N° 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2003 : Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

2003

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2003 de Statistique Canada
Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

APPLICATION DE NOUVELLES TECHNIQUES STATISTIQUES

Richard Valliant[1]

RÉSUMÉ

Nombreux sont les domaines où d'importants progrès d'ordre théorique ont eu lieu dans la dernière décennie. Pour ne citer que quelques exemples, mentionnons l'analyse de données en imputation, l'estimation de variance dans un échantillonnage à plusieurs degrés, la protection de la confidentialité, l'extraction de données et les estimations pour petits domaines. Internet a été source de possibilités nouvelles en matière de diffusion de l'information. Dans la prochaine décennie, nous aurons notamment comme grand défi d'ordre pratique la mise en application de techniques dont nous savons déjà qu'elles sont supérieures à celles qui sont actuellement employées. Nous nous intéresserons entre autres dans le présent document à l'estimation de variance en présence d'imputation ainsi qu'à son application à l'estimation d'indices de prix, à l'échantillonnage à plusieurs degrés et au recours au graphisme dans les publications.

MOTS CLÉS : graphisme; imputation; indices de prix; logiciels.

1. INTRODUCTION

La généralisation de l'usage est habituellement en retard sur l'innovation en procédés. Les praticiens trouvent des solutions adaptées aux problèmes que posent des applications particulières. Les théoriciens peuvent ensuite confirmer ou infirmer la justesse de ces solutions. Enfin, les méthodes s'incarnent dans des progiciels pour devenir accessibles à tous. La décennie 1990 a connu de grands progrès dans bien des secteurs et, parfois, une insistance sur un certain nombre de problèmes. On a ainsi assisté à une flambée d'activités en ce qui concerne l'imputation de valeurs manquantes. À la section 2, nous examinerons certaines des possibilités d'analyse d'ensembles de données contenant des données imputées avec les logiciels et les techniques statistiques appropriés. À la section 3, nous lierons l'imputation aux corrections de qualité dont font l'objet les indices de prix. À la section 4, nous donnerons quelques exemples de graphisme Web exploité par un organisme public. Comme nous le noterons en conclusion, il y a bien d'autres secteurs qui demeureront des domaines d'investigation dynamiques dans un proche avenir.

2. ESTIMATION DE VARIANCE SUITE À L'AJUSTEMENT DE NON-RÉPONSE ET EN PRÉSENCE D'IMPUTATION

Dans les enquêtes, on emploie couramment des méthodes de traitement des données manquantes. Pour la non-réponse complète, la repondération est habituellement employée. En cas de non-réponse à certaines questions, on recourt fréquemment à l'imputation, du moins pour les questions les plus importantes d'une enquête. Nous pourrions un jour procéder de façon routinière à l'imputation de toutes les réponses manquantes à des questions dans ce qui serait autrement une réponse complète et, de ce fait, l'imputation sera autant monnaie courante que la repondération. S'il est fréquent d'appliquer des méthodes d'amélioration d'estimations ponctuelles, on a moins l'habitude d'employer des techniques d'estimation de variance qui tiennent bien compte de ces méthodes. En principe, des méthodes itératives comme la technique jackknife ou celle de la répétition compensée (« balanced repeated replication » ou BRR) tiennent compte de la repondération dans la mesure où l'estimateur obtenu est une fonction lisse de sommes pondérées. La méthode jackknife corrigée et l'imputation multiple peuvent tenir compte de l'incidence des données imputées sur la variance. Certains logiciels dont nous ferons mention sont déjà disponibles

[1] Richard Valliant, Université du Michigan, et Joint Program for Survey Methodology, Université du Maryland, 1218 Lefrak Hall, College Park MD 20742, rvalliant@survey.umd.edu.

pour l'application de ces méthodes spécialisées, mais il peut être difficile de les intégrer aux systèmes de traitement des enquêtes. Voilà pourquoi on optera pour la voie de la facilité dans certaines enquêtes, c'est-à-dire qu'on ne fera rien pour tenir compte de la repondération ni de l'imputation. Dans cette section, nous passerons en revue certaines des possibilités susceptibles d'être exploitées.

2.1 Ajustements aux poids

La repondération en fonction de la non-réponse complète peut être le moyen le plus simple de compenser l'absence de certaines données. De multiples ajustements aux poids sont habituellement employés dans les enquêtes pour tenir compte des unités inadmissibles (UI) d'une base de sondage, de la non-réponse d'un certain nombre d'unités admissibles (NRUA) et de l'utilisation de données auxiliaires à des fins d'estimation. Il se pose la question pratique de savoir comment se comportent les estimateurs de variance courants lorsque la pondération se fait en plusieurs étapes. En guise d'illustration, considérons un échantillonnage stratifié à un degré. Après ajustements aux poids, un total estimé se présente sous la forme

$$\hat{T} = \sum_{(hi)\in s_{ER}\cup s_{IN}} w_{hi}^{*} Y_{hi}$$

où *h* désigne une strate, *i* une unité d'une strate et Y_{hi}, une valeur liée à l'unité *hi* de l'échantillon et où

- w_{hi}^{*} : poids final tenant compte de l'admissibilité inconnue, de la non-réponse et de l'utilisation de données auxiliaires;
- s_{ER} : ensemble de répondants admissibles de l'échantillon;
- s_{IN} : ensemble des unités de l'échantillon que l'on sait inadmissibles.

Les valeurs *Y* des unités inadmissibles reçoivent habituellement le code zéro tant pour l'estimation ponctuelle que pour l'estimation de variance. Les estimateurs de variance linéarisés et les estimateurs de même famille qui utilisent le carré des résidus permettent normalement de tenir compte au mieux de la dernière étape de l'estimation qui consiste à intégrer des données auxiliaires par la poststratification, l'estimation par régression ou par d'autres méthodes semblables. Dans certains progiciels du marché, on tient uniquement compte de caractéristiques du plan d'échantillonnage comme la stratification et la formation de grappes. Il y a plusieurs estimateurs de variance qui, comme variantes de l'estimateur linéarisé, utilisent le carré des résidus définis par $r_{hi} = Y_{hi} - \hat{Y}_{hi}$. Le terme $\hat{Y}_{hi}$ est une valeur prédite fondée sur le modèle implicite qui est sous-jacent à un estimateur. L'estimateur de régression général, $\hat{Y}_{hi} = \mathbf{x}'_{hi}\hat{\mathbf{B}}$ par exemple, où $\mathbf{x}_{hi}$ est un vecteur de variables auxiliaires, $\hat{\mathbf{B}} = \mathbf{A}^{-1}\sum_{(hi)\in s_{ER\cup IN}} w_{2hi}\mathbf{x}_{hi}\, Y_{hi}/v_{hi}$, avec $\mathbf{A} = \sum_{(hi)\in(s_{ER}\cup s_{IN})} w_{2hi}\mathbf{x}_{hi}\,\mathbf{x}'_{hi}/v_{hi}$, est un estimateur de pente basé sur le plan d'échantillonnage selon le modèle $Y_{hi} = \mathbf{x}'_{hi}\boldsymbol{\beta} + \varepsilon_{hi}v_{hi}$, $\varepsilon_{hi} \sim \left(0, \sigma_{hi}^2\right)$. Le poids w_{2hi} est le poids de base, ajusté en fonction de l'admissibilité inconnue et de la non-réponse (mais non pour la poststratification ni pour toute autre utilisation de données auxiliaires en dernière étape). Dans une étude de Valliant (2003), l'estimateur avec résidus au carré le plus performant était celui qui repondérait en moindre résidus:

