Statistics Canada Statistique Canada

11-613F
no 2001-001
c. 1

NOT FOR LOAN
NE S'EMPRUNTE PAS

Methodology Branch

Social Survey
Methods Division

Direction de la méthodologie

Division des méthodes
d'enquêtes sociales

Statistics Canada Statistique Canada
MAR 11 2016
LIBRARY
BIBLIOTHÈQUE

Canada

DOCUMENT DE TRAVAIL
DIRECTION DE LA MÉTHODOLOGIE

COEFFICIENTS DE VARIATION POUR FICHIERS DE MICRODONNÉES À GRANDE DIFFUSION: LE CAS DE L'ENQUÊTE SUR LA DYNAMIQUE DU TRAVAIL ET DU REVENU

SSMD-2001-001F

Johane Dufour et Christian Nadeau

Statistique Canada

Octobre 2000

Le travail présenté dans ce document est sous la responsabilité des auteurs et ne représente pas nécessairement les opinions ou politiques de Statistique Canada

COEFFICIENTS DE VARIATION POUR FICHIERS DE MICRODONNÉES À GRANDE DIFFUSION: LE CAS DE L'ENQUÊTE SUR LA DYNAMIQUE DU TRAVAIL ET DU REVENU

Johane Dufour[1]
et
Christian Nadeau[2]

SOMMAIRE

L'Enquête sur la dynamique du travail et du revenu (EDTR) est une enquête-ménages longitudinale à panels rotatifs munie d'un plan de sondage complexe. Un des principaux objectifs de cette enquête consiste à mieux comprendre le bien-être économique de la société canadienne. Le premier panel de cette enquête a débuté en 1993, et les répondants de ce panel sont suivis à tous les ans (ou à chaque vague de l'enquête) pour une période de six ans. Quoique conçue pour produire des estimations longitudinales, l'EDTR sert également de véhicule pour obtenir des estimations transversales.

Jusqu'à maintenant, l'EDTR a traité les données des six premières vagues qu'elle a recueillies de 1994 à 1999. Des fichiers de microdonnées à grande diffusion pour les vagues 1 et 2 ont été diffusées, alors que ceux des vagues 3 à 6 le seront bientôt. À l'aide de ces fichiers de microdonnées à grande diffusion, il est possible de dériver une grande variété d'estimations. À chacune de ces estimations est associée une erreur d'échantillonnage qu'il est utile de connaître afin d'en évaluer la qualité. Ce document présente tout d'abord les pratiques qui ont été utilisées lors des deux premières vagues de l'EDTR afin d'estimer les erreurs d'échantillonnage ainsi que différentes méthodes qui pourraient être envisagées pour les diffusions futures. Vient ensuite un bref survol des développements dans le domaine à Statistique Canada. Finalement, les auteurs présentent une méthode multicritère permettant d'identifier, parmi les options décrites pour estimer l'erreur d'échantillonnage, celle qui conviendrait le mieux selon l'enquête en pondérant divers critères selon leur importance.

[1]Johane Dufour, Division des méthodes d'enquêtes auprès des ménages, Statistique Canada, Ottawa, Ontario, Canada, K1A 0T6

[2]Christian Nadeau, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa, Ontario, Canada, K1A 0T6

COEFFICIENTS OF VARIATION FOR PUBLIC USE MICRODATA FILES: THE CASE OF THE SURVEY OF LABOUR AND INCOME DYNAMICS

Johane Dufour[1]
and
Christian Nadeau[2]

SUMMARY

The Survey of Labour and Income Dynamics (SLID) is a longitudinal household survey consisting of rotating panels based on a complex sample design. One of the main objectives of this survey is to better understand the economic well-being of Canadian society. The first panel of this survey started in 1993, and its respondents are followed every year (or at each wave of the survey) for a six-year period. Although conceived to produce longitudinal estimates, SLID is also used as a vehicle to produce cross-sectional estimates.

So far, SLID has processed the data from its first six waves, which were collecte from 1994 to 1999. Public Use Microdata Files (PUMF) for waves 1 and 2 have been released; those from waves 3 to 6 will soon be available. The PUMFs permit a wide variety of estimates to be derived. Each estimate is associated with a sampling error that is used to assess quality. This document presents the methods that were used to derive sampling errors for the first two waves of SLID as well as other options that could be adopted for future releases. A brief overview of development in this domain at Statistics Canada is also given. Finally, the authors present a multi-criterion method that weights different criteria by their importance in a given survey, thereby permitting one to identify the best option for estimating sampling error.

[1]Johane Dufour, Household Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6

[2]Christian Nadeau, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6

DANS CE DOCUMENT

1. INTRODUCTION 7

2. MÉTHODES ADOPTÉES AUX VAGUES 1 ET 2 de l'EDTR 8
   2.1 Coefficients de variation approximatifs - variables qualitatives 8
   2.2 Coefficients de variation approximatifs - variables quantitatives 9
      2.2.1 Modèles loglinéaires 9
      2.2.2 Ajustement à l'aide d'un effet de plan de sondage 11

3. MÉTHODES COURAMMENT UTILISÉES À STATISTIQUE CANADA 12
   3.1 Fonctions de variance généralisées 12
      3.1.1 Variables qualitatives: approche conservatrice dans 75 % des cas 13
      3.1.2 Variables qualitatives: approche sans contrainte 13
      3.1.3 Variables quantitatives 13
      3.1.4 Évaluation des fonctions de variance généralisées 14
   3.2 Télé-accès 15
      3.2.1 Télé-accès via le jackknife 15
      3.2.2 Télé-accès via le bootstrap 15
      3.2.3 Diffusion d'un fichier fictif 16
      3.2.4 Évaluation des approches via télé-accès 16
   3.3 Développements dans le domaine à Statistique Canada 17

4. DÉMARCHE MULTICRITÈRE POUR LE CHOIX D'UNE MÉTHODE DE DIFFUSION 18
   4.1 Cotation de la méthode selon chacun des critères 19
   4.2 Calcul d'un score pondéré 21
   4.3 Application de la méthode à l'EDTR 21

5. CHOIX D'UNE STRATÉGIE POUR L'EDTR 22

REMERCIEMENTS 23

BIBLIOGRAPHIE 23

ANNEXE 26

1. INTRODUCTION

L'Enquête sur la dynamique du travail et du revenu (EDTR) est une enquête-ménages longitudinale dont l'échantillon est tiré du plan complexe de l'Enquête sur la population active (EPA). Le but principal de l'EDTR consiste à retracer les changements qui touchent la situation économique des personnes et des familles, au fil des ans, ainsi que les facteurs qui ont une influence déterminante à cet égard. Deux interviews sont réalisées: une interview sur le travail en janvier et une interview sur le revenu en mai. Les données recueillies en janvier sont majoritairement à caractère qualitatif tandis que celles collectées en mai sont de type quantitatif. De plus, quoique l'EDTR ait été conçue pour produire des estimations longitudinales, elle sert également de véhicule pour obtenir des estimations transversales. De même, les utilisateurs s'intéressent aux estimations de niveaux, de ratios et aux changements entre les estimations de niveaux, et entre les ratios au fil des ans. Il est donc possible de dériver une grande variété d'estimations à partir des données de l'EDTR.

À l'EDTR, différents produits de diffusion sont mis à la disposition des utilisateurs afin d'accéder aux données. Tout d'abord un fichier maître est créé à l'intention des employés de Statistique Canada. De ce fichier est créé le principal produit de diffusion, soit un ensemble de fichiers de microdonnées à grande diffusion (FMGD) qui, pour les vagues 1 et 2, comprenait des fichiers pour l'analyse transversale de même que des fichiers pour l'analyse longitudinale. À partir de la vague 3, seuls les FMGD à caractères transversaux sont diffusés. À partir de la vague 6, s'ajoutent aussi une publication papier et une publication électronique comprenant de nombreux tableaux statistiques.

