Techniques d'enquête, décembre 1976
Vol. 2, nº 2
Statistique Canada, nº 12-001-X au catalogue

Enquête sur les logements de 1974

H. Hofmann[1]

Résumé

L'Enquête de 1974 sur les logements, qui a été menée par Statistique Canada au cours de l'automne 1974, était parrainée par la Société centrale d'hypothèques et de logement. Statistique Canada était chargé d'élaborer et de mettre en œuvre toutes les phases de l'enquête, y compris la production de bandes magnétiques de données « sans erreur ». Pour sa part, la Société était responsable de l'établissement des objectifs et des besoins en matière de données, ainsi que de l'analyse des résultats.

Ce rapport est une modification du sommaire préparé par l'équipe responsable, une fois le projet terminé; il consiste en un exposé général de l'enquête et de l'apport de Statistique Canada.

Mots-clés : Enquête sur les logements; microdonnées; plan de sondage.

1 Objectifs de l'enquête

Pour mieux comprendre certains marchés du logement urbain, les objectifs de recherche de la Société centrale d'hypothèques et de logement (SCHL) comprenaient notamment :

i) une description des besoins en logement par niveaux de besoins;

ii) la détermination des raisons pour lesquelles les ménages ont modifié leur consommation de logements et, ce faisant, ce qui a motivé leur choix d'un logement en particulier;

iii) une indication du processus de détérioration ou de revitalisation des logements et des quartiers;

iv) une description des caractéristiques, des coûts et des charges financières du logement, avec le temps.

L'objectif pour Statistique Canada était de fournir à la SCHL les renseignements nécessaires pour atteindre les objectifs de recherche précisés. Comme l'ensemble de données particulier requis n'était pas disponible auprès des sources existantes, ou d'une combinaison de sources, une enquête devait être réalisée à intervalles au moyen du même échantillon de base de logements. Pour fournir des données longitudinales sur le logement au moment du premier cycle de l'enquête, l'échantillon de base devait être sélectionné à partir du Recensement du Canada de 1971, et

1. H. Hofmann, Division des méthodes d'enquêtes auprès des ménages.

des éléments de données particuliers pour certains logements recueillis en 1971 devaient être couplés avec les renseignements recueillis en 1974.

Encore une fois à la suite de l'enquête, la SCHL a décidé d'utiliser les renseignements recueillis en 1974 pour produire une publication sur les statistiques sur le logement.

2 Plan d'échantillonnage

La présente section donne une description générale du plan d'échantillonnage pour le premier cycle de l'enquête.

2.1 Généralités

Le plan d'échantillonnage de cette enquête reposait sur les quatre exigences suivantes, précisées par la SCHL :

A) Des données concernant les caractéristiques du ménage et du logement qui sont statistiquement fiables au niveau de certains marchés du logement urbain étaient requises à des moments précis. Ces données permettraient de mener une analyse transversale entre les cycles.

B) Malgré A), une analyse détaillée des données recueillies à propos de ces concepts devait être effectuée en accordant plus d'importance aux ménages à faible revenu qu'aux ménages à revenu moyen et élevé.

C) Certains sous-ensembles de ces données, statistiquement fiables au niveau de certaines régions urbaines, étaient requis aux fins de comparaison avec le temps. Ces sous-ensembles permettraient de mener une analyse longitudinale entre les cycles.

D) Au premier cycle, un ensemble de données secondaire (de préférence le Recensement de 1971) concernant les caractéristiques des ménages et des logements était requis. Ces données permettraient de mener une analyse longitudinale entre les cycles.

Ces exigences générales ont été résumées conjointement par Statistique Canada et la SCHL pour former un plan d'échantillonnage incorporant les caractéristiques suivantes :

i) Une enquête-ménage serait menée trois fois, le premier cycle débutant à l'automne 1974 et permettant de recueillir des renseignements sur le même ensemble de variables à chaque cycle sur un ensemble commun de ménages; en outre, l'échantillonnage serait plus lourd à l'extrémité inférieure de l'échelle de revenu des ménages.

ii) À chaque cycle après le premier, l'échantillon de logements serait mis à jour par un échantillon supplémentaire de logements construit depuis le cycle précédent. Cela ferait en sorte que l'échantillon total à chaque cycle soit représentatif de la population des logements existants au moment de ce cycle.

iii) L'exigence D) serait satisfaite en couplant les caractéristiques des ménages et des logements mesurées au premier cycle avec les caractéristiques des ménages et des logements mesurées par le Recensement de 1971 pour le même ensemble de logements.

2.2 Population

La population d'intérêt de cette enquête-ménage était celle des logements privés existants pendant la période de référence (l'automne 1974) et situés dans les municipalités désignées des 23 plus grandes régions métropolitaines.

2.3 Bases de sondage

Une base de sondage est la liste des unités dans la population à un moment précis à partir de laquelle un échantillon peut être sélectionné. Compte tenu des exigences précisées par le commanditaire, trois bases de sondage ont été utilisées pour cette enquête. Les voici :

1) Le fichier du Recensement du Canada de 1971 des logements privés occupés (le 1er juin 1971).
2) Le fichier du Recensement du Canada de 1971 des logements privés vacants (le 1er juin 1971).
3) Les enregistrements sommaires de Statistique Canada des permis de bâtir délivrés.

Les bases de sondage 1 et 2 servent de base pour ce cycle et les cycles suivants de cette enquête, car elles constituent une liste de tous les logements privés existants avant le 1er juin 1971. Comme ces fichiers ont

été créés et stockés séparément, il a été nécessaire, pour diverses raisons, de les considérer comme deux bases de sondage indépendantes. Ces bases de sondage couvraient approximativement 85 % de la population des logements privés existants au premier cycle. La base de sondage 3 a permis d'inclure dans l'échantillon les logements construits entre le 1er juin 1971 et la période de référence, à savoir l'automne 1974.

