Théories, paradigmes et courants explicatifs en démographie.
Chaire Quetelet 1997.
Institut de Démographie, Université catholique de Louvain,
Louvain-la-Neuve, Academia-Bruylant/L'Harmattan, 1999, pp. 93-116.

De l'intérêt des analyses multi-niveaux pour l'explication en démographie

Daniel COURGEAU

Institut National d'Etudes Démographiques
Paris, France.

Résumé

Après avoir montré les inconvénients d'une analyse faite à un seul niveau, individuel ou agrégé, nous présentons les modèles multi-niveaux sous diverses conditions d'utilisation. A un moment donné, ces modèles distinguent l'effet de caractéristiques individuelles et agrégées, en introduisant des aléas propres à chaque niveau. L'effet du temps conduit à des modèles biographiques multi-niveaux, qui doivent faire intervenir simultanément les phénomènes étudiés et la mobilité entre aires d'un même niveau d'agrégation. Enfin, des modèles plus complexes permettent d'étudier des groupes dans lesquels les biographies individuelles ne sont plus indépendantes les unes des autres. La conclusion montre que l'on peut pousser plus loin l'analyse, en ne considérant plus le niveau individuel comme prépondérant, mais n'importe lequel des niveaux d'agrégation.

Summary

After showing the problems associated to the use of analysis performed at a single level, individual or aggregated, we present multilevel models under various conditions. In a cross-sectional approach, these models distinguish the effects of individual and aggregated characteristics by introducing error terms at each level. Taking time effects into account leads to multilevel event-history-models where the phenomena under study and the mobility between areas at the same need to be considered simultaneously. Finally, more complex models allow us to study groups in which individual biographies are no longer independent from each other. The conclusion

shows that it is possible to go further in the analysis, when the individual level is no longer considered as the leading level.

Introduction

En démographie, comme dans les autres sciences sociales, les approches individuelles et agrégées impliquent des perspectives théoriques tellement différentes qu'un rapprochement entre elles a longtemps semblé vain, même si de nombreuses tentatives ont été faites dans le passé.

Le démographe a longtemps travaillé sur des données *agrégées* et cherchait, de ce fait, à mettre en évidence des causes sous-jacentes globales qui agissaient sur une population dans son ensemble. Cela revient à suivre l'approche de E. Durkheim, qui écrivait : "Si... cette synthèse *sui generis* qui constitue toute société dégage des phénomènes nouveaux, différents de ceux qui se passent dans les consciences solitaires, il faut bien admettre que ces faits spécifiques résident dans la société même qui les produit, et non dans ses parties, c'est-à-dire dans ses membres. Ils sont donc, en ce sens, extérieurs aux consciences individuelles..." (Durkheim, 1967, p. XVII (1e éd., 1895)). Du fait que tous les membres de la société sont soumis aux mêmes risques, indépendants d'eux-mêmes, il devient inutile de les observer séparément pour mettre en évidence ces faits sociaux. L'utilisation de sources de données agrégées, comme l'étaient les données recueillies à des fins administratives (statistiques de l'état civil, recensements, registres de population, etc.), suffit pour vérifier leur existence.

Ainsi, par exemple, pour chaque région d'un pays on va chercher à expliquer son taux d'émigration par diverses caractéristiques agrégées : taux de chômage, pourcentage d'agriculteurs, part de population de plus de 65 ans, etc. On fait alors l'hypothèse sous-jacente, rarement explicitée clairement, que la probabilité d'émigrer de chaque région, qui est la même pour tous ses habitants, est modulée par les diverses caractéristiques agrégées de cette région. Un modèle de régression permettra de mettre clairement en évidence un tel effet. Ainsi ce n'est pas parce qu'un individu est chômeur qu'il aurait, par exemple, une plus faible probabilité d'émigrer d'une zone, mais parce que le pourcentage de chômeurs est élevé que tout individu vivant dans cette zone aura une plus faible probabilité d'en émigrer. Malheureusement de nombreux chercheurs vont sauter ce pas, en particulier du fait que les données nécessaires auprès des individus ne sont pas disponibles (Lazarsfeld et Menzel, 1961) et en conclure à un effet individuel, alors qu'on ne peut parler que d'un effet agrégé.

Cela conduit à ce que l'on a maintenant coutume d'appeler l'*erreur écologique*, qui a été mise en évidence par W. Robinson (1950) et reprise par de nombreux auteurs depuis (Alker, 1969 ; Firebaugh, 1978 ; Langbein et Lichtman, 1978 ; Piantadosi et al., 1988 ; Courgeau, 1994 ; Baccaïni et Courgeau, 1996). En fait, l'hypothèse d'homogénéité des comportements individuels n'est généralement pas vérifiée et il n'est

guère possible à partir d'une observation agrégée d'en tirer une information fiable sur les comportements individuels. Nous reprendrons cet argument plus en détail par la suite, mais nous développerons d'abord l'utilité d'une approche individuelle.

L'utilisation de données d'enquêtes détaillées et même maintenant de données individuelles mais rendues anonymes, de recensements ou de registres, a permis le développement d'une approche micro, qui lève l'hypothèse d'homogénéité de la population. On travaille maintenant sur les comportements *individuels* que l'on va chercher à expliquer à l'aide de diverses caractéristiques des individus concernés. Les modèles log-linéaires et logistiques, par exemple, permettent d'exploiter de telles données et fournissent des résultats précis sur ces effets. Ainsi, pour reprendre l'exemple précédent, on va maintenant chercher à expliquer une migration interrégionale individuelle, par diverses caractéristiques mesurées au niveau micro : l'individu est-il chômeur, agriculteur, a-t-il plus de 65 ans, etc. ? On pourra dès lors savoir si le fait d'être chômeur va augmenter ou diminuer les probabilités de migrer d'un individu, cela en contrôlant l'effet des autres caractéristiques.

Cependant si l'on considère que le comportement d'un individu est uniquement lié à ses propres caractéristiques, on risque de tomber sur ce que l'on a coutume d'appeler l'*erreur atomiste*. On peut en effet penser qu'un individu sera en plus soumis à des normes et des contraintes du milieu dans lequel il vit, qui vont influer sur ses propres comportements. Ainsi, par exemple, on peut concevoir que la présence d'un fort pourcentage de chômeurs dans une région puisse influencer la probabilité individuelle de migrer, tant des actifs que des chômeurs, en les sensibilisant aux problèmes posés par le chômage. Qui plus est, un tel effet peut être à l'opposé de celui que le chômage individuel joue sur les probabilités de migrer.

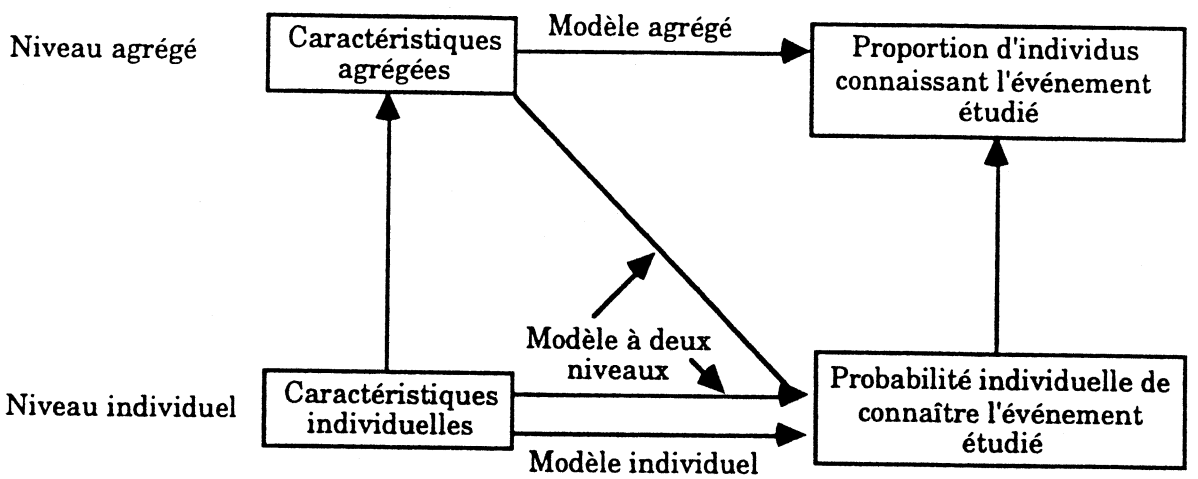
Contrairement à ce qui se passait dans le cas de données agrégées, où il n'était plus possible de revenir aux comportements individuels, les modèles log-linéaires ou logistiques permettent de faire intervenir diverses caractéristiques des milieux dans lesquels l'individu vit et agit, sur ses propres comportements. Cela permet de lever le risque d'erreur atomiste. On se situe alors dans une approche multi-niveaux, qui intègre autant de niveaux d'agrégation que l'on désire. Il faut cependant bien voir que c'est un comportement individuel que l'on explique maintenant, en faisant intervenir des caractéristiques à la fois des personnes et des groupes auxquels elles sont reliées. La figure 1 permet de voir plus clairement ce qu'il en est.

