

Introduction à  
l'économétrie des  
données de panel

Sébastien DESSUS



CODESRIA



# **Introduction à l'économétrie des données de panel**

**Réseau de recherche sur les politiques industrielles en  
Afrique ( RPI)**

**Sébastien Dessus**

*Centre de développement de l'OCDE  
94, rue Chardon-Lagache, 75016 Paris, France  
Tel : 45 24 84 88 / Fax : 45 24 79 43 / E Mail : dessus@oecd.org*

**Document Spécial n° 5**

**Juillet 1996**

## **Introduction à l'économétrie des données de panel**

Copyright ©CODESRIA 1996

Mise en page : Emilie SARR  
Impression : CODESRIA  
Publication : RPI

Codesria en coédition avec NENA  
isbn : 978-2-86978-959-3

Avec le soutien du CNL



# Table des matières

Introduction-----	1
<b>1<sup>ère</sup> partie revue des principaux modèles-----</b>	<b>3</b>
La spécification générale -----	3
Les principaux modèles d'hétérogénéité -----	5
<i>L'hypothèse de comportement uniforme dans le temps et parmi les individus-----</i>	<i>5</i>
<i>L'hypothèse de comportements individuels indépendants les uns des autres -----</i>	<i>5</i>
<i>L'hypothèse de comportements individuels dépendants les uns des autres -----</i>	<i>6</i>
<i>L'hypothèse de l'existence d'effets spécifiques déterministes-----</i>	<i>7</i>
<i>L'hypothèse de l'existence d'effets spécifiques aléatoires-----</i>	<i>7</i>
<b>2<sup>e</sup> partie modèles à effets fixes et modèles à erreurs composées -----</b>	<b>9</b>
Le modèle à effets fixes-----	9
Le modèle à erreurs composées-----	11
Quel modèle choisir ?-----	12
Logiciels et pratique -----	14
<b>3<sup>e</sup> partie le modèle à coefficients variants -----</b>	<b>15</b>
La méthode d'estimation -----	15
Un exemple d'application -----	17
Bibliographie -----	19



## Introduction

L'utilisation de base de données combinant les dimensions temporelles et individuelles — les données de panel — s'est considérablement développée au cours de ces dernières années en économie appliquée. Plusieurs raisons permettent d'expliquer ce phénomène. La première est la disponibilité croissante de ce type de base de données. De nombreux instituts statistiques collectent à intervalles plus ou moins rapprochés des données individuelles, comme les enquêtes de ménage par exemple. L'accumulation au cours du temps de ces photographies détaillées permet aujourd'hui de disposer de bases de données comportant une très large dimension individuelle et une information dynamique suffisante à l'élaboration de modèles statistiques permettant de tester de nombreuses hypothèses. La seconde raison tient aux progrès réalisés en informatique, à la fois en termes de puissance de calcul et d'offre de logiciels économétriques appropriés au traitement des données de panel. Ce qui pouvait apparaître comme un problème insoluble il y a encore quinze ans, par exemple l'inversion d'une matrice de grande taille, est aujourd'hui devenu routinier. Enfin, la réflexion sur la manière pertinente de traiter l'information disponible en fonction des questions posées et des hypothèses retenues s'est considérablement enrichie depuis les travaux précurseurs de Zellner (1978) ou de Balestra et Nerlove (1966).

Mais, plus que toute autre raison, c'est l'intérêt de traiter des bases de données riches en information qui suscite leur utilisation massive. Chacune des deux dimensions procure une information que l'autre ne possède pas. La combinaison des deux dimensions permet alors d'obtenir des résultats à la fois plus fiables et plus généraux que ceux que l'on obtiendrait en n'utilisant qu'une seule des deux dimensions. Le grand nombre d'observations dont on dispose, et donc le grand nombre de degrés de liberté, permet par ailleurs de tester des hypothèses ou modèles beaucoup plus sophistiqués que ceux que l'on peut tester avec de petits échantillons. On peut ainsi par exemple tenir explicitement compte de variables non observables, que l'on nomme généralement les *effets fixes*, dont l'omission pourrait biaiser l'estimation. Le grand nombre d'observations permet de les identifier et de les mesurer. Ce nombre exhaustif d'observations, notamment dans sa dimension individuelle, permet enfin de considérer que les propriétés asymptotiques des estimateurs sont empiriquement vérifiées. Il permet enfin de réduire sensiblement les problèmes de multicolinéarité, puisque les variables explicatives, qui varient dans les deux dimensions, ont moins de chance d'être corrélées.

