Explorer et décrire les parcours de vie

Les typologies de trajectoires

Explorer et décrire les parcours de vie

Les typologies de trajectoires

Nicolas ROBETTE

INED

Paris **2011**

Le Bureau d'Appui à la Recherche (BAR) suscite des groupes de travail rassemblant des chercheurs du Nord et du Sud autour de questions émergentes. Ceux-ci développent un programme d'activité annuel ou pluri-annuel, le plus souvent sous forme d'ateliers. Ces activités sont valorisées par des publications et une diffusion large des résultats. Les sujets sont thématiques, géographiques ou méthodologiques, en particulier le développement, la mise en œuvre et l'évaluation d'outils de collecte et d'analyse.

Les *Collections du CEPED* comportent trois séries :

– la série « les Clefs pour »

– la série « Regards sur »

– la série « Les numériques du CEPED »

Ces publications permettent une diffusion rapide et validée des résultats de recherche ainsi qu'une meilleure connaissance des nouvelles méthodes, techniques et concepts en matière de *Population et développement*.

Les chercheurs, et particulièrement les chercheurs du Sud, qui y ont un accès privilégié, y trouvent une validation scientifique et une bonne dissémination de leurs travaux.

Rédactrice en chef : Éva Lelièvre

Assistante de rédaction : Yvonne Lafitte

Directeur de la publication : Yves Charbit

Responsable du BAR : William Molmy

Maquette de couverture : Christine Tichit

Photo de couverture : © IRD – Laure Emperaire

Conception graphique : sbgraphik – www.sbgraphik.com

ISSN : 1777-4551 – ISBN : 978-2-87762-184-7

CEPED

UMR 196 Université Paris Descartes-INED-IRD

19, rue Jacob – 75006 Paris – France

Tél. : 33 (0)1 78 94 98 70 – Fax : 33 (0)1 78 94 78 79

Courriel : contact@ceped.org

Web : http://www.ceped.org

LES CLEFS POUR …

La série « *les Clefs pour* » des Collections du CEPED se donne pour objectif de faire partager l'expérience, de ménager les échanges en assurant la diffusion des méthodes et des concepts. Cette série se présente sous forme de petits manuels qui n'ont pas pour ambition de faire le tour de la question mais plutôt de proposer soit une introduction, soit un manuel pratique permettant de se familiariser avec le sujet présenté et d'accéder aux publications plus élaborées le cas échéant.

Répondre à une demande concrète

Les chercheurs, les praticiens et les étudiants qui travaillent sur les questions de population, sont confrontés à des contraintes spécifiques de terrain, à l'adaptation d'outils et de concepts, au manque de données qui rend nécessaire une valorisation de données existantes ou des collectes spécifiques et à la nécessité d'innover et d'utiliser au mieux les avancées développées dans des contextes divers.
Ces contraintes appellent des solutions précises, des innovations méthodologiques et des discussions conceptuelles dont la diffusion permet l'avancée de la recherche.

Faire circuler concepts et méthodes

Du point de vue des concepts, les échanges entre les chercheurs travaillant sur des terrains du Sud comme du Nord sont primordiaux. La nécessaire adaptation et parfois la critique de concepts historiquement ancrés ailleurs permet de cerner les particularités et les ressemblances faisant progresser les termes de la comparaison. En effet, l'ethnocentrisme conceptuel génère des catégories de collecte et d'analyse qui peuvent entrer en contradiction avec les catégories de pensée des populations, ou masquer la complexité de l'organisation sociale.

Les méthodes et outils de **collecte** (questionnaires, modes d'observation) provenant de systèmes d'observation standardisés « universels » sont souvent inadaptés. Cela se traduit par des imprécisions, voire une mésinterprétation réciproque des questions et des réponses, et pervertit les données. Pour éviter ces écueils, des outils sont expérimentés dans le recueil de l'activité, de la composition familiale des ménages, de la complexité résidentielle… Il s'agit de favoriser la diffusion des expériences et expertises méthodologiques multilatérales et interdisciplinaires.

Du point de vue de l'**analyse** s'appuyant sur la présentation d'outils développés au Sud comme au Nord, la série « *les Clefs pour* » vise à diffuser des méthodes nouvelles ou qui ont fait leurs preuves dans d'autres disciplines que la démographie, ménageant ainsi les transferts interdisciplinaires.

Table des matières

Introduction

L'étude des parcours de vie, démarche descriptive ou causale ?

Au cours des dernières décennies, l'évolution de la mobilité résidentielle, des structures familiales, des études et des parcours professionnels, ainsi que l'importance de l'interdépendance entre ces différentes sphères de la vie, a suscité un intérêt croissant pour les trajectoires biographiques. Progressivement, l'analyse des parcours de vie (*Life Course Analysis*) est devenue une perspective majeure des sciences sociales, entraînant un passage de la structure au processus, du macro au micro, de l'analyse à la synthèse, du certain à l'incertain ("*from structure to process, from macro to micro, from analysis to synthesis, from certainty to uncertainty*", Willekens, 1999 : 26). Le développement de cette perspective est lié simultanément à des questions théoriques et aux progrès des techniques de collecte et d'analyse statistique des données longitudinales. Du point de vue de la collecte, les sources de données longitudinales se sont multipliées, sous forme de panels ou d'enquêtes biographiques (GRAB, 1999). Du point de vue méthodologique, le développement et la diffusion de nouvelles techniques statistiques d'analyse des parcours de vie ont été lents mais cumulatifs et le corpus des méthodes disponibles est maintenant très substantiel. Depuis le début des années 1980, l'approche centrale dans l'analyse des données longitudinales en sciences sociales est l'analyse biographique ou *Event History Analysis* (Kalbfleisch et Prentice, 1980 ; Allison, 1984 ; Courgeau et Lelièvre, 1986 et 1989 ; Mayer et Tuma, 1990). Ces techniques, comme le célèbre modèle de Cox (1972), généralisent les tables de survie. Des modèles économétriques toujours plus perfectionnés (Wu, 2003) ouvrent la voie à la prise en compte des interactions, de l'hétérogénéité inobservée et des biais de sélection (Lillard, 1993 ; Heckman *et al.*, 1998), ou de différents niveaux d'agrégation (Courgeau et Baccaïni, 1997 ; Steele, 2008).

L'un des principaux attraits de l'analyse des biographies tient sans doute au fait que la réflexion sur la causalité qui y est associée correspond relativement bien à la manière dont le déroulement des événements au cours du temps est ressenti par les individus (Halpin, 2003) : en combien de temps peut-on espérer trouver du travail ? Quel serait l'effet d'une année d'étude supplémentaire ?... On retrouve cette homologie avec les modèles de régression en général mais l'analyse biographique présente l'avantage d'intégrer le temps de manière réaliste. Les modèles de durée, construits dans un cadre statistique probabiliste, permettent de se soustraire à un déterminisme caricatural

(Courgeau et Lelièvre, 1989). Il devient possible d'étudier les interdépendances entre des trajectoires individuelles parallèles, comme la carrière professionnelle et la nuptialité (Courgeau et Lelièvre, 1986), ou entre des individus potentiellement liés, comme c'est le cas des couples, ainsi que d'intégrer différents niveaux d'agrégation – du micro au macro – à l'analyse des comportements individuels (Blossfeld et Rohwer, 2002). L'analyse biographique est centrée sur l'occurrence (ou la non-occurrence) d'événements spécifiques du parcours de vie. Elle modélise les probabilités de transition ou de durée, en faisant l'hypothèse que le parcours de vie est le résultat d'un processus stochastique complexe. C'est donc une approche paramétrique, à vocation explicative, causale.

L'analyse biographique se concentre sur les événements. Pourtant, l'ambition des recherches sur les parcours de vie est aussi d'appréhender les trajectoires dans leur ensemble. En effet, la théorie souligne depuis les origines l'importance de la trajectoire en tant que concept théorique (Sackmann et Wingens, 2003) : les événements ne doivent pas être étudiés indépendamment les uns des autres, mais dans leur enchaînement. Dans la pratique, une grande partie des travaux empiriques sur les parcours de vie en sciences sociales sont basés sur des méthodes centrées sur les transitions. Il est cependant possible d'adopter une autre approche, offrant la possibilité d'étudier les parcours de vie en tant qu'unité d'analyse, comme un tout. Billari (2001 : 440) identifie deux raisons d'adopter un point de vue qu'il qualifie de « holiste », qui appréhende la trajectoire dans son ensemble comme unité conceptuelle ("*a 'holistic' perspective that sees life courses as one meaningful conceptual unit*"). La première, dite « forte », fait l'hypothèse que les parcours sont le résultat des projets de vie des individus, qui souhaitent par exemple maximiser leur utilité. Les individus adopteraient eux-mêmes une perspective holiste lorsqu'ils « planifient » leur vie future : ils considèrent leur parcours de vie à venir dans sa globalité. Le second point de vue, dit « pragmatique », est basé sur l'idée que le parcours de vie en tant qu'unité conceptuelle est le résultat contingent d'une séquence d'événements. Une approche holiste des parcours doit alors permettre de décrire et de résumer le calendrier et la séquence des événements, ainsi que la durée passée dans les différents états et celle séparant les différents événements (Settersten et Mayer, 1997). L'importance de l'ordre des événements a d'ailleurs été soulignée de longue date (Hogan, 1978 ; Marini, 1984 ; Rindfuss *et al.*, 1987) : par exemple, la probabilité de se marier et ses déterminants varient selon que le mariage intervient avant ou après la fin des études. De plus, de nombreux événements impliquent une transition réversible (le mariage en est encore un exemple), d'autres peuvent se révéler difficiles à définir de manière précise et appartiennent à un « temps flou » (GRAB, 2006). Le départ de chez les parents apparaît ainsi plus comme un processus complexe que comme une transition univoque (Villeneuve-Gokalp, 1997 ; Diagne, 2006 ; Diagne et Lessault, 2007).

L'ensemble de ces remarques plaident pour l'adoption d'une approche holiste des parcours de vie, afin d'explorer les données longitudinales individuelles, d'y « découvrir des structures cachées » (Roux, 1993) et d'en réduire la complexité en la synthétisant, en particulier au moyen de typologies. Contrairement au cas de l'analyse biographique, les méthodes adoptant cette approche sont le plus souvent non paramétriques : elles ne font pas d'hypothèse stochastique sur la genèse des parcours

de vie et appartiennent de ce fait à la culture dite « algorithmique » des statistiques (Breiman, 2001). Leur objectif consiste principalement à décrire et explorer les parcours de vie, à en identifier les régularités ou les différences.

En pratique, les méthodes d'exploration des parcours sont nombreuses. Elles peuvent s'appliquer à des terrains et des thématiques variées : de l'insertion urbaine en Afrique de l'Ouest à la mobilité résidentielle en Amérique du Sud, en passant par les histoires matrimoniales et génésiques, le passage à l'âge adulte, les trajectoires foncières ou de santé.

Nous nous concentrerons essentiellement dans ce manuel sur la démarche la plus répandue, la construction de typologies de trajectoires. Nous décrirons tout d'abord l'ensemble de la démarche et les choix successifs qu'elle nécessite, du codage des données au choix final du nombre de classes de la typologie. Nous nous intéresserons ensuite à la caractérisation des résultats, ainsi qu'à quelques approches complémentaires[1], puis nous passerons en revue les méthodes de mesure de la similarité entre trajectoires, qui constitue une étape cruciale pour obtenir une typologie. Enfin, la construction de typologies de trajectoires sera illustrée par une application à des parcours d'activité. Si la mise en œuvre des méthodes présentées n'est pas immédiate, celles-ci sont en réalité plus accessibles que les apparences ne pourraient le laisser craindre et se révèlent d'une utilité précieuse, notamment dans les cas où la richesse des données et/ou la complexité des trajectoires étudiées ne permettent pas de classer ces trajectoires « à la main » ou selon des critères simples.

[1] Cette revue des méthodes ne prétend pas à l'exhaustivité mais entend offrir au lecteur intéressé par leur utilisation un large panorama des possibilités existantes et (plus ou moins) accessibles.

1. Réaliser une typologie de trajectoires

1.1 Les données

Les typologies sont particulièrement appropriées lorsque l'on dispose de données d'enquête permettant de reconstituer, au moins partiellement, les trajectoires des individus enquêtés. Dans certains cas, seuls quelques événements de vie (premier mariage, premier enfant...) sont collectés et l'on analysera des trajectoires simplifiées. Mais de nombreuses enquêtes longitudinales enregistrent dans le détail les trajectoires individuelles, faites de transitions, d'allers et de retours dans des situations variées. Ces enquêtes peuvent être rétrospectives – c'est par exemple le cas des enquêtes biographiques telles que celles sur l'insertion urbaine en Afrique (Antoine *et al.*, 2006) ou la santé vécue et perçue de malades du Sida en Thaïlande (Lelièvre et Lecœur, 2010) – ou prospectives (panels, systèmes de suivi démographiques voir Bringé et Laurent, 2005).

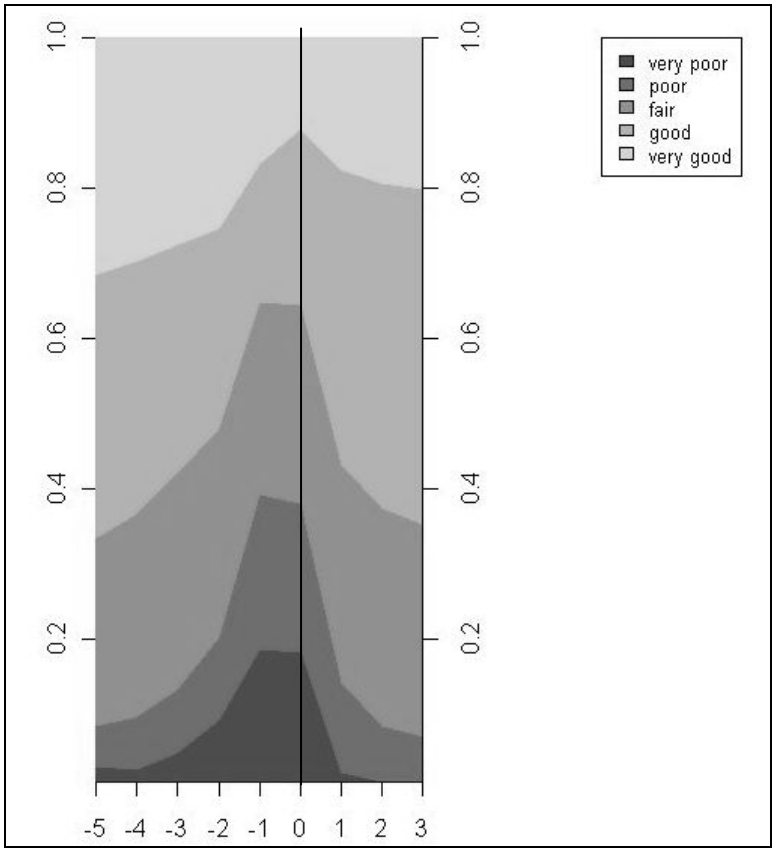
1.2 Choix d'une population

Une première étape dans la construction d'une typologie de trajectoires consiste à circonscrire la population étudiée. Si elle dépend avant tout de la question de recherche, un aspect technique doit tout de même être pris en compte. En effet, l'analyse simultanée de plusieurs sous-populations dont les types de parcours sont nettement différenciés présente le risque de masquer dans les résultats les régularités propres aux parcours de chacune des sous-populations. Par exemple, admettons que l'on souhaite analyser le calendrier de constitution de la famille (mise en couple, naissance des enfants…) dans le but d'identifier les parcours caractérisés par des transitions précoces et ceux dont les événements interviennent plus tardivement. Si dans la population étudiée, les femmes vivent ces transitions plus tôt que les hommes, l'analyse conjointe des femmes et des hommes rendra floue la distinction entre trajectoires précoces et tardives. Il serait alors plus efficace d'analyser les deux populations séparément.

1.3 Choix d'une période d'observation

Le choix d'une période d'observation implique celui d'un début et d'une fin. Dans le cas de parcours de vie, ce sont le plus souvent des âges, par exemple des parcours d'activité entre 14 et 65 ans. Mais on peut tout à fait imaginer borner la période entre deux dates, entre deux événements, depuis un événement fondateur (par exemple, la trajectoire qui suit la fin des études) ou *n* années avant un autre (la trajectoire qui précède le passage à la retraite), voire « encadrer » un événement (l'inflexion d'une trajectoire observée avant un événement marquant et après). La figure 1 présente la trajectoire de santé perçue déclarée par les personnes touchées par le VIH ; on a choisi ici de prendre en compte les séquences des huit années encadrant la mise sous traitement antirétroviral (cinq ans avant l'initiation du traitement, trois ans après)[2].

Figure 1 − Distribution de la santé perçue autour de la mise sous traitement antirétroviral de personnes infectées par le VIH en Thaïlande



Données : Enquête LIWA (*LIving With Antiretrovirals*, 2007)
Lecture : La mise sous traitement correspond à l'année 0.
On constate une forte inflexion de la santé perçue à la faveur du traitement.

[2] Je remercie les responsables de ce projet financé par l'Agence Nationale de Recherches sur le Sida et les Hépatites Virales (ANRS – 12 141), S. Le Cœur et E. Lelièvre, pour la reproduction de ces résultats.

1.3.1 Longueur des séquences

La période d'observation peut être de même longueur pour l'ensemble des individus ou non. Cette question reste toutefois relativement problématique. Tout d'abord, et comme discuté dans la troisième partie concernant *l'Optimal Matching*, la différence de longueur des séquences n'a pas le même sens selon qu'elle est liée à la nature du processus ou à la collecte des données. En effet, une séquence (trouver un emploi, par exemple) n'a aucune raison d'être de même durée pour l'ensemble des individus alors que si l'on enquête la trajectoire résidentielle des 18-35 ans, les parcours seront de longueurs différentes mais ne résulteront pas de la durée différentielle du processus en jeu mais juste de l'observation qui sera tronquée, censurée, pour une partie de l'échantillon.

Dans le premier cas, le fait que la transition entre le système scolaire et le marché du travail soit plus ou moins longue, selon les personnes, constitue une information sur les transitions. Il est alors intéressant que cette information soit prise en compte dans le regroupement des trajectoires individuelles, c'est-à-dire que les jeunes « à trajectoire courte » soient considérés comme relativement similaires – et relativement distincts des jeunes « à trajectoire longue » – lors de la construction de la typologie de trajectoires. Le problème est tout autre lorsque les différences de longueur sont la conséquence du mode de collecte des données, c'est-à-dire lorsque l'on est en présence de censure. Par exemple, si une enquête auprès de personnes entre 30 et 60 ans reconstitue les parcours d'activité à partir de l'âge de 14 ans, la longueur des parcours variera entre 16 et 46 ans, pour des raisons externes au processus observé. Lors de la construction d'une typologie, il sera difficile de distinguer ce qui, dans la composition des classes, relève d'une différence « réelle » des parcours de ce qui est lié à l'hétérogénéité de leurs longueurs.

D'autre part, toutes les mesures de dissemblance entre trajectoires (voir chapitre 3) ne permettent pas de prendre en compte des longueurs différentes : la distance de Hamming ou celle basée sur un codage disjonctif-complet sont basées sur la simultanéité et, à moins d'avoir recours à des artifices de codage (ajout d'états « non observé »…), ne sont pas compatibles avec des longueurs variables. Mais avant tout, il faut se poser la question du sens que porterait la similarité entre trajectoires de longueurs différentes, en relation avec l'objet de la recherche : des trajectoires composées du même enchaînement d'événements mais se déroulant sur des durées très variables doivent-elles être considérées comme ressemblantes ?

1.4 Choix des états

D'une manière générale, il est préférable de débuter en définissant un espace des situations possibles relativement restreint, c'est-à-dire de coder les parcours à partir d'un nombre d'états limité. Dans le cas contraire, les résultats risquent d'être moins robustes et plus difficiles à interpréter. Toutefois, certaines analyses reposent sur un nombre d'états élevé, notamment lorsque l'on traite de parcours multidimensionnels et

que l'on souhaite utiliser des états combinant les diverses dimensions : par exemple l'état matrimonial, le nombre d'enfants et le statut d'activité. Dans cette optique, certains travaux ont montré qu'une fois surmontée la question de leur interprétation, les résultats se révélaient relativement robustes (Robette, 2010).

Il n'existe pas de conduite systématique en présence de non-réponses. Il est par exemple possible d'imputer les données manquantes ou de créer un état supplémentaire « valeur manquante ». Si les non-réponses sont peu nombreuses et ne sont pas liées à un type particulier de trajectoires, cela n'a que peu d'incidence sur les résultats. Lorsqu'à l'inverse les non-réponses sont nombreuses, on court le risque de voir émerger des classes d'individus qui ont principalement en commun l'absence d'information sur leur trajectoire. Une autre possibilité consiste à supprimer les éléments manquants, ce qui raccourcit d'autant les séquences et ramène à la question des séquences de longueurs différentes. Mais comme pour la plupart des autres choix, il est conseillé de faire différents essais pour en comparer les résultats et en évaluer la robustesse.

