ENJEUX ET LIMITES DE L'ANALYSE MULTINIVEAU EN DEMOGRAPHIE

Valérie Golaz (*), Arnaud Bringé (**)

(*) INED-CEPED, UMR 196 Paris Descartes – INED- IRD
(**) INED, Service des méthodes statistiques

Introduction

Les méthodes de l'analyse multiniveau se développent depuis 20 ans à travers toutes les sciences sociales. Après un démarrage incisif dans les sciences de l'éducation [Goldstein,1986, Raudenbusch et Bryk, 1986], l'épidémiologie [Chaix et Chauvin, 2002] , la sociologie, et la démographie sont devenues des disciplines d'application de l'analyse multiniveau [Di Prete et Forristal, 1994]. L'articulation dans un même modèle de données collectées à des niveaux différents (individu, groupes d'individus, sous populations,…) permet d'outrepasser les limites des modèles classiques [Courgeau, 2002, Hoem, 2007]. La modélisation multiniveau cache néanmoins des difficultés particulières. L'adaptation à la démographie de cet outil a des conséquences sur le cadre théorique dans lequel se place l'étude [Courgeau, 2002]. La définition du modèle entraîne une réflexion sur les niveaux d'agrégation pertinents en relation à un objet donné. L'approche multiniveau permet non seulement de dépasser le niveau individuel, mais aussi de mesurer la part des phénomènes étudiés dont l'explication réside dans chacun des niveaux considérés. Pour la démographie, une discipline de plus en plus analytique et explicative [Tabutin 2007], qui cherche à dépasser les contraintes d'une explication trop individu-centrée des phénomènes humains, c'est une avancée importante. Cependant, la mise en œuvre de ces modèles est tributaire de l'existence ou de la conception de données appropriées, ce qui explique un développement relativement limité jusqu'à présent en dehors de quelques objets d'étude particuliers.
A titre illustratif, faisant le bilan des papiers et posters acceptés à la PAA en 2007, seuls 26 sur plus de 1000 communications mobilisent l'analyse multiniveau. Abordant des sources de données et des thèmes variés, les analyses portent dans une très grande majorité de cas sur des groupes correspondant à des unités territoriales, du quartier au pays. Depuis 2000 jusqu'à 2008, la revue *Population* n'a publié que 3 articles de ce type. En bref, depuis la revue effectuée par Nick Parr en 1999 [Parr, 1999], les applications de l'analyse multiniveau continuent à se développer, mais très lentement. Malgré le nombre croissant de manuels sur le sujet [Courgeau 2004, Goldstein, 2003, Hox, 2002, Raudenbusch et Bryk, 2002], ou de manuels de statistique appliquée incluant ces modèles [Bressoux, 2008], malgré une certaine prolixité sur l'avancée que représentent ces méthodes pour les sciences sociales [Courgeau, 2002] et la démographie en particulier [Hoem, 2007], les applications tardent à venir.
Ce papier a pour objectif de rappeler les principes de l'analyse multiniveau, en ramenant systématiquement la discussion aux enjeux liés à ces méthodes pour le démographe, en termes de collecte de données, d'analyse et d'interprétation des résultats. Cela permettra d'aborder non seulement les points forts de cette approche, mais aussi les raisons pour lesquelles son développement en démographie reste limité.

1. Une contextualisation nécessaire

La raison d'être de tout modèle statistique est de schématiser au mieux une réalité complexe, variée et changeante. Lorsqu'ils ont pour objet les caractéristiques d'hommes ou de groupes sociaux, les modèles n'ont en aucun cas ni le pouvoir ni la prétention de saisir la totalité des phénomènes étudiés. En revanche, ils indiquent des relations de dépendance entre caractéristiques ou processus. Des modèles de plus en plus complexes se développent en mathématiques appliquées, qui ont des échos tout à fait favorables en sciences sociales. On sait désormais modéliser le temps, distinguer les effets de contexte des effets individuels… Mais cela n'a pas toujours été le cas, et l'application de ces modèles à la démographie a nécessité non seulement le développement de l'informatique mais aussi

celui de données adaptées. Les approches mises en œuvre dans le passé étaient tributaires des moyens disponibles.

1.1. L'erreur écologique empêche le passage du macro au micro

Ce sont les analyses sur données agrégées au niveau de populations qui ont été les plus développées jusqu'aux années 1950 en démographie. Au niveau macro, la démographie porte sur l'étude des caractéristiques d'une population, définie selon des critères simples. L'exemple le plus courant est celui de la planification qui doit prendre en compte les caractéristiques de la population vivant dans chacun des territoires défini par le maillage administratif d'un pays. Les populations étudiées peuvent aussi être définies, outre l'unité territoriale, par des critères démographiques (âge, sexe) ou sociaux (religion, statut matrimonial, profession, etc.). A l'approche transversale très courante s'est ajoutée, dans l'après guerre, l'approche longitudinale, qui en suivant une population donnée, a permis l'intégration du temps dans l'analyse démographique.
Que ce soit en transversal ou en longitudinal, le niveau d'analyse est alors un niveau agrégé. Ce niveau d'observation aboutit à des résultats sur des populations entières, qu'il serait incorrect d'interpréter au niveau individuel. C'est ce que l'on appelle l'erreur écologique [Courgeau 2002] ou le biais d'agrégation [Bressoux, 2008]. Un exemple classique est fourni par Durkheim, qui constatant que les régions de Prusse au plus fort pourcentage de protestants sont aussi celles où les taux de suicide sont les plus élevés, fait implicitement le lien entre suicide et protestantisme. Aux Etats-Unis, en 1950, l'étude de référence dans le domaine est celle de Robinson. Il remarque une forte corrélation entre proportion de noirs et taux d'illettrisme, alors qu'au niveau individuel la corrélation est beaucoup plus faible.
Il est donc nécessaire, si l'on s'intéresse aux faits des individus, de porter l'analyse à leur niveau.

