DECRIRE DES DONNEES SEQUENTIELLES EN SCIENCES SOCIALES : PANORAMA DES METHODES EXISTANTES

Laurent LESNARD (*), Thibaut DE SAINT POL (**)

(*) Sciences Po, Centre de données socio-politiques
Crest, Laboratoire de sociologie quantitative
(**) Insee, Conditions de vie des ménages
Crest, Laboratoire de sociologie quantitative

Introduction

Que l'objectif soit de décrire les trajectoires d'insertion sur le marché du travail, les carrières professionnelles ou les emplois du temps, disposer d'outils adaptés pour décrire les données séquentielles ou longitudinales est essentiel pour le statisticien et le chercheur en sciences sociales. Ce texte a pour objectif de présenter les principales techniques qui permettent de dresser des typologies empiriques de séquences.

À côté des analyses factorielles et harmoniques, une attention particulière sera accordée aux Méthodes d'Appariement Optimal, technique nouvelle qui s'impose comme la méthode de référence dans les pays anglo-saxons. Issues des travaux en théorie du signal dans les années 1950 et 1960, les Méthodes d'Appariement Optimal (en anglais *Optimal matching analysis*) permettent de construire une distance entre les séquences fondée sur leur comparaison au moyen de trois opérations (insertion, suppression ou substitution d'un élément par un autre). Cette distance est établie comme le coût minimal pour transformer une séquence en une autre au moyen de ces trois opérations. La question du coût affecté aux opérations sera particulièrement discutée. Le coût de ces trois opérations est en effet un paramètre qui donne une grande souplesse à ces analyses.

Les caractéristiques des Méthodes d'Appariement Optimal seront ainsi mises en regard des potentialités des autres techniques habituellement utilisées pour décrire des données séquentielles. Les hypothèses sous-jacentes sur lesquelles reposent ces méthodes seront comparées, en s'intéressant en particulier aux types de données et de régularités pour lesquels ces différentes approches sont le plus adaptées.

1. Les méthodes d'appariement optimal

Bien qu'issues des recherches menées dans les années 1950 et 1960 en informatique où elles sont connues sous le nom de distance de Levenshtein [1], Hamming [2], ou encore *edit distance* [3], les Méthodes d'Appariement Optimal (M.A.O), traduction que nous avons proposée pour *Optimal Matching Analysis* [4], sont plus connues en biologie où elles ont contribué au séquençage du génome[1]. De manière plus générale, les M.A.O. permettent de comparer le degré de similarité de séquences, autrement dit d'évaluer leur proximité : les Méthodes d'Appariement Optimal peuvent donc être vues comme une extension séquentielle des outils de la statistique non inférentielle. C'est Andrew Abbott, de l'Université de Chicago, [5,6] qui se trouve principalement à l'origine de l'introduction des M.A.O. en sciences sociales au travers de l'étude de processus historiques. Principes que Andrew Abbott a ensuite approfondis dans deux articles [7,8].

Les Méthodes d'Appariement Optimal ont pour finalité de bâtir une typologie de séquences, c'est-à-dire rapprocher des suites d'éléments. Alors qu'il est impossible à l'œil humain de comparer des milliers d'éléments et la manière dont ils s'enchaînent, les M.A.O. permettent de les regrouper et de dégager des idéaux-types. La première étape de cette procédure consiste à calculer une distance

[1] Le début de ce texte a été repris et adapté de l'article d'introduction aux Méthodes d'Appariement Optimal [4].

entre les séquences. La seconde étape est la classification proprement dite des séquences mais d'autres méthodes peuvent également être utilisées, comme le *Multidimensional Scaling* [9].

1.1. Comparer des séquences avec les Méthodes d'Appariement Optimal

Dans cette première étape, il s'agit d'arriver à comparer des séquences qui peuvent être de longueurs différentes et contenir des éléments divers. La construction de la distance entre ces séquences est réalisée au moyen de trois opérations (l'insertion d'un élément dans la séquence, la suppression d'un élément dans la séquence ou la substitution d'un élément par un autre) qui correspondent aux trois modifications élémentaires que nous appliquons instinctivement aux séquences quand nous tentons de les comparer à l'œil nu. Les M.A.O. reposent sur la considération de tous les chemins possibles pour passer d'une séquence à l'autre au moyen de ces trois opérations. Il s'agit de trouver pour chaque couple de séquences comment on peut transformer l'une en l'autre le plus facilement possible, c'est-à-dire, en termes mathématiques, pour le coût minimum.

Soient par exemple deux séquences qui représentent les engagements successifs de deux militants X et Y dans les associations A, B, C et D par plages de 5 ans.

