No 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2005 : Défis méthodologiques reliés aux besoins futurs d'information

2005

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2005 de Statistique Canada
Défis méthodologiques pour les besoins futurs d'information

CONTOURNER UNE DIFFÉRENCE DE CONCEPTS ENTRE DEUX SOURCES DE DONNÉES POUR LA PRODUCTION D'ESTIMATIONS

Serge Godbout et Chantal Grondin[1]

RÉSUMÉ

L'Enquête sur l'emploi, la rémunération et les heures est une enquête mensuelle utilisant deux sources de données, soit un recensement de dossiers administratifs et une enquête auprès d'établissements. La source administrative, provenant des comptes de déductions sur la paye de l'Agence du revenu du Canada, contient les variables d'emploi et de paye mensuelle brute. Les données d'enquête contiennent également l'emploi et la paye mensuelle brute, en plus d'un ensemble de variables connexes non disponibles sur la source administrative. Les données d'enquête permettent ainsi de construire des modèles qui servent à imputer massivement un éventail de variables dérivées sur la source administrative. Ce plan de sondage repose sur le fait que les concepts d'emploi et de paye mensuelle brute sont les mêmes sur les deux sources. Cela n'est cependant pas toujours le cas en pratique, causant une certaine instabilité dans les estimations. Dans cet article, nous décrirons différentes solutions apportées au plan de sondage et au modèle d'imputation massive pour permettre de contourner cette différence de concepts et ainsi produire des estimations plus stables dans le temps. Des résultats sur l'estimation des gains hebdomadaires moyens à l'aide des différents scénarios complèteront l'article.

MOTS CLÉS : Sources de données; plan de sondage; estimation par régression.

1. INTRODUCTION

L'Enquête sur l'emploi, la rémunération et les heures (EERH) est une enquête mensuelle utilisant deux sources de données : un recensement de dossiers administratifs et une enquête auprès d'un échantillon d'établissements, soit l'Enquête sur la rémunération auprès des entreprises (ERE). L'EERH a pour objectif de produire des estimations de niveaux et de tendances pour l'emploi, les gains, les heures et autres variables connexes et ce, par province et par industrie (Grondin et Lavallée, 2001).

La source administrative est constituée de formulaires de retenues à la source fournis par l'Agence du revenu du Canada (ARC). Puisque ces formulaires ne peuvent être associés à une province ou une industrie en particulier, ces données doivent être agrégées au niveau de l'entreprise, puis désagrégées au niveau des établissements qui, eux, sont associés à une seule province et une seule industrie. Nous avons ainsi, pour chaque établissement j de l'univers des établissements U, le nombre d'employés $E_{A,j}$ et la paye mensuelle brute $P_{A,j}$. La portion enquête, quant à elle, consiste en un échantillon S stratifié d'environ 11 000 établissements choisis à partir d'une base liste, soit le Registre des entreprises (RE) de Statistique Canada. Ces établissements peuvent être reliés à la source administrative. Parmi les variables recueillies pour chaque unité $i \in S$, nous comptons le nombre d'employés $E_{E,i}$, la paye mensuelle brute $P_{E,i}$ et les gains hebdomadaires $G_{E,i}$. Le poids de sondage des unités de la strate h est $w_i = 1 / p_i$, où p_i est la probabilité de sélection de l'unité i. Ces données d'enquête servent à construire des modèles qui seront utilisés pour imputer massivement un éventail de variables dérivées sur la source administrative, dont les gains hebdomadaires. Les estimations finales sont produites à partir des données imputées massivement sur la source administrative.

Ce plan de sondage suppose que les concepts d'emploi et de paye mensuelle brute sont les mêmes sur les deux sources. Mais le temps a démontré que cela n'est pas toujours le cas en pratique, causant une certaine instabilité dans les estimations.

1. Serge Godbout, Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, 100 promenade du Pré Tunney, Immeuble R.-H.-Coats, 11e étage, Ottawa (Ontario), Canada, K1A 0T6, Serge.Godbout@statcan.ca.
Chantal Grondin, Division des enquêtes spéciales, Statistique Canada, 150 promenade du Pré Tunney, Immeuble Principal, pièce 2500-G, Ottawa (Ontario), Canada, K1A 0T6, Chantal.Grondin@statcan.ca.

