L'APPORT DE L'ANALYSE DES BIOGRAPHIES AUX SCIENCES SOCIALES

Philippe ANTOINE, Philippe BOCQUIER
Démographes, UR 55 : "Enjeux de l'urbanisation"

L'analyse statistique des biographies est réputée d'accès difficile. Pourtant de nombreux progrès ont été faits en ce domaine, de la collecte de données rétrospectives jusqu'aux techniques de recherche. Nous essaierons ici de présenter quelques grands principes de base de l'analyse des biographies du point de vue probabiliste, voie actuellement suivie par la plupart des chercheurs en sciences sociales pratiquant ce type de recherche. Cette approche se distingue de celle qui a été exposée dans un précédent article par Olivier Barbary (*Chroniques du SUD*, n°13) concernant les techniques de l'analyse des données appliquées à l'étude des itinéraires biographiques.

Prenant leur origine dans le traitement de données épidémiologiques sur de petits échantillons, les techniques d'analyse quantitative des biographies connaissent à l'heure actuelle une large diffusion dans toutes les sciences sociales. On parle alors d'analyse des durées de séjour, ou encore d'analyse des événements ou des transitions (*event history analysis*). Plus précisément, lorsqu'on fait intervenir des événements de différente nature, on parle alors d'analyse des biographies.

Nous verrons que les techniques d'analyse des biographies sont de plus en plus accessibles aux chercheurs, même non-statisticiens, mais qu'elles demandent un travail préalable de conceptualisation très rigoureux. Le traitement du temps est absolument central dans ces analyses. Les atouts essentiels de l'approche probabiliste sont la maîtrise des variables explicatives et l'évaluation des interférences entre événements de nature différente.

Le point sur les progrès technologiques (micro-informatique)

Pendant longtemps, les difficultés de programmation et les moyens de calcul importants (centre de calcul) nécessaires ont constitué un frein à la pratique de l'analyse des biographies. Le développement rapide des capacités de calcul sur micro-ordinateurs et l'amélioration considérable des logiciels statistiques ont rendu les programmes de calculs accessibles. On ne compte plus les logiciels qui, accompagnant les progrès de la micro-informatique, proposent des procédures de plus en plus simples pour l'analyse des biographies. Nous travaillons actuellement avec le logiciel STATA, qui est particulièrement pratique pour ce style d'analyse.

Reconnaître au recueil rétrospectif ses mérites

L'amélioration des outils de collecte des données a toujours été une des préoccupations des démographes. L'outil le plus connu, le recensement, ne donne qu'une photographie de la population à un instant donné, il ne rend pas suffisamment compte de la dynamique de la population. La plupart des études démographiques s'intéressent à des événements majeurs de la vie de l'individu : mariage, naissance des enfants, décès. Une grande part des acquis méthodologiques a concerné l'observation suivie, dont un bon exemple est l'observatoire de population de Niakhar (au Sénégal). Il s'agit de suivre pendant plusieurs années un échantillon de population et d'enregistrer, au fur et à mesure, tous les événements démographiques. La collecte des données est fiable, mais il faut attendre plusieurs années avant de disposer de résultats montrant l'évolution des comportements.

De nombreuses enquêtes rétrospectives recueillent les événements passés, mais ne considèrent bien souvent que les caractéristiques socio-économiques de l'individu au moment de l'enquête (exemple les enquêtes démographiques et de santé dites EDS). Les enquêtes biographiques au contraire mettent en relation les événements démographiques, l'itinéraire professionnel, l'itinéraire résidentiel. Le recueil des biographies s'appuie sur un bon repérage dans le temps des événements vécus par l'enquêté. Peu de personnes mémorisent les dates des événements qu'ils ont vécus, mais en revanche, l'enchaînement des événements familiaux est facilement gardé en mémoire. Pour aider les personnes enquêtées à placer dans le temps les principaux moments de leur vie, nous avons eu recours à la fiche AGEVEN (âge-événement). Avant de commencer à remplir le questionnaire, l'enquêteur demande à l'enquêté de situer dans le temps les principaux événements de sa vie familiale, puis de sa vie migratoire et résidentielle, et de sa vie professionnelle. Ces événements sont replacés au fur et à mesure de l'entretien sur une fiche où figurent une échelle de temps (années calendaires) et la durée écoulée depuis l'événement.

