Document généré le 12 oct. 2024 01:32

L'Actualité économique

Le problème des données longitudinales incomplètes : une nouvelle approche

Marie-France Paquet et Denis Bolduc

Volume 80, numéro 2-3, juin–septembre 2004

Hommage à Marcel Dagenais

URI : https://id.erudit.org/iderudit/011390ar
DOI : https://doi.org/10.7202/011390ar

Aller au sommaire du numéro

Éditeur(s)

HEC Montréal

ISSN

0001-771X (imprimé)
1710-3991 (numérique)

Découvrir la revue

Citer cet article

Paquet, M.-F. & Bolduc, D. (2004). Le problème des données longitudinales incomplètes : une nouvelle approche. *L'Actualité économique*, *80*(2-3), 341–361. https://doi.org/10.7202/011390ar

Résumé de l'article

Dans ce travail, nous suggérons l'utilisation de l'échantillonnage de Gibbs combiné à l'augmentation des données pour estimer des modèles à données longitudinales incomplètes, qui dans le cas extrême où l'échantillon est composé de coupes transversales indépendantes, correspond au cas de modèle de type pseudo-panel. Cette idée peut être appliquée dans plusieurs contextes : modèles statiques ou dynamiques de type linéaires, non linéaires, de choix discrets, avec régresseurs endogènes, *etc*. Pour présenter la méthode proposée, nous l'appliquons dans le cas d'un modèle linéaire à variable dépendante continue. Comme point de comparaison, nous utilisons les estimations par l'approche conventionnelle dite de pseudo-panel basée sur des moyennes calculées sur des cohortes. La technique proposée dans ce travail donne des résultats supérieurs, en terme d'efficacité, à la technique conventionnelle. Cette conclusion demeure valide quelle que soit la proportion des observations manquantes.

Tous droits réservés © HEC Montréal, 2004

Ce document est protégé par la loi sur le droit d'auteur. L'utilisation des services d'Érudit (y compris la reproduction) est assujettie à sa politique d'utilisation que vous pouvez consulter en ligne.

https://apropos.erudit.org/fr/usagers/politique-dutilisation/

Cet article est diffusé et préservé par Érudit.

Érudit est un consortium interuniversitaire sans but lucratif composé de l'Université de Montréal, l'Université Laval et l'Université du Québec à Montréal. Il a pour mission la promotion et la valorisation de la recherche.

https://www.erudit.org/fr/

L'Actualité économique, Revue d'analyse économique, vol. 80, n°s 2-3, juin-septembre 2004

LE PROBLÈME DES DONNÉES LONGITUDINALES INCOMPLÈTES : UNE NOUVELLE APPROCHE*

Marie-France PAQUET
Université d'Ottawa

Denis BOLDUC
Université Laval

RÉSUMÉ – Dans ce travail, nous suggérons l'utilisation de l'échantillonnage de Gibbs combiné à l'augmentation des données pour estimer des modèles à données longitudinales incomplètes, qui dans le cas extrême où l'échantillon est composé de coupes transversales indépendantes, correspond au cas de modèle de type pseudo-panel. Cette idée peut être appliquée dans plusieurs contextes : modèles statiques ou dynamiques de type linéaires, non linéaires, de choix discrets, avec régresseurs endogènes, *etc*. Pour présenter la méthode proposée, nous l'appliquons dans le cas d'un modèle linéaire à variable dépendante continue. Comme point de comparaison, nous utilisons les estimations par l'approche conventionnelle dite de pseudo-panel basée sur des moyennes calculées sur des cohortes. La technique proposée dans ce travail donne des résultats supérieurs, en terme d'efficacité, à la technique conventionnelle. Cette conclusion demeure valide quelle que soit la proportion des observations manquantes.

ABSTRACT – In this paper, we suggest to use a Gibbs sampler with data augmentation to estimate models based on incomplete longitudinal data which, in the extreme case where the sample is composed of independent cross-sections, corresponds to a situation that normally calls for pseudo-panel modeling. The idea suggested here can be applied in several contexts: static and dynamic models linear or nonlinear type, discrete choice models, models with endogenous regressors, *etc*. To present the suggested method, we apply it to a linear model with continuous dependent variable. For comparison purpose, we also use the conventional pseudo-panel approach which is based on averages computed on cohorts. In terms of efficiency, the technique suggested in this work gives better results than the conventional pseudo-panel technique. This conclusion remains valid for any proportion of missing observations in the sample.

* Cette recherche a bénéficié du soutien financier du FQRSC (FCAR) et du CRSH. Nous tenons à remercier Stephen Gordon pour ses multiples conseils judicieux.

« Cette idée de travailler sur la théorie de la régression lorsqu'on dispose d'observations incomplètes, ça m'est venu effectivement d'un problème extrêmement pratique. (...) Il y avait des observations manquantes, et on avait l'impression que si on laissait tomber les questionnaires incomplets, on laissait tomber à peu près la moitié de l'échantillon, qui contenait pourtant beaucoup d'information. »

Marcel Dagenais

Introduction

Les données de panel, ou données longitudinales complètes, sont généralement cumulées à partir d'enquêtes répétées à travers le temps sur un même échantillon d'unités de base comme des individus, des ménages ou encore des entreprises. Ces données sont très utiles pour étudier la dynamique intertemporelle des comportements individuels. L'avantage principal des panels résulte du caractère désagrégé des observations et de la grande richesse d'information qui en découle.

Dans le cas où l'information est incomplète à travers le temps, et selon l'importance de l'information manquante, nous sommes en présence de panels dits incomplets ou non balancés. (Voir Baltagi, 1995a; Hirano, Imbens, Ridder et Rubin, 1998). Pour analyser l'évolution de la consommation des individus dans le temps, nous pouvons très bien nous retrouver avec une situation où l'information est complète pour la majeure partie des individus de l'échantillon, c.-à-d. l'information pour ces individus est disponible pour toutes les périodes de l'enquête, alors que pour un sous-groupe d'individus, les données sont absentes pour diverses raisons.

Nous pouvons imaginer le cas extrême, mais très répandu en pratique, où les informations sur chacune des unités de la dimension transversale ne sont présentes qu'à une seule période. Dans ce cas, nous sommes en présence de séries temporelles composées de coupes transversales indépendantes. Plusieurs auteurs ont traité ces cas limites, mais le premier chercheur qui a formalisé la méthodologie appropriée pour les traiter est Deaton (1985). Beaucoup d'autres études ont utilisé l'approche développée par Deaton, notamment Browning, Deaton et Irish (1985), Moffit (1993), Gardes, Langlois et Richeaudeau (1995), Alessie, Devereux et Weber (1997), Gardes et Loisy (1997), et Beaudry et Green (2000).

