Techniques d'enquête, décembre 1978
Vol. 4, nº 2
Statistique Canada, nº 12-001-X au catalogue

L'application d'une méthode systématique de contrôle et d'imputation automatiques au Recensement de la population et des logements du Canada de 1976[1]

C.J. Hill[2]

Résumé

À partir de la méthode systématique de vérification et d'imputation proposée par I.P. Fellegi et D. Holt, on a mis au point un système général de vérification et d'imputation par la méthode du hot-deck et on l'a appliqué aux données du recensement de la population et du logement de 1976. Le présent article vise à étudier cette application de la méthode Fellegi-Holt et à évaluer les points forts et les points faibles de la méthodologie à partir de certains exemples empiriques. La présentation du système est faite dans un contexte plus vaste, celui des grands problèmes posés par la vérification et l'imputation des données d'enquête. L'auteur énumère aussi quelques avenues de développement possibles.

Mots-clés : Contrôle et imputation automatiques; Fellegi-Holt; système de traitement; recensement canadien.

1 Introduction

Le présent article traite de l'application d'une méthode systématique de contrôle et d'imputation automatiques, à l'origine mise au point par I.P. Fellegi et D. Holt [1], au Recensement de la population et des logements du Canada de 1976. La mise en œuvre de cette méthodologie en tant que système informatique au sein de Statistique Canada correspond au système appelé « CAN-EDIT ». Une description en a été faite par Graves [2]. Le système Can-Edit, est à son tour devenu une composante du « Système canadien de contrôle et d'imputation du recensement », qui comprenait plusieurs autres modules personnalisés [3]. Certains de ces modules ont permis de traiter des problèmes particuliers de contrôle et d'imputation. D'autres ont fonctionné de pair avec le système CAN-EDIT et ont permis de régler des problèmes méthodologiques qui n'avaient pas été abordés par Fellegi et Holt. Une discussion sur la méthodologie appliquée dans ces modules figure dans le présent document, car ces derniers se sont avérés essentiels à l'application de la méthode Fellegi-Holt.

1. Adapté d'un document présenté lors de la réunion annuelle de l'American Statistical Association, qui a eu lieu à San Diego en Californie, aux États-Unis, du 14 au 17 août 1978.
2. C.J. Hill, Recensement, Division des méthodes d'enquête, Statistique Canada.

Le recensement est une enquête polyvalente qui comprend des questions sur la population et le logement. Les questions sur le logement en 1976 visaient principalement à déterminer le type et le mode d'occupation du logement. Les questions sur la population ont été divisées en deux parties : une série de questions de base posées à toutes les personnes et une série de questions types posées aux personnes de 15 ans et plus dans le tiers de tous les ménages privés et dans tous les logements collectifs. Les questions de base étaient des questions démographiques sur l'âge, le sexe et l'état matrimonial; une question portait sur la relation avec le chef du ménage et une autre sur la langue maternelle. Les questions types portaient sur la scolarité, la situation d'activité et la mobilité. Le système « CAN-EDIT » a été utilisé pour le contrôle et l'imputation de la plupart des variables. Les seules variables non traitées par ce système étaient la langue maternelle et la mobilité.

Le présent article présente le bien-fondé des opérations de contrôle et d'imputation des données du recensement et fournit une brève description non technique de la méthodologie aux sections 2 et 3. Une évaluation de la méthode est ensuite présentée à la section 4, et une dernière section propose des orientations pour l'élaboration de méthodes de contrôle et d'imputation à partir des leçons tirées de l'application au recensement canadien de 1976.

2 Le bien-fondé du contrôle et de l'imputation des données du recensement

Les termes « contrôle et imputation » (C et I) utilisés dans le présent article en référence au recensement sont deux facettes d'une même opération. Le « contrôle » désigne la détection d'une erreur, et l'« imputation » se rapporte à la correction d'une erreur. Le contrôle peut être considéré séparément de l'imputation en ce sens qu'il peut être utilisé pour amorcer une mesure corrective comportant un retour à un état antérieur dans le traitement. Le contrôle peut également être effectué simplement pour signaler les enregistrements erronés. L'imputation, en tant que correction d'une erreur, signifie toute modification des données qui produit un enregistrement qui passera les contrôles, autrement que par renvoi à la source des données pour obtenir une réponse « vraie ». Cette opération de contrôle et d'imputation est entreprise dans le but

de réduire au minimum le nombre d'erreurs dans les données au microniveau.

La raison pour laquelle on procède à l'imputation plutôt qu'à la correction pour tenter d'obtenir une valeur « vraie », est qu'après une certaine étape de l'opération, il devient coûteux, voire impossible, de revenir sur ses pas. À cette étape, le choix consiste soit à modifier et à imputer les données, soit à publier des données qui comprennent des renseignements non précisés ou erronés.

Entre autres, les trois raisons importantes suivantes ont influencé l'application du contrôle et de l'imputation dans le recensement de 1976.

1) Pour obtenir les estimations requises, des ajustements doivent être effectués concernant les erreurs au macroniveau ou au microniveau. La correction (par contrôle et imputation) au microniveau peut utiliser au maximum l'information disponible et, en principe, obtenir la meilleure estimation.

2) Les opérations subséquentes du recensement, par exemple, la formation de familles seraient beaucoup plus compliquées, voire impossibles, avec des données incomplètes et incohérentes. Dans certains cas, le nombre d'enregistrements invalides augmenterait considérablement.

3) Des estimations officielles cohérentes sont essentielles en tant que service aux utilisateurs aussi bien à l'extérieur qu'au sein de Statistique Canada. Peu d'utilisateurs voudront assumer la responsabilité d'ajuster les estimations, et des difficultés pourraient survenir en raison d'estimations non officielles divergentes.

3 La méthodologie et sa mise en œuvre

3.1 Objectifs de la méthodologie

Fellegi et Holt énoncent trois objectifs pour la méthodologie qui sous-tend le système de contrôle et d'imputation.

1) Un plus grand nombre possible de données originales doivent être conservées en modifiant le nombre minimal de champs dans un

enregistrement imparfait donné afin de produire un enregistrement parfait.