Le fichier du Recensement du Canada de 1971 (2B) est essentiellement une liste mise à jour par ordinateur des logements privés dénombrés lors du Recensement du Canada de 1971 qui étaient occupés au moment du dénombrement. Les caractéristiques détaillées des logements et des ménages sont disponibles pour un tiers de l'échantillon des logements de la base de sondage, et chaque logement est identifié géographiquement au niveau du secteur de dénombrement (SD) du recensement.

Le fichier du Recensement du Canada de 1971 des logements privés vacants dresse la liste de tous les logements privés dénombrés lors du Recensement du Canada de 1971 qui étaient vacants au moment du dénombrement. La base de sondage est essentiellement stockée dans les registres des visites (RV) du recensement. Du fait que les logements sont vacants, aucune caractéristique détaillée n'était disponible pour ces logements dans la base de sondage. Chaque logement peut être identifié géographiquement au niveau du SD.

À la fin de chaque mois, toutes les municipalités considérées dans la population pour cette enquête (et d'autres) soumettent à Statistique Canada un enregistrement de tous les permis de bâtir délivrés dans cette municipalité pendant ce mois. Il s'agit d'une source permettant d'obtenir l'emplacement de tous les logements construits depuis le Recensement du Canada de 1971. Pour chaque permis, le type et l'emplacement de la construction prévue ainsi que le nombre de logements qu'elle accueillera sont déclarés. Tous les logements dans la base de sondage peuvent être identifiés géographiquement au niveau de la municipalité.

2.4 Fiabilité

Comme mentionné à la section 2.1, les estimations statistiques produites pour chaque cycle de cette enquête (estimations transversales) ne devaient être significatives (fiables) qu'au niveau de chacune des 23 régions de l'enquête. Toutes les estimations importantes produites pour chacune des régions de l'enquête devaient donc avoir une variabilité qui n'était pas supérieure à une certaine valeur précisée au préalable, et il n'était pas nécessaire d'avoir des estimations pour tout niveau géographique inférieur (ou supérieur).

La mesure la plus adéquate pour préciser avec succès les critères de fiabilité est le coefficient de variation (α) d'une estimation ($\hat{X}$) qui exprime l'erreur-type de l'estimation sous forme de fraction ou de pourcentage de la valeur réelle.

2.5 Échantillonnage

2.5.1 Stratification

Pour fournir un plan de sondage plus efficace en ce qui a trait aux variables susceptibles d'être fortement corrélées avec les concepts à mesurer, les unités dans la base de sondage 1 ont été stratifiées comme suit dans 40 strates par i) mode d'occupation, ii) revenu total du chef du ménage (et de l'époux ou de l'épouse) et iii) âge du chef du ménage.

Mode d'occupation	Revenu du chef du ménage (et de l'époux ou de l'épouse)	Âge du chef du ménage
Propriétaire (ou en processus d'achat)	Moins de 5 000 $	Moins de 25 ans
Locataire	5 000 $ à 6 999 $	25 à 44 ans
	7 000 $ à 8 999 $	45 à 64 ans
	9 000 $ à 10 999 %	65 ans et plus
	11 000 $ et plus	

Comme aucune variable de stratification pertinente n'était disponible, la base de sondage 2 (et l'échantillon tiré de cette base) n'a pas été stratifiée.

Dans la base de sondage 3, les unités ont été stratifiées comme suit en 4 strates par i) type de construction et ii) période de délivrance du permis de bâtir.

Strate	Type de construction	Période de délivrance
11	Immeubles d'appartements et maisons en rangée	Octobre 1969 à mai 1971
12	Immeubles d'appartements et maisons en rangée	Juin 1971 à décembre 1973
21	Tous les autres types de logements	Juin 1970 à mai 1971
22	Tous les autres types de logements	Juin 1971 à décembre 1973

Ces strates ont été choisies principalement à des fins opérationnelles, pour permettre la sélection d'unités à partir de trois bases de sondage indépendamment de la sélection à partir des bases de sondage 1 et 2.

2.5.2 Taille d'échantillon

Les tailles d'échantillon nécessaires ont été calculées indépendamment pour chaque région d'enquête afin de fournir un coefficient de variation de 6 % ou moins pour les caractéristiques représentant 10 % ou plus de la population au niveau de la région d'enquête.

La taille de base a été déterminée en résolvant pour n_o la relation

$$\alpha(\hat{X}) = \frac{Q}{\sqrt{n_o P}} \quad \text{où } P = 0{,}10$$

Pour un α de 6 %, cela donne $n_o = 2\ 500$ pour chaque région d'enquête.

Cette taille de base a ensuite été corrigée en fonction de ce qui suit :

i) La correction de la population finie au moment t(≡ 1978) pour chaque région d'enquête;
ii) Le taux d'attrition pour les logements échantillonnés;
iii) Un taux de non-réponse estimé pour chaque région d'enquête;
iv) Un taux d'erreur non due à l'échantillonnage supplémentaire pour chaque région d'enquête.

L'application de ces quatre facteurs a donné lieu à une taille d'échantillon pour la région d'enquête r, $n^{(r)}$, qui a ensuite été répartie proportionnellement dans chacune des régions d'enquête pour un α de 6 %, alors que P = 0,10.

