QUELQUES TRAVAUX DE RECHERCHE SUR LA COLLECTE POUR LES ENQUETES AUPRES DES ENTREPRISES A STATISTIQUE CANADA

Wesley Yung[1], Yanick Beaucage[2] et Claude Turmelle[3]

[1] Statistique Canada, 100, promenade Tunney's Pasture, Ottawa, K1A 0T6, Wesley.Yung@statcan.gc.ca
[2] Statistique Canada, 100, promenade Tunney's Pasture, Ottawa, K1A 0T6, Yanick.Beaucage@statcan.gc.ca
[3] Statistique Canada, 100, promenade Tunney's Pasture, Ottawa, K1A 0T6, Claude.Turmelle@statcan.gc.ca

Résumé. La recherche sur la collecte est un sujet très populaire, surtout dans les agences qui mènent des enquêtes cherchant des façons d'augmenter les taux de réponse et/ou de diminuer les coûts associée à la collecte. Statistique Canada reconnaît les avantages de faire ces recherches et est actif dans le domaine depuis quelques années. Comme pour la plupart des recherches qui sont faites, Statistique Canada a mis l'accent sur les enquêtes auprès des ménages et il n'y a pas beaucoup de travail qui a été fait dans le domaine des enquêtes auprès des entreprises. Cette présentation discutera de quelques travaux qui ont été faits récemment sur la collecte pour les enquêtes auprès des entreprises. La présentation couvrira deux projets visant la gestion de la collecte pour les enquêtes auprès des entreprises ayant pour but d'assurer l'utilisation efficace des ressources de collecte et la qualité des statistiques produites. Un des projets de recherche touche le remaniement de nos enquêtes annuelles auprès des entreprises alors que l'autre est un projet plus académique.

1.0 Introduction

La recherche sur la collecte est un sujet très populaire récemment dans les agences statistiques nationales cherchant des façons d'être plus efficace. Avec des coûts de collecte très dispendieux, une épargne d'un petit ordre de grandeur peut se traduire par des économies financières non-négligeables. Statistique Canada reconnaît ce potentiel et s'implique dans ce domaine depuis quelques années. Cependant, la majorité du travail fait dans ce domaine est centré sur les enquêtes auprès des ménages (Laflamme, 2008). Très peu de recherche est faite pour les enquêtes auprès des entreprises même si celles-ci ont des défis différents des enquêtes auprès des ménages. Par exemple, en général, le meilleur temps pour contacter un répondant n'est pas un problème puisque les entreprises ont des heures de travail assez standard. Par contre, combien unités devraient être suivies, lesquelles et dans quelle ordre sont toujours des questions qui demeurent sans réponse. Avec des populations très asymétriques, est-ce qu'il est toujours mieux de mettre beaucoup de ressources afin d'avoir les données des unités économiquement les plus significatives, comme il est fait aujourd'hui ? Évidemment, les données des grandes entreprises sont importantes pour l'estimation du niveau de l'économie, mais est-ce que la stratégie de faire du suivi seulement sur ces unités ne risque pas d'augmenter le potentiel de biais de non-réponse, et donc, de réduire la qualité des estimations finales ?

Le but de cet article est de présenter quelques travaux de recherche sur la collecte pour les enquêtes auprès des entreprises à Statistique Canada. On y discutera deux projets: 1) La collecte pour les enquêtes annuelles auprès des entreprises et 2) Une étude empirique sur le suivi des non-répondants. Dans la section 2, on discutera du travail sur la collecte pour les enquêtes annuelles alors que l'étude su la non-réponse sera présentée dans la section 3. Finalement, quelques conclusions seront présentées dans la section 4.

2.0 La collecte sur les enquêtes annuelles auprès des entreprises

2.1 Le contexte

À Statistique Canada, l'Enquête unifiée auprès des entreprises (EUE) inclut environ 60 enquêtes annuelles. L'EUE a été mise en place à la fin des années 90 afin de répondre aux besoins d'obtenir des estimations plus détaillées au niveau provincial. Un des buts de l'EUE était d'harmoniser les concepts et les méthodes et de rendre le traitement des données plus efficace. Pour plus d'information sur l'EUE, voir Statistique Canada (1999). Au début, l'EUE a réussi à atteindre ces objectifs, mais quand de nouvelles enquêtes se sont intégrées, différents processus ont été introduits dans le système. En introduisant ces nouveaux processus, le système de l'EUE est devenu plus fragile ; il était difficile d'introduire des changements sans risquer une panne du système. En plus, le logiciel utilisé pour bâtir le système n'est plus supporté par l'organisation. Donc il y a quatre ans, il a été décidé de remanier l'EUE afin de rendre les systèmes et les méthodes plus flexibles et efficaces. Le nouveau programme annuel a été nommé le Programme intégré de la statistique des entreprises (PISE). À la fin des cycles d'intégration envisagés, le PISE couvrira plus de 150 enquêtes auprès des entreprises. Ces enquêtes seront un mélange d'enquêtes annuelles, infra-annuelles et d'enquêtes ponctuelles. Ces dernières collecteront de l'information financière, de l'information sur les caractéristiques des entreprises et sur l'achat et la vente de marchandises. Pour plus d'information sur le PISE, voir Statistique Canada (2010).

Parmi les nombreux objectifs du PISE, cet article s'attardera sur trois en particulier :

- Augmenter l'efficacité opérationnelle en comparaison avec aujourd'hui.
- Réduire la période de collecte des données et gérer ses activités de façon plus efficace.
- Réduire le nombre de vérifications des micro-données.

Pour atteindre ces trois objectifs, le modèle de traitement des données a été évalué et quelques améliorations ont été identifiées. L'évaluation a remis en question le lien entre la collecte et le traitement des données ; un nouveau modèle, les Estimations en continu (EC), a été développé et sera mis en place dans le PISE.

2.2 Les estimations en continu

Présentement le processus d'enquête de l'EUE est très linéaire (voir figure 1.).

Figure 1. Processus d'enquête de l'EUE

Le processus commence avec un échantillon qui est tiré du Registre des entreprises de Statistique Canada. Puis, la collecte de données commence et peut durer jusqu'à 8 mois. Après la collecte, le traitement de données est fait, suivi par l'analyse des données par les analystes. Finalement, les résultats des enquêtes sont diffusés environ 18 mois après le début du processus. Dans ce modèle, les analystes doivent attendre jusqu'à la fin de l'étape de traitement avant de voir les estimations et leurs indicateurs de qualité. Ça peut prendre de 10 à 13 mois après que l'échantillon ait été tiré. Une fois que les analystes ont accès aux estimations et aux indicateurs de qualité, un autre 5 à 7 mois sont nécessaires pour faire l'analyse des données. Pendant cette période, les analystes regardent les estimations et modifient les micro-données si nécessaire. Clairement ce travail est très dispendieux