Passage du recueil des données à l'exploitation statistique

Recueillir des biographies nécessite un questionnaire relativement long et complexe. Les événements sont classés dans le temps, et le plus pratique est de recueillir l'information par grands thèmes : activités, migrations, etc. A chaque changement de profession, par exemple, l'enquêteur remplit une nouvelle "ligne" d'informations. A l'issue du terrain, on se retrouve avec une série d'informations biographiques concernant un même individu. Le travail le plus fastidieux et le plus délicat est le passage du questionnaire aux fichiers informatiques. Dans le cas de l'enquête de Dakar, nous disposons, parallèlement à de l'itinéraire migratoire, de l'itinéraire professionnel à Dakar, de l'itinéraire résidentiel en ville, de l'itinéraire matrimonial, de la descendance. Pour chaque thème nous constituons un fichier où, pour un même individu, les événements sont datés et classés dans le temps. Ensuite

il faut fusionner ces différents fichiers, afin d'obtenir dans un fichier unique, l'ensemble des événements vécus par l'individu. Ce travail informatique nécessite un certain nombre de procédures qui ont été testées lors de l'enquête de Dakar. Ces procédures mises au point par Ph. Bocquier ont été proposées aux concepteurs du logiciel STATA qui les ont acceptées et elles seront prochainement diffusées par STATA à ses utilisateurs (voir le STB22, à paraître).

Pour un même individu, le fichier comprend autant de lignes que de changements d'état, et ainsi on peut savoir à chaque instant de sa vie, sa profession, son lieu de résidence, sa situation matrimoniale, et les renseignements afférents à ces divers statuts, et mettre ces éléments en relation. L'ensemble de ces informations sont datées, et l'on connaît pour chaque période de la vie de l'individu, le temps passé dans un état.

Le temps comme variable déterminante pour l'explication causale

Pour qu'il y ait causalité, il faut nécessairement une cause (appelée conventionnellement X) et un effet (Y). Si X cause Y, alors Y ne peut simultanément causer X (principe de l'asymétrie causale). Dans cette formulation, le terme "simultanément" est en fait essentiel : c'est parce que le processus se déroule dans le temps que la relation causale est asymétrique. On voit que **le temps est à la base de la perception que l'on a de la causalité**. Pour percevoir une relation, il est nécessaire de faire évoluer cause et effet dans le temps. C'est dès la conceptualisation que l'on doit se poser la question du temps.

Actuellement, le principe de priorité temporelle de la cause sur l'effet n'est guère remis en question. Cela semble un principe épistémologique adopté par l'ensemble de la communauté scientifique, quelle que soit la discipline. Cela ne signifie pas pour autant que toutes les analyses se conforment maintenant à ce principe fondamental. Sans même que parfois les auteurs en aient conscience, le temps est mal défini. Or, l'indétermination du temps d'observation est à l'origine de bien des erreurs d'interprétation. Dès lors que l'on étudie un événement, on devrait toujours utiliser un moyen d'observation qui permette de tenir compte du temps, afin de bien définir les causalités possibles.

En somme, à chaque fois que l'on doit évaluer un problème, il faut **se poser la question du moment d'observation et du rapport au temps de la caractéristique étudiée dans la population**. Chaque fois qu'une caractéristique variant dans le temps est analysée par un modèle qui, lui, ne tient pas compte du temps, il faut questionner la validité des résultats présentés.

