Recueil du symposium 2002 de Statistique Canada
Modélisation des données d'enquête pour la recherche économique et sociale

EXEMPLES D'IMPUTATION MULTIPLE

Nicholas T. Longford[1]

RÉSUMÉ

Je citerai trois exemples d'imputation multiple portant sur des données ou de l'information incomplètes dans des enquêtes à grande échelle. Chacun illustre la façon d'améliorer une méthode d'imputation simple en y tenant compte de l'incertitude qui s'attache aux valeurs manquantes. Mon propos n'est pas de présenter de nouvelles méthodes ni de démontrer des résultats théoriques, mais de faire valoir à l'intention du praticien la manière d'appliquer couramment une méthode d'imputation multiple à l'analyse de données d'enquête à grande échelle.

MOTS CLÉS : information auxiliaire, erreur de mesure, données manquantes, imputation multiple, valeurs plausibles.

1. INTRODUCTION

Les valeurs manquantes sont une lacune qu'accusent inévitablement en majeure partie les données d'enquête à grande échelle. Qu'on puisse recueillir toutes les données prévues par le plan d'échantillonnage dans la plupart des enquêtes a tout d'un impossible idéal, puisqu'on ne peut s'attendre à ce que les sujets éprouvent l'obligation de s'en tenir au protocole de collecte de données et aux attentes qui y sont liées, qu'il s'agisse de la disponibilité des enquêtés ou de la possession et de la communication volontaire des renseignements demandés. Cela ne devrait pas décourager tout effort de réduction de la non-réponse et d'autres sources d'inachèvement de l'information. Cependant, il ne convient pas non plus de faire fi du problème de l'inachèvement au stade de l'analyse en prétextant qu'on a fait tous les efforts voulus pour réunir les données exhaustives prévues.

Évitons les longs préambules en posant que les données d'enquête à analyser sont recueillies par l'application d'un certain plan d'échantillonnage et qu'on dispose déjà d'un programme d'analyse bien arrêté, c'est-à-dire d'une liste d'éléments d'analyse devant chacun nous donner une estimation $\hat{\theta}$ avec sa variance estimée d'échantillonnage, $\hat{s}^2 = \hat{\text{var}}(\hat{\theta})$, ou la forme pluridimensionnelle de cette paire, à savoir le vecteur $\hat{\boldsymbol{\theta}}$ et la matrice $\hat{\text{var}}(\hat{\boldsymbol{\theta}})$. L'inconvénient est que ces quantités relatives à l'échantillon seraient d'un calcul simple seulement si l'ensemble de données réunies par voie d'enquête était complet ou avait la même forme qu'un ensemble complet. On disposerait alors habituellement d'une matrice rectangulaire de *n* sujets et de *p* variables complètement observées.

Pour nous, chaque $\hat{\theta}$ est un *estimateur de données complètes*, et nous supposons que $\hat{\theta}$ serait une valeur efficiente et exempte de biais si les données étaient complètes. Nous posons en outre que $\hat{s}^2$ serait sans biais pour $\text{var}(\hat{\theta})$ si les données étaient complètes. Une notation plus rigoureuse, même s'il ne s'agit pas d'une notation standard, permet d'éviter certaines expressions longues et maladroites. Nous désignons par $\mathbf{X}^*$ les données complètes, c'est-à-dire l'ensemble de données que nous prévoyions recueillir, et par $\mathbf{X}$ l'ensemble de données effectivement réunies (données *incomplètes* d'observation). À chaque estimateur $\hat{\theta}$, nous associons l'ensemble de données auquel il s'applique. Ainsi, avec des ressources limitées de collecte de données (exécution de l'enquête), l'obtention de l'estimation $\hat{\theta}(\mathbf{X}^*)$ des données complètes est l'idéal visé. Il est impossible d'évaluer le « substitut » $\hat{\theta}(\mathbf{X})$ à moins

[1]Université De Montfort, immeuble James Went 2–8, The Gateway, Leicester, Angleterre, LE1 9BH (ntl@dmu.ac.uk).

de définir des règles appropriées d'exécution des opérations sur les valeurs manquantes. Une fois ces règles définies, nous devrions réévaluer l'efficience de $\hat{\theta}(\mathbf{X})$.

Des deux façons naturelles de « réparer » les données pour qu'elles puissent s'analyser par les méthodes (algorithmes, programmes informatiques, etc.) conçues pour des données complètes, à savoir les méthodes de *réduction* et d'*achèvement* des données, nous ne retiendrons que la seconde. Dans une réduction, l'échantillon est ramené à la dimension des sujets dont les enregistrements sont complets, d'où l'obtention de l'ensemble de données $\mathbf{X}_{-}$. Si on devait écarter les données des sujets dont les enregistrements sont presque complets (là où seule une légère partie des *p* éléments d'information manque), il y aurait gaspillage d'information (pourtant recueillie à un coût appréciable). Il se peut aussi que le sous-échantillon de réponses complètes ne soit pas représentatif de la population observée, bien que l'échantillon le soit dans son ensemble. Bref, $\hat{\theta}(\mathbf{X}_{-})$ pourrait être très peu efficient (et entaché d'un biais), bien que $\hat{\theta}(\mathbf{X}^{*})$ soit efficient.

