« L'analyse exploratoire de trajectoires professionnelles : analyse harmonique qualitative ou appariement optimal ? »

Nicolas ROBETTE
Nicolas THIBAULT

(Draft version, please do not quote)

Résumé
Les enquêtes biographiques permettent d'analyser un grand nombre de carrières professionnelles individuelles dans leur intégralité. Diverses méthodes statistiques ont été développées pour mesurer les durées de séjour dans un état considéré en fonction de caractéristiques individuelles. Jusqu'aux années 1990, le traitement exploratoire des données avec l'objectif de décrire les parcours dans leur complexité n'avait fait l'objet que de peu d'attention dans la littérature. L'analyse harmonique qualitative et l'appariement optimal sont deux méthodes exploratoires qui permettent de dresser des typologies de parcours individuels complexes en prenant en compte la séquence des événements et leur durée. On les utilise ici pour reconstituer des typologies de carrières professionnelles des hommes de l'enquête *Biographies et entourage* (INED, 2001) afin de comparer l'intérêt respectif de chacune de ces techniques[1].

L'analyse démographique des biographies permet notamment d'estimer des durées de séjour[2] dans un état considéré par régression : on utilise classiquement des modèles non-paramétriques (Kaplan et Meier, 1958), semi-paramétriques (Cox, 1972) ou paramétriques[3] qui permettent de mesurer l'impact de caractéristiques individuelles (qui n'en sont pas moins des propriétés sociales) sur la durée de séjour dans un état. On se représente alors les biographies comme des processus stochastiques complexes (Courgeau et Lelièvre, 1989), c'est-à-dire des ensembles d'événements[4] liés entre eux par des lois de probabilité. Ces modèles sont pertinents dès lors qu'on mesure une durée bien définie par une date de début et une date de fin non problématiques. Mais ils n'ont pas pour finalité de décrire des parcours individuels caractérisés par une chronologie complexe de changements d'états c'est-à-dire lorsqu'on étudie une succession d'événements renouvelables et qu'il existe de multiples transitions possibles entre les états au cours de la trajectoire (GRAB, 2006), comme c'est par exemple le cas des parcours professionnels. Se pose alors la question de l'exploration de ce type de données.

Notre travail cherche à déterminer la façon de traiter ces trajectoires que nous qualifierons de complexes au sens où tous les états qui les caractérisent peuvent se reproduire au cours du temps et où il existe de multiples transitions[5] possibles entre ces états. Une réponse conforme au paradigme de l'analyse des biographies (Courgeau et Lelièvre, 1996) est d'utiliser des méthodes de statistique exploratoire pour dégager des typologies de parcours, qui appréhendent les trajectoires individuelles dans leur globalité et non plus simplement sous l'angle des événements qui les composent (Billari, 2001). Différentes méthodes typologiques existent (Grelet, 2002) – distance du khi-2, distance euclidienne (Espinasse, 1993), indicateurs synthétiques, etc. Deux d'entre elles émergent dans la littérature comme particulièrement adaptées à ce type de données c'est-à-dire capables de décrire de manière

[1] Nous tenons à remercier Valérie Golaz, Maryse Marpsat et Thibaut de Saint-Pol pour leur relecture de la première version de cet article. La présente version doit beaucoup à leurs commentaires. Nous remercions Vincent Cardon pour sa relecture de la présente version.
[2] Aussi appelé "survie".
[3] Les modèles de durée paramétriques supposent que le séjour dans un état considéré suit une loi, qui est elle-même fonction du temps.
[4] Qui peuvent toucher ego ou son entourage (LELIÈVRE, BONVALET, BRY, 1997)
[5] Par exemple, dans notre travail, il existe 9 états, il y a donc 81 transitions possibles.

systématique la séquence des événements et leur durée. D'une part, l'analyse harmonique qualitative (AHQ) est une technique d'analyse de données qui tient compte du temps, mise au point par des statisticiens français au tournant des années 1980. D'autre part, la méthode d'appariement optimal[6] (MAO) est une technique algorithmique importée des sciences de la vie par des sociologues américains dans la seconde moitié des années 1980. Nous nous proposons ici de discuter l'intérêt comparé de ces deux techniques à partir de l'exemple des histoires professionnelles des hommes dans l'enquête *Biographies et entourage* (INED, 2001).

1. La nécessité d'une analyse exploratoire des parcours professionnels

1.1 Le problème des événements renouvelables

Les méthodes longitudinales modernes, qui traitent des parcours individuels, ont déjà montré leur intérêt dans l'étude des parcours sur le marché du travail. À partir du fichier historique de l'ANPE, on peut par exemple étudier les parcours de réinsertion des chômeurs sur le marché du travail en fonction de caractéristiques individuelles (Degenne et Lebeaux, 1999) et en fonction des prestations et des mesures dont chaque chômeur a pu bénéficier (Crépon, Gurgand et Dejemeppe, 2005). L'utilisation de modèles de durée s'impose puisque on estime le temps qui sépare l'enregistrement dans les fichiers de l'ANPE du retour à l'emploi. Cette définition de la durée du chômage peut permettre de mesurer l'efficacité des politiques d'emploi. Néanmoins, elle néglige le fait que ce retour à l'emploi peut n'être que temporaire et aboutir rapidement à une nouvelle période de chômage. Les modèles de durée ne peuvent pas, en effet, prendre en compte la réversibilité des événements[7].

