ANALYSE DE BIOGRAPHIES FRAGMENTAIRES

L'analyse biographique est sans doute déjà née lorsque J. Hajnal tire parti de l'information sur l'état matrimonial dans les recensements pour caractériser la nuptialité des générations ou lorsque Louis Henry procède de même pour la fécondité avec le nombre total d'enfants nés d'un couple. Elle se développe lorsque l'on ajoute à cette information unique des éléments sur les événements correspondants (âge au premier mariage, date de naissance des enfants successifs); le recensement britannique de 1946 marque ici un tournant important. Mais de ce tronc naissent deux branches : l'une, à base d'enquêtes, recueille rétrospectivement auprès des intéressés les informations permettant de baliser leur vie familiale, leur carrière professionnelle, leur parcours résidentiel, etc.; l'autre constitue des biographies individuelles compilant les actes administratifs qui en ont marqué les étapes (documents d'état civil, bulletins de recensement, déclarations de domicile, etc.). Chaque méthode a ses limites; en particulier, la seconde est tributaire des informations collectées par l'administration : par exemple, les changements professionnels ne sont jamais enregistrés et en France les déménagements ne le sont pas plus. On doit se contenter alors d'informations recueillies dans les recensements ou à l'occasion d'actes d'état civil. Daniel COURGEAU* *et* Jamal NAJIM** *valident ici les renseignements fournis par ces biographies incomplètes en les comparant à ceux extraits d'enquêtes* ad hoc.

L'utilisation de données d'enquêtes rétrospectives ou de registres de population a permis de nombreuses applications de l'analyse des biographies dans tous les domaines de la démographie. Bien que ces données soient en partie incomplètes, puisqu'elles ne suivent la vie d'un individu que jusqu'à la date de l'enquête ou de consultation du registre, des méthodes d'analyses fiables ont été élaborées pour tenir compte de ce fait (Courgeau et Lelièvre, 1989, 1993). Il existe cependant d'autres types de recueil de l'information qui conduisent à un récit plus fragmenté de la biographie des individus et nécessitent de ce fait des méthodes d'analyse spécifiques.

Éclairons notre propos par des exemples plus précis.

* INED.
** Institut de Statistique de l'Université Pierre et Marie Curie/INED.

En premier lieu, l'« Échantillon démographique permanent » (« E.D.P. ») de l'INSEE[1] (Sautory, 1987) fournit un fichier exceptionnel tant par sa taille (un peu plus élevée que celle d'un échantillon au 1/100 de la population présente en France entre 1968 et 1990) que par les informations qu'il comporte. Elles correspondent au couplage des données issues des bulletins individuels des recensements et des bulletins statistiques de l'état civil depuis 1968. On voit sans peine que l'analyse de la biographie familiale des individus nés après 1953 est facile à entreprendre, car tous les événements familiaux sont datés avec précision. En revanche, si l'on s'intéresse à la vie migratoire ou professionnelle, on ne dispose plus d'une datation précise, car ce fichier ne saisit la résidence et la profession des individus qu'à certaines dates, celles des recensements et des événements familiaux. Il ne permet pas de connaître la date d'occurrence des migrations et des changements professionnels, mais seulement de la situer entre les divers recensements ou événements familiaux. Les méthodes habituelles d'analyse des biographies ne sont pas directement applicables à de telles *biographies fragmentaires*[2].

En second lieu, l'enquête sur la mobilité sociale, géographique et patrimoniale en France aux XIX^{e} et XX^{e} siècles, réalisée par Jacques Dupâquier et Denis Kessler (Dupâquier, 1981), fournit un fichier historique du plus grand intérêt. Cette enquête reconstitue la généalogie descendante, en ligne masculine, de 3 000 couples repérés sous le premier Empire par leur patronyme commençant par les trois lettres « TRA ». Ils ont été répartis sur tout le territoire national dans les limites actuelles des départements, proportionnellement à la population présente au recensement de 1806. A nouveau, on dispose pour chaque acte d'état civil de la profession et de la résidence des intéressés, sans connaître la date exacte de leurs migrations ou changements professionnels. Les données sont donc de même type que dans l'exemple précédent.

Nous allons proposer et tester dans cet article un certain nombre de méthodes qui tiennent compte de cette fragmentation de l'information pour estimer le mieux possible la distribution des migrations et des changements professionnels dans le temps et montrer l'effet de certaines caractéristiques individuelles sur cette distribution.

Le test de ces méthodes sera possible car nous disposons des données rétrospectives complètes de l'enquête sur la biographie familiale, professionnelle et migratoire (« 3B »), réalisée par l'INED en 1981. En fragmentant ces données, de façon semblable à ce que l'on observe dans l'« EDP » ou dans l'enquête « TRA », nous pouvons tester la perte de précision et les erreurs entraînées par ces méthodes. Nous avons supposé ici que l'on observe les résidences ou les professions des enquêtés aux dates des re-

[1] Il existe des échantillons semblables dans d'autres pays : en Grande-Bretagne, par exemple, la « Longitudinal Study » (Social Science Research Unit, 1990) a été mise en place de façon similaire.

[2] On utilise parfois le terme beaucoup plus lourd de « biographies tronquées par intervalle ».

censements passés (1926, 1936, 1946, 1954, 1962, 1968 et 1975) et aux dates des divers événements familiaux connu par les enquêtés. Pour effectuer ces tests, nous travaillons sur les générations féminines nées entre 1911 et 1935.

Nous commencerons par poser et discuter des hypothèses qu'il est nécessaire de faire pour réaliser l'analyse de biographies fragmentaires. Nous passerons ensuite à l'estimation de modèles non-paramétriques pour mesurer l'effet de la durée sur les probabilités d'occurrences des migrations ou des changements professionnels et de modèles semi-paramétriques pour introduire l'effet de diverses caractéristiques individuelles. Nous présenterons également quelques modèles paramétriques, parmi les plus souvent utilisés dans l'analyse de ces phénomènes tout en indiquant les inconvénients qu'ils présentent. Tout au long de l'article, nous vérifierons la validité des divers modèles et des hypothèses faites à l'aide des données de l'enquête « 3B ».

Cet article généralise certains résultats partiels que nous avons présentés par ailleurs (Courgeau, 1993).

I. – Deux hypothèses fondamentales

L'information sur les lieux de résidence ou les professions à diverses dates peut laisser dans l'ombre certaines migrations ou changements professionnels. Il suffit pour cela que plusieurs des événements étudiés se produisent dans l'intervalle entre deux observations. Il faut donc faire l'hypothèse qu'un seul au maximum des événements étudiés (migrations, par exemple) peut arriver entre deux observations. Ainsi par exemple, si l'on cherche à étudier le premier changement de département après le mariage, il faut faire l'hypothèse qu'un départ et un retour dans le département de mariage ne peuvent intervenir entre deux dates d'observation.

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. On peut donc penser que l'utilisation des positions de l'individu, à la fois aux recensements et aux événements familiaux, donnera une meilleure estimation de la mobilité que lorsque l'on utilise ces positions aux recensements seulement ou aux événements familiaux seulement. Les tests que nous pourrons réaliser sur les données de l'enquête « 3B » permettront de juger de cette amélioration. Nous pourrons également mettre en évidence la mobilité qui échappe lorsque l'on utilise des biographies fragmentaires.

Une autre hypothèse est également nécessaire pour estimer les durées de séjour à partir de biographies fragmentaires : les événements qui permettent de localiser l'individu dans l'espace physique ou social doivent être indépendants de la mobilité géographique et professionnelle que l'on

entend mesurer. Sinon les interactions entre les divers phénomènes viendraient troubler cette estimation et conduiraient à des résultats incorrects.

Pour montrer plus clairement ce qu'il en est, supposons que nous cherchions à estimer la durée de séjour dans le premier logement occupé après le mariage. Les résultats obtenus se généralisent sans peine à d'autres situations.

