

# ESTIMATION DE LA VARIANCE DANS LE CADRE D'ENQUÊTES COMPLEXES LIÉES À L'UTILISATION DE DONNÉES ADMINISTRATIVES

Daniel Hurtubise<sup>1</sup>

## RÉSUMÉ

Le fardeau de réponse des entreprises aux enquêtes statistiques économiques constitue une préoccupation importante dans le développement des enquêtes. Différentes approches sont prises afin de réduire ce fardeau de réponse, dont l'utilisation de données administratives. Cependant, les concepts derrière ces données ne correspondent pas toujours aux concepts des enquêtes. Une solution consiste à combiner les données administratives à des données d'enquêtes. L'Enquête sur l'Emploi, la Rémunération et les Heures (EERH) de Statistique Canada est un exemple d'une telle enquête. Elle combine les données de deux sources indépendantes à l'aide d'estimateurs synthétiques. L'estimation de la variance doit tenir compte de ces deux sources de données, de même que du traitement fait à ces données. Dans cet article, une formule générale d'estimation de même que l'estimateur de la variance sont proposés. Des exemples tirés de l'EERH sont utilisés pour illustrer cette approche.

**MOTS CLÉS :** Estimateurs synthétiques; fardeau de réponse; indépendance des sources de données; modèles de régression; ratio de variables.

## ABSTRACT

The response burden of enterprises to the economic statistical surveys constitutes a major preoccupation in the development of surveys. Different approaches are taken in order to reduce this response burden, of which the use of administrative data. However, concepts behinds these data don't always correspond to survey concepts. A solution consists in combining the administrative data to survey data. The Statistics Canada Survey on Employment, Payroll and Hours (SEPH) is an example of such a survey. It combines data from two independent sources using synthetic estimators. The variance estimation must take into account these two sources of data, as well as the processing done to these data. In this paper, a general estimation formula is suggested as well as the variance estimator. Examples from SEPH are also given using this approach.

**KEY WORDS :** Independence of data sources; Response burden; Regression models; Ratios of variables, Synthetic estimators.

## 1. INTRODUCTION

Depuis quelques années, les enquêtes auprès des entreprises se sont adaptées afin de garder le fardeau de réponse à un niveau acceptable. En effet, pour les différentes enquêtes économiques, la même base de sondage est souvent utilisée, ce qui cause un certain chevauchement entre les échantillons, et augmente ainsi le fardeau de réponse. Différentes approches ont été mises en place à Statistique Canada afin de réduire ce fardeau de réponse; entre autre, une enquête unifiée auprès des entreprises, une étude sur la coordination des échantillons (Rubin-Bleuer, 2002), ainsi que l'utilisation de données auxiliaires, plus particulièrement des données administratives. Ces dernières, combinées à des données d'enquête via des modèles statistiques, permettent de réduire la taille d'échantillon des enquêtes tout en gardant une bonne qualité des données. Parmi les données administratives disponibles, on retrouve les données fiscales. Ces dernières proviennent de l'Agence des Douanes et du Revenu du Canada (ADRC), qui collecte et gère ces données fiscales.

L'Enquête sur l'Emploi, la Rémunération et les Heures (EERH) de Statistique Canada est une enquête qui combine à la fois des données d'une enquête statistique et des données administratives. Les premières servent à estimer différents ratios de variables pour certains sous-ensembles prédéfinis de la population, et les secondes servent à estimer l'emploi total et la paie totale pour différents domaines d'estimation. L'EERH utilise des estimateurs synthétiques qui combinent ces deux

---

<sup>1</sup> Daniel Hurtubise (daniel.hurtubise@statcan.ca), méthodologiste, Division des méthodes d'enquêtes auprès des entreprises, Immeuble R.-H.-Coats, 11<sup>e</sup> étage, Ottawa, Ontario, Canada, K1A 0T6.

sources de données. La variance de ces estimateurs est calculée en tenant compte du traitement des données (imputation), des différentes sources de données et cette variance utilise la technique du jackknife et de la variance d'estimateurs post-stratifiés. À ce sujet, Hurtubise et coll. (2000) ont développé une première version de cette variance.

Dans cet article, le plan de sondage de l'EERH est décrit à la section 2. Les différentes sources de données sont étudiées à la section 3, alors que la section 4 présente la formule générale d'estimation. La formule générale de variance est présentée quant à elle à la section 5.

