No 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2005 : Défis méthodologiques reliés aux besoins futurs d'information

2005

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2005 de Statistique Canada
Défis méthodologiques reliés aux besoins futurs d'information

TRAITEMENT DES DONNÉES D'ENQUÊTE MANQUANTES DANS LE CADRE D'ANALYSE LONGITUDINALE

Robert M. Baskin[1]

RÉSUMÉ

L'article passe en revue des techniques pouvant être utilisées pour résoudre le problème des données manquantes dans les enquêtes complexes lors de la réalisation d'une analyse longitudinale. Dans ce contexte, les données longitudinales sont des observations répétées (plusieurs cycles) sur une même unité de mesure. En plus de présenter les mêmes types de données manquantes que les données transversales, les observations longitudinales perdent des données supplémentaires lorsqu'il y a de la non-réponse aux cycles suivants, ce que nous appellerons érosion. En analyse longitudinale, les modèles à effets aléatoires sont ceux utilisés le plus fréquemment pour tenir compte de la nature longitudinale des données. Toutefois, l'intégration de plan de sondage complexe dans des modèles de type multiniveaux utilisés dans ce genre d'analyse longitudinale pose des difficultés, surtout en présence de données manquantes dues à l'érosion. On traite souvent cette forme de données manquantes par la modélisation en série de « l'érosion des échantillons » avec ou sans utilisation des poids longitudinaux. Une autre solution consiste à remplacer les données manquantes par imputation. En présence d'imputation, les estimateurs standard de la variance sous-estiment cette dernière. Donc, il convient d'envisager l'utilisation de méthodes permettant de tenir compte de la variance d'imputation, comme l'imputation multiple des données manquantes causées par l'érosion.

MOTS-CLÉS : érosion de l'échantillon; enquête complexe; données « missing at random » (MAR); modèles multiniveaux; imputation multiple

1. INTRODUCTION

L'article passe en revue des techniques pouvant être utilisées pour résoudre le problème des données manquantes dans les enquêtes complexes lors de la réalisation d'une analyse longitudinale. Il s'agit d'un sujet très général et, afin de limiter l'étendue de l'exposé, les sujets des séries chronologiques et des techniques de modélisation ne seront pas abordés. Seulement les techniques permettant de traiter les données manquantes dans le contexte de l'analyse longitudinal d'une enquête complexe seront examinées. Cependant, certains modèles longitudinaux particuliers seront utilisés à titre d'exemple. L'article a été conçu en supposant que l'analyste sait quel genre d'analyse il convient d'appliquer, mais qu'il a besoin d'aide pour résoudre les problèmes engendrés par les données manquantes.

Puisque l'exposé traite des données d'enquête manquantes, nous supposons que le lecteur est familier avec les méthodes d'estimation fondées sur des données d'enquête décrites dans les ouvrages de référence classiques tel que celui de Cochran (1977). Ces méthodes font souvent référence aux méthodologies sur une population finie, mais ici, elles feront généralement référence aux méthodologies *fondées sur le plan de sondage*. Par contraste, les méthodes parfois qualifiées de *statistiques traditionnelles* ou de méthodologies sur une population infinie seront désignées comme méthodologies *fondées sur un modèle* ou dépendantes d'un modèle. De plus, l'expression *inférence fondée sur la vraisemblance* sera utilisée pour désigner l'inférence s'appuyant sur une fonction de vraisemblance. Enfin, une combinaison des deux méthodes, telle qu'elle est présentée dans Sarndal, Swensen et Wretman (1992), sera appelée méthodologie *assistée par un modèle*, ce qui est différent de *l'approche en superpopulation* citée, entre autre, dans Ghosh et Meeden (1997).