1.5 Choix d'une mesure de dissemblance et codage

Le choix d'une distance, c'est-à-dire d'une mesure de dissemblance (ou dissimilarité) entre les trajectoires peut s'opérer parmi l'ensemble des méthodes qui seront décrites dans la troisième partie. Il peut lui-même impliquer d'autres arbitrages, comme le choix des coûts pour *l'Optimal Matching*.

Ces décisions ne sont pas neutres et doivent se faire au regard des spécificités des différentes méthodes, des données à analyser et de l'objectif de la recherche. Une méthode ne peut être considérée comme « meilleure » qu'une autre que dans la perspective de son application dans un contexte précis. En l'occurrence, l'une des qualités de ces méthodes (et, à ce titre, en particulier de *l'Optimal Matching*) est justement qu'elles rendent les choix explicites et imposent donc de s'interroger sur leur sens d'un point de vue théorique (Lesnard et de Saint Pol, 2009).

Le codage des trajectoires découle du choix de la mesure de dissemblance. Le plus souvent, on aura une variable par élément de la trajectoire : par exemple, une histoire familiale objectivée par l'observation annuelle du statut matrimonial des individus pendant 35 ans sera codée en 35 variables, la n-ième variable ayant pour valeur le statut matrimonial lors de la n-ième année d'observation. Mais d'autres types de codages seront présentés dans le chapitre 2.

1.6 Choix d'une méthode de classification

À partir de la matrice de distances entre trajectoires, obtenue à l'aide de la mesure de dissimilarité choisie, la dernière étape de la construction d'une typologie repose sur une procédure de classification, puis une partition. Cela permet de répartir la population en un nombre limité de groupes relativement homogènes et distincts les uns des autres, en identifiant ainsi un ensemble de « parcours-types ».

Les méthodes de classification sont nombreuses et appartiennent essentiellement à deux familles : les classifications hiérarchiques et les partitions autour des centres mobiles. Parmi les classifications hiérarchiques, la classification descendante divise pas à pas la population en groupes plus petits. À l'opposé, la classification ascendante hiérarchique regroupe de manière itérative les individus qui se ressemblent le plus, selon un critère de ressemblance (ou d'agrégation) prédéfini. Il existe de nombreux critères d'agrégation, comme l'indice de saut minimum (*single linkage*), l'indice de saut maximum (*complete linkage*), l'indice de saut entre centre de gravité (*centroid method*) ou l'indice de saut moyen (*average linkage*). Le plus communément employé est le critère de Ward qui, à chaque étape, cherche à minimiser l'hétérogénéité à l'intérieur des classes (inertie intra-classe), ce qui est équivalent à maximiser l'hétérogénéité entre les classes (inertie inter-classe). Par ailleurs, des analyses ont montré que les critères WPGMA flexible (*Flexible Weighted Pair Group using arithMetic Averages*) ou UPGMA flexible (*Flexible Unweighted Pair Group using arithMetic Averages*) étaient particulièrement efficaces sur des données empiriques en présence de bruit ou d'observations aberrantes (Milligan, 1981 ; Belbin *et al.*, 1992 ; Lesnard et de Saint Pol, 2009).

Les classifications hiérarchiques, ascendantes ou descendantes, aboutissent à un arbre de classification, appelé dendrogramme, dont chaque niveau correspond à une partition de l'ensemble des individus. C'est l'utilisateur qui fait le choix du nombre de classes de la typologie, en s'aidant éventuellement pour cela d'indices statistiques, comme le saut d'inertie (voir sous-partie suivante sur le choix du nombre de classes).

La classification autour des centres mobiles et ses variantes (méthode des *k-means*, nuées dynamiques) consistent à définir des noyaux, dont le nombre est défini par l'utilisateur, puis à agréger chaque individu au noyau dont il est le plus proche. On répète alors plusieurs fois l'opération en prenant pour noyaux les centres de gravité des classes de la partition obtenue. Les itérations s'arrêtent lorsque l'on obtient une partition stable. Le nombre de classes correspond au nombre de noyaux initialement choisi par l'utilisateur. Les avantages de cette famille de méthodes sont la rapidité de calcul et la possibilité de facilement détecter et éventuellement supprimer les individus atypiques (*outliers*). En revanche, les résultats dépendent du choix des noyaux initiaux. De plus, le fait de fixer *a priori* le nombre de classes limite l'exploration des données, alors que les classifications hiérarchiques permettent de choisir le nombre de classes, selon des critères statistiques ou propres à l'utilisateur, ou d'analyser aisément les résultats à différents niveaux de partition. Dans la pratique, les classifications hiérarchiques sont le plus souvent préférées. Cependant, elles sont parfois complétées

par des classifications autour des centres mobiles, soit en amont pour simplifier les données lorsque le nombre d'observations est important et engendre un temps de calcul trop élevé, soit en aval pour rendre les classes plus homogènes une fois leur nombre déterminé par l'utilisateur à l'issue de la classification hiérarchique.

Une possibilité, inspirée des *k-means*, consiste à définir *a priori* quelques trajectoires-types, puis à regrouper chaque trajectoire individuelle avec la trajectoire-type dont elle est la plus similaire (en se basant sur la matrice de distance calculée à l'étape précédente). Cette option est particulièrement intéressante lorsque l'on a avant l'analyse des hypothèses fortes sur les régularités existant parmi les parcours étudiés (Elzinga et Liefbroer, 2007).

Quelques travaux ont aussi utilisé les arbres de décision. Billari et Piccarreta (2005) ont ainsi introduit le *Monothetic Divisive Algorithm*. Le point de départ consiste à coder les trajectoires sous forme de variables binaires. Ces variables binaires représentent le fait qu'un individu, à un âge donné, a vécu ou non un événement donné (par exemple, avoir quitté ses études à 20 ans ou avoir eu un enfant à 25 ans ; dans l'article, l'analyse prend en compte six types d'événements et l'observation porte sur seize années : il y a donc 6*16 = 96 variables binaires). L'ensemble des trajectoires est ensuite divisé en deux groupes, selon la valeur de la variable binaire qui maximise l'homogénéité à l'intérieur des groupes et minimise l'hétérogénéité entre les groupes (*splitting variable*)[3]. Cette étape est ensuite répétée à partir des deux groupes, et ainsi de suite de manière itérative, jusqu'à produire un arbre de classification. Le principal atout de cette méthode est qu'elle rend visible ce qui a décidé des divisions, par l'intermédiaire des variables binaires. Les classes obtenues sont donc plus facilement interprétables. En revanche, cette approche est limitée à des trajectoires composées d'événements non-renouvelables. Depuis, les mêmes auteurs ont en partie assoupli cette limite, les données de départ étant la matrice de distance entre trajectoires, obtenue par exemple avec *l'Optimal Matching* (Piccarreta et Billari, 2007). Mais les variables au principe du découpage de l'arbre de classification restent binaires donc correspondant à des événements non-renouvelables.

On notera aussi le développement récent des méthodes de classification neuronale associées aux cartes de Kohonen (Cottrell et Ponthieux, 2002 ; Delaunay et Lelièvre, 2006 ; Giret et Rousset, 2007). Ces cartes, dites aussi d'auto-organisation, offrent une intéressante visualisation de la proximité entre classes, à la manière des projections d'une analyse factorielle.

Une approche appelée *Stage Latent Class Analysis* a aussi été utilisée pour obtenir une classification de parcours de vie (MacMillan et Eliason, 2003). Cette méthode est cependant paramétrique ; elle implique donc des hypothèses plus fortes que les autres techniques décrites dans ce manuel. Elle semble de plus relativement limitée en termes de nombre d'individus et de nombre d'observations dans le temps.

Au final, si les méthodes de classification sont diverses et qu'il est conseillé d'en essayer plusieurs pour s'assurer de la robustesse des résultats obtenus, on remarquera

[3] Cette méthode ne nécessite donc pas de calcul d'une matrice de distance.

que la classification ascendante hiérarchique (CAH) associée au critère de Ward est la plus largement utilisée – avec succès – dans les travaux réalisant des typologies de trajectoires.

1.7 Choix du nombre de classes de la typologie

Le choix du nombre de classes de la typologie de trajectoires fait l'objet de critiques récurrentes. Ce choix étant laissé à la discrétion du chercheur, il est considéré comme arbitraire et, de là, la méthode est supposée manquer de robustesse. Cette critique ne concerne d'ailleurs pas spécifiquement l'analyse exploratoire des parcours mais les procédures de classification automatique en général. Elle est malheureusement le résultat d'un malentendu quant à la nature de ces méthodes et émane souvent d'une perspective économétrique de l'analyse quantitative.

La classification est exploratoire et non-paramétrique : son objectif n'est pas la mesure précise d'un phénomène ou la quantification de l'effet d'une caractéristique sur une autre, mais l'identification de régularités, avec un minimum d'hypothèses sur les données. C'est dans cette flexibilité que réside sa pertinence et sa puissance analytique. De ce fait, vouloir déterminer le « meilleur » nombre de classes d'une typologie à partir de critères statistiques indépendamment de la question de recherche est un non-sens : « *les classifications produites ne peuvent être vraies ou fausses, ni même probables ou improbables ; elles ne peuvent être que fructueuses ou infructueuses* » (Williams et Lance, 1965)[4]. La création d'une taxinomie en sciences sociales devrait être guidée par des fondements théoriques, la portée heuristique des résultats et/ou un arbitrage entre parcimonie et homogénéité des classes.

Dans la pratique, il est fortement recommandé d'observer les typologies à différents niveaux de partition[5]. Il est souvent intéressant d'examiner la nature et l'homogénéité d'une classe par l'intermédiaire des sous-classes qui la composent. Un nombre de classes trop élevé sera difficile à interpréter et décrire ; en revanche, s'il est trop faible, l'hétérogénéité des classes (ou intra-classe) risque d'être importante et l'on ne pourra guère dégager de parcours-types. Le choix d'une typologie suppose donc un arbitrage, le critère principal étant que la typologie finalement sélectionnée soit cohérente et porteuse d'enseignements du point de vue de la recherche qui est menée[6].

Ceci dit, des indicateurs statistiques peuvent tout de même se révéler utiles, en particulier pour guider les premiers pas dans l'analyse des résultats. Autrement dit, puisque le nombre de partitions possibles à l'issue d'une classification est important,

---

[4] "*Classifications so produced can never be true or false, or even probable or improbable ; they can only be profitable or unprofitable*".

[5] Cela est toutefois impossible avec certaines méthodes, comme les classifications autour des centres mobiles par exemple.

[6] On peut, par exemple, souhaiter voir émerger un type particulier de trajectoires, même si celui-ci est rare et que son apparition en tant que classe de la typologie implique un nombre de classes relativement élevé.

un indicateur peut orienter vers la première partition à observer, avant d'explorer des typologies aux nombres moins et/ou plus élevés de classes. Ces indicateurs sont nombreux (sauts d'inertie, critère de Calinski-Harabasz, d'Hartigan…[7]), le plus souvent basés sur la comparaison des variances inter- et intra-classe : il s'agit de minimiser les variations à l'intérieur des classes et de maximiser la différence entre les classes.

[7] Pour une revue et un test empirique de ces critères, voir Milligan et Cooper (1985).

2. Décrire et représenter des trajectoires

L'hétérogénéité d'un échantillon de trajectoires individuelles est souvent très importante, ce qui en rend la description ardue, voire impossible. En ce sens, une typologie de trajectoires, construite à l'issue d'analyses factorielle ou séquentielle, constitue une représentation simplifiée de l'échantillon, donc plus aisée à décrire. Toutefois, chaque classe de la typologie enfermant elle-même une certaine hétérogénéité, la question de la caractérisation et de la présentation des classes d'une manière simple et précise n'est pas triviale (voir par exemple Halpin et Chan, 1998)[8].

2.1 Représentations graphiques

L'examen des représentations graphiques de la typologie obtenue offre bien souvent un moyen rapide et relativement intuitif d'interpréter les résultats. Il existe deux types principaux de représentations graphiques permettant l'observation synthétique des classes d'une typologie de trajectoires.

2.1.1 Chronogramme

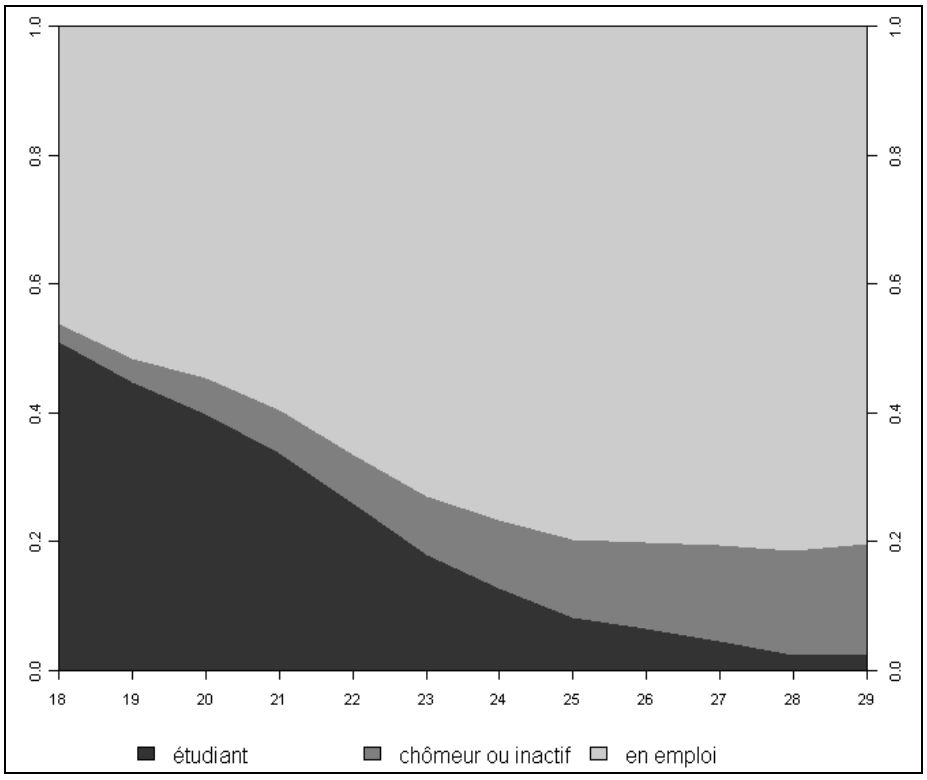
Le premier type de graphique, parfois appelé chronogramme (en anglais, *State Distribution Plot*), est constitué d'une succession de coupes instantanées, indiquant la distribution des individus de la classe entre les différents états à chaque instant d'observation. Autrement dit, à chaque moment de la trajectoire, le graphique présente les proportions cumulées d'individus dans chacune des situations[9]. À titre d'illustration, on construit la trajectoire de sortie du système scolaire d'un échantillon de 500 individus, tirés au hasard parmi les enquêtés de *Biographies et entourage* (voir présentation de l'enquête dans la partie 5). La trajectoire est observée annuellement de

[8] Les représentations graphiques et les indicateurs présentés dans cette partie ont aussi leur utilité avant même le début de la démarche typologique, afin d'avoir de premiers éléments de description des trajectoires pour l'ensemble de la population.

[9] Ces mêmes proportions peuvent être représentées de manières alternatives, sous forme de proportions cumulées avec, par exemple, des diagrammes en barres (histogrammes), ou non cumulées avec de simples courbes.

18 à 29 ans (inclus) et à chaque observation, les individus sont dans l'une des situations suivantes : étudiant, chômeur ou inactif, en emploi. Le graphique montre clairement qu'à 18 ans, les individus se répartissent à peu près à parts égales entre les études et l'emploi (figure 2). Puis la proportion d'étudiants chute, rapidement jusqu'à 25 ans puis plus faiblement, pour être proche de zéro à 29 ans. À la fin de la trajectoire, la proportion de chômeurs ou d'inactifs, qui augmente principalement après 23 ans, atteint presque 20 % alors que celle d'individus en emploi est de près de 80 %.

Figure 2 − Chronogramme de trajectoires d'insertion



Source : *Biographies et entourage* (2001) ; Champ : échantillon aléatoire de 500 individus.

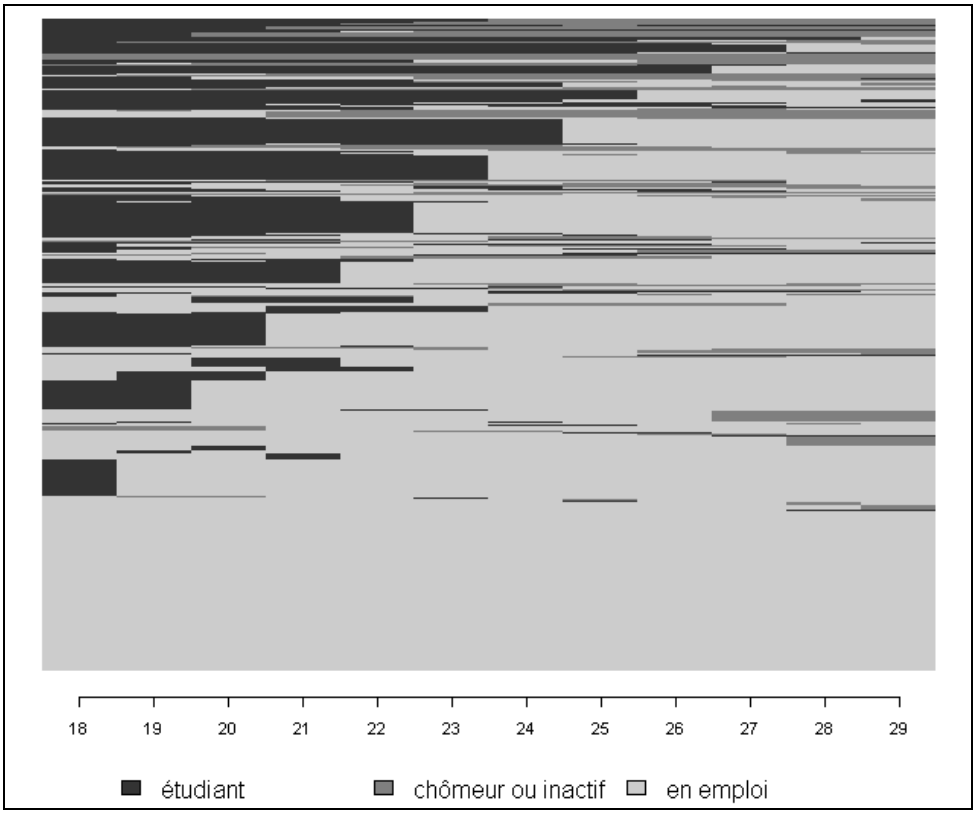
Le principal avantage de ce type de représentation graphique est qu'il est d'une lecture aisée et favorise la comparaison entre les classes. Sa limite est qu'en présentant une succession de coupes, **le chronogramme occulte la dimension individuelle des trajectoires** : on ne peut pas connaître l'enchaînement des situations qui composent les parcours. Par exemple, le graphique ne permet pas de savoir si les individus au chômage à la fin de la trajectoire étaient majoritairement en emploi ou étudiants auparavant. Les transitions, qui sont à la base des processus, ne sont pas présentées.

D'autres exemples de chronogrammes sont présentés dans la partie 5.3.

2.1.2 Tapis

Le second type de représentation graphique, proposé par Stefani Scherrer (2001) et parfois appelé « tapis » (en anglais, *Index Plot*), conserve en revanche la dimension individuelle des parcours. L'abscisse correspond ici encore à l'axe temporel des trajectoires. Chaque individu est représenté par une ligne et chaque ligne est composée de segments de couleurs différentes, la couleur des segments correspondant à la situation occupée et l'abscisse à la longueur des segments respectivement au moment et à la durée de la situation. À partir du même échantillon d'individus, on trace le « tapis » des trajectoires d'insertion (figure 3). On constate que les individus chômeurs ou inactifs en fin de trajectoire sont pour la plupart en emploi auparavant. De plus, la grande majorité des individus passent directement des études à l'emploi sans période intermédiaire de chômage ou d'inactivité. Enfin, une part non négligeable des individus possède une trajectoire parfaitement stable en occupant un emploi du début à la fin de leur parcours.

Figure 3 − Tapis de trajectoires d'insertion



Source : *Biographies et entourage* (2001) ; Champ : échantillon aléatoire de 500 individus.

On voit ici l'intérêt d'une telle représentation graphique : **le tapis conserve la dimension longitudinale des trajectoires individuelles** et, de ce fait, permet de mieux rendre compte des processus. L'interprétation en est toutefois moins aisée que celle des chronogrammes, surtout en présence d'un grand nombre de trajectoires. Afin

de rendre les tapis plus lisibles, il peut être utile de trier les trajectoires, selon les situations en début de parcours, en fin de parcours ou mieux, selon le premier facteur d'un échelonnement multidimensionnel (*Multi-Dimensional Scaling*[10]) sur la matrice de dissimilarité entre trajectoires, comme c'est le cas ici.