Nous n'observons pas directement la date du départ de ce logement mais nous la situons entre des dates entremêlant à la fois migration, naissances d'enfants et recensements. En effet, si nous connaissons avec précision le début du séjour (la date de mariage), nous savons seulement que sa fin s'est produite entre deux dates de naissance d'enfant ou de recensement, ou ne s'est pas produite avant le dernier recensement considéré ou le décès de l'individu.

Certains de ces événements, recensements par exemple, sont à l'évidence indépendants de la migration considérée. Leur prise en compte n'introduira donc aucun biais dans l'estimation faite. En revanche d'autres événements, tels que les naissances successives d'enfants, pourront plus fréquemment en dépendre. On peut, en effet, penser qu'un couple, qui s'est installé après son mariage dans un logement de petite dimension, doive déménager dans un plus grand logement lorsque la taille de son ménage augmente. Dans ce cas, l'observation des dates de naissances successives risque de ne plus être indépendante du fait qu'une migration se soit produite ou non entre elles[3]. Cela introduira un biais d'autant plus important que l'on travaille sur des migrations à courte distance, très liées à l'accroissement de la taille des familles. En revanche, si l'on travaille sur des migrations à plus longue distance (changements de départements ou de régions), ce biais sera plus réduit.

On trouvera en annexe une formulation probabiliste de ces problèmes qui permet de préciser les conditions sous lesquelles l'estimation de la durée de séjour est correcte.

Du fait que nous utilisons ces méthodes avec des données rétrospectives complètes (enquête « 3B ») et artificiellement fragmentées, nous aurons la possibilité, tout au long de cet article, de tester avec précision la dépendance entre événements familiaux et mobilité spatiale ou professionnelle.

II. – Estimation des durées de séjour

Pour effectuer cette estimation, plaçons-nous dans le cas le plus général où l'on part de la $k^{\text{ième}}$ migration spatiale ou professionnelle qui se produit à l'instant aléatoire T_k et où l'on observe la durée de séjour entre cette migration et la suivante ($T = T_{k+1} - T_k$). Nous supposons ici que les

[3] Ce concept d'indépendance locale ou unilatérale a été introduit par Schweder (1970) et repris par Aalen *et al.* (1980) et Courgeau et Lelièvre (1986, 1989, 1993).

variables aléatoires T_k et T sont discrètes (annuelles, par exemple) et indépendantes entre elles. Nous cherchons donc à mesurer les probabilités $m_h = P\ (T_k = t_h)$ pour que la $k^{\text{ième}}$ migration se produise l'année t_h et $v_i = P\ (T = t_i)$ pour que la migration suivante se produise après une durée t_i ($1 \leq h \leq r$, $1 \leq i \leq s$ où r et s sont les durées maximum observées).

Étant donné le mode d'observation nous ne pouvons que situer T_k et T_{k+1} par rapport aux divers événements familiaux ou dates des recensements successifs. Ce que nous observons se présente donc sous la forme de quatre dates (t^j, t^{j+1}, $t^{j'}$, $t^{j'+1}$) telles que $t^j \leq T_k \leq t^{j+1}$ et $t^{j'} \leq T_{k+1} \leq t^{j'+1}$. Bien entendu les données peuvent être tronquées à droite, de sorte que la $(k+1)^{\text{ième}}$ migration peut ne pas se produire avant le dernier événement observé.

L'estimation des probabilités m_h et v_i est réalisée grâce au calcul de la vraisemblance des observations, en prenant les valeurs des probabilités qui la rendent maximum. Pour écrire cette vraisemblance, introduisons des paramètres $\alpha^l_{h,i}$ qui pour l'individu, l, sont égaux à l'unité lorsque $t^j \leq t_h \leq t^{j+1}$ et $t^{j'} \leq t_h + t_i \leq t^{j'+1}$, sinon ils sont nuls. Dans ces conditions, la vraisemblance des observations peut s'écrire :

$$L = \prod_{l=1}^{N} \left(\sum_{h=1}^{r} \sum_{i=1}^{s} \alpha^l_{h,i}\, m_h\, v_i \right) \qquad [1]$$

où N est le nombre d'individus ayant effectué k migrations antérieures. Les valeurs de m_h et v_i qui rendent cette vraisemblance maximum, sont obtenues par une procédure d'ajustement itératif de même type que celle proposée par De Gruttola et Lagakos (1989). Pour ce faire, introduisons une variable aléatoire égale à 1 si la vraie valeur non observée de (T_k, T) pour le $i^{\text{ème}}$ individu est égale à (t_h, t_i) et égale à 0 sinon. Alors l'espérance mathématique de cette variable conditionnée par α^l est égale à :

$$\mu^l_{h,i} = \frac{\alpha^l_{h,i}\, m_h\, v_i}{\sum_{h,i} \alpha^l_{h,i}\, m_h\, v_i} \qquad [2]$$

Il en résulte les estimations de m_h et v_i suivantes :

$$\hat{m}_h = \frac{\sum_{l,i} \mu^l_{h,i}}{N} \qquad \text{et} \qquad \hat{v}_i = \frac{\sum_{l,h} \mu^l_{h,i}}{N} \qquad [3]$$

Pour obtenir ces estimateurs, on va partir de valeurs initiales arbitraires de m_h et de v_i, puis calculer la valeur des $\mu^l_{h,j}$ correspondantes à l'aide de [2] et recalculer de nouvelles valeurs de $\hat{m}_h$ et $\hat{v}_i$ à l'aide de [3]. On répète ces itérations jusqu'à ce que les écarts entre les valeurs successives de m et de v deviennent négligeables. De Gruttola et Lagakos (1989) ont montré que l'on arrive ainsi à un maximum local ou à un point selle

de la vraisemblance. Le calcul de l'opposée de la matrice des dérivées secondes de log L, permet, à la fois de vérifier que l'on n'est pas sur un point selle[4] et fournit grâce à son inverse la matrice des variances et covariances de $(\hat{m}, \hat{v})$[5].

III. – Application aux données de l'enquête «3B» artificiellement fragmentées

L'application de ces méthodes aux données de l'enquête «3B» artificiellement fragmentées, va nous permettre de voir dans quelle mesure les hypothèses faites précédemment sont vérifiées.

Changements de logements après mariage

Lorsque l'on connaît le lieu de résidence de l'individu juste après son mariage (il est normalement enregistré sur le bulletin de mariage), on peut étudier le départ de ce lieu. On se trouve dans le cas le plus simple où la date initiale est connue (elle est prise comme instant 0) et où la date finale peut être cernée par divers types d'événements.

Supposons que l'on ne veuille ou l'on ne puisse (c'est ce qui se produit pour l'enquête «TRA») utiliser que les dates des événements familiaux suivant le mariage. Travaillons d'abord sur les changements de logements qui vérifient bien la première hypothèse : la probabilité pour un individu de revenir dans un logement antérieurement occupé est très faible (Courgeau, 1973, 1979). La figure 1 porte la fonction de répartition cumulée obtenue avec l'ensemble des données et avec les données utilisant seulement l'information fournie aux événements familiaux successifs avec l'intervalle de confiance à 95 % autour des estimations. Cette fonction de répartition donne la probabilité pour qu'un individu ait quitté son logement au mariage avant une durée donnée en années écoulées depuis ce mariage. Ainsi on peut voir que si la probabilité de rester sédentaire après un séjour de 50 ans n'est pas altérée par les troncatures (0,15), on observe une distribution au cours du temps très différente entre données observées et données fragmentées : entre 3 et 20 ans la courbe obtenue avec l'information sur les événements familiaux se situe au-dessus de la courbe réelle. On peut, dès lors, penser que la seconde hypothèse n'est pas vérifiée : il n'y a pas indépendance entre changements de logements et événements familiaux. Voyons donc plus précisément les interactions qui existent entre migration et fécondité. Nous pouvons, par exemple, calculer la probabilité

(4) Si les valeurs propres de cette matrice sont toutes positives, alors le point considéré correspond à un maximum.