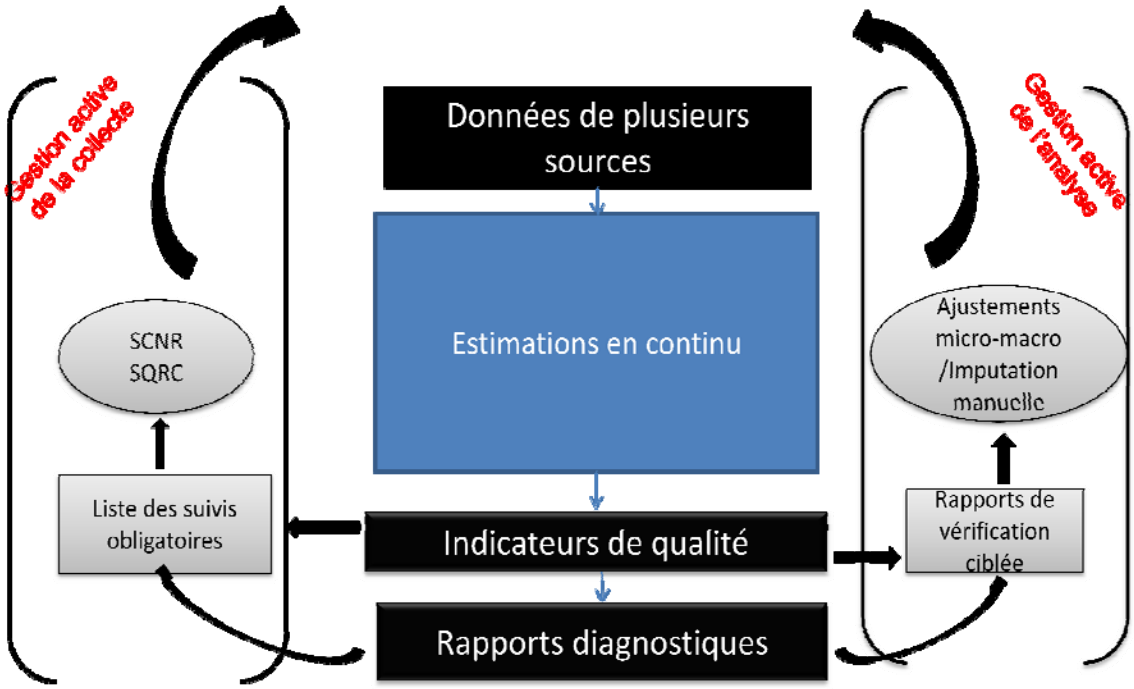
et prend beaucoup de temps.

Dans l'EUE, la collecte est gérée avec une fonction de score et le taux de réponse pondéré économiquement. La fonction de score est basée sur une mesure de taille (normalement le revenu provenant du registre) et assure que les unités importantes (selon la fonction de score) sont suivies par le personnel de la collecte. La décision de terminer la collecte est basée sur le taux de réponse pondéré. Quand ce taux atteint un taux ciblé, la collecte peut se terminer. Si la cible n'est pas atteinte, la collecte continuera et les unités à suivre seront gérées par la fonction de score. En évaluant le modèle actuel, quelques faiblesses ont été identifiées :

- La fonction de score dépend du revenu seulement
- La fermeture de la collecte est basée sur un seul indicateur de qualité (taux de réponse)
- Les analystes ne peuvent voir les estimations et les indicateurs de qualité qu'à la fin des étapes de traitement.

En réponse à ces faiblesses, un nouveau modèle de traitement de données a été développé. Le nouveau modèle, qu'on appelle Estimations en continu (EC), combine la collecte, le traitement et l'analyse des données. L'information de multiples sources sera utilisée pour gérer la collecte et l'analyse dans l'EC.

Figure 2. – Modèle d'estimations en continu



Le modèle EC comprend deux volets (voir figure 2.):

- Les activités de la collecte, c.-à-d.le Suivi des Questionnaires Rejetés au Contrôle (SQRC) et le Suivi des Cas de Non-Réponse (SCNR) seront gérés d'une manière plus dynamique et efficace. Le but principal de cette *Gestion active de la collecte* est de faire seulement des suivis sur les unités qui auront un impact significatif sur la qualité des résultats.
- Le travail des analystes (ou la *Gestion active de l'analyse)* sera aussi géré d'une façon efficace en mettant l'accent sur les unités qui auront un impact significatif sur la qualité des

résultats.

Ces deux activités de gestion seront faites en utilisant des Indicateurs de Qualité (IQ) et des scores de Mesure d'Impact (MI).

2.2.1 Les indicateurs de qualité

Les indicateurs de qualité jouent un rôle clé dans l'EC parce qu'ils seront utilisés pour contrôler la collecte active et l'analyse active. Une fois que les IQs atteindront les cibles spécifiées à priori, la collecte active se terminera. Aujourd'hui, l'EUE utilise le taux de réponse pondéré pour décider si la collecte active pourrait être terminée. Le problème avec l'utilisation de taux de réponse est qu'il ne donne pas une bonne indication de la qualité. Il est reconnu qu'un taux de réponse élevé n'implique pas nécessairement que les estimations produites soient de bonnes qualités. Donc, dans le PISE, pour déterminer si la collecte active pourrait être terminée, les IQs utilisés seront plus centrés sur des aspects de la qualité directement reliés aux estimations, comme la variance et le biais. Les IQs seront utilisés d'une façon similaire pour diriger l'analyse. Une fois que les estimations atteindront un niveau de qualité acceptable, l'analyse sera terminée et les estimations pourront être publiées.

Les IQs seront aussi utilisés afin d'aider à allouer et à prioriser les efforts de collecte et de vérification. Les scores de mesure d'impact (voir section 2.2.2) sont définis à partir des IQs et servent à gérer la collecte active et l'analyse active. Dans l'EUE, la gestion de la collecte est gérée avec une fonction de score basée sur le revenu des unités collectées. La fonction de score est utilisée pour prioriser les unités pour le suivi, mais étant donné que la fonction de score ne dépend que du revenu, le suivi se fait la plupart du temps seulement sur les grandes entreprises. Cette stratégie pourrait introduire un biais si un ajustement de poids est fait pour tenir compte des non-répondants. Mais dans l'EUE, l'imputation est utilisée pour traiter la non-réponse totale de sorte que le risque de biais est grandement réduit (Xie, Godbout, Youn et Lavallée, 2011). Dans l'avenir, la fonction de score sera remplacée avec un sous-échantillon, peut-être basé sur les IQs, de façon à minimiser les risques de biais même si un ajustement de poids est utilisé pour traiter la non-réponse totale.

Le PISE utilisera trois IQs de base : le coefficient de variation (CV) total, le taux de non-réponse pondéré et l'erreur prédite absolue (EPA). Le CV total qui sera utilisé par le PISE tiendra compte de la variance d'échantillonnage et la variance due à l'imputation. Disons que l'estimateur d'un total Y est $\hat{Y}$, le CV total de $\hat{Y}$ est donné par

$$\widehat{CV}(\hat{Y}) = \frac{\left(\hat{V}_{SAM}(\hat{Y}) + \hat{V}_{NR}(\hat{Y}) + \hat{V}_{MIX}(\hat{Y})\right)^{1/2}}{\hat{Y}}$$

où $\hat{V}_{SAM}(\hat{Y})$ est la variance d'échantillonnage de $\hat{Y}$, $\hat{V}_{NR}(\hat{Y})$ est la variance due à l'imputation, et $\hat{V}_{MIX}(\hat{Y})$ est la covariance entre la variance d'échantillonnage et le modèle d'imputation. Pour plus de détails sur l'estimation de chacune de ces composantes, voir Turmelle, Godbout et Bosa (2012).