La solution proposée par Deaton est de créer des panels (au sens légitime du terme) à partir de moyennes prises sur des groupes d'unités classifiées selon des critères assurant une certaine homogénéité. Ces moyennes sur les informations des groupes d'unités, calculées à chaque période, constituent ce que nous appelons des pseudo-panels. Cette façon de faire permet de retrouver certains des avantages attribués aux panels, comme la possibilité de modéliser les effets dynamiques, tout en évitant certains des inconvénients qui leur sont propres. En particulier, les pseudo-panels permettent l'étude de comportements dynamiques sans avoir à être confronté aux problèmes d'attrition ou d'apprentissage. Nous qualifions cette technique de *conventionnelle* pour faire ressortir le fait que c'est l'approche la plus souvent utilisée pour traiter ces données et parce que les chercheurs font référence à cette technique lorsqu'ils parlent de pseudo-panels.

Les pseudo-panels sont générés en exploitant des séries de coupes tranversales provenant d'enquêtes indépendantes accumulées au fil des ans. De façon à obtenir un ensemble de données utiles, certaines règles doivent être respectées lors de la construction du pseudo-panel. Parmi ces règles, mentionnons la nécessité de conduire les enquêtes à partir d'une même population en utilisant la même méthodologie d'échantillonnage.

Une fois que les différentes enquêtes sont réunies, les unités, individus ou entreprises, qui ont des caractéristiques communes sont regroupées en cohortes de façon à ce que chaque unité n'appartienne qu'à une seule cohorte. De plus, cette appartenance à une cohorte doit être invariante dans le temps. La moyenne des unités à l'intérieur d'une cohorte est ensuite interprétée comme étant une observation du pseudo-panel. Un pseudo-panel peut donc être vu comme le résultat d'un ensemble de données longitudinales non balancées où chaque individu n'est présent qu'à une seule période dans les différentes enquêtes.

Nous proposons une méthodologie alternative à l'approche conventionnelle des pseudo-panels qui utilise, entre autre, la technique à augmentation de données (Tanner et Wong, 1987) afin de contourner le problème des données longitudinales incomplètes. La méthode résulte en des gains importants au plan de l'efficacité des estimateurs des paramètres des modèles. Nous considérons à la fois les panels partiellement incomplets et les combinaisons de coupes transversales indépendantes. Dans la première partie, nous présentons la problématique. Par la suite, nous expliquons les étapes spécifiques touchant l'estimation des paramètres. Nous terminons par une démonstration de la méthodologie proposée qui repose sur des données simulées.

1. Le problème

L'objectif de base vise à contourner les problèmes économétriques qui découlent généralement du phénomène de données manquantes. Bien que l'approche développée s'applique tout aussi bien dans des situations de modèles linéaires simples que dans le cas de modèles de choix discrets dynamiques pouvant même comporter des régresseurs endogènes, pour des fins pédagogiques, nous focalisons dans le présent texte sur le modèle de régression linéaire avec variable dépendante continue. Dans un premier temps, nous présentons la notation du modèle linéaire utlisé. Le modèle comprend des effets individuels aléatoires. Ces effets sont souvent nécessaires dans les études empiriques car ils permettent de tenir compte de l'hétérogénéité individuelle non observable. Nous expliquons dans un premier temps la technique conventionnelle et les estimateurs qui s'y rattachent, ainsi que la construction des cohortes utilisées dans notre application empirique. Finalement, nous présentons la technique que nous proposons, laquelle exploite l'échantillonage de Gibbs couplé avec le principe d'augmentation de données.

1.1 *Le modèle économétrique*

Le modèle utilisé dans ce travail pour illustrer la technique proposée pour l'analyse des pseudo-panels est le suivant :

$$y_{nt} = x_{nt}\,\beta + u_{nt}, \tag{1}$$
$$u_{nt} = \theta_n + \nu_{nt},$$
$$n = 1, 2, ..., N \quad \text{et} \quad t = 1, 2, ..., T.$$

Dans cette version du modèle, sans perte de généralité, x_{nt} représente une seule variable explicative. La structure des erreurs est la suivante :

$$\theta_n \sim N(0, \sigma_\theta^2), \tag{2}$$
$$\nu_{nt} \sim N(0, \sigma_\nu^2)$$

où θ_n est un effet aléatoire invariant dans le temps et ν_{nt} représente le terme d'erreur résiduel. Les termes d'erreurs θ_n et ν_{nt} sont considérés comme étant indépendants entre eux. De plus, x_{nt} est postulé comme étant indépendant de θ_n et de ν_{nt} pour tout n et tout t. Les modèles avec effets aléatoires sont appropriés lorsque qu'il est raisonnable de considérer que les individus de l'échantillon sont tirés aléatoirement d'une large population.

Sous les hypothèses faites, la matrice de variances-covariances de l'erreur composée u_{nt} est définie par :

$$\text{var}(u_{nt}) = \sigma_\theta^2 + \sigma_\nu^2 \qquad \text{pour tout } n \text{ et tout } t \tag{3}$$
$$\text{cov}(u_{nt}, u_{ms}) = \begin{cases} \sigma_\theta^2 + \sigma_\nu^2 & \text{pour } n = m,\ t = s, \\ \sigma_\theta^2 & \text{pour } n = m,\ t\ \pi\ s, \\ 0 & \text{autrement.} \end{cases}$$

Pour assurer un bon contrôle dans le cadre de notre application empirique, nous exploitons des données simulées générées selon un modèle bien précis. Pour simuler les variables explicatives, nous utilisons le modèle de régression auxiliaire suivant :

$$x_{nt} = z_{nt}\,\gamma + \varepsilon_{nt} \tag{4}$$

avec $\varepsilon_{nt} \sim N(0, \tau^2)$.

1.2 *Approche à pseudo-panels*

La solution au problème de la non-disponibilité des données de panel, proposée par Deaton (1985), est de créer des panels (au sens légitime du terme) à partir de moyennes prises sur des groupes d'unités classifiées selon des critères assurant une certaine homogénéité. Ces moyennes sur les informations des groupes d'unités, calculées à chaque période, constituent les pseudo-panels. Cette façon de procéder permet de retrouver certains des avantages attribués aux panels tout en évitant certains des inconvénients qui leur sont propres.

