Quelques travaux de recherche sur la collecte pour les enquêtes auprès des entreprises à Statistique Canada

Wesley Yung, Yanick Beaucage et Claude Turmelle
Statistique Canada

Plan de la présentation

- Introduction
- Deux projets
  - La collecte pour les enquêtes annuelles auprès des entreprises
  - Étude sur le suivi des non-répondants
- Conclusions

Introduction

- La recherche sur la collecte est un sujet relativement nouveau mais très actuel
- Statistique Canada est impliqué dans le domaine depuis quelques années
- La collecte pour les enquêtes auprès des entreprises a des défis différents
  - Par exemple, combien et quelles unités devraient être suivis? Dans quel ordre?

Introduction

- La présentation touchera deux projets reliés à la collecte auprès des entreprises
  - La gestion de la collecte dans les enquêtes annuelles auprès des entreprises
  - Comment sélectionner les unités pour le suivi de non-réponse ?

LA COLLECTE POUR LES ENQUÊTES ANNUELLES AUPRÈS DES ENTREPRISES

Introduction

- L'Enquête unifiée auprès des entreprises (EUE) de Statistique Canada inclut environ 60 enquêtes annuelles
- L'EUE fut mise en place à la fin de la décennie 1990
- Récemment il a été décidé de remanier l'EUE afin de la rendre plus flexible et efficace
- Le nouveau programme est le Programme intégré de la statistique des entreprises (PISE)

Introduction

- Les objectifs du PISE incluent
  - Terminer la collecte active plus tôt
  - Augmenter l'efficacité opérationnelle
  - Réduire la micro vérification
- Afin d'atteindre ces objectifs, on a regardé le modèle de traitement des enquêtes
  - On a trouvé des aspects pour lesquels on pourrait être plus efficace

Les estimations en continu

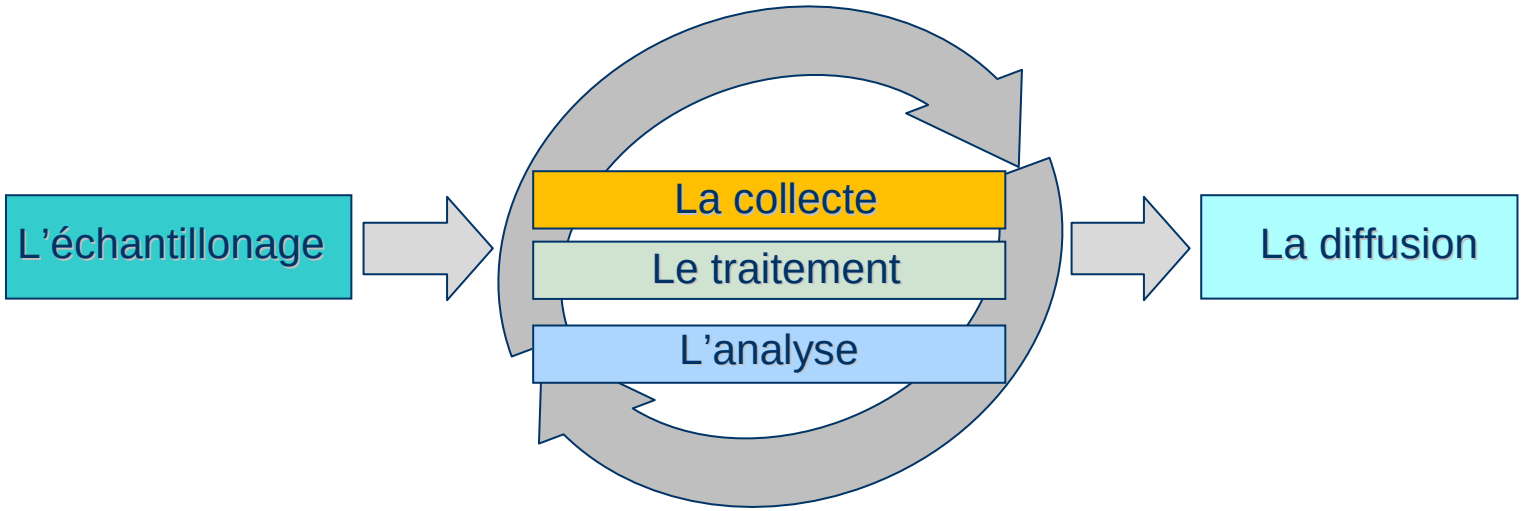
- Le remaniement remet en question le lien entre la collecte et le traitement des données
- Le modèle des estimations en continu (EC) forme un nouveau modèle de traitement qui sera mis en œuvre dans le PISE
- Présentement, le processus est très linéaire

