# ANALYSE MULTI-NIVEAUX EN SCIENCES SOCIALES

Daniel COURGEAU*, Brigitte BACCAÏNI*

Le démographe, comme d'autres chercheurs en sciences sociales, peut se placer à des niveaux d'agrégation différents pour étudier les comportements humains. Mais une fois situé à un niveau d'agrégation donné, il lui est difficile d'en changer en cours d'analyse, car les mesures, les méthodes et les techniques utilisées dans un domaine ne recouvrent pas celles de l'autre.

La démographie a longtemps privilégié l'analyse des grandeurs agrégées. Cela est possible sous l'hypothèse souvent non formulée, que l'on travaille sur des sous-populations, chacune homogène vis-à-vis du comportement étudié. Dans ce cas il n'est guère utile de tenir compte des comportements et des caractéristiques individuels, mais on peut rechercher les relations qui existent entre les taux démographiques classiques, correspondant au phénomène étudié dans chaque sous-population, et les valeurs moyennes des caractéristiques calculées également dans chaque sous-population. Ainsi, pour analyser les taux d'émigration de diverses régions, on va chercher à les relier au taux de chômage, au salaire moyen, au pourcentage de personnes à charge, etc., de ces mêmes régions. On tombe alors sur des modèles de migration classiques, mis en place depuis fort longtemps (Young, 1924), mais toujours utilisés (Puig, 1981 ; Jacquot, 1994). On peut, de la même façon, mettre en place des modèles de fécondité ou de mortalité régionales. Les modèles multi-régionaux de Willekens et Rogers (1978) font la synthèse de ces différentes approches sous l'hypothèse supplémentaire d'indépendance entre les divers phénomènes démographiques.

On peut dire qu'une telle analyse permet de tenir compte de l'effet que les groupes étudiés peuvent avoir sur leurs propres comportements démographiques : les caractéristiques agrégées que l'on peut mesurer sont supposées approcher un certain nombre de contraintes que chaque sous-population impose à ses membres, et qui vont influencer leurs comportements. Ainsi une telle analyse peut montrer une liaison positive entre le taux de chômage dans une région et son taux d'émigration (Puig, 1981).

* Institut national d'études démographiques.

Le risque est grand, à partir de ce résultat, d'en conclure que les individus au chômage ont une plus forte probabilité d'émigrer d'une région, alors que l'on sait seulement qu'un fort taux de chômage conduit à un fort taux d'émigration qui peut concerner aussi bien des actifs occupés, des chômeurs que des inactifs. Ce type d'inférence erronée conduit à ce que l'on a coutume d'appeler l'*erreur écologique*, lorsque l'on cherche à déceler des comportements individuels à partir de mesures agrégées.

Il y a plus de quarante-cinq ans maintenant, Robinson (1950) abordait ces problèmes avec des arguments statistiques. Il a montré que les corrélations entre deux caractéristiques mesurées de façon binaire sur des individus, ou par des proportions sur des régions, n'étaient en général pas égales entre elles. Ainsi la corrélation entre proportions de population noire et d'illettrés aux États-Unis, en 1930, était de 0,95 en travaillant sur neuf divisions géographiques, alors que la corrélation entre le fait d'être noir et le fait d'être illettré, pour un individu, n'était que de 0,20. Il concluait son article de façon très tranchée : on ne peut pas utiliser une corrélation écologique, mesurée au niveau agrégé, comme substitut d'une corrélation individuelle.

Un certain nombre d'auteurs ont étendu ces résultats à des analyses par des méthodes de régression linéaire ou logistique (Alker, 1969 ; Firebaugh, 1978 ; Piantadosi *et al.*, 1988 ; Courgeau, 1994 ; Baccaïni et Courgeau, 1996a). La conclusion est toujours la même : dans la majorité des cas, travailler à un niveau agrégé conduit à des biais si l'on veut raisonner au niveau individuel. Ces biais sont d'autant plus importants que la variabilité entre individus d'un même groupe dépasse à la variabilité entre les groupes considérés, cela pour toutes les caractéristiques étudiées.

Il est, dès lors, utile de travailler également au niveau individuel si l'on veut comprendre les comportements humains. L'approche biographique s'est ainsi mise en place, il y a plus de quinze ans maintenant. Elle a entraîné l'élaboration d'enquêtes recueillant les événements qui surviennent dans tous les domaines de la vie des individus avec leur datation précise. Elle a conduit à la mise en place de méthodes d'analyse originales, qui permettent de relier entre eux divers événements pouvant survenir dans des domaines différents, et de mesurer l'effet de diverses caractéristiques individuelles sur ces mêmes événements. Elle introduit enfin un nouveau paradigme en démographie (Courgeau et Lelièvre, 1996), car l'intérêt ne va plus porter sur des sous-populations homogènes et sur des événements indépendants les uns des autres mais, au contraire, sur l'ensemble de la biographie individuelle, considérée comme un processus stochastique complexe. Ce nouveau paradigme peut être approché par l'hypothèse suivante : un individu parcourt, tout au long de sa vie, une trajectoire complexe qui dépend, à un instant donné, de sa trajectoire antérieure, des informations qu'il a pu acquérir dans son passé et des conditions qui prévalent dans la société où il vit.

Dans ce cas on va rechercher les relations qui existent entre un comportement individuel complexe, dépendant du temps, et diverses caractéristiques

de ces individus. Il peut s'agir de caractéristiques fixées une fois pour toutes (origine sociale des parents, nombre de frères et sœurs, lieu et rang de naissance, etc.) ou de caractéristiques dépendant du temps, indiquant les étapes importantes de leur existence. C'est donc bien l'hétérogénéité des populations qui intervient là et les relations entre les divers phénomènes démographiques seront au cœur de l'analyse. Ainsi, pour étudier les chances d'émigrer de diverses régions, on va faire intervenir le fait que l'individu soit chômeur ou non, son salaire, le nombre de personnes à sa charge, etc. On pourra également faire intervenir d'autres caractéristiques plus permanentes, telles que son lieu de naissance, pour mettre en évidence les chances d'y retourner lorsque l'individu en est parti.

Une telle approche réussit à unifier un certain nombre de méthodes utilisées en sociologie (Tuma et Hannan, 1984), en économie (Lancaster, 1990) et en démographie (Courgeau et Lelièvre, 1989). Il faut cependant prendre garde au fait que l'on considère souvent que les caractéristiques de l'individu ou de personnes proches (son ménage ou sa famille corésidente, par exemple) sont les seules à jouer sur les comportements individuels. On risque de commettre, dans ce cas, ce que l'on appelle l'*erreur atomiste*, car on ignore alors le contexte dans lequel les conduites humaines se produisent. En fait, ce contexte doit certainement jouer sur les comportements individuels et il paraît fallacieux d'isoler l'individu des contraintes imposées par la société et le milieu dans lequel il vit.

D'où l'idée de travailler simultanément à divers niveaux d'agrégation, en vue d'expliquer un comportement qui est toujours individuel, et non plus agrégé, comme précédemment. Cela élimine le risque d'erreur écologique, car la caractéristique agrégée va mesurer une construction différente de son équivalent au niveau individuel. Elle n'intervient plus comme un substitut, mais comme une caractéristique de la sous-population qui va affecter le comportement d'un individu qui en fait partie. Simultanément l'erreur atomiste disparaît à partir du moment où l'on fait intervenir correctement le contexte dans lequel l'individu vit.

Dans le passé, cette possibilité d'analyse multi-niveaux ou contextuelle a été l'objet de nombreux débats méthodologiques en sociologie (Lazarsfeld et Menzel, 1961 ; Hauser, 1974), sans que des applications précises de ces méthodes aient pu apporter les preuves de leurs potentialités. Ce n'est que récemment que l'utilisation de fichiers conçus pour réaliser de telles analyses a permis leur mise en place dans diverses sciences humaines : en épidémiologie (Van Korff *et al.*, 1992), en sciences de l'éducation (Goldstein, 1987 ; 1995), en géographie humaine (Jones, 1993), en sociologie (Entwistle et Mason, 1985), en économie (Geronimus *et al.*, 1996) et en démographie (Courgeau, 1994 ; Baccaïni et Courgeau, 1996a). Nous allons montrer ici de façon plus précise les objectifs, les hypothèses, les méthodes utilisées et les problèmes rencontrés lorsque l'on cherche à réaliser une analyse multi-niveaux.

## I. – Mise en place de modèles multi-niveaux

L'objectif d'une analyse multi-niveaux est d'étudier des processus individuels qui prennent place dans un espace différencié. En effet, différentes actions individuelles ou collectives conduisent à la mise en place de structures spatiales, telles que les bassins d'emploi, les communes ou les départements d'un pays, qui évoluent d'ailleurs au cours du temps. Les individus vivant dans ces unités spatiales vont agir en fonction de leurs propres caractéristiques, mais connaîtront des contraintes imposées par les conditions de vie de chacune d'entre elles : taux de chômage, salaire moyen, densité de population, présence d'une école, etc. On voit ainsi comment les caractéristiques individuelles et les caractéristiques agrégées pourront jouer de façon différente sur les comportements des individus vivant dans chaque zone.

Il faut bien voir maintenant que c'est l'individu qui se trouve au centre d'une telle approche. C'est par lui et à travers lui-même que les divers niveaux d'agrégation existent, mais cela n'empêche pas que les contraintes imposées par ces niveaux puissent l'amener à suivre un comportement différent de celui qu'il aurait suivi hors de ces conditions.

Pour mettre en place une telle analyse il nous faut également distinguer les traits à analyser selon le niveau d'agrégation considéré.

La caractéristique à analyser sera toujours considérée ici comme individuelle. Il pourra s'agir d'une caractéristique binaire : être marié ou non ; d'une caractéristique polytomique : être actif occupé, chômeur ou inactif ; d'une caractéristique pouvant être considérée comme continue : la taille d'un individu, son revenu, etc.

Les caractéristiques explicatives pourront être plus diverses. On pourra d'abord faire intervenir des caractéristiques individuelles, telles que décrites plus haut. Ensuite, pour un niveau d'agrégation donné on pourra agréger simplement ces caractéristiques individuelles et estimer des pourcentages ou des moyennes (Loriaux, 1989) : le pourcentage de mariés, les pourcentages d'actifs occupés et de chômeurs, la taille moyenne d'un individu de chaque unité spatiale. Des procédures analytiques plus complexes pourront également être utilisées : en même temps que le revenu moyen, on peut faire intervenir simultanément l'écart type du revenu, ou la corrélation entre revenu et quotient intellectuel dans chaque région.

D'autres caractéristiques sont plus globales et caractérisent les unités comme un tout : densité de population, nombre de lits d'hôpital, par exemple. Aucune caractéristique individuelle ne leur correspond, mais elles peuvent cependant être agrégées à divers niveaux : le nombre de lits d'hôpital d'une région est la somme des nombres de lits de chacun des départements qui la constituent.

D'autres caractéristiques collectives sont bien définies à un niveau d'agrégation donné, mais ne peuvent guère être agrégées à des niveaux plus larges. Ainsi la coloration politique d'une commune, définie par exemple par le parti d'affiliation de son maire, ne peut guère être agrégée à celle des communes voisines qui peuvent couvrir un large spectre. Elle n'existe donc pas au niveau individuel, ni au niveau départemental ou régional.

Une telle analyse va enfin nécessiter la définition des divers niveaux auxquels on va se placer et des types d'emboîtement pouvant exister entre eux.

L'emboîtement le plus simple et le plus usité est hiérarchique : l'individu appartient à une commune, elle-même partie d'un département, etc. Chaque niveau est constitué de la réunion d'unités de niveau inférieur. Le découpage utilisé peut être administratif, comme dans le cas précédent, ou de tout autre type : élèves situés dans des classes, elles-mêmes situées dans des écoles, elles-mêmes de type public ou privé, etc.

Les emboîtements peuvent être plus complexes : individus situés dans des villes de taille croissante, mais distinguées également en villes administratives, industrielles, touristiques, etc. On a, dans ce cas, une classification croisée, selon que les villes sont classées par taille ou par fonction. Bien entendu on peut avoir des emboîtements pour une part hiérarchiques, et pour une autre part croisés. Ainsi, on peut avoir des individus classés par type de voisinage résidentiel et par type de lieu de travail (classification croisée), eux-mêmes considérés dans un classement hiérarchique en départements et régions.

Ayant ainsi posé les objectifs, les hypothèses, les types de caractéristiques à considérer et les divers emboîtements de niveaux d'agrégation possibles, nous allons passer maintenant aux méthodes d'analyse à proprement parler.

## II. – Effet des caractéristiques individuelles et agrégées sur les comportements sans faire intervenir d'aléas

Nous allons d'abord envisager l'effet simultané de caractéristiques des individus et des divers niveaux d'agrégation sur un comportement donné, sans faire intervenir d'aléas correspondant à ces niveaux. Nous identifions, dans ce cas, les facteurs contextuels qui traduisent les conditions d'insertion des individus dans les divers niveaux d'agrégation. Nous travaillons ici, à titre d'exemple, sur les migrations d'une région norvégienne à l'autre.

### *Analyse des flux d'émigration*

L'utilisation des données du *Registre de population norvégien* nous a permis, lors d'un précédent travail, de montrer l'importance des effets d'agrégation. Nous avions alors pu vérifier les liens, tout d'abord établis de manière théorique, entre des estimations réalisées à partir de divers types

de modèles : régression exponentielle expliquant les taux d'émigration hors des régions, modèles logit ou biographique expliquant les chances individuelles de migrer par les caractéristiques des zones considérées et des individus eux-mêmes (Baccaïni, Courgeau, 1996a).