$$v_J^{*} = \sum_h \left(1 - f_{h,ER\cup IN}\right) \sum_{s_{h,ER\cup IN}} \left(\frac{w_{hi}^{*} r_{hi}}{1-\Delta_{hi}} - \frac{1}{n_{h,ER\cup IN}} \sum_{s_{h,ER\cup IN}} \frac{w_{hi}^{*} r_{hi}}{1-\Delta_{hi}} \right)^2, \qquad (1)$$

où

$n_{h,ER\cup IN}$ = taille de l'échantillon dans la strate *h* de répondants admissibles et d'unités inadmissibles connues et où

$\Delta_{hi} = w_{2hi}\mathbf{x}'_{hi}\mathbf{A}^{-1}\mathbf{x}_{hi}/v_{hi}$.

L'estimateur v_J^* est très semblable à celui du jackknife. Si Δ_{hi} est fixé à zéro, v_J^* est l'estimateur jackknife linéarisé v_{JL} que décrivent Yung et Rao (1996). Le cas particulier de v_{JL} approprié pour la poststratification est celui que présente SUDAAN (Shah et coll., 1996) à l'aide des options POSTWGT et POSTVAR disponibles pour quelques-unes des procédures, mais pas pour toutes. Si on fixe non seulement $\Delta_{hi} = 0$ mais aussi le poids de v_J^* à w_{2hi} plutôt qu'à w_{hi}^*, l'estimateur de variance se ramène à ce qu'on a l'habitude d'appeler l'estimateur linéarisé, v_L. Si des données auxiliaires servent à l'estimation, ce n'est pas un choix que l'on peut exploiter d'emblée dans les progiciels, puisqu'il s'agit d'un poids qui diffère du poids final. Dans un cas de repondération avec plusieurs ajustements comme nous venons de décrire, Valliant (2003) a constaté par simulation que v_L et v_{JL} étaient l'un et l'autre des sous-estimations qui menaient à des intervalles de confiance recouvrant moins que le taux nominal.

Dans certaines applications utilisant des estimateurs de variance avec répliques, on essaie de tenir expressément compte de toutes les étapes de l'estimation en reprenant chaque ajustement séparément pour chaque réplique de l'échantillon. L'estimateur avec répliques jackknife donne généralement des intervalles de confiance qui recouvrent strictement ou bien largement le taux nominal, et ce, parfois, au prix d'une surestimation considérable des erreurs quadratiques moyennes empiriques. L'estimateur de variance en moindres résidus en (1), qui est lié à l'estimateur jackknife, se caractérise par un léger biais positif et un recouvrement presque strict du taux nominal.

La théorie de l'échantillonnage à plusieurs degrés est aussi applicable aux cas de non-réponse. Si l'ensemble d'unités admissibles peut être considéré comme un sous-échantillon aléatoire (peut-être à l'intérieur de groupes) de la base de sondage et que les unités répondantes forment un échantillon aléatoire de l'ensemble des unités admissibles de l'échantillon, alors, l'obtention de l'échantillon de répondants admissibles peut être considéré comme un échantillon à trois degrés. Les estimations de variance tirées de cette hypothèse ou d'autres semblables sont étudiées par Fuller (1998, 2003), Särndal, Swensson, Wretman (1992), Rao et Sitter (1995) et Kim et Sitter (2003). Dans un grand nombre d'échantillons, on procède aussi directement par étapes multiples en stratifiant l'échantillon du premier degré pour accroître le rendement de l'échantillon final.

Il faut des formules spécialisées de variance pour tenir compte de cette pluralité d'étapes. Si on stratifie en deuxième étape l'échantillon du premier degré, le résultat net peut être une réduction de la variance par rapport à l'absence de stratification, mais si les deuxième et troisième étapes sont dues à la non-réponse, la variance sera plus grande que pour un échantillon de la même taille dans un échantillonnage à un degré. Lundström et Särndal (1999) font intervenir des notions de l'échantillonnage à deux phases pour estimer la variance de l'estimateur de régression généralisé dans le cas de la non-réponse.

Il faut également modifier les estimateurs de variance avec répliques pour des échantillons à plusieurs degrés. Fuller (1998) a conçu une méthode de correction de l'estimateur jackknife pour tenir compte de l'échantillonnage à plusieurs degrés. Il ne suffit pas de recalculer les estimations en fonction de l'admissibilité inconnue et de la non-réponse de chaque réplique pour obtenir de justes estimations de variance. Opsomer et coll. (2003) ont appliqué cette méthode à des estimations de régression de l'utilisation des sols dans l'enquête « National Resources Inventory ». Kim et Sitter (2003) ont présenté une méthode de réduction du nombre de répliques jackknife nécessaires pour des échantillons à deux degrés.

Dans aucun des progiciels plus répandus de traitement de données d'enquête, on ne trouve de procédures qui produisent directement des estimations de variance en échantillonnage à plusieurs degrés. SAS[MD], Stata[MD] et SUDAAN[MD] appelable en SAS comprennent effectivement des caractéristiques de programmation permettant à l'utilisateur d'écrire son propre code en faisant usage autant que possible de fonctions intégrées. À l'aide de WesVar[MD], on peut créer des poids de répliques qui tiendront compte de l'échantillonnage à deux degrés dans la mesure où l'estimateur recourt uniquement à des poids de base, ajustés pour la non-réponse. Des estimations plus complexes comme GREG ne peuvent être traitées de la même manière dans ce logiciel.