2.2 Indicateurs

Un autre moyen de décrire les classes d'une typologie de trajectoires consiste à calculer un certain nombre d'indicateurs pour chacune des classes. Des différences significatives apparaissent alors clairement entre les classes, qui viennent questionner la pertinence de la typologie. Outre l'effectif de la classe, les indicateurs envisageables sont de divers types.

2.2.1 Variables décrivant la trajectoire

On peut caractériser les classes en décrivant les états et les transitions qui composent les trajectoires individuelles. De nombreuses possibilités existent (qui rejoignent les indicateurs synthétiques proposés dans le chapitre 3).

- La distribution des situations initiale et finale permet souvent de différencier les classes.
- L'état modal, c'est-à-dire la situation la plus souvent occupée au sein de la classe, ou la transition modale, c'est-à-dire le passage d'un état à un autre le plus fréquent dans la classe, sont des outils utiles en particulier lorsque l'univers des états possibles est grand (Pollock, 2007).
- La durée totale moyenne dans chacun des états, par exemple la durée moyenne du chômage dans les trajectoires individuelles de la classe.
- Le nombre moyen de transitions au cours des trajectoires, qui distingue les classes stables de celles qui rassemblent des parcours plus chaotiques.
- Pour les classes stables, la proportion moyenne de la durée de la trajectoire passée dans la situation principale.
- Le nombre moyen d'épisodes dans chaque état, par exemple le nombre moyen de périodes de chômage au sein des trajectoires individuelles de la classe, ou plus simplement la proportion d'individus de la classe à avoir connu un état particulier.

[10] Le *Multi-Dimensional Scaling* (MDS) est une technique d'analyse des données proches de l'analyse en composantes principales (ACP) mais qui utilise en entrée une matrice de distance. Elle permet de « réduire » l'information contenue dans la matrice initiale en en extrayant les principales dimensions.

- Le temps moyen d'accès à un état particulier, par exemple la durée moyenne avant d'occuper un premier emploi stable.
- etc.

Il existe aussi plusieurs indicateurs décrivant la complexité des parcours, tels que la « turbulence » (Elzinga et Liefbroer, 2007), l'entropie longitudinale (Gabadinho *et al.*, 2009) ou la « complexité » (Gabadinho *et al.*, 2010a), qui seront présentés plus en détail dans le chapitre 4.

2.2.2 Homogénéité des classes

Une typologie issue d'une classification est construite de manière à obtenir des classes à la fois homogènes et distinctes des autres classes. L'homogénéité à l'intérieur des classes et l'hétérogénéité entre les classes sont toutefois variables d'une classe à l'autre. Elles peuvent être mesurées par le biais de divers indicateurs. Les distances issues d'une analyse factorielle ou de *l'Optimal Matching* permettent de calculer pour chaque classe la distance moyenne entre les trajectoires de la classe (distance intra-classe), la distance moyenne entre les trajectoires d'une classe et celles des autres classes (distance inter-classe), la distance maximale ou la distance moyenne à une trajectoire-type caractérisant la classe (Aassve *et al.*, 2007). D'autres indicateurs d'homogénéité existent, comme l'entropie (Fussel, 2005) ou l'indice de Gini.

Certains de ces indicateurs seront illustrés dans l'exemple traité dans le chapitre 5.

2.3 Les trajectoires-types

La caractérisation des classes d'une typologie peut aussi répondre à la question de savoir quelles sont les trajectoires typiques d'une classe. On cherche alors à réduire chaque classe à une ou plusieurs trajectoires particulièrement représentatives. Il est envisageable d'identifier la trajectoire individuelle la plus fréquente dans la classe. Néanmoins, les trajectoires d'une même classe, même si elles se ressemblent, sont le plus souvent complexes et trop diverses pour laisser émerger une trajectoire significativement plus fréquente. Un autre moyen de présenter une trajectoire-type est de construire une trajectoire moyenne, constituée de l'enchaînement des situations modales à chaque instant d'observation. Toutefois, cette trajectoire moyenne ne correspond pas nécessairement à une trajectoire réellement observée. Elle peut même engendrer des aberrations, notamment lorsque certaines transitions sont irréversibles : par exemple, dans le cas de l'étude de trajectoires parentales, une trajectoire moyenne pourrait comporter une situation « a eu au moins un enfant » précédant une situation « n'a jamais eu d'enfant ». Une meilleure solution consiste à utiliser une trajectoire réelle, observée plutôt qu'artificielle, ce qui peut être effectué en considérant la trajectoire qui constitue le parangon de la classe (*medoid*). Cette trajectoire est la plus proche du centre de gravité de la classe, autrement dit la moins distante de l'ensemble

des autres trajectoires de la classe. Elle correspond donc bien à une trajectoire vécue par un individu. Elle constitue dès lors un moyen particulièrement parlant de caractériser une classe de manière synthétique, de l'« incarner », et est largement adoptée dans les travaux utilisant les typologies de trajectoires (Aassve *et al.*, 2007 ; Wiggins *et al.*, 2007). Lorsque l'hétérogénéité d'une classe est trop importante, il est parfois souhaitable de présenter plusieurs trajectoires-types au lieu d'une seule afin de rendre compte de la diversité des parcours d'une même classe.

3. Mesurer la dissemblance entre trajectoires

Les techniques de mesure de la dissemblance (ou dissimilarité) entre trajectoires sont nombreuses, chacune présentant ses avantages et ses limites. On peut identifier deux familles principales : l'une liée à l'analyse factorielle des données et l'autre à l'analyse séquentielle.

3.1 Une famille de méthodes liées à l'analyse factorielle

Une première famille de méthodes de construction de typologies de trajectoires est basée sur l'usage des techniques d'analyse factorielle des données (dite aussi analyse « géométrique » des données)[11]. Elle est principalement le résultat des nombreux travaux sur l'emploi et la formation effectués par le CEREQ (Centre d'Études et de Recherches sur les Qualifications)[12]. Le CEREQ a réalisé sa première grande enquête de cheminement professionnel en 1980, puis d'autres enquêtes de suivi longitudinal se sont succédé, nécessitant et favorisant le développement de méthodes spécifiques pour exploiter ces données au fur et à mesure de la complexification de l'insertion professionnelle : le processus finement observé devenant à la fois plus difficile à définir et à repérer (Fénelon *et al.*, 1997). L'insertion n'est plus un simple changement d'état, une transition irréversible entre une situation initiale (les études) et une situation finale (un contrat de travail stable). Mais, elle s'est également progressivement transformée en un processus complexe à la faveur notamment de la diversification de la formation et ce parcours observé sur un temps relativement long est fréquemment qualifié de « cheminement ».

Pratiquement, la majeure partie des travaux du CEREQ sur les parcours mettent à profit les données des calendriers récapitulatifs qui accompagnent les questionnaires des enquêtes de suivi et collectent la situation professionnelle des individus le plus souvent mensuellement. La nomenclature des états recensés dans les calendriers varie

[11] Pour une recension détaillée de ces méthodes, voir Grelet (2002).

[12] On en trouve toutefois quelques rares exemples dans la littérature anglo-saxonne (Van der Heijden, 1987 ; Martens, 1994 ; Van der Heijden *et al.*, 1997).

d'une enquête à l'autre et se concentre sur une ou plusieurs dimensions de la situation d'emploi (statut d'activité, contrat de travail, temps de travail, salaire, Profession et Catégorie Socioprofessionnelle [PCS]...) avec un niveau de précision variable.

Afin de rendre les développements méthodologiques qui suivent plus intelligibles, nous utiliserons un exemple d'application purement fictif. On considère le parcours d'activité d'un ensemble d'individus de l'âge de 18 ans jusqu'à 29 ans inclus. La situation des individus est observée chaque année : on a donc douze situations successives pour chaque personne. Chaque année d'observation, la situation d'un individu est résumée par l'un des trois états suivants : étudiant ou en formation (E), chômeur ou inactif (I), en activité (A). Si on prend l'exemple d'une personne, que l'on appellera Calvin, se trouvant au chômage à l'arrêt de ses études à 20 ans, avant de trouver un emploi à 23 ans, le calendrier de son parcours peut être représenté comme suit :

Tableau 1 − Exemple de calendrier de parcours individuel, celui de Calvin

18	19	20	21	22	23	24	25	26	27	28	29
E	E	I	I	I	A	A	A	A	A	A	A

Les parcours des individus sont souvent relativement divers. Potentiellement, le nombre de parcours distincts au sein d'une population est très élevé : dans notre exemple, il serait de $3^{12} = 531\,441$ (3 états et 12 années d'observation). Même si dans la pratique certains parcours, en particulier les plus stables, concernent souvent un nombre important d'individus, tenter de regrouper les individus se ressemblant par la simple observation des données peut rapidement s'avérer insurmontable. On peut en effet souhaiter réaliser une typologie *a priori* en fixant un certain nombre de critères qui vont déterminer la construction des classes d'individus : par exemple en formant une classe de personnes n'ayant jamais connu une situation donnée, ou une autre regroupant des parcours contenant une longue période en un état donné (Degenne *et al.*, 1994). Mais la variété des parcours risque de poser rapidement des problèmes : certains parcours atypiques peuvent ne correspondre à aucun critère, d'autres au contraire en remplir plusieurs. Une autre possibilité consiste à centrer l'analyse sur les parcours stables (la même situation est occupée du début à la fin de la période) ou comportant peu de transitions, mais cela implique de laisser de côté une partie de la population étudiée. L'emploi d'outils statistiques plus élaborés apparaît donc indispensable : il s'agit alors de construire des typologies *a posteriori*. Ainsi, les travaux du CEREQ utilisent majoritairement les techniques d'analyse factorielle des données, issues du courant français des statistiques exploratoires multidimensionnelles (par exemple, voir Rouanet et Le Roux, 1993 ; Bry, 1995 et 1996 ; Lebart *et al.*, 2000), qu'ils adaptent et appliquent à des données longitudinales. L'idée directrice de la famille de méthodes expérimentées par le CEREQ est de classifier les individus, représentés par leur parcours, selon leur ressemblance, en un nombre fini de groupes homogènes et disjoints.

Il existe plusieurs variantes appliquant les techniques d'analyse factorielle à des trajectoires. La différence entre ces variantes réside essentiellement dans le codage des données et dans le type de mesure de la ressemblance utilisé (Grelet, 2002).

3.1.1 Disjonctif complet

Une première méthode consiste à recoder le calendrier du parcours sous forme disjonctive complète : une variable binaire est associée à chaque état pour chaque moment d'observation. Dans notre exemple, pour chaque année, on crée trois variables correspondant à chacun des états possibles et codées 1 lorsque l'individu est dans l'état concerné cette année-là, et 0 sinon. On a alors 12*x*3 = 36 variables dichotomiques pour chaque individu. Le parcours de Calvin est ainsi résumé dans le tableau 2.

Tableau 2 − Tableau disjonctif-complet du parcours de Calvin

18E	18I	18A	19E	19I	19A	20E	20I	20A	...	29E	29I	29A
1	0	0	1	0	0	0	1	0	...	0	0	1

Lecture : À 18 ans, Calvin est en études ou formation et non au chômage, ni en inactivité ou en emploi, de même à 19 ans…

Le tableau ainsi constitué peut être considéré comme un tableau de fréquences et soumis à une Analyse Factorielle des Correspondances (AFC). La distance utilisée est alors celle du χ^2. Mais il est aussi possible de soumettre le même tableau à une Analyse en Composantes Principales (ACP) non normée, qui utilise la distance euclidienne usuelle. Cette méthode est parfois appelée « méthode du Lhire[13] » (Espinasse, 1993 ; Béduwé *et al.*, 1995 ; Béduwé, 2001), du nom du laboratoire dont sont issus certains des premiers travaux utilisant cette approche. Avec l'AFC et donc la métrique du χ^2, la distance entre deux parcours individuels est pondérée par l'inverse de la fréquence des variables : les situations peu fréquentes une année donnée contribuent plus au calcul de l'écart entre deux parcours que les situations majoritaires. En d'autres termes, on donne plus d'importance aux états rares. Au contraire, la distance euclidienne égalise la contribution des états[14] et se base sur le nombre de discordances entre les trajectoires. Dans notre exemple (tableau 3), les parcours de Calvin et Hobbes diffèrent à 20, 22 et 23 ans, le nombre de discordances entre les parcours est donc égal à 3.

Tableau 3 − Les parcours professionnels de Calvin et Hobbes

	18	19	20	21	22	23	24	25	26	27	28	29
Calvin	E	E	I	I	I	A	A	A	A	A	A	A
Hobbes	E	E	E	I	A	I	A	A	A	A	A	A

[13] Laboratoire Interdisciplinaire de recherche sur les Ressources Humaines et l'Emploi.

[14] Pour plus de développements sur la formalisation mathématique de ces distances et leurs implications, voir Grelet (2002).

*La mesure de la dissemblance entre parcours individuels par l'utilisation d'un codage disjonctif-complet associé à une analyse factorielle met donc l'accent sur la **contemporanéité** de situations identiques, que ces situations identiques se suivent ou se situent à des moments éloignés de la trajectoire. C'est la **durée** de la simultanéité dans un état commun qui est au principe du rapprochement des individus, la notion de simultanéité impliquant aussi la prise en compte du **moment**, du "timing", des situations. En revanche, la nature des transitions et l'ordre de leur enchaînement n'est pas intégré à l'analyse.*

On notera que le recours à l'analyse factorielle n'est pas indispensable. Il est envisageable d'effectuer une classification directement à partir des distances calculées avec la distance euclidienne[15] ou celle du χ^2. Toutefois, l'utilisation d'une analyse factorielle est fortement recommandée. Cela permet en effet de concentrer l'analyse sur les dimensions captant l'essentiel de l'information (ou inertie) apportée par les variables. On élimine ainsi les dimensions résiduelles, trop faiblement corrélées aux variables et par suite sans robustesse (on peut les concevoir comme des dimensions de « bruit »). Cette remarque est aussi valable pour les méthodes qui seront présentées dans la suite de cette partie 3.1.

Par ailleurs, l'analyse factorielle est basée sur les corrélations (ou covariances) entre variables. Dans le cas des parcours codés sous forme de tableau disjonctif complet, ces variables correspondent au fait d'être dans un état donné une année donnée. Or les parcours de vie individuels ne sont souvent composés que d'un nombre limité de changements d'état. Il en résulte que le fait d'être dans un état à un moment t sera probablement corrélé au fait d'être dans le même état au moment $t+1$. Cela implique que l'analyse factorielle assouplit l'accent mis sur la contemporanéité de situations identiques. En d'autres termes, deux trajectoires identiques mais légèrement décalées dans le temps (par exemple, deux trajectoires composées d'une transition des études à l'emploi, mais à un an de différence) seront tout de même considérées comme relativement similaires.

3.1.2 Analyse harmonique qualitative

Une alternative au codage disjonctif-complet consiste à décrire les parcours en résumant les calendriers individuels, méthode parfois appelée Analyse Harmonique Qualitative (AHQ). Plus précisément, on découpe la période étudiée en sous-périodes, puis on mesure pour chacune d'elles le temps passé dans chacun des états. Il est aussi possible d'y adjoindre des variables dénombrant les transitions d'un état à un autre. On emploie ensuite une analyse factorielle des correspondances (distance du χ^2).

[15] Dans ce cas, la méthode est équivalente à la distance de Hamming qui sera évoquée dans la partie sur l'analyse de séquences.

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 et 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. Pendant longtemps, la méthode n'a que rarement été utilisée (Degenne *et al.*, 1996 ; Grelet, 2002). Finalement, il faut attendre le renouveau des collectes biographiques pour voir des applications se concrétiser, d'abord sur des données latino-américaines (Dureau *et al.*, 1994 ; Barbary, 1997 ; Barbary et Pinzon Sarmiento, 1998) puis, plus récemment, françaises puisque l'enquête *Biographies et entourage* va permettre de suivre les trajectoires familiales (Robette et Thibault, 2006) ou géographiques (Robette *et al.*, 2011) mais aussi les carrières professionnelles (Robette et Thibault, 2008).

La construction du tableau de données pour l'AHQ se fait en plusieurs étapes. Une fois la période d'étude fixée (de 18 à 29 ans dans notre exemple), on doit la découper en sous-périodes pour l'analyse. Un nombre de sous-périodes égal à la longueur de la trajectoire serait équivalent au codage sous forme de disjonctif complet. À l'opposé, un nombre trop réduit de sous-périodes entraînerait la perte d'une partie importante de la richesse de l'information disponible. Il y a donc un arbitrage à effectuer pour établir le nombre des sous-périodes.

Un autre arbitrage concerne l'amplitude des sous-périodes. 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 une forte densité d'événements, d'autres par une mobilité plus faible[16]. Choisir des amplitudes plus courtes pour les moments de plus grande mobilité permet de leur donner plus d'importance dans la construction de la typologie des parcours individuels. Dans notre exemple, si l'on décide de découper la période étudiée en trois sous-périodes, on peut choisir de le faire avec des amplitudes toutes égales à quatre ans. Les trois sous-périodes seront alors : de 18 à 21 ans, de 22 à 25 ans, de 26 à 29 ans. Mais on peut aussi estimer qu'une majeure partie des événements, c'est-à-dire des transitions d'un état à un autre, sont susceptibles d'arriver au début du parcours et décider d'accentuer l'importance de ce début de parcours en découpant la période comme suit, par exemple : de 18 à 20 ans, de 21 à 23 ans et de 24 à 29 ans, avec des amplitudes de respectivement 3, 3 et 6 ans.

Ensuite, pour chaque individu, on calcule la proportion de la durée de chaque sous-période passée dans chacun des états possibles. Le nombre de variables créées est égal

[16] Empiriquement, un moyen d'évaluer la concentration des événements selon le moment du parcours consiste à calculer les déciles de la distribution par âge des changements d'états.

au nombre de sous-périodes multiplié par le nombre d'états. Dans notre exemple, avec trois états et un découpage de la période en trois sous-périodes d'amplitude égale, le nombre total de variables calculées est de 3*x*3 = 9. Pour Calvin, elles prendraient les valeurs présentées dans le tableau 4.

Tableau 4 − Codage de l'AHQ du parcours de Calvin

1E	1I	1A	2E	2I	2A	3E	3I	3A
0.5	0.5	0	0	0.25	0.75	0	0	1

Lecture : Calvin passe la moitié de la première sous-période (de 18 à 21 ans) en étude ou formation, l'autre moitié au chômage ou en inactivité...

Certains auteurs proposent d'ajouter aux données analysées une autre série de variables actives, relatives aux transitions (Degenne *et al.*, 1995). Dans notre exemple, le nombre de types de transitions possibles est égal au carré du nombre d'états[17], soit ici $3^2 = 9$, et le nombre total de transitions au cours du parcours correspond au nombre d'années de la période -1, soit ici 12-1 = 11. On dénombre alors le nombre de transitions de chaque type présentées dans le tableau 5.

Tableau 5 − Exemple de variables de transition pour le parcours de Calvin

E→E	E→I	E→A	I→E	I→I	I→A	A→E	A→I	A→A
1	1	0	0	2	1	0	0	6

Lecture : Calvin a vécu au cours de son parcours une transition de E vers E, de E vers I et de I vers A, deux de I vers I et 6 de A vers A.

On peut aussi restreindre les transitions aux seuls changements d'états (Robette et Thibault, 2006). Le nombre de transitions possibles est ici égal à :

(nombre d'états) x (nombre d'états -1), soit ici 3*x*2 = 6.

On dénombre chaque type de transition (changement d'état) et on le rapporte au nombre total de transitions. Dans le cas de Calvin, le parcours comprend deux changements d'état : l'un de E vers I et l'autre de I vers A. Les valeurs des variables de transition créées sont indiquées dans le tableau 6.

Tableau 6 − Exemple de variables de changement d'état pour le parcours de Calvin

E→I	E→A	I→E	I→A	A→E	A→I
½	0	0	½	0	0

Lecture : La moitié des changements d'état au cours du parcours de Calvin sont de type E vers I et l'autre de type I vers A.

[17] La notion de transition étant ici entendue comme le passage de la situation au moment *t* à la situation au moment *t*+1, que ces situations soient différentes ou non.

Les variables de transition ne bouleversent pas les résultats obtenus avec une AHQ simple mais permettent souvent d'obtenir une typologie des parcours avec des classes plus homogènes (Robette et Thibault, 2006). D'autre part, ainsi construit, le tableau de données intègre à l'analyse trois dimensions cruciales des parcours (Degenne *et al.*, 1995) : la **durée** et le **moment**, comme avec le disjonctif complet, mais aussi les **transitions**, car un processus n'existe que s'il y a changement d'état. Toutefois, la séquence des changements d'état, vue comme un tout, n'est pas prise en compte.