En complétant l'ensemble de données, c'est-à-dire en imputant une valeur appropriée à chaque élément manquant, nous produisons un ensemble de données $\mathbf{X}_{+}$ paraissant plus riche en information que l'ensemble de données effectivement réunies. Ainsi, $\hat{s}^2(\mathbf{X}_{+})$ se trouvera à sous-estimer la variance d'échantillonnage de $\hat{\theta}(\mathbf{X}_{+})$. Il se peut en outre que, selon la méthode d'imputation employée, $\hat{\theta}(\mathbf{X}_{+})$ soit biaisé. Les difficultés d'analyse de $\mathbf{X}_{+}$ s'expliquent par le défaut d'avoir tenu compte dans $\hat{\theta}$ des différences d'état entre valeurs observées et valeurs imputées. Les valeurs imputées sont des réponses devinées qui ajoutent de l'incertitude (variation) aux éventuelles irrégularités de l'échantillonnage.

Entre la population et un ensemble de données (incomplet), il y a un échantillonnage et une non-réponse qui, tous deux, réduisent l'information, le premier de la population à un ensemble de données complet et la seconde, d'un ensemble complet à un ensemble incomplet. Nous sommes maîtres de l'échantillonnage (nous le définissons par un plan d'échantillonnage), moyen premier de production d'estimateurs efficients des quantités d'intérêt. La non-réponse pourrait être décrite comme un échantillonnage dont nous ne connaissons pas le détail. L'ensemble de données incomplet ne renseigne nullement sur certaines de ses caractéristiques. Nous nous trouvons seulement à observer la composition des deux, c'est-à-dire de l'échantillonnage (délibéré, conçu pour répondre aux limites des ressources disponibles) et de la non-réponse (lacune créée par une collaboration imparfaite des sujets).

Pour combler cette lacune (perte d'information) causée par la non-réponse, nous devrions essayer de récupérer les inférences que des données complètes auraient permis d'obtenir. C'est la raison d'être d'une imputation tant simple que multiple. L'imputation simple (IS) est une tentative de refaire des données complètes en « réparant » le mieux possible chacune des valeurs manquantes. Dans une imputation multiple (IM), la récupération d'un ensemble complet est subordonnée à des objectifs :

1. d'estimation efficiente de θ par une estimation de $\hat{\theta}(\mathbf{X}^{*})$;
2. d'estimation sans biais de la variance d'échantillonnage par rapport à la composition de l'échantillonnage et de la non-réponse.

Tenir compte de l'incertitude qui s'attache aux valeurs manquantes est ce qui caractérise principalement l'imputation multiple.

Les méthodes hybrides où on impute seulement une partie des valeurs manquantes et où on réduit ensuite l'ensemble achevé pour qu'il présente la forme standard, héritent des insuffisances des méthodes tant de réduction que d'imputation simple.

1.2 Arithmétique de l'incertitude

On peut illustrer les insuffisances de toute méthode d'imputation simple par l'exemple élémentaire suivant. Supposons qu'un ensemble de données d'enquête n'a de valeurs manquantes que pour une variable. L'application

d'une méthode IS à un élément manquant quelconque *x* peut donner une valeur nulle. En admettant que nous soyons tout à fait sûrs que la réponse aurait été 0 – bien que +1 et –1 soient aussi des possibilités bien distinctes –, on peut exprimer le tout en associant la valeur nulle à la probabilité 0,8 et les valeurs +1 et –1 à la probabilité 0,1 chacune. Le problème étant ainsi présenté, il pourrait être difficile de songer à imputer une valeur autre que zéro, mais une fois que nous considérons les opérations auxquelles la valeur serait soumise, les défauts de ce choix deviennent évidents. Si on se sert uniquement de fonctions linéaires pour évaluer $\hat{\theta}$ et $\hat{s}^2$, il convient de choisir 0, mais s'il s'agit d'une fonction quadratique comme $\sum_i x_i^2$, ce choix devient irrationnel, puisqu'il y aurait substitution de la plus petite valeur possible dans le cas de x^2 ! L'espérance de la distribution calculée pour x^2, qui correspond à 0,2, représente nettement le « bon » choix, mais comment régler le paradoxe $0^2 = 0{,}2$?

L'identité $\mathrm{E}(x^2) = \{\mathrm{E}(x)\}^2 + \mathrm{var}(x)$ nous offre la réponse pour toute variable aléatoire *x* ou, plus généralement, $\mathrm{E}\{f(x)\} \neq f\{\mathrm{E}(x)\}$ pour la plupart des fonctions non linéaires *f* (qui sont toutes strictement convexes ou concaves, par exemple) si *x* n'est pas une valeur dégénérée. Ainsi, s'il y a imputation à l'égard d'un élément d'information, la valeur en question peut convenir à certaines fonctions, mais non aux autres (pour la plupart).

Si nous « réparons » les données par des imputations appropriées pour une fonction, nous ne réparons rien pour une autre fonction. Comme les estimateurs d'une quantité relative à la population et de sa variance d'échantillonnage comportent d'ordinaire des fonctions différentes des données (fonction linéaire pour l'estimateur de quantité et quadratique pour la variance), il est impossible de réparer les données même pour une analyse simple. Une des quantités évaluées habituellement, $\hat{s}^2$, devra faire l'objet d'une certaine correction. Dans le cas de données d'enquête soumises à des analyses nombreuses d'une nature et d'une complexité diverses, une correction propre à ces diverses analyses n'a rien de viable, surtout si les analystes connaissent peu les questions à traiter et ne disposent ni des logiciels ni du matériel voulus pour effectuer des analyses non standard.

