VERS UNE ANALYSE BIOGRAPHIQUE MULTINIVEAU

D. COURGEAU

INED

1. Introduction

Dès que le démographe quitte le terrain purement descriptif, dans lequel il s'est longtemps cantonné, et recherche les causes des phénomènes qu'il observe, il doit examiner avec la plus grande circonspection les hypothèses à la base de ses raisonnements. En effet, s'il ne prend pas cette précaution, il risque d'aboutir à des conclusions au moins mal assurées sinon illusoires. Ainsi, avec un point de vue du moment, il a été possible de montrer que les Français au chômage n'ont pas une plus forte vulnérabilité à la dissolution de leur couple que les autres (Herpin 1990), alors qu'un point de vue de plus longue durée montre que les risques de rupture conjugale sont très élevés chez les couples ayant eu une forte précarité professionnelle dans le passé (Paugam, 1993). Il importe donc de voir plus en détail les hypothèses à la base de ces démonstrations, apparemment contradictoires, avant de pouvoir conclure à un effet du chômage sur la dissolution des couples.

Il ne peut être question dans cette courte communication de développer le sujet dans toute sa complexité (voir à ce sujet, Duchêne et al., 1989; Franck, 1994; Piaget, 1967; Wunsch, 1988), mais seulement de montrer l'évolution de quelques hypothèses, souvent passées inaperçues, qui sont à la base de raisonnements tenus en démographie. Cela permettra de voir l'intérêt d'une approche biographique multiniveau, qui lève un certain nombre d'incertitudes dans les résultats obtenus sous ces hypothèses.

2. Données utilisées pour la démonstration

Pour rendre la démonstration plus explicite, nous allons travailler sur des données du registre de population norvégien, qui furent centralisées et informatisées en 1964. Les données biographiques de ces registres ont été ici couplées aux informations recueillies lors des recensements de 1970 et 1980[1]. L'histoire résidentielle complète de chacun de ces individus est connue à partir de 1964, mais le souci des services statistiques norvégiens de préserver l'anonymat des individus, nous a conduit à travailler au niveau régional, la Norvège étant divisée en 19 régions (Baccaïni et Courgeau, 1996).

Nous observons ici tous les hommes nés en 1948, résidant en Norvège en 1991 et n'ayant fait aucune migration vers l'étranger, soit un total de 28 462 individus. Nous considérons les changements de région effectués sur une courte période de deux ans suivant le recensement de 1970.

Enfin nous ne prendrons, dans cet exemple simplifié, qu'une seule caractéristique de l'individu au recensement : le fait qu'il soit agriculteur ou non. Egalement, comme nous le verrons plus loin, le registre norvégien ne nous permet pas de considérer simultanément migrations et changements de profession, car ces derniers ne sont pas enregistrés de façon continue au cours du temps.

3. D'une approche agrégée à une approche multiniveau

Plaçons-nous d'abord dans le cadre d'une analyse démographique classique, qui va chercher à voir l'effet d'une caractéristique, ici le fait d'être agriculteur, sur un comportement démographique, ici la migration interrégionale. Que va-t-on faire pour y arriver?

La solution usuellement choisie consiste à décomposer le pays sur lequel on travaille en régions et montrer que la probabilité d'émigrer d'une région est linéairement liée à son pourcentage d'agriculteurs. C'est la méthode que Durkheim préconisait déjà en 1895, sous le nom de méthode des variations concomitantes, reprise par Landry (1945) sous le nom de rapport de concomitance et que nous appelons maintenant *méthode de régression linéaire* sur caractéristiques agrégées. Voyons plus en détail les hypothèses qui soutiennent cette approche.

[1] Nous remercions ici les services statistiques norvégiens qui nous ont permis d'avoir accès à des fichiers créés à partir de ces données par Kjetil Sørlie et Øjsten Kravdal.

Supposons qu'il existe, dans un pays, des groupes d'individus homogènes vis à vis d'un phénomène donné. Ainsi on peut penser que les membres d'une même profession (ici les agriculteurs) ont la même propension à la migration, et qu'elle est différente de celle des autres professions. On peut considérer que ces tendances collectives ont une existence qui leur est propre, indépendamment des individus qui constituent ces groupes : dans ce cas, ce sont des forces aussi réelles que les forces physiques, bien qu'elles soient d'une autre nature (Durkheim, 1930).

Cela permet de définir une première forme de causalité, qui trouve son origine dans la société elle-même et non dans l'individu et qui s'impose à lui, indépendamment de sa volonté propre. C'est en fonction du système social dans lequel l'individu vit qu'un fait donné est la cause d'un effet social : cela permet d'éviter une analyse sur des données individuelles car les données agrégées, dont on dispose plus facilement, suffisent pour la faire. On peut également penser que ces forces ont une assez grande stabilité dans le temps et ne changent qu'avec une extrême lenteur : cela conduit à une analyse du moment. On reconnaît là une forme extrême du holisme méthodologique.

Voyons plus précisément comment expliquer l'action causale exercée respectivement par le fait d'être agriculteur ou non, sur les taux de migration d'un pays donné. Si l'on connaît, par exemple, les pourcentages d'individus qui migrent ainsi que les pourcentages d'agriculteurs, dans chaque région du pays, il sera possible de porter sur un graphique la propension à la migration des régions selon leur part d'agriculteurs. Dans ce cas, on va montrer que, si les hypothèses initiales sont vérifiées, on aura une relation linéaire entre ces deux caractéristiques.