Ce schéma montre qu'il n'est généralement pas possible d'estimer un modèle individuel lorsqu'on ne dispose que des données agrégées. En revanche, à partir des caractéristiques individuelles on peut estimer les trois types de modèles. Il est bien entendu possible d'étendre ce schéma à un nombre plus élevé de niveaux d'agrégation, qu'il s'agisse de niveaux hiérarchisés ou croisés (Courgeau et Baccaïni, 1997).

Il nous faut maintenant, pour terminer cette introduction, introduire la dimension temporelle. Cette dimension absente des modèles

de régression est, en effet, indispensable à considérer en démographie. Dès la fin des années 1970, les démographes et d'autres chercheurs en sciences sociales ont mis en place les méthodes d'analyse des biographies, qui permettent de montrer comment la trajectoire antérieure d'un individu influençait les chances d'arrivée de divers événements de son existence (Tuma et Hannan, 1984 ; Courgeau et Lelièvre, 1989). Les sources nécessaires pour mettre en place cette analyse deviennent de plus en plus fréquentes dans de nombreux pays (Antoine et al., 1997) : enquêtes biographiques rétrospectives ou prospectives, données de "panel". Les méthodes d'analyse sont désormais bien intégrées dans les diverses sciences sociales (Keilman, 1993). Il faut dès lors essayer d'y introduire divers niveaux d'agrégation.

Figure 1. Relations entre divers types de modèles : individuel, agrégé, à deux niveaux



Cet article envisagera d'abord les raisons des divergences entre modèles individuel et agrégé, pour montrer ensuite les solutions proposées par les modèles multi-niveaux, selon que l'on travaille sur des groupes indifférenciés ou non et que l'on fait intervenir la dimension temporelle ou non.

1. Divergences entre modèles individuels et agrégés

Jusqu'à un passé récent, la plupart des auteurs ont essentiellement discuté le cas de phénomènes dépendants mesurés par des variables continues ou considérées comme telles, tant au niveau individuel qu'agrégé (Langbein et Lichtman, 1978 ; Piantadosi et al., 1988 ; Amrhein, 1995) et le cas où il n'y a qu'une seule variable explicative considérée (Robinson, 1950 ; Alker, 1969). Nous envisagerons ici plus en détail le cas de phénomènes mesurés par des caractéristiques binaires sur lesquelles interviennent de nombreuses variables, ce qui est le cas en démographie : l'individu a connu ou non un événement

donné, qui est influencé par de nombreuses caractéristiques. Il est utile de partir de cas relativement simples, pour faire intervenir ensuite des situations plus complexes.

Observons d'abord une cohorte à un moment donné, une année par exemple, et supposons que l'arrivée du phénomène étudié ne dépendra que de deux caractéristiques, mesurées par des variables binaires[1]. Les résultats obtenus sont aisément généralisables à un nombre quelconque de caractéristiques. Pour être plus concret, supposons que le phénomène étudié soit le changement de commune d'un individu qui vient d'entrer dans la vie active.

1.1. Le phénomène étudié ne dépend que de caractéristiques individuelles

Supposons dans un premier temps que la migration ne dépend que de la situation de chômage (1 s'il est chômeur, 0 s'il est actif occupé) et de la situation d'isolement (1 s'il vit seul, 0 si non) de l'individu. Elle ne dépend pas de sa région de résidence, ni de ses caractéristiques agrégées.

Si l'on observe les données individuelles, la probabilité pour qu'un individu i, vivant dans la région j, soit un migrant pourra dépendre des quatre états dans lesquels il peut se trouver. On va les représenter par des variables binaires égales à 1 si l'individu est dans cet état, 0 si non : x_{00}^{ij} (actif occupé, non isolé), x_{10}^{ij} (chômeur, non isolé), x_{01}^{ij} (actif occupé, vivant seul), x_{11}^{ij} (chômeur, vivant seul). Si les probabilités de migrer dans les divers états sont respectivement p_{00} p_{10} p_{01} p_{11} et si y^{ij} est une variable binaire indiquant si l'individu est migrant (1) ou sédentaire (0), on peut alors écrire :

$$y^{ij} = p_{00}\, x_{00}^{ij} + p_{10}\, x_{10}^{ij} + p_{01}\, x_{01}^{ij} + p_{11}\, x_{11}^{ij} + u^{ij} \qquad [1]$$

où u^{ij} est un terme aléatoire d'espérance mathématique nulle, n'ayant pas une distribution normale. Comme on dispose de données individuelles, on connaît également la valeur de la variable $y_{kl}^{ij}(k = 0,1; l = 0,1)$ égale à 1 si l'individu est à la fois migrant et dans l'état kl. Dès lors, les variables x_{kl}^{ij} étant orthogonales entre elles, puisque chaque individu ne peut être que dans un et un seul des quatre états, on peut remplacer la relation [1] par quatre autres[2], dont le terme général s'écrit :

---

1 Ces caractéristiques peuvent cependant être polytomiques (p valeurs) et seront dans ce cas remplacées par (p-1) variables binaires. Elles peuvent même être continues et seront dans ce cas discrétisées en regroupant leurs valeurs en p classes, ce qui nous ramène au cas précédent.

2 Nous remercions ici Xavier Bry de l'idée de cette substitution qui facilite les comparaisons entre modèle individuel et modèle agrégé.

$$y_{kl}^{ij} = p_{kl} x_{kl}^{ij} + u_{kl}^{ij} \qquad [2]$$

u_{kl}^{ij} étant un terme aléatoire. On peut également dire que pour les $N_{kl}^{\cdot\cdot}$ individus dans l'état (kl), la variable aléatoire $y_{kl}^{\cdot\cdot}$ est la somme de $N_{kl}^{\cdot\cdot}$ variables aléatoires indépendantes y_{kl}^{ij} prenant chacune la valeur 1 avec la probabilité p_{kl} et la valeur 0 avec la probabilité $(1 - p_{kl})$.

L'estimation des paramètres p_{kl} par la méthode du maximum de vraisemblance conduit aux estimateurs[3] :

$$\hat{p}_{kl} = \frac{y_{kl}^{\cdot\cdot}}{x_{kl}^{\cdot\cdot}} \qquad \text{et} \qquad \text{var}\left(\hat{p}_{kl}\right) = \frac{y_{kl}^{\cdot\cdot}\left(x_{kl}^{\cdot\cdot} - y_{kl}^{\cdot\cdot}\right)}{N\left(x_{kl}^{\cdot\cdot}\right)^3} \qquad [3]$$

où $y_{kl}^{\cdot\cdot}$ et $x_{kl}^{\cdot\cdot}$ sont les valeurs moyennes sur l'ensemble de N observations de y_{kl}^{ij} et de x_{kl}^{ij}.

Si l'on observe maintenant les données agrégées, on disposera dans chacune des r régions des pourcentages de migrants, $y^{\cdot j}$, des pourcentages de chômeurs, $x_{1\cdot}^{\cdot j}$, et des pourcentages d'individus vivant seuls, $x_{\cdot 1}^{\cdot j}$. On peut alors écrire :

$$x_{1\cdot}^{\cdot j} = x_{10}^{\cdot j} + x_{11}^{\cdot j} \text{ (chômeurs)} \quad \text{et} \quad x_{\cdot 1}^{\cdot j} = x_{01}^{\cdot j} + x_{11}^{\cdot j} \text{ (isolés).} \qquad [4]$$

Dans le cas agrégé, on a donc la possibilité d'estimer les paramètres de la relation linéaire suivante :

$$y^{\cdot j} = a_{00} + a_{1\cdot} x_{1\cdot}^{\cdot j} + a_{\cdot 1} x_{\cdot 1}^{\cdot j} + V^j \qquad [5]$$

où V^j est un nouveau terme aléatoire d'espérance mathématique nulle et qui a maintenant une distribution que l'on peut considérer comme normale.