On peut reformuler cette argumentation dans le langage de l'algorithme EM espérance-maximisation (Dempster, Laird et Rubin, 1977), bien que celui-ci vise une fonction de maximum de vraisemblance (MV). Il est impossible d'exprimer facilement comme estimateurs MV certains estimateurs utilisés dans des enquêtes. À l'étape « espérance » de l'algorithme EM, on estime les contributions respectives des valeurs manquantes à la fonction logarithmique de vraisemblance des données complètes par leurs espérances conditionnelles compte tenu des données et des estimations existantes des paramètres. Les estimations des valeurs manquantes (en espérance conditionnelle) livreraient des estimations différentes de leurs apports respectifs, et on obtiendrait un estimateur de données incomplètes qui serait peu efficient. L'algorithme EM est applicable à des problèmes à valeurs manquantes, mais il se prête mal à une analyse poussée de données d'enquête, car un grand nombre d'espérances conditionnelles doivent être évaluées.

2. IMPUTATION MULTIPLE

Dans l'application d'une méthode IM, on génère un certain nombre d'achèvements plausibles (parallèles) de $\mathbf{X}$ par une opération aléatoire traduisant l'incertitude qui s'attache aux valeurs manquantes. Il y a quatre étapes :

1. ajustement d'un modèle pour les valeurs manquantes;
2. production (en simulation) de plusieurs jeux (M) de valeurs plausibles à partir du modèle;
3. analyse des ensembles de données achevés;
4. résumé des résultats des ensembles achevés.

Voici les hypothèses posées :

a. le modèle pour les valeurs manquantes est correctement spécifié;
b. les valeurs plausibles sont correctement produites;

c. l'analyse à données complètes comporte un estimateur efficient et sans biais $\hat{\theta}(\mathbf{X}^*)$ avec sa variance d'échantillonnage, $s^2(\mathbf{X}^*) = \text{var}\{\hat{\theta}(\mathbf{X}^*)\}$, estimée sans biais par $\hat{s}^2(\mathbf{X}^*)$;

d. la variance d'échantillonnage de l'estimateur $\hat{s}^2(\mathbf{X}^*)$ est inférieure en ordre de grandeur à s^4.

L'*imputation* sera juste en b si les sources d'incertitude au sujet des paramètres du modèle pour les valeurs manquantes et de ces mêmes valeurs manquantes compte tenu des paramètres du modèle se retrouvent comme facteurs dans l'opération de production de valeurs plausibles. Dans la plupart des spécifications, il y aura deux sources d'incertitude, l'une au sujet des paramètres du modèle posé pour la non-réponse et l'autre au sujet des résultats compte tenu des valeurs des covariables et des paramètres du modèle.

Désignons par $\mathbf{X}_m$ l'ensemble de données achevé par le m^{e} jeu de valeurs plausibles et par $\hat{\theta}_m = \hat{\theta}(\mathbf{X}_m)$ l'estimation fondée sur le m^{e} ensemble achevé. Si les conditions a à d posées sont remplies, l'estimateur IM

$$\tilde{\theta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m$$

est approximativement sans biais et sa variance d'échantillonnage se trouve estimée avec au plus un léger biais par l'équation :

$$\tilde{s}^2 = \frac{1}{M}\sum_{m=1}^{M}\hat{s}^2(\mathbf{X}_m) + \sum_{m=1}^{M}(\hat{\theta}_m - \tilde{\theta})^2$$

(Rubin, 1987).

Sauf pour a, les conditions posées sont naturelles et habituellement remplies, bien que la condition b prescrive la façon de produire les valeurs plausibles. Beaucoup disent par ailleurs que la condition a nuit à une application efficace de la méthode IM. Il serait difficile en particulier de justifier l'hypothèse des « valeurs manquantes au hasard » sur laquelle reposent la plupart de ces méthodes. Après tout, la méthode d'imputation simple n'est explicitement liée à aucune hypothèse semblable. En fait, l'argument est spécieux : une imputation simple équivaut à *M* imputations identiques en imputation multiple, ce qui implique l'existence d'un modèle sans variance interimputation, c'est-à-dire sans incertitude au sujet des valeurs manquantes. Dans la plupart des spécifications, ce choix d'un modèle d'imputation laisse très nettement à désirer et va à l'encontre d'une stratégie d'imputation simple tant que le modèle de valeurs manquantes n'est pas correctement spécifié. L'imputation simple n'enlève pas les problèmes de valeurs manquantes au hasard, ni le conditionnement à prévoir (modèle). Sa spécification est une tâche commune à l'imputation simple et à l'imputation multiple.

Pour combattre l'hésitation à recourir à une imputation multiple, nous proposons une stratégie d'amélioration de la méthode IS « par défaut » en formulant le modèle impliqué. Comme simple exemple général de l'application d'une telle stratégie, citons le remplacement de la méthode de « report de la dernière valeur » pour une variable dichotomique longitudinale par des tirages multiples d'un modèle où on suppose des probabilités (faibles) de variation entre deux points qui se succèdent dans le temps. On estime ces probabilités à partir des enregistrements complets, et chaque jeu de valeurs plausibles vient de tirages par probabilités plausibles. On tire une probabilité plausible de la distribution (estimée) de l'estimateur de cette probabilité. Dans la plupart des spécifications, la normalité de l'estimateur tient à une approximation acceptable. Dans la méthode d'imputation multiple, les jeux de valeurs plausibles (à imputer) diffèrent à deux égards, c'est-à-dire par les irrégularités d'une opération de Bernoulli et par les différences de probabilités entre les jeux (imputations). On peut résoudre le problème de valeurs manquantes au hasard par un modèle plus fin où on suppose que les probabilités diffèrent selon les groupes de sujets entre les points successifs dans le temps en fonction non seulement des résultats précédents, mais aussi des résultats antérieurs (si on les connaît), etc. Avec une variable multinomiale, on peut modéliser de la même manière la distribution conditionnelle du nouvel état (compte tenu de la variation depuis le point précédent dans le temps).