2) Les données après imputation devraient conserver, dans la mesure du possible, les propriétés distributionnelles des enregistrements parfaits.

3) L'imputation devrait provenir directement des règles de contrôle.

Ces objectifs visent clairement à assurer la qualité des données; leur validité sera abordée ci-dessous dans la section sur l'évaluation. Le troisième objectif est une considération pratique, car cela permet de simplifier grandement l'élaboration d'une définition de l'imputation.

3.2 Mise en œuvre de ces objectifs

La première tentative de mise en œuvre de la méthodologie a été effectuée par un système composé de deux sous-systèmes de base :

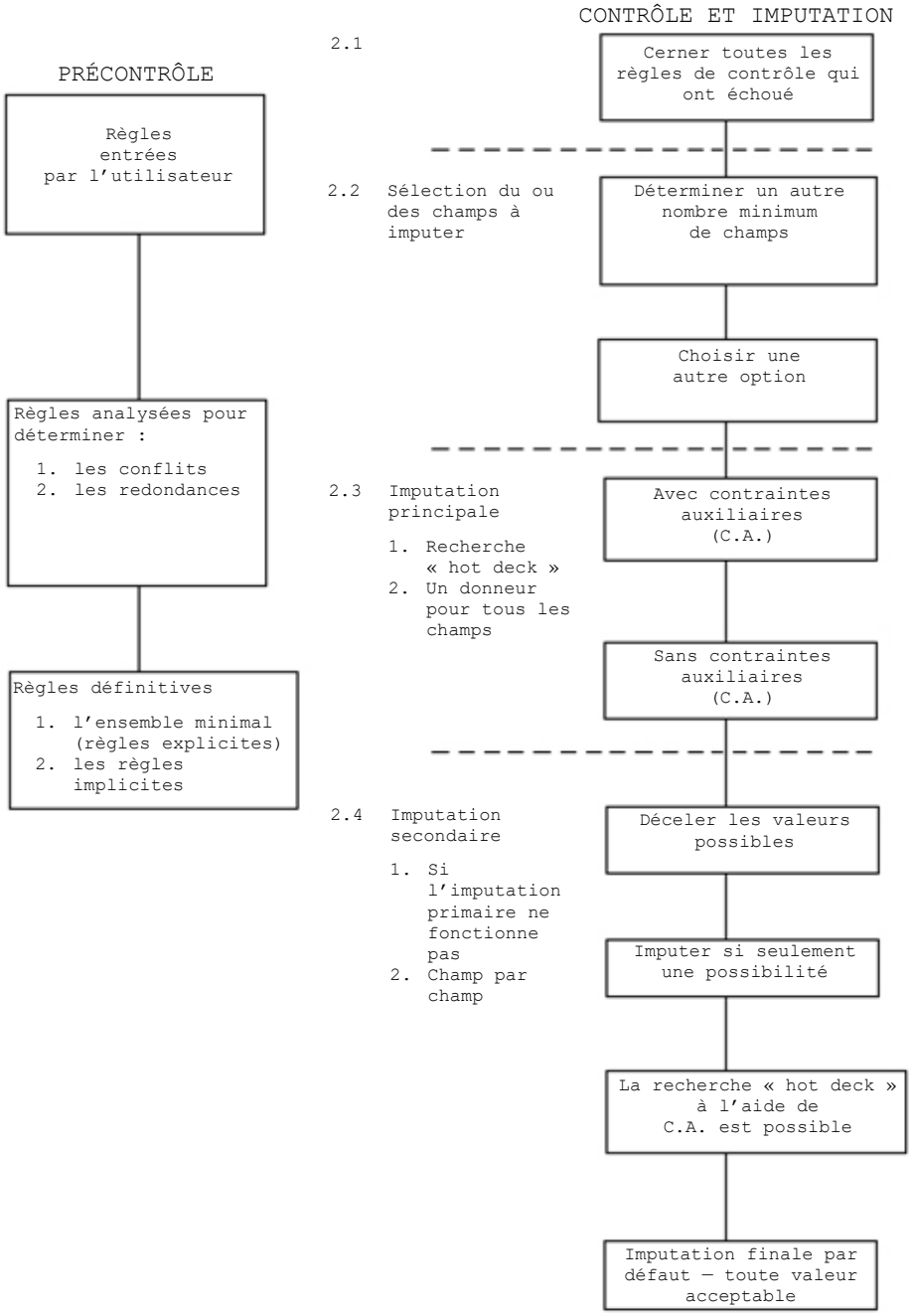
1) Un système pour analyser les règles de contrôle.

2) Le système de contrôle et d'imputation qui fonctionne à partir des données.

Ces opérations sont présentées dans le diagramme 1.

Diagramme 1

Organigramme des processus de « CAN-EDIT » à entreprendre pour chaque strate*



*La stratification et les contraintes auxiliaires sont expliquées à la section 3.3.

(1) Le système pour analyser les règles de contrôle

La première étape de l'opération de contrôle et d'imputation est l'analyse des règles de contrôle. Cette étape comprend les phases suivantes :

Les contrôles sont rédigés sous forme de conflit. Il peut s'agir de contrôles au niveau de la même personne ou entre des personnes.

> Voici un exemple de contrôle au niveau de la même personne :
>
> « Il y a un conflit si la troisième personne du ménage est mariée et est âgée de moins de 15 ans ».
>
> Voici un exemple de contrôle entre des personnes :
>
> « Il y a un conflit si la sixième personne dans le ménage est le parent du chef du ménage et un homme, et que la neuvième personne dans le ménage est le parent du chef du ménage et un homme ».

Il est important de mentionner qu'un même concept nécessite de nombreuses règles de contrôle. Si, par exemple, un contrôle est nécessaire pour exclure la possibilité que le chef du ménage ait deux parents du même sexe, les règles de contrôle doivent être écrites entre toutes les paires de personnes possibles. Cette fonction essentielle impose certaines limites au système, lesquelles seront abordées dans une section ultérieure.

Les règles de contrôle sont ensuite analysées et le résultat définit :

i) toute incohérence ou tout conflit entre les règles;

ii) toute redondance dans les règles.