*Par rapport aux méthodes utilisant un codage disjonctif-complet, le découpage en sous-périodes **assouplit la sensibilité à la simultanéité** dans un état commun. En d'autres termes, deux séquences d'événements identiques mais légèrement décalées dans le temps auront tendance à être considérées comme plus similaires par l'AHQ que par les variantes précédentes. Par ailleurs, la possibilité de choisir un découpage en sous-périodes d'amplitudes différentes confère à la méthode une certaine souplesse, en permettant de donner plus d'importance à des parties du parcours considérées comme portant la part la plus intéressante de l'information.*

3.1.3 Indicateurs synthétiques

Une autre méthode consiste à résumer les parcours de manière plus drastique qu'avec l'AHQ. Les parcours individuels sont synthétisés par un certain nombre d'indicateurs simples de durée ou de comptage (Charlot et Pottier, 1987), par exemple :

- la durée passée dans chacun des états E, I et A (en années ou en % du parcours),
- le temps d'accès à la première situation d'emploi A,
- le fait d'avoir ou non passé au moins 1 an dans un état (E, I ou A),
- le nombre de périodes dans chacun des états E, I et A,
- le nombre total de périodes au cours de la trajectoire (ou le nombre de transitions).

Cette liste n'est pas exhaustive et on peut imaginer de nombreux indicateurs selon la problématique de la recherche, éventuellement en les combinant. Le tableau construit à partir de ces variables est ensuite soumis à une ACP (distance euclidienne).

3.1.4 Analyse de tableaux multiples

Une cinquième variante est liée aux méthodes d'analyse tableaux multiples, telles que l'Analyse Factorielle Multiple (AFM) ou STATIS (Escofier et Pagès, 2008). L'AFM consiste ainsi à effectuer une ACP globale de l'ensemble des tableaux en pondérant chaque groupe de variables pour en équilibrer l'influence. Une application de l'AFM au traitement des enquêtes de cheminement propose ainsi la construction de trois tableaux (Grelet, 1994). Le premier correspond au calendrier résumé, dont un exemple

est donné dans le tableau 4. Le second tableau présente les durées totales, c'est-à-dire le temps passé dans chaque état au cours de la période. Pour Calvin, on obtiendrait les valeurs présentées dans le tableau 7.

Tableau 7 − Tableau des durées totales du parcours de Calvin

E	I	A
2	3	7

Lecture : Calvin a passé 2 années en études ou en formation, 3 années au chômage ou en inactivité et 7 années en emploi au cours de son parcours.

Le troisième tableau est appelé « minimal ». Il retient uniquement, pour chaque état, le fait pour un individu d'être passé (codé 1) ou non (codé 0) au moins une fois par cet état au cours de son parcours. Dans notre exemple, cela donnerait ceci :

Tableau 8 − Tableau minimal pour le parcours de Calvin

E	I	A
1	1	1

Lecture : Calvin a passé au moins une année en études ou en formation au cours de son parcours, de même qu'au chômage ou en inactivité ou en emploi.

Cette méthode, bien que peu connue et rarement utilisée, présente plusieurs avantages: elle permet de représenter les mouvements globaux entre situations et les trajectoires individuelles sur des plans factoriels, de projeter les informations du tableau minimal en éléments illustratifs et d'avoir des indications sur les liens entre facteurs de l'analyse globale et partielle.

3.1.5 Analyse textuelle

On peut enfin signaler quelques applications de la statistique textuelle aux parcours professionnels (Houzel et Le Vaillant, 1994 ; Jalaudin et Moreau, 1995 ; Courgeau et Guérin-Pace, 1998 ; Briard, 2007). Traduit dans le langage des méthodes d'analyse textuelle, chaque situation combinée à sa durée forme un mot, l'ensemble des mots existant composant le vocabulaire relatif aux données étudiées. Une suite de mots est appelée segment et la succession des mots représentant la trajectoire individuelle est appelée phrase. Les parcours individuels peuvent être décomposés en plusieurs segments de longueur variable. Le tableau croisant les individus et l'ensemble des segments construits est soumis à une Analyse de Correspondances Multiples (ACM).

3.2 Une autre famille de méthodes : l'analyse de séquences

La seconde famille est basée sur la notion de « séquences ». Une grande variété de travaux en sciences sociales s'intéresse aux séquences d'événements ou de situations. Ils concernent par exemple l'étude des événements qui composent les parcours de vie (relatifs aux études, au travail, à la famille, à la mobilité résidentielle, etc.), des carrières professionnelles, de l'évolution des politiques ou des lois, des changements culturels, etc. Pratiquement, une séquence se définit comme une liste d'éléments ordonnés, ces éléments pouvant être de n'importe quelle nature (événements, nombres, etc.). L'analyse de séquences (*Sequence Analysis*) est un corpus de techniques analytiques traitant les données sous forme de séquences telles qu'elles viennent d'être définies. Les recherches utilisant l'analyse séquentielle tentent principalement de répondre à trois questions (Macindoe et Abbott, 2004).

- Existe-t-il des modèles (*patterns*), des séquences typiques parmi un ensemble de séquences donné ?
- Lorsqu'elles existent, comment ces séquences typiques sont-elles produites, quels facteurs les déterminent ?
- Quelles sont les conséquences de ces séquences typiques ?

Les données séquentielles peuvent être analysées de nombreuses manières. L'analyse de séquences se caractérise par le fait qu'elle considère comme unité d'analyse la séquence dans son ensemble, comme un tout, et non une série de points ou d'observations. Contrairement à l'analyse biographique (*Event History Analysis*) ou aux séries temporelles (*Time Series*), la séquence n'est pas vue comme un processus stochastique, généré pas à pas, mais comme une unité d'analyse à part entière (Abbott et Tsay, 2000)[18]. Il existe différentes méthodes d'analyse séquentielle, appliquées à des champs de recherche en sciences humaines et sociales variés, comme la psychologie, l'archéologie, la linguistique, les sciences politiques ou la sociologie (Abbott, 1995). L'une de ces méthodes est discutée, employée et diffusée beaucoup plus largement que les autres : il s'agit de l'*Optimal Matching Analysis* (OMA)[19].

3.2.1 L'Optimal Matching Analysis (OMA)

Description

L'usage de l'*Optimal Matching* s'est principalement développé en biologie moléculaire pour l'analyse des protéines et des séquences d'ADN. L'objectif est de rechercher dans d'importantes bases de données des séquences ressemblant à une séquence particulière, par exemple à une protéine donnée. Des algorithmes permettant

[18] Développées et théorisées par A. Abbott à propos de l'analyse de séquences, ces différentes remarques s'appliquent en fait aussi à l'analyse des trajectoires avec les méthodes factorielles qui viennent d'être présentées.

[19] *Optimal Matching Analysis* peut se traduire en français par « Méthodes d'Appariement Optimal » (Lesnard et de Saint Pol, 2004).

d'effectuer cette tâche sont apparus au début des années 1970, puis se sont multipliés jusque dans les années 1980 (Sankoff et Kruskall, 1983). La première utilisation d'un algorithme d'*Optimal Matching* en sciences sociales est réalisée dans un article consacré aux séquences de figures dans diverses danses folkloriques, dans le but de caractériser les modèles de solidarité dans l'Angleterre rurale du XIX^{e} siècle (Abbott et Forrest, 1986). Depuis, de nombreux travaux ont exploité cette méthode, le plus souvent pour l'étude des carrières professionnelles (Abbott et Hrycak, 1990 ; Stovel *et al.*, 1996 ; Halpin et Chan, 1998 ; Blair-Loy, 1999 ; Robette et Thibault, 2008) et du début de la vie active[20] (Scherer, 2001 ; McVicar et Anyadike-Danes, 2002 ; Brzinsky-Fay, 2007). Mais des domaines de recherche variés ont été explorés, comme le passage à la retraite (Han et Moen, 1999), le passage à l'âge adulte (Aassve *et al.*, 2007 ; Robette, 2010), les carrières militantes (Blanchard, 2010), les emplois du temps (Wilson, 1998 ; Lesnard et de Saint Pol, 2004) ou les trajectoires résidentielles (Stovel et Bolan, 2004), voire des thèmes plus inattendus, tels que la succession d'épisodes de lynchage aux États-Unis (Stovel, 2001), la structure rhétorique des articles de sociologie (Abbott et Barman, 1997), le contenu de manuels scolaires (Levitt et Nass, 1989) ou les réseaux d'entreprise (Stark et Vedres, 2006).

L'*Optimal Matching* ne constitue en pratique qu'une étape d'une analyse séquentielle. Son principe consiste à mesurer la dissemblance (ou dissimilarité) entre chaque paire de séquences constituant l'échantillon[21]. La matrice de dissimilarité ainsi calculée sert alors de point de départ à la seconde étape de la démarche analytique. Celle-ci consiste le plus souvent à construire une typologie de séquences, en créant des groupes de séquences similaires, à l'aide de techniques telles que les classifications hiérarchiques[22]. Cela répond à la première question de l'analyse séquentielle, à savoir s'il existe des séquences typiques au sein de la base de données. La typologie peut ensuite être utilisée comme variable dépendante ou indépendante pour de nouvelles analyses, qui constituent la troisième et dernière étape de l'analyse séquentielle, afin de répondre aux questions suivantes : quelles sont respectivement les causes et les conséquences de l'existence de séquences typiques[23] ? L'*Optimal Matching*, comme la plupart des autres méthodes d'analyse séquentielle, est donc une méthode avant tout descriptive et non explicative.

Les algorithmes d'*Optimal Matching* définissent une mesure afin de calculer une distance entre des séquences. L'idée générale consiste à mesurer la dissimilarité entre deux séquences en transformant l'une en l'autre au moyen d'opérations élémentaires[24].

---

[20] *School-to-work transitions*.

[21] Pour un ensemble de N séquences, le nombre total de distances calculées est donc de N(N-1)/2.

[22] On pourrait, préalablement à la classification automatique, soumettre la matrice de dissimilarité à un échelonnement multidimensionnel (MDS), afin d'éliminer le « bruit » statistique à la manière des méthodes factorielles présentées en 3.1. Cette voie n'a cependant jamais été empruntée (à notre connaissance).

[23] Pratiquement, l'OM n'implique pas nécessairement la construction de typologies. Dans certains travaux, la démarche se limite au calcul de distances entre les séquences sans y adjoindre de classification (voir par exemple Scherer, 2001). Cette question sera développée dans la quatrième partie.

[24] Pour désigner la transformation d'une séquence en une autre, on pourra parler d'« appariement » ou d'« alignement » d'une paire de séquences.

Les trois opérations élémentaires sont : 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). Il existe de nombreuses manières de transformer une séquence en une autre au moyen des opérations élémentaires. En conséquence, la distance − mesurée par la dissimilarité − entre deux séquences correspond au nombre minimum d'opérations élémentaires nécessaires à la transformation d'une séquence en l'autre. Cette distance est appelée distance de Levenshtein I (Levenshtein, 1965).

Pour illustrer cette démarche, reprenons l'exemple de la partie précédente. La représentation des parcours adoptée alors correspond en fait déjà à une représentation séquentielle. Ajoutons au parcours de Calvin celui de Hobbes, légèrement différent (voir tableau 9). L'appariement de ces deux séquences au moyen des trois opérations élémentaires peut se faire de plusieurs manières. Une première possibilité consiste à supprimer une année d'études (E) au début de la séquence de Hobbes, à ajouter une année d'activité (A) à la fin et à remplacer l'activité (A) par l'inactivité (I) entre les deux étapes d'inactivité : cela nécessite trois opérations. Une seconde possibilité consiste à remplacer E par I à 20 ans, A par I à 22 ans et I par A à 23 ans : de nouveau trois opérations sont nécessaires. Les deux alternatives sont donc équivalentes en termes de distance entre la séquence de Calvin et celle de Hobbes.

Tableau 9 − Séquences des parcours professionnels de Calvin et Hobbes[25]

	18	19	20	21	22	23	24	25	26	27	28	29
Calvin	E	E	I	I	I	A	A	A	A	A	A	A
Hobbes	E	E	E	I	A	I	A	A	A	A	A	A

Cette version simple de l'*Optimal Matching* correspond au cas où l'on considère que les opérations d'insertion, de suppression et de substitution sont d'égale importance. Cependant, de nombreuses raisons théoriques peuvent amener à penser que l'une ou l'autre des opérations élémentaires a plus de poids que les autres. Or il est possible d'associer un coût spécifique à chacune des opérations élémentaires. Une série d'opérations aura un coût équivalent à la somme des coûts des opérations élémentaires. La distance entre deux séquences sera alors définie comme le coût minimal nécessaire à la transformation d'une séquence en l'autre. Plusieurs algorithmes dynamiques existent pour le calcul du coût minimal (Sankoff et Kruskal, 1983), le plus utilisé en sciences sociales étant celui de Needleman-Wunsch (1970)[26].

[25] Ce tableau est identique au tableau 3 et est reproduit pour plus de clarté.

[26] Cet algorithme a été implémenté dans plusieurs logiciels de statistiques : TDA (Rohwer et Pötter, 2005), Stata (Brzinsky-Fay *et al.*, 2006) ou R (Gabadinho *et al.*, 2009). Voir en annexe pour une revue plus complète des possibilités logicielles.

Le choix des coûts

Le choix des coûts des opérations élémentaires constitue une étape essentielle des techniques d'*Optimal Matching* : il « hante » les analyses utilisant cette méthode (Stovel *et al.*, 1996). C'est la possibilité de détermination des coûts par le chercheur qui confère à la méthode sa souplesse et sa capacité à s'adapter à l'objet étudié (Lesnard et de Saint Pol, 2004). Dans la pratique, insertion et suppression sont considérées comme une seule et même opération dans la mesure où, pour l'appariement d'une paire de séquence, la suppression d'un élément dans une séquence est équivalente à l'insertion d'un élément dans l'autre[27]. L'opération d'insertion-suppression est appelée *indel*, par contraction des termes anglais *insertion* et *deletion*. Les opérations *indel* privilégient l'ordre des événements en rapprochant des parties de séquences identiques mais situées à des moments différents. L'insertion ou la suppression d'un élément induisent en contrepartie une altération de la structure temporelle des séquences comparées et elles déforment le temps. À l'inverse, les opérations de substitution préservent la structure temporelle des séquences, puisqu'elles comparent des situations situées au même moment de la séquence, mais altèrent l'enchaînement des événements (Lesnard et de Saint Pol, 2004).

Il y a donc deux catégories de coûts : les coûts de substitution et le coût *indel*. Dans la pratique, on choisit en général les coûts de substitution en premier, le choix du coût *indel* venant ensuite, selon l'importance que l'on souhaite donner à l'ordre des éléments des séquences. Les coûts de substitution peuvent être identiques quels que soient les éléments substitués : la substitution de E par A aura le même coût que celle de E par I. Mais ils peuvent aussi être spécifiques à chaque paire d'éléments : on définit alors une matrice des coûts de substitution. Cette matrice est symétrique, dans la mesure où pour l'appariement de deux séquences AEA et AIA, il est équivalent de remplacer E par I dans la première séquence ou I par E dans la seconde[28]. Dans notre exemple, une matrice des coûts de substitution pourrait prendre la forme suivante :

Tableau 10 − Exemple de matrice de coûts de substitution

	E	I	A
E	0	1	2
I	1	0	3
A	2	3	0

Lecture : Le remplacement de I par E a un coût de 1, celui de A par E de 2, celui de A par I de 3.

[27] Par exemple, si l'on souhaite apparier deux séquences EIA et EA, il est équivalent de supprimer I dans la première séquence ou d'insérer I entre le E et le A dans la seconde séquence.

[28] D'un point de vue mathématique, la symétrie des coûts de substitution garantit que la distance réponde aux propriétés d'une métrique : séparation, symétrie et inégalité triangulaire (MacIndoe et Abbott, 2004).

La valeur des coûts de substitution de la matrice peuvent être déterminés de différentes manières. L'enjeu est d'adapter le mode de détermination des coûts aux données et aux hypothèses de recherche. 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. Cette stratégie se base le plus souvent sur une structure hiérarchique des éléments, préexistante ou construite pour l'analyse. 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 (Halpin et Chan, 1998 ; Blair-Loy, 1999 ; Scherer, 2001 ; Solis et Billari, 2002). Une solution alternative consiste à laisser les données diriger la détermination des coûts, en dérivant les coûts de substitution des probabilités de transition entre les éléments (Rohwer et Pötter, 2005). Le coût de substitution entre deux éléments est alors d'autant plus élevé que la probabilité de transition entre ces éléments est faible (Han et Moen, 1999 ; Pollock *et al.*, 2002 ; Aassve *et al.*, 2007 ; Robette et Thibault, 2008). On peut enfin adopter des stratégies plus complexes, par exemple en utilisant conjointement une hiérarchie des éléments et les probabilités de transition (Abbott et Hrycak, 1990 ; Stovel et Bolan, 2004).

Le choix du coût *indel* est lui aussi important dans sa relation avec les coûts de substitution. Certains chercheurs préfèrent adopter des coûts de substitution et *indel* d'une même et unique valeur, arguant du manque de justifications théoriques à opérer différemment (Dijkstra et Taris, 1995). Dans ce cas, on est en présence de la version simple de l'algorithme d'*Optimal Matching* (distance de Levenshtein I, voir tableau 11) : la distance entre deux séquences est égale au nombre minimum d'opérations élémentaires nécessaires à leur appariement.

Tableau 11 − Coûts de substitution et d'insertion-suppression des distances de Hamming et Levenshtein

Distance	Coût de substitution	Coût *indel*
Hamming	1	-
Levenshtein I	1	1
Levenshtein II	-	1

Les premières applications de l'*Optimal Matching* avaient tendance à fixer les coûts *indel* à un niveau assez élevé. Cependant, avec des séquences de longueur égale, si les coûts *indel* ont une valeur supérieure au coût de substitution maximal multiplié par la moitié de la longueur des séquences, les opérations d'insertion-suppression ne sont jamais utilisées. On est alors en présence de la distance de Hamming (Hamming, 1950), qui se base sur la simultanéité d'éléments identiques : la dissimilarité entre deux séquences est équivalente au nombre de substitutions nécessaires pour les apparier, c'est-à-dire − en termes de parcours − le nombre d'unités de temps pour lesquelles la situation est différente. Lorsque les séquences ont des longueurs différentes, un tel coût *indel* a pour conséquence une utilisation des opérations d'insertion-

suppression uniquement pour compenser la différence de longueur. C'est pourquoi lorsque, comme souvent, l'un des objectifs de l'analyse est d'identifier des portions de séquences identiques mais décalées dans le temps, le coût *indel* devrait être fixé à une valeur nettement inférieure (Macindoe et Abbott, 2004). Toutefois, avec un coût *indel* inférieur ou égal à la moitié du coût de substitution minimal, seules les opérations *indel* seront utilisées. La dissimilarité entre deux séquences correspond alors à la longueur de leur plus longue sous-séquence commune, aussi appelée distance de Levenshtein II. Finalement, le choix des coûts revient à « positionner le curseur entre les deux cas limites des distances de Hamming et de Levenshtein II », selon que l'on privilégie la contemporanéité des situations ou la présence de sous-séquences communes (Lesnard et de Saint Pol, 2009), la temporalité ou l'ordre au sein des trajectoires.

3.2.2 Critiques et alternatives

Parallèlement à la multiplication des travaux mettant à profit leurs qualités, les techniques d'*Optimal Matching* doivent faire face à un certain nombre de critiques. Celles-ci sont souvent dues au fait que l'analyse séquentielle est mise en regard des approches stochastiques, dont la large diffusion a validé la pertinence mais a aussi tendance à orienter le questionnement des chercheurs. Toutefois, ces deux familles de méthodes ne répondent pas aux mêmes objectifs et ne sont pas concurrentes mais complémentaires. Si l'on s'en tient à l'ambition modeste mais indispensable qui consiste à explorer et décrire des données longitudinales complexes, la plupart des problèmes soulevés par l'*Optimal Matching* sont tout à fait tolérables (Halpin, 2003). Diverses critiques pointent cependant précisément certaines faiblesses de l'*Optimal Matching*, et ont parfois engendré le développement d'une « deuxième vague » d'analyse séquentielle (Aisenbrey et Fasang, 2010), tentant d'en dépasser les limites.

- On reproche ainsi à l'*Optimal Matching* de ne pas offrir de moyen convaincant de validation des choix adoptés durant l'analyse, relatifs au codage des données, aux coûts des opérations élémentaires ou à la sélection d'une typologie (Levine, 2000 ; Wu, 2000 ; Elzinga, 2003). Tout d'abord, les nomenclatures construites en sciences sociales, comme celles des professions, seraient trop faibles ou floues pour permettre un codage pertinent des éléments composant les séquences (Levine, 2000). Ce n'est toutefois pas un problème spécifique à l'analyse séquentielle, mais à l'ensemble des sciences sociales. D'autre part, certains travaux montrent que les différences de codage n'ont que peu d'impact sur les résultats (Forrest et Abbott, 1990).
- Ensuite, la détermination des coûts est jugée arbitraire et peu liée à des hypothèses théoriques, du fait que les opérations élémentaires n'ont pas d'interprétation sociologique claire (Levine, 2000 ; Wu, 2000 ; Elzinga, 2003). Il en résulte que les distances calculées n'auraient pas de signification intrinsèque. De nouveau, cela ne constitue pas réellement un problème dans l'optique d'une analyse exploratoire. Les opérations élémentaires sont avant tout des procédures techniques visant à produire un indice de similarité, mettant plus ou moins l'accent sur la simultanéité ou l'ordre des états. De plus, la détermination des

coûts de substitution à partir des données, en fonction des probabilités de transition entre états (Rohwer et Pötter, 2005), répond en partie à la réserve concernant l'arbitraire du choix des coûts[29].

