Composante du produit nº 11-522-X au catalogue de Statistique Canada
La série des symposiums internationaux de Statistique Canada : recueil

Article

Symposium 2008 :
Collecte des données : défis, réalisations et nouvelles orientations

Les défis de l'utilisation de données administratives dans l'Enquête sur l'emploi, la rémunération et les heures de travail

par Anthony Yeung, Anne-Marie Houle et Sharon Wirth

2009

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2008 de Statistique Canada
Collecte des données : défis, réalisations et nouvelles orientations

Les défis de l'utilisation de données administratives dans l'Enquête sur l'emploi, la rémunération et les heures de travail

Anthony Yeung, Anne-Marie Houle et Sharon Wirth[1]

Résumé

L'Enquête sur l'emploi, la rémunération et les heures de travail (EERH) est une enquête mensuelle qui utilise deux sources de données : un recensement des formulaires de retenues sur la paye (PD7) (données administratives) et une enquête auprès des établissements. Le présent document est axé sur le traitement des données administratives, de la réception hebdomadaire des données de l'Agence du revenu du Canada à la production d'estimations mensuelles par les responsables de l'EERH.

Les méthodes de contrôle et d'imputation utilisées pour traiter les données administratives ont été révisées au cours des dernières années. Les objectifs de ce remaniement étaient principalement d'améliorer la qualité des données et l'uniformité avec une autre source de données administratives (T4), qui constitue une mesure repère pour les responsables du Système de comptabilité nationale de Statistique Canada. On visait en outre à s'assurer que le nouveau processus serait plus facile à comprendre et à modifier, au besoin. Par conséquent, un nouveau module de traitement a été élaboré pour contrôler et imputer les formulaires PD7, avant l'agrégation des données au niveau mensuel.

Le présent document comporte un aperçu des processus actuel et nouveau, y compris une description des défis auxquels nous avons fait face pendant l'élaboration. L'amélioration de la qualité est démontrée à la fois au niveau conceptuel (grâce à des exemples de formulaires PD7 et à leur traitement au moyen de l'ancien et du nouveau systèmes) et quantitativement (en comparaison avec les données T4).

Mots clés : Données administratives, contrôle, imputation.

1. Introduction

L'Enquête sur l'emploi, la rémunération et les heures de travail (EERH) est conçue pour produire des estimations mensuelles des niveaux et des tendances d'un mois à l'autre de l'emploi, de la rémunération, des heures rémunérées et des gains, au niveau des industries, pour le Canada, les provinces et les territoires. Les industries dont il est question sont tirées du Système de classification des industries de l'Amérique du Nord (SCIAN). La population cible est composée de tous les employeurs au Canada qui effectuent des retenues sur les salaires et traitements des employés aux fins de l'impôt sur le revenu, des cotisations au Régime de pensions du Canada et des cotisations d'assurance-emploi. Tous les secteurs industriels sont inclus, sauf l'agriculture, la pêche et le piégeage, les services aux ménages privés, les organismes religieux et le personnel militaire des services de défense.

Les données de l'EERH sont recueillies à partir de deux sources différentes : un recensement de données administratives pour les variables de l'emploi et de la rémunération totale et les données obtenues dans une enquête mensuelle. Certaines variables sont tirées directement de sources administratives, tandis que d'autres sont estimées à partir des deux sources.

Comme c'est le cas pour toutes les enquêtes, il faut procéder à une validation et une imputation pour s'assurer que les données finales utilisées pour le calcul des estimations sont complètes, cohérentes et valides. Cela est

[1]Anthony Yeung, Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa, K1A 0T6, anthony.yeung@statcan.gc.ca; Anne-Marie Houle, Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa, K1A 0T6, anne-marie.houle@statcan.gc.ca ; Sharon Wirth, Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa, K1A 0T6, sharon.wirth@statcan.gc.ca

particulièrement pertinent pour la source de données administratives utilisées par l'EERH, parce que ces données sont recueillies à des fins non statistiques. Les processus actuels ont été analysés et une nouvelle approche a été proposée pour déterminer et corriger les erreurs des formulaires PD7 avant l'agrégation des données en valeurs mensuelles.

