



BD/MSPS

Guide de création de la base de données

Le Guide de création de base de données décrit les techniques utilisées pour la construction de la BDSPS.



Statistics
Canada

Statistique
Canada

Canada

Table des matières

Sommaire	i
Introduction.....	1
Objectifs, sources de données et techniques	2
Objectifs.....	2
Sources de données.....	3
Techniques	4
Les données hôtes	6
Pondérations de l'EDTR.....	7
Vérification des variables de l'EDTR.....	7
Imputation des réponses « ne sait pas »	7
Enfants sous tutelle	8
Chef de famille.....	8
Arrondissement du revenu	8
Troncation de l'âge	8
Ajout des personnes âgées en établissement.....	8
Conversion	9
Fractionnement de la base de données.....	10
Appariement par catégorie.....	10
Ajustement du revenu élevé.....	11
Agrégation des micro-enregistrements	11
Appariement par catégorie.....	14
Imputation des données historiques d'assurance-emploi.....	14
Fichier donneur de l'a.-e.	15
Appariement par catégorie.....	17
Duplication des ménages	17
Imputation stochastique des renseignements de l'impôt sur le revenu.....	18
Les données du donneur.....	20
Transformation des données	20
Calcul des statistiques de distribution.....	22
Imputation	24
Imputation des données de l'Enquête sur les dépenses des ménages	25
Détermination des variables d'appariement et habitudes de consommation	25
Appariement par catégorie.....	27
Autres sujets :	28
Frais de garde d'enfants	28
Imputation des semaines travaillées aux demandes de prestations d'a.-e.	28
Références.....	28

Sommaire

Le présent guide décrit la construction de la base de données fournie avec le système **Base de données et modèle de simulation de politiques sociales (BD/MSPS)**. Cette base de données a été conçue spécialement pour permettre l'analyse de politiques fiscales et de transfert ainsi que de taxes de vente. Ces politiques exigent de plus en plus une analyse intégrée qui franchit les limites traditionnelles des programmes et des paliers d'administration. La base de données du système BD/MSPS a été construite en vue de permettre la modélisation de microanalyses par la combinaison des données administratives individuelles obtenues des déclarations de revenus personnelles et des dossiers historiques des prestataires d'assurance-emploi avec les données d'enquêtes sur les revenus des familles et les régimes de dépenses des familles.

Des données administratives agrégées supplémentaires ont été utilisées pour la création des éléments tant base de données que modèle de la BD/MSPS. Les données d'entrées-sorties ont aussi été appliquées à la modélisation des droits et des taxes de vente qui touchent la consommation personnelle. Les techniques utilisées pour la création de la base de données et le respect de la confidentialité des données comprennent diverses formes d'imputation stochastique et d'appariement par catégorie.

Introduction

Au Canada, un petit nombre de ministères du gouvernement fédéral ont eu un monopole virtuel sur la capacité d'analyses détaillées des répercussions qu'ont sur les impôts et les transferts les changements survenant dans les politiques. Le public est vivement intéressé à connaître quels sont les groupes de familles ou d'individus qui gagneront ou perdront du fait d'un projet de politique en particulier. Les gens intéressés, à l'extérieur des ministères en particulier (y compris d'autres ministères du gouvernement fédéral et les gouvernements provinciaux), n'ont pas les moyens nécessaires pour évaluer les estimations publiées de ces répercussions des projets de politique sur la distribution, aucune façon d'explorer plus en détail ou de développer des chiffres comparables pour leurs propres projets. Cette situation ne ressemble pas à celle des États-Unis où divers organismes indépendants, comme la Urban Institute et la Mathematica Policy Inc., ont des capacités de microsimulation perfectionnées. Et cela ne ressemble pas non plus à la situation du secteur des politiques macroéconomique où de nombreux organismes des deux pays fournissent régulièrement des analyses et des prévisions indépendantes.

Avec le système **Base de données et modèle de simulation de politiques sociales (BD/MSPS)** de Statistique Canada, n'importe qui, avec des efforts suffisants, peut faire sur son propre ordinateur personnel des analyses de microsimulation des répercussions des changements apportés au programme fiscaux et de transfert. Le niveau de perfectionnement approche et, dans certains cas dépasse, celui des ministères et du gouvernement fédéral.

La BD/MSPS cadre dans une philosophie différente des produits traditionnels d'un organisme de statistiques national - habituellement des publications imprimées avec des nombreuses tables remplies de chiffres. Le projet BD/MSPS avait, au départ, l'objectif de rendre disponible au public une capacité d'analyse de programmes fiscaux et de transfert entourant les politiques. Étant donné cet objectif, une base de données a été conçue spécialement à cet effet avec un progiciel d'analyse et de récupération de données.

La base de données a été adaptée explicitement aux logiciels et aux applicables d'analyse, contrairement à ce qui se passe dans des situations plus courantes où l'analyse est soumise aux contraintes des données déjà disponibles. Parmi les autres contraintes de développement, la base de données devait être non confidentielle, au sens de la *Loi sur la statistique*, et le progiciel et la base de données devaient être utilisables dans une vaste gamme d'environnements informatiques, particulièrement les PC. Ces contraintes sont nécessaires pour que la BD/MSPS atteigne l'objectif d'une vaste accessibilité du public.

Une analyse des politiques entourant les programmes fiscaux et de transfert ne peut se faire effectivement que par microsimulation. Pour estimer les répercussions probables d'un changement dans les exemptions d'impôt sur le revenu pour divers types de familles par plage de revenus, par exemple, le ministère des Finances du Canada emploie un modèle de microsimulation qui recalcule l'impôt à payer sur le revenu pour un large échantillon de contribuables selon leur déclaration de revenus réelle pour une année récente. Essentiellement, le logiciel analyse une à la fois les déclarations de revenu d'un échantillon représentatif et, pour chacune, il calcul l'impôt correspondant à un autre scénario de politique. De même, le ministère des Ressources humaines et du développement a son propre

modèle de microsimulation, pour le système de l'assurance-emploi, qui utilise un échantillon de ses propres dossiers de données administratives internes.

Dans pratiquement tous les cas au Canada, ce sont seulement (mais pas nécessairement simplement) des calculs comptables; aucune réaction de comportement n'est supposée. La BD/MSPS est semblable de ce point de vue - le logiciel de modélisation ne fait que des calculs comptables.

Un aspect unique et important de la BD/MSPS est le cadre d'analyse fiscale et de transferts qui est intégré dans le système. La BD/MSPS fournit en un progiciel, intégré sur le plan des microdonnées, suffisamment de données pour modéliser l'impôt sur le revenu des particuliers, l'assurance-emploi, les programmes de transferts majeurs (à l'exception des revenus reliés à la pension et au bien-être) et les taxes de consommation. L'analyse au niveau individuel et au niveau de la famille est possible.

Un grand défi de la construction du volet base de données de la BD/MSPS a été l'assemblage et la fusion d'un certain nombre d'ensembles de microdonnées. Il est essentiel que la plus grande partie de la richesse des détails de chacun des ensembles de microdonnées donneurs soit préservée. La fusion des ensembles de microdonnées doit aussi donner des enregistrements de microdonnées qui sont communs ou fusionnés -- chacun d'eux est réaliste ou plausible, même s'il se révèle synthétique et artificiel. D'autre part, l'ensemble de microdonnées résultant doit respecter la Loi sur la statistique et ne pas permettre l'identification de toute personne réelle.

Le présent guide décrit la façon dont la base de données de simulation de politiques sociales a été construite. Nous commençons par les objectifs généraux de la BDSPS et par le caractère des données des sources. Ensuite, la plus grande partie du document décrit les nombreuses étapes de l'assemblage de la BDSPS.

Nous recommandons fortement à l'utilisateur d'étudier ce manuel en profondeur. La validité de l'analyse faite avec la BD/MSPS dépend de la compréhension qu'à l'utilisateur des microdonnées sur lesquelles le modèle repose.

Dans ce guide, l'année de base réfère à l'année sur laquelle toutes les bases de données utilisées pour construire la BDSPS sont alignées.

Objectifs, sources de données et techniques

OBJECTIFS

Pour le développement de la BDSPS, nous avons fait tous les efforts possibles pour conserver la variété et l'utilité des données de la source originale tout en veillant à leur non-confidentialité, de façon que la base de données et le modèle résultant puissent être mis à la disposition du grand public. Quatre grands objectifs ont guidé le choix des techniques, des sources de données et des variables ainsi que du processus :

- **Non-confidentialité et accessibilité au public**

Le premier objectif a été de veiller à ce qu'aucune personne présente dans l'une ou l'autre des bases de données ne puisse être identifiée par une divulgation explicite ou résiduelle. Cette condition préalable devait être respectée si l'on voulait rendre la BD/MSPS accessible au public. Aussi liée à l'accessibilité du public, la nécessité que la base de données et le modèle puissent fonctionner sur un PC de prix moyen.

- **Exactitude de la distribution et de l'agrégation**

La BD/MSPS a été conçue de façon à reproduire aussi étroitement que possible des agrégats « connus », comme le nombre total des prestataires de l'assurance-emploi. En outre, on a déployé des efforts particuliers afin de représenter avec précision la distribution des agrégats entre les diverses classifications essentielles à l'analyse des politiques publiques au Canada, comme la province, l'âge, le revenu, le type de famille et le sexe. Enfin, il importe que, au niveau des microdonnées, la courbe de distribution de variables spécifiques soit bien représentée.

- **Exhaustivité et détails des données**

La sélection et l'agrégation des variables des principales sources de données ont tenté de prévoir les options de politiques probables ainsi que de combler les besoins des actuels modèles fiscaux et de transferts.

- **Uniformité des micro-enregistrements**

Pour des raisons de confidentialité, on a utilisé des techniques stochastiques plutôt que des techniques d'appariement exact. En retour, il a fallu songer à éviter la création d'enregistrements de microdonnées irréalistes - par exemple un couple de personnes âgées sans enfants ayant de pleines déductions d'impôt pour les frais de garde d'enfants.