Les difficultés engendrées par le fait de ne pas tenir compte du plan de sondage lors de la modélisation des données d'enquête ont été bien décrites, mais il est conseillé au lecteur de consulter Skinner, Holt et Smith (1987) ou Brogan (1998) pour une revue des questions telles que la robustesse des modèles et l'estimation sans biais de la variance

[1] Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850. Les opinions exprimées dans le présent document sont celles de l'auteur et ne sont pas sanctionnées officiellement par le Department of Health and Human Services ni l'Agency for Healthcare Research and Quality.

d'après les données d'une enquête complexe. Pour une introduction à l'utilisation des poids de sondage en modélisation des données d'enquête, il peut être intéressant de consulter Korn et Graubard (1999) et pour une analyse en profondeur de la pondération et des données d'enquête, Pfeffermann (1993) est la référence suggérée. Une description des difficultés engendrées par le fait de ne pas tenir compte des données manquantes, particulièrement dans l'analyse de cas complets, figure dans Korn et Graubard (1999) ou dans Little et Rubin (2002). Little et Rubin (2002) discutent également des concepts de données « missing completely at random » (MCAR), de données « missing at random » (MAR) et de données « not missing at random » (NMAR), dont il sera question à l'occasion dans le présent article.

Le plan de l'exposé fait l'objet des prochaines lignes. En premier lieu, il sera question des méthodes fondées purement sur le plan de sondage pour traiter les données manquantes. En deuxième lieu, des méthodes dépendantes d'un modèle qui intègrent certains aspects des idées fondées sur le plan de sondage seront présentées. Par la suite, les méthodes bayesiennes seront abordées. Puis, les méthodes comportant une imputation seront introduites. Finalement, un exemple de modélisation de données manquantes sera donné. La plupart des sujets présentés ici sont traités dans Skinner, Holt et Smith (chapitres 12 à 14, 1989) ou dans Chambers et Skinner (2003).

2. MÉTHODES FONDÉES SUR LE PLAN DE SONDAGE

Une analyse plus détaillée des sujets abordés dans cette section figure dans Skinner, Holt et Smith (1989, chapitre 14). Dans le cas des méthodes fondées purement sur le plan de sondage, la seule option, mis à part le remplacement des données manquantes par imputation, est le recours à la repondération. En analyse transversale, les méthodes habituellement choisies sont la repondération pour la non-réponse totale (unité manquante) et l'imputation pour la non-réponse partielle (question(s) manquante(s)). Cependant, en analyse longitudinale, une unité peut manquer pour une période mais non pour une autre. Cette situation arrive lorsqu'une unité est présente lors des premiers cycles de collecte des données, puis cesse plus tard de faire partie de l'enquête. L'érosion est souvent une des préoccupations majeures. Les analystes pensent souvent que ces cas d'unités manquantes diffèrent de façon systématique des cas complets disponibles pour l'analyse. Cependant, d'autres scénarios de données manquantes sont possibles. Une nouvelle unité peut entrer dans l'échantillon après les premiers cycles de collecte. Cette situation, parfois appelée *nouvelle unité statistique* ou *échantillon supplémentaire*, n'est habituellement pas considérée comme une source importante de biais. Souvent, on émet l'hypothèse que les autres modèles de données manquantes correspondent à des cas de données MCAR.

Afin d'aborder l'analyse longitudinale d'une façon fondée purement sur le plan de sondage, l'ensemble des observations longitudinales faites pour chaque unité est traitée comme un vecteur d'observations sur l'unité en question et une matrice de covariance est estimée en appliquant une méthode de pondération. Si aucune unité n'a d'observation manquante, cette méthode peut être appliquée directement aux modèles linéaires ou conjointement à la linéarisation pour estimer de nombreux modèles non linéaires. En présence de données manquantes causées par l'érosion, si l'on dispose des poids de sondage selon le cycle d'une enquête, on peut utiliser la méthode qui suit pour modéliser les échantillons perdus par érosion au cours du temps. Pour chaque période, disons *t*, les unités qui n'ont pas décroché jusqu'au temps t sont modélisées en utilisant les poids de sondage selon le cycle pour estimer la matrice des covariances des observations pour les unités. On obtient ainsi pour chaque période *t* un ensemble de modèles qui vise à tenir compte des données manquantes causées par l'érosion.