Les estimations en continu

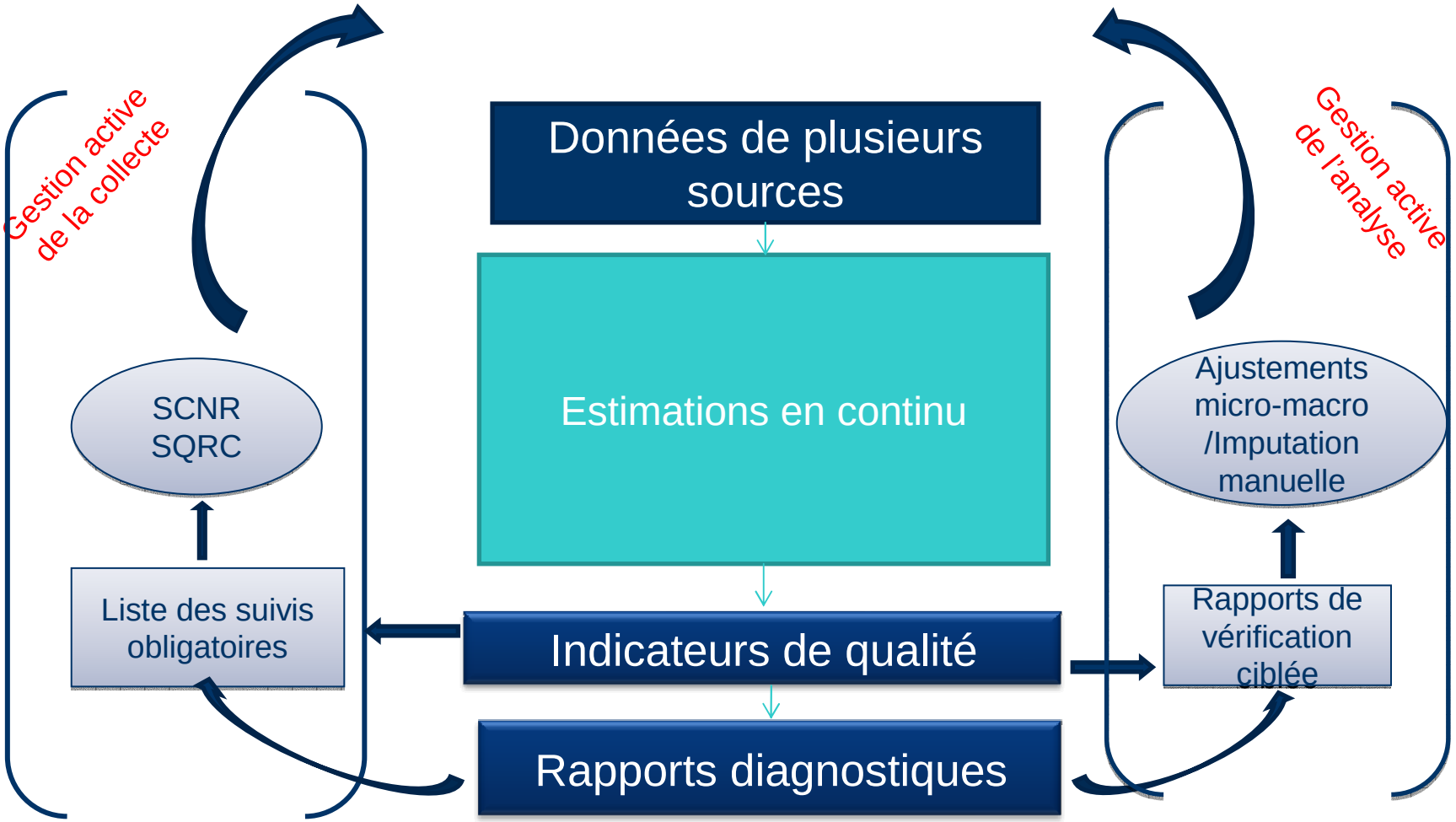
- Dans le modèle actuel, les estimations et les indicateurs de qualité sont disponibles seulement vers la fin de l'étape de traitement
- La collecte est gérée avec une fonction de score et les taux de réponse
  - La fonction de score est basée sur le revenu disponible sur la base de sondage
  - La collecte se termine quand le taux de réponse pondéré atteint le taux ciblé

Les estimations en continu

- Dans le PISE, la collecte, le traitement et l'analyse seront combinés
- La rétroaction sera utilisée pour gérer la collecte et l'analyse dans les EC



Les estimations en continu



Les estimations en continu

- Le modèle des EC comprendra :
  - La gestion active de la collecte
    - Gérer les activités de la collecte, le suivi des questionnaires rejetés au contrôle et des cas de non-réponse, d'une façon efficace
    - Suivre les unités qui auront un impact sur la qualité des résultats
  - La gestion active de l'analyse
    - Gérer efficacement le travail des analystes
    - Analystes travaillent sur les unités qui ont un impact sur la qualité des résultats
- Cette gestion sera faite en utilisant des indicateurs de qualité et des scores de mesure d'impact

Les indicateurs de qualité

- Ils jouent un rôle clé dans les EC en
  - Servant au contrôle de la collecte et l'analyse
    - Quand les indicateurs de qualité atteindront les niveaux pré-spécifiés, la collecte **active** et la vérification se termineront
  - Aidant à allouer et à prioriser les efforts de collecte et de vérification
    - En utilisant les scores de mesure d'impact (MI)

Les indicateurs de qualité

- Les indicateurs de qualité mesurent la qualité d'une variable dans un domaine donné
- Exemples
  - Taux de réponse pondéré économiquement

$$\hat{R}_d = \frac{\sum_{i \in s} w_i r_i x_i \delta_{di}}{\sum_{i \in s} w_i x_i \delta_{di}}$$

où w_i est le poids de sondage, r_i est un indicateur de réponse, x_i est le revenu associé à l'unité i et δ_{di} est un indicateur de domaine

Les indicateurs de qualité

- CV total

$$CV = Y^{-1}\left(V_{SAM} + V_{IMP} + V_{MIX}\right)^{1/2}$$

- Les indicateurs de qualité seront comparés avec des cibles pré-spécifiées afin de décider si la collecte active peut être fermée

Les scores de la mesure d'impact

- Afin d'aider la gestion de suivis et la vérification, les scores de mesure d'impact vont mesurer l'influence d'une unité sur une estimation ou un indicateur de qualité
- Le score MI d'**une** unité *k* sur un paramètre 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

$$MI_k = (\tilde{\theta} - \hat{\theta}) / \varphi$$

Les scores de la mesure d'impact

- Le paramètre θ peut être l'estimation d'un total ou un indicateur de qualité
  - Par exemple, si θ est le total Y

$$\tilde{\theta} = \sum_{i \in s} w_i \tilde{y}_i, \qquad \hat{\theta} = \sum_{i \in s} w_i y_i$$

$$\Rightarrow MI_k = (\tilde{\theta} - \hat{\theta}) / \varphi = w_k (\tilde{y}_k - y_k) / \varphi$$

où

$$\tilde{y}_i = \begin{cases} \tilde{y}_k & \text{si } i = k \\ y_k & \text{sinon} \end{cases}$$

Les scores de la mesure d'impact

- Si θ est le CV d'un estimateur

$$\tilde{\theta} = \hat{Y}^{-1}\left[\tilde{V}_{sam} + \tilde{V}_{NR} + \tilde{V}_{mix}\right]^{1/2},$$

$$\hat{\theta} = \hat{Y}^{-1}\left[\hat{V}_{sam} + \hat{V}_{NR} + \hat{V}_{mix}\right]^{1/2}$$

et

$$MI_k = (\tilde{\theta} - \hat{\theta})/\varphi$$

$$= \hat{Y}^{-1}\left[\left(\tilde{V}_{sam} - \hat{V}_{sam}\right) + \left(\tilde{V}_{NR} - \hat{V}_{NR}\right) + \left(\tilde{V}_{mix} - \hat{V}_{mix}\right)\right]^{1/2}/\varphi$$

Les scores de mesure d'impact

- Les scores MI sont « locaux » (c.-à.-d. au niveau variable/unité) donc ils doivent être combinés afin de produire un score « global » pour chaque unité
- Il y a plusieurs façons de combiner les scores locaux (Hedlin, 2008)
- Les scores globaux peuvent être utilisés pour prioriser le suivi pour les questionnaires rejetés au contrôle et pour la non-réponse

Suivi des questionnaires rejetés

- Les unités qui ont été rejeté au contrôle seront suivies si leur score MI est au-dessus d'un seuil
- Les valeurs prédites seront basées sur les données historiques ou d'autres méthodes d'imputation simple
- Les scores MI seront calculés pour quelques variables et domaines clés seulement
- Des seuils détermineront quelles unités seront suivies
  - Les seuils dépendront du budget, des contraintes de qualité, etc.