Ces objectifs centraux sont interdépendants et l'on a fait des compromis entre eux. Le processus utilisé pour en arriver à ces compromis comprenait la consultation d'un groupe de travail spécial formé de représentants de quatre ministères fédéraux qui sont intéressés dans la BD/MSPS obtenu et qui ont de l'expérience dans leurs propres modèles de microsimulation. Le produit final représente donc un compromis entre des questions de méthodologie, d'information, de technologie, de ministères et de politiques publiques.

Un autre objectif a pu être ajouté à posteriori. Dans le domaine des comptes nationaux, on se préoccupe de plus en plus du manque de bases de microdonnées pour les agrégats macroéconomiques, par exemple, dans les écrits de Ruggles. Bien que cela ne cadre pas dans l'intention originale, il advient que la BDSPS peut aussi être considérée comme une base de microdonnées pour le secteur des ménages canadiens, comme l'ont écrit Adler et Wolfson (1988).

SOURCES DE DONNÉES

La BDSPS a été construite à partir de quatre grandes sources de microdonnées.

- **L'Enquête sur la dynamique du travail et du revenu (EDTR)** : principale source de

données de Statistique Canada sur la distribution du revenu entre les individus et les familles a servi de fichiers de données hôte. Elle est riche en données sur la structure des familles et les sources de revenu, mais il lui manque de l'information détaillée sur les dossiers historiques du chômage, les déductions fiscales et les dépenses des consommateurs. Elle remplace l'Enquête sur les finances des consommateurs (EFC) qui a été effectué la dernière fois en 1997. La BDSPS débute avec le fichier de microdonnées à grande diffusion (FMGD);

- Les données des déclarations d'impôt des particuliers : un échantillon de plus de 400 000 déclarations de revenu (T1) des particuliers utilisé comme base de la publication annuelle de l'Agence du revenu du Canada, **Statistiques sur les statistiques du revenu** (aussi connu sous le nom de Livre vert et anciennement connu en tant que Statistiques fiscales);
- Les dossiers historiques des demandes de prestations d'assurance-emploi (a.-e.) : un échantillon de 10 % des dossiers historiques du système administratif de Développement des ressources humaines; et
- L'**Enquête sur les dépenses des ménages** (EDM) : enquête périodique de Statistique Canada contenant des données très détaillées sur les revenus des Canadiens et les régimes de dépenses des ménages, comprenant de l'information sur les changements nets dans l'actif et le passif (épargnes annuelles). La BD/SPS débute avec le fichier de microdonnées à grande diffusion (FMGD).

Aux fins de la base de données de simulation de politique sociale (BDSPS), ces quatre sources de données ont été transformées en un ensemble de microdonnées non confidentielles et publiques. En outre, ces microdonnées ont été améliorés par des références à diverses données agrégées qui ont servi principalement de jalons ou de totaux de contrôle.

TECHNIQUES

La mise en commun des quatre fichiers de microdonnées initiaux, l'ajout de nouveaux renseignements et le remplacement ou l'ajustement des mesures biaisées dépendaient largement de cinq techniques employées de façon intensive pour la création de la BDSPS : conversion, méthode itérative du quotient par régression, imputation stochastique, agrégation de micro-enregistrements et appariement par catégorie.

- La **conversion** est une méthode d'ajustement des microdonnées permettant de contourner le problème de la non-réponse à des éléments. Elle implique l'identification des personnes qui n'ont déclaré aucun paiement d'un programme en particulier (c.-à-d. prestations d'assurance-emploi) et l'imputation de paiements à ces personnes (c.-à-d. qu'elles sont « converties » en répondants).
- La **méthode itérative du quotient par régression** désigne une technique permettant de réduire le biais en forçant la correspondance entre des données et des totaux de contrôle connus. Les poids des ménages sont forcés (de façon itérative) à correspondre à des totaux de contrôle connus au niveau de l'individu. Les totaux de contrôle employés sont la population par âge, par sexe et par province; distribution des salaires par province; distribution d'emploi indépendant non agricole par province; statistiques d'emploi

indépendant agricole par province; taille de la famille par province; ainsi que taille des ménages par province.

- L'**imputation stochastique** est la génération de valeurs de données synthétiques pour les particuliers dans un ensemble de données hôte par l'extraction aléatoire des fonctions de densité ou de distribution dérivées de l'ensemble de données de la source.
- L'**agrégation de micro-enregistrements** est le processus qui consiste à créer des micro-enregistrements synthétiques par la mise en groupe d'enregistrements similaires. Ainsi, des micro-enregistrements de contribuables à revenu élevé sont mis en groupes de cinq selon des critères correspondant à la politique. Dans chaque groupe de cinq enregistrements, des valeurs de variables pertinentes (p. ex., les gains en capital) sont pondérées (en moyenne) pour créer des enregistrements non identifiables qui ressemblent aux microdonnées, mais qui sont, en réalité, synthétiques.
- L'**appariement par catégorie** implique en premier lieu la classification des enregistrements des ensembles de données hôte et donneur selon les critères propres à la politique qui sont communs aux deux ensembles de données (p. ex., le type de logement, le statut d'emploi, la classe de revenu). L'information des enregistrements donneurs ainsi classifiée peut alors être attribuée à des enregistrements ayant des caractéristiques similaires de notre ensemble de données hôte sans qu'elle ne puisse être identifiée.

La figure 1 donne un aperçu du processus de création de la BDSPS. Les ellipses représentent les fichiers de données (p. ex., l'EDTR, le Livre vert) et les rectangles représentent les processus. La section suivante du guide décrit chaque étape de la construction de la BDSPS, comme elle est représentée à la figure 1.

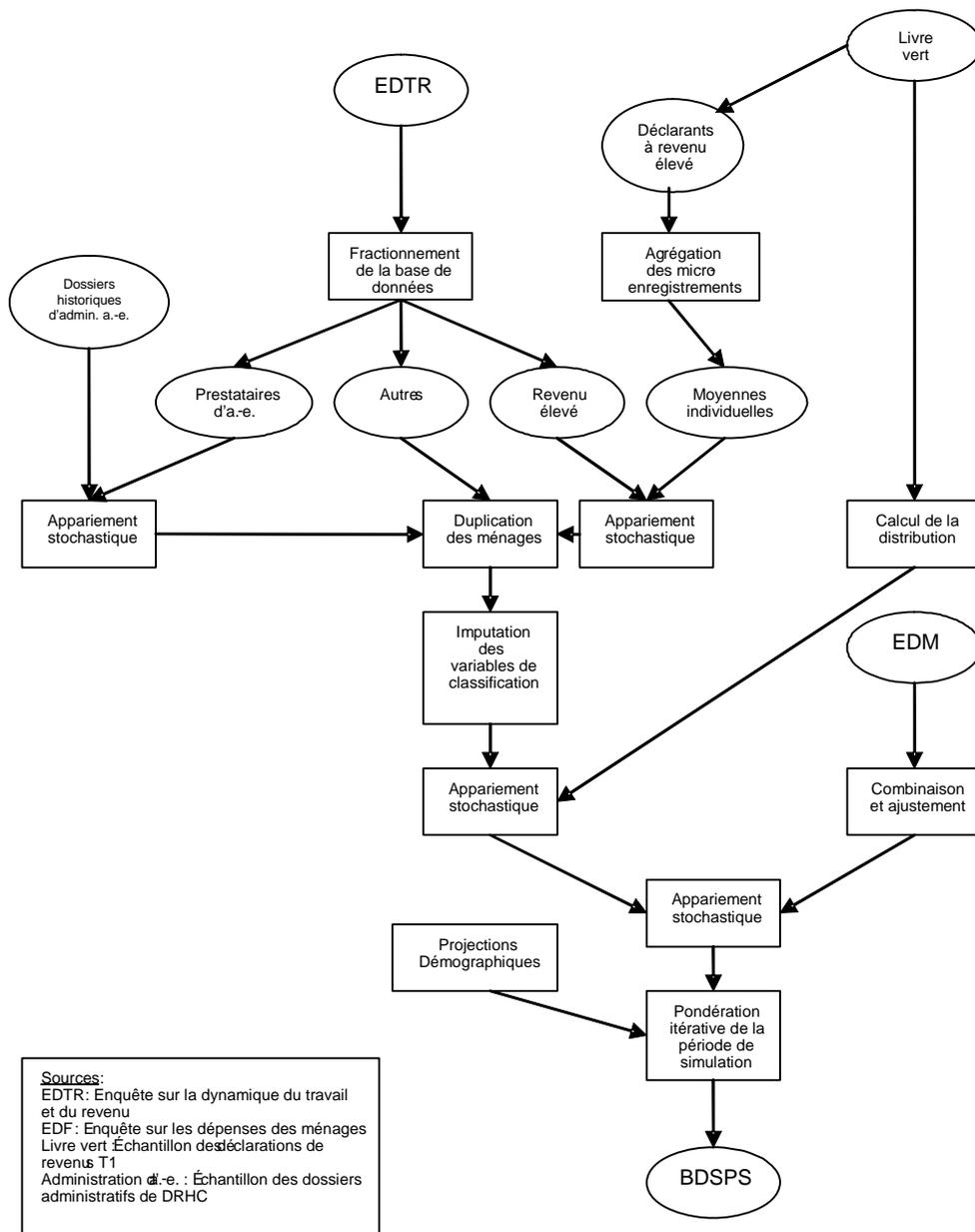


Figure 1 : Processus de création de la base de données BDSPS

Les données hôtes

Le fichier de données « hôte » ou « cible » est dérivé de l'**Enquête sur la dynamique du travail et du revenu** de Statistique Canada (EDTR) pour l'année de base. L'EDTR est une enquête annuelle auprès de ménages sélectionnés extraits du cadre d'enquête de l'Enquête

sur la population active (EPA). Il s'agit d'une enquête longitudinale, les répondants faisant partie de l'échantillon pendant six ans. En janvier, on leur pose des questions sur leur expérience du marché du travail durant l'année précédente ainsi que tout changement survenu sur le plan des études ou de la famille. En mai, on recueille les données sur le revenu de l'année précédente. On utilise deux méthodes pour recueillir des renseignements sur les revenus. Une forte proportion de répondants (environ 80 %) ont autorisé Statistique Canada à consulter leurs déclarations de revenus T1. Le reste des répondants sont interviewés en mai, ce qui leur permet de s'appuyer sur leurs déclarations de revenus pour répondre. La BDSPS débute avec le fichier de microdonnées à grande diffusion de l'EDTR.

