L'apport de l'analyse biographique en démo-économie

Daniel COURGEAU et Eva LELIÈVRE
I.N.E.D.
27 rue du Commandeur - 75675 Paris Cedex 14
Tél. : (1) 42 18 21 07 - Fax : (1) 42 18 21 95

Il convient en tout premier lieu de faire une mise au point du vocabulaire utilisé dans les diverses disciplines pour nommer ces méthodes d'analyses et les mesures utilisées. En effet, en français du moins, les diverses sciences humaines utilisent des termes différents pour désigner la même démarche, ce qui contribue à masquer le fait que de l'économétrie à la démographie en passant par les biostatistiques l'utilisation de ces méthodes s'est rapidement développée depuis la fin des années 1980. Quelques indications lexicographiques ne sont donc pas superflues, même si elles ne prétendent pas être exhaustives, elles ont simplement pour but de rappeler une partie du vocabulaire utilisé, afin de dissiper les ambiguïtés ou les doutes.

En premier lieu ce que nous avons nommé, analyse démographique des biographies ou analyse biographique (Courgeau et Lelièvre, 1989), est également désignée sous les noms d'analyse de survie, d'analyse de durée de séjour, d'analyse de fiabilité, d'analyse d'histoires de vie. Néanmoins l'emploi qui en est fait en démographie permet de dépasser une simple analyse longitudinale classique qui constitue elle même une avancée par rapport à l'analyse transversale (Courgeau et Lelièvre, 1990).

De l'analyse transversale à l'analyse biographique ...

L'analyse transversale consiste en l'étude de "phénomènes" caractérisés par des événements que l'on a observés à un moment donné, à une certaine "période". Les "comportements" correspondent alors au calendrier et à la structure de chaque phénomène supposé agir de façon indépendante des autres. Ce cadre de l'analyse transversale s'est mis en place à la fin du XVII^{e} siècle et au cours du XVIII^{e} siècle (Table de Halley en 1693, Table de Wargentin en 1776) et a été utilisé presque exclusivement jusqu'à la fin de la seconde guerre mondiale. En analyse transversale, il est aisé de faire intervenir des caractéristiques en vue de tester leur effet sur la survenue du phénomène étudié. Néanmoins c'est en essayant de construire des indices synthétiques que l'analyse transversale rencontre des difficultés. Supposons par exemple que l'on désire analyser les variations de l'activité féminine selon l'âge. Il est toujours possible de combiner les différents indicateurs par âge, mais on voit rapidement que ce calcul ne fait pas référence au suivi d'une génération réelle, car elle mesure l'effet des conditions du marché de l'emploi et des contraintes familiales sur une *génération fictive*. Si l'on étudie la nuptialité, c'est un phénomène aux fluctuations importantes qui présente des périodes d'ajournement suivies de récupération à la suite des périodes de crises. Ainsi à la fin de la seconde guerre mondiale, la somme des quotients de primo-nuptialité par âge dépasse l'unité, alors que cet indicateur devrait toujours rester inférieur à un dans le cas où l'on suit une génération puisque une part plus ou moins grande des individus reste définitivement célibataire.

Supposons maintenant que la population observée ne le soit plus en un instant mais consiste en une génération née une année donnée ou ayant connu un même événement-origine. L'analyse longitudinale consiste alors à étudier la survenue au ***cours du temps*** d'un événement (mariage, migration, changement d'emploi, etc.) au sein d'un groupe initialement bien défini ou cohorte. Le temps écoulé depuis l'entrée dans la cohorte définit la ***durée*** que l'on va analyser, en essayant d'éliminer l'effet de phénomènes perturbateurs. On aboutit ainsi à l'étude des divers phénomènes à ***l'état pur***, qui apparaissent comme des chapitres séparés dans les manuels de démographie classique (Henry, 1972 ; Pressat, 1961). Cette analyse longitudinale n'est pas sans poser un certain nombre de problèmes que l'analyse biographique a en partie au moins résolus. Ainsi, lorsque l'on analyse un phénomène, le fait d'éliminer l'effet des autres phénomènes perturbateurs en faisant l'hypothèse -visiblement erronée- de leur indépendance, empêche toute analyse sérieuse d'interaction entre eux (Courgeau et Lelièvre, 1994).

L'analyse biographique, qui prolonge l'analyse longitudinale classique, présente par rapport à cette dernière des apports essentiels à la compréhension des phénomènes humains dans toute leur complexité [Courgeau et Lelièvre, 1989].

Le premier problème est lié à la formalisation statistique insuffisante de l'analyse longitudinale. En effet, celle-ci en travaillant sur des cohortes de taille importante n'avait pas à se préoccuper de calculer des ***variances*** pour les quotients estimés. Néanmoins dès que l'on travaille sur des populations de plus faible taille, sous-populations spécifiques ou échantillons d'enquêtes rétrospectives ou prospectives, il devient nécessaire d'estimer ces variances pour pouvoir comparer des populations différentes et pour avoir une idée de la précision des estimations ainsi faites. L'analyse biographique en replaçant ces calculs dans un cadre statistique strict et qui est beaucoup plus complexe que l'on peut le penser *a priori* [Andersen et al., 1993], permet de résoudre ce problème.

Le second problème vient de ***l'hétérogénéité*** des cohortes sur lesquelles on travaille. Il faut bien voir que l'une des hypothèses de base de l'analyse longitudinale classique, est que l'on observe une cohorte homogène. Cette hypothèse n'étant généralement pas vérifiée, l'analyse classique va devoir désagréger la population initiale en sous-populations plus homogènes, selon divers critères qui peuvent parfois se référer à l'avenir. L'analyse biographique offre une possibilité de prendre en compte cette hétérogénéité entre les individus du groupe étudié, bien entendu sous certaines hypothèses. Elle permet en plus de faire dépendre du temps cette hétérogénéité qui peut changer au cours de la trajectoire des individus. Les comportements sont décrits et expliqués grâce aux éléments que l'on connaît de leur passé sans référence à leur avenir.