À chacune des nombreuses estimations pouvant être produites à partir des données de l'EDTR est associée une erreur d'échantillonnage qu'il est utile de connaître afin d'en évaluer la qualité. Or le calcul des erreurs d'échantillonnage pour les enquêtes à plan complexe n'est pas une tâche facile pour les utilisateurs de microdonnées. De plus, lors de la diffusion de FMGD, afin de satisfaire aux politiques en matière de confidentialité de Statistique Canada, les variables relatives au plan d'échantillonnage sont omises. La méthodologie a donc comme mandat de développer une méthode qui répond le mieux aux besoins des utilisateurs des différents produits (fichier maître, FMGD, publications) tout en respectant les lignes directrices et les politiques de Statistique Canada. Le présent document traitera plus particulièrement de telles méthodes dans le cadre de la diffusion des FMGD de l'EDTR.

Dans un premier temps, les pratiques utilisées aux vagues 1 et 2 de l'EDTR pour le calcul de coefficients de variation approximatifs sont présentées. Deux stratégies différentes avaient été adoptées pour traiter les variables quantitatives et qualitatives. Dans un deuxième temps, certaines pratiques qui pourraient être envisagées pour les diffusions subséquentes, sont présentées. Encore une fois, on traite de variables qualitatives et quantitatives. On aborde ensuite la possibilité d'avoir recours au télé-accès via les méthodes du jackknife et du bootstrap pour la dérivation de variances, et la possibilité de fournir aux utilisateurs des fichiers fictifs de poids jackknife ou bootstrap. Pour chacune des méthodes exposées, on retrouve une brève évaluation sommaire selon cinq critères: la flexibilité, la fiabilité, la facilité d'utilisation, l'autonomie et le développement méthodologique requis. On traite ensuite brièvement des nouveaux développements dans le domaine à Statistique Canada. Suit ensuite une démarche multicritère visant à faciliter le choix d'une méthode à partir des cinq critères précédents. Finalement, la dernière section de ce rapport discute du choix de la stratégie pour les diffusions subséquentes de l'EDTR.

2. MÉTHODES ADOPTÉES AUX VAGUES 1 ET 2 de l'EDTR

Dans cette section, les pratiques utilisées lors de la diffusion des vagues 1 et 2 de l'EDTR sont présentées. Pour les variables qualitatives, la méthode présentée est celle qui se trouve dans le "Guide de l'utilisateur des microdonnées de l'EDTR" (Statistique Canada 1997). Pour les utilisateurs, l'utilisation de telles méthodes se résume à regarder dans un tableau la valeur du coefficient de variation (CV) en fonction de l'estimation obtenue et de la caractéristique d'intérêt. Pour les variables de type quantitatif, un modèle loglinéaire a été utilisé pour dériver les coefficients de variation des variables de revenu. Ce modèle ne fait pas partie du guide de l'utilisateur mais a été utilisé par quelques utilisateurs.

Ayant en main une estimation du CV, l'utilisateur se réfère à un tableau (similaire au tableau A1 de l'annexe) pour obtenir une idée de l'acceptabilité de l'estimation.

2.1 Coefficients de variation approximatifs - variables qualitatives

Les coefficients de variation approximatifs pour les variables qualitatives de l'EDTR s'obtiennent des tableaux de variabilité d'échantillonnage approximative dérivés pour cette enquête. Cette méthode s'inspire grandement de celle adoptée par la Division des enquêtes spéciales à Statistique Canada (Hunter 1991). Chaque tableau de CV est valide pour un niveau différent de géographie (Canada ou province) et selon trois groupes de variables. Les variables qualitatives ont été regroupées après discussion avec les spécialistes du sujet selon trois groupes de thèmes: thème 2 - caractéristiques ethnoculturelles, thème 7 - caractéristiques de l'activité sur le marché du travail et tous les thèmes qui restent (plus de détails sur les différents thèmes sont donnés au tableau A2 en annexe).

Flexibilité

Ce genre de tableaux ne permet de traiter que les variables de type qualitatif. Une seule série de tableaux permet le calcul de CV pour les estimations de niveaux, de proportions et de ratios, de même que pour les différences de niveaux, de proportions et de ratios. Ces tableaux ne permettent cependant pas le calcul de CV pour des statistiques plus complexes.

Fiabilité

Les CV calculés sont en général conservateurs (environ 75% du temps) pour les estimations de niveaux et de proportions. Il est toutefois difficile d'en connaître l'incidence sur les résultats de tests d'hypothèse. La qualité des CV produits dépend en partie de la méthode utilisée pour effectuer le calcul des variances. La méthode du jackknife avec regroupement de strates, qui mène en général à une estimation conservatrice de la variance (surestimation), a été utilisée à l'EDTR. Dans le cas des estimations de ratios, une corrélation positive entre le numérateur et le dénominateur mènera à une estimation conservatrice du CV alors qu'une corrélation négative mènera à une sous-estimation du CV. De même, les CV pour les estimations de différences seront biaisées si les estimations des caractéristiques utilisées sont corrélées.

Facilité d'utilisation

Ces tableaux sont très faciles à utiliser pour le calcul de CV relatifs aux estimations de niveaux et de proportions. Leur utilisation devient un peu plus difficile pour l'obtention de CV d'estimations combinant plusieurs composantes (ratios, différences de niveaux, de proportions ou de ratios). Le niveau de complexité demeure tout de même acceptable. Les utilisateurs de données de Statistique Canada sont généralement familiers avec ceux-ci. L'augmentation du niveau de détail géographique ou du nombre de regroupements de variables utilisés lors de la création de tels tableaux peut vite devenir un inconvénient en raison du nombre de tableaux à consulter pour l'utilisateur.

Autonomie

Une fois les directives comprises, l'utilisateur de données peut calculer ses propres CV sans aucune autre forme d'aide. Comme mentionné à l'item flexibilité, il ne pourra cependant pas calculer de CV pour des estimations faites à partir de variables de type quantitatif ni pour des estimations plus complexes à partir de variables qualitatives.

Développement méthodologique requis

Cette méthode est facile à appliquer puisque les programmes existent déjà. Pour ce faire, il a suffi de déterminer quels thèmes pouvaient être regroupés, décision qui a été prise par les spécialistes du sujet. Par la suite, il a fallu calculer plusieurs effets de plan de sondage ainsi que le 75ième percentile afin d'alimenter le programme de coefficients de variation disponible à Statistique Canada. (Ce percentile est utilisé afin d'assurer une surestimation pour environ 75 % du temps, et une sous-estimation 25 % du temps.)

2.2 Coefficients de variation approximatifs - variables quantitatives

À l'EDTR, les coefficients de variation approximatifs pour les variables quantitatives s'obtiennent de deux méthodes différentes. Pour certaines variables d'intérêt de l'EDTR, des modèles loglinéaires ont été appliqués et les résultats semblent satisfaisants. Ce type de modèles fait partie des fonctions de variance généralisées (FVG) décrites dans Wolter (1985). Ces fonctions permettent d'obtenir une approximation du coefficient de variation d'une estimation quelconque comme fonction simple de l'estimation. De telles fonctions sont obtenues généralement en modélisant les variances estimées pour un certain nombre de caractéristiques en fonction des estimations de ces mêmes caractéristiques. Pour l'ensemble des variables non couvert par ces modèles, l'EDTR suggère à l'utilisateur de calculer une variance (se référer à la section 2.2.2) et ensuite d'utiliser le tableau des effets de plan de sondage fournis par l'EDTR pour pouvoir dériver une estimation d'un CV.