Cette méthode offre de nombreux avantages de la méthodologie fondée sur le plan de sondage, mais elle a aussi ses inconvénients. Une des difficultés vient du fait que l'on tient compte, dans toute analyse, que d'un seul modèle de données manquantes, habituellement les données manquantes dues à l'érosion. Un autre problème qui empêche souvent d'utiliser cette méthode est que l'on ne dispose pas pour toutes les enquêtes des poids de sondage pour chacun des cycles de collecte. Le problème, peut-être le plus important ici, est que toute comparaison des modèles est faite de façon ponctuelle.

3. MODÈLES À EFFETS ALÉATOIRES

Les modèles à effets aléatoires sont devenus la norme *de facto* pour l'analyse longitudinale en statistique traditionnelle. Verbeke et Molenberghs (2000), par exemple, décrivent l'utilisation des modèles à effets aléatoires pour l'analyse longitudinale avec données manquantes. Les modèles à effets aléatoires ont la propriété de produire une inférence valide même si les données sont MAR, tandis que d'autres modèles nécessitent parfois l'hypothèse que les données soient MCAR pour assurer que l'inférence soit valide. Cependant, comme dans les traités de statistique traditionnelle en général, l'approche de Verbeke et Molenberghs (2000) n'aborde pas les questions que soulève la modélisation de données provenant d'une enquête complexe.

La construction de modèles à effets aléatoires longitudinaux qui tiennent compte d'un plan de sondage complexe est une tâche difficile. On pourrait soutenir, d'un point de vue technique, qu'une approche fondée sur le plan de sondage ne permet pas de produire un modèle à effets aléatoires. Toutefois, il est possible d'élaborer un modèle à effets aléatoires *convergent par rapport au plan de sondage* qui est fondé sur une fonction de vraisemblance. L'expression « convergent par rapport au plan de sondage » utilisée ici fait référence à un modèle produisant des estimations qui pourraient ne pas être sans biais par rapport au plan de sondage, mais qui sont convergentes au sens de l'absence asymptotique de biais par rapport au plan de sondage. Cette démarche est celle proposée pour les données transversales dans Pfeffermann et Lavange (1998) et modifiée dans Pfeffermann et coll. (1998). Skinner et Holmes (2003) étendent les travaux antérieurs afin d'y inclure les modèles à effets aléatoires pour les données longitudinales. Il est probable, puisque les modèles à effets aléatoires de Skinner et Holmes sont fondés sur la vraisemblance, qu'ils retiennent la propriété d'être valides pour des données MAR par opposition à des données MCAR. Cette série d'articles montre une évolution des moindres carrés généralisés (MCG) aux MCG pondérés par les probabilités, puis aux moindres carrés généralisés itérés (MCGI) et enfin aux moindres carrés généralisés itérés pondérés par les probabilités (MCGIPP). Le lecteur trouvera dans Skinner et Holmes (2003) des références concernant chaque étape.

Les idées qui sous-tendent la technique des MCGIPP sont illustrées par le scénario qui suit. Si nous voulons construire un modèle à effets aléatoires sur nos données, alors ce modèle aura des paramètres de régression, indiqués par $\boldsymbol{\beta}$, et des composantes de la variance indiquées par $\boldsymbol{\theta}$. Si nous ne tenons pas compte du poids de sondage et que nous construisons un modèle par la méthode des MCGI, nous devons émettre une hypothèse initiale quant à $\boldsymbol{\theta}$, disons $\boldsymbol{\theta}_0$. En utilisant la valeur de $\boldsymbol{\theta}_0$, nous produisons une estimation de $\boldsymbol{\beta}$, disons $\boldsymbol{\beta}_1$, par les moindres carrés généralisés. Puis, en traitant $\boldsymbol{\beta}_1$ comme une constante, nous retournons en arrière et produisons une nouvelle estimation de $\boldsymbol{\theta}$, disons $\boldsymbol{\theta}_1$. Ce processus est itéré, de sorte que chaque ensemble de paramètres est mis à jour consécutivement jusqu'à ce que l'on atteigne la convergence.