Que devrait être le niveau de détail du modèle que nous devrions formuler? (Jusqu'où devrions-nous pousser le conditionnement ou la stratification?) Rubin (1996) recommande un conditionnement qui soit le plus détaillé possible, mais nous ne devrions pas y voir un impératif. Un conditionnement moindre vaut bien mieux que l'absence de conditionnement. La peur de ne pouvoir répondre à un ultimatum ne devrait donc pas nous contraindre à la résignation. De toute manière, on en viendra à un point où le modèle de données manquantes sera si riche en paramètres que l'incertitude qui s'attache à ces valeurs deviendra tout simplement excessive. Il faut garder le souci que l'analyste soit à l'aise avec tous ces détails.

Dans les trois sections qui suivent, nous évoquerons trois applications de la méthode d'imputation multiple dans la perspective que nous venons d'exposer. Chacune est la version condensée d'une description d'étude de cas.

3. LA UK LABOUR FORCE SURVEY

La UK Labour Force Survey, l'enquête sur la population active du Royaume-Uni, est une enquête permanente portant sur les adresses résidentielles dans ce pays. Chaque adresse échantillonnée à un moment quelconque est conservée pendant un an. Il y a prise de contact peu après la sélection, ainsi qu'aux 3, 6, 9 et 12 mois. La principale variable de résultat observée est la situation d'emploi ou d'activité – selon la définition de l'Organisation internationale du travail (OIT) – avec ses quatre catégories : enfant (de 1 à 15 ans inclusivement), personne occupée, chômeur et personne économiquement inactive. L'analyse médiatisée qui émane de cette enquête livre les taux estimatifs de chômage de la population nationale en âge de travailler (16 à 64 ans pour les hommes et 16 à 59 ans pour les femmes).

L'unité fondamentale d'échantillonnage est l'adresse, mais les données sont recueillies auprès (et au sujet) des personnes. Ainsi, une adresse peut comporter une liste d'occupants semblables ou encore légèrement ou entièrement différents d'une prise de contact à l'autre. On dispose d'une courte période de deux semaines pour la prise de contact, car les délais de publication sont serrés. Il peut donc être impossible de déterminer s'il n'y a pas d'occupants à une adresse ou si l'intervieweur n'a pu prendre contact avec les occupants. Bien sûr, des sujets peuvent refuser toute collaboration (actuelle ou future) à l'enquête au nom d'une partie ou de la totalité des occupants à une adresse. On applique un protocole bien établi pour réduire la non-réponse à des questions. On accepte les réponses par procuration (des adultes), aussi est-il possible d'obtenir des données sur un sujet absent au moment de l'interview. Parfois, on ne pourra même pas compter sur un répondant par procuration.

Dans le cadre de cette opération, on impute les données de la situation d'emploi (OIT) par report de la dernière valeur si cet état a été déterminé trois mois auparavant. Notons en particulier une absence d'imputation de valeurs manquantes de situation d'emploi au premier contact planifié. La base de données est structurée par trimestres, et on y trouve les enregistrements de tous les sujets avec qui on a pris contact (directement ou par procuration) dans le trimestre. On relève les variables sociodémographiques de base pour tous les sujets de la base (sinon ceux-ci ne sont pas pris en compte). Dans la base d'un trimestre, on dénombre quelque 140 000 personnes, dont 80 000 sont en âge de travailler. Sur les 26 369 sujets de la base de mars-mai 2001 dont les adresses étaient en suivi d'enquête pour la cinquième fois, 1 418 (5,4 %) ont fait l'objet d'une imputation et 562 (2,1 %) ont été laissés sans imputation.

3.1 Du report de la dernière valeur à une imputation par donneur

Considérons une imputation de valeurs manquantes une situation d'emploi (OIT) de la population adulte. Le modèle impliqué de report de la dernière valeur est celui d'une transition entre les trois états (personne occupée, chômeur et personne économiquement inactive) avec une forte probabilité d'absence de changement. Un simple regard sur les données nous révèle que les changements de situation d'emploi sont bien plus fréquents (de 10 % à 20 % selon la définition des sous-populations) chez les jeunes adultes de 16 à 24 ans que dans la population d'âge moyen ou avancé. Le taux de non-réponse des jeunes adultes est bien supérieur au taux global et, dans la population âgée, ce taux devient très faible.