- De même, si l'on substitue à la dernière étape de l'analyse (la construction d'une typologie de séquences) le calcul de la distance entre la séquence de chaque individu et une unique séquence référence ou idéal-typique (Scherer, 2001 ; Malo et Munoz-Bullon, 2003 ; Kogan, 2004), la distance aura alors un sens facilement interprétable, par exemple : l'écart d'un individu par rapport à la trajectoire « normale » d'une population.
- La sélection d'une typologie de trajectoire pertinente peut aussi être vue comme problématique (Wu, 2000), même si cette question n'est pas spécifique à l'analyse séquentielle mais concerne l'ensemble des méthodes classificatoires. Il est toutefois possible d'utiliser des indicateurs, par exemple à partir des distances inter- et intra-classes (Abbott et Hrycak, 1990), pour aider au choix d'un nombre de classes pertinent[30]. Mais le plus souvent, ce choix est guidé par l'arbitrage entre la nécessité de ne pas trop simplifier la diversité des trajectoires individuelles et celle d'obtenir un ensemble de classes interprétables et significatives.
- En revanche, le fait que la différence de longueur entre les séquences influence le calcul des distances est un enjeu important de l'*Optimal Matching*. Lorsque la différence de longueur des séquences est directement liée au processus étudié (par exemple, insertion professionnelle courte *versus* longue), il est légitime que la distance reflète cette différence : il s'agit alors simplement de choisir des coûts appropriés, notamment par l'intermédiaire d'un coût *indel* relativement élevé. En revanche, la question apparaît plus décisive lorsque les différences de longueur des séquences sont le résultat d'une censure indépendante du processus étudié (Wu, 2000). Une première alternative consiste à contourner le problème en recodant les données, soit en imputant une valeur aux données manquantes, soit en créant un nouvel état « valeur manquante ». Il est aussi possible de standardiser la distance entre une paire de séquences en fonction de leur longueur, par exemple en divisant la distance par la longueur de la plus longue des deux séquences (Abbott et Hrycak, 1990 ; Stovel *et al.*, 1996). Enfin, Stovel et Bolan (2004) proposent d'introduire un coût *indel* variable selon la longueur des séquences : le coût *indel* est fixe lorsque les séquences comparées sont de même longueur, et égal à environ un quart du coût fixe lorsque les séquences sont de longueur différente.
- Un ensemble de critiques concernent la manière dont l'*Optimal Matching* traite les interdépendances complexes au cours du temps. L'analyse séquentielle aborde la question de l'interdépendance des causes en faisant l'hypothèse que les processus sont hétérogènes. Elle ne fait pas de distinction entre causes et effets

[29] Plusieurs travaux ont par ailleurs montré que différents systèmes de coûts n'engendraient pas d'importantes différences de résultats (Levitt et Nass, 1989 ; Chan, 1995 ; McVicar et Anyadikes-Danes, 2002).

[30] Il est aussi notable que la technique de classification utilisée n'a qu'une influence marginale sur les résultats (McVicar et Anyadikes-Danes, 2002).

au cours de la séquence mais appréhende la séquence comme le produit de processus potentiellement multiples et étroitement liés (Aisenbrey et Fasang, 2010). Certains processus sont pourtant multidimensionnels ou se déroulent en interaction, comme c'est le cas du parcours d'activité et de l'histoire familiale. La multidimensionnalité peut être prise en compte en combinant *a priori* les différentes dimensions dans le codage des états (Abbott et Hrycak, 1990 ; Stovel *et al.*, 1996 ; Blair-Loy, 1999 ; Aassve *et al.*, 2007 ; Pollock, 2007 ; Robette, 2010) – un état sera par exemple « étudiant, célibataire, sans enfant » − et éventuellement en combinant les coûts de substitution de chaque dimension, ou en utilisant séparément l'*Optimal Matching* pour chacune des dimensions – par exemple : professionnelle, conjugale et parentale − puis en sommant les distances obtenues (Han-Moen, 1999 ; Blanchard, 2010).

- Une autre critique est liée à l'ordre des événements (Wu, 2000). En effet, les opérations de substitution sont symétriques, c'est-à-dire que substituer A à B dans une séquence a le même coût que substituer B à A, et de ce fait ne tiennent pas compte de l'ordre des événements. Il convient tout d'abord de noter que si les probabilités de transition entre deux éléments ne sont pas symétriques (la probabilité de passer des études à l'emploi n'est pas nécessairement la même que celle de passer de l'emploi aux études), cela n'impose en rien aux opérations de substitution d'être asymétriques, dans la mesure où une substitution ne constitue pas une transition (Halpin, 2003) : les substitutions sont des opérations techniques, constitutives de la méthode, alors que les transitions sont des unités conceptuelles qui composent les trajectoires. Par ailleurs, de nouvelles formes d'analyse séquentielle, appelées *Non Alignment Techniques*, ont été développées pour traiter cette question. Elles n'emploient pas les opérations élémentaires de l'*Optimal Matching*. L'idée est de déterminer la similarité entre les séquences en comparant des paires d'éléments ordonnés des séquences. Les coefficients DT (Dijkstra et Taris, 1995) calculent le nombre de paires d'éléments ordonnés communes à deux séquences. Les différentes métriques introduites par H. Elzinga (2003 et 2006) peuvent en être vues comme une extension et sont basées sur les relations d'ordre entre paires d'éléments : plus long préfixe (début de séquence), longueur de la plus longue sous-séquence commune, nombre de sous-séquences communes…
- Une dernière critique concerne le fait que l'*Optimal Matching* ne prend pas en compte la direction du temps (Wu, 2000). En effet, la dépendance non linéaire des séquences par rapport au temps est négligée si les coûts de transformation sont identiques à chaque point de la séquence. L. Lesnard (2010) a proposé une mesure, appelée *Dynamic Hamming Matching*, qui résout ce problème en liant les coûts de substitution au temps. Une matrice de coûts de substitution distincte est calculée pour chaque moment de la séquence, à partir des probabilités de transition entre états à ce moment précis. Pour une paire de séquences, on obtient des distances pour chaque moment ; celles-ci sont ensuite additionnées pour obtenir la mesure globale de la distance entre les deux séquences. Cette technique a la particularité de ne pas utiliser d'opérations *indel*, donc de traiter uniquement des séquences de longueur identique. Elle se révèle particulièrement appropriée à

l'analyse d'emplois du temps, pour identifier les régularités des rythmes sociaux (Lesnard, 2010). La question de la dépendance non-linéaire au temps peut aussi être abordée par le biais du codage : par exemple, K. Stovel (2001) recode les trajectoires étudiées de manière à ce que la situation à un moment donnée prenne en compte les événements passés.

On signalera enfin quelques autres alternatives axées sur la manière de déterminer les coûts des opérations élémentaires de l'*Optimal Matching*. B. Halpin (2010) propose ainsi de modifier l'algorithme d'OM de façon à pondérer les opérations élémentaires par l'inverse de la longueur des épisodes. Le coût *indel* peut aussi être majoré selon la ressemblance entre l'élément inséré ou supprimé et les éléments voisins, ressemblance mesurée par le coût de substitution entre ces éléments (Hollister, 2009). J.A. Gauthier *et al.* (2009) utilisent quant à eux une procédure itérative pour déterminer les coûts.

4. Sans typologie, quel salut ?

Même si ce manuel est principalement consacré aux typologies de trajectoires, une analyse exploratoire des parcours de vie n'implique pas nécessairement la construction d'une typologie ; d'autres voies sont possibles, pour la compléter ou s'y substituer.

Par exemple, dans le cas des méthodes basées sur une analyse factorielle des trajectoires, l'examen des facteurs issus d'une analyse en composantes principales (ACP) ou d'une analyse des correspondances (AFC) est, à lui seul, porteur d'enseignements (Degenne *et al.*, 1995).

Ensuite, les dissimilarités entre trajectoires, calculées au moyen de l'une des mesures décrites dans la partie 3, peuvent être analysées pour elles-mêmes, sans avoir recours à une classification. Certains travaux s'intéressent à la distance entre les séquences individuelles et une séquence de référence (Scherer, 2001 ; Malo et Munoz-Bullon, 2003 ; Kogan, 2004). Celle-ci peut correspondre à la trajectoire la plus fréquente ou à une trajectoire « normale » construite théoriquement. Il s'agira alors de tenter d'expliquer l'écart à la normale : par exemple, quels sont les facteurs qui influent sur l'écart à une carrière professionnelle continue et à temps-plein ? On peut aussi comparer la distribution de la dissimilarité entre les parcours dans différentes sous-populations (selon le sexe, la cohorte de naissance…) : les parcours des femmes sont-ils plus divers que ceux des hommes ? Observe-t-on une « dé-standardisation » des parcours au fil des générations (Elzinga et Liefbroer, 2007; Robette, 2010) ?

R. Piccarreta et O. Lior (2010) proposent d'utiliser les « tapis » (*index plots*) associés à un échelonnement multidimensionnel (*Multi-Dimensional Scaling* ou MDS, voir note 10) comme outils graphiques pour l'exploration des séquences. À partir de la matrice de distance entre séquences, le MDS va permettre de réduire l'information contenue dans cette matrice pour en extraire les principaux facteurs propres. Les premiers facteurs sont ensuite utilisés pour trier les séquences, rendant ainsi leur représentation sous forme de « tapis » plus facilement interprétable. Cette approche peut être utile à différentes étapes de l'analyse (choix de la longueur des séquences, codage des états…), et notamment pour explorer les relations entre plusieurs dimensions des parcours de vie (familiale et professionnelle par exemple) ou entre les parcours et des variables externes (sexe, génération…).

Il existe par ailleurs plusieurs indicateurs de la complexité des séquences. L'*entropie longitudinale* (Gabadinho *et al.*, 2009) mesure l'occurrence et la distribution des durées des différents états dans la trajectoire. Elle est minimale lorsque l'intégralité de la trajectoire est passée dans le même état, et maximale lorsque la trajectoire est passée

par l'ensemble des états avec des durées identiques. C'est donc la prévalence des situations qui est considérée, mais l'entropie transversale ne tient pas compte de la fréquence des transitions ou de l'ordre des événements. La *turbulence* (Elzinga et Liefbroer, 2007), renommée depuis complexité (Elzinga, 2010), dénombre les sous-séquences distinctes en les pondérant éventuellement par la variance des durées de séjour dans les états successifs. L'indice de *complexité* (Gabadinho *et al.*, 2010a) est la moyenne géométrique de l'entropie longitudinale et du nombre de transitions dans la séquence.

De plus, il est possible d'extraire, à partir d'un ensemble de séquences, des sous-séquences fréquentes (ou motifs séquentiels) puis des règles d'association entre événements (Blockeel *et al.*, 2001; Ritschard et Oris, 2005 ; Gabadinho *et al.*, 2010b). Le but est alors principalement de comparer les séquences de plusieurs groupes d'individus.

Enfin, les arbres de décision (*decision trees*) ou les jeux de règles (*rule sets*) forment un ensemble d'outils susceptibles de s'adapter à l'exploration des parcours[31]. F.C. Billari *et al.* (2000) comparent ainsi le passage à l'âge adulte en Italie et en Autriche à partir des caractéristiques des trajectoires, avec des indicateurs décrivant l'occurrence, le calendrier et l'ordre des événements. A. Gabadinho *et al.* (2010c) utilisent une généralisation des principes de l'analyse de la variance (ANOVA) puis un arbre d'induction pour déterminer quelles sont les caractéristiques sociodémographiques qui discriminent le plus les dissimilarités entre parcours de vie.

On le voit, lorsqu'il ne s'agit pas d'explorer les régularités au sein des parcours au moyen de typologies, l'objectif est le plus souvent de comparer les caractéristiques des trajectoires entre plusieurs populations définies *a priori*.

[31] Comme l'illustrait déjà la classification de séquences par le biais du *monothetic divisive algorithm* dans la partie 1.6. Mais les méthodes présentées ici n'aboutissent pas au découpage de la population étudiée en groupes d'individus (autrement dit à une typologie de parcours). Par ailleurs, ces techniques sont aussi applicables aux trajectoires dans une perspective stochastique, centrée sur les événements : ce sont par exemple les *survival trees* (De Rose et Pallara, 1997).

5. Illustration sur des trajectoires professionnelles

5.1 Contexte et données

On s'intéresse dans cette application à la mobilité professionnelle intra-générationnelle. En particulier, on souhaite identifier les régularités au sein des carrières, en distinguant les trajectoires de stabilité de celles de mobilité.

Les données utilisées sont issues de l'enquête *Biographies et entourage*, collectée par l'Institut National d'Études Démographiques (INED) en 2001. Celle-ci retrace les histoires familiale, résidentielle et professionnelle de 2 830 Franciliens nés entre 1930 et 1950, ainsi que des membres de leur entourage (Lelièvre et Vivier, 2001). Le volet professionnel de l'enquête reconstitue la succession des activités occupées au cours de la vie, du premier emploi jusqu'au moment de l'enquête. Les différentes occupations, dont l'inactivité, ont été relevées selon un calendrier rétrospectif de dimension annuelle. Chaque étape a ensuite été codée selon une nomenclature des Professions et Catégories Socioprofessionnelles (PCS) s'inspirant largement de celle de l'Institut National de la Statistique et des Études Économiques (INSEE).

5.2 Les choix successifs

5.2.1 La population d'étude

Seules les carrières masculines seront analysées ici (n=1341). Une classification de l'ensemble de la population est envisageable mais présente plusieurs inconvénients. Tout d'abord, l'espace des états est différent pour les hommes et les femmes : le service militaire n'intervient ainsi que dans les itinéraires masculins. Certaines professions sont de plus fortement sexuées, par exemple les ouvriers pour les hommes ou les employées pour les femmes. Enfin, parmi ces générations, les femmes sont nombreuses à connaître des carrières incomplètes, avec des arrêts ou des interruptions d'activité, ce qui n'est pas le cas des hommes. Ces différences selon le sexe risquent alors de « brouiller » les résultats de la procédure de classification des trajectoires et la typologie de carrières construite perdrait en portée heuristique.

5.2.2 La longueur des séquences

Les individus sortent de l'observation à la date d'enquête : les données sont donc tronquées à droite. Si les modèles de durée sont aptes à contrôler l'effet des troncatures de manière satisfaisante, ce n'est pas le cas des méthodes exploratoires que nous employons. On choisit donc d'analyser les carrières individuelles observées sur une période identique, délimitée par les mêmes bornes. L'étude portera sur la mobilité professionnelle entre l'âge de 14 ans, qui marque la fin de la scolarité obligatoire pour les générations étudiées, et celui de 50 ans, qui est l'âge des enquêtés les plus jeunes au moment de l'enquête. On aurait pu continuer l'analyse au-delà de 50 ans en travaillant sur une sous-population de l'enquête. Toutefois, cette procédure apparaît peu intéressante ; 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. De plus, la plupart des transitions professionnelles intervient avant l'âge de 50 ans.

5.2.3 Les états retenus

Nous décrivons les différents états constituant les trajectoires à partir des groupes socioprofessionnels[32]. Par construction, l'échantillon ne contient pas de retraité puisque la description s'arrête à 50 ans. On a donc raisonné sur les six groupes d'actifs – agriculteurs exploitants ; artisans, commerçants et chefs d'entreprise ; cadres et professions intellectuelles supérieures ; professions intermédiaires ; employés ; ouvriers – 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 professionnelle pour d'autres raisons. On a ensuite ajouté un dernier groupe[33], les militaires du contingent, nécessaire pour rendre compte des parcours masculins pour ces générations qui ont notamment connu les guerres de décolonisation.

Notre échantillon se compose finalement de 1 341 trajectoires professionnelles, observées annuellement de 14 à 50 ans (donc 37 observations successives), et prenant pour valeur à chaque observation l'un des neufs états que l'on vient de définir. D'un point de vue purement théorique, il existe donc 937 trajectoires distinctes possibles. Empiriquement, certains enquêtés ont toutefois des trajectoires identiques, mais le nombre de trajectoires distinctes reste important : dans notre échantillon de 1 341 carrières, on observe ainsi 1 097 carrières distinctes. On voit bien que la diversité est ici trop grande pour espérer pouvoir classer « à la main » les trajectoires : il est nécessaire de faire appel à des méthodes plus élaborées.

[32] Qui correspondent aux huit postes de la nomenclature des PCS (Professions et Catégories Socioprofessionnelles.

[33] Issu de la nomenclature de 1954 mais aujourd'hui supprimé.

5.2.4 Choix et mise en œuvre de la méthode

Du fait que l'objectif est ici notamment de différencier les trajectoires de mobilité de celles de stabilité, on fait le choix d'employer l'*Optimal Matching Analysis*. Cette technique a en effet montré dans de précédents travaux qu'elle était mieux à même d'identifier les régularités liées aux *transitions* qu'une méthode factorielle, qui met plus l'accent sur les durées, même si les différences restent relativement marginales (Robette et Thibault, 2008).

Ayant choisi l'*Optimal Matching*, l'étape suivant consiste à fixer les coûts de substitution et d'insertion-suppression. Contrairement à d'autres nomenclatures de professions (par exemple, en Grande-Bretagne), la nomenclature des PCS n'implique pas de hiérarchie entre les catégories. Il semble donc difficile de baser les coûts de substitution sur une échelle *théorique* entre les états, comme on pourrait l'envisager avec des niveaux de diplôme dans une trajectoire scolaire, par exemple. On décide donc de déterminer les coûts de substitution en fonction des estimateurs *empiriques*, à partir des probabilités de transition entre états observées dans l'échantillon (tableau 12). Les transitions entre les états « agriculteurs » et « ouvriers » étant relativement plus fréquentes que celles entre « ouvriers » et « cadres », le coût de substitution entre « agriculteurs » et « ouvriers » est ainsi moins élevé (1,895) que celui entre « ouvriers » et « cadres » (1,997). Pratiquement, le coût de substitution entre un état A et un état B est ici égal à 2 auquel on soustrait les probabilités de transition entre A et B et entre B et A observées dans l'échantillon de trajectoires (Rohwer et Pötter, 2005). Les probabilités de transition de « cadre » à « ouvrier » ou d'« ouvrier » à « cadre » étant très faibles, le coût de substitution en ces deux états est très proche de 2 (1,997). À l'inverse, la mobilité entre les états « agriculteurs » et « ouvriers » est relativement forte, le coût de substitution correspondant est donc plus faible (1,895).

Tableau 12 − Matrice des coûts de substitution

	Agriculteurs	Artisans etc.	Cadres	Prof. interm.	Employés	Ouvriers	Étudiants	Inactifs	Service militaire
Agriculteurs	0,000	1,992	2,000	2,000	1,990	1,895	1,999	2,000	1,954
Artisans, etc.	1,992	0,000	1,990	1,986	1,987	1,978	1,997	1,987	1,991
Cadres	2,000	1,990	0,000	1,971	1,990	1,997	1,972	1,964	1,912
Prof. interm.	2,000	1,986	1,971	0,000	1,961	1,976	1,966	1,948	1,853
Employés	1,990	1,987	1,990	1,961	0,000	1,970	1,972	1,962	1,896
Ouvriers	1,895	1,978	1,997	1,976	1,970	0,000	1,947	1,905	1,782
Étudiants	1,999	1,997	1,972	1,966	1,972	1,947	0,000	1,986	1,947
Inactifs	2,000	1,987	1,964	1,948	1,962	1,905	1,986	0,000	1,980
Service militaire	1,954	1,991	1,912	1,853	1,896	1,782	1,947	1,980	0,000

Le coût *indel* (insertion et suppression) permet d'arbitrer entre les différents types de régularités que l'on va pouvoir observer, c'est-à-dire entre contemporanéité des situations et ordre des événements qui composent les trajectoires. On fait ici un choix intermédiaire : le coût *indel* est fixé à une valeur relativement peu élevée, légèrement supérieure à la moitié du coût de substitution minimal (soit 1,1), afin de prendre en compte conjointement la contemporanéité et l'ordre.

En ce qui concerne la méthode de classification, l'option adoptée ici est la plus « usuelle » : on utilise une Classification Ascendante Hiérarchique (CAH)[34] et le critère d'agrégation de Ward.

L'analyse est réalisée à l'aide du logiciel R. En annexe 2 figure le programme commenté correspondant, en espérant que sa brièveté convaincra le lecteur de la relative facilité de mise en œuvre des approches typologiques des trajectoires.

5.3 Résultats

On examine ensuite différentes typologies de trajectoires, en commençant avec deux classes puis en augmentant graduellement ce nombre. Dans un premier temps, une typologie en cinq classes est jugée satisfaisante : ces dernières semblent relativement homogènes et clairement distinctes les unes des autres.