Tableau 1
Taille et répartition de l'échantillon pour $\alpha = 0{,}06$ lorsque P = 0,10

Région d'enquête (r)	Nom de la région d'enquête	Précision: α = 0,06, P = 0,10			
		$n^{(r)}$	Base de sondage 1 (f = 1)	Base de sondage 2 (f = 2)	Base de sondage 3 (f = 3)
01	Calgary	3 380	2 540	162	678
02	Charlottetown	1 324	1 026	32	266
03	Chicoutimi	3 238	2 512	78	648
04	Edmonton	3 388	2 600	108	680
05	Halifax	3 330	2 558	106	666
06	Hamilton	3 374	2 618	80	676
07	Kitchener	3 338	2 536	134	668
08	London	3 350	2 520	160	670
09	Montréal	3 432	2 524	220	688
10	Ottawa	3 388	2 656	54	678
11	Québec	3 402	2 558	164	680
12	Regina	3 324	2 526	132	666
13	St. Catharines	3 348	2 544	134	670
14	Saint-Jean	3 238	2 538	52	648
15	St. John's	3 210	2 518	50	642
16	Saskatoon	3 294	2 424	210	660
17	Sudbury	3 286	2 602	26	658
18	Thunder Bay	3 220	2 498	78	644
19	Toronto	3 410	2 618	110	682
20	Vancouver	3 396	2 608	108	680
21	Victoria	3 328	2 528	134	666
22	Windsor	3 330	2 584	80	666
23	Winnipeg	3 398	2 608	108	682
Total		74,726	57 294	2 520	14 962

2.5.3 Répartition de l'échantillon

Au sein de chaque base de sondage, l'échantillon de logements a été réparti dans chacune des strates et sélectionné indépendamment à partir de chaque strate de la manière suivante.

Fichier du Recensement du Canada de 1971 (2B) des logements privés occupés

L'échantillon sélectionné à partir de cette base de sondage a été stratifié en fonction des variables mises en évidence dans la section 2.5.1 de la manière suivante : pour une valeur donnée de P (et Q), calculer $n^{(r)}$; déterminer ensuite n_{oh}, de sorte que lorsqu'il est corrigé par la correction de la population finie au sein de chaque strate (h = 1, ..., 40) selon

$$n_h^{(r)} = \frac{n_{oh}^{(r)}}{1 + \frac{n_{oh}^{(r)}}{N_h^{(r)}}}$$

la valeur de $\left|\Sigma_{h=1}^{40} n_h^{(r)} - n^{(r)}\right|$ est un minimum, où $\alpha^{(r)}$ est le coefficient de variation r; $n^{(r)}$ est la taille de l'échantillon pour la région d'enquête r; $n_h^{(r)}$ est la taille de l'échantillon de la strate h dans la région r; $N_h^{(r)}$ est la taille de la population de la strate h dans la région r.

Cette procédure garantissait qu'un coefficient de variation uniforme puisse être escompté pour toutes les estimations des strates dans une région d'enquête donnée, même si le coefficient de variation variait d'une région à l'autre.

Fichier du Recensement du Canada de 1971 des logements privés vacants

Comme aucune stratification des logements dans cette base de sondage n'a été effectuée, aucune répartition de l'échantillon n'a été nécessaire.

Enregistrements de Statistique Canada des permis de bâtir délivrés

L'échantillon pour la base de sondage a été réparti proportionnellement.

2.5.4 Sélection de l'échantillon

Dans la base de sondage 1, un programme de sélection d'échantillons généralisée automatisée a été utilisé pour sélectionner un échantillon aléatoire simple stratifié (sans remise) de codes d'identification des logements à partir du fichier 2B du recensement. Ces codes d'identification ont ensuite été jumelés aux RV pour récupérer l'adresse. Pour environ 50 % des adresses de l'échantillon, ce jumelage avait déjà été réalisé précédemment dans le cadre du projet de Registre national des adresses (RNA) mené par le Secteur du recensement de Statistique Canada. Pour ces adresses, il n'a fallu que récupérer l'adresse à partir de ce fichier. Dans la base de sondage 2, un échantillon aléatoire systématique (sans remise) a été directement sélectionné manuellement à partir des RV. Dans la base de sondage 3, un échantillon aléatoire systématique (sans remise) a été sélectionné manuellement à partir des rapports sommaires des permis de bâtir.

2.5.5 Traitement de l'échantillon

Pour faciliter le contrôle et la mise à jour des logements sélectionnés pour tous les cycles de l'enquête, une liste de contrôle de toutes les unités d'échantillonnage a été créée. Cette liste centrale comprend pour

chaque logement l'identificateur de l'enquête (code de la région et numéro de ménage), les codes de base de sondage et de strates, les identificateurs du recensement (pour les bases de sondage 1 et 2), l'adresse municipale.

Pour créer cette liste centrale, toutes les unités sélectionnées ont été transcrites sur des formulaires préimprimés à partir desquels les renseignements ont ensuite été saisis au moyen de la technique « taper et numériser ». Cela a été suivi d'une étape de révision et de correction conçue précisément pour cette liste de contrôle.

2.5.6 Estimation

Toutes les estimations obtenues à partir des données recueillies dans cette enquête sont basées sur des enregistrements pondérés; les poids d'échantillonnage sont par la suite formés de l'inverse de la base de sondage corrigé en fonction de la non-réponse complète; aucune variable externe n'est disponible afin de préciser davantage ces estimations.

Des coefficients de variation sont aussi calculés afin de fournir des estimations de la variabilité des totaux, des moyennes et de la proportion estimés. De plus, des tables de variabilité d'échantillonnage brutes ont été construites pour servir d'indicateurs de la variabilité d'échantillonnage approximative.

