

# Les limites des méthodes de régression appliquées aux données hiérarchiques : principe et application des modèles multi-niveaux

## 1. Introduction

Les données collectées chez l'homme répondent très fréquemment, sinon toujours, à une structure hiérarchique (voir figure 1). Selon les cas, il peut s'agir de :

- données groupées (*clustered*) ;
- répétition des mesures chez les mêmes sujets (données longitudinales) ;
- appartenance à une entité géographique (dimension spatiale) ;
- appartenance à un groupe social, professionnel, institutionnel, scolaire, familial (génétique) ;
- sondage en grappe ;
- randomisation par groupe.

**Brahim Chedati**

ISESCO, Rabat  
(chedati@yahoo.fr)

Figure 1  
**Structure hiérarchique des données**

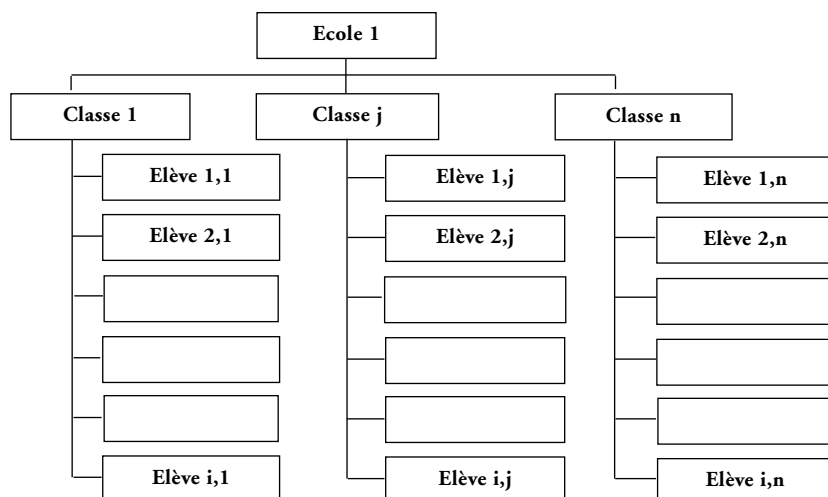
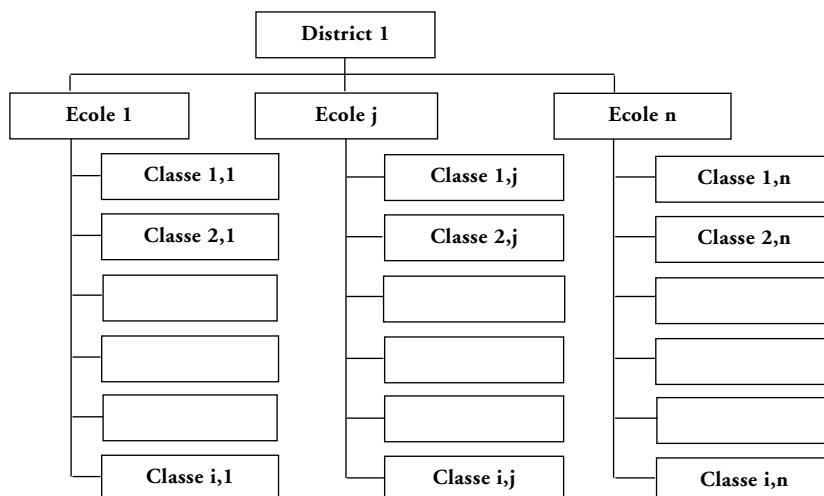


Figure 2  
Structure hiérarchique des données (suite)



Les unités observées qui appartiennent à une même entité ou groupe et qui forcément partagent ses caractéristiques auront tendance à se ressembler beaucoup plus que celles qui appartiennent à des groupes différents. La conséquence en est une non-indépendance des observations, ce qui est, en fait, une violation du principe de base des régressions linéaires classiques où toutes les observations sont observées à un seul niveau.

Cette structure de corrélation doit être prise en compte au stade de l'analyse statistique afin d'éviter des inférences statistiques incorrectes quant aux coefficients de régression et afin d'obtenir une meilleure précision des effets.

Les deux figures ci-dessus se complètent. Ils visualisent une structure hiérarchique à trois niveaux :

- les élèves (niveau 1)
- les classes (niveau 2)
- les écoles (niveau 3).