(5) L'estimation de ces probabilités et de la matrice des variances et covariances peut être réalisée à l'aide de deux logiciels (npara1.C et npara2.C) mis en forme par J. Najim (1994).

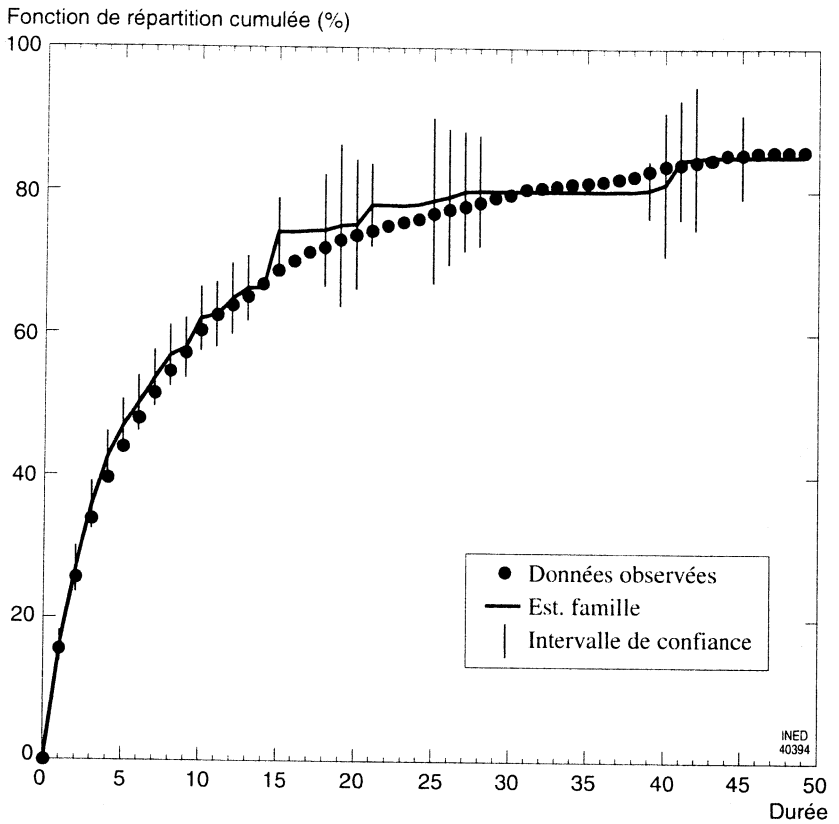

Figure 1. – Fonctions de répartition cumulées des changements de logement après le mariage, estimées à l'aide des données observées ou fragmentées par les événements familiaux, en %, avec les intervalles de confiance à 95 % pour les données fragmentées

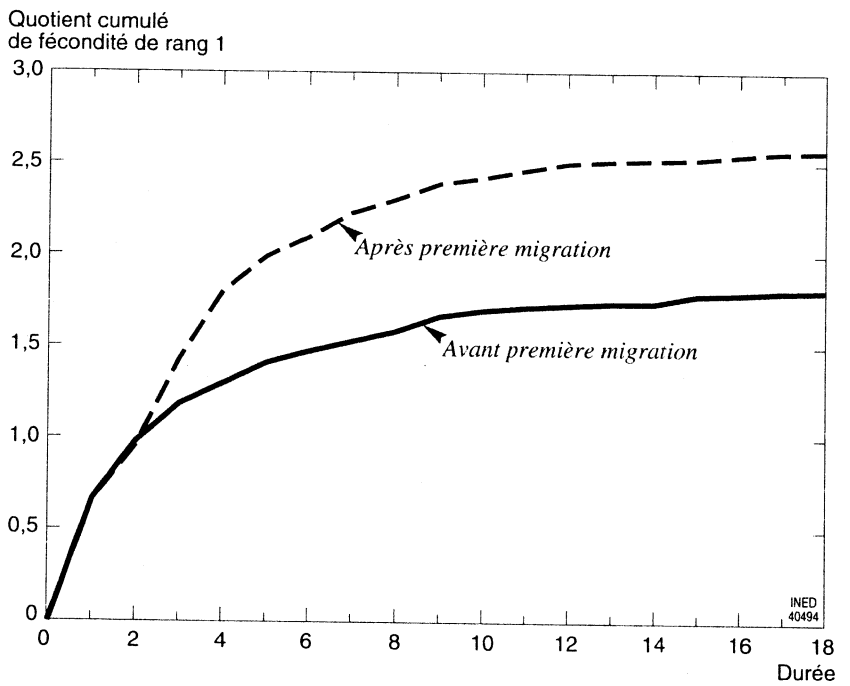

Figure 2. – Quotients cumulés de fécondité de rang 1 selon que cette naissance se produit avant ou après la première migration après le mariage

d'une première naissance selon que l'individu ait ou non changé de logement avant cette naissance. Si ces probabilités sont identiques, on vérifie que l'observation de la naissance est indépendante du phénomène étudié ici, le premier changement de logement. Pour tester cette hypothèse, nous avons porté sur la figure 2 les quotients cumulés de première naissance selon que l'individu ait migré ou non avant elle. On voit sans peine que pour les durées supérieures à 2 ans la migration antérieure favorise très fortement la venue de la première naissance. Comme nous l'avons indiqué en annexe 1 cette dépendance conduit bien à une fonction de répartition cumulée estimée avec les données fragmentées supérieure à celle estimée avec les données complètes, du moins pour des durées de 3 à 20 ans après le mariage.

Si l'on utilise maintenant les dates des recensements, pour déterminer le premier changement de logement après le mariage, on a forcément indépendance entre changements de logement et recensements. La figure 3 porte la fonction de répartition cumulée ainsi obtenue toujours comparée avec l'observée. On voit que les distributions sont étroitement mêlées tout au long du temps. Cependant, la fonction obtenue avec les observations des recensements est beaucoup plus chahutée que l'observée. Si l'on utilise maintenant simultanément les dates des recensements et celles des événe-

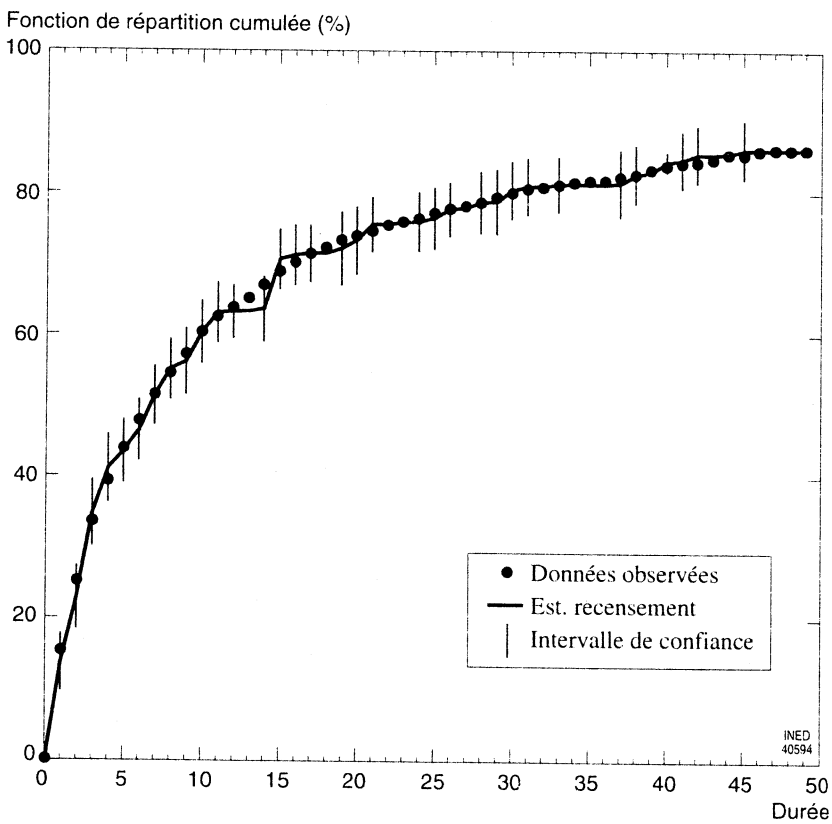

Figure 3. – Fonctions de répartition cumulées des changements de logement après le mariage, estimées à l'aide des données observées ou fragmentées par les recensements, en %, avec les intervalles de confiance à 95 % pour les données fragmentées