Dans une amélioration du modèle de report de la dernière valeur, la première étape consiste à estimer les probabilités de transition entre les trois états. Une amélioration évidente sera de stratifier par âge, car les probabilités varient

selon les groupes d'âge (16 à 24, 25 à 39 et 40 ans et plus). On donne un peu de complexité à l'analyse en ajoutant une stratification selon le sexe et l'état matrimonial (célibat ou non), mais ces facteurs ne sont pas aussi utiles que celui de la situation d'emploi (deux, trois ou quatre trimestres auparavant), si on le connaît. Comme le nombre de strates (catégories) ainsi introduites est peut-être excessif, on peut regrouper les états antérieurs en comptant les changements ou en classant les sujets dans des catégories « changement pendant l'enquête » et « absence de changement ». Par convention, un état inconnu à un point antérieur dans le temps est considéré comme un changement.

Une application pratique de cette méthode est l'imputation par donneur. Dans chaque strate (combinaison de catégories), nous reconnaissons des donneurs, c'est-à-dire des sujets présentant la combinaison en question et dont l'état actuel est connu, et des receveurs, c'est-à-dire des sujets présentant la même combinaison et dont l'état actuel est inconnu. Pour chaque receveur, nous choisissons un donneur au hasard et imputons son état.

L'imputation par donneur est aléatoire, puisqu'on impute une valeur d'état avec des probabilités correspondant à la représentation dans le groupe de donneurs. L'inconvénient avec cette méthode, c'est que nous nous reportons aux probabilités multinomiales *estimées* plutôt qu'aux probabilités sous-jacentes, ce à quoi on peut remédier en tirant les probabilités de la distribution estimée d'échantillonnage par approximation normale. On s'assure ainsi que les valeurs plausibles ont une variation appropriée entre imputations en dehors des différences par tirage aléatoire d'une distribution multinomiale. On trouvera dans Longford (2002a) d'autres renseignements généraux et particuliers sur le plan d'imputation et l'analyse (statistiques régionales).

On peut faire valoir que l'imputation de valeurs manquantes de situation d'emploi dans la base de données de l'enquête sur la population active sera moins prioritaire que l'imputation de non-réponse à des adresses. S'il n'y a pas de prise de contact à une adresse échantillonnée, celle-ci ne figure pas dans la base de données. Le taux de non-réponse à des adresses est de plus de 30 %. Il est facile de constater que les jeunes et les célibataires sont sous-représentés dans cette base. Comme leur taux de chômage est supérieur à celui de la population d'âge moyen ou avancé, la non-réponse introduit un biais dans plusieurs analyses importantes. Dans une imputation multiple de non-réponse à des adresses, celle-ci porterait sur des ménages entiers à chaque adresse sans prise de contact. On tiendrait compte de la possibilité qu'il n'y ait pas d'occupants, que les occupants soient les mêmes que trois mois auparavant ou qu'ils aient légèrement ou entièrement changé. La spécification d'un tel modèle est bien plus laborieuse, et c'est le prix à payer pour s'approcher de l'idéal de l'estimation efficiente (à l'aide de toutes les données disponibles) et d'une estimation honnête de la précision (estimation sans biais de la variance d'échantillonnage). Il s'agit notamment de distinguer les plans d'échantillonnage conçus et réalisés. Il s'agit aussi d'un effort de modélisation des principales insuffisances de la collecte de données, et plus particulièrement du facteur de non-réponse à des adresses.

4. LA NATIONAL SURVEY OF HEALTH AND DEVELOPMENT, ALCOHOL CONSUMPTION

La National Survey of Health and Development, l'enquête nationale sur la santé et le développement, qui est financée par le Medical Research Council (Royaume-Uni), est une enquête longitudinale portant sur les sujets nés dans ce pays dans la semaine du 3 au 9 mars 1946. On a échantillonné l'ensemble des naissances légitimes chez les femmes des travailleurs non-manuels ou agricoles et une naissance légitime sur quatre chez les femmes des travailleurs manuels. On a régulièrement suivi les sujets de la naissance à l'adolescence et à l'âge adulte. En 1989, la 19^{e} prise de contact a eu lieu. La 20^{e} reprise est récente. L'échantillon initial de 5 362 sujets a été ramené à 3 262 (61 %) au fil des ans par perte de contact, refus, émigration ou décès.

Dans une des questions, on a demandé aux sujets en 1989 de tenir un journal de toute leur consommation d'aliments et de boissons dans une semaine. Celui-ci a été rempli pour les deux premiers jours pendant la visite d'un professionnel de la santé, quoiqu'un certain nombre de sujets aient refusé de collaborer ou que l'interview ait pris fin pour d'autres raisons en cours de consignation ou même avant. L'étude que décrivent Longford et coll. (2000) porte sur la consommation d'alcool en général et la consommation excessive en particulier. Dans cette étude, le traitement des données incomplètes s'est limité à l'univers de 3 262 sujets collaborant partiellement ou entièrement à l'enquête. Il ne convient certes pas d'oublier les sujets perdus dans les reprises antérieures de l'enquête, bien que ceux qui ont

été perdus par décès ou émigration puissent être considérés comme n'appartenant plus à la population d'intérêt, mais il conviendrait encore moins de limiter l'analyse aux 2 002 sujets à journal complet.

Les habitudes de consommation d'alcool sont variables : on peut ne jamais consommer, consommer occasionnellement en petite ou en grande quantité, consommer régulièrement en quantité variable, etc. Il est donc possible d'« apprendre » par une courte période de tenue d'un journal ce que peut être la consommation les autres jours. On dispose en outre d'une information auxiliaire solide et assez abondante sur des variables qui sont en corrélation étroite avec la consommation d'alcool. Les variables du sexe et de la masse corporelle sont un choix qui s'impose d'emblée, mais on peut avoir intérêt à rechercher des données auxiliaires par-delà les variables qui seraient les covariables de la spécification classique d'une régression. Dans l'enquête nationale sur la santé et le développement de 1989, on a demandé aux sujets d'indiquer combien de quatre types de boissons alcooliques (bière, vin, sherry et liqueur) ils avaient consommé la semaine précédente.

