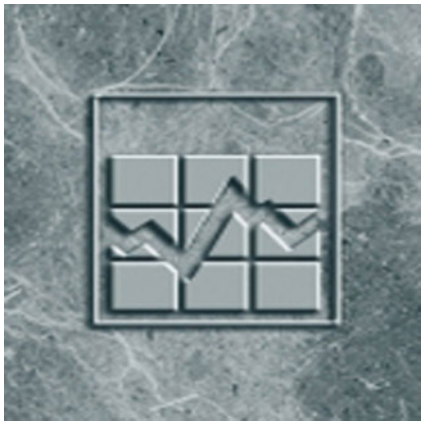
Nº 75F0002MIF au catalogue — Nº 010
ISSN: 1707-2867
ISBN: 0-662-74645-7

Document de recherche

Série de documents de recherche - Revenu

Imputation générale des variables du logement (sans les services publics) dans l'Enquête sur la dynamique du travail et du revenu (EDTR)



par Georgina House

Division de la statistique du revenu
Division des méthodes d'enquêtes sociales
Immeuble Jean-Talon, Ottawa, K1A 0T6

Téléphone: 1 613 951-7355

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Toute demande de renseignements au sujet du présent produit ou au sujet de statistiques ou de services connexes doit être adressée à : Division de la statistique du revenu, Statistique Canada, Ottawa, Ontario, K1A 0T6 (téléphone : (613) 951-7355; (888) 297-7355 : revenu@statcan.ca).

Pour obtenir des renseignements sur l'ensemble des données de Statistique Canada qui sont disponibles, veuillez composer l'un des numéros sans frais suivants. Vous pouvez également communiquer avec nous par courriel ou visiter notre site Web.

Service national de renseignements	1 800 263-1136
Service national d'appareils de télécommunications pour les malentendants	1 800 363-7629
Renseignements concernant le Programme des services de dépôt	1 800 700-1033
Télécopieur pour le Programme des services de dépôt	1 800 889-9734
Renseignements par courriel	infostats@statcan.ca
Site Web	www.statcan.ca

Renseignements pour accéder au produit

Le produit nº 75F0002MIF au catalogue est disponible gratuitement. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.ca et de choisir la rubrique Nos produits et services.

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois, et ce, dans la langue officielle de leur choix. À cet égard, notre organisme s'est doté de normes de service à la clientèle qui doivent être observées par les employés lorsqu'ils offrent des services à la clientèle. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1 800 263-1136. Les normes de service sont aussi publiées dans le site www.statcan.ca sous À propos de Statistique Canada > Offrir des services aux Canadiens.

Statistique Canada
Division de la statistique du revenu

Série de documents de recherche - Revenu

Imputation générale des variables du logement (sans les services publics) dans l'Enquête sur la dynamique du travail et du revenu (EDTR)

Publication autorisée par le ministre responsable de Statistique Canada

Juillet 2005

N° 75F0002MIF au catalogue, vol. 10

ISSN: 1707-2867

ISBN: 0-662-74645-7

Périodicité : hors-série

Ottawa

This publication is available in English upon request (Catalogue no. 75F0002MIE).

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population, les entreprises, les administrations canadiennes et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques précises et actuelles.

Table des matières

Introduction

Depuis un certain temps, la Société canadienne d'hypothèques et de logement (SCHL[1]) utilise les données du Recensement de la population sur les caractéristiques du logement et les dépenses liées au logement. Bien que la source de données que constitue le recensement réponde dans une large mesure aux besoins de la SCHL, cet organisme du gouvernement fédéral a exploité les enquêtes-ménages annuelles de Statistique Canada pour obtenir des renseignements plus fréquents. Cela devait lui permettre d'avoir une image plus fidèle des tendances annuelles, et peut-être d'avoir un plus grand choix d'autres caractéristiques pour le recoupement des données sur le logement des ménages canadiens. En 2001, la SCHL a commencé à commanditer des questions supplémentaires à la fois dans l'Enquête sur la dynamique du travail et du revenu (EDTR) et dans l'Enquête sur les dépenses des ménages (EDM), à compter de l'année de référence 2002[2].

L'Enquête sur la dynamique du travail et du revenu (EDTR) est une enquête longitudinale qui date de 1993. Elle est destinée à mesurer les variations du bien-être économique des Canadiens ainsi que les facteurs touchant ces changements. La population cible est formée de toutes les personnes vivant au Canada, sauf les suivantes : les résidents du Yukon, des Territoires du Nord-Ouest et du Nunavut, les résidents des réserves, les personnes institutionnalisées et le personnel militaire en casernement.

L'échantillon de l'EDTR est formé de 2 panels. Chaque panel demeure dans l'enquête pour 6 années consécutives, et un nouveau panel y entre tous les 3 ans. En janvier suivant l'année de référence, les préposés aux interviews téléphoniques prennent contact avec les ménages de l'échantillon de l'EDTR. Ils recueillent des données démographiques pour chaque membre du ménage. Ils recueillent des données d'enquête complètes pour chaque personne admissible de plus de 15 ans qui se trouve dans le ménage. Ils posent des questions sur le travail (l'activité sur le marché du travail, l'expérience de travail, les épisodes de chômage et l'emploi même), le niveau de scolarité et les sources de revenu. À la fin de l'interview de janvier, ils disent aux répondants qu'on recommuniquera avec eux en mai pour obtenir des données sur leur revenu ainsi que sur certains postes de dépenses. Cependant, le répondant peut dès lors accorder à Statistique Canada la permission de récupérer toutes les données requises dans le fichier des données fiscales T1, ce qui élimine la nécessité d'une deuxième interview. La collecte des données sur le revenu est reportée en mai, de sorte que le répondant (venant alors juste de produire une déclaration de revenus) connaîtra mieux les données requises.