Le présent document est axé sur la réception et le traitement des données administratives (PD7). La section 2 comprend des descriptions des données recueillies au moyen des formulaires PD7, du traitement actuel des données et des raisons qui sous-tendent la décision de modifier le processus. La section 3 est axée sur les défis de l'élaboration d'un nouveau système de traitement de contrôle. La section 4 comprend une description détaillée du nouveau processus et certains résultats. Enfin, un sommaire des réalisations et des prochaines étapes de l'élaboration figure à la section 5.

2. Traitement des données administratives

2.1 Sources de données de l'EERH

La première source de données, c'est-à-dire les formulaires de retenues sur la paye (PD7), représente un recensement des données administratives reçues de l'Agence du revenu du Canada (ARC). Deux variables d'intérêt figurent dans ces formulaires : le nombre d'employés rémunérés au cours de la dernière période de paie du mois (E_F), et la rémunération brute versée aux employés au cours de la période de versement (P_F). Une troisième variable, à savoir le montant versé à l'ARC par l'employeur aux fins de l'impôt sur le revenu, du Régime de pensions du Canada et de l'assurance-emploi, à l'égard de la rémunération totale des employés (R_F), est aussi disponible à partir des formulaires PD7 et est fortement corrélée avec P_F.

L'autre source de données est une enquête mensuelle auprès des établissements, l'Enquête sur la rémunération auprès des entreprises (ERE). Un échantillon d'environ 11 000 établissements sélectionnés à partir du Registre des entreprises de Statistique Canada sert à recueillir des données sur l'emploi, la rémunération, le total des heures régulières et des heures supplémentaires, les paiements spéciaux, ainsi qu'une ventilation de ces variables selon les différentes catégories d'employés (rémunérés sur une base horaire, salariés et autres).

Les estimations finales de l'EERH utilisent une combinaison des deux sources de données. Les estimations de la rémunération mensuelle brute et du nombre d'employés sont obtenues directement à partir des données administratives. Des modèles de régression sont utilisés pour prédire les heures et les gains hebdomadaires, à partir de l'échantillon des répondants de l'ERE. Les coefficients de régression estimés qui en résultent sont par la suite appliqués à chaque enregistrement de la source de données administratives, afin de prédire les heures et les gains hebdomadaires. D'autres variables, comme le nombre d'employés rémunérés sur une base horaire, sont obtenues en multipliant l'emploi, les heures ou les gains hebdomadaires par un ratio estimé à partir de l'échantillon de l'ERE. De plus amples détails concernant la méthodologie de l'EERH sont disponibles dans Rancourt et Hidiroglou (1998).

2.2 Description de la source de données administratives

Comme il est mentionné dans la section 2.1, trois variables sont obtenues à partir des formulaires PD7 : le nombre d'employés rémunérés de la dernière période de paie (E_F), le montant versé aux employés pour la période de versement (P_F) et le montant versé à l'ARC par l'employeur (R_F). Un ensemble de règles d'agrégation complexes sert à obtenir les valeurs mensuelles requises pour l'EERH. La complexité est le résultat des divers modèles de remise.

Chaque formulaire PD7 correspond à un seul versement et est lié à une seule entreprise. Toutefois, une entreprise peut soumettre des formulaires PD7 distincts pour différents groupes d'employés (et pour des périodes de remise de durées différentes), en fonction de ses pratiques et de ses antécédents au chapitre de la rémunération. Selon les règles de l'ARC, les employeurs qui ont toujours versé des remises annuelles importantes doivent effectuer des versements au moins deux fois par mois. Les employeurs de taille moyenne doivent effectuer des versements une fois par mois, tandis que ceux qui ont les remises les plus faibles peuvent les effectuer sur une base trimestrielle.

2.3 Traitement actuel des données administratives après agrégation

Dans le système actuel de traitement de la production, seules quelques validations simples de données sont effectuées au niveau du formulaire, comme la détermination des formulaires comportant un ratio inhabituel de remise sur rémunération. La majeure partie du traitement des données administratives est effectué après l'agrégation au niveau mensuel. Le système actuel de traitement comprend des corrections automatiques, une détection des valeurs aberrantes, des corrections manuelles et l'imputation.

Parmi les exemples de règles de contrôle appliquées au niveau mensuel figure la détermination des ratios P_A/E_A non plausibles[2], selon des limites propres à chaque industrie, et la détermination des tendances extrêmes de P_A ou de E_A, d'un mois à l'autre.