3 Collecte des données

L'étape de la collecte des données a eu lieu pendant les mois d'octobre, de novembre et de décembre 1974 dans les 23 régions d'enquête. La section qui suit du rapport met en évidence certaines des principales caractéristiques de cette étape de collecte.

3.1 Dénombrement

Le dénombrement a été réalisé au moyen d'interviews en personne (sur place) dans chaque région d'enquête. On a demandé aux agents recenseurs de tenter d'obtenir une interview avec le chef du ménage ou son époux ou son épouse. Les interviews avec un autre membre du ménage (âgé de 18 ans ou plus) étaient permises uniquement s'il était possible de déterminer

que le chef du ménage et son époux ou son épouse seraient absents pendant toute la période de référence de l'enquête (2 mois et demi).

3.2 Visites de rappel

Un nombre raisonnable de visites de rappel a été suggéré pour tenter d'établir un contact initial ou de terminer un questionnaire ayant été partiellement rempli précédemment. Tous les suivis pour obtenir des renseignements manquants ou supplémentaires devaient se faire en personne, à moins que l'agent recenseur n'ait obtenu la permission explicite du répondant en question d'obtenir l'information par téléphone. Les suivis étaient particulièrement importants pour les données sur le revenu et pour les données sur la dimension des pièces.

Les renseignements sur le revenu personnel devaient être obtenus par interview sur place de chaque bénéficiaire d'un revenu au sein du ménage. Si cela n'était pas possible parce que les membres ne pouvaient pas être interrogés directement, une réponse par personne interposée devait être permise pour tous les membres du ménage. Dans tous les cas, les données sur le revenu pour le chef du ménage et l'époux ou l'épouse devaient être considérées plus importantes que les données sur le revenu pour les autres membres du ménage.

Les renseignements sur les dimensions des pièces devaient être obtenus en demandant au répondant de fournir les renseignements à partir de plans ou d'autres documents ou en mesurant la longueur et la largeur de chaque pièce. Des rubans à mesurer ont été fournis à cette fin, au besoin. Dans les cas où le répondant refusait ou était incapable d'accomplir cette tâche, au moment de l'interview, l'agent recenseur déposait un document sur lequel le répondant pouvait saisir les mesures à sa convenance. L'intervieweur effectuait un suivi (en personne ou par téléphone) pour récupérer ces données.

3.3 Procédures en cas de non-réponse

Si, après le nombre maximal de visites de rappel, le questionnaire n'avait pas été terminé, on demandait aux intervieweurs d'enregistrer le dernier rappel, en indiquant la date, l'heure et l'état d'achèvement sur la première page du questionnaire et la raison pour laquelle l'interview n'avait pas été terminée. Dans les cas où aucun renseignement n'avait été obtenu, on

a aussi demandé aux intervieweurs de remplir un formulaire « Non-interview » comportant des renseignements comme l'adresse du ménage, et s'ils étaient connus, le nom et le numéro de téléphone, le nom des intervieweurs, les dates et les heures de toutes les visites de rappel, et la raison de la non-interview. Ce formulaire était alors joint au questionnaire approprié et envoyé à leurs superviseurs respectifs. On demandait aux superviseurs des intervieweurs de vérifier chaque non-interview reçue et de tenter de la convertir en interview réussie.

Si toutes les tentatives échouaient, les documents étaient envoyés aux bureaux régionaux respectifs. En général, à ce stade, une mesure supplémentaire n'aurait pas donné lieu à une interview terminée; par conséquent, ces cas étaient établis comme étant des non-interviews.

3.4 Procédures de contrôle de la qualité

Pour maintenir un niveau de qualité adéquat des données obtenues pendant l'étape de collecte, un échantillon de tous les questionnaires remplis a été révisé une fois par semaine par le personnel de niveau supérieur des bureaux régionaux. Dans chaque questionnaire, les points clés étaient vérifiés et si l'on constatait que les réponses n'étaient pas enregistrées correctement ou étaient enregistrées de manière illisible, on communiquait avec l'intervieweur en question pour lui demander d'améliorer la qualité du travail; tous ses questionnaires étaient examinés et corrigés au besoin.

De plus, un échantillon de 3 % de tous les ménages interrogés faisait l'objet d'une deuxième communication par téléphone par le personnel du bureau régional qui posait un minimum de questions pour déterminer si, en fait, un intervieweur avait communiqué avec le ménage.

4 Traitement des données

Le produit de cette enquête était un ensemble de fichiers de microdonnées « nettoyées » comprenant des éléments de données prédéterminés par la division cliente. Le système qui générait ces fichiers comprenait une série d'étapes de traitement des données manuelles et automatisées; la philosophie de base à l'origine de la conception de ce système était d'intégrer suffisamment de généralités pour garantir un système de longue

durée pour les cycles supplémentaires anticipés de cette enquête. Certaines techniques et méthodes comme l'entrée de données par reconnaissance optique de caractères, les bases de données, les dictionnaires de données et une procédure de détection et de correction d'erreurs automatisées (GEISHA)[2] ont été utilisées. La section qui suit présente un aperçu de chacune des étapes de traitement.

4.1 Préparation du questionnaire

Lorsqu'on les avait reçus des bureaux régionaux, tous les documents étaient vérifiés par rapport aux feuilles de contrôle principales et groupés par lots. Tous les documents faisaient l'objet d'une vérification de l'assurance de la qualité limitée pour s'assurer que toutes les entrées requises étaient dans un format qui pouvait être facilement saisi. Des corrections étaient apportées au besoin. Les questionnaires étaient ensuite envoyés, par lot, au secteur de la saisie des données.