Le deuxième IQ qui sera utilisé par le PISE est le taux de non-réponse pondéré économiquement. Cet IQ sera utilisé seulement s'il est décidé que le CV total n'est pas approprié. Ça pourrait, entre autre, arriver dans des cas où la composante de variance due à la non-réponse est trop complexe à calculer. Dans ces cas, l'IQ est défini comme

$$\hat{R}_{NR} = \frac{\sum_{k \in s} w_k x_k I_k}{\sum_{k \in s} w_k x_k}$$

où w_k est le poids d'échantillonnage de l'unité k, x_k est le revenu de l'unité k (disponible pour toutes les unités) et I_k est un indicateur de non-réponse défini comme

$$I_k = \begin{cases} 1 & \text{si l'unité } k \text{ est non-répondante} \\ 0 & \text{sinon.} \end{cases}$$

Le dernier IQ qui sera employé dans le PISE est l'erreur prédite absolue relative (EPAR) qui sera utilisé pour gérer l'analyse active. Cet IQ est basé sur la méthodologie de vérification sélective (Hedlin, 2008) et dépend de valeurs prédites. Si $\tilde{y}_k$ est une valeur prédite disponible avant le commencement de la collecte, l'erreur prédite absolue (EPA) est définie comme

$$d_k = |\hat{y}_k^* - \tilde{y}_k|$$

où

$$\hat{y}_k^* = \begin{cases} y_k & \text{si l'unité } k \text{ est répondante} \\ y_k^* & \text{si l'unité } k \text{ est non-répondante} \end{cases}$$

et y_k^* est une valeur imputée si l'unité k est non-répondante. Donc, l'IQ pour l'EPAR pour une variable donnée est défini par

$$\hat{R}_{EPAR} = \frac{\sum_{k \in s} w_k d_k I_{EDIT,k}}{\sum_{k \in s} w_k \hat{y}_k^*}$$

où $\sum_{k \in s}$ indique une sommation sur toutes les unités k dans l'échantillon s et $I_{EDIT,k}$ est une variable indiquant si la valeur a été vérifiée par l'analyste et est définie comme

$$I_{EDIT,k} = \begin{cases} 1 & \text{si la valeur } \hat{y}_k^* \text{ n'a pas été vérifiée} \\ 0 & \text{si la valeur } \hat{y}_k^* \text{ a été vérifiée.} \end{cases}$$

Donc, l'IQ pour l'EPA est simplement le ratio entre la sommation pondéré d'EPA et l'estimation du total Y.

Au début, le PISE se servira de ces trois IQs de base mais le travail sur de nouveaux IQs va continuer. Ces trois IQs ont été évalués pendant l'été de 2012 en utilisant les données de l'EUE ; les résultats préliminaires semblent démontrer qu'il y a des efficacités qui pourraient être réalisées avec la mise en place de cette approche.

2.2.2 Les scores de mesure d'impact

Tel que mentionné, les IQs seront utilisés afin de décider si la collecte active peut être terminée. En plus, les IQs jouent un rôle dans les scores de mesure d'impact (MI). Les scores de MI vont mesurer l'influence d'une unité sur une estimation ou un IQ. L'idée est d'utiliser les scores de MI afin d'identifier les unités qui auront un effet sur les estimations ou les IQs et d'entreprendre les démarches nécessaires pour obtenir l'information requise.

Si θ représente un paramètre (une estimation ou un IQ), le score MI d'une unité k sur θ est la différence standardisée entre le paramètre estimé, $\hat{\theta}$, et sa valeur prédite, $\tilde{\theta}$, quand l'unité k passe d'une valeur prédite à une valeur observée. C'est-à-dire

$$MI_k = \frac{\left(\tilde{\theta} - \hat{\theta}\right)}{\varphi}$$

où φ est une constante de standardisation.

Pour le PISE, le score MI associé au CV total de $\hat{Y}$ est

$$\widehat{MI}\left[\widehat{CV}(\hat{Y})\right] = \frac{\left(\delta_k\left[\hat{V}_{SAM}(\hat{Y})\right] + \delta_k\left[\hat{V}_{NR}(\hat{Y})\right] + \delta_k\left[\hat{V}_{MIX}(\hat{Y})\right]\right)^{1/2}}{\hat{Y}}$$

où $\delta_k\left[\hat{V}_{SAM}(\hat{Y})\right]$ est la différence entre $\hat{V}_{SAM}(\tilde{Y})$ et $\hat{V}_{SAM}(\hat{Y})$ étant donné que $\hat{V}_{SAM}(\tilde{Y})$ est égal à $\hat{V}_{SAM}(\hat{Y})$ lorsque l'unité k passe d'un statut de non-répondant à répondant. Les autres composantes sont définies de façon semblable. Pour les expressions des $\delta_k\left[\hat{V}_{SAM}(\hat{Y})\right]$, $\delta_k\left[\hat{V}_{NR}(\hat{Y})\right]$ et $\delta_k\left[\hat{V}_{MIX}(\hat{Y})\right]$, voir Turmelle, Godbout et Bosa (2012).

Le score MI pour le deuxième IQ, soit le taux de non-réponse pondéré économiquement, est

$$\widehat{MI}_k(\hat{R}) = \frac{w_k x_k}{\sum_{k\in s} w_k x_k}$$

où x_k est la valeur du revenu sur la base de sondage pour l'unité k et w_k est le poids d'échantillonnage associé avec l'unité k. Le dernier score de MI dans le PISE, associé à l'EPAR, est

$$\widehat{MI}_k(EPAR) = \frac{w_k d_k I_{EDIT,k}}{\sum_{k\in s} w_k y_k}.$$

Pour plus de détails voir Turmelle, Godbout et Bosa (2012)

2.2.3 La gestion active de la collecte

La gestion active de la collecte aura deux parties : le suivi des questionnaires rejetés au contrôle et le suivi des cas de non-réponse. Les deux parties seront gérées en utilisant les scores de MI afin de suivre seulement les unités qui auront un impact significatif sur les résultats finaux. Aujourd'hui dans l'EUE, pratiquement tous les questionnaires échouant aux règles de contrôle sont suivis peu importe leur impact sur les estimations. Il est fort possible qu'une grande portion de ces suivis ne soit pas essentielle à l'obtention de bons résultats. En utilisant les scores de MI, le PISE ne fera du suivi qu'auprès des unités avec un score au-dessus d'un seuil, indiquant que l'unité aura un impact important sur θ, qui pourrait être une estimation ou un IQ. Les seuils ne sont pas encore définis mais ils dépendront du budget disponible, des contraintes de qualité et/ou des contraintes opérationnelles.

Pour le suivi des cas de non-réponse, le PISE se servira des scores de MI pour identifier les non-répondants à suivre qui devraient améliorer de façon significative les IQs si ces unités deviennent répondantes. Le plan original était de faire un suivi pour un sous échantillon des non-répondants, mais à cause de contraintes opérationnelles il a été décidé pour les premières années du PISE de continuer d'utiliser un processus similaire à celui de l'EUE, c.-à-d l'utilisation d'une fonction score et d'un suivi des unités avec les plus grands scores. Cependant, au lieu d'utiliser une fonction score qui dépend seulement du revenu, le PISE se servira des scores de MI afin que les unités avec le plus grand impact sur la qualité réelle soient suivies et non nécessairement celles avec les plus grands revenus. Comme la gestion des questionnaires rejetés au contrôle, le nombre de suivis des cas de non-réponse dépendra du budget, des contraintes de qualité et/ou des contraintes opérationnelles.

2.2.4 La gestion active de l'analyse

Le deuxième volet de l'EC est la gestion active de l'analyse. Aujourd'hui dans l'EUE, il y a un sentiment que les analystes regardent presque tous les enregistrements afin de s'assurer qu'ils sont sans erreurs. Il est clair que cette vérification prend beaucoup de temps et de ressources, mais il est moins clair qu'elle a toujours un impact important sur la qualité des estimations. En utilisant les scores MI, la gestion active de l'analyse aidera les analystes à identifier les unités qui auront un impact significatif sur la qualité des résultats. Ainsi, les unités à vérifier ne seront pas nécessairement toujours les plus grandes unités de l'échantillon. L'objectif global est de vérifier seulement les unités qui en valent la peine. Le but ultime est de réussir à passer moins de temps à inspecter tous les enregistrements de l'échantillon afin de publier les résultats plus rapidement et de dépenser moins d'argent en analysant les données.

2.3 Résumé

Un objectif clé du PISE est de rendre les processus de la collecte et du traitement de données plus efficaces. Avec le modèle d'EC, on s'attend à ce qu'on puisse épargner de l'argent, réduire le fardeau de réponse et publier les résultats plus rapidement qu'aujourd'hui. Au cœur du modèle d'EC se trouvent les IQs et les scores de MI, sur lesquels la gestion active de la collecte et de l'analyse seront basés.