Bien que conçue à l'origine comme une enquête longitudinale, l'EDTR a toujours conservé la capacité de produire des estimations transversales. Toutes les personnes qui sont membres des ménages choisis pour l'EDTR au début de la première année de

1. La Société canadienne d'hypothèques et de logement (SCHL) est un organisme du gouvernement fédéral dont le mandat est de favoriser : la construction résidentielle, et la réparation et la modernisation des habitations existantes; l'accès à une diversité de logement abordables; l'amélioration des conditions de logement; la disponibilité de financement à moindre coût; et la prospérité du secteur de l'habitation.
2. Au moment de la rédaction du présent rapport, cette commandite se poursuit chaque année.

l'existence d'un panel font partie de l'échantillon longitudinal pour l'EDTR. À ce titre, ce sont ces particuliers qui font l'objet d'une observation longitudinale. Toute autre personne vivant dans un ménage avec une personne de l'échantillon longitudinal est appelée cohabitant. Les cohabitants vivant avec des personnes de l'échantillon longitudinal admissibles à l'échantillon transversal font aussi partie de l'échantillon transversal.

Pour plus de renseignements sur les concepts, les définitions et la conception de l'enquête, voir la publication de Statistique Canada intitulée « *Enquête sur la dynamique du travail et du revenu – Un aperçu de l'enquête* », http://www.statcan.ca :8096/bsolc/francais/bsolc?catno=75F0011X.

Pour répondre au besoin de données plus abondantes de la SCHL, Statistique Canada a ajouté plus de 20 questions sur le logement à l'interview de l'EDTR sur le travail. Certaines questions s'adressent aux propriétaires-occupants et aux locataires, certaines uniquement aux propriétaires-occupants, et d'autres uniquement aux locataires. L'EDTR a commencé à recueillir des renseignements sur les paiements hypothécaires, les impôts fonciers, les frais de copropriété, les loyers et ce qui est inclus dans le loyer.

À cause de la non-réponse à certaines questions particulières, il a fallu instituer l'imputation des variables du logement dans l'EDTR[3]. L'objet de l'imputation est de remplacer des données manquantes ou absentes par des valeurs qui donnent des estimations raisonnables. Deux méthodes d'imputation ont été utilisées : l'imputation longitudinale et l'imputation transversale par donneur. C'est ce que nous expliquons dans les deux sections suivantes.

Imputation longitudinale

L'imputation longitudinale tient compte, dans la mesure du possible, des renseignements du cycle précédent. L'efficacité de l'imputation est considérablement accrue lorsque les cycles consécutifs sont en étroite corrélation.

L'imputation longitudinale des variables du logement a été effectuée dans toute la mesure du possible. L'imputation longitudinale est préférable, surtout lorsque le ménage n'a pas déménagé. Nous supposons que, si le ménage n'a pas changé de résidence depuis l'année précédente, les renseignements pour la nouvelle année risquent d'être les mêmes que ceux de l'année précédente.

Puisque l'année de référence 2002 était l'année d'un nouveau panel, l'imputation longitudinale a été effectuée différemment, selon que le ménage faisait ou pas partie du premier cycle du nouveau panel. Pour les ménages qui ne faisaient pas partie du premier cycle du nouveau panel, les renseignements pour les trois variables du logement (mode d'occupation, type de logement et nombre de chambres à coucher) pouvaient s'obtenir des années précédentes. C'étaient les trois seules variables du logement qui n'étaient pas

3. L'EDTR a toujours fait l'imputation des valeurs manquantes pour les variables du revenu.

nouvelles dans l'EDTR pour l'année de référence 2002. (Les données sur le prêt hypothécaire sur le logement pouvaient aussi s'obtenir des années antérieures, mais il a été décidé de ne pas effectuer d'imputation longitudinale de cette variable, parce que la situation hypothécaire risque davantage de changer même s'il semble que l'adresse est restée la même.) On a imputé les renseignements manquants pour les ménages à l'aide des renseignements du même ménage pour les années précédentes si le code postal était toujours le même. C'était considéré comme une bonne indication que le ménage n'avait pas changé de logement.

L'Enquête sur la population active (EPA) et l'Enquête sur le loyer (un supplément de l'EPA) ont été exploitées comme source de renseignements pour l'imputation longitudinale pour les ménages faisant partie du premier cycle du nouveau panel[4]. On a imputé les renseignements manquants pour les ménages en utilisant les renseignements du même ménage selon les données de l'EPA si le code postal était le même. Pour utiliser l'Enquête sur le loyer, il a fallu vérifier si la variable du mode d'occupation dans l'EDTR indiquait que le ménage était toujours locataire.

L'imputation longitudinale a permis d'abaisser le nombre de ménages avec valeurs manquantes pour les variables qui étaient accessibles dans les années antérieures de l'EDTR. Pour d'autres, cependant, elle a eu une incidence faible ou nulle. Les nouvelles variables du logement pour lesquelles la seule source de renseignements était l'enquête sur le loyer ont donné très peu d'imputations longitudinales valides. C'est surtout parce que l'enquête sur le loyer n'était accessible que pour les ménages du cycle 4 qui étaient considérés comme locataires. Sur ce faible nombre, seuls ceux ayant des données valides dans l'Enquête sur le loyer et ceux dont le code postal concordait ont alors fait l'objet d'une imputation.