L'information des fichiers d'assurance-emploi, du Livre vert et de l'EDM ont alors été « ajoutés » à l'EDTR. Afin de permettre l'exploitation de toute la variété de cette information qui est imputée d'autres sources, de nombreux enregistrements originaux de l'EDTR ont fait l'objet d'une copie ou d'une duplication. Par exemple, les enregistrements représentant les prestataires d'a.-e. ou les personnes qui pourraient recevoir de l'a.-e. ont été reproduits. Les enregistrements représentant les particuliers à revenu élevé (ceux dont le revenu dépasse 135 000 \$) ont été reproduits de façon que leur nombre corresponde à celui des enregistrements de personnes à revenu élevé dérivés de l'agrégation des micro-enregistrements de l'échantillon de l'Agence du revenu du Canada. Pour conserver la structure de famille et la somme globale des poids, les enregistrements de toutes les autres personnes des ménages comprenant des individus en chômage ou à revenu élevé ont été reproduits en double de façon similaire. Le poids attribué à un enregistrement a été réduit de façon à tenir compte du nombre de fois où il y a eu duplication.

PONDÉRATIONS DE L'EDTR

Traditionnellement, les salaires et traitements de l'EDTR produisaient des salaires totaux qui étaient plus grands que les estimés des T1 ou T4 correspondants. Les estimations du fichier T1 sont des estimés des salaires et traitements qui sont dérivés de la déclaration de revenus des individus. Les estimations du fichier T4 sont des estimés qui proviennent des formulaires T4 que l'employeur remplit et qui est envoyé aux employés. Une comparaison avec le fichier T4 démontrait une surreprésentation de la population dans le groupe intermédiaire et une sous représentation de la population à bas salaires. Puisque les salaires représentent la plus grande source de revenu marchand au Canada, l'EDTR utilise le fichier T4 comme référence pour étalonner la répartition des salaires.

VÉRIFICATION DES VARIABLES DE L'EDTR

Le fichier de microdonnées à grande diffusion de l'EDTR est le point de départ de la BDSPS. De nombreuses modifications ont été apportées aux variables dans ce fichier, dont certaines des plus importantes sont exposées ici.

Imputation des réponses « ne sait pas »

Pour certaines variables de l'EDTR qui sont utilisées dans la BDSPS, la réponse est « ne sait pas ». Ces variables sont la province de résidence, l'état matrimonial, la profession, la situation d'emploi l'année précédente, le niveau de scolarité et la situation d'études l'année précédente. On a procédé à des régressions ou des tirages aléatoires à partir des répartitions figurant dans le FMGD pour imputer des réponses pour ces personnes.

Enfants sous tutelle

Dans l'EDTR, les enfants de moins de 18 ans qui n'habitent pas avec un parent ne sont pas inclus dans les familles de recensement. Dans la BDSPS, ces enfants ont été considérés comme étant les enfants du soutien économique principal.

Chef de famille

La définition de chef de famille utilisée dans l'EDTR a été modifiée pour la BDSPS. Les familles de recensement sont les couples (mariés ou en union libre), les parents seuls et leurs enfants de moins de 25 ans qui habitent avec eux. Dans la BDSPS, les personnes hors famille sont considérées être des familles de recensement d'une personne. Dans le cas d'un couple, l'homme est réputé être le chef de la famille de recensement. Autrement, le parent seul ou la personne hors famille est le chef de famille.

Les familles économiques se composent de personnes qui habitent dans le même ménage et qui sont liées par le sang, le mariage, l'union libre ou l'adoption. Dans la BDSPS, les personnes hors famille sont considérées être des familles économiques d'une personne. Le chef de la famille économique est défini dans la BDSPS comme étant le chef de la famille de recensement qui comprend le soutien économique principal.

Le chef de la famille économique qui comprend le soutien économique principal du ménage devient le chef du ménage.

Arrondissement du revenu

Les variables de revenu dans le FMGD de l'EDTR sont arrondies. Comme des variables « non arrondies » sont nécessaires aux fins de la modélisation des données fiscales, les variables de revenu étaient toutes « non arrondies ». La méthodologie utilisée pour « désarrondir » les données n'était pas la même que celle utilisée pour les « arrondir ». Cette façon de procéder visait à assurer la confidentialité des techniques de perturbation. On a « désarrondi » les variables à l'intérieur des intervalles tout en respectant la répartition dans le Livre vert, un échantillon de déclarations de revenus T1. Pour garantir la confidentialité du Livre vert et s'assurer que des valeurs uniques ne seraient pas trouvées, on a fait la moyenne des valeurs dans le Livre vert en groupes de cinq avant de créer les répartitions.

Troncation de l'âge

Dans l'EDTR, l'âge est tronqué à 80 ans. La structure par âge de la population selon le Recensement de 2001 est utilisée pour définir des tranches d'âge plus diverses pour les personnes âgées. L'imputation de l'âge aux personnes âgées tient compte également des répartitions dans le recensement pour les personnes en établissement, l'état matrimonial et le sexe. Il convient de signaler qu'on procède préalablement à l'imputation des personnes âgées en établissement, de sorte que ces enregistrements clonés peuvent indiquer des âges différents que les enregistrements originaux.

AJOUT DES PERSONNES ÂGÉES EN ÉTABLISSEMENT

La base de sondage de l'EDTR n'inclut pas les personnes âgées en établissement. Toutefois, comme les personnes âgées constituent un groupe nombreux et touché par les politiques, on

l'inclut de nouveau dans la BDSPS. À cette fin, on repère les personnes dans l'EDTR qui habitent seules et qui n'ont pas travaillé durant l'année précédente. Un certain nombre de ces observations sont alors copiées et étiquetées comme correspondant à des personnes en établissement. On continue de faire des copies jusqu'à ce que le nombre d'enregistrements dans la BDSPS corresponde à la proportion de la population de personnes âgées en établissement selon la province, l'âge et le sexe dans le recensement.

CONVERSION

Des preuves de l'extérieur laissent entendre que la sous-déclaration des prestations d'assurance-emploi, d'aide sociale et les paiements du RPC/RRQ sont susceptibles d'être des non-réponses à des points. Le problème ne réside pas dans le fait que les prestataires sont sous-représentés dans l'échantillon; c'est plutôt qu'ils oublient ou négligent de déclarer les paiements. Le fait que la sous-déclaration de ces points est nettement moins importante dans l'EDTR que dans l'Enquête sur les finances des consommateurs vient étayer ces preuves. Les variables de revenu d'une partie des répondants à l'EDTR ont été imputées à partir de leurs déclarations de revenus, ce qui s'est traduit par une diminution de la non-réponse.

La technique de conversion tente de régler ce problème de non-réponse à des points par l'identification des individus appropriés qui n'ont pas déclaré de paiements et l'imputation d'un paiement à ces individus (c.-à-d. qu'ils sont « convertis » en répondants). Cette étape d'ajustement de la base de données est entreprise, comme méthode itérative du quotient par régression (voir la section suivante), afin de faire en sorte que la base de données soit en équilibre avec des totaux de contrôles pour les éléments qui font l'objet de l'itération. Cette conversion est entreprise non pas dans la préparation des microdonnées de la BDSPS, mais pendant l'exécution du MSPS comme l'indiquent les paragraphes suivants. Nous décrivons ce processus ici parce que, comme les autres aspects de la création de la base de données, il influe sur la nature des microdonnées de la BDSPS et peut présenter un intérêt quand on interprète les résultats de l'analyse.

Il peut y avoir une tendance dans les occurrences de non-réponse. Par exemple, les non-répondants de l'a.-e. peuvent comprendre ceux dont les prestations se sont terminées au cours des quelques premières semaines de l'année civile. Dans ce cas, toute tentative d'identification des vrais non-répondants devrait comprendre un examen des individus qui peuvent avoir connu quelques semaines de chômage par année.

En l'absence d'information auxiliaire sur les tendances de non-réponse, une tentative d'identification et de conversion des vrais non-répondants peut introduire des distorsions dans la base de données. Comme dans l'exemple ci-dessus, les non-répondants peuvent avoir des caractéristiques passablement différentes de répondants.

La stratégie de conversion qui a été adoptée était conçue de façon à introduire aussi peu de distorsions que possible. La première étape implique le calcul d'une régression logistique de l'état des réponses (c.-à-d., répondant/non-répondant) pour attribuer une probabilité de réponse à chaque individu. En effet, ceci permet de classer les non-répondants au chapitre de la similarité avec les répondants.

L'identification des personnes qui seront converties a été effectuée : par la méthode du

classement. La méthode du classement garantit que les totaux de contrôle sont vérifiés et ne convertit que ceux qui sont semblables aux répondants.

Méthode du classement : - Dans des classes déterminées par des totaux de contrôle, elle convertit les non-répondants des classes plus hautes jusqu'à ce que les totaux de contrôle soient vérifiés.

FRACTIONNEMENT DE LA BASE DE DONNÉES

Le fractionnement est une étape de préparation mécanique de données qui répartit l'EDTR en quatre sous-ensembles : les individus à revenu élevé, les prestataires d'a.-e., les prestataires potentiels d'a.-e. et les autres. Il est à noter qu'une personne peut faire partie à la fois du sous-ensemble pour revenu élevé et du sous-ensemble d'a.-e. Les individus à revenu élevé sont ceux qui sont définis comme des déclarants à revenu élevé, tandis que les prestataires d'a.-e. sont ceux qui (i) ont déclaré avoir reçu certaines prestations dans l'Enquête sur les sur la dynamique du travail et du revenu, (ii) qui ont été convertis en prestataires du fait des non-réponses imputées, ou (iii) qui ont été considéré en tant que prestataires potentiels d'a.-e.