En effet, si m_a et $m_{\bar{a}}$ sont les taux de migration des agriculteurs et des non-agriculteurs, qui gardent la même intensité dans toutes les régions du pays considéré, si N_{aj} et $N_{\bar{a}j}$ sont les effectifs des deux populations dans une région j donnée, alors le nombre de migrations observées de cette région, M_j, sera :

$$M_j = N_{aj}m_a + N_{\bar{a}j}m_{\bar{a}} = N_{aj}(m_a - m_{\bar{a}}) + N_j m_{\bar{a}}$$

où N_j est la population totale de la région j. Il en résulte bien que le taux de migration, m_j, sera fonction linéaire de la proportion d'agriculteurs dans chaque région, a_j :

$$m_j = a_j(m_a - m_{\bar{a}}) + m_{\bar{a}}$$

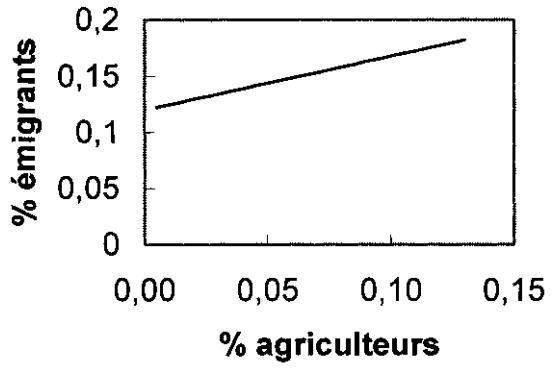
L'observation des résultats sur les données agrégées par région nous permet de voir quel est l'effet du pourcentage d'agriculteurs sur la probabilité de migrer (tableau 1, colonne intitulée " Régression linéaire agrégée ").

Paramètres	Régression linéaire agrégée		Régression logistique (résultats agrégés)	
m_a	0,597	(0,197)	0,094	(0,069)
$m_{\bar{a}}$	0,119	(0,014)	0,150	(0,012)
$m_a - m_{\bar{a}}$	0,478	(0,209)	-0,056	(0,070)

Tableau 1.- Paramètres estimés et écart type entre parenthèse, pour les modèles agrégés

On voit qu'il y a une liaison positive tout à fait significative, entre le taux de migration et la proportion d'agriculteurs (Figure 1). Sous les hypothèses posées plus haut, les paramètres de la régression nous permettent de dire que les agriculteurs ont une plus forte propension à la migration (plus de cinq fois plus élevée) que les autres professions.

Figure 1: Modèle linéaire agrégé



Mais si ces hypothèses ne sont pas valides, le résultat de ces observations ne permet plus de conclure que les agriculteurs ont de plus fortes chances de migrer que les autres. Tout ce qu'elles permettent de dire, c'est qu'un fort pourcentage d'agriculteurs dans la population, conduit à un plus fort taux d'émigration d'ensemble, qui peut concerner aussi bien les agriculteurs que les autres professions, ou bien même seulement les autres professions. On reconnaît là le risque d'*erreur écologique*, que l'on peut rencontrer lorsque l'on cherche à déceler des comportements individuels à partir de données agrégées : ce risque avait été mis en évidence dès 1950 par Robinson. Très souvent, les chercheurs n'ont pas la possibilité de vérifier ce risque, car les données agrégées dont ils disposent ne le permettent pas.

Puisque nous disposons ici des données individuelles, nous devrions nous attendre à une confirmation de la plus forte probabilité d'émigrer des agriculteurs, si les hypothèses précédentes sont vérifiées, sinon nous avons commis une erreur écologique. Nous allons donc estimer un modèle de régression logistique, qui fait intervenir pour chaque individu sa profession. Si l'on suppose qu'un *modèle logit* s'applique correctement à nos données individuelles, on peut écrire la probabilité pour que l'individu i présent dans la région j soit un migrant ($\mu_{ij} = 1$) en fonction du fait qu'il soit agriculteur ($a_{ij} = 1$) ou non ($a_{ij} = 0$), sous la forme suivante :

$$P(\mu_{ij} = 1 \mid a_{ij}) = (1 + exp-[\alpha_0(1 - a_{ij}) + \alpha_1 a_{ij}])^{-1}$$

Dans ce cas, lorsque l'individu est agriculteur, sa probabilité de migrer s'écrit :

$$P(\mu_{ij} = 1 \mid a_{ij} = 1) = (1 + exp[-\alpha_1])^{-1} = m'_a$$

et lorsqu'il a une autre profession :

$$P(\mu_{ij} = 1 \mid a_{ij} = 0) = (1 + exp[-\alpha_0])^{-1} = m'_{\bar{a}}$$

On peut, à partir de ces probabilités, calculer le nombre total de migrants prévus dans la zone j, comme la somme des probabilités de migrer de chaque individu de cette zone :

$$M'_j = \sum_j P(\mu_{ij} = 1 \mid a_{ij}) = N_{aj} m'_a + N_{\bar{a}j} m'_{\bar{a}}$$

Cela conduit à une formule identique à celle de la régression précédente, mais les probabilités de migrer sont cette fois-ci calculées à l'aide des paramètres

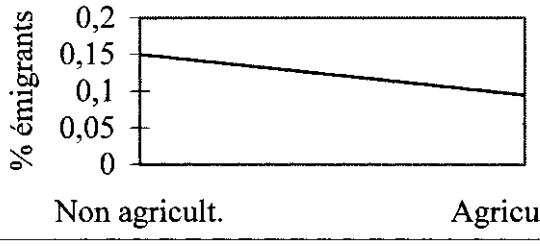
α_0 et α_1, estimés à partir des données individuelles et non à partir des données agrégées. Le tableau 2 porte l'estimation des paramètres du modèle logit simple.