Un troisième problème vient des ***interactions*** qui existent entre les divers phénomènes étudiés. L'hypothèse faite lors d'une analyse longitudinale classique, d'une indépendance entre phénomènes perturbateurs et phénomènes étudiés, doit être dépassée. L'analyse biographique offre une possibilité de prendre en compte cette perturbation et d'étudier comment la survenue d'un événement (la migration vers les métropoles par exemple) peut modifier la probabilité de survenue d'un autre événement (la nuptialité et la fécondité dans l'exemple cité) [Courgeau, 1987]. En incorporant la dimension temporelle des divers événements les uns par rapport aux autres, cette analyse respecte aussi leur ordre dans le temps ; notons cependant que dans certains cas cet ordre n'est pas celui des prises de décision.

La dimension fondamentale d'une analyse biographique, appelée aussi analyse de durée, est donc le ***temps*** qui marque la définition de la cohorte étudiée et la survenue des divers événements les uns par rapport aux autres. Cette dimension introduit des problèmes statistiques particuliers et très complexes qui n'ont été résolus, en partie que récemment grâce aux théories des martingales en temps continu, à l'intégration stochastique et aux théories des processus de comptage.

La particularité des données de durée de séjour est qu'elles peuvent s'interpréter comme résultant d'un processus stochastique (probabiliste) sous-jacent, processus dont témoignent les dates de changement d'état. Le cadre théorique des méthodes présentées ici est donc le suivant :

1. un processus stochastique dans un espace fini ;
2. un paramètre ***temps*** qui correspond à la durée écoulée depuis un événement de référence ;
3. les états de l'espace fini qui sont identifiés par le statut dans lequel se trouve l'individu ;
4. les transitions, caractérisées par les passages d'un état à un autre.

Les phénomènes que l'on étudie sont ainsi caractérisés par l'intensité des passages d'un état à un (aux) autre(s). Dans la situation la plus simple on a seulement deux états : ("vivant" / "mort" par exemple), mais la plupart des étapes du parcours des individus sont des états transients avec un passage toujours possible vers l'état final (la disparition). Le modèle peut donc se compliquer à loisir. Malgré les hypothèses simplificatrices que l'on va être amené à poser dans les situations les plus complexes, cette approche permet une analyse non déterministe de phénomènes trop souvent envisagés en termes de causalité.

Une telle analyse démographique des biographies a pris forme au début des années 80 avec les premiers articles de Menken et al. (1981) et de Trussel et Hammerslough (1983) dont la publication fut simultanée à la réalisation en 1981 par l'INED de l'enquête "Triple biographie: Familiale, Professionnelle

et Migratoire" (appelée encore "3B") et à ses premiers résultats utilisant les méthodes d'analyse des biographies présentés à la Chaire Quetelet de 1983 (Courgeau, 1985). Devant l'importance prise par ces méthodes en démographie, nous avons réalisé un stage de formation en 1987 et rédigé un manuel (Courgeau et Lelièvre, 1989). De façon parallèle, les premiers articles d'économétrie sur les transitions entre états (emploi, chômage et inactivité, par exemple) ont été publiés au début des années 80 (Flinn & Heckman, 1982a et 1982b). Leur développement a conduit à la rédaction d'un manuel par Lancaster (1990) qui applique ces méthodes aux problèmes rencontrés par les économistes.

Dans cette présente communication nous allons aborder une présentation plus détaillée des méthodes utilisées et des diverses hypothèses qui sont nécessaires pour les appliquer à des données démo-économiques. Nous examinerons ensuite les difficultés d'obtention de sources de bonne qualité et donc la nécessité de mettre en place des analyses susceptibles d'utiliser des sources imparfaites. Nous verrons enfin les nouvelles avancées dans le domaine : soit pour analyser des unités plus complexes (couples, familles etc.), soit pour travailler simultanément à des niveaux d'agrégation différents.

I - L'ANALYSE BIOGRAPHIQUE EN DEMO-ECONOMIE

Devant l'unification des méthodes d'analyse, il devient possible d'aborder les liens entre les différentes sciences sociales, en particulier entre l'économie et la démographie. Si l'on dispose d'enquêtes suffisamment riches, recueillant à la fois les phénomènes démographiques (fécondité, nuptialité, migrations, etc.) et les phénomènes économiques (mobilité entre activités, chômage et inactivité, mobilité professionnelle, etc.), on peut mettre en évidence les interactions qui existent entre ces phénomènes tout en tenant compte de l'hétérogénéité des populations soumises aux risques (niveau d'éducation, profession des parents, scolaire etc.).

1) Analyse des interactions entre phénomènes

Nous allons montrer ici en nous appuyant sur des analyses de ces interactions (Courgeau 1993 ; 1995), les différents problèmes rencontrés et les solutions proposées.

Prenons le cas par exemple des liens entre la mobilité géographique vers les métropoles et la mobilité des individus entre diverses catégories socioprofessionnelles. Ces phénomènes sont relativement complexes à étudier car ces mobilités peuvent intervenir plusieurs fois dans la vie d'un individu (phénomènes récurrents), et la mobilité professionnelle peut théoriquement se produire d'un état donné vers tous les autres (risques compétitifs). Pour les générations de l'enquête "3B" (nées entre 1911 et 1936), on a pu considérer que la migration vers les métropoles était souvent définitive, du moins jusqu'à l'âge de la retraite. Ceci simplifiait un peu le problème en permettant de ne considérer qu'un seul phénomène récurrent (la mobilité professionnelle), mais pour les générations plus récentes, il devient nécessaire de considérer également la mobilité vers les métropoles comme un phénomène récurrent. C'est ce que nous faisons ici pour traiter le problème dans toute sa généralité.