2.2 Imputation de certaines questions

Le recours à l'imputation pour remplacer des données manquantes à certaines questions est monnaie courante dans bien des enquêtes. Ceci facilite l'analyse de microdonnées. L'imputation permet d'estimer des agrégats de

population comme les moyennes ou les totaux sans faire d'ajustements aux poids qui aurait été différents pour chaque variable. Souvent, on soumet à l'imputation un sous-ensemble de questions offrant un intérêt analytique particulier. L'imputation vise généralement à estimer des paramètres de population plutôt que des valeurs manquantes particulières. Un autre but est la conservation des relations univariées et multivariées.

L'imputation a pour avantage pratique de contourner certaines limites des logiciels par la création d'ensembles de données rectangulaires. Une autre possibilité est la suppression de cas, mais on peut ainsi se trouver à écarter un grand nombre de cas. Dans une suppression de cas ou dans une analyse des cas complets, le nombre d'éléments écartés peut s'accroître considérablement au gré du passage d'une analyse à une variable à une analyse à plusieurs variables.

Si les valeurs imputées sont traitées comme des valeurs réelles et qu'on emploie des méthodes types d'estimation de variance, la variance obtenue sera généralement trop petite. Le tableau 1 livre certains résultats de simulation pour illustrer l'éventuelle gravité de ce problème. Dans le cadre d'une étude réalisée par le National Center for Education Statistics, j'ai pu comparer avec certains collègues de Westat un certain nombre de méthodes d'estimation de variance pour des ensembles de données avec imputation (Jones et coll., 2003). Cette étude est trop vaste pour que je puisse la décrire en détail ici, mais j'évoquerai les effets de l'imputation en mentionnant quelques résultats sommaires. Sur une population approximative de 12 000 districts scolaires, nous avons prélevé des échantillons stratifiés à un degré comportant un millier d'unités. Pour une variable, en l'occurrence le nombre d'administrateurs en équivalence « plein temps », nous avons attribué au hasard des valeurs manquantes en utilisant des taux différents de non-réponse pour les 12 strates du plan d'échantillonnage. Nous avons créé plusieurs combinaisons de taux de non-réponse, et le pourcentage moyen de non-réponse s'établissait à 5, 10, 20, 30 ou 50. Nous avons procédé à une imputation par donneur « hot-deck » pondérée avec des cellules « hot-deck » définies par les strates ou un autre ensemble de cellules interstrates. Le tableau 1 indique uniquement les résultats des cellules définies par les strates. Nous nous sommes ensuite reportés à l'estimateur π pour estimer le nombre total d'administrateurs dans la population. Nous avons fait deux choix d'estimateur de variance jackknife avec 120 groupes, c'est-à-dire dans une estimation où les valeurs imputées étaient traitées comme des valeurs réelles et dans une estimation jackknife de Rao-Shao (Rao et Shao, 1992).

Il est faux d'assimiler les valeurs imputées aux valeurs réelles et cette fausseté s'accroît progressivement à mesure qu'augmente le pourcentage de cas d'imputation. Ce problème est bien connu aujourd'hui. Rao et Shao (1992) en ont fait la démonstration théorique et le tableau 1 illustre le phénomène pour la simulation. Si les groupes de création de valeurs manquantes sont les mêmes que les groupes d'imputation par donneur, le rapport entre la simple estimation jackknife et l'erreur quadratique moyenne tombe de 0,75 pour un taux d'imputation de 5 % à 0,23 pour un taux d'imputation de 50 %. Le recouvrement des intervalles de confiance (IC) à 95 % varie de 90,9 % à 62,2 %. L'estimation jackknife de Rao-Shao est plus exempte de biais, mais demeure trop petite en moyenne pour un recouvrement d'intervalle de confiance qui varie de 91,3 % à 94,5 %.

Tableau 1. Résultats de simulation pour l'estimateur jackknife groupé avec valeurs par donneur assimilées à des valeurs réelles et pour la méthode jackknife de Rao-Shao; 1 000 échantillons stratifiés de taille *n* = 1 020 prélevés sur une population de 11 941 districts scolaires; nombre total d'administrateurs calculé à l'aide de l'estimateur π

Pourcentage d'imputation	Simple méthode jackknife à assimilation des valeurs imputées aux valeurs réelles		Méthode jackknife de Rao-Shao	
Probabilités moyennes de réponse	$v_J \big/ mse\left(\hat{T}\right)$	Recouvrement des intervalles de confiance à 95 %	$v_{J-RS} \big/ mse\left(\hat{T}\right)$	Recouvrement des intervalles de confiance à 95 %
5	0,75	90,9	0,88	92,8
10	0,70	88,3	0,94	92,3
20	0,59	85,2	0,95	93,6
30	0,51	82,2	1,00	94,5
50	0,23	62,2	0,94	91,3

Shao (2001) est une description limpide du mode d'application de la méthode jackknife de Rao-Shao à différentes situations d'imputation. Ce document traite de l'imputation par régression déterministe, dont les imputations par la moyenne et par le quotient constituent des cas particuliers. Une autre catégorie est celle de l'imputation par régression probabiliste, dont l'imputation par donneur « hot deck » pondérée est un cas particulier. Shao (2001, p. 146) étend l'application de la méthode jackknife corrigée de Rao-Shao aux situations générales d'imputation de la manière suivante. Soit *I* désignant une méthode d'imputation particulière, $\tilde{y}_i$ la valeur imputée pour le non-répondant *i* et $\tilde{y}_i^{(r)}$ la valeur imputée de l'unité *i* en fonction des répondants de la r^{e} réplique. La valeur jackknife corrigée de $\tilde{y}_i$ dans la réplique *r* est

$$\tilde{y}_{ir}^{adj} = \tilde{y}_i + E_I\left(\tilde{y}_i^{(r)}\right) - E_I\left(\tilde{y}_i\right), \tag{2}$$

où E_I est l'espérance d'imputation. Les valeurs des répondants sont inchangées. Ainsi, si *k* désigne une classe d'imputation et R_k, l'ensemble des répondants de la classe *k* et qu'on recourt à l'imputation par donneur hot deck pondérée, alors $E_I\left(\tilde{y}_i^{(r)}\right) = \sum_{i\in R_k} w_i^{(r)} y_i \Big/ \sum_{i\in R_k} w_i^{(r)}$, c'est-à-dire la moyenne pondérée pour les répondants de la classe *k* et la réplique *r*. $E_I\left(\tilde{y}_i\right)$ est la moyenne pondérée des répondants pour tout l'échantillon dans la classe *k*. Cette méthode s'applique aux estimateurs π et aux fonctions lisses et non linéaires de ces estimateurs. L'expression (2) est une formule générale qui peut se programmer pour les cas particuliers d'imputation déterministe par moyennes, quotients ou régression, ainsi que pour l'imputation par régression probabiliste. Voilà une autre méthode utile que n'offrent pas les logiciels du marché.