2.2.1 Modèles loglinéaires

Les modèles loglinéaires ont été développés après la première vague de l'EDTR mais n'ont pas été offerts comme partie intégrante du Guide de l'utilisateur (Statistique Canada 1997) étant donné le temps de production requis pour leur développement. Pour utiliser ces tableaux, il s'agissait d'avoir en main l'estimation de la variable d'intérêt et le secteur géographique (province) et de lire dans un tableau papier les valeurs de deux paramètres. Ces deux paramètres (A et B), de concert avec l'estimation (X), sont

ensuite utilisés par l'utilisateur dans une formule mathématique ($CV = A*(X)^B$). Toutefois, ces modèles ne sont disponibles que pour un sous-ensemble de variables quantitatives (se référer au tableau A3 de l'annexe). De plus, pour une estimation à l'échelle nationale, le travail est un peu plus ardu. Il faut obtenir une estimation et un CV pour chacune des provinces, dériver la variance correspondante (variance $= (CV * X)^2$), sommer les 10 variances obtenues pour obtenir la variance nationale et finalement redériver le CV. Toutefois, il serait possible de donner les coefficients pour le Canada.

Flexibilité

De tels modèles ont été dérivés pour un nombre restreint de variables quantitatives et seuls des CV pour des estimations de totaux sont traités. Les modèles sont disponibles à l'échelle provinciale ainsi que pour le Canada. Il est impossible d'utiliser une telle méthode pour le calcul de CV de statistiques complexes tel que les coefficients de régression.

Fiabilité

La fiabilité de tels modèles pour des variables de type quantitatif est généralement mis en doute dans la littérature, bien qu'ils donnent généralement une meilleure estimation du CV que ne le permet la méthode décrite en 2.2.2. L'utilisation de ces modèles pour le calcul de CV des estimations de ratios, de différences ou de différences de ratios entraîne un biais si les estimations utilisées lors de ces calculs sont corrélées. La fiabilité dépend aussi en grande partie de la qualité de la méthode d'estimation de la variance utilisée, soit celle du jackknife avec regroupement de strates dans le cas de l'EDTR.

Facilité d'utilisation

La formule à utiliser est assez simple même s'il faut choisir les paramètres à appliquer. Le calcul de CV pour le Canada exige beaucoup de manipulations mathématiques. De même, l'utilisation de ces modèles pour le calcul de CV des estimations de ratios, de différences ou de différences de ratios comporte des manipulations supplémentaires qui rendent cette tâche un peu plus difficile.

Autonomie

Le fait que ces modèles n'aient pas été fournis dans le guide d'utilisateur a fait en sorte que la majorité des utilisateurs ignorait leur existence et n'ont pu les utiliser. Ceux-ci ont donc dû utiliser la méthode présentée à la section 2.2.2.

Développement méthodologique requis

Ce modèle a exigé plusieurs étapes du point de vue de la méthodologie. Tout d'abord il faut mentionner que lors de son développement, après la vague 1, le nombre de données était très restreint. Afin de modéliser, il a fallu dans un premier temps se définir différents domaines dans le but d'avoir plusieurs estimations pour une même caractéristique d'intérêt (exemple pour le revenu total). Dix-huit domaines ont été définis (voir tableau A4 en annexe). Pour chacune des variables d'intérêt (voir le tableau A3 de l'annexe), 18 données étaient disponibles et le but était de tenter de modéliser le CV de l'estimation en fonction de l'estimation. Plusieurs modèles ont été envisagés, et c'est un modèle loglinéaire qui a donné les

meilleurs résultats. Les régions géographiques ont été regroupées (par exemple les Maritimes) afin de présenter les résultats en utilisant le moins de tableaux possibles tout en étant le plus fidèle possible aux données. Les valeurs des R^2 et les résidus ont été utilisés comme critère pour déterminer si les regroupements étaient acceptables.

2.2.2 Ajustement à l'aide d'un effet de plan de sondage

Pour les variables pour lesquelles aucun modèle loglinéaire n'a pu être ajusté, on suggère à l'utilisateur de calculer une variance issue d'un tirage aléatoire simple et d'appliquer un effet de plan de sondage pour obtenir une estimation de la variance sous le plan complexe (variance = effet de plan de sondage x variance sous un plan aléatoire simple). Par la suite, le CV est dérivé de la formule suivante ($CV = (\text{variance})^{.5} / X$). L'EDTR fournit aux utilisateurs un tableau d'effets de plan de sondage pour certaines variables d'intérêt.

Flexibilité

Une telle procédure peut s'appliquer à toutes les variables et ce pour plusieurs types d'estimations. Dans les faits, les effets de plan calculés étaient limités à ceux relatifs à des estimations de niveaux pour le Canada et pour les différentes provinces. L'utilisateur est donc limité aux caractéristiques pour lesquelles un effet de plan est fourni. Comme dans le cas des deux méthodes précédemment décrites, il est possible d'utiliser cette méthode pour obtenir le CV d'estimations de ratios et de différences sous l'hypothèse d'une covariance nulle entre les composantes de telles estimations. Il est cependant impossible d'utiliser cette méthode pour le calcul de CV de statistiques complexes tel que les coefficients de régression.

Fiabilité

Les CV calculés ainsi sont peu précis. Si des CV sont calculés pour des estimations tels des ratios ou des différences à partir d'effets de plan pour des estimations de niveaux, les CV estimés seront biaisés en raison de la corrélation entre les différentes estimations. De plus, puisque l'utilisateur dérive lui-même une estimation de la variance, il est impossible de contrôler la qualité du CV obtenu.

Facilité d'utilisation

Cette méthode paraît simple a priori mais elle exige tout de même de la part de l'utilisateur le calcul d'une variance à partir d'un plan d'échantillonnage aléatoire simple.

Autonomie

L'utilisateur peut habituellement se débrouiller seul avec cette méthode à condition de pouvoir calculer une variance lors d'un tirage aléatoire simple.

Développement méthodologique requis

Cette méthode n'est pas très exigeante du point de vue méthodologique. Il suffit de produire des effets de plan de sondage pour plusieurs caractéristiques, et ces valeurs sont essentielles qu'importe la méthode utilisée. Aucune évaluation relative à la qualité des données ne peut être réalisée puisque l'estimation finale

est sous la responsabilité de l'utilisateur.

3. MÉTHODES COURAMMENT UTILISÉES À STATISTIQUE CANADA

Les tableaux de CV approximatifs et les effets de plan de sondage présentés dans le guide des utilisateurs de l'EDTR pour les vagues 1 et 2 , de même que les modèles loglinéaires décrits à la section 2, n'étaient basés que sur les données de la vague 1. Considérant l'évolution de l'échantillon, l'introduction d'un nouveau panel à la vague 4, les changements à la méthode de pondération de même qu'à la stratégie de diffusion, l'EDTR se doit d'analyser les différentes méthodologies disponibles afin de produire des CV pour la diffusion des prochains fichiers de microdonnées. De plus, étant donné que les fichiers de microdonnées à grande diffusion sont diffusés sur cédérom, il est possible d'envisager l'inclusion d'un module interactif qui permettrait aux utilisateurs d'obtenir en un rien de temps les CV désirés. Ce module, qui pourrait également être fourni via Internet, permettrait aux utilisateurs d'obtenir les CV désirés uniquement en répondant à une série de questions.