Une fois les incohérences éliminées, le résultat final est :

i) un ensemble minimal de règles de contrôle (règles explicites).

ii) un ensemble de règles de contrôle implicites, qui sont générées à partir de l'ensemble minimal.

Ces deux ensembles combinés constituent l'ensemble complet.

(2) Le système de contrôle et d'imputation

L'analyse des règles étant terminée, le contrôle et l'imputation des données peuvent être entrepris. Cette opération se divise en quatre étapes :

(2.1) Le contrôle qui définit les règles qui ont échoué pour chaque enregistrement.

(2.2) La sélection des champs à imputer. Cette fonction comporte deux parties :

i) la détermination du ou des champs qui représentent le nombre minimal de champs à modifier pour assurer un dossier parfait;

ii) la sélection aléatoire parmi les options s'il y a plus d'un ensemble minimal. L'information qui existait dans les champs sélectionnés pour l'imputation n'est maintenant pas prise en compte et n'aura aucune incidence sur l'action d'imputation.

Il y a deux étapes d'imputation, l'imputation primaire et l'imputation secondaire.

(2.3) L'imputation primaire est une méthode par laquelle un enregistrement donneur donne à un « enregistrement imparfait » toutes les valeurs nécessaires pour effectuer l'imputation. Pour ce faire, le donneur doit mettre en correspondance l'enregistrement imparfait pour les champs qui ne seront pas modifiés et qui sont liés par une règle de contrôle aux champs à imputer. Ces conditions garantissent qu'un nouvel enregistrement est parfait. (Les améliorations apportées à ce principe sont abordées ci-dessous.) On trouve un enregistrement donneur en sélectionnant au hasard un enregistrement acceptable à partir d'un fichier d'environ 2 000 enregistrements. Il s'agit d'une forme de la méthode d'imputation connue sous le nom d'imputation « hot deck ». Si aucun enregistrement acceptable n'est trouvé, la recherche se poursuit par la méthode de l'imputation secondaire.

(2.4) L'imputation secondaire est une méthode d'imputation « hot deck » champ par champ. Dans cette méthode, certaines conditions

d'appariement peuvent être appliquées pendant la recherche d'un donneur. Cependant, la condition cruciale pour accepter un donneur n'est pas un appariement parfait qui s'est déjà révélé impossible, mais plutôt le fait que le nouvel enregistrement satisfait aux règles de contrôle concernant les champs laissés inchangés ou déjà imputés. Une fois qu'un champ est imputé, il est intégré à l'enregistrement pour que la recherche se poursuive de façon à imputer le champ suivant.

Une découverte importante qui a été faite pendant les essais de mise au point de « CAN-EDIT » est que, pour l'imputation primaire, seul l'ensemble minimal de règles est requis, tandis que l'imputation secondaire nécessite l'ensemble complet de règles. Si l'ensemble complet n'est pas utilisé, il pourrait survenir une situation dans laquelle un enregistrement partiellement imputé deviendrait impossible à remplir.

3.3 Modifications et améliorations conformes à la méthodologie initiale

Grâce à l'expérience acquise dans l'application du système, diverses modifications et améliorations ont été apportées. Certaines d'entre elles étaient conformes à la méthodologie, dont quatre sont examinées dans la présente section. La section 3.4 porte sur deux modifications qui entraient en conflit avec les objectifs initiaux. Les deux sont des améliorations importantes aux principes d'imputation dans le système « CAN-EDIT ». Il s'agit : 1) « des contraintes auxiliaires » et 2) « de la dissociation dépendante des données ». Les deux autres sont des éléments du « système de recensement » qui traitent des problèmes méthodologiques non abordés par Fellegi et Holt. Il s'agit : 3) du sous-système de stratification et (4) du choix entre le contrôle à une ou à plusieurs unités.

(1) Contraintes auxiliaires

Les contraintes auxiliaires sont des champs utilisés pour l'appariement pendant la recherche d'un enregistrement donneur, qu'elles soient nécessaires ou non comme condition d'appariement pour assurer un enregistrement parfait. Elles sont utilisées dans l'imputation primaire et secondaire. Les champs utilisés comme contraintes auxiliaires seront

normalement ceux qui sont fortement corrélés avec les champs à imputer. Fellegi et Holt ont suggéré cette amélioration dans leur article.

Dans l'imputation primaire, ces champs doivent être utilisés comme un ensemble complet ou pas du tout. Le système a été conçu de cette façon parce qu'il n'y a pas d'algorithme très évident pour assouplir les contraintes lorsque l'enregistrement entier est imputé simultanément. En effet, par conséquent, l'imputation primaire comporte deux niveaux d'appariement : les conditions optimales d'appariement qui comprennent des contraintes auxiliaires et une option dégradée d'appariement dans les champs nécessaires seulement.

Dans l'imputation secondaire, avec l'imputation champ par champ, on peut tenter d'apparier autant de champs qu'il est précisé et choisir le meilleur appariement.

(2) Dissociation dépendante des données

Lors de l'essai d'une version antérieure de « CAN-EDIT », des conditions d'appariement excessives ont forcé le traitement d'un grand nombre d'enregistrements au moyen de l'imputation secondaire. Une analyse du problème a révélé que les conditions d'appariement dans la recherche d'un donneur étaient trop restrictives.

Dans la version originale, un appariement a été fait avec chaque champ lié aux champs à imputer par les règles de contrôle. Cependant, comme deux champs sont liés par des règles de contrôle, cela ne signifie pas nécessairement que la valeur dans le champ qui doit rester inchangé restreint les valeurs acceptables dans le champ à imputer. Un exemple de cela se trouve dans le champ « relation avec le chef du ménage », en référence à la règle susmentionnée empêchant que les deux parents du chef du ménage soient du même sexe. De toute évidence, s'il y a une personne dans le ménage qui est codé comme étant le chef du ménage et un homme, cela restreint l'imputation du code parent à un autre homme. Si, par contre, aucune personne n'est associée à ce code, cette restriction n'est plus nécessaire.