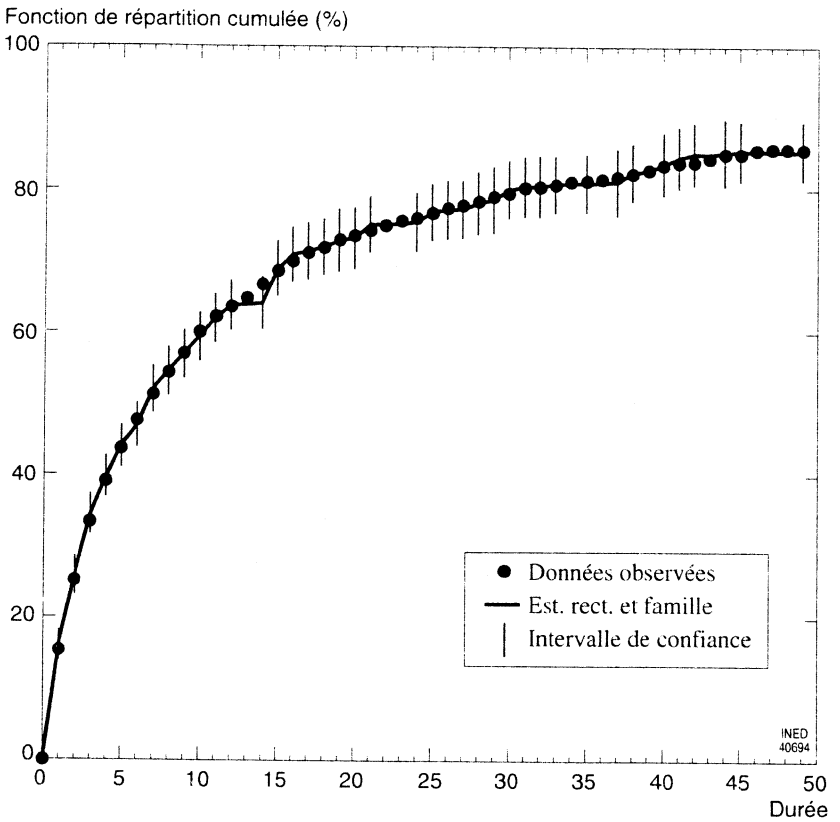

Figure 4. – Fonctions de répartition cumulées des changements de logement après le mariage, estimées à l'aide des données observées ou fragmentées à la fois par les recensements et les événements familiaux, en %, avec les intervalles de confiance à 95 % pour les données fragmentées

ments familiaux, la figure 4 montre une nouvelle amélioration de l'estimation, en dépit de la dépendance que nous avons mise en évidence plus haut. Il semble que mêler événements indépendants et événements dépendants, en augmentant la densité dans le temps des événements qui permettent de localiser l'individu, permette de compenser les erreurs dues à la dépendance. L'estimation en utilisant toutes les informations de l'« EDP » sera donc à préférer à celle qui utilise uniquement les informations des recensements.

Changements de département après mariage

Passons maintenant à l'étude des changements de département qui semblent mieux devoir vérifier la seconde hypothèse : nous avons en effet déjà montré par ailleurs (Courgeau, 1985) que la mobilité liée aux événements familiaux était essentiellement une mobilité à courte distance, tandis que la mobilité à plus longue distance se produit pour des raisons économiques. On peut donc penser que les événements familiaux seront peu influencés par les migrations interdépartementales. En revanche, la première hypothèse ne sera plus vérifiée : nous avons déjà montré (Courgeau, 1973, 1979) que parmi les nouvelles migra-

tions effectuées par les individus, 16 % environ de celles-ci correspondent à un retour dans le département de résidence antérieur.

Cette fois-ci, que l'on utilise les événements familiaux seuls (figure 5), les recensements seuls (figure 6) ou simultanément les deux types d'événements (figure 7), on aura toujours une fonction de répartition cumulée qui sera en dessous de celle réellement observée. Certains changements de département sont définitivement omis du fait d'un retour : la probabilité de ne pas changer de département tout au long d'une durée de 50 ans est de 0,605 lorsqu'on observe toutes les migrations ; elle monte à 0,636 lorsqu'on l'estime avec les événements familiaux seuls et retombe à 0,620 pour les deux autres modes d'estimation. Le calendrier des changements de département est également modifié : les différences entre ce que l'on observe et ce que l'on estime est maximum pour les courtes durées de séjour, pour se réduire ensuite et même récupérer des migrations tardives faites après 40 années de séjour. A nouveau, l'estimation avec, à la fois les départements de résidence aux recensements et aux événements familiaux est la meilleure.

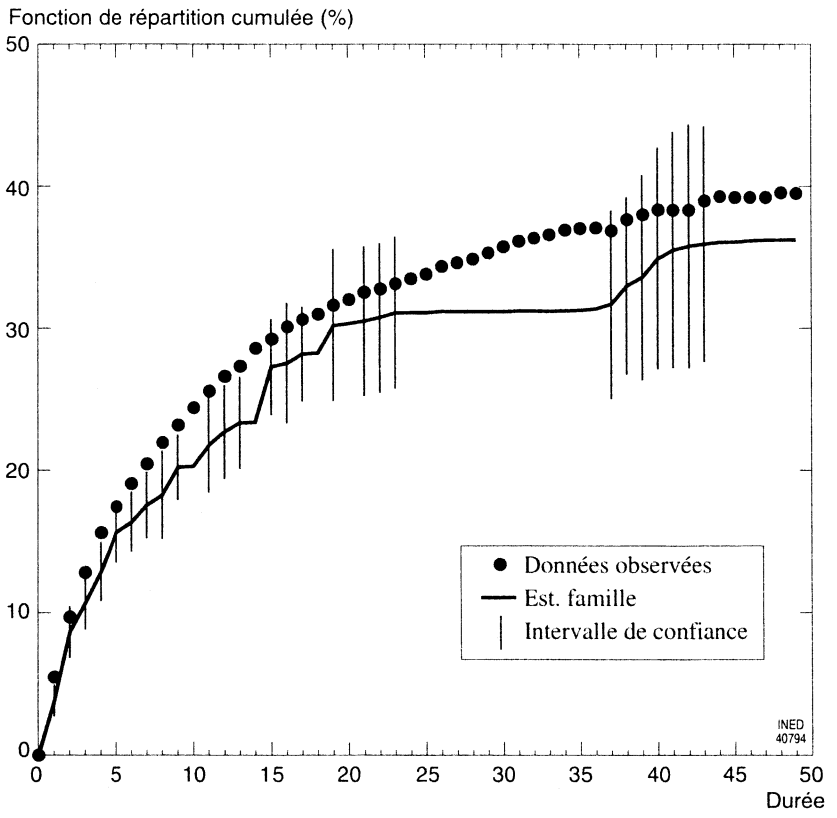

Figure 5. – Fonctions de répartition cumulées des changements de département après le mariage, estimées à l'aide des données observées et fragmentées par les événements familiaux, en %, avec les intervalles de confiance à 95 % pour les données fragmentées