3.0 Étude du suivi des cas de non-réponse

3.1 Introduction

Tel que mentionné précédemment, le plan original du PISE était de faire le suivi d'un échantillon de non-répondants et de ne pas utiliser la fonction de score qui est utilisée dans l'EUE. Afin de trouver la meilleure façon de tirer l'échantillon, une étude par simulation a été menée pour comparer quelques options. En utilisant l'hypothèse d'un budget fixe pour les suivis, l'étude avait pour objectif de répondre aux questions suivantes :

- Est-il mieux de faire un suivi ciblé (c'est-à-dire, choisir un échantillon de non-répondants) ou de faire un suivi de tous les non-répondants ?
- Si un suivi ciblé est mieux, quelle est la meilleure façon de sélectionner les unités à suivre ?

Notez que pour l'étude 'meilleure' est évaluée en terme d'erreur quadratique moyenne (EQM). Étant donné que l'étude était conçue pour le PISE, la simulation a été menée dans le contexte des enquêtes auprès des entreprises. Au lieu de générer une population par simulation, on a utilisé des données provenant de l'enquête mensuelle sur les services de restauration et débits de boisson (EMSRDB). L'étude s'est servie d'un mois de données. Ces données, rapportées ou imputées, étaient disponibles pour toutes les unités de l'échantillon.

Comme les enquêtes auprès des entreprises typiques, l'EMSRDB utilise un plan d'échantillonnage stratifié aléatoire simple avec des strates à tirage complet et à tirage partiel. Les unités à tirage complet sont les grandes entreprises importantes et sont normalement toutes suivies. Ces unités ont donc été exclues de l'étude. Un envoi postal a été simulé avec une probabilité de réponse de chaque unité fixe. Après l'envoi postal, le suivi des non-répondants a été simulé en utilisant différentes méthodes de sélection des unités à suivre. Il y a trois résultats possibles associés à un suivi : une réponse, un refus ou un cas toujours en cours (non-résolu). Chacun de ces résultats a sa propre probabilité et ces trois dernières peuvent toutes être différentes. De plus, ces probabilités peuvent également être différentes de la probabilité de répondre à l'envoi postal. Finalement, un coût est associé à chacun des trois résultats et ces coûts peuvent aussi être différents les uns des autres. Le

suivi a continué jusqu'à ce que le budget pour le suivi soit épuisé ou que toutes les unités aient été résolues (une réponse ou un refus). Un estimateur de Horvitz-Thompson a été utilisé avec un ajustement des poids pour la non-réponse résiduelle (les cas toujours en cours). Pour chaque méthode de sélection des unités à suivre, le biais relatif et l'EQM empiriques ont été calculés et comparés.

3.2 Notation

Plus précisément, supposons qu'il y a une population d'unités *i*, U, stratifiées en L strates, $U = \cup U_h$, $h=1, \ldots L$, où U_h représente la population de la strate *h*. Un échantillon aléatoire simple stratifié des unités, *s*, est sélectionné avec les probabilités de sélection π_{1hi} pour l'unité *i* dans la strate *h*, $i=1, \ldots, n_h$ et $h=1, \ldots, L$. On envoie un questionnaire aux unités sélectionnées dans l'échantillon et n_r unités répondent avec les probabilités de réponse p_{1hi}. Puis, après l'envoi postal simulé, un échantillon de non-répondants est sélectionné de s_{nr} pour un suivi avec probabilités $\pi_{2hi,}$ où s_{nr} est l'ensemble des non-répondants à l'envoi postal. Les suivis sont simulés et les unités suivies ont une probabilité de réponse de p_{2hi}, qui peut être différente de leur probabilité de répondre à l'envoi postal. Le suivi continue jusqu'à ce que le budget fixe soit épuisé ou que toutes les unités de suivi aient été résolues (réponse ou refus). Si le budget est épuisé et qu'il reste encore des non-répondants, un ajustement de poids est fait pour tenir compte de leur contribution aux estimations.

L'estimateur stratifié du total de la population utilisé dans l'étude est

$$\hat{Y} = \sum_h \sum_{i \in s_{hr}} w_{1hi} y_{hi} + \sum_h \sum_{i \in s_{hFr}} \widetilde{w}_{2hi} y_{hi} \qquad (1)$$

où s_{hr} est l'ensemble des unités dans la strate *h* qui répondent à l'envoi postal, s_{hFr} est le groupe des unités dans la strate *h* qui répondent au suivi, $w_{1hi}=1/\pi_{1hi}$ est le poids associé avec la sélection de l'échantillon original, $\widetilde{w}_{2hi} = w_{1hi} \times 1/\pi_{2hi} \times a_{2hi}$ est le poids ajusté pour la non-réponse pour les unités qui répondent au suivi et y_{hi} est la variable d'intérêt pour les unités *i* dans la strate *h*. Pour des raisons de simplicité, on a supposé que les classes de repondération correspondent aux strates et l'ajustement de non-réponse dans l'estimateur (1), au niveau des strates, est

$$a_{2hi} = \frac{\sum_{i \in s_{hF}} w_{1hi}/\pi_{2hi}}{\sum_{i \in s_{hFr}} w_{1hi}/\pi_{2hi}}$$

où s_{hFr} est le groupe des unités dans la strate *h* qui répondent au suivi.

Dans le cas extrême où il n'y a aucun répondant au suivi dans la classe de repondération, l'estimateur (1) devient

$$\hat{Y} = \sum_h \sum_{i \in s_{hr}} w_{1hi} a_{1hi} y_{hi}$$

ou

$$a_{1hi} = \frac{\sum_{i \in s_{hr}} w_{1hi} + \sum_{i \in s_{hF}} w_{1hi}/\pi_{2hi}}{\sum_{i \in s_{hr}} w_{1hi}}$$

et s_{hF} est l'ensemble des unités dans la strate *h* sélectionnées pour un suivi.

Les probabilités de répondre à l'envoi postal ou au suivi étaient soit uniformes, ou modérément ou fortement corrélées à la variable d'intérêt (les ventes). Les plans d'échantillonnage considérés pour l'échantillon de suivi étaient :

- Suivi de toutes les unités non-répondantes (un recensement)
- Un échantillon aléatoire simple (EAS) des unités non-répondantes
- Un échantillon aléatoire simple (EAS) stratifié des unités non-répondantes
- Un échantillon avec probabilité proportionnelle à la taille (PPT) des unités non-répondantes.

Les tailles d'échantillon pour le suivi des non-répondants étaient de 100, 200, 300, 400, 500, 700, 900 et 1188 pour le suivi de tous les non-répondants. Pour comparer les différents plans d'échantillonnage, on a calculé le biais relatif et l'EQM empiriques.

Avant de montrer les résultats de l'étude, on voudrait ajouter quelques précisions concernant l'échantillonnage avec probabilité proportionnelle à la taille. On a choisi le revenu de l'unité comme mesure de taille mais on ne dit pas que cette mesure est la meilleure option. En fait, on a choisi le revenu et le revenu pondéré par le poids issu de l'envoi postal comme mesures de taille parce que le revenu était disponible pour toutes les unités de la population. Au lieu d'utiliser le revenu, on a aussi discuté l'utilisation d'autres mesures comme la probabilité de répondre, la contribution à l'indicateur de représentativité (Schouten, Cobben et Bethlehem, 2009), le biais conditionnel ou les scores de MI discutés dans le contexte du PISE.