La sous-section qui suit présente des approches **informatisées** basées sur des modèles, pour dériver des CV de variables qualitatives et quantitatives. La complexité du modèle n'est plus un facteur à considérer puisque celui-ci devient transparent pour l'utilisateur. Le modèle peut donc être plus raffiné que lors de l'utilisation de tableaux papier et être développé pour une plus vaste gamme de variables. Suit ensuite une autre approche, totalement différente, soit celle du télé-accès. Dans ce cas, aucune limite n'est imposée quant à la dérivation des CV mis à part la dépendance de l'utilisateur envers Statistique Canada pour l'obtention de ceux-ci, le porte-feuille de celui-ci et la disponibilité du personnel à Statistique Canada pour satisfaire de telles requêtes. Finalement, on traite brièvement d'autres méthodes qui sont présentement en développement à Statistique Canada.

3.1 Fonctions de variance généralisées

Les méthodes proposées dans cette section réfèrent aux deux modèles proposés par Wolter (1985). Il s'agit d'un modèle loglinéaire du type:

$$\ln(CV) = a + b \ln (X) \qquad (1)$$

et d'un modèle non linéaire de la forme:

$$CV^2 = a + b/X \qquad (2)$$

où CV représente le coefficient de variation d'une estimation de niveau X pour une caractéristique donnée. Il est bon de mentionner ici que plusieurs études empiriques réalisées à Statistique Canada ont montré que le modèle (1) donnait de meilleurs résultats (Ganghurde 1981, Dufour et Lemaître 1988, Bourque et Brisebois 1998 et Brisebois et Bourque 1998). Pour ajuster un tel modèle, il existe deux types d'approches.

3.1.1 Variables qualitatives: approche conservatrice dans 75 % des cas

La première approche, décrite dans Phillips et Kaushal (1998), garde en tête une approche similaire à celle décrite à la section 2.1, c'est-à-dire qu'elle a comme objectif de surestimer les CV dans environ 75 % des cas. Pour ce faire, un premier modèle de régression est appliqué à l'ensemble de la série des caractéristiques d'intérêt pour un niveau géographique désiré (comme par exemple à l'échelle nationale, provinciale, région métropolitaine de recensement). Avec un tel modèle, on s'attend à ce qu'environ 50 % des estimations se retrouvent de part et d'autre de la courbe de régression. On élimine donc toute donnée se trouvant sous la courbe de régression et un deuxième modèle de régression est appliqué aux données restantes. Le modèle retenu est utilisé pour produire les tableaux approximatifs.

3.1.2 Variables qualitatives: approche sans contrainte

La deuxième approche décrite dans Dufour et Lemaître (1988) et reprise par Boyer et Gwet (1991) ne pose pas la même contrainte de surestimation dans 75 % des cas. La qualité des estimations de CV obtenues repose plutôt sur 2 facteurs: le regroupement des caractéristiques et le modèle lui-même. Plusieurs modèles sont ajustés puisque les caractéristiques sont combinées en fonction des effets de plan de sondage similaires. Ces auteurs favorisent un modèle loglinéaire avec différentes interventions dans le modèle pour refléter des changements dans le plan de sondage (exemple: une réduction importante de la taille d'échantillon dans le temps). Comme méthodes de regroupements, on peut songer à une analyse visuelle ou encore faire appel à des tests de comparaisons multiples (tests de Scheffe ou Tukey), ou à des méthodes d'analyse de regroupement après vérification de certaines hypothèses.

Qu'importe l'approche proposée (3.1.1 ou 3.1.2), il est possible d'utiliser des regroupements de caractéristiques et de raffiner les modèles. On peut tout simplement combiner les caractéristiques selon les recommandations des spécialistes du sujet (comme suggéré en 2.1) ou faire appel à des techniques de regroupement plus sophistiquées (par exemple: "cluster analysis"). Il est également possible de dériver des facteurs d'ajustement qui permettraient, après avoir obtenu le CV pour un total, d'obtenir des CV pour une variation annuelle ou pour des niveaux géographiques plus détaillés. L'utilisation d'un module informatique permet également d'obtenir une foule de combinaisons sans toutefois compliquer la vie aux utilisateurs.

3.1.3 Variables quantitatives

Une façon d'aborder le cas des variables quantitatives consiste à rendre ces variables catégoriques et à utiliser un des modèles décrits en 3.1.1 ou en 3.1.2. Toutefois, une telle approche se limite aux estimations d'effectifs pour les catégories retenues. L'adoption d'un module informatique rend cette procédure transparente à l'utilisateur et peut donner des résultats satisfaisants d'autant plus que les catégories peuvent couvrir une étendue plus ou moins grande. Il faut cependant s'arrêter pour trouver un bon compromis entre le nombre de catégories et des données de qualité. On peut également songer à utiliser les modèles (1) et (2) décrits au début de la section 3.1 directement sur les variables quantitatives. Ce genre de modèles combinés à une approche conservatrice a d'ailleurs été utilisé pour les vagues 1 et 2 de l'EDTR (voir section 2.2.1).

3.1.4 Évaluation des fonctions de variance généralisées

L'utilisation d'une **approche conservatrice dans 75 % des cas et d'une approche sans contrainte,** autant pour les variables à caractère qualitatif que quantitatif, est évaluée ci-dessous selon les cinq critères pré-établis.

Flexibilité

Ce genre de modèles permet l'utilisation d'une approche similaire autant pour les variables de type qualitatif que quantitatif. En plus de permettre le calcul de CV pour des estimations de niveaux, il est aussi possible d'étendre leur application à des estimations de proportions et de ratios, de même qu'à des estimations de différences de niveaux, de proportions et de ratios en calculant certains facteurs d'ajustement. Pour tenir compte du niveau de détail géographique, on peut soit en tenir compte lors de la modélisation ou encore calculer certains facteurs d'ajustement. Ces modèles ne permettent toutefois pas le calcul de CV pour des estimations plus complexes comme des coefficients de régression.

Fiabilité

Dans le cas de caractéristiques de type quantitatif, on sait, selon la littérature (Wolter 1985), que ces modèles peuvent être non performants mais qu'ils donnent de meilleurs résultats que de proposer aux utilisateurs d'estimer une variance sous l'hypothèse d'un tirage aléatoire simple et d'appliquer un effet de plan. De plus, on exerce un meilleur contrôle sur la qualité des estimations obtenues. En ce qui a trait aux variables qualitatives, les CV obtenus pour des estimations de niveaux sont plutôt fiables. Ces résultats dépendent cependant de la qualité des regroupements effectués lors de la modélisation. L'approche 3.1.1 peut constituer un compromis acceptable entre la fiabilité et simplicité du point de vue du développement méthodologique. De plus la fiabilité des CV pour des estimations de proportions et de ratios de même que pour des estimations de différences de niveaux, de proportions et de ratios dépendra du travail effectué pour calculer les facteurs visant à tenir compte de la covariance entre les estimations. L'utilisation d'un logiciel interactif devrait permettre de minimiser les erreurs de calcul que pourrait engendrer l'application de formules mathématiques par les utilisateurs de données. Il est à noter que comme pour toutes les méthodes présentées jusqu'à maintenant, la fiabilité des CV calculés dépend en partie de la méthode d'estimation de variance utilisée.

Facilité d'utilisation

Le développement d'un module informatique rend la procédure transparente aux utilisateurs qui auraient pu sursauter à la vue d'une formule mathématique munie d'une fonction logarithmique. L'utilisateur n'aura alors qu'à répondre à une série de questions lui permettant d'obtenir un CV pour son estimation.

Autonomie

L'utilisateur pourra lui-même calculer les CV pour les estimations couvertes par une telle méthode, sans aucun recours à Statistique Canada.