On procède aussi à la détection des valeurs aberrantes à l'intérieur de chaque strate (les strates sont fondées sur l'industrie, les groupes de provinces et le nombre d'employés). La méthode des quartiles est utilisée pour identifier les valeurs importantes au chapitre du nombre d'employés et de la rémunération totale; les enregistrements ainsi identifiés sont exclus du calcul des tendances et des ratios utilisés pour l'imputation. La méthode d'Hidiroglou-Berthelot (1986) sert à déterminer les valeurs aberrantes dans les tendances et les ratios. Si des enregistrements sont identifiés comme comportant des valeurs aberrantes pour le ratio P_A/E_A, les valeurs P_A et E_A sont par la suite considérées comme manquantes et étiquetées pour l'imputation; les autres valeurs aberrantes de ratios et de tendances sont seulement étiquetées en vue d'être exclues. Enfin, les entreprises comportant les différences les plus importantes quant au niveau déclaré de rémunération et/ou d'emploi entre le mois courant et le mois précédent sont identifiées à l'intérieur de chaque industrie. Les analystes valident ces enregistrements et les corrigent manuellement au besoin.

On procède à l'imputation pour E_A, P_A ou les deux. Quelques méthodes sont utilisées selon les renseignements disponibles. Si l'unité semble active au cours du mois courant et si les données du mois précédent sont de bonne qualité, on a recours à l'imputation par la tendance. Si on ne dispose pas de données utilisables pour le mois précédent, mais que d'autres variables pour le mois courant sont disponibles (P_A si E_A est imputé ou vice versa), on a recours à l'imputation par le ratio. En dernier ressort, on calcule des moyennes de strate pour le mois courant et on les utilise pour l'imputation par la moyenne.

2.4 Raisons de la modification du traitement des données administratives

Les responsables du Système de comptabilité nationale (SCN) ont indiqué des incohérences entre les taux de croissance annuels de la rémunération totale obtenus à partir des données T4 et des données annuelles agrégées PD7. Les données T4 comprennent le revenu annuel total versé aux employés par chaque employeur. Afin d'examiner cette situation, on a produit le niveau et les tendances de rémunération, à partir des données brutes reçues de l'ARC et à partir des données après le traitement. Cette comparaison a fait ressortir une réduction du niveau de rémunération et une augmentation plus faible d'un mois à l'autre, après le traitement. Une évaluation du traitement des données administratives de l'EERH a par conséquent été entreprise, afin de déterminer la ou les étapes de traitement qui ont donné lieu à ces réductions.

Cette évaluation a amené les analystes de l'enquête à suggérer un certain contrôle au niveau des formulaires, avant l'agrégation au niveau mensuel. En fait, il existe une corrélation étroite entre la rémunération brute versée aux employés (P_F) et les remises (R_F), ainsi qu'un rapport connu entre ces variables en raison des dispositions législatives dans le domaine de l'impôt sur le revenu. Par ailleurs, certaines erreurs de déclaration évidentes sont plus facilement (ou seulement) identifiables et explicables au niveau du formulaire.

Un autre objectif était d'uniformiser et d'automatiser le processus de contrôle et d'imputation dans la plus large mesure possible, étant donné le volume important de données à traiter chaque mois. Le processus existant comprend

[2] La notation de variable utilisée est similaire à celle décrite dans la section 2.1. L'indice inférieur A correspond aux valeurs une fois les règles d'agrégation appliquées, tandis que l'indice inférieur F correspond aux valeurs au niveau du formulaire.

de nombreuses étapes, est très complexe et n'est pas facile à modifier. La modification du système a fourni l'occasion d'apporter des améliorations dans ces domaines.

Le traitement des données administratives a par conséquent été revu complètement. Cela a eu pour résultat un nouveau module de contrôle et d'imputation au niveau du formulaire. Ce nouveau module sert principalement au contrôle et à l'imputation de la variable de la rémunération. Comme il est expliqué dans la section 4, la majeure partie du traitement de la variable de l'emploi se fait au niveau agrégé.

3. Défis de l'élaboration du module de contrôle et d'imputation au niveau du formulaire

Plusieurs défis se sont posés au moment de l'élaboration du nouveau système de contrôle et d'imputation (C et I) au niveau du formulaire. Il s'agit notamment des modèles de déclaration différents, des divers types d'erreurs qui peuvent se produire pendant la collecte des données, ainsi que de l'utilisation du revenu annuel des T4 comme référence pour évaluer le nouveau processus.