*Les données utilisées*

La Norvège dispose de registres décentralisés de population, dans lesquels sont enregistrés les événements démographiques des individus vivant dans le pays, en particulier leurs migrations internes (changements de municipalités). Ce registre fut centralisé et informatisé en 1964, pour toutes les personnes résidant en Norvège au recensement du 1er novembre 1960. Les données biographiques de ce registre ont alors été couplées aux informations recueillies lors des recensements de 1960, 1970 et 1980.

Nous avons travaillé sur un fichier regroupant les 54 814 individus nés en 1958, vivant en Norvège en 1991 et n'ayant pas fait de migration vers l'étranger. Pour chacun de ces individus, nous connaissons les changements de régions successifs (la Norvège est divisée en 19 régions, voir figure 1). Nous n'avions alors considéré que les flux d'émigration des régions, observés sur une courte période de deux ans, 1980 et 1981, les individus étant alors âgés de 22-23 ans.

Un recensement ayant eu lieu en 1980, nous connaissions les diverses caractéristiques des individus à cette date et nous avons également pu déterminer la durée écoulée depuis l'installation dans la région où l'individu réside début 1980.

Huit caractéristiques avaient été retenues, au niveau individuel, comme pouvant avoir un effet sur les chances de quitter la région : le statut matrimonial (marié/non marié), l'exercice d'une activité (actif/non actif), le type d'activité professionnelle (agriculteur/non-agriculteur), le niveau d'éducation (plus de 12 ans d'études/moins de 13 ans d'études), la présence d'enfants (au moins un enfant/pas d'enfant), et le niveau de revenu (haut revenu/bas revenu/aucun revenu).

Nous avions alors pu reconstituer les caractéristiques agrégées des 19 régions (pourcentage d'individus ayant quitté la région en 1980-1981, pourcentage d'individus mariés, pourcentage d'agriculteurs, etc.).

Considérant, dans un premier temps, l'effet des seules caractéristiques agrégées, nous avons montré la proximité des résultats obtenus à partir des trois modèles suivants : une régression exponentielle expliquant les taux d'émigration, un modèle logit et un modèle biographique expliquant les probabilités individuelles de migrer. Le modèle biographique présentait cependant une meilleure précision en faisant intervenir la durée de séjour dans la région de départ.

L'effet des caractéristiques individuelles est, par contre, indépendant de celui des caractéristiques agrégées. Ainsi, si l'on fait intervenir simultanément les caractéristiques des régions et les caractéristiques individuelles dans un

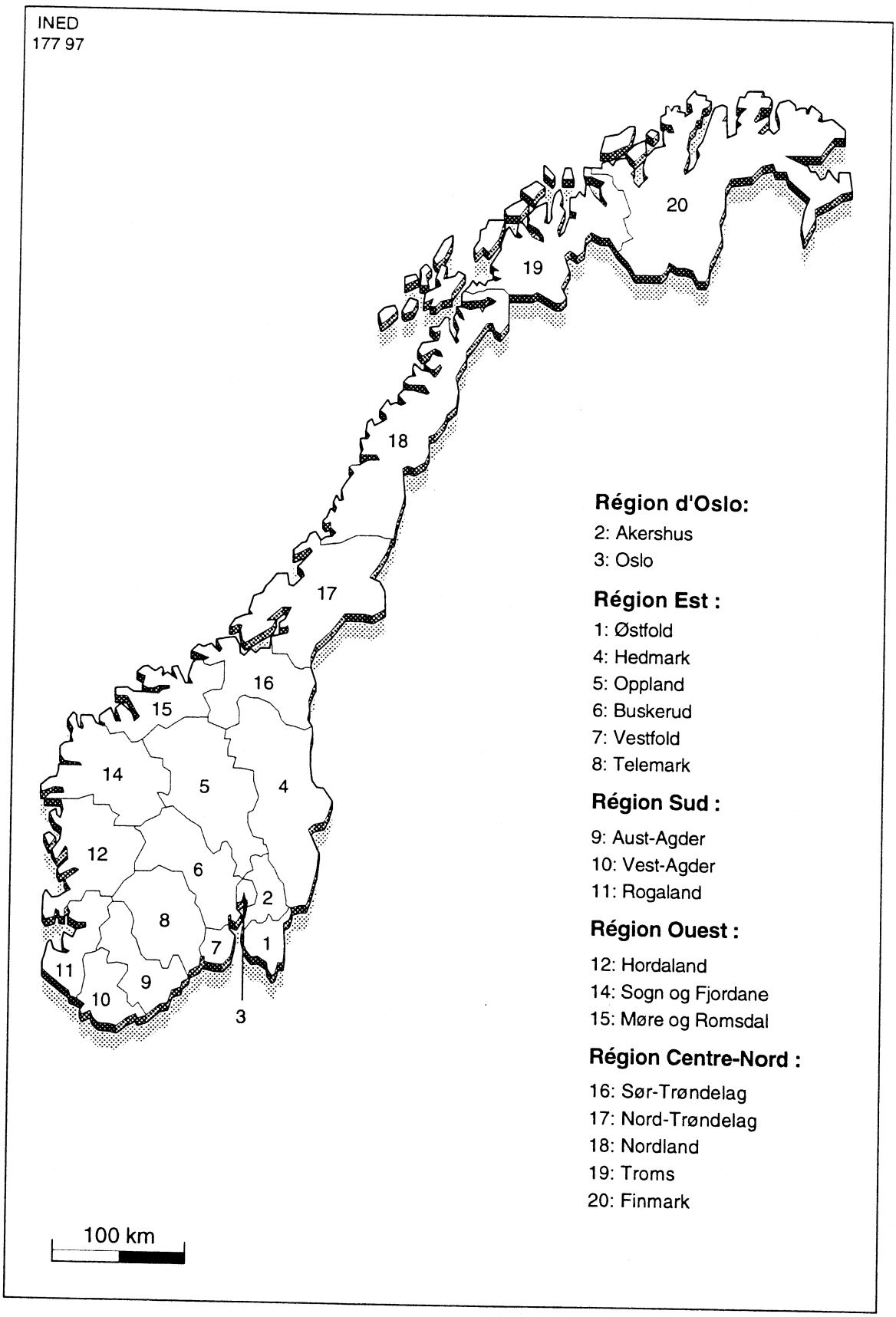


*Figure 1.* – Le découpage régional de la Norvège

modèle logit ou biographique, on observe des effets très différents des deux types de caractéristiques, jouant parfois dans un sens contraire (paramètres de signe opposé). Ainsi, si le fait d'être marié, pour un homme, accroît sa probabilité de quitter sa région de résidence à 22 ans, cette probabilité d'émigration diminue quand la proportion d'hommes mariés, dans leur région, augmente (tableau 9).

À l'inverse, le fait d'être agriculteur, pour un homme, réduit ses chances de quitter sa région, mais cette probabilité augmente avec la proportion d'agriculteurs dans la région.

Nous avons montré comment cet apparent paradoxe s'explique dès que l'on décompose la population soumise au risque en deux groupes disjoints : les mariés et les non-mariés, les agriculteurs et les non-agriculteurs, etc. (voir III. et la figure 6, schéma 1.b).

L'indépendance entre l'effet des caractéristiques « macro » et celui des caractéristiques « micro » a pu être confirmée avec le calcul des coefficients de corrélation entre les paramètres micro et macro : ces corrélations sont en général très faibles, de l'ordre de – 0,10.

***Le cas plus complexe des flux entre régions***

Nous avons poursuivi ce travail sur les données norvégiennes en considérant les flux entre régions plutôt que l'émigration seule. Il s'agit donc non seulement d'expliquer les chances de quitter sa région, mais également celles de choisir une destination plutôt qu'une autre.

Les 19 régions ont, pour ce faire, été regroupées en 5 grandes régions (Oslo, Est, Sud, Ouest, Centre-Nord), afin de limiter le nombre de flux (figure 1).

Par rapport à notre travail antérieur, certaines améliorations ont été apportées. Nous avons, en particulier, considéré les migrations effectuées au cours des années 1981 et 1982 (et non 1980-1981). Le recensement a, en effet, eu lieu au mois de novembre 1980 et si l'on veut connaître les caractéristiques exactes des individus avant leur migration, il est plus judicieux de considérer les années 1981-1982.

De nouvelles caractéristiques peuvent intervenir au niveau individuel, en particulier les informations que les individus peuvent avoir sur les autres régions, leurs relations avec les destinations potentielles. Nous avons ainsi considéré le fait d'avoir ou non vécu antérieurement dans les régions de destination potentielles, la durée du (ou des) séjour(s) passé(s) antérieurement dans ces régions, et la durée écoulée depuis la fin du dernier séjour (âge de l'individu si aucun séjour).

Nous posons ainsi comme hypothèse qu'un individu a d'autant plus de chances de se diriger vers une région qu'il y a déjà résidé antérieurement, que ce séjour a été long et qu'il s'est produit il y a peu de temps.

Au niveau agrégé, de nouvelles variables doivent également intervenir, pour expliquer les flux entre régions.

On peut agréger les trois variables définies plus haut et introduire les pourcentages d'individus de la région d'origine ayant vécu dans les autres régions (destinations potentielles), la durée moyenne passée dans les diverses régions de destination par les individus vivant dans la région d'origine, la durée moyenne écoulée depuis la fin du dernier séjour dans les diverses régions de destination, pour tous les individus de la région d'origine.

Si l'on suppose que les individus ont des relations privilégiées avec des individus présentant les mêmes caractéristiques socio-démographiques qu'eux, il faudra alors faire intervenir les proportions d'individus ayant les mêmes caractéristiques, dans les régions de destination potentielles (pourcentage d'individus de même statut matrimonial, pourcentage d'individus de même profession, pourcentage d'individus de même niveau d'éducation). Ces variables combinent ainsi une dimension macro (elles sont mesurées au niveau des régions) et une dimension micro (elles dépendent des caractéristiques propres des individus). Elles ne peuvent donc être utilisées que dans les modèles estimant des probabilités individuelles de migrer (nous utiliserons ici le modèle biographique).

On pourra enfin faire intervenir la distance géographique entre les régions, son rôle d'obstacle à la migration ayant souvent été montré[1].

Dans un premier temps, nous n'avons considéré, dans un modèle biographique, que les caractéristiques individuelles, afin, en particulier, de tester l'effet des séjours antérieurs sur le choix d'une destination. L'analyse a donc été désagrégée en cinq modèles à risque compétitif (un modèle par région d'origine, les individus ayant le choix entre quatre destinations). Nous ne présentons ici que les résultats concernant la région d'Oslo et le Centre-Nord (tableaux 1a et 1b).

La variable binaire « avoir effectué un séjour antérieur dans la région » et les variables discrètes « durée des séjours antérieurs » et « durée écoulée depuis la fin du dernier séjour » sont fortement redondantes. Ainsi, la première, qui joue très significativement lorsqu'elle est introduite seule, n'a plus d'effet lorsque les deux autres sont également prises en compte. Ces trois caractéristiques, très corrélées, ne doivent donc pas être introduites simultanément dans les modèles. Les résultats présentés dans les tableaux ne considéreront que l'effet du séjour antérieur mais les effets des durées ont également été testés.

Avoir vécu antérieurement dans une région donnée accroît de manière significative les chances d'y retourner à 23-24 ans, et cela d'autant plus que le (ou les) séjour(s) antérieur(s) ont été longs. Les individus gardent donc des liens privilégiés avec les régions dans lesquelles ils ont vécu plus jeunes et y retournent plus facilement.

Dans de nombreux cas, la durée écoulée depuis la fin du dernier séjour dans une région de destination potentielle a également un effet signi-

[1] Afin d'estimer les distances entre régions, nous avons considéré les distances kilométriques séparant les villes les plus importantes de chaque région.

ficatif sur les chances d'y retourner : plus le séjour eu lieu il y a longtemps, moins l'individu a de chances de retourner dans la région, les liens s'étant alors fortement relâchés. Cet effet est particulièrement significatif pour les migrations d'Olso vers l'Est ou l'Ouest, d'une part, et du Nord vers Oslo ou l'Ouest, d'autre part.

TABLEAU 1a. – EFFET DES CARACTÉRISTIQUES INDIVIDUELLES SUR UNE MIGRATION INTERRÉGIONALE EN 1981-1982 (GÉNÉRATION 1958, RÉSIDANT DANS LA RÉGION D'OSLO FIN 1980)

| Caractéristiques des individus | Région destination Est | | Région destination Sud | | Région destination Ouest | | Région destination Nord | |
|---|---|---|---|---|---|---|---|---|
| | Paramètre estimé | Écart type | Paramètre estimé | Écart type | Paramètre estimé | Écart type | Paramètre estimé | Écart type |
| Homme | -0,35*** | 0,09 | -0,04 | 0,20 | -0,27 | 0,18 | -0,11 | 0,14 |
| Actif | -0,10 | 0,12 | -0,15 | 0,23 | -0,15 | 0,22 | -0,36** | 0,16 |
| Marié | -0,06 | 0,12 | -0,14 | 0,24 | -0,11 | 0,22 | 0,08 | 0,17 |
| Présence d'enfant(s) | 0,11 | 0,16 | -0,71 | 0,54 | -0,04 | 0,36 | -0,44 | 0,27 |
| Agriculteur | 0,83*** | 0,25 | -0,23 | 1,01 | 1,05*** | 0,43 | 0,90*** | 0,35 |
| < 10 ans d'études | 0,02 | 0,12 | -0,40 | 0,38 | 0,12 | 0,26 | -0,15 | 0,20 |
| > 12 ans d'études | -0,29** | 0,13 | -0,30 | 0,24 | 0,24 | 0,21 | -0,07 | 0,17 |
| Aucun revenu | -0,10 | 0,25 | -0,34 | 0,63 | -0,40 | 0,63 | 0,41 | 0,31 |
| Revenu < 20 000 cour. | 0,08 | 0,14 | 0,13 | 0,25 | 0,68*** | 0,24 | 0,15 | 0,18 |
| Revenu > 50 000 cour. | 0,15 | 0,11 | -0,44* | 0,24 | 0,15 | 0,22 | -0,11 | 0,16 |
| Séjour antérieur dans région de destination | 1,65*** | 0,14 | 3,24*** | 0,27 | 2,58*** | 0,22 | 2,26*** | 0,19 |

*** significatif au seuil de 1 % ; ** significatif au seuil de 5 % ; * significatif au seuil de 10 %.
*Source* : *Registre de population norvégien*, Central Bureau of Statistics, Oslo.