## **2. L'ENQUÊTE SUR L'EMPLOI, LA RÉMUNÉRATION ET LES HEURES**

L'EERH est une enquête mensuelle couvrant tous les employeurs dans à peu près toutes les industries au Canada. Cette enquête collecte et publie des informations sur l'emploi, la paie, les heures payées et ce, pour différentes catégories d'employés et différents domaines d'intérêt. Les objectifs principaux de l'EERH sont d'estimer mensuellement : i) le nombre total d'employés rémunérés; ii) la rémunération hebdomadaire moyenne; iii) la moyenne des heures payées hebdomadairement, ainsi que d'autres variables connexes.

À ses débuts, l'EERH était une enquête stratifiée d'établissements dont la taille d'échantillon était d'environ 70 000 établissements, enquêtés mensuellement. Les procédures d'estimation étaient basées sur le plan de sondage et sa pondération associée, les données étant recueillies directement auprès des répondants. Depuis le milieu des années 1990, l'EERH a été remaniée et est maintenant basée sur deux sources de données indépendantes : des données administratives provenant de l'ADRC et des données d'enquête provenant de l'Enquête sur la rémunération auprès des entreprises (ERE). L'utilisation des données administratives a permis de réduire la taille de l'échantillon à environ 11 000 établissements.

De plus amples informations sur l'EERH sont disponibles dans *Le guide d'utilisation des données de l'Enquête sur l'emploi, la rémunération et les heures* publié par Statistique Canada, et également dans Rancourt et Hidirogrou (1998).

## **3. SOURCES DES DONNÉES**

### **3.1. Données administratives**

Chaque entreprise opérant au Canada et ayant des employés doit remettre à l'ADRC toutes les retenues d'impôt à la source prélevées sur le salaire de ses employés. Elle le fait pour chacune de ses listes d'employés. Le résultat constitue un fichier de tous les comptes de Numéros d'Entreprise (NE) contenant les trois variables suivantes : la remise, nombre total d'employés (emploi) et les salaires totaux (paie). Un NE est un identificateur unique assigné par l'ADRC à chaque entité légale participant dans l'un ou l'autre des programmes fiscaux suivants : l'impôt sur les sociétés, la Taxe sur les Produits et Services (TPS), les retenues salariales, la taxe sur l'importation et l'exportation. Il y a deux types possibles de remises. Le premier type représente les remises mensuelles (lorsque la moyenne mensuelle des remises de l'année précédente était de moins de 15 000 \$). Le second type représente les remises accélérées (lorsque la moyenne mensuelle des remises de l'année précédente était d'au moins 15 000 \$). Il peut y avoir de 1 à 4 remises mensuelles, selon la fréquence de paie (mensuelle, hebdomadaire, toutes les deux semaines, etc.). Un recensement de tous les comptes NE qui sont dans le champ de l'enquête est réalisé, ce qui représente environ un million de comptes mensuellement.

Un processus de vérification et d'imputation est en place afin de vérifier la validité des valeurs et d'imputer les valeurs manquantes ou invalides, le cas échéant. L'imputation est basée sur les données historiques lorsque celles-ci sont disponibles, en utilisant une tendance. Un ratio est utilisé si au moins une des trois variables est disponible. En dernier lieu, l'imputation est réalisée en utilisant la moyenne de la classe d'imputation. Cette dernière est définie selon la géographie, l'industrie et la taille (nombre d'employés) des comptes NE.

### **3.2. L'Enquête sur la rémunération auprès des entreprises**

L'EERH s'intéresse particulièrement à deux des trois variables disponibles sur les fichiers administratifs : l'emploi et la paie. Cependant, cette enquête nécessite l'utilisation de plusieurs autres variables, non disponibles sur la source administrative, telles que le nombre d'heures payées, les gains totaux, les heures supplémentaires et ce, pour différentes catégories d'employés, telles que les employés payés à l'heure, les salariés. L'ERE est donc utilisée pour obtenir des données qui servent à modéliser les variables désirées, et imputer massivement les dossiers administratifs.