3.3 Résultats

3.3.1 Biais Relatif

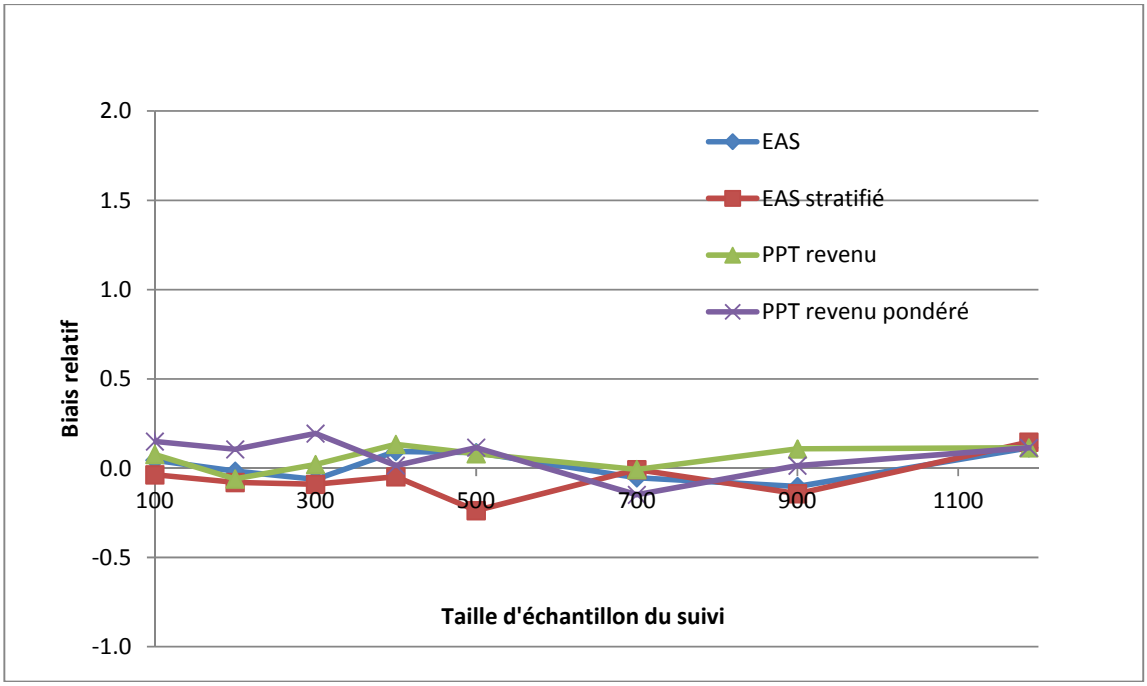
La première mesure d'évaluation est le biais relatif qui est définie comme

$$BR = \frac{\frac{\sum_{r=1}^{R} \hat{Y}_r}{R} - \tilde{Y}}{\tilde{Y}} * 100\%$$

où R est le nombre de répliques dans la simulation, $\hat{Y}_r$ est l'estimation obtenue pour la réplique r, et $\tilde{Y}$ est l'estimation obtenue à partir de l'échantillon original.

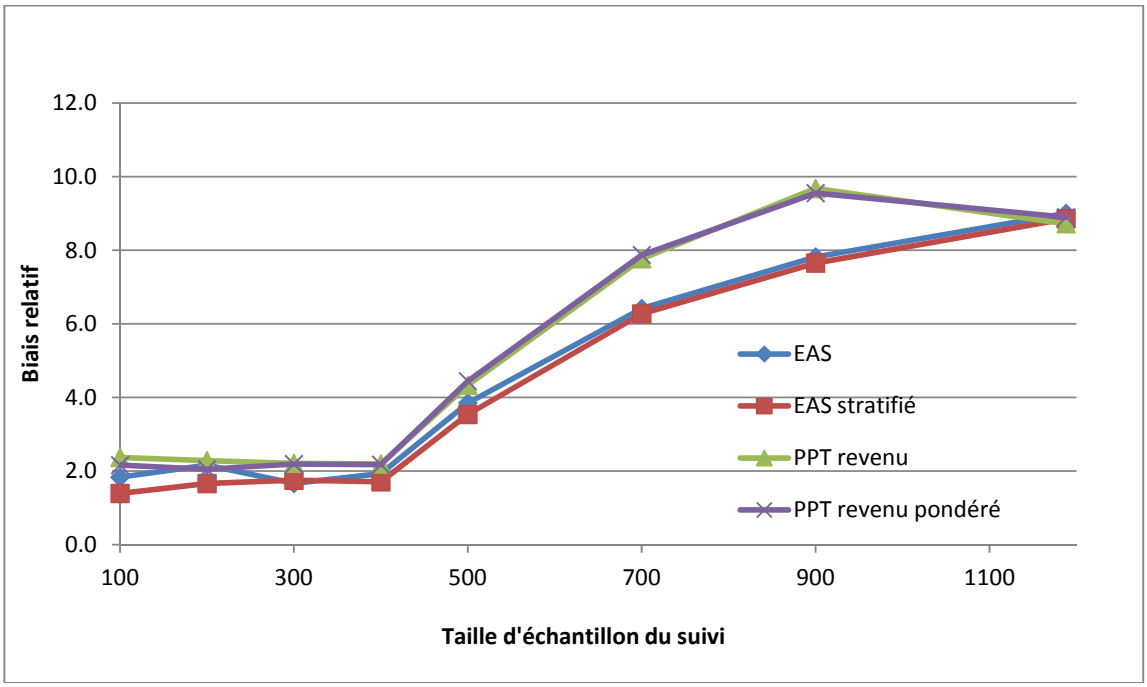
Les résultats pour le biais relatif sont présentés aux figures 3 à 6. La figure 3 montre les résultats pour la situation où les probabilités de répondre à l'envoi postal et aux suivis sont uniformes. Comme on s'y attend, on note qu'il n'y a pas de biais dans cette situation et qu'il n'y a pas une grande différence entre les quatre plans d'échantillonnage.

Figure 3 – Biais Relatif : Envoi postal – Uniforme / Suivi – Uniforme



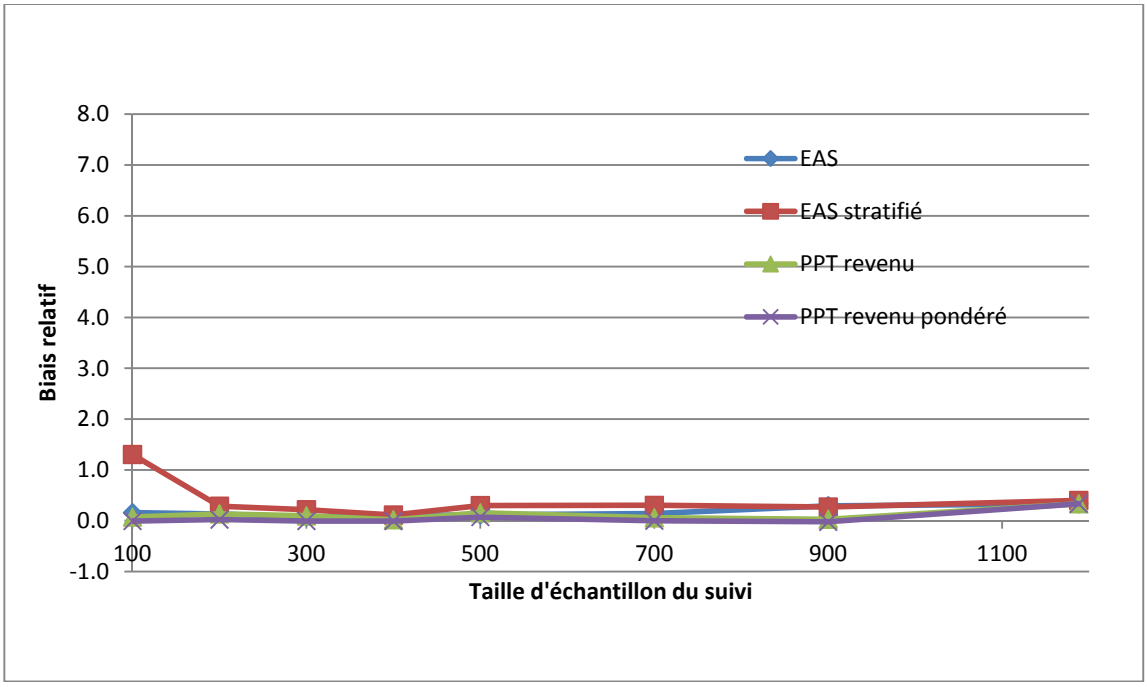
La figure 4 présente les biais relatifs dans la situation où la probabilité de réponse à l'envoi postal est uniforme mais la probabilité de réponse au suivi est modérément corrélée aux ventes. Il y a un petit biais pour tous les plans d'échantillonnage et le biais est plus petit quand la taille d'échantillon du suivi est de 400 ou moins. Le biais augmente à mesure que la taille d'échantillon du suivi augmente. L'augmentation dans le biais est due au budget limité qui ne permet pas assez de suivis afin de résoudre tous les cas. Dans cette situation, les répondants du suivi ne sont pas un échantillon représentatif et l'utilisation d'un ajustement de non-réponse cause le biais d'augmenter. La situation où tous les non-répondants sont suivis montre un biais de presque 9%. Les deux plans PPT semblent avoir un biais un peu plus grand, mais la différence n'est probablement pas significative.

