Par conséquent, choisir une taille d'échantillon supérieure à $n_2(C,\infty)$ aurait l'effet de réduire le taux de réponse espéré.

Nous pouvons également obtenir le nombre espéré d'unités résolues d'une façon similaire à (3.4) :

$$\tilde{n}_{2,\mathrm{res}}(C,K) = C \frac{\sum_{h=1}^{L} \sum_{i \in s_{1h,\mathrm{nr}}} \pi^*_{2hi} \left(1 - \left(P^{(3)}_{2hi}\right)^K\right)}{\sum_{h=1}^{L} \sum_{i \in s_{1h,\mathrm{nr}}} \pi^*_{2hi} \tilde{c}_{hi}(K)}. \tag{3.6}$$

Nous pouvons aisément constater que $\tilde{n}_{2,\text{res}}(C,K) \le n_2(C,K)$ et $\tilde{n}_{2,\text{res}}(C,\infty) = n_2(C,\infty)$. Si on choisit une taille d'échantillon de suivi inférieure à $n_2(C,\infty)$, le coût espéré est $\sum_{h=1}^{L}\sum_{i\in s_{1h,\text{nr}}} \pi_{2hi}\tilde{c}_{hi}(\infty) = C^*$, avec $C^* < C$, et en fonction de (3.4) et (3.6), les nombres espérés de répondants et d'unités résolues diminuent.

Si la probabilité $P_{2hi}^{(3)}$ est très proche de 1 pour quelques unités $i \in s_{1h,\text{nr}}$, $h = 1,\ldots,L$, la taille minimale d'échantillon $n_2(C,\infty)$ pourrait devenir très petite. Dans cette situation, il peut être approprié de choisir une valeur finie de K pour éviter de dépenser une trop grande portion du budget pour quelques unités. Cela réduirait le taux de réponse espéré, tel que souligné ci-dessus, et augmenterait possiblement le biais des estimateurs. Toutefois, utiliser une valeur finie de K peut également accroître de façon importante le nombre espéré de répondants et réduire la variance des estimateurs. Représenter graphiquement le taux de réponse espéré et le nombre espéré de répondants sous forme de fonction de K peut être utile pour déterminer un compromis acceptable entre la maximisation du taux de réponse espéré $(K=\infty)$ et la maximisation du nombre de répondants espéré, qui pourraient être atteints pour une valeur finie de K. Une faible réduction du taux de réponse espéré pourrait être tolérée, si elle entraîne une hausse importante du nombre espéré de répondants.

Dans le contexte d'une réponse de suivi uniforme, nous avons : $P_{2hi}^{(1)} = P_2^{(1)}$, $P_{2hi}^{(2)} = P_2^{(2)}$ et $P_{2hi}^{(3)} = P_2^{(3)}$, pour chaque unité $i \in s_{1h,\text{nr}}$, $h = 1,\ldots,L$. La taille de l'échantillon de suivi (3.3), le nombre espéré de répondants (3.4), le taux de réponse espéré (3.5) et le nombre espéré d'unités résolues (3.6) se réduisent comme suit :

$$n_2(C,K) = \frac{C}{\left(c^{(1)}P_2^{(1)} + c^{(2)}P_2^{(2)} + c^{(3)}P_2^{(3)}\right)} \frac{1-P_2^{(3)}}{1-\left(P_2^{(3)}\right)^K}, \tag{3.7}$$

$$\tilde{n}_{2r}(C,K) = \frac{C}{\left(c^{(1)}P_2^{(1)} + c^{(2)}P_2^{(2)} + c^{(3)}P_2^{(3)}\right)} P_2^{(1)}, \tag{3.8}$$

$$\frac{\tilde{n}_{2r}(C,K)}{n_2(C,K)} = P_2^{(1)} \frac{1-\left(P_2^{(3)}\right)^K}{1-P_2^{(3)}}, \tag{3.9}$$

et

$$\tilde{n}_{2,\text{res}}(C,K) = \frac{C}{\left(c^{(1)}P_2^{(1)} + c^{(2)}P_2^{(2)} + c^{(3)}P_2^{(3)}\right)} \left(1-P_2^{(3)}\right), \tag{3.10}$$

respectivement. Il convient de souligner que le nombre espéré de répondants (3.8) et le nombre espéré d'unités résolues (3.10) ne dépendent plus de K. Le nombre espéré d'unités résolues, $\tilde{n}_{2,\text{res}}(C,K)$, est donc égal à la taille minimale de l'échantillon permettant de dépenser le budget C, $n_2(C,\infty)$, pour chaque valeur de K. Comme nous l'avons mentionné pour le taux de réponse espéré général (3.5), le taux de réponse espéré (3.9) ne dépend pas du budget C et diminue à mesure que K diminue. Selon les observations ci-dessus, la valeur de K maximisant à la fois le taux de réponse espéré et le nombre espéré

de répondants est $K = \infty,$ dans le contexte d'une réponse uniforme, ce qui nous amène à choisir la taille d'échantillon $n_2\left(C, \infty\right).$

Les probabilités $P_{2hi}^{(1)}$, $P_{2hi}^{(2)}$ et $P_{2hi}^{(3)}$ sont inconnues. Dans la pratique, ces probabilités doivent être remplacées par des estimations des expressions ci-dessus. Parce qu'elles sont nécessaires avant de procéder à la sélection de l'échantillon de suivi et à la collecte des données, les estimations de $P_{2hi}^{(1)}$, $P_{2hi}^{(2)}$ et $P_{2hi}^{(3)}$ pourraient être obtenues de données d'enquête antérieures.

4. Étude par simulations

Nous avons mené une étude par simulations, afin d'évaluer les propriétés de l'estimateur ajusté pour la non-réponse (2.4), $\hat{Y}_{\text{HH}-\text{NA}},$ selon différents scénarios de réponse et plans de sondage du suivi.

4.1 Cadre de simulation

Données utilisées pour créer l'échantillon s_1

Les données utilisées pour l'étude par simulations sont les données-échantillons provenant d'une enquête-entreprise réelle : l'Enquête mensuelle sur les services de restauration et les débits de boissons (EMSRDB) de Statistique Canada. Comme cela est typique des enquêtes-entreprises, l'EMSRDB est stratifiée par province, industrie et revenu (une strate à tirage complet et une ou plusieurs strates à tirage partiel au sein de chaque combinaison de province et industrie). Pour obtenir de plus amples précisions sur l'EMSRDB, consulter Statistique Canada (2017). Chaque strate à tirage complet au sein d'une combinaison province-industrie comprend les grandes entreprises importantes, qui font généralement toutes l'objet d'un suivi. Ces unités sont exclues de la présente étude par simulations, afin de mettre l'accent sur la stratégie de suivi pour la strate « à tirage partiel ». L'ensemble des unités d'échantillonnage comprises dans l'étude par simulations est par conséquent l'échantillon initial de 2 375 unités sélectionné dans $L = 63$ strates à tirage partiel.

Deux variables sont utilisées pour l'étude par simulations : le « revenu » et les « ventes ». La première variable, le revenu, provient de la base de sondage (Registre des entreprises de Statistique Canada) et est disponible pour toutes les unités sélectionnées dans l'échantillon de l'EMSRDB. Nous utilisons le revenu comme variable auxiliaire, $x,$ pour échantillonner les non-répondants pour l'envoi par la poste (voir ci-dessous). La deuxième variable, les ventes, est l'une des variables recueillies lors de l'enquête; il s'agit de la variable d'intérêt $y.$ La réponse totale et la réponse partielle sont traitées par imputation dans l'EMSRDB; les ventes sont donc disponibles pour toutes les unités dans l'étude par simulations et sont imputées pour 15 % des unités d'échantillonnage. La corrélation entre le revenu et les ventes est d'environ 83 %, tant pour les données des répondants que pour les données entièrement imputées.