L'étude des relations entre l'individu et son milieu a toujours préoccupé les chercheurs en sciences sociales, particulièrement les sociologues (1) et les démographes. L'idée que les comportements individuels sont déterminés (au sens causal du terme) par des variables contextuelles en plus des variables personnelles (âge, genre, niveau d'instruction, état de santé, situation matrimoniale, etc.) est communément admise dans les rangs des chercheurs en sciences sociales. Si le principe général de l'interaction est largement acquis, le problème qui restait posé, pratiquement jusqu'au début des années 80, tenait aux techniques statistiques capables d'effectuer l'analyse dite multiniveaux. Cela ne veut pas dire que les variables contextuelles n'ont pas fait l'objet d'analyse statistique, mais l'application informatique des

(1) Durkheim (1897) ;  
Merton (1968) ; Weber  
(1964) ; Blau (1960) ;  
Boudon (1963).

méthodes spécifiques d'estimation des paramètres qui permettent d'isoler les effets à différents niveaux n'était pas encore au point avant les années 80.

Prenons un exemple scolaire : les élèves inscrits au titre d'une année scolaire « t » constituent le niveau le plus bas de l'analyse (niveau 1). Ces élèves appartiennent à des classes (niveau 2), lesquelles classes appartiennent à des établissements scolaires (niveau 3) qui, à leur tour, forment des circonscriptions (niveau 4), etc. (voir schémas ci-dessus).

Le problème posé ici consiste à déterminer l'effet de chacune des variables explicatives sur la variable dépendante qui peut être, dans cet exemple scolaire, la réussite ou la note à une examen intermédiaire ou final, etc. Il est essentiel de signaler que les variables explicatives n'appartiennent pas toutes à un même niveau. En effet, si les caractéristiques socio-démographiques des élèves sont par excellence les variables liées à un même niveau (les élèves), il n'en est pas de même des variables dites contextuelles qui, elles, appartiennent à des niveaux différents (i.e. l'âge des enseignants, leur qualification, leurs salaires... l'infrastructure scolaire, le milieu d'implantation de l'école, etc.).

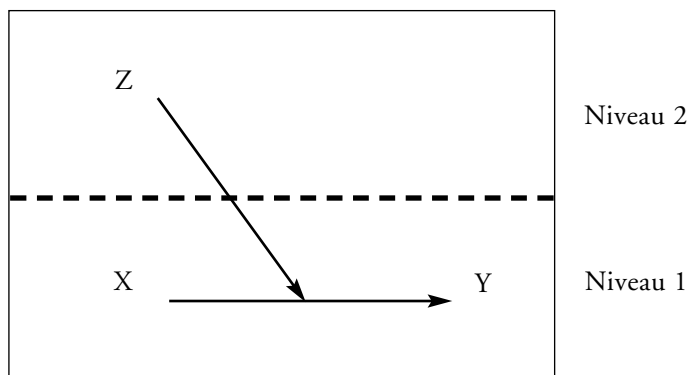
Dans un pareil cas, le recours aux méthodes classiques de régression pour déterminer l'effet de chacune des variables explicatives sur la variable dépendante n'est pas une solution valable et ce pour deux raisons principales :

1. l'hypothèse fondamentale d'indépendance des termes d'erreur se trouve violée (les MCO (2) supposent en effet que les résidus sont indépendants) ;

(2) Moindres carrés ordinaires.

2. les coefficients estimés par les MCO représentent des effets fixes, étant donné que les observations sont « nichées » (nested) à l'intérieur d'un même niveau. Or, dans le cas où les observations appartiennent à des niveaux hiérarchisés, les effets deviennent aléatoires.

Figure 3  
**Imputation (agrégation)  
des variables du niveau 2**

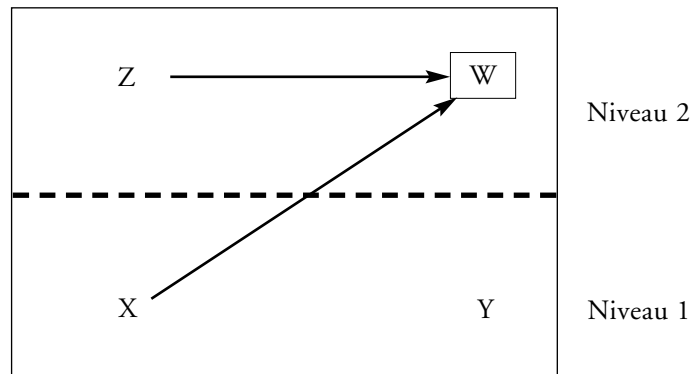


Afin de contourner le problème des niveaux d'analyse, l'estimation des paramètres se fait par application de l'une des méthodes suivantes :

- La première consiste à imputer au niveau inférieur les données des niveaux supérieurs de la hiérarchie puis à estimer le modèle sur la base des observations du premier niveau. Cette approche pose des problèmes de fond puisque l'imputation des données démultiplie artificiellement les données du niveau supérieur et entraîne la violation de l'hypothèse d'indépendance des termes d'erreur (point 1 *supra*).

