Techniques d'enquête, juin 1986
Vol. 12, n° 1, pp. 53–65
Statistique Canada

Méthodes d'imputation dans un système généralisé

P. GILES et C. PATRICK[1]

RÉSUMÉ

Afin de répondre aux exigences de traitement de la plupart de ses enquêtes, Statistique Canada a mis sur pied un projet visant à élaborer un système généralisé de contrôle et d'imputation. Les auteurs de ce document analysent les diverses méthodes d'imputation qui ont été proposées dans le cadre de ce projet, pour traiter la non-réponse partielle. Ils se penchent aussi sur les aspects importants de l'application de ces propositions dans un système généralisé.

MOTS CLÉS: Modularité; prototype; imputation par enregistrement donneur; modèles de régression.

1. SYSTÈMES GÉNÉRALISÉS

Comme les ressources consacrées aux enquêtes sont de plus en plus limitées depuis quelques années, particulièrement en ce qui a trait à la recherche, on parle de plus en plus maintenant de logiciel général. Un logiciel général est une série de programmes machines intégrés en un système et qui permet à l'utilisateur de résoudre son problème par la solution la plus appropriée. Par exemple, un utilisateur doit prélever un échantillon d'enregistrements dans un fichier de données. Un système généralisé d'échantillonnage offrirait à l'utilisateur divers plans de sondage, par exemple l'échantillonnage aléatoire simple ou l'échantillonnage avec probabilités inégales (avec ou sans remise), l'échantillonnage systématique, l'échantillonnage stratifié ou encore le sondage en grappes.

Un système véritablement généralisé est presque par définition une chose complexe. La modularité est un moyen essentiel de réduire la complexité du système en divisant la tâche globale en un certain nombre de sous-tâches moins complexes. Chacune des sous-tâches ou fonctions est exécutée séquentiellement. De plus, à l'intérieur de chaque sous-tâche, l'utilisateur peut choisir parmi diverses méthodes d'exécution celle qui lui convient le mieux. Ainsi, non seulement est-il possible de diviser la tâche globale en tâches plus simples, mais il existe aussi plusieurs façons d'exécuter chaque sous-tâche.

La figure 1 montre comment le contrôle et l'imputation peuvent être répartis en trois sous-tâches. Ces trois sous-tâches sont le contrôle, la définition des zones à imputer et l'imputation proprement dite. Chaque section ou module d'une sous-tâche correspond à une méthode d'exécution particulière. Par exemple, C1 pourrait correspondre à une méthode quelconque d'imputation par enregistrements donneurs, C2 pourrait correspondre à une méthode d'imputation par la moyenne, et ainsi de suite. L'utilisateur choisirait un module par sous-tâche (A, B et C).

Il convient de souligner que le système représenté ci-dessus n'est pas le seul système de contrôle et d'imputation qui puisse exister. En fait, le système proposé dans le projet de Statistique Canada comporte cinq sous-tâches plutôt que trois. Si nous avons choisi d'illustrer un système à trois sous-tâches, ce n'est que pour des raisons de simplicité.

Chaque sous-tâche illustrée à la figure 1 représente une fonction trés précise. Les fichiers d'entrée nécessaires et les fichiers résultant doivent avoir une structure préétablie. L'utilisateur peut ainsi concentrer son attention sur le choix d'un module dans chaque sous-tâche, sachant

[1] Philip Giles et Charles Patrick, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Parc Tunney, Ottawa (Ontario), Canada, K1A 0T6.

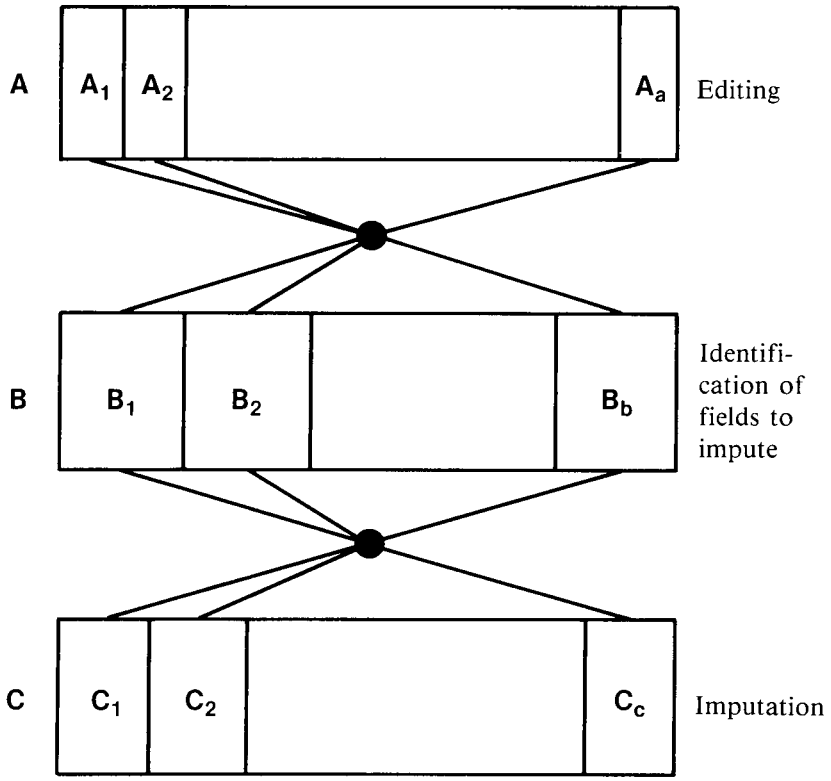

Figure 1. Exemple de système généralisé - Contrôle et imputation

que le système peut assurer les opérations de servitude (c'est-à-dire la gestion des fichiers et autres fonctions accessoires dont l'utilisateur ne souhaite pas avoir à se préoccuper). Même si le système peut accepter toutes les combinaisons de modules possibles, il se peut que certaines combinaisons ne soient pas acceptables ou même qu'elles soient incorrectes au point de vue logique. En général, c'est à l'utilisateur qu'il incombe de veiller à la compatibilité des modules.