Pour l'année de référence 2003, l'imputation longitudinale a été effectuée sur toutes les variables du logement, pour les deux panels. On a imputé les renseignements manquants pour les ménages à l'aide des renseignements du même ménage de l'année de référence 2002 si le code postal était demeuré le même. Si le mode d'occupation était connu pour l'année de référence 2003, alors les renseignements manquants n'ont été imputés que si le mode d'occupation concordait également.

Le tableau 1 indique le pourcentage pondéré des valeurs manquantes pour chaque variable. Là où l'imputation longitudinale était possible, le pourcentage pondéré des valeurs manquantes après imputation longitudinale est également indiqué. Puisque l'imputation longitudinale n'était possible que pour certaines variables en 2002, certaines valeurs sont toujours manquantes. Certaines variables ne s'appliquent qu'à certains ménages (p. ex., propriétaires-occupants et locataires, propriétaires-occupants seulement, et locataires seulement); par conséquent, un grand nombre n'ont pas le même dénominateur. Ceux pour qui il faut imputer une valeur autre que « sans objet » constituent les différents dénominateurs. Pour les définitions des variables, voir l'annexe 1.

4. L'EDTR utilise les ménages qui sortent de l'EPA en décembre de l'année précédant l'année de référence de l'EDTR et en janvier de l'année de référence de l'EDTR.

Tableau 1 – Effet de l'imputation longitudinale pour 2002 et 2003

Variable	Groupe de ménages nécessitant l'imputation pour cette variable	Valeurs manquantes (%)				Nombre de ménages soumis à l'imputation longitudinale	
		Avant		Après			
		2002	2003	2002	2003	2002	2003
dwldet25	Propriétaires et locataires	6,9	7,4	0,6	0,5	1646	1871
dwtenr25	Propriétaires et locataires	6,6	7,4	0,4	0,5	1637	1908
dwltyp25	Propriétaires et locataires	6,9	7,4	0,6	0,5	1646	1871
rooms25	Propriétaires et locataires	6,9	8,0	2,7	0,6	1215	2031
opbu25	Propriétaires et locataires	8,1	9,4	8,0	0,8	9	2649
rnre25	Locataires	8,8	10,8	8,7	0,9	7	798
rnpk25	Payent un loyer	8,0	9,5	7,9	0,6	6	689
rnht25	Payent un loyer	8,0	9,5	7,9	0,6	4	689
rnwa25	Payent un loyer	8,0	9,5	7,9	0,6	5	689
rnec25	Payent un loyer	8,0	9,5	7,9	0,6	2	689
rntv25	Payent un loyer	8,0	9,5	7,9	0,6	1	689
rnfg25	Payent un loyer	8,0	9,5	7,9	0,6	5	689
rnst25	Payent un loyer	8,0	9,5	7,9	0,6	6	689
rnwd25	Payent un loyer	8,0	9,5	7,9	0,6	3	689
rnfu25	Payent un loyer	8,0	9,5	8,0	0,6	0	689
rnno25	Payent un loyer	8,0	9,5	8,0	0,6	0	689
rentm25	Locataires	12,1	14,4	10,2	2,6	109	946
repa25	Propriétaires et locataires	7,0	7,8	s.o.[2]	0,6	s.o.	1987
heat25	Propriétaires et locataires	11,6	12,8	s.o.	2,3	s.o.	2703
heatg25	Propriétaires et locataires	11,6	12,8	s.o.	2,3	s.o.	2703
opfm25	Propriétaires et locataires	3,4[1]	7,0	s.o.	0,7	s.o.	1716
mortg25	Propriétaires	7,7	8,1	s.o.	0,8	s.o.	1426
cond25	Propriétaires	6,5	7,1	s.o.	0,6	s.o.	1267
mortgn25	Paient une hypothèque	9,5	10,3	s.o.	1,9	s.o.	861
mortgm25	Paient une hypothèque	29,4	29,5	s.o.	8,6	s.o.	2297
prtxm25	Propriétaires	25,7	26,7	s.o.	5,4	s.o.	4546
condm25	Membres d'un condominium	13,8	15,7	s.o.	5,4	s.o.	95
rnbs25	Locataires	9,6	11,3	s.o.	1,2	s.o.	834

1. Dans l'année de référence 2002, seuls les ménages (de l'extérieur du Nouveau-Brunswick) qui ont un code postal indiquant une région rurale (deuxième chiffre du code postal = 0) se sont vu demander si le ménage exploitait une ferme dans sa propriété (OPFM25); tous les autres sont fixés à « sans objet ». Le vrai taux de non-réponse est de 16,6 %, avec dénominateur de 11 018.
2. Imputation longitudinale non disponible en 2002.

Imputation transversale par donneur

L'imputation par donneur consiste à définir un groupe de ménages partageant plusieurs caractéristiques avec le ménage soumis à l'imputation, puis à choisir l'un d'entre eux comme donneur. La valeur déclarée par le donneur remplace la zone manquante du ménage soumis à l'imputation. Ainsi, l'EDTR a imputé le revenu d'un particulier en choisissant le revenu d'une personne du même sexe et du même niveau d'instruction vivant dans la même province, faisant partie de la même tranche d'âge et ayant le même niveau d'instruction, le même type d'emploi (salarié, travailleur indépendant) et la même profession.