L'imputation multiple est une méthode qui est plus disponible dans les logiciels. Rubin (1996) et les observations qui l'accompagnent passent en revue les éléments de cette méthode, ses applications et certains de ses avantages et de ses inconvénients. L'avantage est la généralité, puisqu'un mode valide d'imputation multiple s'applique à l'estimation de statistiques descriptives simples et de statistiques analytiques bien plus complexes. Pour résumer cette méthode, posons que l'unité de l'échantillon *i* comporte *M* imputations – $y_{i1}, y_{i2}, \ldots, y_{iM}$ – et que nous voulons estimer la moyenne de population. S'il n'y a pas de données manquantes, y_{im} est simplement la valeur déclarée. L'estimation tirée de chacune des valeurs *M* est

$$\hat{\bar{Y}}_m = \sum_{i\in s} w_i y_{im} \Big/ \sum_{i\in s} w_i \ ; \ m = 1, \ldots, M \ ,$$

où *s* désigne l'ensemble d'unités de l'échantillon. L'estimateur de la moyenne de population est alors la moyenne des moyennes *M*,

$$\hat{\bar{Y}} = \sum_{m=1}^{M} \hat{\bar{Y}}_m \Big/ M \ .$$

On calcule la variance de $\hat{\bar{Y}}$ à l'aide de formules propres à l'imputation multiple. Désignons un estimateur de la variance de $\hat{\bar{Y}}_m$ par U_m. Ainsi, il est possible d'estimer chaque U_m par v_J^* en (1) ou un estimateur approprié de variance avec répliques. Une composante « intravariance » est $U^* = \sum_{m=1}^{M} U_m / M$ et une composante « intervariance » se calcule comme $B = (M-1)^{-1} \sum_{m=1}^{M} \left(\hat{\bar{Y}}_m - \hat{\bar{Y}}\right)^2$. Pour l'estimation moyenne VP, le calcul de variance est alors

$$v\left(\hat{\bar{Y}}\right) = U^* + B\left(1 + M^{-1}\right).$$

Il y a plusieurs options logicielles d'imputation multiple et d'analyse de données en imputation multiple, dont certaines sont examinées par Horton et Lipsitz (2001). IVEware (Raghunathan, Solenberger et Van Hoewyk, 2003) est un logiciel appelable en SAS et écrit en macrolangage SAS avec un ensemble de programmes en C et en Fortran. Le système SAS est d'un emploi obligatoire, mais celui-ci est si souvent présent dans les organismes d'enquête que ce n'est pas là une limitation digne de mention. IVEware exécutera des imputations simples et multiples à l'aide de l'algorithme de régression séquentielle que décrivent Raghunathan, Lepkowski, Van Hoewyk et Solenberger (2001). Au nombre des autres options logicielles en question, on peut ranger Solas^{MD}, les programmes autonomes élaborés par Schafer (1997) et un certain nombre de procédures en SAS. Toutes ne sauraient convenir à des données complexes d'enquête.

Comme il ressort des observations qui accompagnent Rubin (1996), il y avait un certain débat à l'époque sur la validité des estimations de variance avec la formule indiquée plus haut pour les échantillons en grappes ou les situations où le modèle d'imputation omet des variables importantes. Il semblerait que ces zones d'ombre n'ont pas encore été dissipées et que la recherche se poursuivra dans la présente décennie sur les forces et les faiblesses de l'imputation multiple. Comme le fait voir Judkins (1996), les praticiens auront encore besoin de règles précises pour juger si cette méthode convient.

3. IMPUTATION ET ESTIMATION D'INDICES DE PRIX

Les indices de prix comptent parmi les grandes statistiques diffusées par les gouvernements nationaux. Ils servent d'indicateurs économiques, de déflateurs d'autres séries économiques (produit intérieur brut, ventes au détail, pouvoir d'achat du dollar de consommation, révision des prestations de revenu, échelles mobiles des conventions collectives, etc.) et d'indicateurs de politique monétaire.

Nous nous attacherons à l'Indice des prix à la consommation (IPC), mais les mêmes questions se posent souvent dans le cas des indices de prix à la production ou autres. Parmi les programmes statistiques des gouvernements, la publication de cet indice est des plus importantes dans maints pays à cause des vastes mouvements monétaires qui peuvent être liés à l'ampleur des variations indiciaires. Aux États-Unis par exemple (voir http://bls.gov/cpi.htm), l'IPC influe sur le revenu de presque 75 millions de gens par mesure réglementaire, qu'il s'agisse des 48,4 millions de bénéficiaires du régime de sécurité sociale, des 4,2 millions de retraités et de survivants des régimes applicables aux membres militaires et civils de la fonction publique fédérale ou des 19,8 millions de bénéficiaires du régime « Food Stamps » de bons d'alimentation. Mentionnons aussi que le coût des repas servis à l'école à 26,5 millions d'enfants est indexé.

Plus de deux millions de travailleurs américains relèvent de conventions collectives où les salaires sont en indexation à l'IPC. Il y a souvent aussi indexation des loyers, des redevances, des pensions alimentaires et des prestations d'aide à l'enfance. Aux États-Unis, les tranches d'imposition fédérales sont modifiées chaque année en vue de prévenir l'alourdissement fiscal par l'inflation (phénomène de la dérive fiscale). Le Congressional Budget Office (O'Neill, 1995, tableau 1) a estimé qu'une baisse hypothétique de 0,5 point de l'IPC aurait contribué pour 26,2 milliards au total à la diminution du déficit fédéral dans le seul exercice 2000 compte tenu des baisses des dépenses fédérales, des hausses de recettes et des allégements de service de la dette.

Les systèmes d'enquête nécessaires à l'estimation indiciaire sont souvent complexes et obligent à résoudre un certain nombre de problèmes statistiques : plan d'échantillonnage, estimation des composantes de la variance, répartition de l'échantillon en fonction d'objectifs multiples, corrections de données particulières aux indices, imputation de valeurs manquantes, estimation de variance pour des estimateurs non linéaires, etc. Schultze et Mackie (2002) et Leaver et Valliant (1995) décrivent un grand nombre de domaines d'investigation statistique et économique. Une application importante de l'imputation, qui n'est peut-être pas reconnue comme telle, est celle des corrections de qualité.