Élaborer un système de traitement par une méthode modulaire a des conséquences importantes. On peut dire que le système est en quelque sorte toujours en développement puisqu'on peut et qu'en principe on devrait sans cesse y incorporer de nouveaux modules qui renferment des méthodes inédites et apportent des perfectionnements aux anciens modules. Cette extensibilité fait que le système se prête bien à la création de prototypes, notion très importante dans le domaine. La création de prototypes est une méthode par laquelle on crée un premier sous-ensemble de modules. Le système devient alors accessible à un certain nombre d'utilisateurs. D'autres modules viennent s'ajouter par la suite afin de répondre aux besoins d'autres utilisateurs. Ainsi, le principal avantage de la modularité et de la création de prototypes est qu'elles permettent d'envisager des améliorations graduelles du système et facilitent ces améliorations. Il y a toutefois une condition minimale mais indispensable à l'application de telles méthodes. Il est en effet nécessaire de définir soigneusement un système (comme celui représenté à la figure 1) et un cadre d'utilisation central (structure des fichiers de données et langage de programmation) et de les définir au tout début du processus général d'élaboration.

Outre les avantages liés à l'élaboration du système, il y a ceux qui peuvent découler de sa mise en application. Les possibilités qui s'offrent à l'utilisateur sont nombreuses. Si celui-ci doit choisir parmi plusieurs solutions qui semblent toutes aussi viables les unes que les

autres, il peut orienter son choix par des données chronologiques, en testant les diverses solutions avant de recueillir les données. Cette opération n'entraîne pas d'effort superflu. Une fois que le système généralisé est sur pied, les utilisateurs n'ont plus à mobiliser autant de ressources et le délai d'application s'en trouve raccourci.

L'application d'un système généralisé a toutefois ses inconvénients. Dans un contexte de production, un logiciel général pourra s'avérer moins efficace que le système sur mesure qui aurait autrement été utilisé. Au début, un système généralisé nécessitera plus de ressources qu'un système personnalisé. Mais le coût plus élevé doit être considéré à la lumiére du fait qu'il est encore plus cher de créer constamment des programmes sur mesure. Par ailleurs, il n'est pas raisonnable de penser qu'un système généralisé puisse répondre à toutes les conditions. Deux possibilités s'offrent alors à l'utilisateur. D'une part, il peut concevoir son propre module. Cette solution n'exige pas autant d'effort qu'une personnalisation complète. Toutefois, si cela se produit fréquemment, le système généralisé n'a plus sa raison d'être. D'autre part, l'utilisateur peut adapter ses exigences au système généralisé. Si le système a été bien conçu, une telle adaptation ne devrait pas avoir d'effets notables sur la qualité des données. Il convient aussi de souligner qu'il est souvent nécessaire de modifier les exigences originales dans l'élaboration d'un système personnalisé.

2. DONNÉES DE BASE SUR L'IMPUTATION

Dans le présent document, le mot "imputation" se rapporte à une certaine catégorie de procédures utilisées pour traiter les cas de non-réponse. Les données en entrée sont constituées d'un fichier de données saisies. La procédure d'imputation crée un fichier dont chaque enregistrement est "épuré"; un enregistrement "épuré" est un enregistrement où aucune donnée ne manque et qui est acceptée par tous les contrôles précisés. Si l'on veut créer un enregistrement épuré, une valeur doit être estimée pour toute valeur manquante.

Les contrôles précisés par l'utilisateur sont des contraintes logiques imposées aux valeurs que chaque variable peut prendre. Tous les contrôles, ensemble, définissent la région d'acceptation des données. Pour les données catégorielles, un contrôle est précisé comme un ensemble de combinaisons de valeurs de données acceptables. La région d'acceptation peut être représentée comme un ensemble de sommets dans un espace à N dimensions. Pour les données numériques, un contrôle est une égalité ou une inégalité linéaire. Le fait d'imposer la linéarité ne constitue pas une exigence trop restrictive, car un contrôle non linéaire peut être rendu linéaire au moyen de manipulations algébriques ou par l'addition de variables supplémentaires qui sont des fonctions non linéaires, définies de façon appropriée, des variables d'enquête. La région d'acceptation, dans le cas des données numériques, est un ensemble de régions convexes dans un espace à N dimensions. Il peut y avoir plus d'une région convexe parce qu'il est possible qu'il existe des contrôles conditionnels. Ces contrôles ne se rapportent qu'à un sous-ensemble d'enregistrements; par exemple les contrôles pertinents à un enregistrement particulier peuvent être fort différents selon que la valeur "Masculin" ou "Féminin" est donnée à la variable "Sexe".

Si un enregistrement particulier est rejeté à un ou plusieurs contrôles, il se peut qu'on ne puisse déterminer quelles variables sont erronées et doivent, par conséquent, faire l'objet d'une imputation. Par exemple, la combinaison $A + B \leq C$ sera rejetée au contrôle si l'enregistrement considéré possède les valeurs $A = 10$, $B = 5$, $C = 12$. Sept combinaisons de variables peuvent être modifiées pour obtenir un enregistrement épuré. Ce sont A, B, C, A & B, A & C, B & C, et A & B & C. Sans autre renseignement ou régle de décision, chacun de ces choix est également valide. Nous n'étudierons pas dans le présent document le problème de la détermination des variables à soumettre à l'imputation. Nous supposerons que, pour chaque enregistrement, ces variables ont été identifiées. Aucune distinction n'est faite ici entre les variables soumises à l'imputation à cause de valeurs manquantes et celles qui le sont à cause de rejets au contrôle.