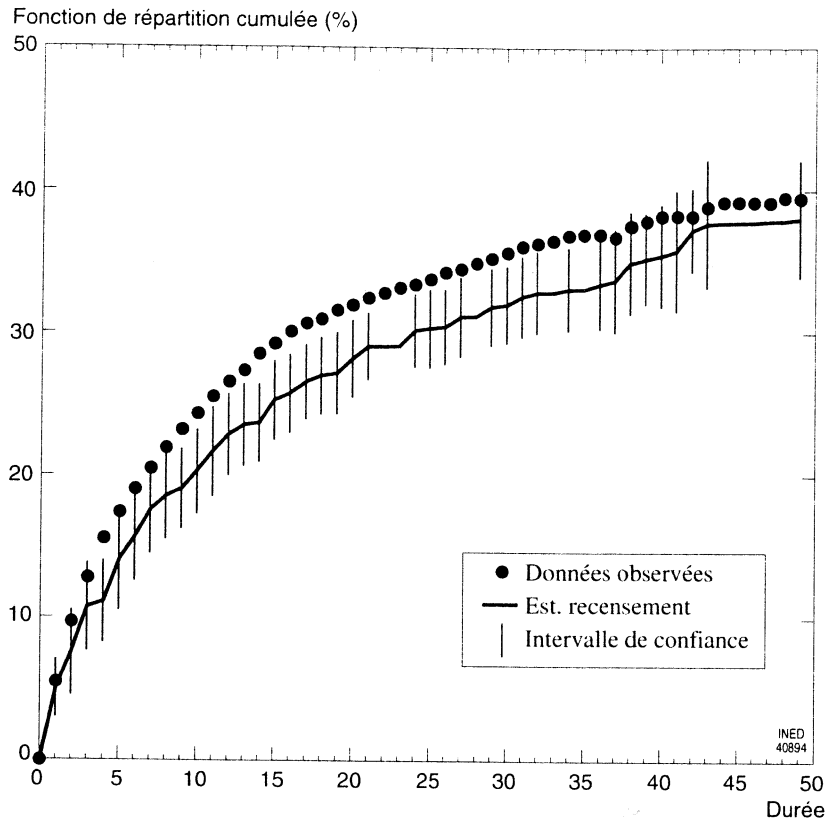

Figure 6. – Fonctions de répartition cumulées des changements de département après le mariage, estimées à l'aide des données observées et fragmentées par les recensements, en %, avec les intervalles de confiance à 95 % pour les données fragmentées

Départ du domicile parental suivi d'une première migration

Pour terminer cette première série d'exemples, considérons maintenant la survenue de deux événements dont les datations sont fragmentaires : le départ de chez les parents et la première migration après ce départ. La figure 8 donne l'estimation de la fonction de répartition cumulée de départ de chez les parents à partir des données observées et fragmentées (à la fois à partir de l'observation aux recensements et aux événements familiaux), avec l'intervalle de confiance à 95 % autour des estimations. Notons que l'on travaille ici sur une population de 1 995 femmes. Bien entendu, toutes les femmes de l'échantillon quittent leurs parents, ce qui donne une intensité finale égale à l'unité pour cet événement. La figure 9 donne de la même façon le calendrier et l'intensité de la première migration après le départ de chez les parents avec le même intervalle de confiance. L'intensité du phénomène est parfaitement estimée (0,67) et les données observées sont toujours dans l'intervalle de confiance des données estimées. Notons cependant que cet intervalle de confiance est beaucoup plus important que celui observé pour le départ de chez les parents. Cela vient de la double imprécision sur l'instant de début et de fin de ce séjour.

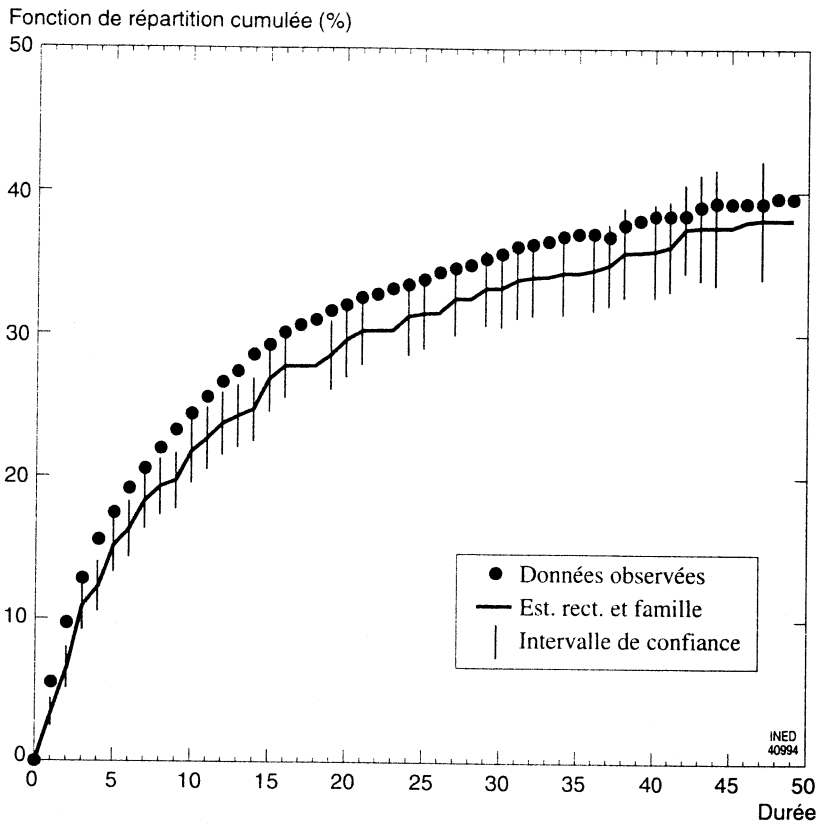

Figure 7. – Fonctions de répartition cumulées des changements de département après le mariage, estimées à l'aide des données observées et fragmentées à la fois par les recensements et les événements familiaux, en %, avec les intervalles de confiance à 95 % pour les données fragmentées

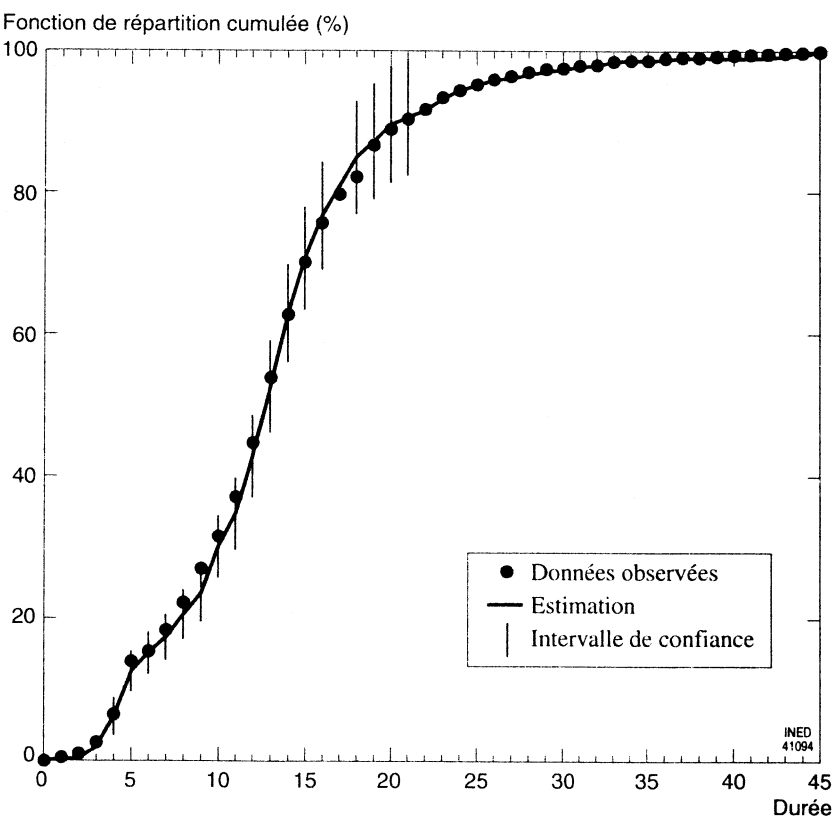

Figure 8. – Fonctions de répartition cumulées des départs de chez les parents estimées à l'aide des données observées ou fragmentées, en .%, avec les intervalles de confiance à 95.% pour les données fragmentées