3.1 Détermination des modèles de déclaration

Afin d'élaborer un processus de C et I au niveau du formulaire et de déterminer les enregistrements dont les données comportent des irrégularités, il a été nécessaire de comprendre le rapport entre les variables des formulaires PD7 et d'autres sources administratives. Du fait du million de formulaires PD7 environ reçus chaque mois de l'ARC, la détermination des nombreux modèles de déclaration et d'erreurs a représenté une tâche énorme. Étant donné qu'il n'existe pas de norme quant au nombre de formulaires et à la couverture de chaque formulaire pour le versement de sommes à l'ARC (mais seulement sur la fréquence de remise), il existe de nombreuses façons possibles de verser la remise. Par exemple, un employeur qui doit effectuer des versements mensuels peut choisir d'envoyer les formulaires PD7 à l'ARC chaque fois que ses employés sont payés (deux fois par mois, toutes les deux semaines, toutes les semaines, etc.). En outre, si l'employeur compte des catégories différentes d'employés, comme des employés à temps plein et à temps partiel ou des employés salariés et rémunérés sur une base horaire, il peut remplir des formulaires distincts pour chaque catégorie. Selon les règles de l'impôt sur le revenu, on peut s'attendre à ce que le taux de remise pour certaines catégories d'employés soit beaucoup plus faible que pour les autres catégories dans la même entreprise.

3.2 Types d'erreurs dans le processus de collecte des données

Étant donné que les données recueillies représentent un recensement des formulaires PD7, les problèmes déterminés dans chaque formulaire sont attribuables à des erreurs non dues à l'échantillonnage de différentes sources. Les types suivants d'erreurs ont été identifiés au cours de l'élaboration du processus de C et I au niveau du formulaire.

a. Erreurs de transcription

Ces types d'erreurs se produisent lorsque les employeurs remplissent les formulaires PD7. Il se peut que des employeurs entrent les données de façon incorrecte (p. ex. qu'ils intervertissent les champs de paie et de remise, qu'ils incluent les cents dans la paie et/ou les remises, qu'ils intervertissent les mois et les jours, etc.). Ces types d'erreurs peuvent être déterminés grâce au système de contrôle proposé au niveau du formulaire.

b. Interprétation conceptuelle erronée des variables du formulaire PD7

Les lignes directrices pour remplir le formulaire PD7 ne figurent pas clairement dans le formulaire. Pour ce qui est de la valeur de la paie, les employeurs doivent inclure toute la rémunération qu'ils versent aux employés, y compris les avantages et les allocations imposables, de même que les paiements spéciaux, mais ces paiements sont parfois exclus des formulaires PD7. Ces types d'erreurs ne peuvent être décelés par le système proposé de contrôle au niveau du formulaire, celui-ci étant conçu pour déterminer les erreurs évidentes. Ces problèmes peuvent être décelés uniquement au moyen d'autres sources administratives, comme les données T4. En outre, parmi les employeurs qui versent leurs remises sur une base trimestrielle, certains peuvent inclure

le montant complet versé aux employés au cours de l'ensemble de la période de trois mois, même si le montant demandé touche la paie du dernier mois.

c. Erreurs de saisie de données à l'ARC

Tous les formulaires envoyés à l'ARC sont saisis là et des erreurs peuvent se produire pendant la saisie des données. À l'occasion, des erreurs se produisent dans les remises, mais une vérification comptable interne est effectuée par l'ARC, afin de comparer les remises déclarées dans le formulaire et les montants réels en dollars versés par les employeurs. Toutefois, aucune validation n'est faite pour l'emploi et la rémunération à l'ARC. Par conséquent, les erreurs de remises sont plus susceptibles d'être déterminées et corrigées par l'ARC que les erreurs dans les champs de l'emploi et de la paie des formulaires PD7. Dans le cas de ces deux champs, certaines erreurs évidentes peuvent être déterminées grâce au système proposé de contrôle au niveau du formulaire.

d. Erreurs introduites par des intermédiaires

Certains employeurs recrutent des compagnies de services de paie ou utilisent un logiciel financier pour remplir les formulaires de remise. On a découvert qu'à certaines occasions, des erreurs systématiques de données se produisent en raison d'erreurs de programmation ou de transcription. Certaines de ces erreurs sont décelées grâce au processus actuel d'assurance de la qualité et des mesures sont prises pour les corriger.

e. Erreurs de transmission des données (transmission en double pour des raisons inconnues)

Des formulaires en double sont parfois transmis de façon erronée à Statistique Canada. Le système proposé de contrôle peut les repérer.

f. Toutes les variables ne sont pas déclarées

Même si le formulaire PD7 sert à recueillir des données sur la paie et le nombre d'employés, un certain nombre d'employeurs indiquent uniquement les montants de remise sur ce formulaire. À cause de cela, une méthode d'imputation doit être appliquée pour produire des estimations des variables qui manquent dans le formulaire. Avec le système proposé de C et I au niveau du formulaire, la paie brute est imputée pour chaque formulaire, et le nombre d'employés est imputé au niveau mensuel agrégé.