(3) Le système de stratification

La fonction du système de stratification était de diviser les données en sous-ensembles qui (1) partageaient un ensemble commun de règles de

contrôle et (2) manifestaient un degré d'homogénéité au-delà de celui du partage des règles de contrôle. Le contrôle et l'imputation sont ensuite entrepris indépendamment dans chaque strate.

Les variables de contrôle[1], le type de document et le type de logement collectif ont été utilisés à cette fin, pour les données intégrales, ainsi qu'une variable définie en fonction de la composition du ménage. L'âge, le sexe et le type de logement collectif ont servi à stratifier les données de l'échantillon.

Une compréhension exhaustive de la nature de la stratification doit être considérée de pair avec la question du contrôle des unités simples et multiples, puisque l'une des dimensions de la stratification pour le contrôle des unités multiples était le nombre de personnes dans le ménage.

(4) Contrôle d'une ou de plusieurs unités

Dans un certain sens, le recensement représentait trois, voire quatre enquêtes regroupées en une seule, et la complexité de la tentative de contrôle à son égard tient en partie à cette nature multiple. Les données sur les logements sont indépendantes et ne présentent que des problèmes mineurs. La difficulté réside dans l'interrelation entre les données sur les personnes, les familles et les ménages. Au début des opérations de contrôle, le nombre de personnes (l'unité de bas niveau) dans les ménages (l'unité de haut niveau) a été figé. Bien entendu, la taille des ménages varie. À cette étape, la famille n'a pas encore été définie. IL faut maintenant choisir entre traiter la personne ou le ménage comme une unité modifiable.

[1] L'opération précédant le contrôle et l'imputation a permis de déterminer si un ménage était un logement privé ou collectif, occupé ou inoccupé et s'il faisait partie ou non de l'échantillon. Elle a également permis de s'assurer que tous les logements collectifs avaient un type déterminé, p. ex. un hôpital, un orphelinat, un hôtel. Cette information a été figée en tant que variables de contrôle « type de document » et « type de logement collectif ».

Ce problème, qui n'a pas été abordé par Fellegi et Holt, représentait un problème pratique majeur lors de l'intégration de « CAN-EDIT » dans le « système du recensement ». La méthodologie est fondée sur un espace de données cartésiennes qui, dans un cas précis, c'est-à-dire un ménage d'une certaine taille, possède un nombre fixe de dimensions. Il n'était pas possible d'avoir des ensembles de règles de contrôle qui traitaient des espaces de dimensions différentes, parce que chaque règle englobe toutes les dimensions de l'espace. Par conséquent, s'il doit y avoir des règles de contrôle entre les personnes, chaque taille de ménage nécessite un ensemble unique de règles de contrôle.

Le contrôle d'une seule unité est la méthode de contrôle selon laquelle la personne correspond à l'unité modifiable. Cela signifie qu'il ne peut y avoir de règles de contrôle entre les personnes.

Le contrôle d'unités multiples est une méthode de contrôle selon laquelle le ménage correspond à l'unité modifiable. Cette méthode permet de modifier les règles entre les personnes. Cependant, cela ne se fait pas sans heurts.

i) Les données doivent être stratifiées selon la taille du ménage.

ii) La taille potentielle de l'unité modifiable devient très grande.

iii) Il y a un seuil au-delà duquel il est totalement irréaliste d'effectuer un contrôle d'unités multiples, ce qui signifie qu'il doit y avoir un contrôle d'unité unique pour les autres personnes restantes dans les grands ménages.

En 1976, le contrôle d'unités multiples a été utilisé pour le contrôle des données intégrales dans les ménages privés, principalement parce qu'il était nécessaire d'établir des données familiales parfaites. Le contrôle d'une seule unité a été utilisé pour le contrôle de la plupart des personnes dans les logements collectifs, la 13^{e} personne et les suivantes et les ménages très importants ainsi que toutes les données de l'échantillon.

3.4 Modifications et améliorations incompatibles avec la méthodologie initiale

Lors de l'élaboration du système du recensement, deux caractéristiques qui entraient en conflit avec l'objectif initial, établi par Fellegi et

Holt, ont été incluses. Ces caractéristiques consistaient à modifier le nombre minimal de champs de données. Ces deux fonctions étaient des systèmes externes au système « CAN-EDIT », mais elles s'appuyaient sur une propriété particulière de ce système pour obtenir leur effet. Il s'agissait des systèmes suivants : 1) un système dérivé utilisé avant le contrôle et l'imputation et 2) une structure hiérarchique de contrôle et d'imputation. La méthodologie de Fellegi-Holt précisait que la variation des données observées devrait être réduite au minimum. Par conséquent, tous les champs sont des candidats égaux au changement. Le système « CAN-EDIT » a reconnu, pour de très bonnes raisons, qu'il y avait des variables de contrôle établies avant le contrôle et que le système devait inclure la possibilité de faire la distinction entre les champs « imputables » et « non imputables ».

Le système *Derive* : Ce logiciel est un système semi-généralisé qui crée un environnement dans lequel des variables supplémentaires peuvent être dérivées pour l'opération de contrôle et d'imputation.

i) Combiner deux champs ou plus en un seul.

ii) Dériver une variable pour la stratification.

iii) Créer des valeurs de classe pour une variable.

iv) Servir de moyen de forcer une action d'imputation.

C'est cette dernière fonction qu'il est important de considérer dans le cas présent, car elle entre en conflit avec les objectifs initiaux. La variable dérivée a été figée en tant que variable non imputable. Cela signifie que lorsqu'un contrôle concernait ce champ et d'autres, certains des autres champs ont été forcés de changer. Cela a été utilisé pour forcer un résultat d'imputation précis. En général, cela signifiait qu'il fallait changer plus que le nombre minimum de champs. Ce principe est expliqué en détail à la section 4.3.3.

Contrôle hiérarchique : Le contrôle hiérarchique est un système de contrôle dans lequel un ensemble de champs est modifié, imputé et figé avant qu'un autre ensemble de champs ne soit modifié, et dans lequel il existe au moins une règle de contrôle permettant de coupler les deux ensembles. S'il n'y a pas de règles permettant de coupler les deux ensembles, l'ordre n'est pas pertinent. Toutefois, s'il y a des règles de couplage, le gel de certains champs dans une hiérarchie antérieure

peut forcer un changement plus important que le changement minimal dans l’ensemble de l’enregistrement. Le principe du changement minimal ne s’applique qu’à une seule hiérarchie.