4.2 Saisie des données

La technique « taper et numériser » avec la méthode de saisie au clavier de chaînes de caractères a été employée pour saisir les questionnaires et les mises à jour. Cette méthode exige de taper les données à saisir dans des formulaires particuliers en une suite de chaîne. Les données tapées sont ensuite numérisées par le lecteur de reconnaissance IBM 1,288 et transformées en format lisible par machine. Pour s'assurer que les dactylos ne dépassent pas un taux d'erreurs de 5 %, chaque lot de documents tapés était vérifié par échantillonnage en utilisant les mécanismes de vérification sur clavier. Lors de cette procédure, il fallait saisir de nouveau au clavier un nombre précis de documents pour chaque lot à l'aide d'un équipement d'enregistrement sur disque. Si le nombre d'erreurs relevées dépassait le maximum permis pour ce lot, tous les documents du lot étaient vérifiés et toutes les erreurs corrigées. Pour maintenir un taux d'erreurs de moins de 5 %, la vérification de 28 % des documents était requise.

2. GEISHA est l'acronyme pour « generalized edit and imputation system using the hot deck approach » (système généralisé de vérification et d'imputation reposant sur la méthode hot deck) mise au point par le Secteur des services statistiques de Statistique Canada.

4.3 Vérification et mise à jour de base

Ce sous-système permettant l'extraction et la validation des données brutes a été conçu de sorte que la vérification externe puisse se faire pour tous les types et tous les niveaux de vérification.

La validation de base des données saisies au clavier était contrôlée par l'intermédiaire de la liste maîtresse centrale (seules les identifications de questionnaires valides étaient acceptées) et les spécifications de vérification par tableaux (pour garantir que seules les données valides étaient saisies). Chaque enregistrement de ménage créé à partir des procédures de validation était ensuite traité à une étape de vérification de la construction où l'on vérifiait la cohérence de base de la séquence des données logiques existantes. Les incohérences étaient inscrites sur une liste avec les erreurs de validation et traitées par une interface manuelle aux fins de correction par l'entremise d'un document aller-retour. Les corrections étaient appliquées dans l'ensemble des modules existants jusqu'à ce que tous les enregistrements soient exempts de toute donnée incohérente ou invalide détectable.

4.4 Traitement des données sur les logements et les ménages actuels

Après l'achèvement de l'étape de révision de base, les fichiers de données existants étaient divisés en deux parties : i) ceux qui comportaient les variables liées aux caractéristiques des logements et des ménages actuels; ii) ceux qui comportaient les variables liées aux caractéristiques des logements et des ménages précédents et les données sur la mobilité. Cette séparation était nécessaire, car les deux parties devaient être traitées différemment — la première avec imputation, la dernière sans imputation — et aussi parce que des fichiers distincts devaient être produits pour chacune. Les étapes de traitement supplémentaires suivantes renvoient au traitement des données sur les logements et les ménages actuels seulement.

4.5 Vérification et imputation

L'étape de vérification et d'imputation du traitement des données comportait elle-même trois étapes. La première consistait à réviser et à imputer, au moyen d'un module personnalisé, un ensemble limité de variables

clés auxquelles la plupart des autres variables étaient logiquement liées. Ces quelques variables étaient corrigées séparément, en raison de leur grande influence sur le schéma de réponse dans les variables restantes. À la suite de cette étape, les variables restantes étaient révisées et imputées au moyen d'un logiciel généralisé reposant sur une routine d'imputation hot deck (GEISHA). Finalement, en raison des limites techniques de ce logiciel généralisé, il était nécessaire d'effectuer quelques imputations supplémentaires sur des variables quantitatives.

4.6 Formations des familles

À la suite de l'étape de vérification et d'imputation, les variables démographiques étaient combinées pour créer un certain nombre de variables dérivées liées à la taille et aux types de familles.

4.7 Pondération

Tous les enregistrements recevaient une pondération en fonction de la probabilité d'inclusion de l'enregistrement dans l'échantillon et corrigés en fonction de la non-réponse complète.

4.8 Masquage de la confidentialité

Certaines variables dans certains enregistrements étaient jugées suffisamment uniques de sorte qu'elles pouvaient révéler des renseignements à propos de l'identité des répondants individuels et ainsi contrecarrer les mesures de sécurité en matière de confidentialité. Pour surmonter cet obstacle, des procédures et des programmes ont été adoptés pour masquer ces données (voir la section 8).

Création des fichiers de logements actuels

Finalement, des fichiers de microdonnées comportant des variables dérivées existantes précisées préalablement par la division cliente ont été créés.

Les données précédentes sur les logements et les ménages ont dû passer par de nombreuses étapes supplémentaires, car i) elles ne devaient pas être imputées et ii) les poids à utiliser pour ces données étaient ceux qui avaient été fournis avec les données sur les logements et les ménages actuels.

Des fichiers de microdonnées comprenant ces variables ont été créés, jumelés et fusionnés avec les fichiers des données sur les logements et les ménages précédents, ce qui a donné lieu à la création d'un seul enregistrement pour chaque questionnaire.

Certaines variables issues des données de recensement ont aussi été masquées pour satisfaire les exigences en matière de confidentialité.

5 Relations publiques

Des lettres de présentation ont été postées aux répondants environ une semaine avant l'interview et, au moment de l'interview, des brochures ont été remises. Des lettres de remerciement ont été postées à tous les ménages répondants.

Les bureaux de la SCHL et les chefs de police locaux ont été informés de la réalisation de l'enquête.

6 Prétests

Au printemps 1973, un prétest a été mené sur 550 ménages à Toronto, en Ontario et à Hull, au Québec. Cette étude visait à tester certains des concepts à mesurer dans l'enquête principale et à déterminer les difficultés qui pouvaient survenir en demandant de se souvenir de renseignements pouvant remonter jusqu'à trois ans auparavant.