1.2. Erreur atomiste

Les analyses au niveau de l'individu, souvent appelé niveau micro, ont dû attendre les progrès de l'informatique pour se développer chez les quantitativistes. La démographie a alors progressivement pu utiliser des modèles statistiques de plus en plus complexes, sur un nombre élevé d'observations. En transversal, la grande famille des modèles logistiques place l'individu au cœur de l'analyse. Les données nécessaires sont individuelles, et ne doivent plus être agrégées. La démographie peut alors relier au niveau individuel les événements qui sont au cœur de ses intérêts avec les autres caractéristiques des individus à un moment donné. Il est aussi désormais possible de prendre en compte le temps dans des modèles individuels. Le développement de l'utilisation de modèles de durée a nécessité la production de données individuelles adaptées, comprenant les durées entre différents événements, ou leurs dates, au cours de la vie des individus étudiés. L'observation, toujours au niveau individuel, peut accompagner les individus (collecte prospective) ou les interroger sur leur vie passée (collecte rétrospective). Les enquêtes biographiques jouent un rôle particulier dans ce groupe, en articulant les trajectoires familiale, résidentielle et professionnelle d'un individu, de sa naissance au moment de l'enquête, dans un questionnaire rétrospectif.
Ce niveau micro d'observation et d'analyse ouvre la porte à l'erreur atomiste [Courgeau 2002]. Expliquer des événements au niveau individuel par les caractéristiques de l'individu uniquement revient à le sortir de son contexte, des mécanismes d'influence qui peuvent avoir un rôle important dans son comportement. Replacer un événement dans la trajectoire individuelle, rechercher des facteurs déterminants dans la vie passée de l'individu, comme le fait l'analyse biographique par exemple, constitue déjà une contextualisation relative. Mais pour échapper à l'erreur atomiste, il est nécessaire d'aller au-delà de la trajectoire individuelle, de replacer l'individu dans un contexte social plus large, celui des groupes humains dans lesquels il évolue, famille, réseaux sociaux, milieu social, société, nation, etc.

1.3. Différentes formes de contextualisation

Dans la modélisation, contextualiser les données individuelles revient à intégrer de manière pertinente un certain nombre de caractéristiques des groupes sociaux ou spatiaux auxquels l'individu appartient. On peut distinguer deux types de variables contextuelles, selon leur source : celles issues de la base de données individuelles sont dites endogènes, celles issues d'autres sources de données sont dites exogènes.
L'agrégation de données individuelles à un niveau où la représentativité de l'échantillon est assurée, permet de construire des variables contextuelles. Ces variables ne sont pas redondantes avec les

variables individuelles. Au contraire, leur prise en compte simultanée permet de distinguer les effets individuels de la caractéristique des effets liés au groupe. Par exemple, les chômeurs auront des comportements différents selon le taux de chômage de leur région. Daniel Courgeau [2004] a montré que la propension à migrer des agriculteurs et des non-agriculteurs entre les différentes régions qui composent la Norvège dépend de la proportion d'agriculteurs dans la région d'origine, qui est une variable contextuelle analysée au niveau individuel. Ainsi l'observation individuelle produit des données agrégées, qui peuvent être analysées au niveau macro ou être réinjectées dans des modèles d'analyse micro, comme des caractéristiques individuelles.
Suivant la même démarche, au niveau micro, il est possible d'adjoindre aux données collectées des caractéristiques contextuelles provenant d'autres sources (par exemple les caractéristiques de l'infrastructure de chaque zone géographique considérée, …), pour une description plus complète du contexte dans lequel se situent les individus. Ainsi, des données agrégées construites à partir d'autres enquêtes, des données environnementales approchées par images satellite, des données exhaustives issues de différents ministères ou d'autres organismes, peuvent être réinjectées dans une base micro et prises en compte aux côtés des variables individuelles.
La prise en compte de caractéristiques contextuelles permet ainsi, au niveau micro, de mesurer les effets des caractéristiques individuelles, mais aussi des caractéristiques contextuelles sur les comportements individuels, et au niveau macro, de pointer les différences entre des populations définies par des caractéristiques contextuelles différentes (régions au faible pourcentage de chômeurs / autres régions, etc.).

2. Du contextuel au multiniveau

Depuis quelques années, il est désormais possible d'articuler dans les mêmes modèles d'analyse différents niveaux d'observation. Contrairement aux modèles contextuels, où toutes les caractéristiques sont analysées au niveau individuel, les modèles multiniveau font référence à une structuration des données, et donc des comportements analysés selon différents niveaux d'observation et d'analyse. Des groupes d'individus y sont définis (modèles à 2 niveaux), ainsi qu'éventuellement des groupes de groupes (modèles à 3 niveaux ou plus) formant les niveaux méso et macro. La structuration en groupes est marquée dans les modèles multiniveau par l'introduction d'aléas au niveau de ces groupes. Ainsi, par rapport à un modèle contextuel au niveau individuel, la variance entre individus est décomposée en deux (ou plus) variances, l'une entre individus au sein des groupes prédéfinis, et l'autre entre groupes. Les trois aspects fondamentaux de la modélisation multiniveau sont la définition des groupes, la modélisation des effets fixes et la modélisation des effets aléatoires. Le choix d'un modèle pour les effets fixes se pose dans toute modélisation statistique et ne sera par conséquent pas développé ici. En revanche, il est important de préciser que la modélisation multiniveau en démographie s'inscrit parfaitement dans le paradigme général des sciences sociales.