3.3 Utilisation des données T4 comme référence

L'augmentation de la cohérence entre les données annuelles agrégées PD7 et les données T4 constitue l'un des principaux objectifs du remaniement du traitement des données administratives. En outre, les données T4 sont la seule source de comparaison qui peut servir à évaluer les avantages de l'utilisation d'un module de C et I au niveau du formulaire. Même s'il est largement reconnu que ces deux sources permettent une comparaison exacte ou très proche du revenu annuel, cela n'est pas entièrement correct.

Une étude effectuée pour mesurer l'écart attendu entre les deux sources a montré que la paie annuelle selon les formulaires PD7 est de 1,4 % à 1,8 % plus faible que la paie annuelle selon les données T4 pour les années 2003 à 2005. L'étude a été fondée sur des enregistrements qui n'ont pas été modifiés du tout par le traitement des données administratives. Autrement dit, ces enregistrements comportaient toutes les données nécessaires et répondaient à l'ensemble des hypothèses de données au niveau du formulaire.

Cette étude a aussi aidé à expliquer les différences entre les deux sources. Tout d'abord, la paie annuelle agrégée selon les formulaires PD7 peut ne pas inclure les données pour les 12 mois d'une année donnée, tandis que la paye selon les données T4 représentent un total annuel véritable. Par exemple, une valeur mensuelle du formulaire PD7 peut avoir été supprimée du fichier de l'EERH parce que l'unité ne faisait pas partie du champ de cette enquête pour un mois donné. En deuxième lieu, dans le cas des enregistrements qui figurent dans les univers des deux sources, il existe plusieurs raisons pour expliquer pourquoi le total des données T4 est supérieur au total annuel agrégé selon les formulaires PD7. Les raisons identifiées sont énumérées ci-après.

a. Il se peut que certains paiements spéciaux inclus dans la paie des T4 ne le soient pas dans la rémunération mensuelle brute déclarée dans les formulaires PD7. Par exemple, on a trouvé des

enregistrements comportant une valeur de remise très importante dans les formulaires PD7 pour un mois particulier chaque année, tandis que la paie déclarée est seulement la paie régulière.

b. Même si certains avantages imposables, comme l'utilisation de l'automobile de l'employeur à des fins personnelles, devraient être inclus dans le montant de rémunération brute déclaré dans les formulaires PD7, certaines compagnies ne les incluent pas. On a trouvé des exemples de paie selon les formulaires PD7 inférieure à la paie selon les T4 en raison de cette situation.

c. Les employeurs ont jusqu'à la fin de février de l'année Y+1 pour déclarer les remises à l'ARC pour l'année Y. Si certains montants de remise pour l'année Y sont déclarés au cours des deux premiers mois de l'année Y+1, cela peut donner lieu à une sous-estimation de la paie pour l'année Y dans les formulaires PD7.

d. Si des entreprises ne reçoivent pas leur formulaire de remise de l'ARC pour un mois, si elles le perdent ou si elles découvrent qu'elles ont fait des erreurs dans le versement de leurs retenues sur la paie, on leur demande d'envoyer un chèque ou un mandat à leur centre fiscal, avec une note indiquant la période de paie à laquelle le montant s'applique. Tous les montants versés à l'ARC sans formulaire PD7 contribuent à la sous-estimation de la rémunération selon les formulaires PD7 par rapport à la rémunération selon les données T4.

Il n'a pas toujours été possible de déterminer les enregistrements qui font partie de ces cas, mais certains l'ont été. Si ces enregistrements sont exclus de l'analyse, la différence entre les deux sources diminue de façon significative et disparaît même pour certaines années. Ainsi, même si on croit que les deux sources de données sont identiques, il existe certaines différences non négligeables entre elles. Ces différences doivent être prises en compte au moment de l'évaluation du nouveau module de C et I.