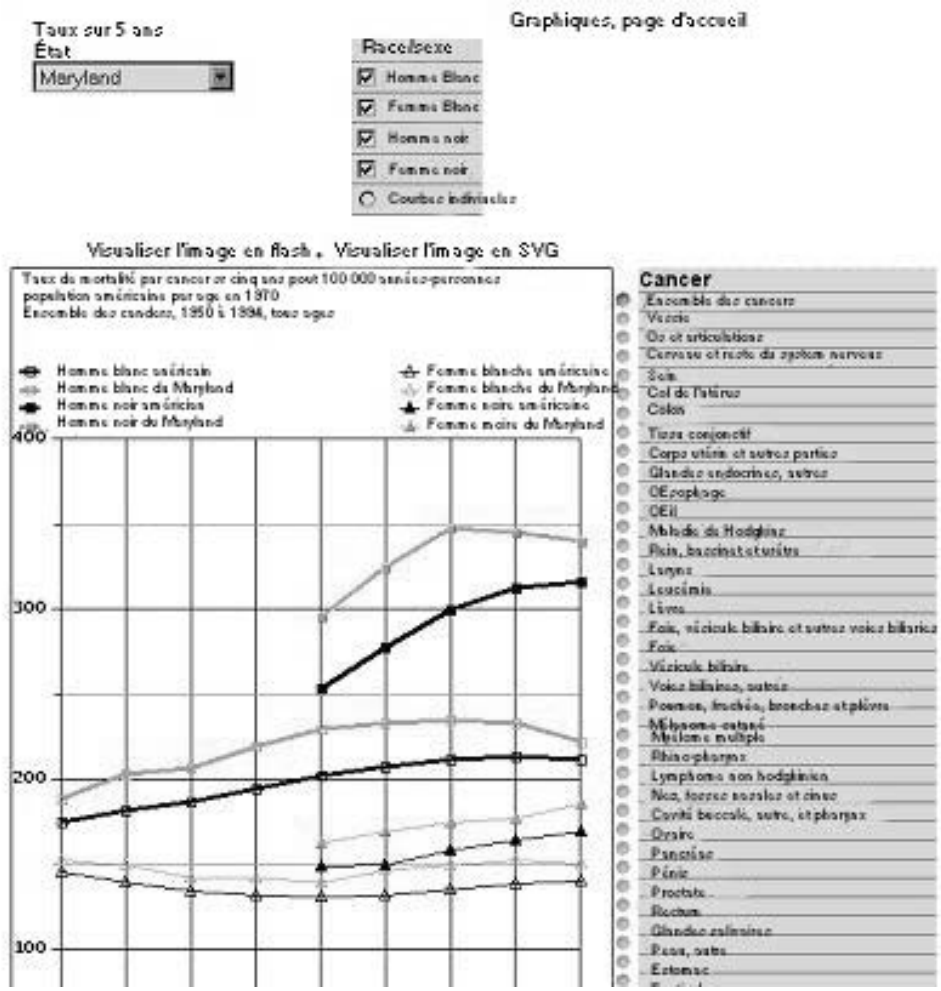
La qualité de certains articles peut varier amplement dans le temps. Un exemple patent en est le matériel informatique, mais bien d'autres produits subissent des variations de qualité : appareils audioélectroniques (lecteurs MP3 et DVD), réfrigérateurs, laveuses et sécheuses, etc. En raison de cette évolution, il est fréquemment impossible de relever les prix en toute comparabilité dans des périodes consécutives. Les économistes procèdent à des

corrections de qualité à l'aide de modèles par souci de comparabilité. Une méthode économétrique type à cet égard consiste à prévoir le prix d'un article en fonction de ses caractéristiques, puis à comparer le prix ainsi dégagé au prix réel. Ces méthodes dites « hédonistes » sont examinées dans Triplett (1987). En 1998 par exemple, on a réduit de 6,5 % l'indice des micro-ordinateurs à la suite de corrections hédonistes (Schultze et Mackie, 2002, p. 129-130).

À l'instar d'autres imputations, ces corrections ne reposent pas sur des constantes connues, mais sont traitées comme s'il y en avait. Il serait acceptable de considérer les imputations mêmes comme déterministes compte tenu des données des échantillons si (1) le résultat est un estimateur statistiquement sans biais ou cohérent de l'IPC et que (2) une estimation convenable des erreurs-types peut se calculer. Dans les autres cas, l'imputation multiple reste une possibilité.

Dans son message de président au Joint Statistical Meetings de 1998, Moynihan (1999) s'est dit d'avis que l'affinement des méthodes de correction de qualité constituait un des grands enjeux statistiques pour le gouvernement américain à cause de l'incidence de ces méthodes sur le budget des États-Unis. On continue à s'interroger sur le degré de sensibilité indiciaire à l'ordre de grandeur des corrections, sur la façon de tenir compte de ces mêmes corrections dans la variance et sur le mode de conception d'études de sensibilité valables d'un point de vue économique. Pour des statistiques primordiales comme celles de l'IPC, il est bon de consacrer temps et efforts à l'étude de ces questions.

Figure 1. Fragment du site Web du National Cancer Institute qui présente les taux de mortalité par cancer aux États-Unis; http://www3.cancer.gov/atlasplus (l'axe temporel horizontal est escamoté)