5.3.1 Une typologie en cinq classes

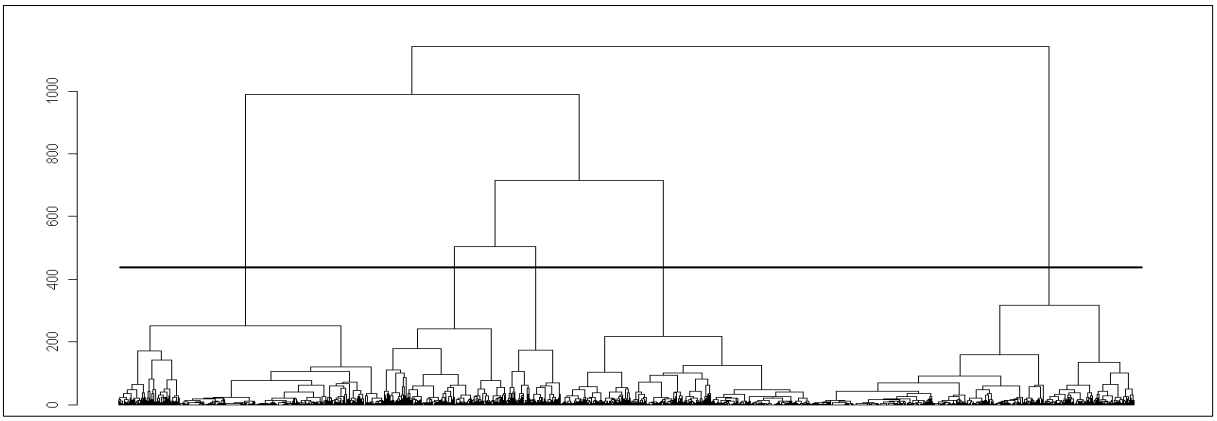
On s'aide pour cela de l'examen du dendrogramme (l'arbre de la classification ; voir figure 4a) et des sauts d'inertie entre les différents niveaux de partition (voir figure 4b) : lorsque la différence d'inertie entre deux niveaux voisins de partition est importante, cela signifie qu'une typologie avec une classe supplémentaire apporte un supplément d'information non négligeable. À l'inverse, lorsque la différence d'inertie est peu élevée, la plus-value à augmenter le nombre de classes de la typologie est statistiquement faible.

Ces classes semblent se caractériser par une certaine stabilité dans la profession occupée le plus longtemps : c'est la durée passée dans les différents états qui a dicté les regroupements (voir figure 5). La première classe se compose principalement d'hommes ayant occupé un emploi de cadre et profession intellectuelle supérieure tout au long de leur carrière (33 %, voir tableau 13), la deuxième classe est constituée de professions intermédiaires (27 %), la troisième d'ouvriers (26 %), la quatrième d'employés (9 %) et la cinquième d'artisans, commerçants ou chefs d'entreprise (5 %).

[34] Sans consolidation par les centres mobiles ou les *k-means*, afin de pouvoir explorer différents niveaux de partition.

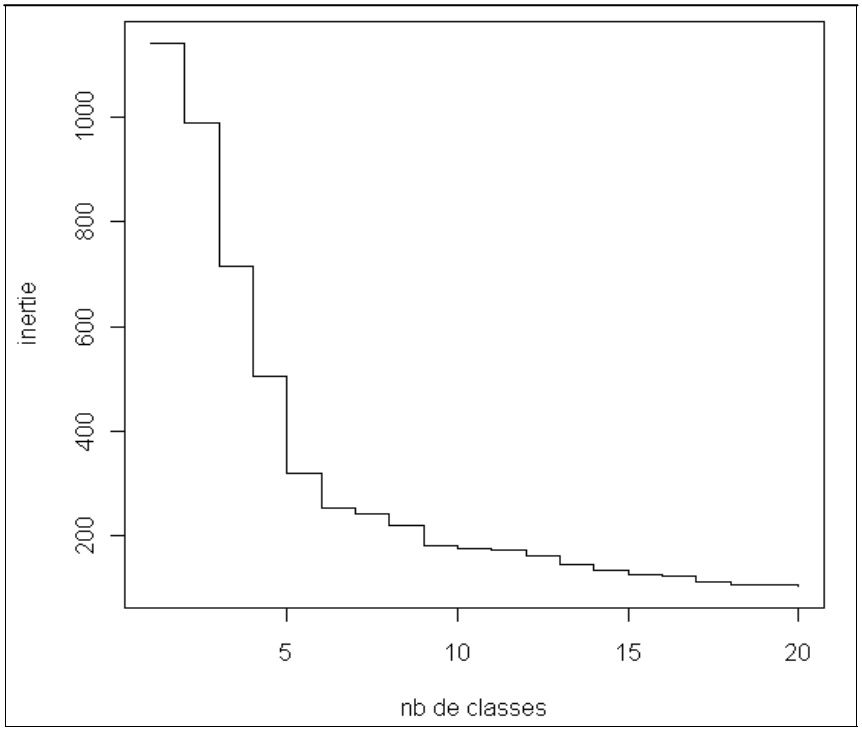
Cette dernière classe présente une particularité : les individus qui en font partie ont occupé au début de leur carrière une profession différente de celle qui représente la majeure partie de la trajectoire.

Figure 4a − Dendrogramme de la classification



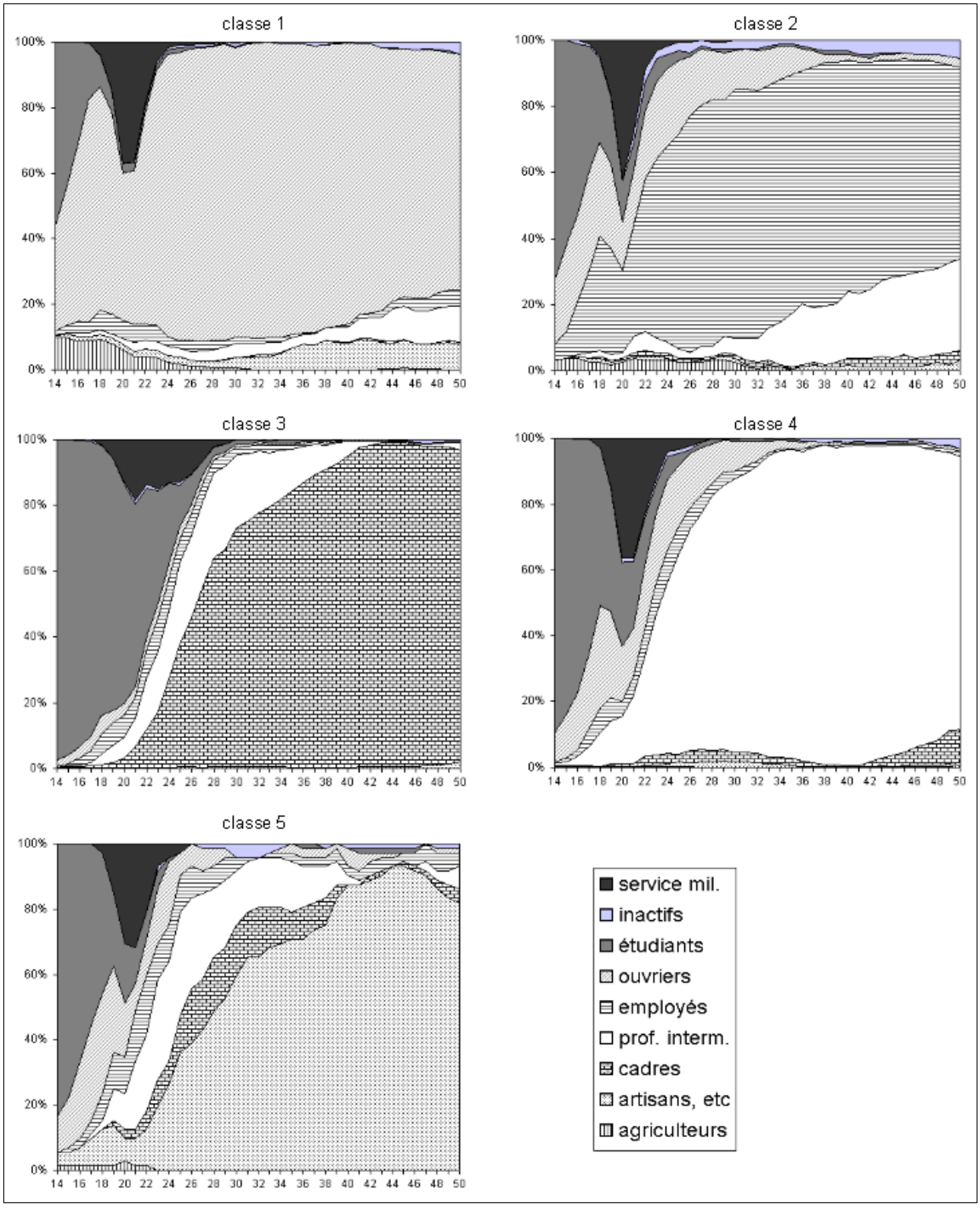
Source : *Biographies et entourage* (2001) ; Champ : 1 341 hommes.

Figure 4b − Inertie de la partition selon le nombre de classes



Source : *Biographies et entourage* (2001) ; Champ : 1 341 hommes.

Figure 5 − Chronogrammes de la typologie en cinq classes[35]



Source : *Biographies et entourage* (2001) ; Champ : 1 341 hommes.

[35] Du fait du nombre élevé d'états (neuf), l'impression en noir et blanc d'un graphique en aplats de couleurs est problématique. Les chronogrammes présentés ici ont donc été retravaillés sous Excel afin d'utiliser des motifs (hachures…). Les *index plots* sont quant à eux pratiquement illisibles en noir et blanc au-delà de trois ou quatre états : on ne les présentera pas dans ce manuel. Toutefois, chronogrammes et *index plots* en couleur, relatifs au présent exemple et obtenus avec le logiciel R, sont téléchargeables à l'adresse suivante :
http://nicolas.robette.free.fr/Docs/Graphes_couleur_manuel.zip
Pour l'utilisateur, le choix de graphiques en couleur ou en noir et blanc est donc à considérer selon le type de compte rendu de recherche envisagé (diaporama, article pour une revue…).

Tableau 13 − Typologie des carrières professionnelles en cinq classes

Classe	Effectif	%
cadres et professions intellectuelles supérieures	437	33
professions intermédiaires	362	27
ouvriers	352	26
employés	117	9
artisans, commerçants et chefs d'entreprise	73	5
Total	*1 341*	*100*

Source : *Biographies et entourage* (2001) ; Champ : 1 341 hommes.

Un moyen de donner plus de « chair » à cette typologie consiste à incarner les classes en décrivant la trajectoire du parangon de chaque classe, c'est-à-dire de l'individu le plus proche du centre de la classe. Pratiquement, on le trouve à partir de la matrice de distance, en identifiant l'individu de l'échantillon dont la distance moyenne aux autres membres de sa classe est la moins élevée. On peut alors « incarner » les parangons en proposant leur « portrait » à partir des données de l'enquête.

Classe des cadres et professions intellectuelles supérieures

À la fin de son service militaire en coopération au Sénégal, à l'âge de 26 ans, Michel y devient enseignant contractuel en mathématiques. Il est titularisé neuf années plus tard, puis part au Cameroun. À 45 ans, il rentre en France, en région parisienne, où il exercera au collège puis au lycée.

Classe des professions intermédiaires

Francis commence sa carrière à Paris, à 19 ans, en tant qu'outilleur dans une entreprise de téléphonie. Il devient chef d'atelier dans une bijouterie à 23 ans, puis rapidement professeur d'atelier.

Classe des ouvriers

Né en Espagne, Alberto y commence à travailler en tant qu'ouvrier dans une scierie. Deux ans plus tard, il quitte son pays pour Nancy, où il travaille toujours dans une scierie. Après deux ans de service militaire en Espagne, il emménage en région parisienne, où il exercera les professions de manutentionnaire dans une usine agro-alimentaire, monteur de cloisons et chauffeur magasinier dans l'assistance publique, après un cours intermède de chômage suite à un licenciement économique.

Classe des employés

Bernard est maroquinier à Paris dès l'âge de 17 ans. Sept années plus tard, il devient employé de presse dans une entreprise de messagerie. Suite à une invalidité, il arrêtera de travailler quelques années avant la retraite.

Classe des artisans, commerçants et chefs d'entreprise

Alain exerce toute sa carrière dans l'entreprise familiale de boutons ; il y commence comme représentant à 21 ans, avant d'en prendre la direction et la gérance à 28 ans.

5.3.2 Description à l'aide d'indicateurs

Un examen plus détaillé des classes par le biais d'indicateurs apporte un éclairage supplémentaire sur la typologie (tableau 14). On retrouve avec les indicateurs de durée moyenne ou d'occurrence des états (au moins un épisode dans l'état concerné), la caractérisation des classes par une profession principale, occupée la majeure partie de la trajectoire. Mais alors que les individus de la classe 3 sont ouvriers 28 années, en moyenne entre 14 et 50 ans, ceux de la classe 1 ne sont cadres ou professions intellectuelles supérieures que 22 ans en moyenne. Cela s'explique principalement par les études : la classe des cadres y a passé huit ans, pour moins de deux années pour les ouvriers (ce qui apparaissait déjà sur les chronogrammes). Ces derniers ne sont d'ailleurs que 57 % à avoir poursuivi des études après 14 ans. Par ailleurs, deux tiers des hommes de la classe des employés ont aussi connu un ou plusieurs épisodes en tant qu'ouvriers. Ceux sont également eux qui sont le plus fréquemment et le plus longtemps en inactivité. Enfin, les chronogrammes indiquaient que les individus de la classe 5 ne rejoignent la catégorie d'artisan, commerçant ou chef d'entreprise qu'après une période dans une autre catégorie. Les indicateurs montrent qu'aucune autre profession ne se détache significativement concernant le début de carrière : 52 % des hommes de cette classe ont connu un ou plusieurs épisodes de professions intermédiaires, 51 % d'ouvriers, 34 % d'employés et 25 % de cadres. Les chemins qui mènent vers la profession d'artisan, commerçant ou chef d'entreprise sont multiples, ce qui explique l'hétérogénéité de la classe – indiquée par l'entropie et la distance intra-classe – ou le niveau relativement élevé du nombre moyen de transitions au cours de la trajectoire.

Tableau 14 − Description de la typologie en cinq classes par des indicateurs

Classes		1	2	3	4	5
Durée moyenne (en années)	Agriculteurs	0,0	0,1	0,8	0,8	0,1
	Artisans, etc.	0,2	0,2	1,7	0,4	**19,9**
	Cadres	**22,1**	0,7	0,0	0,4	2,2
	Prof. intermédiaires	3,8	**23,8**	1,8	0,9	4,3
	Employés	1,1	2,9	1,0	**23,5**	2,1
	Ouvriers	0,7	3,0	**28,3**	5,5	2,7
	Étudiants	8,0	4,7	1,8	3,2	4,1
	Inactivité	0,1	0,3	0,3	1,2	0,4
	Service militaire	1,1	1,3	1,2	1,1	1,2
au moins 1 épisode (%)	Agriculteurs	0	1	12	7	3
	Artisans, etc.	5	5	18	12	**100**
	Cadres	**100**	17	1	7	25
	Prof. intermédiaires	48	**100**	22	22	52
	Employés	18	35	20	**97**	34
	Ouvriers	11	46	**100**	67	51
	Étudiants	98	90	57	69	84
	Inactivité	5	10	10	18	11
	Service militaire	62	70	61	56	66
nombre de transitions		3,0	3,5	2,9	3,5	3,9
entropie		0,31	0,36	0,35	0,43	0,47
distance intra-classe		20,8	23,2	21,9	28,0	31,8

Source : *Biographies et entourage* (2001) ; Champ : 1 341 hommes.

Ces différents moyens de description de la typologie font apparaître le profil des classes. Il faut cependant garder à l'esprit que chaque classe présente toujours une certaine hétérogénéité, comme l'illustre le cas de la classe 5. Cette hétérogénéité est aussi visible à l'examen des *index plots* (les « tapis », non représentés ici, voir note 35). On constate par exemple que dans la classe 2, les hommes semblent effectivement avoir occupé une profession intermédiaire pendant la majeure partie de leur trajectoire, mais pas tous pendant l'intégralité de leur carrière (hors études et service militaire). En effet, une partie des individus de la classe (en haut du graphique) ont commencé leur carrière en tant qu'ouvrier, pendant une durée plus ou moins longue. On retrouve ce

phénomène dans les autres classes, avec d'autres catégories socioprofessionnelles. Cette typologie en cinq classes fait donc avant tout émerger les régularités en termes de durée dans une catégorie de professions, qui structurent le plus fortement les trajectoires de l'échantillon. En revanche, elle laisse de côté la distinction entre trajectoires stables et trajectoires en mobilité. Pour voir apparaître ces différences, il est indispensable d'analyser la classification à un niveau plus fin, autrement dit de construire une typologie en un nombre plus élevé de classes.

5.3.3 Une typologie en dix classes

Une typologie en dix classes permet ainsi nettement mieux d'appréhender la distinction entre trajectoires stables et trajectoires de mobilité (tableau 15). Les quatre premières classes de la typologie en cinq classes se scindent en effet en deux ou trois sous-classes : la première, la plus importante en effectif, regroupe des trajectoires stables dans la catégorie socioprofessionnelle ; l'autre (ou les autres) étant composée(s) de trajectoires de mobilité professionnelle, le plus souvent *vers* la catégorie principale, c'est-à-dire occupée le plus longtemps. Par exemple, la classe de cadres regroupe une sous-classe d'hommes ayant été cadres tout au long de leur carrière (26 %) et une autre d'individus ayant entamé leur trajectoire professionnelle par une profession intermédiaire avant de devenir cadre (6 %). De même, la classe des professions intermédiaires se décompose en trajectoires stables (16 %) et en trajectoires de mobilité d'ouvrier (6 %) ou d'employé (5 %) vers profession intermédiaire ; la classe des ouvriers en trajectoires stables (19 %) et en trajectoires de mobilité d'ouvrier vers artisan, commerçant ou chef d'entreprise ou profession intermédiaire (7 %) ; la classe des employés en trajectoires stables (6 %) et en trajectoires de mobilités d'ouvrier vers employé. Au final, en prenant en compte la classe 5 d'artisans, commerçants ou chefs d'entreprise qui ont connu une courte période dans une autre catégorie en début de carrière, la mobilité professionnelle représente 31 % des trajectoires, et elle est le plus souvent ascendante.

Par ailleurs, on peut remarquer qu'on trouve certaines formes de mobilité dans plusieurs sous-classes. Par exemple, les trajectoires d'ouvriers à profession intermédiaire sont présentes dans des sous-classes issues des classes de professions intermédiaires mais aussi d'ouvriers. Elles ne sont cependant pas identiques : dans le premier cas, la mobilité ascendante intervient en début de carrière, le plus souvent avant 35 ans, alors que dans le second la mobilité a lieu après 40 ans. On voit donc que cette nouvelle typologie en dix classes prend à la fois en compte l'existence ou non de mobilité professionnelle, le type de mobilité mais aussi son calendrier.

Tableau 15 − Typologie des carrières professionnelles en dix classes

Typologie en cinq classes	Typologie en dix classes	Effectif	%
Cadres et professions intellectuelles supérieures	Cadres	351	26
	Prof. interm. → cadres	86	6
Professions intermédiaires	Professions intermédiaires	208	16
	Ouvriers → prof. interm.	84	6
	Employés → prof. interm.	70	5
Ouvriers	Ouvriers	258	19
	ouvriers → artisans, etc./prof. interm.	94	7
Employés	Employés	87	6
	Ouvriers → employés	30	2
Artisans, commerçants et chefs d'entreprise	Artisans, commerçants et chefs d'entreprise	73	5
Total		*1 341*	100

Source : *Biographies et entourage* (2001) ; Champ : 1 341 hommes.

6. Conclusion

Les méthodes d'analyse séquentielle, en particulier l'*Optimal Matching*, se sont largement diffusées depuis leur introduction dans les sciences sociales dans les années 1980, spécialement depuis une dizaine d'années et un numéro spécial de la revue *Sociological Methods and Research* en 2000[36]. C'est beaucoup moins le cas des techniques basées sur une analyse factorielle, sans doute surtout du fait du peu d'écho rencontré par l'analyse des données « à la française » dans la recherche anglo-saxonne. D'une manière générale, les approches exploratoires restent en retrait dans les travaux quantitatifs en sciences sociales. Les méthodes présentées ici constituent pourtant un ensemble d'outils fort utiles pour toute recherche s'intéressant à des données longitudinales complexes telles que les parcours de vie individuels. Elles permettent en particulier d'explorer les données, de les décrire et d'identifier des régularités, ainsi que d'effectuer des comparaisons entre plusieurs populations. Elles sont à même de traiter des types de données extrêmement variés, des trajectoires les plus simples à des parcours multidimensionnels ou des trajectoires liées[37]. On notera d'ailleurs qu'elles peuvent s'appliquer à toute suite d'éléments ordonnés, et pas seulement à des parcours individuels, comme le montrent des travaux sur le lynchage (Stovel, 1991) ou les danses folkloriques (Abbott et Forrest, 1986). Cette souplesse d'utilisation et les nombreux choix que ces méthodes impliquent en forment l'un des principaux atouts : ces choix obligent à s'interroger et à rendre explicites l'adéquation entre les hypothèses et questions de recherches et la démarche adoptée.