Dans cet article, nous décrirons différentes solutions apportées pour permettre de contourner cette différence de concepts entre les deux sources. Nous présenterons d'abord certains des problèmes qui alimentent ou causent en partie la différence entre les deux sources. Nous proposerons différentes améliorations au plan de sondage afin d'augmenter la cohérence entre les deux sources. Nous comparerons ensuite divers modèles et scénarios d'estimation des variables dérivées et verrons leur impact sur la stabilité des estimations. Dans le cadre de cet article, nous allons nous limiter à l'étude de l'estimation des gains hebdomadaires moyens même si l'EERH s'intéresse à d'autres estimations de formes similaires, comme les heures hebdomadaires moyennes payées.

2. PROBLÈMES RENCONTRÉS ET AMÉLIORATIONS AU PLAN DE SONDAGE

Le plan de sondage de l'EERH repose sur l'hypothèse que les concepts communs aux deux sources sont équivalents. Cependant, plusieurs différences existent. D'abord, l'emploi et la paye des établissements de la source administrative ne correspondent pas toujours à ceux des établissements sélectionnés dans l'ERE. Cela arrive lorsque les données sont manquantes et doivent être imputées, par exemple. Une autre raison est que les valeurs d'emploi et de paye des établissements de la source administrative ne sont pas déclarées directement par les répondants, mais proviennent d'un processus d'agrégation (à l'entreprise) et de désagrégation (à l'établissement). La désagrégation est basée sur une proportion approximative du nombre d'employés de chaque établissement provenant du RE. Bien que cette désagrégation produise généralement des valeurs d'emploi assez fiables, nous ne pouvons en dire autant de la paye mensuelle. En effet, puisque nous utilisons une proportion d'employés pour désagréger la paye, ceci revient à supposer que tous les employés des différents établissements d'une entreprise gagnent le même salaire mensuel moyen, ce qui peut s'avérer faux dans la réalité. En ce moment, aucune autre information n'est disponible pour permettre de faire la désagrégation de la paye autrement. Toutefois, afin de rendre les données des deux sources plus cohérentes, nous envisageons de recueillir l'information de l'ERE au niveau des entreprises, et d'utiliser le même processus de désagrégation qui est utilisé sur la source administrative pour générer l'information au niveau des établissements dans l'enquête. Nous aurons donc une meilleure cohérence entre les valeurs d'emploi et de paye des deux sources.

La collecte d'information dans l'ERE est présentement faite au niveau des établissements. Mais il arrive que certaines unités déclarent à notre insu une information qui se rapporte à un niveau différent (au niveau de l'entreprise ou d'un groupe d'établissements, par exemple). Conséquemment, les données déclarées par ces unités ont un poids relatif trop important lorsque vient le temps de construire les modèles de régression. Ce problème particulier pourra toutefois être résolu par la collecte d'information au niveau de l'entreprise, à condition que les répondants arrivent effectivement à déclarer à ce niveau. Mais le problème survient aussi lorsque nous avons affaire à des unités qui changent de classification ou de province dans le temps. Leurs valeurs, en particulier leur nombre d'employés, une fois pondérées deviennent trop influentes. Une solution proposée ici est de réduire le poids de ces unités jugées trop influentes. Cet ajustement de la pondération est fait en tentant d'ajuster la distribution du nombre d'employés de l'ERE par groupe de taille à celle observée sur la source administrative.

Un autre problème particulier à la variable de paye se pose également. Dans la source administrative, la paye mensuelle brute correspond aux salaires payés durant le mois de référence, incluant généralement des montants pour des jours travaillés au cours du mois précédent, dépendamment de la définition des périodes de paye de l'établissement. Du côté de l'ERE cependant, la paye mensuelle brute, bien qu'elle soit définie de la même façon que sur la source administrative, semble être déclarée différemment. Certains répondants déclarent les sommes payables pour les jours travaillés entre le premier et le dernier jour du mois, alors que d'autres se contentent de déclarer deux périodes bihebdomadaires ou quatre périodes hebdomadaires de paye sans tenir compte des mois contenant une période de paye « en extra ». L'effet est que la paye mensuelle brute dans l'enquête ne montre pas la même tendance entre deux mois que celle de la source administrative. Une solution proposée ici est de redéfinir les modèles de régression de sorte à ne plus utiliser la variable de paye mensuelle brute provenant de l'enquête comme variable explicative.

3. MODÈLES D'ESTIMATION

Dans cette étude, nous nous attardons à l'estimation des gains hebdomadaires moyens (GHM). Nous verrons d'abord l'estimation des GHM à partir des données de l'ERE seulement. Nous verrons ensuite deux façons de produire les estimations des GHM à l'aide des données combinées des deux sources. Nous verrons ainsi

comment un changement dans la façon de modéliser les données permet de contourner le problème d'incohérence entre les deux sources.