On répond à de telles questions en très peu de temps après un bref exercice de remémoration. Les réponses obtenues sont donc moins sûres, et on sait par d'amples indices qu'il y a sous-déclaration. Il reste que la non-réponse est un phénomène rare. On sait par ailleurs que les données de journal sont bien plus fiables si un tel relevé est complet. On peut donc dire que, si les données de remémoration ne sauraient très bien remplacer les données manquantes de journal (surtout s'il s'agit de journaux remplis mais incomplets), elles gardent leur utilité comme variables auxiliaires, c'est-à-dire qu'elles ont valeur d'*information*. La façon pratique d'exploiter cette association est de formuler un modèle pluridimensionnel des quantités en remémoration pour les sept jours de tenue du journal et de tirer des valeurs plausibles pour les jours « manquants » de la distribution conditionnelle plausible des valeurs manquantes compte tenu de la partie remplie du journal et des quantités remémorées. On peut introduire une régression dans ce modèle à variables multiples par une stratification relative à d'autres variables d'intérêt comme la masse corporelle et le tabagisme, ainsi qu'aux quatre questions du questionnaire CAGE sur les problèmes d'alcool.

Dans cette étude, la production de valeurs plausibles s'est faite par étapes. Premièrement, on a produit de telles valeurs pour la taille et la masse corporelle. Deuxièmement, on en a produit pour les quantités en remémoration dans un conditionnement par la masse corporelle, les questions CAGE et le tabagisme. Comme la distribution de ces quantités peut être bien approchée par une distribution lognormale à zéros valorisés, on a produit séparément le signe (nul ou positif) de la consommation et la valeur logarithmique (omise en cas de signe nul) à l'aide de modèles distincts à quatre variables (pour les quatre types de boissons).

Troisièmement, on a produit des signes et des valeurs logarithmiques plausibles pour la consommation d'alcool de chaque jour « manqué ». On trouvera les détails dans Longford et coll. (2000). Rubin (1996) et Schafer (1997) font valoir qu'il n'est pas essentiel que les données employées comme résultats d'un modèle de valeurs manquantes soient en distribution normale. Si on suit leur avis, on peut simplifier quelque peu l'opération de production de valeurs plausibles.

Le but de l'étude est d'estimer le pourcentage de Britanniques d'âge moyen qui font une consommation excessive d'alcool. Dans l'analyse, nous devons oublier la variation de cette consommation de semaine en semaine, tout comme les différences systématiques entre le groupe de 43 ans et les cohortes d'âge voisines. Selon la façon de spécifier la consommation excessive (quantités qui diffèrent selon le sexe ou quantités correspondant à diverses valeurs de masse corporelle), un relevé incomplet peut nous livrer le résultat d'intérêt en toute certitude : si un sujet a consommé plus que son « quota » hebdomadaire d'alcool les deux premiers jours, la consommation des autres jours de la semaine importe peu.

Les méthodes courantes d'imputation simple (report de la dernière valeur, imputation par moyenne ou par zéro pour chaque valeur manquante, etc.) appliquées dans des spécifications semblables ne sont pas appropriées et leurs insuffisances se remarquent tout particulièrement s'il s'agit d'estimer des probabilités en queue de distribution comme dans le cas présent. On trouvera les détails dans Ely (2003).

Pour une imputation à l'égard des sujets perdus dans des reprises antérieures, il y a la difficulté logistique considérable de trier une information abondante (et incomplète) qui peut ne pas être informatisée ou qui peut se présenter sous une forme électronique se prêtant mal à un traitement. Il faut exclure les sujets perdus par décès ou émigration qui n'appartiennent plus à la population d'intérêt. À noter que, si un sujet revient d'émigration au

Royaume-Uni, il n'est pas réintégré à l'enquête. On ignore d'habitude la situation (sujets habitant au Royaume-Uni, décédés ou émigrés) des gens avec qui on a perdu le contact. Le traitement de cette non-réponse par « données ne manquant pas au hasard » pose un problème épineux et, à cet égard, tant un conditionnement poussé par l'information des reprises antérieures qu'une analyse de sensibilité sont essentiels.

5. LA SCOTTISH HOUSE CONDITION SURVEY

La Scottish House Condition Survey, l'enquête sur l'état des logements en Écosse, porte sur le parc d'habitations écossais. Elle a eu lieu en 1991, 1996 et 2002. En 1996, on a eu recours à un plan d'échantillonnage par grappes stratifiées, avec des corrections (valorisations) garantissant des tailles suffisantes de sous-échantillon dans certaines régions. Dans cette enquête, on fait appel à des intervieweurs professionnels pour l'évaluation des logements échantillonnés. Celle-ci comporte un grand nombre d'éléments d'information : type de logement, chauffage central, degré de délabrement de diverses parties ou caractéristiques du logement, etc. On note ces dernières variables sur une échelle de 11 points (0 à 10) qui s'interprètent comme des proportions de 0, 10, …, 100 % du coût de remplacement devant permettre d'aligner les parties ou les caractéristiques en question sur les normes établies. On convertit ces valeurs en coût total pour le logement (*coût de réparation visible*) et en coût de remise en état pour 10 ans (*coût de réparation dans l'ensemble*). On se fonde à cette fin sur des tableaux et des formules détaillés de conversion qui tiennent compte de la taille, de la nature et du lieu d'un logement, des économies d'échelle (lorsque les réparations à prévoir sont nombreuses), etc.