Appariement par catégorie

L'appariement par catégorie implique la création d'enregistrements composites « fusionnés » à partir de deux bases de données de microdonnées. Supposons deux bases de données, une base de données **A** hôte et une base de données **B** donneuse. Diverses méthodes permettent d'attribuer, en tout ou en partie, l'information d'un enregistrement de la base de données **B** à tout enregistrement donné de la base de données **A**. Tout repose sur l'idée que nous désirons trouver un enregistrement de la base de données **B** qui est dans un certain sens similaire à l'enregistrement donné de la base de données **A**. La détermination de la similarité repose sur des variables communes aux deux bases des données et est influencée par l'utilisation prévue des enregistrements fusionnés. Divers algorithmes « du plus proche voisin », qui utilisent des méthodes similaires à celles de l'analyse typologique, peuvent être utilisés si l'on veut déterminer un appariement mathématiquement « optimal », étant donné une méthode particulière de détermination de la distance dans un espace à N dimensions. Les complications surgissent dans la pratique du fait des limites de taille de l'ensemble d'enregistrements « du donneur » (base de données **B** dans notre exemple) et le désir d'utiliser des variables non continues (p. ex., discrètes ou de classe).

Dans la BDSPS, une technique différente, plus heuristique, est utilisée. Elle implique le partitionnement de deux bases de données dans des « cases » d'enregistrements définis de façon identique qui sont par la suite triés selon une des variables continues communes aux deux bases de données (habituellement le revenu total dans la BDSPS). Les enregistrements qu'il y a dans une case donnée sont alors appariés directement entre les deux bases de données (c.-à-d. l'enregistrement **n** dans la case **m** de la base de donnée **A** est apparié avec l'enregistrement **n** de la case **m** de la base de données **B**). Les complications surviennent parce que le nombre d'enregistrements dans une case donnée n'est habituellement pas égal dans les deux bases de données et, en outre, parce qu'il y a des poids d'enregistrements dans l'une des bases des données ou dans les deux. Ces problèmes sont résolus par la duplication sélective d'enregistrements d'une des bases de données ou des deux.

La BDSPS utilise l'appariement par catégories pour ajouter les données de l'a.-e. et les données sur le revenu du Livre vert pour les personnes qui ont un revenu élevé ainsi que les données de l'EDM de tous les ménages. Cette technique permet de préserver les corrélations entre éléments de l'enregistrement donneur. Chacune des procédures d'appariement est décrite plus en détail dans les sections suivantes.

Ajustement du revenu élevé

Les estimés de l'EDTR pour les individus à revenu élevé (de même que pour le revenu de ces individus) sont plus bas que n'en indiquent les enregistrements d'impôt sur le revenu des particuliers. La sous-déclaration et la non-déclaration de nombreux éléments de déduction et de revenu sont tous deux pris en compte dans la création de la BDSPS. La figure 2 donne un aperçu de ce processus d'ajustement du revenu élevé.

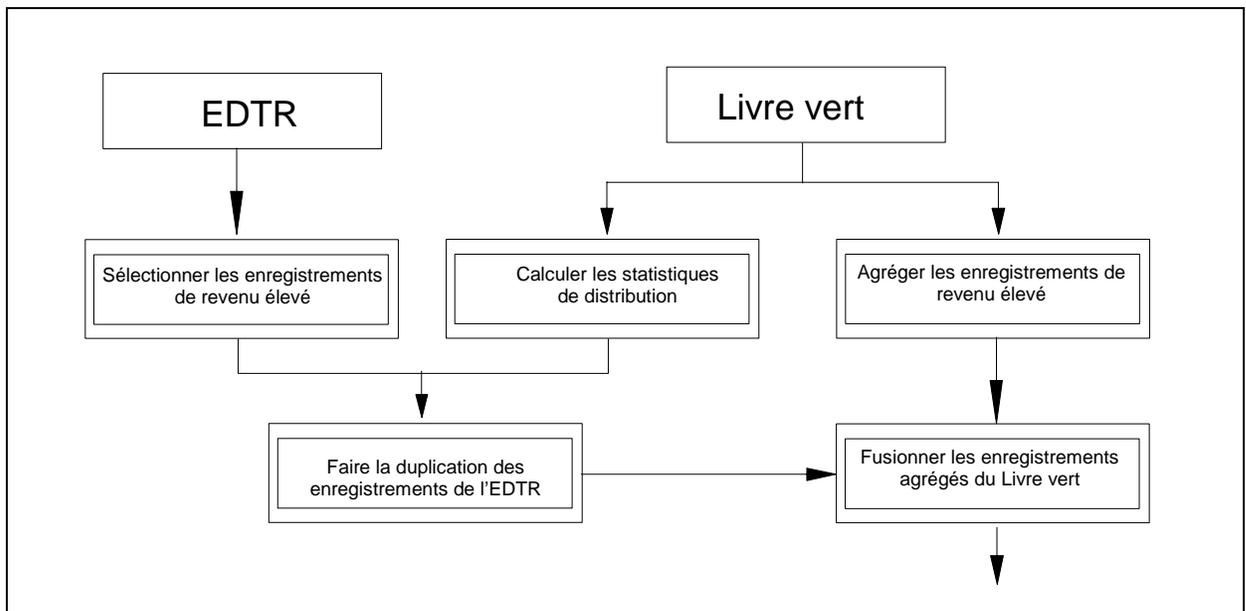


Figure 2 : Processus d'ajustement du revenu élevé

AGRÉGATION DES MICRO-ENREGISTREMENTS

La non-déclaration des individus à revenu élevé dans l'EDTR est compensée par l'utilisation comme totaux de contrôle des comptes du Livre vert pour les déclarants à revenu élevé. Le poids de chaque enregistrement de personne à revenu élevé dans l'EDTR est ajusté de façon que la somme des poids corresponde au Livre vert. Ces enregistrements de l'EDTR sont utilisés comme « hôtes » pour l'acceptation de l'information plus précise du Livre vert. Cela fournit ensuite la base d'un ajustement des éléments de revenu pour le groupe à revenu élevé.

Même avec un ajustement à la hausse des poids pour les enregistrements de personne à revenu élevé de l'EDTR, il y a quand même une sous-déclaration substantielle des revenus dans ce groupe. Comme seconde étape, le biais de la sous-déclaration est corrigé par le remplacement des éléments de revenu ainsi que quelques éléments de déductions de ces enregistrements par des ensembles d'éléments de revenu plausibles, mais non identifiables,

du Livre vert.

Éléments de revenu et de déduction de l'EDTR remplacés pour les individus à revenu élevé

- **Revenus liés à l'emploi**

- idiemp - Revenus d'emploi
 - idisefm - Revenu net de travail agricole
 - idisenf - Revenu d'emploi indépendant - Non agricole

- **Revenus**

- ididiv - Montant imposable de dividendes canadiens
 - idiinvnd - Autres revenus d'investissement
 - idicapg - Gains/pertes en capital imposables pour l'année
 - idipens - Revenu de pension
 - iditrsp - Retraits de REER

- **Autres revenus**

- idialimo - Pension alimentaires
 - idiworkc - Indemnités d'accidenté du travail
 - iditoth - Autre revenu imposable

- **Retenues sur le revenu total**

- idrpp - Cotisations à un régime enregistré de pensions (207)
 - idrrsp - Déduction pour REER (208)
 - iddues - Cotisations annuelles syndicales, professionnelles et semblables (212)
 - idiloss - Pertes au titre d'un placement d'entreprise (217)
 - idmovexp - Frais de déménagement (219)
 - iddalimo - Pension alimentaire versée (220)
 - idcarry - Frais financiers et frais d'intérêt (221)
 - idexplor - Frais d'exploration et d'aménagement (224)
 - idalexp - Autres dépenses d'emploi (229)
 - idothded - Autres déductions du revenu total (232)

- **Retenues sur le revenu net**

- idpartlo - Pertes comme commanditaire d'autres années (251)
 - idnclos - Pertes autres que des pertes en capital d'autres années (252)
 - idcloss - Déduction pour gains en capital (253)
 - idcapgex - Déduction pour gains en capital (254)
 - idadddd - Autres déductions du revenu net (256)

- **Information sur les crédits d'impôt non remboursables et remboursables**

- idtuin - Frais de scolarité pour soi (323)
 - idmedgro - Frais médicaux bruts (330)

- idcgless - Gains en capital imposables résultant de dons (339)
- idcharit - Dons de charité (340)
- idgifts - Dons - Canada/provinces/culture (342)
- idpolcon - Contributions politiques fédérales (409)
- idlabtxg - Fonds de travailleurs brut acheté (414)
- idmincar - Report d'impôt minimum (427)
- idfortx - Impôt payé à un pays étranger (431)
- idforinc - Revenu étranger net (433)
- idgstreb - Remboursement de la T.P.S. (457)
- idprvftc - Crédit d'impôt étranger provincial

Pour chaque province, les enregistrements du Livre vert sont regroupés en ensembles d'au moins cinq enregistrements. Ces enregistrements regroupés sont considérés comme étant une table non confidentielle bien qu'ils conservent un grand nombre des caractéristiques des micro-enregistrements. Les groupes représentent les individus ayant des gains en capital, des revenus de dividendes, des revenus d'investissements, des revenus d'emploi et un âge similaires. Pour ces groupes, une moyenne pondérée est calculée des éléments indiqués ci-dessus. Une fois regroupés, les enregistrements sont considérés comme non confidentiels puisqu'ils représentent au moins cinq individus. Cela équivaut à publier une table dans laquelle chaque cellule contient au moins cinq individus. En plus, l'addition d'un poids à une moyenne ajoute de l'incertitude

L'agrégat résultant contient des milliers de pseudo-enregistrements de microdonnées représentant plusieurs dizaines de milliers d'enregistrements du Livre vert qui représentent à leur tour plus de trois centaines de milliers de déclarants à revenu élevé. Ces enregistrements agrégés, dérivés de microdonnées autrement confidentielles peuvent maintenant faire partie d'un ensemble de données public dans lequel il y a peu de pertes d'information.