Paramètres	Modèle logit		
	simple	contextuel sans interaction	contextuel avec interaction
α_0 (non agriculteur)	-1,736 (0,017)	-1,984 (0,032)	-1,996 (0,033)
α_1 (agriculteur)	-2,260 (0,084)	-2,614 (0,093)	-2,155 (0,209)
α_2 (% agriculteurs)		4,266 (0,453)	4,469 (0,461)
α_3 (agric. × % agric.)			-5,774 (2,447)

Tableau 2 .- Paramètres estimés et leur écart type entre parenthèse, par les modèles individuels logit

On voit maintenant que les agriculteurs ont une bien plus faible probabilité de migrer que les autres professions (Figure 2). Il est possible d'estimer les pourcentages théoriques de migrants prévus dans chaque région, si ce modèle s'applique à l'ensemble de la population. Cela conduit à un résultat inverse de ce que nous avions observé à partir des données agrégées : voir tableau 1, colonne intitulée " Régression logistique (résultats agrégés) ". Cela fournit une probabilité de migrer des agriculteurs (0,094) plus d'un tiers inférieure à celle des autres profession (0,15), et à une pente de la droite de régression négative (-0,056), bien que non significative. Ce résultat, en parfaite contradiction avec ce que le modèle précédent montrait, nous conduit à modifier les hypothèses des deux modèles pour rétablir une cohérence entre eux.

Figure 2: Modèle logit



Observons d'abord que le résultat obtenu avec les données individuelles semble a priori plus clair que celui obtenu avec les données agrégées. Nous avions en effet indiqué les difficultés d'interpréter les résultats obtenus avec les données agrégées, avec le risque d'erreur écologique. Mais il faut cependant prendre garde à un autre risque d'inférence erronée lorsque l'on travaille sur des données individuelles. Le danger est alors de commettre l'*erreur atomiste,* lorsqu'on ignore l'influence du contexte dans lequel les comportements humains ont lieu. En fait ce contexte peut avoir lui aussi un rôle et dans ce cas il est erroné de considérer les individus isolés des contraintes et des possibilités imposées par le milieu dans lequel ils vivent.

D'où l'idée d'introduire l'influence de caractéristiques contextuelles et de travailler simultanément à divers niveaux d'agrégation avec l'objectif d'expliquer un comportement qui est maintenant considéré comme individuel et non plus collectif, comme dans la première approche. Cela lève le risque de commettre l'erreur écologique, car une caractéristique agrégée n'est plus utilisée 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.

Cela conduit à des *modèles contextuels*, qui introduisent les deux types de caractéristiques simultanément. Le modèle logit correspondant peut alors s'écrire :

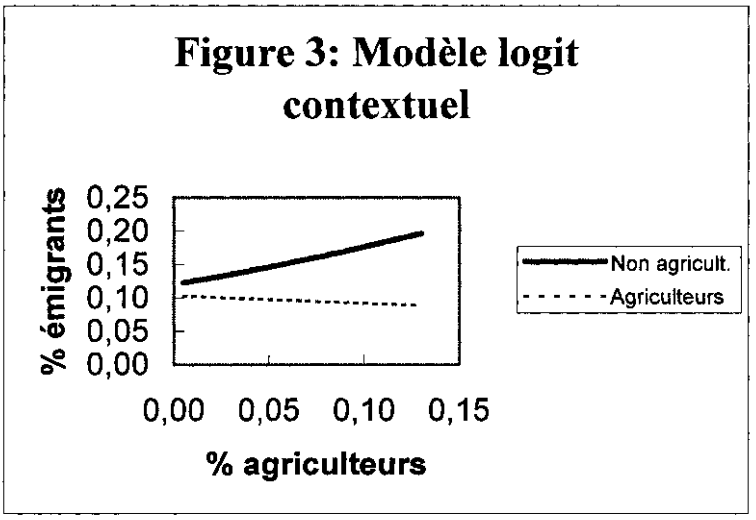
$$P(m_{ij} = 1 \mid a_{ij}, a_{.j}) = (1 + exp-[\alpha_0(1 - a_{ij}) + \alpha_1 a_{ij} + \alpha_2 a_{.j} + \alpha_3 a_{ij} a_{.j}])^{-1}$$

où $a_{.j}$ est la proportion d'agriculteurs observée dans la région *j*. Ainsi le terme α_2 représente l'effet contextuel du pourcentage d'agriculteurs sur la probabilité de migrer, tant des agriculteurs que des autres professions, et le terme α_3 représente un effet d'interaction, qui permet de faire dépendre l'effet contextuel de la profession de l'individu.