Pour chaque individu, nous disposons des dates des changements socioprofessionnels successifs, représentées par les variables aléatoires $T_0 \leq T_1 \leq T_2 \ldots$ et les états professionnels successifs que nous caractérisons par une série de variables aléatoires $\{S_k; k = 0,1,2\ldots\}$, $S_k \in \{0,1,\ldots m$ où m est le nombre d'états possibles. Simultanément pour les individus ayant débuté leur vie professionnelle hors des métropoles, nous observons les dates successives de migration entre les aires non-métropolitaines et métropolitaine, représentées par les variables aléatoires $T^0 \leq T^1 \leq T^2 \ldots$ et les lieux de résidence successifs $\{S^l; l = 0,1,2\ldots\}$, $S^l \in \{0,1$ où 0 représente le non-métropolitain et le 1 le métropolitain. Notons ici que $T^0 = T_0$ puisqu'il s'agit de la date du début d'observation des individus (entrée sur le marché du travail) et que les séries complètes de dates ne seront pas observées du fait que l'on travaille sur les données d'une enquête rétrospective. Nous disposons également d'un certain nombre de

caractéristiques des individus (origine sociale, niveau d'éducation, nombre de frères et soeurs, etc.) z_k dont certaines pourront changer au cours de l'observation.

On voit ainsi que l'on dispose de deux chaînes d'événements professionnels et migratoires entre lesquelles existent des dépendances que nous devons chiffrer. Il nous faudra cependant résoudre un certain nombre de problèmes avant de pouvoir le faire.

Un premier problème vient de la prise en compte du temps. En effet, il peut paraître plus simple de considérer un temps unique qui se déroule à partir à partir de la date d'entrée dans le monde du travail, sans rupture lorsqu'un événement professionnel ou migratoire se produit. Une telle prise en compte du temps paraît peu satisfaisante dans la mesure où un changement professionnel ou une migration, peuvent être considérés comme suffisamment importants pour changer le cours de la vie d'un individu. Il paraît donc préférable de remettre l'horloge à zéro lorsqu'un événement professionnel se produit, dans l'étude des changements professionnels, ou lorsqu'une migration se produit dans l'étude des mouvements entre métropoles et hors métropoles. Aussi pour un ouvrier qui passe à son compte, la durée dans la nouvelle profession marque mieux la réussite de cette "promotion", que la durée écoulée depuis l'entrée dans la vie active. Un retour rapide au statut d'ouvrier marque plus clairement l'échec de cette promotion, quelle que soit la durée de l'activité antérieure à celle-ci. Pour aller plus loin, il est possible de tenir compte à la fois du temps écoulé depuis l'entrée dans le monde du travail et du temps écoulé depuis le dernier changement en travaillant sur un temps multidimensionnel (Anderson, et al., 1993).

C'est là une nouvelle voie de recherche de grand intérêt, qui pose cependant des problèmes théoriques complexes à résoudre, avant de pouvoir être utilisée en sciences humaines ; nous choisirons donc ici la seconde solution qui consiste à remettre l'horloge à zéro chaque fois qu'un événement se produit.

Un second problème vient des simultanéités entre événements, qui peuvent être fréquents si la liaison entre eux est forte. Ainsi on peut penser qu'une migration vers les métropoles est une occasion pour un individu de connaître une "promotion" : il peut d'ailleurs l'effectuer dans ce but si cela est le cas, il nous paraît préférable de décaler arbitrairement ces événements dans le temps : pour étudier les migrations qui entraînent une "promotion", il est utile de considérer ces migrants "simultanés" dans la population soumise au risque de "promotion". D'autres options sont cependant possibles et il est du plus grand intérêt de les explorer, pour voir comment une prise en compte différente de ces "simultanéités" influe sur les quotients estimés. Cela revient à l'analyse des effets qu'entraînent sa prise en compte. Pour aller plus en avant, il serait nécessaire de travailler avec des psychologues qui exploreraient le mécanisme des prises de décision intervenant antérieurement à l'événement que l'on mesure.

La formulation plus précise du modèle peut dès lors intervenir. Il nous faut pour ce faire, considérer distinctement les deux phénomènes étudiés tout en introduisant dans les quotients correspondant à l'un, l'effet de l'autre en interaction. Considérons par exemple que le quotient instantané de passage de l'état professionnel S_{k-1} à l'état S_k selon le lieu de résidence de l'individu à l'instant considéré, t. Si l'individu réside hors zone métropolitaine ($T^l \le t < T^{l+1}$ avec l pair égal à $2n$), on peut écrire ce quotient :

$$h_{k-1,k}^{l=2n}(t;z_k) = \lim_{dt \to 0} \frac{P(T_k - T_{k-1} < t+dt\ ,\ S_k = s_k | T_k - T_{k-1} \ge t\ ,\ S_{k-1} = s_{k-1}\ ,\ z_k\ ,\ T^l \le t < T^{l+1},\ l = 2n)}{dt}$$

et si au contraire l'individu réside dans une métropole (l impair égal à $2n + 1$)

$$h_{k-1,k}^{l=2n+1}(t;z_k) = \lim_{dt\to 0} \frac{P(T_k - T_{k-1} < t+dt\ ,\ S_k = s_k \big| T_k - T_{k-1} \geq t\ ,\ S_{k-1} = s_{k-1}\ ,\ z_k\ ,\ T^l \leq t < T^{l+1},\ l = 2n+1)}{dt}$$

On peut de façon symétrique calculer des quotients de migration de rang l selon l'état professionnel dans lequel se trouve l'individu au cours de la période séparant la migration de rang *l* de celle de rang *l* - *1*. On voit que les diverses populations soumises au risque ne sont pas uniformément décroissantes, puisqu'elles peuvent augmenter, dans le premier cas par exemple, par l'apports d'individus revenus en zone non-métropolitaine. Les logiciels habituellement utilisés (TDA, STATA, SAS, etc.) ne permettent pas de traiter de tels cas et il est nécessaire de mettre en place de nouveaux programmes pour ce faire.