- Selon la seconde approche, les données du premier niveau sont agrégées aux niveaux supérieurs de la hiérarchie et constituent la base de l'estimation du modèle de régression. Cette agrégation engendre néanmoins d'énormes problèmes car, en agissant ainsi, on accepte de perdre l'information concernant la variation du premier niveau qui, dans certains cas, peut être très importante.

Figure 4  
**Agrégation des variables du niveau 1**



- Analyse de covariance : on utilise dans ce cas une régression habituelle (MCO) au niveau individuel en utilisant les groupes comme facteur (variables muettes).

Le désavantage de cette procédure est que l'inférence vise ces groupes particuliers et non la population des groupes dont les présents n'en constituent qu'un des échantillons. Il s'agit donc de modèle à effets fixes et non aléatoires.

$$Y = a_0 + a_1X_1 + a_2X_2 + \dots + b_1C_1 + b_2C_2 + \dots + b_{(p-1)} C_{(p-1)}$$

où : Y est la variable dépendante

$X_i$  les variables indépendantes individuelles

$C_i$  les variables muettes « classe » avec  $C_p$  comme référence

$a_i$  et  $b_v$  les coefficients du modèle.

- Procédure à deux étapes :
  - on régresse au niveau de chaque groupe (régression MCO),
  - on utilise les coefficients estimés pour chaque groupe comme variable dépendante dans une analyse inter-groupe. Ceci est une forme plus subtile d'agrégation. Mais le risque encouru par une telle procédure est le peu de fiabilité des coefficients à cause du nombre réduit d'observations (groupes).

De plus, on ne tient pas compte de la différence de taille des groupes.

Pour résoudre de tels problèmes, les statisticiens ont conçu des modèles appropriés dénommés « modèles multi-niveaux » (3) qui permettent de mieux analyser les bases de données à structure hiérarchique (Bryk et Raudenbuch, 1992 ; Goldstein, 1986, 1987, 1995 ; Goldstein et McDonald, 1989 ; Hox et Kreft, 1994 ; Prosser, Rabach et Goldstein, 1991 ; T. Snijders, 1999, 1994).

(3) Selon le domaine de recherche, ces modèles sont appelés modèles linéaires multi-niveaux ou hiérarchiques (sociologie), modèles à effets mixtes ou modèles à effets aléatoires (biométrie), modèles de régression à coefficients aléatoires (économétrie), etc.

## 2. Formalisation

Considérons, pour simplifier, une hiérarchie à deux niveaux où des élèves (niveau 1) dotés de caractéristiques (X) sont « nichés » dans des classes (niveau 2) dans lesquelles on s'intéresse aux variables explicatives contextuelles (Z). Supposons qu'on veuille expliquer la note finale (Y) obtenue par chacun des élèves de l'échantillon par la note obtenue tout au début de l'année (note initiale, X).

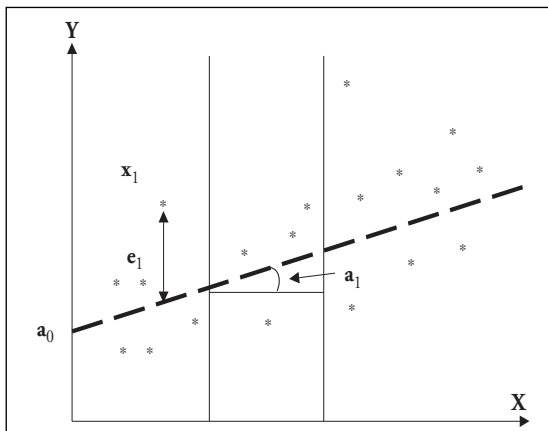
Si tous les élèves de l'échantillon étaient dans une même classe et partageaient ainsi les mêmes conditions d'apprentissage (méthode pédagogique, matériel d'enseignement, qualification des enseignants, composition sociale, etc.), une simple régression suffirait pour pouvoir apprécier l'effet de la note initiale sur la note finale. Une telle équation de régression s'écrirait :

$$Y_i = a_0 + a_1 \cdot x_i + e_i \quad (1)$$

où  $Y_i$  représente la note finale obtenue par le  $i^{\text{e}}$  élève,  
 $a_0$  et  $a_1$  représentent les coefficients estimés de la droite par la méthode des MCO.  
 $e_i$  est le résidu ou l'écart entre la note réelle et la note estimée par le modèle pour l'élève  $i$ .