APPARIEMENT PAR CATÉGORIE

Les enregistrements de personnes à revenu élevé originaux de la BDSPS sont copiés en double de façon que le nombre total corresponde au nombre d'enregistrements agrégés du Livre vert de personnes à revenu élevé par province. Ces enregistrements ne fournissent pas une base suffisante pour les caractéristiques démographiques de la population des déclarants à revenu élevé. Donc, un appariement détaillé par âge, par sexe, par province et par revenu total n'est pas possible. Plutôt, les enregistrements en double de la BDSPS se sont vu imputer une nouvelle valeur de revenu total basée sur l'âge (six groupes), le sexe et la région par la même procédure qui a été décrite dans une section subséquente (**imputation stochastique des données de l'impôt sur le revenu**). Pour chacune des province, cette nouvelle valeur imputée du revenu total a été utilisée comme clé pour le tri des enregistrements de la BDSPS avant la fusion des pseudo-enregistrements de microdonnées du Livre vert qui ont été agrégés et triés de la même façon.

Afin d'améliorer l'appariement relativement à la situation fiscale, au revenu total, à la province, au sexe et à l'âge, il faudrait disposer d'un échantillon original de l'EDTR beaucoup plus gros.

Imputation des données historiques d'assurance-emploi

Le régime d'assurance-emploi (a.-e.) est un régime complexe dont l'administration exige le contrôle des activités hebdomadaires des prestataires sur le marché du travail. Les données administratives recueillies dans le cadre du programme servent (i) à assurer le suivi des activités hebdomadaires des prestations et de demandes de prestations des prestataires d'a.-e., (ii) à déterminer l'admissibilité et les droits par la surveillance de la participation antérieure au programme dans le cas de demandes répétées ou renouvelées et (iii) à surveiller les antécédents d'emploi par l'intermédiaire du « dossier d'emploi ». En outre, de nombreuses modifications ont été apportées au programme, notamment le remplacement de l'assurance-chômage par l'assurance-emploi. L'une des répercussions les plus importantes a été que l'admissibilité au programme, auparavant fondée sur le nombre de semaines de travail, était dorénavant fondée sur le nombre d'heures de travail. Dans le MSPS, nous voulons pouvoir modéliser l'un et l'autre programme.

Les prestations d'a.-e. constituent un élément important du revenu imposable autant que du revenu disponible. Les prestations déclarées et simulées servent à indiquer les coûts du programme, la taille de la clientèle ainsi que les gagnants et les perdants dans d'autres structures de programme. Pour permettre l'analyse uniforme ainsi que l'entrée au module d'impôt sur le revenu, le versement des prestations est nécessaire pour une année civile plutôt que pour une demande de prestations. Par conséquent, la tâche initiale de la construction de cet élément de la base de données a exigé l'élaboration simultanée du module de simulation d'a.-e. et l'identification d'un ensemble restreint de variables d'a.-e. pertinent au programme (table 2) qui pourrait servir comme entrée du module de simulation d'a.-e.

Étant donné que le nombre de personnes qui touchent des prestations d'a.-e. ou d'a.-c. variera durant la période de référence du MSPS, ce modèle doit être doté d'un mécanisme permettant d'augmenter et de diminuer le nombre de prestataires en sus des changements entraînés par les petites modifications apportées aux programmes. À cette fin, on procède à une régression pour déterminer lesquelles des personnes qui n'ont pas reçu des prestations d'a.-e. sont les plus susceptibles d'en recevoir à l'avenir. Pour permettre une augmentation de 80 % du nombre de prestataires d'a.-e., on fera des copies des enregistrements des candidats les plus probables et on leur imputera des antécédents d'a.-e. Pour plus de renseignements, voir les parties de ce guide qui portent sur la conversion et le fractionnement de la base de données.

FICHER DONNEUR DE L'A.-E.

Les dossiers historiques de l'administration de l'a.-e. imputés à la BDSPS étaient basés sur un échantillon de 10 % des dossiers administratifs de la population ayant une certaine activité de demandes de prestations d'a.-e. au cours de l'année de base.

L'échantillon comprend plus de 200 000 individus et représente près de 250 000 demandes de prestations. Les individus dans la BSSPS peuvent avoir jusqu'à 2 demandes par année civile. L'information retenue de ce fichier assure la confidentialité des données tout en étant suffisamment riche pour permettre la saisie des données historiques sur la population active pertinentes à l'application des règlements du programme d'a.-e. La liste suivante donne un ensemble de variables employées comme entrées au modèle d'a.-e.

Variables historiques a.-e.

Numéro de la demande de prestations (1er ou 2e dans l'année en cours)
 Drapeau de réitérant
 Type de prestations initiales
 Type de prestation principale
 Drapeau de changement de type
 Semaines de prestations (demande courante)
 Heures de travail (avant la demande courante)
 Semaines de dénominateur minimum
 Gains hebdomadaires moyens (avant la demande)
 Semaine où la demande a été produite
 Semaines de prestations d'a.-e. pour chaque année au cours des cinq ans précédant la Réclamation (une ou deux demandes de prestations)
 Taux hebdomadaire de prestations en vigueur
 Taux de chômage local
 Drapeau prestataire ayant épuisé ses prestations d'a.-e.
 Semaines de prestations de formation
 Taux de semaines de prestations de formation
 Semaines d'autres prestations (autres prestations incluent les éléments tels que la création d'emploi ou les prestations pour travail partagé)
 Autres taux de prestations hebdomadaire
 Drapeau d'entrant et de ré-entrant
 Prestations parentales reçues

Nous ajoutons à cette liste une variable de semaines de travail avant la présentation de la demande de prestations. Cette variable n'est pas disponible dans le fichier administratif de l'a.-e. puisqu'elle était nécessaire seulement pour calculer les prestations d'a.-c. On estime cette variable en imputant une variable « heures de travail hebdomadaires » à la demande de prestations puis en l'utilisant ainsi que le nombre total d'heures travaillées avant la présentation de la demande de prestations pour calculer le nombre de semaines travaillées avant la présentation de la demande.

Chaque enregistrement de la BDSPS dans lequel un revenu d'a.-e. durant l'année civile était déclaré ou qui était étiqueté comme correspondant à un prestataire futur éventuel a été apparié par catégorie à quatre prestataires sélectionnés à partir de l'échantillon de 10 % des prestataires de l'a.-e. Les clés d'appariement sont l'âge du demandeur, le type de prestation, le nombre total de prestations durant l'année de référence, la province et le sexe.

Les types de demande sont un élément important dans l'appariement, étant donné les différences importantes qui existent actuellement entre les types de demande en ce qui a trait aux règles d'admissibilité et aux droits. On a construit une classification des types de demande de prestations fondée sur l'ensemble de données de l'EDTR en :

- (i) identifiant les prestataires d'a.-e. dont la profession était codée comme « chasse, pêche, piégeage » (prestations de pêche),
- (ii) identifiant les prestataires d'a.-e. de sexe féminin ayant un enfant âgé de 0 à 1 an, (prestations de maternité),

- (iii) identifiant les prestataires d'a.-e. qui fréquentent l'école (prestations de formation).

Il a été impossible d'établir une distinction entre les autres types de prestations dans le fichier de données de l'EDTR.

APPARIEMENT PAR CATÉGORIE

L'appariement a commencé par le partitionnement des fichiers de données administratives donneur (a.-e.) et hôte (EDTR) par groupe d'âge, province, sexe et type de demande. La duplication des enregistrements au sein de ces cellules a été effectuée afin de faire en sorte que les cellules correspondantes des fichiers de données a.-e. et EDTR aient un nombre égal d'enregistrements. Si, dans toute cellule donnée, le nombre d'enregistrements hôtes dépassait le nombre d'enregistrements d'a.-e., alors les enregistrements a.-e. étaient uniformément copiés (les données d'a.-e. étaient un simple échantillon aléatoire). Ce processus est fait deux fois, une fois pour les enregistrements de l'EDTR qui reçoivent l'a.-e. (ou qui ont été converti pour devenir des prestataires d'a.-e), et une fois pour les enregistrements de l'EDTR qui ont été étiquetées en tant que prestataires d'a.-e. potentiels.

Le résultat des étapes d'appariement de cellules et de duplication a été une augmentation du nombre d'enregistrements représentant la population des prestataires d'assurance-emploi. Au départ, le fichier de données EDTR contenait environ 7 500 enregistrements de ce genre, tandis que, après la duplication, il y avait environ 30 000 enregistrements qui étaient prestataires de l'a.-e. et 22 000 enregistrements supplémentaires qui pourraient potentiellement être prestataires de l'a.-e. dans le futur.

À l'intérieur des cellules, les enregistrements appariés hôte et d'a.-e. ont été identifiés comme les enregistrements dont le rang correspondait dans les deux fichiers de données. Les enregistrements ont été mis dans l'ordre des prestations d'a.-e. reçues (en dollars).

Duplication des ménages

Il y a eu duplication des enregistrements de ménages de l'EDTR dans trois situations. Ce sont : (1) l'imputation des données d'imposition des gagnants à salaire élevé, (2) l'appariement par catégorie des données d'a.-e. et (3) la création d'un groupe synthétique de personnes âgées en établissement. Ce dernier groupe a été « créé » parce que le cadre d'échantillonnage sous-jacent du fichier de données hôte, l'EDTR, exclut la population des personnes en établissement, et parce que les personnes âgées constituent la partie la plus grande et la plus touchée par les politiques de cette population exclue.

Dans le cas des données d'a.-e. ou d'imposition, la raison de la duplication des ménages était d'utiliser autant que possible la richesse et la variété des fichiers de microdonnées administratives donneurs dans toute la mesure du possible. La duplication, ou copie, des fichiers hôtes de l'EDTR fournit la base pour l'absorption de cette variété dans les fichiers de données donneurs. Il faut noter que, dans ces deux cas, les copies d'individus sont faites en premier. Par la suite, les autres individus de leur ménage sont aussi copiés. Dans les cas où il

y a plus d'un membre d'un même ménage qui est copié (p. ex., si plusieurs membres du ménage ont reçu des prestations d'a.-e.), la duplication supplémentaire est nécessaire si l'on veut que chaque individu soit correctement représenté. La duplication, plutôt que le changement des poids individuels, est nécessaire si les poids de tous les membres du ménage doivent demeurer les mêmes.