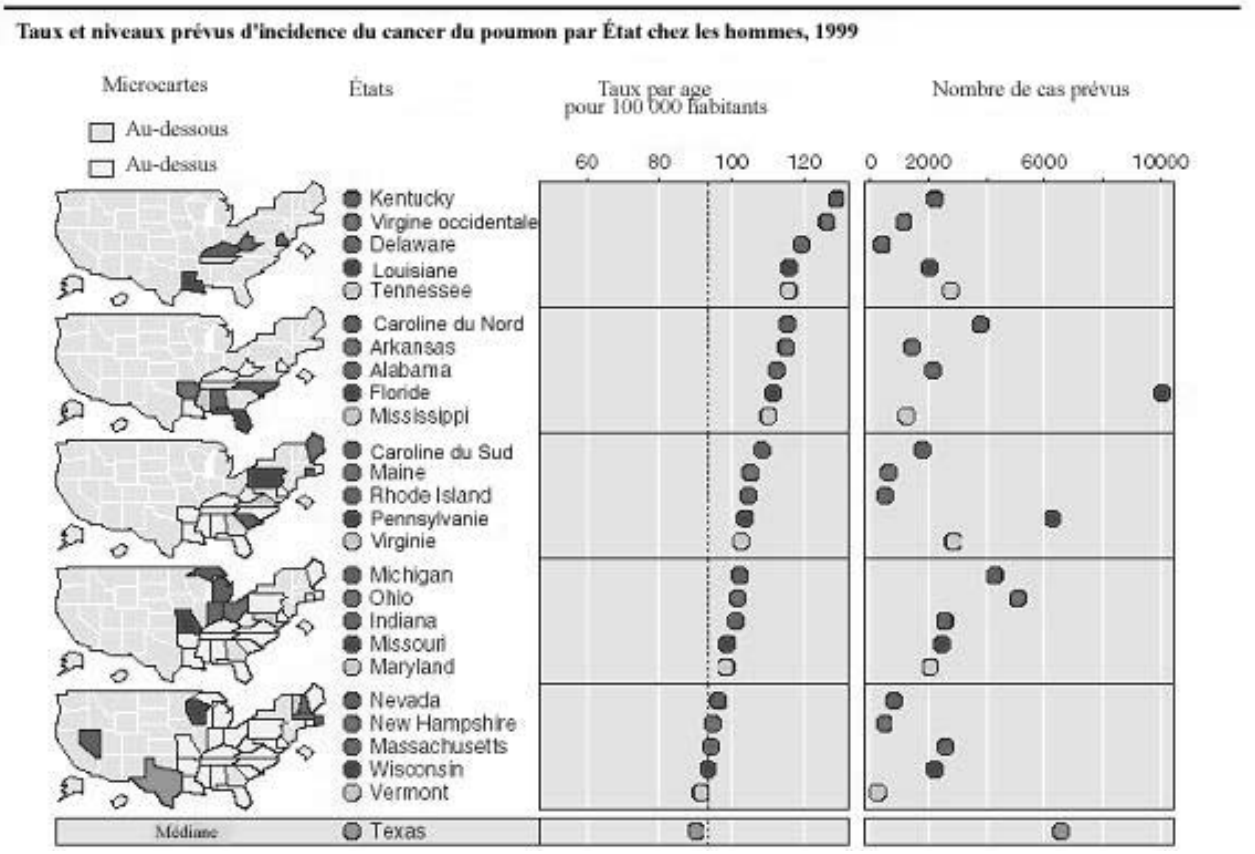


4. RECOURS AU GRAPHISME DANS LES PUBLICATIONS GOUVERNEMENTALES

Il y a encore dix ans, les éléments graphiques étaient restreints dans les publications gouvernementales, mais depuis l'évolution a été rapide. Dans cette section, nous présenterons quelques éléments graphiques qui, à mon avis du moins, offrent de remarquables exemples de ce que la technologie apporte de nos jours comme capacités graphiques de présentation de données. Les exemples viennent du National Cancer Institute (NCI) des États-Unis, mais il y a d'autres organismes dans le monde qui font des travaux graphiques tout aussi intéressants. Les éléments graphiques des figures 1 à 4 ont été conçus par la Statistical Research and Applications Branch du NCI et comportent des cartes et des graphiques interactifs des taux de mortalité par cancer aux États-Unis (National Cancer Institute, 2003a, 2003b). Ils ont été créés à l'aide de POPCHART^{MD} (Corda Technologies, 2003). Pickle, White, Mungiole, Jones (1996) et Carr (2001) examinent certains des aspects de ce graphisme de présentation statistique.

La figure 1 présente un fragment de page Web du National Cancer Institute. L'utilisateur peut sélectionner un État ou tout le territoire américain, la race/sexe, le type de cancer (ensemble, vessie, os et articulations, cerveau et autres parties du système nerveux, etc.), et une série chronologique de taux de mortalité sur cinq ans est mise en graphique à l'écran. (L'axe temporel horizontal est escamoté à la figure 1.) En cliquant à droite, l'usager peut sauvegarder un graphique dans un format d'animation SVG ou en GIF. Il peut également visualiser les données précises d'un point en y faisant passer la souris.

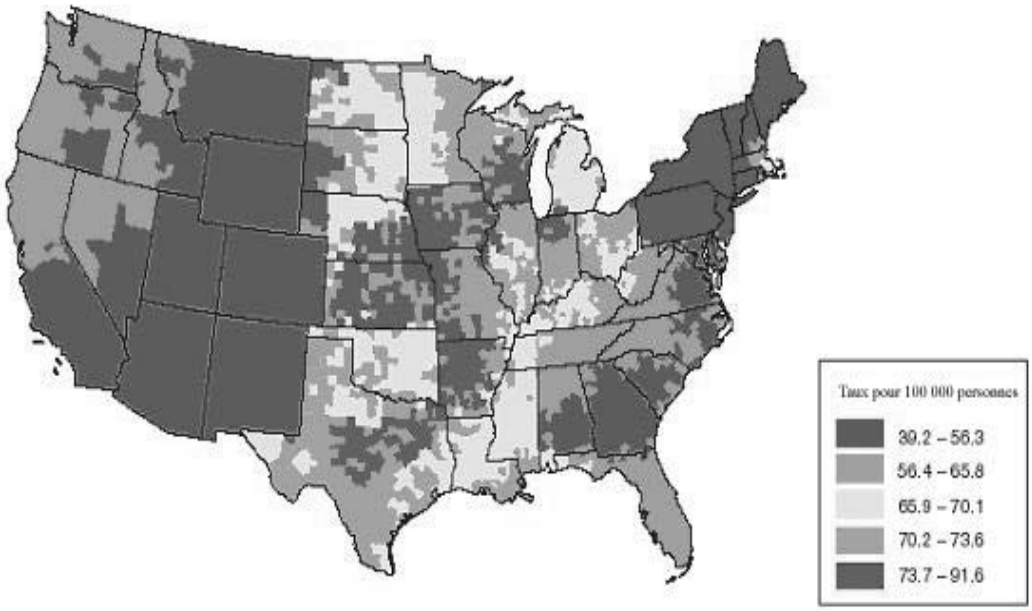
Figure 2. Fragment de microcartes des taux et des niveaux prévus d'incidence du cancer du poumon par État; http://srab.cancer.gov/



La figure 2 relie aux cartes les taux par âge d'incidence du cancer du poumon des États. Les pointillés du côté droit indiquent les taux des États et les chiffres de prévision de cas par taux ordonné. Les points sont en raccordement couleur avec les petites cartes des États-Unis à gauche. Les États forment des groupes de cinq et, à mesure que l'œil parcourt la page vers le bas, les États s'ajoutent.