3.1 Estimation à partir des données d'enquête seulement (méthode E)

L'estimation des GHM peut se faire à partir des données d'enquête seulement. Nous nous servons alors de l'estimateur d'Horvitz-Thompson en utilisant le poids de sondage correspondant à l'inverse de la probabilité de sélection. Pour un domaine *d*, les gains hebdomadaires moyens ($G\hat{H}M$) sont estimés à l'aide d'un quotient de sommes pondérées : $G\hat{H}M_{\mathrm{E}} = \sum_{i \in d \cap S} w_{E,i} G_{E,i} \Big/ \sum_{i \in d \cap S} w_{E,i} E_{E,i}$ où $w_{E,i}$ est le poids de sondage, $G_{E,i}$ sont les gains hebdomadaires et $E_{E,i}$ est le nombre d'employés de chaque établissement *i* échantillonné par l'enquête.

L'estimation des gains hebdomadaires moyens se fait ici selon une méthode connue, mais la taille d'échantillon de l'enquête ne permet pas d'obtenir des estimations suffisamment précises pour plusieurs domaines (niveau provincial par industrie).

3.2 Estimation à partir des deux sources sans appariement des microdonnées (méthode A1)

Pour estimer les GHM à l'aide des deux sources sans appariement des microdonnées, nous nous servons d'estimateurs synthétiques. C'est-à-dire que nous modélisons les gains hebdomadaires par employé en fonction de la paye mensuelle par employé des données d'enquête, puis nous appliquons les coefficients obtenus à la paye mensuelle par employé de la source administrative. Pour faciliter le calcul de ces estimations synthétiques, nous procédons d'abord à l'imputation massive de la variable des gains hebdomadaires $G_{A,j}$ pour chaque établissement *j* de la source administrative comme suit. Nous divisons les établissements de l'ERE en groupes homogènes (appelés groupes modèles) afin de modéliser les gains hebdomadaires par employé en fonction de la paye mensuelle par employé. Le modèle de régression pour un groupe modèle donné *g* est le suivant : $(G/E)_{E,i(g)} = \beta_{1(g)} + \beta_{2(g)}(P/E)_{E,i(g)} + \varepsilon_{i(g)}$, où $(G/E)_{E,i} = \mathrm{G}_{E,i} / \mathrm{E}_{E,i}$ sont les gains hebdomadaires par employé et $(P/E)_{E,i} = \mathrm{P}_{E,i} / \mathrm{E}_{E,i}$ est la paye mensuelle par employé. Les paramètres $\beta_{1(g)}$ et $\beta_{2(g)}$ sont estimés selon la méthode des moindres carrés (Lohr, 1999) en utilisant les données de l'enquête. Ces paramètres sont ensuite appliqués à l'ensemble des établissements *j* de la source administrative (par groupe modèle) pour imputer massivement les gains hebdomadaires par employé $(\hat{G/E})_{A,j}$. De là, nous dérivons les gains hebdomadaires de l'établissement *j* comme étant $\hat{G}_{A,j} = (\hat{G/E})_{A,j} \times E_{A,j}$. Finalement, l'estimation des $G\hat{H}M$ pour le domaine *d* est obtenue à l'aide du quotient $G\hat{H}M_{\mathrm{A1}} = \sum_{j \in d \cap U} \hat{G}_{A,j} \Big/ \sum_{j \in d \cap U} E_{A,j}$.

Lorsque les groupes modèles sont très homogènes, cette méthode permet de diminuer considérablement l'erreur quadratique moyenne des estimations pour de très petits domaines, car la variance est grandement réduite et le biais petit. De plus, si le modèle est bien ajusté, nous aurons des coefficients de détermination R^2 très élevés. Cependant, ce modèle est particulièrement sensible aux différences entre les concepts d'emploi et de paye des deux sources.