Développement méthodologique requis

Les deux approches proposées dans cette section nécessitent l'application de nombreux modèles de régression, que ce soit pour satisfaire la contrainte de 75 % de surestimation ou encore pour raffiner la formation des groupes de modélisation. L'utilisation de groupements peut également nécessiter beaucoup de travail méthodologique dépendant de la complexité envisagée. La validation des modèles retenus exige également des ressources méthodologiques. La détermination des facteurs d'ajustement pour dériver à partir du CV d'un total toute une foule d'autres estimations requiert aussi beaucoup d'efforts. Finalement, le développement d'un module informatique exige de la part de la méthodologie la rédaction de spécifications et une stratégie de mise à l'essai. Cependant des modules informatiques semblables existent déjà et pourraient servir de base.

3.2 Télé-accès

Le télé-accès permet aux utilisateurs d'obtenir les CV désirés par l'intermédiaire de Statistique Canada. Pour le moment, deux modes de télé-accès sont en vigueur pour les enquêtes auprès des ménages. Le premier exige que les utilisateurs soumettent une requête pour toute demande de CV. Les programmes sont rédigés (par exemple la méthode du jackknife ou du bootstrap) et soumis par du personnel de l'Agence. Le deuxième mode offre plus de flexibilité aux utilisateurs. Des fichiers fictifs constitués de "paramètres fictifs" à la place des variables sont fournis aux utilisateurs. Les utilisateurs peuvent ainsi développer leurs propres programmes et communiquer avec Statistique Canada au moment venu pour une soumission finale sur le fichier maître de façon à obtenir le CV approprié.

3.2.1 Télé-accès via le jackknife

L'EDTR utilisait aux vagues 1 et 2 la méthode du jackknife, soit une méthode basée sur des répétitions afin d'estimer toutes les variances désirées. L'idée de cette méthode consiste à créer un certain nombre de sous-échantillons de l'échantillon maître (en soutirant une réplique), et pour chacun de ces sous-échantillons d'obtenir une estimation de la caractéristique d'intérêt. Une estimation de la variance s'obtient de l'estimateur de la variabilité entre les différentes estimations des sous-échantillons. À chacune des itérations de la méthode un ensemble de poids est créé. Ce fichier est conservé à Statistique Canada comme ensemble de poids jackknife officiel de l'EDTR pour une vague donnée. Par la suite, il est très facile et rapide d'obtenir des estimations de variance ou de CV pour des estimations de niveaux et de ratios, pour des changements de niveaux ou de ratios entre deux vagues et pour des statistiques complexes. Ce fichier de poids dits jackknife ne peut être diffusé à l'extérieur de l'Agence pour des raisons de confidentialité. Mayda, Mohl et Tambay (1996) ont tenté de contourner le problème en créant des "super-strates" et des "super-grappes" selon la méthode de Rust (1986) mais leurs efforts n'ont pas convaincu le Comité de diffusion de microdonnées de Statistique Canada. Conséquemment, le télé-accès est utilisé pour répondre à différentes demandes dans un temps relativement court.

3.2.2 Télé-accès via le bootstrap

La méthode du bootstrap peut aussi être utilisée pour le calcul de la variance. Cette méthode est basée sur des répétitions tout comme la méthode du jackknife. Elle consiste à construire un certain nombre de sous-échantillons à partir de l'échantillon maître et à calculer une estimation de la caractéristique d'intérêt pour

chaque sous-échantillon. Une estimation de la variance est ensuite calculée à partir de l'estimation de la variabilité entre les estimations des différents sous-échantillons. Les sous-échantillons sont construits en tirant un échantillon aléatoire de répliques à l'intérieur de chaque strate. Pour plus de détails sur cette méthode se référer à Wolter (1985). Pour un exemple pratique se référer à Yeo, Mantel et Liu (1999).

Tout comme la méthode du jackknife, l'ensemble des poids bootstrap officiels pour une vague donnée est conservé. Toutefois, pour des raisons de confidentialité, l'ensemble des poids bootstrap ne peut figurer sur les fichiers de microdonnées. Yung (1998) tente de contourner le problème en matière de confidentialité en introduisant les poids bootstrap moyens évitant ainsi l'identification d'une grappe ou d'un ensemble de grappes. Toutefois, cette méthode n'a pas été approuvée par le Comité de diffusion de microdonnées. Les utilisateurs doivent donc, via le télé-accès, soumettre leurs différentes demandes d'estimations à Statistique Canada.

3.2.3 Diffusion d'un fichier fictif

Une autre approche à Statistique Canada en matière d'estimation de variance consiste à fournir aux utilisateurs un fichier fictif, c'est-à-dire un fichier de même constitution que le fichier maître mais avec des valeurs fictives. Ces fichiers contiennent donc toute la série de variables d'intérêt publiée par l'enquête ainsi qu'une série de poids correspondants (bootstrap, jackknife ou autres). Les chercheurs peuvent donc ajuster eux-mêmes leurs programmes et lorsque ceux-ci sont finaux, ils demandent à Statistique Canada de les soumettre avec le fichier maître. L'Enquête nationale sur la santé de la population emploie une telle approche (pour plus de détails sur le sujet, se référer à Mantel et Nadon 1999).

3.2.4 Évaluation des approches via télé-accès

Flexibilité

Les approches via le télé-accès permettent le calcul de CV pour n'importe quel type d'estimation, et ce autant pour des variables de type qualitatif que quantitatif. Dans le cas des approches sans fichier fictif, la seule limite tient de la disponibilité du personnel à Statistique Canada pour l'écriture de programmes pour effectuer l'estimation de variance pour certaines analyses complexes. Pour ce qui est de l'approche avec fichier fictif, l'utilisateur écrit ces propres programmes pour estimer la variance et n'est limité que par ses capacités techniques. Une autre limite réside dans la diversité des logiciels d'analyse de données ou langages de programmation supportés par Statistique Canada.

Fiabilité

Les CV calculés selon ces méthodes sont de meilleure qualité que ceux calculés à l'aide des FVG, ces derniers requérant les premiers comme intrants. Le jackknife donne cependant des CV instables pour les estimations de mesures non lisses comme par exemple des estimations de percentiles, de la proportion de faible revenu et du coefficient GINI (se référer à Kovacevic, Yung et Pandher 1995). Le bootstrap possède de meilleures justifications théoriques que d'autres méthodes. La méthode du jackknife conduit généralement à une légère surestimation des CV.

Facilité d'utilisation

Les approches via télé-accès sans diffusion de fichiers fictifs ne présentent aucune difficulté pour l'utilisateur qui n'a qu'à présenter sa demande à Statistique Canada. Par contre, l'approche avec diffusion de fichiers fictifs demande, de la part de l'utilisateur, d'écrire ses propres programmes de calcul de variance, ce qui n'est pas à la portée de tous les utilisateurs.

Autonomie

L'utilisateur s'en remet entièrement à Statistique Canada pour le calcul final des CV. Son autonomie est toutefois supérieur dans le cas de l'approche avec diffusion de fichier fictif puisqu'il peut faire une partie du travail, voire même jusqu'à écrire le programme de variance de façon à ce que le personnel de Statistique Canada n'ait qu'à le soumettre sur le fichier maître.

Développement méthodologique requis

Les exigences de développement relatif à un programme de jackknife ou à un programme de bootstrap sont à peu de choses près équivalents. De tels programmes doivent être exécutés de toute façon pour toutes les autres méthodes suggérées et pour créer les ensembles de poids jackknife ou bootstrap officiels pour usage interne. L'utilisation d'un bootstrap requiert toutefois du travail supplémentaire de façon à trouver un nombre adéquat de sous-échantillons à utiliser pour l'EDTR. La création d'un fichier fictif exige le développement d'une stratégie méthodologique ayant pour but de créer un fichier qui ressemble à la réalité sans toutefois divulguer de l'information confidentielle. Cette stratégie peut être toute simple ou aussi complexe que l'on désire. Plus elle est complexe, et plus elle ressemble à la réalité, plus le risque de mauvaises utilisations par les utilisateurs augmente. Il faut donc trouver le compromis souhaitable. Finalement, les méthodologistes sont impliqués dans la soumission (et souvent re-soumission) des programmes reçus des utilisateurs, et ils doivent s'assurer au meilleur de leur connaissance que les résultats obtenus respectent toutes les règles de la confidentialité.