Bien que d'une grande qualité, les évaluations des enquêteurs sont imparfaites. Comme moyen de contrôle de qualité, on observe une seconde fois un sous-échantillon « non informatif » de logements et compare les données des deux observations. En cas de différences entre les évaluations d'un même logement, on peut s'interroger sur la précision estimée des principaux estimateurs produits, puisqu'on évalue en supposant que les observations d'enquête sont parfaites. L'importance relative des différences est difficile à chiffrer, car les divergences sont plus fréquentes et/ou généralement supérieures pour certains éléments d'information et l'incidence sur le coût est plus appréciable dans d'autres cas.

En 1996, l'échantillon de cette enquête était d'environ 16 000 logements, dont 575 ont chacun fait l'objet d'une double observation.

5.1 Erreurs d'attribution et imputation multiple

Pour chaque élément d'information et chaque logement, on définit l'évaluation *idéale*, c'est-à-dire ce qu'elle aurait été si elle avait été faite par un enquêteur parfait. Ce sont là les données manquantes. Les évaluations réalisées sont des données auxiliaires de grande qualité sur les évaluations idéales. Par les paires d'évaluations, on peut définir un modèle approprié dont on tirera des valeurs (évaluations) plausibles.

À une échelle ordinale, on délimite le voisinage d'un résultat d'évaluation d'une manière naturelle. Ainsi, pour les résultats 0 à 10, le voisinage de $0 < k < 10$ comprend k–1 et k+1. On s'attache à deux types d'imperfections des évaluations des enquêteurs :

- divergence et attribution consécutive de l'élément considéré à une catégorie voisine de la catégorie idéale;
- erreur grossière et attribution consécutive de l'élément considéré à une catégorie arbitraire.

On désigne respectivement par p_d et p_g les probabilités propres aux éléments qui ont à voir avec ces deux types d'imperfections. L'information antérieure semble indiquer que p_g est bien moindre que p_d et que cette dernière probabilité est au plus de quelques points en pourcentage. L'évaluation d'un élément peut être entachée tant d'une divergence que d'une erreur grossière. Si une catégorie k a L_k voisins, la probabilité d'attribution d'un logement à la catégorie idéale est de :

$$\mathrm{P}(X = X^* \mid X^* = k) = 1 - L_k p_k - (K-1) p_g + (K-1) p_d p_g \,,$$

où X désigne l'évaluation réalisée et X^*, l'évaluation idéale. On estime séparément les probabilités p_d et p_g pour chaque élément contribuant au coût évalué par moments appariés (comme racine d'une équation quadratique) et les variances d'échantillonnage correspondantes en développement de Taylor. On trouvera les détails dans Longford (2002b). Pour produire un ensemble de données idéales plausibles, il faut une distribution conditionnelle plausible du résultat idéal compte tenu du résultat attribué, $\mathrm{P}(X^* \mid X)$, ce qu'on obtient par le théorème de Bayes par paire plausible de probabilités $(\tilde{p}_d, \tilde{p}_g)$. On tire un ensemble de résultats (idéaux) plausibles par une simulation des divergences et des erreurs grossières par $\tilde{p}_d$ et $\tilde{p}_g$ respectivement. Les simulations sont indépendantes pour chacune des 50 variables et plus des calculs de coûts. On évalue les coûts plausibles à partir de chaque ensemble de résultats plausibles (et de quelques autres variables). Les quantités d'intérêt premier pour la population sont les totaux de grands sous-domaines comme le coût total de réparation dans l'ensemble pour chaque type de logements (maisons individuelles ou jumelées, immeubles d'appartements, etc.) et pour les petites régions (autorités administratives locales). Dans certaines analyses, la variable de résultat (coût) est à l'échelle initiale et, dans d'autres, à l'échelle logarithmique. S'il n'y avait pas d'imputation multiple, il faudrait concevoir des méthodes différentes pour les deux types d'analyse. Avec l'imputation multiple, les valeurs plausibles produites par le programme d'élaboration de données sont d'une application universelle, c'est-à-dire qu'elles conviennent à *toute* analyse.

La variance interimputation nous renseigne sur l'extension de la variance d'échantillonnage qui est imputable à l'imperfection des évaluations. Bien que M=5 ensembles de valeurs plausibles puissent suffire à l'évaluation de l'estimateur IM, il est recommandé d'en prendre plus lorsqu'il s'agit de planifier une future enquête.

On peut examiner les apports relatifs des éléments à la variance interimputation par la méthode suivante : on remplace les valeurs plausibles de l'élément visé par les évaluations réalisées et on procède à l'estimation IM en supposant que les évaluations de cet élément sont parfaites; on compare ensuite la variance estimée d'échantillonnage ou sa composante « entre imputations » à la variance correspondante de l'estimateur initial IM; de même, on peut appliquer les valeurs plausibles pour un seul élément.

Les M ensembles de données plausibles sont appréciables, mais bien moindres que l'ensemble initial comprenant de nombreuses variables dont la plupart n'entrent pas dans la formule de coût.