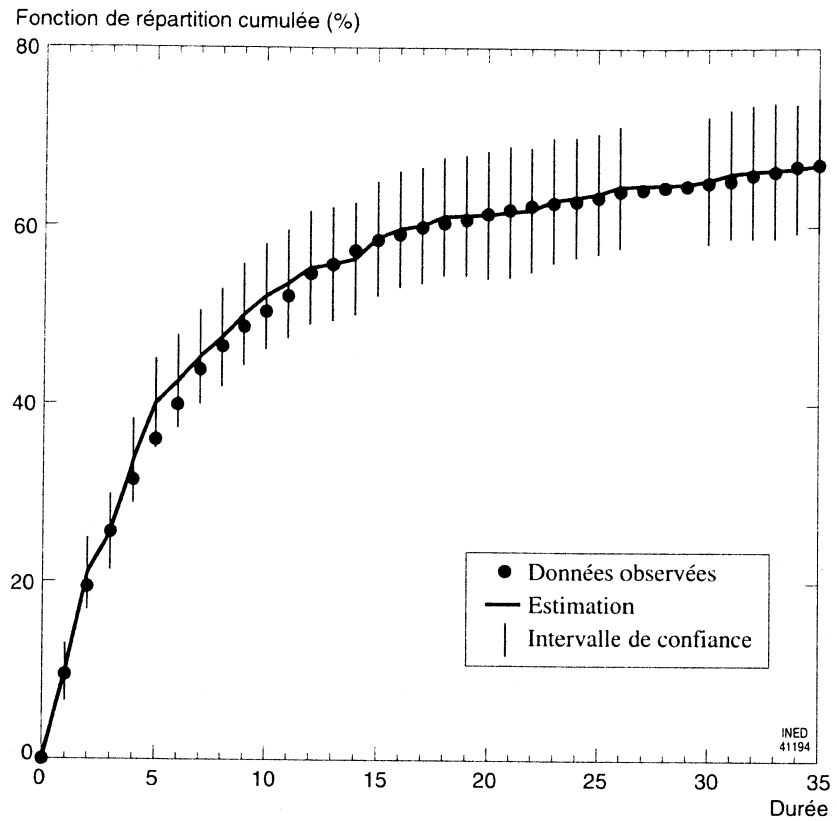

Figure 9. – Fonctions de répartition cumulées de la première migration après le départ de chez les parents, estimées à l'aide des données observées ou fragmentées, en %, avec les intervalles de confiance à 95 % pour les données fragmentées

IV. – Effet des diverses caractéristiques individuelles

Nous voulons maintenant voir l'effet de diverses caractéristiques sur cette durée de séjour. Nous utilisons ici un modèle semi-paramétrique à risques proportionnels (Courgeau et Lelièvre, 1989, 1992).

On dispose toujours pour chaque individu des quatre dates (t^j, t^{j+1}, $t^{j'}$, $t^{j'+1}$) et d'un vecteur de m caractéristiques individuelles qui peuvent être des variables quantitatives (âge à la fin d'études, rang de naissance d'un enfant, etc.), mais le plus souvent des variables qualitatives, que l'on représente de façon binaire (0 si l'individu est célibataire, 1 s'il est marié, par exemple).

On cherche toujours à mesurer la probabilité $m_h = P\ (T_k = t_h)$ dont la signification est la même que précédemment, et la probabilité d'arrivée du second événement à une durée t_i du premier, conditionnée par les ca-

ractéristiques $x_l = (x_l^1, \ldots x_l^m)$ de l'individu, $v_i(x_l) = P(T = t_i \mid x_l)$. Le modèle étant à risques proportionnels et à temps discret, on peut écrire (Kalbfleisch & Prentice, 1980) :

$$P(T > t_i \mid x_l) = [P(T > t_i \mid x = 0)]^{\exp(x_l\beta)} = (p_1 \times p_2 \times \ldots \times p_i)^{\exp(x_l\beta)} \quad [4]$$

où $p_i = P(T > t_i \mid T > t_i - 1,\ x = 0)$ est estimé pour un individu dont toutes les caractéristiques sont nulles. Il en résulte que l'on peut écrire :

$$v_i(x_l) = (p_1 \times p_2 \times \ldots \times p_{i-1})^{\exp(x_l\beta)} (1 - p_i^{\exp(x_l\beta)}) \quad [5]$$

Sous ces conditions, la vraisemblance des observations s'écrit :

$$L = \prod_{l=1}^{N} \left(\sum_{h=1}^{r} \sum_{i=1}^{s} \alpha_{h,i}^{l}\, m_h\, v_i(x_l) \right) \quad [6]$$

L'estimation de m_h, p_i et β peut être réalisée en utilisant, à la fois, l'algorithme précédent et celui de Newton-Raphson (Kim *et al.*, 1993 ; Najim, 1994). De même, on estime la matrice des variances et des covariances de ces divers paramètres[6].

A nouveau, nous appliquons ce logiciel à un fichier issu de l'enquête « 3B » pour tester la qualité des estimations. Nous nous centrons maintenant sur la validité des paramètres β estimés avec toutes les observations ou au contraire des données fragmentaires.

Du fait que l'enquête « 3B » demandait le statut d'occupation à l'arrivée et au départ de chaque logement, on a la possibilité de mettre en évidence des changements de statut, non seulement, lors d'un déménagement, mais également lors du séjour dans un même logement. On s'intéresse ici au premier logement occupé hors de chez les parents et à la migration qui suit cette installation.

On va distinguer le comportement des individus qui changent de statut d'occupation lors du même séjour (logés initialement chez les parents) de ceux qui quittent le domicile de leurs parents soit pour être logés gratuitement par un employeur, soit pour devenir propriétaires de leur logement. C'est l'effet de ces trois caractéristiques sur leur migration suivante que nous allons analyser ici. Le tableau 1 donne cet effet estimé sur le fichier complet et sur le fichier fragmenté à la fois par les événements familiaux et par les recensements.

On observe un effet très proche selon que l'on utilise le fichier complet ou le fichier fragmenté. Les individus qui changent de statut d'occupation lors du même séjour, en grande majorité des agriculteurs, ont la plus faible mobilité. Viennent ensuite les propriétaires, tandis que les individus logés gratuitement par leur employeur, pour la plupart employés de maison (Villeneuve-Gokalp, 1994), ont la plus forte mobilité, les locataires constituant le groupe témoin ($\beta = 0$). On constate une légère aug-

(6) Cette estimation qui est très lourde en calculs peut être réalisée à l'aide du logiciel Semipar.e, rédigé par Najim Jamal (1994).

TABLEAU 1. – EFFET DES CARACTÉRISTIQUES INDIVIDUELLES DANS UN MODÈLE À RISQUE PROPORTIONNEL ESTIMÉ SUR LES DONNÉES OBSERVÉES ET FRAGMENTÉES SELON QU'IL S'AGIT D'UNE ESTIMATION SEMI-PARAMÉTRIQUE OU PARAMÉTRIQUE (MODÈLE MIGRANT-SÉDENTAIRE)

Caractéristique	Données observées		Données fragmentaires			
			Estimation semi-paramétrique		Estimation modèle migrant-sédentaire	
	β	Écart-type de β	β	Écart-type de β	β	Écart-type de β
Logé chez les parents	– 2,298	0,222	– 2,341	0,227	– 2,225	0,222
Logé gratuitement	0,575	0,058	0,662	0,059	0,675	0,061
Propriétaire	– 1,619	0,171	– 1,641	0,173	– 1,637	0,179

mentation de l'écart type des paramètres lorsque l'on passe des données observées aux données fragmentées, tous les effets restant significatifs.

On peut donc conclure que l'estimation des effets des caractéristiques est peu sensible au fait que l'on travaille sur des données fragmentées.