Dans le cas que nous étudions, les résultats sont également portés dans le tableau 2, dans les colonnes intitulées " Modèle logit contextuel sans ou avec interaction ". Il est utile de considérer d'abord le modèle ne faisant pas intervenir de terme d'interaction. On voit que l'effet des deux caractéristiques, « être agriculteur » et « proportion d'agriculteurs », est significatif et que leurs paramètres sont de signe opposé : le fait d'être agriculteur réduit toujours aussi fortement les chances de migrer, mais lorsque le pourcentage d'agriculteurs augmente cela entraîne une augmentation des chances de migrer de l'ensemble de la population, permettant d'expliquer la contradiction apparente précédente. On peut de plus penser que la plus grande rareté des emplois non agricoles dans les régions où le pourcentage

d'agriculteurs est important, entraîne une plus forte mobilité interrégionale des autres professions.

On a la possibilité de vérifier cela, en utilisant le modèle avec interaction (colonne 3 du tableau 2). On peut montrer que, si les chances de migrer des agriculteurs ne dépendent pas significativement du pourcentage d'agriculteurs, le paramètre correspondant étant égal à –1,305 (4,469-5,774) non significativement différent de zéro, car son écart type est égal à 2,403, celles des autres professions croissent très fortement lorsque ce pourcentage augmente.



On peut également penser qu'un effet régional plus général va jouer sur les chances d'émigrer. D'où l'idée de mettre d'abord en évidence cet effet régional avant de faire intervenir des caractéristiques agrégées, comme le pourcentage d'agriculteurs. Nous nous dirigeons alors vers un *modèle multiniveau*, qui permet de faire intervenir des aléas régionaux pour prendre en compte cet effet (Courgeau et Baccaïni, 1997, 1998).

Un tel modèle peut s'écrire :

$$P(\mu_{ij}=1 \mid a_{ij}) = p_{ij} = (1 + exp-[(\alpha_0 + u_{0j})(1 - a_{ij}) + (\alpha_1 + u_{1j})a_{ij}])^{-1}$$

où les paramètres u_{0j} et u_{1j} sont des aléatoires, de moyenne nulle et dont on va estimer les variances et covariance :

$$var(u_{0j}) = \sigma^2_{u0} \qquad var(u_{1j}) = \sigma^2_{u1} \qquad cov(u_{0j}, u_{1j}) = \sigma_{u01}$$

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

$$\mu_{ij} \approx B(p_{ij}, 1)$$

Dans ce cas on a la variance conditionnée suivante :

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

et 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 $\sigma^2_e = 1$

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 régional. Il est cependant possible de libérer la variance individuelle de la contrainte d'égalité à l'unité, pour vérifier qu'un modèle logit s'applique correctement aux données. Ces paramètres et leurs variances- covariance ont été estimés ici à l'aide du logiciel Mln (Goldstein, 1995).

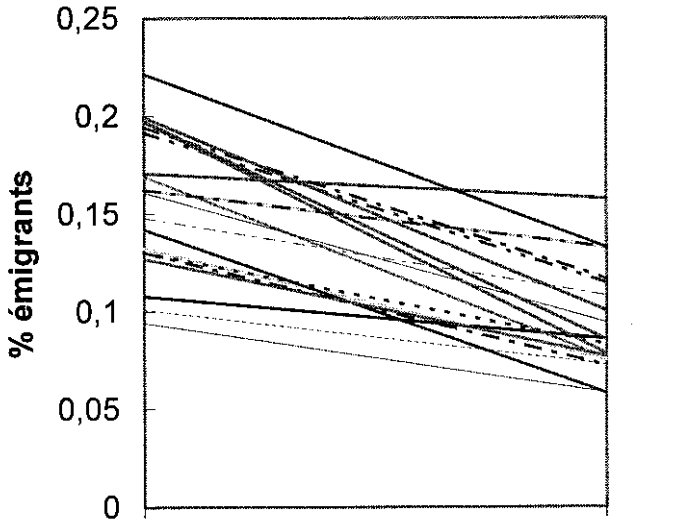
Le tableau 3 porte les résultats de l'application de ce modèle aux données norvégiennes, dans la colonne 2 intitulée « Modèle multiniveau simple ».

Paramètres	Modèle multiniveau		
Fixes :	simple	contextuel sans interaction	contextuel avec interaction
α_0 (non agriculteur)	-1,710 (0,070)	-2,150 (0,110)	-2,067 (0,119)
α_1 (agriculteur)	-2,306 (0,133)	-2,786 (0,200)	-2.017 (0,340)
α_2 (% agriculteurs)		6,654 (0,989)	5,420 (1,209)
α_3 (agric. × % agric.)			-8,691 (3.238)
Aléatoires :			
$\sigma^2_{u_0}$ (non agriculteur)	0,088 (0,032)	0,049 (0,025)	0,047 (0,024)
$\sigma_{u_{01}}$ (covariance)	0,054 (0,044)	0,104 (0,068)	0,085 (0,042)
$\sigma^2_{u_1}$ (agriculteur)	0,167 (0,135)	0,312 (0,238)	0,181 (0,119)

Tableau 3 .- Paramètres estimés et leur écart type entre parenthèse, par les modèles individuels multiniveaux

Les paramètres fixes de ce modèle sont très proches de ceux du modèle logit correspondant. Les paramètres aléatoires montrent une dispersion significative des régions pour les non agriculteurs (σ^2_{u0}), et une dispersion encore plus forte pour les agriculteurs (σ^2_{u1}), cependant non significativement différente de la précédente. La figure 4 montre la position des diverses régions les unes par rapport aux autres et concrétise la faible corrélation entre les aléas correspondant aux agriculteurs et aux non agriculteurs.