3.3 Développements dans le domaine à Statistique Canada

À Statistique Canada, plusieurs projets qui traitent de l'estimation de la variance en présence de fichiers de microdonnées à grande diffusion sont en cours. Beaumont (2000) propose une méthode pour estimer la variance provenant d'une enquête complexe pour des variables quantitatives sans avoir recours à des identificateurs au niveau des strates et des unités primaires d'échantillonnage (cette méthode s'applique également aux variables qualitatives). La méthode proposée, soit le "Taylor avec hypothèse d'indépendance", repose sur l'hypothèse d'une sélection indépendante de ménages. En pratique, cette hypothèse est souvent violée mais elle est toutefois très raisonnable dans certains cas. Cette méthode est présentement utilisée pour l'Enquête sur les dépenses en réparation et en rénovation; la formule de variance à utiliser est fournie à l'utilisateur de même que des étapes de programmation faciles à suivre. Toutefois, il serait facile d'étendre la méthode en fournissant aux utilisateurs une série de poids bootstraps créés selon les hypothèses sous-jacentes au modèle. Il est à noter que cette méthode a été développée pour un fichier de microdonnées au niveau ménage. Cette méthode pourrait être généralisée pour les fichiers d'individus mais des travaux supplémentaires sont nécessaires.

Tambay (2000) présente une méthode pour fournir de l'information relative au plan d'échantillonnage sur le fichier de microdonnées permettant l'estimation tout en maintenant un niveau acceptable de risque de divulgation. Par un niveau acceptable, on entend rendre difficile l'utilisation de l'information reliée au plan d'échantillonnage pour identifier les unités. Cette stratégie s'applique à toute enquête utilisant un plan d'échantillonnage stratifié à plusieurs degrés où les unités primaires d'échantillonnage sont choisies avec remise. L'idée de base de la méthode consiste à séparer et à regrouper les répliques avec deux degrés de sévérité et à fournir les résultats pour chaque séparation/regroupement avec une indication de la sévérité. Une étude de simulation à l'aide des données de l'Enquête nationale sur la santé de la population sera entreprise sous peu.

Dans la littérature, on retrouve d'autres types de méthodes comme par exemple l'encryptage de données. Le National Center for Education Statistics (Jabine 1993) encrypte les données présentes sur le fichier de microdonnées et fournit aux utilisateurs un programme qui génère les estimations et les variances désirées. Statistique Canada n'utilise pas cette méthode.

Statistique Canada est conscient que son engagement auprès des répondants d'assurer la confidentialité de leurs données pose de sérieuses contraintes aux chercheurs qui utilisent les fichiers de microdonnées à grande diffusion. Pour venir à bout de ce problème, Statistique Canada, en collaboration avec le Conseil de recherche en sciences humaines du Canada, ont mis sur pied la création de Centres de données de recherche à travers le Canada. Le but de ces centres est d'offrir aux chercheurs un accès plus facile aux microdonnées tout en assurant la confidentialité de l'information, conformément aux dispositions de la *Loi sur la statistique*. Ces sites universitaires sont en fait des prolongements de Statistique Canada et les mêmes mesures de sécurité y sont appliquées. On y retrouvera d'ailleurs un employé de Statistique Canada à temps plein. Les chercheurs désireux d'utiliser un tel service soumettront une demande à un comité de l'Agence. Pour les demandes retenues, les chercheurs prêteront serment en vertu de la *Loi sur la statistique*, les données retirées physiquement des lieux sûrs seront soigneusement examinées pour assurer la confidentialité et un rapport de recherche sera remis à Statistique Canada pour un examen professionnel et institutionnel. Neuf centres de données devraient être en opération sous peu.

4. DÉMARCHE MULTICRITÈRE POUR LE CHOIX D'UNE MÉTHODE DE DIFFUSION

Les pratiques décrites aux sections 2 et 3 permettent toutes d'informer les utilisateurs de l'erreur d'échantillonnage associée aux estimations faites à partir de fichiers de microdonnées. Il existe cependant plusieurs différences entre ces pratiques en ce qui a trait à la flexibilité, la fiabilité, la facilité d'utilisation, le degré d'autonomie procuré à l'utilisateur et à l'importance des développements méthodologiques requis pour leur mise en oeuvre. Il importe de tenir compte de ces facteurs, de même que de leur importance relative dans un contexte donné, lors du choix d'une méthode. Dans cette section, on propose une démarche en deux étapes visant à identifier, parmi les pratiques décrites, celle qui répond le mieux aux exigences du contexte dans lequel le choix doit s'effectuer. L'utilisation de cette démarche est ensuite illustrée dans le contexte de l'EDTR.

4.1 Cotation de la méthode selon chacun des critères

La première étape de la démarche de sélection consiste simplement à évaluer les pratiques décrites aux sections 2 et 3 en s'appuyant sur cinq critères cotés selon une échelle d'appréciation. Les cinq critères utilisés, de même que les cotes attribuées selon ces critères, reflètent les propos tenus dans les sections précédentes. Ces critères sont:

i) la flexibilité, c'est-à-dire sa capacité à fournir des CV pour des statistiques, des types de variables et des domaines différents;

ii) la fiabilité des indicateurs obtenus;

iii) la facilité d'utilisation de l'outil proposé;

iv) l'autonomie procurée à l'utilisateur en ce qui a trait à l'obtention des indicateurs de qualité;

v) les développements méthodologiques requis, c'est-à-dire la quantité de travail méthodologique requis pour sa mise en oeuvre.

La cotation des diverses méthodes est présentée au tableau 1 alors que les échelles d'évaluation utilisées sont présentées au tableau 2.

À titre d'exemple, la méthode des CV approximatifs sans contrainte via un module informatique présentée dans la sous-section 3.1.2 permet l'estimation de CV autant pour des statistiques reliées à des variables quantitatives que qualitatives. Elle se limite toutefois aux variables utilisées lors de la modélisation et ne permet pas l'estimation de CV pour des statistiques complexes. Pour ces raisons, on la considère relativement *flexible*. La fiabilité des CV obtenus est généralement *bonne* quoique moindre que celle obtenue en utilisant le jackknife ou le bootstrap et ce particulièrement dans le cas des variables quantitatives. Elle est *extrêmement facile à utiliser* peu importe les aptitudes techniques de l'utilisateur puisqu'elle implique l'utilisation d'un logiciel interactif convivial. Cette méthode rend l'utilisateur *très autonome* puisqu'il peut calculer lui-même ces propres indicateurs dans un grand nombre de situations. Enfin, l'application de nombreux modèles de régression, le regroupement des caractéristiques, la détermination de facteurs d'ajustement pour la dérivation de CV pour des estimations autres que des totaux et l'élaboration d'un module informatique exigent cependant *beaucoup de développements* méthodologiques.