De plus, pour l'estimation de ces modèles, des spécifications supplémentaires sont nécessaires. Plutôt que de donner une forme paramétrique à l'effet de la durée de séjour (modèle de Gompertz, modèle log-normal, etc.), nous préférons utiliser un modèle de Cox (1972) pour lequel il existe un quotient de mobilité non-paramétrique sous-jacent, qui est le même pour tous les individus d'une catégorie de départ, et pour lequel les diverses caractéristiques, qu'elles dépendent ou non du temps, jouent de façon multiplicative sur ce quotient. Ceci évite une paramétrisation incorrecte de cette durée et surtout permet de mieux contrôler l'effet de l'hétérogénéité non observée. En effet, il a été possible d'estimer de façon théorique comment l'omission de caractéristiques, indépendantes de celles que l'on fait intervenir, affectait les paramètres estimés dans un tel modèle (Bretagnole et Huber-Carol, 1988). Ils ont montré que cette omission n'affectait pas le signe des paramètres estimés, mais qu'elle entraînait une réduction de leur valeur absolue. Il en résulte que si l'effet d'un caractéristique était apparu comme significatif lorsque l'on en omettait d'autres indépendantes d'elle-même, leur introduction dans le modèle de Cox ne fera que renforcer l'effet de la première caractéristique. En revanche, certaines caractéristiques qui semblaient n'avoir aucun effet significatif, peuvent devenir tout à fait significatives lorsque l'on introduit des caractéristiques initialement non observées. Ces résultats sont très importants en ce qu'ils assurent le sens des effets observés.

2) Existence de situations plus complexes que celles d'interactions entre phénomènes

La mise en oeuvre d'une analyse **d'interaction** nécessite l'identification de phénomènes distincts dont on teste la concurrence à des moments clefs de la trajectoire individuelle. Cette démarche a l'avantage d'offrir une structure à une interaction particulière.

Cependant si le choix de phénomènes distincts s'impose facilement dans les exemples cités ci-dessus, d'autres interactions semblent plus difficiles à caractériser en terme d'enjeux de deux (ou plusieurs) processus distincts. Dans le cas où l'apparition d'un phénomène détermine une modification radicale des conditions d'évolution d'un second processus qui lui-même affecte en retour le premier processus, les décisions résultantes dans l'un ou l'autre des deux domaines peuvent apparaître plus complexes, et leur mécanisme difficile à évaluer.

L'interaction entre la fécondité et l'activité féminine fournit l'exemple d'une telle circularité, dont les nombreuses études (voir Bernhart, 1988 pour un bilan) n'ont pas réussi à déterminer d'axe dominant. Le débat idéologique s'est alors souvent substitué à la recherche empirique ou théorique dans ce domaine, du fait de la complexité de l'interaction et en l'absence d'une approche conceptuelle ou méthodologique capable de proposer son analyse dans un cadre formel scientifiquement satisfaisant. Une analyse d'interaction ne réussit qu'à identifier une forte influence réciproque d'un événement sur l'autre sans que l'on soit à même de déterminer un axe de dépendance dominant. On se trouve alors face à une dépendance circulaire que l'analyse ne permet pas de dépasser.

Il faut néanmoins noter qu'avant la mise en oeuvre du concept d'interaction, ce genre de problème ne pouvait même pas être identifié dans la mesure où chaque phénomène constituait un objet d'étude séparé et que l'influence des autres processus était mesurée en l'absence d'évaluation de son

influence propre sur ces autres processus. Nous sommes bien ici dans un cas limite d'application de l'analyse des interactions entre phénomènes.

Ainsi une réflexion critique sur la nature et les implications de ces limites (Kempeneers et Lelièvre, 1990) a révélé que l'impasse était plutôt de nature conceptuelle et que le problème nécessitait d'être posé en des termes dépassant la simple interaction entre phénomènes. L'existence d'une situation plus complexe a amené à préconiser la saisie de l'interaction à sa source : en amont de l'analyse proprement dite. Ainsi l'amélioration proposée n'est pas du ressort d'une modélisation plus savante des phénomènes en jeu mais plutôt d'une amélioration supplémentaire des données. La collecte du ***travail*** féminin, recueillerait alors simultanément l'interaction entre la sphère domestique, dont dépend la fécondité mais aussi la présence ou l'absence d'un conjoint, et la sphère rémunérée caractérisée non seulement par les modalités de l'activité (salaire, horaires...) mais aussi par les contraintes du marché de l'emploi.

Une analyse biographique appliquée ensuite à ces données complexes sur le travail féminin permettrait d'appréhender les arbitrages dans leur intégralité.