L'ERE est une enquête établissement utilisant un plan de sondage stratifié, dont la taille d'échantillon est d'environ 11 000 établissements sélectionnés parmi 900 000 établissements sur le Registre des Entreprises (RE). Le RE constitue la base de sondage interne principale pour les enquêtes économiques à Statistique Canada. Elle contient la liste des entreprises faisant affaire au Canada. Les établissements considérés dans le champ de l'enquête sont ceux ayant des employés, qui sont actives, qui sont sur le RE et qui opèrent dans une industrie couverte par le champ de l'enquête. Les établissements du RE sont répartis en groupes modèles homogènes et indépendants basé sur l'industrie (et parfois sur la géographie et la taille). Les groupes modèles forment une partition exhaustive du RE. Ils sont choisis de façon à obtenir le meilleur ajustement possible pour la modélisation tout en conservant une signification analytique. Les strates, basées sur la géographie et la taille, sont définies à l'intérieur de chaque groupe modèle. Les strates à tirage complet sont définies à l'aide de la méthode de l'écart-sigma (Bernier et Nobrega, 1998). Ces strates sont séparées en deux, selon la taille, à l'aide de cette même méthode. Les établissements ayant les plus grandes tailles sont exclus de la modélisation et forment un groupe modèle par eux-mêmes. Les établissements ayant les plus petites tailles de la strate à tirage complet sont conservés dans la modélisation du groupe modèle. Les établissements restant font partie des strates à tirage partiel. Ces strates sont formées selon la géographie et la taille des établissements. L'échantillon est choisi selon un échantillonnage colloqué à l'intérieur de chaque strate. Chaque établissement demeure dans l'échantillon 12 mois en moyenne, et un douzième de l'échantillon est remplacé à chaque mois.

#### 4. ESTIMATION

De façon générale, l'estimateur d'une statistique  $Z$  pour un certain domaine ( $d$ ) est :

$$\hat{Z}_{(d)} = \frac{\sum_g \hat{Y}_{g(d)}}{\sum_g \hat{X}_{g(d)}} = \frac{\hat{Y}_{(d)}}{\hat{X}_{(d)}} \quad (1)$$

où

$$\hat{Y}_{g(d)} = \hat{\alpha}_{YE,ERE,g} \times E_{Y,ADM,g(d)} + \hat{\alpha}_{YP,ERE,g} \times P_{Y,ADM,g(d)} \quad (2)$$

$$\hat{X}_{g(d)} = \hat{\alpha}_{XE,ERE,g} \times E_{X,ADM,g(d)} + \hat{\alpha}_{XP,ERE,g} \times P_{X,ADM,g(d)} \quad (3)$$

La base de cette formule correspond à des modèles de régression ayant Emploi et Paie comme variables explicatives. Les paramètres  $E_X$ ,  $E_Y$ ,  $P_X$ , et  $P_Y$  sont donc fonction des variables Emploi et Paie que l'on retrouve sur le fichier administratif, et les paramètres  $\hat{\alpha}$  sont les paramètres du modèle, soient des fonctions des paramètres de régression et/ou des ratios de variables obtenus de l'ERE. On peut donc voir  $\hat{Z}_{(d)}$  comme un ratio de variables aléatoires dépendantes ou encore un ratio d'estimateurs synthétiques dépendants. Ces estimateurs sont synthétiques car ils combinent des informations d'un certain domaine d'intérêt ( $d$ ) avec des informations provenant de groupes modèles  $g$ , ces derniers pouvant couvrir plus qu'un domaine d'intérêt.

De façon particulière, on peut considérer les deux variables Heures payées (HRS) et Gains totaux (GAINS), qui sont dérivées par des modèles de régression classiques. Ces deux modèles considèrent l'Emploi et la Paie comme variables explicatives. Dans ce cas, les formules (2) et (3) s'écrivent de la façon suivante :

$$\hat{HRS} = \hat{\beta}_{HE,ERE,g} \times EMPLOI_{ADM,g(d)} + \hat{\beta}_{HP,ERE,g} \times PAIE_{ADM,g(d)}$$

$$\hat{GAINS} = \hat{\beta}_{GE,ERE,g} \times EMPLOI_{ADM,g(d)} + \hat{\beta}_{GP,ERE,g} \times PAIE_{ADM,g(d)}$$

Les paramètres  $\hat{\beta}$  peuvent être utilisés par la suite dans les formules (2) ou (3) selon les différentes définitions des variables.