En 1976, il y avait deux hiérarchies principales : une pour les données intégrales et l’autre pour les données de l’échantillon. De toute évidence, cette structure n’a eu d’incidence que sur le questionnaire-échantillon, principalement en ce qui concerne la question sur l’âge. L’âge était figé dans la première hiérarchie et pouvait ne pas correspondre aux données sur la scolarité, la situation d’activité et la mobilité. En pratique, ces incohérences étaient rares et l’effet sur les données était négligeable. Une hiérarchie mineure supplémentaire a été utilisée pour les questions dans les filtres dans les données de l’échantillon.

4 Une évaluation de la méthodologie de contrôle et d’imputation

4.1 Introduction

La méthode peut être évaluée en tant qu’instrument permettant la réussite du contrôle et de l’imputation des données et, objectivement, par une évaluation externe par rapport à une source de données réelles. Un projet est en cours pour atteindre ce dernier objectif. Les résultats de ce projet seront présentés dans une publication du recensement [4]. Cependant, la présente discussion porte sur l’examen du système en tant qu’instrument de production d’une base de données parfaite.

4.2 L’évaluation de la méthode comme instrument pour le contrôle et l’imputation des données

Les points suivants seront pris en compte dans l’évaluation du système généralisé comme moyen de réussir une opération de contrôle et d’imputation.

(1) La portée méthodologique du système, c’est-à-dire la gamme de types de variables et de conditions de contrôle que le système est conçu pour traiter.

(2) Le caractère fini, c'est-à-dire les limites pratiques auxquelles le système est conforme.

(3) La pertinence des trois objectifs énoncés par Fellegi et Holt.

4.2.1 La portée de la méthode

Dans leur article, Fellegi et Holt écrivent : [traduction] « Au début, limitons-nous aux enregistrements ne contenant que des données qualitatives (codées), c'est-à-dire des données qui ne sont pas assujetties à une mesure significative ».

Lors de l'élaboration d'un système généralisé de contrôle et d'imputation, il a fallu limiter la portée des types de données qu'il pouvait traiter. Comme l'ont indiqué Fellegi et Holt, la méthodologie portait principalement sur les données qualitatives.

Les champs quantitatifs peuvent, bien sûr, être traités comme s'il s'agissait de variables qualitatives et donc être traités dans le même système. Cependant, il y a deux objections importantes à cet égard :

(1) La perte d'information du fait que l'on délaisse la mesure.

(2) Le nombre potentiellement important de règles de contrôle qui peuvent être générées en tentant de traiter les règles arithmétiques comme des règles logiques entre les catégories.

Malgré ces objections, le système a été appliqué, dans le cadre du recensement, aux enregistrements qui contenaient une combinaison de données quantitatives et qualitatives. Ce choix était justifié dans la mesure où les variables étaient principalement qualitatives et que les contrôles appliqués aux variables quantitatives étaient de nature limitée. Cependant, comme le recensement tentait de contrôler des variables qui ne faisaient pas partie du champ d'application pour lequel le système de contrôle a été conçu, les résultats n'étaient pas entièrement satisfaisants.

La seule variable quantitative dans les données intégrales était la date de naissance ou, implicitement, l'âge.

La date de naissance a été définie par trois variables, soit la décennie, l'année et le mois de naissance, cette dernière étant plus exactement les deux périodes de janvier à mai, de juin à décembre. Chacun de ces

éléments pris séparément pouvait être utilisé comme variable qualitative et a effectivement été traité de la sorte. Il y avait deux problèmes principaux :

(1) Une barrière d'âge cruciale survient à l'âge de 15 ans. Seules les personnes de 15 ans ou plus devaient répondre aux questions types. De plus, certaines conditions n'étaient permises qu'à cet âge ou plus, par exemple être le chef du ménage ou être marié. Le problème était qu'après le contrôle et l'imputation, il y avait un nombre supérieur au nombre prévu de certains groupes de personnes proches de la limite d'âge de 15 ans, en particulier des personnes veuves ou divorcées. La seule consolation, c'est que le problème a été grandement réduit par rapport aux données de 1971.

(2) Il était impossible de rédiger des contrôles pour assurer un intervalle d'âge raisonnable entre les parents et les enfants. Le nombre de contrôles requis pour assurer une différence minimale de 15 ans était très élevé, car cela aurait nécessité une règle de contrôle pour chaque différence d'âge individuelle. La décision a donc été la suivante :

   i) limiter ces contrôles aux différences d'âge entre le chef du ménage et le conjoint ou la conjointe et leurs enfants (le principal groupe de contrôle ainsi exclu était celui entre le chef du ménage et ses parents);

   ii) utiliser seulement la décennie de naissance dans les contrôles;

   iii) veiller à ce qu'au moins un parent soit né au cours d'une décennie antérieure à celle de tous les enfants. (Il est théoriquement possible qu'un beau-parent soit plus jeune qu'un enfant adulte.)

L'application de ces règles a permis de supprimer certaines données erronées, mais pas toutes. Pour trouver une solution efficace à ce problème, il faudra pouvoir élaborer une méthodologie qui peut être mise en œuvre en tant que système permettant non seulement de contrôler et d'imputer des données quantitatives, mais aussi des données quantitatives combinées à des données qualitatives complexes.

4.2.2 Caractère fini

La population du Canada est de 23 millions d'habitants. Le nombre de ménages est de 7 millions. L'espace de données complet représentant les ménages comporte beaucoup plus de cellules que le nombre total de ménages. Pour les ménages de taille « n », cet espace contient environ $(2000)^n$ cellules. Le nombre de règles de contrôle nécessaires pour partitionner cet espace est également potentiellement très élevé. Une condition particulière de contrôle entre les personnes qui pourrait s'appliquer à la plupart des personnes dans le ménage, dans presque toutes les positions, aurait généré 100 millions de règles de contrôle. Un tableau des données indique qu'en fait, seulement 1 700 personnes au Canada pourraient ne pas cadrer dans ces règles.