6. QUESTIONS MARGINALES

Il n'y a pas si longtemps, on aurait été fondé de se soucier du stockage des données et de l'ampleur des calculs nécessaires à l'exécution des deux premières étapes de l'application de la méthode d'imputation multiple. De nos jours, les puissances de calcul et les capacités de mémoire sont si grandes et coûtent si peu que de tels problèmes ne se posent plus sauf pour les bases de données les plus considérables. Dans une imputation multiple, on a besoin d'un surcroît de mise en mémoire pour M ensembles de valeurs plausibles, peut-être en classement par sujet et variable. Ainsi, pour M=5 et une proportion de 14 % d'éléments d'information manquants, les M ensembles de données supplémentaires contiennent autant d'éléments que la base de données initiale (et incomplète). On tiendra bien plus compte dans l'organisation de la base de données des besoins d'un utilisateur secondaire insuffisamment pourvu en logiciels si on lui fournit les M ensembles achevés. Si le client désire des imputations pour quelques variables seulement, on peut ajouter comme variables leurs valeurs plausibles à la base de données. Un autre moyen commode de procurer des valeurs plausibles à des utilisateurs secondaires est un programme qui produirait un jeu de valeurs plausibles et compléterait la base de données en effectuant les imputations.

Pour appliquer M fois une méthode (à données complètes), il faut bien moins de ressources que le multiple M de ressources à prévoir pour une analyse sans itération, car la dépense principale est en conception, élaboration et mise au point d'un programme. En fait, le grand avantage de l'imputation multiple est que les utilisateurs secondaires n'ont besoin de logiciels et de compétences que pour procéder à l'analyse à données complètes. Le travail même de calcul ou le temps réel passé à M–1 analyses supplémentaires n'est pas un facteur d'importance.

Le nombre d'ensembles de valeurs plausibles influe non seulement sur la quantité de calculs, mais aussi sur la variance d'échantillonnage de l'estimateur IM. Cette variance est $W + B(1+1/M)$, où W est la variance intraimputation, c'est-à-dire l'estimation de la variance d'échantillonnage à données complètes, et où B est la variance interimputation, c'est-à-dire la quantité d'information perdue par inachèvement. Le reste de la perte, B/M, s'explique par le nombre limité de M imputations. Dans la plupart des enquêtes, M=5 suffit, mais le fait de ne pas en avoir plus ne se justifie pleinement que pour les immenses bases de données où abondent les données manquantes.

7. CONCLUSION

La méthode d'imputation multiple est un remède pratique à l'information incomplète d'enquêtes à grande échelle. Idéalement, elle se distingue par son efficience et son intégrité. Elle exploite toutes les données d'enregistrements incomplets; elle évalue la précision des estimateurs et n'introduit guère de biais. Si on ne se trouve pas dans des circonstances idéales où on disposerait d'un juste modèle de traitement de la non-réponse, on ne s'en tire pas aussi bien, mais avec les outils types de modélisation, on a plus de chances d'approcher de cet idéal. Avec toute méthode d'imputation simple, le chemin est bien plus long devant soi, parce qu'on peut l'améliorer en passant à une imputation multiple, c'est-à-dire en tenant compte des sources d'incertitude au sujet des valeurs imputées dans la production des estimations et en reprenant l'opération un certain nombre de fois. Dans certains cas particuliers, il est possible d'estimer directement et d'exprimer (ou approcher) analytiquement la variance interimputation. La chose est faisable lorsqu'on n'a que quelques analyses à effectuer. Si on se trouve à appliquer une grande diversité de méthodes à données complètes, l'imputation multiple se révèle une méthode supérieure, car on fait faire le travail à l'ordinateur.

L'étendue des problèmes auxquels peut s'appliquer la méthode d'imputation multiple est toute celle du traitement des données manquantes, comme l'évoquent Dempster, Laird et Rubin (1977). Dans cette imputation, il y a une perte résiduelle d'efficience parce que le nombre d'itérations est limité, mais l'analyste secondaire n'a pas à posséder de compétences en traitement de données manquantes. Par ailleurs, l'algorithme EM espérance-maximisation est efficient et ne demande rien d'autre que des conditions moyennes de régularité, mais il est spécifique aux analyses et bien plus difficile à appliquer et cette application peut poser des problèmes de convergence.

RÉFÉRENCES

Dempster, A. P., Laird, N. M., et Rubin, D. B. (1977), "Maximum likelihood for incomplete data via the EM algorithm", *Journal of the Royal Statistical Society*, Series B, 39, pp. 1–38.

Ely, M. (2003), "Comparison of methods for dealing with missing values in the National Survey of Development and Health", PhD. thesis. In preparation.

Longford, N. T. (2002a), "Missing data and small area estimation in the UK Labour Force Survey." Submitted.

Longford, N. T. (2002b), "Surveyor inconsistency in the Scottish House Condition Survey", unpublished manuscript.

Longford, N. T., Ely, M., Hardy, R., et Wadsworth, M. E. J. (2000), "Handling missing data in diaries of alcohol consumption", *Journal of the Royal Statistical Society*, Series A, 163, pp. 381–402.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

Rubin, D. B. (1996), "Multiple imputation after 18+ years", *Journal of the American Statistical Association*, 91, pp. 473–489.

Schafer, J. L. (1997), "Analysis of Incomplete Multivariate Data", New York: Chapman and Hall.