Nous allons voir ici les liens qui peuvent exister entre la relation [1] mesurée au *niveau individuel* et la relation [5] mesurée au *niveau agrégé.*

Partons de la formule [1], additionnons celles correspondant à tous les individus d'une région j et divisons ce total par la population de cette région $N_{\cdot\cdot}^{j}$. On obtient ainsi la relation suivante :

---

3 La méthode des moindres carrés conduit au même estimateur, mais sa variance a cependant un dénominateur légèrement différent et égal à $\left(Nx_{kl}^{\cdot\cdot} - 2\right)\left(x_{kl}^{\cdot\cdot}\right)^2$. Dès que la population observée est suffisamment importante, la différence entre les deux dénominateurs devient négligeable.

$$y^{\cdot j} = p_{00}x^{j}_{00} + p_{10}x^{j}_{10} + p_{01}x^{j}_{01} + p_{11}x^{j}_{11} + u^{\cdot j} \quad [6]$$

où $u^{\cdot j}$ est un terme d'erreur dont la distribution tend vers une loi normale dès que $N^{j}_{\cdot\cdot}$ est important et dont la variance sera inversement proportionnelle à $N^{j}_{\cdot\cdot}$. On voit également qu'il est possible de faire la même opération sur la formule [2], à condition de connaître les pourcentages de migrants dans chaque catégorie, y^{j}_{kl}, ce qui conduit à la relation agrégée suivante :

$$y^{j}_{kl} = p_{kl}x^{j}_{kl} + u^{j}_{kl} \quad [7]$$

où u^{j}_{kl} est un aléa de variance inversement proportionnelle à $N^{j}_{\cdot\cdot}$. Il est donc à nouveau possible d'estimer p_{kl}, mais la méthode du maximum de vraisemblance donnera des estimations différentes de la méthode des moindres carrés pondérés par les populations régionales. La méthode du maximum de vraisemblance fournit en effet des estimations identiques à [3] :

$$\hat{p}_{kl} = \frac{\sum_j N^{j}_{\cdot\cdot}\, y^{j}_{kl}}{\sum_j N^{j}_{\cdot\cdot}\, x^{j}_{kl}} \quad \text{et} \quad \text{var}(\hat{p}_{kl}) = \frac{\hat{p}_{kl}\,(1-\hat{p}_{kl})}{\sum_j N^{j}_{\cdot\cdot}\, x^{j}_{kl}} \quad [8]$$

car on vérifie facilement que les numérateurs et dénominateurs de ces formules sont identiques. En revanche, la méthode des moindres carrés pondérés conduit aux estimations :

$$\tilde{p}_{kl} = \frac{\sum_j N^{j}_{\cdot\cdot}\, x^{j}_{kl} y^{j}_{kl}}{\sum_j N^{j}_{\cdot\cdot}\, (x^{j}_{kl})^2} \quad \text{et} \quad \text{var}(\tilde{p}_{kl}) = \frac{1}{r-2}\left[\frac{\sum_j N^{j}_{\cdot\cdot}\, (y^{j}_{kl})^2}{\sum_j N^{j}_{\cdot\cdot}\, (x^{j}_{kl})^2} - \tilde{p}^{2}_{kl}\right] \quad [9]$$

qui sont différentes des précédentes. Voyons plus précisément en quoi. Pour cela on peut réécrire la formule [3] de la façon suivante :

$$\hat{p}_{kl} = \frac{\sum_{i,j} (y_{kl}^{ij} - y_{kl}^{j} + y_{kl}^{j})(x_{kl}^{ij} - x_{kl}^{j} + x_{kl}^{j})}{\sum_{i,j} (x_{kl}^{ij} - x_{kl}^{j} + x_{kl}^{j})^2}$$

$$= \frac{\sum_{j} N_{..}^{j} \left[y_{kl}^{j} x_{kl}^{j} + \mathrm{cov}(y_{kl}^{ij}, x_{kl}^{ij}) \right]}{\sum_{j} N_{..}^{j} \left[(x_{kl}^{j})^2 + \mathrm{var}(x_{kl}^{j}) \right]} \qquad [10]$$

On voit donc que le numérateur de [10] diffère de celui obtenu par la méthode des moindres carrés [9] par la covariance entre y et x, interne aux zones j, et le dénominateur par la variance. L'estimation de ces covariances et variances conduit à :

$$\mathrm{cov}\left(y_{kl}^{ij}, x_{kl}^{ij}\right) = y_{kl}^{j}\left(1 - x_{kl}^{j}\right) \quad \text{et} \quad \mathrm{var}\left(x_{kl}^{j}\right) = x_{kl}^{j}\left(1 - x_{kl}^{j}\right) \qquad [11]$$

ce qui permet de passer des formules donnant $\tilde{p}_{kl}$ à celle donnant $\hat{p}_{kl}$. Cela permet également de voir le biais commis en utilisant la formule [9] pour estimer p_{kl}.

Ainsi, lorsque l'on dispose de données agrégées, tant pour la variable dépendante que pour les variables explicatives, décomposées par catégories de population (4 dans notre cas), les estimations avec les données individuelles et les données agrégées seront identiques, si l'on utilise la méthode du maximum de vraisemblance. En revanche elles vont différer si l'on utilise la méthode des moindres carrés (pondérés par les populations dans le cas agrégé). Il est possible d'estimer le biais commis en utilisant cette méthode.

Dans le cas habituel, où l'on ne dispose plus de la décomposition de la variable dépendante selon les catégories de population, $y_{..}^{j}$, et où l'on ne distingue que les proportions d'individus ayant chacune des caractéristiques $\left(x_{1.}^{j} \text{ et } x_{.1}^{j}\right)$, il ne sera généralement plus possible[4] de relier les formules [5] et [6].

Au vu de la relation [5], on peut réécrire [6] de la façon suivante :

$$y^{.j} = p_{00} + (p_{10} - p_{00})x_{1.}^{j} + (p_{01} - p_{00})x_{.1}^{j} + (p_{00} + p_{11} - p_{10} - p_{01})x_{11}^{j} + u^{.j} \qquad [12]$$

4 Ce n'est que dans le cas où la migration ne dépend que d'une caractéristique et d'une seule, que l'on pourra estimer p_1 et p_0 à l'aide de telles données agrégées. En revanche, dans le cas où n caractéristiques jouent sur le phénomène étudié, on aura 2^n paramètres à estimer, alors que l'on ne disposera que de n proportions.

Ce n'est donc que lorsque l'avant dernier terme est nul que l'on peut relier les paramètres estimés[5] des relations [5] et [6] :

$$\hat{p}_{00} = \hat{a}_{00},\ \hat{p}_{10} = \hat{a}_{1.} + \hat{a}_{00},\ \hat{p}_{01} = \hat{a}_{.1} + \hat{a}_{00} \quad \text{et} \quad \hat{p}_{11} = \hat{a}_{00} + \hat{a}_{1.} + \hat{a}_{.1} \qquad [13]$$

Mais on voit que les paramètres a étant le plus souvent des différences entre paramètres p, leurs valeurs seront beaucoup plus faibles et les estimations pourront dans certains cas conduire à des valeurs non significativement différentes de zéro. Il en résulte que la régression optimale peut conduire à des estimations erronées des paramètres p.

Un autre cas permet encore d'estimer les paramètres p, mais à l'aide d'une relation différente de [5]. En effet, s'il y a indépendance entre les deux caractéristiques dans chaque région $\left(x_{11}^{\cdot j} = x_{1.}^{\cdot j} x_{.1}^{\cdot j}\right)$, on peut écrire [14] :

$$y^{\cdot j} = p_{00} + (p_{10} - p_{00}) x_{1.}^{\cdot j} + (p_{01} - p_{00}) x_{.1}^{\cdot j} + (p_{00} + p_{11} - p_{10} - p_{01}) x_{1.}^{\cdot j} x_{.1}^{\cdot j} + u^{\cdot j}$$

ce qui revient à estimer une régression du type suivant :

$$y^{\cdot j} = a_{00} + a_{1.} x_{1.}^{\cdot j} + a_{.1} x_{.1}^{\cdot j} + a_{11} x_{1.}^{\cdot j} x_{.1}^{\cdot j} + u^{\cdot j} \qquad [15]$$

Dans ce dernier cas on peut vérifier que :

$$\hat{p}_{00} = \hat{a}_{00},\ \hat{p}_{10} = \hat{a}_{00} + \hat{a}_{1.},\ \hat{p}_{01} = \hat{a}_{00} + \hat{a}_{.1} \text{ et } \hat{p}_{11} = \hat{a}_{11} + \hat{a}_{00} + \hat{a}_{1.} + \hat{a}_{.1} \qquad [16]$$

La remarque faite sur les solutions [12] et [13] s'applique encore dans ce cas, qui peut conduire à des estimations erronées des paramètres p lors de la régression optimale.

Dans le cas général, il ne sera pas possible d'estimer les valeurs des probabilités, du fait que l'on ne connaît pas les proportions d'individus ayant simultanément les deux caractéristiques. On pourra cependant toujours estimer des probabilités selon les formules [13] et [16], en sachant que ces estimations pourront être très éloignées des probabilités sous-jacentes.

Notons enfin que, dans les cas considérés jusqu'ici, il n'y a aucun effet des caractéristiques agrégées ni des diverses régions sur les probabilités de migrer : le modèle au niveau agrégé donne dans ces cas des estimations imparfaites des paramètres du modèle au niveau individuel.