On l'a vu, les diverses méthodes de construction de typologies de trajectoires présentent chacune des spécificités, des qualités et des limites. Se pose alors la question de la robustesse des résultats. À l'heure actuelle, il n'existe pas de travaux confrontant de manière systématique la majeure partie des techniques disponibles. Cependant, dans les applications empiriques, plusieurs alternatives sont souvent comparées, du choix de la mesure de dissimilarité à celui d'une méthode de classification (par exemple, voir Grelet, 2002 ; Robette et Thibault, 2008 ; Anyadike-Danes et McVicar, 2010). Les résultats apparaissent le plus souvent cohérents : les

---

[36] Abbott prédisait alors de manière provocatrice que les sceptiques doutant que ces techniques s'imposent parmi les techniques de base des sciences sociales dans un horizon de 25 ans seraient fort surpris (Abbott, 2000). L'avenir dira si ces paroles tenaient de la prophétie ou du vœu pieu.

[37] Sur les trajectoires individuelles composées de plusieurs dimensions (familiale, professionnelle…), voir partie 3.2.2. Sur l'analyse conjointe de trajectoires de plusieurs individus liés, voir par exemple Robette *et al.* (2009) ou Lelièvre et Robette (2010) pour les trajectoires de couples, ou Robette *et al.* (2011) pour les trajectoires de parents et de leurs enfants.

parcours de vie sont suffisamment structurés pour faire émerger de manière relativement similaire les régularités essentielles qu'ils renferment. Autrement dit, les principaux types de parcours des typologies sont généralement identiques. Les différences concernent à la marge les effectifs des classes (les individus les plus périphériques d'une classe pouvant plus aisément passer dans une classe voisine), ou la nature des classes les plus marginales[38]. Certains comportements atypiques émergeront parfois selon la méthode employée, d'autant plus que l'on découpera la population en un nombre élevé de classes. Une démarche saine consisterait à choisir la méthode la plus appropriée à la problématique de la recherche et aux données disponibles, puis à s'assurer de la validité des résultats en adoptant des choix méthodologiques légèrement différents. Mais il convient de garder à l'esprit que le problème de la robustesse ne se pose pas avec la même acuité lorsqu'il s'agit d'explorer que lorsqu'il s'agit d'expliquer et de prédire. Et pour achever de démystifier cette question : « il vaut bien mieux une réponse approximative à la bonne question, qui est souvent vague, qu'une réponse exacte à une mauvaise question, que l'on peut toujours rendre précise ("*Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise*" », Tukey, 1962 :13).

Lorsque l'on a achevé l'exploration des parcours de vie individuels et caractérisé la typologie retenue, un nouvel enjeu apparaît : comment poursuivre l'analyse ? Macindoe et Abbott (2004) identifiaient deux directions : déterminer, d'une part, les facteurs qui produisent les parcours typiques et, d'autre part, les conséquences de ces facteurs. Intuitivement, l'on est tenté de répondre à ces questions au moyen de régressions. Dans le premier cas de figure, la classe de trajectoires sera considérée comme variable dépendante dans une régression logistique, par exemple. On modélise ainsi la probabilité d'appartenir à une classe de trajectoires donnée : par exemple, Anyadikes et McVicar (2010) montrent l'impact du contexte familial et des expériences scolaires sur le début des trajectoires d'activité de femmes en Grande-Bretagne. Cependant, les variables explicatives potentielles sont en nombre limité : du fait de la dimension causale de l'analyse, elles ne peuvent être que des caractéristiques constantes (sexe, lieu de naissance, etc.) ou antérieures au début de la trajectoire (la profession des parents pendant l'enfance de l'enquêté si l'on s'intéresse aux carrières professionnelles, etc.). Si l'on souhaite prendre en compte un spectre de facteurs plus large, il est nécessaire de renoncer en partie à la causalité, pour adopter des techniques d'analyse plus descriptives, des simples distributions aux analyses factorielles. L'antagonisme entre description et inférence est toutefois à relativiser, tant bien souvent la relation de causalité sous-jacente aux modèles est rendue délicate par les anticipations ou le « temps flou » (GRAB, 2006) : il est alors plus juste de parler d'interdépendance (Courgeau et Lelièvre, 1989). Si l'on aborde maintenant les

[38] Robette et Thibault (2008) montrent ainsi que l'*Optimal Matching* semble particulièrement efficace pour distinguer stabilité et mobilité des trajectoires, alors que l'Analyse Harmonique Qualitative donne plus de poids aux durées passées dans les états et fait émerger des trajectoires atypiques. Cette différence peut d'ailleurs aisément s'expliquer par le fonctionnement de chacune de ces techniques, l'ordre des éléments qui composent une trajectoire étant explicitement pris en compte par l'*Optimal Matching* à travers les opérations d'insertion et de suppression, et la durée passée dans les situations par l'Analyse Harmonique Qualitative à travers le codage des données.

conséquences d'un type de parcours donné, la classe de trajectoires est alors considérée comme variable explicative d'une caractéristique ou d'un phénomène postérieurs à la trajectoire, par exemple la position professionnelle en fin de carrière en fonction de la trajectoire d'insertion. Cette direction n'a encore été que peu suivie (Mouw, 2005 ; Aassve *et al.*, 2010).

Pour finir, l'opposition entre approches longitudinales dites « atomistes », qui analysent une transition particulière, et « holistes », qui étudient l'ensemble de la trajectoire, est réelle dans le sens où l'une est causale et probabiliste alors que l'autre est descriptive et (souvent) exploiratoire. Elles répondent à des questions différentes : la première s'interroge de l'impact sur le risque d'occurrence d'un événement donné, de caractéristiques ou d'autres événements ; la seconde vise à identifier les principaux éléments qui différencient les parcours de vie pris comme un tout (Billari, 2005). Mais, on l'a vu dans le paragraphe précédent, les applications empiriques de ces deux approches ne s'excluent pas nécessairement. Tout d'abord, les questions de recherche et la nature des données disponibles dictent souvent le choix de la méthode la plus adaptée. Mais surtout, les deux approches peuvent (et gagneraient sans doute à) être envisagées comme complémentaires (Bry et Antoine, 2004). Par exemple, il est parfois utile d'explorer et de synthétiser des parcours complexes au moyen de méthodes descriptives avant d'énoncer des hypothèses, que des méthodes explicatives viendront valider ou réfuter. L'exploration des parcours de vie peut avantageusement compléter des techniques stochastiques, en amont ou en aval de la recherche des causes et des effets des processus. Les modalités de cette articulation restent encore très largement à défricher, laissant la porte ouverte à de futures recherches d'innovations méthodologiques en sciences sociales.

Références bibliographiques

Aassve Arnstein, Billari Francesco C., Piccarreta Raffaella, 2007 - "Strings of adulthood: a sequence analysis of young british women's work-family trajectories". *European Journal of Population*, 23 (3-4) : 369-388.

Aassve Arnstein, Piccarreta Raffaella, Robette Nicolas, 2010 - "Consequences of new demographic behaviour on intergenerational relationships". Communication au *Multilinks Meeting*, 25 mars, Milan, Università Bocconi.

Abbott Andrew, 1995 - "Sequence analysis: new methods for old ideas". *Annual review of sociology*, 21 : 93-113.

Abbott Andrew, 2000 - "Reply to Levine and Wu" *Sociological methods et research*, 29 (1) : 65-76.

Abbott Andrew, Barman Emily, 1997 - "Sequence comparison via alignment and Gibbs sampling". *Sociological methodology*, 27 : 47-87.

Abbott Andrew, Forrest John, 1986 - "Optimal Matching Methods for Historical Sequences". *Journal of Interdisciplinary History*, 16 (3) : 471-494.

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

Abbott Andrew, Tsay Angela, 2000 - "Sequence analysis and optimal matching methods in sociology: Review and prospect". *Sociological methods et research*, 29 (1) : 3-33.

Aisenbrey Silke, Fasang Anette E., 2010 - "New Life for Old Ideas: The ''Second Wave'' of Sequence Analysis Bringing the ''Course'' Back Into the Life Course". *Sociological Methods et Research*, 38 (3) : 420-462.

Allison Paul David, 1984 - *Event history analysis: regression for longitudinal event data*. Beverly Hills, CA, Sage (coll. Quantitative applications in the social sciences), 46, 87 p.

Antoine Philippe, Bocquier Philippe, Marcoux Richard, Piché Victor - 2006, *L'expérience des enquêtes biographiques en Afrique*. Chaire Quételet « Les systèmes d'information en démographie et en sciences sociales. Nouvelles questions, nouveaux outils ? », Louvain-la-Neuve, Belgique, 17 p.

Anyadike-Danes Michael, McVicar Duncan, 2010 - "My Brilliant Career: Characterizing the Early Labor Market Trajectories of British Women From Generation X". *Sociological Methods et Research*, 38 (3) : 482-512.

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

Barbary Olivier, Pinzon Sarmiento Luz Mary, 1998 - « L'analyse harmonique qualitative et son application à la typologie des trajectoires individuelles ». *Mathématiques, Informatique et Sciences Humaines*, (144) : 29-54.

Béduwé Catherine, 2001 - Trajectoires-type : une méthode pour l'étude des mobilités professionnelles, *in 8èmes Journées d'études Céreq-Lasmas IdL « Construction et usage des catégories d'analyse »*, 17 et 18 mai, Marseille : 1-14.

Béduwé Catherine, Dauty Françoise, Espinasse Jean-Michel, 1995 - Trajectoires types d'insertion professionnelle. Application au cas des bacheliers professionnels de Midi-Pyrénées, *in Deuxièmes journées d'étude Céreq-Lasmas-IdL « L'analyse longitudinale du marché du travail »*, 28 et 29 juin, Caen, Céreq : 7-29.

Belbin L., Faith D., Milligan G.W., 1992 - "A Comparison of Two Approaches to Beta-Flexible Clustering". *Multivariate Behavioral Research*, 27 : 417-433.

Billari Francesco C., 2001 - "Sequence analysis in demographic research". *Canadian Studies in Population*, 28 (2) : 439-458.

Billari Francesco C., 2005 - "Life course analysis: two (complementary) cultures? Some reflections with examples from the analysis of the transition to adulthood". *Advances in life course research*, 10 : 261-281.

Billari Francesco C., Fürnkranz J., Prskawetz A., 2000 - *Timing, sequencing, and quantum of life events: A machine learning approach*. Rostock, Max-Planck-Institute for Demographic Research, Working Paper n° 10.

Billari Francesco C., Piccarreta Raffaella, 2005 - "Analyzing demographic life courses through sequence analysis ». *Mathematical population studies*, (12) : 81-106.

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

Blanchard Philippe, 2010 - *Analyse séquentielle et carrières militantes*. Rapport de recherche, 166 p. http://halshs.archives-ouvertes.fr/hal-00476193/

Blockeel H., Fürnkranz J., Prskawetz A., Billari Francesco C., 2001 - "Detecting temporal change in event sequences: An application to demographic data", *in* Raedt L.D., Siebes A. (eds), *Principles of data mining and knowledge discovery: 5th European conference*, PKDD 2001, Freiburg in Brisgau, Springer, vol. 2168 of LNCS : 29-41.

Blossfeld Hans-Peter, Rohwer Götz, 2002 - *Techniques of event history modeling: new approaches to causal analysis*? Mahwah, NJ, L. Erlbaum, 310 p.

Bocquier Philippe, 1996 - « Antériorité, causalité et corrélation », *in* Bocquier Philippe, *Analyse des enquêtes biographiques*, (coll. Documents et manuels du CEPED).

Breiman Leo, 2001 - "Statistical Modeling: The Two Cultures". *Statistical Science*, 16 (3) : 199-231.

Briard Karine, 2007 - « Profils types des salariés du secteur privé : approche par une classification des carrières ». *Economie et prévision*, (180-181) : 59-85.

Bringé Arnaud, Laurent Raphaël, 2005 - *Reconstituer des histoires individuelles à partir de données de suivi démographique*. Collection du CEPED, série « les Clefs pour », 85 p.

Bry Xavier, 1995 - Analyses factorielles simples, Paris, Economica (coll. Techniques quantitatives - poches), 112 p.

Bry Xavier, 1996 - *Analyses factorielles multiples*. Paris, Economica (coll. Techniques quantitatives - poches), 112 p.

Bry Xavier, Antoine Philippe, 2004 - « Explorer l'explicatif : application à l'analyse biographique ». *Population*, (6) : 909-946.

Brzinsky-Fay Christian, 2007 - "Lost in transition: labour market entry sequences of school leavers in Europe". *European Sociological Review*, 23 (4) : 409-422.

Brzinsky-Fay Christian, Kohler Ulrich, Luniak Magdalena, 2006 - "Sequence analysis with Stata". *The Stata Journal*, 6 (4) : 435-460.

Chan Tak Wing, 1995 - "Optimal matching analysis: a methodological note on studying career mobility". *Work and occupations*, 22 (4) : 467-490.

Charlot A., Pottier F., 1987 - « L'université et l'emploi : des relations stables entre deux milieux en évolution ». *Formation Emploi*, (18) : 82-100.

Cottrell Marie, Ponthieux Sophie, 2002 - « Classification neuronale et analyse des données "traditionnelle" : quelques applications aux conditions de vie des ménages ». *INSEE Méthodes*, (101) : 7-37.

Courgeau Daniel, Baccaïni Brigitte, 1997 - « Analyse multi-niveaux en sciences sociales ». *Population*, (4) : 831-863.

Courgeau Daniel, Guérin-Pace France, 1998 - « Le suivi des itinéraires professionnels des couples par les méthodes de la statistique textuelle. Lecture des parcours professionnels des couples », *in JADT 1998, 4èmes journées internationales d'analyse des données Textuelles*, Université de Nice - Sophia Antipolis : 221-232.

Courgeau Daniel, Lelièvre Eva, 1986 - « Nuptialité et agriculture ». *Population*, (2) : 303-326.

Courgeau Daniel, Lelièvre Eva, 1989 - *Analyse démographique des biographies*. Paris, INED, 269 p.

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

Degenne Alain, Lebeaux Marie-Odile, Mounier Lise, 1994 - « Essai d'une typologie des cheminements d'entrée dans la vie active », *in L'analyse longitudinale du marché du travail*. CEREQ (coll. Documents séminaires; n° 99) : 287-296.

Degenne Alain, Lebeaux Marie-Odile, Mounier Lise, 1995 - « Construction d'une typologie de trajectoires à partir de l'enquête de suivi des jeunes des niveaux V, Vbis et VI », *in Deuxièmes journées CEREQ-LASMAS-IDL sur l'analyse longitudinale du marché du travail* ; CEREQ CNRS.

Degenne Alain, Lebeaux Marie-Odile, Mounier Lise, 1996 - « Typologies d'itinéraires comme instrument d'analyse du marché du travail », *in* Degenne Alain, Mansuy Michèle, Podevin Gérard, Werquin Patrick, (eds), *Typologie des marchés du travail, suivi et parcours*, 23 et 24 mai, Rennes, (coll. Documents séminaire CEREQ), 115 : 27-42.

Delaunay Daniel, Lelièvre Eva, 2006 - « Examen topographique des transitions biographiques complexes à l'aide des cartes de Kohonen », *in* GRAB (Groupe de Réflexion sur l'Approche Biographique), *Etats flous et trajectoires complexes. Observation, modélisation, interprétation*. Paris, INED-CEPED (coll. Méthodes et savoirs), 5 : 219-238.

De Rose Alessandra, Pallara Alessandro, 1997 - "Survival Trees: An Alternative Non-Parametric Multivariate Technique for Life History Analysis". *European Journal of Population*, 13 (3) : 223-241.

Deville Jean-Claude, 1974 - « Méthodes statistiques et numériques de l'analyse harmonique ». *Annales de l'INSEE*, (15) : 3-101.

Deville Jean-Claude, 1977 - « Analyse harmonique du calendrier de constitution des familles en France. Disparités sociales et évolution de 1920 à 1960 ». *Population*, (1) : 17-63.

Deville Jean-Claude, 1982 - « Analyses de données chronologiques qualitatives : comment analyser des calendriers ? ». *Annales de l'INSEE*, (45) : 45-104.

Deville Jean-Claude, Saporta Gilbert, 1980- « Analyse harmonique qualitative », *in* Diday Edwin (éds), *Data analysis and informatics*. Amsterdam, North Holland Publishing : 375-389.

Diagne Alioune, 2006 - *L'entrée en vie adulte à Dakar*. Thèse de doctorat en démographie, Université de Paris I, Institut de démographie, 380 p.

Diagne Alioune, Lessault David, 2007 - *Emancipation résidentielle différée et recomposition des dépendances intergénérationnelles à Dakar*. Les Collections du CEPED, série « Regards sur », 42 p.

Dijkstra W., Taris T., 1995 - "Measuring the agreement between sequences". *Sociological methods et research*, (24) : 214-231.

Dureau Françoise, Barbary Olivier, Elisa Flores 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", *in Atelier du CEDE (Montevideo), Nuevas modalidades y tendencias de la migracion entre paises fronterizos y los processos de integracion*, 27-29 octobre 1993, Paris, ORSTOM, p. 31.

Elzinga Cees H., 2003 - "Sequence similarity: a nonaligning technique". *Sociological methods et research*, 32 : 3-29.

Elzinga Cees H., 2006 - "Sequence analysis: metric representations of categorical time series". *Sociological methods et research*, under revision

Elzinga Cees H., 2010 - "Complexity of Categorical Time Series". *Sociological Methods et Research*, 38(3) : 463-481.

Elzinga Cees H., Liefbroer Aart C., 2007 - "De-standardization of family-life trajectories of young adults: a cross-national comparison using sequence analysis". *European Journal of Population*, 23 (3-4) : 225-250.

Escofier Brigitte, Pagès Jérome, 2008 - *Analyses factorielles simples et multiples* Dunod (coll. Sciences sup.), 318 p.

Espinasse Jean-Michel, 1993 - « Enquêtes de cheminement, chronogrammes et classification automatique ». *Note du Lhire*, 19 (159).

Fénelon Jean-Pierre, Grelet Yvette, Houzel Yvette, 1997 - « Modéliser l'insertion ». *Formation Emploi*, (60) : 37-47.

Forrest John, Abbott Andrew, 1990 - "The optimal matching method for studying anthropological sequence data". *Journal of Quantitative Anthropology*, 2 : 151-170.

Fussell Elizabeth, 2005 - "Measuring the Transition to Adulthood in Mexico: An Application of the Entropy Index", *in* Macmillan Ross (ed.), *Advances in Life Course Research* : 91-122.

Gabadinho Alexis, Studer Matthias, Müller Nicolas, Ritschard Gilbert, 2009 - *Mining sequence data in R with the TraMineR package: A user's guide*. Geneva, Department of Econometrics and Laboratory of Demography, University of Geneva, 100.

Gabadinho Alexis, Studer Matthias, Müller Nicolas, Ritschard Gilbert, 2010a - « Indice de complexité pour le tri et la comparaison de séquences catégorielles », *in Extraction et gestion des connaissances (EGC 2010), Revue des nouvelles technologies de l'information RNTI*, E-19 : 61-66.

Gabadinho Alexis, Studer Matthias, Müller Nicolas, Ritschard Gilbert, 2010b - « Extraction de règles d'association séquentielle à l'aide de modèles de durée », *in Extraction et gestion des connaissances (EGC 2010), Revue des nouvelles technologies de l'information RNTI*, E-19 : 25-36.

Gabadinho Alexis, Studer Matthias, Müller Nicolas, Ritschard Gilbert, 2010c - « Classer, discriminer et visualiser des séquences d'événements », *in Extraction et gestion des connaissances (EGC 2010), Revue des nouvelles technologies de l'information RNTI*, E-19 : 37-48.

Gauthier Jacques-Antoine, Widmer Éric D., Bucher Philipp, Notredame Cédric, 2009 - "How Much Does It Cost?: Optimization of Costs in Sequence Analysis of Social Science Data". *Sociological Methods et Research*, 38 (1) : 197-231.

Giret Jean-François, Rousset Patrick, 2007 - Une analyse de la diversité des itinéraires professionnels en début de carrière. *XIV*^{èmes} *Journées d'étude sur les données longitudinales dans l'analyse du marché du travail*, Orléans, 30 et 31 mai, 13 p.

GRAB, 1999 - *Biographies d'enquêtes : bilan de 14 collectes biographiques*. Paris, INED (coll. Méthodes et savoirs), 3, 340 p.

GRAB, 2006 - *Etats flous et trajectoires complexes : observation, modélisation, interprétation*. Paris, INED (coll. Méthodes et savoirs), 5, 301 p.