Tableau 1. Cotation des méthodes de calcul et de diffusion de coefficients de variation selon les cinq critères

Méthodes	Flexibilité	Fiabilité	Facilité	Autonomie	Développement
Tableaux papier	2	3	3	4	2
Modèles loglinéaires (papier)	2	3	2	4	2
Effets de plan (papier)	2	2	1	4	4
CV approximatifs via un module informatique (sans contrainte)	3	3	5	4	1
CV approximatifs via un module informatique (conservatrice)	3	3	5	4	1
jackknife (télé-accès)	5	5	5	0	4
bootstrap (télé-accès)	5	5	5	0	4
Fichier fictif (télé-accès)	5	5	4	1	3

Tableau 2. Échelles d'évaluation

	Flexibilité	Fiabilité	Facilité	Autonomie	Développement
0	extrêmement limité	médiocre	extrêmement difficile	totalement dépendant	trop de développement
1	très limité	faible	très difficile	très dépendant	beaucoup de développement
2	limité	passable	difficile	dépendant	moyennement de développement
3	flexible	bonne	facile	autonome	peu de développement
4	très flexible	très bonne	très facile	très autonome	très peu de développement
5	extrêmement flexible	excellente	extrêmement facile	totalement autonome	aucun développement

4.2 Calcul d'un score pondéré

Une fois chaque méthode cotée selon les différents critères, la deuxième étape consiste à attribuer une note globale à chaque méthode. Pour ce faire, on attribue d'abord des poids w_i, $i \in$ {*flexibilité, fiabilité, facilité, autonomie et développement*} d'une valeur de 0 à 4 à chaque critère de façon à refléter son importance relative dans un contexte donné. Plus un critère *i* est jugé important quant au choix de la méthode, plus le poids w_i sera élevé. On suggère d'utiliser les définitions suivantes:

i) $w_i = 0$ pour un critère sans importance;
ii) $w_i = 1$ lorsqu'il est peu important;
iii) $w_i = 2$ lorsqu'il est important;
iv) $w_i = 3$ lorsqu'il est très important;
v) $w_i = 4$ pour un critère primordial.

Il est à noter que les valeurs des w_i peuvent être établies en collaboration avec la division cliente. Le score (S) attribué à une méthode est donc la somme pondérée des cotes (c_i) associées aux différents critères pour celle-ci, c'est-à-dire:

$$S = w_{\text{fiabilité}} * c_{\text{fiabilité}} + w_{\text{flexibilité}} * c_{\text{flexibilité}} + w_{\text{facilité}} * c_{\text{facilité}} + w_{\text{autonomie}} * c_{\text{autonomie}} + w_{\text{développement}} * c_{\text{développement}}.$$

Plus le score est élevé, mieux la méthode répond aux exigences pour le contexte dans lequel le choix doit être effectué

4.3 Application de la méthode à l'EDTR

Dans le cas de l'EDTR, des fichiers de microdonnées à grande diffusion à caractère transversal seront diffusés. Puisque les ressources méthodologiques sont restreintes et que la méthode utilisée devra être disponible dans un temps relativement court, il devient *primordial* d'envisager une méthode qui ne requiert pas trop de développements méthodologiques. ($w_{\text{développement}}=4$). On considérera ensuite la fiabilité des CV calculés et la flexibilité de l'outil comme *très important* ($w_{\text{flexibilité}}=w_{\text{fiabilité}}=3$) puisqu'on veut permettre à l'utilisateur d'avoir accès à des CV précis pour un large éventail d'estimations. La facilité d'utilisation de même que l'autonomie procurée à l'utilisateur sont jugées *importantes* ($w_{\text{facilité}}=w_{\text{autonomie}}=2$), les méthodes répondant à ces critères ayant de plus grandes chances d'être utilisées. On remarque, au tableau 3 sous scénario 1, que les approches via télé-accès sans fichier fictif sont préférées, et ce principalement à cause des développements méthodologiques impliqués.

Le choix d'une telle méthode répondrait toutefois aux besoins des utilisateurs expérimentés au détriment des utilisateurs qui accordent moins d'importance à l'exactitude des CV calculés ($w_{\text{fiabilité}}=2$). Le développement d'une telle méthode pourrait être envisageable à plus long terme, ce qui diminuerait aussi l'importance accordée aux développements méthodologiques ($w_{\text{développement}}=1$). L'importance accordée à la facilité d'utilisation et à l'autonomie dans une telle situation se voit donc augmentée ($w_{\text{facilité}}=w_{\text{autonomie}}=4$) alors que l'importance accordée à la flexibilité demeure la même ($w_{\text{flexibilité}}=3$). On remarque, au tableau 3 sous le scénario 2, que les modèles loglinéaires avec module informatique sont préférables dans une telle situation. Selon le contexte, d'autres pondérations pourraient être utilisées pour le choix des méthodes à utiliser. Une discussion avec les différents intervenants du projet s'impose afin de déterminer les

paramètres appropriés.

Tableau 3. Évaluation des méthodes de calcul et de diffusion de CV - cas de l'EDTR

Méthodes	Scénario 1 $W_{flexibilité} = 3$ $W_{fiabilité} = 3$ $W_{facilité} = 2$ $W_{autonomie} = 2$ $W_{développement} = 4$	Scénario 2 $W_{flexibilité} = 3$ $W_{fiabilité} = 2$ $W_{facilité} = 4$ $W_{autonomie} = 4$ $W_{développement} = 2$
Tableaux papier	37	42
Modèles loglinéaires (papier)	35	38
Effets de plan envisagés (papier)	38	34
CV approximatifs via un module informatique (sans contrainte)	40	52
CV approximatifs via un module informatique (conservatrice)	40	52
Jackknife (télé-accès)	56	49
Bootstrap (télé-accès)	56	49
Fichier fictif (télé-accès)	52	48

5. CHOIX D'UNE STRATÉGIE POUR L'EDTR

Il faut tout d'abord mentionner que pour se soumettre aux standards et lignes directrices de Statistique Canada à l'égard de la qualité des données, l'EDTR se doit de produire des CV pour les vagues 3 à 6. La méthode retenue se doit de tenir compte des différents types d'utilisateurs (des moins expérimentés au plus expérimentés) de même que des différents produits de diffusion mis à leur disposition. Elle doit également prendre en considération les contraintes de temps et de budget imposées par les horaires de production serrés de l'EDTR.

Pour satisfaire plusieurs types de clientèle, la méthodologie envisage l'adoption de deux options. Tout d'abord, l'option du télé-accès sera mise en place. Cette méthode permet non seulement d'obtenir tous les indicateurs de qualité voulus mais également d'en contrôler la qualité. Il faudra toutefois être prudent dans l'administration d'un tel programme pour éviter toute divulgation d'information confidentielle. En l'absence de standards et de lignes directrices établis au sein de l'Agence pour le télé-accès, l'EDTR devra se munir de règles rigoureuses. La méthode du bootstrap sera préférée à celle du jackknife, principalement en raison de l'importance de l'estimation de statistiques non lisses à l'EDTR (médiane, quintiles, coefficients GINI).

Pour ce qui est de la deuxième option envisagée, la méthodologie suggère d'innover en insérant sur le cédérom (ou via Internet) un module permettant d'obtenir des CV basés sur un modèle tel que présenté à la section 3.1. Ce module permettrait aux utilisateurs de dériver toute une gamme de CV sans toutefois nécessiter de manipulation de papier, sans interprétation de formule ou sans recherche quelconque. Il suffirait pour les utilisateurs de se laisser guider par les questions du programme interactif.

On pense qu'une combinaison de ces deux options implantée de façon progressive répondrait aux exigences d'une plus vaste clientèle et offrirait une plus grande flexibilité dans l'analyse et l'interprétation des données. La méthode du bootstrap via le télé-accès sera d'abord implantée, puis, suivront les modèles loglinéaires via cédérom.

REMERCIEMENTS

Les auteurs aimeraient remercier Michel Latouche, Joseph Duggan ainsi que Guy Laflamme pour leurs précieux commentaires qui ont permis d'améliorer la qualité de ce document.