TABLEAU 1b. – EFFET DES CARACTÉRISTIQUES INDIVIDUELLES SUR UNE MIGRATION INTERRÉGIONALE EN 1981-1982 (GÉNÉRATION 1958, RÉSIDANT DANS LA RÉGION CENTRE-NORD FIN 1980)

| Caractéristiques des individus | Région destination Oslo | | Région destination Est | | Région destination Sud | | Région destination Ouest | |
|---|---|---|---|---|---|---|---|---|
| | Paramètre estimé | Écart type | Paramètre estimé | Écart type | Paramètre estimé | Écart type | Paramètre estimé | Écart type |
| Homme | -0,26** | 0,11 | -0,13 | 0,15 | 0,10 | 0,21 | -0,16 | 0,16 |
| Actif | -0,35*** | 0,12 | -0,22 | 0,16 | -0,22 | 0,23 | -0,33* | 0,18 |
| Marié | -0,25* | 0,15 | 0,07 | 0,17 | 0,17 | 0,25 | -0,19 | 0,21 |
| Présence d'enfant(s) | -1,11*** | 0,21 | -0,41* | 0,21 | -0,32 | 0,29 | -0,88*** | 0,27 |
| Agriculteur | -0,06 | 0,25 | 0,27 | 0,29 | -0,19 | 0,52 | 0,03 | 0,35 |
| < 10 ans d'études | -0,43*** | 0,15 | -0,10 | 0,18 | -0,47 | 0,30 | -0,08 | 0,21 |
| > 12 ans d'études | 0,32** | 0,15 | -0,09 | 0,23 | -0,08 | 0,34 | 0,13 | 0,22 |
| Aucun revenu | -0,19 | 0,23 | -0,29 | 0,34 | -0,40 | 0,44 | 0,10 | 0,31 |
| Revenu < 20 000 cour. | 0,05 | 0,14 | 0,13 | 0,19 | -0,01 | 0,26 | 0,13 | 0,21 |
| Revenu > 50 000 cour. | -0,06 | 0,12 | 0,21 | 0,17 | -0,21 | 0,24 | -0,14 | 0,19 |
| Séjour antérieur dans région de destination | 1,26*** | 0,18 | 1,73*** | 0,21 | 2,32*** | 0,28 | 1,93*** | 0,23 |

*** significatif au seuil de 1 % ; ** significatif au seuil de 5 % ; * significatif au seuil de 10 %.
*Source* : *Registre de population norvégien*, Central Bureau of Statistics, Oslo.

Lorsqu'elles ont un effet significatif, les diverses caractéristiques individuelles jouent le plus souvent dans le même sens, quelles que soient les régions d'origine et de destination : faibles probabilités de changer de région pour les hommes, les actifs, les individus mariés avec des enfants, les individus avec un faible niveau d'éducation et ceux qui disposent d'un haut revenu. On retrouve les effets déjà observés pour la probabilité de quitter sa région (Baccaïni et Courgeau, 1996a ; 1996b).

On observe, cependant, quelques effets qui jouent différemment selon la direction de la migration. Ainsi, avoir des enfants, qui est en général un frein à la migration, accroît les chances des individus d'Oslo de se diriger vers la région Est (déconcentration de la région urbaine d'Oslo). De la même manière, alors que les individus de faible niveau d'éducation ont en général de faibles chances de quitter leur région, c'est l'inverse pour les migrations de la région Sud vers la région Est.

Le fait d'être agriculteur a également un effet très différent selon la région d'origine des individus. Les agriculteurs originaires de la région d'Oslo ont de fortes chances d'en partir (pour se diriger vers l'Est, l'Ouest ou le Nord), alors que dans les autres régions, au contraire, être agriculteur est un frein à la migration.

Certaines caractéristiques individuelles jouent donc dans un sens opposé, selon que l'on considère la migration de la région $i$ vers la région $j$ ou celle de $j$ vers $i$. Si l'on considère les deux flux symétriques entre la région d'Oslo et celle du Centre-Nord, quatre caractéristiques jouent en sens inverse selon la direction de la migration : être marié, être agriculteur, avoir fait de longues études et ne disposer d'aucun revenu.

Dans un deuxième temps, nous avons élaboré divers types de modèles prenant en compte les caractéristiques des régions d'origine et/ou de destination.

L'introduction de ces caractéristiques agrégées oblige à poser des hypothèses sur les processus pouvant conduire un individu à quitter une région $i$ pour se rendre dans une région $j$. La question essentielle peut se résumer ainsi : la décision de quitter sa région de résidence est-elle antérieure ou postérieure au choix d'une région de destination ? Autrement dit, la décision de quitter la région $i$ est-elle une conséquence du souhait de vivre dans la région $j$ ou est-ce le choix de la région $j$ qui est une conséquence du souhait de quitter la région $i$ ?

Les deux processus se combinent probablement et nous considérerons que ce sont les avantages ou les inconvénients que présente la région d'origine par rapport aux régions de destination potentielles qui vont ou non pousser l'individu à effectuer sa migration, la décision dépendant également de ses caractéristiques individuelles.

L'objectif est donc d'étudier les chances que les individus ont de se diriger vers les diverses destinations qui leur sont offertes, en fonction de leurs caractéristiques individuelles, et des avantages que peut présenter cette région par rapport à celle dans laquelle ils vivent.

Pour ce faire, nous avons tout d'abord considéré un modèle par région de destination, la population soumise au risque de s'y diriger étant celle des quatre autres régions. Les variables macro à faire intervenir, afin de mettre en évidence l'effet des relations avec des individus présentant des caractéristiques socio-démographiques proches, seront les rapports entre le pourcentage d'individus de même caractéristique dans la région de destination et ce pourcentage dans la région d'origine de l'individu.

Nous présentons les résultats pour deux régions de destination : la région d'Oslo et le Centre-Nord.

Introduites seules dans un modèle biographique, la plupart des caractéristiques agrégées ont un effet très significatif sur les chances de migrer vers Oslo, un peu moins sur celles de migrer vers le Centre-Nord (tableau 2).

TABLEAU 2. – EFFET DES CARACTÉRISTIQUES DES RÉGIONS SUR UNE MIGRATION INTERRÉGIONALE EN 1981-1982, SELON LA RÉGION DE DESTINATION (GÉNÉRATION 1958)

| Caractéristiques des régions d'origine et de destination | Destination Oslo | | Destination Centre-Nord | |
|---|---|---|---|---|
| | Paramètre estimé | Écart type | Paramètre estimé | Écart type |
| Rapport % individus de même statut matrimonial (a) | 2,37*** | 0,17 | 1,35*** | 0,23 |
| Rapport % individus de même niveau éducation (a) | 0,81*** | 0,07 | – 0,63*** | 0,20 |
| Rapport % individus de même profession (a) | 0,36*** | 0,06 | 0,21*** | 0,06 |
| Distance origine-destination | 0,001** | 0,000 | – 0,001* | 0,00 |
| % population ayant vécu dans région destination | 0,16*** | 0,02 | 0,00 | 0,02 |

(a) : rapport entre le % dans la région de destination et le % dans la région d'origine.
*** significatif au seuil de 1 % ; ** significatif au seuil de 5 % ; * significatif au seuil de 10 %.
*Source* : *Registre de population norvégien,* Central Bureau of Statistics, Oslo.

Les chances de se diriger vers chacune de ces deux régions sont d'autant plus élevées que le poids relatif des individus de même statut matrimonial et celui des individus exerçant la même profession sont importants par rapport à ce qu'ils sont dans la région d'origine de l'individu, ce qui semble confirmer nos hypothèses.

Par contre, l'effet du poids relatif des individus de même niveau d'éducation joue différemment selon la région de destination. Oslo attire préférentiellement des individus dont le niveau d'éducation est moins représenté dans leur région d'origine qu'à Oslo, alors que la région Centre-Nord attire préférentiellement des individus dont le niveau de diplôme était plus fréquent dans leur région d'origine qu'il ne l'est dans le Centre-Nord. Les individus ayant un niveau d'éducation élevé ont, en effet, plus de chances de migrer que les autres, or, le poids des individus hautement qualifiés est plus important dans la région d'Oslo que dans le Centre-Nord.

La distance est un obstacle pour les migrants vers le Centre-Nord, cette région attirant donc préférentiellement des individus originaires de régions proches, ce qui n'est pas le cas pour les migrants vers Oslo, qui semblent peu sensibles à la distance.

Par contre, l'importance et la proximité dans le temps des séjours antérieurs effectués par la population de la région d'origine (expression de la force des liens entre les deux régions) n'ont un effet significatif que pour les migrations vers la région d'Oslo.

Les effets de certaines de ces caractéristiques agrégées changent lorsqu'on introduit également les caractéristiques individuelles dans le modèle (tableau 3). Ainsi, en particulier, l'effet du poids relatif des individus de même niveau d'éducation dans la région de destination et dans celle d'origine s'inverse, une fois pris en compte le niveau d'éducation des individus.

TABLEAU 3. – EFFET DES CARACTÉRISTIQUES DES INDIVIDUS ET DES RÉGIONS SUR UNE MIGRATION INTERRÉGIONALE EN 1981-1982, SELON LA RÉGION DE DESTINATION (GÉNÉRATION 1958)

| Caractéristiques des individus et des régions d'origine et de destination | Destination Oslo | | Destination Centre-Nord | |
|---|---|---|---|---|
| | Paramètre estimé | Écart type | Paramètre estimé | Écart type |
| Homme | – 0,18*** | 0,05 | – 0,04 | 0,08 |
| Actif | – 0,33*** | 0,06 | – 0,37*** | 0,08 |
| Marié | – 0,36** | 0,16 | – 0,24* | 0,14 |
| Présence d'enfant(s) | – 1,00*** | 0,12 | – 0,59*** | 0,15 |
| Agriculteur | – 0,12 | 0,15 | 0,13 | 0,21 |
| < 10 ans d'études | – 0,57*** | 0,09 | – 0,46*** | 0,13 |
| > 12 ans d'études | 0,78*** | 0,20 | 0,59*** | 0,19 |
| Aucun revenu | 0,00 | 0,12 | 0,29* | 0,17 |
| Revenu < 20 000 cour. | 0,39*** | 0,06 | 0,64*** | 0,10 |
| Revenu > 50 000 cour. | – 0,30*** | 0,06 | – 0,14 | 0,10 |
| Séjour antérieur dans région destination | 0,76*** | 0,09 | 2,01*** | 0,11 |
| Rapport % individus de même statut matrimonial (a) | 1,21*** | 0,35 | 0,49 | 0,34 |
| Rapport % individus de même niveau éducation (a) | – 0,55** | 0,25 | 1,35*** | 0,48 |
| Rapport % individus de même profession (a) | 0,36*** | 0,08 | 0,07 | 0,11 |
| Distance origine-destination | 0,001* | 0,000 | 0,000 | 0,00 |
| % population ayant vécu dans région destination | 0,16*** | 0,02 | – 0,01 | 0,02 |

(a) : rapport entre le % dans la région de destination et le % dans la région d'origine.
*** significatif au seuil de 1 % ; ** significatif au seuil de 5 % ; * significatif au seuil de 10 %.
*Source* : *Registre de population norvégien,* Central Bureau of Statistics, Oslo.

Nous avons ensuite changé de perspective, en reprenant le modèle à risques compétitifs. La question est alors de voir comment, pour les individus d'une région donnée, les avantages relatifs offerts par les autres régions peuvent le pousser à migrer. On peut alors comparer les comportements des migrants originaires des diverses régions.

TABLEAU 4. – EFFET DES CARACTÉRISTIQUES INDIVIDUELLES ET AGRÉGÉES SUR UNE MIGRATION INTERRÉGIONALE EN 1981-1982 (GÉNÉRATION 1958 RÉSIDANT DANS LA RÉGION D'OSLO FIN 1980)

| Caractéristiques des individus et des régions | Région destination Est | | Région destination Sud | | Région destination Ouest | | Région destination Nord | |
|---|---|---|---|---|---|---|---|---|
| | Paramètre estimé | Écart type | Paramètre estimé | Écart type | Paramètre estimé | Écart type | Paramètre estimé | Écart type |
| Homme | -0,31*** | 0,09 | 0,01 | 0,20 | -0,36* | 0,18 | -0,12 | 0,13 |
| Actif | -0,14 | 0,12 | -0,24 | 0,25 | -0,08 | 0,22 | -0,43*** | 0,16 |
| Marié | -0,17 | 0,14 | -0,24 | 0,33 | -0,15 | 0,27 | 0,05 | 0,20 |
| Présence d'enfant(s) | 0,18 | 0,16 | -0,99* | 0,54 | -0,21 | 0,36 | -0,45 | 0,27 |
| Agriculteur | 1,22*** | 0,27 | -0,41 | 1,05 | 0,91* | 0,50 | 0,39 | 0,37 |
| < 10 ans d'études | 0,01 | 0,12 | -0,61 | 0,38 | 0,08 | 0,26 | -0,18 | 0,20 |
| > 12 ans d'études | 0,02 | 0,15 | -0,32 | 0,34 | 0,27 | 0,27 | -0,22 | 0,23 |
| Aucun revenu | 0,00 | 0,25 | -0,48 | 0,63 | -0,53 | 0,63 | 0,20 | 0,49 |
| Revenu < 20 000 cour. | 0,08 | 0,14 | 0,26 | 0,25 | 0,60** | 0,24 | 0,11 | 0,19 |
| Revenu > 50 000 cour. | 0,20* | 0,11 | -0,54** | 0,23 | 0,13 | 0,22 | -0,14 | 0,16 |
| Séjour antérieur dans région destination | 0,43*** | 0,09 | 0,37* | 0,21 | 0,37** | 0,18 | 0,47*** | 0,14 |
| Rapport % individus même statut matrimonial (a) | 0,40 | 0,30 | 1,18** | 0,55 | 0,19 | 0,54 | -0,72 | 0,48 |
| Rapport % individus même niveau éducation (a) | 1,54*** | 0,50 | -0,87 | 0,87 | -0,18 | 0,71 | -0,51 | 0,62 |
| Rapport % individus même profession (a) | -0,32* | 0,17 | -0,39 | 0,44 | 0,08 | 0,21 | 0,38*** | 0,08 |

(a) rapport entre le % dans la région de destination et le % dans la région d'origine.
*** significatif au seuil de 1 % ; ** significatif au seuil de 5 % ; * significatif au seuil de 10 %.
*Source : Registre de population norvégien,* Central Bureau of Statistics, Oslo.