Exemple : la définition intuitive pour le nombre de salariés rémunérés à l'heure est donnée par

$$H\_EMPLOI_{(d)} = EMPLOI_{ADM,g(d)} \times \frac{H\_EMPLOI_{ERE,g}}{EMPLOI_{ERE,g}} \quad (4)$$

où  $H_{EMPLOI_{ERE,g}}$  correspond au nombre d'employés à salaire horaire dans le groupe modèle  $g$  obtenu de l'ERE,  $EMPLOI_{ERE,g}$  correspond au nombre total d'employés dans le même groupe modèle et obtenu de l'ERE, et  $EMPLOI_{ADM,g(d)}$  correspond au nombre total d'employés du groupe modèle  $g$  pour le domaine  $(d)$  obtenu de la source administrative. Sous la forme des équations (2) et (3), les différents paramètres prennent les valeurs suivantes :

$$\hat{\alpha}_{YE,ERE,g} = \frac{H_{EMPLOI_{ERE,g}}}{EMPLOI_{ERE,g}} = \hat{R}_{1,ERE,g}, E_{Y,ADM,g(d)} = EMPLOI_{ADM,g(d)}, \hat{\alpha}_{YP,ERE,g} = 0, P_{Y,ADM,g(d)} = 0, \quad (5)$$

$$\hat{\alpha}_{XE,ERE,g} = 1, E_{X,ADM,g(d)} = 1, \hat{\alpha}_{XP,ERE,g} = 0, P_{X,ADM,g(d)} = 0.$$

Dans le cas où un domaine couvre plusieurs groupes modèles et où le paramètre  $E_{X,ADM,g(d)}$  égale 1, les paramètres  $\hat{\alpha}_{XP,ERE,g}$  et  $P_{X,ADM,g(d)}$  sont égaux à 0, le paramètre  $\hat{\alpha}_{XE,ERE,g}$  doit être égal à  $1/g$ , où  $g$  est le nombre de groupes modèles dans le domaine. Il est à noter que ce nombre  $g$  est connu et fixe pour tous les domaines. De plus, une somme sur tous les groupes modèles doit alors être ajoutée à la formule (4).

## 5. ESTIMATION DE LA VARIANCE

### 5.1. Caractéristiques du plan de sondage

Les données utilisées dans l'EERH proviennent de deux sources différentes : un recensement des données administratives, dont l'unité de base est le comptes NE ainsi que l'ERE, dont l'unité d'échantillonnage est l'établissement. Ces deux concepts différents nous forcent à considérer les deux sources de données comme indépendantes. De plus, la variabilité due à l'échantillonnage est nulle dans le cas administratif. Seule la variabilité due à l'échantillonnage de la source de l'enquête est à considérer. L'indépendance des deux sources permet de conclure que la covariance entre les données administratives et les données de l'enquête est nulle.

Du côté administratif, les données manquantes sont imputées (ce qui est appelé ci-après l'imputation classique). Les méthodes employées sont décrites dans la section 3.1. Du côté de l'ERE, les observations ayant des valeurs manquantes ne sont pas conservées dans le processus d'estimation des différents ratios.

Une dernière caractéristique à considérer est l'imputation massive des différentes variables du fichier administratif. Cette imputation est faite en utilisant différents ratios et paramètres de régression provenant de la source d'enquête. La variabilité de ces paramètres doit être considérée dans le calcul de la variance totale.

En résumé, la variabilité totale dépend de : l'échantillonnage de l'ERE, l'imputation classique des variables administratives et l'estimation des différents paramètres et ratios provenant de l'enquête, utilisés pour l'imputation massive.