---

5 On utilise dans ce cas-là la méthode des moindres carrés pondérés pour estimer les paramètres a et leur matrice de variances et de covariances. L'utilisation de cette matrice permet d'estimer la variance ou l'écart-type des paramètres p. Ainsi on peut écrire :

$$var(\hat{p}_{10}) = var(\hat{a}_{1.}) + var(\hat{a}_{00}) + 2\,cov(\hat{a}_{1.}, \hat{a}_{00})$$

Pour montrer sur des exemples chiffrés ce qu'il en est, nous avons simulé un échantillon de 40 000 individus répartis dans 40 régions de 1 000 habitants chacune. Nous avons pris diverses répartitions régionales des variables explicatives, pour montrer les différences entre paramètres estimés au niveau individuel et au niveau agrégé, selon les diverses formules présentées plus haut. Les probabilités individuelles de migrer dans les divers états sont : $p_{00} = 0,07$; $p_{10} = 0,10$; $p_{01} = 0,15$; $p_{11} = 0,20$. Les proportions de chômeurs sont aléatoires dans les régions et varient entre 10 % et 40 %.

Supposons d'abord qu'il y ait indépendance dans chaque région entre le fait que l'individu soit chômeur et le fait qu'il soit isolé $\left(x_{11}^j = x_{1.}^j x_{.1}^j\right)$, mais qu'en revanche le pourcentage d'isolés dans une région soit négativement lié à son pourcentage de chômeurs $\left(x_{.1}^j = 0,75 - 1,5 x_{1.}^j + \varepsilon^j\right)$: les régions où il y a beaucoup de chômeurs sont des régions où les individus restent regroupés, qu'ils soient chômeurs ou non. Les estimations des probabilités sous les divers modèles présentés plus haut sont données dans le tableau 1. Dans le cas n° 1, on voit que lorsque l'on dispose à la fois des effectifs de migrants et des proportions de population dans les 4 cas possibles, les estimations individuelles [3] et agrégées, selon le maximum de vraisemblance [8] ou selon les moindres carrés pondérés [9], conduisent à des valeurs très proches des probabilités théoriques, avec des écarts types quasiment identiques : les différences théoriques entre maximum de vraisemblance et moindres carrés sont dans ce cas minimes. En revanche lorsqu'on ne dispose plus que de l'effectif total de migrants, mais toujours des proportions de population dans les 4 cas possibles dans chaque région, la formule [6] donne des estimations avec un écart type près de 20 fois plus élevé. Certaines des probabilités estimées sont très différentes des réelles $(p_{11} = 0,286$ *contre* $p_{11} = 0,200)$, même si l'intervalle de confiance englobe ces valeurs réelles. Qui plus est, certains paramètres n'étant plus significativement différents de zéro, cela conduit à un modèle optimal sans p_{00} qui fournit des estimations des probabilités n'incluant plus les réelles dans leur intervalle de confiance ($\hat{p}_{10} = 0,173 \pm 0,044$ avec un intervalle de confiance de 95 % contre $p_{10} = 0,10$). Enfin lorsque l'on ne dispose plus que des effectifs de chômeurs et d'individus isolés dans chaque région, l'estimation par la formule [16] conduit a des valeurs proches de ce que donnait la formule [6], ce qui permet de vérifier la condition d'indépendance dans chaque région entre le fait que l'individu soit chômeur et qu'il soit isolé. Cependant le modèle optimal dans ce cas conduit à des probabilités finales différentes à la fois des précédentes et des théoriques : il ne fait plus apparaître de différences significatives entre p_{00} et p_{10}, ni entre p_{01} et p_{11}, tout comme le modèle [13] optimal. Notons ici que tous ces modèles ont un coefficient de détermination pratiquement identique et élevé (0,988).

Tableau 1. Estimation des probabilités p_{00}, p_{10}, p_{01} et p_{11} de leur écart-type et du coefficient de détermination lorsqu'elles se font par régression, selon les divers modèles donnés dans le texte

	Estimation des probabilités par les formules		Estimations des probabilités par régression, formule [6]		Estimation des probabilités par régression, formule [16]		Estimations des probabilités, formule [13]	
	[3] et [8]	[9]	*modèle avec toutes les probabilités*	*modèle optimal*	*modèle avec toutes les probabilités*	*modèle optimal*	*modèle avec toutes les probabilités*	*modèle optimal*
Cas n°1								
p_{00}	0,070 (0,002)	0,071 (0,002)	0,062 (0,042)	-	0,062 (0,042)	0,085 (0,005)	0,051 (0,038)	0,085 (0,005)
p_{10}	0,098 (0,003)	0,097 (0,003)	0,096 (0,062)	0,173 (0,022)	0,095 (0,062)	0,085 (0,005)	0,122 (0,042)	0,085 (0,005)
p_{01}	0,149 (0,003)	0,149 (0,004)	0,142 (0,030)	0,180 (0,026)	0,141 (0,030)	0,150 (0,013)	0,158 (0,013)	0,150 (0,008)
p_{11}	0,207 (0,007)	0,209 (0,007)	0,286 (0,131)	0,345 (0,127)	0,290 (0,131)	0,150 (0,013)	0,229 (0,089)	0,150 (0,008)
R^2 (ajusté)	-	-	0,987	0,987	0,988	0,988	0,988	0,988
Cas n°2								
p_{00}	0,070 (0,002)	0,070 (0,002)	0,085 (0,028)	0,093 (0,025)	0,080 (0,028)	0,082 (0,005)	0,091 (0,027)	0,082 (0,005)
p_{10}	0,091 (0,004)	0,093 (0,004)	0,106 (0,050)	0,114 (0,048)	0,118 (0,049)	0,082 (0,005)	0,072 (0,030)	0,082 (0,005)
p_{01}	0,155 (0,003)	0,156 (0,003)	0,170 (0,021)	0,176 (0,019)	0,174 (0,040)	0,156 (0,008)	0,153 (0,011)	0,156 (0,013)
p_{11}	0,201 (0,006)	0,200 (0,008)	0,066 (0,101)	-	0,037 (0,103)	0,156 (0,008)	0,134 (0,055)	0,156 (0,013)
R^2 (ajusté)	-	-	0,991	0,992	0,992	0,992	0,992	0,992
Cas n°3								
p_{00}	0,066 (0,002)	0,065 (0,010)	0,209 (0,038)	-	0,215 (0,033)	0,236 (0,007)	0,215 (0,032)	0,236 (0,007)
p_{10}	0,093 (0,004)	0,089 (0,004)	-0,224 (0,083)	-	-0,248 (0,067)	-0,267 (0,020)	-0,245 (0,038)	-0,267 (0,020)
p_{01}	0,160 (0,003)	0,172 (0,012)	0,251 (0,034)	-	0,241 (0,014)	0,236 (0,007)	0,242 (0,011)	0,236 (0,007)
p_{11}	0,204 (0,006)	0,203 (0,006)	-0,249 (0,126)	-	-0,198 (0,497)	-0,267 (0,020)	-0,219 (0,076)	-0,267 (0,020)
R^2 (ajusté)	-	-	0,991	-	0,991	0,992	0,992	0,992

Dans le cas n° 2, nous avons introduit, en plus des conditions précédentes, une dépendance dans chaque région entre le fait que l'individu soit chômeur et le fait qu'il soit isolé $\left(x_{11}^j = x_1^j x_1^j + 0,01 + \varepsilon_1^j\right)$: cette hypothèse semble plus proche de la réalité que la précédente, car le chômage des individus doit les conduire à rester dans leur famille. Lorsque l'on dispose des effectifs de migrants et des proportions de population dans les quatre possibilités, les estimations sont plus dispersées que dans le cas précédent, menant parfois la vraie valeur à la limite de l'intervalle de confiance de certaines estimations ($\hat{p}_{01} = 0,156 \pm 0,006$ alors que $p_{01} = 0,150$). Lorsque l'on ne dispose plus que de l'effectif total de migrants de chaque région, les estimations sont encore plus erratiques : ainsi $\hat{p}_{11} = 0,066 \pm 0,202$ est non significativement différente de zéro, alors qu'il s'agit de la probabilité maximale de migration $p_{11} = 0,20$. L'application des formules [16] et [13] conduit à des estimations aussi erronées que celles données par la formule [6].

On peut dire, à partir de ces deux cas, que l'estimation au niveau agrégé, lorsque l'on ne dispose que de l'effectif total des migrants de chaque région, non décomposé par type, fournit une information fortement biaisée des probabilités individuelles de migrer.

1.2. Le phénomène étudié dépend de caractéristiques individuelles et agrégées

Supposons maintenant qu'en plus de l'effet des deux caractéristiques précédentes, il y ait également un effet du pourcentage de chômeurs présents dans la région sur les probabilités de migrer : plus ce pourcentage de chômeurs augmente, plus la crainte de retrouver un chômage aussi important ailleurs avec un réseau de relations moindre retiendra les individus dans leur région de résidence. C'est donc un effet en sens inverse du chômage individuel, qui va s'exercer sur l'ensemble de la population.