De la même façon, lorsqu'on analyse les interruptions de carrières professionnelles, on définit généralement la durée étudiée comme la différence entre le début de la N^{ième} période d'inactivité et la reprise d'activité suivante. Mais lorsqu'on s'intéresse à la complexité des parcours caractérisés par des intermittences d'activité, l'analyse d'un événement unique n'est pas suffisante car les individus peuvent connaître plusieurs interruptions de carrière. Divers aménagements ont pu être proposés pour traiter des épisodes répétés. D'une part, on peut différencier les sorties d'activité par rang. La distinction par rang de naissance est une méthode traditionnelle pour l'étude de la fécondité, mais au contraire des enfants, les sorties d'activité n'ont pas un sens particulier selon leur rang, c'est plutôt leur longueur voire leur longueur cumulée qui a une influence sur une carrière professionnelle (Desplanques et Saboulin, 1986 ; Lelièvre, 1987 ; Cambois et Lelièvre, 1988). *In fine*, cette méthode n'est pas pleinement satisfaisante car elle ramène la durée des étapes d'inactivité à la première définition (la différence entre le début de la N^{ième} période d'inactivité et la reprise d'activité suivante). D'autre part on peut raisonner sur l'ensemble des périodes en considérant les personnes qui ont connu une ou plusieurs périodes d'inactivité comme un niveau d'agrégation dans un modèle multi-niveaux (Courgeau, 2000). Mais ce genre d'analyse est encore difficile à mener actuellement.

1.2 Le problème de la multiplicité des états et des transitions

Le problème de définition de la durée se pose avec encore plus d'acuité lorsque l'état considéré ne se mesure pas de manière binaire mais peut se caractériser par une multiplicité de situations : différentes catégories socioprofessionnelles, temps plein, temps partiel, chômage, inactivité …

Les démographes cherchent par exemple à mesurer la durée de l'activité et à la mettre en relation avec d'autres événements, comme la naissance des enfants. Ils posent la question classique de l'interaction entre activité, notamment féminine, et fécondité. Les méthodes

[6] En anglais, O*ptimal Matching Analysis.*

[7] La solution adoptée par GURGAND, CREPON et DEJEMEPPE (2005) est de considérer qu'on n'est pas réellement sorti du chômage en dessous d'un certain seuil de temps en emploi.

biographiques semblent plus à même de la traiter que l'analyse par cohorte qui nécessite le recours à l'hypothèse d'indépendance entre événements. Toutefois, elles se heurtent selon M. Kempeneers et É. Lelièvre (1991) à trois problèmes imbriqués : définition de l'état dont on estime la durée et des variables exogènes, définition de la population soumise au risque, définition de l'intervalle d'étude[8]. Les choix de définitions hypothèquent la solidité du raisonnement et rendent les conclusions tributaires d'hypothèses de départ, souvent implicites. Les auteures recommandent donc une analyse plus descriptive qui rendrait compte de toutes les activités depuis l'âge de 15 ans.

1.3 Des données sur des trajectoires complexes

Cela les conduit à requérir la constitution de nouvelles sources permettant, non seulement de reconstituer les dates de sortie d'activité, mais de suivre les biographies dans leur intégralité sur le modèle de l'enquête *Triple biographie* (INED, 1981). Les enquêtes biographiques collectent de manière rétrospective les trajectoires complètes d'individus, le plus souvent année par année. Ainsi, ces données appellent une description statistique des parcours de vie.

L'enquête *Biographies et entourage* (INED, 2001), sur laquelle nous nous appuyons ici, est constituée de 2 830 individus des générations 1930 à 1950, représentatifs de la population francilienne des générations considérées à la date d'enquête (Lelièvre et Vivier, 2001). Elle a pour objectif principal d'aider à la compréhension de la mobilité résidentielle et familiale des enquêtés en interaction avec celle de leur entourage. Toutefois, le volet professionnel contient des données suffisamment intéressantes pour mériter une exploitation pour lui-même (Thibault, 2008). Les différentes professions occupées par chaque enquêté au cours de sa vie sont relevées selon un calendrier rétrospectif de dimension annuelle. Chaque étape a ensuite été codée selon la nomenclature des PCS de l'INSEE[9]. Si les histoires professionnelles peuvent en théorie se répartir sur tout le champ des possibles, elles n'en dessinent pas moins de grandes tendances qui sont liées aux mutations de la société (Marchand et Thélot, 1997).

2. Construire une typologie de trajectoires professionnelles complexes

2.1 Les trajectoires étudiées

L'unité statistique sur laquelle porte la classification est la trajectoire individuelle : les classes sont constituées en fonction de la ressemblance entre les parcours. Un certain nombre de choix sont nécessaires: (i) la population d'étude, (ii) l'intervalle d'étude, (iii) la construction des variables de l'analyse.

2.1.1 La population d'étude

Nous n'analyserons ici que les carrières masculines (n=1 341). Une classification sur l'intégralité de la population est possible mais elle a deux inconvénients. D'une part, la liste

[8] Les auteures traitent de l'interaction entre vie familiale et vie professionnelle des femmes. D'une part, l'analyse suppose une relation univoque : faut-il alors considérer que l'activité a une influence sur la fécondité ou au contraire que la fécondité a une influence sur l'activité ? Le choix de cette problématique de départ n'est aucunement neutre, et les conclusions en dépendent nécessairement. D'autre part, le choix de l'échantillon d'étude oblige aussi à poser un certain nombre d'hypothèses fortes. Doit-on n'étudier que les femmes qui ont eu des enfants ? De la même façon, doit-on limiter l'échantillon aux femmes qui sont entrées sur le marché du travail ou considérer que le fait de ne pas y entrer peut relever d'un projet anticipé de fécondité ? Enfin, le choix de l'événement initial de l'étude est aussi déterminant. Si on retient le projet d'une étude sur la durée d'activité avant interruption, plusieurs événements initiaux sont également légitimes selon la question que l'on se pose ; ils n'en sont pas moins problématiques du fait de la complexité des biographies. Le choix du moment de l'entrée sur le marché du travail comme événement initial néglige les femmes qui ont eu un enfant avant de travailler. Le mariage (ou le début de l'union) pose un problème du même type puisque les enfants peuvent naître avant le début de l'union considérée.

[9] Pour une présentation plus complète des données professionnelles, notamment sur la façon dont les principes de la nomenclature des PCS s'appliquent, on se rapportera à Thibault (2008).

des états est différente pour les hommes et les femmes : le service militaire n'a d'influence que sur les itinéraires masculins. D'autre part, lorsqu'on traite ensemble des deux sexes, les différences dans le rythme de carrière entre hommes et femmes peuvent parfois être masquées par la simplification impliquée par la classification[10].