Le nombre total de règles de contrôle dépend de la taille du ménage et de l'ensemble des conditions de contrôle à appliquer. Une utilisation réaliste des ressources informatiques établit une limite de 2 048 sur le nombre total de règles de contrôle. Cette limite a été mise en œuvre en restreignant le contrôle des unités multiples aux ménages de 12 personnes ou moins, ou aux 12 premières personnes dans les grands ménages, et en excluant certains types de conditions de l'ensemble des règles de contrôle. Un programme spécial de « nettoyage » a été utilisé pour corriger et imputer ces problèmes résiduels.

On observe aussi des limites de données lorsqu'on essaie de pousser la méthode trop loin. L'imputation s'est faite selon une méthode « hot deck ». En tentant de contrôler et d'imputer de grands ménages, le système s'est heurté à la limite des données, à savoir que le nombre d'enregistrements disponibles pour le « hot-deck » était devenu très faible. Dans le cas des très grands ménages, on atteint un point où l'opération est très coûteuse, le nombre d'enregistrements est très petit et la qualité de l'imputation est grandement réduite par la petite taille du « hot-deck ». Les limites finies du système sont probablement une entrave à l'efficacité de la méthode étant donné la nature limitée des données.

4.2.3 Le fondement méthodologique

Le contrôle est une opération essentiellement très simple et passive par rapport aux données définitives. Le seul problème que présente le contrôle concerne la nécessité de s'assurer que les règles de contrôle

sont parfaites et cohérentes. Il s'agit donc ici de se pencher sur le fondement méthodologique de l'action d'imputation. Les trois critères énoncés par Fellegi et Holt ont été décrits ci-dessus dans la description de la méthodologie et seront maintenant évalués.

4.2.3.1 Modification du moins d'éléments de données possible

Fellegi et Holt considèrent que le principe selon lequel il faut changer le moins d'éléments de données possible (champs) est d'une importance primordiale. Cette position est plus que justifiée en tant que réaction à la correction excessive de données qui, on le sait, a eu lieu. Toutefois, leur formulation est un cas particulier et relève d'un principe général selon lequel la modification des données devrait être réduite au minimum. Le problème, c'est que le nombre de champs est en quelque sorte arbitraire. Le nombre de champs comportant les mêmes renseignements peut être modifié par des changements dans le questionnaire ou dans la saisie des données. Un concept simple et facile à définir peut être saisi de façon fiable par une seule question, tandis qu'un certain nombre de questions peuvent être utilisées pour définir un seul concept potentiellement ambigu. Par ailleurs, on ne peut pas commencer à compter des concepts comme s'ils avaient la même existence concrète qu'une question.

Fellegi et Holt reconnaissent implicitement ce problème en suggérant que les poids puissent être rattachés aux champs en fonction de leur fiabilité. Cette suggestion n'a pas été mise en œuvre dans le système appliqué au recensement de 1976. Toutefois, une analyse minutieuse s'impose avant que des solutions de rechange ne soient mises en place.

D'autres formulations du principe selon lequel il faut assurer un minimum de changements peuvent être envisagées.

(1) Changer le moins d'éléments de données possible.

(2) Changer un nombre minimum pondéré d'éléments de données.

(3) Déplacer la distance minimale dans un espace conceptuel.

La première de ces formulations est donnée par Fellegi et Holt, et la deuxième est une solution de rechange qu'ils proposent. La justification du recours à la deuxième solution peut toutefois se rapporter aux intentions conceptuelles du questionnaire plutôt qu'à la fiabilité de

chaque champ. Ce principe peut être illustré par des références aux questions sur l'éducation.

Dans une question portant sur les études, on demande au répondant de donner son plus haut niveau de scolarité, alors que dans trois autres questions, on demande au répondant d'indiquer ses études postsecondaires et ses qualifications. Dans le recensement, on entendait par « postsecondaire » un enseignement de nature avancée nécessitant un certain niveau de scolarité minimum comme condition d'admission. Malheureusement, une proportion étonnamment élevée de répondants ont interprété cela comme toute éducation obtenue après la fin des études. En général, les répondants qui faisaient cette erreur donnaient deux mauvaises réponses qui concordaient l'une avec l'autre, mais qui entraient en conflit avec le niveau le plus élevé qui était trop faible pour entrer dans les études postsecondaires. Dans ce cas-ci, le changement minimal a fait en sorte que le niveau le plus élevé a été coté à la hausse par erreur. On a finalement décidé que la meilleure stratégie consistait à modifier certaines règles pour éviter le risque de distorsion grave de la réponse la plus élevée par imputation.

4.2.3.2 Règles d'imputation dérivées des règles de contrôle correspondantes

Voici quelques-uns des avantages spécialisés du système énumérés par Fellegi et Holt :

(1) [traduction] « Compte tenu de la disponibilité d'un système généralisé de contrôle et d'imputation, les spécialistes peuvent facilement mettre en œuvre diverses spécifications expérimentales de contrôle dont l'incidence peut donc être évaluée sans effort supplémentaire pour la mise au point de systèmes. Cette capacité est particulièrement importante compte tenu de la nature généralement heuristique des spécifications de contrôle ».

(2) « Seuls les contrôles doivent être précisés à l'avance, puisque les imputations sont dérivées des contrôles eux-mêmes pour chaque enregistrement courant. Il s'agit d'une simplification importante pour les spécialistes de la charge de travail liée à la spécification d'un système complet de contrôle et d'imputation ».

Le premier de ces deux avantages, une approche « paramétrique » du contrôle, en était clairement un. Toutefois, le deuxième de ces deux avantages n'est pas nécessairement un avantage absolu.

Le fait que les actions d'imputation découlent directement des règles de contrôle empêche toute correction de données propre à une erreur. La méthodologie contrôle la flexibilité dont dispose l'utilisateur. Elle facilite l'expérimentation des règles de contrôle, mais supprime tout contrôle que l'utilisateur pourrait avoir autrement sur l'imputation. Un moyen d'utiliser une fonction particulière dans le système a toutefois été trouvé et a donné lieu à un certain contrôle de l'imputation. Il s'agissait de l'utilisation d'une variable dérivée non imputable mentionnée à la section 3.4. La variable utilisée pour forcer les changements dans les ménages comportant des personnes vivant en union libre est abordée ci-dessous.