Figure 4. – Biais Relatif : Envoi postal - Uniforme / Suivi - Corrélée aux ventes



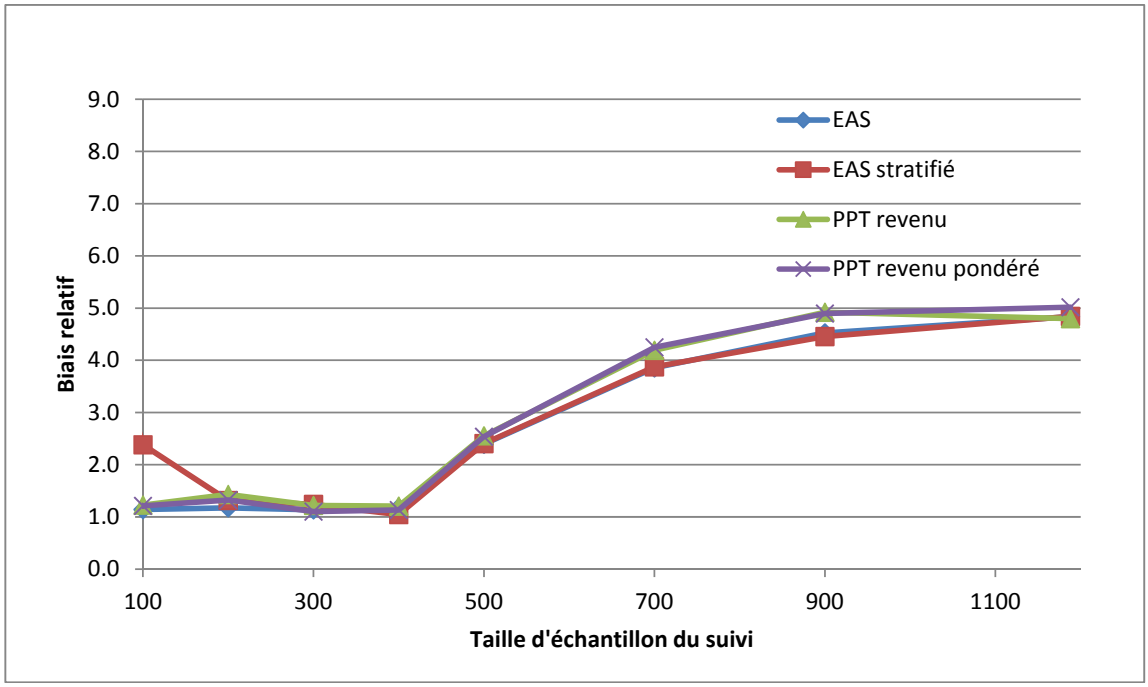
La figure 5 présente les résultats quand la probabilité de répondre à l'envoi postal est corrélée fortement aux ventes et la probabilité de répondre au suivi est uniforme. Même si la probabilité de répondre à l'envoi postal est corrélée aux ventes, le fait de faire des suivis des non-répondants a enlevé le biais potentiel. Le petit biais qui apparait dans le plan EAS stratifié avec une taille de 100 est probablement dû à l'erreur Monte Carlo et il n'est probablement pas significativement différent de zéro.

Figure 5 – Biais Relatif : Envoi postal - Corrélée aux ventes / Suivi – Uniforme



La figure 6 montre les résultats pour la situation où les probabilités de répondre à l'envoi postal et au suivi sont corrélées aux ventes (fortement pour l'envoi postal et modérément pour le suivi). Comme prévu, il y a un biais pour tous les plans d'échantillonnage. Le scénario où l'on suit tous les non-répondants a le plus grand biais. Le biais est minimisé quand la taille d'échantillon du suivi est d'environ 400 unités, ce qui correspond à la situation où il y a assez de budget pour résoudre presque tous les cas.

Figure 6 – Biais Relatif : Envoi postal et suivi- Corrélée aux ventes



3.3.2 Racine de l'EQM relative (REQMR)

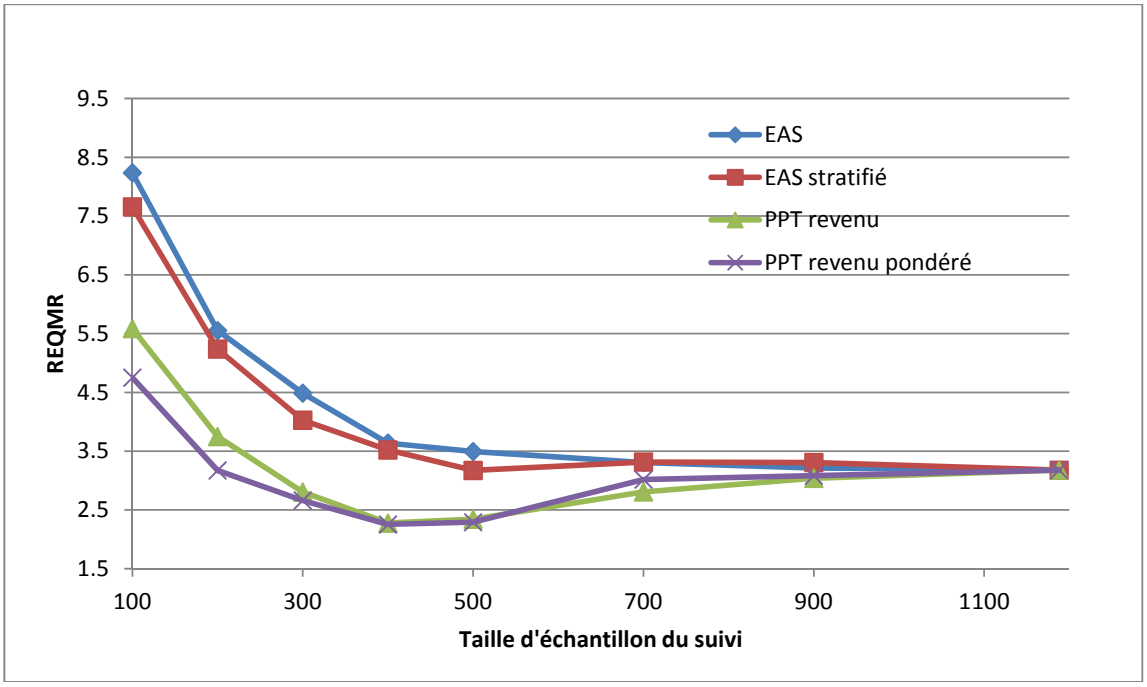
La deuxième mesure d'évaluation est la racine de l'EQM relative qui est définie comme

$$REQMR = \frac{\sqrt{\frac{\sum_{r=1}^{R}\left(\hat{Y}_r - \tilde{Y}\right)^2}{R}}}{\tilde{Y}}$$

où R est le nombre de répliques dans la simulation, $\hat{Y}_r$ est l'estimation obtenue pour la réplique r, et $\tilde{Y}$ est l'estimation obtenue à partir de l'échantillon original.