Enfin, on a créé un pseudo-échantillon des personnes âgées en établissement. Cela s'est fait simplement par la duplication des enregistrements des personnes âgées seules qui ne sont pas en établissement (âgées de 65 ans et plus) et qui ne font pas partie de la population active. La raison de la sélection de cette population donneuse est que ces individus sont plus susceptibles de ressembler à la population en établissement. Les poids de ces enregistrements ont été ajustés de façon à refléter les comptes obtenus du recensement pour ce qui est de la population en établissement par âge, sexe et province. Lorsque l'année de base n'est pas celle d'un recensement, le recensement le plus près est utilisé et une proportion de la population en établissement de ce recensement est appliqué aux données de l'EDTR.

Imputation stochastique des renseignements de l'impôt sur le revenu

La présente section décrit l'imputation stochastique, la méthode utilisée pour l'attribution des renseignements de l'impôt sur le revenu des particuliers aux enregistrements de la BDSPS. Les renseignements, dans ce cas, diffèrent de l'information qui a fait l'objet d'un appariement utilisé pour l'amélioration de la représentation des personnes à revenu élevé. Dans le premier cas, l'information ajoutée provenait principalement des revenus par source. Dans ce cas, l'information ajustée provient principalement de divers crédits d'impôt ainsi que de diverses exemptions et déductions détaillées qui sont nécessaires au calcul des impôts sur le revenu à payer. La liste suivante a été extraite du Livre vert et a été entrée dans la BDSPS version 15.0. Ce sont les postes qui ne sont pas bien représentés dans l'EDTR (p. ex., les gains en capital), qui sont absents complètement (comme les frais financiers) ou qui ne sont pas facilement modélisables (p. ex., les déductions pour personnes handicapées). Seulement les éléments qui ont un échantillon assez grand, avec habituellement au moins 100 observations, sont imputés.

1. Montant imposable des dividendes de sociétés canadiennes imposables (120)
2. Revenus de placements autres que dividendes
3. Déduction pour RPA (207)
4. Déduction pour RÉER (208)
5. Frais de garde d'enfants (Formulaire T778, ligne 214)
6. Perte au titre d'un placement d'entreprise (217)
7. Frais de déménagement (219)
8. Pension alimentaire payée (220)
9. Frais financiers et frais d'intérêt (221)
10. Frais d'exploration et d'aménagement (224)
11. Autres dépenses d'emploi (229)
12. Autres déductions du revenu total (232)
13. Déduction pour prêts à la réinstallation d'employés (248)
14. Déductions pour options d'achat de titres (achat d'actions) (249)
15. Pertes comme commanditaire d'autres années (251)
16. Pertes autres que des pertes en capital d'autres années (252)
17. Pertes en capital nettes d'autres années (253)
18. Déduction pour gains en capital (254)
19. Déductions pour les habitants de régions éloignées (255)
20. Déductions supplémentaires du revenu net (256)
21. Montant pour personnes à charge âgées de 18 ans ou plus et ayant une déficience (306)
22. Montant pour aidants naturels (315)
23. Montant pour personnes handicapées (316)
24. Montant pour personnes handicapées transféré d'une personne à charge autre que votre conjoint (318)
25. Intérêts payés sur vos prêts étudiants (319)
26. Frais de scolarité (320)
27. Mois aux études à temps partiel (321)
28. Mois aux études à temps plein (322)
29. Frais médicaux bruts (330)
30. Gains en capital imposable moins les déductions de gains en capital sur les dons d'immobilisations (339)
31. Dons de charité (340)
32. Dons de biens culturels ou écosensibles (342)
33. Contributions politiques fédérales totales (409)
34. Crédit d'impôt à l'investissement (412)
35. Crédit d'impôt relatif à un fonds de travailleurs (414)
36. Report d'impôt minimum (427)
37. Impôt payé à un pays étranger non commercial (431)
38. Revenu non commercial étranger net (433)
39. Remboursement de la TPS/TVH (457)
40. Crédit d'impôt étranger provincial
41. Contributions politiques provinciales totales

42. Total du loyer payé au Manitoba et en Ontario (6110)
43. Impôt foncier net payé au Manitoba et en Ontario (6112)
44. Résidence collégiale d'étudiants en Ontario, aide fiscale aux propriétaires-occupants au Manitoba (6114)
45. Taxes scolaires versées au Manitoba (6122)

Ces éléments, combinés à d'autres dispositions qui peuvent être facilement calculées à partir des données disponibles (p. ex., les exemptions personnelles), permettent le calcul complet du revenu imposable et de l'impôt à payer.

Dans le cas des personnes au revenu élevé dont le revenu a été remplacé par une moyenne tirée du Livre vert, certaines déductions seront remplacées par une moyenne tirée du Livre vert. Pour plus de détails, voir la section portant sur le revenu élevé.

LES DONNÉES DU DONNEUR

Les données de source pour l'imputation ont été calculées à partir d'un échantillon des déclarations de revenu des particuliers de l'année de base, de l'Agence du revenu du Canada. L'échantillon est stratifié par source de revenu, le lieu de résidence, la situation fiscale et par plage de revenu. L'échantillon inclus les strates supplémentaires pour les salariés qui ont un revenu total plus grand que 250 000 \$, les cas particuliers et les non résidents.

L'information de cet échantillon comporte la plus grande partie de l'information qu'il y avait dans les déclarations de revenu des particuliers fédérales et provinciales de l'année de base ainsi que dans les annexes qui les accompagnaient. Cet échantillon n'a aucune structure de famille explicite (c.-à-d. que les déclarations du chef de famille, du conjoint et des personnes à charge ne peuvent être analysées ensemble à l'intérieur d'une unité familiale identifiable).

TRANSFORMATION DES DONNÉES

Pour établir le lien entre les données de l'impôt sur le revenu du Livre vert et l'échantillon hôte découlant de l'EDTR, on a défini un ensemble de caractéristiques de classification communes. Les attributs suivants ont été choisis autant pour leur degré de pertinence à la politique que pour leur disponibilité et la similarité de leurs définitions dans les deux fichiers de données :

1. Province d'imposition
2. Groupe d'âge
3. Sexe
4. Situation matrimoniale aux fins d'impôt
5. Groupe de revenu
6. Groupe de salaire
7. Nombre d'enfants
8. Taille des cotisations syndicales

Les sous-échantillons définis par la classification mixte de ces postes sont supposés avoir suffisamment de distributions différentes pour se mériter le maintien du caractère unique de ces distributions. Par exemple, à la figure 3, une comparaison est fournie de la distribution de

la taille des dons de charité pour deux groupes de revenus différents en 2003.

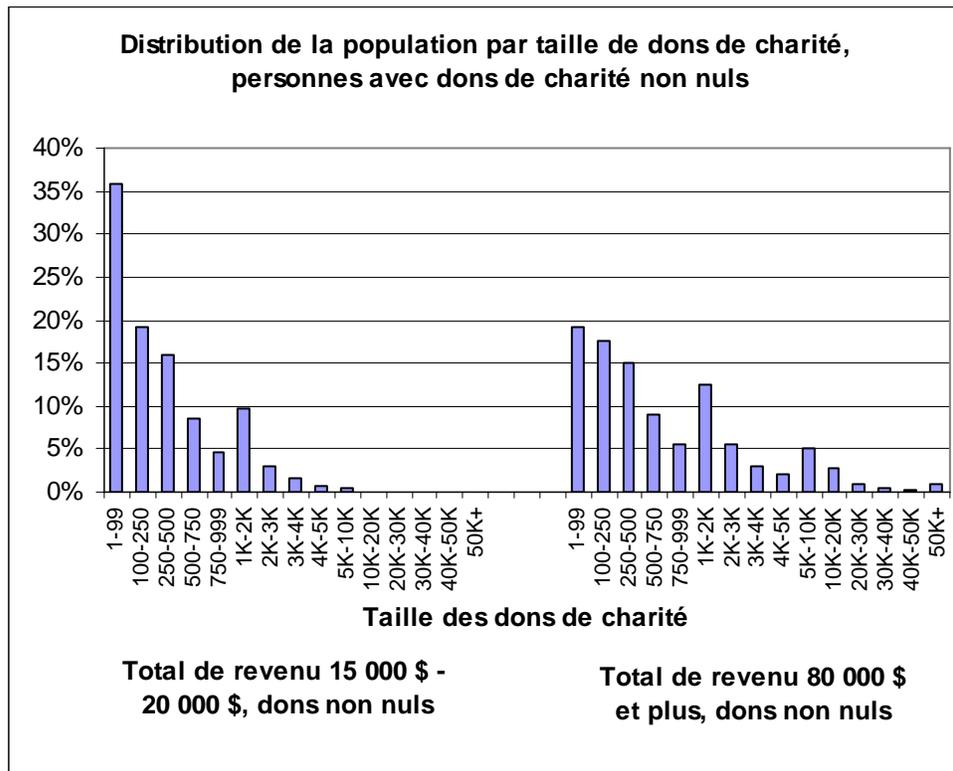


Figure 3 : Distribution des dons de charité selon le Livre vert en 2003

Avant l'imputation, le fichier de données hôte a été préparé par l'identification des déclarants possibles, l'établissement de l'admissibilité pour certains postes visés (p. ex. déductions pour frais d'éducation, de scolarité et de garde d'enfants) et la création d'un mode de classification parallèle à l'intérieur des fichiers de données hôte, la BDSPS, et donneur, le Livre vert.