Les concepts de population soumise au risque de troncature et de risques concurrents : les bases de l'analyse des biographies

Il s'agit de prendre en considération le temps qui s'écoule entre un instant de référence commun à tous les individus analysés et la date de l'événement observé ou bien la date de sortie de l'observation. Cette méthode nécessite surtout un effort important de conceptualisation rigoureuse de la question étudiée. Il faut définir précisément la population soumise au risque, l'événement étudié (le risque), les risques concurrents qui amèneront l'individu à sortir de l'observation. Par exemple, si l'on étudie la transition du premier mariage au divorce pour les hommes à Dakar, la population soumise au risque sera composée des hommes en première union qui résident à Dakar depuis le début de leur union ; le temps qui s'écoule sera mesuré depuis la date de cette union jusqu'à la date de divorce. Toutefois l'observation peut-être tronquée si l'individu quitte Dakar (il émigre avant son éventuel divorce) ou si son épouse décède (il devient veuf). S'il reste présent et toujours marié (l'individu est donc toujours soumis au risque), la date de troncature sera la date de fin d'observation, c'est-à-dire la date de l'enquête. Ce type d'analyse permet de dépasser l'analyse transversale et de prendre en considération les différents états qu'a connus un individu. Ainsi on peut étudier l'itinéraire matrimonial, l'itinéraire professionnel et tenir compte des influences de l'un sur l'autre. Le mariage des femmes accélère-t-il ou non leur entrée sur le marché du travail ? Le divorce change-t-il ce rythme d'entrée ? Autant de questions auxquelles il devient possible de répondre.

D'un point de vue descriptif, on calcule une probabilité de connaître l'événement à chaque âge, et l'on peut ainsi obtenir une courbe qui s'interprète simplement comme la proportion de "survivants" pour chaque durée de séjour dans un état donné. Cette courbe dite de séjour, ou encore de Kaplan-Meier, est un des outils exploratoires les plus efficaces de l'analyse des biographies. La courbe décrit le comportement hypothétique d'une cohorte qui aurait connu les mêmes conditions de vie pour que l'événement étudié, éventuellement, se réalise.

L'avantage des modèles probabilistes : tenir compte des variables explicatives

Comment passe-t-on de la description à l'analyse plus approfondie ? Le premier stade de la statistique est généralement la construction d'un tableau croisé de deux variables. Les résultats de ce croisement sont souvent considérés, à tort, comme "neutres", parce qu'ils sont "simples" à produire. Or, dans un tableau à deux dimensions, on cherchera à expliquer la distribution d'une variable par une autre variable, et par cette variable seule. Les autres variables qui n'apparaissent pas dans le tableau sont supposées ne pas avoir d'effet. Un tableau croisé est donc implicitement un "modèle" comportant une seule variable explicative. On peut compliquer ce

modèle en y ajoutant une variable ou deux, mais il faut bien comprendre que **l'analyse descriptive contient toujours un modèle d'explication implicite**, même si elle tient compte du temps (comme dans les courbes de séjour), et quel que soit le nombre de variables prises en compte.

Il y a trois inconvénients majeurs à augmenter le nombre de dimensions dans une tabulation croisée. D'abord, l'effet des variables explicatives sur la variable expliquée est combiné : on ne peut clairement identifier l'effet propre d'une des variables explicatives, puisqu'elles jouent toutes en interaction sur la variable expliquée. Ensuite, un tableau à plus de quatre dimensions est souvent incompréhensible pour une personne normalement constituée (y compris un statisticien). Enfin, en multipliant les dimensions d'un tableau, on réduit le nombre d'observations dans chacune des cases de ce tableau.

Comment fait-on, quand on ne peut commenter ni les tableaux univariés (variable par variable), ni un grand tableau incompréhensible croisant toutes les variable disponibles, avec des effectifs très faibles dans chaque case du tableau ? On doit faire appel à des techniques d'analyse de régression qui permettent de **mesurer l'influence d'une variable explicative tout en contrôlant l'influence des autres variables explicatives sur la variable expliquée**, quel que soit leur nombre. C'est le sens de la phrase rituelle : "toutes choses égales par ailleurs".