Plusieurs travaux (Heckman et Robbs, 1985; Deaton, 1985; Moffit, 1993) ont suggéré qu'il n'était pas indispensable d'avoir de vrais panels pour arriver à identifier et à estimer les effets de dynamique des modèles traditionnels.

L'idée proposée par Deaton consiste à regrouper ensemble les individus ayant des caractéristiques communes de façon à former des cohortes. Une fois les observations groupées, ce sont les moyennes des différentes variables explicatives qui font office d'unités et qui constituent le pseudo-panel. De façon à bien comprendre la procédure de formation des pseudo-panels, exploitons le modèle linéaire présenté à l'équation (1). Avec un ensemble de coupes transversales indépendantes, il faut réaliser que les N individus présents à chaque période diffèrent. Maintenant, définissons un ensemble de C cohortes dont chacune d'elle possède des critères d'appartenance invariables dans le temps. L'année de naissance et le sexe de l'individu constituent de bon critères pour former les cellules. Chaque individu n ne peut appartenir qu'à une seule cohorte c. Lorsque nous agrégeons les individus en cohortes et que nous calculons les moyennes intracohortes, cela donne :

$$\begin{aligned} \bar{y}_{ct} &= \bar{x}_{ct}\beta + \bar{u}_{ct}, \\ \bar{u}_{ct} &= \bar{\theta}_{ct} + \bar{\nu}_{ct}, \\ c &= 1, 2, ..., C \quad \text{et} \quad t = 1, 2, ..., T \end{aligned} \tag{5}$$

où $\bar{y}_{ct}$ représente la moyenne de tous les y_{nt} qui font partie de la cohorte c à la période t et il en va de même pour les autres variables du modèle. Il est important de constater que l'effet spécifique aux cohortes devient maintenant variable dans le temps. Ceci est dû au fait que nous n'observons pas les mêmes individus d'une période à l'autre et, par conséquent, la moyenne devient dépendante de l'indice temporel. Vient s'ajouter à cela le fait que, la plupart du temps, $\bar{\theta}_{ct}$ est corrélé avec x_{nt}. Le choix du type d'effets considérés, à savoir fixes ou aléatoires, amène aussi son lot de problèmes. Alors que les effets fixes nous conduisent au problème d'identification, les effets aléatoires peuvent mener à des estimations non convergentes, si la corrélation potentielle entre $\bar{\theta}_{ct}$ et x_{nt} n'est pas prise en compte. Pour ces raisons, ce dernier modèle n'est pas celui qui sera utilisé afin d'obtenir des estimateurs convergents, à moins que la taille des cohortes permette de dire que $\bar{\theta}_{ct}$ est une bonne approximation de $\bar{\theta}_c$. Une façon de s'assurer que $\bar{\theta}_{ct} = \bar{\theta}_c$ est d'avoir un grand nombre d'observations dans chacune des cohortes (Verbeek et Nijman, 1992).

Dans le cas où le nombre d'observations dans chaque cohorte n'est pas très grand, nous pouvons utiliser la version du modèle avec la vraie population des cohortes dans le but d'avoir des estimateurs convergents, soit :

$$y^*_{ct} = x^*_{ct}\beta + \theta_c + \nu^*_{ct} \tag{6}$$

avec y^*_{ct} et x^*_{ct} représentant les moyennes non observables de la population des cohortes et θ_c l'effet spécifique aux cohortes. Comme la vraie population d'une cohorte est la même dans le temps, θ_c est invariant dans le temps. Deaton propose l'utilisation d'un estimateur corrigé pour tenir compte d'erreurs sur les variables. Ces erreurs proviennent du fait que les moyennes intracohortes du modèle (5) sont en réalité des mesures imparfaites de y^*_{ct} et x^*_{ct}. Ici, toutes les variables sont estimées avec erreurs de mesures à l'exception de θ_c.

Cependant, dans la pratique, le problème d'erreurs de mesure sur les variables est ignoré dès que le nombre d'observations à l'intérieur de chacune des cohortes est élevé (voir Browning *et al.*, 1985; Blundell *et al.*, 1990; Moffit, 1993). Dans ce cas, l'estimateur le plus souvent utilisé est le suivant :

$$\hat{\beta}_W = \left(\frac{1}{CT} \sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)' (\bar{x}_{ct} - \bar{x}_c) \right)^{-1} \left(\frac{1}{CT} \sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)' (\bar{y}_{ct} - \bar{y}_c) \right) \quad (7)$$

où $\bar{x}_c = \frac{1}{T} \sum_{t=1}^{T} \bar{x}_{ct}$ et $\bar{y}_c = \frac{1}{T} \sum_{t=1}^{T} \bar{y}_{ct}$. Cet estimateur est connu sous le nom de « *within* » parce qu'il utilise l'information qui provient de la variabilité entre les observations de la dimension individuelle contenue dans une cohorte en particulier. Dans le cas où les effets sont fixes, l'estimateur de θ_c est $\hat{\theta}_c = \bar{y}_{ct} - \bar{x}_{ct} \hat{\beta}_W$. Nous pouvons aussi définir un estimateur qui privilégiera l'information qui provient de la variabilité entre les cohortes. C'est l'estimateur « *between* » qui s'écrit :

$$\hat{\beta}_B = \left(\frac{1}{CT} \sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_c' \bar{x}_c) \right)^{-1} \times \left(\frac{1}{CT} \sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_c' \bar{y}_c) \right). \quad (8)$$

Afin que ces estimateurs soient convergents, Verbeek et Nijman (1992) montrent que la taille des cohortes doit tendre vers l'infini et que la convergence de l'estimateur « *within* » sera assurée si les conditions suivantes sont respectées :

$$\operatorname*{plim}_{n_c \to \infty} \frac{1}{CT} \sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c) \bar{\nu}_{ct} = 0 \quad (9)$$

et $\operatorname*{plim}_{n_c \to \infty} \frac{1}{CT} \sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c) \bar{\theta}_{ct} = 0$.

Par ailleurs, ils montrent pour certains modèles assez simples que le problème d'erreurs de mesure devient négligeable lorsque la taille des cellules est assez grande (quelques centaines d'individus par cellule).

1.3 *Les cohortes dans la présente application*

La construction même des cohortes doit respecter certaines règles bien précises. Par exemple, les cellules doivent être définies sur la base d'une variable distribuée de façon continue que nous appelons *s*. Cette variable doit être distribuée de façon indépendante entre les individus et doit posséder une variance normalisée à l'unité. De plus, les cohortes sont construites de telle sorte que la probabilité non conditionnelle d'appartenir à une cohorte donnée est la même pour tous les individus.