Le suivi des cas de non-réponse

- Les unités non-répondantes avec un score global au-dessus du seuil seront sujet au suivi
- L'indicateur de qualité qui sera utilisé pour le suivi de la non-réponse sera la variance due à l'imputation
- Il n'est pas nécessaire d'avoir une valeur prédite
  - L'impact de passer d'un statut de non-répondant à répondant peut-être estimé à l'aide de formules

La gestion active de l'analyse

- Basé sur les scores MI, les analystes valideront les unités qui ont un impact significatif sur la qualité des résultats
  - 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

Où en sommes-nous?

- Les tests du prototype EC en 2011 et 2012
  - Aucune rétroaction aux activités de la collecte
- À l'été 2011
  - Capacité de produire les estimations pendant la collecte, en n'incluant que des indicateurs de qualité de base (taux de réponse)
- À l'été 2012
  - Trois indicateurs de qualité : le taux de réponse, l'erreur prédite et la variance due à l'imputation
  - Les résultats sont attendus à l'automne

ÉTUDE DU SUIVI DES CAS DE NON-RÉPONSE

Les objectifs de l'étude

- Sous l'hypothèse d'un budget fixe pour le suivi
  - Est-il mieux de faire un suivi ciblé ou de faire un suivi de tous les non-répondants?
  - Si un suivi ciblé est mieux, qu'elle est la meilleure façon de sélectionner les unités à suivre?
- Notez que, « meilleure » est évaluée en terme d'erreur quadratique moyenne (EQM)
- Une simulation a été menée dans le contexte des enquêtes auprès des entreprises

Description de la simulation

- Échantillon de l'Enquête mensuelle sur les services de restauration et débits de boisson (EMSRDB)
- On a les données pour toutes les unités échantillonnées (mélange de données rapportées et imputées)
- L' EMSRDB utilise un plan stratifié aléatoire simple avec des strates à tirage complet et à tirage partiel

Description de la simulation

- Les unités à tirage complet sont les grandes unités importantes et elles sont normalement toutes suivies, donc elles ont été exclues de l'étude
- Un envoi postal a été simulé avec une probabilité de réponse fixe

Description de la simulation

- Après l'envoi postal, le suivi a commencé
- Un suivi peut se solder en une réponse, un refus ou un cas toujours en cours
  - On a fait varier les probabilités de chaque possibilité
- Le coût pour chaque résultat possible était différent
  - Réponse: 5 unités
  - Refus: 2 unités
  - Toujours en cours: 1 unité

Description de la simulation

- Le suivi a continué jusqu'à ce que le budget soit épuisé ou que toutes les unités aient été résolues
- Un ajustement des poids pour la non-réponse résiduelle a été utilisé

Notation

- Population d'unités, U, stratifié en L strates
- Dans la strate h, un EAS d'unités, s_h, de taille n_h est sélectionné avec probabilités π_{1hi}
- Les unités choisies reçoivent un questionnaire et alors, s_{hr} unités répondent avec probabilités p_{1hi}
- Disons qu'un échantillon de non-répondants est sélectionné de s_{hnr} avec probabilités π_{2hi} pour le suivi

Notation

- Les probabilités de réponse au suivi sont p_{2hi}, qui peuvent-être différente de p_{1hi}
- S'il y a quelques unités qui ne répondent pas (c.-à-d. le budget est épuisé), un ajustement pour la non-réponse est utilisé pour tenir compte de leur contribution

Notation

- L'estimateur stratifié du total de la population

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

où $w_{1hi} = 1/\pi_{1hi}$ et $\tilde{w}_{2hi} = w_{1hi} \times 1/\pi_{2hi} \times a_{2hi}$ est le poids ajusté des unités qui répondent au suivi. L'ajustement pour la non-réponse est l'inverse de la probabilité de répondre au niveau de la strate.

Notation

- Les probabilités de répondre à l'envoi postal ou au suivi étaient soit uniformes, ou modérément ou fortement corrélées à la variable d'intérêt (ventes)
- Les plans étudiés pour l'échantillon de suivi étaient;
  - Suivi de toutes les unités non-répondantes
  - Un EAS des unités non-répondantes
  - Un EAS stratifié des unités non-répondantes
  - Un échantillon PPT des unités non-répondantes

Notation

- Les mesures d'évaluation étaient; le biais relatif et l'EQM empirique de l'estimateur