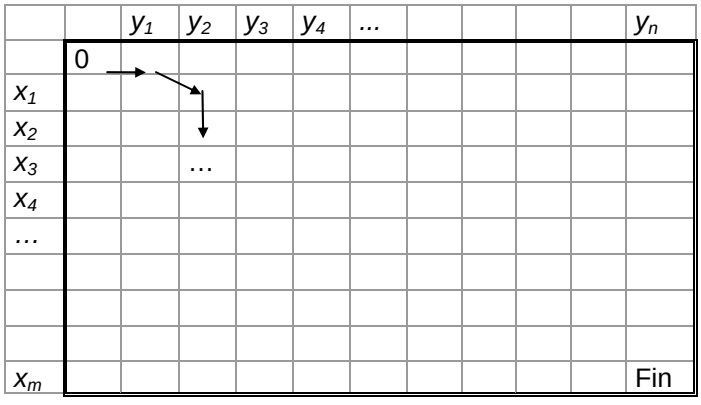
Figure 1 – Deux séquences à comparer

X : C – A – B – D – D
Y : A – B – C – D

Pour passer de la séquence X à la séquence Y, il suffit de supprimer le C en 1re position dans la séquence X et de transformer le D alors en 3e position dans X en un C. Le coût de passage de la séquence X à la séquence Y selon ce chemin est le coût d'une suppression de C et d'une transformation d'un D en C.

Mais ce n'est pas la seule manière de passer de la première séquence à la seconde. On peut aussi supprimer le C en 1re position puis le D en dernière position et insérer un C entre le B et le D. Le coût du passage de X à Y sera alors la somme des coûts des deux suppressions et de l'insertion. Il s'agit donc de considérer tous les moyens de passer de X à Y. La distance entre les deux séquences sera le coût du chemin le moins cher.

Figure 2 – Représentation matricielle de la comparaison de deux séquences par les M.A.O.

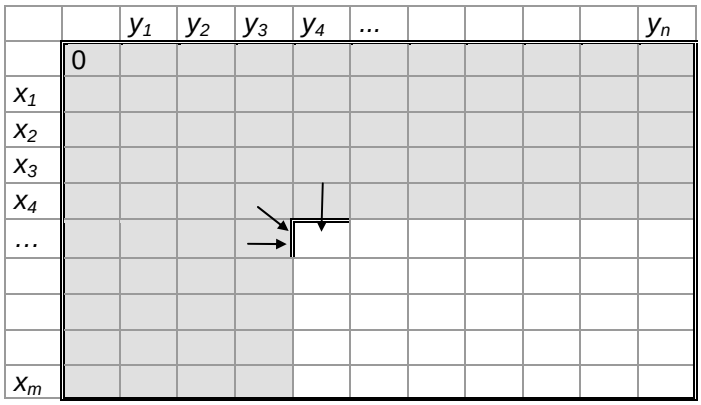


Si on généralise ce processus à deux séquences de taille *m* et *n*, on peut représenter cette procédure sous la forme d'une matrice de taille *m*,*n*. Ainsi, si on compare les séquences X = (x_1,…,x_m) et Y = (y_1,…y_n), on obtient la matrice représentée ci dessous. Passer de X à Y, c'est passer de la cellule en haut à gauche à celle en bas à droite. Descendre verticalement d'une ligne, c'est supprimer l'élément de X correspondant. Passer à la colonne de droite, c'est insérer un élément de Y dans X. Descendre en diagonale, c'est transformer l'élément de X en l'élément de Y correspondant. A titre d'exemple, on a représenté ici l'insertion de y_1, la transformation de x_1 en y_2 et la suppression de x_2[2].

[2] Ce graphique et le suivant sont inspirés de Chan [13].

Dès lors qu'on connaît le coût initial et le coût affecté à chaque opération, il est possible d'obtenir le coût en chaque case. Comme le montre la figure 3, il n'y a que trois façons de parvenir sur une case. On peut ainsi déterminer l'appariement optimal, c'est à dire celui qui fournit le coût minimum. La distance entre nos deux séquences sera donc le coût du chemin le moins onéreux pour transformer l'une en l'autre.

Figure 3 – Représentation matricielle du processus de minimisation de la distance entre deux séquences par les M.A.O.



Cette procédure de minimisation permet ainsi de calculer la distance de chaque séquence à toutes les autres séquences de l'échantillon. Il s'agit ensuite de mettre en œuvre des techniques de classification pour rassembler les séquences qui sont les plus proches au regard de la distance qui vient d'être construite. On passe à la seconde étape de la Méthode d'Appariement Optimal.

1.2. Regrouper les séquences voisines

Il existe de nombreuses techniques de classifications qui reposent sur des algorithmes plus ou moins complexes. Elles ont pour but de construire des classes qui doivent être les plus homogènes possibles. Si on distinguait autrefois deux grands types de méthodes, les méthodes hiérarchiques et les méthodes de partitionnement, d'autres approches ont vu le jour récemment, comme les réseaux de neurones par exemple.