4. Processus de contrôle et d'imputation au niveau du formulaire

4.1 Méthodologie

Dans les formulaires PD7, les données sur la remise sont plus précises que celles sur l'emploi ou la rémunération, parce que l'ARC procède à une validation, conformément à la description figurant dans la section 3.2. Étant donné la corrélation étroite entre la remise (R_F) et la paie (P_F), et du fait du rapport fiscal connu entre ces variables, le contrôle de la paie au niveau du formulaire pourrait être très efficace.

Par contre, il est difficile de contrôler l'emploi au niveau du formulaire. Pour une entreprise donnée, le nombre total d'employés est habituellement assez stable d'un mois de référence à l'autre, et les erreurs dans l'emploi sont souvent faciles à déterminer après l'agrégation au niveau mensuel. Toutefois, ce n'est pas toujours le cas d'un formulaire PD7 à un autre pour une entreprise donnée. Cela vient de ce qu'une entreprise peut soumettre des formulaires distincts pour différentes catégories d'employés. Étant donné que le rapport entre E_F et P_F, ou entre E_F et R_F, peut être très différent d'une catégorie à l'autre, cela rend difficile la détermination des erreurs dans E_F au niveau du formulaire. Par conséquent, les règles de contrôle au niveau du formulaire sont axées sur le contrôle de la paie, mais ce ne sont pas tous les problèmes qui peuvent être décelés à ce niveau. Un processus de contrôle et d'imputation est requis au niveau mensuel agrégé, particulièrement pour déterminer les problèmes possibles au chapitre de l'emploi.

4.1.1 Hypothèses relatives aux données

Le système élaboré est fondé sur un ensemble d'hypothèses énoncées de façon explicite concernant le contenu des variables P_F, E_F et R_F et les rapports entre elles. Par exemple, nous partons du principe que E_F est un nombre entier et que $E_F > 0$ et $P_F > E_F$. Si on considère que la remise annuelle versée est définie par les lois fédérale et provinciales sur l'impôt, des hypothèses additionnelles concernant les ratios R_F/P_F et P_F/E_F peuvent être définies. Par exemple, pour le Québec, on utilise l'hypothèse suivante : $0,05 < R_F/P_F < 0,4$. Pour les provinces à l'extérieur du Québec, on utilise l'hypothèse suivante : $0,1 < R_F/P_F < 0,6$.

Chaque formulaire est évalué, afin de déterminer s'il satisfait ou non chaque hypothèse. La population des formulaires est par la suite divisée en groupes qui s'excluent mutuellement et qui sont exhaustifs. Pour chaque groupe, on détermine le niveau de changement minimum requis pour modifier le moins grand nombre de variables afin de satisfaire les hypothèses, et un ensemble de règles de contrôle est appliqué.

4.1.2 Règles de contrôle au niveau du formulaire

Il existe deux catégories de règles de contrôle : les contrôles de validité et les contrôles de cohérence. Les contrôles de validité ne portent que sur une variable d'un formulaire et servent à déterminer les situations où la variable se situe à l'extérieur d'une fourchette acceptable. Par exemple, comme la remise dans le formulaire est toujours disponible et positive, le champ de la paie doit être supérieur à 0. Par conséquent, si les données sur la paie du formulaire sont égales à 0, on considère le champ comme manquant et on l'impute plus tard. Les contrôles de cohérence ont trait aux rapports entre les variables d'un formulaire. Certains des contrôles de cohérence utilisent uniquement les renseignements du formulaire contrôlé, tandis que d'autres utilisent des données historiques (p. ex., les données des formulaires précédents pour le même groupe d'employés). Parmi les contrôles de cohérence qui n'utilisent pas de données historiques figure l'établissement de P_F à manquant et son identification en vue de l'imputation, si $P_F < R_F$.

