Guide de l'utilisateur des microdonnées pour l'enquête transversale

L'enquête sur la dynamique du travail et du revenu (EDTR)

Année de référence: 2003

TABLE DES MATIÈRES

1. INTRODUCTION

Le fichier de microdonnées transversales à grande diffusion pour l'Enquête sur la dynamique du travail et du revenu (EDTR) est un recueil de variables des domaines du revenu, du travail et de la famille sur les personnes au Canada et leur famille. L'EDTR est une enquête auprès des ménages couvrant la population des dix provinces canadiennes à l'exception des réserves indiennes, des résidents d'institutions et des casernes militaires.

L' EDTR a commencé à recueillir des données pour l'année de référence 1993. Au départ, l'EDTR a été conçue pour être avant tout une enquête longitudinale portant principalement sur le travail et le revenu ainsi que sur les relations qu'il y a entre ces données et la composition des familles. Par la suite, il a été décidé d'étendre les objectifs de l'EDTR de façon qu'elle soit la principale source de données transversales sur les revenus des ménages.

Pendant de nombreuses années, l'Enquête sur les finances des consommateurs produisit des fichiers de microdonnées à grande diffusion (FMGD) permettant de combler les besoins des utilisateurs de données transversales sur le revenu des ménages. Les FMGD de l'EFC ont été publiés jusqu'à l'année de référence 1997. Dans le cadre de son engagement à l'endroit des principaux utilisateurs de données, Statistique Canada a promis que l'EDTR continuerait de produire des fichiers de microdonnées à grande diffusion (FMGD) permettant de combler les besoins des utilisateurs de FMGD de l'EFC. Pour ce qui est des tableaux types, Statistique Canada a décidé de passer de l'EFC à l'EDTR entre 1995 et 1996. Par conséquent, les fichiers de microdonnées transversales à grande diffusion de l'EDTR ont été publiés pour la première fois pour l'année de référence 1996. Les fichiers de l'EDTR ont été conçus pour être analogues à ceux produits pour l'EFC. Les données sur le revenu recueillies par l'EDTR étaient du même genre que celles de l'EFC, la principale exception étant que les répondants à l'EDTR avaient le choix d'une interview traditionnelle sur le revenu et la permission accordée à Statistique Canada d'utiliser les données de l'impôt tirées des T1.

Il est possible d'obtenir plus d'information sur la comparabilité de l'EDTR et de l'EFC dans les ouvrages D'une enquête à l'autre : une série intégrée de données sur le revenu de l'EFC et de l'EDTR, 1989-1997, ou Comparaison des résultats de l'Enquête sur la dynamique du travail et du revenu (EDTR) et de l'Enquête sur les finances des consommateurs (EFC) 1993-1997 (voir aussi Produits et services connexes).

Comment citer l'EDTR dans les publications

Lors de la publication de tous renseignements basés sur ces fichiers de microdonnées sur CD-ROM de l'EDTR (75M0010XCB), nous recommandons la citation suivante:

«Cette analyse est fondée sur les Microdonnées à grande diffusion de l'Enquête sur la

dynamique du travail et du revenu de Statistique Canada, qui contiennent des données anonymes de l'Enquête sur la dynamique du travail et du revenu. Tous les calculs effectués à l'aide de ces microdonnées sont la responsabilité de (Nom de l'utilisateur). L'utilisation et l'interprétation de ces données sont uniquement la responsabilité des auteurs».

2. STRUCTURES DES FICHIERS

Bien qu'il soit souvent désigné comme un seul fichier, le FMGD de l'EDTR comporte quatre fichiers distincts : CLÉ, PERSONNE, FAMILLE ÉCONOMIQUE et FAMILLE DE RECENSEMENT. Les structures de fichier utilisées pour les FMGD de l'EFC ont été en grande partie conservées.

Dans le fichier CLÉ, il y a un enregistrement par personne de l'échantillon et les indicateurs nécessaires afin de l'apparier aux fichiers personne, familles économiques et familles de recensement.

Dans le fichier PERSONNE, il y a un enregistrement par personne de 16 ans et plus de l'échantillon. Les caractéristiques d'emploi, comme l'industrie, le taux de traitement et l'horaire de travail, sont comprises dans le fichier personne. Ces caractéristiques d'emploi ont trait à l'emploi principal de la personne pendant l'année de référence (l'emploi auquel la personne a consacré le plus d'heures de travail pendant l'année). Bien que l'EDTR recueille des données sur tous les emplois occupés par chaque personne âgées de moins de 70 ans pendant l'année, les caractéristiques de tous les autres emplois ne sont pas comprises dans le FMGD de l'EDTR. Le fichier PERSONNE contient des identificateurs qui permettent à un chercheur de regrouper les personnes en ménages, en familles économiques et en familles de recensement.