II - VERS DE NOUVELLES SOURCES D'OBSERVATION MOINS COMPLETES

L'analyse préconisée dans la partie précédente nécessite des données très complètes sur l'histoire de vie d'un échantillon de taille importante. Nous avons pu utiliser les données de l'enquête "3B" pour la réaliser, mais la faible taille de l'échantillon observé (4 602 personnes nées entre 1911 et 1936) ne nous a pas permis de le faire dans toute sa complexité. Nous avons en particulier dû regrouper les divers séjours sans tenir compte du rang de l'événement considéré. Nous travaillons également pour ce faire sur une enquête rétrospective, où des problèmes de mémoire peuvent perturber les résultats car certains événements se sont produit plus de 50 ans avant l'enquête. Conscients de ces problèmes, nous avons essayé de voir dans quelle mesure ils affectent les résultats d'analyses biographiques. Nous avons donc réalisé un premier test dès 1983 sur un faible échantillon de 50 personnes (Duchêne, 1985 ; Courgeau, 1985) et en 1990, nous avons pu interroger un échantillon plus important de 500 personnes (Poulain et al., 1991 ; Courgeau, 1991), en Belgique, pays où existe un registre de population. En effet, ce registre qui enregistre au jour le jour les événements familiaux (mariages, naissances d'enfants), et les événements migratoires permet parfaitement de réaliser ce test. Si les erreurs de mémoire sont loin d'être négligeables, en particulier pour les migrations, les résultats d'analyses biographiques effectuées avec des données remémorées par les personnes enquêtées ne sont guère altérés par cette imperfection. Il semble que les erreurs portent sur la datation exacte des événements, mais ne modifient pas l'ordre logique des divers événements qui est correctement remémoré. La mémoire semble donc fiable là où l'analyse l'exige.

Devant la lourdeur, les risques d'erreur et le coût d'une telle observation détaillé de la vie passée des enquêtés, on peut chercher à utiliser des observations moins complètes qui fournissent cependant un fichier biographique. Nous avons ainsi travaillé sur les enquêtes sur l'emploi qui interrogent les habitants d'un même logement pendant trois années consécutives (Courgeau, 1995). De tels fichiers peuvent être utilisés pour étudier des phénomènes assez fréquents pour que l'on observe suffisamment d'événements en trois ans en vue d'une analyse sérieuse. C'est ainsi que nous avons étudié les durées de chômage, en prenant cependant un certain nombre de précautions. Les individus étaient interrogés lors de chaque passage annuel sur leur situation vis à vis de l'emploi chaque mois de l'année précédente. Dans ce cas, il est préférable de prendre comme événement de départ le premier changement de situation observé : en effet la datation de l'événement antérieur à celui qui pouvait avoir lieu de nombreuses années auparavant, était mal saisie par les questions posées. Il est dès lors possible d'étudier à partir de ce point de départ les changements successifs dans l'emploi des enquêtés et de les relier à divers événements démographiques enregistrés : migrations, naissances d'un enfant, etc.. Notons

cependant que, si l'enregistrement des changements de situation d'emploi était mensuel, celui des autres événements démographiques est annuel, introduisant un flou dans leur datation.

Ce flou sera encore plus important si l'on utilise une autre source, par ailleurs de très grand intérêt : l'échantillon démographique permanent de l'INSEE. Ce fichier est en effet exceptionnel tant par sa taille (un peu plus élevée que celle d'un échantillon au 1/100 de la population présente en France de 1968 à 1982, malheureusement réduit au 1/200 pour le recensement de 1990) que par les informations qu'il comporte. Elles correspondent au couplage de données issues des bulletins individuels des recensements et des bulletins statistiques de l'état civil depuis 1968. Si ce fichier permet de suivre les individus restés en France sur 22 ans, il ne fournit pas toute l'information sur leurs changements de résidence et la profession. Il ne saisit en fait que la profession et la résidence des individus qu'à certaines dates, celles des recensements et des événements familiaux. Il ne permet pas de connaître la date exacte d'occurrence des migrations et des changements professionnels mais seulement de la situer entre divers recensements ou événements familiaux. Les méthodes habituelles d'analyse de biographies ne sont pas directement applicables à de telles biographies fragmentaires.

En dépit de cela, nous avons pu montrer (Courgeau et Najim, 1995) que l'on pouvait analyser cette source par des méthodes biographiques sous certaines conditions. Deux hypothèses fondamentales doivent être vérifiées pour que cela soit possible. En premier lieu, il est nécessaire que les intervalles ne contiennent pas plus d'un événement étudié, sinon l'un des deux échappera à l'analyse : il sera impossible à détecter. Ainsi si un changement de département suivi d'un retour dans le département de résidence initial se produit entre deux observations, cette migration échappera complètement à l'observation. On voit que pour éviter au maximum cet inconvénient, il faut que la densité dans le temps des événements qui permettent de localiser l'individu soit suffisamment importante pour ne laisser échapper qu'un petit nombre de migrations ou de changements professionnels. L'utilisation des positions de l'individu, à la fois aux recensements et aux événements familiaux, devait donner la meilleure estimation de la mobilité. Il faut cependant voir qu'une deuxième hypothèse à faire peut venir contrecarrer cette première constatation. Il est en effet également nécessaire que les événements qui permettent de localiser l'individu dans l'espace physique ou social soient indépendants de la mobilité géographique et professionnelle que l'on cherche à mesurer. Sinon, les interactions entre les divers phénomènes viendraient troubler cette estimation et conduiraient à des résultats incorrects. Nous avons pu montrer ce résultat de façon théorique en utilisant une formulation probabiliste des interactions entre dates de mobilité, de recensements et d'événements familiaux (Courgeau et Najim, 1995) et avons pu vérifier certaines dépendances entre changement de logements et événements familiaux en tronquant artificiellement les données de l'enquête "3B". Il est évident que les dates des recensements vérifient bien cette condition d'indépendance, mais que les événements familiaux peuvent entraîner des migrations, en général à courte distance, qui viendront troubler les résultats obtenus. Dans la mesure où ces deux hypothèses ne sont pas parfaitement vérifiées, il va être indispensable d'explorer à l'avenir les biais auxquels elles peuvent conduire et voir s'il y a une possibilité de les corriger.