Pour certains éléments de déductions, il était possible d'identifier l'admissibilité de l'hôte et parfois celui des données du donneur. Voici quelques exemples :

Déduction pour gains en capital	Revenu de gains en capital reçus
Déduction pour frais de garde d'enfants	Présence d'enfants
Déduction pour pension alimentaire payée	N'ayant pas reçu de revenu de pension alimentaire
Montant relatif aux études, pour soi	A fréquenté l'école
Programme de réinstallation pour employés	A touché un traitement ou salaire
Déduction pour résidence collégiale d'étudiants en Ontario / aide fiscale aux propriétaires-occupants au Manitoba	Étudiants à temps plein dans un établissement post-secondaire en Ontario ou un propriétaire au Manitoba
Revenu étranger (en tant que pourcentage du revenu total)	Revenu total doit être non nul
Impôt étranger payé en tant que pourcentage du revenu étranger	Revenu étranger reçu

En visant l'imputation aux individus admissibles à ces déductions, on fait en sorte qu'il y ait un certain degré d'uniformité interne dans les enregistrements synthétiques. Par exemple, seules les personnes ayant des enfants se verront imputer la déduction pour frais de garde d'enfants. Malheureusement, il n'est pas aussi simple de déterminer l'admissibilité à toutes les déductions et à tous les postes de revenu imputés.

La distribution commune des cotisations à des REP (régime enregistré de pension) et à des REER (régime enregistré d'épargne-retraite) a posé un problème du fait que la Loi de l'impôt empêche que le total des deux dépasse une certaine limite. L'imputation distincte de ces deux valeurs ne ferait pas en sorte que ce seuil ne soit pas dépassé. Pour contourner cette difficulté, nous avons imputé séparément les REP, les REER pour les individus sans cotisations au REP, ainsi que les REER pour les individus avec des cotisations au REP.

CALCUL DES STATISTIQUES DE DISTRIBUTION

Un objectif de ce processus d'imputation vise à faire en sorte que le montant moyen des divers crédits ainsi que des diverses déductions et exemptions réclamés dans la BDSPS reflètent avec précision les moyennes réelles (p. ex., publiées) pour les sous-groupes définis, par exemple, par province, par âge, par plage de revenus, etc. Un objectif supplémentaire, plus rigoureux, est que la BDSPS reproduise la distribution de ces postes de la façon dont ils se trouvent dans le fichier du Livre vert. Cela exige une méthode permettant de représenter des fonctions de densité arbitraires. Ainsi, la méthode devrait représenter aussi bien les distributions bimodales, tronquées et non tronquées.

Un autre facteur qui a prévalu dans le choix de la méthode était l'intensité des calculs à effectuer. Comme le fichier de données de source contient plus de 400 000 enregistrements, l'algorithme devant servir à générer ces représentations devait être raisonnablement efficace.

La méthode que l'on a finalement choisi consistait d'abord à désagréger la population

globale de façon hiérarchique en utilisant les variables de classe indiquées ci-dessus. Par la suite, à l'intérieur de chacun de ces sous-groupes définis de façon hiérarchique, les distributions à variable unique des postes particuliers étaient représentées en premier par la proportion dans tout sous-groupe donné avec une valeur non-zéro pour le poste. Ensuite, pour le sous-sous-groupe avec des valeurs non-zéro, la fonction de densité a été représentée par les seuils déciles, un traitement particulier étant donné aux extrémités de la courbe de distribution.

Une contrainte a été imposée à la procédure de désagrégation hiérarchique afin d'assurer la non-confidentialité des statistiques ainsi obtenues. Cette contrainte consistait à exiger un nombre minimal d'observations dans chacun des sous-groupes ou des sous-sous-groupes. Afin d'utiliser le plus possible les données, le processus de désagrégation a été appliqué de façon indépendante pour les statistiques de pourcentage et de distribution (c.-à-d. déciles). Les statistiques de pourcentage pourraient être basées sur un nombre beaucoup plus petit d'observations que les seuils de déciles, de façon que l'on puisse utiliser l'information d'un degré de désagrégation beaucoup plus fin.

Les statistiques de pourcentage ont été conservées si la somme des poids de la cellule dépassait 400 et si le nombre d'enregistrements ayant une valeur non-zéro dépassait 20. Si ces critères n'étaient pas respectés, on leur substituait les statistiques d'un niveau d'agrégation supérieur.

Les critères de statistiques de distribution devaient être plus rigoureux. La taille minimale des cellules était de 100 enregistrements, c'est-à-dire que, si une cellule qui ne contenait pas au moins 100 enregistrements non-zéro, les statistiques de cette cellule n'étaient pas calculées. Les statistiques de distribution étaient plutôt calculées pour un niveau d'agrégation supérieur.

Pour chaque poste à imputer (tous ceux qui se retrouvent sur la liste du début de la présente section), les quelque 400 000 enregistrements de déclarations de revenus ont été classifiés dans des cellules pertinentes (p. ex., groupe de revenus par âge, situation matrimoniale, sexe, province).

Pour chacun de ces groupes, si l'échantillon était suffisant, les statistiques suivantes étaient calculées :

- les valeurs pour les seuils de déciles 1 à 9;
- la moyenne des déciles supérieur et inférieur;
- la moyenne des cinq valeurs les plus élevées et la moyenne des cinq valeurs les plus basses; et
- le pourcentage à l'intérieur de la cellule donnant une valeur non-zéro pour le poste.

Ces statistiques sont bien adaptées pour représenter une distribution arbitraire et elles sont simples à calculer.

Pour des raisons de confidentialité, les valeurs minimum et maximum réelles d'une cellule ne pouvaient pas être utilisées. La moyenne des cinq valeurs les plus élevées et la moyenne des cinq valeurs les plus basses de la cellule étaient utilisées pour les remplacer.

Les mêmes statistiques ont alors été générées par agrégation de cellules, dans ce cas, pour le groupe de revenu par âge, situation matrimoniale, sexe et région. Le regroupement des dix provinces en cinq régions rehausse le niveau d'agrégation et, par conséquent, accroît le nombre d'individus à l'intérieur d'une cellule. Il y aura alors plus de cellules à respecter le critère de taille minimale de cellule pour le calcul des ensembles de statistiques de distribution.

L'idéal serait que toutes les valeurs soient imputées depuis le niveau d'agrégation le plus bas. Cependant, à cause de la faible densité d'un grand nombre de postes de données, cela est rarement possible.

Pour remplir ces cellules vides ou peu denses, on substitue aux valeurs les statistiques du niveau d'agrégation plus élevée. Si, par exemple, la cellule représentant la classification suivante

<u>Groupe de revenu</u>	35 000 \$ à 39 999 \$
<u>Groupe d'âge</u>	25 à 35
<u>État matrimonial</u>	Personne seule, mariée aux fins de l'impôt
<u>Sexe</u>	Femme
<u>Province</u>	Québec

était vide ou refusée à cause du critère de taille de cellule, les statistiques utilisées pour remplacer la valeur serait le niveau d'agrégation suivant

<u>Groupe de revenu</u>	35 000 \$ à 39 999 \$
<u>Groupe d'âge</u>	25 à 35
<u>État matrimonial</u>	Personne seule, mariée aux fins de l'impôt
<u>Sexe</u>	Femme

représentant ce groupe de revenu, ce groupe d'âge, cet état matrimonial et ce sexe pour le Canada. Si cette cellule était elle aussi vide ou peu dense, on prendrait les statistiques du niveau d'agrégation suivant. Dans le pire cas, les statistiques d'une cellule seraient calculées à partir de l'échantillon complet, c'est-à-dire tous les groupes de revenu, tous les groupes d'âge, tous les états matrimoniaux, les deux sexes et toutes les provinces.

Les statistiques de pourcentage et de distribution ainsi obtenues ne sont pas confidentielles puisqu'elles ne révèlent jamais des valeurs de données brutes. Les valeurs extrêmes sont synthétisées par le calcul de la moyenne des cinq valeurs les plus élevées et de la moyenne des cinq valeurs les plus basses.

IMPUTATION

En utilisant cet ensemble complexe de statistiques de distribution générées à partir du fichier du Livre vert des déclarations de revenu, il est possible de recréer la même distribution de valeurs dans le fichier de données hôte. Pour chaque individu admissible du fichier de données hôte, une valeur synthétique est obtenue d'une distribution représentant les déclarations de revenu d'un groupe de gens semblables.

Les valeurs pour les huit déciles du milieu sont générées en supposant une distribution uniforme entre les seuils de déciles. (Des fonctions de densité plus complexes ont été

essayées à l'intérieur de ces déciles. Cependant, des tests ont laissé voir que le gain de précision était marginal, particulièrement en regard de la très grande augmentation des coûts computationnels.)

Les déciles du haut et du bas sont traités de façon particulière de manière que la forme et la taille des extrémités de la courbe soient représentées de façon précise. La préservation des extrémités de la courbe de distribution est essentielle au maintien des totaux et des moyennes globales, particulièrement pour des postes dont les extrémités de la courbe de distribution sont longues, comme c'est le cas pour les gains en capital ou les pertes commerciales.

En imputant les déciles du haut et du bas, les valeurs sont obtenues en supposant une distribution de Pareto pour générer des extrémités ayant la forme appropriée. La distribution de Paréto spécifique utilisé dans chaque cas est telle que la moyenne de décile est conservée. Les valeurs extrêmes sont tronquées à la moyenne des cinq valeurs les plus élevées ou des cinq valeurs les plus basses du groupe.

Imputation des données de l'Enquête sur les dépenses des ménages

Les données de dépenses des ménages visent à assurer le soutien des simulations exigeant de l'information sur les frais de logement, des simulations entourant les frais de garde d'enfants ainsi que des simulations des taxes à la consommation. Du fait du nombre limité d'enregistrements du fichier de données sur les dépenses des ménages (environ 15 000), on a décidé de faire une imputation à la structure de la consommation en utilisant le plus de données possible sur les catégories de ménage. Un nombre limité d'observations est fixé par classe.

Il y avait deux grandes étapes pour l'imputation des tendances de la consommation :

- L'analyse à variables multiples qui crée la variable d'appariement.
- L'appariement par catégorie (duplication pondérée)

DÉTERMINATION DES VARIABLES D'APPARIEMENT ET HABITUDES DE CONSOMMATION

On a élaboré une solution d'appariement basée sur une analyse à variables multiples pour l'EDM. Les ménages d'une tranche sont regroupés selon la ressemblance de leur modèle de consommation et non selon leur niveau de consommation. Les variables utilisées pour définir les modèles de consommation sont les 47 catégories de dépenses et quelques autres variables supplémentaires (p. ex. les épargnes et autres dépenses).