Mais il faut être conscient de ce que signifie la réalisation d'une classification pour nos séquences. Si nous possédons à ce stade une distance deux à deux entre séquences, il nous faut désormais définir une distance entre groupes de séquences. En effet, l'enjeu des procédures de classification est de passer d'une distance entre des individus à une distance entre des groupes. Ainsi, pour pouvoir faire des classes, les algorithmes de classification utilisent la distance entre une séquence et un groupe, ou entre deux groupes. C'est ce qu'on appelle le critère d'agrégation. On retient à chaque étape la réunion entre les deux éléments qui ont la distance la moins importante. Puis on recalcule à nouveau les distances et on retient encore la plus faible. Appliquer une classification à notre matrice de distance ne pose pas de grands problèmes techniques. Le logiciel SAS propose par exemple une dizaine de méthodes de classification.

Toutes ces méthodes reposent sur des algorithmes différents (certaines considèrent la moyenne, d'autres la variance, d'autres encore utilisent directement la distance de chacune des séquences qui composent le groupe). Le choix de la « bonne » méthode est parfois difficile et dépend de la nature des variables, de la problématique posée et souvent des habitudes du domaine d'étude. Les classifications, notamment ascendantes hiérarchiques (CAH), occupent une place de choix dans la boîte à outil classique du chercheur en sciences sociales et du statisticien. Utilisées dans de nombreux travaux, elles permettent de regrouper des individus selon un critère prédéfini et de former des classes. La première partie des M.A.O. a donné ce critère. Il suffit de retenir une méthode et de regrouper les séquences.

Proches de la distance de Hamming, qui se trouve être elle-même assimilable à la distance de Manhattan ou L_1 dans certain cas, les M.A.O. s'accommodent mal *a priori* de la mesure d'agrégation de CAH euclidienne (la méthode de Ward). Par ailleurs, des analyses ont montré que les méthodes

WPGMA flexible (*Flexible Weighted Pair Group using arithMetic Averages*), ou mieux UPGMA flexible (*Flexible Unweighted Pair Group using arithMetic Averages*), sont les plus performantes sur les données empiriques, en particulier en présence de bruit ou d'observations aberrantes [10,11, 12]. La méthode WPGMA flexible est disponible dans R, SAS (sous le nom de *beta-flexible*) et ClustanGraphics mais reste indisponible dans la version 17 de SPSS et 11 de Stata.

1.3. La question des coûts

Nous avons présenté le principe des Méthodes d'Appariement Optimal en laissant jusqu'ici sous silence la détermination des coûts de chacune des trois opérations fondamentales. En effet, le problème de la fixation des coûts est l'aspect central des M.A.O., et aussi ce qui lui confère une grande souplesse. Le coût relatif à chaque opération détermine directement le calcul des distances. Le choix des coûts est donc le point le plus délicat, mais c'est aussi le plus essentiel des techniques d'Appariement Optimal. Cet aspect est souvent laissé de côté dans les applications des M.A.O. publiées par le passé, le choix des coûts étant présenté comme un choix uniquement technique donc secondaire. Nous considérons au contraire que la détermination des coûts est fondamentale d'un point de vue théorique puisque, comme nous allons le montrer maintenant, c'est en jouant sur les coûts qu'il est possible d'adapter la méthode à l'objet traité et au type de régularité recherché.

D'un point de vue théorique, les méthodes de séquençage ne reposent en fait que sur deux types d'opérations : les opérations d'insertion-suppression d'un côté (*insertion* et *deletion* en anglais, ce qui donne, par combinaison des premières lettres de ces deux mots, l'acronyme *indel*), et les opérations de substitution de l'autre. Les premières opérations décalent les séquences de manière à faire émerger des enchaînements communs, donc privilégient l'identification de suites d'états codées de la même manière au détriment de leurs localisations respectives dans les deux séquences considérées. Autrement dit, les opérations d'insertion-suppression déforment les structures temporelles des séquences comparées (insérer un *événement*, c'est insérer du *temps*) et permettent ainsi d'accélérer ou de ralentir le temps de chaque séquence pour mieux mettre en regard leurs points communs. Au contraire, les opérations de substitution conservent les structures temporelles des séquences puisqu'elles privilégient la comparaison d'événements situés aux mêmes points des séquences comparées, ce qui revient à faire pencher la balance de la comparaison en faveur des différences entre des événements qui sont identiques du point de vue de l'échelle du temps utilisée, qui sont donc *comparables* du point de vue du temps.