Dans nos études par simulation, l'échantillon s_1 n'est pas généré aléatoirement plusieurs fois à partir des données de l'EMSRDB. L'échantillon s_1 est fixe et comprend l'ensemble de toutes les $n_1 = 2\,375$

unités de l'échantillon initial de l'EMSRDB. L'identifiant de strate, le poids de sondage, la variable d'intérêt y (les ventes) et la variable auxiliaire x (le revenu) pour chaque unité de s_1 proviennent du fichier de l'échantillon de l'EMSRDB. Les unités ayant des valeurs y imputées sont incluses dans s_1 et les valeurs imputées sont traitées comme des valeurs observées. Cela nous permet de calculer l'estimation basée sur l'échantillon complet, $\hat{Y}_{\text{COMPLET}}$, donnée au (2.1). Cette estimation sert d'estimation de référence pour évaluer les propriétés de $\hat{Y}_{\text{HH}-\text{NA}}$ pour différents scénarios de réponse et plans de sondage du suivi, comme cela est détaillé ci-après.

Génération de l'ensemble $s_{1,\text{nr}}$ des non-répondants à l'envoi postal

Ensuite, à partir de s_1, la réponse à l'envoi par la poste est générée indépendamment d'une unité à une autre à partir d'une distribution de Bernoulli avec probabilité p_{1hi}, $i \in s_{1h}$, $h = 1, \ldots, L$. Deux scénarios de probabilité de réponse sont envisagés :

1. Uniforme : $p_{1hi} = 50\,\%$ pour toutes les unités d'échantillonnage. Selon ce scénario, le nombre espéré de répondants à l'envoi par la poste est de 2 375/2 = 1 187,5.
2. Corrélée à la variable d'intérêt : p_{1hi} déterminée au moyen de la fonction logit :

$$\log\left(\frac{p_{1hi}}{1-p_{1hi}}\right) = -0{,}31 + 0{,}000004\, y_{hi}.$$

Les constantes -0,31 et 0,000004 sont choisies par essais et erreurs, de telle sorte que le nombre espéré de non-répondants à l'envoi par la poste correspond de nouveau à environ la moitié de la taille de s_1. Il convient de mentionner que le nombre espéré de répondants à l'envoi par la poste peut s'exprimer sous la forme $\sum_{h=1}^{L}\sum_{i\in s_{1h}}\left(1-p_{1hi}\right)$. Par conséquent, les constantes sont telles que $\sum_{h=1}^{L}\sum_{i\in s_{1h}} p_{1hi} \approx 1187{,}5$, où $p_{1hi} = \left[1+\exp\left(0{,}31-0{,}000004\, y_{hi}\right)\right]^{-1}$.

Sélection de l'échantillon de suivi s_2

L'étape suivante de la simulation consiste à sélectionner un échantillon de suivi s_2 à partir de l'ensemble des non-répondants à l'envoi postal, $s_{1,\text{nr}}$, généré dans l'un des deux scénarios de probabilité de réponse ci-dessus. Cinq plans de sondage différents sont utilisés pour la sélection de l'échantillon de suivi :

1. le recensement des non-répondants à l'envoi postal;
2. un échantillonnage aléatoire simple (EAS) sans remise, ne tenant pas compte de la stratification initiale;
3. un EAS stratifié sans remise basé sur la stratification initiale, avec une répartition de l'échantillon dans les strates proportionnelle au nombre de non-répondants à l'envoi postal;
4. un échantillonnage systématique avec probabilité proportionnelle au revenu, x_{hi}, ne tenant pas compte de la stratification initiale;

5. un échantillonnage systématique avec probabilité proportionnelle au revenu multipliée par le poids de sondage initial, $w_{1hi}x_{hi}$, ne tenant pas compte de la stratification initiale.

Il convient de mentionner que les variables de taille utilisées pour les deux plans de sondage avec probabilité proportionnelle à la taille (PPT) sont élaguées en dessous du 5^{e} centile, afin de supprimer les observations dont la valeur est nulle et certaines valeurs extrêmement faibles, causant une instabilité. En moyenne, l'envoi par la poste compte 1 188 non-répondants. Pour le premier plan de sondage, tous les non-répondants font l'objet d'un suivi. Pour les quatre autres plans de sondage, nous choisissons 100, 200, 300, 400, 500, 700 et 900 comme taille d'échantillon de suivi pour la simulation.

Génération des résultats des appels

Les résultats de la procédure par collecte de suivi téléphonique sont simulés au niveau de la tentative d'appel. Pour chaque unité d'échantillonnage $i \in s_{1h}$, $h = 1, \ldots, L$, les probabilités $P_{2hi}^{(1)}$, $P_{2hi}^{(2)}$ et $P_{2hi}^{(3)}$ des trois résultats possibles (voir la section 2) sont attribuées avant de lancer la simulation et ne varient pas à mesure que progresse la collecte des données. Deux scénarios de réponse sont envisagés :

1. Uniforme : $P_{2hi}^{(1)} = 25\,\%$, $P_{2hi}^{(2)} = 5\,\%$, et $P_{2hi}^{(3)} = 70\,\%$ pour toutes les unités. Ces valeurs sont tirées de Xie, Godbout, Youn et Lavallée (2011).
2. Corrélée à la variable d'intérêt : la probabilité d'une « réponse » repose sur la fonction logit suivante :

$$\log\left(\frac{P_{2hi}^{(1)}}{1-P_{2hi}^{(1)}}\right) = -1{,}29 + 0{,}000002\, y_{hi} + 0{,}3\, z_{hi},$$

où z_{hi} est généré à partir de la distribution normale centrée réduite. Les constantes -1,29, 0,000002 et 0,3 sont choisies en procédant par essais et erreurs, de sorte que la moyenne de $P_{2hi}^{(1)}$ pour toutes les unités de l'échantillon s_1 est d'environ 25 %, c'est-à-dire $n_1^{-1}\sum_{h=1}^{L}\sum_{i\in s_{1h}} P_{2hi}^{(1)} \approx 0{,}25$, où $P_{2hi}^{(1)} = \left[1+\exp\left(1{,}29 - 0{,}000002\, y_{hi} - 0{,}3\, z_{hi}\right)\right]^{-1}$. Il convient de souligner que le coefficient de corrélation entre la probabilité de réponse $P_{2hi}^{(1)}$ et la variable d'intérêt y_{hi} est de 61 %. Les deux autres probabilités se définissent sous la forme : $P_{2hi}^{(2)} = \frac{0{,}05}{0{,}75}\left(1 - P_{2hi}^{(1)}\right)$ et $P_{2hi}^{(3)} = \frac{0{,}70}{0{,}75}\left(1 - P_{2hi}^{(1)}\right)$. Ainsi $P_{2hi}^{(1)} + P_{2hi}^{(2)} + P_{2hi}^{(3)} = 1$.

Pour une unité d'échantillonnage de suivi donnée, les probabilités $P_{2hi}^{(1)}$, $P_{2hi}^{(2)}$ et $P_{2hi}^{(3)}$ servent à générer aléatoirement l'issue de chaque appel. Après une tentative d'appel, l'unité est placée à la fin de la liste d'appels, sauf si l'on a terminé de la traiter et qu'un résultat « réponse » ou « non-réponse finale » est obtenu. Les résultats sont générés indépendamment d'un appel à un autre. Le nombre de tentatives d'appels à une même unité dans notre étude par simulations n'a pas de limite supérieure explicite $(K = \infty)$.

Notons que, pour le scénario de réponse comportant diverses probabilités de réponse, les unités répondant à la première tentative d'appel sont généralement des unités à probabilité de réponse plus

élevée. Par conséquent, les unités demeurant dans la liste d'appels pour la deuxième tentative tendent à être des unités ayant une probabilité de réponse moins élevée. Il s'ensuit que la proportion d'unités répondantes lors de la deuxième tentative d'appel tend à être plus petite que celle des unités répondantes lors de la première. De façon similaire, la proportion d'unités répondantes lors de la troisième tentative d'appel tend à être inférieure à la deuxième, etc. La proportion d'unités qui répondent décroît à chaque tentative d'appel, puisque les unités qui demeurent dans la liste d'appels sont celles qui sont plus difficiles à joindre. Les estimations peuvent par conséquent présenter un biais substantiel, si la collecte de données se termine de façon prématurée et si les unités plus difficiles à joindre tendent à présenter des valeurs de y supérieures ou inférieures à celles des autres unités d'échantillonnage.