2.1.2 La période d'étude

Les individus sortent d'observation à la date d'enquête : les données sont donc tronquées à droite[11]. Les modèles de durée sont aptes à contrôler l'effet des troncatures mais pas les statistiques descriptives qui nécessitent de raisonner sur la même population au cours du temps[12]. La description de trajectoires individuelles suppose donc que tous les individus soient observés sur une période identique, délimitée par les mêmes bornes. Notre étude portera sur la mobilité professionnelle entre l'âge de 14 ans, qui marque la fin de la scolarité obligatoire pour les générations étudiées[13], et celui de 50 ans, qui est l'âge des enquêtés les plus jeunes[14]. On aurait pu continuer l'analyse au-delà de 50 ans en travaillant sur une sous-population de l'enquête. Toutefois, cette procédure nous est apparue fallacieuse ; en effet, le nombre d'enquêtés décroît rapidement avec l'âge : à 55 ans, on ne raisonne plus que sur 65,3 % de la population, 40,7 % à 60 ans, 20,7 % à 65 ans et 4,8 % à 70 ans.

2.1.3 La variable d'état

Nous décrivons les différents états constituant les trajectoires à partir des groupes socioprofessionnels c'est-à-dire la nomenclature des PCS en huit postes. Un découpage trop fin des états n'est pas souhaitable car passer d'employé administratif à employé de commerce n'a pas le même sens que de passer cadre. Par construction, l'échantillon ne contient pas de retraité puisque la description s'arrête à 50 ans. On aurait donc dû raisonner sur sept groupes socioprofessionnels, les six groupes d'actifs et les « autres personnes sans activité professionnelle ». Toutefois, il nous est apparu intéressant de scinder cette dernière catégorie entre les étudiants et ceux qui sont sans activité pour des raisons familiales ou de santé[15]. On ajoute un dernier groupe issu de la nomenclature de 1954 mais aujourd'hui supprimé, les militaires du contingent : il est nécessaire pour restituer les parcours masculins pour ces générations qui ont notamment connu la Guerre d'Algérie.

2.2 Constitution d'une typologie par analyse harmonique qualitative

2.2.1 L'analyse harmonique qualitative

---

[10] Une description des parcours à partir de toute la population a pourtant l'avantage de montrer les proximités entre certaines carrières masculines et féminines. Mais cette objection vient plutôt renforcer notre parti pris de traiter séparément les hommes et les femmes. Si les carrières sont ressemblantes on dégagera des types proches, comme c'est le cas chez les cadres. Si les itinéraires sont dissemblables, on évite de lisser les disparités liées au genre.

[11] On dit qu'elles sont "censurées" à droite. Par construction des enquêtes, qui sont rétrospectives, la date d'enquête est la seule modalité de censure à droite : les individus décédés ou ayant migrés ne sont pas observés. Le relevé a été suffisamment exhaustif pour qu'il n'y ait pas de censure à gauche, c'est-à-dire des individus dont la biographie ne serait connue qu'à partir d'une date donnée.

[12] Le fait que certaines enquêtes portent sur des échantillons représentatifs de la population adulte dans son ensemble comme *Mobilités spatiales dans l'aire métropolitaine de Bogota* (ORSTOM, Université des Andes, 1993) et non sur certaines générations rend les résultats des méthodes typologiques difficiles d'interprétation (BARBARY et PINZON SARMIENTO, 1998).

[13] Certains enquêtés ont connu des étapes professionnelles dès l'âge de huit ans, par exemple parce qu'ils étaient bergers dans une ferme et ne se rendaient qu'épisodiquement à l'école, nous négligerons ces étapes dans l'analyse car la consigne donnée à l'enquêté faisait commencer la datation à 14 ans. Ce choix est cohérent avec les principes de la nomenclature des PCS qui ne s'applique qu'aux personnes de plus de 15 ans.

[14] On raisonne en âge atteint (en différence d'années) et non en âge exact. Nous connaissons l'âge exact au moment de l'enquête mais les événements biographiques sont connus en âge atteint.

[15] La catégorie inclut aussi les chômeurs n'ayant jamais travaillé (qui sont actifs au sens de l'INSEE), mais ce cas peut être négligé puisqu'il ne concerne que trois étapes sur les 19 930 étapes de l'enquête. Ce qui est bien sûr la conséquence du caractère annuel des données.

L'analyse harmonique est une branche des mathématiques qui a connu de nombreuses applications en sciences physiques ou en biologie. Son utilisation dans les sciences sociales est plus récente et date des années 1970 (Deville, 1974, 1977). Il s'agissait alors d'introduire la durée dans l'explication des phénomènes sociaux grâce à des données sur les histoires individuelles. « Devant des données d'une telle richesse le statisticien éprouve une certaine perplexité. Des tableaux de plus en plus complexes deviennent ininterprétables sans le secours de méthode d'analyse "automatique". Il cherche alors à définir une méthode d'analyse qui lui permette de tirer l'essentiel des données dont il dispose. Le mot "essentiel" prend alors un sens précis, quantifiable, lié à la méthode qu'il met en œuvre. » (Deville, 1977). Cette technique a été ensuite adaptée pour en faire une technique de statistique exploratoire des trajectoires complexes (Deville et Saporta, 1980 ; Deville, 1982) appelée analyse harmonique qualitative.

L'analyse harmonique qualitative consiste à déterminer une période d'observation, à la diviser en un nombre fini d'intervalles puis à mesurer pour chaque individu la proportion du temps passé dans chacun des états dans chaque intervalle. Une analyse factorielle des correspondances sur la matrice[16] ainsi constituée permet de résumer l'information en sélectionnant les facteurs portant le plus d'inertie (Deville, 1982). On élimine de cette façon le "bruit" statistique sans éliminer d'individu. Dans les années 1990, les progrès des méthodes de classification permettent d'utiliser par exemple une classification ascendante hiérarchique à partir des facteurs pour dresser des typologies de parcours (Barbary, 1996 ; Degenne, Lebeaux et Mounier, 1996 ; Barbary et Pinzon Sarmiento, 1998). La typologie ainsi construite tient compte d'une part de la succession des états, du moment auquel les états interviennent mais aussi des durées passées dans les différents états et permet d'autre part de conserver tous les individus dans l'analyse.