3. TECHNIQUES D'IMPUTATION PROPOSÉES

La présente section comprend quatre sous-sections qui définissent toutes les techniques d'imputation proposées. Ce sont l'imputation déterministe, l'imputation par enregistrement donneur, les modèles de régression et les autres estimateurs pour l'imputation. L'utilisation des modèles de régression ainsi que la sous-section sur les autres estimateurs ne s'appliquent qu'aux données numériques. Les deux autres sous-sections s'appliquent aux données numériques et aux données catégorielles.

Presque toutes les techniques d'imputation peuvent être formulées selon un cadre de prévisions décrit de la façon suivante par Rubin (1976). On précise une distribution à plusieurs variables, $f(X_1, \ldots, X_N)$, qui résume le comportement statistique de la population des enregistrements complets. Cela peut être fait, que les variables individuelles soient quantitatives ou qualitatives. Sans perte de généralité, pour une enregistrement i pour lequel il faut effectuer une imputation, les N variables peuvent être divisées de la façon suivante: $X_1, \ldots, X_{m_i}$, sont les variables pour lesquelles il faut effectuer une imputation et $X_{m_{i+1}}, \ldots, X_N$, celles pour lesquelles aucune imputation n'est nécessaire. On peut alors obtenir une distribution conditionnelle $f(X_1, \ldots, X_{m_i} \mid x_{m_{i+1}}, \ldots, x_N)$. Les valeurs imputées, $y_1, \ldots, y_{m_i}$, sont choisies pour $X_1, \ldots, X_{m_i}$ à partir de l'ensemble

$$\{y_1, \ldots, y_{m_i} : f(y_1, \ldots, y_{m_i} \mid x_{m_{i+1}}, \ldots, x_N) > 0\}$$

On peut employer divers mécanismes de sélection. Cependant, comme nous l'avons mentionné ci-dessus, quelques-uns ne s'appliquent qu'à certains types de données.

Il faut remarquer que ces propositions ne contiennent rien de neuf ou de radicalement différent. Elles sont basées sur des travaux effectués auparavant à Statistique Canada et à l'extérieur. La discussion de l'imputation par enregistrement donneur est basée sur l'article de Fellegi et Holt (1976). La façon de déterminer une valeur à imputer au moyen d'un modèle est étudiée par Little (1982). D'autres articles connexes qui présentent un certain intérêt sont ceux de Sande (1976), de Kalton et Kasprzyk (1982) et de Kalton et Kish (1981).

3.1 Imputation déterministe

Le premier type d'imputation est appelé imputation déterministe. C'est le cas où une seule valeur peut être acceptée au contrôle. Si plus d'une variable doit faire l'objet d'une imputation dans un enregistrement particulier, une solution déterministe peut être possible pour certaines variables ou pour toutes. On doit vérifier si l'imputation déterministe peut être utilisée avant de passer à d'autres procédures d'imputation.

L'imputation déterministe peut se présenter dans des situations très simples et faciles à déterminer. Supposons, par exemple, qu'il existe un contrôle $A + B = 10$. Pour l'enregistrement étudié, A doit recevoir une valeur imputée et B a la valeur 6. Il est évident que $A = 4$ est la seule valeur qui puisse être acceptée au contrôle. Un autre exemple illustre les cas de ce genre pour les variables catégorielles. Supposons qu'un contrôle est énoncé de la façon suivante: "Si le lien avec la personne repère du ménage est épouse, alors le sexe doit être féminin". Si l'enregistrement repère contient "épouse" comme valeur pour le "lien avec la personne repère du ménage" et s'il faut faire une imputation pour la variable "Sexe", la seule valeur imputée valide est Sexe = Féminin.

Cependant, dans une enquête typique, on trouvera plusieurs contrôles plutôt qu'un seul. Cela peut signifier qu'une solution déterministe existante peut ne pas être apparente. La procédure utilisée pour vérifier si une imputation déterministe est possible consiste à trouver la région d'acceptation réduite définie par les contrôles actifs et les "bonnes" valeurs pour

les données. Les contrôles actifs sont définis comme étant le sous-ensemble de contrôles dans lequel figurent les variables à soumettre à l'imputation. Cela peut aussi être exprimé dans la notation du cadre de prévisions donné au début de la section 3. La distribution conditionnelle $f(X_1, \ldots, X_{m_i} \mid x_{m_{i+1}}, \ldots, x_N)$ déterminera une valeur unique pour certaines ou pour toutes les variables $X_1, \ldots, X_{m_i}$.

Donnons un exemple pour illustrer la procédure utilisée pour identifier l'imputation déterministe. Il faut remarquer que bien que l'exemple utilise des variables numériques, une situation analogue existe pour les variables catégorielles.

Considérons les trois contrôles suivants:

$$X + Y \leq 16,$$

$$Y + Z \leq 4,$$

$$X - 3Z \leq 8.$$

L'enregistrement considéré possède les valeurs suivantes:

$$X = 11 \textit{ et } Y = 3.$$

Il faut faire une imputation pour la variable Z.