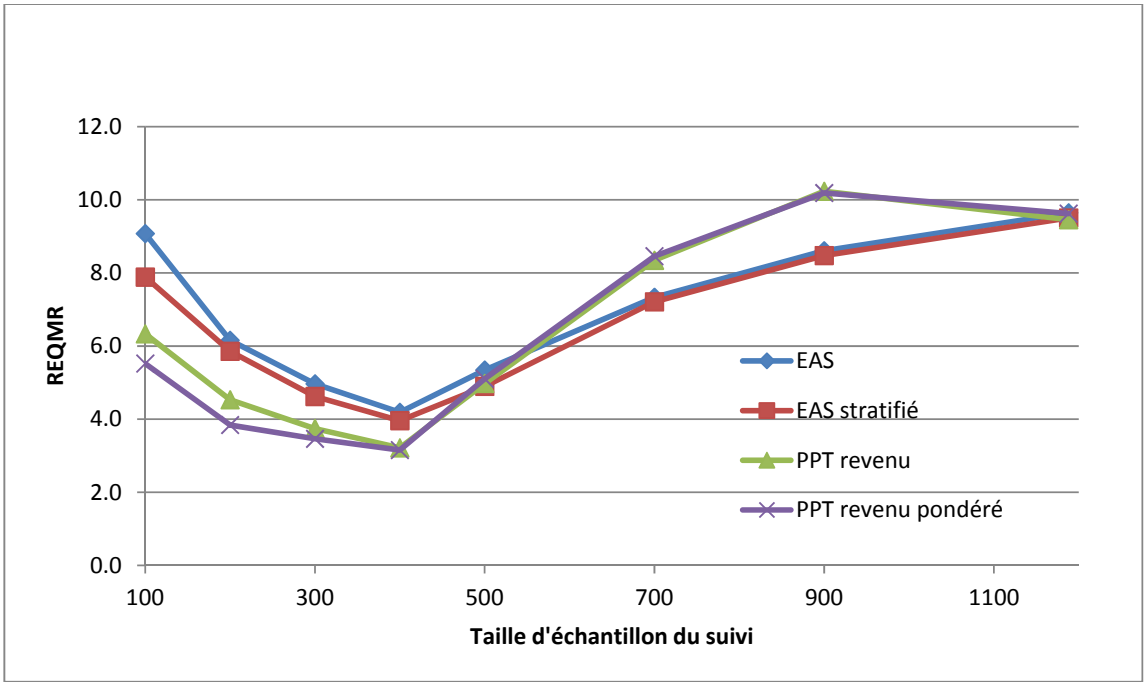
La figure 7 présente les REQMRs pour les quatre plans d'échantillonnage quand les deux probabilités de répondre sont uniformes. Parce qu'il n'y a pas de biais dans cette situation, la REQMR est égale à la variance. Il est intéressant de noter que la REQMR est minimisée quand tous les non-répondants sont suivis. Quand un échantillon de non-répondants est tiré, la deuxième phase d'échantillonnage introduit de la variabilité dans les poids, ce qui résulte en un accroissement de la variance. Quand la taille de l'échantillon des non-répondants devient assez grande, environ 300 unités, la variabilité dans les poids diminue et la variance s'approche de celle du recensement. Les pertes d'efficacité sont négligeables par rapport à un recensement pour des tailles d'échantillon supérieures à 700. En général, les plans PPT sont un peu plus efficaces que les deux plans EAS.

Figure 7. REQMR: Envoi postal - Uniforme / Suivi – Uniforme



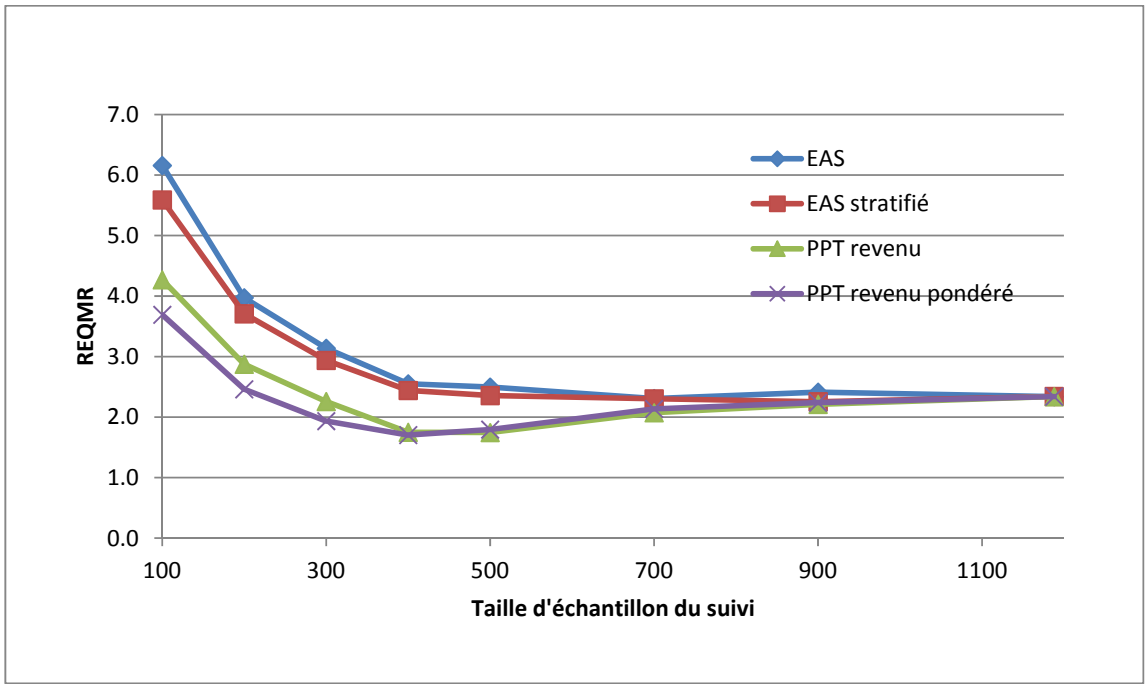
Quand la probabilité de répondre à l'envoi postal est uniforme et la probabilité de répondre au suivi est corrélée aux ventes, on voit que les plans PPT sont un peu plus efficaces que les plans EAS et que la REQMR est minimisée quand la taille d'échantillon est d'environ 400 unités (voir la figure 8). Au-delà de cette taille, la REQMR augmente de façon monotone. Dans cette situation, il vaut donc mieux faire le suivi d'un échantillon des non-répondants qu'un recensement.