Dans l'application de cette méthode au recensement de 1976, il est devenu évident qu'il y avait des situations dans lesquelles le contrôle plutôt que l'imputation aurait pu produire des résultats plus appropriés. Certains types d'erreurs de réponse ont entraîné des rejets au contrôle pour lesquels une procédure de correction clairement indiquée pourrait être précisée. En voici les principaux exemples :

(1) Le mauvais codage de la relation avec le chef du ménage en inversant la relation, c'est-à-dire que le fils ou la fille du chef du ménage était codé « Père » ou « Mère » du chef du ménage. Une proportion de ces erreurs, en particulier pour les groupes d'âge plus jeunes, a été imputée correctement.

(2) Codage incorrect de la relation avec le chef du ménage lorsqu'il y a des enfants dans le ménage, le chef du ménage et le conjoint ou la conjointe étant codés « Père » et « Mère ». Le problème, c'est que la première personne a été remplacée par le chef du ménage, mais que la deuxième personne est demeurée « père » ou « mère ».

Le principal problème avec les données dans ces deux exemples est qu'il s'agit de cas d'erreurs peu fréquentes sur des conditions courantes qui sont mal attribuées à des conditions peu fréquentes.

(3) Un type d'erreur qui a causé des problèmes particuliers concernait les réponses erronées associées aux unions libres.

L'intention du recensement était que les unions consensuelles soient traitées de la même façon que les unions légales, permettant ainsi l'identification des familles. Toutefois, dans ces cas, la tendance de réponse fréquente consistait à donner l'état matrimonial légal, c'est-à-dire « non marié », ainsi que l'union libre avec le chef du ménage, soit l'époux(se) ou le(la) conjoint(e) de fait.

Voici les réponses types :

Personne 1.	Chef du ménage	Divorcé(e)
Personne 2.	Conjoint(e) du chef	Célibataire

Dans un tel cas, le changement minimal des champs de données consistait à modifier la relation avec le chef du ménage de la personne 2 plutôt que l'état matrimonial des deux personnes. Des problèmes de cette nature ont été relevés pendant le test du recensement. Il a été décidé que la meilleure stratégie consistait à forcer l'utilisation des données à l'aide d'une variable dérivée non modifiable. La valeur « Conjoint(e) confirmé(e) » était attribuée chaque fois que des cas comme ceux qui précèdent se produisaient.

Les réponses ont ensuite été forcées de suivre le modèle suivant :

Personne 1.	Chef du ménage	Marié(e)
Personne 2.	Conjoint(e) du chef	Marié(e)

Il restait un problème résiduel quant à la façon de procéder au contrôle des enfants du partenaire en union libre dans ces cas. Certaines distorsions dans les données ont été jugées trop importantes pour être laissées sans correction. D'autres stratégies de correction ont donc été adoptées, soit avant l'application de la méthodologie Fellegi-Holt, comme pour les conjoints de fait, soit dans certains cas par un nettoyage subséquent. Une évaluation est en cours pour déterminer l'exactitude des mesures prises pendant toute l'opération de contrôle et d'imputation.

Ces problèmes particuliers qui peuvent être résolus par des corrections systématiques doivent toutefois être évalués par rapport aux avantages

de la méthode. Il y a de très nombreuses règles auxquelles les données devraient être conformes, chacune ayant entraîné un échec par un petit nombre d'enregistrements. Des règles d'imputation distinctes pour chacun de ces éléments auraient nécessité un système beaucoup plus complexe.

Le premier des deux avantages, « l'approche paramétrique » mentionnée ci-dessus, doit également être soupesé par rapport à la perte de flexibilité dans la spécification de l'imputation. Toutefois, pour ces contrôles, même une action d'imputation très imparfaite aurait eu un effet négligeable sur les données définitives.

Le système a créé un cadre dans lequel d'autres spécifications de contrôle pouvaient être examinées, évaluées et modifiées très facilement. Il a fallu un certain travail de la part du personnel spécialisé pour se familiariser avec le système et son langage. Cela fait, cependant, des progrès considérables pourraient être réalisés pour comprendre les problèmes liés aux données et peaufiner les contrôles.

Un incident a permis d'illustrer la souplesse du système. Des totalisations ont été exécutées sur les données à des étapes de la production. Une totalisation indiquait qu'une règle avait été omise dans un ensemble particulier de règles. La condition erronée détectée était une condition rare qui ne s'était pas produite dans les données d'essai, mais qui aurait néanmoins causé moins de difficultés dans le programme subséquent de formation des familles. Cette omission a été corrigée dans un délai de 48 heures. Le système ne peut naturellement pas garantir que l'utilisateur a inclus un ensemble complet de contrôles, mais il peut s'assurer que l'ensemble existant est parfait et uniforme. Il a fallu beaucoup plus de temps pour apporter des corrections à des programmes sur mesure, toujours avec le risque qu'une correction introduise une nouvelle erreur.

4.2.3.3 Conservation des propriétés distributionnelles des données parfaites

En l'absence de renseignements supplémentaires, le maintien des propriétés distributionnelles des données parfaites est la stratégie la plus appropriée à adopter pendant l'imputation. L'efficacité du système pour y parvenir a été accrue par l'utilisation de contraintes auxiliaires, c'est-à-dire les champs utilisés comme critères d'appariement dans la recherche « hot-deck » en raison de leur

corrélation avec le champ à imputer sans égard aux liens créés par les règles de contrôle. Cependant, il y a eu des situations où les enregistrements imparfaits ont clairement été tirés d'une distribution très différente de celle des enregistrements parfaits. Ces situations sont également vraies pour tous les sous-ensembles de la population définis par d'autres champs des enregistrements. Dans ce cas, l'inadéquation de l'imputation telle qu'elles se reflètent dans les données définitives est fonction de la différence entre les deux répartitions et de la proportion d'enregistrements imparfaits.