L'imputation par donneur a été effectuée pour les autres variables et pour tous les ménages où l'imputation longitudinale était impossible. L'imputation par donneur a été divisée en sept modules. Il s'agissait de faciliter l'utilisation des variables imputées dans un module ou comme variables auxiliaires d'appariement dans un autre. Les enregistrements imputés peuvent aussi servir de donneurs dans des cycles successifs d'imputation.

Chaque module a des listes différentes de variables auxiliaires et d'appariement. Les variables auxiliaires servent à créer des groupes d'imputation[5], alors qu'on utilise les variables d'appariement dans une fonction de pointage pour chercher le donneur le plus approprié dans un groupe d'imputation. Dans certains cas, les variables auxiliaires ont créé des groupes de receveurs[6] sans groupes de donneurs correspondants[7]; en l'occurrence, les variables auxiliaires ont été regroupées. Par exemple, dans tous les modules sauf un, la variable « province » a été regroupée en cinq régions. Lorsque le regroupement n'était pas suffisant pour créer des donneurs possibles pour tous les receveurs, on a transformé les variables auxiliaires en variables d'appariement. Dans chaque module, les variables figurant sur la liste des « variables à imputer » ont aussi servi de variables d'appariement lorsqu'une valeur pour cette variable était disponible. Par exemple, dans le module des types de logement et des modes d'occupation, si le type de logement était manquant et que le mode d'occupation ne l'était pas, on a tenté de trouver un donneur ayant une valeur concordante pour le mode d'occupation. Dans chaque groupe d'imputation, les donneurs étaient évalués par une fonction de pointage. Cette fonction de pointage était basée sur la fonction de pointage utilisée dans l'imputation du revenu pour l'EDTR :

$$s(X,Y) = \sum_{k=1}^{K} p_k \, \mathrm{I}(X_k, Y_k) \text{, où } \mathrm{I}(X_k, Y_k) = \begin{cases} 1 & \text{si } X_k = Y_k \\ 0 & \text{autrement.} \end{cases}$$

5. Les donneurs et les receveurs (enregistrements nécessitant une imputation) sont séparés en groupes d'imputation. Chaque receveur au sein d'un groupe d'imputation est comparé à chaque donneur au sein du même groupe d'imputation.
6. Les groupes de receveurs sont les groupes, fondés sur les variables auxiliaires, de chaque module d'enregistrements nécessitant une imputation.
7. Les groupes de donneurs sont les groupes, fondés sur les variables auxiliaires, de chaque module d'enregistrements de donneurs possibles.

Noter que p_k est un poids nous permettant d'attribuer plus ou moins d'importance à la variable d'appariement k. Il a été décidé que toutes les variables d'appariement ont la même importance, d'où $p_k=1$. X_k est la valeur de la variable k du receveur et Y_k la valeur de la variable k du donneur.

On choisit alors un donneur en fonction de ce pointage. C'est le donneur ayant le plus haut pointage qui est choisi. S'il y a plusieurs donneurs ayant la valeur du pointage élevé, alors un donneur est choisi au hasard dans la liste. Pour les modules où il n'y a pas de variables d'appariement, un donneur est choisi au hasard dans l'ensemble du groupe d'imputation.

Les paragraphes suivants sont un résumé de chaque module d'imputation. Chaque tableau renferme une liste de variables à imputer avec les listes des variables auxiliaires et d'appariement. Il a été décidé que, pour les modules où il fallait imputer des variables numériques, celles qui se trouvaient dans le centile supérieur seraient exclues du bassin des donneurs. Cela se produit dans quatre modules différents; le module des propriétaires et des locataires, le module des prêts hypothécaires, le module des impôts fonciers et des frais de copropriété et, enfin, le module des loyers. C'était en partie pour diminuer le risque d'imputer une valeur supérieure au revenu du ménage pour des montants comme le loyer ou le paiement hypothécaire.

Module des types de logement et des modes d'occupation

Le module des types de logement et des modes d'occupation est le premier appliqué. De cette façon, chaque module appliqué utilise soit le type de logement soit le mode d'occupation comme variable auxiliaire ou d'appariement. Noter que le nombre de codes est indiqué entre crochets.

Enregistrements dans le module :	Tous les ménages
Variables à imputer :	Type de logement détaillé, type de logement, mode d'occupation
Variables auxiliaires :	Province (10), groupe de taille de région urbaine (3), nombre de personnes dans le ménage au 31 décembre (3)

Module des propriétaires et des locataires

Le module des propriétaires et des locataires vient ensuite. Les variables à imputer dans ce module sont les variables pour lesquelles tous les enregistrements exigent une valeur autre que « sans objet ». Dans le choix des donneurs possibles dans ce module, on a exclu tous ceux ayant une valeur de 9 ou plus pour le nombre de chambres à coucher.