V. – Estimations paramétriques

Lorsque l'on cherche à modéliser de façon paramétrique les durées de séjour tant dans des lieux de résidence que dans des professions, on utilise le plus souvent soit un modèle de Gompertz, soit un modèle migrant-sédentaire (Blumen *et al.*, 1955 ; Courgeau, 1973, 1979 ; Ginsberg, 1979 ; Myers *et al.*, 1967). Il s'agit maintenant de modèles en temps continu et les probabilités annuelles précédentes sont à remplacer par des densités de probabilité $m(\theta)$ pour le premier événement et $v(t)$ pour le second rapporté au premier. On peut écrire dans ce cas la vraisemblance des observations :

$$L = \prod_{l=1}^{n} \int_{t_l^j}^{t_l^{j+1}} m(\theta) \int_{t_l^{j'}}^{t_l^{j'+1}} v(t) dt \, d\theta \qquad [7]$$

En fait, le plus souvent, on estime $m(\theta)$ de façon non-paramétrique en supposant toujours T^k comme une variable aléatoire discrète. C'est la densité $v(t)$ qui sera supposée suivre un modèle de Gompertz ou un modèle migrant-sédentaire. Dans le cas d'un modèle migrant-sédentaire, on écrira :

$$v(t) = \rho k \exp(-\rho t) \exp(x\beta) [1 - k(1 - \exp(-\rho t))]^{\exp(x\beta)-1} \qquad [8]$$

où k, ρ et β sont les paramètres à estimer. Dans le cas d'un modèle de Gompertz, on a :

$$v(t) = \lambda\mu \exp(x\beta) \exp\{\mu t + \lambda \exp(x\beta) [1 - \exp \mu t]\} \qquad [9]$$

où λ, μ et β sont les paramètres à estimer.

L'estimation des paramètres à partir de la vraisemblance [7] se fait à l'aide de la méthode de Newton-Raphson et ne soulève pas de problèmes particuliers. Elle est bien entendu beaucoup plus rapide que l'estimation des modèles non-paramétriques ou semi-paramétriques[7]. Rappelons cependant que ces modèles sont beaucoup plus sensibles que les modèles semi-paramétriques précédemment présentés, aux effets de l'hétérogénéité non observée (Courgeau et Lelièvre, 1989, 1992). Cela restreint leur utilité pour analyser des données biographiques.

Nous avons porté, à titre d'exemple, dans le tableau 1, les paramètres estimés pour les individus logés chez leurs parents, logés gratuitement ou propriétaires à l'aide du modèle migrant-sédentaire. Ils sont indiscernables des paramètres estimés avec les autres formulations et leur écart-type est du même ordre de grandeur. L'avantage de ces modèles se trouve dans la rapidité des calculs. En revanche, ils imposent une distribution paramétrique dont il faut vérifier l'adéquation aux données utilisées.

Conclusion

Nous avons résolu ici les problèmes théoriques posés par l'utilisation de fichiers contenant des données fragmentaires, sous certaines conditions dont il importe de vérifier la validité lorsque l'on utilise de telles données.

La première condition est liée à la densité dans le temps des événements qui permettent de situer l'individu dans l'espace géographique ou professionnel. Plus cette densité est élevée, plus l'estimation évitera l'omission d'événements proches dans le temps. On a vu combien cette omission pouvait être importante lorsque l'on travaillait sur les changements de département. Cela entraîne également une déformation de la courbe donnant la fonction de répartition cumulée, qui va enregistrer aux durées élevées des événements qui ne devraient pas intervenir dans cette estimation (mobilité de rang élevé suivant des retours dans le département de résidence initial).

La seconde condition est liée à la dépendance qui peut exister entre événements qui situent l'individu dans l'espace et sa mobilité géographique ou professionnelle. Cette dépendance entraîne essentiellement un changement de calendrier mais affecte peu l'intensité du phénomène étudié. Nous avons vu qu'elle affectait surtout les données de démographie historique (enquête « TRA » de Dupâquier) qui ne saisissent les localisations des individus que lors des divers événements familiaux. Cet effet est très prononcé pour la mobilité à courte distance qui est très sensible aux

[7] Cette estimation peut être réalisée à l'aide des logiciels migsed2.c, migsed4.2, gomp2.c pour les modèles ne faisant pas intervenir de caractéristiques et migsed6.c pour le modèle migrant-sédentaire faisant intervenir des caractéristiques, mis en forme par J. Najim (1994).

événements familiaux. Il devient négligeable pour la mobilité à plus longue distance, tels que les changements de département ou de région. Des études plus approfondies sur les interactions entre les divers phénomènes démographiques restent à faire pour permettre des estimations plus précises du calendrier de la mobilité à courte distance.

Cette recherche a permis la mise au point de logiciels informatiques écrits en langage C que nous tenons à la disposition de tout chercheur intéressé. Ils permettent d'effectuer toutes les estimations non-paramétriques, semi-paramétriques ou paramétriques à partir de données biographiques fragmentaires. Comme nous l'avons indiqué dans l'introduction, ces données sont présentes dans de nombreux fichiers et il est indispensable de tenir compte de ces troncatures pour estimer correctement les durées de séjour et l'effet de diverses caractéristiques sur ces durées.

Avant d'utiliser ces logiciels, il importe également de vérifier la qualité de l'information portée dans ces fichiers. Ainsi, pour l'« EDP » cette vérification se révèle être indispensable car les renseignements recueillis lors des recensements peuvent être incorrectement déclarés ou omis par les intéressés ; certaines personnes ou même certains ménages peuvent échapper au recensement (Coeffic, 1993).

La poursuite de ce travail se trouve non seulement dans l'exploitation de données biographiques fragmentaires existantes, mais dans la mise au point de modèles qui permettent de prendre en compte le fait que les hypothèses de base ne sont pas vérifiées. L'utilisation de données de registres de population ou d'enquêtes biographiques portant sur des effectifs élevés, devrait permettre de corriger ces erreurs. Un travail important reste encore à faire sur ce sujet.

Daniel COURGEAU, Jamal NAJIM

RÉFÉRENCES BIBLIOGRAPHIQUES

AALEN (O.), BORGAN (O.), KEIDING (N.), THORMAN (J.), (1980), « Interaction between life history events : Non parametric analysis for prospective & retrospective data in presence of censoring », *Scandinavian Journal of Statistics*, 7, pp. 161-171.

BLUMEN (I.), MARVIN (K.), Mc CARTHY (P.J.), (1955), *The industrial mobility of labour as a probability process*, Cornell Studies of Industrial and Labour Relations, vol. VI, New York.

COEFFIC (N.), (1993), « L'enquête post-censitaire de 1990. Une mesure de l'exhaustivité du recensement », *Population*, 48, 6, p. 1655-1682.

COURGEAU (D.), (1973), « Migrants et migrations », *Population*, 28, 1, pp. 95-129.

COURGEAU (D.), (1979), « Migrants and migrations », *Population, Selected papers*, 3.

COURGEAU (D.), (1985), « Changements de logement, changements de départements et cycle de vie », *L'Espace géographique*, 4, pp. 289-306.

COURGEAU (D.), (1993), « An attempt to analyse individual migration histories from data on place of usual residence at the time of certain vital events. France during the nineteenth century », in *Methods in Historical Demography*, D. Reher and R. Schofield eds., Clarendon Press, Oxford, pp. 206-222.

COURGEAU (D.), LELIÈVRE (E.), (1986), « Nuptialité et agriculture », *Population*, 41, 2, pp. 303-326.

COURGEAU (D.), LELIÈVRE (E.), (1989), *Analyse démographique des biographies*, Editions de l'INED, Paris.