En 2003, il n'y pas eu de changement apporté à la structure des fichier du FMGD par rapport à l'année de référence précédente.

Les fichiers à grande diffusion de 2003 ont les tailles suivantes :

Fichiers	Nombre d'enregistrements	Nombre de variables	Longueur de l'enregistrement
Fichier personne	57,233	130	491
Fichier famille économique	29,846	67	371
Fichier famille de recensement	33,461	64	368
Fichier clé	71,418	14	45

3. COMMENT SE SERVIR DU CLICHÉ D'ENREGISTREMENT, DU DICTIONNAIRE DE DONNÉES ET DES DISTRIBUTIONS À UNE VARIABLE

Trois autres fichiers sont fournis pour aider les utilisateurs du fichier à grande diffusion (FGD). Pour chacun des quatre fichiers de données (clé, personne, famille économique et famille de recensement) un cliché d'enregistrement, un dictionnaire de données et des distributions à une variable sont fournis. Ces fichiers sont organisés par thèmes et, dans certains cas, par sous-thèmes.

A. Les colonnes du cliché d'enregistrement sont les suivantes:

Nom de variable. Il s'agit du nom de variable dans le fichier de microdonnées.

Type. Le type indique si la variable est numérique (utilisable dans les opérations mathématiques) ou de type caractère.

Séquence. Ordre d’apparence des variables.

Position de début. Il s'agit de l'emplacement de la variable dans le fichier à grande diffusion.

Longueur. Désigne à la fois le nombre d'espaces et le nombre de décimales, le cas échéant. Ainsi, le format d'une variable qui peut avoir une valeur allant de zéro (00,0) à 99,9 sera exprimé de la façon suivante : 4,1. Le format d'une variable qui peut avoir une valeur allant de zéro (00) à 99 sera exprimé de la façon suivante : 2.

Nombre de catégories. Il s'agit du nombre de catégories que renferme l'ensemble des valeurs relatives à la variable en question. Cette colonne s'applique uniquement aux variables de type « caractère ». Les variables numériques comportent des intervalles, qui sont précisés dans le dictionnaire des données.

Long nom de variable. Un nom normalisé comporte au maximum 26 caractères et peut être utilisé pour identifier rapidement les variables, étiqueter les tableaux, etc. Tout en étant encore passablement cryptiques, ces caractères sont considérablement plus révélateurs que le nom de la variable. Toutefois, ce nom plus long exclut évidemment beaucoup de renseignements importants compris dans la description de la variable figurant dans le dictionnaire de données. En résumé, les analystes doivent être prudents lorsqu'ils font des hypothèses concernant la définition de la variable en se basant sur le long nom de variable.

B. Dictionnaire des données

Le dictionnaire des données comprend des renseignements complets au sujet de chaque

variable de l'enquête sur les trois fichiers. Il fournit, pour chacune d'elles : le nom ainsi que la description ou la définition de la variable, des listes de codes avec des descriptions ou encore la gamme de valeurs qui peuvent être attribuées à la variable, le type de variable, sa longueur (ou son format), et la population à laquelle elle se rapporte, c'est-à-dire à laquelle elle est applicable.

C. Distributions à une variable

Ces distributions sont fournies aux utilisateurs des fichiers de microdonnées à grande diffusion afin qu'ils puissent vérifier leur totalisation. Ces distributions se rapportent au fichier à grande diffusion et non à la base de données interne; les distributions sont semblables mais non identiques.

Pour les variables caractères, les fréquences pondérées et non pondérées pour chaque code, incluant les codes réservés, sont fournis. Pour les variables numériques, les valeurs sont divisées en de nombreuses tranches et les fréquences pondérées et non pondérées sont fournies pour chaque tranche. Les valeurs minimums, les valeurs maximums ainsi que la moyenne pondérée sont aussi fournies à l'exception des codes réservés.

Valeurs manquantes et codes réservés

Dans le cadre de l'EDTR des codes réservés ayant une signification particulière ont été adoptés. Il est important de porter attention à ces codes réservés notamment avec les variables numériques. Si vos calculs donnent des résultats qui semblent trop élevés, vérifiez que les codes réservés non pas été inclus dans vos calculs. Les codes réservés, à quelques exceptions près, sont les valeurs les plus élevées qu'une variable peut prendre. Les codes réservés font l'objet d'une brève explication ci-dessous.