Il n'est pas évident si l'imputation déterministe peut être utilisée. Dans une première étape, on doit considérer tous les contrôles actifs. Dans l'exemple, deux contrôles contiennent la variable Z.

$$Y + Z \leq 4,$$

$$X - 3Z \leq 8.$$

Il faut ensuite introduire les valeurs connues de X et Y dans ces contrôles pour déterminer la région d'acceptation réduite.

$$3 + Z \leq 4,$$

$$11 - 3Z \leq 8.$$

En réduisant ces inégalités nous obtenons la solution suivante.

$$Z \leq 1,$$

$$Z \geq 1.$$

Il est maintenant évident que $Z = 1$ est la seule valeur imputée valide qui soit possible.

Dans la plupart des situations réelles, l'incidence de l'imputation déterministe devrait être faible. Le contraire indiquerait que les contrôles sont plus restrictifs qu'il n'est nécessaire ou souhaitable, et cela devrait entraîner un réexamen des spécifications de contrôle. Cependant, comme l'imputation déterministe permet de réduire le problème de l'imputation, elle constitue une première étape nécessaire.

3.2 Imputation par enregistrement donneur

L'imputation par enregistrement donneur est une méthode qui apparie chaque enregistrement pour lequel on doit effectuer une imputation, l'enregistrement receveur, avec un enregistrement d'une population d'enregistrements donneurs définie. Une des méthodes utilisées pour déterminer la valeur à imputer consiste à reproduire directement dans l'enregistrement receveur la valeur tirée de l'enregistrement donneur. Pour les variables numériques, si les renseignements auxiliaires appropriés sont disponibles, on peut utiliser des méthodes plus complexes pour déterminer la valeur à imputer. Les estimateurs pour l'imputation par enregistrement donneur sont étudiés plus en détail dans la sous-section 3.3.

Habituellement, la population des enregistrements donneurs est définie comme l'ensemble de tous les enregistrements de l'enquête en cours pour lesquels aucune variable ne doit être soumise à l'imputation. Nous reportant au cadre de prévisions décrit au début de la section 3, cette situation implique que $f(X_1, \ldots, X_N)$ est la fonction de probabilité empirique. Cependant, il est possible d'utiliser d'autres méthodes pour définir la population d'enregistrements donneurs. Pour la suite de la discussion sur l'imputation par enregistrement donneur, nous supposerons tout simplement qu'une population d'enregistrements donneurs a été définie.

Les paires d'enregistrement donneur - enregistrement receveur sont formées au moyen de variables d'appariement. Les variables d'appariement sont définies comme des variables pour lesquelles aucune imputation ne doit être effectuée pour l'enregistrement receveur et pour lesquelles il existe une corrélation élevée avec la ou les variables à soumettre à l'imputation. De préférence, la corrélation entre les variables d'appariement doit en outre être faible. Deux variables d'appariement fortement corrélées auraient le même pouvoir discriminatif qu'une seule, mais elles auraient pour effet de doubler le poids donné à l'une quelconque de ces variables.

Dans le cas des variables catégorielles, un enregistrement donneur est choisi, au moyen d'une méthode aléatoire quelconque, parmi les enregistrements donneurs potentiels qui ont, pour les variables d'appariement, les mêmes valeurs que celles de l'enregistrement receveur. Comme les variables numériques peuvent prendre un nombre beaucoup plus considérable de valeurs que les variables catégorielles, il est très peu probable qu'il existera un appariement exact pour les variables d'appariement. Par conséquent, pour les données numériques, on utilise une fonction de distance pour définir la similarité. Cette fonction de distance est une fonction des variables d'appariement des enregistrements receveurs et des enregistrements donneurs potentiels. L'enregistrement donneur choisi est celui qui se trouve à la plus faible distance de l'enregistrement receveur. Habituellement, les variables d'appariement sont transformées pour le calcul des distances afin de supprimer l'effet de l'échelle dans laquelle la variable est mesurée. Par exemple, ce serait très ennuyeux pour l'utilisateur si la formation du couple enregistrement donneur - enregistrement receveur dépendait du fait que la variable de longueur est mesurée en mètres ou en pieds. Les transformations et fonctions de distance proposées sont étudiées plus loin.

Les variables d'appariement à utiliser peuvent être définies par l'utilisateur ou déterminées au moyen d'une procédure automatisée. Habituellement, à cause de contraintes de temps, toutes les décisions doivent être prises avant la collecte des données. Par conséquent, si la détermination des variables d'appariement est effectuée par l'utilisateur, ce dernier doit définir les variables d'appariement pour chaque combinaison possible de variables à soumettre à l'imputation. Si le fichier contient N variables, l'utilisateur doit définir $(2^{**}N) - 2$ spécifications. Il est évident que la valeur N n'a pas à être très grande pour que cette méthode devienne impossible à appliquer. Pour réduire ce nombre, il est possible de définir des variables d'appariement par strates. Tous les enregistrements receveurs d'une strate particulière utiliseraient les mêmes

variables d'appariement. Dans une telle situation, il est possible (selon le soin apporté par l'utilisateur pour définir les variables d'appariement) qu'un enregistrement donneur particulier ait une variable d'appariement pour laquelle il faut effectuer une imputation. L'utilisateur qui introduit les spécifications des variables d'appariement doit savoir que cette décision peut entraîner une augmentation considérable de la charge de travail.

Nous proposons une méthode qui peut être utilisée pour déterminer automatiquement les variables d'appariement. Cette procédure peut être utilisée de façon analogue pour les données numériques ainsi que pour les données catégorielles. Voici en substance en quoi cette procédure consiste. L'ensemble des variables d'appariement doit contenir au moins les variables qui se trouvent dans les règles de contrôle avec les variables à soumettre à l'imputation. D'après la définition que nous avons déjà donnée, ce sont les contrôles actifs. Cette méthode semble intuitivement raisonnable car il est souhaitable qu'il existe une corrélation entre les variables d'appariement et les variables à soumettre à l'imputation. Les variables dans les contrôles actifs limitent la gamme de valeurs possibles à imputer. Cela implique un certain type de dépendance ou de structure de corrélation.