Grelet Yvette, 1994 - Les trajectoires professionnelles dans les enquêtes du CEREQ : esquisses de traitement par l'analyse des données, *in L'analyse longitudinale du marché du travail*, CEREQ (coll. Documents séminaires, n° 99) : 219-236.

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

Halpin Brendan, 2003 - *Tracks through time and continuous processes: transitions, sequences, and social structure, paper prapared for the conference 'Frontiers in social and economic mobility'*. Cornell University, March 27-29, 18 p.

Halpin Brendan, 2010 - "Optimal Matching Analysis and Life-Course Data: The Importance of Duration". *Sociological Methods et Research*, 38 (3) : 365-388.

Halpin Brendan, Chan Tak Wing, 1998 - "Class careers as sequences: an optimal matching analysis of work-life histories". *European Sociological Review*, 14 (2) : 111-130.

Hamming R.W., 1950 - "Error-detecting and error-correcting codes". *Bell System Technical Journal*, 29 (2) : 147-160.

Han Shin-Kap, Moen Phyllis, 1999 - "Clocking out: temporal patterning of retirement". *American journal of sociology*, 105 (1) : 191-236.

Heckman James T., Ichimura Hidehiko, Todd Petra, 1998 - "Matching as an Econometric Evaluation Estimator". *The Review of Economic Studies*, 65 (2) : 261-294.

Hogan Dennis P., 1978 - "The variable order of events in the life course". *American sociological review*, 43 : 573-586.

Hollister Matissa, 2009 - "Is Optimal Matching Suboptimal?". *Sociological Methods et Research*, 38 (2) : 235-264.

Houzel Yvette, Le Vaillant Marc, 1994 – « Analyse statistique de données textuelles et traitement des données de calendriers : application à l'analyse de l'insertion professionnelle des élèves issus des écoles d'art », *in* Ourteau Maurice, Werquin Patrick *L'analyse longitudinale du marché du travail*. CEREQ (coll. Documents séminaires, n° 99) : 237-255.

Jalaudin Christophe, Moreau Gilles, 1995 - Le phrasé des parcours : contribution de l'analyse statistique de données textuelles à l'étude des itinéraires d'insertion professionnelle, *in* Degenne Alain, Mansuy Michèle, Werquin Patrick, *L'analyse longitudinale du marché du travail*. CEREQ (coll. Documents séminaires, n° 112) : 239-257.

Kalbfleisch John D., Prentice Robert L., 1980 - *The statistical analysis of failure time data.* New York, Wiley (coll. Wiley series in probability and mathematical statistics), 322 p.

Kogan Irena, 2004 - "Labour market careers of immigrants in Germany and the United Kingdom". *Journal of International Migration and Integration*, 5 (4) : 419-449.

Lebart Ludovic, Morineau Alain, Piron Marie, 2000 - *Statistique exploratoire multidimensionnelle*. Paris, Dunod (coll. Sciences sup.), 439 p.

Lelièvre Éva, 1999, « Collecter des données de mobilité : des histoires migratoires aux biographies d'entourage ». *Espace, populations, société*, (2) : 195-205.

Lelièvre Éva, Lecœur Sophie, 2010 - « Relations intergénérationnelles dans un contexte d'accès généralisé au traitement du sida en Thaïlande : parents âgés, enfants adultes ». *Autrement*, 53 (1) : 147-166.

Lelièvre Éva, Robette Nicolas, 2010 - « Les trajectoires spatiales d'activité des couples ». *Temporalités*, (11).

Lelièvre Éva, Vivier Géraldine, 2001 - « Évaluation d'une collecte à la croisée du quantitatif et du qualitatif : l'enquête Biographies et entourage ». *Population*, (6) : 1043-1073.

Lesnard Laurent, 2010 - "Setting Cost in Optimal Matching to Uncover Contemporaneous Socio-Temporal Patterns". *Sociological Methods et Research*, 38 (3) : 389-419.

Lesnard Laurent, de Saint Pol Thibaut, 2004 - « Introduction aux méthodes d'appariement optimal (Optimal Matching Analysis) ». *Document de travail INSEE*, (15), 30 p.

Lesnard Laurent, de Saint Pol Thibaut, 2009 - « Décrire des données séquentielles en sciences sociales : panorama des méthodes existantes ». Communication aux X^e *Journées de Méthodologie Statistique*, 23-25 mars, Paris, INSEE.

Levenshtein V.I., 1966(1965) - "Binary codes capable of correcting deletions, insertions, and reversals". *Soviet Physics Doklady*, 10 : 707-710.

Levine Joel H., 2000 - "But what have you done for us lately? Commentary on Abbott and Tsay". *Sociological methods et research*, 29 (1) : 34-40.

Levitt Barbara, Nass Clifford, 1989 - "The lid on the garbage can". *Administrative Science Quaterly*, 34 : 190-207.

Lillard Lee A., 1993 - "Simultaneous equations for hazards: marriage duration and fertility timing". *Journal of Econometrics*, 56 :189-217.

Macindoe Heather, Abbott Andrew, 2004 - "Sequence analysis and optimal matching techniques for social science data", *in* Hardy Melissa, Bryman Alan, *Handbook of Data Analysis*. London, Sage : 387-406.

Macmillan Ross, Eliason Scott R., 2003 - "Characterizing the life course as role configurations and pathways", *in* Mortimer Jeylan T., Shanahan Michael J. (eds), *Handbook of the Life Course* : 529-554.

Malo Miguel A., Munoz-Bullon Fernando, 2003 - "Employment status mobility from a lifecycle perspective: A sequence analysis of work-histories in the BHPS". *Demographic research*, 9 : 119-162.

Marini Margaret Mooney, 1984 - "The order of events in the transition to adulthood". *Sociology of education*, 57 : 63-84.

Martens Bernd, 1994 - "Analyzing event history data by cluster analysis and multiple correspondence analysis: an example using data about work and occupations of scientists and engieneers", *in* Greenacre Michael, *Blasius Jorg Correspondence analysis in the social sciences: recent developments and applications*. New-York : 233-251.

Mayer Karl Ulrich, Tuma Nancy Brandon, 1990 - *Event history analysis in life course research. Madison*. University of Wisconsin Press, 297 p.

McVicar Duncan, Anyadike-Danes Michael, 2002 - "Predicting successful and unsuccessful transitions from school to work by using sequence methods". *Journal of royal statistical society A*, (165) : 317-334.

Milligan G.W., 1981 - "A Monte-Carlo Study of Thirty Internal Criterion Measures for Cluster Analysis". *Psychometrika*, 46 :187-199.

Milligan G.W., Cooper M.C., 1985 - "An examination of procedures for determining the number of clusters". *Psychometrika*, 50 : 159-179.

Mouw Ted, 2005 - "Sequences of early adult transitions: A look at variability and consequences", *in* Settersten Jr Richard A., Furstenberg Jr Frank F., Rumbaut Ruben G., *On the Frontier of Adulthood: Theory, Research, and Public Policy*. Chicago, The University of Chicago Press : 256-291.

Needleman Saul B., Wunsch Christian D., 1970 - "A general method applicable to the search for similariries in the amino acid sequence of two proteins". *Journal of Molecular Biology*, (48) : 443-453.

Piccarreta Raffaella, Billari Francesco C., 2007 - "Clustering work and family trajectories by using a divisive algorithm". *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170 (4) : 1061-1078.

Piccarreta Raffaella, Lior Orna, 2010 - "Exploring sequences: a graphical tool based on multi-dimensional scaling". *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173 (1) : 165-184.

Pollock Gary, 2007 - "Holistic trajectories: a study of combined employment, housing and family careers by using multiple-sequence analysis". *Journal of royal statistical society*, (170) : 167-183.

Pollock Gary, Antcliff Valerie, Ralphs Rob, 2002 - "Work orders: analysing employment histories using sequence data". *International Journal of Social Research Methodology*, 5 (2) : 91-105.

Rindfuss Ronald R., Swicegood C. Gray, Rosenfeld Rachel A., 1987 - "Disorder in the life course: how common and does it matter?". *American sociological review*, 52 (6) : 785-801.

Ritschard Gilbert, Oris Michel, 2005 - "Life course data in demography and social sciences: statistical and data-mining approaches". *Advances in life course research*, 10 : 283-314.

Robette Nicolas, 2010 - "The diversity of pathways to adulthood in France: evidence from a holistic approach". *Advances in life course research*, 15 (2-3) : 89-96.

Robette Nicolas, Bonvalet Catherine, Bringé Arnaud, 2011 - « Les trajectoires géographiques des Franciliens depuis leur départ de chez leurs parents », *in* Bonvalet Catherine, Lelièvre Eva, *De l'entourage à l'espace résidentiel*. Paris, INED (coll. Grandes enquêtes), à paraître.

Robette Nicolas, Lelièvre Eva, Bry Xavier, 2011 - « Des trajectoires d'activité sur deux générations: continuités et singularités entre mères et filles », *in* Bonvalet Catherine, Lelièvre Eva, *De l'entourage à l'espace résidentiel*, Paris, INED (coll. Grandes enquêtes), à paraître.

Robette Nicolas, Solaz Anne, Pailhé Ariane, 2009 - "Work and family over the life-cycle: a typology of couples". Poster à la *XXVI*^{ème} *Conférence Internationale de la Population*, UIESP, Marrakech, 2 Octobre.

Robette Nicolas, Thibault Nicolas, 2006 - Les itinéraires familiaux des Franciliens nés entre 1930 et 1950 : analyser et classifier des trajectoires démographiques complexes. Communication au *Colloque AISLF « L'État social à l'épreuve des parcours de vie »*, Liège, 26 Septembre.

Robette Nicolas, Thibault Nicolas, 2008 - « L'analyse exploratoire de trajectoires professionnelles : analyse harmonique qualitative ou appariement optimal ? ». *Population-F*, 64 (3) : 621-646.

Rohwer Götz, Pötter Ulrich, 2005 - *TDA's user manual*, 1021 p.

Rouanet Henry, Le Roux Brigitte, 1993 - *Analyse des données multidimensionnelles : Statistique en sciences humaines*. Paris, Dunod, 310 p.

Roux M., 1993 - « Classification des données d'enquêtes », *in* Grange D., Lebart L., *Traitements statistiques des enquêtes*. Dunod.

Sackmann Reinhold, Wingens Matthias, 2003 - "From transitions to trajectories: Sequence types", *in* Heinz Walter R., Marshall Victor W., *The Life Course: Sequences, Institutions and Interrelations*. New-York, Aldine de Gruyter : 93-112.

Sankoff David, Kruskal Joseph, (dir), 1983 - *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison*. Reading, Addison-Wesley, 408 p.

Scherer Stefani, 2001 - "Early career patterns: a comparison of Great Britain and West Germany". *European Sociological Review*, 17 (2) : 119-144.

Settersten Jr. Richard A., Mayer Karl Ulrich, 1997 - "The measurement of age, age structuring and the life course". *Annual review of sociology*, 23 : 233-261.

Solis Patricio, Billari Francesco C., 2002 - "Work lives amid social change and continuity: occupational trajectories in Monterrey, Mexico", MPIDR Working Paper, (9), 52 p.

Stark David, Vedres Balázs, 2006 - "Social Times of Network Spaces: Network Sequences and Foreign Investment in Hungary". *American journal of sociology*, 111 (5) : 1367-1411.

Steele Fiona A., 2008 - "Multilevel Models for Longitudinal Data". *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171 (1) : 5-19.

Stovel Katherine, 2001 - "Local sequential patterns: The structure of lynching in the Deep South, 1882-1930". *Social Forces*, 79 (3) : 843-880.

Stovel Katherine, Bolan Marc, 2004 - "Residential trajectories. Using optimal alignment to reveal the structure of residential mobility". *Sociological methods et research*, 32 (4) : 559-598.

Stovel Katherine, Savage Michael, Bearman Peter, 1996 - "Ascription into achievement: models of career systems at Lloyds Bank, 1890-1970". *American Journal of Sociology*, 102 (2) : 358-399.

Tukey John W., 1962 - "The Future of Data Analysis". *The Annals of Mathematical Statistics*, 33 (1) : 1-67.

Van der Heijden Peter G.M., 1987 - *Correspondence analysis of longitudinal categorical data.* Leiden, DSWO Press.

Van der Heijden Peter G.M., Teunissen Joop, van Orlé Charles, 1997 - "Multiple correspondence analysis as as tool for quantification or classification of career data". *Journal of Educational and Behavioral Statistics*, 22 (4) : 447-477.

Villeneuve-Gokalp Catherine, 1997 - « Le départ de chez les parents : définitions d'un processus complexe ». *Economie et statistique*, (304-305) : 149-162.

Wiggins Richard D., Erzberger Christian, Hyde Martin, Higgs Paul, Blane David, 2007 - "Optimal Matching Analysis using ideal types to describe the lifecourse: an illustration of how histories of work, partnerships and housing relate to quality of life in early old age". *International Journal of Social Research Methodology*, 10 (4) : 259-278.

Willekens F.J., 1999 - "The life course: Models and analysis", *in* van Wissen L.J.G., Dykstra P. (eds), *Population issues: An interdisciplinary focus*. New-York, Plenum Press, p. 23-51.

Williams W.T., Lance G.N., 1965 - "Logic of computer-based intrinsic classifications". *Nature*, 207 (4993) : 159-161.

Wilson W.C., 1998 - "Activity pattern analysis by means of sequence-alignment methods". *Environment and Planning A* 30 (6) : 1017-1038.

Wu Lawrence L., 2000 - "Some comments on 'Sequence analysis and optimal matching methods in sociology: Review and prospect'". *Sociological methods et research*, 29 (1) : 41-64.

Wu Lawrence L., 2003 - "Event history models for life course analysis", *in* Mortimer Jeylan T., Shanahan Michael J., *Handbook of the life course*. New-York, Kluwer academic/Plenum : 477-502.

Annexe 1

Les logiciels

Le codage des parcours sous forme de disjonctifs complets ou de calendriers simplifiés, de même que les analyses factorielles ou les procédures de classification, sont réalisables à partir de la plupart des logiciels de statistiques généralistes (SAS, Stata, Spss, R…). Ce n'est en revanche pas le cas de l'analyse séquentielle en général et de l'*Optimal Matching* en particulier.

SAS et Spss ne permettent pas à ce jour d'effectuer des analyses de séquences de manière satisfaisante.

Stata possède un module nommé « sq » dédié à l'analyse séquentielle et plus particulièrement à l'*Optimal Matching* (Brinsky-Fay *et al.*, 2006), qui permet aussi la représentation de « tapis » (*index plots*). Il est relativement lent, cependant Brendan Halpin propose une version plus rapide de l'algorithme d'OMA.
http://teaching.sociology.ul.ie/seqanal/

TDA, un logiciel libre de statistiques plus particulièrement orienté vers l'analyse biographique, possède des fonctions rapides et relativement simples pour l'analyse de séquences et l'*Optimal Matching* (Rohwer et Pötter, 2005).
http://www.stat.ruhr-uni-bochum.de/tda.html

Le *Dynamic Hamming* a été implémenté par son auteur sous Stata et sous SAS (extension « Seqcomp »).
http://laurent.lesnard.free.fr/rubrique.php3?id_rubrique=4

Elzinga a développé son propre logiciel, Chesa, pour mettre à disposition ses métriques et le calcul de la turbulence des séquences.
http://home.fsw.vu.nl/ch.elzinga/

D'autres logiciels spécifiquement consacrés à l'analyse de séquences et à l'*Optimal matching* existent aussi : Optimize, développé par Andrew Abbott (mais plus mis à jour depuis 1997), ou des logiciels issus de la biologie (Pile-up, ClustalG…).
http://home.uchicago.edu/~aabbott/om.html
http://www.hku.hk/bruhk/gcgdoc/pileup.html

Mais le moyen le plus performant et le plus complet de faire de l'analyse de séquences est sans doute à l'heure actuelle le package « TraMineR » du logiciel libre R (Gabadinho *et al.*, 2009). Celui-ci offre en effet de nombreuses possibilités : *Optimal Matching*, mais aussi *Dynamic Hamming*, certaines métriques d'Elzinga (*longest common subsequence*…) ; diverses représentations graphiques dont les « tapis » (*index plots*) ; calculs d'entropie, de turbulence et de complexité, etc. Et il est gratuit !
http://cran.r-project.org/
http://mephisto.unige.ch/traminer/

Annexe 2

Programme R de l'analyse des trajectoires professionnelles

Le fichier se nomme ici « Trajectoires.txt » et est placé dans le répertoire « c:\ ». Il est composé d'une ligne par trajectoire individuelle ; les 37 états successifs prennent une valeur comprise entre 1 et 9 et sont séparés par des virgules. A titre d'exemple, les 3 premières lignes sont reproduites ici :

1,1,1,1,1,1,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,4,4,4,4,4,4,4,4

7,7,7,7,7,7,9,9,6,6,6,6,6,4,4,4,4,4,4,4,4,4,4,6,6,6,6,6,6,6,6,6,6,6,6,6,6

6,6,6,6,6,6,6,9,9,9,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6

Le programme est le suivant (pour la typologie en cinq classes) :

```
## ================
## Etapes préliminaires
## ================

# Importation des trajectoires
traj <- read.csv("C:/Trajectoires.txt",header=FALSE,sep=",")

# Chargement en mémoire des bibliothèques de fonctions nécessaires au programme
library(TraMineR)
library(cluster)

# Mise en mémoire des paramètres graphiques par défaut
par.def <- par()

# Définition de la palette de couleurs pour les graphiques
palette <- c("#D19392","#DB843D","#AA4643","#71588F","#B9CD96","#89A54E","#93A9CF","#4198AF","#4572A7")
```

```
# Définition des labels des états
labels <- c("agriculteurs","acce","cadres","prof. int.","employes","ouvriers","etudiants","inactifs","serv. mil.")

# Mise en forme des trajectoires sous forme d'un "objet" séquence
seq <- seqdef(traj, lab=labels, cpal=palette)



## ========================
## Optimal matching et typologie
## ========================

# Définition des coûts de substitution pour l'optimal matching
couts <- seqsubm(seq,method="TRATE", cval=2)

# Calcul des dissimilarités avec l'optimal matching
seq.om <- seqdist(seq, method="OM", indel=1.1, sm=couts)

# Mise en forme des dissimilarités sous forme de matrice de distance
seq.dist <- as.dist(seq.om)

# Classification ascendante hiérarchique
seq.agnes <- agnes(seq.dist, method="ward", keep.diss=FALSE)

# Dendrogramme de la classification
par(par.def)
plot(as.dendrogram(seq.agnes), leaflab="none")

# Inertie des partitions selon le nombre de classes
par(par.def)
plot(sort(seq.agnes$height, decreasing=TRUE)[1:20], type='s', xlab="nb de classes", ylab="inertie")
```

```
## ================================
## Représentations graphiques de la typologie
## ================================

# Choix d'une partition en 5 classes
nbcl <- 5
seq.part <- cutree(seq.agnes, nbcl)

# Tapis de la typologie, triés par multidimensional scaling mds <- cmdscale(seq.om, k=1, eig=F) par(mfrow=c(2,3),mar=c(5,4,4,2)) seqiplot(seq, sortv=mds, group=seq.part, xtlab=14:50, tlim=0, space=0, border=NA, withlegend=T, yaxis=FALSE, title="classe")

# Chronogrammes de la typologie
par(mfrow=c(2,3),mar=c(5,4,4,2))
seqdplot(seq, group=seq.part, xtlab=14:50, border=NA, withlegend=T, title="classe")

## ====================
## Description de la typologie
## ====================

# Distribution de la partition (effectif et pourcentage)
distri.eff <- table(seq.part)
distri.pct <- round(distri.eff/sum(distri.eff)*100,1)
distri.eff
distri.pct

# Durées dans les états
dur <- seqistatd(seq)
durees <- aggregate(dur,by=list(seq.part),FUN=mean)
durees
```

```
# Au moins un épisode dans les états
epi <- ceiling(dur/ncol(seq))
episodes <- aggregate(epi,by=list(seq.part),FUN=mean)*100
episodes

# Nombre de transitions
trans <- seqtransn(seq)
transitions <- aggregate(trans,by=list(seq.part),FUN=mean)
transitions

# Entropie transversale moyenne par classe
entropie <- vector()
for(i in 1:nbcl) entropie[i] <- round(mean(seqstatd(seq[seq.part==i,])$Entropy),2)
entropie

# Distance intra-classes
Dintra <- integer(length=nbcl)
for(i in 1:nbcl) Dintra[i] <- round(mean(seq.om[seq.part==i,seq.part==i]),1)
Dintra
```

.

Liste des figures

Liste des tableaux

Légende de la photo de couverture

© IRD – Laure Emperaire
Gros plan de tamis à mailles fines servant à la préparation
des galettes de manioc (beijus). Motifs de vannerie. Amazonie, Brésil.

Imprimé en France
par PRÉSENCE GRAPHIQUE
2, rue de la Pinsonnière - 37260 MONTS
N° d'imprimeur :

Dépôt légal 3e trimestre 2011