### 5.2. Formule générale de la variance

L'estimateur de la variance de  $\hat{Z}_{(d)}$  est obtenu par la formule générale d'un ratio de deux variables aléatoires dépendantes, appliquée sur des groupes modèles, soit

$$\hat{V}(\hat{Z}_{(d)}) = \frac{1}{\hat{X}_{(d)}^2} \left[ \sum_g \hat{V}(\hat{Y}_{g(d)}) + \hat{Z}_{(d)}^2 \sum_g \hat{V}(\hat{X}_{g(d)}) - 2\hat{Z}_{(d)} \sum_g \text{cov}(\hat{Y}_{g(d)}, \hat{X}_{g(d)}) \right]. \quad (6)$$

Le terme général de la covariance entre  $\hat{Y}_{g(d)}$  et  $\hat{X}_{g(d)}$  est donné par

$$\begin{aligned}
\text{cov}(\hat{Y}_{g(d)}, \hat{X}_{g(d)}) &= E_{Y,ADM,g(d)} E_{X,ADM,g(d)} \text{cov}(\hat{\alpha}_{YE,ERE,g}, \hat{\alpha}_{XE,ERE,g}) + \hat{\alpha}_{YE,ERE,g} \hat{\alpha}_{XE,ERE,g} \text{cov}(E_{Y,ADM,g(d)}, E_{X,ADM,g(d)}) \\
&\quad - \text{cov}(\hat{\alpha}_{YE,ERE,g}, \hat{\alpha}_{XE,ERE,g}) \text{cov}(E_{Y,ADM,g(d)}, E_{X,ADM,g(d)}) \\
&\quad + E_{Y,ADM,g(d)} P_{X,ADM,g(d)} \text{cov}(\hat{\alpha}_{YE,ERE,g}, \hat{\alpha}_{XP,ERE,g}) + \hat{\alpha}_{YE,ERE,g} \hat{\alpha}_{XP,ERE,g} \text{cov}(E_{Y,ADM,g(d)}, P_{X,ADM,g(d)}) \\
&\quad - \text{cov}(\hat{\alpha}_{YE,ERE,g}, \hat{\alpha}_{XP,ERE,g}) \text{cov}(E_{Y,ADM,g(d)}, P_{X,ADM,g(d)}) \\
&\quad + P_{Y,ADM,g(d)} E_{X,ADM,g(d)} \text{cov}(\hat{\alpha}_{YP,ERE,g}, \hat{\alpha}_{XE,ERE,g}) + \hat{\alpha}_{YP,ERE,g} \hat{\alpha}_{XE,ERE,g} \text{cov}(P_{Y,ADM,g(d)}, E_{X,ADM,g(d)}) \\
&\quad - \text{cov}(\hat{\alpha}_{YP,ERE,g}, \hat{\alpha}_{XE,ERE,g}) \text{cov}(P_{Y,ADM,g(d)}, E_{X,ADM,g(d)}) \\
&\quad + P_{Y,ADM,g(d)} P_{X,ADM,g(d)} \text{cov}(\hat{\alpha}_{YP,ERE,g}, \hat{\alpha}_{XP,ERE,g}) + \hat{\alpha}_{YP,ERE,g} \hat{\alpha}_{XP,ERE,g} \text{cov}(P_{Y,ADM,g(d)}, P_{X,ADM,g(d)}) \\
&\quad - \text{cov}(\hat{\alpha}_{YP,ERE,g}, \hat{\alpha}_{XP,ERE,g}) \text{cov}(P_{Y,ADM,g(d)}, P_{X,ADM,g(d)})
\end{aligned} \tag{7}$$

Pour obtenir la variance de  $\hat{Y}_{g(d)}$ , il suffit de prendre la covariance entre  $\hat{Y}_{g(d)}$  et  $\hat{Y}_{g(d)}$ ; le même principe s'applique pour  $\hat{X}_{g(d)}$ . Cette formule est basée sur l'article de Goodman (1960).

### 5.3. Variabilité due à la source administrative

La seule source de variabilité provenant du fichier administratif est l'imputation classique due aux valeurs manquantes de certaines observations. Selon l'article de Felx et Rancourt (2000), la variance totale est exprimée selon l'expression suivante :

$$V_{\text{TOT}} = V_{\text{SAM}} + V_{\text{IMP}} + 2V_{\text{MIX}}. \tag{8}$$

Le premier terme réfère à la variabilité due à l'échantillonnage, le second à l'imputation classique et le dernier terme réfère à un effet combiné de l'échantillonnage et de l'imputation. Dans le cas présent, un recensement des données administratives est effectué. La variance totale correspond donc seulement à la variance due à l'imputation. Felx et Rancourt (2000) définissent les différentes formules à utiliser lorsque différentes méthodes d'imputation sont utilisées, ce qui s'applique ici.