Les variables de classification possibles qui sont utilisées pour grouper les habitudes de consommation similaires sont : le type de ménage, la tenure, le sexe du chef du ménage, la tranche de revenu, le groupe d'âge, le travail à temps plein du chef, le travail à temps plein du conjoint, la région, la présence d'enfants d'âge préscolaire et la présence d'enfants d'âge scolaire sont tous utilisés en tant que variables de classification. L'analyse à variables multiples évalue la capacité d'explication de chaque variable de classification ainsi, les meilleures variables sont utilisées. Le processus est répété avec les variables de classification

choisies jusqu'à ce qu'aucun autre sous-groupe ne puisse être formé et qu'il contient un nombre minimum d'observations. La première variable de classification est obligatoirement le type de ménage afin de garantir la cohérence des modèles de dépenses : on s'attend à ce qu'une famille monoparentale avec deux enfants n'ait pas le même modèle de consommation que celui d'un ménage avec deux adultes et un enfant.

Voici un exemple qui sert à illustrer le processus. La première variable de classification est obligatoirement le type de ménage. Alors, pour chaque type de ménage, on exécute une régression pour déterminer lesquelles des variables de classification restantes ont la plus grande capacité d'explication. Par exemple, dans le cas des célibataires, la variable de classification qui s'applique ensuite peut être le mode d'occupation du logement tandis que dans le cas des couples mariés avec enfants, elle peut être la présence d'enfants d'âge préscolaire. Les célibataires qui sont propriétaires de leur logement peuvent ensuite être réparties selon le sexe tandis que celles qui sont locataires de leur logement peuvent être réparties selon le groupe d'âge. On continue de processus jusqu'à ce qu'une autre subdivision donnerait une case contenant un nombre d'observations inférieur au minimum, soit dans l'EDM, soit dans la BDSPS. On attribue alors la variable hdevmv aux catégories finales.

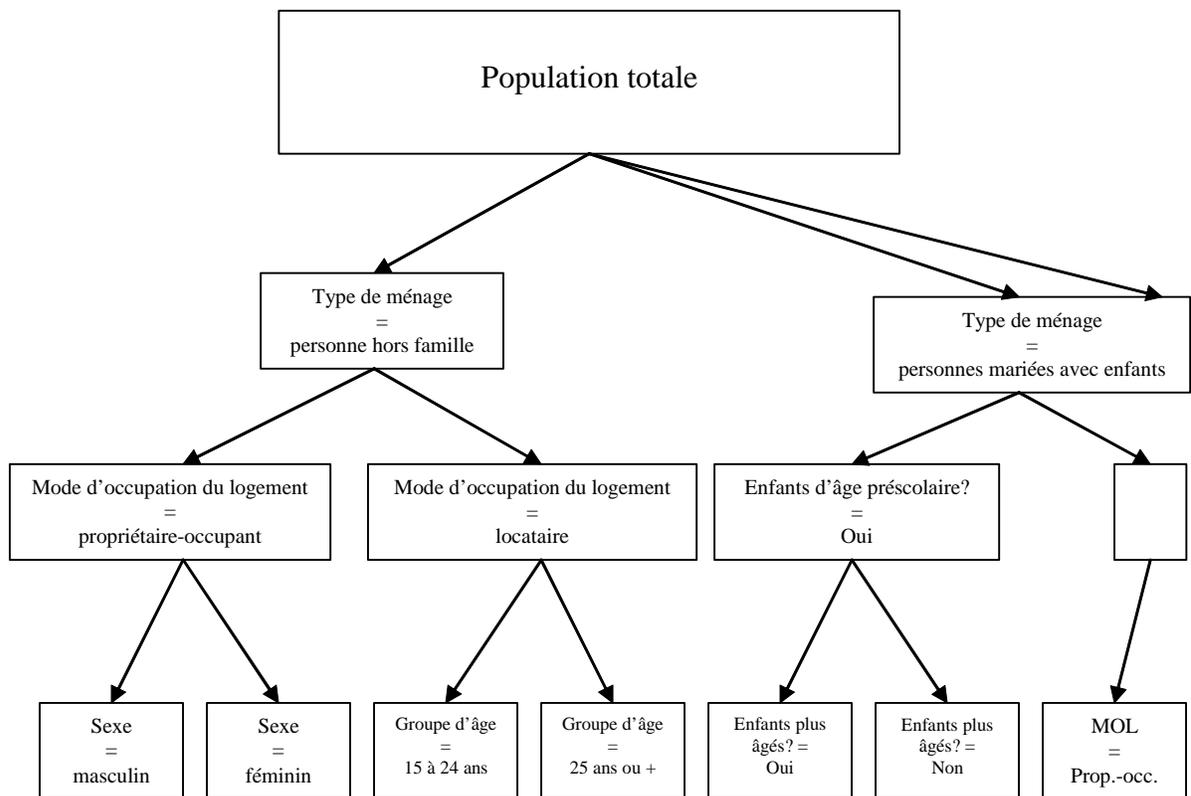


Figure 4 : Exemple indiquant comment l'appariement est créé

APPARIEMENT PAR CATÉGORIE

Une fois la clé d'appariement produite, on peut répartir les données de la BDSPS et de l'EDM dans les cases appropriées. On attribue ensuite un vecteur des dépenses établi à partir de l'EDM aux ménages compris dans la BDSPS qui figurent dans ces cases. On fait alors des copies pondérées des enregistrements de l'EDM pour s'assurer qu'un vecteur des dépenses tiré de l'EDM est attribué à chaque observation dans la BDSPS. À l'intérieur des cases, les observations figurant dans l'un et l'autre ensemble de données sont réparties selon le revenu total. En moyenne, il peut y avoir environ six fois plus d'enregistrements de la BDSPS que d'enregistrements de l'EDM. Cependant, il ne conviendrait pas de tout simplement faire cinq copies de chaque enregistrements de l'EDM parce que cela aurait pour effet réel de traiter l'EDM comme un échantillon simple plutôt qu'un échantillon aléatoire stratifié; on ne prendrait pas en compte les poids de l'échantillon de l'EDF. Plutôt, les enregistrements de l'EDM ayant un poids plus élevé sont copiés dans une proportion plus grande que ceux qui ont un poids plus faible.

Plus précisément, on calcule une probabilité pondérée d'occurrences du ménage EDM i dans la case j . En multipliant cette probabilité par la taille désirée de l'échantillon dans la case hôte, on obtient une estimation du nombre de fois qu'un ménage EDM donné devrait apparaître dans le fichier de données hôte. Si le nombre est inférieur à un, cela veut alors dire que le vecteur de dépense tiré de l'EDM n'est pas utilisé dans le processus d'appariement. Si la probabilité ainsi déterminée est simplement arrondie ou tronquée à son équivalent entier, l'erreur d'arrondissement peut produire un compte total incorrect dans la case hôte. Pour corriger cette erreur, un total cumulé des fréquences des cellules hôtes D est calculé ().

$$D_{ij} = \sum_{k=1}^i \left[\left[\frac{W_{ij}}{\sum_{i=1}^n W_{ij}} \right] \times (N_j^h - N_j^d) \right]$$

où :

i = le i^e ménage EDM

j = la j^e case d'appariement

W = le poids de l'enregistrement donneur de l'EDM

N^h = la taille de l'échantillon dans la case hôte de la BDSPS

N^d = la taille de l'échantillon dans la case donneuse EDM

Chaque enregistrement EDM est alors copié par la valeur arrondie du total cumulatif moins la valeur arrondie du total cumulatif de l'enregistrement précédent plus un. De cette façon, l'erreur arrondie est répartie à l'intérieur de la cellule, chaque enregistrement EDM se voit assurer au moins un appariement et on atteint les totaux corrects de cellules d'hôtes.

Cette procédure permet largement de préserver les distributions pondérées des données de l'EDM, au moins jusqu'à ce que les poids de la BDSPS soient associés à elles. L'écart entre les poids de l'EDTR et les poids de l'EDM peut cependant créer des distorsions dans les distributions appariées.

Autres sujets :

FRAIS DE GARDE D'ENFANTS

Les frais de garde d'enfants sont imputés à partir de l'Enquête sur les dépenses des ménages ainsi que de l'échantillon T1. L'imputation à partir de l'EDM est indépendante de l'imputation du vecteur des dépenses totales. On a procédé ainsi afin de pouvoir utiliser les variables plus appropriées pour appairer des enregistrements similaires. La méthodologie utilisée ressemble à celle utilisée pour imputer le vecteur des dépenses. On a imputé d'abord les frais de garde d'enfants de l'échantillon T1 de la même façon que les autres déductions imputées à partir du formulaire T1. Dans le cas des deux types de frais de garde d'enfants, les dépenses imputées ont été attribuées aux enfants de la famille.

IMPUTATION DES SEMAINES TRAVAILLÉES AUX DEMANDES DE PRESTATIONS D'A.-E.

Le vecteur de la situation d'a.-e. n'inclut pas le nombre de semaines travaillées durant les 52 semaines précédant la présentation d'une demande de prestations d'a.-e. Le FMGD de l'EDTR, par contre, inclut le nombre de semaines travaillées par les personnes l'année précédente. Pour imputer cette variable pour chaque demande de prestations d'a.-e., on a utilisé un processus d'appariement par catégorie pour appairer les personnes dans le FMGD de l'EDTR aux personnes ayant un fichier historique de demandes d'a.-e. dans la BDSPS. On a ensuite apparié les personnes et les prestataires en utilisant une combinaison de caractéristiques comme l'âge, la province, le sexe, l'industrie, le nombre d'heures de travail rémunérées l'année dernière et les prestations pour enfant reçues.

Références

Adler, H.J. et M.C.Wolfson (1988), « Projet-pilote de raccordement micro-macro pour le secteur des ménages au Canada », *The Review of Income and Wealth*, volume 34, n° 4.

Ruggles, R. and N. Ruggles (1986), « The Integration of Micro and Macro Data for the Household Sector », *The Review of Income and Wealth*, series 32, no. 2. 2.