Bien que les tableaux croisés ne soient pas généralement appelés des "modèles" statistiques, il faut souligner qu'ils ne sont pas moins des modèles que les modèles de régression statistique. Simplement, les schémas de causalité ne sont pas les mêmes. Dans un cas, les variables explicatives jouent en interactions, tandis que dans l'autre, elles jouent indépendamment.

Le modèle semi-paramétrique dit de Cox

L'idée (géniale) qu'a eue D.R. Cox en 1972, fut de combiner deux types d'analyse : régression et tables de survie. **On peut voir le modèle de Cox comme le contrôle par la régression de l'effet des variables explicatives dans l'analyse de survie, ou bien comme l'introduction de la dimension temporelle dans la régression.** Les avantages d'une technique permettent de combler les lacunes de l'autre.

Pour résoudre le problème de la durée et des facteurs explicatifs, l'idée de Cox fut de faire une régression non pas sur la caractéristique acquise par l'individu à l'issue de sa vie, mais sur la caractéristique acquise chaque année de son existence jusqu'au moment de l'enquête. En quelque sorte, chaque année vécue par chaque membre de l'échantillon constitue une observation. La modalité de référence, telle que l'exige le modèle de régression, n'est pas unique pour l'ensemble de l'échantillon, mais elle est

propre à chaque durée d'observation. Cette série de probabilités permet d'établir une courbe de séjour de référence (par exemple en l'état de "non encore actif" s'il s'agit de l'analyse du premier emploi) appelée encore fonction de séjour de base : c'est la composante non paramétrique du modèle.

Ce modèle de régression calcule alors l'effet des variables explicatives sur le risque annuel de connaître l'événement. C'est la composante paramétrique du modèle, qui s'ajoute à la composante non paramétrique, pour former un modèle dit semi-paramétrique. A chaque variable est associé un coefficient de régression qui mesure l'influence moyenne de cette variable sur le risque annuel.

La spécificité de l'analyse des biographie : les interférences entre événements de nature différente

Les variables explicatives les plus communément utilisées sont celles qui caractérisent l'individu à sa naissance : il s'agit par exemple, du sexe, de l'appartenance à un groupe ethnique, à une caste, etc. On suppose que l'effet de chacune de ces variables est constant tout au long de la vie de l'individu : c'est ce qu'on appelle des conditions permanentes. Leur effet est supposé proportionnel à la probabilité annuelle de connaître l'événement étudié comme on l'a expliqué plus haut.

Dans un modèle semi-paramétrique, les événements qu'a connus l'individu depuis sa naissance, et qui ont pu influer sur ses chances de connaître l'événement étudié, peuvent aussi être introduits sous la forme de variables explicatives. Par rapport au modèle simple où n'intervient qu'une condition permanente (par exemple l'appartenance à une génération), on peut introduire des **variables explicatives qui évoluent dans le temps**, c'est-à-dire des événements qui peuvent modifier le cours de la vie d'un individu : on est donc là au plus près de la relation causale élémentaire, qui pose comme principe la priorité temporelle de la cause sur l'effet. Par exemple, l'influence de l'itinéraire de formation (études, apprentissage, chômage, etc.) sur la probabilité d'obtenir un emploi peut être analysée en tant que rendement du capital humain, c'est-à-dire comme une **variable interne** (ou endogène) au processus d'entrée ou de mobilité sur le marché du travail.

La possibilité de faire intervenir des variables explicatives au cours du temps d'observation est surtout intéressante lorsqu'on fait intervenir des **variables externes** (ou exogènes) au processus : on pense par exemple au mariage, à l'itinéraire d'une autre personne du ménage (le conjoint, les enfants...), à la fermeture d'une usine, à un changement de législation du travail, etc. Ces événements explicatifs ont pour particularité de n'avoir pas forcément lieu avant l'événement étudié, au contraire des événements internes au processus.