À ce stade, la méthode semble prometteuse parce qu'elle est fondée sur la vraisemblance, mais elle ne tient pas compte des poids de sondage. L'étape suivante consiste donc à intégrer ces poids. Pfeffermann et coll. (1998) ont modifié la méthode MCGI de façon à ce que les poids de sondage soient inclus de la manière suivante. Pour commencer, les poids sont normalisés ingénieusement comme il est décrit dans Pfeffermann et coll. (1998). Puis, les poids normalisés sont inclus dans le modèle à effets aléatoires MCGI sous forme de covariables, mais non sous forme de poids de sondage (poids p). On obtient ainsi des estimations ponctuelles des paramètres $\boldsymbol{\beta}$ qui sont simultanément sans biais par rapport au modèle et convergentes par rapport au plan. Malheureusement, si le plan de sondage n'est pas équipondéré (autopondération), cette modification produit des estimations de la variance qui ne sont convergentes ni par rapport au modèle, ni par rapport au plan de sondage, de sorte qu'une modification supplémentaire est nécessaire. À ce stade, à la deuxième étape de l'approche MCGI, il faut utiliser un estimateur de la variance convergent par rapport au plan de sondage, ce qui a le malencontreux effet d'obliger l'analyste à arrêter le logiciel standard de calcul des MCGI à chaque itération afin d'insérer une nouvelle estimation de la variance dans le processus ou d'effectuer l'itération manuellement. Néanmoins, cette méthode offre un moyen d'ajuster des modèles à effets aléatoires à des données d'enquête complexe fondés sur la vraisemblance, mais demeurant convergents par rapport au plan de sondage.

Skinner et Holmes (2003) font une modification supplémentaire pour tenir compte de l'aspect longitudinal des données. La correction vise à tenir compte de la corrélation en série des données dans le modèle longitudinal. Ensemble, ces modifications produisent des modèles ayant les propriétés théoriques souhaitées, mais dont l'ajustement est assez fastidieux en pratique.

Un dernier aspect technique de la méthode qu'il convient d'éclaircir est la façon de déterminer l'estimation originale de la variance. Pfeffermann et coll. (1998) proposent de partir d'un estimateur convergent par rapport au plan de sondage des composantes de la variance $\boldsymbol{\theta}_0$ sachant une estimation convergente par rapport au plan de sondage de $\boldsymbol{\beta}$ provenant d'un modèle basé sur le plan de sondage standard.

4. MODÈLES BAYESIENS

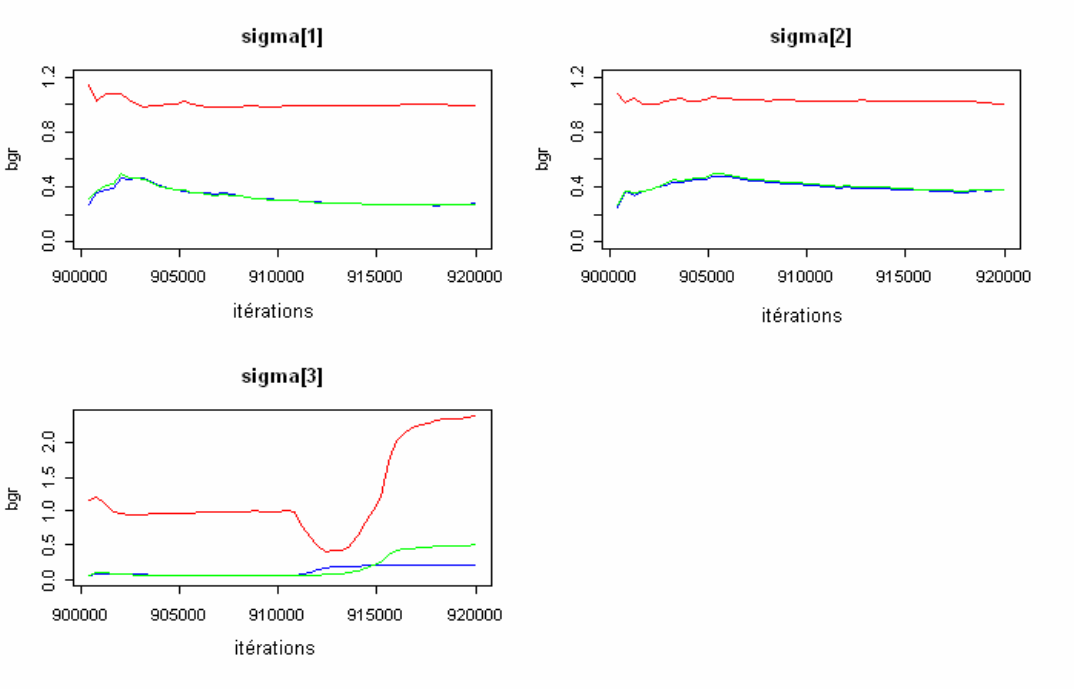
Les méthodes bayesiennes sont de plus en plus utilisées et ont de bonnes propriétés théoriques. Le gain de popularité de ces méthodes est dû en grande partie à l'utilisation des méthodes de Monte Carlo comme l'échantillonnage de Gibbs. Ces méthodes pour modèles bayesiens, requérant une grande puissance de calcul, sont offertes dans plusieurs progiciels, en particulier WinBUGS. La version libre de ce dernier, OpenBUGS, devient très répandue. Le progiciel R, de la R Development Core Team (2005), offre l'accès direct à OpenBUGS ainsi qu'à de nombreux outils diagnostiques.