Dans la pratique, la variable *s* est généralement basée sur plus d'une variable sous-jacente. Cependant, il faut bien comprendre que les critères d'association des cellules excluent certaines variables de choix. Par exemple, une variable qui ne serait pas constante pour chaque individu à chaque période ne pourrait être utilisée car cela impliquerait que certains individus perdraient leur droit d'appartenance à une cellule bien définie. D'autre part, la variable de choix se doit d'être observable pour tous les individus de l'échantillon. Ceci implique que des variables tels le revenu du chef de famille ou la taille de la famille ne sont pas de bons candidats à la construction de la variable *s* permettant de définir les cohortes. Pour plus de détails, consulter Verbeek et Nijman (1992).

Cette façon de procéder permet de conserver exactement le même nombre de séries temporelles mais a le grand désavantage de réduire considérablement le nombre d'observations de chaque tranche transversale présente dans le pseudo-panel. Par ailleurs, la définition des cohortes joue aussi un rôle important dans l'efficacité des estimateurs : la détermination même des cohortes, c.-à-d. la taille et le nombre, a un impact important sur la taille du biais et de la variance des estimateurs dans un contexte d'échantillons finis (Verbeek et Nijman, 1993).

Les observations sont regroupées en cellules à l'aide de la variable *s* qui est invariante dans le temps et qui est observable pour toutes les personnes à toutes les périodes. Pour fins d'illustration, nous n'avons utilisé qu'une seule variable pour construire les cellules. Afin de pouvoir appliquer la méthodologie des cohortes telle que définie précédemment, l'ensemble complet est divisé en 6 cohortes sur la base de la distribution d'une variable z_{nt} appartenant à la régression auxiliaire (4).

Dans cet ensemble d'expériences, nous nous plaçons dans un contexte à deux périodes où, pour chaque expérience, le nombre d'observations manquantes est approximativement de 5 %, 10 % et de 20 % . Expliquons le principe retenu pour l'échantillon à 5 %. Pour le créer, nous avons enlevé de façon aléatoire 5 % des 5 000 observations de chacune des périodes, en nous assurant par contre qu'une observation ne peut être manquante dans les deux périodes. Le même principe est appliqué pour générer les échantillons à 10 % et à 20 % d'observations manquantes. Les décomptes du nombre résultant d'observations par cohortes sont donnés aux tableaux 1 et 2. Le tableau 1 concerne la première période alors que le tableau 2 concerne la deuxième période.

TABLEAU 1

NOMBRE DE COHORTES = 6, t = 1

Numéro de cohorte	NOMBRE D'INDIVIDUS			
	Échantillon complet	**Troncature 5 %**	**Troncature 10 %**	**Troncature 20 %**
1	783	723	710	635
2	759	675	676	617
3	950	864	875	798
4	986	864	892	802
5	744	705	671	598
6	703	699	703	642
Total	5 000	4 530	4 527	4 092

TABLEAU 2

NOMBRE DE COHORTES = 6, t = 2

Numéro de cohorte	NOMBRE D'INDIVIDUS			
	Échantillon complet	**Troncature 5 %**	**Troncature 10 %**	**Troncature 20 %**
1	798	687	742	695
2	729	691	681	622
3	952	884	869	815
4	979	873	882	836
5	785	698	711	665
6	757	734	680	635
Total	5 000	4 567	4 565	4 268

Par la suite, le même ensemble de données est divisé en 8 et 12 cohortes et la même méthodologie est appliquée pour l'estimation des paramètres. Le nombre d'observations pour ces deux autres structures d'ensembles de données est présenté à l'annexe A.

1.4 *Échantillonnage de Gibbs et augmentation de données*

L'approche bayésienne repose sur la distribution *a posteriori* :

$$p(\theta \mid Y) = \frac{p(Y \mid \theta)\, p(\theta)}{\int_\theta p(Y \mid \theta)\, p(\theta)\, d\theta} \tag{10}$$

$$\propto L(\theta \mid Y)\, p(\theta)$$

où θ représente le vecteur des paramètres. $L(\theta \mid Y)$ qui constitue une réécriture de $p(Y \mid \theta)$ représente la fonction de vraisemblance et $p(\theta)$ la densité qui correspond à la distribution *a priori* de θ. L'échantillon observable est constitué des variables y_{nt} et x_{nt} auquel nous ajoutons les observations qui étaient manquantes et qui sont indexées d'un m. Par conséquent, l'échantillon complet s'écrit $Y = \{Y^{ob}, Y^*\}$ où $Y^* = \{y_{mt}, x_{mt}\}$ et $Y^{ob} = \{y_{nt}, x_{nt}\}$.

La densité conditionnelle qui nous intéresse s'obtient en intégrant par rapport aux données manquantes :

$$p(Y^{ob} \mid \theta) = \int_{Y^*} p(Y^{ob}, Y^* \mid \theta)\, d\, Y^*. \tag{11}$$

Il est toutefois fréquent que la forme de cette densité ne soit pas connue. Par ailleurs, l'évaluation de la fonction de vraisemblance, à cause de la présence des observations manquantes, peut nécessiter un temps de calcul considérable et ce, malgré la performance croissante des ordinateurs.

Pour contourner ces différents problèmes, nous utilisons la technique dite d'augmentation de données. Cette technique permet de traiter tous les éléments non observables du modèle comme s'ils étaient des paramètres à estimer. L'idée est la suivante : si nous connaissons Y^*, nous pouvons calibrer une valeur raisonnable pour le vecteur des paramètres θ; si nous connaissons θ, nous pouvons estimer les données manquantes, c.-à-d. Y^*. Pour arriver à mettre ce principe en application, il est essentiel de pouvoir tirer des données des distributions suivantes :

$$p(\theta \mid Y^*, Y^{ob}) \tag{12}$$

et $p(Y^* \mid Y^{ob}, \theta)$.

Dans la plupart des cas, effectuer des tirages à partir des distributions conditionnelles ne pose pas de problème grâce à la technique d'échantillonnage de Gibbs. Pour plus de détails sur l'échantillonnage de Gibbs, consulter Casella et George (1992), Gordon et Bélanger (1996) ou Paquet (2002).