Le budget total du suivi est fixé à 3 000 unités (unités monétaires ou temporelles) dans le cadre de notre étude. Un coût est facturé pour chaque tentative d'appel. Le montant facturé dépend du résultat de la tentative : un résultat de « réponse » a un coût de 5 unités $\left(c^{(1)}=5\right)$, un résultat de « non-réponse finale » représente un coût de 2 unités $\left(c^{(2)}=2\right)$ et un résultat « toujours en cours », un coût de 1 unité $\left(c^{(3)}=1\right)$. La collecte prend fin lorsque le budget est épuisé ou lorsqu'il ne reste plus de cas dans la liste d'appels (c'est-à-dire que toutes les unités sont résolues), selon ce qui se produit en premier. On a choisi les valeurs de coût et le budget quelque peu arbitrairement puisqu'ils dépendent de l'enquête. Cependant, nous avons veillé à ce que $c^{(1)}>c^{(2)}>c^{(3)}$, car nous nous attendons généralement à ce que cette relation soit vérifiée dans le cas d'enquêtes téléphoniques.

Mesures de Monte Carlo

La génération de réponses à l'envoi par la poste, la sélection de l'échantillon de suivi et la génération de réponses au suivi sont répétées de façon indépendante $R=1\,000$ fois pour chaque combinaison de scénario de réponse à l'envoi par la poste, de plan de sondage du suivi et de scénario de réponse de suivi décrits précédemment. L'estimateur ajusté pour la non-réponse (2.4), $\hat{Y}_{\mathrm{HH-NA}}$, est calculé pour chaque itération. Les ajustements de poids pour la non-réponse a_{2hi} sont calculés à l'aide de (2.5) comme l'inverse du taux de réponse pondéré global. Nous utilisons $a_{2hi}=a_2$, donné au (2.5), plutôt que $a_{2hi}=a_{2h}$, donné au (2.6), pour éviter quelques cas où certains des ensembles s_{2hr} sont vides, ce qui pourrait entraîner des valeurs infinies de a_{2h}. L'ajustement de poids pour la non-réponse (2.5) peut être considéré comme une forme extrême de regroupement. Un regroupement moins extrême pourrait être appliqué en pratique et pourrait présenter de meilleures propriétés. Nous choisissons (2.5) dans la présente étude par simulations pour sa simplicité.

À l'aide des 1 000 répliques de $\hat{Y}_{\mathrm{HH-NA}}$, le biais relatif Monte Carlo (BR) et la racine de l'erreur quadratique moyenne relative (REQMR) Monte Carlo de $\hat{Y}_{\mathrm{HH-NA}}$ sont calculés comme suit :

$$\mathrm{BR}=\frac{1}{R}\sum_{r=1}^{R}E_r\times 100\,\% \quad \text{et} \quad \mathrm{REQMR}=\sqrt{\frac{1}{R}\sum_{r=1}^{R}E_r^2}\times 100\,\%,$$

où $E_r=\left(\hat{Y}_{\mathrm{HH-NA}}^{r}-\hat{Y}_{\mathrm{COMPLET}}\right)\Big/\hat{Y}_{\mathrm{COMPLET}}$ est l'erreur relative pour la r^{e} réplique de simulation et $\hat{Y}_{\mathrm{HH-NA}}^{r}$ est l'estimateur de Hansen-Hurwitz ajusté pour la non-réponse pour la r^{e} réplique; $r=1,\ldots,1\,000$.

Comme cela a été souligné ci-dessus, l'échantillon initial s_1 est fixe pour chacune des 1 000 répliques, afin de nous permettre de nous concentrer sur les mécanismes de réponse à l'envoi par la poste et au suivi

ainsi qu'au plan de sondage du suivi. Alors qu'il aurait été possible de créer une population artificielle et de tirer un échantillon initial différent à chaque réplique, il a été déterminé que cette complexité supplémentaire ne changerait pas nos principales conclusions, sauf pour une augmentation systématique de la variance de $\hat{Y}_{\mathrm{HH-NA}}$. Notre cadre de simulation présente également l'avantage d'être conditionnel à des données-échantillons réelles.

4.2 Résultats de la simulation

Dans la présente section, nous traitons des résultats de la simulation pour quatre scénarios de réponse par la poste et de réponse de suivi :

1. la probabilité de réponse est uniforme pour l'envoi par la poste et le suivi; il s'agit du scénario de référence avec lequel comparer les autres scénarios;
2. la probabilité de réponse est corrélée aux ventes pour l'envoi par la poste et uniforme pour le suivi;
3. la probabilité de réponse est uniforme pour l'envoi par la poste et corrélée aux ventes pour le suivi;
4. la probabilité de réponse est corrélée aux ventes pour l'envoi par la poste et le suivi; ce scénario est probablement le plus réaliste.

Scénario de réponse 1 : Probabilité de réponse uniforme pour l'envoi par la poste et le suivi

La figure 4.1 présente le biais relatif par rapport à la taille de l'échantillon de suivi pour les cinq plans de sondage. La figure 4.2 présente la REQMR par rapport à la taille de l'échantillon de suivi. Il convient de mentionner que les résultats du suivi de tous les non-répondants à l'envoi postal sont indiqués par le dernier point des figures (c'est-à-dire une taille d'échantillon de 1 188).

Figure 4.1 Biais relatif par rapport à la taille de l'échantillon de suivi pour le scénario 1.

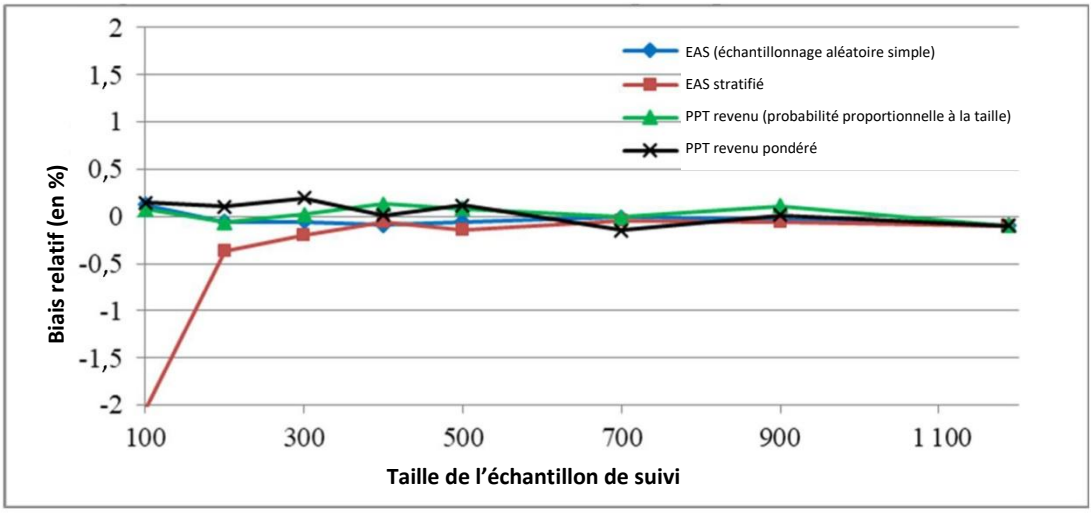
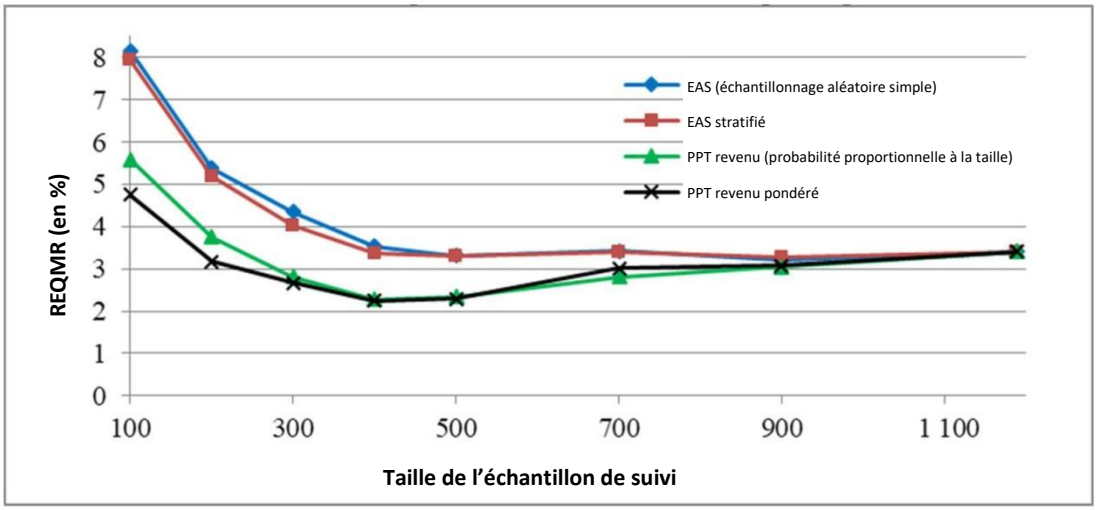