L'analyse harmonique qualitative a été jusqu'à maintenant peu utilisée. La raison en est certainement le manque de données adaptées. Ainsi, J-C. Deville (1982) constitue un échantillon *ad hoc* à partir des questionnaires éliminés lors des enquêtes sur les familles de l'INSEE de 1962 et 1975[17]. L'échantillon n'est par conséquent aucunement représentatif et n'a pour ambition que de fournir un terrain d'application à une méthode qui n'en avait pas. Sur plus de vingt ans, la méthode n'est utilisée qu'une fois sur des données françaises (Degenne, Lebeaux et Mounier, 1996). Ces travaux portent sur une partie du parcours professionnel pour interroger l'insertion sur le marché du travail. Finalement, il faut attendre le renouveau des collectes biographiques pour voir des applications se concrétiser, d'abord sur des données latino-américaines (Dureau, Barbary, Florez et Hoyos, 1994 ; Barbary, 1997 ; Barbary et Pinzon Sarmiento, 1998) puis plus récemment françaises puisque l'enquête *Biographies et entourage* va nous permettre de suivre les parcours migratoires (Bonvalet, Bringé et Robette, 2008) mais aussi l'intégralité des carrières professionnelles[18].

2.2.2 La constitution de la matrice harmonique

Une fois la période d'étude fixée, on doit la découper en intervalles pour l'analyse. Les données étant de dimension annuelle, le découpage en périodes d'un an peut sembler le plus naturel et le plus apte à conserver l'information biographique. Toutefois, une telle procédure

[16] Avec en lignes les individus et en colonnes le temps passé dans à chacun des états pendant chaque intervalle.

[17] L'échantillon est constitué d'un certain nombre de femmes mariées trois fois ou plus dont l'histoire conjugale avait été considérée comme trop complexe pour être traitée.

[18] L'enquête *Carrière et mobilité* (INSEE, 1989) procède à un relevé biographique incomplet : on demande la profession occupée à sept dates différentes. Elle permet donc de reconstituer partiellement les carrières mais elle ne permet pas de suivre les individus tout au long de leur parcours professionnel. Elle a donné lieu à un traitement par reconstitution d'indicateurs transversaux : par exemple comparant le premier emploi à l'emploi actuel. On a pu aussi dresser une typologie des parcours par le principe des nuées dynamiques (GOUX, 1991). Cette procédure de classification n'est applicable que parce qu'on ne dispose que des professions occupées (ou de l'inactivité) à des dates données : le problème du traitement de la durée dans la typologie ne se pose donc pas.

n'est pas optimale car elle engendre un tableau dont la grande majorité des cases sont nulles, ce qui hypothèque la qualité de l'analyse[19]. A l'opposé, un nombre trop réduit d'intervalles entraînerait la perte d'une partie de la richesse de l'information disponible. Il y a donc un arbitrage à effectuer pour établir le nombre des intervalles.

Un autre arbitrage concerne l'amplitude des intervalles. Rien n'oblige à ce que les amplitudes soient égales ; bien au contraire, certains moments de la vie, le plus souvent la jeunesse, sont caractérisés par un nombre important de changements professionnels, d'autres par une mobilité plus faible. Nous avons choisi de découper en dix intervalles correspondant aux déciles de la distribution des changements d'étapes d'activité en fonction de l'âge[20].

Pour chaque individu, on calcule la proportion de la durée de chaque intervalle passée dans chacun des états possibles[21]. Nous réalisons une analyse factorielle de la matrice obtenue, puis une classification ascendante hiérarchique à partir des 25 premiers facteurs qui portent 70 % de l'inertie. Cela permet de réduire l'hétérogénéité des données tout en conservant une part essentielle de l'information[22].

2.3 Constitution d'une typologie par appariement optimal

2.3.1 L'appariement optimal

La Méthode d'Appariement Optimal s'appuie sur un ensemble d'algorithmes dynamiques utilisés principalement par la biologie moléculaire pour analyser les similarités entre chaînes d'ADN. Elle a ensuite été introduite dans les sciences sociales par Andrew Abbott dans les années 1980 (Abbott et Forrest, 1986 ; Abbott et Hrycak, 1990). Son principe est fondé sur la mesure de la similarité ou de la dissimilarité entre des paires de séquences. L'idée de base consiste à mesurer la dissimilarité entre deux séquences en évaluant le coût représenté par la transformation de l'une des séquences en l'autre. La transformation est effectuée au moyen de trois opérations élémentaires : l'insertion (un élément est inséré dans la séquence), la suppression (un élément est supprimé de la séquence) et la substitution (un élément est substitué à un autre). On peut assigner un coût spécifique à chacune de ces opérations élémentaires. Une série d'opérations a un coût équivalent à la somme des coûts des opérations élémentaires. La distance entre deux séquences est alors définie comme le coût minimal de la transformation d'une séquence en l'autre. Des algorithmes dynamiques spécifiques garantissent l'obtention du coût minimal (Sankoff et Kruskal, 1983). L'appariement de l'ensemble des paires de séquences permet la création d'une matrice de distances, que l'on peut ensuite utiliser pour regrouper les séquences les plus similaires, au moyen de méthodes de classification par exemple, et obtenir une typologie.

Le choix des coûts des opérations élémentaires constitue une étape essentielle des techniques d'Appariement Optimal. C'est la possibilité de détermination des coûts qui confère à la méthode sa souplesse et sa capacité à s'adapter à l'objet étudié (Lesnard et Saint-Pol, 2004). Dans la pratique, l'appariement optimal ne repose que sur deux opérations : l'insertion-suppression, appelée *indel* par contraction des termes anglais *insertion* et *deletion*, et la substitution.