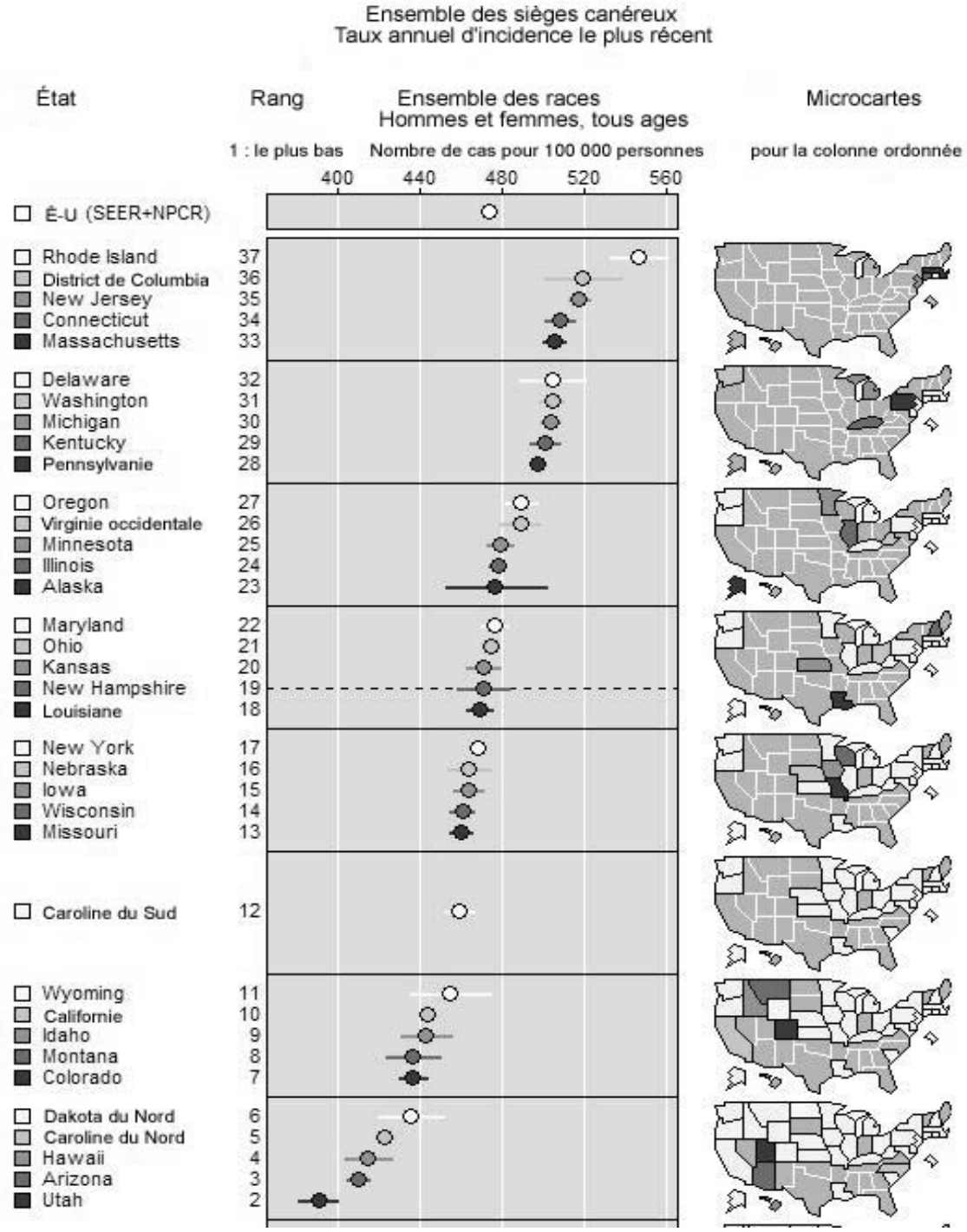
La troisième figure est une carte couleur des comtés américains où on s'est servi d'un programme de lissage non paramétrique (Mungiole, Pickle et Simonson, 1999) pour établir les prévisions des divers comtés. La figure 4 présente également en microcartes les taux d'incidence du cancer par État. Les États visés sont ordonnés du taux d'incidence le plus haut au taux le plus bas. Les cartes permettent de voir les tendances régionales. La figure présente en outre les intervalles de confiance particuliers aux États. On peut ainsi constater que les estimations relatives à l'Alaska sont moins précises que les autres.

Figure 3. Carte couleur des taux prévus de cancer colo-rectal des comtés chez les hommes en 1999;



http://srab.cancer.gov/.

Figure 4. Fragment de microcartes des taux annuels d'incidence du cancer par État; http://srab.cancer.gov/.



5. CONCLUSION

Nous avons évoqué plusieurs domaines – analyse de données avec imputation, échantillonnage à plusieurs degrés et graphisme de présentation statistique – où des techniques et des outils nouveaux sont disponibles, mais sans y être universellement employés. On pourrait dresser une longue liste d'autres domaines où il est possible de faire meilleur usage des méthodes disponibles ou où des travaux de recherche-développement s'imposent. En voici quelques-uns :

- Estimations pour petits domaines
  - Il faut élaborer des normes permettant de juger si des estimations sont publiables
  - On a besoin de logiciels généralisés pour calculer de façon empirique le meilleur estimateur linéaire sans biais ainsi que les estimations de Bayes hiérarchisées et empiriques pour l'estimation ponctuelle et l'estimation d'erreur quadratique moyenne.
- Il faut des logiciels d'analyse en ligne qui permettent des recherches définies par l'utilisateur dans les fichiers de microdonnées
  - On doit protéger la confidentialité des données, tout en laissant de la souplesse aux analystes.
- Extraction de données – comment utiliser ces méthodes au mieux?
- Méthodes de protection de la confidentialité
  - Il faut ménager un juste équilibre entre la protection des renseignements personnels des enquêtés et l'exploitabilité des ensembles de données pour les analystes.