TABLEAU 5. – EFFET DES CARACTÉRISTIQUES INDIVIDUELLES ET AGRÉGÉES SUR UNE MIGRATION INTERRÉGIONALE EN 1981-1982 (GÉNÉRATION 1958, RÉSIDANT DANS LA RÉGION CENTRE-NORD FIN 1980)

| Caractéristiques des individus et des régions | Région destination Oslo | | Région destination Est | | Région destination Sud | | Région destination Ouest | |
|---|---|---|---|---|---|---|---|---|
| | Paramètre estimé | Écart type | Paramètre estimé | Écart type | Paramètre estimé | Écart type | Paramètre estimé | Écart type |
| Homme | -0,40*** | 0,11 | -0,16 | 0,15 | 0,11 | 0,21 | -0,11 | 0,16 |
| Actif | -0,24** | 0,12 | -0,09 | 0,16 | -0,26 | 0,24 | -0,30 | 0,19 |
| Marié | -0,28* | 0,17 | 0,19 | 0,19 | -0,16 | 0,30 | 0,08 | 0,22 |
| Présence d'enfant(s) | -1,11*** | 0,21 | -0,37* | 0,21 | -0,09 | 0,29 | -0,88*** | 0,27 |
| Agriculteur | -0,21 | 0,25 | -0,01 | 0,31 | -0,31 | 0,53 | 0,11 | 0,36 |
| < 10 ans d'études | -0,37** | 0,15 | -0,15 | 0,19 | -0,50 | 0,30 | -0,12 | 0,21 |
| > 12 ans d'études | -0,48** | 0,24 | 0,30 | 0,28 | -0,25 | 0,44 | 0,38 | 0,28 |
| Aucun revenu | -0,20 | 0,23 | -0,34 | 0,34 | -0,59 | 0,45 | 0,26 | 0,31 |
| Revenu < 20 000 cour. | 0,01 | 0,14 | 0,05 | 0,19 | -0,08 | 0,26 | 0,14 | 0,21 |
| Revenu > 50 000 cour. | -0,08 | 0,12 | 0,23 | 0,17 | -0,28 | 0,24 | -0,13 | 0,19 |
| Séjour antérieur dans région destination | -0,88*** | 0,13 | -0,44*** | 0,15 | -0,91*** | 0,23 | -0,44*** | 0,17 |
| Rapport % individus même statut matrimonial (a) | 0,46 | 0,47 | 0,14 | 0,50 | 2,58*** | 0,55 | -1,37* | 0,74 |
| Rapport % individus même niveau éducation (a) | 1,21*** | 0,27 | -1,38** | 0,62 | -0,66 | 0,88 | -0,14 | 0,43 |
| Rapport % individus même profession (a) | -0,68** | 0,34 | -2,13*** | 0,49 | -0,18 | 0,70 | 0,29 | 0,49 |

(a) rapport entre le % dans la région de destination et le % dans la région d'origine.
*** significatif au seuil de 1 % ; ** significatif au seuil de 5 % ; * significatif au seuil de 10 %.
*Source : Registre de population norvégien,* Central Bureau of Statistics, Oslo.

Nous présentons les résultats pour deux régions d'origine : la région d'Oslo et le Centre-Nord (tableaux 4 et 5).

L'introduction des caractéristiques agrégées, dans ces modèles à risques compétitifs, modifie la valeur de certains paramètres associés aux caractéristiques individuelles (par rapport à ce qu'ils étaient dans le modèle micro vu plus haut). Pour les migrants originaires d'Oslo, par exemple, le fait d'être agriculteur n'a plus d'effet significatif sur la propension à se diriger vers les régions Ouest ou Centre-Nord. Pour les individus originaires du Centre-Nord, l'effet d'un niveau d'éducation élevé sur les chances de migrer vers Oslo passe de positif à très négatif, une fois introduites les caractéristiques agrégées.

Les diverses caractéristiques agrégées jouent très différemment, selon la région de destination, pour les individus originaires d'une région donnée. Ainsi, l'attraction de la région Est sur les originaires d'Oslo s'exerce préférentiellement sur ceux dont le niveau d'éducation est fortement représenté dans l'Est par rapport à ce qu'il est à Oslo, alors que c'est l'inverse pour les migrations d'Oslo vers les autres régions. La région Est attire, par contre, préférentiellement les originaires d'Oslo dont la profession est peu représentée dans l'Est, par rapport à son poids dans la région d'Oslo, l'inverse étant observé pour les migrations d'Oslo vers le Centre-Nord.

Le même type d'observations peut être fait à propos des migrations d'individus originaires du Centre-Nord.

Ces premières analyses, qui devront être poussées plus avant, montrent l'intérêt que peut présenter, pour la compréhension des processus migratoires, la prise en compte simultanée des caractéristiques des individus et de celles des zones d'origine et de destination dans les modèles. Elles mettent également en évidence la prudence qu'impose l'interprétation des résultats.

## III. – Analyse avec des aléas multi-niveaux

L'analyse précédente ne considérait que cinq régions, que l'on aurait pu étudier séparément, étant donné la taille de l'échantillon. Supposons maintenant que l'on augmente le nombre de régions considérées. On peut, dans ce cas, considérer les régions étudiées comme un échantillon, qui va fournir des informations sur les caractéristiques des régions, considérées comme une population de régions. Voyons, plus précisément, ce qu'il en est pour une analyse de régression.

***Analyse de régression*** Reprenons pour ce faire l'exemple très didactique donné par Woodhouse *et al.* (1996). Il s'agit d'une étude longitudinale faite sur une cohorte d'élèves anglais entrant dans un cycle primaire (*junior classes*) à l'âge de 8 ans et suivis

jusqu'à l'entrée en secondaire à l'âge de 11 ans. Ces élèves étudiaient dans une cinquantaine d'écoles primaires tirées aléatoirement parmi 650 écoles londoniennes. L'objectif de cette étude était de déceler si certaines écoles étaient plus efficaces que d'autres pour faciliter les progrès scolaires de leurs élèves. Pour ce faire les auteurs considèrent les progrès en mathématiques, mesurés par des tests passés aux âges de 8 et 11 ans, qu'ils vont considérer à la fois au niveau individuel et au niveau des écoles.

Soit $y_{ij}$ la note à 11 ans obtenue par le $i^{\text{ème}}$ élève de l'école $j$, et $x_{1ij}$ sa note à 8 ans. Ajustons d'abord des régressions linéaires pour chaque école :

$$y_{ij} = a_{oj} + a_{1j}\, x_{1ij} + e_{ij} \qquad [1]$$

où $a_{oj}$ et $a_{1j}$ sont des paramètres ajustés sur la $j^{\text{ème}}$ école, $e_{ij}$ étant le résidu aléatoire de moyenne nulle et de variance $\sigma^2_{ej}$.

S'il y avait peu d'écoles on pourrait estimer autant de paramètres $a_{oj}$ et $a_{1j}$ qu'il y a d'écoles, en supposant par exemple une variance de $e_{ij}$ indépendante de l'école. La comparaison de ces paramètres caractérisant chaque école est possible, mais ne permet guère de généralisations, car ils s'appliquent aux écoles de l'échantillon et ne fournissent pas d'information sur l'ensemble des écoles.

Si, au contraire, on considère la cinquantaine d'écoles comme un échantillon tiré d'une population de 650 écoles, on peut en déduire des informations de type statistique sur cette plus grande population. Voyons donc comment formaliser plus avant cette analyse faisant intervenir deux niveaux d'agrégation : l'élève et l'école.

On voit qu'une façon d'introduire les écoles dans l'équation [1] est de supposer que les paramètres $a_{oj}$ et $a_{1j}$ sont aléatoires et vont donc varier d'une école à l'autre, ce qui revient à poser :

$$\begin{aligned} a_{oj} &= a_o + e_{oj} \\ a_{1j} &= a_1 + e_{1j} \end{aligned} \qquad [2]$$

où $a_o$ et $a_1$ sont les paramètres moyens ajustés sur toutes les écoles, $e_{oj}$ et $e_{1j}$ des variables aléatoires, de moyenne nulle dont on va estimer les variances et la covariance :

$$\begin{aligned} \text{var}\,(e_{oj}) &= \sigma^2_{eo} \\ \text{var}\,(e_{1j}) &= \sigma^2_{e1} \\ \text{cov}\,(e_{oj}\,,\, e_{1j}) &= \sigma_{eo1} \end{aligned} \qquad [3]$$

On peut dès lors réécrire la formule [1] sous la forme :

$$y_{ij} = a_o + a_1\, x_{1ij} + (e_{oj} + e_{1j}\, x_{1ij} + e_{ij}) \qquad [4]$$

qui est formée d'une partie indépendante de l'école $(a_o + a_1 x_{1ij})$, et d'une partie aléatoire qui dépend à la fois de l'école et de l'individu.

L'estimation des divers paramètres ainsi que des variances et covariances est possible à l'aide de méthodes généralisant celle des moindres carrés (Goldstein, 1986 ; 1995). On arrive ainsi aux estimations du tableau 6, calculées à l'aide du logiciel MLn.

TABLEAU 6. – PARAMÈTRES ET ÉCARTS TYPES ESTIMÉS DANS LE MODÈLE MULTI-NIVEAUX RELIANT LA NOTE DES ÉLÈVES À 11 ANS À CELLE OBTENUE À 8 ANS

| Paramètres | Estimation | Écart type |
|---|---|---|
| **Non aléatoires** | | |
| Constante | 15,040 | 1,318 |
| Note à 8 ans | 0,612 | 0,043 |
| **Aléatoires** | | |
| Niveau école | | |
| $\sigma^2_{e0}$ (constante) | 44,990 | 16,360 |
| $\sigma^2_{e01}$ (covariance) | – 1,231 | 0,521 |
| $\sigma^2_{e1}$ (note à 8 ans) | 0,034 | 0,017 |
| Niveau élève | | |
| $\sigma^2_e$ | 26,960 | 1,343 |

*Source* : Woodhouse, 1996.

Ce tableau permet de constater que tous les effets sont significatifs. En premier lieu, on voit que plus la note de l'individu à 8 ans est importante plus sa note à 11 ans le sera, quelle que soit l'école où il se trouve. Mais, selon l'école, les variances et covariances des variables aléatoires $e_{oj}$ et $e_{1j}$ seront également significatives : le fait que la covariance entre $e_{oj}$ et $e_{1j}$ soit négative indique que plus l'école a une note moyenne élevée, moins la note à 11 ans dépendra de la note obtenue à 8 ans. Cela revient à dire que certaines écoles arriveront bien à mettre tous les élèves d'une classe donnée à un bon niveau mathématique, quelle que soit la note initiale de ces différents élèves; en revanche, d'autres écoles laisseront à la traîne des élèves dont le niveau mathématique de départ est déjà faible.