2. Estimation économétrique

L'approche proposée est divisée en deux étapes. Dans la première, nous utilisons la technique d'augmentation de données afin de simuler les données manquantes. Dans la seconde, nous simulons le vecteur des paramètres étant donné les données augmentées.

2.1 *L'étape de l'augmentation des données*

Étant donné les hypothèses faites sur le modèle, cette étape s'effectue comme suit :

- Étape 1 : simulation de y_{mt}

Étant donné x_{mt}, où m dénote une observation individuelle manquante, et les paramètres du modèle, nous remarquons à partir de l'équation (1) qu'il est possible de simuler y_{mt} comme suit :

$$y_{mt} = \beta\, x_{mt} + u_{mt} \tag{13}$$

et $u_{mt} = \theta_m + \nu_{mt}$.

Les valeurs pour y_{mt} sont simulées à partir d'une distribution normale

$$Y_{[m]} \sim NMV(\beta\, x_{mt}, \Omega_{[m]}) \tag{14}$$

où $Y_{[m]}$ est un vecteur de tous les y_{mt} et $\Omega_{[m]}$ est la sous-matrice de Ω qui correspond uniquement aux observations manquantes. La matrice Ω est définie à l'équation (24) de l'annexe B.

- Étape 2 : simulation de x_{mt}

Maintenant, nous devons simuler les données manquantes de la variable x_{mt}. Pour ce faire, nous utilisons la régression auxiliaire présentée à l'équation (4). La distribution conditionnelle d'intérêt est obtenue à partir des hypothèses faites sur le modèle :

$$p(Y_{[m]} \mid X_{[m]}, \theta_{[m]}) \propto \exp\left(-\frac{1}{2}(Y_{[m]} - X_{[m]}\beta)'\, \Omega_{[m]}^{-1}(Y_{[m]} - X_{[m]}\beta)\right) \tag{15}$$

et $$p(X_{[m]} \mid Z_{[m]}, \theta_{[m]}) \propto \exp\left(-\frac{1}{2}(X_{[m]} - Z_{[m]}\gamma)'\, V_{[m]}^{-1}(X_{[m]} - Z_{[m]}\gamma)\right) . \tag{16}$$

En combinant (15) et (16) nous pouvons isoler la contribution de $X_{[m]}$, ce qui donne

$$p(X_{[m]} \mid Y_{[m]}, Z_{[m]}, \theta_{[m]}) = \frac{p(X_{[m]}, Y_{[m]} \mid Z_{[m]})}{p(Y_{[m]})} , \tag{17}$$

$$p(X_{[m]} \mid Y_{[m]}, Z_{[m]}, \theta_{[m]}) \propto p(X_{[m]}, Y_{[m]} \mid Z_{[m]})$$

et $$p(X_{[m]} \mid Y_{[m]}, Z_{[m]}, \theta_{[m]}) \propto p(Y_{[m]} \mid X_{[m]})\, P(X_{[m]} \mid Z_{[m]}) .$$

L'annexe B présente tous les calculs associés à cette distribution conditionnelle, dont le résultat dans le cas d'une seule variable x_{mt} s'écrit :

$$p(x_{mt} \mid y_{mt}, z_{mt}, \theta) \propto N\left(-\frac{\left[\frac{2}{\sigma_v^2}(-\beta\, y_{mt} + \theta\beta) + \frac{2}{\tau^2}(-2\gamma x_{mt} z_{mt})\right]}{2\left(\frac{\beta^2}{\sigma_v^2} + \frac{1}{\tau^2}\right)}, \left(\frac{\beta^2}{\sigma_v^2} + \frac{1}{\tau^2}\right)^{-1} \right) . \quad (18)$$

Étant donné les autres éléments du modèle, ceci nous fournit la possibilité de simuler les valeurs pour tous les x_{mt}.

2.2 *L'étape de la simulation du vecteur des paramètres*

Une fois que les tirages pour les données manquantes sont obtenus, l'estimation du vecteur des paramètres découle de l'application de résultats bien connus en économétrie bayésienne touchant le modèle de régression linéaire.

- Étape 3 : simulation de β

Si les croyances *a priori* sur β sont décrites comme suit $p(\beta) = N(\bar{\beta}, \bar{A}^{-1})$ et si σ_u^2 est connu, alors le modèle devient une simple régression avec variance connue. Dans ce cas, nous pouvons montrer que

$$p(\beta \mid X, Y) = N(\hat{\beta}, \hat{A}^{-1}) \quad (19)$$

où $\hat{A} = (X'\,X / \sigma_u^2) + \bar{A}$ et $\hat{\beta} = \hat{A}^{-1}\left((X'\,Y / \sigma_u^2) + \bar{A}\,\bar{\beta}\right)$.

Par conséquent, simuler β implique le tirage d'une valeur à partir d'une distribution normale multivariée.

- Étape 4 : simulation de σ_u^2

Étant donné β et les données, il est possible de récupérer directement le vecteur d'erreur $u = Y - X\,\beta$. Si les croyances *a priori* sur σ_u^2 sont décrites par $p(\sigma_u^{-2}) = G(\bar{a}, \bar{b})$, alors les valeurs de σ_u^2 sont simulées à partir de la distribution *a posteriori*

$$\begin{aligned} p(\sigma_u^{-2} \mid \beta, X, Y) &= G\left(\bar{a} + \frac{N}{2}, \bar{b} + \frac{u'u}{2}\right) \quad (20) \\ &= G(\tilde{a}, \tilde{b}) \end{aligned}$$

qui représente une distribution Gamma avec moyenne $\tilde{a}\,\tilde{b}$ et variance $\tilde{a}\,\tilde{b}^2$ connues.

- Étape 5 : simulation de γ

Étant donné Z_{nt} et τ^2, la régression auxiliaire présentée à l'équation (4) est encore une fois un modèle de régression avec variance connue. Lorsque les croyances *a priori* pour γ sont représentées par $p(\gamma) = N(\bar{\gamma}, \bar{C}^{-1})$, alors

$$p(\gamma \mid \tau, Z) = N(\hat{\gamma}, \hat{C}^{-1}) \tag{21}$$

où $\hat{C} = (Z' Z / \tau^2) + \bar{C}$ et $\hat{\gamma} = \hat{C}^{-1}\left((Z' X / \tau^2) + \bar{C}\,\bar{\gamma}\right)$.