Enregistrements dans le module :	Tous les ménages
Variables à imputer :	Nombre de chambres à coucher, réparations nécessaires au logement, combustible principal utilisé pour le

	chauffage, groupe de combustibles principaux utilisés pour le chauffage, le ménage exploite une ferme, le ménage exploite une entreprise
Variables auxiliaires :	Province (5), groupe de taille de région urbaine (3), groupe de types de logement (3), nombre de personnes dans le ménage au 31 décembre (3)
Variables d'appariement :	Propriété du logement (2)

Module des propriétaires

Vient ensuite le module des propriétaires. Ce module utilise les variables imputées des deux modules précédents à titre de variables auxiliaires. Tous les ménages qui ont une valeur de mode d'occupation indiquant qu'ils sont propriétaires du logement entrent dans ce module. Tous les autres enregistrements ayant des valeurs manquantes sont fixés à « sans objet ».

Enregistrements dans le module :	Propriétaires
Variables à imputer :	Prêt hypothécaire sur le logement, plus d'un prêt hypothécaire sur le logement, le logement est un logement en condominium
Variables auxiliaires :	Province (5), groupe de taille de région urbaine (3), groupe de types de logement (2), nombre de chambres à coucher (4)
Variables d'appariement :	Nombre de personnes au 31 décembre (3)

Module des prêts hypothécaires

Le module des prêts hypothécaires suit le module des propriétaires. C'est pour faire en sorte que seuls les ménages qui ont un prêt hypothécaire sur le logement (MORTG25) entrent dans ce module. Tous les autres pour lesquels les valeurs sont manquantes sont fixés à « sans objet ». Dans le choix des donneurs possibles dans ce module, tous ceux qui avaient une valeur pour un paiement hypothécaire mensuel courant supérieur à 3 000 $ ont été exclus.

Enregistrements dans le module :	Propriétaires ayant un prêt hypothécaire sur le logement
Variables à imputer :	Montant mensuel des paiements hypothécaires courants
Variables auxiliaires :	Province (5), type de logement (2), nombre de chambres à coucher (4), plus d'un prêt hypothécaire sur le logement (2)

Module des impôts fonciers et des frais de copropriété

Le module des impôts fonciers et des frais de copropriété peut être appliqué après le module de l'hypothèque. Tous les propriétaires-occupants entrent dans ce module. Tous les autres pour lesquels il y a des valeurs manquantes sont fixés à « sans objet ». Puisque la variable « logement en condominium » (COND25) est l'une des variables d'appariement des frais de copropriété, les propriétaires-occupants qui ne font pas partie d'un condominium verront leurs frais mensuels de copropriété remplacés, par imputation, par « sans objet ». En choisissant les donneurs possibles dans ce module, on a exclu tous ceux qui avaient un montant mensuel d'impôts fonciers supérieur à 500 $ ou une valeur de frais mensuels de copropriété supérieure à 800 $.

Enregistrements dans le module :	Propriétaire
Variables à imputer :	Montant mensuel des impôts fonciers, frais mensuels de copropriété
Variables auxiliaires :	Province (5), groupe de types de logement-propriétaires (2), logement en condominium (2)
Variables d'appariement :	Groupe de taille de région urbaine (3), nombre de chambres à coucher (4)

Module des locataires

Le module des locataires doit suivre le module des propriétaires et des locataires. Tous les ménages qui ont une valeur de mode d'occupation indiquant qu'ils sont locataires entrent dans ce module. Pour tous les autres pour lesquels des valeurs sont manquantes, on indique « sans objet ». Ceux qui sont locataires mais ne paient pas de loyer mensuel (RENTM25=0), ont des variables comme « l'électricité, le stationnement, etc., sont inclus dans le loyer » ont une valeur qui est fixée à « sans objet ».

Enregistrements dans le module :	Locataires
Variables à imputer :	… inclus dans le loyer (plusieurs éléments), la totalité ou une partie des meubles sont inclus, raison du loyer réduit, loyer calculé en fonction du revenu
Variables auxiliaires :	Province (5), groupe de taille de région urbaine (3), groupe de types de logement-locataires (3)
Variables d'appariement :	Nombre de chambres à coucher (4)

Module des loyers

Enfin, le dernier module d'imputation est celui des loyers. Tous les ménages qui payent un loyer mensuel entrent dans ce module. Tous les autres ménages pour lesquels des valeurs sont manquantes se voient imputer la valeur « sans objet ». En choisissant les

donneurs possibles dans ce module, on a exclu tous ceux qui avaient un montant mensuel de loyer supérieur à 1 500 $.

Enregistrements dans le module :	Locataires
Variables à imputer :	Montant mensuel du loyer
Variables auxiliaires :	Province (5), groupe de types de logement-locataires (3), nombre de chambres à coucher (4), raison du loyer réduit (2)
Variables d'appariement :	……inclus dans le loyer (plusieurs éléments), nombre de chambres à coucher (4)

Il y a eu une vaste gamme de pourcentages de valeurs manquantes avant l'imputation par donneur pour l'année de référence 2002. Certaines variables ont un pourcentage relativement élevé de valeurs manquantes, de sorte que les bassins de donneurs sont restreints. Le tableau 1 indique le pourcentage pondéré des valeurs manquantes avant l'imputation transversale pour les années de référence 2002 et 2003.

Évaluation des méthodes d'imputation pour l'année de référence 2002

Pour évaluer le succès des méthodes d'imputation, il a fallu examiner les fréquences avant et après imputation pour les variables nominales et les moyennes avant et après imputation pour les variables continues. Les variations de moins de 1 % des pourcentages avant et après imputation et une variation de moins de 6 % des moyennes avant et après imputation ont été comptées comme résultat positif. On a aussi effectué une analyse de validation par recoupement à l'aide de 1 000 enregistrements de données complètes. Ces enregistrements ont été imputés. Puis, on a comparé les données initiales aux données imputées.