Un moyen de bien montrer ces différences entre écoles est de porter sur un même graphique les relations linéaires estimées à l'aide d'un modèle multi-niveaux entre les notes à 8 ans et celles à 11 ans dans chaque école. Cela revient à écrire pour chaque école la relation suivante :

$$\hat{y}_{ij} = a_o + e_{oj} + (a_1 + e_{1j})\, x_{1ij} \qquad [5]$$

où les $e_{oj}$ et les $e_{1j}$ sont les résidus par rapport au modèle, calculés sur chaque région $j$. La figure 2 porte ces résultats. On y voit clairement apparaître les écoles extrêmes : dans la partie supérieure de la figure se trouvent celles où la note à 11 ans dépend peu de la note initiale, qui réussissent donc à mettre tous les élèves à un bon niveau ; dans la partie inférieure

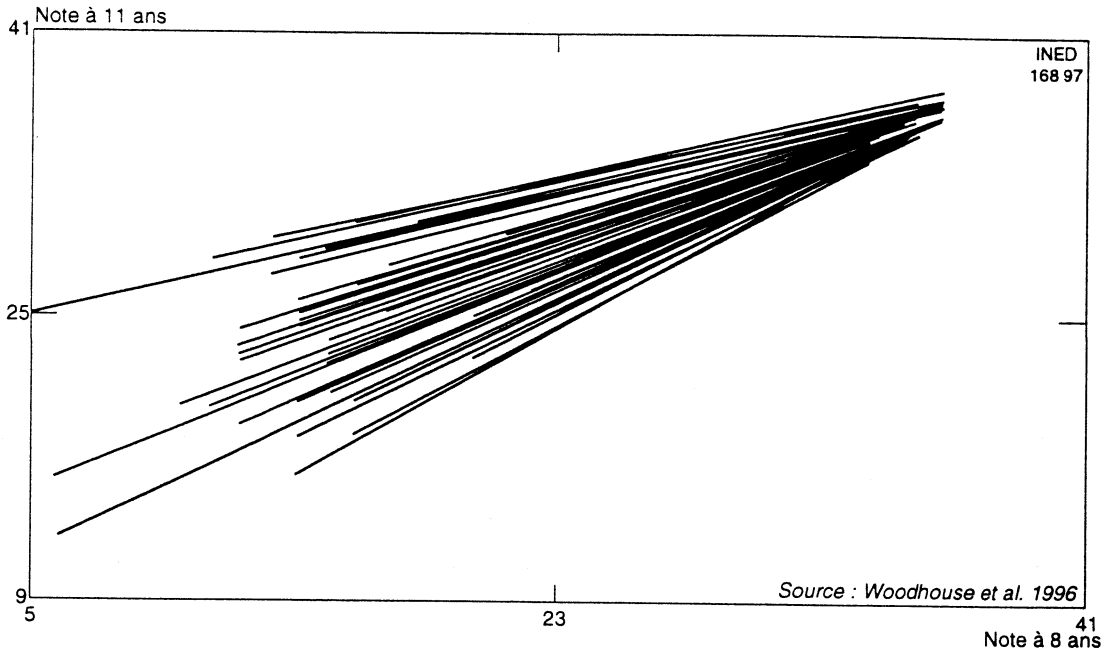


*Figure 2.* – Relations estimées entre les notes à 8 ans et à 11 ans dans chaque école, pour un modèle multi-niveaux appliqué à un échantillon d'écoles londoniennes

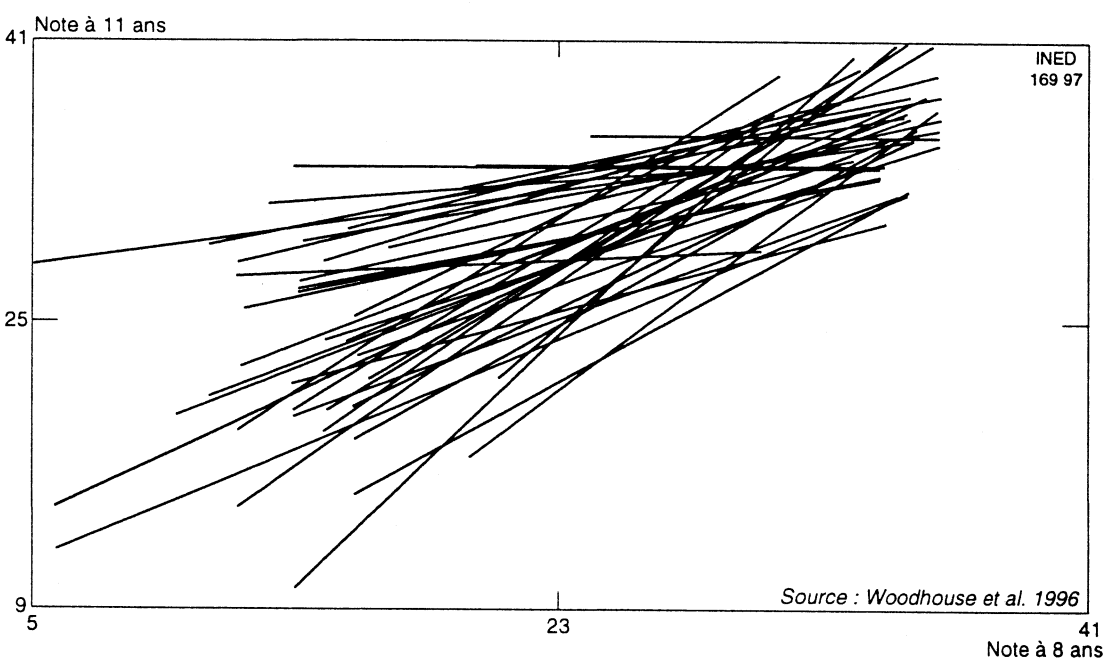


*Figure 3.* – Relations estimées entre les notes à 8 ans et à 11 ans dans chaque école, pour des modèles de régression linéaires appliqués à chaque école londonienne de l'échantillon

se trouvent celles où la note à 11 ans est encore fort dépendante de la note initiale. Il est intéressant de comparer ces résultats à ceux obtenus en appliquant le modèle de régression [1] à chaque école, portés sur la figure 3. Cette figure est beaucoup plus confuse que la précédente, car elle fait intervenir des régressions estimées séparément sur chaque école, certaines d'entre elles ayant un très petit nombre d'élèves, ce qui entraîne une mauvaise estimation des valeurs de $a_{oj}$ et $a_{1j}$.

Bien entendu, il est possible d'introduire diverses autres caractéristiques des individus ou des écoles pour affiner l'analyse (Goldstein, 1995). Ici cependant, nous n'irons pas plus avant dans cette analyse, pour examiner plus en détail les hypothèses à la base de ces modèles.

***Risque d'inférence erronée*** Supposons qu'une seconde caractéristique, $x_{2ij}$, *indépendante* de la première, joue également sur la note à 11 ans. Il pourrait s'agir, par exemple, du soutien assidu des parents en mathématiques, mesuré par une variable binaire, égale à 1 si ce soutien existe et 0 sinon. Supposons enfin qu'il existe une relation indépendante de l'école dans laquelle se trouve l'élève entre sa note à 11 ans et les deux caractéristiques (note à 8 ans et soutien des parents). Cette relation fait donc intervenir des variables aléatoires qui sont également indépendantes de l'école dans laquelle se trouve l'élève. On peut dès lors l'écrire :

$$y_{ij} = a_o + e_{oij} + (a_1 + e_{1ij})\, x_{1ij} + (a_2 + e_{2ij})\, x_{2ij} + (a_{12} + e_{12ij})\, x_{1ij} \times x_{2ij} + e_{ij} \qquad [6]$$

où $e_{oij}$, $e_{1ij}$, $e_{2ij}$, $e_{12j}$, et $e_{ij}$ sont des variables aléatoires de moyenne nulle, de variance $\sigma^2_{eo}$, $\sigma^2_{e1}$, $\sigma^2_{e2}$, $\sigma^2_{e12}$, et $\sigma^2_e$, dont toutes les covariances sont nulles puisqu'elles sont indépendantes de la zone $j$ et indépendantes entre elles. Nous supposons ici que l'effet direct du soutien parental est positif, mais que l'interaction entre note à 8 ans et soutien parental est négative : l'effet est d'autant plus important que l'enfant a obtenu une note faible à 8 ans.

Sous ces conditions, il est possible de simuler des échantillons les vérifiant toutes, dont la taille et le nombre de parents soutenant leurs enfants sont tirés aléatoirement de façon indépendante de l'école. Nous avons ainsi simulé divers échantillons dont les paramètres sont situés dans les intervalles suivants :

$$2 \le a_o + e_{oij} \le 7 \; ; \; 0{,}25 \le a_1 + e_{1ij} \le 1{,}25 \; ;$$

$$21 \le a_2 + e_{2ij} \le 29 \; ; \; -0{,}57 \le a_{12} + e_{12ij} \le -0{,}43$$

Nous présentons ici les résultats obtenus sur un de ces échantillons, très proches de tous les autres.

Si l'on ne dispose d'aucune mesure du soutien parental, on ne peut estimer qu'un modèle où intervient la note à 8 ans, $x_{1ij}$. Dans ce cas, on vérifie facilement que les relations suivantes sont vérifiées :

$$a_o + e_{oij} \leq a'_o + e'_{oj} \leq a_o + a_2 + e_{oij} + e_{2ij}$$

$$a_1 + e_{1ij} \leq a'_1 + e'_{1j} \leq a_1 + a_{12} + e_{1ij} + e_{12ij}$$

où $e'_{oj}$ et $e'_{1j}$ sont des termes aléatoires qui vont donc maintenant dépendre de l'école. On peut dès lors réécrire le modèle précédent sous la forme :

$$y_{ij} = a'_o + e'_{oj} + (a'_1 + e'_{1j})\, x_{1ij} + e_{ij} \qquad [7]$$

Le tableau 7 porte ces résultats : il montre bien un effet de l'école tout à fait significatif et très proche de ce qu'on obtenait dans le tableau 6. Certaines écoles semblent bien mettre tous les élèves à un bon niveau mathématique, quelle que soit leur note initiale, d'autres semblent laisser à la traîne les élèves dont le niveau de départ est faible. Comme précédemment la figure 4 porte l'estimation des notes à 11 ans obtenues pour chaque école en fonction de la note à 8 ans : elle est très proche de la figure 2 et montre un fort effet de l'école, alors qu'en fait il n'existe pas. La figure 5 est parallèle à la figure 3 et montre, comme elle, un effet plus confus de l'école, mais en est également très proche.

TABLEAU 7. – PARAMÈTRES ET ÉCARTS TYPES ESTIMÉS DANS LE MODÈLE MULTI-NIVEAUX SIMULÉ RELIANT LA NOTE DES ÉLÈVES À 11 ANS À CELLE OBTENUE À 8 ANS

| Paramètres | Estimation | Écart type |
|---|---|---|
| **Non aléatoires** | | |
| Constante | 16,720 | 1,189 |
| Note à 8 ans | 0,503 | 0,033 |
| **Aléatoires** | | |
| Niveau école | | |
| $\sigma^2_{e0}$ (constante) | 57,000 | 14,080 |
| $\sigma^2_{e01}$ (covariance) | – 1,298 | 0,373 |
| $\sigma^2_{e1}$ (note à 8 ans) | 0,030 | 0,011 |
| Niveau élève | | |
| $\sigma^2_e$ | 91,730 | 2,977 |

Si l'on essaye de faire intervenir une variable supplémentaire, la note moyenne à 8 ans dans chaque école, parmi les caractéristiques non aléatoires, son effet n'est absolument pas significatif et laisse le modèle inchangé. En revanche dès que l'on fait intervenir les deux caractéristiques individuelles et leur interaction, laissant les termes aléatoires à estimer inchangés, toutes les variances et covariances au niveau école deviennent nulles comme on pouvait s'y attendre. Les paramètres estimés et leur écart type sont portés dans le tableau 8, qui devient parfaitement cohérent avec les simulations faites.

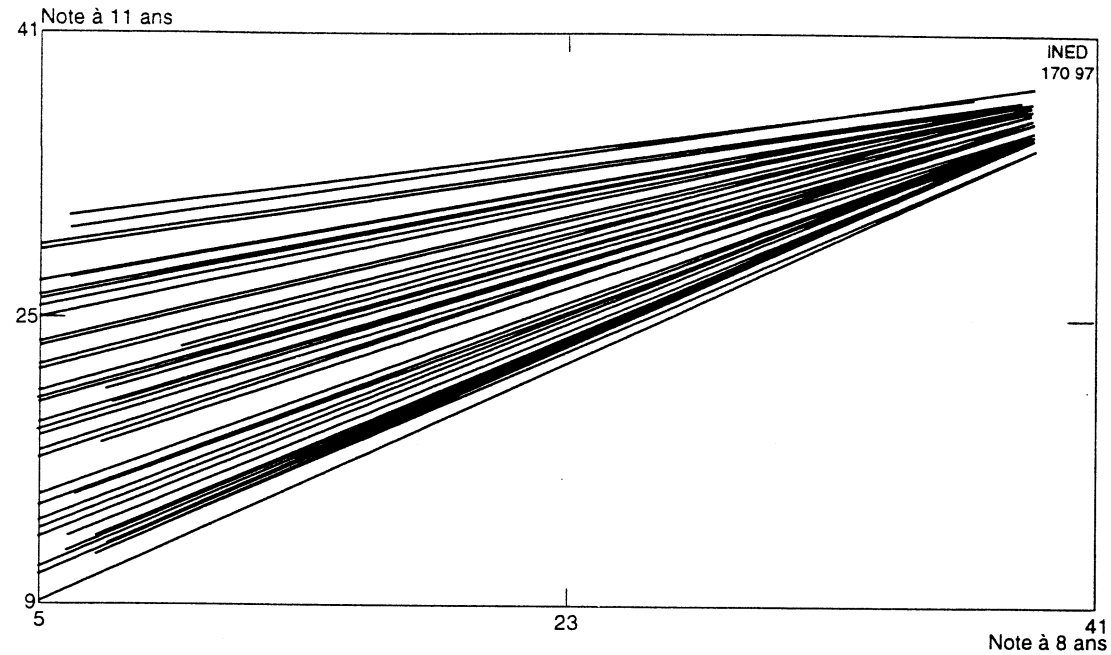


*Figure 4.* – Relations estimées entre les notes à 8 ans et à 11 ans dans chaque école, pour un modèle multi-niveaux appliqué à un échantillon simulé d'écoles (modèle 7)