Graphique 1

**Représentation des individus en fonction de la variable explicative par la droite des MCO (effet fixe)**



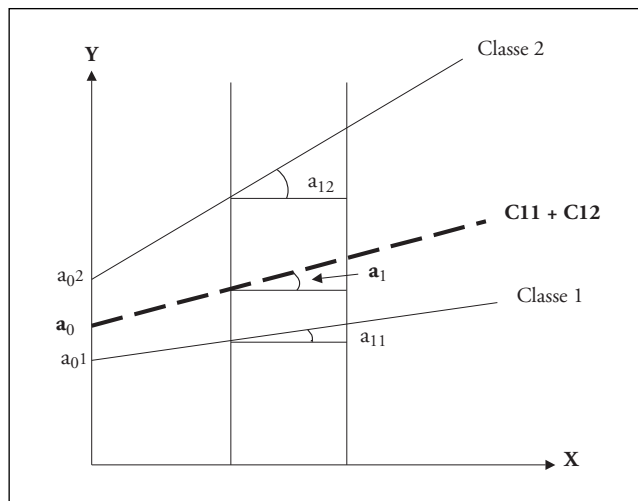
Supposons à présent que les groupes d'élèves fréquentent des classes différentes (pédagogiquement, matériellement et peut-être même socialement). Dans ce cas, la note finale sera influencée par la note initiale obtenue par l'élève (niveau 1) et par les variables du contexte (effet classe).

$$Y_{ij} = a_{0j} + a_{1j} \cdot x_{1ij} + e_{ij} \quad (2)$$

où  $Y_{ij}$  représente la note finale obtenue par le  $i^e$  élève de la  $j^e$  classe ;  
 $a_{0j}$  et  $a_{1j}$  sont des paramètres estimés sur la  $j^e$  classe ;  
 $e_{ij}$  le résidu aléatoire de moyenne égale à zéro et de variance  $\sigma_e^2$ .

Graphique 2

**Représentation multiniveaux (effets fixes et aléatoire)**



En analyse multiniveaux, les classes (niveau 2) sont considérées comme des échantillons tirés d'un ensemble plus vaste de classes. De ce fait, les paramètres  $a_{0j}$  et  $a_{1j}$  sont aléatoires et vont changer d'une classe à l'autre, ce qui nous amène à écrire les deux égalités suivantes :

$$a_{0j} = a_0 + \mu_{0j} \quad (3)$$

$$a_{1j} = a_1 + \mu_{1j} \quad (4)$$

où :

$a_0$  et  $a_1$  sont des paramètres moyens fixes estimés sur l'ensemble des classes ;

$\mu_{0j}$  et  $\mu_{1j}$  sont aléatoires de moyenne nulle et de variances (à estimer) égales à :

$$\text{var}(\mu_{0j}) = \sigma_{\mu 0}^2$$

$$\text{var}(\mu_{1j}) = \sigma_{\mu 1}^2 \text{ et } \text{Cov}(\mu_{0j}, \mu_{1j}) = \sigma_{\mu 01}^2$$

Le graphique 2, bien qu'il soit trop simplifié (intentionnellement) visualise les écarts aléatoires entre les coefficients de l'ensemble des classes (droite en pointillés gras) et ceux de chaque classe.

En remplaçant les coefficients de l'équation (2) par leur valeurs dans (3) et (4), on peut écrire le modèle bi-niveau comme suit :

$$Y_{ij} = (a_0 + a_1 * x_{1ij}) + (\mu_{1j} * x_{1ij} + \mu_{0j} + e_{ij}) \quad (5)$$



Partie fixe



Partie aléatoire

La méthode d'estimation des paramètres (MCO) cesse d'être valide étant donné que les termes d'erreur ne sont plus indépendants (2 élèves appartenant à un même groupe auront tendance à se ressembler plus que deux élèves de groupes différents).

L'estimation des paramètres se fait grâce à des programmes informatiques spécialisés.