Si un tel effet ne peut être mis en évidence par un modèle individuel ne faisant pas intervenir de données agrégées, il pourra devenir prépondérant lorsque l'on travaille sur un modèle agrégé ne décomposant pas la population de chaque région par type. Ainsi les modèles [12] et [14] montrent que l'effet des caractéristiques individuelles est mesuré dans l'estimation par des différences entre probabilités, alors que l'effet de la caractéristique agrégée sera parfaitement pris en compte. Il peut dès lors facilement l'emporter et effacer entièrement celui des caractéristiques individuelles. Dans ce cas le modèle agrégé mesurera bien l'effet des caractéristiques agrégées.

Pour voir plus précisément ce qu'il en est, nous avons poursuivi la simulation, précédemment engagée, en supposant en plus la dépendance suivante entre les probabilités d'émigrer et les pourcentages de chômeurs, dans chaque région :

$$p_{00}^{j} = \frac{0,46 - x_{1.}^{j}}{3} + \varepsilon_{00}\,,\; p_{10}^{j} = 0,2 - 0,4\,x_{1.}^{j} + \varepsilon_{10}\,,\; p_{01}^{j} = 0,275 - 0,5\,x_{1.}^{j} + \varepsilon_{01}$$

et

$$p_{11}^{j} = \frac{1,1 - 2x_{1.}^{j}}{3} + \varepsilon_{11}$$

Lorsque le pourcentage de chômeurs varie entre 10 % et 40 %, ces probabilités sont toujours centrées sur les valeurs de p données précédemment. Ces conditions constituent le cas n° 3 porté dans le tableau 1.

On voit en premier lieu que lorsque l'on dispose à la fois des effectifs de migrants et des proportions dans les quatre catégories possibles, les estimations restent cohérentes avec les valeurs attendues mais sont plus dispersées que dans le cas n° 2 (modèle [9]). Cela est dû au fait que dans chaque catégorie, la probabilité de migrer est toujours centrée sur la même valeur, mais sa dispersion est beaucoup plus élevée. En revanche, l'effet de la caractéristique agrégée n'est pas mis en évidence.

Dès que l'on ne distingue plus les migrants selon la catégorie à laquelle ils appartiennent, on va tomber sur des estimations des probabilités totalement irréalistes : certaines estimations deviennent fortement négatives de façon tout à fait significative. On voit que dans ce cas le fait d'être chômeur ou de vivre seul (caractéristiques individuelles) n'a plus qu'un effet négligeable sur la probabilité de migrer dans une région donnée, et que l'effet prépondérant est joué ici par le pourcentage de chômeurs présents (caractéristique agrégée). Cela apparaît clairement lorsque l'on estime le modèle [5], qui donne un effet fortement négatif (-0,503) du pourcentage de chômeurs, l'effet du pourcentage d'isolés étant non significatif.

Ainsi, dans ce cas, travailler au niveau individuel permet d'estimer les probabilités de migrer des diverses catégories de population, avec cependant un biais non négligeable, mais ne met pas en évidence l'effet de la caractéristique agrégée sur ces probabilités. A l'inverse, si l'on travaille au niveau agrégé, on peut arriver à des résultats complètement divergents selon que le pourcentage de migrants est décomposé en fonction de la catégorie de population à laquelle ils appartiennent, ou qu'il ne l'est pas. Dans le premier cas, on tombe sur des résultats proches de ce que l'on avait au niveau individuel. Dans le second cas, c'est l'effet de la caractéristique agrégée qui peut devenir prépondérant, laissant dans l'ombre celui des caractéristiques individuelles.

Pour conclure, on peut dire que les modèles individuels donnent généralement des résultats d'interprétation claire sur les comportements des individus, en fonction des diverses caractéristiques qu'ils ont. En revanche, les modèles agrégés sont d'interprétation beaucoup plus délicate : dans certains cas, ils peuvent donner une idée des comportements individuels avec une marge importante d'erreur ; dans d'autres cas, ils fournissent une information sur les comportements

agrégés ; dans le cas le plus général ils fournissent une information sur le mélange des deux comportements. Il ne sera guère possible en face de telles données de dire dans quelle situation on se trouve, sans disposer simultanément des données individuelles ou au moins d'une information sur les comportements individuels[6].

Pour démêler cet écheveau d'effets, il devient nécessaire d'aborder de front une analyse qui fasse intervenir simultanément divers niveaux d'agrégation.

2. Analyse faisant intervenir des groupes non structurés dans un modèle atemporel

L'existence de regroupements spatiaux, tels que les bassins d'emploi pour la mobilité géographique ou professionnelle, peut entraîner des comportements très différents d'une aire à l'autre. Ainsi par exemple, certains de ces bassins fortement spécialisés pousseront les chômeurs ou ceux qui recherchent un autre type d'emploi à migrer, alors que d'autres bassins, par la plus grande variété des emplois qui y sont offerts, maintiendront sur place beaucoup plus d'actifs occupés ou de chômeurs. On peut penser que cet effet joue sur tous les habitants d'une même zone, sans y introduire une quelconque structure, mais en faisant apparaître une hétérogénéité entre les divers bassins. Bien entendu, les caractéristiques individuelles joueront toujours sur cette mobilité. On voit ainsi apparaître à la fois un effet du chômage individuel et un effet du taux de chômage de la zone dans laquelle vit l'individu sur sa probabilité de migrer. Également les mêmes individus peuvent être impliqués dans d'autres types de groupements non structurés, qui peuvent jouer sur leurs comportements : quartier, ville, région linguistique, etc. Enfin nous observons ici les individus à un moment donné, laissant pour le prochain chapitre l'analyse de l'effet du temps sur leurs comportements. Comment, dès lors, faire intervenir de façon sensée ces divers niveaux d'agrégation, pour ne pas retomber sur les incertitudes décrites plus haut ?

En premier lieu, il semble utile de se placer à un niveau d'agrégation donné de façon à pouvoir faire intervenir l'effet des diverses caractéristiques individuelles et agrégées, sur l'unité de base de ce niveau. Dans ces conditions, la solution a priori la plus satisfaisante est de privilégier l'individu pour faire intervenir sur son comportement l'effet de ses propres caractéristiques et de celles des divers niveaux plus agrégés. Cette solution que nous envisagerons ici plus en détail, n'est cependant pas la seule possible et nous discuterons de ce problème dans la conclusion.

---

6 C'est ainsi que certains auteurs cherchent à mieux cerner des comportements considérés comme purement individuels, alors qu'ils disposent essentiellement de données agrégées avec cependant un petit échantillon au niveau individuel (Holt et al., 1996 ; Steel et al., 1996). Cela leur permet d'améliorer leurs estimations et de réduire l'erreur écologique afférente à l'observation agrégée.

En second lieu, nous avions fait intervenir précédemment, soit un aléa individuel, u^{ij} (formule [1]), soit un aléa au niveau plus agrégé, $u^{.j}$ (formule [6]). Il nous faut donc maintenant introduire simultanément divers types d'aléas et résoudre les problèmes que pose l'estimation de ces termes aléatoires.

Enfin, dans la mesure où l'on veut faire intervenir de façon synthétique l'effet des diverses caractéristiques individuelles et agrégées sur le phénomène étudié, il est nécessaire d'utiliser ici un modèle plus formalisé que les précédents : il doit permettre de décrire les relations entre les changements dans les caractéristiques mesurées à divers niveaux d'agrégation et les changements dans la probabilité individuelle de connaître l'événement étudié, ici la migration. Parmi les divers types de modèles possibles[7], nous privilégierons ici le modèle logit.

La réponse y^{ij} d'un individu (1 s'il migre, 0 si non) étant supposée distribuée selon une loi binominale, $B\left(p^{ij},1\right)$, on peut écrire :

$$p^{ij}=\left(1+exp\left[-\left\{\left(\beta_{00}+\varepsilon_{00}^{j}\right)x_{00}^{ij}+\left(\beta_{10}+\varepsilon_{10}^{j}\right)x_{10}^{ij}+\left(\beta_{01}+\varepsilon_{01}^{j}\right)x_{01}^{ij}+\left(\beta_{11}+\varepsilon_{11}^{j}\right)x_{11}^{ij}+\gamma_{1}x_{1.}^{.j}\right\}\right]\right)^{-1} \quad [17]$$

ce qui conduit au modèle : $y^{ij}=p^{ij}+e^{ij}\sqrt{p^{ij}\left(1-p^{ij}\right)}$ [18]

où la variance de e^{ij} au niveau individuel est égale à un et les variances et covariances entre $\varepsilon_{00}^{j},\varepsilon_{10}^{j},\varepsilon_{01}^{j}$ et ε_{11}^{j} au niveau régional peuvent être estimées par des méthodes appropriées (Goldstein, 1995). On peut facilement montrer que ce modèle d'où l'on a éliminé l'effet du pourcentage de chômeurs $\left(x_{1.}^{.j}\right)$ est identique au modèle [1], sous réserve de poser : $\hat{p}_{kl}=\left[1-exp\left(-\hat{\beta}_{kl}\right)\right]^{-1}$ [19]