COURGEAU (D.), LELIÈVRE (E.), (1992), *Event history analysis in demography*, Clarendon Press, Oxford.
DUPÂQUIER (J.), (1981), « Une grande enquête sur la mobilité géographique et sociale du XIXè et XXè siècles », *Population*, 36, 6, pp. 1164-1167.
GINSBERG (R.), (1979), « Timing and duration effects in residence histories and other longitudinal data. II Studies of duration effects in Norway, 1965-1971 », *Regional Sciences and Urban Economics*, 9, pp. 369-392.
De GRUTTOLA (V.), LAGAKOS (S.W.), (1989), « Analysis of doubly-censored survival data, with application to AIDS », *Biometrics*, 45, pp. 1-11.
KALBFLEISCH (J.), PRENTICE (R.), (1980), *The statistical analysis of failure time data*, Wiley, New York.
KIM (M.Y.), De GRUTTOLA (V.G.), LAGAKOS (S.W.), (1993), « Analysing doubly censored data with covariates, with application to AIDS », *Biometrics*, 49, pp. 13-22.
MYERS (G.C.), Mc GINNIS (R.), MASNICK (G.), (1967), « The duration of residence approach to a dynamic stochastic model of internal migration : a test of the axiom of cumulative inertia », *Eugenics Quartely*, 14, 2, pp. 21-126.
NAJIM (J.), (1994), *Les méthodes d'estimation de la probabilité d'arrivée d'un événement et leurs utilisations en logiciels*, Mémoire de l'Institut de Statistique de l'Université Pierre et Marie Curie, Paris.
SAUTORY (O.), (1987), « L'échantillon démographique permanent de l'INSEE », *Courrier des Statistiques*, 41, pp. 1-4.
SCHWEDER (T.), (1970), « Composable Markov processes », *Journal of Applied Probabilities*, 7, pp. 400-410.
S.S.R.U. (1990), *OPCS longitudinal study. User manual*, London.
VILLENEUVE-GOKALP (C.), (1994), « Les gens de maison », *Population*, 4-5.

ANNEXE
Formulation probabiliste des interactions entre dates de mobilité, de recensement et d'événements familiaux

Nous formalisons ici plus avant les probabilités que nous estimons comparées à celles que nous devrions estimer.

Travaillons pour simplifier les formulations sur la première migration après le mariage. Les résultats que nous obtiendrons se généralisent sans peine aux cas plus complexes. Notons T_1 la variable aléatoire correspondant à la durée entre le mariage et la première migration qui le suit. Soient T^1, T^2,...T^m, les durées entre le mariage et les divers événements qui permettent de localiser l'individu (naissances et recensements). Ces variables aléatoires sont positives et ordonnées.

Bien que nous ne puissions pas estimer la probabilité d'arrivée de la première migration après le mariage, nous pouvons estimer celle d'événements plus complexes faisant intervenir, à la fois, migrations, naissances et recensements. Si dans ce cas, nous connaissons avec précision le début du séjour (la date de mariage), nous savons seulement que sa fin s'est produite entre deux dates de naissances ou de recensements, ou ne s'est pas produite avant le dernier recensement considéré ou le décès de l'individu. On peut dans ce cas estimer des probabilités conditionnelles. Ainsi, si la première migration s'est produite entre le $j^{\text{ième}}$ et le $(j+1)^{\text{ième}}$ événements observés aux durées t et t', on peut estimer la probabilité suivante :

$$P(t \le T_1 \le t' \mid T^j = t \cap T^{j+1} = t') = \frac{P(t \le T_1 \le t' \cap T^j = t \cap T^{j+1} = t')}{P(T^j = t \cap T^{j+1} = t')} =$$

$$P(t \le T_1 \le t') \frac{P(T^j = t \cap T^{j+1} = t' \mid t \le T_1 \le t')}{P(T^j = t \cap T^{j+1} = t')} \qquad [1]$$

la dernière relation ayant été obtenue en appliquant deux fois le théorème des probabilités composées. Nous avons ainsi isolé la probabilité que nous désirons estimer,

$P(t \leq T_1 \leq t')$, pour montrer par quel terme il faut la multiplier pour obtenir la probabilité que nous pouvons en fait estimer. On voit que lorsque la probabilité d'arrivée des deux événements qui localisent différemment l'individu, est indépendante du fait qu'une première migration se soit produite entre eux, alors la probabilité que l'on estime est égale à celle que nous désirons estimer.

Si la première migration se produit entre deux recensements cette hypothèse est parfaitement vérifiée. En revanche, si elle se produit entre deux événements familiaux, il nous faudra vérifier s'il y a indépendance entre naissances et première migration.

De façon semblable si l'on sait que la première migration n'a pas eu lieu avant le dernier événement observé, à la durée t, on estime la probabilité :

$$P(T_1 \geq t \mid T^m = t) = P(T_1 \geq t) \frac{p(T^m = t \mid t \leq T_1)}{P(T^m = t)} \qquad [2]$$

On voit à nouveau que la fonction de séjour, $P(T_1 \geq t)$, est correctement estimée si la probabilité du dernier événement ne dépend pas du fait que l'individu ait migré ou non antérieurement à lui.

Dans le test que nous avons réalisé sur les interactions entre migration et fécondité (partie 3 de l'article) on peut réécrire la formule [1], où $t = 0$ puisque l'on travaille sur l'intervalle entre mariage et première naissance :

$$P(T_1 \leq t' \mid T^1 = t') = P(T_1 \leq t') \frac{P(T^1 = t' \mid T_1 \leq t')}{P(T_1 = t')} \qquad [3]$$

On voit dans ce cas que la fonction de répartition cumulée que l'on estime sera supérieure à celle que l'on observe, car $P(T^1 = t' \mid T_1 \leq t') > P(T^1 = t')$ pour les durées auxquelles surviennent les premières naissances dans le mariage. Ce résultat peut se compléter en observant les naissances de rang plus élevé.

COURGEAU (Daniel), NAJIM (Jamal). – **Analyse de biographies fragmentaires**

Certaines sources de données longitudinales, telles que l'échantillon démographique permanent de l'INSEE ou l'enquête sur la mobilité sociale, géographique et patrimoniale de Dupâquier et Kessler, comportent des informations biographiques fragmentaires. Cet article propose des méthodes d'analyse de telles biographies et discute les hypothèses qui doivent être vérifiées pour qu'une telle approche donne des estimations correctes. La validité de ces hypothèses est testée à l'aide des données rétrospectives complètes de l'enquête sur la biographie familiale, professionnelle et migratoire, artificiellement fragmentées de façon semblable à ce que l'on observe dans les sources incomplètes. Cela ouvre la possibilité d'utiliser des données fragmentaires pour une estimation des durées de séjour correcte, en particulier, dans le domaine de la mobilité géographique et professionnelle.

COURGEAU (Daniel), NAJIM (Jamal). – **Analysis of fragmentary biographies**

Some sources of longitudinal data, such as INSEE's permanent demographic sample and the social, geographic and wealth mobility survey by Dupaquier and Kessler, contain fragmentary demographic information. This article describes methods for analysing such biographies and dicusses what assumptions are required to give correct estimations. The validity of these assumptions is tested using complete retrospective data from the survey on family, professional and migratory data artificially fragmented in the same way as observed in incomplete sources. This opens up the possibility of using fragmentary data to estimate the correct lengths of residence, particularly in the field of geographic and professional mobility.

COURGEAU (Daniel), NAJIM (Jamal). – **Análisis de biografías fragmentadas**

Algunas fuentes de datos longitudinales, como la encuesta demográfica permanente del INSEE o la encuesta de movilidad social, geográfica y patrimonial de Dupaquier y Kessler, ofrecen informaciones demográficas fragmentadas. Este artículo propone métodos para analizar tales biografías y discute las hipótesis que deben cumplirse para que los resultados sean correctos. La validez de las hipótesis se verifica utilizando datos retrospectivos completos procedentes de la encuesta sobre la biografía familiar, profesional y migratoria; estos datos se fragmentan para hacerlos comparables a los de fuentes incompletas. Este análisis hace posible la utilización de datos fragmentados para estimar duraciones de estancia sin sesgos, concretamente en el caso la movilidad geográfica y profesional.