Tel que mentionné, les méthodes bayesiennes ont de bonnes propriétés théoriques. Une introduction à l'application de ces méthodes en population finie se retrouve dans Ghosh et Meeden (1997). Les méthodes décrites par ces auteurs correspondent à des modèles en superpopulation en ce sens qu'ils tiennent compte de l'échantillonnage en population finie et d'un modèle aléatoire pour les valeurs observées. Les modèles qu'ils utilisent pour les valeurs observées ont des propriétés semblables à celles des modèles à effets aléatoires, de sorte qu'ils émettent l'hypothèse que les données sont MAR si le mécanisme de création des données manquantes n'est pas intégré directement dans le modèle. Une distinction doit être faite entre les hypothèses qui sous-tendent les modèles bayesiens et celles qui sous-tendent les modèles de la théorie classique. Afin de faire des inférences valides d'après les modèles à effets aléatoires, l'analyste doit supposer que les données sont MAR et que tout paramètre du modèle est distinct de tout paramètre du mécanisme de non-réponse. En revanche, pour qu'une inférence à partir d'un modèle bayesien soit valide, en plus de l'hypothèse des données MAR , l'analyste doit supposer que les paramètres du modèle et ceux du mécanisme de non-réponse ont des *distributions a priori indépendantes*. Cette hypothèse est plus forte que celle de l'existence de paramètres distincts et si l'analyste n'est pas convaincu de l'indépendance des distributions, il devrait, afin de faire des inférences valides à partir du modèle bayesien, modéliser les données en supposant qu'elles sont NMAR, ce qui soulève également la question des lois a priori des paramètres. Toutefois, la discussion des lois a priori et de toute controverse à ce sujet dépasse de loin le cadre du présent article. Le lecteur intéressé par une discussion sur les lois a priori dans une perspective bayesienne est invité à consulter Bernardo et Smith (2000).

Si l'analyste souhaite réaliser une analyse bayesienne, il existe aujourd'hui des logiciels qui rendent la construction de modèles bayesiens beaucoup plus facile que par le passé. Toutefois, outre les problèmes habituels de vérification des modèles, l'analyste doit être conscient que la modélisation par la méthode de Monte Carlo offerte dans les progiciels tels que BRugs, l'interface R avec OpenBUGS, exige la vérification de la convergence. Des outils diagnostiques permettant cette vérification sont intégrés dans BRugs.