2.1. Mesure et paradigme

Les modèles multiniveau ne sortent plus l'événement étudié ou l'individu enquêté de son contexte. Ils permettent d'articuler des observations de nature différente. La disponibilité de telles méthodes plaide pour des collectes simultanées aux différents niveaux d'analyse pertinents pour la question étudiée. L'analyse quantitative menée dans ce sens répond au paradigme général des sciences sociales qui interprète les comportements des individus comme le produit de leur propre histoire et de l'histoire des groupes sociaux et des territoires qu'ils ont traversés, rejoignant en cela des disciplines plus qualitatives.
Mobiliser différents niveaux d'observation dans une même analyse statistique est une nouveauté méthodologique qui apporte à l'approche quantitative une nouvelle dimension. Observer et analyser simultanément à différents niveaux permet, comme la plupart des modèles au niveau individuel, de mesurer les effets des caractéristiques de l'individu, du groupe, et éventuellement leur interaction (en les croisant). Ceci est aussi possible avec des modèles contextuels au niveau individuel. Cependant dans un modèle multiniveau, certains des éléments contextuels sont rattachés à un niveau supérieur, ce qui permet de hiérarchiser les contextes. En quantifiant et décomposant la partie non expliquée du modèle selon les niveaux mobilisés, les modèles multiniveau permettent une meilleure compréhension de la distance qui sépare le modèle des faits observés. En effet, aucun modèle statistique ne peut être en adéquation parfaite avec la réalité des phénomènes sociaux, mais l'intérêt de la modélisation est de se rapprocher le plus possible de cette réalité, aussi complexe soit elle, avec un nombre limité de variables, de caractéristiques. Dans les modèles multiniveau, l'hétérogénéité non expliquée par le modèle est quantifiée à chacun des niveaux du modèle. Elle peut aussi être associée à certaines

caractéristiques du modèle. Ainsi le démographe peut non seulement mesurer l'adéquation du modèle à la réalité exprimée par les données qu'il utilise, et de ce fait tenter de la réduire au mieux, mais aussi savoir à quel niveau ou pour quelles sous populations le modèle pourrait éventuellement être affiné.
Ainsi, l'approche multiniveau, par la finesse de modélisation qu'elle implique, non seulement s'inscrit dans le paradigme général des sciences sociales, mais permet en théorie d'en mesurer les différentes composantes : la part d'explication du phénomène étudié qui revient à chacun des niveaux pertinents d'observation. Dans ce type d'approche, on aura tendance à privilégier une approche parcimonieuse en termes de variables explicatives, et avoir comme objectif une diminution significative des variances calculées à chacun des niveaux.

2.2. Définir des groupes cohérents

Identifier les niveaux d'observation, ou les groupes, à prendre en compte est un véritable enjeu pour la qualité du modèle. Définir les groupes à prendre en compte dans le modèle multiniveau nécessite un compromis entre données et cohérence explicative. Les comportements humains s'inscrivent dans des contextes complexes, familiaux, sociaux, sont contingents des contraintes environnementales ou politiques. Différents types de contextes entrent en compte mais ne peuvent pas tous être modélisés. Cependant, même si les modèles statistiques ne peuvent que caricaturer la réalité, en particulier lorsqu'il s'agit d'étudier des comportements humains, tout l'enjeu est d'approcher le plus simplement possible cette complexité. L'analyse des comportements humains inclut forcément le niveau individuel, mais différents niveaux d'analyse peuvent être envisagés, selon l'objet d'étude et la source de données, au dessous ou au dessus de ce niveau. La règle fondamentale est que les groupes doivent former un ensemble d'unités disjointes qui englobent la totalité de la population étudiée. La représentativité des données individuelles au niveau des groupes n'est pas nécessaire, car le modèle est calculé sur l'ensemble de la base. Les groupes ainsi définis doivent avoir un sens pour le phénomène étudié. Tous les regroupements ne sont pas pertinents.

Deux principaux types de niveaux peuvent être définis en dessous du niveau individuel. Cette démarche nécessite des données de panel ou des données longitudinales, qui pour la plupart se structurent spontanément avec l'individu comme premier niveau d'agrégation [Singer et Willett, 2003]. En effet, l'individu statistique, dans ce type de base de données, est une mesure répétée à diverses échéances (suivi démographique, panel) ou bien un événement qui se reproduit au fil de la trajectoire individuelle (naissances, déménagements, changements professionnels…).
Entre l'individu et l'ensemble de la population étudiée, les groupes eux-mêmes peuvent être définis à différents niveaux : famille, entourage [Bonvalet et Lelièvre, 1995], réseau social, territoire, etc. Certains de ces niveaux peuvent s'emboîter (commune – département – région ou bien ménage – entourage ou encore ménage – commune), mais d'autres non (réseau social et commune ?).

Les analyses menées conjointement sur différents niveaux mettent en lumière ceux d'entre eux qui sont pertinents, eu égard au phénomène étudié. Selon l'objet de la recherche, les niveaux à retenir pour l'observation peuvent différer. Ainsi pour l'étude de la mobilité spatiale entre départements, le département s'impose tout naturellement comme niveau supérieur. Pour l'étude de la réussite scolaire, la classe, l'école, le territoire défini par la carte scolaire constituent des niveaux potentiellement intéressants. Dans le domaine de la santé, le médecin traitant, la structure d'accueil sont des niveaux auxquels il est bon de porter attention. Sur le monde du travail, l'entreprise et le secteur d'activité constituent des niveaux d'observation possibles. De plus, quel que soit le comportement social étudié, il est évident qu'un certain nombre de facteurs sont à rechercher dans l'entourage présent et passé des individus, de même que dans les milieux dans lesquels il a vécu.

En longitudinal, il est nécessaire que les groupes définis soient stables autour de l'événement étudié, ou bien que les modifications de leur composition ou de leurs caractéristiques au cours du temps soient prises en compte. Le temps introduit donc un double problème, qui se porte sur chacun des niveaux d'observation : celui des changements du contexte, et celui des changements de contexte des individus. Comment prendre en compte des contextes évolutifs ? Le plus simple exemple est le cas d'une zone rurale qui au fil du temps s'urbanise. A partir de quand est elle urbaine ? Question qui alimente les débats géographiques, et très concrètement revient à la difficulté de définir un seuil précis [GRAB, 2006]. Comment prendre en compte les changements de groupe des individus ? Cette question soulève des problèmes méthodologiques importants, et entraine le plus souvent la création de modèles adaptés aux changements de groupe des individus au fil de l'observation [Goldstein et al., 2000].