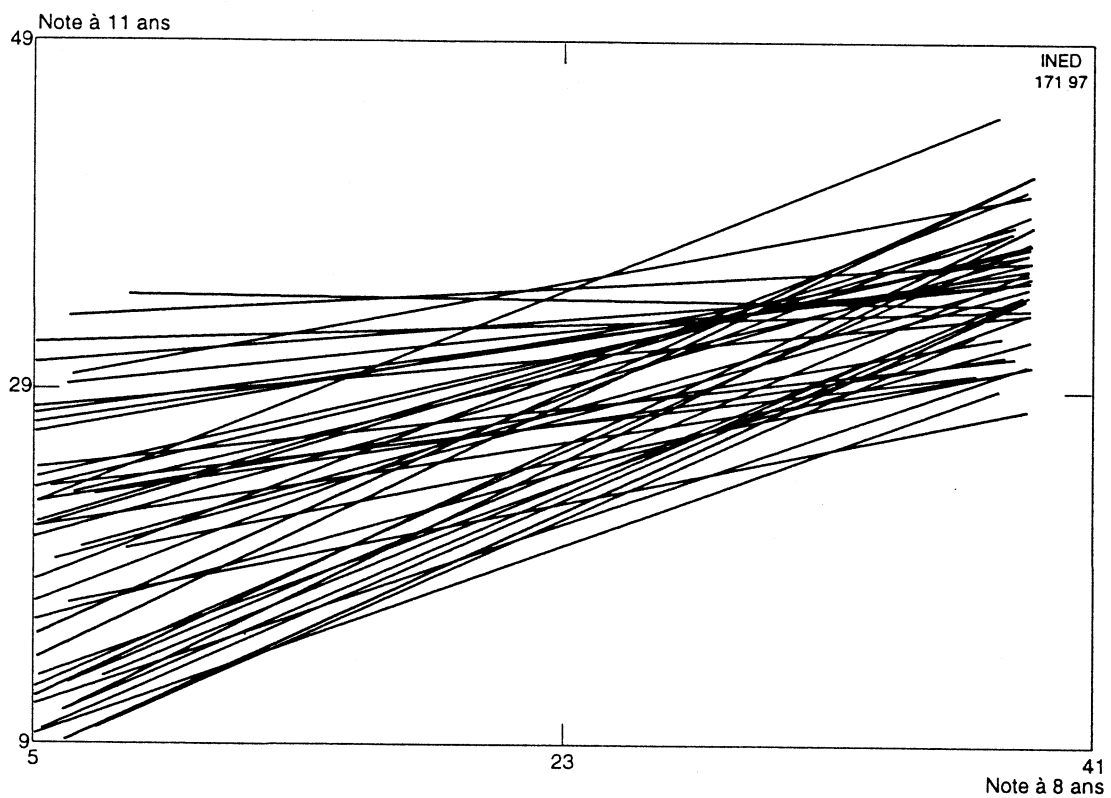


*Figure 5.* – Relations estimées entre les notes à 8 ans et à 11 ans dans chaque école, pour un modèle de régression linéaire appliqué à chaque école de l'échantillon simulé (modèle 7)

TABLEAU 8. – PARAMÈTRES ET ÉCARTS TYPES DANS LE MODÈLE MULTI-NIVEAUX SIMULÉ RELIANT LA NOTE DES ÉLÈVES À 11 ANS À CELLE À 8 ANS, AU SOUTIEN PARENTAL ET À L'INTERACTION ENTRE SOUTIEN ET NOTE À 8 ANS

| Paramètres | Estimation | Écart type |
|---|---|---|
| **Non aléatoires** | | |
| Constante | 4,410 | 0,545 |
| Note à 8 ans | 0,766 | 0,023 |
| Soutien des parents | 25,170 | 0,783 |
| Interaction | – 0,529 | 0,033 |
| **Aléatoires** | | |
| Niveau élève | | |
| $\sigma_e^2$ | 54,840 | 1,757 |

Ainsi les risques d'inférence erronée paraissent être importants lorsque l'on utilise un modèle multi-niveaux, même lorsque la caractéristique omise est indépendante de celles introduites dans le modèle initial. Il est donc nécessaire de poursuivre des recherches dans ce domaine pour mieux assurer les résultats d'une telle analyse. Une bonne précaution est de faire intervenir, dans la partie fixe, le plus grand nombre de caractéristiques ayant un effet sur le phénomène, de façon à éviter au maximum le risque de conclure à un effet d'un niveau d'agrégation alors qu'il n'existe pas dans les faits.

### *Analyse par des modèles dichotomiques*

Un grand nombre de caractéristiques démographiques sont observées sous la forme de variables dichotomiques ou polytomiques : l'individu est marié ou non, l'individu peut migrer entre $n$ régions données, par exemple. Nous développons plus en détail ici le cas binaire, qui peut être étendu avec un certain nombre de modifications au cas polytomique.

Supposons que l'on travaille sur un modèle logit, où la probabilité pour que la caractéristique à estimer, $y_{ij}$, soit égale à 1 s'écrit en fonction de la caractéristique explicative, $x_{ij}$, supposée ici binaire :

$$P(y_{ij} = 1 \mid x_{ij}) = p_{ij} = \left[1 + \exp\left(-\left[a_o + u_{oj} + (a_1 + u_{1j})\, x_{ij}\right]\right)\right]^{-1} \qquad [8]$$

Il s'ensuit que les réponses $y_{ij}$ sont distribuées selon une loi binomiale de paramètres :

$$y_{ij} \sim B(p_{ij}, 1) \qquad [9]$$

Dans ce cas on a la variance conditionnée suivante :

$$\text{var}(y_{ij} \mid p_{ij}) = p_{ij}(1 - p_{ij})$$

Le modèle devient alors un modèle non linéaire :

$$y_{ij} = p_{ij} + e_{ij}\, z_{ij}$$

où :

$$z_{ij} = \sqrt{p_{ij}\,(1 - p_{ij})}$$

et où :

$$\sigma_e^2 = 1$$

Dans ce cas la variance est égale à l'unité au niveau individuel, et l'on travaillera essentiellement sur les variances et covariances au niveau 2 (Goldstein, 1991).

***Une application aux migrations norvégiennes*** Nous travaillons ici sur les flux d'émigration des 19 régions norvégiennes, pour les individus nés en 1958 et ayant migré en 1980-1981 (voir pour plus de détails : Baccaïni et Courgeau, 1996a). Nous disposons, pour expliquer les comportements, des 8 caractéristiques individuelles et agrégées que nous avons déjà définies en II.

Dans la mesure où caractéristiques individuelle et agrégée ont un effet propre sur les chances d'émigrer des régions, nous les faisons intervenir d'abord pour chaque type de caractéristique dans un modèle logit simple et dans un modèle logit multi-niveaux. Le tableau 9 porte ces résultats, pour les hommes.

Les paramètres non aléatoires estimés avec un modèle multi-niveaux, sont en général très proches de ceux que l'on obtient avec un modèle logit simple. Cela vient confirmer les résultats (Baccaïni, Courgeau, 1996a) que nous rappelions en II. Mais lorsque l'effet des aléas liés à la caractéristique est non nul au niveau régional, on observe une forte augmentation de la dispersion de ces paramètres, de l'ordre du doublement de leur écart type. En dépit de cela, la plupart des effets significatifs au seuil de 5 % dans le modèle logit simple le restent dans le modèle multi-niveaux. Échappent seulement à cette règle deux effets de caractéristiques agrégées : l'influence positive qu'avait le fait de vivre dans une région de faibles revenus sur les chances de migrer n'est plus significative dans le modèle multi-niveaux ; en revanche on voit apparaître un effet de rétention significatif des régions où le niveau d'études est élevé, lorsqu'on utilise un modèle multi-niveaux alors que le modèle logit simple ne montrait rien de tel.

Voyons maintenant plus en détail l'effet conjoint des paramètres fixes et des paramètres aléatoires au niveau région. La fonction logit de la probabilité d'émigrer de $j$ des individus qui n'ont pas la caractéristique étudiée, $\prod_{oj}$, est donnée par $a_o + u_{oj}$ ; sa variance entre régions est égale à $\sigma_{eo}^2$ .La fonction logit pour les individus qui l'ont, $\prod_{1j}$, est la somme $a_o + a_1 + u_{oj} + u_{1j}$ ; sa variance entre régions est donc égale à : $\sigma_{eo}^2 + 2\sigma_{eo1} + \sigma_{e1}^2$. Il sera également intéressant de comparer ces variances selon que l'on fait intervenir la caractéristique agrégée (tableau 9) ou non : pour ne pas alourdir le tableau ces dernières

TABLEAU 9. – ESTIMATION DES PARAMÈTRES ET DE LEUR ÉCART TYPE (ENTRE PARENTHÈSES) DES MODÈLES LOGIT SIMPLE ET MULTI-NIVEAUX FAISANT INTERVENIR SIMULTANÉMENT UNE CARACTÉRISTIQUE INDIVIDUELLE ET LA CARACTÉRISTIQUE AGRÉGÉE CORRESPONDANTE EN 1980 (GÉNÉRATION MASCULINE NÉE EN 1958)

| Paramètres | Marié | | Actif | | Agriculteur | | Plus de 12 ans d'études | |
|---|---|---|---|---|---|---|---|---|
| | Logit simple | Multi-niveaux | Logit simple | Multi-niveaux | Logit simple | Multi-niveaux | Logit simple | Multi-niveaux |
| **Non aléatoires** | | | | | | | | |
| Constante | – 1,465 (0,061) | – 1,563 (0,114) | 1,586 (0,684) | 2,978 (1,625) | – 2,190 (0,043) | – 2,291 (0,149) | – 2,216 (0,076) | – 1,725 (0,217) |
| Caractéristique | 0,418 (0,054) | 0,393 (0,079) | – 0,540 (0,042) | – 0,588 (0,074) | – 0,401 (0,097) | – 0,406 (0,096) | 0,531 (0,058) | 0,648 (0,117) |
| Caractéristique agrégée | – 0,057 (0,005) | – 0,050 (0,008) | – 0,044 (0,009) | – 0,062 (0,021) | 0,028 (0,018) | 0,028 (0,018) | 0,002 (0,008) | – 0,058 (0,024) |
| **Aléatoires niveau région** | | | | | | | | |
| $\sigma^2_{e0}$ (constante) | | 0,018 (0,015) | | 0,045 (0,032) | | 0,064 (0,029) | | 0,099 (0,055) |
| $\sigma_{e01}$ (covariance) | | 0,013 (0,012) | | – 0,020 (0,027) | | 0,000 | | 0,107 (0,072) |
| $\sigma^2_{e1}$ (caractéristique) | | 0,058 (0,045) | | 0,060 (0,037) | | 0,000 | | 0,178 (0,146) |
| | Au moins un enfant | | Faibles revenus | | Hauts revenus | | Sans revenus | |
| | Logit simple | Multi-niveaux | Logit simple | Multi-niveaux | Logit simple | Multi-niveaux | Logit simple | Multi-niveaux |
| **Non aléatoires** | | | | | | | | |
| Constante | – 1,307 (0,077) | – 1,373 (0,180) | – 3,053 (0,150) | – 2,590 (0,382) | 0,562 (0,306) | – 0,698 (0,670) | – 2,240 (0,083) | – 2,313 (0,290) |
| Caractéristique | – 0,133 (0,079) | – 0,165 (0,098) | 0,096 (0,051) | 0,125 (2,103) | – 0,195 (0,039) | – 0,256 (0,099) | – 0,065 (0,132) | – 0,074 (0,124) |
| Caractéristique agrégée | – 0,110 (0,080) | – 0,099 (0,026) | 0,053 (0,009) | 0,025 (0,021) | – 0,004 (0,005) | – 0,022 (0,011) | 0,038 (0,029) | 0,082 (0,099) |
| **Aléatoires niveau région** | | | | | | | | |
| $\sigma^2_{e0}$ (constante) | | 0,033 (0,014) | | 0,100 (0,035) | | 0,035 (0,024) | | 0,067 (0,029) |
| $\sigma_{e01}$ (covariance) | | 0,012 (0,022) | | – 0,116 (0,034) | | – 0,032 (0,038) | | 0,000 |
| $\sigma^2_{e1}$ (caractéristique) | | 0,055 (0,093) | | 0,156 (0,054) | | 0,152 (0,068) | | 0,000 |

*Source* : *Registre de Population Norvégien,* Central Bureau of Statistics, Oslo.

estimations n'y sont pas portées et nous les citerons directement dans le texte lorsque cela sera utile.

Examinons plus en détail l'effet de trois caractéristiques que nous avons schématisé dans la figure 6.

Ainsi, les agriculteurs ont une plus faible probabilité de migrer que les autres professions. Mais on voit que les variances au niveau régions des logits $\prod_{oj}$ et $\prod_{1j}$ sont égales, que l'on fasse intervenir ou non d'ailleurs la caractéristique agrégée. Le schéma 1.a de la figure 6 porte les valeurs de $\prod_{oj}$ et $\prod_{1j}$ correspondant respectivement aux non-agriculteurs et aux agriculteurs, reliées pour chaque région par des lignes, qui sont dans ce cas parallèles entre elles. Indiquons que lorsque l'on fait intervenir le pourcentage d'agriculteurs, la variance entre régions décroît légèrement de 0,070 à 0,064, mais surtout, lorsque ce pourcentage augmente, la probabilité de migrer tant des agriculteurs que des autres va augmenter. Notons cependant qu'ici cet effet n'est pas très significatif, mais le deviendra lorsqu'on fera intervenir d'autres caractéristiques. Le schéma 1b porte les valeurs moyennes de $\prod_{oj}$ et de $\prod_{1j}$ en fonction du pourcentage d'agriculteurs de chaque région : on a, dans ce cas, deux lignes parallèles entre elles, conduisant à un schéma similaire à celui que nous avons observé pour la même catégorie en France (Courgeau, 1994). Les agriculteurs ont une beaucoup plus faible probabilité de migrer que les non-agriculteurs, mais lorsque le pourcentage d'agriculteurs présents dans une région augmente, cela accroît les chances de migrer tant des uns que des autres. Ce résultat éclaire bien le danger d'inférer des résultats sur l'individu à partir de résultats obtenus à un niveau plus agrégé : la présence de nombreux agriculteurs entraîne une plus forte probabilité d'émigrer de toutes les catégories de la population, du fait sans doute de la plus grande rareté d'emplois non agricoles dans ces régions. Cela n'implique en rien que les agriculteurs aient de plus fortes chances d'émigrer que les autres, puisque c'est exactement l'inverse que l'on observe.