Tableau 1 – Signification des deux opérations de base des Méthodes d'Appariement Optimal

	Insertion-Suppression	**Substitution**
Ce qui est préservé	Événements	Temps
Ce qui est simplifié	Temps	Événements

Le modèle de comparaison de séquences proposé par les M.A.O. consiste donc à distordre une des deux dimensions fondamentales des séquences, le temps ou les événements, pour mieux comparer les séquences du point de vue de la dimension qui est préservée (voir Tableau 1) : les opérations d'insertion-suppression déforment le temps pour mieux comparer les événements identiquement codés des séquences tandis que les opérations de substitution distordent les événements pour mieux comparer leur dimension temporelle. Les M.A.O. alternent donc ces deux types de simplifications que permet de visualiser la représentation matricielle du processus (voir Figure 2 ci-dessus) : la seule possibilité de conserver les temporalités des séquences est de passer par la diagonale, tout détour vertical ou horizontal correspondant à une suppression du temps d'une séquence qui est en même temps une insertion de temps dans l'autre[3]. Au final, les M.A.O. sont donc une combinaison d'accélérations, de ralentissements et d'écoulements normaux[4] du temps qui permettent de comparer

[3] C'est la raison pour laquelle le même coût est attribué à ces opérations symétriques, symétrie qui apparaît clairement dans la représentation matricielle des M.A.O..

[4] Par « écoulement normal du temps » il faut entendre « conformément au rythme de l'échelle de temps des séquences ».

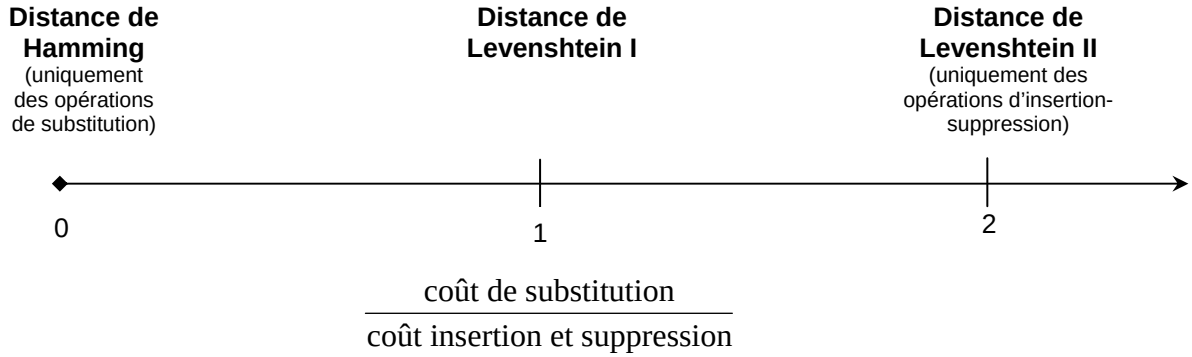
des séquences d'événements. Cette combinaison est par définition optimale et déterminée par l'algorithme mais peut cependant être orientée par le choix des coûts.

Tableau 3 – Distances de Hamming et de Levenshtein

	Operations utilisées	
	Substitution	Insertion et suppression
Hamming	Oui (coût = 1)	Non
Levenshtein I	Oui (coût = 1)	Oui (coût = 1)
Levenshtein II	Non	Oui (coût = 1)

Du choix des coûts associés aux trois opérations des M.A.O. dépendent en effet l'équilibre entre les insertions-suppressions et les substitutions mais également le degré de simplification que ces opérations induisent. C'est pourquoi nous avons choisi de parler « des » Méthodes d'Appariement Optimal, alors que l'anglais privilégie le singulier. Ce n'est que conditionnellement aux choix des coûts que l'appariement est optimal : l'usage du pluriel indique bien qu'il n'existe pas une unique façon de comparer des séquences. Affecter des coûts aux opérations d'insertion-suppression et de substitution, c'est arbitrer entre la distance temporelle qui sépare des mêmes événements et la distance entre événements qui se déroulent sur les mêmes unités de temps : choisir des coûts d'insertion-suppression inférieurs aux coûts de substitution, c'est faire ainsi le choix de ne pas utiliser les opérations de substitution, d'asseoir la comparaison uniquement sur le rapprochement temporel d'événements identiques, plus exactement sur le nombre d'unités temporelles séparant des événements identiques. N'utiliser que des opérations d'insertion-suppression, c'est en effet réduire deux séquences à leurs éléments communs, leur distance s'élevant au nombre d'éléments écartés pondérés par le coût de leur suppression. Le choix des coûts permet de donner plus ou moins d'importance aux décalage dans le temps. Dans le cas extrême de la distance de Hamming (voir Tableau 3), aucune opération d'insertion-suppression n'est utilisée (l'utilisation d'un coût *indel* infiniment grand reviendrait au même). C'est justement pour introduire un peu plus de souplesse dans la comparaison des séquences que Vladimir Levenshtein a suggéré l'utilisation d'opérations d'insertion – suppression (distance de Levenshtein I), puis proposé que dans certains cas il soit intéressant de ne pas utiliser de substitutions (distance de Levenshtein II), ce qui revient à identifier la plus longue suite d'états commune aux deux séquences comparées. Au final, le choix des coûts revient positionner le curseur entre les deux cas limites des distance de Hamming et de Levenshtein II (voir Figure 4). Plus le coût de substitution est faible comparé au coût d'insertion – suppression, plus la contemporanéité des événements est privilégiée. Dans le cas inverse, l'intérêt se portera plus sur la recherche des plus longues sous-séquences communes[5].