Ainsi l'application d'un tel modèle au cas n°1 du tableau 1 conduit à des variances et covariances nulles ou non significativement différentes de zéro entre $\varepsilon_{00}^{j},\varepsilon_{10}^{j},\varepsilon_{01}^{j}$ et ε_{11}^{j} et à des estimations $\hat{\beta}_{kl}$, conduisant, en utilisant [14], à des valeurs de $\hat{p}_{kl}$ identiques à celles données par la formule [3] dans le tableau 1 (voir tableau 2 : cas n° 1)

En revanche, son application au cas n° 3, permet une estimation de l'effet des différentes régions par les variances et covariances entre $\varepsilon_{00}^{j},\varepsilon_{10}^{j},\varepsilon_{01}^{j}$ et ε_{11}^{j}, toujours lorsqu'on ne fait pas intervenir le pourcentage de chômeurs dans chaque région (tableau 2, cas n° 3, modèle 1). Ces diverses variances et covariances sont significativement différentes de zéro, montrant un effet régional important, qui n'avait pas pu être décelé en utilisant le modèle [1]. Les valeurs des coefficients non aléatoires correspondent toujours aux valeurs de p estimées par la formule [3]. Si l'on fait maintenant intervenir simultanément le pour-

7 On trouvera une présentation plus détaillée de ces modèles et les avantages d'utiliser le modèle logit dans Mc Cullagh et Nelder (1989).

Tableau 2. Estimation des paramètres des modèles logit, de leur écart-type et des probabilités correspondantes $p_{00}, p_{10}, p_{01}, p_{11}$ dans les cas n° 1 et n° 3

	Cas n°1		Cas n°3		
			Modèle 1		Modèle 2
Caractéristiques	*Paramètres p du modèle logit*	*Probabilités correspondantes*	*Paramètres p du modèle logit*	*probabilités correspondantes*	*Paramètres p (modèle avec le % des chômeurs)*
Fixes :					
00 Actif occupé, non isolé	-2,582 (0,028)	0,070	-2,670 (0,054)	0,065	-1,639 (0,078)
10 Chômeur, non isolé	-2,215 (0,038)	0,098	-2,224 (0,053)	0,098	-1,170 (0,091)
01 Actif occupé, seul	-1,747 (0,035)	0,149	-1,763 (0,052)	0,146	-0,751 (0,074)
11 Chômeur, seul	-1,346 (0,045)	0,207	-1,375 (0,054)	0,202	-0,354 (0,082)
% de chômeur	-				-4,102 (0,298)
Aléatoires :	*Variances et covariances*		*Variances et covariances*		*Variances et covariances*
niveau 2					
σ^2_{00}	0		0,081 (0,026)		0,002 (0,008)
$\sigma_{10/00}$	0		0,060 (0,020)		0
σ^2_{10}	0		0,030 (0,025)		0
$\sigma_{01/00}$	0		0,080 (0,022)		0,005 (0,006)
$\sigma_{01/10}$	0		0,057 (0,020)		0
σ^2_{01}	0,016 (0,011)		0,080 (0,024)		0,008 (0,008)
$\sigma_{11/00}$	0		0,067 (0,021)		0
$\sigma_{11/10}$	0		0,036 (0,019)		0
$\sigma_{11/01}$	-0,006 (0,012)		0,069 (0,021)		0
σ^2_{11}	0,003 (0,022)		0,055 (0,026)		0
niveau 1					
σ^2_e	1,000 (0,022)		1,000 (0,007)		1,000 (0,007)

centage de chômeurs dans chaque région, on voit disparaître l'effet aléatoire régional (tableau 2, cas n° 3, modèle 2). Celui-ci est donc entièrement expliqué par les pourcentages de chômeurs dans chaque région.

On voit apparaître là un risque d'inférence erronée : l'omission d'une caractéristique indépendante de celles observées peut conduire à une forte différenciation des diverses régions, alors que l'intervention de cette caractéristique élimine entièrement cet effet (Courgeau et Baccaïni, 1997). Il importe dès lors d'observer et de faire intervenir correctement le maximum de caractéristiques jouant sur le phénomène étudié, pour éviter de conclure à un effet régional incorrect. C'est là le risque de toute analyse multi-niveaux : on ne peut jamais être certain qu'un effet régional mis en évidence ne soit pas dû à une caractéristique non observée qui vient entièrement expliquer cet effet. Elle permet cependant d'indiquer la possibilité d'existence d'une telle caractéristique.

3. Vers des modèles biographiques multi-niveaux

Nous avons jusqu'à présent omis de faire intervenir le temps dans notre analyse et nous avons supposé que les groupes n'étaient pas structurés. Il nous faut maintenant lever ces deux hypothèses pour mettre en œuvre des modèles biographiques multi-niveaux. Faisons d'abord intervenir le temps tout en ne faisant pas jouer la structure des groupes.

3.1. Analyse biographique dans des groupes non structurés

Nous suivons maintenant dans une cohorte donnée l'arrivée d'un événement au cours du temps, qui dépendra à la fois de caractéristiques individuelles pouvant se modifier dans le temps, et de la situation de l'individu dans les divers groupes qui peuvent influencer son comportement. Ces groupes d'appartenance n'ont aucune raison de rester semblables pendant la durée d'observation : leur composition et leurs caractéristiques peuvent se modifier à chaque instant. Qui plus est, l'individu n'a aucune raison de rester dans les mêmes groupes tout au long du temps. On voit dès lors que les conditions de collecte et les méthodes d'analyse de telles biographies multi-niveaux seront beaucoup plus complexes que dans le cas classique où l'on ne distingue pas ces niveaux.

Si des enquêtes biographiques rétrospectives permettent de recueillir des données individuelles, avec une précision suffisante pour l'analyse (Poulain et al., 1991 ; Courgeau, 1991), le recueil de données à la fois biographiques et multi-niveaux sera beaucoup plus lourd.

En premier lieu, pour pouvoir réaliser une telle analyse, la taille de l'échantillon devra être beaucoup plus importante que celle des enquêtes biographiques habituelles : il faut en effet pouvoir disposer d'un effectif suffisant dans chacune des régions considérées pour mettre en

évidence l'effet de chaque niveau d'agrégation. Si quelques milliers d'individus suffisent pour une enquête biographique, il sera nécessaire de pouvoir disposer de plusieurs dizaines de milliers d'individus si l'on veut que cette enquête soit également multi-niveaux.

En second lieu, une telle observation doit recueillir non seulement des biographies individuelles, mais suivre au cours du temps de nombreuses caractéristiques des régions dans lesquelles les individus vivent : variation de la densité du peuplement, présence d'un hôpital, etc. (Lazarsfeld et Menzel, 1961). Pour ce faire il faudrait "mettre en œuvre des systèmes d'observation représentatifs des 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, 1989, p.364). Nous entendons plus précisément ici sous le terme d'"enquête sociologique individuelle", une enquête biographique. Il ne fait aucun doute que l'on est encore loin de pouvoir disposer d'un système d'observation aussi complet. Nous devons donc, pour le moment, nous contenter de sources moins riches pour effectuer une telle analyse.

Voyons maintenant ce qu'il en est des méthodes d'analyse. S'il est possible sans trop de difficultés de généraliser les méthodes d'analyse biographique classiques (modèle de Cox, par exemple) à une analyse faisant intervenir divers niveaux d'agrégation (Godstein, 1995), un problème complexe se pose dès que l'on veut réaliser de telles estimations.

En effet, à partir du moment où l'on fait intervenir le temps dans l'analyse, comment peut-on intégrer la mobilité des individus entre les diverses régions ? Supposons que l'on étudie en multi-niveaux la fécondité d'une cohorte. On sait faire intervenir correctement diverses caractéristiques individuelles pouvant changer au cours du temps, pour expliquer ce comportement : effet des diverses situations d'activité ou d'inactivité dans lesquelles la femme peut se trouver, effet d'un changement de statut matrimonial, etc. On peut également faire intervenir les zones et leurs caractéristiques propres pour voir leur effet sur les comportements féconds : on sait qu'il y a des comportements féconds différents selon les départements en France, que ces comportements dépendent de leurs conditions socio-économiques, etc. Mais qu'advient-il lorsque l'individu migre entre divers départements au cours de sa vie féconde ? La fécondité va-t-elle être immédiatement affectée par les conditions socio-économiques des départements où il réside, ou au contraire cet effet joue-t-il avec une période d'adaptation. La première hypothèse conduit à un modèle de Markov : la nouvelle venue prend immédiatement les comportements de la zone d'accueil, en oubliant les conditions de vie antérieures. Cette hypothèse, qui résoudrait facilement le problème posé par les migrations, semble malheureusement peu vraisemblable. Il est dès lors utile de vérifier à partir d'enquêtes biographiques quelles autres hypothèses sont plus proches de la réalité. Il est en particulier possible de tester la rapidité d'adaptation aux conditions de la région d'accueil, si celle-ci se pro-

duit, ou les conditions de sélection des migrants dans la région d'origine, si cette autre hypothèse est vérifiée (Courgeau, 1987). Cela conduit à des modèles non-markoviens beaucoup plus complexes, dont la mise en place en est à ses débuts.