BIBLIOGRAPHIE

Beaumont, J.-F. (2000). Estimation de variance pour un fichier de microdonnées à grande diffusion provenant d'une enquête complexe, Statistique Canada, Document de travail de la direction de la méthodologie, HSMD-2000-002F/A.

Bourque, A. et Brisebois, F. (1998). Redesign of the Labour Force Survey's Monthly CV Tables, Statistique Canada, Division des méthodes d'enquêtes auprès des ménages, Document interne.

Boyer, R. et Gwet, J.-P. (1991). Generalized Variance Functions, Statistique Canada, Division des méthodes d'enquêtes auprès des ménages, Document interne.

Brisebois, F. et Bourque, A. (1998). Modification of the LFS CV look-up Table, Statistique Canada, Division des méthodes d'enquêtes auprès des ménages, Document interne.

Dufour, J. et Lemaître, G. (1988). Modèles des erreurs d'échantillonnage, Statistique Canada, Document de travail de la direction de la méthodologie, SSMD-88-024E.

Ghangurde, P.D. (1981). Modèles des erreurs d'échantillonnage, Techniques d'enquête, Vol:7, No:2, 1981, pages 187-203.

Hunter, L. (1991). Variance for Supplementary Surveys, Statistique Canada, Document interne.

Jabine, T. (1993). Procedures for Restricted Data Access, Journal of Official Statistics, 9, 537-589.

Kovacevic, M.S., Yung, W. et Pandher, G.S. (1995). Estimating the sampling variances of measures of income inequality and polarization - An empirical study, Stastique Canada, Document de travail de la direction de la méthodologie, HSMD-95-007E.

Mantel, H. et Nadon, S. (1999). Dummy File Creation for the Remote Access Program of the National Population Health Survey, Conférence 1999 de la SSC, à paraître.

Mayda, J.E., Mohl, C. et Tambay, J.-L. (1996). Variance Estimation and Confidentiality: They Are Related!, Actes de la conférence de la SSC, p: 135-141.

Phillips, O. et Kaushal, R. (1998). Methodology for CV Look-up Tables, Statistique Canada, Document de travail de la direction de la méthodologie, HSMD-98-001E.

Rust, K. (1986). Efficient Replicated Variance Estimation, Actes de la conférence de l'ASA, p: 81-87.

Statistique Canada (1997). Guide de l'utilisateur de microdonnées de l'EDTR - Vague 2 - 1994, No Catalogue 75M0001GPF.

Tambay, J.-L. (2000). Strategies for the Treatment of PUMF Variances, Statistique Canada, Division des méthodes d'enquêtes auprès des ménages, Document interne.

Wolter, K.M. (1985). Introduction to Variance Estimation, Springer-Verlag New York Inc.

Yeo, D., Mantel, H. et Liu, T.P. (1999). Bootstrap variance estimation for the National Population Health Survey, Actes de la conférence de l'ASA - 1999, à paraître.

Yung, W. (1997). Variance Estimation for Public Use Microdata Files, Actes du Symposium de Statistique Canada, 91-95.

ANNEXE

Tableau A1 : Niveaux acceptables du coefficient de variation

Coefficient de variation approximatif (%)	Restriction quant à l'utilisation
0,0 à 16,5	On peut envisager une diffusion générale non restreinte des estimations.
16,6 à 25,0	On peut envisager une diffusion générale non restreinte des estimations en y joignant une mise en garde aux utilisateurs ultérieurs quant à la variabilité d'échantillonnage élevée liée aux estimations.
25,1 à 33,3	On peut envisager une diffusion générale non restreinte des estimations seulement quand les coefficients de variation ont été calculés de manière exacte et qu'ils ont été jugés acceptables. Autrement, on ne doit pas utiliser ces estimations ni les diffuser.
33,4 ou plus	Les estimations ne peuvent être utilisées ou diffusées en aucun cas.

Tableau A2 : Variables de l'EDTR regroupées selon différents thèmes

Thème	Variables
Thème 1 : Caractéristiques démographiques	Année de naissance Âge Sexe État matrimonial
Thème 2 : Caractéristiques ethnoculturelles	Langue maternelle Pays de naissance Immigrant Antécédents ethniques Niveau de scolarité des parents Minorité visible Antécédent de minorité visible Groupe cible des autochtones Antécédent autochtone
Thème 3 : Limitations d'activité	Durée et existence Empêchement à la recherche ou occupation d'emploi Mobilité professionnelle Satisfaction du nombre de semaines travaillées Nombre de membres 16+ ayant un problème de santé
Thème 4 : Information sur les enfants de la personne	-
Thème 5 : Géographie	Région géographique Mobilité
Thème 6 : Caractéristiques de la famille et du ménage	Taille et type de ménage Taille et type de famille Échelle d'équivalence familiale Sources de revenu (de la famille économique) Caractéristiques relatives au travail (de la fam. écon.) Incapacité Prestation d'a.-e., aide sociale et indemnité accident
Thème 7 : Caractéristiques de l'activité sur le marché du travail	Nombre d'emplois occupés chaque semaine de l'année Nombre d'emplois occupés pendant l'année Mois au cours desquels la personne a travaillé Nombre total d'heures travaillées pendant l'année Nombre total d'heures rémunérées pendant l'année Situation vis-à-vis de l'activité Périodes sans travail et absences du travail Rémunération Recherche d'emploi Variables familiales (situation vis-à-vis de l'activité)
Thème 8 : Expérience de travail	-
Thème 9 : Périodes sans travail	Dates et durées Recherche de travail Vecteur indiquant pendant quels mois

Thème	Variables
Thème 10: Caractéristiques de l'emploi	Catégorie de travailleurs Horaire de travail Revenu (emplois rémunérés) Heures supplémentaires Profession Caractéristiques de l'employeur Absence du travail durant l'emploi
Thème 11 : Source de revenu	Principales sources de revenu Variables du revenu total Autres variables liées au revenu Variables de faible revenu
Thème 12 : Prestations mensuelles reliées à l'assurance-emploi, aux accidents du travail et à l'aide sociale	-
Thème 13 : Activité scolaire	Statut d'étudiant Type d'école fréquentée Indications sur les diplômes
Thème 14 : Niveau de scolarité	Niveau Certificat d'études postsecondaires

Tableau A3 : Liste des variables pour lesquelles un modèle loglinéaire s'est avéré une bonne technique

Groupes	Variables
1.1	Salaires et traitements
1.2	Revenu net d'un emploi autonome agricole
1.3	Revenu net d'un emploi autonome non agricole
2.1	Revenu de placements
2.2	Revenu de placements imposable
2.3	Gains en capital imposables
3.1	Prestation fiscale pour enfants
3.2 3.3	Pension de sécurité à la vieillesse, supplément de revenu garanti, allocation au conjoint RPC/RRQ Prestation d'assurance-emploi Aide sociale
3.4	Indemnités pour accidents de travail Crédit de la taxe sur les produits et services
4	Revenu de pension
5	Pension alimentaire Autre revenu imposable
6.1	Revenu total monétaire
6.2	Revenu après l'impôt
7.1	Impôt sur le revenu (provincial + fédéral)
7.2	Impôt fédéral sur le revenu Impôt provincial sur le revenu

Tableau A4 : Domaines utilisés lors de la modélisation loglinéraire

Variables	Domaines
Sexe	Homme Femme
État matrimonial	Marié Célibataire Union libre Veuf Séparé Divorcé
Groupe d'âge	15-19 20-24 25-34 35-44 45-54 55-64 65 +
Classe de travailleurs	Salarié Travailleur autonome Autre

STATISTICS CANADA LIBRARY
BIBLIOTHÈQUE STATISTIQUE CANADA
1010723652