Figure 4: modèle multiniveau



Le fait de laisser varier la variance individuelle nous montre que le modèle logit s'applique parfaitement car on a : $\sigma_e^2 = 0{,}9998$, valeur non discernable de l'unité.

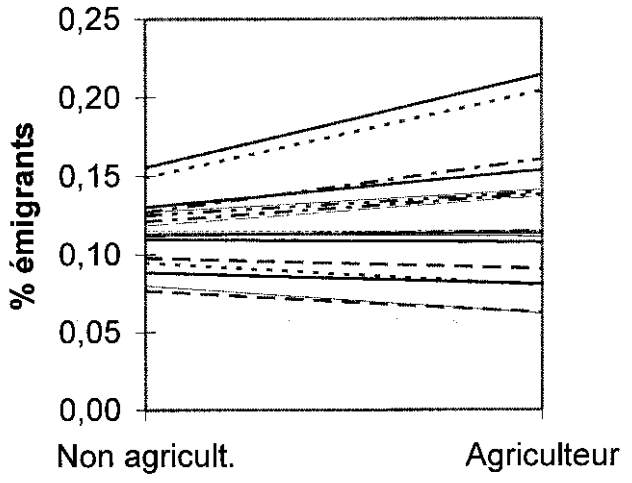
Il reste maintenant à introduire la caractéristique agrégée, la proportion d'agriculteurs dans chaque région. Le modèle s'écrit dans ce cas :

$$P(\mu_{ij}=1 \mid a_{ij}, a_{.j}) = p_{ij} = (1-exp-[(\alpha_0+u_{0j})(1-a_{ij})+(\alpha_1+u_{1j})a_{ij}+\alpha_2 a_{.j}+\alpha_3 a_{ij}a_{.j}])^{-1}$$

Ses résultats sont portés dans le tableau 3, dans les colonnes 3 et 4 intitulées " Modèle multiniveau contextuel sans ou avec interaction ". Le modèle sans interactions a des paramètres fixes qui sont de nouveau proches de ceux du modèle logit contextuel et sont tous significatifs. Les paramètres aléatoires marquent une forte réduction de variance pour les non-agriculteurs, montrant clairement que la prise en compte des pourcentages régionaux d'agriculteurs joue sur cette variance, alors que pour les agriculteurs l'effet est encore plus élevé. Le modèle avec interaction conduit à nouveau à des paramètres fixes assez proches de ceux du modèle contextuel correspondant. Il marque en revanche un retour de la variance correspondant aux

agriculteurs à ses valeurs obtenues pour le modèle multiniveau simple, montrant une prise en compte de cette variance par le terme d'interaction. Il est également intéressant de voir que la corrélation entre les termes aléatoires est très proche de l'unité, montrant, qu'une fois pris en compte l'effet du pourcentage d'agriculteurs, les régions où les probabilités de migrer des agriculteurs sont importantes sont également des régions où les probabilités de migrer des non agriculteurs seront importantes, et vice versa (Figure 5).

Figure 5: Modèle multiniveau contextuel



Un tel modèle permet de réconcilier les résultats contradictoires des approches agrégée et individuelle et permet d'introduire un effet régional plus général dans l'estimation, qui bien que réduit par l'intervention de la caractéristique agrégée, reste toujours significatif.

4. D'une approche du moment à une approche biographique

Nous avons jusqu'à présent travaillé sur des données du moment, comme dans l'exemple proposé dans l'introduction (Herpin, 1990), sans la possibilité de faire intervenir des dépendances entre événements. Cette approche est à relier au paradigme prévalant antérieurement en démographie, selon lequel le démographe ne peut étudier que l'arrivée d'un événement et d'un seul sous l'hypothèse qu'il est indépendant des autres phénomènes et qu'il se produit dans une population supposée homogène (Henry, 1959). Un tel paradigme ne paraît plus tenable à partir du moment où l'on peut observer au cours du temps, grâce à des enquêtes détaillées, l'histoire de vie complète des individus, dans le domaine familial, professionnel et migratoire, par exemple (Courgeau, 1999). Le second exemple, proposé dans l'introduction, illustre bien cette possibilité de dépendance entre événements survenus à des instants très éloignés dans le temps (Paugam, 1993).

Nous avons donc proposé un nouveau paradigme (Courgeau et Lelièvre, 1996, 1997), qui permet d'appuyer l'approche plus synthétique que nous préconisons : 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.

Ce paradigme permet de faire intervenir les divers événements à l'œuvre dans une vie humaine, comme des processus en interaction les uns avec les autres, et constitue une approche intégrant les divers niveaux d'agrégation, allant de l'individu à la société dans laquelle il vit. Ces divers événements peuvent maintenant être dépendants les uns des autres et il est possible de faire intervenir l'hétérogénéité de la population sur laquelle on travaille. Voyons rapidement comment cela est possible.

Supposons que l'on travaille toujours sur les liens entre migration et agriculture. Cette fois-ci, l'individu peut changer de profession plusieurs fois au cours de sa vie et effectuer un grand nombre de migrations. Pour simplifier la formulation, nous faisons ici l'hypothèse que l'individu effectue au cours de sa vie au plus une migration interrégionale et une mobilité entre l'agriculture et les autres professions. Il est bien entendu possible de généraliser cette formulation au cas le plus général.