Après l'imputation à l'aide des méthodes décrites plus haut, plusieurs analyses ont été effectuées. Ces analyses avaient pour objet d'établir si les données obtenues étaient ou n'étaient pas semblables aux données d'origine. Les fréquences avant et après imputation ont été reprises sur toutes les variables catégoriques imputées. À titre d'exemple, le tableau 2 renferme des pourcentages avant et après imputation pour le type de logement. Avant l'imputation par donneur, 4 % des ménages étaient sans cette variable. Le tableau 2 indique que les pourcentages avant et après imputation sont très semblables. Toute différence qui existe est inférieure à 1 %, ce qui est très raisonnable.

Tableau 2 : Pourcentage avant et après imputation par donneur pour le type de logement

DWLDET25 Code fixé	Avant imputation par donneur (sans les valeurs manquantes dans le dénominateur)	Après imputation par donneur
01 Maison individuelle non attenante	66,5	65,9
02 Double	3,53	3,50
03 En rangée ou en terrasse	4,13	4,14
04 Duplex	4,02	4,02
05 Appartement de faible hauteur	14,0	14,4
06 Tour d'appartements	5,01	5,17
07 Hôtel, maison de chambres, camping	0,26	0,27
08 Maison mobile	2,42	2,42
09 Autres	0,15	0,17

Il est également important de regarder de près les variables pour lesquelles le pourcentage de valeurs manquantes est beaucoup plus élevé. Ainsi, le tableau 3 présente les pourcentages avant et après imputation pour la source principale de chauffage. Avant l'imputation par donneur, cette variable était manquante pour 24 % des ménages. Là encore, cependant, le tableau 3 indique que la différence des pourcentages est toujours inférieure à 1 %. Ces résultats sont typiques de ce que l'on a observé pour toutes les variables catégoriques.

Tableau 3 : Pourcentage avant et après imputation par donneur pour la source principale utilisée pour le chauffage

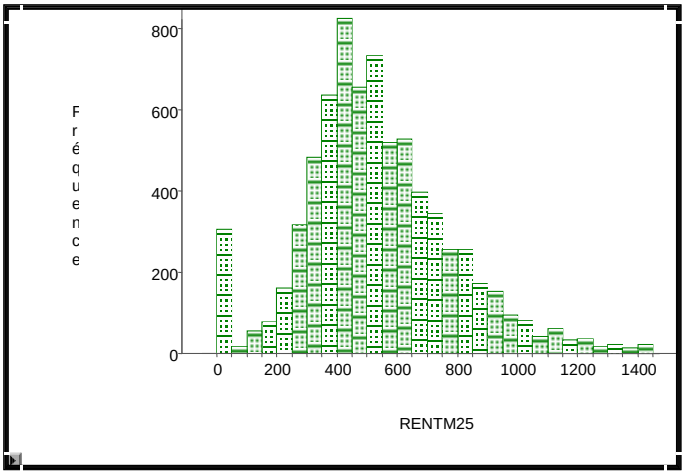
HEAT25 Code fixé	Pourcentage avant imputation par donneur (sans les valeurs manquantes dans le dénominateur)	Pourcentage après imputation par donneur
01 Mazout ou autre combustible liquide	17,1	16,7
02 Gaz canalisé (gaz naturel)	41,0	41,8
03 Gaz en bouteille (propane)	0,90	0,88
04 Électricité	33,0	33,5
05 Bois	7,39	7,07
06 Autres	0,06	0,06

Pour les variables continues, on a examiné les moyennes avant et après imputation. Cela a donné une bonne indication quant à la question de savoir si les données avaient ou non changé pour la peine. À titre d'exemple, voyons les moyennes avant et après imputation du montant du loyer mensuel (RENTM25) et des impôts fonciers (PRTXM25). La moyenne avant imputation de RENTM25 était de 604,12 $, et cette variable était manquante pour 29 % des ménages; après imputation, la moyenne était tombée

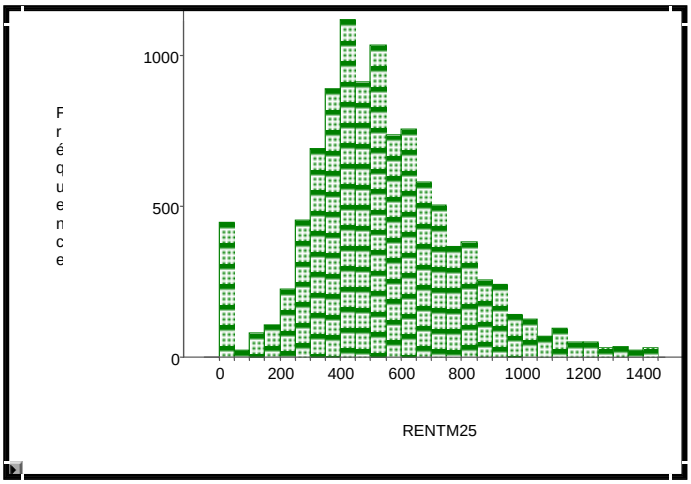
légèrement à 589,53 $. C'est une diminution de 2 %. La moyenne avant imputation PRTXM25 était de 160,68 $, et cette variable était manquante pour 35 % des ménages; après imputation, la moyenne a encore tombé légèrement, à 154,53 $. C'est une diminution de 4 %. Ces résultats sont caractéristiques de toutes les variables continues. Les réductions vont de 2 % à 6 %. Il faut prévoir une baisse de la moyenne, surtout parce que le centile supérieur du bassin de donneurs a été exclu.