Figure 4.2 Racine de l'erreur quadratique moyenne relative par rapport à la taille de l'échantillon de suivi pour le scénario 1.



Nous faisons les observations suivantes après examen des figures 4.1 et 4.2 :

- Le BR est d'environ zéro pour toutes les tailles d'échantillon de suivi et tous les plans de sondage. La seule exception est l'EAS stratifié avec une taille d'échantillon de suivi de 100. La stratégie d'allocation proportionnelle pour l'échantillon de suivi n'assure pas la sélection d'au moins une unité de chaque strate. Par conséquent, pour des tailles d'échantillon de suivi plus petites (par exemple 100), certaines strates finissent sans échantillon de suivi, même si elles peuvent comprendre des non-répondants à l'envoi postal. Cela entraîne un biais négatif pour l'estimation du chiffre de population total.
- À mesure que la taille de l'échantillon augmente, passant de 100 à 400, la REQMR diminue pour tous les plans de sondage. Cela peut s'expliquer par une augmentation du nombre moyen de répondants à mesure que la taille de l'échantillon augmente (non présenté dans les figures).
- Pour les tailles d'échantillon supérieures à 400, la REQMR demeure relativement constante pour les plans de sondage avec EAS et d'EAS stratifié. Pour ces tailles d'échantillon, le nombre moyen de répondants demeure relativement constant. Ce constat concorde avec l'équation (3.8). Il indique que, dans le cas d'une réponse uniforme au suivi, le nombre espéré de répondants ne varie pas en fonction de K, et par conséquent en fonction de la taille d'échantillon de suivi, tant que le budget est dépensé.
- Les plans de sondage avec PPT semblent plus efficaces que les plans de sondage avec EAS et EAS stratifié. Toutefois, pour des tailles d'échantillon supérieures à 400, les gains d'efficacité diminuent à mesure que la taille de l'échantillon augmente.

Scénario de réponse 2 : Probabilité de réponse corrélée aux ventes pour l'envoi par la poste et uniforme pour le suivi

Les figures 4.3 et 4.4 présentent respectivement le biais relatif et la REQMR pour le scénario 2.

Figure 4.3 Biais relatif par rapport à la taille de l'échantillon de suivi pour le scénario 2.

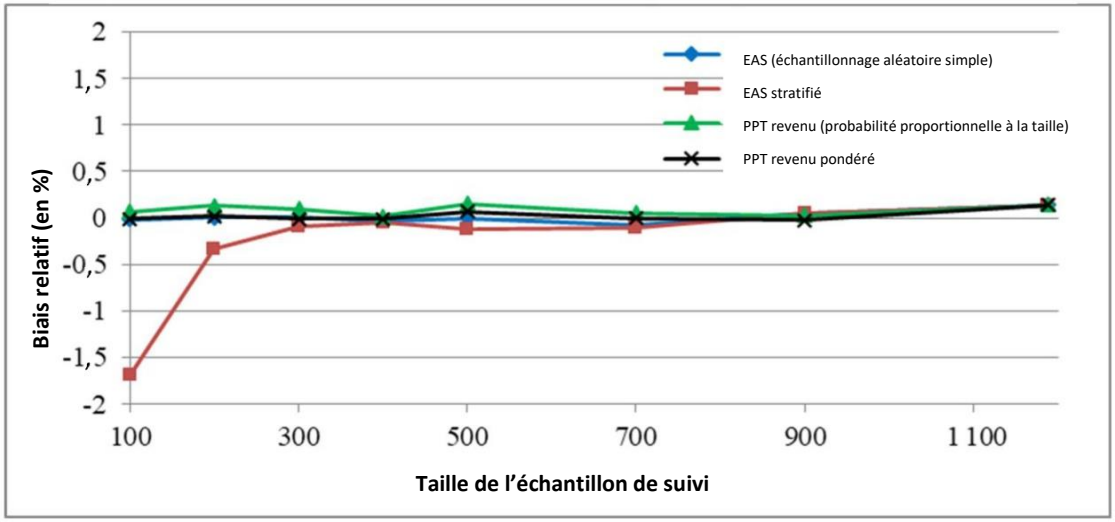
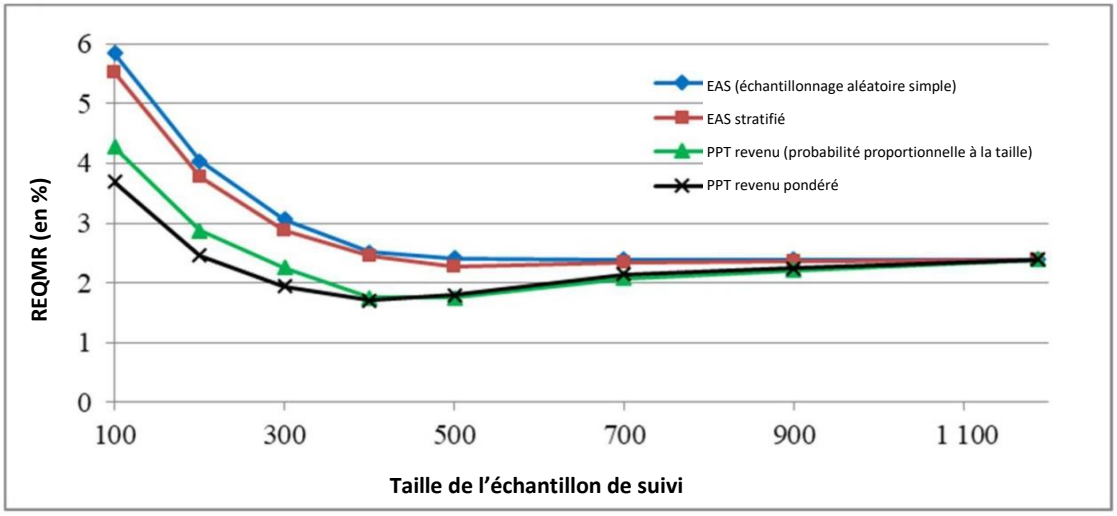


Figure 4.4 Racine de l'erreur quadratique moyenne relative par rapport à la taille de l'échantillon de suivi pour le scénario 2.



Nous faisons les observations suivantes après examen des figures 4.3 et 4.4 :

- Les résultats indiquent que si la probabilité de réponse à l'envoi par la poste est corrélée aux ventes, mais que la probabilité de réponse au suivi est uniforme, le biais peut être pratiquement

éliminé par le plan de sondage du suivi. Cela peut être expliqué en observant que l'estimateur de Hansen et Hurwitz (1946) (2.3) ne comporte pas de biais pour tout mécanisme de réponse à l'envoi par la poste.

- Les observations relatives au scénario 1 s'appliquent également au scénario 2.

Scénario de réponse 3 : Probabilité de réponse uniforme pour l'envoi par la poste et corrélée aux ventes pour le suivi

Les figures 4.5 et 4.6 présentent respectivement le biais relatif et la REQMR pour le scénario 3.