- Étape 6 : simulation de τ^2

Tout comme dans le cas de la simulation de σ_v^2, étant donné X, Z et γ, nous pouvons récupérer $\in = X - Z\gamma$. Si les croyances *a priori* sur τ^2 sont représentées par une loi gamma inverse $p(\tau^{-2}) = G(\bar{c}, \bar{d})$, alors τ^{-2} est simulé à partir de la distribution gamma suivante :

$$\begin{aligned} p(\tau^{-2} \mid \gamma, X, Z) &= G\left(\bar{c} + \frac{N}{2}, \bar{d} + \frac{\in' \in}{2}\right) \\ &= G(\tilde{c}, \tilde{d}). \end{aligned} \tag{22}$$

Encore une fois, les deux premiers moments de la distribution sont connus, soit la moyenne $\tilde{c}\,\tilde{d}$ et la variance $\tilde{c}\,\tilde{d}^{\,2}$.

3. Les résultats d'estimation

Pour comparer les deux techniques d'estimation, le point de référence est l'estimation par échantillonnage de Gibbs sans augmentation de données obtenue à partir de l'ensemble complet de données de panel. Dans ce cas, la banque de données simulées est composée de 5 000 observations pour chacune des deux périodes. À la lumière du tableau 3, il est possible de retirer certaines constatations quant aux performances respectives des deux techniques d'estimation, soit la méthodologie à cohortes représentatives (la méthode à pseudo-panels) et l'échantillonnage de Gibbs avec augmentation de données.

Les résultats obtenus sont très bons et la petitesse des écart-types suggère que les paramètres sont estimés avec un haut degré de précision. Pour la méthodologie des cohortes, les résultats sont affectés par le nombre de cellules utilisées dans la création des différents échantillons. Les résultats basés sur 6, 8 ou 12 cellules sont présentés au bas du tableau 3. Au fur et à mesure que le nombre de cellules augmente, les résultats des estimations *between* s'améliorent mais la précision des estimations demeure bien en deçà de celle obtenue à l'aide de l'échantillonnage de Gibbs. De la même façon, les résultats des estimations *within* deviennent moins précis lorsque le nombre de cellules augmente car cela implique que moins d'observations sont utilisées dans chacune d'elles. Étant donné que l'estimateur *within* utilise la variation à l'intérieur de la cellule, ce résultat n'est pas du tout surprenant.

TABLEAU 3

RÉSULTATS D'ESTIMATION EN PANEL (N = 5 000, t = 2)

			Paramètre	β	σ^2_ν	σ^2_θ
			Vraie valeur	1	0,36	0,25
Gibbs (5 000 tirages)			Est.	1,028	0,334	0,243
			É-T	0,002	0,073	0,012
Approche conventionnelle	6 cohortes	β_{within}	Est.	1,023		0,356
			É-T	0,803		
		$\beta_{between}$	Est.	1,903		0,00031
			É-T	13,336		
	8 cohortes	β_{within}	Est.	0,852		0,419
			É-T	3,482		
		$\beta_{between}$	Est.	1,903		0,000961
			É-T	17,761		
	12 cohortes	β_{within}	Est.	0,841		0,566
			É-T	1,838		
		$\beta_{between}$	Est.	1,904		0,000961
			É-T	11,448		

Pour l'échantillonnage de Gibbs avec augmentation de données, c.-à-d. dans le contexte où nous procédons à une troncature des échantillons, les résultats demeurent très intéressants. Cependant, lorsque le nombre de données manquantes augmente, c.-à-d. lorsque nous passons d'une troncature de 5 %, à 10 % puis à 20 %, les résultats perdent quelque peu de leur précision (voir les tableaux 4, 5 et 6). Par contre, pour la méthodologie des cohortes, le nombre de cellules formées à l'aide de l'échantillon de départ ne semble plus être le facteur critique. C'est plutôt le niveau de troncature de l'échantillon qui devient le facteur clé.

TABLEAU 4

RÉSULTATS D'ESTIMATION EN PSEUDO-PANEL TRONQUÉ À 5 %

			Paramètre	β	σ_ν^2	σ_θ^2
			Vraie valeur	1	0,36	0,25
Gibbs (5 000 tirages)			Est.	1,016	0,311	0,281
			É-T	0,00169	0,071	0,0115
Approche conventionnelle	6 cohortes	β_{within}	Est.	1,192		0,234
			É-T	1,621		
		$\beta_{between}$	Est.	1,903		
			É-T	13,344		
	8 cohortes	β_{within}	Est.	0,855		0,419
			É-T	1,466		
		$\beta_{between}$	Est.	1,903		0,000872
			É-T	12,292		
	12 cohortes	β_{within}	Est.	0,726		0,697
			É-T	3,419		
		$\beta_{between}$	Est.	1,904		0,000872
			É-T	11,445		

TABLEAU 5

RÉSULTATS D'ESTIMATION EN PSEUDO-PANEL TRONQUÉ À 10 %

			Paramètre	β	σ_ν^2	σ_θ^2
			Vraie valeur	1	0,36	0,25
Gibbs (5 000 tirages)			Est.	1,00491	0,284	0,312
			É-T	0,00207	0,0701	0,0118
Approche conventionnelle	6 cohortes	β_{within}	Est.	1,110		0,288
			É-T	1,00462		
		$\beta_{between}$	Est.	1,897		0,000442
			É-T	13,166		
	8 cohortes	β_{within}	Est.	0,834		0,433
			É-T	1,677		
		$\beta_{between}$	Est.	1,894		0,00118
			É-T	12,0734		
	12 cohortes	β_{within}	Est.	0,761		0,652
			É-T	2,852		
		$\beta_{between}$	Est.	1,901		0,00118
			É-T	11,373		

TABLEAU 6

RÉSULTATS D'ESTIMATION EN PSEUDO-PANEL TRONQUÉ À 20 %

			Paramètre	β	σ_ν^2	σ_θ^2
			Vraie valeur	1	0,36	0,25
Gibbs (5 000 tirages)			Est.	0,761	0,408	0,440
			É-T	0,0302	0,0672	0,0189
Approche conventionnelle	6 cohortes	β_{within}	Est.	1,121		0,281
			É-T	1,049		
		$\beta_{between}$	Est.	1,911		0,000365
			É-T	13,542		
	8 cohortes	β_{within}	Est.	0,981		0,319
			É-T	0,694		
		$\beta_{between}$	Est.	1,909		0,00111
			É-T	12,429		
	12 cohortes	β_{within}	Est.	0,782		0,624
			É-T	2,556		
		$\beta_{between}$	Est.	1,914		0,00111
			É-T	11,676		

CONCLUSION

L'objectif de ce travail a été de proposer une approche alternative à la technique conventionnelle des pseudo-panels en utilisant les techniques à augmentation de données (Tanner et Wong, 1987). L'approche suggérée utilise au maximum toute l'information disponible au lieu de réduire l'information des unités à un niveau de définition qui correspond à des moyennes de groupes. Elle est d'ailleurs si flexible qu'elle peut être utilisée pour traiter tant le cas extrême soulevé par Deaton, soit la combinaison de coupes transversales indépendantes, que les situations de panels incomplets proprement dits.