On a aussi effectué un contrôle visuel. On a porté les variables continues sur un histogramme pour comparer le bassin des donneurs avec le fichier nouvellement imputé. Pour simplifier la comparaison, il a fallu exclure le centile supérieur des données de l'histogramme du fichier nouvellement imputé. Les histogrammes des graphiques 1 et 2 montrent que la distribution avant et après imputation est très semblable. On a ainsi comparé les histogrammes de toutes les variables continues, et l'exemple choisi est typique de ce que nous avons observé. C'est une bonne indication que les méthodes d'imputation choisies étaient bonnes, en ce qu'elles n'ont pas modifié les données pour la peine.

Graphique 1 : Histogramme du bassin des donneurs pour RENTM25



Graphique 2 : Histogramme du fichier après imputation (RENTM25<1500) pour RENTM25



À titre d'analyse complémentaire, on a effectué une validation par recoupement. Une validation par recoupement est une comparaison, appliquée aux ménages pour lesquels nous avons des données complètes, entre les valeurs imputées par les méthodes décrites plus haut et les valeurs déclarées. Une validation par recoupement est utile parce qu'elle simule le processus d'imputation d'une manière qui facilite la vérification de son exactitude. Un ensemble de 1 000 enregistrements pour lesquels il y avait des données complètes a été soumis à l'imputation par les méthodes décrites plus haut. Les données initiales de ces enregistrements ont été comparées aux données nouvellement imputées. Cette comparaison par validation par recoupement n'a été effectuée que pour les

variables catégoriques. Une valeur incorrecte a été imputée entre 10 % et 40 % du temps. La validation par recoupement a été effectuée plusieurs fois et les résultats qui suivent sont caractéristiques de ce qui a été observé.

Lorsqu'on considère les résultats de la validation par recoupement, il faut retenir que l'imputation s'effectue en modules séquentiels. Le module des types de logement et des modes d'occupation est appliqué d'abord, si bien que les 11 % de ménages dont le mode d'occupation est incorrect seront alors incorrects pour la plupart des autres modules. Ainsi, ceux qui sont considérés comme propriétaires sont les seuls pour lesquels serait imputée une valeur pour le prêt hypothécaire sur le logement (MORTG25); tous les autres se verraient imputer la valeur « sans objet ».

Pour fins de comparaison, on a effectué des imputations aléatoires pour les 1 000 mêmes ménages. On a imputé une valeur aléatoire à chaque variable, en fonction des fréquences initiales.

Puisque l'imputation est séquentielle, l'imputation d'un module s'effectue avant l'imputation d'un autre module. Cela génère deux façons de procéder avec l'analyse. L'imputation aurait pu utiliser les valeurs créées pendant la validation par recoupement dans le fichier précédent, ou aurait pu utiliser les nouvelles valeurs qui ont été attribuées aléatoirement dans le fichier précédent. Pour maintenir une analyse logique des variables imputées, nous utilisons les données de la validation par recoupement. De cette façon, les répondants qui étaient au départ des locataires sont demeurés locataires, si bien que les variables applicables aux locataires, comme « chauffage inclus dans le loyer » (RNHT25) auraient une valeur imputée valide plutôt qu'une valeur « sans objet », comme ce serait le cas si la valeur imputée aléatoirement dans un module antérieur en faisait des propriétaires. Par conséquent, ceux qui sont considérés comme incorrectement imputés par la méthode aléatoire sont fondés sur le même sous-ensemble de la population que ceux de la validation par recoupement.

Il ressort de ces résultats qu'il y a une différence entre les variables pour lesquelles l'imputation longitudinale a donné un résultat positif et les variables pour lesquelles il n'y a pas eu de résultat positif. Par exemple, l'imputation aléatoire a été effectuée sur DWTENR25. Elle l'a été plusieurs fois et, typiquement, les données ont été imputées incorrectement 40 % du temps, ce qui est beaucoup plus qu'avec les méthodes d'imputation exposées ici. À titre d'exemple d'imputation aléatoire d'une variable où l'imputation longitudinale n'était pas disponible, on peut considérer la variable MORTG25. Les résultats révèlent que la méthode d'imputation ne donne qu'une légère amélioration par rapport à la méthode d'imputation aléatoire. Typiquement, la valeur incorrecte a été imputée 43 % du temps. Il semblerait, au vu de ces résultats, que la méthode d'imputation longitudinale est bien meilleure que l'imputation aléatoire, mais que la méthode d'imputation transversale par donneur n'est que légèrement meilleure. Le tableau 4 est un tableau de la différence en pourcentage entre les données initiales et les données imputées pour la validation par recoupement, avec les résultats typiques des imputations aléatoires.