Il est primordial de disposer de logiciels d'application de techniques avancées d'estimation. Dans les grandes enquêtes périodiques, on peut compter sur des budgets de développement pour l'élaboration de logiciels sur mesure, mais dans un grand nombre de petites enquêtes il faut se rabattre sur les logiciels commerciaux. Sinon, on risque de continuer à employer des formules qui ne conviennent pas, notamment pour les estimations de variance. Les choix qui s'offrent à nous sont bien plus riches qu'il y a quelques décennies, mais des lacunes subsistent. Les progiciels analytiques du marché doivent comporter des caractéristiques par lesquelles on pourra tenir compte des imputations dans l'estimation de la variance de données d'enquête. Il faut des options logicielles pour l'imputation multiple et les méthodes avec répliques de conception récente dans le traitement de données d'imputation. Du point de vue de l'utilisateur, les logiciels doivent fréquemment être mis à jour, commander un bon soutien technique et se vendre à un prix abordable. Du point de vue du concepteur commercial, il faut que les débouchés soient suffisants pour la rentabilité des produits logiciels. Si un logiciel spécialisé est mis au point dans un organisme, il faut pouvoir en attendre des gains d'efficience ou de validité par rapport aux méthodes en usage dans l'organisme. Une direction et un appui fermes des gestionnaires en matière d'innovation représentent un facteur clé de changement.

RÉFÉRENCES

Corda Technologies, Inc. (2003), POPCHART, http://www.corda.com.

Carr, D. B. (2001), "Designing Linked Micromap Plots for States with Many Counties," *Statistics in Medicine*, **20**, John Wiley & Sons, 1331-1339.

Fuller, W.A. (1998), "Replication Variance Estimation for Two-Phase Samples", *Statistica Sinica*, **8**, 1153-1164.

Fuller, W.A. (2003), "Estimation for Multiphase Samples" in *Analysis of Survey Data*, New York: John Wiley & Sons.

Horton, N.J. et Lipsitz, S.P. (2001), "Multiple Imputation in Practice: Comparison of Software Packages for Regression Models with Missing Variables," *American Statistician*, **55**, 244-254.

Jones, M.E., Brick, J.M., Kalton, G., et Valliant, R. (2003), "A Simulation Study of Two Methods of Variance Estimation with Hot Deck Imputation," *Proceedings of the Section on Survey Research Methods*, Washington DC: American Statistical Association, to appear.

Judkins, D. (1996), Comment on "Multiple Imputation after 18+ Years" by D. Rubin, *Journal of the American Statistical Association*, **91**, 507-510.

Kim, J.K., et Sitter, R.R. (2003), "Efficient Replication Variance Estimation for Two-Phase Sampling," *Statistica Sinica*, **13**, 641-653.

Leaver, S.G. et Valliant, R. (1995), "Statistical Problems in Estimating the U.S. Consumer Price Index," Chapter 28, pp. 543-566, in Business Survey Methods, B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge, P. Kott eds., New York, John Wiley.

Lundström, S. et Särndal, C.-E. (1999), "Calibration as a Standard Method for Treatment of Nonresponse", *Journal of Official Statistics*, **15**, 305-327.

Moynihan, D.P. (1999), "Data and Dogma in Public Policy," *Journal of the American Statistical Association*, **94**, 359-364.

Mungiole, M. Pickle, L.W., et Simonson, K.H. (1999), "Application of a Weighted Head-banging Algorithm to Mortality Data Maps," *Statistics in Medicine*, **18**, 3201-3209.

National Cancer Institute (2003a), *Cancer Mortality Maps and Graphs*, http://www3.cancer.gov/atlasplus/.

National Cancer Institute (2003b), *Statistical Tables, Maps, and Graphs*, http://surveillance.cancer.gov/statistics/stat_sources.

Opsomer, J., Fuller, W.A., et Li, X. (2003), "Replication Variance Estimation for the National Resources Inventory", *Proceedings of the Section on Survey Research Methods*, American Statistical Association, to appear.

O'Neill, J. (1995), "Prepared Statement on the Consumer Price Index," Hearings before the Committee on Finance, U.S. Senate, 194th Congress, First Session, Washington DC: U.S. GPO.

Pickle, L.W., White A.A., Mungiole, M., Jones, G.K. (1996), "Atlas of United States Mortality," *Proceedings of the Statistical Graphics Section*, American Statistical Association 40-44.

Raghunathan, T., Solenberger, P., et Van Hoewyk, J. (2003), IVEware: Imputation and Variance Estimation Software, http://www.isr.umich.edu/src/smp/ive/.

Raghunathan, T., Lepkowski, J., Van Hoewyk, J. et Solenberger, P. (2001), "A multivariate technique for multiply imputing missing values using a sequence of regression models", *Survey Methodology*, **27**, 85-96.

Rao, J.N.K. (1999), "Some Current Trends in Sample Survey Theory and Methods", *Sankhyā* , **61**, 1-25.

Rao, J.N.K. et Shao, J. (1992), "Jackknife Variance Estimation with Survey Data under Hot Deck Imputation", *Biometrika*, **79**, 811-822.

Rao, J.N.K. et Sitter, R.R. (1995), "Variance Estimation under Two-phase Sampling with Applications to Imputations for Missing Survey Data," *Biometrika*, **82**, 453-460.

Rubin, D.B. (1996), "Multiple Imputation after 18+ Years", *Journal of the American Statistical Association*, **91**, 473-506.

Särndal, C.-E., Swensson, B., Wretman, J. (1992), *Model Assisted Survey Sampling*, New York: Springer-Verlag.

Schultze, C.L., et Mackie, C. eds. (2002), *At What Price? Conceptualizing and Measuring Cost-of-Living and Price Indexes*, National Academy Press: Washington DC.

Schafer, J.L. (1997), *Analysis of Incomplete Multivariate Data*, London: Chapman & Hall.

Shao, J. (2001), "Replication Methods for Variance Estimation in Complex Surveys with Imputed Data", Chapter 20 in *Survey Nonresponse*, eds. R.M. Groves, D.A. Dillman, J.L. Eltinge, et R.J.A. Little, New York: John Wiley & Sons, 303-314.

Shah, B.V., Barnwell, B., Bieler, G., Boyle, K., Folsom, R., Lavange, L., Wheeless, S., et Williams, R. (1996). Technical Manual: Statistical Methods and Algorithms Used in SUDAAN, Research Triangle Park, NC: Research Triangle Institute.

Triplett, J.E. (1987), "Hedonic Functions and Hedonic Indexes," in *The New Palgrave: A Dictionary of Economics*, M.M. Eatwell and P. Newman, eds., Vol. 2, London: Macmillan, 630-634.

Valliant, R. (1999), "Uses of Models in the Estimation of Price Indexes: A Review", *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 94-102.

Valliant, R. (2003), "The Effect of Multiple Weighting Steps on Variance Estimation", *Journal of Official Statistics*, **19**, to appear.

Yung, W., et Rao, J.NK. (1996). Jackknife Linearization Variance Estimators under Stratified Multi-stage Sampling. *Survey Methodology*, **22**, 23-31.