Figure 4.5 Biais relatif par rapport à la taille de l'échantillon de suivi pour le scénario 3.

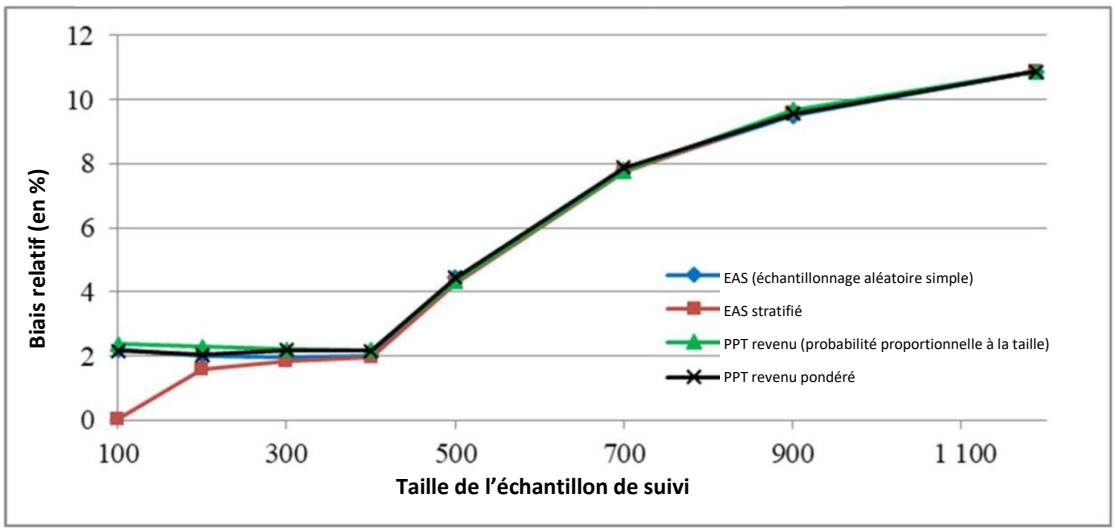
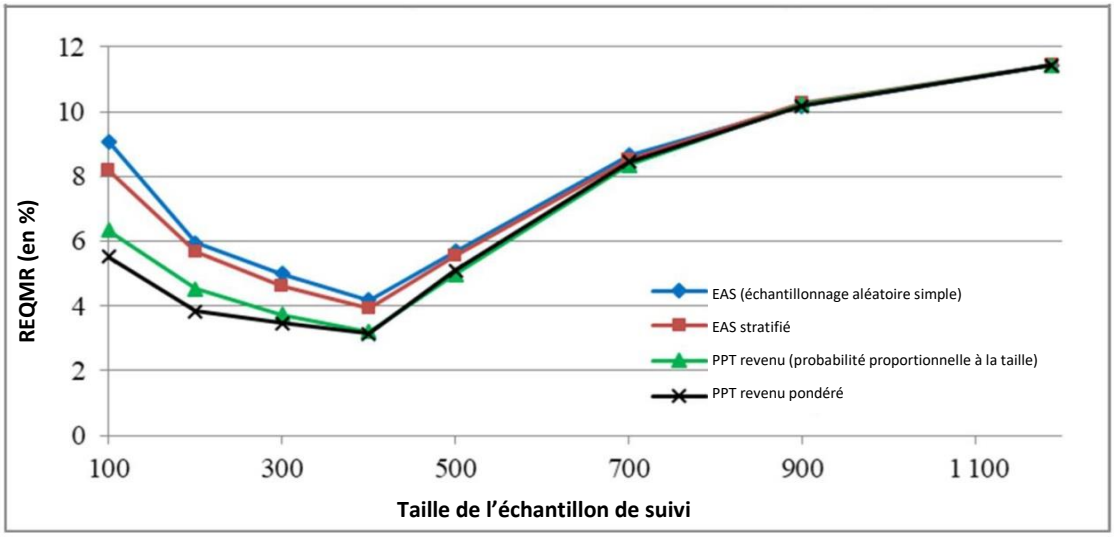


Figure 4.6 Racine de l'erreur quadratique moyenne relative par rapport à la taille de l'échantillon de suivi pour le scénario 3.



Nous faisons les observations suivantes après examen des figures 4.5 et 4.6 :

- Le BR est le plus bas pour les tailles d'échantillon inférieures ou égales à 400, pour lesquelles nous observons que le traitement de toutes les unités s'est terminé avant la dépense totale du budget. Le BR plus bas pour l'EAS stratifié avec une taille d'échantillon de suivi de 100 est attribuable à une strate ne comportant aucun échantillon de suivi (voir le scénario de réponse 1).
- La REQMR est minimisée pour une taille d'échantillon de 400.
- Pour des tailles d'échantillon supérieures à 400, le BR et la REQMR augmentent tous deux à mesure que la taille de l'échantillon augmente. Pour ces tailles d'échantillon, nous avons observé une diminution du taux de réponse moyen à mesure que la taille de l'échantillon augmentait (voir l'exposé sous l'équation (3.5) pour une justification théorique). Cela explique l'augmentation du BR et de la REQMR au fur et à mesure de l'augmentation de la taille de l'échantillon.
- Les plans de sondage avec PPT semblent de nouveau plus efficaces que les plans de sondage avec EAS et EAS stratifié. Toutefois, pour des tailles d'échantillon supérieures à 400, les gains d'efficacité diminuent à mesure que la taille de l'échantillon augmente.

Scénario de réponse 4 : Probabilité de réponse corrélée aux ventes pour l'envoi par la poste et le suivi

Les figures 4.7 et 4.8 présentent respectivement le biais relatif et la REQMR pour le scénario 4.

Les figures 4.7 et 4.8 sont similaires aux figures 4.5 et 4.6. Les observations relatives au scénario 3 s'appliquent également au scénario 4.

Figure 4.7 Biais relatif par rapport à la taille de l'échantillon de suivi pour le scénario 4.

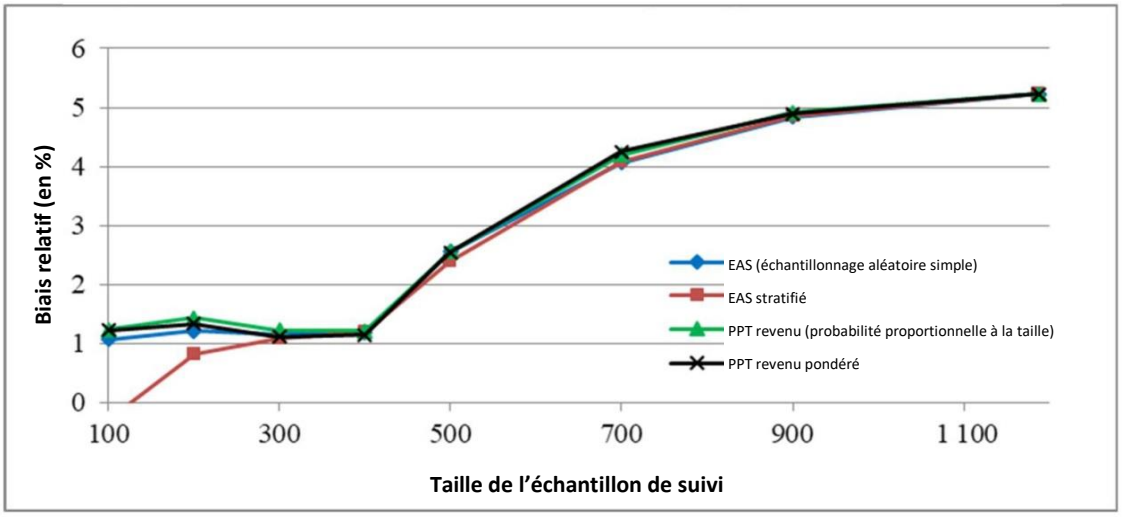
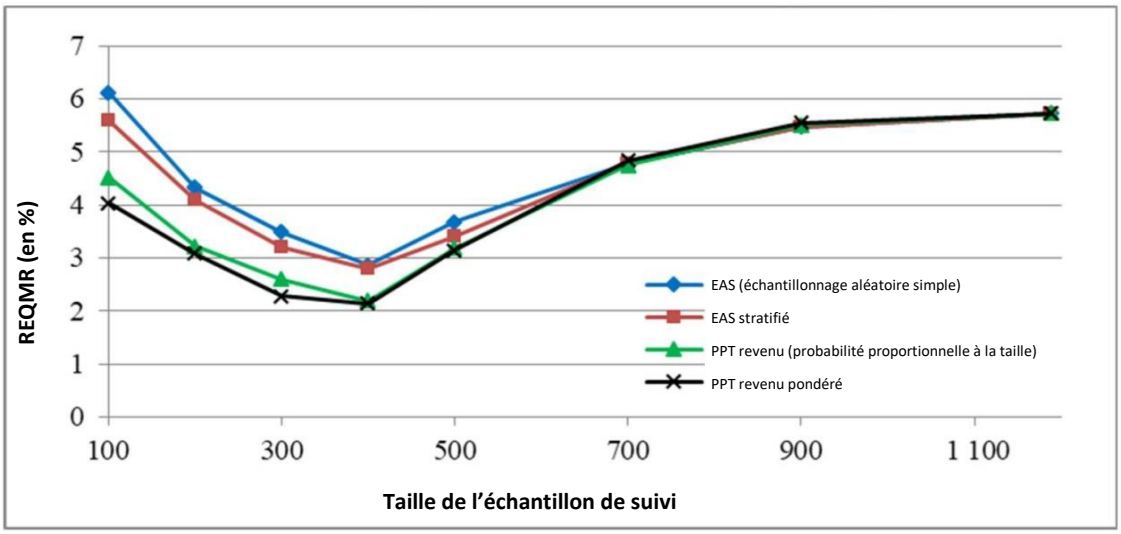