Si le champ de la variable ne s'étend pas à un sous-groupe particulier de population, il n'y a pas de valeur valide pour ce sous-groupe, et les valeurs fournies prennent la forme suivante : 9, 99, 9,9 et ainsi de suite, ce qui indique que la variable ne s'applique pas. La population admissible pour chaque variable du fichier est énoncée dans le dictionnaire des données.

Des valeurs peuvent être absentes de certains enregistrements, du fait qu'aucune valeur valide n'est disponible, même si la variable s'applique. Il se peut que le répondant n'ait pas fourni les renseignements, ou encore que ceux-ci aient été rejetés en cours de traitement, et que la valeur n'ait pas été imputée. Ces valeurs manquantes apparaissent avec un code comme 7, 97, 9,7 et ainsi de suite, selon le format. Pour certaines variables, le nombre de valeurs manquantes a été réduit au moyen de l'imputation. Les valeurs manquantes relatives aux variables sur le revenu ont été entièrement imputées, mais la plupart des autres variables comportent des valeurs manquantes.

Le traitement des valeurs manquantes de cette dernière catégorie dépend du type d'analyse effectué et de la portée des données manquantes. Même si la solution finale pourrait consister à exclure de l'analyse les enregistrements auxquels il manque des

valeurs, on devrait tout d'abord procéder à un examen pour évaluer les répercussions des valeurs manquantes sur la représentativité globale des données. Se peut-il qu'un biais découle des données manquantes, par exemple, les (autres) caractéristiques des personnes pour lesquelles il manque des valeurs diffèrent-elles de celles de la partie observée de l'échantillon? Il peut être nécessaire de tenir compte, d'une façon ou d'une autre, des répercussions possibles. Dans tous les cas, lorsque les analystes publient leurs résultats, ils devraient indiquer pour quelles variables les enregistrements qui ont des valeurs manquantes ont été exclus.

Enfin, on aura attribué à quelques valeurs le code 8, 98, 9,8, etc. Il s'agit des refus de répondre à certaines questions de l'interview.

4. LIGNES DIRECTRICES POUR L'APPLICATION DES POIDS

Les microdonnées des fichiers à grande diffusion ne sont pas pondérées. Il est du ressort des utilisateurs des données d'appliquer les poids appropriés compte tenu des résultats qu'ils veulent produire. Si l'on n'applique pas les poids appropriés, les estimations effectuées à partir des microdonnées ne peuvent être considérées comme représentatives de la population observée, et ne correspondront pas à celles que produirait Statistique Canada. On retrouve les poids sous la variable «contrôle de l'échantillon». Dans le guide de l'utilisation des microdonnées de l'enquête transversale de l'EDTR, la variable de poids est ICSWT26.

5. LIGNES DIRECTRICES POUR LA DIFFUSION (QUALITÉ DES DONNÉES ET ARRONDISSEMENT)

Les utilisateurs de microdonnées devraient appliquer les règles d'évaluation de la qualité des données figurant ci-dessous à toutes les estimations qu'ils produisent et ne devraient retenir que celles qui répondent aux critères s'appliquant à la diffusion. Les estimations qui ne répondent pas à ces critères ne sont pas fiables.

Introduction
Les lignes directrices pour la diffusion et la publication s'appuient sur le concept de la "variabilité d'échantillonnage" afin de déterminer si les estimations tirées des microdonnées sont fiables. La variabilité d'échantillonnage peut être définie comme l'erreur dans les estimations qui découle du fait qu'on effectue l'enquête auprès d'un échantillon plutôt que de l'ensemble de la population. Le concept de l' «écart-type» et les mesures connexes du «coefficient de variation» et de
l' «intervalle de confiance» fournissent une indication de la taille de la variabilité d'échantillonnage.

L’écart-type et le coefficient de variation ne servent pas à mesurer les biais systématiques des données d'enquêtes qui pourraient avoir des répercussions sur les estimations. Ils sont plutôt fondés sur l'hypothèse que les erreurs d'échantillonnage suivent une distribution

normale de probabilités.

Sous réserve de cette hypothèse, il est possible d'estimer dans quelle mesure les divers échantillons qui ont le même plan et le même nombre d'observations pourraient aboutir à des résultats différents. Cela donne une idée de la marge d'erreur susceptible d'être comprise dans les estimations dérivées de notre échantillon unique.