À titre d'exemple, considérons l'analyse longitudinale suivante des données sur l'État de Santé Autodéclaré (ESAD) provenant de l'Enquête par Panel sur les Dépenses Médicales (EPDM) de 2002. L'EDPM est une enquête nationale par sondage probabiliste complexe et parrainée par l'agence de recherche et de qualité sur les soins de santé. L'EDPM est conçue afin de fournir des estimations nationalement représentatives de l'utilisation des services de santé, des dépenses en services de santé, des sources de paiement et de la couverture par une assurance-maladie pour la population civile non placée en établissement des États-Unis. L'EDPM est composée de trois enquêtes étroitement liées, la Composante des Ménages (CM) étant l'enquête de base. L'EDPM-CM est une enquête par panels et les personnes admissibles font partie du panel pendant cinq cycles consécutifs de collecte des données. Le fait que les données de l'EDPM-CM soient recueillies auprès de panels continus offre l'avantage de permettre la réalisation d'une analyse longitudinale des données, mais pose aussi le défi de résoudre le problème des données manquantes dans un ensemble de données longitudinales. Les données sur l'ESAD sont recueillies au cours de cinq cycles consécutifs, sur une période de deux ans, et présentent des cas d'érosion, des échantillons supplémentaires, ainsi que d'autres schémas de génération de données manquantes. Nous avons ajusté à ces données un simple modèle de moyenne avec et sans covariables temporelles en utilisant des distributions a priori non informatives. Nous nous sommes servis du logiciel BRugs, et avons soumis le modèle à 900 000 itérations de rodage avant

d'activer les outils diagnostiques. Ce nombre d'itérations peut paraître extrême pour un simple modèle de moyenne, mais nous l'avons utilisé pour démontrer les faits. Une fois les diagnostics activés, nous avons soumis le modèle à 20 000 itérations supplémentaires, puis nous avons examiné les résultats diagnostiques afin de déterminer la convergence. De façon générale, les outils diagnostiques ont indiqué la convergence de tous les paramètres de moyenne du modèle. Par contre, ils ont révélé un problème en ce qui concerne la convergence des paramètres de variance. La statistique de Brooks-Gelman-Rubin (BGR) pour les composantes de la variance est présentée à la figure 1. Le graphique comporte trois courbes diagnostiques. La courbe supérieure devrait converger vers 1,0 et les deux courbes inférieures devraient converger ensemble vers 0,4. Comme le montre le graphique diagnostique qui suit, la troisième composante de la variance semble converger au départ, mais après 10 000 itérations, la courbe supérieure s'écarte soudainement du point idéal de 1,0 et diverge fortement. Si l'analyste n'avait surveillé le graphique que pendant 10 000 itérations, il aurait pu penser que la convergence était atteinte. Donc, il convient d'être extrêmement prudent lorsque l'on effectue le diagnostic de la convergence des méthodes de Monte Carlo appliquées à des modèles bayesiens. En particulier, il est bien connu que les composantes de la variance posent des problèmes. Pour des travaux récents sur les paramètres de variance dans les modèles bayesiens, consulter notamment Gelman (2005).



5. MÉTHODES D'IMPUTATION

L'imputation est une technique statistique permettant de combler les données manquantes utilisée fréquemment pour traiter la non-réponse partielle lors des enquêtes par sondage. Elle a de nombreux avantages, mais également des inconvénients. Une grande variété de méthodes d'imputation ont été mises au point, mais elles se regroupent en deux classes générales ayant des propriétés distinctes. L'imputation déterministe, comme l'imputation par la moyenne et certaines formes d'imputation par la méthode du plus proche voisin, ne comportent pas de composante aléatoire. L'imputation déterministe peut être sans biais par rapport à la distribution des réponses, mais atténue la structure de covariance des données. Il ne s'agit pas nécessairement d'un problème dans le cas de certaines formes d'analyse très simplifiée, mais peut l'être dans le cas de l'analyse longitudinale, pour laquelle il est habituellement important de préserver les relations entre les variables et la structure de covariance.

Une deuxième classe de méthodes d'imputation comporte une composante aléatoire, comme l'imputation hot deck ou certaines formes d'imputation par le plus proche voisin. Ce type de procédures d'imputation peut être sans biais par rapport à la distribution des réponses et maintenir la covariance sous certaines hypothèses. Toutefois, il présente

un inconvénient important que les statisticiens ont du mal à surmonter. En ce qui concerne les formes d'imputation comportant une composante aléatoire, les méthodes standard d'estimation de la variance qui traitent les données imputées comme s'il s'agissait de données observées produisent une sous-estimation de la variance.