Les contrôles de validité sont appliqués en premier et sont suivis par les contrôles de cohérence utilisant des données historiques, puis par les contrôles de cohérence utilisant des données du formulaire. L'ordre de priorité de l'application des contrôles historiques de cohérence est fondé sur le concept d'une mesure de distance. Cette mesure de distance est définie du point de vue de la fiabilité perçue des trois variables et du nombre de variables utilisées pour l'appariement. Des règles de contrôle historiques qui utilisent trois variables sont appliquées en premier. Celles qui utilisent deux variables sont appliquées en deuxième lieu. Étant donné que R_F est considérée comme plus fiable que P_F, les règles de contrôle fondées sur l'appariement avec R_F et E_F comportent un ordre de priorité plus grand que les règles fondées sur l'appariement avec P_F et E_F. Des règles de contrôle comprenant un appariement avec une variable sont appliquées par la suite. On était d'avis que R_F était la seule variable suffisamment fiable pour les critères d'appariement avec une variable seulement. Enfin, tous les formulaires comportant des problèmes évidents pour les trois variables P_F, E_F et R_F sont automatiquement envoyés au traitement manuel.

Certaines recherches ont été effectuées afin de déterminer le meilleur choix pour le nombre maximum de formulaires à utiliser dans les tentatives d'appariement, ainsi que le délai maximal entre les formulaires appariés. Il a été décidé de permettre les tentatives d'appariement à un maximum de 48 formulaires distant d'au plus un an.

Plus de renseignements concernant la méthodologie du processus de contrôle au niveau du formulaire se trouvent dans Wirth (2007).

4.1.3 Processus d'imputation au niveau du formulaire

On a recours à l'imputation au niveau du formulaire pour trouver la meilleure estimation de P_F, lorsque la paie ne peut être déterminée pendant le processus de contrôle. On impute uniquement P_F pour chaque formulaire, tandis que le nombre d'employés est imputé au niveau mensuel agrégé au besoin.

Sept méthodes d'imputation de P_F sont suggérées au niveau du formulaire. Comme c'est le cas pour le processus de contrôle, la priorité des méthodes à appliquer est fondée sur le concept d'une mesure de distance (fiabilité). Les méthodes qui utilisent des données historiques plus exactes sont appliquées en premier, étant donné qu'on croit qu'elles produisent une meilleure estimation de la paie. Par exemple, si on trouve un appariement proche pour les valeurs de l'emploi et de la remise parmi les formulaires antérieurs qui n'ont été rejetés à aucun contrôle, les données (E_F, P_F, R_F) du formulaire apparié servent à imputer la paie du formulaire courant (méthode 1). S'il n'y a pas d'appariement proche pour les valeurs de l'emploi et de la remise, la méthode suivante d'imputation est fondée sur un appariement proche de la remise seulement (méthode 2). Si aucun appariement proche n'est trouvé, les trois méthodes d'imputation suivantes utilisent des ratios fondés sur des données historiques agrégées sur la paie, la paie attendue ou la remise attendue (méthodes 3, 4 et 5) pour imputer la paie du formulaire courant. Pour chaque formulaire, la paie attendue pour un niveau d'emploi et de remise donné est calculée selon le modèle d'imposition

provincial. Si une entreprise n'a jamais déclaré la paie dans ses formulaires, celle-ci est imputée à partir de la paie attendue (méthode 6). La dernière méthode d'imputation est utilisée uniquement lorsqu'il n'y a pas d'antécédents de déclaration, ni aucune donnée sur l'emploi dans le formulaire courant. Avec cette méthode, P_F est imputée au moyen de la valeur de remise divisée par un des deux ratios de la remise sur la paie, selon la province d'emploi.

Il convient de souligner qu'une fois P_F imputée, les contrôles postérieurs à l'imputation font en sorte que la paie imputée respecte les hypothèses présentées à la section 4.1.1.

4.2 Comparaison avec les données T4

Comme il a déjà été mentionné, l'un des principaux objectifs de l'ajout d'un module de traitement au niveau du formulaire est d'augmenter la qualité des données, ainsi que la cohérence avec les données T4. Afin d'évaluer si cet objectif est atteint, une comparaison de la paie annuelle du fichier PD7 traité et du fichier T4 a été effectuée au niveau du Canada en combinant toutes les industries. À cette fin, on a calculé la différence en pourcentage entre la paie totale annuelle des deux sources pour les enregistrements communs à ces deux sources. Ce pourcentage est calculé de la façon suivante : $(P_{PD7} - P_{T4}) \times 100 / P_{T4}$, où P_{PD7} et P_{T4} correspondent respectivement à la paie annuelle du fichier PD7 et du fichier T4.

La comparaison a été faite pour la paie à deux étapes du processus : i) après l'agrégation au niveau mensuel (autrement dit, après le traitement au niveau du formulaire, mais avant le traitement au niveau mensuel), et ii) après le traitement au niveau mensuel.