L'emploi de cette procédure d'appariement avec la transcription directe a une conséquence importante pour les variables catégorielles. Nous sommes assurés que toutes les variables imputées seront acceptées aux contrôles. Cela est très important car c'est une condition requise pour créer un enregistrement épuré. Sans cette assurance, l'utilisateur doit effectuer un nouveau contrôle des enregistrements et peut-être adopter une procédure d'imputation secondaire. Pour les données numériques, la similarité, telle que définie par une fonction de distance, ne garantit pas ce résultat. Cependant, plus la distance entre les enregistrements donneur et receveur approche de zéro, plus la probabilité que les valeurs imputées seront acceptées aux contrôles est élevée.

Un exemple nous permettra d'illustrer la détermination des variables d'appariement au moyen de cette procédure automatisée.

Supposons les cinq contrôles suivants:

$$\text{I.} \quad A + B \leq \alpha_1,$$

$$\text{II.} \quad B - E \leq \alpha_2,$$

$$\text{III.} \quad C + 2D + 3E \leq \alpha_3,$$

$$\text{IV.} \quad A + C + D \leq \alpha_4,$$

$$\text{V.} \quad A - 2B + C \leq \alpha_5.$$

Il y a cinq variables d'enquête A, B, C, D, E et α_1, α_2, α_3, α_4, α_5 sont des scalaires connus.

Pour l'enregistrement receveur à l'étude, il faut faire une imputation seulement pour la variable B.

La première étape consiste à identifier les contrôles actifs. Dans cet exemple, il y en a trois. Ce sont les contrôles I, II et V.

La deuxième étape consiste à déterminer les variables actives. On entend par variable active toute variable contenue dans au moins un des contrôles actifs. Dans l'exemple, il y a quatre variables actives: A, B, C, E. Il faut remarquer que, par définition, toutes les variables à soumettre à l'imputation se trouvent parmi les variables actives.

La troisième étape consiste à déterminer les variables d'appariement; ce sont les variables actives pour lesquelles aucune imputation n'est requise. Dans cet exemple, les variables d'appariement sont A, C, E.

L'imputation par enregistrement donneur pour les données numériques exige, en plus de la détermination des variables d'appariement, le choix d'une transformation des données ainsi que d'une fonction de distance.

Deux types de transformations des données sont proposés ici. Dans les deux cas, chaque variable doit être transformée indépendamment. Les deux transformations proposées sont la transformée des valeurs de rang et la transformée d'échelle de localisation.

Pour la transformée des valeurs de rang, les valeurs de chaque variable sont triées, les valeurs de rang sont ensuite divisées par une constante appropriée afin que toutes les valeurs se trouvent dans l'intervalle de zéro à un. Les valeurs transformées sont réparties uniformément sur cet intervalle.

La transformée d'échelle de localisation a la forme suivante:

$$y^T = \frac{1}{b}(y - a),$$

où y^T est la valeur transformée,
y est la valeur originale de la donnée,
a, b sont des paramètres précisés par l'utilisateur.

Deux choix fréquents pour ces constantes sont, premièrement, que a soit la moyenne de l'échantillon et b l'écart-type de l'échantillon et, deuxiémement, que a soit la valeur minimale de l'échantillon et b l'étendue des valeurs de l'échantillon. D'autres options sont possibles.

Dans le choix d'une transformation pour les données, il faut tenir compte de la robustesse et des observations extrêmes aberrantes. La transformée des valeurs de rang est très robuste par rapport aux changements dans les valeurs des données et elle ramène les observations extrêmes aberrantes plus près des autres valeurs des données. Cela peut être ou ne pas être souhaitable. Aucune limite n'est imposée aux valeurs transformées quand on utilise la transformée d'échelle de localisation avec la moyenne et l'écart-type. Ces paramètres sont également sensibles aux valeurs extrêmes aberrantes. Le choix de la valeur minimale et l'étendue des valeurs permettrait de limiter les valeurs transformées entre zéro et un. Cependant, ces paramètres sont très sensibles aux valeurs extrêmes. Une valeur très grande pourrait donner à toutes les valeurs transformées, sauf une, une valeur très rapprochée de zéro.

Pour le choix de la fonction de distance, nous proposons une famille de fonctions de distance. Ce sont les normes $\mathcal{L}^p$ pondérées, où p est une constante précisée par l'utilisateur. Voici la forme générale de ces fonctions:

$$D(X, Y) = \left[\sum_{k=1}^{r} w_k |x_k - y_k|^p \right]^{1/p},$$

où x_k, y_k sont les r variables d'appariement dans les deux enregistrements,
w_k sont des poids précisés par l'utilisateur,
p est une constante précisée par l'utilisateur.

Les poids sont utilisés si l'utilisateur désire que certaines des variables d'appariement contribuent plus que d'autres au calcul de la distance. Les valeurs implicites servent à donner à tous les poids la valeur un.