---

[19] On observe 1 341 individus pouvant connaître 9 états sur p périodes. La matrice étudiée aura 1 341 lignes. Si on choisi une période annuelle entre 14 et 50 ans, la matrice aura 9x37=333 colonnes. Par construction des données, l'individu ne connaît qu'un état par an : chaque ligne comporte donc 37 uns et 296 zéros. Sur les 1 341x333=446 553 cellules du tableau, seul un neuvième seront non nulles.

[20] On ne peut évidemment choisir que les valeurs annuelles les plus proches des déciles théoriques. Les déciles sont alors 18, 20, 22, 24, 26, 29, 33, 38 et 43 ans.

[21] On a dix intervalles et neuf états, on crée donc quatre-vingt dix variables d'état. La matrice finale comporte 1 341 lignes et 90 colonnes.

[22] Les tests effectués montrent que les classes de la typologie obtenue sont relativement stables lorsque l'on conserve entre 50 et 90 % de l'inertie. Peu d'individus changent de classes quand on modifie la part de l'information retenue.

2.3.2 La constitution de la matrice de coûts

Il est possible de différencier les coûts de substitution selon la combinaison des éléments substitués. Si certains chercheurs préfèrent adopter des coûts de substitution fixes du fait du manque de théorie sur le sujet (Dijkstra et Taris, 1995), de nombreux travaux adoptent des coûts de substitution différenciés selon des hypothèses propres à l'objet étudié : plus les éléments sont similaires, plus le coût de substitution est faible. Ainsi dans le cas de travaux sur les carrières professionnelles, les coûts de substitution sont fixés en fonction des positions relatives des catégories socioprofessionnelles au sein d'une hiérarchie de ces catégories (Stovel et al, 1996; Halpin et Chan, 1998 ; Blair-Loy, 1999 ; Scherer, 2001 ; Solis et Billari, 2002). Une solution alternative consiste à dériver les coûts de substitution des probabilités de transition entre les éléments : le coût de substitution entre deux éléments est d'autant plus élevé que la probabilité de transition entre ces éléments est faible (Rohwer et Pötter, 2005). Dans la mesure où il n'existe pas de hiérarchie a priori entre les états observés, nous avons adopté la seconde solution (Annexe 1).

On s'intéresse ensuite à la relation entre coût de substitution et coût *indel*. Une substitution étant équivalente à la combinaison d'une insertion et d'une suppression, on peut fixer le coût *indel* à la moitié du coût de substitution. Avec des coûts de substitution variables, un coût *indel* supérieur à la moitié du coût de substitution maximal évite l'utilisation des insertions-suppressions excepté pour tenir compte de la différence de longueur des séquences. Cette approche se justifie dans le cas où l'on privilégie la position des éléments au sein de la séquence. En revanche, si l'on privilégie l'ordre des épisodes, un coût *indel* égal à 1/10 du coût de substitution maximal est plus approprié (Macindoe et Abbott, 2004). Dans un itinéraire de mobilité sociale courant de 14 à 50 ans, le moment auquel interviennent les différents états est fondamental. Nous avons donc privilégié la première option (Annexe 1).

Encadré 1 : représentation des trajectoires en MAO et en AHQ

Pour l'exemple, prenons 7 ans de la carrière de Jean-Louis, de 14 à 20 ans : de 14 à 15 ans il est étudiant puis de 16 à 18 ans il est ouvrier, enfin de 19 à 20 ans il cadre. Pour simplifier l'exposé nous nous limiterons à ces trois états (E, O et C). Son parcours peut-être représenté sous les deux formes différentes.

- En AHQ, si on découpe en deux intervalles, un de quatre ans (14-17 ans) et un de trois ans (18-20 ans), on obtient la matrice :

14-17 ans			18-20 ans		
E	O	C	E	O	C
0,5	0,5	0	0	0,33	0,67

- En MAO, on obtient la séquence : EEOOOCC dans laquelle chaque lettre représente l'état occupé annuellement.

3 Comparaison des deux techniques sur les itinéraires professionnels

3.1 Des typologies proches

En travaillant sur l'intégralité des trajectoires professionnelles masculines, nous avons retenu deux partitions en six et dix classes de parcours qui semblent un bon compromis entre les exigences de synthèse de l'information et celle de présentation de l'hétérogénéité des trajectoires individuelles. Les deux techniques donnent des résultats proches, ce qu'on peut montrer à l'aide de leur matrice de correspondance qui croise les répartitions de la population selon deux partitions distinctes. On peut ainsi voir si les individus d'une classe issue de la première partition sont plutôt concentrés dans une seule classe de la seconde partition ou s'ils

sont au contraire ventilés entre plusieurs classes. Ici, nous présentons en ligne la classification obtenue par appariement optimal et en colonne celle résultant de l'analyse harmonique qualitative. Les matrices de correspondance en effectifs pour les partitions en six et dix classes sont reproduites en annexes 2 et 4. A titre d'exemple, nous présentons, le tableau des proportions en ligne qui en résulte pour une partition en six classes.

	En %	AHQ						
		1	2	3	4	5	6	Total
MAO	1	88,6	4,8	5,1	1,4	0,0	0,0	100,0
	2	1,1	80,4	12,4	4,4	1,4	0,3	100,0
	3	0,9	1,7	83,8	9,4	0,0	4,3	100,0
	4	0,0	2,0	1,4	90,9	3,7	2,0	100,0
	5	12,3	17,8	2,7	0,0	67,1	0,0	100,0
	6	26,7	68,6	2,3	1,2	1,2	0,0	100,0

Source : Biographies et entourage (INED, 2001)
Champ : 1341 carrières professionnelles d'hommes des générations 1930-1950 résidant en Île-de-France à la date d'enquête

Figure 1 : Correspondance entre la typologie obtenue par appariement optimal et celle obtenue par analyse harmonique qualitative

La lecture intuitive de la diagonale du tableau montre une forte proximité entre les classifications obtenues par les deux méthodes. Il est possible de calculer le taux de correspondance, indice synthétisant le degré de similitude entre les deux partitions. Formellement (1), avec n_{ij} l'effectif appartenant simultanément à la $i^{ème}$ classe de la première partition et à la $j^{ème}$ classe de la seconde partition et N la population totale, on fait la somme de l'effectif du mode de chacune des lignes et de chacune des colonnes qu'on rapporte au double de l'effectif total. On calcule ainsi la moyenne de la correspondance de la classification obtenue par MAO à celle obtenue par AHQ et de la classification obtenue par AHQ à celle obtenue par MAO.

$$(1)\ \sigma = \frac{\sum_i \max_j(n_{ij}) + \sum_j \max_i(n_{ij})}{2N}$$

La correspondance entre les typologies obtenues selon les deux méthodes est de 82% dans le cas d'une partition à 6 classes et de 75% dans le cas d'une partition à 10 classes. Les deux techniques donnent donc des résultats grandement convergents, ce qui est une bonne indication de leur solidité.