Deux raisons principales expliquent ce type de problème :

1) Certains sous-groupes de la population ont de la difficulté à choisir la bonne réponse et sont donc plus susceptibles de ne pas répondre.

(2) De nombreuses questions comportent une catégorie « nul » ou « aucun ». Aucun appareil n'a encore été inventé pour prévenir la non-réponse relativement élevée des personnes qui font partie de ce groupe.

Le tableau 1 illustre ce problème. Il présente sous forme de tableau la situation d'activité définie à partir des données non révisées. De toute évidence, la non-réponse a tendance à augmenter à mesure que la proportion de personnes ne faisant pas partie de la population active augmente. Cela donne à penser que les non-répondants ont tendance à provenir davantage de la population non participante. Il est possible de contrôler l'imputation en ce qui concerne les variables du recensement, mais pas pour toute relation au-delà de celles-ci.

Une évaluation de ce problème est en cours. Des améliorations possibles à la méthodologie ont aussi été envisagées pour tenir compte de cette différence dans le nombre de non-réponses. Toutefois, pour utiliser ces améliorations, il faut des renseignements externes pour estimer les taux de non-réponse différentiels par rapport à la variable cible.

Situation d'activité déterminée à partir des données pondérées du recensement non révisées								
		Situation définie				Situation non définie		
		Personnes actives		Personnes inactives		Situation d'activité		
		Nombre	%	Nombre	%	Nombre	%	Total
LES DEUX SEXES	15 à 19 ans	346 243	42,39	424 653	51,99	45 872	5,72	816 768
	20 à 24 ans	534 746	71,99	175 551	23,63	32 554	4,38	742 851
	25 à 54 ans	2 065 994	68,46	851 709	28,22	100 203	3,32	3 017 906
	55 à 64 ans	340 744	50,57	304 740	45,23	28 304	4,20	673 788
	65 ans et plus	74 373	9,46	675 876	85,93	36 298	4,61	786 547
	Total	3 362 100	55,68	2 432 529	40,29	243 231	4,03	6 037 860
HOMMES	15 à 19 ans	194 651	46,50	200 104	47,80	23 892	5,71	418 647
	20 à 24 ans	300 829	30,33	56 673	15,13	16 979	4,53	374 841
	25 à 54 ans	1 320 626	86,85	152 289	10,01	47 745	3,14	1 520 660
	55 à 64 ans	231 330	70,99	82 003	25,16	12 533	3,85	325 866
	65 ans et plus	52 347	15,73	264 988	79,61	15 541	4,67	332 876
	Total	2 099 783	70,64	756 057	25,43	116 690	3,93	2 972 530
FEMMES	15 à 19 ans	151 592	38,08	224 549	56,50	21 980	5,52	398 121
	20 à 24 ans	233 917	63,50	118 878	32,27	15 575	4,23	368 370
	25 à 54 ans	745 368	49,78	699 420	46,71	52 458	3,50	1 497 246
	55 à 64 ans	109 414	31,45	222 737	64,02	15 771	4,53	347 922
	65 ans et plus	22 026	4,86	410 888	90,57	20 757	4,58	453 671
	Total	1 262 317	41,18	1 676 472	54,69	126 541	4,13	3 065 330

5 Conclusions

Le système de contrôle et d'imputation élaboré à partir de la méthodologie décrite par Fellegi et Holt a été conçu pour être un système généralisé. Le principal motif de l'élaboration, cependant, était les besoins du recensement, comme en témoignent les problèmes vécus pendant le contrôle et l'imputation du recensement de 1971. Il s'agissait d'une tentative visant à mettre de l'ordre dans une opération complexe et potentiellement chaotique.

Le système a très bien réussi à atteindre cet objectif. Les données révisées étaient disponibles relativement plus tôt que les données de 1971. Il n'a pas été nécessaire d'apporter des correctifs après le contrôle. Les problèmes résiduels dans les données en général semblent moins graves que ceux relevés en 1971. Il y a beaucoup plus de connaissances sur les problèmes de données et les moyens de les corriger.

Ce système a en fait permis une analyse beaucoup plus critique des données et a permis de cerner les secteurs problématiques, comme l'erreur de réponse systématique et le biais de non-réponse. Les travaux à venir peuvent être axés sur une meilleure gestion de ces problèmes au sein d'une structure contrôlée.

Les quatre enjeux suivants sont quelques-uns des enjeux clés qui doivent être abordés, s'ils ne le sont pas déjà :

(1) Il faut trouver un moyen de traiter les erreurs systématiques qui peuvent être intégrées au système existant.

(2) Les solutions de rechange au principe de changement d'un nombre minimal de champs doivent être examinées. Ces solutions de rechange peuvent s'avérer peu utiles comparativement au traitement des erreurs systématiques.

(3) Il faudrait envisager des stratégies de traitement de la non-réponse pour tenir compte des différences entre la population répondante et la population non répondante.

(4) Un système expérimental de contrôle arithmétique et d'imputation est déjà en cours d'élaboration. L'intégration dans ce système de moyens de gestion des variables quantitatives et qualitatives fait partie des plans à long terme possibles.

Les erreurs ne peuvent pas être évitées, peu importe la minutie avec laquelle l'enquête est conçue. La pertinence de la stratégie de contrôle et d'imputation repose sur sa capacité de récupérer les valeurs « vraies ». Pour ce faire, il faut davantage de données empiriques sur la nature des erreurs dans les données.

Bibliographie

[1] Fellegi, I.P. et Holt, D., "A Systematic Approach to Automatic Edit and Imputation". *Journal of the American Statistical Association*, mars 1976. Volume 71, numéro 353.

[2] Graves, R.B., "Can-Edit, A Generalized Edit and Imputation System In A Data Base Environment". A report to the working party on electronic data processing, Conference of European Statisticians, (CES/WP.9/142). Février 1976.

[3] Turner, H.J., "The Use of Data Base Technology in Large Systems". A report to the working party on electronic data processing, Conference of European Statisticians, avril 1977.

[4] Quality of Date, '76 Census, Series 1: Sources of Error - Effect of Edit and Imputation Procedures on the Quality of Data from the 1976 Census of Population and Housing. N° 99-843 au catalogue.