Trois choix particuliers d'une valeur de p présentent un intérêt spécial, $p = 1$, $p = 2$, et $p = \infty$. Quand $p = 1$, cette fonction calcule les distances entre les îlots. Quand $p = 2$, c'est la distance euclidienne qui est calculée. Le cas limite de cette fonction, quand $p = \infty$, donne la distance minimax. Pour ce choix de p, la fonction s'écrit de la façon suivante:

$$D(X, Y) = \max_{1 \le k \le r} [w_k |x_k - y_k|].$$

Le dernier point à mentionner à propos de l'imputation par enregistrement donneur est le concept d'une "pénalité" pour utilisation d'enregistrements donneurs. Cette pénalité réduirait le nombre de fois où un enregistrement donneur particulier est utilisé. Pour l'imputation par enregistrement donneur de données catégorielles, un enregistrement donneur est choisi sans remplacement, dans la population d'enregistrements donneurs. Cette stratégie doit être légérement modifiée si la taille de la population d'enregistrements receveurs est supérieure à celle de la population d'enregistrements donneurs.

Dans le cas des données numériques, on modifie la fonction de distance en augmentant la distance selon le nombre de fois qu'un enregistrement donneur particulier est utilisé. Une façon d'atteindre ce résultat consiste à utiliser $D'(X, Y)$ pour calculer les distances, où

$$D'(X, Y) = D(X, Y) \times (1 + ud),$$

où u représente la "pénalité" imposée par l'utilisateur,

d représente le nombre de fois qu'un enregistrement donneur a été choisi.

L'imposition d'une pénalité pour la fonction de distance implique que le choix d'un enregistrement donneur pour chaque enregistrement receveur dépend maintenant de l'ordre des enregistrements receveurs.

3.3 Modèles de régression

La présente sous-section traite les estimateurs d'imputation obtenus à l'aide de modèles de régression. Pour cette étude, nous n'utiliserons que deux modèles. Ce sont:

MODÈLE I : $y_i = \alpha + \epsilon_i, \qquad \text{Var}(\epsilon_i) = \sigma^2,$

MODÈLE II: $y_i = \beta x_i + \epsilon_i, \qquad \text{Var}(\epsilon_i) = \sigma^2 x_i.$

Il faut remarquer que ces modèles constituent des cas spéciaux de la forme plus générale des modèles de régression, qui est la suivante:

$$\underset{\sim}{y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{\epsilon},$$

$$\text{où } E(\underset{\sim}{\epsilon}) = \underset{\sim}{0}, \; V(\underset{\sim}{\epsilon}) = \underset{\sim}{V}.$$

Le modèle II est employé quand des données auxiliaires sont disponibles. Sinon, le modèle I est utilisé. Dans les deux modèles, un paramètre doit être estimé. La méthode des moindres carrés permet d'obtenir les estimations suivantes pour les paramètres:

$$\hat{\alpha} = \bar{y},$$

$$\hat{\beta} = \frac{\bar{y}}{\bar{x}}.$$

Avant d'énoncer les divers estimateurs proposés, nous présenterons une notation particulière.

Soit t l'indice inférieur pour le temps t, l'enquête actuelle;

y_{it} la variable à l'étude pour l'unité i et le temps t; c'est la valeur à imputer pour les enregistrements receveurs;

x_{it} la variable auxiliaire (corrélée avec Y) pour l'unité i et le temps t;

R l'indice inférieur pour tous les répondants au temps t (c'.-à-d. que y_{it} est connu);

NR l'indice inférieur pour tous les non-répondants au temps t (c.-à-d. que y_{it} doit faire l'objet d'une imputation),

C, D les indices supérieurs qui indiquent soit un enregistrement receveur (C), soit un enregistrement donneur (D), chaque fois que la distinction est requise.

Plusieurs notes explicatives doivent accompagner la notation. D'abord, les indices R et NR sont tels que définis dans l'enquête, peu importe que chaque enregistrement corresponde à un répondant ou à un non-répondant. Deuxièmement, les valeurs des variables $y_{i(t-1)}$, x_{it}, $x_{i(t-1)}$ peuvent elles-mêmes avoir été imputées. La seule contrainte est qu'aucune d'entre elles ne peut être omise. Troisièmement, la notation n'inclut pas le concept de classes d'imputation. Les classes d'imputation sont essentiellement définies après les strates, en ce sens qu'elles définissent des ensembles d'enregistrements jugés homogènes à l'intérieur des groupes et hétérogènes entre les groupes. Cependant, la notation ainsi que les estimateurs d'imputation peuvent facilement être étendus ou augmentés pour inclure les classes d'imputation.

Les estimateurs peuvent donc être classés selon:

(i) le choix du modèle, I ou II,

(ii) le groupe d'imputation,

(iii) les variables dans la régression utilisée pour estimer le paramètre.

Les données des enregistrements qui font partie du groupe d'imputation précisé sont exactement celles utilisées pour estimer le(s) paramètre(s) dans le modèle. Ce concept permet une grande flexibilité. Par exemple, il permet d'exclure les valeurs extrêmes aberrantes du calcul de la valeur estimative du paramètre. Une fois le paramètre estimé, ce dernier est utilisé à des fins de prévision pour déterminer la valeur imputée. Dans la notation employée, Y_t représente toujours la variable prévue.

D'après ces deux modèles, nous proposons huit estimateurs d'imputation et cette liste peut éventuellement être augmentée. On pourrait obtenir des estimateurs additionnels en choisissant, par exemple, d'autres modèles qui pourraient comprendre plus de variables.

On peut voir, en parcourant la liste de ces estimateurs, qu'il s'agit des estimateurs d'imputation familiers utilisés traditionnellement.