L'échantillon du prétest a été sélectionné à partir de l'Enquête sur les finances des consommateurs de 1968 et l'enquête a été menée par interview sur place. Un compte rendu des agents recenseurs a eu lieu à l'achèvement des interviews; une analyse des résultats et une évaluation du prétest ont eu lieu manuellement à l'aide des questionnaires réels.

Les principales conclusions tirées de cette étude étaient que i) il ne semblait y avoir aucune difficulté évidente, les répondants comprenant bien les concepts demandés; ii) il ne semblait y avoir aucune hostilité envers ce sujet de la part des répondants; iii) le rappel ne semblait entraîner aucun taux de non-réponse exceptionnellement élevé.

À l'hiver 1974, dans le cadre de l'élaboration du questionnaire pour l'enquête de 1974, un deuxième prétest a été réalisé sur 300 ménages

dans la région d'Ottawa dans le but de déterminer la faisabilité d'obtenir des données sur les dimensions des pièces dans le cadre d'une interview. Encore une fois, l'étude a été menée par interview sur place et suivie d'un compte rendu des agents recenseurs et d'une analyse et d'une évaluation manuelles des données recueillies. La principale recommandation découlant de ce prétest était que, pour obtenir des données fiables sur les dimensions des pièces, on devait demander aux répondants de mesurer les pièces dans leurs logements et demander à l'intervieweur d'attendre que cela se fasse ou de faire un rappel par téléphone pour obtenir l'information.

7 Disponibilité des données

Comme mentionné précédemment, les données de l'Enquête de 1974 sur les logements ont été séparées en deux parties pendant l'étape de traitement et chaque partie a été stockée différemment. Au sein de Statistique Canada, les données sur les ménages et les logements actuels sont stockées individuellement pour chacune des 23 régions d'enquête dans une base de données à accès direct à l'état imputé et non imputé. Elle est accessible aux fins de totalisations à l'aide de STATPAK et disponible dans des fichiers sur bande magnétique séquentiels créés selon les spécifications en matière de format de la SCHL[3].

La SCHL a aussi créé une base de données pour ces données et a fourni un système d'extraction interactif aux fins de totalisation.

Les fichiers sur bande magnétique séquentiels devraient pouvoir être utilisés par des tiers en 1977[4].

Les données sur les ménages et les logements précédents et l'extraction des données du Recensement de 1971 aux fins de couplage ne peuvent être utilisées qu'au sein de Statistique Canada et de la SCHL dans le format des fichiers sur bande magnétique séquentiel à l'état non imputé.

Ces fichiers ne seront pas mis à la disposition d'utilisateurs tiers.

---

3. Ces bandes magnétiques ne comprennent pas les enregistrements mis de côté en raison de contraintes relatives à la confidentialité.
4. Il est possible d'obtenir des renseignements supplémentaires au sujet du coût et de la disponibilité de ces fichiers en communiquant avec la Division de coordination des enquêtes spéciales de Statistique Canada.

8 Vérification en vue de la diffusion des microdonnées

Pour préserver la confidentialité, les données sur les fichiers de microdonnées diffusées à partir de cette enquête ont été vérifiées pour supprimer toute possibilité d'identification des personnes. Les étapes suivantes ont été suivies :

I. Données de 1974

1. Tous les enregistrements indiquant un revenu du ménage supérieur à 100 000 $ ont été supprimés des fichiers.
2. Tous les âges des personnes supérieurs à 75 ans ont été codés 76 dans le fichier.
3. Tous les capitaux impayés sur les hypothèques de 75 000 $ et plus ont été codés 75 000 $ dans le fichier.
4. Toutes les valeurs marchandes de 150 000 $ et plus ont été codées 150 000 $ dans le fichier.

Pour chaque région d'enquête, le sommaire suivant des résultats de la vérification de diffusion des microdonnées est accessible :

i) Montant pondéré total du revenu supprimé.

ii) Nombre total d'enregistrements supprimés.

iii) Poids d'enregistrement total supprimé.

iv) Montant pondéré total des capitaux impayés sur les hypothèques supérieures à 75 000 $, c'est-à-dire la différence pondérée entre le montant réel et 75 000 $.

v) Nombre total d'enregistrements indiquant des capitaux impayés sur l'hypothèque de 75 000 $ ou plus.

vi) Montant pondéré total des valeurs marchandes supérieures à 150 000 $, c'est-à-dire la différence pondérée entre la valeur réelle et 150 000 $.

vii) Nombre total d'enregistrements indiquant des valeurs marchandes de 150 000 $ ou plus.

II. Données sur les logements précédents

1. Tous les renseignements pour les enregistrements indiquant un revenu du ménage supérieur à 100 000 $ ont été supprimés des fichiers.
2. Tous les âges des personnes supérieurs à 75 ans ont été codés 76 dans le fichier.
3. Tous les capitaux impayés sur les hypothèques de 75 000 $ et plus ont été codés 75 000 $ dans le fichier.
4. Tous les prix de vente de 150 000 $ et plus ont été codés 150 000 $ dans le fichier.

III. Données de recensement

1. Tous les âges des personnes supérieurs à 75 ans ont été codés 76 dans le fichier.
2. Toutes les données de recensement pour les enregistrements indiquant un revenu du ménage supérieur à 100 000 $ ont été supprimées.
3. Le revenu de tous les chefs de ménage de 75 000 $ ou plus est codé 75 000 $ dans le fichier à l'exception des femmes chefs de ménage dans les régions d'enquête dans la région de l'Atlantique où tout revenu de 50 000 $ ou plus est codé 50 000 $ dans les fichiers.