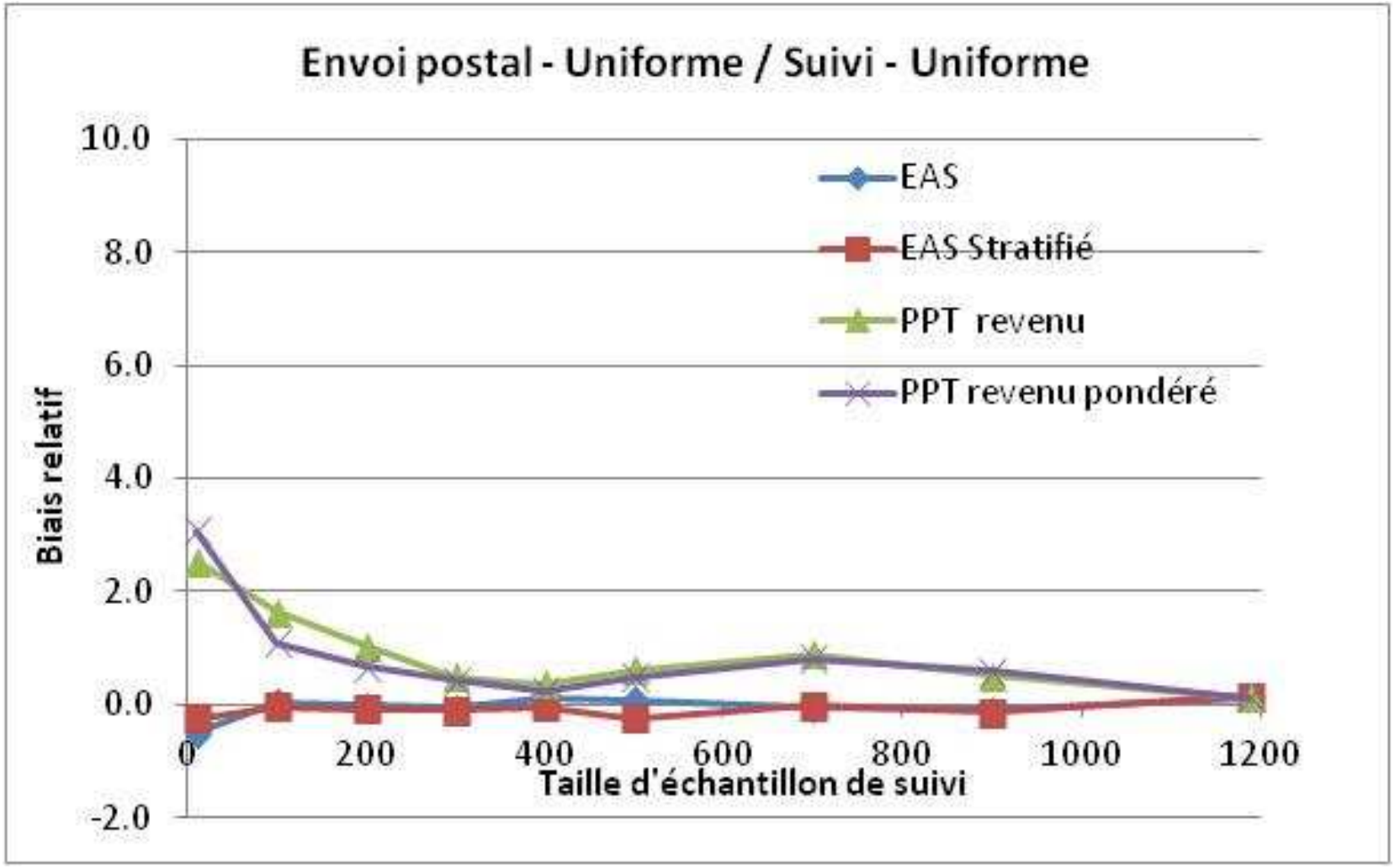
Notes

- PPT présente des défis
  - Quelques revenus sont nul. C'est sans doute pourquoi nous observons un biais dans le cas uniforme/uniforme – On regarde ça
  - On ne dit pas que le PPT avec le revenu comme mesure de taille est la meilleure option. Peut-être qu'on devrait utiliser la probabilité de répondre, la contribution à l'indicateur R, le biais conditionnel ou les scores MI comme mesure de taille

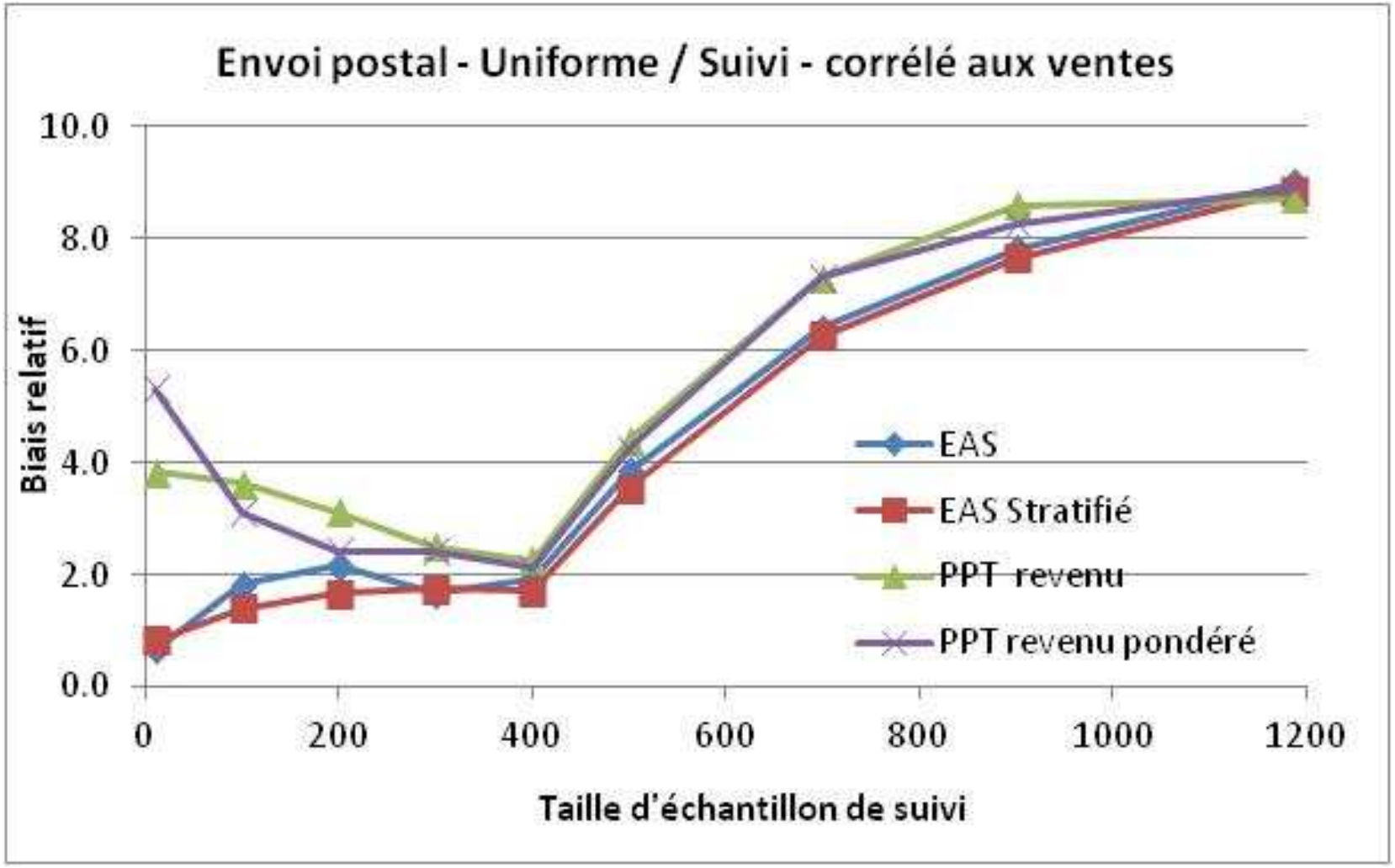
Notes

- L’erreur Monte-Carlo peut-être très grande quand la taille d’échantillon de suivi est petite. Donc, certains biais pourraient ne pas être significativement différents de 0