### 5.4. Variabilité due à la source d'enquête

La variabilité due à la source de l'ERE est la variabilité des différents ratios et paramètres de régression qui entrent dans les estimateurs synthétiques. La méthode du « jackknife » est utilisée à l'intérieur de chaque groupe modèle pour calculer la variance de même que la covariance entre les différents ratios. Soit  $B_g$  une fonction de ratios et/ou de paramètres de

régression. La variance de  $\hat{B}_g$  est donnée par  $\hat{V}_J(\hat{B}_g) = \sum_{h=1}^{H_g} \frac{(L_{gh} - 1)}{L_{gh}} \sum_{l=1}^{L_{gh}} (\hat{B}_{g(hl)} - \hat{B}_g)^2$ , où  $H_g$  est le nombre de strates

dans le groupe modèle  $g$ ;  $L_{gh}$  est le nombre de répliques dans la strate  $h$  ( $h=1, 2, \dots, H_g$ ) pour le groupe modèle  $g$  (dans notre cas,  $L_{gh}=2$ );  $\hat{B}_{g(hl)}$  est l'estimation de  $B_g$  lorsque la réplique  $l$  de la strate  $h$  est enlevée des données et  $\hat{B}_g$  est l'estimation de  $B_g$  avec l'échantillon complet. L'indice  $J$  réfère à l'utilisation de la méthode du « jackknife ».

### 5.5. Exemple

Pour l'exemple du nombre de salariés rémunérés à l'heure, dont l'estimateur est donné par l'équation (4) dans la section 4, et les différents paramètres par l'équation (5), la formule de variance est

$$\begin{aligned}
\hat{V}(H\_EMPLOI_{(d)}) &= \sum_g \left[ \hat{V}_{IMP}(EMPLOI_{ADM,g(d)}) \times \hat{R}_{1,ERE,g}^2 + \hat{V}_J(\hat{R}_{1,ERE,g}) \times EMPLOI_{ADM,g(d)}^2 \right. \\
&\quad \left. - \hat{V}_J(\hat{R}_{1,ERE,g}) \times \hat{V}_{IMP}(EMPLOI_{ADM,g(d)}) \right]
\end{aligned} \tag{9}$$

## 6. CONCLUSION

Les estimateurs synthétiques présentés dans cet article combinent les résultats de deux sources indépendantes de données. L'hypothèse sous-jacente est que les coefficients de régression et les ratios calculés pour un groupe modèle sont valides pour chaque domaine d'intérêt faisant partie de celui-ci.

Bien que la formule semble complexe, il suffit de décomposer les estimateurs désirés en terme de variables administratives disponibles et de fonctions de variables d'enquête.

## REMERCIEMENTS

L'auteur aimerait remercier Chantal Grondin et Martin Provost pour leurs commentaires et suggestions. Également, des remerciements vont à Pierre Lavallée, Yves Morin et Michel Hidirolou pour leur participation dans l'élaboration initiale de cet article.

## RÉFÉRENCES

- Bernier, J. et K. Nobrega (1998), Outlier detection in asymmetric samples: A comparison of an inter-quartile range method and a variation of a sigma gap method. *Recueil de la section des méthodes d'enquête*, Congrès annuel de la Société Statistique du Canada 1998, pp. 137-141.
- Felx, P. et E. Rancourt (2000), Applications of Variance Due to Imputation in the Survey of Employment, Payrolls and Hours. Article non publié, Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada
- Goodman, L. A. (1960), On the exact variance of products. *Journal of the American Statistical Association*, **55**, pp. 708-713
- Hurtubise D., Y. Morin, P. Lavallée et M.A. Hidirolou (2000), Variance estimation for synthetic estimators in the context of an establishment survey. International Conference on Establishment Survey II, Buffalo, Juin 2000
- Rancourt, E. et M. A. Hidirolou (1998), Use of administrative records in the Canadian Survey of Employment, Payrolls and Hours. *Recueil de la section des méthodes d'enquête*, Congrès annuel de la Société Statistique du Canada 1998, pp. 39-47
- Rubin-Bleuer, S. (2002), Report on Rivière's Random Permutations Method of Sampling Co-ordination. Document interne, Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada.
- Statistique Canada, *Le guide d'utilisation des données de l'Enquête sur l'emploi, la rémunération et les heures*, numéro au Catalogue 72-620-GIF.