Lorsque les groupes pertinents ne s'emboitent pas, ce qui est souvent le cas lorsque l'on veut prendre en compte des contextes territoriaux et des contextes sociaux, des modèles spécifiques peuvent être mis en œuvre. Il est aussi possible de construire un niveau méso d'analyse à partir des intersections des deux niveaux (fractions de groupes sociaux situées dans des unités spatiales, ou regroupements géographiques au sein du groupe social).

Selon Daniel Courgeau [2004], les groupes sociaux jouent un rôle plus important dans les phénomènes humains que le milieu physique. Néanmoins la plupart des recherches appliquées menées en démographie utilisent des contextes géographiques. Seules quelques études lancent les pistes d'autres applications du multiniveau. Ainsi, Jani Erola et ses collègues [Erola et al., 2008] ont utilisé la famille ou la fratrie comme niveau d'analyse. Le ménage, qui souvent joue le rôle d'individu statistique en économie, n'est que rarement mobilisé comme niveau d'agrégation de caractéristiques individuelles, et ce même dans les sociétés où leur taille moyenne est relativement importante.

Identifier les groupes pertinents pour l'analyse d'un phénomène pose la base de la modélisation. La plupart des exemples connus portent sur des regroupements territoriaux d'individus, et concernent des problématiques spécifiques, liées au territoire (équipement, …). Alors que les groupes sociaux sont reconnus comme fondamentaux, du point de vue sociologique, rares sont les bases de données qui permettent la mise en œuvre d'analyses à leur niveau. Le modèle mis en œuvre est par conséquent fortement tributaire des données existantes. En revanche, lorsqu'il s'agit d'envisager une nouvelle collecte, il est intéressant en amont de considérer les groupes sociaux tout autant que les unités territoriales afin, dans le futur, de pouvoir aller plus loin dans cette direction de recherches.

2.3. Différents modèles, différentes prises en charge du contexte

La modélisation elle-même nécessite une démarche prudente et structurée. Il est important de procéder par étapes, dans un ordre bien défini, du plus simple au plus complexe. En premier lieu, le modèle choisi doit avoir été mis au point au niveau micro, sur les données individuelles au moins. Voyons ici l'exemple d'un modèle logit, modèle couramment utilisé en démographie, où les variables étudiées sont souvent qualitatives. Par soucis de clarté, nous nous arrêterons ici au modèle à deux niveaux, mais le raisonnement et la modélisation suivent les mêmes principes lorsque des niveaux supplémentaires sont pris en compte.

2.3.1.1. L'aléa sur la constante

La première étape de la modélisation consiste à ajouter au modèle logistique simple un aléa sur la constante. Soit Y la variable dépendante, V_1 à V_n les caractéristiques individuelles, la probabilité que le Y de l'individu i dans le groupe j vaille 1 est donnée, dans un modèle logistique simple, par la formule (1).

$$(1) \qquad P(Y_{ij} = 1) = (1 + \exp-[\alpha_0 + \alpha_1 V_{1ij} + \alpha_2 V_{2ij} + \ldots + \alpha_n V_{nij}])^{-1}$$

Autorisons la constante α_0 à prendre une valeur différente dans chaque groupe. Le modèle se transforme en (2), où u_{0j} est un terme aléatoire, dont on fait l'hypothèse qu'il suit une loi normale. Sa variance σ^2_{u0} donne la mesure du résidu au niveau groupe, c'est-à-dire qu'elle donne une indication de la dispersion des groupes, en d'autres termes, de l'hétérogénéité entre groupes qui n'est pas expliquée par les variables prises en compte.

$$(2) \quad P(Y_{ij} = 1) = (1 + \exp-[\alpha_0 + u_{0j} + \alpha_1 V_{1ij} + \alpha_2 V_{2ij} + \ldots + \alpha_n V_{nij}])^{-1} \quad \text{avec } \mathrm{var}(u_{0j}) = \sigma^2_{u0}$$

Pour cette première étape, il est conseillé d'ajouter une à une au modèle ainsi défini les variables explicatives identifiées au préalable, en particulier de travailler progressivement depuis le niveau le plus micro jusqu'au niveau le plus macro des variables explicatives. Ici les variables contextuelles sont donc incorporées délicatement après les variables individuelles (3)

$$(3) \qquad P(Y_{ij} = 1) = (1 + \exp-[\alpha_0 + u_{0j} + \alpha_1 V_{1ij} + \alpha_2 V_{2ij} + \ldots + \alpha_n V_{nij} + \beta_1 C_{1j} + \beta_2 C_{2j} + \ldots + \beta_m C_{mj}])^{-1} \quad \text{avec } \mathrm{var}(u_{0j}) = \sigma^2_{u0}$$

Une définition pertinente des groupes utilisés dans le modèle conduit forcément à une variance σ^2_{u0} significativement différente de 0 en début de modélisation (modèle limité aux caractéristiques explicatives individuelles). A l'inverse, si la variance n'est pas significative, les groupes ne se distinguent pas les uns des autres, et par conséquent aucune partie du phénomène étudié ne repose sur une différentiation selon ce niveau. La démarche de l'analyste consiste à tenter de réduire la variance inter-groupes, en particulier par l'ajout de caractéristiques contextuelles pertinentes.