Remerciements

L'auteur tient à souligner la contribution des membres de l'équipe de projet, R. Haddow, S. Porter, C. Scott et G. Smith, durant la préparation et la rédaction du rapport sur lequel le présent article est fondé.

Annexe

Renseignements détaillés sur les procédures d'estimation

1. Estimations

Chacune des bases de sondage utilisées dans cette enquête diffère des autres en ce qui a trait aux renseignements qu'elle comporte à propos des unités qui y sont présentes et à la manière de les stocker. C'est la raison pour laquelle un plan de sondage différent a été utilisé pour sélectionner un échantillon d'unités à partir de chaque base de sondage. Cela a fait en sorte que trois ensembles différents de techniques d'estimation étaient nécessaires.

La section qui suit présente les formules pour calculer les poids, les estimations de valeurs de domaine et leur coefficient de variation estimé pour chacune des trois sous-populations.

Même si l'unité d'échantillonnage utilisée dans cette enquête est l'unité de logement, l'unité pour laquelle des estimations sont souhaitées est le logement occupé (ménage), un domaine de chacune des trois sous-populations. Par conséquent, toutes les estimations abordées dans cette section renvoient à ce domaine.

2. Fichier du Recensement du Canada de 1971 des logements privés occupés

Les unités de cette base de sondage ont été stratifiées en 40 strates au moyen des variables du mode d'occupation, du revenu du chef du ménage (et de l'époux ou de l'épouse) et de l'âge du chef du ménage. À l'intérieur de chaque strate, un échantillonnage aléatoire simple d'unités sans remise a été sélectionné pour chaque région d'enquête.

Notation

h: indice inférieur indiquant la h^e strate (h = 1, ..., 40)

i: indice inférieur indiquant la i^e unité (enregistrement)

N: nombre d'unités de la sous-population (c'est-à-dire nombre de logements privés occupés au 1er juin 1971)

N' : nombre (inconnu) d'unités de la sous-population dans le domaine d'intérêt (c'est-à-dire nombre de logements occupés pendant la période de référence dans la sous-population)

$\hat{N}'$: nombre estimatif d'unités de la sous-population dans le domaine d'intérêt (c'est-à-dire nombre de logements occupés pendant la période de référence dans la sous-population)

N_h : nombre d'unités de la sous-population dans la strate h

N'_h : nombre d'unités de la sous-population dans la strate h dans le sous-échantillon (c'est-à-dire dans le fichier du recensement 2B)

Les valeurs de l'échantillon correspondant sont exprimées en remplaçant N par n.

X : total du domaine pour la caractéristique X

$\bar{X}$: moyenne du domaine pour la caractéristique X

x_{hi} : valeur de la caractéristique X prise par la i^e unité d'échantillonnage dans la strate h

$\bar{x}_h$: moyenne de l'échantillon dans le domaine d'intérêt pour la caractéristique x pour la strate h

$V(\hat{X})$: variance du total du domaine estimé

$V(\hat{\bar{X}})$: variance de la moyenne du domaine estimée

$\alpha(\hat{X})$: coefficient de variation du total du domaine estimé

$\alpha(\hat{\bar{X}})$: coefficient de variation de la moyenne du domaine estimée

Poids

Trois facteurs doivent être pris en compte dans le calcul des poids pour les unités de cette sous-population :

i) Le facteur de sous-échantillonnage (poids des ménages du recensement) associé à chaque unité de la base de sondage

ii) La probabilité de sélection de base pour chaque unité sélectionnée à partir de la base de sondage

iii) Le facteur de non-réponse appliqué à chaque unité sélectionnée à la suite de la non-réponse à l'enquête

Le premier poids est rendu nécessaire par le fait que le fichier du recensement utilisé représente seulement environ le tiers de tous les

logements privés occupés. Ainsi, chaque enregistrement dans le fichier comporte un poids d'une valeur de 3. W_{h_1} désigne le premier poids.

Les deuxième et troisième poids sont calculés séparément au niveau de la strate et sont appliqués à chaque enregistrement de cette strate. Le deuxième poids est représenté par :

$$W_{h_2} = \frac{N_h'}{n_h}, \quad h = 1, \ldots, 40.$$

Le troisième poids est représenté par :

$$W_{h_3} = \begin{cases} \dfrac{n_h}{n_h'} & \text{si } n_h' \neq 0 \\ 0 & \text{si } n_h' = 0 \end{cases}$$

La combinaison de W_{h_1}, W_{h_2} et W_{h_3} donne lieu à un poids de strate de base :

$$W_h = W_{h_1} W_{h_2} W_{h_3} = \begin{cases} \dfrac{N_h}{n_h'} & \text{si } n_h' \neq 0 \\ 0 & \text{si } n_h' = 0 \end{cases}$$

Estimations

Total du domaine : $\hat{X} = \sum_{h=1}^{40} W_h \sum_{i=1}^{n_h''} x_{hi}$

Moyenne du domaine : $\hat{\bar{X}} = \dfrac{\hat{X}}{\hat{N}_d}$ où $\hat{N}_d = \sum_{h=1}^{40} W_h n_h''$

Coefficient de variation

Pour calculer le coefficient de variation estimé d'une estimation, il faut d'abord calculer la variation estimée. Celle-ci est obtenue au moyen de la formule ci-dessous.