Figure 4.8 Racine de l'erreur quadratique moyenne relative par rapport à la taille de l'échantillon de suivi pour le scénario 4.



4.3 Remarques sur les résultats de la simulation

Nous avons observé que pour des tailles d'échantillon de suivi inférieures ou égales à 400 et pour tous les plans de sondage et les scénarios de réponse, toutes les unités étaient traitées définitivement et avaient une issue de « réponse » ou de « non-réponse finale » avant que le budget ne soit entièrement dépensé, à l'exception de deux répliques de simulation. Par conséquent, le taux de réponse de suivi est demeuré relativement constant, alors que le nombre de répondants a augmenté à mesure que la taille d'échantillon de suivi augmentait pour passer de 100 à 400, ce qui a réduit la variance et l'erreur quadratique moyenne de l'estimateur $\hat{Y}_{\mathrm{HH-NA}}$.

Pour des tailles d'échantillon de 500 ou plus, le budget de suivi était toujours entièrement dépensé avant le traitement définitif de toutes les unités. À mesure que la taille d'échantillon de suivi augmentait, les nombres de répondants et d'unités traités définitivement demeuraient relativement constants. En moyenne, entre 430 et 445 cas étaient clos à la fin de la collecte de données selon le plan de sondage et le scénario de réponse; les autres unités demeurant dans la liste d'appels avec une issue « toujours en cours ». Il semble donc que le budget de suivi utilisé pour l'étude par simulations était suffisamment important pour traiter définitivement environ 440 unités pour des tailles d'échantillon supérieures ou égales à 500. Étant donné que le nombre de répondants est demeuré relativement constant au fur et à mesure de l'augmentation de la taille, le taux de réponse a diminué. La diminution du taux de réponse peut s'expliquer par un plus petit nombre moyen de tentatives d'appel par unité d'échantillonnage au fur et à mesure de l'augmentation de la taille de l'échantillon de suivi. Cela entraîne la conséquence peu souhaitable d'accroître le biais et l'erreur quadratique moyenne de $\hat{Y}_{\mathrm{HH-NA}}$ pour le mécanisme de réponse de suivi non uniforme.

Dans les figures 4.2, 4.4, 4.6 et 4.8, nous observons également que la REQMR atteint un minimum pour une taille d'échantillon de 400 ou 500 quelque soit le scénario de réponse et le plan de sondage. La taille d'échantillon réduisant la REQMR au minimum semble approximativement correspondre à la taille d'échantillon minimale permettant de dépenser le budget de suivi en moyenne. Comme nous l'avons exposé plus haut, une taille d'échantillon réduite augmente la variance de $\hat{Y}_{\mathrm{HH-NA}}$, du fait d'un nombre de répondants moins élevé, alors qu'une taille d'échantillon plus élevée peut accroître le biais dû à un taux de réponse réduit. La taille d'échantillon minimale pour dépenser le budget de suivi semble être la même que le nombre espéré d'unités résolues, qui était d'environ 440 dans notre étude par simulations pour des tailles d'échantillon de 500 ou plus.

La théorie élaborée à la section 3 concorde avec les observations empiriques ci-dessus pour une réponse au suivi uniforme. Le tableau 4.1 présente les valeurs de la taille de l'échantillon (3.7), le nombre espéré de répondants (3.8), le taux de réponse espéré (3.9) et le nombre espéré d'unités résolues (3.10) pour différentes valeurs de K et pour les valeurs de C, $c^{(1)}$, $c^{(2)}$, $c^{(3)}$, $P_2^{(1)}$, $P_2^{(2)}$ et $P_2^{(3)}$ utilisées dans l'étude par simulations : $C = 3\,000$, $c^{(1)} = 5$, $c^{(2)} = 2$, $c^{(3)} = 1$, $P_2^{(1)} = 0{,}25$, $P_2^{(2)} = 0{,}05$ et $P_2^{(3)} = 0{,}70$. La taille d'échantillon minimale $n_2(C, \infty)$ et le nombre espéré d'unités résolues $\tilde{n}_{2,\mathrm{res}}(C, K)$ égalent 439; cela concorde avec les résultats de la simulation.

Comme le montre le tableau 4.1, une faible valeur de K peut réduire le taux de réponse espéré de façon importante, alors que le nombre espéré de répondants ne varie pas en fonction de K, tant que le budget est dépensé. Par conséquent, dans un contexte de réponse au suivi uniforme, il ne semble exister aucun avantage à utiliser une taille d'échantillon de suivi supérieure à $n_2(C, \infty)$, la taille d'échantillon minimale permettant de dépenser le budget en moyenne, qui est de 439 dans ce scénario. Ce choix maximise le taux de réponse espéré sans réduire le nombre espéré de répondants. Dans des cas s'écartant modérément de la réponse uniforme, choisir une taille d'échantillon proche de $n_2(C, \infty)$ (ou une valeur élevée de K) permettrait de s'assurer de mieux contrôler le biais de non-réponse.

Nos résultats de simulation indiquent que les conclusions tirées du tableau 4.1 se vérifient approximativement pour une réponse au suivi non uniforme. En particulier, la taille d'échantillon minimale permettant de dépenser le budget était proche de 439 et les nombres espérés de répondants et d'unités résolues demeuraient relativement constants lorsque la taille de l'échantillon de suivi augmentait. Par conséquent, supposer de façon incorrecte une réponse uniforme lorsqu'elle n'est pas uniforme se traduit par une taille d'échantillon qui demeure appropriée dans notre cadre de simulation. Une autre conclusion de notre étude par simulations est que choisir une taille d'échantillon de suivi proche de $n_2(C, \infty)$ semble réduire à la fois le biais de non-réponse et l'erreur quadratique moyenne de $\hat{Y}_{\mathrm{HH-NA}}$. Cependant, nous montrerons dans les deux exemples suivants que nos conclusions peuvent ne pas toujours se vérifier en cas de déviations plus importantes d'une réponse uniforme.