Afin de contourner ce problème de sous-estimation de la variance totale, deux catégories de méthodes ont été proposées. D'une part, des ajustements dans les répliques, soit les répliques jackknife ou les répliques BRR (Balanced Repeated Replication), qui ont été proposées par Rao et Shao (1992) et par Shao (1996) pour les moyennes et les totaux. Il n'est pas certain que ces méthodes puissent s'appliquer facilement à l'analyse longitudinale de données provenant d'une enquête complexe. D'autre part, on a proposé des méthodes consistant à imputer des données plus d'une fois. La plus connue est l'imputation multiple décrite par Rubin (1987), mais Fay (1996) a proposé une méthode d'imputation fractionnée pondérée et Shao et Sitter (1996) ont préconisé d'utiliser l'imputation conjointement au bootstrap. Dans le cas de l'imputation bootstrap, l'imputation doit être répétée dans chaque réplique bootstrap, ce qui en fait une méthode informatiquement coûteuse qui pourrait ne pas être une solution satisfaisante, particulièrement dans le contexte de l'analyse longitudinale. Il reste à déterminer si l'imputation fractionnée pondérée de Fay peut être étendue à l'analyse longitudinale. Bien que la controverse au sujet de l'imputation multiple se poursuive depuis presque deux décennies, l'article de Kim, Brick, Fuller et Kalton (2006) démontre que le traitement des données d'enquête par imputation multiple biaise la variance, ce qui est dommage, car, à part un estimateur de la variance présentant un biais, l'imputation multiple offre de nombreuses propriétés souhaitables pour l'analyse longitudinale en présence de données manquantes, tout spécialement la facilité relative de son utilisation.

Il convient de souligner que le recours à l'imputation dans le contexte de l'analyse longitudinale pose un problème unique. La plupart des méthodes d'imputation ont été élaborées pour traiter la non-réponse partielle et reposent sur une prédiction de la réponse manquante fondée sur d'autres valeurs courantes disponibles pour l'unité échantillonnée pour laquelle la réponse manque. Dans le cas de l'analyse longitudinale, comme nous l'avons mentionné plus haut, les données manquantes dues à l'érosion constituent une préoccupation importante. Pour ces cas, il n'existe aucune autre réponse courante à d'autres questions en vue de prédire la valeur manquante, si bien que les techniques types ne peuvent pas nécessairement être adaptées à cette situation. Certaines études ont porté sur l'imputation en vue de remplacer des données manquantes dues à des cas d'érosion par des techniques telles que le report de la dernière observation, mais à l'heure actuelle, l'évaluation de ces techniques n'est pas prometteuse.

Bien que l'imputation soit une technique fréquemment utilisée et ayant de nombreuses caractéristiques souhaitables, dans le cas de l'analyse longitudinale de données d'enquête présentant de nombreuses observations manquantes, elle pose de multiples problèmes que l'analyste devrait prendre minutieusement en considération avant de s'engager dans une analyse avec des données imputées.

6. MODÉLISATION DES DONNÉES MANQUANTES

Les données manquantes peuvent être modélisées sous des hypothèses particulières dans le cadre du processus de modélisation. Si les données sont NMAR, il s'agit du seul moyen de traiter correctement leur absence. Little et Rubin (2002) illustrent ces techniques pour des données ne provenant pas d'enquêtes. Les exemples de modélisation des données manquantes dans le cas de données d'enquête ne sont pas fréquents, mais il en existe quelques-uns. Beckett et coll. (1993) appliquent une idée intéressante de modélisation des données manquantes au moyen de modèles markoviens. Les individus observés au cours du temps sont modélisés par régression logistique comme étant dans l'état 0, 1 ou manquant. Les modèles logistiques utilisés sont basés sur le plan de sondage. Cette étude fournit un exemple intéressant dans le cas de données catégoriques, mais les techniques ne s'appliquent pas aux données continues ni aux données de type ordinal.