Avec les variables indépendantes qui évoluent au cours du temps, on tente de suivre le processus au fur et à mesure qu'il se déroule, en contrôlant à chaque étape les changements de situation dans la vie des enquêtés. Là encore, c'est dans le **contrôle du temps** que réside l'intérêt de l'analyse des biographies.

En somme, le modèle semi-paramétrique, dit de Cox, nous permet de contrôler le temps à deux niveaux : par sa composante non paramétrique (qui tient compte de l'interruption de l'observation à la date d'enquête), et par l'utilisation qui est faite des variables indépendantes, fonction du temps. On peut donc se situer au plus près de l'analyse causale en construisant des systèmes de relation entre variables où le temps est introduit explicitement (principe de priorité temporelle), et en confrontant ces systèmes de causalité avec les données biographiques. On peut parler de modèles dynamiques dans le sens où c'est bien une chaîne causale qu'on tente de vérifier.

Quelques résultats en guise d'illustration

C'est lorsque l'analyse des biographies porte sur l'interférence entre des événements de nature différente qu'elle donne toutes ses potentialités. Ainsi, on peut étudier des phénomènes aussi divers que l'effet des changements matrimoniaux sur la carrière professionnelle (en particulier chez les femmes), l'effet de l'arrivée d'un nouvel enfant sur la survie du dernier-né, l'effet de la polygamie sur le divorce, l'effet de l'entrée dans la vie active sur le départ du domicile parental, l'effet d'un changement de législation sur l'accès au logement ou à l'emploi, etc.

Par exemple, nous voulions connaître les facteurs de divorce. Dans le cas d'une analyse descriptive classique, nous aurions constaté que sur les 511 hommes ayant été mariés au moins une fois, 25 ont un statut de divorcé au moment de l'enquête. Parmi eux 16% sont chômeurs. Peut-on en conclure pour autant que le chômage est un facteur de divorce ? A priori non, car la situation au moment l'enquête ne correspond pas à celle observée au moment de l'événement. Le recueil biographique nous permet de connaître les diverses caractéristiques qu'a connues l'individu au cours de sa vie. Ainsi dans l'enquête, 106 des premières unions des hommes se terminent par un divorce. Deux facteurs majeurs accélèrent le divorce. D'une part les situations de précarité : perte de logement et surtout chômage multiplient par trois le risque de divorcer. D'autre part l'entrée en polygamie : l'arrivée d'une seconde épouse multiplie le risque de divorce par quatre, c'est-à-dire qu'elle constitue à Dakar le principal facteur de divorce de la première épouse.

Autre exemple concernant cette fois les femmes, où l'on constate que l'itinéraire matrimonial influe fortement sur l'accès au salariat. Le mariage est incontestablement un frein à l'accès au salariat : une fois mariées, les femmes voient divisées par quatre leurs chances d'accès à ce type d'emploi, par rapport au moment où elles étaient célibataires. Le divorce les place en

revanche au-dessus des célibataires en ce qui concerne l'accès au salariat : une période de divorce après le premier mariage multiplie leurs chances par plus de deux par rapport à une période de célibat, et par plus de sept s'il s'agit d'une période de divorce après un deuxième mariage. C'est dans l'analyse des interférences entre événements que réside surtout l'intérêt des biographies.

Conclusions

La convivialité de l'outil informatique d'une part, et l'amélioration des techniques statistiques d'autre part, ont permis la convergence des sciences sociales. La modélisation n'est plus accessible au seul statisticien, tandis que les systèmes de causalités souvent complexes des chercheurs peuvent maintenant être testés de manière fiable. Le reproche adressé à la technique statistique et au chiffre en général, de "réduire" la réalité, n'est plus valide : les raisonnements élaborés à l'aide de l'outil statistique rejoignent en complexité, mais aussi en finesse, les raisonnements induits à partir d'observations dites qualitatives.