2.2 Les nuances apportées par chacune des méthodes

Toutefois, quelques nuances apparaissent, ce qui nous amène à formuler des hypothèses issues de nos observations sur l'intérêt comparé de chacune des méthodes. La comparaison des deux typologies en six et en dix classes révèle quelques différences. Sans entrer dans le détail des classes constituées qui ne nous intéresse pas ici et qui est présenté en annexes 3 et 5, nous pouvons noter que la non concordance entre les deux typologies correspond à deux cas.

D'une part, les classes obtenues par MAO sont plus sensibles aux transitions, notamment dans le début de la trajectoire où les changements d'états professionnels sont les plus nombreux.

Les classes distinguent ainsi particulièrement les parcours stables des parcours de mobilité ; elles sont de ce point de vue plus homogènes qu'en utilisant l'AHQ. Cela vient du fait que la MAO retient la séquence année par année alors que l'analyse harmonique qualitative est fondée sur la mesure de la proportion de chaque intervalle de temps passée dans chaque état.
D'autre part, l'analyse harmonique qualitative fait apparaître quelques classes d'effectifs plus réduits regroupant des trajectoires très typées c'est-à-dire différant fortement des autres par la nature et la séquence des états sur tout le parcours. On observe par exemple une classe regroupant des individus ayant connu des épisodes courts et diffus en interruption d'activité. Au contraire, la partition obtenue par MAO tend à agréger ces trajectoires avec d'autres avec lesquelles elles possèdent une ou plusieurs transitions en commun.

2.3 Hypothèses sur l'intérêt comparé de chacune des méthodes

D'une part, la possibilité que laisse l'analyse harmonique qualitative de ne retenir pour la classification que les principaux facteurs issus de l'analyse factorielle permet d'éliminer une partie du « bruit », autrement dit de faire ressortir les caractéristique les plus structurantes de l'information ce qui en facilite l'interprétation. Mais l'analyste n'est pas à même de contrôler quelle partie de l'information statistique est ainsi négligée. D'autre part, le découpage de la période d'étude en intervalles, d'amplitudes non nécessairement égales, permet d'insister sur les moments des trajectoires dans lesquels se concentrent les événements qui intéressent l'analyste. Cela peut notamment se révéler pertinent dans le cas des parcours professionnels, pour lesquels la majorité des événements se concentrent avant 30 ans. Ainsi, l'analyse harmonique qualitative permet d'insister particulièrement sur la période d'apparition d'un événement et sur la durée des états. Les professions occupées le plus longtemps ont notamment une influence plus forte sur la typologie que dans l'appariement optimal, ce qui permet d'insister sur la stabilité socioprofessionnelle qui reste aujourd'hui forte (Goux, 1991).
Au contraire, l'analyse par appariement optimal conserve l'intégralité du détail de la séquence au lieu de la simplifier (Lesnard et Saint-Pol, 2004). C'est alors à l'analyste de fixer lui-même les "coûts" en fonction de ses hypothèses théoriques, ce qui détermine ensuite le rapprochement entre certaines trajectoires plutôt qu'entre certaines autres. La fixation du coût *indel* permet notamment de faire porter l'intérêt plutôt sur le type de transition ou plutôt sur le moment d'apparition d'un événement. Comparée à l'analyse harmonique qualitative, l'analyse par appariement optimal insiste sur la séquence des événements et le type de transition qui apparaissent dans celle-ci. Par conséquent, elle donne plus de poids aux transitions d'une catégorie socioprofessionnelle à une autre, ce qui permet de mettre en évidence des parcours de mobilité intragénérationnelle.

Conclusion

Les enquêtes biographiques doivent permettre d'étudier statistiquement les parcours de vie. Pourtant, même les promoteurs de ces enquêtes avouent que les résultats sont souvent inférieurs aux espérances. Ainsi, dans un livre consacré aux collectes biographiques quantitatives publié par le groupe de réflexion sur l'approche biographique, nous pouvons lire que « […] nous sommes en deçà de l'utilisation potentielle de telles données. On peut dire que l'effort énorme qui a été développé pour produire ces données et acquérir les compétences nécessaires pour les traiter de façon adéquate, pouvait et devait produire davantage, avoir une application et une diffusion très large » (GRAB, 1999, p 46, citation de M. Bottai). En effet, les questionnaires sont assez longs ce qui rend la collecte coûteuse. Les échantillons sont nécessairement limités à quelques milliers d'individus. L'information est par conséquent difficile à traiter car elle retrace une multiplicité d'événements pour un petit nombre d'individus.

Les méthodes typologiques d'exploration de trajectoire permettent de pallier cette difficulté en fournissant un outil de traitement permettant de restituer quantitativement la logique des parcours individuels. Elles sont ainsi adaptées à ces sources de données prometteuses pour les démographes et plus généralement pour les sciences sociales. Encore faut-il se repérer parmi les différentes méthodes possibles, encore peu connues et dont les avantages comparés n'avaient pas été testés. Il ressort de notre travail que l'analyse harmonique qualitative semble plus adaptée si on cherche à insister sur la durée dans certaines étapes et donc à mettre l'accent sur les états dans lesquels l'individu reste le plus longtemps. Au contraire, l'analyse par appariement optimal a un avantage si on cherche à rapprocher les trajectoires sur la base du type de transition et insister ainsi sur les parcours de mobilité.