Figure 4 – Effet des coûts sur le type de régularité statistique privilégié



Prenons un exemple de deux séquences largement semblables mais dont le calendrier est décalé (voir Figure 5). Avec le système de coût qui était traditionnellement utilisé dans lequel une insertion-suppression coûte une unité contre deux pour toute substitution, l'appariement optimal est obtenu

[5] Lorsque le coût d'insertion suppression est d'une unité contre deux pour la substitution (Levenshtein II), alors les opérations de substitution ne sont plus utilisées puisqu'elle peut être remplacée par une insertion et une suppression pour le même coût.

pour un coût de quatre unités (deux insertions de C et deux suppressions de B) contre huit pour un appariement composé uniquement d'opérations de substitution[6].

Figure 5 – Deux séquences décalées

X : A – A – A – A – B – B – B – B
Y : C – C – A – A – A – A – B – B

Plus précisément, les éléments qui apparaissent communs dépendent de l'ordre des événements dans chacune des séquences[7], autrement dit, le temps n'est pas aboli mais réduit à sa dimension de succession : ce qui est recherché avec l'utilisation intensive d'opérations d'insertion-suppression, ce sont des suites d'événements identiques quelles que puissent être les différences de leurs positions respectives dans chaque séquence. La simplification du temps sous-jacente aux opérations d'insertion-suppression apparaît donc clairement : le temps est considéré comme uniforme, comme simple support de classement des événements qui peut donc être manipulé afin de faciliter le rapprochement de suites d'événements identiques.

Au contraire, préserver toute l'échelle de temps de l'action requiert des coûts d'insertion-suppression très élevés[8] mais pose la question de la distance entre événements, question que la stratégie classique supprime au prix d'une simplification temporelle qui passe bien souvent inaperçue faute d'en voir toutes les conséquences. Conserver la structure temporelle passe, nous l'avons vu, par la simplification de la comparaison entre les événements, autrement dit par la transformation de toutes les différences entre deux événements par un seul chiffre, le coût de substitution. Bien que la solution classique suggère d'affecter un coût de deux unités à toute opération de substitution, il est également possible de faire varier le coût de substitution selon les couples d'états, et de les déterminer théoriquement ou selon des critères empiriques. Par exemple, il est possible d'utiliser l'information diachronique sur les transitions entre états pour l'ensemble des séquences de manière à comparer synchroniquement les séquences deux à deux. Autrement dit, la matrice des transitions entre tous les états constituée à partir de l'ensemble des séquences à comparer est utilisée comme matrice des coûts pour substituer un état à un autre, c'est-à-dire pour comparer la proximité des états de deux séquences[9].

La comparaison de séquences par la technique de l'Appariement Optimal nécessite donc d'arbitrer l'une ou l'autre de leurs dimensions, temps ou événements. C'est le choix de la complexité et des valeurs des différents coûts qui permet de jouer sur ces deux simplifications afin de décrire au mieux les ressemblances entre les séquences, c'est-à-dire selon la nature des séquences étudiées et l'objectif de l'analyse. Joel Levine [14] voit dans le choix des coûts le signe d'une faiblesse intrinsèque de la méthode : la statistique, affirme-t-il, n'est qu'un moyen de « discriminer un signal même en présence d'un degré considérable de bruit » et n'a pas à s'adapter aux sciences sociales. Cette critique minimise la portée sociologique des choix qui se trouvent derrière toute procédure statistique, inférentielle ou descriptive : discriminer un signal du bruit, c'est, au travers des hypothèses sur lesquelles s'appuie toute méthode statistique, choisir d'ignorer une partie de l'information (qui devient du bruit selon les hypothèses choisies) pour mieux amplifier et analyser ce qui reste. Outre que les Méthodes d'Appariement Optimal ne sont pas des modèles statistiques et reposent donc sur des hypothèses moins fortes, ce n'est pas la présence de choix qui les en distingue, mais leur plus grande visibilité. Mieux, il est légitime de considérer que les M.A.O. sont plus transparentes : loin d'être un désagrément, l'obligation de rendre compte des choix de l'analyse est au contraire un avantage puisqu'il ne devient plus possible d'appliquer machinalement une méthode sans s'interroger sur les choix sociologiques qui se trouvent engagés. La grande nouveauté ici est que, contrairement à nombre de méthodes statistiques classiques, les Méthodes d'Appariement Optimal rendent visible les