Ce document ne vise en aucun cas à faire une revue exhaustive des problèmes, et solutions qui s'y rapportent, que l'on peut rencontrer lors du traitement économétrique des données de panel. Une revue de cet ordre est proposée de manière claire et détaillée dans la seconde édition du recueil d'articles «*The Econometrics of Panel Data*» édité en 1996 sous la direction de L. Mátyás et P. Sevestre. On peut aussi se référer à l'ouvrage plus ancien de C.

Hsiao (1986) ou à celui de B. H. Baltagi (1995) qui présentent aussi de manière claire les principaux résultats théoriques de l'économétrie des données de panel. On se bornera ici à présenter de manière concise les premières questions que doit à notre avis se poser tout utilisateur d'une base de données de panel (Balestra 1996) Lorsque les observations comportent la double dimension individuelle et temporelle, il est nécessaire — et c'est le point crucial de l'économétrie des données de panel — de définir clairement dans quel mesure les différences de comportement entre individus et / ou dans le temps doivent être prises en compte, ou modélisées. En montrant comment à diverses hypothèses concernant les différences de comportement correspondent diverses spécifications, on espère donner au lecteur une vue globale de l'économétrie des données de panel (1ère partie). La seconde partie présente plus en détail les deux types de spécification généralement retenues, comment choisir celle qui semble la plus adaptée au problème rencontré et comment les estimer pratiquement.

La troisième partie du document est une illustration pratique, parmi d'autres, de la richesse d'information contenue dans les données de panel, et des possibilités de spécification qui en découlent. On présente un modèle à coefficients variants, qui permet d'expliquer les différences structurelles entre pays quant à la contribution de leur capital humain à la croissance. Le type de spécification retenu dans ce cas n'est utilisable qu'avec des données possédant la double dimension individuelle et temporelle. Il constitue donc une bonne illustration de la spécificité des modèles élaborés de manière à tenir compte des deux dimensions.

# 1<sup>ère</sup> partie Revue des principaux modèles

## La spécification générale

Un panel peut être défini comme un ensemble d'individus dont l'observation est répétée au cours du temps. Au sens strict du terme, le terme panel s'applique lorsque l'on parle d'un groupe de personnes, sans que la dimension temporelle soit nécessairement présente, mais en économétrie sa signification est différente. Outre le fait qu'il comporte obligatoirement une information dynamique, c'est à dire au moins deux points dans le temps, la notion d'individu est prise au sens large : elle peut en effet se référer à une personne physique, mais aussi à une entreprise, une région, un pays, un secteur d'activité, une catégorie de dépenses ou à un type de polluants etc. L'important étant de disposer de plusieurs individus homogènes dans la base de données.

On considérera par la suite que l'on dispose d'observations statistiques pour  $N$  individus durant  $T$  périodes de temps. Que ces dernières soient annuelles, hebdomadaires ou autres n'a pas d'importance. Seul compte le fait que les périodes considérées soient homogènes au sein de l'échantillon. Si pour chaque individu on dispose du même nombre d'observations (ici  $T$  observations par individus), on parlera d'échantillon cylindré. Dans le cas contraire, les résultats présentés par la suite restent valables. Néanmoins, le traitement économétrique d'un échantillon non cylindré tendra à accorder plus d'information à certains individus qu'à d'autres. Les résultats obtenus refléteront alors principalement les caractéristiques des individus les plus représentés dans l'échantillon, autrement dit, ceux pour lesquels on dispose du plus grand nombre d'informations dans le temps.