Bibliographie :

Abbott A., Forrest J., 1986, « Optimal Matching Methods for Historical Sequences », *Journal of Interdisciplinary History,* 16(3), p. 471-494.

Abbott A., Hrycak A., 1990, « Measuring ressemblance in sequence data : an optimal matching analysis of musicians' careers », *American journal of sociology*, (96), p. 144-185.

Barbary O., 1996, *Analisis tipologico de datos biograficos en Bogota*, Bogota, Universidad Nacional de Colombia, 254 p.

Barbary O., 1997, « Analisis estadistico de datos biograficos : metodos, ejemplos y perspectivas en el estudio de itinerarios migratorios » in J. A. Bustamante, D. Delaunay, J. Santibanez, *Medicion de la migracion internacional*, Tijuana, Documento de trabajo del Colegio de la Frontera Norte

Barbary O., Pinzon Sarmiento L.M., 1998, « L'analyse harmonique qualitative et son application à la typologie des itinéraires individuelles », *Mathématiques informatiques et sciences humaines*, n°144, p. 29-54

Billari F., 2001, « Sequence analysis in demographic research », *Canadian Studies in Population*, 28(2), p. 439-458 (Dossier Special Issue on Longitudinal Methodology).

Blair-Loy M., 1999, « Career patterns of executive women in finance: an optimal matching analysis », *American Journal of Sociology*, 104(5), p. 1346-1397.

Bonvalet C., Bringé A., Robette N., 2008, « Les itinéraires géographiques des Franciliens depuis leur départ de chez les parents », in *Famille et entourage dans la société urbaine*, C. Bonvalet et É. Lelièvre (eds), Paris, PUF/INED, à paraître

Cambois M.A., Lelièvre É, 1988, « Durée d'activité et interruption de carrière des femmes âgées de 45 ans à 64 ans en 1981 », *Population*, 3, p. 669-675

Courgeau D., 2000, « Vers une analyse biographique multiniveau », communication aux Journées de méthodologie statistique, INSEE, 4-5 décembre

Courgeau D., Lelièvre É., 1989, *Analyse démographique des biographies*, Paris, PUF/INED, 268 p.

Courgeau D., Lelièvre E., 1996, « Changement de paradigme en démographie », *Population,* n°3, vol. 51, pp 645-654

Cox D. R ., 1972, « Regression models and life tables (with discussion) », *Journal of royal statistical society*, B34, p. 187-220

Crépon B, Gurgand M., Dejemeppe M., 2005, « Counseling the unemployed : does it lower unemployment duration and recurrence ? », *Document de travail*, Centre d'Étude sur l'Emploi, n°40

Degenne A., Lebeaux M.-O., 1999, *Étude sur les sorties du chômage, comparaison jeunes et adultes*, Caen, LASMAS, Rapport pour le Commissariat Général du Plan

Degenne A., Lebeaux M.-O., Mounier L., 1996, « Typologies d'itinéraires comme instrument d'analyse du marché du travail », in A. Degenne, M. Mansuy, G. Podevin, P. Werquin

(eds.), *Typologie des marchés du travail, suivi et parcours*, Marseille, Document du CÉREQ n°115, p. 27-42

Desplanques G., Saboulin M. de, 1986, « Activité féminine : carrières continues et discontinues », *Économie et statistiques*, n°192-193, p 51-62

Deville J-C., 1974, « Méthodes statistiques et numériques de l'analyse harmonique », *Annales de l'INSEE*, n°15, p. 3-101

Deville J-C., 1977, « Analyse harmonique du calendrier de constitution des familles en France », *Population*, n°1, p. 17-63

Deville J-C., 1982, « Analyse de données chronologiques qualitatives : comment analyser des calendriers ? », *Annales de l'INSEE*, n°45, p. 45-104

Deville J-C., Saporta G., 1980, « Analyse harmonique qualitative », in *Data analysis and informatics*, E. Diday (éd.), Amsterdam, North Holland Publishing, p. 375-389

Dijkstra W., Taris T., 1995, « Measuring the agreement between sequences », *Sociological methods & research,* (24), p. 214-231.

Dureau F., Barbary, O., Elisa Florez C., Hoyos M. C., 1994, *La observacion de las diferentes formas de movilidad : propuestas metodologicas experimentadas en la encuesta de movilidad espacial en el area metropolitana de Bogota,* Paris, ORTSOM, Atelier du CEDE (Montevideo) du 27-29 octobre 1993 : « Nuevas modalidades y tendencias de la migracion entre paises fronterizos y los processos de integracion », 31 p.

Espinasse J.-M., 1993, « Enquêtes de cheminement, chronogrammes et classification automatique », Note du LHIRE, 19(159).

Goux D., 1991, « Coup de frein sur les carrières », *Économie et statistiques*, n°249, p. 75-87

GRAB, 1999, *Biographies d'enquête. Bilan de quatorze collectes biographiques*, Paris, PUF/INED, 340 p.

GRAB, 2006, *États flous et trajectoires complexes. Observation, modélisation, interprétation*, Paris, PUF/INED, 301 p.

Grelet Y., 2002, « Des typologies de parcours. Méthodes et usages », Document Génération 92, (20), 47 p.

Halpin B., Chan T. W., 1998, « Class careers as sequences: an optimal matching analysis of work-life histories », *European Sociological Review*, 14(2), p. 111-130.

Kaplan E., Meier P., 1958, « Nonparametric estimation from incomplete observations », *Journal of american statistical association,* vol. 53, pp. 457-481

Kempeneers M., Lelièvre É., 1991, « Analyse biographique du travail féminin », *Revue européenne de démographie*, 7, p.377-400

Lelièvre É., 1987, « Activité professionnelle et fécondité : les choix et les déterminations chez les femmes françaises, de 1930 à 1960 », *Cahiers québécois de démographie*, 16, p.209-236

Lelièvre É., Bonvalet C., Bry X., 1997, « Analyse démographique des groupes : les avancées d'une recherche en cours », *Population*, n°4, n°spécial : *Nouvelles approches méthodologiques en démographie*, p. 803-830

Lelièvre É., Vivier G., 2001, « Évaluation d'une collecte à la croisée du quantitatif et du qualitatif », *Population*, 56, p. 1043-1074

Lesnard L., Saint-Pol T. de, 2004, « Introduction aux méthodes d'appariement optimal (Optimal Matching Analysis) », *Document de travail INSEE,* (15), 30 p.