[6] Lorsque les séquences comparées sont de même longueur, l'utilisation des seules opérations de substitution revient à appliquer la distance de Hamming.
[7] Ce n'est donc pas un simple dénombrement des éléments communs de chaque séquence.
[8] Voire de réduire l'analyse aux seules opérations de substitution, solution qui est présentée dans le second exemple de Thibaut de SAINT POL, Laurent LESNARD, Décrire des données séquentielles en sciences sociales : mise en pratique des Méthodes d'Appariement Optimal, Session « Analyses longitudinales ».
[9] Dans ce cas les coûts de substitution sont inversement proportionnels aux fréquences de transition ce qui permet d'assigner des coûts faibles aux états associées à de fortes transitions et inversement.

enjeux sociologiques de la statistique : elles permettent de véritablement réfléchir et de choisir ce qui convient théoriquement le mieux[10].

2. Analyse factorielle et données séquentielles

Le fort développement et ancrage de l'analyse factorielle en France suite aux travaux de Benzécri [15] explique très certainement l'utilisation dès les années 1970 de ces méthodes pour décrire des données séquentielles. S'il existe en théorie de nombreuses façons d'utiliser les techniques de l'analyse factorielle pour décrire les séquences, en pratique deux stratégies peuvent être identifiées : l'analyse harmonique qualitative [16, 17, 18, 19, 20] et l'analyse des correspondances multiples [21].

2.1. L'analyse harmonique qualitative

Issue de recherches menées en mathématique, l'analyse harmonique qualitative (A.H.Q.) a été introduite dans les sciences sociales par Jean-Claude Deville dans les années 1970. En pratique, l'A.H.Q revient à réduire les trajectoires à des durées moyennes passées dans les différents états dans lesquelles elles se déploient et à les soumettre à une Analyse Factorielle des Correspondances (AFC)[11]. La première étape consiste donc à diviser la période d'observation en groupes d'épisodes, ou sous-périodes, qui peuvent être de taille inégale. Cette opération est plus délicate qu'il n'y paraît puisque le choix d'un trop grand nombre de sous-périodes est susceptible de produire de nombreuses durées nulles et de fragiliser ainsi l'analyse. Inversement, un petit nombre de sous-périodes risque de faire disparaître artificiellement l'essentiel des variations temporelles qui sont précisément l'objet de l'étude. Le choix du nombre et des bornes des intervalles peut être guidé par la distribution des changements d'états (déciles par exemples). Ainsi, dans leur récente analyse des parcours professionnels, Nicolas Robette et Nicolas Thibault ont choisi de découper la période d'observation en dix classes et de s'inspirer des déciles de la distribution des changements de PCS en fonction de l'âge pour déterminer leurs bornes [20].

La seconde étape consiste à calculer la proportion du temps passé dans chaque état pour toutes les sous-périodes définies à la première étape. À l'issue de ces deux premières étapes, les trajectoires se trouvent donc simplifiées par une série de durées relatives passées dans les différents états. Plus exactement, chaque trajectoire individuelle est décrite par P x K variables, avec P le nombre de périodes considérées dans l'analyse et K le nombre d'états (voir Tableau 4).

Tableau 4 – Matrice harmonique

	Période 1				…	Période *p*				…
	État 1	État 2	…	État *K*		État 1	État 2	…	État *K*	
Individu 1	x_{111}	x_{112}	…	x_{11K}		x_{1p1}	x_{1p2}	…	x_{1pK}	
…										
Individu *i*	x_{i11}	x_{i12}	…	x_{i1K}		x_{ip1}	x_{ip2}	…	x_{ipK}	
…										

Cette matrice harmonique est ensuite soumise à une AFC pour réduire la dimensionnalité des trajectoires. Les trajectoires individuelles harmoniques simplifiées sont donc remplacées par leurs coordonnées sur les facteurs issus de l'AFC. Le propre des analyses factorielles étant de se placer dans un espace euclidien, il est dès lors possible d'utiliser la distance euclidienne pour évaluer le degré de proximité des trajectoires individuelles et de construire une typologie à l'aide d'une classification ascendante hiérarchique et du critère d'agrégation de Ward. Il est d'usage de ne pas retenir la totalité des facteurs de manière à éliminer le bruit, contenu dans les tout derniers facteurs. Il est possible d'affiner l'analyse en introduisant en plus des P x K variables décrivant la proportion de

[10] Thibaut de SAINT POL, Laurent LESNARD, Décrire des données séquentielles en sciences sociales : mise en pratique des Méthodes d'Appariement Optimal, Session « Analyses longitudinales ».
[11] Pour une présentation détaillée de l'analyse harmonique et de son application à un processus qualitatif, voir [16, 17, 19].

chaque sous-période passée dans les K états, une série de variables qui prennent en compte le nombre de transitions [18].