2.3.1.2. Aléas sur une caractéristique individuelle

Même si la réalité est bien sûr toujours plus complexe que ce que la modélisation permet, il est encore possible de s'en rapprocher. Il est possible de placer l'aléa sur l'une des caractéristiques explicatives, ce qui revient à envisager que cette caractéristique a un effet différent selon le groupe. En général, on suppose également dans ce cas que la constante varie elle aussi selon le groupe considéré. L'équation (4) comporte alors deux coefficients aléatoires, u_{0j} et u_{1j}, que l'on suppose suivre une loi normale centrée, de variances respectives σ^2_{u0} et σ^2_{u1}, et de covariance σ_{u01}.

$$(4)\quad P(Y_{ij} = 1) = (1 + \exp-[\ \alpha_0 + u_{0j} + (\alpha_1 + u_{1j})\ V_{1ij} + \alpha_2\ V_{2ij} + \ldots + \alpha_n\ V_{nij} + \beta_1\ C_{1j} + \beta_2\ C_{2j} + \ldots + \beta_m\ C_{mj}])^{-1}$$
$$\text{var}\ (u_{0j}) = \sigma^2_{u0}$$
$$\text{var}\ (u_{1j}) = \sigma^2_{u1}$$
$$\text{cov}\ (u_{0j}, u_{1j}) = \sigma_{u01}$$

2.3.1.3. Effets d'interaction

Il faut noter que la modélisation multiniveau, tout comme la modélisation simple, prend en charge les variables explicatives construites à partir du croisement de variables individuelles et de variables contextuelles. Un exemple clair est celui développé par Daniel Courgeau [2004], sur les migrations norvégiennes, qui repose sur un modèle du type du type de celui donné par l'équation (5) ci-dessous. Il montre que l'émigration touche différemment les agriculteurs et les non agriculteurs (caractéristique individuelle), que la proportion d'agriculteurs de la région (caractéristique contextuelle) est également significative. Mais il montre aussi que le fait d'être agriculteur ou non a un effet différent sur l'émigration selon la proportion d'agriculteurs de la région (croisement individuel-contextuel).

$$(5)\quad P(Y_{ij} = 1) = (1 + \exp-[\ \alpha_0 + u_{0j} + (\alpha_1 + u_{1j})\ V_{ij} + \beta\ C_j + \gamma\ V_{ij}\ C_j])^{-1}$$
$$\text{var}\ (u_{0j}) = \sigma^2_{u0}$$
$$\text{var}\ (u_{1j}) = \sigma^2_{u1}$$
$$\text{cov}\ (u_{0j}, u_{1j}) = \sigma_{u01}$$

Ces quelques modèles illustrent la complexité de l'approche multiniveau. Le principe relativement simple de la modélisation des aléas ne peut cependant pas être décliné outre mesure, car chaque nouvel aléa introduit entraine un nombre de plus en plus important de coefficients à estimer (variance et covariances avec les autres aléas). Il n'est donc pas question de modéliser, même en présence de données relativement fouillées, la plus grande complexité, mais au contraire de sélectionner les éléments les plus forts, les plus pertinents, quitte à les étudier séparément les uns des autres lorsque le besoin se présente. Néanmoins, il est généralement possible de modéliser des effets d'interaction entre niveaux d'observations, des effets de caractéristiques correspondants à plusieurs types de groupes, et de rendre aléatoires un petit nombre de variables.

La construction d'un modèle multiniveau repose sur différents choix : la modélisation des effets fixes, l'identification des niveaux pertinents à prendre en compte et la modélisation des aléas. Les groupes considérés et la modélisation des aléas permettent d'affiner le modèle simple (celui sur les effets fixes) en l'adaptant un peu plus à la réalité observée. Cependant, cette démarche est limitée par la complexification rapide des calculs sous jacents, et le modèle multiniveau, s'il permet d'approcher cet objectif, ne l'atteint généralement pas. Même si elle est relative, c'est néanmoins une amélioration notable par rapport au modèle simple.

3. L'idéal-type de la modélisation des comportements humains : croiser groupes sociaux et spatiaux dans une perspective diachronique

En théorie, l'approche multiniveau permet d'aller très loin dans la modélisation des phénomènes humains. Le paradigme auquel correspondent les applications des modèles multiniveau en démographie stipule que les caractéristiques d'un individu à un moment donné sont le produit de sa propre histoire comme de celle des groupes sociaux et spatiaux, de tous les « contextes » que l'individu a traversé tout au long de sa vie. Une adhésion maximale à ce paradigme consisterait à non seulement inclure plusieurs niveaux de contexte, de nature différents, mais aussi à dépasser le transversal en prenant en compte le temps dans les modèles mis en œuvre. Si l'idée est séduisante, la mise en œuvre de cette approche dans l'analyse est loin d'être aisée. Elle repose sur l'existence de données micro longitudinales, et d'éléments de contexte correspondant à chacun des « moments » de la trajectoire. La reconstitution à posteriori, au moment de l'analyse, des caractéristiques des contextes vécus par l'individu est délicate et généralement très coûteuse en temps de travail. Elle nécessite l'utilisation de sources secondaires pertinentes, lorsqu'elles existent. Voyons, à travers deux types d'enquêtes biographiques, les solutions adoptées lorsque la reconstitution à partir de sources secondaires n'est pas possible, avant d'aborder les pistes restant à explorer en la matière.

3.1. Des enquêtes spécifiques sur les contextes spatiaux

Le projet Chitwan Valley Family Study, conduit en 1996-1997 dans une région rurale du Népal par l'équipe de W.Axinn (Université du Wisconsin) comprend l'articulation d'une enquête biographique (Individual Life History) et d'une enquête rétrospective sur l'histoire des localités enquêtées (Neighborhood History)[Axinn et al., 1997].

L'enquête burkinabè EMIUB (Enquête Migration et Insertion Urbaine au Burkina Faso) est une enquête biographique nationale, conduite en 2000. Un échantillon important des lieux vécus par les enquêtés, sur la base des parcours résidentiels complets recueillis dans l'enquête, ont fait l'objet d'une post-enquête : une collecte de données contextuelles rétrospectives [Shoumaker et al., 2006].

Ces deux enquêtes, aux objectifs très différents, ont adopté une méthodologie similaire pour répondre au même besoin : apporter des éléments contextuels pertinents pour mieux comprendre et analyser les comportements individuels. Ces éléments de contexte sont d'ordre spatial, il s'agit des caractéristiques des localités de résidence des individus étudiés : infrastructure, équipement, taille de la localité, etc. Dans le cas du Népal comme du Burkina Faso dans son ensemble, les individus qui au cours de leur trajectoire sortent de la zone d'observation sont perdus pour l'analyse, puisque l'on ne dispose pas d'informations sur les autres lieux. La mise en œuvre d'une enquête 'villages' ou 'localités' est donc d'autant plus valorisable que la mobilité de la population étudiée vers les régions/ pays voisins, et vice versa, est faible. Sinon, des précautions particulières devront être prises dans l'analyse.