Pour une description détaillée des mesures de la variabilité d'échantillonnage, voir A. Satin et W. Shastry, L'échantillonnage: un guide non mathématique, Statistique Canada, produit no 12-602F au catalogue.

Taille minimum des estimations destinées à la diffusion

Les seuils de suppression, ou les mesures de qualité des données, sont établis en se basant sur la taille de l'échantillon à partir duquel les estimés sont produits. De façon générale, un échantillon composé d'au moins vingt-cinq observations est requis pour que l'estimé soit publiable. Le seuil de suppression peut varier légèrement selon le type d'estimés produits. Ces seuils nous permettent d'assurer la confidentialité des répondants et la qualité des données.

Seuils de suppression

ESTIMÉ	SUPPRIMÉ SI:
Pourcentage, distribution, proportion/part :	
% sous le seuil de faible revenu (SFR) Distribution du revenu Proportion des familles ayant un revenu =0	Dénominateur* taille de l'échantillon < 25 ou Dénominateur* taille de l'échantillon < 100 et numérateur de la taille de l'échantillon < 5
Ratios:	
Gains femmes/hommes	Numérateur de la taille de l'échantillon < 25 ou Dénominateur de la taille de l'échantillon < 25
Quintiles (parts, moyennes et limites supérieures du revenu)	
Part du revenu par quintile Moyenne du revenu par quintile Limites supérieures du revenu	Taille de l'échantillon de tous les quintiles /5 < 25 ou Limites supérieures du revenu pour le quintile de revenu supérieur ou l'ensemble des quintiles
Autres mesures	
Comptes Moyennes Médianes Coefficients de Gini	Taille de l'échantillon < 25

*La taille de l'échantillon du dénominateur réfère à la taille de l'échantillon de la population totale à partir duquel la distribution, les pourcentages, les proportions ou les parts sont dérivés.

Estimation d’agrégats et de moyennes pour les provinces

Lors du calcul d’estimés pour les agrégats et les moyennes au niveau provincial, il importe de noter que pour un nombre restreint d’enregistrement la province de résidence a été supprimée. Cela devrait engendrer un léger biais dans les estimés provinciaux.

Lignes directrices pour l'arrondissement

Une fois qu'il a été déterminé que les résultats obtenus sont fiables, le niveau d'arrondissement correspond au niveau de précision des données. Les lignes directrices qui suivent devraient être utilisées pour l'arrondissement :

- Les estimations de sous-groupes de population devraient être arrondies à la centaine près ;

- Les taux et les pourcentages devraient être arrondis à la décimale près.

- Il convient de souligner que tous les calculs doivent être faits à partir d'éléments non arrondis, puis arrondis au moyen de la technique d'arrondissement classique.

- Dans la technique d'arrondissement, si le premier ou le seul chiffre à supprimer se situe entre 0 et 4, le dernier chiffre à conserver ne change pas. Si le premier ou le seul chiffre à supprimer se situe entre 5 et 9, on augmente de 1 la valeur du dernier chiffre à conserver. Par exemple, selon la technique d'arrondissement classique à la centaine près, une estimation de 49 448 serait arrondie à 49 400, et une estimation de 49 252, à 49 300. Le chiffre 1,78 % serait arrondi à 1,8 %.

Test d'hypothèse compris dans les progiciels statistiques

Nous rappelons aux utilisateur de microdonnées que les résultats des test d'hypothèse (p.ex., valeurs de p du test t ou statistiques de Pearson) fournis automatiquement par les progiciels de statistiques courants sont erronés lorsque les données analysées proviennent d'enquêtes complexes comme l'EDTR. Ces progiciels supposent au départ qu'on a procédé à un échantillonnage aléatoire simple; ils ne tiennent pas compte des caractéristiques spéciales du plan de sondage de l'EDTR comme la stratification, la mise en grappes et les probabilités inégales de sélection.

Nombre de progiciels courants tiennent compte des probabilités inégales de sélection en autorisant le recours à la pondération pour la production des estimations, mais ils ne prennent pas correctement en compte le plan de sondage lors du calcul des estimation de la variance, un élément essentiel de la plupart des tests statistiques.