Voyons maintenant le cas des individus ayant au moins un enfant. On constate qu'ils ont toujours une plus faible probabilité de migrer que les sans enfant, que l'on fasse intervenir ou non le pourcentage ayant au moins un enfant. On voit également que la variance, entre régions, des logits de la probabilité régionale de migrer pour ceux qui ont au moins un enfant (0,174), est plus de trois fois plus élevée que celle des sans enfant (0,061), lorsqu'on ne fait pas intervenir les pourcentages ayant au moins un enfant.

Il en résulte, dans ce cas, le schéma 2 de la figure 6, qui porte toujours les valeurs de $\prod_{oj}$ et de $\prod_{1j}$ : la corrélation positive et proche de l'unité (0,95) entre les effets aléatoires caractérise la forme en éventail de ce schéma. Lorsque ce pourcentage intervient il a un effet tout à fait significatif et, surtout, il réduit de moitié les variances et covariances entre les régions, leur faisant perdre leur effet significatif : de 0,06 à 0,03 pour la variance des individus sans enfant, de 0,17 à 0,11 pour celle des indi-

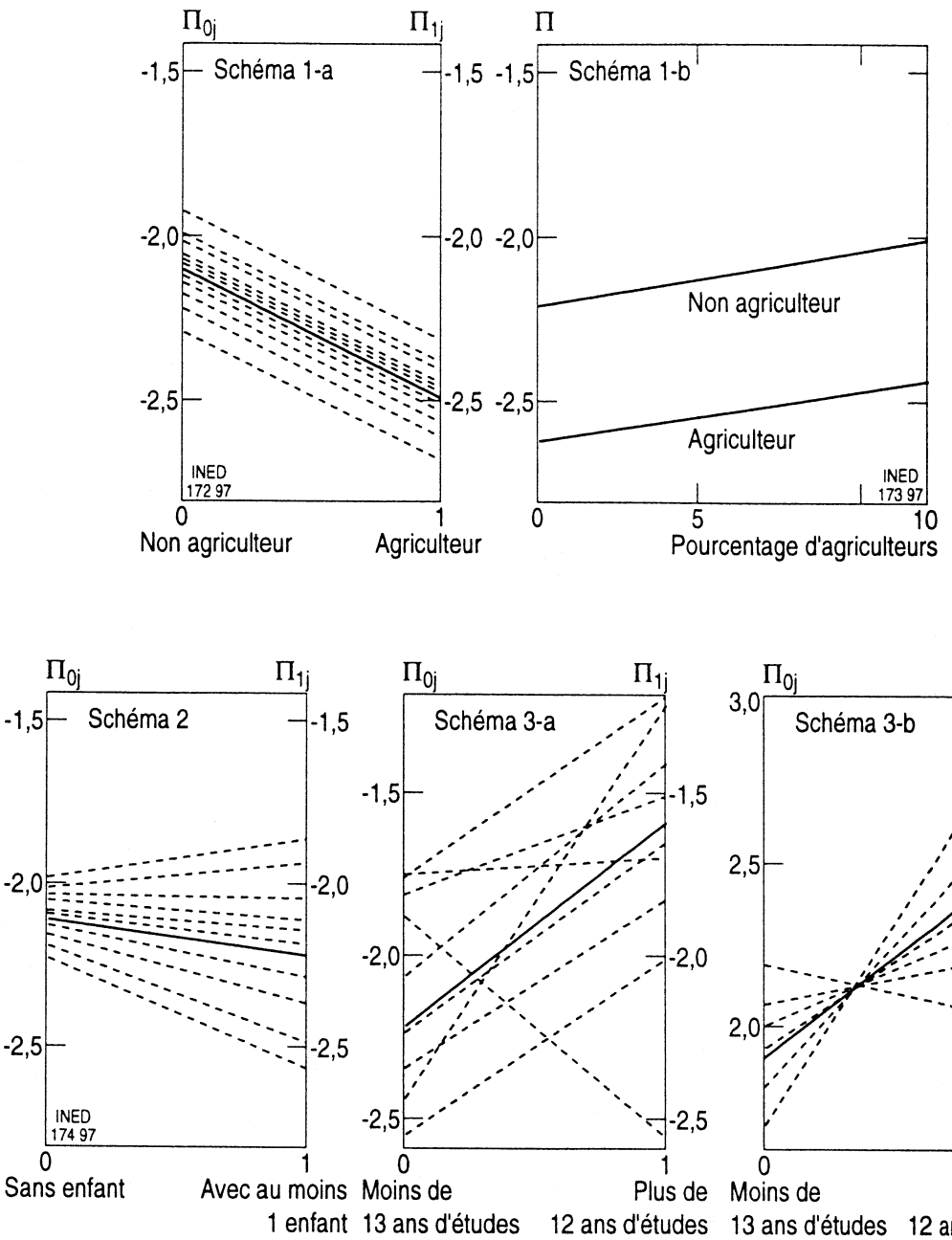


Schéma 1-a : sans faire intervenir le pourcentage d'agriculteurs
Schéma 1-b : en faisant intervenir simultanément le pourcentage d'agriculteurs
Schéma 2 : sans faire intervenir le pourcentage d'individus ayant au moins un enfant
Schéma 3-a : sans faire intervenir le pourcentage d'individus ayant plus de 12 ans d'études
Schéma 3-b : en faisant intervenir toutes les autres caractéristiques

*Figure 6.* – Effet de trois caractéristiques (être agriculteur, avoir au moins un enfant et avoir plus de 12 ans d'études) sur le logit de la probabilité de migrer en Norvège (génération 1958 entre 1980 et 1981)

vidus avec enfants et de 0,10 à 0,04 pour la covariance. On peut donc dire qu'il explique bien une partie de ces effets aléatoires.

Voyons, pour terminer ces exemples, le cas des individus ayant plus de 12 ans d'études qui ont une plus forte probabilité de migrer que les autres. Dans ce cas la corrélation entre les aléas régionaux des individus ayant plus ou moins 12 ans d'études est pratiquement nulle (0,07) : cela entraîne le schéma 3 de la figure 6, dans lequel les droites correspondant aux diverses régions semblent être tirées aléatoirement autour de la tendance moyenne reliant $a_o$ à $(a_o + a_1)$. Bien entendu les autres caractéristiques conduisent à autant d'interprétations différentes des situations régionales.

Considérons enfin un modèle où interviennent simultanément toutes les caractéristiques qui ont un effet sur la probabilité régionale de migrer. Le tableau 10 donne les résultats d'un modèle logit simple, comparé à un modèle multi-niveaux où seules les caractéristiques du niveau d'éducation sont considérées comme aléatoires entre régions. Les effets des caractéristiques non aléatoires sont très proches, que l'on utilise le premier ou le second modèle. Le cas des agriculteurs devient pleinement significatif : le fait d'être agriculteur diminue toujours la probabilité de migrer, alors qu'un pourcentage élevé d'agriculteurs dans une région va augmenter les chances de migrer des agriculteurs et des non-agriculteurs.

TABLEAU 10. – ESTIMATION DES PARAMÈTRES ET DE LEUR ÉCART TYPE (ENTRE PARENTHÈSES) DES MODÈLES LOGIT SIMPLE ET MULTI-NIVEAUX FAISANT INTERVENIR LES DIVERSES CARACTÉRISTIQUES INDIVIDUELLES ET AGRÉGÉES AYANT UN EFFET SIGNIFICATIF SUR LA PROBABILITÉ DE MIGRER EN 1980-1981 (GÉNÉRATION MASCULINE NORVÉGIENNE NÉE EN 1958)

| Paramètres | Logit simple | Multi-niveaux (Plus de 12 ans d'études) |
|---|---|---|
| **Non aléatoires** | | |
| Constante | 2,467 (0,856) | 1,711 (1,152) |
| Marié | 0,641 (0,061) | 0,653 (0,070) |
| Actif | – 0,595 (0,046) | – 0,598 (0,085) |
| Agriculteur | – 0,226 (0,100) | – 0,208 (0,100) |
| Plus de 12 ans d'étude | 0,520 (0,063) | 0,621 (0,082) |
| Au moins un enfant | – 0,467 (0,089) | – 0,467 (0,102) |
| Faibles revenus | – 0,256 (0,063) | – 0,261 (0,067) |
| Hauts revenus | – 0,107 (0,051) | – 0,102 (0,084) |
| Sans revenus | – 0,610 (0,140) | – 0,619 (0,133) |
| Part d'actifs | – 0,042 (0,011) | – 0,034 (0,014) |
| Part d'agriculteurs | 0,070 (0,007) | 0,074 (0,010) |
| Part ayant au moins un enfant | – 0,155 (0,012) | – 0,138 (0,010) |
| Part sans revenus | – 0,078 (0,033) | – 0,100 (0,037) |
| **Aléatoires niveau région** | | |
| $\sigma^2_{e0}$ (constante) | | 0,019 (0,009) |
| $\sigma_{e01}$ (covariance) | | – 0,057 (0,030) |
| $\sigma^2_{e1}$ (plus de 12 ans d'études) | | 0,150 (0,108) |

*Source* : *Registre de Population Norvégien,* Central Bureau of statistics, Oslo.

En revanche, le fait d'avoir des revenus élevés devient non significatif dans un modèle multi-niveaux. Ce sont les paramètres aléatoires au niveau région qui sont les plus modifiés, par rapport au modèle où seule la caractéristique non aléatoire, niveau d'éducation, intervenait. La variance entre régions $\sigma^2_{e0}$ est réduite au cinquième de ce qu'elle était (0,018 contre 0,99) du fait de l'intervention des autres caractéristiques. En revanche, la variance entre régions, $\sigma^2_{e0} + 2\sigma_{e01} + \sigma^2_{e1}$, reste proche de ce qu'elle était (0,055 contre 0,063), mais surtout la corrélation entre les aléas régionaux des individus ayant plus ou moins 12 ans d'études devient égale à – 0,99 alors qu'elle était pratiquement nulle dans le modèle antérieur (0,07). Cela conduit au schéma 4 de la figure 6, qui est à comparer au schéma 3 : les régions où la probabilité de migrer des moins de 13 ans d'études est la plus faible, ayant tenu compte de l'effet de toutes les caractéristiques, seront celles où la probabilité de migrer des plus de 12 ans d'études est la plus forte, et inversement. On voit ainsi apparaître une relation qui était brouillée par les autres caracactéristiques, lorsque celles-ci n'étaient pas prises en compte dans le modèle.

Dans l'exemple pris ici, on peut conclure que l'introduction d'un modèle utilisant des aléas multi-niveaux ne change pas l'essentiel des conclusions obtenues avec un modèle logit simple, faisant cependant intervenir des caractéristiques mesurées à divers niveaux d'agrégation. En revanche, ces aléas fournissent des éléments d'information intéressants sur les liens entre probabilités d'émigrer des diverses régions des individus ayant ou n'ayant pas une caractéristique donnée. L'illustration sur des schémas du type de ceux donnés dans la figure 6 permet de visualiser de tels liens. Notons, finalement, que dans notre exemple l'effet de ces aléas est souvent statistiquement non significativement différent de zéro, rendant les conclusions plus ténues.

## IV. – Vers une analyse biographique multi-niveaux

Les analyses présentées jusqu'ici n'étaient que partiellement biographiques : les régressions individuelles faisaient bien intervenir une note des élèves en début et en fin d'une période de 4 ans, mais ne permettaient pas d'étudier son évolution tout au long de leur carrière scolaire ; les analyses logit ou les modèles biographiques appliqués aux migrations norvégiennes, considéraient les événements qui se produisaient pendant deux ans, sans étudier une migration de rang donné tout au long du séjour des individus dans diverses régions. Il nous faut donc maintenant envisager les possibilités et les difficultés de réaliser une analyse biographique multi-niveaux.

Dans la mesure où l'on va suivre les individus tout au long de leur séjour dans un état donné, certaines de leurs caractéristiques individuelles peuvent changer à des instants donnés (ils se marient, changent de profession, deviennent inactifs, etc.) et les caractéristiques des régions dans

lesquelles ils vivent vont changer de façon continue au cours du temps (augmentation du pourcentage de mariés ou de ceux ayant au moins un enfant, changements du pourcentage d'actifs dans la génération, etc.).

Il faut donc pouvoir disposer de ces caractéristiques tout au long de la vie des individus ; le registre de population norvégien, s'il suit bien le mariage et les naissances d'enfants des membres de la cohorte, ne suit pas leur activité, ni leur profession, ni leur revenu, etc. On ne dispose de ces informations que lors d'un recensement et nous avons utilisé les informations données lors du recensement de 1980 pour effectuer nos analyses sur la courte période 1981-1982. Nous n'avons donc pas la possibilité de réaliser cette même analyse sur toute la durée de séjour. Un problème de disponibilité de données se pose donc si l'on veut faire une analyse biographique multi-niveaux. Il serait également nécessaire de pouvoir disposer de données d'enquêtes biographiques portant sur des effectifs beaucoup plus importants que ceux habituellement considérés, qui permettent de calculer des caractéristiques régionales de façon continue au cours du temps, avec une précision suffisante. La mise en place de sondages contextuels qui soient capables de rétablir les ponts entre les comportements individuels et les structures sociales, paraît indispensable pour réaliser des analyses biographiques multi-niveaux. Pour ce faire, il faudrait « mettre en œuvre des systèmes d'observation représentatifs de contextes sociaux diversifiés et hiérarchisés, en combinant dans un système d'indicateurs intégrés multi-niveaux les apports de l'analyse écologique, de l'enquête sociologique individuelle et de l'analyse contextuelle » (Loriaux, 1987). Il s'agit d'un domaine encore peu exploré, mais essentiel.