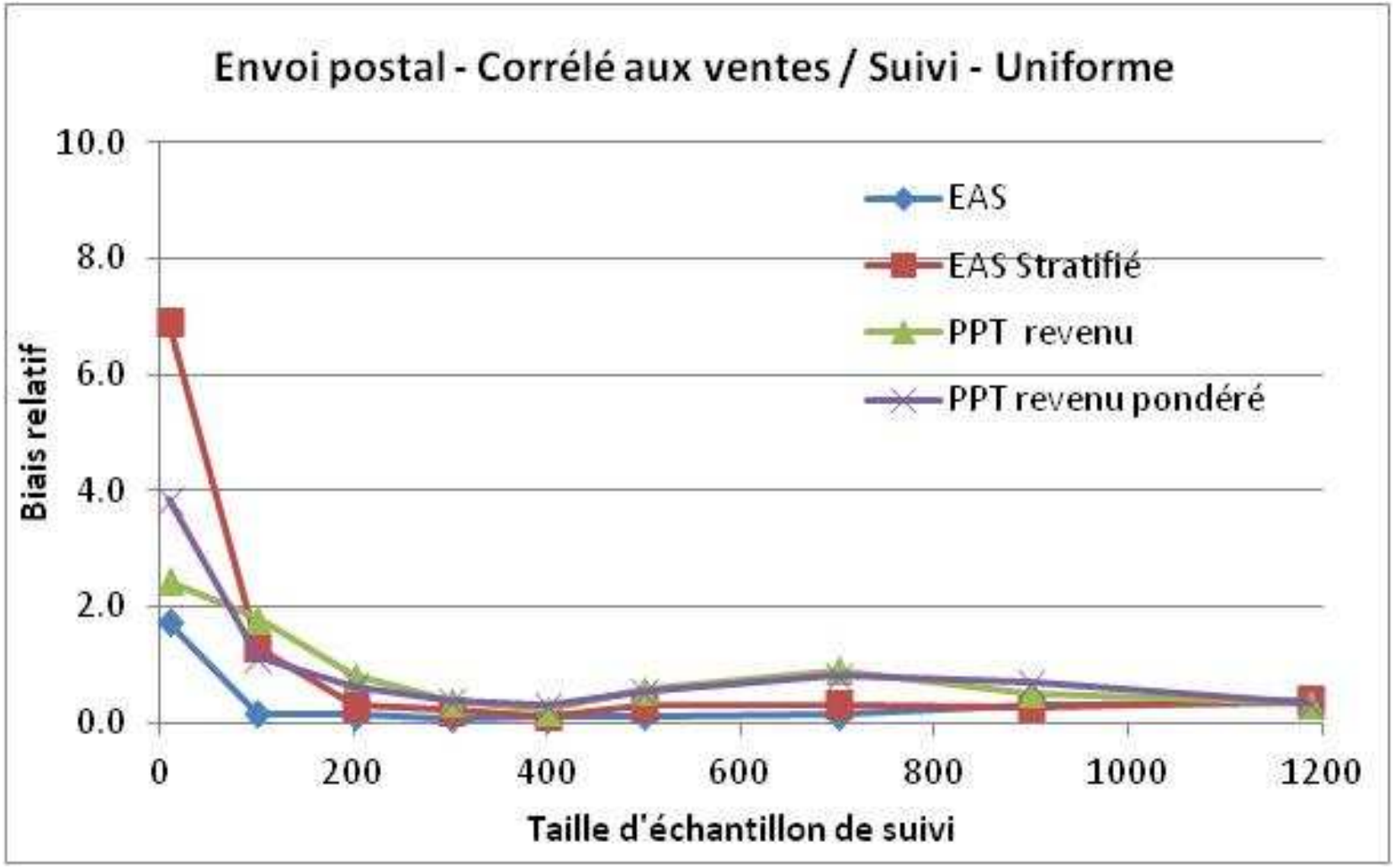
Le biais relatif



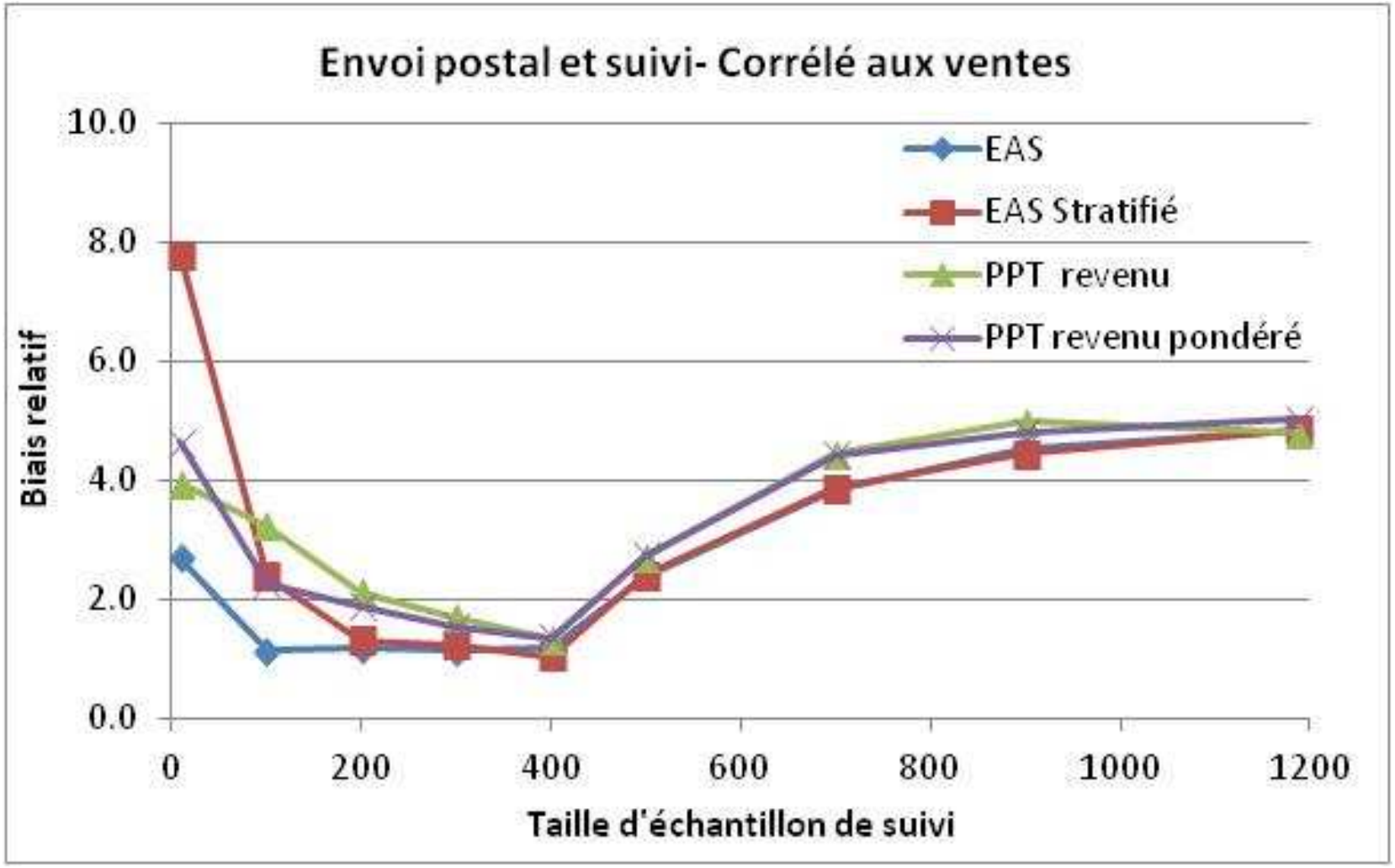
Le biais relatif



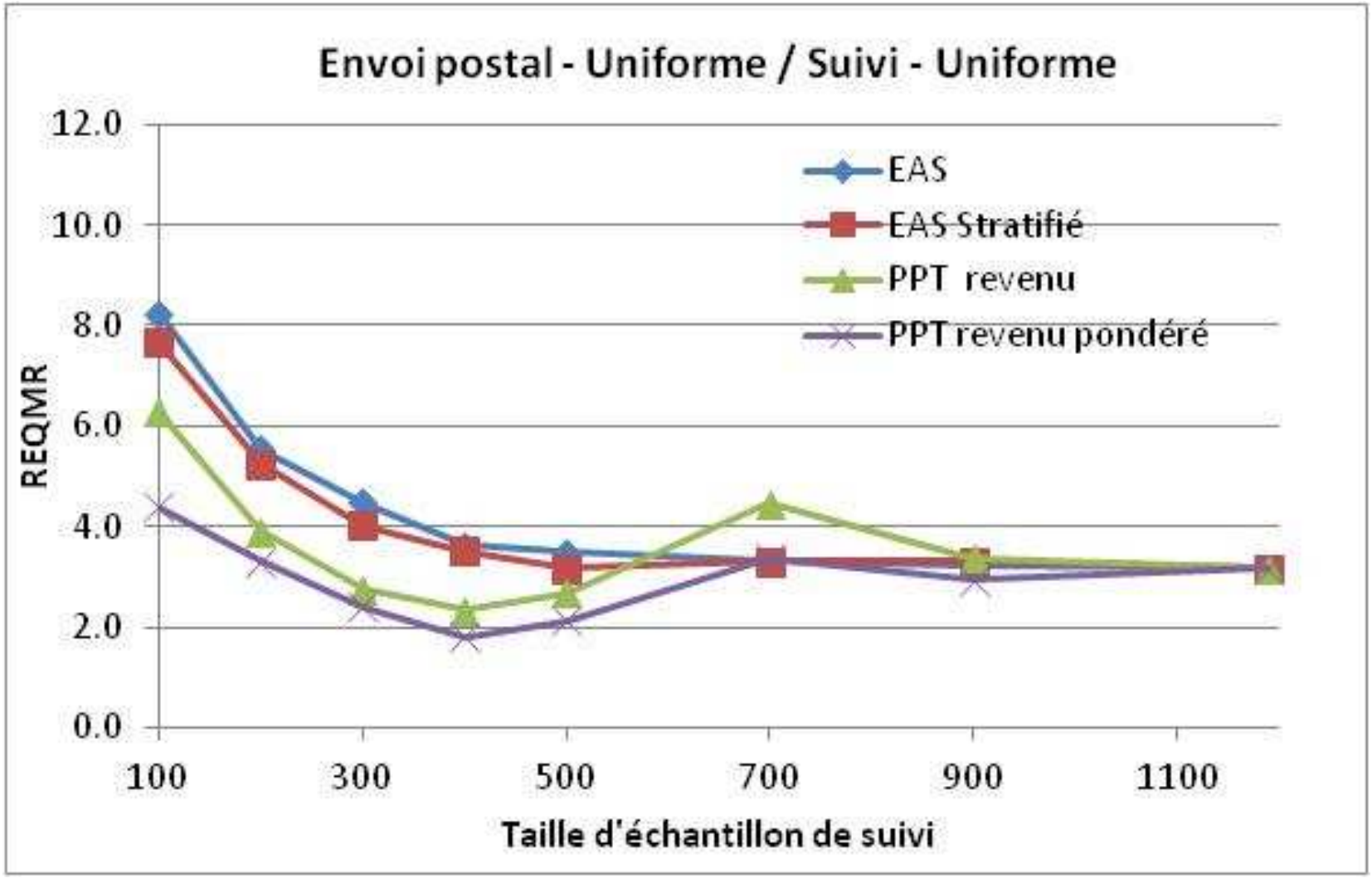
Le biais relatif



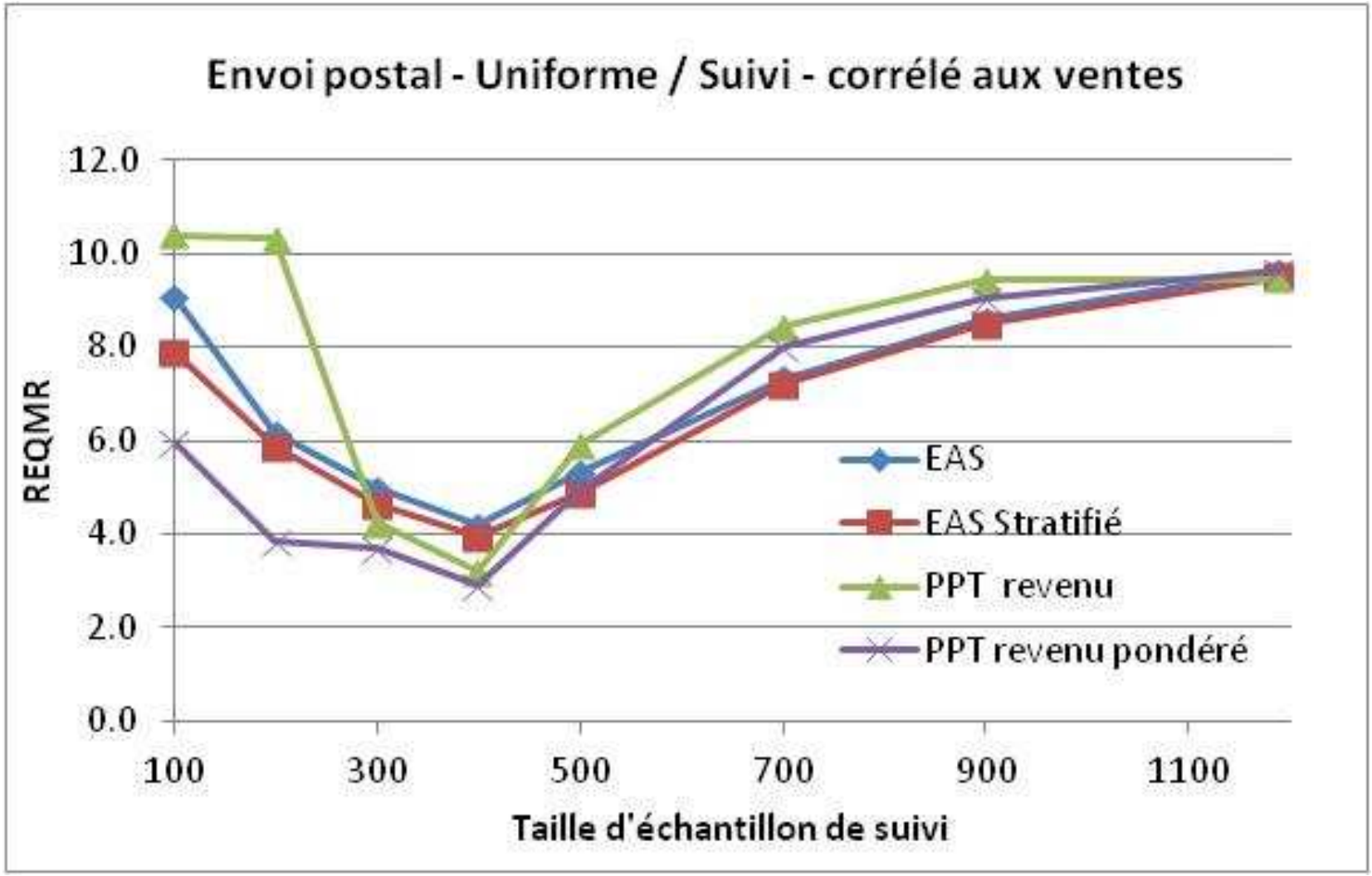
Le biais relatif



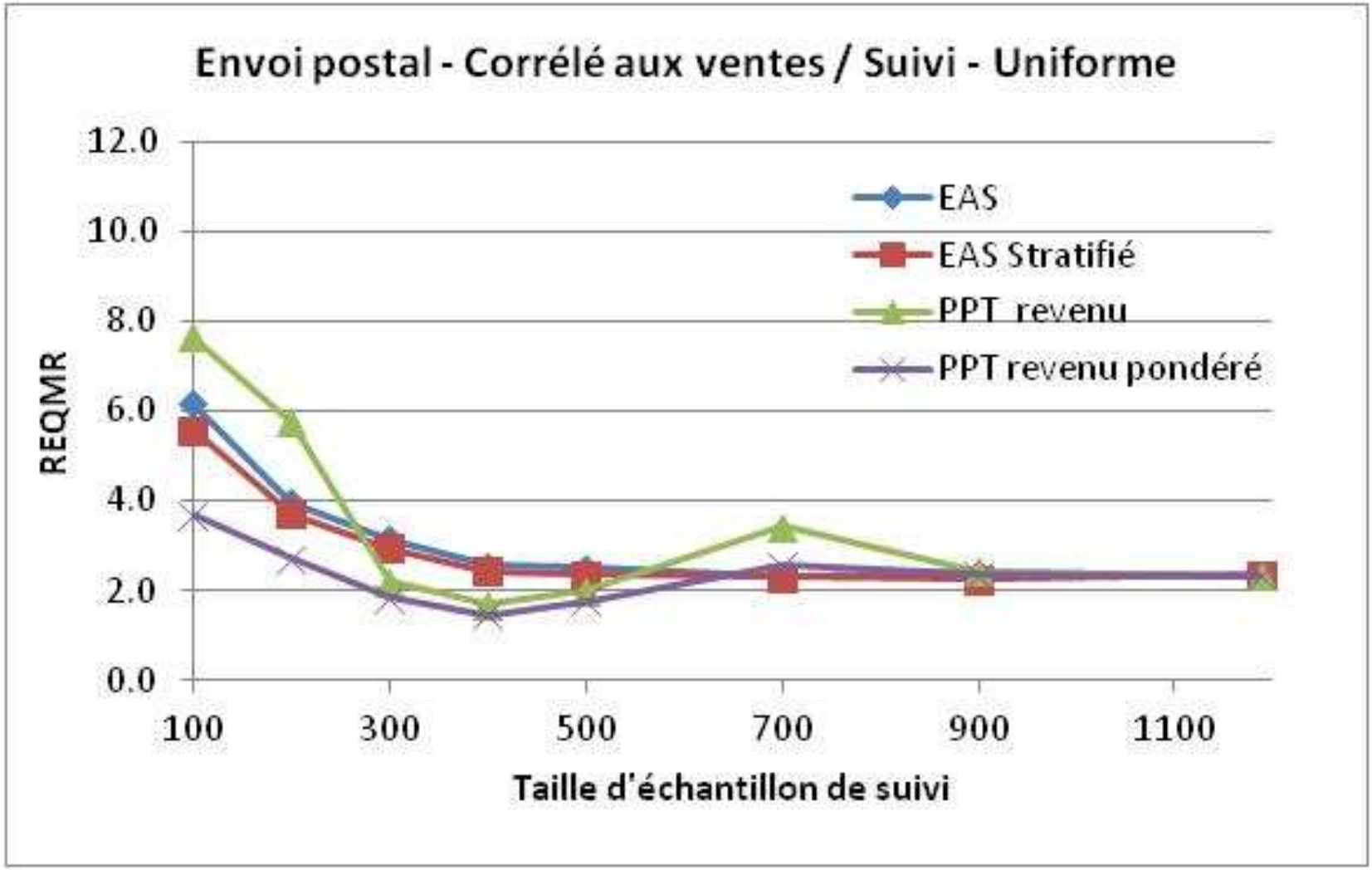
Racine de l'EQM relative



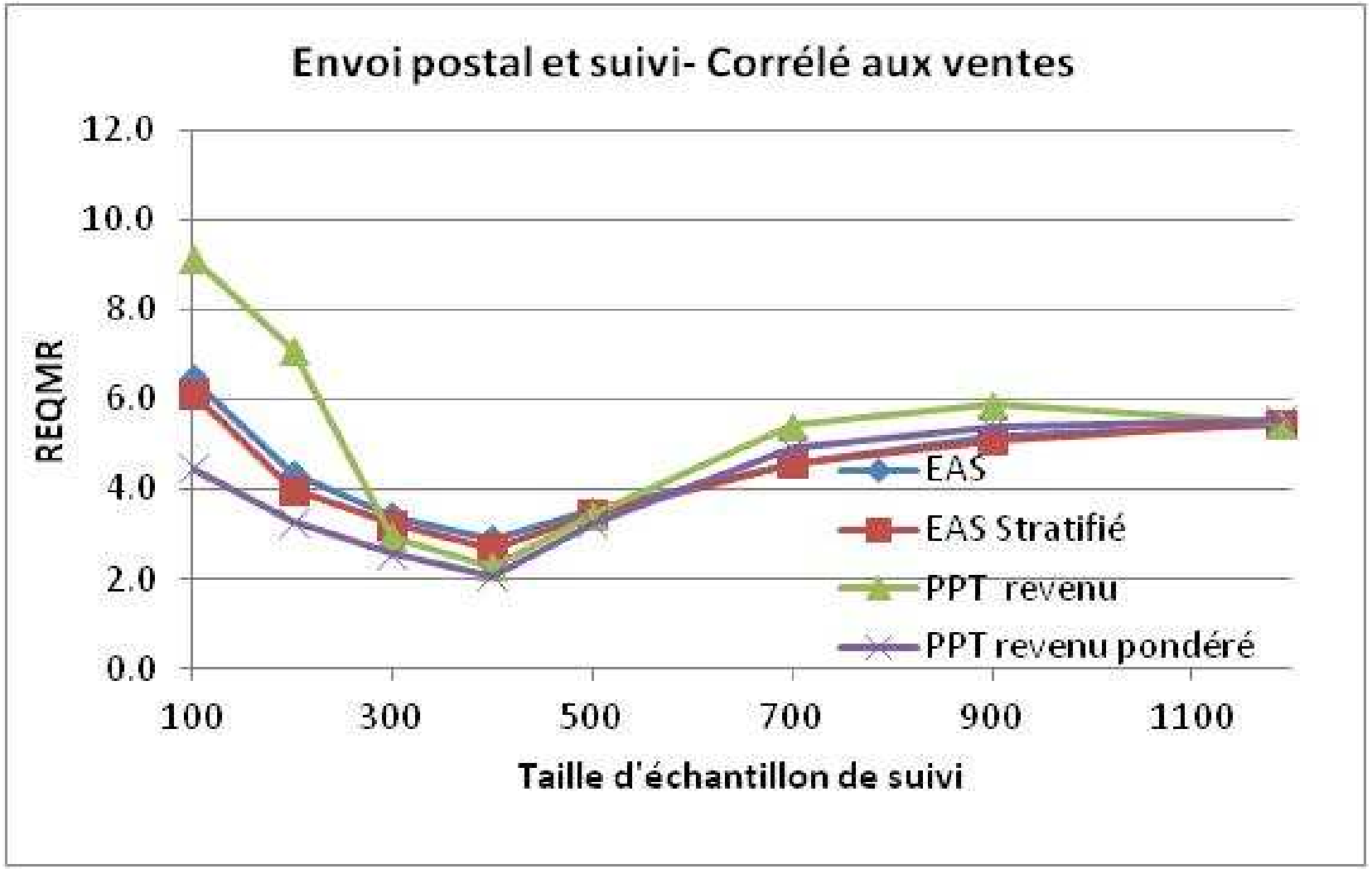
Racine de l'EQM relative



Racine de l'EQM relative



Racine de l'EQM relative



Conclusions

- PISE
  - Les EC jouent un rôle clé dans la gestion de la collecte
  - En utilisant les indicateurs de qualité et les scores MI, on s'attend a ce que les EC puissent rendre la collecte plus efficace sans détériorer significativement la qualité des estimations
  - Le suivi des cas de non-réponse sera géré avec les MI mais dans le futur on regardera si on peut faire le suivi d'un échantillon d'unités
    - Peut-être un échantillon PPT avec le MI comme mesure de taille

Conclusions

- Le suivi des cas de non-réponse
  - Les résultats sont plus ou moins comme on s'y attendait
  - En général, un sous-échantillon produit les meilleurs résultats
  - Il faut faire attention à la taille du sous-échantillon
    - La taille devrait être assez grande pour réduire la variance mais pas trop grande pour éviter un biais

- For more information please contact:
- Pour plus d'information veuillez contacter :

Wesley Yung
Statistique Canada

Wesley.Yung@statcan.gc.ca