Estimateur 1: La valeur tirée de l'enquête antérieure pour la même unité est imputée.
$y_{i(t-1)}$

Estimateur 2: La valeur moyenne est tirée de l'enquête précédente et est imputée.
$\bar{y}_{(t-1)}$

Estimateur 3: La valeur moyenne, pour tous les répondants, pour l'enquête courante est imputée. $\bar{y}_{tR}$

Estimateur 4: La valeur est reproduite directement dans l'enregistrement receveur à partir de l'enregistrement donneur. y_{it}^{D}

Estimateur 5: Une estimation, par la méthode du quotient, qui utilise des valeurs tirées de l'enquête courante, est imputée.
$$\frac{\bar{y}_{tR}}{\bar{x}_{tR}} x_{it}$$

Estimateur 6: Une estimation, par la méthode du quotient, basée sur les valeurs des enregistrements donneur et receveur est imputée.
$$\frac{y_{it}^{D}}{x_{it}^{D}} x_{it}$$

Estimateur 7: La valeur tirée de l'enquête antérieure pour la même unité, avec un ajustement de la tendance calculé à partir d'une variable auxiliaire, est imputée.
$$\frac{y_{i(t-1)}}{x_{i(t-1)}} x_{it}$$

Estimateur 8: La valeur tirée de l'enquête antérieure pour la même unité, avec un ajustement de la tendance calculé à partir du changement dans les valeurs déclarées pour la variable Y, est imputée.
$$\frac{\bar{y}_{tR}}{\bar{y}_{(t-1)R}} y_{i(t-1)}$$

Il est intéressant de faire ressortir les différences dans les estimateurs quant on fixe tous les éléments de classification sauf un. Par exemple, la différence entre les estimateurs 1 et 2 n'est due qu'à la différence dans le choix du groupe d'imputation, ce qui est aussi le cas pour les estimateurs 3 et 4 et pour les estimateurs 5 et 6. La différence entre les estimateurs 1 et 7 n'est due qu'au choix du modèle. Cela vaut aussi pour les estimateurs 3 et 5 et pour les estimateurs 4 et 6. Il faut aussi remarquer que les estimateurs 4 et 6 sont ceux utilisés lors de l'imputation par enregistrement donneur, sujet traité dans la sous-section 3.2.

3.4 Autres estimateurs d'imputation

Le choix des techniques d'imputation dépend des hypothèses faites par l'utilisateur à propos de l'ensemble des non-répondants. Quand on utilise l'imputation par enregistrement donneur, on suppose qu'il existe des répondants qui ont des caractéristiques semblables à celles de chaque non-répondant. Si on impute la moyenne de l'enquête courante, on suppose que la valeur moyenne pour les répondants est la même que celle qui s'applique aux non-répondants. De même, on peut étudier tous les estimateurs et dresser la liste des hypothèses implicites. Le premier estimateur proposé dans la présente sous-section vise à assouplir les hypothèses quelque peu restrictives (et habituellement fausses) imposées dans la sous-section précédente; par contre cet estimateur est plus complexe. Cet estimateur, dit "chain link" (estimateur en chaîne), a été proposé par Madow et Madow (1978).

Voici comment cet estimateur est obtenu. Nous supposons d'abord que le taux de variation (tendance) des populations de non-répondants et de répondants sont les mêmes que ceux qui ont été observés lors de l'enquête précédente, ce qui nous permet d'estimer la moyenne de la variable Y pour la population dans le cas de la population des non-répondants lors de l'enquête courante.

$$\bar{y}_{NRt} = \frac{\bar{y}_{NR(t-1)}}{\bar{y}_{R(t-1)}} \bar{y}_{Rt}.$$

On peut alors déterminer la valeur imputée d'après la variable auxiliaire.

$$y_{it} = \frac{\bar{y}_{NRt}}{\bar{x}_{NRt}} x_{it}$$

$$= \frac{\bar{y}_{NR(t-1)}}{\bar{x}_{NRt}} \quad \frac{\bar{y}_{RT}}{\bar{y}_{R(t-1)}} x_{it}$$

Il faut remarquer que cela équivaut à une application plus complexe de la méthode du modèle de régression étudiée à la sous-section 3.3. D'abord, il faut imputer temporairement $y_{it} = \bar{y}_{NRt}$, comme cela est précisé ci-dessus. Il faut ensuite utiliser le modèle II et définir le groupe d'imputation comme l'ensemble de tous les enregistrements des non-répondants lors de l'enquête courante pour la variable Y. La variable de réponse est Y_t. La variable explicative est X_t. L'estimateur résultant correspond au résultat de la formule donnée plus haut.

Le second estimateur proposé dans la présente sous-section peut être utilisé quand on possède des données sur la variable Y pour plusieurs enquêtes antérieures. Cette méthode n'utilise pas de variables auxiliaires ou de données provenant d'autres enregistrements. On considère le comportement de chaque non-répondant indépendamment des autres. Cette

méthode est appelée lissage exponentiel. Il s'agit d'une technique de prévision utilisée couramment en économétrie. Un paramètre doit être précisé par l'utilisateur. Ce paramètre permet de modifier l'apport relatif des diverses valeurs des données. Algébriquement, l'estimateur est donné par la formule suivante:

$$y_{it} = \frac{1-A}{1-A^t} \sum_{r=0}^{t-1} A^r y_{i(t-r-1)},$$

où $0 < A < 1$, est précisé à l'avance.

Plus A est près de zéro, plus le poids des données récentes est important. Si $t = 1$, cela revient à imputer la valeur déclarée lors de l'enquête précédente.

4. TRAVAUX ANTÉRIEURS EFFECTUÉS À STATISTIQUE CANADA

Statistique Canada a essayé, dans le passé, d'élaborer un système généralisé de contrôle et d'imputation. Deux de ces tentatives seront soulignées car elles forment la base du présent document. Ce sont le système CAN-EDIT et le système de contrôle et d'imputation des données numériques (NEIS).