Supposons que le nombre de non-répondants à l'envoi postal soit exactement de 1 188 et que les valeurs de C, $c^{(1)}$, $c^{(2)}$, $c^{(3)}$, $P_2^{(1)}$, $P_2^{(2)}$ et $P_2^{(3)}$ soient exactement les mêmes que celles de l'étude par simulations et du tableau 4.1. Toutefois, pour l'une des 1 188 unités (disons l'unité j), les probabilités

$P_{2j}^{(1)} = 0{,}25$, $P_{2j}^{(2)} = 0{,}05$ et $P_{2j}^{(3)} = 0{,}70$ sont respectivement remplacées par $P_{2j}^{(1)} = 0{,}000005$, $P_{2j}^{(2)} = 0{,}000001$ et $P_{2j}^{(3)} = 0{,}999994$. Le mécanisme de réponse est presque uniforme, à l'exception d'une unité exhibant une très faible probabilité d'être résolue. Par souci de simplicité, nous supposons que l'échantillon de suivi est sélectionné par échantillonnage aléatoire simple sans remise. Dans ce scénario, le tableau 4.2 indique la taille de l'échantillon (3.3), le nombre espéré de répondants (3.4), le taux de réponse espéré (3.5) et le nombre espéré d'unités résolues (3.6) pour diverses valeurs de K.

Tableau 4.1
Taille de l'échantillon, taux de réponse espéré et nombres espérés de répondants et d'unités résolues pour diverses valeurs de K dans le cas d'une réponse au suivi uniforme

K	Taille de l'échantillon espéré (3.7)	Taux de réponse espéré (3.9)	Nombre espéré de répondants (3.8)	Nombre espéré d'unités résolues (3.10)
∞	439	83,3 %	366	439
20	439	83,3 %	366	439
10	452	81,0 %	366	439
6	498	73,5 %	366	439
5	528	69,3 %	366	439
4	578	63,3 %	366	439
3	668	54,8 %	366	439
2	861	42,5 %	366	439
1*	1 188	25,0 %	297	356

* L'application directe de (3.7) mène à $n_2(C,1) =$ 1 463. Toutefois, cette valeur est supérieure au nombre espéré de non-répondants à l'envoi postal dans l'étude par simulations. En supposant un nombre précis de 1 188 non-répondants à l'envoi postal et en les englobant tous dans l'échantillon de suivi $(n_2 = 1{,}188)$, nous pouvons calculer le coût espéré du suivi (3.1) comme étant $\tilde{C}(n_2, 1) =$ 2 435,4, ce qui est inférieur au budget total de 3 000. En utilisant un budget révisé de 2 435,4, les nombres espérés de répondants et d'unités résolues sont de 297 et de 356, respectivement.

Tableau 4.2
Taille de l'échantillon, taux de réponse espéré et nombres espérés de répondants et d'unités résolues pour diverses valeurs de K lorsqu'une unité présente une probabilité très faible d'être résolue

K	Taille de l'échantillon espéré (3.3)	Taux de réponse espéré (3.5)	Nombre espéré de répondants (3.4)	Nombre espéré d'unités résolues (3.6)
∞	20	83,3 %	17	20
20	439	83,2 %	365	438
10	452	80,9 %	365	438
6	498	73,5 %	366	439
5	528	69,3 %	366	439
4	578	63,3 %	366	439
3	668	54,7 %	366	439
2	861	42,5 %	366	439
1*	1 188	25,0 %	297	356

* L'application directe de (3.3) mène à $n_2(C,1) =$ 1 464, ce qui est supérieur au nombre de non-répondants à l'envoi postal (1 188). De façon similaire au tableau 4.1, nous pouvons calculer le coût espéré du suivi (3.1) comme étant $\tilde{C}(n_2 = 1\,188, K = 1) =$ 2 434,4, ce qui est inférieur au budget total de 3 000. En utilisant un budget révisé de 2 434,4, les nombres espérés de répondants et d'unités résolues sont de 297 et de 356, respectivement.

La taille d'échantillon minimale permettant de dépenser le budget, en moyenne, est $n_2(C,\infty) = 20$ dans ce scénario. Elle est nettement inférieure à 439 (la valeur correspondante pour une réponse uniforme

indiquée au tableau 4.1). Comme cela a été exposé à la section 3, avoir recours à une valeur finie de K peut permettre d'éviter de dépenser une trop grande portion du budget pour quelques unités présentant une très faible probabilité de résolution (unité j dans le présent exemple). En effet, le tableau 4.2 montre que le taux de réponse espéré diminue légèrement en réduisant la valeur de K de l'infini à 20, alors que le nombre espéré de répondants augmente de façon importante de 17 à 365. Utiliser une valeur finie de K semble souhaitable dans ce scénario, car elle peut réduire de façon substantielle la variance de $\hat{Y}_{\mathrm{HH-NA}}$. L'incidence sur le biais de non-réponse est probablement négligeable, sauf si la valeur y de l'unité j est extrêmement différente des autres unités. Supposer de façon incorrecte une réponse uniforme pour toutes les unités entraînerait le choix d'une taille d'échantillon de 439, comme le montre le tableau 4.1. Ce choix semble demeurer approprié pour ce mécanisme de réponse au suivi non uniforme.

Supposons de nouveau que le nombre de non-répondants à l'envoi postal est de 1 188, les valeurs de C, $c^{(1)}$, $c^{(2)}$ et $c^{(3)}$ sont les mêmes que celles utilisées dans l'étude par simulations et le tableau 4.1, et l'échantillon de suivi est sélectionné par échantillonnage aléatoire simple sans remise. Supposons maintenant que les 1 188 non-répondants à l'envoi postal peuvent être répartis en deux groupes de réponse homogènes, chacun ayant une taille de 594. Les probabilités sont $P_{2hi}^{(1)} = 0{,}45$, $P_{2hi}^{(2)} = 0{,}09$ et $P_{2hi}^{(3)} = 0{,}46$ pour les 594 unités du premier groupe et $P_{2hi}^{(1)} = 0{,}05$, $P_{2hi}^{(2)} = 0{,}01$ et $P_{2hi}^{(3)} = 0{,}94$ pour les 594 unités restantes. Le mécanisme de réponse n'est pas uniforme; il est uniforme dans chacun des deux groupes de réponse homogènes. Les probabilités moyennes pour les 1 188 non-répondants à l'envoi postal sont les mêmes que celles fournies par le scénario de réponse uniforme. Le tableau 4.3 présente la taille de l'échantillon (3.3), le nombre espéré de répondants (3.4), le taux de réponse espéré (3.5) et le nombre espéré d'unités résolues (3.6) pour diverses valeurs de K.

Tableau 4.3
Taille de l'échantillon, taux de réponse espéré et nombres espérés de répondants et d'unités résolues pour diverses valeurs de K dans le cas d'une réponse uniforme au sein de groupes

K	Taille de l'échantillon espéré (3.3)	Taux de réponse espéré (3.5)	Nombre espéré de répondants (3.4)	Nombre espéré d'unités résolues (3.6)
∞	235	83,3 %	196	235
20	305	71,2 %	217	261
10	409	60,9 %	249	299
6	519	54,2 %	281	338
5	566	51,9 %	294	352
4	629	48,9 %	308	370
3	727	44,7 %	325	390
2	914	37,7 %	344	413
1*	1 188	25,0 %	297	356

* L'application directe de (3.3) mène à $n_2(C,1) =$ 1 463, ce qui est supérieur au nombre de non-répondants à l'envoi postal (1 188). De façon similaire au tableau 4.1, nous pouvons calculer le coût espéré du suivi (3.1) comme étant $\tilde{C}(n_2 =$ 1 188, $K = 1) =$ 2 435,4, ce qui est inférieur au budget total de 3 000. En utilisant un budget révisé de 2 435,4, les nombres espérés de répondants et d'unités résolues sont de 297 et de 356, respectivement.