Ces deux enquêtes illustrent la mise en œuvre réussie d'une collecte multi-niveau biographique. Néanmoins, il n'est pas possible à travers ces projets de reconstituer les groupes sociaux vécus par les enquêtés. L'analyse est donc limitée à la prise en charge d'un seul type de contexte : le village.

3.2. Vers une contextualisation sociale des trajectoires individuelles

Reconstituer a posteriori les groupes sociaux est extrêmement difficile. Certaines configurations simples, du type famille proche, fratrie, peuvent parfois être reconstituées à partir de données historiques. C'est le cas de l'enquête TRA par exemple [Bourdelais, 2004]. Les sites de suivi démographiques, courants dans les pays du Sud, permettent un suivi dans le temps, prospectif, d'unités domestiques [Bringe et Laurent, 2005]. Dans le cas des projets de recherche qui reposent sur une enquête rétrospective, il est nécessaire de penser à la reconstitution du contexte social dès la préparation de la collecte. Recueillir des informations auprès d'ego concernant d'autres individus, aussi proches aient-ils été d'ego dans le passé, n'est pas aisé, mais possible.

C'est ce qu'illustre l'enquête *Biographies et entourage*, conduite par l'Ined en 2001, auprès de 2800 individus des générations 1930-1950 résidents en Ile de France. La spécificité de *Biographies et entourage*, conçue dans la lignée de *Proches et parents*, découle de l'effort investi pour capter, au fil de la vie de l'enquêté, non seulement l'ensemble de ses corésidents, c'est-à-dire les différents ménages auxquels il a appartenu, mais aussi l'ensemble des personnes qui lui ont été importantes,

c'est-à-dire son entourage[1] [Lelievre et Vivier, 2001]. Il est ainsi possible de reconstituer le ménage et l'entourage d'ego à chaque âge [Golaz et Lelièvre 2006]. Cela permet de définir autour des individus différents types de groupes sociaux qui ouvrent la possibilité d'une analyse biographique multiniveau prenant en charge un contexte social.
La complexité de telles analyses nécessite en particulier de définir la notion de corésidence de manière précise, en prenant en compte les étapes de la vie où l'individu « navigue » entre deux résidences. On peut ainsi citer l'étudiant qui passe la semaine dans sa chambre universitaire, et qui rentre chez ses parents le week-end, et qui de fait constitue un cas de corésidence partielle.
Une contextualisation spatiale à partir de sources secondaires sur les communes d'Ile de France (données de recensements) est également à l'étude, mais l'analyse est ici limitée aux parties d'épisodes résidentiels en Ile de France, ce qui renforce l'hétérogénéité des trajectoires, du fait de la forte attractivité de la région parisienne pour les générations étudiées, et engendre des données lacunaires pour les périodes passées hors Ile de France.

3.3. Des compromis possibles ?

Ces trois exemples d'enquêtes sont tous des exemples d'enquêtes biographiques incluses dans des projets où une certaine forme de contextualisation est envisagée dès la collecte. On peut constater à travers ces trois cas que si l'analyse multiniveau butte encore sur un certain nombre d'écueils, et ce en particulier pour les modèles prenant en compte le temps, la collecte biographique multiniveau n'est pas non plus très simple. Plus les durées en question dans la collecte comme dans l'analyse sont longues, moins les résultats sont probants. Du point de vue de la collecte, les travaux sur l'approche biographique montrent que la profondeur des trajectoires n'est pas un obstacle à leur qualité [GRAB, 1999]. En revanche, plus on remonte dans le temps, dans une enquête rétrospective, plus les éléments collectés sont hétérogènes : par exemple, seulement 40% des enquêtés de *Biographies et entourage* sont nés en Ile de France, alors que l'enquête est représentative de la population de cette région en 2001. Du point de vue de l'analyse, plus la durée d'observation est longue, plus les changements de groupes sont nombreux. Or, les modèles permettant ces changements sont très rapidement complexes, et peu adaptés à la prise en compte de plusieurs types de niveaux [Goldstein, 2000].

En fin de compte, même si à travers ces enquêtes, la collecte porte sur l'ensemble de la trajectoire des enquêtés, prise depuis la naissance, la plupart des analyses multiniveau auxquels elles ont donné lieu portent généralement sur des périodes beaucoup plus courte, soit précédent l'enquête, soit au tout début des trajectoires (petite enfance). D'autres types de collectes rétrospectives peuvent être envisagés, plus ciblées, permettant la mise en œuvre d'une approche multiniveau diachronique.

4. Apports et enjeux du multiniveau

Les modèles multiniveau constituent un pas en avant dans la modélisation de la réalité sociale. L'apport exclusif du multiniveau réside dans trois points particuliers, qui illustrent l'intérêt de mettre en œuvre une approche a priori plus lourde, en particulier en démographie. Le premier avantage de l'approche multiniveau est qu'elle peut permettre de mesurer l'importance relative de différents niveaux, et en cela, d'identifier les niveaux pertinents pour l'analyse. Le second est que la mise en œuvre de modèle multiniveau ne nécessite pas de base de données très lourde, et permet d'envisager des collectes sur de petits effectifs. Le troisième, enfin, réside dans l'analyse conjointe de données d'origines différentes. Paradoxalement, ces points forts de l'approche multiniveau sont encore rarement mis en œuvre, les analyses menées restant le plus souvent sur des chemins mieux balisés.