Macindoe H., Abbott A., 2004, « Sequence analysis and optimal matching techniques for social science data », in Hardy Melissa, Bryman Alan*, Handbook of Data Analysis*, London, Sage, p. 387-406.

Marchand O., Thélot C., 1997, *Le Travail en France. 1800-2000*, Paris, Nathan, 269 p.

Rohwer G., Pötter U., 2005, « TDA's user manual », *http://www.stat.ruhr-uni-bochum.de/tman.html.*

Sankoff D., Kruskal J., (dir), 1983, *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison*, Reading, Addison-Wesley, 408 p.
Scherer S., 2001, « Early career patterns: a comparison of Great Britain and West Germany », *European Sociological Review*, 17(2), p. 119-144.
Solis P., Billari F., 2002, « Work lives amid social change and continuity: occupational trajectories in Monterrey, Mexico », *Max Planck IDR Working paper*, 2002-009, 52 p.
Stovel K., Savage M., Bearman P., 1996, « Ascription into achievement: models of career systems at Lloyds Bank, 1890-1970 », *American Journal of Sociology*, 102(2), p. 358-399.
Thibault N., 2008, « La mobilité sociale, une construction biographique : l'exemple des enfants d'indépendantes », in *Famille et entourage dans la société urbaine*, C. Bonvalet et E. Lelièvre (eds), Paris, PUF/INED, à paraître

Annexes

Annexe 1 : matrice des coûts de substitution et coût *indel* de l'appariement optimal

	agri	*acce*	*cadre*	*pi*	*empl*	*ouvr*	*serv mil*	*inact*	*etu*
agri	0,000	1,992	2,000	2,000	1,990	1,895	1,999	2,000	1,954
acce	1,992	0,000	1,990	1,986	1,987	1,978	1,997	1,987	1,991
cadre	2,000	1,990	0,000	1,971	1,990	1,997	1,972	1,964	1,912
pi	2,000	1,986	1,971	0,000	1,961	1,976	1,966	1,948	1,853
empl	1,990	1,987	1,990	1,961	0,000	1,970	1,972	1,962	1,896
ouvr	1,895	1,978	1,997	1,976	1,970	0,000	1,947	1,905	1,782
serv mil	1,999	1,997	1,972	1,966	1,972	1,947	0,000	1,986	1,947
inact	2,000	1,987	1,964	1,948	1,962	1,905	1,986	0,000	1,980
etu	1,954	1,991	1,912	1,853	1,896	1,782	1,947	1,980	0,000

indel=1,01

Annexe 2 : matrice de correspondance pour les typologies en six classes

Effectifs		AHQ						
		1	2	3	4	5	6	Total
MAO	1	311	17	18	5	0	0	*351*
	2	4	291	45	16	5	1	*362*
	3	1	2	98	11	0	5	*117*
	4	0	7	5	320	13	7	*352*
	5	9	13	2	0	49	0	*73*
	6	23	59	2	1	1	0	*86*
	Total	*348*	*389*	*170*	*353*	*68*	*13*	*1341*

Source : Biographies et entourage (INED, 2001)
Champ : 1341 carrières professionnelles d'hommes des générations 1930-1950 résidant en Île-de-France à la date d'enquête

Annexe 3 : caractéristiques des trajectoires des parangons de chaque classe des typologies en six classes

Le parangon d'une classe est l'individu le plus proche du centre de la classe.

Classe	MAO	AHQ

1	longtemps cadre	longtemps cadre
2	longtemps profession intermédiaire	longtemps profession intermédiaire
3	longtemps employé	longtemps employé
4	longtemps ouvrier	longtemps ouvrier
5	longtemps patron	longtemps patron
6	profession intermédiaire devient cadre	agriculteur devient ouvrier

Annexe 4 : matrice de correspondance pour les typologies en dix classes

Effectifs		AHQ										
		1	2	3	4	5	6	7	8	9	10	Total
MAO	1	304	17	18	0	9	0	0	0	0	3	*351*
	2	4	187	0	0	15	0	0	0	0	2	*208*
	3	1	2	83	0	0	0	0	0	0	1	*87*
	4	0	0	1	231	0	0	0	5	20	1	*258*
	5	0	8	0	0	72	1	1	1	1	0	*84*
	6	1	10	0	0	0	41	21	0	0	0	*73*
	7	0	6	4	27	24	24	6	0	3	0	*94*
	8	0	0	13	5	3	0	0	4	2	3	*30*
	9	22	51	2	0	9	2	0	0	0	0	*86*
	10	0	18	45	0	6	0	1	0	0	0	*70*
	Total	*332*	*299*	*166*	*263*	*138*	*68*	*29*	*10*	*26*	*10*	*1341*

Source : Biographies et entourage (INED, 2001)
Champ : 1341 carrières professionnelles d'hommes des générations 1930-1950 résidant en Île-de-France à la date d'enquête

Annexe 5 : caractéristiques des trajectoires des parangons de chaque classe des typologies en dix classes

Classe	MAO	AHQ
1	longtemps cadre	longtemps cadre
2	longtemps profession intermédiaire	longtemps profession intermédiaire
3	longtemps employé	longtemps employé
4	longtemps ouvrier	longtemps ouvrier
5	ouvrier devient profession intermédiaire	ouvrier devient profession intermédiaire
6	devient patron	devient patron
7	ouvrier devient profession intermédiaire ou patron	commence patron
8	ouvrier devient employé	agriculteur devient ouvrier avant 25 ans
9	profession intermédiaire devient cadre	agriculteur devient ouvrier après 25 ans
10	employé devient profession intermédiaire	interruption d'activité