3.2. Analyse biographique dans des groupes structurés

Il reste maintenant à faire intervenir la structure des groupes, pour voir comment celle-ci peut agir sur les comportements de leurs membres. Il s'agira essentiellement de groupes de faible taille, tels que la famille, le ménage ou l'entourage. On va alors chercher à mettre en évidence leur action sur le devenir de chacun de leurs membres et réciproquement de montrer le rôle d'un acteur individuel sur le devenir du groupe (Lelièvre et al., 1997).

Si l'on peut suivre l'individu tout au long de sa vie, du moins jusqu'au moment de l'enquête, le ménage qui a un sens clair à un instant donné, devient impossible à suivre de façon longitudinale sans utiliser des critères arbitraires (Keilman et Keyfitz, 1988). Il paraît dès lors utile de privilégier la notion d'entourage centrée sur un individu, qui peut être plus facilement suivi tout au long de son existence (Bonvalet et Lelièvre, 1995). On rattache, dans ce cas, à l'individu les membres des différents ménages dans lesquels il a vécu et quelques membres clefs de sa famille avec lesquels il n'a pas forcément cohabité. Ces divers membres sont alors suivis, qu'ils cohabitent ou non avec l'individu : on peut penser qu'en plus des diverses caractéristiques de l'enquêté et de sa vie passée, cet entourage doit avoir un rôle important dans les décisions qu'il peut prendre et réciproquement.

La collecte de données permettant de suivre l'entourage d'un individu va à nouveau poser des problèmes.

On peut, en premier lieu, utiliser des enquêtes biographiques classiques qui ont coutume d'interroger l'individu sur diverses personnes faisant partie de son entourage. Mais ces enquêtes recueillent peu de renseignements, et cela sur certains membres de l'entourage seulement, et ne les suivent plus après le départ du domicile de l'enquêté.

Il est dès lors préférable d'utiliser le suivi par enquête prospective de tous les individus composant un ménage initial et y ajouter au cours du temps tous les autres individus qui ont vécu avec ses divers membres (Ott, 1995). Cela permet de reconstituer la majeure partie de l'entourage d'un individu tout au long du temps et de repérer tous les événements qui l'affectent. Peuvent encore y échapper des membres clefs de la famille de l'enquêté avec lesquels il n'a pas cohabité. Une telle enquête n'est cependant pas parfaite. Elle peut perdre lors de leur suivi des individus ayant rompu avec les membres du ménage observé et étant partis sans laisser d'adresse en refusant de poursuivre l'enquête. Surtout il faut un très long délai avant de pouvoir disposer d'une observation longitudinale suffisamment remplie pour commencer l'analyse. C'est là le principal inconvénient de ces enquêtes prospectives.

Une troisième possibilité est d'interroger rétrospectivement un individu sur son entourage et sur l'histoire de vie de ses différents membres. On ne peut malheureusement pas s'attendre dans ce cas à une histoire aussi détaillée que ce que donnerait une enquête prospective. Il y aura également de nombreuses erreurs de mémoire qui viendront fausser ses résultats. Une telle enquête *Biographie et Entourage* est en cours à l'INED (Lelièvre, Bonvalet et Courgeau) : elle complète une enquête biographique classique, par la localisation au cours du temps des enfants et des parents (filiation directe) dès qu'ils ont quitté le domicile de l'enquêté, ainsi que certains autres renseignements sur leur profession. Elle fournira ainsi la structure du réseau de parenté direct tout au long de la vie de l'enquêté et devrait permettre de déduire l'existence d'une influence lorsqu'il y a proximité sociale ou résidentielle. Enfin une telle enquête étant rétrospective, il n'y a pas de risque de perdre des enquêtés lassés par un passage annuel, et surtout elle peut être immédiatement analysée.

Voyons maintenant comment analyser l'évolution de ces groupes. Nous présenterons ici deux pistes en allant de la plus simple à la plus complexe.

Une première approche envisage la biographie de l'entourage considéré comme un individu composite. Il sera quand même décrit à la fois par des caractéristiques collectives (type de logement, effectif du ménage ou de l'entourage à un moment donné, etc.) et par des caractéristiques individuelles des différentes membres du groupe. Les événements étudiés seront par contre essentiellement de type collectif (changement de composition du ménage, changement de localisation, etc.), les événements individuels n'étant considérés que de façon implicite (l'arrivée ou le départ d'un ou plusieurs membres du ménage n'étant pas individualisé, mais considéré seulement comme un changement de composition du ménage). Dans ces conditions, les méthodes d'analyse biographique classiques du groupe considéré comme un individu peuvent s'appliquer sans modifications majeures (Lelièvre et al., 1997).

Une telle approche n'est cependant pas parfaite car le processus étudié est un processus multivarié. Il faut donc pousser plus avant l'analyse et étudier la distribution jointe des différentes durées individuelles des membres de l'entourage : dans ce cas les biographies des divers membres du groupe ne sont plus indépendantes. La modélisation statistique de tels processus a déjà été entamée (Clayton et Cuzick, 1985 ; Bandeen-Roche et Liang, 1996) dans des cas simples de dépendance. Il est nécessaire de la poursuivre en vue d'une application plus précise à des biographies liées entre elles. Une voie prometteuse considère que la dépendance entre biographies individuelles provient de l'influence de caractéristiques observées de l'entourage. Dans ce cas, il est possible d'étudier les biographies d'individus en interaction comme si elles étaient indépendantes : il s'agit d'un artifice purement formel, puisque la dépendance est prise en compte par les caractéristiques de l'entourage, mais qui permet d'étudier des proces-

sus multivariés à l'aide des outils classiques (Lelièvre et al., 1997). Il s'agit là d'une nouvelle voie de recherches très prometteuse.

Conclusions

Nous avons essayé dans cette communication de donner une vue d'ensemble de l'intérêt des analyses multi-niveaux en sciences sociales.

Après un examen théorique des résultats obtenus en travaillant sur des données individuelles ou sur des données agrégées, nous avons d'abord montré les risques d'inférence erronée lorsque l'on ne se situe qu'à un seul niveau d'agrégation : risque d'ignorer l'effet de caractéristiques agrégées lorsque l'on travaille sur des données individuelles ; incapacité de savoir si l'on met en évidence un effet agrégé du niveau auquel on se place ou un effet individuel lorsque l'on travaille sur des données agrégées.

L'utilisation de modèles multi-niveaux permet d'éviter de telles conclusions erronées : ils distinguent correctement l'effet de caractéristiques individuelles de l'effet des caractéristiques des divers niveaux d'agrégation considérés, et d'un effet aléatoire propre à chaque niveau. Il est cependant nécessaire de mesurer toutes les caractéristiques jouant sur le phénomène étudié, pour être certain de l'existence de ce dernier effet aléatoire. Il en résulte, à notre avis, qu'il faut le prendre plus comme un résidu non expliqué par le modèle mis en œuvre, que comme une caractéristique géographique ayant une existence indépendante du contenu social des zones considérées (Jones, 1997). Ces niveaux n'existent pas dans un espace totalement abstrait et vide, mais apparaissent comme la concrétisation spatiale de formes sociales dont on cherche à comprendre la structure (Giddens, 1987).

Une telle analyse multi-niveaux était au départ atemporelle. Il fallait donc, dans une deuxième étape, y introduire le temps. Cela s'est révélé possible à l'aide de modèles biographiques multi-niveaux. Cependant, dans ce cas, il devient indispensable de faire intervenir l'interaction entre la mobilité des individus et les phénomènes étudiés. Les modèles markoviens habituellement utilisés peuvent dès lors se révéler inadéquats et conduire à la mise en place de modèles non markoviens beaucoup plus complexes. La vérification des hypothèses théoriques à leur base nécessite aussi des enquêtes portant sur des échantillons beaucoup plus importants que ceux des enquêtes biographiques habituelles.

Enfin, nous avons introduit une structure dans certains niveaux d'agrégation, tels que l'entourage, pour mieux comprendre leur action sur les comportements individuels. Cela revient à ne plus considérer chaque histoire de vie comme indépendante des autres, mais au contraire à introduire une dépendance entre certaines biographies. Ainsi le départ de certains membres d'un ménage peut-il induire une migration des restants vers un plus petit logement. Des modèles biographiques, qui tiennent compte d'une telle dépendance, mais qui

permettent toujours d'utiliser une modélisation de type classique (modèle de Cox par exemple), constituent une avancée dans ce domaine.