Dans le cas simplifié, nous avons une interaction entre deux types d'événements avec aussi deux états d'origine possibles : l'individu est initialement agriculteur ou non. Le traitement de ces deux cas est dès lors symétrique et nous n'en avons qu'un seul à présenter. Supposons que l'on travaille sur les individus initialement dans le secteur agricole.

Considérons les deux variables aléatoires T_m et $T_{\bar{a}}$, correspondant à l'arrivée de la migration et à celle de la mobilité hors du secteur agricole. On peut dès lors formaliser les distributions conditionnelles à l'aide des quotients instantanés d'émigration à l'instant t :

$$h_{am}(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(T_m < t + \Delta t \mid T_m \geq t, T_{\bar{a}} \geq t)$$

qui est le quotient d'émigration, si la mobilité hors de l'agriculture n'est pas encore survenue, et par :

$$h_{\bar{a}m}(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(T_m < t + \Delta t \mid T_m \geq t, T_{\bar{a}} = u) \qquad u < t$$

si la mobilité hors du secteur agricole est survenue à l'instant u antérieur à t. Des quotients symétriques peuvent être estimés pour la mobilité hors du secteur agricole avant ou après émigration, respectivement $h_{\bar{m}a}(t)$ et $h_{ma}(t)$.

L'estimation de ces divers quotients est possible, sous des conditions très générales (Andersen et al., 1993). Leur comparaison permet de déceler des dépendances complexes entre ces divers événements (Courgeau et Lelièvre, 1989, 1992).

Une *dépendance unilatérale,* par exemple, lorsque les quotients cumulés $H_{am}(t)$ et $H_{\bar{a}m}(t)$ sont significativement différents alors que les quotients cumulés $H_{\bar{m}a}(t)$ et $H_{ma}(t)$ sont identiques, montrant que le fait d'être agriculteur joue sur la migration, alors que le fait d'avoir migré ne joue pas sur le départ de l'agriculture.

Une *dépendance réciproque* lorsque les deux couples de quotients cumulés sont significativement différents et une *indépendance totale* lorsqu'ils sont identiques.

Une telle indépendance totale, posée comme hypothèse lorsqu'on réalise une analyse démographique longitudinale classique, est en fait la plus rarement observée lorsqu'on travaille sur des données biographiques réelles. Bien entendu, d'autres types de dépendances plus complexes sont également possibles, mais nous ne poursuivrons pas l'analyse dans cette direction et renvoyons le lecteur intéressé aux manuels qui développent ces questions (Courgeau et Lelièvre, 1989, 1992).

La comparaison des quotients cumulés permet aussi de vérifier si un *modèle à risques proportionnels* s'applique à ces données, comme c'est souvent le cas en démographie. Si tel est le cas, on peut écrire de façon plus synthétique les deux modèles précédents sous la forme d'un modèle de Cox (1972) généralisé :

$$h_{.m}^{ij}(t) = h_{am}(t)\,exp[\beta_1' \overline{a}_{ij}(t)]$$

où $h_{am}(t)$ est le quotient instantané sous-jacent d'émigration des individus initialement agriculteurs, et $\overline{a}_{ij}(t)$ une caractéristique binaire dépendant maintenant du temps. Initialement nulle, elle devient égale à l'unité si l'individu *i,* vivant dans la région *j* et initialement agriculteur, quitte ce secteur. Il est bien sur possible de faire intervenir dans un tel modèle la caractéristique agrégée $a_{.j}(t)$, correspondant à la part d'individus restés agriculteurs à la date *t,* dans la région *j*.

Enfin on pourra prolonger ce modèle en faisant intervenir différents niveaux d'agrégation, pour estimer un *modèle biographique multiniveau* (Goldstein, 1995; Horowitz, 1999) :

$$h_{.m}^{ij}(t \mid a_{ij}(t), a_{.j}(t)) = h_{am}^{.j}(t)\,exp([\beta_1' + u_{1j}']\overline{a}_{ij}(t) + \beta_2' a_{.j}(t))$$

Le quotient instantané peut maintenant dépendre de la région *j* ainsi que l'effet de la caractéristique binaire fonction du temps $\overline{a}_{ij}(t)$, par l'intermédiaire d'un paramètre qui dépend également de la région u_{1j}'. A nouveau il existe des procédures pour estimer les différents paramètres, tant fixes qu'aléatoires, de ce type de modèle.

Comme nous l'avons indiqué plus haut, il ne nous est pas possible à l'aide des données du registre norvégien d'estimer ces modèles. En effet, on dispose de l'information sur la profession des individus à la date des recensements de 1970 et 1980, mais on n'a pas la possibilité de la suivre de façon continue au cours du temps, comme cela est nécessaire pour estimer ces divers modèles. Il aurait été possible, si nous avions disposé d'un plus grand nombre de générations successives, d'utiliser les méthodes mises au point pour l'analyse biographique de données fragmentaires (Courgeau et Najim, 1995, 1996). Mais même dans ce cas, l'intervalle de dix ans entre recensements paraît trop important pour permettre une estimation suffisamment précise.

En fait, pour pouvoir réaliser une telle analyse de façon correcte, il faudrait disposer d'enquêtes, permettant le suivi au cours de leur existence des membres d'un échantillon de grande taille, pour bien pouvoir distinguer les niveaux d'agrégation. 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 une analyse biographique pleinement multiniveau.

5. Conclusion

Le cheminement suivi dans cet article montre en fait une alternance entre induction et déduction correspondant à deux types de causalité différents, que nous allons maintenant détailler (Courgeau, 2000).