La typologie de trajectoires obtenue par l'analyse harmonique qualitative dépend donc fortement de la première étape. Découper la période d'observation en classes revient à imposer *a priori* un calendrier simplifié commun à l'ensemble des individus. Si deux groupes d'individus passent par le même état pendant une durée identique mais à des dates différentes et que ces dates se trouvent dans le même intervalle, alors cette différence de timing disparaît au moment de la construction de la matrice harmonique. Par ailleurs, l'ordre de passage dans les états au sein d'un même intervalle de recodage est également perdu [19]. Cela apparaît d'autant plus problématique qu'une typologie empirique vise précisément à, comme le dit Henry Rouanet, « dégager ce que les données ont à dire » [22] avec aussi peu d'hypothèses que possible. Par conséquent, la détermination des intervalles d'observation en analyse harmonique qualitative revient à faire des hypothèses très fortes sur l'organisation temporelle des trajectoires étudiées alors même que c'est cette organisation temporelle qui est l'objet de l'analyse. En outre, l'utilisation de l'analyse factorielle sur des variables ordonnées dans le temps ne va pas de soi et doit être questionnée. Cette question se pose également avec l'application directe de l'analyse factorielle aux données séquentielle et sera donc abordée une fois cette stratégie présentée. Cela étant dit, il faut souligner qu'il est possible de tirer profit du découpage de la période d'observation en intervalles pour mettre l'accent sur certaines périodes (intervalles courts) qui seraient, d'un point de vue théorique, critiques pour le phénomène étudié et de gagner ainsi en parcimonie. Dans l'exemple des parcours professionnels, la majeure partie des événements intervient avant 30 ans, ce qui justifie un niveau de détail élevé avant 30 ans et plus fruste après [20].

2.2. L'analyse factorielle de séquences

Généralement, la construction de typologie de trajectoires à l'aide de l'analyse des correspondances multiples (ACM) consiste dans un premier temps à mettre les séquences sous forme d'un tableau disjonctif complet. Pour chaque épisode, cela revient à décrire les *J* épisodes par *J* x *K* variables, *K* étant le nombre d'états. Ainsi, pour décrire les parcours d'insertion (56 mois) à la sortie du système éducatif dans un espace comprenant 8 états, on obtient 446 variables prenant chacune pour valeur 0 ou 1 [21]. Une analyse factorielle des correspondances est ensuite réalisée sur ces 446 variables. Le reste de la procédure est identique à celle mise en œuvre pour l'analyse harmonique qualitative : la distance entre les individus est obtenue par l'utilisation de la distance euclidienne sur leurs coordonnées sur les premiers facteurs et cette matrice de distance est soumise à un algorithme de classification ascendante hiérarchique (critère d'agrégation de Ward).

Tableau 5 – Tableau de Burt et données séquentielles

		…	Épisode j						…	
			État 1	État 2	…	État *k*	…	État *K*		
…										
Épisode j	État 1		n_{jj11}	0		0		0		$K \times n_{jj11}$
	État 2		0	n_{jj22}		0		0		$K \times n_{jj22}$
	…									
	État *k*		0	0		n_{jjkk}		0		$K \times n_{jjkk}$
	…									
	État *K*		0	0		0		n_{jjKK}		$K \times n_{jjKK}$
…										
Épisode j'	État 1		$n_{j'j11}$	$n_{j'j12}$		$n_{j'j1k}$		$n_{j'j1K}$		$K \times n_{j'j'11}$
	État 2		$n_{j'j21}$	$n_{j'j22}$		$n_{j'j2k}$		$n_{j'j2K}$		$K \times n_{j'j'22}$
	…									
	État *k*		$n_{j'jk1}$	$n_{j'jk2}$		$n_{j'jkk}$		$n_{j'jkK}$		$K \times n_{j'j'kk}$
	…									
	État *K*		$n_{j'jK1}$	$n_{j'jK1}$		$n_{j'jKk}$		$n_{j'jKK}$		$K \times n_{j'j'KK}$
…										
			$K \times n_{jj11}$	$K \times n_{jj22}$		$K \times n_{jjkk}$		$K \times n_{jjKK}$		

Pour bien comprendre quelles sont les hypothèses qui se trouvent implicitement engagées dans l'analyse factorielle de séquence, il est peut être utile de considérer plutôt le tableau de contingence de Burt, c'est à dire le tableau croisant tous les états pour tous les épisodes (voir Tableau 5). En effet, si l'analyse factorielle d'un tableau de Burt est équivalente avec l'analyse factorielle du tableau

disjonctif complet auquel il correspond, les conséquences de l'analyse factorielle de séquences apparaissent plus nettement sous cette forme. Lorsqu'il se rapporte à des données séquentielles, le tableau de Burt contient le nombre d'individu ayant connu l'état *k* à l'épisode *j* et *k*' à l'épisode *j*. L'ACM des séquences revient à réaliser une analyse factorielle sur le tableau de Burt, autrement dit à analyser le nuage de *J* x *K* points-modalités dans $\Re^{J \times K}$.