Cette dernière approche nous permet de reposer de façon plus précise une question que nous avions abordée lors de la mise en place de modèles multi-niveaux plus simples. Ces modèles privilégiaient l'individu, pour mettre en évidence l'effet de ses propres caractéristiques et de celles de divers niveaux plus agrégés, sur ses comportements. Mais dans le cas de l'entourage, ne peut-on pas considérer ce niveau agrégé comme l'unité privilégiée de l'analyse ? Nous avons déjà indiqué comment cela était possible.

De façon semblable, ne pourrait-on pas modéliser l'évolution d'un niveau d'agrégation donné, considéré comme l'unité d'analyse privilégiée ? Cette évolution peut dépendre de règles propres à ce niveau d'agrégation, qui peuvent être modifiées par l'action de niveaux encore plus agrégés, mais également par les comportements de niveaux moins agrégés. Nous allons montrer comment cela est possible à l'aide d'un exemple plus précis.

Supposons que l'on étudie les migrations individuelles. Un niveau d'agrégation s'impose dans une telle étude : celui des marchés locaux du logement. Ce marché, par le coût des logements proposés, leur composition, leur nombre, etc., va agir sur les probabilités individuelles de migrer. Mais inversement les individus, en fonction de leurs possibilités d'investir dans un logement de taille et de coût donnés, vont avoir une action réciproque non négligeable sur les règles de ce marché du logement. Ainsi si l'on se centre maintenant sur les divers marchés, on aura à mettre en évidence leurs règles propres de fonctionnement. Mais l'on voit que ces règles peuvent être modifiées par des actions individuelles, du niveau inférieur, et par des décisions des acteurs de l'offre et des politiques plus générales de construction, de niveaux supérieurs.

C'est donc un modèle plus complet qu'il est nécessaire de mettre en place pour tenir compte des actions réciproques des divers niveaux les uns sur les autres. L'élaboration d'un tel modèle devrait amener à une réflexion épistémologique plus poussée sur les modèles multi-niveaux. Bien que cela ne soit pas l'objet de cet article, nous poserons pour le terminer un certain nombre de questions sur les bases de ces modèles.

En premier lieu, quelle signification doit-on accorder aux divers niveaux d'agrégation ? Le niveau individuel est-il marqué par l'expérience passée de l'individu et sa liberté relative, limitée en fait par les contraintes imposées par la société dans laquelle il vit et par les conséquences non intentionnelles de son action ? Les niveaux agrégés sont-ils le reflet de l'organisation sociale complexe dans laquelle nous vivons, telle que la famille, l'entourage, la ville, le pays, etc. ? Dans ce cas, parmi la grande variété de niveaux que l'on peut faire intervenir, n'y en a-t-il pas un certain nombre qui soient plus pertinents que les autres ? Il serait alors nécessaire d'analyser la pertinence des divers niveaux, pour mettre en évidence ceux qui sont à privilégier.

Comme on le voit, nombreuses sont les questions posées par l'analyse multi-niveaux et l'on est encore bien loin de pouvoir leur apporter une réponse. En revanche, ces questions montrent qu'une telle approche peut conduire à une théorie des comportements humains dont l'épistémologie, les méthodes de mesure et d'analyse restent encore largement à établir. Nous espérons que les recherches à venir confirmeront la richesse d'une telle piste.

Bibliographie

ALKER, H.R. (1969), "A typology of ecological fallacies", in : DOGAN et ROKKAN (eds), *Quantitative Ecological Analysis*, MIT Press, Massachussets, pp.69-86.

AMRHEIM, J. (1995), "Searching for the elusive aggregation effect : Evidence from statistical simulation", *Environment and Planning*, n° 27, pp.105-120.

ANTOINE, P., BONVALET, C., COURGEAU, D., DUREAU, F. et LELIEVRE, E., (1998), *L'apport des collectes biographiques pour la connaissance de la mobilité*, L'Harmattan, Paris.

BACCAINI, B. et COURGEAU, D. (1996), "Approche individuelle et approche agrégée : utilisation du registre de population norvégien pour l'étude des migrations", in : J.-P. BOCQUET-APPEL, D. COURGEAU et D. PUMAIN (eds), *Analyse spatiale de données biodémographiques*, John Libbey / INED, Paris, pp. 79-104.

BANDEEN-ROCHE, K. et LIANG, K. (1996), "Modelling failure-time associations in data with multiple levels of clustering", *Biometrika*, n° 83, pp. 29-39.

BONVALET, C. et LELIEVRE, E. (1995), "Du concept de ménage à celui d'entourage : une redéfinition de l'espace familial", *Sociologie et Sociétés*, Numéro spécial, *Une nouvelle morphologie sociale*, vol 27, n° 2, pp. 177-190.

CLAYTON, D. et CUZICK, J. (1985), "Multivariate generalizations of the proportional hazard model", *Journal of the Royal Statistical Society* , pp. 82-117.

COURGEAU, D. (1987), "Constitution de la famille et urbanisation", *Population*, vol. 42, n° 1, pp. 57-82.

COURGEAU, D. (1991), "Analyse de données biographiques erronées", *Population*, vol. 46, n° 1, pp. 89-104.

COURGEAU, D. (1994), "Du groupe à l'individu : l'exemple des comportements migratoires", *Population*, vol. 49, n° 1, pp. 7-26.

COURGEAU, D. (1996), "Towards a multilevel analysis in social sciences" & "Vers une analyse multi-niveaux en sciences sociales", in : J.-P. BOCQUET-APPEL, D. COURGEAU et D. PUMAIN (eds), *Analyse spatiale de données biodémographiques*, John Libbey / INED, Paris, pp. 10-22.

COURGEAU, D. et BACCAINI, B. (1997), "Analyse multi-niveaux en sciences sociales", *Population*, vol. 52, n° 4, pp. 831-864.

COURGEAU D. et LELIEVRE, E. (1989), *Analyse démographique des biographies*, INED, Paris.

DURKHEIM, E. (1967), *Les règles de la méthode sociologique*, PUF, Paris (1ère éd., Paris, Félix Alcan, 1895).

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

GIDDENS, A. (1987), *La constitution de la société*, PUF, Paris.

GOLDSTEIN, H. (1995), *Multilevel Statistical Models*, Edward Arnold, London.

HOLT, I., STEEL, D. et TRANMER, M. (1996), "Adjusting for aggregation effects in ecological regression", in : J.-P. BOCQUET-APPEL, D. COURGEAU et D. PUMAIN (eds), *Spatial Analysis of Biodemographie Data*, John Libbey-INED, Paris, pp. 49-62.

JONES, K. (1997), "Multi-level approaches to modelling contextuality : from nuisance to substance in the analysis of voting behaviour", in : W. VERHOEFF (eds), *Place and People : Multi-Level Modelling in Geographical Research*, Urban Research Centre, Utrecht, pp. 19-40.

KEILMAN, N. (1993), "Emerging issues in demographie methodology" in : J.-L. RALLU et A. BLUM (eds), *European population II. Demographic dynamics*, John Libbey / INED, Paris, pp. 483-508.

KEILMAN, N. et KEYFITZ, N. (1988), "Recurrent issues in dynamic household modelling" in : N. KEILMAN, A. KUIJSTEN et Ad. VOSSEN (eds), *Modelling Household Formation and Dissolution*, Clarendon Press, Oxford, pp. 254-285.

LANGBEIN, L. et LICHTMAN, A. (1978), *Ecological Inference*, Sage, Beverly Hills.

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

LELIEVRE, E., BONVALET, C. et BRY, X. (1997), "Analyse biographique des groupes : les avancées d'une recherche en cours", *Population*, vol. 52, n° 4, pp. 803-830.

LORIAUX, M. (1989), "L'analyse contextuelle : renouveau théorique ou impasse méthodologique", in : J. DUCHÊNE, G. WUNSCH et E. VILQUIN (eds), *L'explication en sciences sociales : la recherche des causes en démographie*, Louvain-la-Neuve, Ciaco, pp. 333-368.

McCULLAGH, P. et NELDER, J. (1989), *Generalized Linear Models*, (2nd ed.) Chapman and Hall, London.

OTT, N. (1995), "The use of panel data in the analysis of household structures", in : E. VAN IMHOFF, A. KUIJSTEN, P. HOOIMEIJER et L. VAN WISSEN (eds), *Household Demography and Household Modelling*, Plenum Press, New York, pp. 163-184.

PIANTADOSI, S., BYAR, D. et GREEN, S. (1988), "The ecological fallacy", *American Journal of Epidemiology*, vol. 127, pp. 893-904.

POULAIN, M., RIANDEY, B. et FIRDION, J-M. (1991), "Enquête biographique et registre belge de population : une confrontation des données", *Population*, vol. 46, n° 1, pp. 65-88.

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

STEEL, D., HOLT, D. et TRANMER, M. (1996), "Inférences au niveau unitaire à partir de données agrégées", *Techniques d'enquêtes*, vol. 22, pp. 3-15.

TUMA, N. et HANNAN, M. (1984), *Social Dynamics : Models and Methods*, Academic Press, Orlando.