La *déduction* permet de dériver des conclusions de prémisses données. Elle part d'hypothèses supposées vérifiées et en dérive des conclusions en utilisant des règles logiques explicites. Lorsque les prémisses sont vérifiées, ces conclusions doivent rester identiques, quelle que soit la façon dont la démonstration est faite.

Lorsque diverses conclusions, issues des mêmes prémisses, sont contradictoires entre elles, on peut alors être certain que ces prémisses sont incorrectes. Mais, même si ces conclusions sont cohérentes entre elles, on ne pourra jamais conclure que les hypothèses de départ sont vraies, car de nombreuses autres prémisses peuvent conduire à la même conclusion.

Ainsi, nous avons d'abord supposé que les agriculteurs avaient tous la même probabilité de migrer, quelle que soit leur position sur le territoire : ils sont donc homogènes vis à vis de la migration. Cela constitue notre hypothèse de départ, identique à celle de Durkheim, qui considérait que tous les membres d'un même groupe (ici les agriculteurs, chez Durkheim les protestants) avaient la même propension à connaître un événement donné (ici la migration, chez Durkheim le suicide), quelle que soit leur province de résidence.

Dans ce cas, nous avons d'abord cherché à vérifier à l'aide de données agrégées, qu'il y avait bien une variation linéaire de la proportion de migrants, lorsque la proportion d'agriculteurs d'une région croît. Il se trouve ici que cette linéarité est bien vérifiée : on peut donc en conclure par déduction que les agriculteurs ont une plus forte probabilité de migrer que les autres professions, sous l'hypothèse d'une homogénéité des comportements sur l'ensemble du territoire.

Si nous disposons de données individuelles, nous devrions nous attendre à une confirmation de cette plus forte probabilité. Il se trouve que, dans le cas considéré ici, les agriculteurs ont une bien plus faible chance de migrer que les autres professions, contredisant ainsi les résultats obtenus avec les données agrégées.

Nous tombons donc sur des conclusions, qui nous montrent que notre hypothèse de départ est incorrecte. Comment dès lors modifier cette hypothèse pour rendre cohérentes entre elles ces conclusions contradictoires? Il nous faut alors nous tourner vers l'*induction*, dont la définition nécessite une discussion plus approfondie.

Si nous suivons l'approche de Popper (1973) et définissons l'induction en disant qu'elle permet de répondre à " la question de savoir comment établir des

énoncés universels fondés sur l'expérience ", force nous est de reconnaître que cette induction ne peut jamais être considérée comme définitive. En effet, il est toujours possible qu'une expérience à venir contredise l'énoncé supposé universel, sans qu'il nous soit possible de le savoir a priori. Dès lors, toute tentative de fonder ce " principe d'induction sur l'expérience échouera puisque celle-ci doit conduire à une régression à l'infini ".

Cependant, d'autres définitions de l'induction ont été proposées, qui permettent d'avancer dans ce domaine. Si l'on se reporte aux définitions données dans différents dictionnaires, on voit que l'induction peut être considérée comme un " raisonnement qui va du particulier au général " (Larousse) ou qui permet de remonter " de cas particuliers à une proposition plus générale " (Robert).

Ces définitions n'impliquent plus la multiplication à l'infini des expériences pour s'assurer de la régularité de la relation d'induction. Mais, au contraire, celle-ci reçoit ici un sens différent de celui que lui donnait Popper : il ne s'agit plus de mettre en place une hypothèse universelle, mais de chercher des hypothèses plus générales que les précédentes, qui permettent de réconcilier des conclusions divergentes.

Ainsi, lorsqu'un chercheur fait des observations inattendues au vu des théories prévalantes, il va être amené à mettre en place d'autres hypothèses pour expliquer ces divergences. On peut qualifier ce raisonnement du nom de processus de " *rétroduction* " (Greenland, 1998). Ces nouvelles hypothèses n'auraient jamais pu voir le jour sans ces observations précises et l'on peut dire que ce sont les données qui les ont générées. Pour les chercheurs, qui identifient induction et rétroduction, le problème n'est plus de savoir si cette dernière existe, car il est évident qu'elle s'est depuis très longtemps révélée être très féconde pour la recherche, mais de montrer comment elle opère.

Ainsi, devant l'incohérence des résultats obtenus sous l'hypothèse que les agriculteurs ont tous la même propension à l'émigration, il nous faut poser de nouvelles conditions, qui permettent de résoudre ces contradictions, cela de façon la plus simple possible. Un modèle contextuel fournit une de ces solutions, en faisant intervenir une propension individuelle à la migration, qui sera modifiée par les conditions particulières de chaque région dans lesquelles ces individus vivent. En particulier, lorsque ces régions auront un pourcentage important d'agriculteurs, la probabilité d'en émigrer sera plus forte pour l'ensemble de la population, du fait sans doute de la plus grande rareté des emplois non agricoles.

De même, les problèmes rencontrés, lorsque l'on suppose que l'action d'événements individuels est seulement immédiate, conduisent à une approche biographique qui étend cette action tout au long de l'existence, tout en permettant un effet du moment de certaines caractéristiques. Enfin, pour accorder les deux généralisations précédentes, il devient nécessaire de mettre en place une approche biographique en multiples niveaux.