4.1 CAN-EDIT

Le système CAN-EDIT lui-même n'est pas un système complètement général, mais la méthode qu'il utilise l'est. Le système est basé sur le document de Fellegi et Holt (1976) sur l'imputation pour les données catégorielles. Il a été élaboré pour dépouiller les données du recensement de la population et du logement effectués en 1976 et 1981 au Canada.

La méthode utilisée dans le système CAN-EDIT est celle de l'imputation par enregistrement donneur. Les variables d'appariemment ont été déterminées automatiquement au moyen des procédures décrites dans la sous-section 3.2. Le système CAN-EDIT emploie ce qui est appelé l'imputation primaire et l'imputation secondaire. Si l'imputation relative à un enregistrement receveur ne peut être effectuée par l'imputation primaire, cet enregistrement passe à l'imputation secondaire.

Dans l'imputation primaire, toutes les valeurs imputées sont tirées du même enregistrement donneur. Les variables d'appariemment ont été déterminées d'après toutes les variables à imputer. Un enregistrement est rejeté lors de l'imputation primaire si aucun enregistrement donneur ne possède des valeurs identiques à celles des variables d'appariemment.

Dans l'imputation secondaire, chacune des variables à imputer est traitée indépendamment et séquentiellement. La procédure utilisée pour déterminer les variables d'appariemment est la même. Cependant, comme on ne considère qu'une variable à la fois, le nombre de variables d'appariemment sera en général moindre que lors de l'imputation primaire. (Il ne peut y en avoir plus, mais il pourrait y en avoir autant.) Cela signifie que la population d'enregistrements donneurs potentiels est plus grande. L'imputation secondaire présente quelques désavantages par rapport à l'imputation primaire. Tout d'abord, il est possible de choisir, comme variable d'appariemment, une variable qui doit être soumise à l'imputation. Il n'existe pas alors de valeur avec laquelle effectuer un appariemment. Deuxièmement, cette méthode n'utilise pas les distributions conjointes des variables. Les valeurs imputées pour les deux variables peuvent être acceptées aux contrôles, chacune peut être une valeur réellement valide, mais il se peut qu'elles ne se trouvent que rarement combinées dans la population.

4.2 Système de contrôle et d'imputation des données numériques (NEIS)

Le système NEIS est le premier prototype d'un système généralisé de C&I pour les données numériques. Ce système a été écrit comme un ensemble de modules du progiciel statistique PSTAT. Aucun autre prototype n'a été élaboré par la suite. Ce système est décrit dans l'ouvrage de Gordon Sande (1979), qui l'a élaboré. Nous pensons que la méthodologie

employée est très valable et qu'elle devrait être incorporée dans un nouveau système. Cependant, il se peut que le logiciel PSTAT ne constitue plus un logiciel approprié. Le système NEIS a été utilisé, dans un cadre d'exploitation, lors de l'Enquête de 1981 sur l'utilisation de l'énergie dans les fermes. La méthodologie a aussi été employée lors de l'élaboration du système de dépouillement du Recensement de l'agriculture de 1981.

Le système NEIS, comme le système CAN-EDIT, utilise la méthode de l'imputation par enregistrement donneur avec des variables d'appariemment déterminées automatiquement au moyen de la procédure décrite dans la sous-section3.2. Cependant, comme nous l'avons expliqué dans cette sous-section, quand les variables d'appariemment sont déterminées de cette façon pour des données numériques, la procédure d'imputation ne produit pas toujours un enregistrement épuré. La stratégie adoptée pour réduire l'importance de ce problème consiste à choisir les enregistrements donneurs les plus rapprochés. Si l'enregistrement donneur le plus rapproché ne permet pas d'imputer des valeurs qui sont acceptées aux contrôles, c'est alors l'enregistrement donneur le plus rapproché suivant qui est considéré et ainsi de suite.

Le système NEIS ne permet pas à l'utilisateur de choisir une fonction de transformation ou une fonction de distance. Il utilise la transformée des valeurs de rang et la norme $\mathcal{L}^\infty$ pondérée pour le calcul des distances.

5. CONCLUSION

Les suggestions présentées ici offrent un choix considérable à tout utilisateur d'un système généralisé de contrôle et d'imputation. Comme nous l'avons dit, ces suggestions n'excluent pas la possibilité d'autres méthodes. Cependant, nous pensons qu'un système élaboré avec ces éléments répondrait aux besoins d'un grand nombre d'utilisateurs. L'expérience des auteurs permet d'affirmer que la puissance et l'utilité ultimes d'un tel système ne sont pas évidentes avant qu'on ne commence à l'utiliser. À mesure que les essais se déroulent, il devient apparent qu'un tel système comporte plus de possibilités et qu'il peut être augmenté plus qu'on ne le pensait d'abord.

BIBLIOGRAPHIE

FELLEGI, I.P., et HOLT, D. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, 71, 17-35.

KALTON, G., et KASPRZYK, D. (1982). Imputing for missing survey responses. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 22-31.

KALTON, G., et KISH, L. (1981). Two efficient random imputation procedures. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 146-151.

LITTLE, R.J.A. (1982). Models for non-response in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.

MADOW, L.H., et MADOW, W.G. (1978). On link relative estimators. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 534-539.

RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

SANDE, G. (1976). Searching for numerically matched records. Rapport technique, Division des méthodes d'enquêtes - entreprises, Statistique Canada.

SANDE, G. (1979). *The Numerical Edit and Imputation Subsystem for PSTAT — A User's Guide.* Sous-division de la recherche et des systèmes généraux, Statistique Canada.