Tableau 4 : Pourcentage de catégories incorrectement imputées pour deux stratégies d'imputation

Variable	Imputation transversale (%)	Imputation aléatoire (%)
Dwtenr25	11	40
Dwldet25	20	54
Mortg25	40	43
Repa25	35	36
Heat25	39	67
Heatg25	39	67
Obpu25	16	16
Mortgn25	38	38
Cond25	13	17
Rnpk25	17	18
Rnwa25	16	19
Rnec25	19	21
Rnht25	20	22
Rntv25	15	15
Rnfg25	16	20
Rnst25	16	21
Rnwd25	16	17
Rnfu25	13	13
Rnno25	12	12
Rnre25	17	18
Rnbs25	17	19

Conclusion

Étant donné la non-réponse initialement élevée pour les variables du logement, il a été utile de mettre en place une méthodologie d'imputation. L'un des points forts de la procédure d'imputation décrite est que le résultat final est un ensemble de données sans valeurs manquantes et avec convergence interne des enregistrements imputés.

Il ressort de ces analyses que les méthodes d'imputation choisies ont donné d'excellents résultats. Les fréquences avant et après imputation pour les variables catégoriques et les moyennes avant et après imputation pour les variables continues ne font ressortir que des variations relativement faibles de moins de 1 % et 6 %, respectivement. Comme confirmation supplémentaire, les histogrammes avant et après imputation se ressemblent beaucoup. Cela confirme que la procédure d'imputation a bien fonctionné pour préserver les distributions à une variable.

Une question plus complexe est celle de savoir dans quelle mesure les distributions à une variable se sont trouvées modifiées. Pour répondre à cette question, nous avons examiné

les estimations pour certains groupes déterminés. Ainsi, puisque l'âge n'est pas explicitement pris en compte dans l'imputation, on risque de perdre la corrélation existante entre l'âge et le mode d'occupation. Selon des analyses effectuées sur les données sur l'âge du membre du ménage le plus âgé, le processus d'imputation ne change rien aux estimations de propriétaires-occupants selon l'âge. La distribution de l'âge du plus vieux membre du ménage chez les propriétaires-occupants est très semblable avant et après imputation. L'âge moyen de la personne ayant le plus haut revenu net dans les ménages ayant un paiement hypothécaire mensuel imputé par opposition à non imputé est très semblable. C'est une preuve de plus que la méthode d'imputation est très bonne. Puisqu'il est impossible de vérifier toutes les variables pour lesquelles il pourrait y avoir une corrélation, on ne peut être sûr qu'aucune des estimations pour les groupes particuliers n'est touchée par ce qui semble être, selon une analyse de validation par recoupement, un taux d'erreur élevé.

En somme, bien que la mesure du taux d'erreur pour certaines variables soit élevée à cause de la faiblesse des modèles d'imputation sous-jacents, l'incidence sur les distributions à une variable et certaines distributions à variables multiples semble plutôt limitée. Autrement dit, bien que les taux de non-réponse aient été au départ élevés, on a pu remplacer les valeurs manquantes sans causer de changements importants à l'ensemble des estimations.

Annexe 1

Liste des variables

Variable	Description de la variable	Module	Groupe de ménages pour lesquels cette variable doit être imputée
dwldet25	Type de logement détaillé	Type de logement/ mode d'occupation	Propriétaires et locataires
dwtenr25	Propriété du logement	Type de logement/ mode d'occupation	Propriétaires et locataires
dwltyp25	Type de logement	Type de logement/ mode d'occupation	Propriétaires et locataires
rooms25	Nombre de chambres à coucher	Propriétaires/locataires	Propriétaires et locataires
Opbu25	Entreprise dans la propriété	Propriétaires/locataires	Propriétaires et locataires
Repa25	Réparations nécessaires	Propriétaires/locataires	Propriétaires et locataires
Heat25	Combustible principal utilisé pour le chauffage	Propriétaires/locataires	Propriétaires et locataires
heatg25	Groupe de combustibles principaux utilisés pour le chauffage	Propriétaires/locataires	Propriétaires et locataires
Opfm25	Ferme dans la propriété	Propriétaires/locataires	Propriétaires et locataires
mortg25	Prêt hypothécaire sur le logement	Propriétaires	Propriétaires
Cond25	Logement en condominium	Propriétaires	Propriétaires
mortgn25	Plus d'un prêt hypothécaire	Propriétaires	Débiteurs hypothécaires
mortgm25	Paiement hypothécaire mensuel	Prêt hypothécaire	Débiteurs hypothécaires
prtxm25	Impôts fonciers mensuels	Impôts fonciers et frais de copropriété	Propriétaires
condm25	Frais mensuels de copropriété	Impôts fonciers et frais de copropriété	Membres d'un condominium
Rnre25	Loyer réduit et raison	Locataires	Locataires
Rnpk25	Stationnement inclus dans le loyer	Locataires	Payent un loyer
Rnht25	Chauffage inclus dans le loyer	Locataires	Payent un loyer
rnwa25	Eau incluse dans le loyer	Locataires	Payent un loyer
Rnec25	Électricité incluse dans le loyer	Locataires	Payent un loyer
Rntv25	Câble inclus dans le loyer	Locataires	Payent un loyer
Rnfg25	Réfrigérateur inclus dans le loyer	Locataires	Payent un loyer

Rnst25	Cuisinière incluse dans le loyer	Locataires	Payent un loyer
rnwd25	Laveuse et sécheuse incluses dans le loyer	Locataires	Payent un loyer
Rnfu25	Meubles inclus dans le loyer	Locataires	Payent un loyer
Rnno25	Pas de commodités incluses dans le loyer	Locataires	Payent un loyer
Rnbs25	Loyer calculé en fonction du revenu	Locataires	Locataires
rentm25	Loyer mensuel	Loyer	Locataires