La taille d'échantillon minimale permettant de dépenser le budget, en moyenne, est $n_2(C,\infty)=235,$ ce qui est bien inférieur à la valeur correspondante de 439 pour une réponse uniforme. Dans ce scénario, utiliser une valeur finie de K ne semble pas avantageux. En diminuant la valeur de K de l'infini à 20, le nombre espéré de répondants augmente uniquement de 21, alors que le taux de réponse espéré diminue de plus de 10 %. Cette faible réduction de la variance pourrait être contrebalancée par une plus forte hausse du biais de non-réponse. L'ampleur du biais de non-réponse dépend de la force de l'association entre la variable y et les groupes de réponse homogènes. Une faible valeur de K (taille d'échantillon importante) peut être appropriée si cette association est faible afin de profiter d'un nombre de répondants plus élevé que prévu. Néanmoins, c'est un choix risqué, puisque le taux de réponse espéré diminuerait de façon importante, offrant ainsi une moindre protection contre un écart par rapport au mécanisme de réponse postulé. Par conséquent, une taille d'échantillon de 439 dans ce scénario peut ne pas être appropriée du fait du risque accru de biais de non-réponse. Ce dernier peut être atténué à l'étape de l'estimation, au moins asymptotiquement, en calculant séparément l'ajustement de poids pour la non-réponse (2.5) pour chaque groupe de réponse homogène. Cette stratégie de pondération est standard et devrait être utilisée lorsqu'il est possible de déterminer des groupes de réponse homogènes; elle n'offre cependant pas de protection complète en cas d'écart par rapport au mécanisme de réponse postulé. C'est pour cette raison qu'une valeur K élevée, voire infinie, peut être préférable dans ce scénario.

Comme cela a été souligné à la section 3, représenter graphiquement le taux de réponse espéré et le nombre espéré de répondants comme une fonction de K peut être utile pour déterminer un compromis acceptable entre la maximisation du taux de réponse espéré $(K=\infty)$ et la maximisation du nombre de répondants espéré, comme cela est illustré dans les exemples ci-dessus. Une valeur infinie de K devrait être la valeur par défaut, car elle réduit au minimum le biais de non-réponse. Toutefois, une valeur finie élevée de K peut être appropriée, si elle accroît de manière marquée le nombre espéré de répondants sans incidence importante sur le taux de réponse espéré.

5. Conclusions

Dans la section 3, nous avons dérivé une expression explicite de $n_2(C,\infty),$ la taille d'échantillon minimale permettant de dépenser le budget $C,$ en moyenne, tout en résolvant toutes les unités d'échantillonnage de suivi. Nous avons montré que cette taille d'échantillon minimale maximisait le taux de réponse espéré, réduisant ainsi au minimum le biais de l'estimateur de Hansen et Hurwitz (1946) ajusté pour la non-réponse. Nos expériences empiriques ont montré que cette taille d'échantillon minimale semblait également réduire au minimum l'erreur quadratique moyenne de cet estimateur. Cela peut s'expliquer par le fait que le nombre espéré de répondants demeure relativement constant à mesure que la taille de l'échantillon augmente, ce qui donne une variance à peu près constante. Pour le mécanisme de réponse au suivi uniforme, il a été possible de montrer théoriquement que le nombre espéré de répondants ne variait pas à mesure que la taille d'échantillon augmentait (ou ne variait pas en fonction de K), ce qui confirme les résultats empiriques.

À première vue, l'idée de maximiser le taux de réponse espéré pour réduire au minimum le biais de non-réponse peut sembler contredire la littérature existante sur la non-réponse. Il est bien connu qu'une procédure de collecte de données tentant de maximiser le taux de réponse pour un échantillon donné augmentera probablement le biais de non-réponse lorsque les répondants faciles à joindre diffèrent des autres unités d'échantillonnage. En d'autres termes, accroître le taux de réponse ne réduit pas nécessairement le biais de non-réponse pour un échantillon donné et peut en fait avoir l'effet contraire. Nos résultats ne contredisent pas cela, puisque nous avons étudié une autre caractéristique de la conception d'une collecte de données : l'effet de la taille de l'échantillon de suivi sur le taux de réponse espéré et le biais de non-réponse. Il semble que cette question n'ait pas été étudiée dans la littérature. Notre principale conclusion est qu'une plus petite taille d'échantillon de suivi contribue à accroître le taux de réponse espéré et à réduire le biais de non-réponse.

Nos conclusions peuvent avoir des répercussions importantes en pratique. Dans les enquêtes-entreprises que mène Statistique Canada, tous les non-répondants à l'envoi postal font actuellement l'objet d'un suivi, et une procédure de collecte adaptative sert à prioriser les cas (voir Bosa et coll., 2018). Nous pensons que le biais de non-réponse pourrait être encore réduit en effectuant un suivi uniquement d'un échantillon de non-répondants à l'envoi postal dans des situations où le budget de suivi est insuffisant pour traiter correctement le volume de non-répondants à l'envoi postal. La procédure de collecte adaptative actuellement en place pourrait continuer à être utilisée pour gérer la collecte de données de l'échantillon de suivi.

Une autre conclusion de nos études empiriques est que les plans de sondage avec PPT semblent être légèrement plus efficaces que les plans de sondage avec EAS et EAS stratifié. Toutefois, nous n'avons pas tenté d'optimiser la stratification ou la répartition du plan de sondage avec EAS stratifié. L'efficacité du plan de sondage stratifié serait probablement amélioré par une utilisation plus efficace de la variable auxiliaire « Revenu » pour la stratification.

Enfin, nous avons observé que, contrairement au mécanisme de réponse au suivi, le mécanisme de réponse à l'envoi par la poste n'avait pas d'incidence sur le biais de l'estimateur de Hansen et Hurwitz (1946) ajusté pour la non-réponse. Par conséquent, le biais de non-réponse à l'envoi par la poste pouvait être éliminé, même si la probabilité de réponse à l'envoi par la poste est corrélée à la variable d'intérêt, tant que la probabilité de réponse au suivi est uniforme. Ce résultat n'est pas étonnant, puisque l'estimateur de Hansen et Hurwitz (1946) est sans biais pour tout mécanisme de réponse à l'envoi postal.

Remerciements

Les auteurs aimeraient remercier trois examinateurs anonymes et le rédacteur associé pour leurs commentaires constructifs, qui ont permis d'améliorer grandement la clarté de ce texte.

Bibliographie

Beaumont, J.-F., Bocci, C. et Haziza, D. (2014). An adaptive data collection procedure for call prioritization. *Journal of Official Statistics*, 30, 607-621.

Beaumont, J.-F., Bocci, C. et Hidiroglou, M. (2014). On weighting late respondents when a follow-up subsample of nonrespondents is taken. Document présenté au Comité consultatif des méthodes statistiques, Statistique Canada, mai 2014, Ottawa.

Bosa, K., Godbout, S., Mills, F. et Picard, F. (2018). Comment décomposer la variance due à la non-réponse : une méthode fondée sur l'erreur d'enquête totale. *Techniques d'enquête*, 44, 2, 319-337. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2018002/article/54957-fra.pdf.

Groves, R.M., et Heeringa, S.G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society, Series A*, 169, 439-457.

Hansen, M.H., et Hurwitz, W.N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association*, 41, 517-529.

Peytchev, A., Riley, S., Rosen, J., Murphy, J. et Lindblad, M. (2010). Reduction of non-response bias in surveys through case prioritization. *Survey Research Methods*, 4, 21-29.

Schouten, B., Cobben, F. et Bethlehem, J. (2009). Indicateurs de la représentativité de la réponse aux enquêtes. *Techniques d'enquête*, 35, 1, 107-121. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2009001/article/10887-fra.pdf.

Statistique Canada (2017). *Enquête mensuelle sur les services de restauration et débits de boissons (EMSRDB)*. Statistique Canada, https://www23.statcan.gc.ca/imdb/p2SV_f.pl?Function=getSurvey&Id=413027.

Thompson, K.J., Kaputa, S. et Bechtel, L. (2018). Stratégies de sous-échantillonnage des non-répondants pour les programmes économiques. *Techniques d'enquête*, 44, 1, 81-107. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2018001/article/54929-fra.pdf.

Tourangeau, R., Brick, J.M., Lohr, S. et Li, J. (2017). Adaptive and responsive survey designs: A review and assessment. *Journal of the Royal Statistical Society, Series A*, 180, 203-223.

Xie, H., Godbout, S., Youn, S. et Lavallée, P. (2011). Collection Follow-Up Operation Using Priority Scores For Business Surveys. Conference of European Statisticians, Work Session on Statistical Data Editing, Ljubljana (Slovénie).