### 3. Les logiciels d'analyse multiniveaux

Il existe à présent plusieurs logiciels permettant l'estimation des modèles multiniveaux. Les programmes les plus utilisés (car plus conviviaux) sont : Hierarchical Linear Models (HLM) conçu par les professeurs A. Bryk et S. Raudenbush en 1986 ; Multilevel Model (MLWin) conçu par l'équipe de l'Institut of Education de Londres sous la direction du professeur H. Goldstein en 1992 ; VARCL mis au point par N. Langford en 1990. D'autres logiciels sont également disponibles tels MIXREG, MIXOR et EGRET. Le lecteur intéressé peut apprendre énormément de choses en consultant l'article de Hox et Kreft (1994).

En plus de ces logiciels spécialisés en analyse multiniveaux, il existe aussi des modules multiniveaux implémentés aux logiciels d'analyse statistique tels que : SAS (procédure "proc mixed"), stata, SPSS 11 et plus (procédure "mixed"), Systat 11 et plus, S-Plus, etc.

### 4. Etude de cas

Afin de montrer en quoi la méthode de régression usuelle ne convient pas aux données hiérarchiques, nous avons constitué une base de données sur 330 élèves de la première année du collège répartis en 10 classes (4).

La base contient entre autres les données suivantes :

- l'âge des élèves ;
- le genre des élèves ;
- la note obtenue à l'examen final de mathématiques ;
- le genre des enseignants ;
- l'ancienneté des enseignants.

Les trois premières informations sont relatives à l'élève (niveau 1), les deux dernières concernent l'enseignant (niveau 2).

(4) Je tiens à remercier le directeur du collège Imik d'avoir mis l'information souhaitée à notre disposition.

Si on régresse la note obtenue par les élèves sur le genre des élèves et l'ancienneté des enseignants par la méthode de régression habituelle (MCO), l'affectation de la valeur de la variable "ancienneté des enseignants" se trouve affectée aux élèves d'une même classe (voir approche de désagrégation *supra*), amplifie "artificiellement" le nombre d'observations du niveau classe et biaise, de ce fait, les coefficients estimés.

Le tableau suivant montre en effet que la variabilité des notes de l'ensemble des élèves (330) n'est expliquée qu'à hauteur de 7,6 % par les variables prises en compte dans le modèle. Ajouté à cela la non-significativité statistique de la constante.

Effect	Coefficient	Std Error	t	P(2 Tail)
Constant	1.585	1.389	1.14	0.255
AnciennetéProf	0.253	0.054	4.65	0.00
GenreElève	- 0.846	0.364	- 2.33	0.021

Dependant Variable : note ; n° 330 ; Squared multiple R : 0.076.

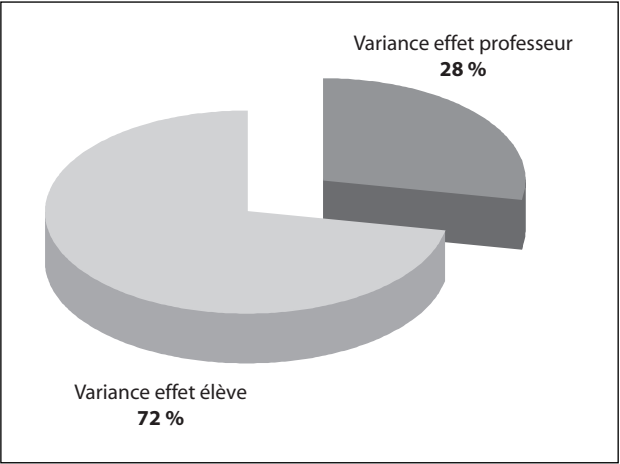
Si à présent on tient compte de la structure hiérarchique des données collectées auxquelles on applique une analyse multiniveaux, la variabilité des 330 notes d'examen est expliquée à hauteur de 28 % par les variables de niveau 1 (les caractéristiques des élèves) et 78 % par les variables inhérentes aux enseignants (niveau 2). Ce dernier pourcentage est connu sous le terme "effet maître" dans la littérature sur l'efficacité en éducation.

Nous reproduisons ci-dessous les résultats de l'analyse multiniveaux que nous avons réalisée par Mlwin et Systat version 11.