L'approche de la démographie défendue ici est à relier au courant de l'individualisme méthodologique en sciences sociales, opposé au holisme méthodologique. Ce dernier peut aller jusqu'à nier l'existence d'unités individuelles, pour ne retenir qu'une explication en termes de relations entre variables globales. L'individualisme méthodologique va considérer le comportement individuel comme primordial, mais peut simultanément, comme dans le modèle multiniveau, considérer l'effet de variables globales sur ce même comportement. On peut dire qu'à la limite il peut contenir le holisme, ce qui amène à le privilégier pour une approche démographique des phénomènes sociaux.

Une réflexion épistémologique plus approfondie s'impose enfin pour définir la signification à accorder aux différents niveaux d'agrégation utilisés, pour tenir compte de la structure sociale des groupes (Lelièvre et al., 1997, 1998), considérés ici comme homogènes, pour analyser les changements qui s'opèrent à des niveaux plus agrégés que le niveau individuel, etc.

Il s'agit de mettre en place une théorie plus complète et non contradictoire 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 les divers niveaux d'agrégation qui interviennent dans les sciences sociales et de faire jouer le temps vécu par les individus.

Bibliographie

Andersen P., Borgan O., Gill R., Keiding N. (1993), *Statistical models based on counting processes*, Springer-Verlag, New York.

Baccaïni B., Courgeau D. (1996), Approche individuelle et approche agrégée : utilisation du registre de population norvégien pour l'étude des migrations, in Bocquet-Appel, Courgeau et Pumain eds. : *Analyse spatiale de données biodémographiques*, John-Libbey/INED, Paris, pp. 79-104.

Courgeau D. (1999), L'enquête " Triple biographie : familiale, professionnelle et migratoire ", in Groupe de réflexion sur l'approche biographique : *Biographies d'enquêtes*, Méthodes et savoirs, PUF diffusion/INED, Paris, pp.59-74.

Courgeau D. (2000), Réflexions sur la causalité en sciences sociales, *Recherches et prévisions*, à paraître.

Courgeau D., Baccaïni B. (1997), Analyse multi-niveaux en sciences sociales, in *Nouvelles approches méthodologiques en sciences sociales*, *Population*, Courgeau ed., vol. 52, n°4, pp. 831-864 (**(1998),** Multilevel analysis in the social sciences, in *New methodological approaches in the social sciences, Population : An English Selection*, Courgeau ed., vol. 10, pp. 39-71).

Courgeau D., Lelièvre E. (1989), *Analyse démographique des biographies*, Éditions de l'INED, Paris.

Courgeau D., Lelièvre E. (1992), *Event history analysis in demography*, Clarendon Press, Oxford.

Courgeau D., Lelièvre E. (1996), Changement de paradigme en démographie, *Population*, vol. 51, n° 2, pp. 645-654 (**(1997),** Changing paradigm in demography, *Population : An English Selection*, vol. 9, pp. 1-10).

Courgeau D., Najim J. (1995), Analyse de biographies fragmentaires, *Population*, vol. 50, n° 1, pp. 149-168 (**(1996),** Interval censored event history analysis, *Population : An English Selection*, vol 8, pp. 191-208).

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

Duchêne J., Wunsch G., Vilquin E. eds. (1989), *L'explication en sciences sociales : la recherche des causes en démographie*, Éditions Ciaco, Bruxelles.

Durkheim E. (1930), *Le suicide*, PUF, Paris (1 ère ed. 1896).

Durkheim E. (1937), *Les règles de la méthode sociologique*, PUF, Paris (1 ère ed. 1895).

Franck R. ed. (1994), *Faut-il chercher aux causes une raison? L'explication dans les sciences humaines*, Librairie Philosophique Vrin, Paris.

Goldstein H. (1995), *Multilevel statistical models*, Kendall's Library of statistics, Arnold, London.

Greenland S. (1998), Induction versus Popper : substance versus semantics, *International Journal of Epidemiology*, 27, pp. 543-548.

Henry L. (1959), D'un problème fondamental de l'analyse démographique, *Population*, vol. 13, n°1, pp. 9-32.

Herpin N. (1990), La famille à l'épreuve du chômage, *Économie et Statistique*, n° 235, pp. 31-42.

Horowitz J.L. (1999), Semiparametric estimation of a proportional hazard model with unobserved heterogeneity, *Econometrica*, vol. 67, n° 5, pp. 1001-1028.

Landry A. (1945), *Traité de démographie*, Payot, Paris.

Lelièvre E., Bonvalet C., Bry X. (1997), Analyse biographique des groupes. Les avancées d'une recherche en cours, in *Nouvelles approches méthodologiques en sciences sociales, Population*, Courgeau ed., vol. 52, n° 4, pp. 803-830 (**(1998)**, Event history analysis of groups. The findings of an on-going project, in *New methodological approaches in the social sciences, Population : an English Selection*, Courgeau ed., vol. 10, pp. 11-37).

Paugam S. (1993), Précarité et risques d'exclusion en France, *Document du CERC*, n° 109, La Documentation Française, Paris.

Piaget J. ed. (1967), *Logique et connaissance scientifique*, Encyclopédie de la Pléiade, Editions Gallimard, Paris.

Popper K.R. (1973), *La logique de la découverte scientifique*, Éditions Payot, Paris (1 ère ed. 1935).

Robinson W.S. (1950), Ecological correlation and the behavior of individuals, *American Sociological Review*, 15, pp. 351-357.

Wunsch G. (1988), *Causal theory and causal modeling*, Leuven University Press, Leuven.