Si nous supposons en revanche que ces deux hypothèses sont vérifiées, il est possible d'estimer de façon correcte les durées de séjour de façon non paramétrique ou semi-paramétrique à l'aide de logiciels que nous avons mis au point. Les estimations en particulier semi-paramétriques sont très lourdes en calcul, car elles utilisent des procédures d'ajustement interactif complexes. Nous avons pu en revanche vérifier que l'écart type des paramètres ainsi estimés étaient très peu affecté par cette observation des données de l'enquête "3B" rendue artificiellement fragmentaire. Il reste maintenant à utiliser ces méthodes sur les données de l'échantillon démographique permanent : de la qualité des informations portées dans ce fichier, qu'il est indispensable au préalable de tester, de la densité dans le temps des points d'observation et de l'indépendance relative entre événements marqueurs et événements étudiés dépendra la précision des estimations ainsi obtenues.

III- DE NOUVELLES AVANCEES

1) Passage des données individuelles à des données de groupes plus complexes

Le passage de l'individu à un groupe plus complexe était présent à notre esprit dès nos premières réflexions conceptuelles. Dans une perspective biographique l'étude d'un groupe plus complexe (ménage ou famille) se propose de révéler la logique d'influence que deux strates exercent l'une sur l'autre. Elle veut mettre en évidence le pouvoir du groupe sur le devenir d'un individu ; et réciproquement comment l'acteur individuel peut influencer une action collective. Cette démarche se place donc dans une optique différente de la description des différents types de familles et de l'étude de l'évolution de leur répartition. En revanche elle partage ses préoccupations avec l'étude des conséquences démographiques de l'évolution de la structure des ménages et de la famille dans ses deux aspects : les changements de l'environnement familial des individus au cours du cycle de vie et réciproquement, les effets des caractéristiques du ménage et de la famille sur les processus démographiques individuels (Bongaarts, 1983 ; Courgeau et Lelièvre, 1993).

L'enquête 3B qui saisit à chaque moment la présence d'un conjoint et de divers enfants auprès de l'enquêté, permet de suivre la composante familiale des ménages des individus. Nous avons donc dans un premier temps analysé de façon approchée les variations de la taille des ménages dans lesquels vivent les enquêtés (Courgeau, 1995). Cette exploration a permis de tester les deux approches différentes (détaillées précédemment) pour modéliser la dépendance temporelle des divers quotients à estimer : (1) utiliser une horloge globale qui démarre à la date de la formation du ménage; (2) utiliser les durées de séjour dans chacun des états, l'horloge étant remise à zéro lorsque le ménage entame une nouvelle étape de son parcours. L'examen des résultats obtenus montre que la prise en compte de la durée passée dans l'état précédent conduit à des estimations plus discriminantes rendant les résultats plus clairs et plus faciles à interpréter que lorsque l'on travaille avec la durée totale du processus. Le choix entre les deux horloges pourrait cependant ne pas être exclusif : en effet la durée totale du processus est la somme des durées des étapes successives. Cependant, comme nous l'avions déjà noté la complexité de tels modèles restreint l'estimation à des cas très simples (Andersen et al., 1992).

Dans un deuxième temps, il nous a semblé que l'approche devait être poursuivie avec des données plus pertinentes car des informations fondamentales échappent au recueil des biographies individuelles : celles qui concernent la structure encadrant l'individu et son impact sur les décisions individuelles. Aussi avons nous entamé une réflexion plus théorique sur l'entité à prendre en considération. L'approche biographique nous avait en effet amené à reformuler les bases de l'analyse démographique en termes d'analyses de processus stochastiques complexes (Courgeau et Lelièvre, 1989), où chaque trajectoire individuelle est replacée dans le contexte le plus large possible. Notre démarche a, cette fois, constitué à identifier l'entourage influant et influencé par l'individu en repérant les agents marquants. L'hypothèse étant que la destinée des individus résulte, dans une mesure variable selon les personnes et les périodes, de l'influence qu'ils subissent de la part des individus de leur entourage et en retour de celle qu'ils exercent sur ces individus.

Notre objectif a été double : réinsérer le ménage dans le groupe familial afin de comprendre son rôle dans les stratégies sociales des individus et d'en saisir la dynamique en réintroduisant la dimension temporelle (Lelièvre et Bonvalet, 1994; Bonvalet et Lelièvre, 1995). En effet, le système d'influence a pour support, d'une part, le cadre des ménages successifs auxquels l'individu a appartenu (ce qui implique une résidence commune) et, d'autre part, hors de ce cadre d'individus clefs, en fonction de liens qui sont centrés sur l'alliance et la filiation ; le tout constituant l'entourage des individus.

Une telle analyse permettra de poursuivre l'interprétation des comportements individuels considérés pour l'instant indépendamment de leur contexte familial et social. Il s'agit, d'une part, de prendre en compte la dimension intergénérationnelle dans l'analyse des pratiques tant résidentielles que professionnelles et même démographiques, et d'autre part, d'identifier les interactions qui s'établissent entre les individus et leur entourage. Cette recherche, qui en est à un stade très expérimental, a pour premier objectif de permettre de générer les données nécessaires à la mise en oeuvre d'une analyse

biographique de l'entourage des individus en proposant les principes de collecte d'une enquête à venir de l'Ined.