4.1. Mesurer la pertinence des niveaux choisis pour le phénomène étudié

Au-delà de la mesure de l'influence de certains groupes sociaux ou spatiaux sur les comportements individuels, l'approche multiniveau donne une indication de la pertinence relative de ces différents niveaux d'analyse. Sous réserve de l'existence de données appropriées, l'approche multiniveau permet ainsi d'identifier les niveaux d'analyse pertinents. Articuler différents niveaux permet d'établir quels sont les niveaux pertinents d'observation, et quelles caractéristiques à chacun de ces niveaux

[1] Selon la définition de C. Bonvalet et E. Lelièvre [1995].

ont une influence sur l'objet de l'étude. L'exemple le plus classique concerne des niveaux administratifs imbriqués les uns dans les autres qui peuvent contribuer plus ou moins au phénomène mesuré. Ainsi, certains de ces niveaux ne seront pas pertinents pour l'analyse ou le seront moins que d'autres.
Autre exemple, un niveau d'observation défini sur un critère spatial (la région, ou toute autre division administrative de la population étudiée), peut être confronté à des niveaux construits sur des critères sociaux (le groupe humain, le secteur d'activité, …). Lors de la mise en œuvre d'une analyse contextuelle, des caractéristiques du groupe social et des caractéristiques spatiales peuvent être significatives dans le même modèle, du fait d'une corrélation notable des variables en question. Un modèle multiniveau associant ces deux types de niveaux d'agrégation permet de déterminer de manière claire la part du phénomène étudié qui est liée aux caractéristiques individuelles, familiales, culturelles, sociales ou régionales.

Cependant, la complexité des modèles est limitée, beaucoup plus limitée que ne l'est la réalité sociale, rapidement plus limitée que ce que le chercheur peut concevoir [DiPrete et Forristal, 1994]. L'introduction d'aléas multiples, en particulier, intensifie les calculs sous jacents. L'estimation des paramètres devient vite problématique. Par ailleurs, la mise en œuvre de modèles conceptuels est tributaire de l'existence des données… et cela non plus n'est pas trivial. Les données existantes, même dans des pays possédant un appareil statistique ancien et dynamique, ne sont pas toujours suffisantes pour répondre aux nouvelles questions de recherche que l'analyse multiniveau permet de soulever. Il serait donc souvent nécessaire de concevoir une approche multiniveau dès la collecte pour avoir les données minimales requises pour la question étudiée.

4.2. Des contraintes faibles pour la collecte

Critère important pour la collecte, l'approche multiniveau permet de travailler sur des données plus limitées que si l'on étudiait séparément chacun des niveaux. Les calculs statistiques sont effectués sur la totalité de l'échantillon et non sur chaque groupe de données séparément, ce qui permet de travailler avec des groupes relativement peu nombreux. Ainsi dans le domaine des sciences de l'éducation, Harvey Goldstein [2003] travaille sur les résultats scolaires de 728 élèves issus de 48 écoles élémentaires. Les écoles sont utilisées comme niveau supérieur d'observation. Certaines ne comportent qu'une dizaine d'élèves, ce qui n'empêche en aucune manière les modèles qu'il utilise de converger. En démographie, cela permet d'envisager un travail au niveau de groupes sociaux de taille relativement petite, comme le ménage, la fratrie ou la famille proche. Cela permet aussi d'envisager de nouvelles collectes de taille raisonnable qui permettent de tester des hypothèses multiniveau.

Cependant, à l'heure actuelle en démographie, les opérations de collecte qui se développent et qui sont vouées à amener le plus grand nombre d'analyses sont des grandes enquêtes nationales standardisées. Dans les pays du Sud, c'est le cas des Enquêtes démographiques et de Santé depuis deux décennies, en Europe, de projets comme *Gender and Generation Survey*, et dans le monde en général, la mise à disposition de données de recensement, avec un effort important d'harmonisation des variables dans l'espace et dans le temps[2], visent à l'analyse comparative de région et de pays différents, et appellent naturellement une approche multiniveau construite sur un critère territorial. Dans les pays nordiques, les registres de population, en donnant des informations localisées, ont déjà fait l'objet d'une telle approche.

Ainsi, alors que la modélisation multiniveau permet d'envisager des collectes reposant, dans chacun des groupes considérés, sur des petits effectifs, ce sont les collectes de grande ampleur qui se développent dans la perspective de mobiliser cette approche. L'hétérogénéité des population étudiée sen est l'une des causes, la profondeur des trajectoires recueillies introduit également un autre source de divergence.

4.3. L'articulation de sources différentes

L'analyse multiniveau permet la prise en compte simultanée de données par nature différentes. C'est là une possibilité de valorisation nouvelle, de sources de données variées. Toutes les informations souhaitées pour répondre à une question de recherche ne sont pas toujours disponibles, mais l'analyse multiniveau permet, du point de vue des données quantitatives, d'articuler une partie de

[2] Il faut noter l'initiative dans ce sens d'IPUMS-International (https://international.ipums.org/international/).

celles qui existent. Cela permet de mobiliser au niveau macro des sources agrégées habituellement peu valorisées (données sectorielles, …) et au niveau micro des sources individuelles ou familiales plus détaillées, pas forcément exhaustives (micro données de recensements et d'enquêtes). C'est un enjeu particulièrement important dans les pays où les micro-données statistiques sont rares ou difficiles d'accès.

Cette articulation de sources différentes est non seulement une opportunité mais aussi une nécessité. C'est ce qui fait la spécificité de l'approche multiniveau, son intérêt particulier. Cependant, il est parfois difficile d'avoir accès à ces autres sources de données. La recherche d'un minimum d'information sur les groupes utiles à l'analyse n'est pas toujours simple. Même dans le cas d'unités administratives, la recherche des caractéristiques de ces unités dans le passé est complexe. Ainsi, collecter les caractéristiques contextuelles nécessaires à l'analyse d'une enquête nécessite souvent un travail de recherche à proprement parler. Les indicateurs souhaités ne sont pas toujours disponibles aux niveaux pertinents pour l'analyse, ou bien, lorsqu'ils émanent de l'administration locale, peuvent être trop hétérogènes. Ces problèmes interfèrent avec la mise en œuvre balisée de modèles multiniveau.

Ainsi, les éléments qui font la force de l'analyse multiniveau en sont aussi les points d'achoppement. C'est le cas en analyse transversale, et plus encore lorsque l'on se place dans une perspective longitudinale.