Dans le système actuel, seuls quelques contrôles automatiques sont effectués au niveau du formulaire. La majeure partie du processus de contrôle et d'imputation est effectuée au niveau mensuel, y compris des corrections manuelles.

Dans le nouveau système, de nombreux contrôles s'ajoutent au niveau du formulaire et l'imputation de la paie est faite à ce niveau si possible. L'imputation au niveau mensuel est par la suite effectuée pour les entreprises qui n'ont envoyé aucun formulaire pour l'ensemble du mois. Aucune correction automatique ou manuelle n'est effectuée à ce niveau.

On considère que le nouveau système produit de meilleurs résultats que le système actuel si la différence en pourcentage entre la paie annuelle agrégée du formulaire PD7 et la paie annuelle T4 est plus faible avec le nouveau système qu'avec le système actuel. Les tableaux 4.2-1 et 4.2-2 présentent respectivement les différences en pourcentage avec le système de traitement actuel et avec le nouveau système.

Tableau 4.2-1
Différence en pourcentage entre la paie des formulaires PD7 et la paie des T4 – Traitement actuel des données administratives

Étape du traitement des données	Année		
	2003	2004	2005
	Différence en pourcentage		
Paie après agrégation	-10,3 %	-10,2 %	-8,1 %
Paie après traitement au niveau mensuel	-9,3 %	-9,5 %	-10,1 %

Tableau 4.2-2
Différence en pourcentage entre la paie des formulaires PD7 et la paie des T4 – Nouveau traitement des données administratives, y compris le module de C et I au niveau du formulaire

Étape du traitement des données	Année		
	2003	2004	2005
	Différence en pourcentage		
Paie après agrégation	-4,7 %	-5,1 %	-3,8 %
Paie après traitement au niveau mensuel	-2,0 %	-1,7 %	-1,1 %

On observe que les différences en pourcentage sont beaucoup plus faibles lorsque l'on applique le nouveau module au niveau du formulaire (tableau 4.2-2) qu'avec le système actuel (tableau 4.2-1). Par ailleurs, les résultats mentionnés dans la section 3.3 montrent que, pour les enregistrements non modifiés par le nouveau traitement des données administratives, la paie annuelle agrégée des formulaires PD7 est plus faible que la paie annuelle des T4, dans une proportion d'environ 1,8 % en 2003 et de 1,4 % en 2004 et 2005. Étant donné que la différence pour les enregistrements non modifiés est très similaire à celle de la dernière ligne du tableau 4.2-2, on croit que le nouveau module de traitement au niveau du formulaire augmente considérablement la cohérence entre les sources de données.

5. Sommaire

Un module de contrôle au niveau du formulaire a été élaboré pour compléter le processus actuel au niveau mensuel afin de déterminer et de corriger les erreurs de déclaration évidentes et explicables, le plus rapidement possible dans le processus de traitement. En outre, un processus d'imputation au niveau du formulaire fournit une valeur réaliste de la paie pour chaque formulaire PD7, selon de nombreux facteurs, comme le modèle de déclaration historique de l'entreprise et les règlements touchant l'impôt sur le revenu. Cette amélioration augmente considérablement la cohérence des estimations mensuelles de la paie en comparaison avec d'autres sources de données. D'autres études devront être effectuées pour améliorer encore davantage la qualité des données concernant le nombre d'employés au niveau mensuel. Malheureusement, il n'existe pas de source similaire aux données T4 pour servir de référence dans ce cas.

L'élaboration de ce nouveau processus fourni de l'information sur les modèles de remise des employeurs. Statistique Canada a maintenant la permission de communiquer avec les employeurs directement pour obtenir des précisions concernant les données fournies. Cela permettra d'améliorer encore davantage les spécifications de contrôle.

Bibliographie

Hidiroglou, M.A. et Berthelot, J.-M. (1986), Contrôle statistique et imputation dans les enquêtes-entreprises périodiques, *Techniques d'enquête*, 12, 79-89.

Rancourt, E. et Hidiroglou, M. (1998). Use of Administrative Records in the Canadian Survey of Employment, Payrolls, and Hours, *Recueil du groupe des méthodes d'enquête*, Société statistique du Canada, 39-49.

Wirth, S. (2007). SEPH PD7 Edit System: Explanation of the Editing Concepts, rapport inédit, Ottawa, Canada: Statistique Canada.