Pour fins de démonstration, la méthodologie suggérée est appliquée dans le contexte d'un modèle linéaire à erreurs composées et elle est comparée à la technique conventionnelle des pseudo-panels. L'estimation est produite à partir de bases de données synthétiques. Les résultats indiquent que la méthode proposée est très prometteuse car les résultats sont de loin supérieurs à ceux obtenus par la méthode conventionnelle des pseudo-panels. Par ailleurs, l'approche est très flexible et est directement applicable à tous les types de modèles linéaires, avec erreurs aléatoires ou à effets fixes. Cette méthode devient d'autant plus intéressante dans un contexte de modèles à choix discrets (Paquet, 2002).

ANNEXE A

Nombre de données dans chaque cellule

Voici le nombre d'observations par cellule pour chaque ensemble de données utilisé pour l'estimation du modèle linéaire.

TABLEAU 7

Nombre de cohortes = 8, $t = 1$

	NOMBRE D'INDIVIDUS			
Numéro de cohorte	**Échantillon complet**	**Troncature 5 %**	**Troncature 10 %**	**Troncature 20 %**
1	794	758	723	657
2	479	454	430	385
3	563	526	501	472
4	665	637	608	539
5	644	610	584	519
6	588	569	540	493
7	495	470	445	100
8	772	731	699	632
Total	5 000	4 755	4 530	4 097

TABLEAU 8

Nombre de cohortes = 8, $t = 2$

	NOMBRE D'INDIVIDUS			
Numéro de cohorte	**Échantillon complet**	**Troncature 5 %**	**Troncature 10 %**	**Troncature 20 %**
1	747	722	687	648
2	472	457	434	396
3	597	565	539	525
4	661	630	602	559
5	663	622	610	558
6	609	589	553	512
7	446	427	408	379
8	805	763	734	683
Total	5 000	4 775	4 567	4 260

TABLEAU 9

NOMBRE DE COHORTES = 12, $t = 1$

Numéro de cohorte	NOMBRE D'INDIVIDUS			
	Échantillon complet	**Troncature 5 %**	**Troncature 10 %**	**Troncature 20 %**
1	330	316	300	266
2	464	442	423	391
3	331	313	297	267
4	420	400	378	349
5	467	441	418	379
6	489	463	446	401
7	512	488	468	415
8	432	418	396	364
9	434	413	392	351
10	349	330	313	282
11	431	408	387	343
12	341	232	312	289
Total	5 000	4 755	4 530	4 097

TABLEAU 10

NOMBRE DE COHORTES = 12, $t = 2$

Numéro de cohorte	NOMBRE D'INDIVIDUS			
	Échantillon complet	**Troncature 5 %**	**Troncature 10 %**	**Troncature 20 %**
1	307	293	281	261
2	440	429	406	387
3	354	344	326	296
4	400	382	365	351
5	478	450	427	413
6	498	476	457	420
7	520	486	482	442
8	447	431	391	368
9	418	400	391	358
10	333	321	307	281
11	466	439	429	400
12	339	324	301	283
Total	5 000	4 775	4 567	4 260

ANNEXE B

Distribution conditionnelle pour le modèle linéaire univarié

Soit le modèle linéaire suivant que l'on exprime sous forme matricielle comme suit :

$$Y = X\beta + u\,, \tag{23}$$

$$u = Z_\theta\,\theta + \nu$$

et $Z_\theta = I_N \otimes \iota_T$

où ι_T est un vecteur de uns de dimension T. La matrice des variances-covariances s'écrit alors :

$$\Omega = E(u\,u') = Z_\theta\,E(\theta\,\theta')\,Z'_\theta + E(\nu\,\nu') \tag{24}$$

$$= \sigma^2_\theta(I_N \otimes J_T) + \sigma^2_\nu(I_N \otimes I_T)\,.$$

La régression auxiliaire pour sa part s'écrit :

$$X = Z\gamma + \in, \tag{25}$$

$$\in \sim N(0, V)$$

et $V = E(\in \in') = \tau^2\,I_{NT}\,.$

Selon les hypothèses faites sur le modèle, nous pouvons déduire les distributions conditionnelles suivantes :

$$P(Y|X) \propto \exp\left(-\frac{1}{2}(Y - X\beta)'\,\Omega^{-1}(Y - X\beta)\right)$$

$$\propto \exp\left(-\frac{1}{2}(Y'\,\Omega^{-1}\,Y - 2\beta'\,X'\,\Omega^{-1}\,Y + \beta'\,X'\,\Omega^{-1}\,X\beta)\right)$$

et

$$P(X|Z) \propto \exp\left(-\frac{1}{2}(X - Z\gamma)'\,V^{-1}(X - Z\gamma)\right) \tag{26}$$

$$\propto \exp\left(-\frac{1}{2}(X'\,V^{-1}\,X - 2\gamma'\,Z'\,V^{-1}\,X + \gamma'\,Z'\,V^{-1}\,Z\gamma)\right).$$

Lors de l'application, nous retenons une seule variable explicative. Par conséquent, le modèle est défini comme suit :

$$y_{nt} = x_{nt}\,\beta + u_{nt} \tag{27}$$

et $u_{nt} = \theta_n + \nu_{nt}$

où $n = 1, 2, ..., N$ et $t = 1, 2, ..., T$, avec la structure suivante pour les termes aléatoires :

$$\theta_n \sim N(0, \sigma_\theta^2) \tag{28}$$

et $\nu_{nt} \sim N(0, \sigma_\nu^2)$.

En ce qui a trait à la régression auxiliaire permettant de simuler les variables manquantes, nous avons, sous l'hypothèse d'une seule variable explicative pour cette régression, la relation suivante :

$$x_{nt} = z_{nt}\,\gamma + \in_{nt} \tag{29}$$

et $\in_{nt} \sim N(0, \tau^2)$.