2) Passage du niveau individuel au niveau agrégé

En parallèle une autre voie de recherche s'ouvre à nous à partir de données généralement disponibles, l'analyse des liens entre les résultats obtenus à divers niveaux d'agrégation. En particulier, les liens entre le niveau individuel et le niveau agrégé.

L'analyse biographique se situe au niveau individuel : elle analyse en finesse l'influence sur le comportement étudié des diverses caractéristiques individuelles disponibles y compris la durée de séjour précédant l'événement. D'autre part on dispose souvent à un niveau plus agrégé (ville, région, etc.) de caractéristiques globales du groupe (proportion de chômeurs, d'agriculteurs, etc.) que l'on essaye de relier à des comportements. Dans cette deuxième approche on essaye de représenter une réalité complexe par un schéma relativement simplifié qui articule les grandes caractéristiques du groupe entre elles. En effet, on ne dispose plus de comportements individuels mais de proportions d'individus ayant ce comportement et l'on en déduit souvent d'hypothétiques trajectoires individuelles. Ainsi les modèles de migration vont expliquer les flux de migrants sous l'hypothèse que leur comportement est influencé par diverses caractéristiques des zones de départ et d'arrivée et par la distance physique et sociale qui sépare ces zones. Cette approche est suivie depuis fort longtemps (Young, 1924), se situe d'emblée à un niveau macro-géographique et ne permet pas de prendre en compte explicitement le temps, sauf à comparer des régressions effectuées sur des périodes successives.

L'intégration de ces deux approches est difficile à plusieurs titres. D'une part, la première prédit un comportement individuel à partir de caractéristiques biographiques tandis que la seconde prédit un comportement collectif à l'aide de caractéristiques du groupe. D'autre part, la première travaille en longitudinal alors que la seconde est essentiellement transversale : dans le cas des changements professionnels on estime soit un risque de perte d'emploi au long de la période, soit un flux de passage au chômage en transversal. On s'est très tôt posé le problème de la comparaison des résultats obtenus sous ces deux approches. L'analyse au niveau agrégé est-elle simplement une somme des résultats obtenus au niveau individuel ou, au contraire, ces résultats sont-ils, en grande mesure, indépendants entre eux ?

Les quelques travaux antérieurs (dont Robinson, 1950) concluaient de façon assez tranchée qu'on ne peut pas utiliser une corrélation écologique, mesurée au niveau agrégé, comme substitut d'une corrélation individuelle. Les cas où cela est possible sont rares et nécessitent une vérification précise de l'égalité des corrélations, ce qui n'est généralement pas réalisable du fait de l'absence de données individuelles dans de nombreux domaines des sciences sociales. En dépit de cette conclusion, de nombreux chercheurs continuent à utiliser des données agrégées, pour en tirer des conclusions au niveau individuel[1]. Disposant de données individuelles, nous avons commencé à comparer de façon systématique modèles utilisés et résultats obtenus à des niveaux différents d'agrégation afin de déterminer les liens entre ces deux analyses (Courgeau, 1994 ; Baccaïni et Courgeau, 1995).

Pour le moment cette analyse va chercher à expliquer soit les probabilités d'émigrer des individus des régions françaises ou norvégiennes, soit les taux d'émigration des mêmes régions à l'aide de caractéristiques individuelles (le fait d'être agriculteur, par exemple) ou de caractéristiques agrégées (la proportion d'agriculteurs présents dans une région). Nous avons pu montrer qu'il était possible de relier les résultats obtenus sur les probabilités d'émigrer à ceux obtenus sur les taux d'émigration, lorsque les caractéristiques étaient de type agrégé. En revanche, lorsque l'on fait intervenir l'effet des caractéristiques individuelles sur les probabilités d'émigrer, cet effet pourra être très différent de celui des caractéristiques mesurées au niveau agrégé. Nous avons pu vérifier que la corrélation entre les

[1] On trouvera des citations de nombreux travaux de ce type dans Firebaugh (1978) et plus récemment dans Piantadosi (1994) et Cohen (1994).

paramètres estimés avec les caractéristiques des régions et ceux estimés avec les caractéristiques individuelles était très faible. Il est dans ce cas possible de faire intervenir simultanément dans un modèle logit ou biographique des deux types de caractéristiques qui peuvent simultanément jouer en sens inverse. Ainsi dans le modèle appliqué aux données françaises l'effet d'être agricultrice est opposé à celui du pourcentage d'agricultrices.

Ce paradoxe apparent peut s'expliquer en décomposant la population soumise aux risques en deux groupes disjoints : les agricultrices et les non agricultrices. Il est dès lors possible d'estimer par régression le logarithme de la probabilité d'émigrer de ces deux groupes en fonction de la part d'agricultrices dans les diverses régions (figure 1).

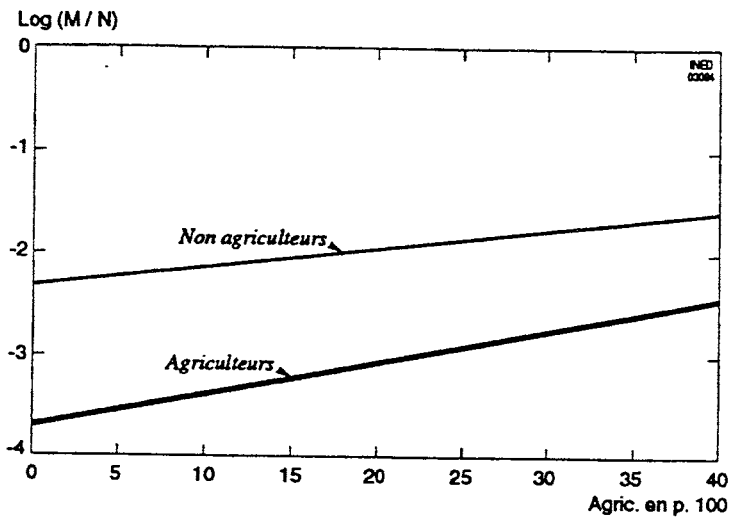

Figure 1. – Logarithme de la probabilité d'émigrer des agriculteurs et des autres catégories en fonction de la part d'agriculteurs dans chaque zone