Dans cet article, nous avons beaucoup insisté sur le temps. Le fil du temps est tout comme un fil d'Ariane qui relie les sciences sociales entre elles. Le temps est au coeur de l'analyse causale. L'analyse biographique, en contrôlant cette dimension essentielle, rapproche au plus près la statistique du raisonnement causal. Certes, les théories générales en sciences sociales ne peuvent être facilement réduites à de simples modèles, mais, au niveau des théories dites auxiliaires (spécifiques ou partielles), l'analyse biographique est appelée à jouer un rôle de plus en plus important en sciences sociales. Elle donne une impulsion supplémentaire au va-et-vient incessant entre la vérification des théories et la recherche de nouvelles explications.

Depuis deux ans un effort est entrepris en vue de mettre en place des mécanismes institutionnels qui rendent possible et facilitent les analyses comparatives de biographies. En effet, à l'instigation de quatre institutions, le CERPOD, l'IFAN, l'ORSTOM et le Département de démographie de l'Université de Montréal, un réseau sur le thème de "l'insertion urbaine en Afrique de l'Ouest" s'est mis en place en 1992. Il a reçu l'appui financier du Réseau Démographie de l'AUPELF-UREF, du CEPED et d'autres bailleurs de fonds. Le premier objectif de ce réseau est de favoriser l'analyse comparative des processus d'insertion urbaine en Afrique à partir d'enquêtes biographiques. La comparaison des études biographiques menées à Dakar et à Bamako constitue une étape (les résultats ont été présentés lors du séminaire d'octobre). Le second objectif du réseau est d'assurer à ses membres une bonne maîtrise des techniques d'analyse des données d'enquêtes biographiques. A cet effet, un stage regroupant vingt-cinq participants a été organisé en avril 1993 au CERPOD, à Bamako. Une seconde réunion du réseau a été organisée en octobre 1994. Elle rassemblait des participants d'une dizaine d'institutions africaines et a permis

d'arrêter des objectifs et une problématique communs aux équipes intéressées. Ce réseau dénommé RIVAS (Réseau insertion dans les villes d'Afrique subsaharienne) sera coordonné par le CERPOD.

Bibliographie

ANTOINE Ph., BOCQUIER Ph., FALL A.S., GUISSE Y.M., 1992 : Etude de l'insertion urbaine des migrants à Dakar. Présentation de la méthodologie d'enquête, *in : La ville en mouvement : Habitat et Habitants,* édité par E. LELIEVRE et C. LEVY-VROELANT, L'Harmattan, Paris, pp. 247-257.

ANTOINE Ph., DJIRE M., LAPLANTE B., 1994: Les déterminants socio-économiques de la sortie du célibat à Dakar. A paraître dans *Population.*

ANTOINE Ph., NANITELAMIO J., 1994 : Peut-on échapper à la polygamie à Dakar ? A paraître dans *Courtyards, Markets, City Streets : Urban Women in Africa.* ed by K. Sheldon, Westview Press, 35 p.

ANTOINE Ph., PICHE V., 1994 : Les jeunes vivent la crise, leurs aînés la supportent. L'insertion urbaine à Bamako et Dakar. *Pop Sahel,* CERPOD, n°21.

BOCQUIER Ph., 1992 : *L'insertion et la mobilité professionnelles à Dakar,* Thèse de Doctorat (nouveau régime) en Démographie, Université Paris V-René Descartes-Sorbonne.

BOCQUIER Ph., NANITELAMIO J., 1993 : Les déterminants familiaux de l'activité professionnelle des femmes de Dakar (Sénégal), Actes du séminaire de l'UIESP sur les *Femmes et les changements démographiques en Afrique au Sud du Sahara,* Dakar, Sénégal, 3-6 mars 1993, 24 p.

BOCQUIER Ph., 1995 : *Manuel d'analyse des biographies pour la micro-informatique,* préface de Daniel COURGEAU, à paraître aux Editions de l'ORSTOM.

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