7. CONCLUSION

Le présent article donne une brève revue de certaines techniques qui ont été utilisées pour résoudre le problème des données manquantes lors de l'analyse longitudinale de données provenant d'une enquête complexe. Ce sujet étant trop vaste pour être examiné en entier dans un seul article, seuls les faits saillants des techniques que l'auteur a

examinées personnellement sont présentés. Il est évident que, de façon générale, le traitement des données manquantes est un sujet très difficile. Si l'on ajoute la question des données longitudinales ainsi que des données provenant d'enquêtes complexes, les méthodes en vue de surmonter les problèmes que causent les données manquantes posent un ensemble unique de défis.

RÉFÉRENCES

Beckett, Laurel A., Brock, Dwight B., Scherr, Paul A. et de Leon, Carlos Mendes (1993), "Markov models for longitudinal data from complex samples", *ASA Proceedings of the Section on Survey Research Methods*, 921-925.

Bernardo, J.M. et Smith, A.F.M. (2000), *Bayesian Theory*, New York: Wiley.

Binder, D.A. (1983), "On the variance of asymptotically normal estimators from complex surveys." *International Statistical Review*, 51, 279-92.

Brogan, Donna (1998), "Pitfalls of using standard statistical software packages for sample survey data", http://www.rti.org/sudaan/pdf_files/brogan.pdf

Chambers, R. et Skinner, C.J. (2003), *Analysis of Survey Data*, New York: Wiley.

Cochran, W.G. (1977), *Sampling Techniques*, Wiley & Sons, New York

Fay, Robert E. (1996), "Alternative paradigms for the analysis of imputed survey data", *Journal of the American Statistical Association*, 91, 490-498.

Gelman, A. (2005) "Prior distributions for variance parameters in hierarchical models", *Bayesian Analysis*, 1, 1-19.

Ghosh, Malay et Meeden, Glen (1997), *Bayesian Methods for Finite Population Sampling*, New York: Chapman Hall.

Kim, J.K, Brick, J.M., Fuller, W.A. et Kalton, G. (2006), "On the bias of the multiple imputation variance estimator in survey sampling", *Journal of the Royal Statistical Society, Series B*, à paraître.

Korn, E.L. et Graubard, B.I. (1999), *Analysis of Health Surveys*, New York: Wiley.

Little, R.J.A. et Rubin, D.B. (2002), *Statistical Analysis With Missing Data* (2e éd.), New York: Wiley.

Pfeffermann, D. (1993), "The role of sampling weights when modeling survey data." *International Statistical Review*, 61, 317-337.

Pfeffermann, D. et Lavange, L. (1989), "Regression Models for Stratified Multi-Stage Cluster Samples", Dans *Analysis of Complex Surveys* (Skinner, C. J., Holt, D., and Smith T.M.F. eds.), chapitre 12, New York: Wiley.

Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H. et Rasbash, J. (1998), "Weighting for Unequal Selection Probabilities in Multi-Level Models", *Journal of the Royal Statistical Society, Series B*, 60, 23-40.

R Development Core Team (2005), R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rao, J. N. K. et Shao, J. (1992), "Jackknife variance estimation with survey data under hot deck imputation", *Biometrika*, 79, 811-822.

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

Shao, Jun (1996), “Resampling methods in sample surveys”, *Statistics*, 27, 203-237.

Shao, Jun et Sitter, Randy R. (1996), “Bootstrap for imputed survey data”, *Journal of the American Statistical Association*, 91, 1278-1288.

Skinner, C.J. et Holmes, D.J. (2003), “Random Effects Models for Longitudinal Survey Data”, Dans *Analysis of Survey Data* (Chambers, R. and Skinner, C.J. eds.), chapitre 14, New York: Wiley.

Skinner, C. J., Holt, D. et Smith, T.M.F (1989), *Analysis of Complex Surveys*, New York: Wiley.

Verbeke, G. et Molenberghs, G. (2000), “Linear mixed models for longitudinal data”, Springer-Verlag Inc (Berlin; New York)