On voit d'abord que les agricultrices ont toujours une plus faible probabilité d'émigrer, quelle que soit leur part dans la région. Cela confirme le paramètre négatif obtenu au niveau individuel (-0,333). Mais, simultanément, on voit que la probabilité d'émigrer, tant des agricultrices que des non agricultrices, croît lorsque la proportion d'agricultrices, augmente. Il en résulte que le paramètre concernant cette variable agrégée est positif (+3,785). Le danger d'inférer certaines hypothèses sur le comportement individuel à partir de résultats obtenus au niveau agrégé, apparaît clairement : la présence de nombreuses agricultrices aient de plus fortes chances d'émigrer que les autres : c'est exactement l'inverse que l'on observe au niveau individuel. De plus quelle que soit la région d'origine, ce résultat est toujours vérifié.

Comme on le voit avec ce rapide survol de l'utilisation des méthodes biographiques, leur application est extrêmement variée et procure des occasions innombrables de réflexion théorique. Au delà de l'instrument lui-même qui apporte de nouvelles ouvertures indéniables, son adaptation aux données disponibles, incomplètes, fragmentaires ou erronées suscite des formalisations nouvelles et les limites de son application conduisent à repenser la collecte.

BIBLIOGRAPHIE

ANDERSEN (P.), BORGAN (O.), GRILL (R.), KERDING (N.), (1993), *Statistical models based on counting processes*, Springer-Verlag, New-York, VII + 768 p.

BACCAINI (B.), COURGEAU (D.), (1995), *Approche individuelle et approche agrégée*, non publié, 20p.

BERNHARDT (B.), (1989), "Fertility and Employment" *Stockholm Research Reports in Demography*, n°55.

BONGAARTS (J.), BURTCH (T.K.), WACHTER (K.W.), (eds) (1986), *"Family demography": Methods and their application*, Oxford, Oxford University Press.

BRETAGNOLE (J.), HUBER-CAROL (C.), (1988), "Effects of omitting covariats in Cox's model for survival data", *Scandinavian journal of Statistics*, 15, pp. 125-138.

COHEN (B.), (1994), "Inorted commentary : in defense of ecologic studies for testing a linear-no threshold theory", *American journal of Epidemology*, vol. 39, n°8, pp. 765-768.

COURGEAU (D.), (1985), "Effet de déclarations erronées sur une analyse de données migratoires", *Chaire Quetelet : Migrations internes*, JEZIERSKI ed., Louvain - la - Neuve, 151-156.

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

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

COURGEAU (D.), (1993), "Nouvelle approche statistique des liens entre mobilité du travail et mobilité géographique", *Revue Economique*, vol. 44, n°4, pp. 791-807.

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

COURGEAU (D.), (1995), "Event history analysis of household formation and dissolution", in *Household demography and household modelling*, Van Imhoff et al., eds. Plenm Publishing Corporation, London, pp. 185-202.

COURGEAU (D.), (1995), "Mobilité : déménagement et emploi", in *Le logement en question*, F Ascher ed., Editions de l'Aube, pp. 141-170.

COURGEAU (D.), LELIEVRE (E.), (1989), *Manuel d'Analyse Démographique des Biographies*, coll. de l'INED, PUF, 268 p.

COURGEAU (D.), LELIEVRE (E.), (1990), "L'approche biographique en démographie", *Revue Française de Sociologie*, n°31, p.55-74.

COURGEAU (D.), LELIEVRE (E.), (1994), "Concurrence et Indépendance entre phénomènes démographiques, *réflexions et commentaires sur un article de X.Thierry*", Population, 2, pp.481-498.

COURGEAU (D.), NAJIM (J.), (1995), "Analyse de biographies fragmentaires", *Population*, vol. 50, n°1, pp. 149-168.

COX (D.), (1972), "Regression models and life tables", *Journal of Royal Statistical Society*, B 34, pp. 187-220.

DUCHENE (J.), (1985), "Un test de fiabilité des enquêtes rétrospectives "Biographie Familiale Professionnelle et Migratoire", *Chaire Quetelet : Migrations internes*, JEZIERSKI ed., Louvain - la - Neuve.

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

FLINN (C.), HECKMAN (J.), (1982 a), "New methods for analysing structural models of labour force dynamics", *Journal of Econometrics*, 18, pp. 115-168.

FLINN (C.), HECKMAN (J.), (1982 b), "Models for the analysis of labour force dynamics", *Advances in Econometrics*, vol. 1, pp.35-95.

HENRY (L.), (1972), *Démographie, Analyse et Modèles*, Larousse, Paris. réed. INED, Paris 1984 p.

KEMPENEERS (M.), LELIEVRE (E.), (1991), "Analyse biographique du travail féminin", *Revue européenne de démographie*, 377-400.

LANCASTER (T.), (1990), *The econometric analysis of transition data*, Cambridge University Press, 352p.

PIANTADOSI (S.), (1994), "Ecologic bras", *American journal of Epidemology*, vol. 139, n°8, pp. 761-764.

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

PRESSAT (H.), (1961), *L'analyse démographique*, Paris, PUF, 321 p.

YOUNG (E.), (1924), "The movement of farm population", Cornell University, Ithaca.