Considérons la variable d'intérêt  $y$  dont on cherche à expliquer l'évolution dans le temps et pourquoi elle diffère d'un individu à l'autre. Cette variable est doublement indicée : l'indice  $i$  représente l'individu tandis que l'indice  $t$  représente la période considérée. L'observation typique s'écrit donc :

$$y_{it} : i = 1, \dots, N ; t = 1, \dots, T$$

Ces observations peuvent être agencées les unes par rapport aux autres de diverses manières. On pourrait ainsi les organiser de manière à disposer en ligne des individus et en colonne des périodes, mais généralement on empile les individus : les observations sont contenues dans un vecteur d'une colonne et de  $N$  fois  $T$  lignes. Les individus (chacun représenté par ses  $T$  périodes) sont agencés les uns à la suite des autres. Dans ce vecteur typique, les  $T$  premières lignes sont les  $T$  observations du premier individu en ordre croissant avec  $T$ . L'avant dernière observation du vecteur est celle correspondant au  $i$ ème individu à l'avant-dernière période, soit en  $T-1$ . Cette méthode présente deux



avantages : le premier est qu'il facilite la présentation générale de l'économétrie des données de panel, car il simplifie au maximum les calculs matriciels ; le second est qu'une fois agencé de la sorte, ce type de vecteur peut être lu et utilisé par des logiciels conçus pour traiter des séries temporelles, ce qui représente la grande majorité des logiciels disponibles actuellement.

$$y = \begin{pmatrix} y_{1,1} \\ \cdot \\ \cdot \\ y_{1,T} \\ \cdot \\ \cdot \\ y_{i,t} \\ y_{i,t+1} \\ \cdot \\ \cdot \\ y_{N,1} \\ \cdot \\ \cdot \\ y_{N,T} \end{pmatrix}$$

Dans la spécification la plus générale qui soit, on explique cette variable  $y$  par un ensemble de  $K$  variables exogènes  $x$  prédéterminées, par une constante  $\alpha$  et par une variable aléatoire non observable  $u$ . Le modèle retenu est linéaire, comme c'est le cas dans de nombreuses applications économiques, et est par ailleurs adapté à une première présentation de l'économétrie des données de panel. Retenir un modèle non linéaire nous entraînerait dans des considérations compliquées, masquant ainsi la spécificité de l'économétrie des données de panel. Ce modèle s'écrit ainsi:

$$y_{it} = \alpha_{it} + \beta_{1it}x_{1it} + \dots + \beta_{Kit}x_{Kit} + u_{it}$$

Ce modèle suppose donc que chaque individu a un comportement spécifique, lui même différencié d'une période à l'autre. Ce modèle n'a pas d'intérêt d'un point de vue économique, puisqu'il est tout au plus descriptif. Il n'a aucun pouvoir explicatif ni prédictif car il ne dispose d'aucun degré de liberté et qu'il ne permet pas d'identifier un comportement commun dans le temps ou entre individus, ce qui est pourtant le but de l'inférence statistique. Il permet toutefois d'exprimer dans sa globalité les différentes dimensions d'un modèle propre aux

le temps propre à chaque individu. Ce modèle-ci peut s'écrire ainsi :

$$H1(5): \begin{cases} \alpha_{it} = \alpha \\ \beta_{kit} = \beta_k \end{cases} \quad \text{et} \quad H2(5): u_{it} = \alpha_i + \varepsilon_{it}$$

La prise en compte d'un effet spécifique n'est effectuée qu'au niveau du résidu et ne se fait sentir que sur les moments du second ordre. Les résidus sont donc hétéroscédastiques ce qui nécessite alors une estimation par la méthode des moindres carrés généralisés (MCG). Nous développerons dans la seconde partie la méthode d'estimation, mais il est déjà clair que si ce modèle est séduisant, il est aussi plus difficile à estimer. En fait, les applications économétriques utilisant des données de panel se concentrent très largement sur les deux derniers modèles décrits, qui allient tous les deux les propriétés de faisabilité économétrique (méthodes relativement aisées à appliquer et peu consommatrices de degrés de liberté), et de spécification simple, compréhensible et cohérente d'hétérogénéité / homogénéité de comportements entre individus. La deuxième partie décrit plus en détail ces deux méthodes d'estimation de l'effet individuel, et présente les arguments qui font peser la balance en faveur de l'un ou l'autre de ces modèles.

Dans certaines applications, on considère aussi un effet spécifique pour la dimension temporelle, ce qui n'a pas été le cas ici. Bien que rendant la tâche plus compliquée, il est possible d'adapter les modèles discutés ci-dessus à cette spécification plus générale. Pour des raisons de présentation, on préfère toutefois conserver l'hypothèse selon laquelle les comportements sont invariants dans le temps.