Pour le total du domaine : $\hat{V}(\hat{X}) = \sum_{h=1}^{40} N_h (N_h - n_h') \dfrac{S_{ha}^2}{n_h'}$

où $$S^2_{ha} = \frac{1}{n'_h - 1} \left[\sum_{i=1}^{n''_h} x^2_{hi} - \frac{\left(\sum_{i=1}^{n''_h} x_{hi} \right)^2}{n'_h} \right]$$

Pour la moyenne du domaine :

$$\hat{V}(\hat{\bar{X}}) = \frac{1}{(\hat{N}')^2} \left[\sum_{h=1}^{40} N_h (N_h - n'_h) \frac{S^2_{hb}}{n'_h} \right]$$

$$+ \frac{1}{(\hat{N}')^2} \left[\sum_{h=1}^{40} \frac{N_h (N_h - n'_h)}{n'_h (n'_h - 1)} n''_h \left(1 - \frac{n''_h}{n'_h}\right) (\bar{x}_h - \hat{\bar{X}})^2 \right]$$

où $$\hat{N}' = \sum_{h=1}^{40} \frac{N_h}{n'_h} n''_h = \sum_{h=1}^{40} W_h n''_h$$

et $$S^2_{hb} = \frac{1}{n'_h - 1} \left[\sum_{i=1}^{n''_h} x^2_{hi} - \frac{\left(\sum_{i=1}^{n''_h} x_{hi} \right)^2}{n''_h} \right]$$

Cela donne lieu aux coefficients de variation estimés suivants :

Pour le total du domaine : $\hat{\alpha}(\hat{X}) = \sqrt{\hat{V}(\hat{X})/\hat{X}}$

Pour la moyenne du domaine : $\hat{\alpha}(\hat{\bar{X}}) = \sqrt{\hat{V}(\hat{\bar{X}})/\hat{\bar{X}}}$

3. Fichier du Recensement du Canada de 1971 des logements privés vacants

Comme aucune des unités de cette base de sondage ne comprend des caractéristiques du logement et du ménage qui lui sont associées, aucune stratification n'a été effectuée sur les unités. Dans chaque région d'enquête, un échantillon aléatoire systématique d'unités a été sélectionné sans remise à partir des registres des visites du Recensement du Canada de 1971.

Par conséquent, la notation et les formules dans cette sous-section sont un cas particulier de celles de la sous-section 2 ayant h = 1 sauf pour l'exception suivante.

Pour le calcul de la variance estimée pour la moyenne d'un domaine,

$$s_b^2 = \frac{1}{n'' - 1} \left[\sum_{i=1}^{n''} x_i^2 - \frac{\left(\sum_{i=1}^{n''} x_i \right)^2}{n''} \right]$$

4. Enregistrements de Statistique Canada des permis de bâtir délivrés

En raison des problèmes liés à l'utilisation de cette base de sondage (voir Enquête de 1974 sur les logements/Un rapport sur le plan d'échantillonnage), les probabilités de sélection des unités d'échantillonnage sont très difficiles à calculer. Pour simplifier les procédures, l'échantillon peut être pondéré, non pas selon les chiffres de délivrance des permis de bâtir, mais selon les chiffres d'achèvement des logements.

Comme les chiffres d'achèvement des logements ne sont pas disponibles au niveau de la municipalité par type de logement, la stratification utilisée dans la procédure de sélection ne sera pas employée dans les procédures d'estimation. Par conséquent, cet échantillon sera considéré comme ayant été sélectionné par échantillonnage aléatoire systématique sans remise et sans réplique à partir d'une population non limitée dans chaque région d'enquête.

Ainsi, la notation et les formules pour les estimations obtenues à partir de cette base de sondage sont identiques à celles de la sous-section 3.

5. Estimations de la population

Comme mentionné dans la section 1, l'hypothèse avancée ici est que les sous-populations sont indépendantes les unes des autres. La validité de cette hypothèse ne sera pas connue avant qu'une évaluation du plan et des procédures d'enquête ne soit réalisée.

En traitant les sous-populations comme étant indépendantes les unes des autres, les estimations pour le domaine d'intérêt sur l'ensemble de la population peuvent être obtenues en traitant les unités de chacune des

bases de sondage 2 et 3 comme appartenant à deux strates indépendantes; par conséquent, toutes les sommes sont supérieures à 42 strates.

Estimations

Total du domaine : $\hat{X} = \sum_{h=1}^{42} \hat{X}_h$

Moyennes de domaines : $\hat{\bar{X}} = \dfrac{\sum_{h=1}^{42} \hat{X}_h}{\sum_{h=1}^{42} \hat{N}'_h}$

Coefficient de variation

Pour calculer les estimations des coefficients de variation d'une estimation, il faut d'abord calculer les variances estimées. Celles-ci sont obtenues au moyen des formules ci-dessous.

Pour le total du domaine : $\hat{V}(\hat{X}) = \sum_{h=1}^{42} \hat{V}(\hat{X}_h)$

Pour les moyennes de domaines :

$$\hat{V}(\hat{\bar{X}}) = \frac{1}{(\hat{N}')^2} \left[\sum_{h=1}^{42} N_h (N_h - n'_h) \frac{S_{hb}^2}{n'_h} \right.$$

$$+ \frac{1}{(\hat{N}')^2} \left[\sum_{h=1}^{42} \frac{N_h (N_h - n'_h)}{n'_h (n'_h - 1)} n''_h \left(1 - \frac{n''_h}{n'_h}\right) (\bar{x}_h - \hat{\bar{X}})^2 \right.$$

où, pour $h = 41, 42$: $S_{hb}^2 = \dfrac{1}{n'_h - 1} \left[\sum_{i=1}^{n''_h} x_{hi}^2 - \dfrac{\sum_{i=1}^{n''_h} x_i^2}{n''_h} \right.$

et pour $h = 1, \ldots, 40$, S_{hb}^2 est comme dans la sous-section 2.