**Structure des effets élève et professeur (Modèle 2)**



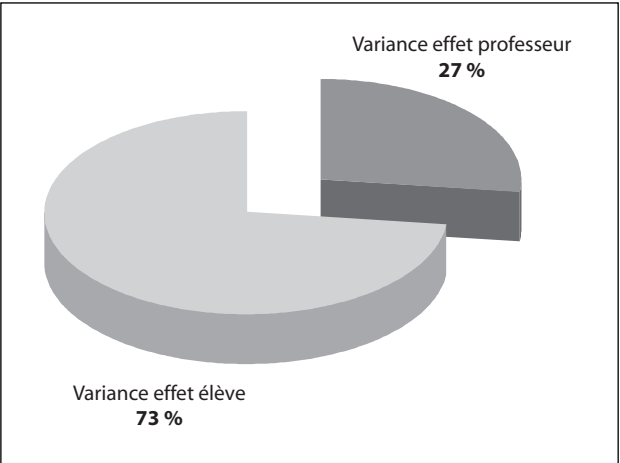
Systat 11.0

Calculation of the intracluster correlation

---

Residual variance	=	7.954	(73 %)
Cluster variance	=	3.002	(27 %)

**Structure des effets élève et professeur  
(Modèle 1-systat 11)**



## Références bibliographiques

- Blau P.M. (1960), « Structural Effect », *American Sociological Review*, 22, p. 178-193.
- Boudon R. (1963), « Propriétés individuelles et propriétés collectives : un problème d'analyse écologique », *Revue française de sociologie*, 4, p. 275-299.
- Bryk A.S., Raudenbush S.W. (1992), *Hierarchical Linear Models : Application and Data Analysis Methods*, Sage Publications, Inc.
- Chedati B. (1997), « L'analyse multiniveaux comme solution aux limites des régressions classiques appliquées aux données hiérarchiques », in *les Méthodes qualitatives en sciences sociales*, Publications de la Faculté des lettres et des sciences humaines, Rabat.
- Courgeau D. et Baccaini B. (1997), « Analyse multi-niveaux en sciences sociales », *Population*, vol. 52, n° 4, juillet-août, p. 831-864.
- De Euw J. (1992), « Series Editor's Introduction to Hierarchical Linear Models », in Bryk A.S. et Raudenbush S.W. (1992), *Hierarchical Linear Models : Application and Data Analysis Methods*, Sage Publications, Inc, XIII-XVI.
- Duncan C., Jones K., Moon G. (1995), « Blood Pressure, Age Gender », in Woodhouse G. (ed), *A Guide to MLN New Users*, Multilevel models project, Institute of education, University of London, 59-85.
- Durkheim E. (1993), *le Suicide*, PUF, Paris (1<sup>re</sup> édition 1897).
- Elston R.C., Grizzle J.E. (1962), « Estimation of Time Response Curves and Their Confidence Bands », *Biometrics*, 18, 148-159.
- Goldstein H. (1986), « Multilevel mixed linear model analysis using interactive generalized least square », *Biometrika*, 73, 43-56.
- Goldstein H. (1987), « Multileveled Models in Educational and Social Research », London, Griffin.
- Goldstein H. (1995), « Multilevel statistical models », London, Edward Arnold.
- Goldstein H., McDonald R. (1988), « A General Model for the Analysis of Multilevel Data », *Psychometrika*, 53, 455-467.
- Hox J.J. (1994), *Applied Multilevel Analysis*. Amsterdam TT- Publikaties.
- Hox J.J., Kreft I.G.G. (1994), « Multilevel Analysis Methods », *Sociological Methods and Research*, 22, 283-299.
- Merton R.K. (1968), *Social Theory and Social Structure*, New York, free press.
- McDonald R.P. (1994), « The Bilevel Reticular Action Model for Path Analysis with Latent Variables », *Sociological Methods and Research*, 22, 399-413.
- Muthen B.O. (1989), « Latent Variable Modelling in Heterogeneous Population », *Psychometrika*, 54, 557-585.
- Muthen B.O. (1994), « Multilevel Covariance Structure Analysis », *Sociological Methods and Research*, 22, (3), 364-375.
- Patterson L. (1991), « Multilevel Logistic Regression », in Prosser R., Rasbash J., Goldstein H., *Data Analysis with ML3*, Institute of education, University of London, 5-18.
- Patterson L. (1995), « Entry to University by School Leavers », in Woodhouse G. (ed). Plewis I. (1991), « Repeated Measures Models », in Prosser R., Rasbash J., Goldstein H., *Data Analysis with ML3*, Institute of education, University of London, 44-58.
- Prosser R., Rasbash J., Goldstein H. (1991), *ML3 Software for Three-Level Analysis Users Guide for V. 2*, Institute of education, University of London.
- Snijders T.A.B., Bosker R.J. (1994), « Modeled Variance in Two-Level Models », *Sociological Methods and Research*, 22, 342-363.