La mise en œuvre de modèles statistiques prenant en compte différents niveaux d'observation permet d'intégrer l'analyse quantitative des comportements humains dans un paradigme large, dans lequel les inflexions et les continuités des trajectoires individuelles sont expliquées à l'aune de la vie passée de chacun, ainsi que de son entourage en perpétuelle redéfinition, des milieux sociaux et des territoires dans lesquels il a vécu. Selon le sujet traité, le niveau inférieur d'observation s'impose de lui-même, et il est désormais possible de prendre en compte de manière quantitative les niveaux supérieurs. Ainsi, la technicité de plus en plus grande des modèles utilisés en démographie non seulement permet mais aussi nécessite un lien fort avec les autres disciplines des sciences sociales afin d'expliquer au mieux les comportements humains. Cependant, comme toujours, la mise en œuvre de ce type d'analyse dépend de la disponibilité de données pertinentes et de la difficulté que pose la quantification de contextes parfois flous.

La démographie associe différents niveaux d'observation, des populations aux individus qui les forment, et de ces individus aux événements qui jalonnent leur vie, de la naissance jusqu'au décès. La description et l'analyse de ces événements individuels, facteurs directs ou indirects de l'évolution de la structure des populations, fait intervenir simultanément le niveau des individus et celui de la population. Les méthodes d'analyse simultanée à différents niveaux des données quantitatives se sont développées récemment. Cette innovation méthodologique, allant de pair avec la généralisation d'approches articulant quantitatif et qualitatif, devrait contribuer à une prise en compte de réalités de plus en plus complexes par le démographe.

Remerciements. Nous remercions Daniel Courgeau pour sa relecture et ses conseils sur une version antérieure de ce texte.

Bibliographie

[1] Axinn William G.; Barber, Jennifer S.; Ghimire, Dirgha J., "The neighborhood history calendar: A data collection method designed for dynamic multilevel modeling." *Sociological Methodology*. 1997, 27, 355 - 392.
[2] Bonvalet C. et Lelièvre E., « Du concept de ménage à celui d'entourage : une redéfinition de l'espace familial », *Sociologie et sociétés*, Vol.XXVII, N°2, p.177-190, 1995
[3] Bourdelais Patrick « L'enquête des 3000 familles : un premier bilan », *Annales de démographie historique*. 2004-1 , pp.XXXXX.
[4] Bressoux P., Modélisation statistique appliquée aux sciences sociales, DeBoeck, Bruxelles, 464p, 2008.
[5] Bringé A. et Laurent R., 2005, *Reconstituer des histoires individuelles à partir de données de suivi démographique*. Série Les Clefs pour…, Paris, CEPED, 85p.

[6] Chaix B, Chauvin P. « L'apport des modèles multiniveaux dans l'analyse contextuelle en épidémiologie sociale : une revue de la littérature ». Rev Epidemiol Sante Publique 2002 ; 50 : 489-499.
[7] Courgeau D., « Evolution ou révolution dans la pensée démographique ? », *Mathématiques et sciences humaines*, n° 160, p. 49-76,2002.
[8] Courgeau D., Du groupe à l'individu. Synthèse multiniveau. Paris. INED.2004
[9] DiPrete T.A. et Forristal J.D., Multilevel Models: Methods and substance, *Annual Review of Sociology*, 20, 1994.
[10] Erola Jani, Härkönen Juho et Jäntti Markus, 2008, Trends in brother correlations in class and incomes in finland: a comparison of cohorts born in 1932-1962, papier présenté aux rencontres de l'ISA, 15-18 mai 2008, Florence.
[11] Goldstein H., Multilevel Mixed Linear Model Analysis Using Iterative Generalised Least Squares. *Biometrika*, 73 (1), p.43-56, 1986
[12] Goldstein H.; Rasbash J.; Browne W.; Woodhouse G.; Poulain M., « Multilevel Models in the Study of Dynamic Household Structures», *European Journal of Population/ Revue européenne de Démographie*, Volume 16, n°4, pp. 373-387, December 2000
[13] Goldstein H. Multilevel statistical models. 3rd ed. London: Arnold, 2003
[14] Groupe de réflexion sur l'approche biographique, 2006, *États flous et trajectoires complexes : observation, modélisation, interprétation*, Groupe de réflexion sur l'approche biographique, E. Lelièvre et P. Antoine (Eds)., Ined / Ceped: Paris. 2006
[15] Hoem J.M., La démographie, aujourd'hui et demain, Population, 62(1), p.53-56, 2007.
[16] Hox J. Multilevel analysis: techniques and applications. Mahwah, NJ: Lawrence Erlbaum Associates, 2002
[17] Lelièvre E, Bonvalet C, Bry X, Analyse biographique des groupes : les avancées d'une recherche en cours, *Population*, 52 (4), p. 803-830, juillet-août 1997
[18] Lelièvre E, Vivier G., 2001, "Évaluation d'une collecte à la croisée du quantitatif et du qualitatif, l'enquête " Biographies et entourage "", Population, n°6, vol 56, pp.1043-1074.
[19] Parr Nick, « Applications of Multilevel Models in Demography: What have we learned? », Macquarie Business Research Papers No. 10, Sydney, Australia, Macquarie University, Department of Business, 26p. Dec.1999.
[20] Raudenbush S, Bryk A. Hierarchical linear models: applications and data analysis methods. Thousand Oaks: Sage, 2002.
[21] Schoumaker B., Dabire H.B., Gnoumou-Thiombiano B., « Collecter des biographies contextuelles pour étudier les déterminants des comportements démographiques. L'expérience d'une enquête au Burkina Faso», Population, vol. 61, n° 1-2, p. 77-106, janvier-avril 2006.
[22] Singer, J.D. et Willett. J., Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence: Oxford University Press. 2003
[23 Snijders T, Bosker R. Multilevel analysis: an introduction to basic and advanced multilevel modeling. London: Sage Publications, 1999.
[24] Tabutin D., « Vers quelle(s) démographie(s) ? Atouts, faiblesses et évolutions de la discipline depuis 50 ans », *Population*, 62(1), p.15-32, 2007.