Pour effectuer des tests d'hypothèses, il existe une méthode en deux étapes qui utilise les progiciels de statistiques courants pour calculer les paramètres du test. Il s'agit d'abord d'estimer les caractéristique d'intérêt en utilisant les poids fournis dans les fichiers de

microdonnées, puis d'obtenir des estimations de la variance approximative de ces caractéristiques en utilisant le progiciel comme pour produire les estimations des caractéristiques, mais en appliquant cette fois un poids relatif correspondant au quotient du poids original par la moyenne des poids originaux de l'ensemble des observations utilisées pour les calculs. Les données obtenues dans ces deux étapes peuvent alors être combinées pour calculer les paramètres du test. Il convient cependant de noter que cette méthode ne donne que des estimations approximatives de l'écart-type.

Il convient de noter qu'il est impossible d'obtenir de meilleures estimations de la variance fondée sur le plan de sondage en utilisant des progiciels conçus spécifiquement pour les données d'enquête. Les informations sur le plan de sondage qui seraient nécessaires à cette fin ne sont en effet toujours pas disponibles dans les fichiers de données de l'EDTR pour des raisons de confidentialité. Toutefois, on peut obtenir de Statistique Canada contre recouvrement des coûts, de meilleures estimations de la variance.

6. CONFIDENTIALITÉ DU FICHIER DES MICRODONNÉES À GRANDE DIFFUSION

La production d'un fichier de microdonnées à grande diffusion comprend de nombreuses mesures de protection visant à prévenir l'identification d'une personne. Les enquêtes longitudinales comportent un défi supplémentaire du point de vue de la confidentialité, étant donné que des données sont recueillies pour une même personne pendant plusieurs années. Pour cette raison, Statistique Canada planifie de diffuser seulement les données de l'enquête transversale de l'EDTR. Le nombre de sujets compris dans l'EDTR augmente aussi le traitement supplémentaire requis pour assurer la confidentialité. La confidentialité du fichier à grande diffusion est assurée principalement par la réduction de l'information, c'est-à-dire la suppression de variables complètes ou de certains détails qu'elles comprennent, ou encore le regroupement de ces détails.

Dans le cadre de l'EDTR, on utilise un certain nombre de techniques pour assure la confidentialité:

- Le fichier des microdonnées à grande diffusion de L’EDTR est composé d’un échantillon provenant de l’échantillon complet de l’EDTR sélectionné de façon aléatoire.

- Toutes les variables permettant l'identification directe de personnes sont évidemment supprimées du fichier. Il s'agit du nom, du numéro de téléphone et d'autres données utilisées pour la collecte.

- Regroupement de catégories. Cette méthode est appliquée aux variables catégoriques (c'est-à-dire qualitatives).

- Codage supérieur et inférieur. Les valeurs très élevées et très faibles sont généralement rares ou uniques au sein d'une population. De telles valeurs

extrêmes sont remplacées par une fourchette supérieure ou inférieure ou par une valeur supérieure ou inférieure.

- Arrondissement. Certaines variables, particulièrement celles de nature pécuniaire, sont arrondies.

- Suppression des caractéristiques. Dans certain cas, les combinaisons de variables peuvent être problématiques. On a procédé au croisement détaillé des caractéristiques afin de discerner ces cas, et on a ensuite supprimé ou regroupé certaines des valeurs impliquées.

- Les enregistrements imputés du fichier et des variables ne sont pas identifiés comme tels.

- Addition du "bruit" (perturbation). Certaines valeurs numériques ont peut-être été ajustées de façon aléatoire à la hausse ou à la baisse par des montants et des proportions inégaux, tout en maintenant l'intégrité des données, afin de permettre la production de statistiques exactes et précises.

7. CONTENU DE L'EDTR, NOTES ET DÉFINITIONS, MÉTHODOLOGIE

Veuillez consulter la section appropriée dans l'Enquête sur la dynamique du travail et du revenu (EDTR) - Un aperçu de l'enquête

8. PRODUITS ET SERVICES CONNEXES

Veuillez consulter la section appropriée dans l'Enquête sur la dynamique du travail et du revenu (EDTR) - Un aperçu de l'enquête

9. QUESTIONS ET COMMENTAIRES

Si vous avez des questions ou des commentaires au sujet des données que contient ce CD-ROM, vous pouvez communiquer avec la Division de la statistique du revenu.

Téléphone : 1 888 297-7355 ou (613) 951-7355
Télécopieur : (613) 951-3012
Internet : revenu@statcan.ca

Division de la statistique du revenu
Statistique Canada
Ottawa (Ontario)
K1A 0T6