3.3 Estimation à partir des deux sources avec appariement des microdonnées (méthode A2)

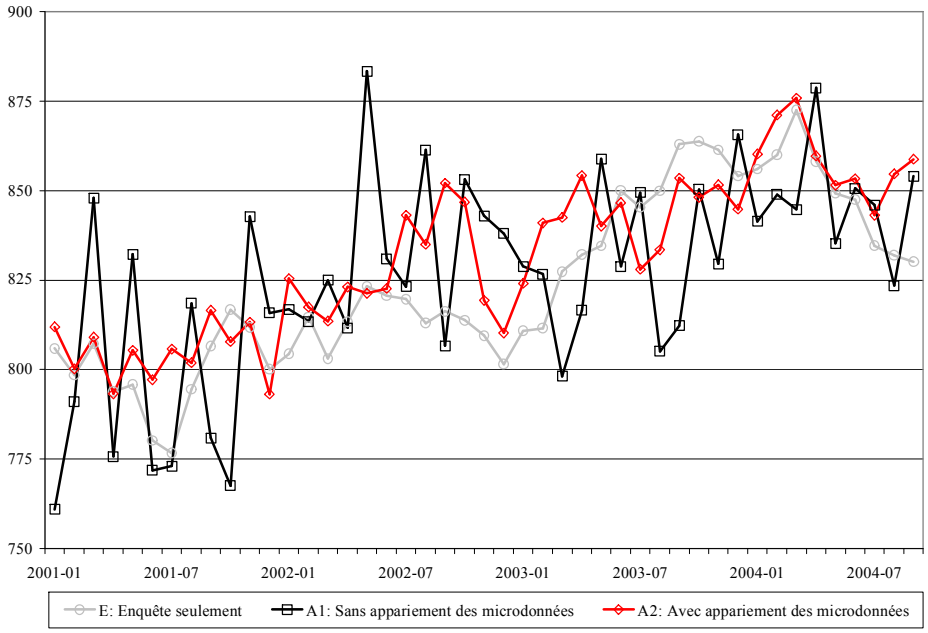
L'estimation des GHM peut aussi être faite en appariant les microdonnées de la source d'enquête à celles de la source administrative. Pour chaque unité *i* échantillonnée de l'ERE, nous estimons les paramètres du modèle de régression en utilisant le rapport des gains hebdomadaires par employé provenant de l'enquête $(G/E)_{E,i} = \mathrm{G}_{E,i} / \mathrm{E}_{E,i}$, ainsi que la paye mensuelle par employé provenant de la source administrative $(P/E)_{A,i} = \mathrm{P}_{A,i} / \mathrm{E}_{A,i}$. Le modèle de régression pour un groupe modèle donné *g* devient donc : $(G/E)_{E,i(g)} = \beta_{1(g)} + \beta_{2(g)}(P/E)_{A,i(g)} + \varepsilon_{i(g)}$. Les coefficients $\beta_{1(g)}$ et $\beta_{2(g)}$ sont encore ici estimés à l'aide de la méthode des moindres carrés et l'imputation massive se fait comme montré à la section 3.2. L'estimation finale des $G\hat{H}M$ pour le domaine *d* est obtenue à l'aide du quotient $G\hat{H}M_{\mathrm{A2}} = \sum_{j \in d \cap U} \hat{G}_{A,j} \Big/ \sum_{j \in d \cap U} E_{A,j}$ où $\hat{G}_{A,j} = (\hat{G/E})_{A,j} \times E_{A,j}$.

Ce modèle a comme avantage de ne plus être affecté par les problèmes liés à la différence de concepts entre les deux sources. En effet, la paye mensuelle par employé utilisée pour estimer les paramètres du modèle provient de la source administrative $(P/E)_{A,j}$ et lors de l'imputation massive, nous appliquons les paramètres estimés à cette même variable.

3.4 Résultats

Nous avons estimé à l'aide des données de l'EERH les GHM entre janvier 2001 et septembre 2004 à l'aide des trois méthodes décrites précédemment. Nous nous attardons ici à un domaine en particulier pour lequel nous avions dénoté un problème d'incohérence marquée entre les deux sources. Pour ce domaine étudié *d*, la taille d'échantillon de la partie enquête est d'environ 1300 unités alors que la source administrative compte plus de 60 000 unités. Nous n'avons utilisé qu'un seul groupe modèle ($g = d$). Lors de l'estimation des paramètres $\beta_{1(g)}$ et $\beta_{2(g)}$ sur les 45 mois, le modèle A1 présentait de meilleurs coefficients de détermination R^2 (entre 0,96 et 0,99) que le modèle A2 (entre 0,62 et 0,94) mais cette mesure est trompeuse puisque le problème avec le modèle A1 provient de la différence de concepts entre les variables des deux sources. Lorsque ces trois séries sont représentées graphiquement (Figure 1), le modèle A1 démontre une grande variabilité par rapport aux deux autres méthodes.