Figure 8 – REQMR : Envoi postal - Uniforme / Suivi - Corrélée aux ventes



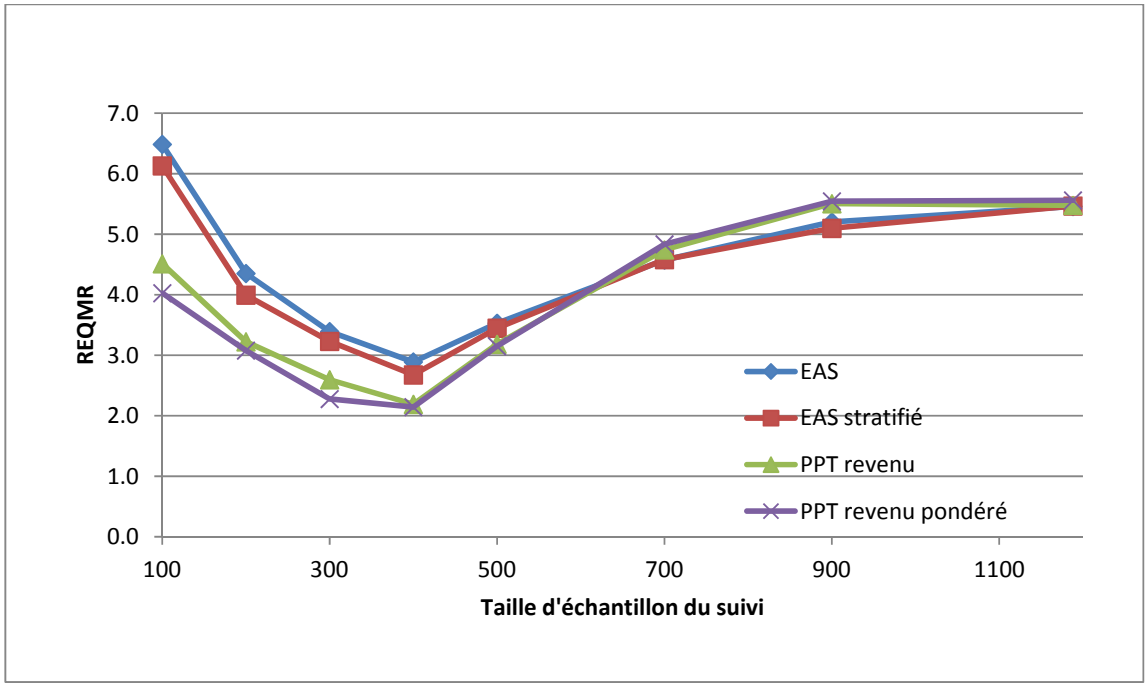
La figure 9 présente les REQMRs pour la situation où la probabilité de répondre à l'envoi postal est corrélée aux ventes mais la probabilité de répondre au suivi est uniforme. Les courbes sont très semblables à celles de la figure 7, où les REQMR sont plus grands pour les suivis avec une taille d'échantillon inférieure à 300 ou 400 unités. Pour les échantillons avec une taille plus grande que 500 ou 600 unités, il n'y a pas beaucoup de différences entre suivre tous les non-répondants et suivre un échantillon de non-répondants.

Figure 9 – REQMR : Envoi postal - Corrélée aux ventes / Suivi - Uniforme



La figure 10 présente les résultats quand les probabilités de répondre à l'envoi postal et au suivi sont corrélées aux ventes. Les courbes sont très semblables à celles de la figure 8 au sens où les plans d'échantillonnage sont plus efficaces que la stratégie de suivre tous les non-répondants, sauf pour les deux plans EAS avec une taille de suivi de seulement 100 unités. Aussi, la REQMR est minimisée quand la taille d'échantillon est d'environ 400 unités. En général, les plans PPT sont plus efficaces des plans EAS quand la taille d'échantillon est suffisamment petite pour que la majorité des suivis soient résolus. Une fois que la taille devient trop grande pour résoudre tous les suivis, les plans PPT deviennent moins efficaces probablement à cause du biais (voir la figure 6).

Figure 10 – REQMR : Envoi postal et suivi - Corrélée aux ventes



3.4 Résumé

En résumé, les résultats de l'étude ont corroboré ce à quoi on s'attendait. C'est-à-dire qu'il y a un biais quand la probabilité de répondre au suivi est corrélée aux ventes. En suivant un échantillon des non-répondants, le biais peut être réduit mais pas éliminé. Si la probabilité de répondre au suivi est uniforme, il n'y a pas de problème de biais.

Pour ce qui est de la REQMR, on note que dans les situations où la probabilité de répondre au suivi est uniforme, l'approche de suivre tous les non-répondants produit des résultats assez raisonnables. Cette approche n'est pas la plus efficace mais elle n'est pas très loin de la méthode la plus efficace. Par contre, si la probabilité de répondre au suivi est corrélée aux ventes, cette approche est souvent beaucoup moins efficace que les autres à cause d'un biais qui pourrait être significatif. Parmi les plans d'échantillonnage, les plans PPT sont normalement plus efficaces que les plans EAS. Une observation importante est que la taille d'échantillon devrait être assez grande pour réduire la variance mais pas au point de résulter en un grand nombre de cas non résolus lors du suivi car cela augmente le biais.

4.0 Conclusions

Le PISE est un grand projet avec beaucoup d'objectifs. L'objectif principal est la réduction des coûts et le modèle d'EC a été conçu afin d'atteindre cet objectif en rendant le processus de collecte plus efficace en terme de rapidité et les processus de suivi des non-répondants et de gestion de l'analyse mieux ciblés. Les IQs et les score de MI sont des éléments fondamentaux de l'EC et vont gérer le suivi des questionnaires rejetés au contrôle, le suivi des cas de non-réponse et le travail des analystes. Le PISE commencera avec trois IQs simples et les scores de MI correspondants, mais le travail sur les autres IQs potentiels continuera.

L'étude par simulation a montré qu'il est plus efficace, en termes de biais et d'EQM, de suivre un échantillon des non-répondants plutôt que tous les non-répondants si la probabilité de répondre est

corrélée à la variable d'intérêt. Si la probabilité de répondre est uniforme, un échantillon n'est pas nécessaire mais cette hypothèse est très forte et, selon notre expérience, n'est pas réaliste. Les résultats de la simulation montrent que les plans d'échantillonnage PPT peuvent être plus efficaces que les plans aléatoires simples (stratifiés ou non), mais on pense qu'il devrait y avoir plus de recherches sur la mesure de taille utilisée. Comme mentionné précédemment, les mesures comme la probabilité de répondre, le biais conditionnel ou les score de MI pourraient donner les meilleurs résultats.

Remerciements

Le travail sur l'étude du suivi des cas de non-réponse a été fait conjointement avec Jean-François Beaumont, David Haziza, Mike Hidiroglou et Elisabeth Neusy de Statistique Canada. Les auteurs voudraient remercier les examinateurs, Elisabeth Neusy et Keven Bosa, pour leurs commentaires qui ont amélioré la version initiale du présent article.

Bibliographie

Hedlin, D. (2008). Local and Global Score Functions in Selective Editing, *Conférence européenne des statisticiens*, Session de travail sur la validation des données statistiques, Vienne (Autriche).

Laflamme, F. (2008). Recherche sur la collecte des données à l'aide de paradonnées à Statistique Canada. Recueil: Symposium 2008, Collecte des données : défis, réalisations et nouvelles orientations. Statistique Canada.

Schouten, B., Cobben, F. et Bethlehem, J. (2009). Indicateurs de la représentativité de la réponse aux enquêtes. *Techniques d'enquête*, **35**, pp. 107-121.

Statistique Canada (1999). *Trousse d'information sur l'Enquête unifiée auprès des entreprises (EUE)***.** Catalogue No 68F0015XIF.

Statistique Canada (2010). *Programme intégré de la statistique des entreprises. Plan directeur.* Rapport non publié, Ottawa, Canada ; Statistique Canada

Turmelle, C., Godbout, S. et Bosa, K. (2012). Methodological Challenges in the Development of Statistics Canada's new Integrated Business Statistics Program, Proceedings of the Fourth International Conference on Establishment Surveys, American Statistical Association, to appear.

Xie, H., Godbout, S., Youn, S. et Lavallée, P. (2011).Collection Follow-Up Operation Using Priority Scores For Business Surveys, *Conférence européenne des statisticiens*, Session de travail sur la validation des données statistiques, Ljubljana (Slovenia).