Nous pouvons maintenant réécrire les équations de probabilité conditionnelle comme suit :

$$P(y_{nt} \mid x_{nt}, \theta_n) \propto \exp\left(-\frac{1}{2\sigma_v^2}(y_{nt} - \theta_n - \beta\, x_{nt})^2\right) \tag{30}$$

et $$P(x_{nt} \mid z_{nt}) \propto \exp\left(-\frac{1}{2\tau^2}(x_{nt} - \gamma\, z_{nt})^2\right). \tag{31}$$

En combinant les équations (30) et (31) nous obtenons la distribution conditionnelle suivante :

$$
\begin{aligned}
P(x_{nt} \mid y_{nt}, z_{nt}) &= \frac{P(x_{nt}, y_{nt} \mid z_{nt})}{P(y_{nt})} \\
&\propto P(x_{nt}, y_{nt} \mid z_{nt}) \\
&\propto P(y_{nt} \mid x_{nt})\, P(x_{nt} \mid z_{nt}) \\
&\propto \exp\left(-\frac{1}{2\sigma_v^2}(y_{nt} - \theta_n - \beta\, x_{nt})^2\right) \exp\left(-\frac{1}{2\tau^2}(x_{nt} - \gamma\, z_{nt})^2\right) \\
&\propto \exp\left(-\frac{1}{2\sigma_v^2}(y_{nt}^2 - 2\theta_n\, y_{nt} - 2\beta\, x_{nt}\, y_{nt} + \theta_n^2 + 2\theta_n\, \beta\, x_{nt} + \beta^2 x_{nt}^2)\right) \\
&\quad \times \exp\left(-\frac{1}{2\tau^2}(x_{nt}^2 - 2\gamma\, x_{nt}\, z_{nt} + \gamma^2 z_{nt}^2)\right) \\
&\propto \exp\left(-\frac{\left(\frac{\beta^2}{\sigma_v^2} + \frac{1}{\tau^2}\right)}{2}\left(x_{nt}^2 + x_{nt}\left[\frac{\left(\frac{1}{\sigma_v^2}(A) + \frac{1}{\tau^2}(-2\gamma\, z_{nt})\right)}{\left(\frac{\beta^2}{\sigma_v^2} + \frac{1}{\tau^2}\right)}\right]\right)\right) \\
&\propto \exp\left(-\frac{\tilde{\varpi}}{2}\left(x_{nt}^2 + \frac{\Lambda}{\tilde{\varpi}}\, x_{nt}\right)\right) \\
&\propto \exp\left(-\frac{\tilde{\varpi}}{2}\left(x_{nt} + \frac{\Lambda}{2\tilde{\varpi}}\right)^2\right), \qquad (32)
\end{aligned}
$$

qui est le noyau d'une distribution normale $N\left(-\frac{\Lambda}{2\varpi}, \varpi^{-1}\right)$ avec

$$\varpi = \frac{\beta^2}{\sigma_v^2} + \frac{1}{\tau^2},$$

$$\Lambda = \frac{A}{\sigma_v^2} + \frac{B}{\tau^2},$$

$$A = -2\beta\, y_{nt} + 2\theta_n\, \beta$$

et $B = -2\gamma\, z_{nt}$.

BIBLIOGRAPHIE

ALESSIE, R., M.P. DEVEREUX et G. WEBER (1997), « Intertemporal Consumption, Durables and Liquidity Constraints: A Cohort Analysis », *European Economic Review*, 41 : 37-59.

BALTAGI, B. (1995a), *Econometric Analysis of Panel Data*, John Wiley Sons, Chischester, England.

BALTAGI, B. (1995b), « Panel Data », *Journal of Econometrics*, 68 : 1-243.

BEAUDRY, P. et D. GREEN (2000), « Cohort Patterns in Canadian Earnings: Assessing the Role of Skill Premia in Inequality Trends », *Canadian Journal of Economics*, 33(4).

BLUNDELL, R. et C. MEGHIR (1990), « Panel Data and Lify-Cycle Models », *in* J. HARTOG, G. RIDDER et J. THEEUWES, (éds), *Panel Data and Labor Market Studies,* North Holland, Amsterdam, p. 231-252.

BROWNING, M. A. DEATON et M. IRISH (1985), « A Profitable Approach to Labor Supply and Commodity Demands Over the Life Cycle », *Econometrica*, 53 : 503-543.

CASELLA, G et E. GEORGE (1992), « Explaining the Gibbs Sampler », *The American Statistician*, 46 : 109-126.

DEATON, A. (1985), « Panel Data from Time Series of Cross Sections », *Journal of Econometrics*, 30 : 109-126.

GARDES, F., S. LANGLOIS et D. RICHAUDEAU (1996), « Cross-Section Versus Time Series Income Elasticities of Canadian Consumption », *Economic Letters*.

GARDES, F. et C. LOISY (1997), « La pauvreté selon les ménages : une évaluation subjective et indexée sur leur revenu », *Économie et Statistique*, 308-309-310 : 95-113.

GORDON, S. et G. BÉLANGER (1996), « Échantillonnage de Gibbs et autres applications des chaînes markoviennes », *L'Actualité économique*, 72(1) : 27-49.

HECKMAN, J.J. et R. ROBBS (1985), « Alternative Models for Evaluating the Impact of Interventions: An Overview », *Journal of Econometrics*, 30 : 239-267.

HIRANO, K., G.W. IMBENS, G. RIDDER et D.B. RUBIN (1998), « Combining Panel Data Sets with Attrition and Refreshment Samples », Technical Working Paper 230, National Bureau of Economic Research, Cambridge, MA.

MOFFIT, R. (1993), « Identification and Estimation of Dynamic Models with a Time Series of Repeated Cross-Sections », *Journal of Econometrics*, 59 : 99-123.

PAQUET, M.-F.(2002), « Une approche à simulation pour le traitement des données longitudinales incomplètes », Thèse de doctorat, Université Laval.

TANNER, M.A. et W.H. WONG (1987), « The Calculation of Posterior Distributions by Data Augmentation », *Journal of the American Statistical Association*, 82(398) : 528-540.

VERBEEK, M. et TH.E. NIJMAN (1992), « Can Cohort Data Be Treated As Genuine Panel Data? », *Empirical Economics*, 17 : 9-23.

VERBEEK, M. et TH.E. NIJMAN (1993), « Minimum MSE Estimation of a Regression Model with Fixed Effects and a Series of Cross-Sections », *Journal of Econometrics*, 59 : 125-136.