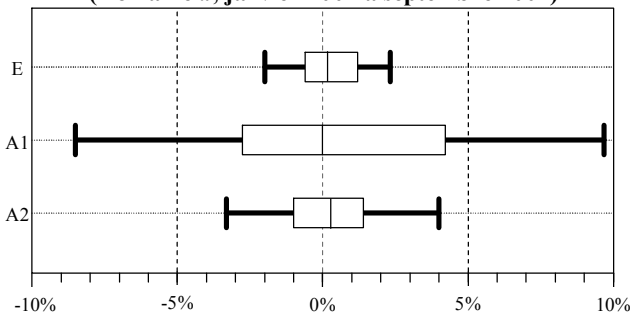
Figure 1
Estimation des gains hebdomadaires moyens en dollars selon trois méthodes
(Domaine *d*, janvier 2001 à septembre 2004)



Dans le cas de la méthode sans appariement des microdonnées (A1), les variations observées dans les GHM semblent être causées par les fortes variations de la paye mensuelle dans la source administrative. Le résultat est une courbe en dents de scie. Ce modèle est trop rigide dans le temps puisqu'il est basé sur la paye mensuelle par employé de l'enquête alors que les coefficients estimés sont appliqués à la paye mensuelle par employé de la source administrative, et cette variable a tendance à être beaucoup plus stable du côté enquête que du côté de la source administrative. Nous pouvons voir par exemple que les pointes élevées dans la série des gains hebdomadaires moyens avec la méthode A1 apparaissent pour les mois comptant cinq jeudis ou cinq vendredis. Nous avons vu ces mêmes pointes lorsque nous avons produit la série pour la paye mensuelle par employé de la source administrative.

Visuellement, les courbes E et A2 sont nettement plus stables que la courbe A1. Les variations mensuelles relatives (Figure 2) de A1 ont un écart-type de 4,35 points de pourcentage alors qu'il est de 1,53 pour A2 et de 1,08 pour E. Sous l'hypothèse d'une tendance linéaire sur les 45 mois, la série A2 serait meilleure que les séries E et A1 puisque son coefficient de détermination R^2 est de 0,76, par rapport à 0,69 et 0,33 pour les deux autres séries.

Figure 2
Distribution des variations mensuelles relatives selon les trois méthodes
(Domaine *d*, janvier 2001 à septembre 2004)



Finalement, l'analyse des modèles A1 et A2 a montré que les résidus sont généralement distribués normalement et exempts d'hétéroscédasticité.

CONCLUSION

Nous avons noté certains problèmes liés à la différence des concepts entre les sources de données de l'EERH. Certaines de ces incohérences seront amoindries en apportant des changements à l'ERE. Nous envisageons donc à l'avenir de collecter l'information au niveau de l'entreprise plutôt qu'au niveau des établissements. L'information sera ensuite désagrégée au niveau des établissements en utilisant le même processus de désagrégation qui est utilisé pour les données administratives. De plus, nous utiliserons à l'avenir une méthode pour réduire l'impact des valeurs influentes qui ramènera la distribution d'emploi de l'enquête à une distribution similaire à celle obtenue des données administratives.

Côté estimation, parmi les trois méthodes examinées pour estimer les gains hebdomadaires moyens, l'approche utilisant les deux sources avec appariement des microdonnées (méthode A2) donne des résultats beaucoup plus stables que les deux autres méthodes. Elle permet de mieux cerner la tendance dans le temps et, contrairement à la méthode basée sur les données d'enquête seulement, elle permet de produire des estimations plus précises pour des niveaux provinciaux par industrie. De plus, la variable de paye mensuelle par employé de la source administrative peut être vue comme un complément aux données d'enquête lors de la modélisation. Cette nouvelle méthode a également été appliquée à l'estimation des heures hebdomadaires moyennes payées et des paiements spéciaux. Les résultats allaient dans le même sens que ceux présentés dans ce document.

REMERCIEMENTS

Nous remercions les réviseurs ainsi que l'équipe des méthodologistes de l'EERH (Pierre Lavallée, Anne-Marie Houle, Richard Belcher, Édith Hovington et Sharon Wirth) pour l'aide apportée, leurs commentaires et leurs suggestions. Nous remercions également Don Royce, John Kovar et Éric Rancourt pour leurs suggestions au sujet des modèles.

RÉFÉRENCES

Grondin, C. et Lavallée, P. (2001), "Current Methodology of the Survey of Employment, Payrolls and Hours", Ottawa, Canada : Statistique Canada.

Kroese, A.H. et Renssen, R.H. (2000), "New Applications of Old Weighting Techniques : Constructing a Consistent Set of Estimates Based on Data from Different Sources", *Proceedings of the Second International Conference on Establishment Surveys : Survey Methods for Businesses, Farms, and Institutions*, Buffalo, New York : American Statistical Association, p. 831-840.

Lohr, S.L. (1999), *Sampling : Design and Analysis*, Pacific Groove: Duxbury Press.