En revanche, les méthodes d'analyse existent, qui permettent de calculer une vraisemblance partielle (Cox, 1972), rapport du quotient instantané de l'individu qui connaît l'événement à un instant donné à la somme des quotients de l'ensemble de la population soumise au risque. Le produit de ces vraisemblances, calculées pour chaque instant où intervient un événement, peut être maximisé en faisant intervenir plusieurs niveaux d'agrégation (Goldstein, 1995). On voit que cela conduit rapidement à des fichiers énormes, qui peuvent dépasser les capacités de mémoire vive disponibles, car ils doivent contenir, pour chaque événement observé, toutes les caractéristiques de chaque membre de la population soumise au risque, ces caractéristiques changeant d'ailleurs d'un événement au suivant.

Par ailleurs, à un niveau d'agrégation donné, un individu peut changer de zone au cours de son séjour dans la population soumise au risque. Prenons, pour le montrer, l'exemple d'une étude de la fécondité entre diverses régions d'un pays. Il est évident que certains individus vont migrer entre ces régions pendant leur période féconde. L'individu devra donc être rattaché à chacune des nouvelles régions chaque fois qu'il migrera et les caractéristiques agrégées de ces régions auront un effet sur son comportement fécond. Une telle hypothèse markovienne (le comportement de l'individu ne dépend que de la région dans laquelle il se trouve et il oublie, dès son installation dans le nouveau lieu, les contraintes des régions anté-

rieurement habitées), paraît peu vraisemblable. Il est nécessaire de la rendre moins stricte. Il est utile dans ce cas de tester la rapidité d'adaptation aux conditions de la région d'adoption, si celle-ci se produit, ou les conditions de sélection des migrants dans la région d'origine, si cette seconde hypothèse est vérifiée (Courgeau, 1987).

On arrive là à des modèles non-markoviens de comportement démographique, dont la complexité va s'ajouter à la considération de niveaux d'agrégation multiples.

La mise en place des modèles multi-niveaux proprement biographiques se heurte donc à des problèmes d'observation et de données disponibles, ainsi qu'à des problèmes d'analyse qui restent en grande partie à résoudre.

## Conclusions

Nous sommes passés, tout au long de cet article, des modèles plus simples, faisant intervenir la multiplicité des niveaux sous la forme de caractéristiques individuelles et agrégées, à des modèles plus complexes mettant en œuvre des aléas propres à chaque niveau, pour aboutir aux modèles biographiques multi-niveaux, les plus satisfaisants mais les plus difficiles à utiliser du fait de problèmes d'observation et d'analyse.

Ces diverses avancées montrent la richesse des modèles multi-niveaux mais, en même temps, le besoin de mettre en place une théorie cohérente de ces modèles.

Pour ce faire, une réflexion épistémologique s'impose en vue de définir la signification à accorder aux différents niveaux d'agrégation que l'on peut utiliser. Peut-on penser que les caractéristiques agrégées sont le reflet de l'organisation sociale dans laquelle nous vivons, tandis que la liberté individuelle se montrerait sous les caractéristiques propres à chaque individu (Courgeau, 1996) ? Dans ce cas quelle signification donner à l'utilisation d'un grand nombre de niveaux d'agrégation (de l'individu, au ménage, au quartier, à la commune, au département, à la région, etc.) ? N'y a-t-il pas lieu d'essayer de mettre en évidence des niveaux privilégiés qui ne soient pas forcément des niveaux d'agrégation administratifs (bassins d'emploi, par exemple), et qui sont à intégrer dans une théorie plus générale ? Il paraît enfin nécessaire d'articuler entre eux ces divers niveaux, qui ne sont pas indépendants les uns des autres.

Ne faut-il pas également dépasser l'approche individuelle prise ici, qui cherche à expliquer les comportements par des caractéristiques mesurées à différents niveaux d'agrégation ? Il semble nécessaire de compléter cette étude par celle des comportements propres aux divers niveaux, qui va ensuite chercher à les relier les uns aux autres. Ainsi des actions individuelles isolées, dans une communauté donnée, amènent la prise de conscience

d'un problème qui affecte l'ensemble de cette collectivité. Dès lors, cela peut conduire à des mesures politiques prises à un niveau plus agrégé. Bien entendu ces mesures vont affecter les conduites individuelles, amener de nouvelles actions pour contrebalancer leurs effets pervers, et ainsi de suite.

Ne faut-il pas, enfin, tenir compte de la structure sociale des groupes considérés ? Le travail présenté par Bonvalet, Bry et Lelièvre dans ce même numéro de *Population*, montre la possibilité de le faire quand on considère des groupes restreints, comme la famille ou le ménage. Pour des groupes plus étendus, l'utilisation de valeurs moyennes des caractéristiques individuelles ou même l'utilisation de variances ou de covariances, permet-elle de le définir correctement ? Il serait, là aussi, nécessaire de tenir compte des interactions qui existent entre les membres du groupe et de leurs changements au cours du temps, pour intégrer pleinement leur structure sociale. C'est une tâche difficile qui nécessite la mise en œuvre de nouveaux moyens d'observation.

Cette approche dépasse donc les méthodes d'analyse proposées ici, pour mettre en place une théorie des comportements humains dont les bases épistémologiques, les méthodes de mesure et d'analyse restent encore largement à établir. Les recherches à venir nous diront la fécondité d'une telle piste, qui permet d'aborder simultanément l'étude des divers niveaux d'agrégations apparaissant dans les sciences sociales.

Daniel COURGEAU, Brigitte BACCAÏNI

***Remerciements*** : Des versions antérieures de cet article ont été présentées et discutées au colloque international sur l'*Analyse en multiples niveaux : problématique générale et méthodologie*, le 11 octobre 1996 à Louvain-la-Neuve, au séminaire *Démodynamiques* de l'Ined, le 16 janvier 1997 et au séminaire franco-hollandais sur la *Mobilité résidentielle et choix du logement*, les 3 et 4 avril 1997. Nous remercions également Dominique Tabutin pour ses commentaires. Nous remercions enfin les services statistiques norvégiens qui nous ont permis d'avoir accès à des fichiers issus du registre de population et créés par Kjetil Sørlie et Øjsten Kravdal.

## BIBLIOGRAPHIE

ALKER H.-R., (1969), « A typology of ecological fallacies », in *Quantitative ecological analysis*, Dogan and Rokkan eds., MIT Press, Massacussets, pp. 69-86.

BACCAÏNI B., COURGEAU D., (1996a), « Approche individuelle et approche agrégée : utilisation du Registre de population norvégien pour l'étude des migrations », in *Analyse spatiale de données biodémographiques*, Bocquet-Appel, Courgeau et Pumain éds., John Libbey/Ined, Paris, pp. 79-104.

BACCAÏNI B., COURGEAU D., (1996b), « The spatial mobility of two generations of young adults in Norway », *International journal of population geography*, vol. 2, n° 4, p. 333-359.

COURGEAU D., (1987), « Constitution de la famille et urbanisation », *Population*, 42, pp. 57-82.

COURGEAU D., (1994), « Du groupe à l'individu : l'exemple des comportements migratoires », *Population*, 49, pp. 7-26.

COURGEAU D., (1996), « Towards a multilevel analysis in social sciences »/« Vers une analyse multi-niveaux en sciences sociales », in *Analyse spatiale de données biodémographiques*, Bocquet-Appel, Courgeau et Pumain éds., John Libbey/Ined, Paris, pp. 10-22.

COURGEAU D., LELIÈVRE É., (1989), *Analyse démographique des biographies*, Éditions de l'Ined, Paris, 268 p.

COURGEAU D., LELIÈVRE É., (1996), « Changement de paradigme en démographie », *Population*, 51, pp. 645-654.

COX D.-R., (1972), « Regression models and life tables (with discussion) », *Journal of the Royal Statistical Society*, B34, pp. 187-220.

ENTWISTLE B., MASON W.-M., (1985), « Multilevel effects of socio-economic development and family planning programs on children ever born », *American Journal of Sociology*, 91, pp. 616-649.

FIREBAUGH G., (1978), « A rule for inferring individual-level relationships from aggregate data », *American Sociological Review*, 43, pp. 557-572.

GERONIMUS A.-T., BOUND J., NEIDERT L.-J., (1996), « On the validity of using census geocode characteristics to proxy individual socio-economic characteristics », *Journal of the American Statistical Association*, 91, pp. 529-537.

GOLDSTEIN H., (1986), « Multilevel mixed linear model analysis using iterative generalized least squares », *Biometrika*, 73, pp. 43-56.

GOLDSTEIN H., (1987), « Multilevel covariance component models », *Biometrika*, 74, pp. 430-431.

GOLDSTEIN H., (1991), « Nonlinear multilevel models, with an application to discrete response data », *Biometrika*, 78, pp. 45-51.

GOLDSTEIN H., (1995), *Multilevel Statistical Models*, Edward Arnold, 178p.+XIV.

HAUSER R.-M., (1974), « Contextual analysis revisited », *Sociological Methods and Research*, vol. 2, n° 3, pp. 365-375.

JACQUOT A., (1994), « 1982-1990 : un modèle de déséquilibre pour les marchés régionaux du travail en France », *Revue d'Économie Régionale et Urbaine*, 3.

JONES K., (1993), *Everywhere is nowhere : multilevel perspectives on the importance of place*, The University of Portsmouth Inaugural Lectures, 12 p.

LANCASTER T., (1990), *The Econometric Analysis of Transition Data*, Econometric Society Monographs, Cambridge University Press,

LAZARSFELD P.-F., MENZEL H., (1961), « On the relation between individual and collective properties », in *Complex Organizations*, Etzioni ed., Holt, Reinhart and Winston, New York, pp. 422-440.

LORIAUX M., (1989), « L'analyse contextuelle : renouveau théorique ou impasse méthodologique », in *L'explication en sciences sociales : la recherche des causes en démographie*, Duchêne, Wunsch et Vilquin éds, Éditions Ciaco, Louvain-la-Neuve, pp. 333-368.

PIANTADOSI S., BYAR D., GREEN S., (1998), « The ecological fallacy », *American Journal of Epidemiology*, 127, pp. 893-904.

PUIG J.-P., (1981), « La migration régionale de la population active », *Annales de l'INSEE*, n° 44, pp. 41-79.

ROBINSON W.-S., (1950), « Ecological correlations and the behaviour of individuals », *American Sociological Review*, 15, pp. 351-357.

TUMA N.-B., HANNAN M.-T., (1984), *Social Dynamics. Models and Methods*, Academic Press, Orlando, 580p.+XX.

VON KORFF M., KOEPSELL T., CURRY S., DIEHR P., (1992), « Multilevel analysis in epidemiologic research on health behaviors and outcomes », *American Journal of Epidemiology*, 135, pp. 1077-1082.

WILLEKENS F., ROGERS A., (1978), *Spatial Population Analysis : Methods and Computer Programs*, Research Report, IIASA, Laxenburg, Austria, 302 p.

WOODHOUSE G., RASBASH J., GOLDSTEIN H., YANG M., (1996), « Introduction to multilevel modelling », in *Multilevel Modelling Applications*, Woodhouse ed., Institute of Education, London, pp. 9-57.

YOUNG E.-C., (1924), *The Movement of Farm Population*, Cornell Universtiy, Ithaca, New York.

COURGEAU (Daniel), BACCAÏNI (Brigitte). – **Analyse multi-niveaux en sciences sociales**

L'approche multi-niveaux permet d'aborder les comportements humains, en tenant compte non seulement des caractéristiques individuelles, mais également du fait que ces individus font partie d'unités géographiques plus larges telles que les communes ou les régions. Une présentation détaillée et critique est faite ici des objectifs et des diverses formulations de ces modèles. Cet article va des modèles les plus simples, qui font intervenir la multiplicité des niveaux sous la forme de caractéristiques individuelles et agrégées, à des modèles plus complexes mettant en oeuvre des aléas propres à chaque niveau, pour aboutir à des modèles biographiques multi-niveaux. Il ouvre à une réflexion épistémologique plus générale sur l'apport de ces modèles.

COURGEAU (Daniel), BACCAÏNI (Brigitte). – **Multilevel analysis in the social sciences**

The multilevel approach can be used to study human behaviour taking into account not only individual characteristics but also the fact that these individuals belong to larger geographical units such as communes and regions. This article gives a detailed critical presentation of the aims and formulations of these models. Attention ranges from the most basic models, which introduce the many different levels in the form of individual and aggregated characteristics, to more complex models which operate with the random characteristics specific to each level, and culminates with multilevel event history models. The article concludes with a more general epistemological reflection on the contribution of these models.

COURGEAU (Daniel), BACCAÏNI (Brigitte). – **Análisis multi-nivel en ciencias sociales**

El análisis multi-nivel permite abordar los comportamientos humanos teniendo en cuenta no únicamente las características individuales, sino también el hecho de que los individuos forman parte de unidades geográficas tales como los municipios o las regiones. El artículo hace una presentación detallada y crítica de los objetivos y formulaciones diversas de tales modelos. Se va desde los modelos más simples, que hacen intervenir la multiplicidad de niveles bajo forma de características individuales y agregadas, a modelos más complejos que incluyen factores de heterogeneidad en cada nivel, hasta llegar a modelos biográficos multi-nivel. Finalmente, se realiza una reflexión epistemológica general sobre la aportación de estos modelos.

Daniel COURGEAU, Institut national d'études démographiques, 27 rue du Commandeur, 75675 Paris Cedex 14, France, tél. (33) 01 42 18 21 07, fax. (33) 01 42 18 21 99, E-mail: courgeau@ined.fr