Tableau 6 – Tableau des profils de Burt et données séquentielles

		…	Épisode j						…	
			État 1	État 2	…	État *k*	…	État *K*		
…										
Épisode j	État 1		1/ *K*	0		0		0		*1*
	État 2		0	1/ *K*		0		0		*1*
	…									
	État *k*		0	0		1/ *K*		0		*1*
	…									
	État *K*		0	0		0		n_{jjKK}		*1*
…										
Épisode j'	État 1		$n_{j'j11} / K\, n_{j'j'11}$	$n_{j'j12} / K\, n_{j'j'11}$		$n_{j'j1k} / K\, n_{jj'11}$		$n_{j'j1K} / K\, n_{jj'11}$		*1*
	État 2		$n_{j'j21} / K\, n_{j'j'22}$	$n_{j'j22} / K\, n_{j'j'22}$		$n_{j'j2k} / K\, n_{jj'22}$		$n_{j'j2K} / K\, n_{jj'22}$		*1*
	…									
	État *k*		$n_{j'jk1} / K\, n_{j'j'kk}$	$n_{j'jk2} / K\, n_{j'j'kk}$		$n_{j'jkk} / K\, n_{jj'kk}$		$n_{j'jkK} / K\, n_{jj'kk}$		*1*
	…									
	État *K*		$n_{j'jK1} / K\, n_{j'j'KK}$	$n_{j'jK1} / K\, n_{j'j'KK}$		$n_{j'jKk} / K\, n_{jj'KK}$		$n_{j'jKK} / K\, n_{jj'KK}$		*1*
…										

Par conséquent, l'ACM sur données séquentielles consiste à réaliser l'analyse factorielle du tableau des profils de Burt qui se trouve être proportionnel au facteur 1/*K* près aux fréquences de transition entre tous les états et pour toutes les combinaisons d'épisodes (voir Tableau 6). Par exemple, hors matrice diagonale correspondant aux croisements des épisodes avec eux-mêmes, le profil correspondant à l'épisode *j'* et l'état *k* se compose de la série des transitions vers les autres états pour tous les autres épisodes conditionnellement à l'épisode *j'* et l'état *k*. Par conséquent, le tableau des profils de Burt est en fait proportionnel aux matrices de Markov correspondantes au croisement de tous les épisodes. La distance entre deux lignes est la somme des carrés des différences de leurs fréquences de transition pondérés par l'inverse des marges des colonnes. L'analyse factorielle de séquences consiste donc à comparer les transitions vers tous les états pour tous les épisodes conditionnellement à deux états qui peuvent correspondre au même épisode ou à deux épisodes distincts. L'analyse factorielle de séquence consiste donc à identifier les différentes ressemblances et dissemblances qui émergent en termes de transitions et à les ordonner selon leur importance, mesurée par leur inertie. La phase de classification ramène au niveau individuel ces oppositions structurelles, mais l'essentiel de la comparaison des trajectoires se trouve donc dans l'étape factorielle où la mesure de similarité entre séquences est définit en termes de probabilités de transition.

Deux états seront d'autant plus éloignés que les transitions vers les autres états seront différentes, c'est-à-dire vers des états différents pour une même date, ou vers des états semblables à des dates différentes, et ce d'autant plus que les états et dates vers lesquels les transitions s'effectuent sont peu fréquentés. On retrouve ici le biais bien connu des ACM qui consiste à donner plus de poids aux modalités rares, ici les états peu fréquentés à une date donnée. Le risque est donc de voir des séquences être considérées comme proches uniquement en raison du fait qu'elles partagent une transition vers des états peu fréquentés à certains moments quel que soit leur degré de ressemblance par ailleurs. Ce biais de l'ACM est bien connu et considéré comme indésirable puisqu'il risque de perturber les directions des premiers axes factoriels, et avec eux toute l'analyse factorielle [23]. La recommandation usuelle est alors de limiter les modalités rares, éventuellement en procédant à des regroupements. Cette stratégie est évidemment plus difficile à mettre en œuvre ici puisque le regroupement d'états à certaines dates revient à rapprocher arbitrairement des états du seul fait qu'à certains moments ils se trouvent peu fréquentés. S'il est toujours possible de définir l'espace d'états de manière à prévenir ce genre de problème, il n'en reste pas moins que, pour éviter complètement ce biais, il serait nécessaire de considérer la totalité de la distribution temporelle des états. En outre, le fait que certains états soient très fréquentés à certaines dates et moins, voire pas du tout, à d'autres est l'objet même d'une analyse de trajectoire, et à ce titre ne devrait pas être corrigé. C'est, sous une forme différente, le même problème que le découpage de la période d'observation en intervalles pose pour l'analyse harmonique qualitative.

Cependant, l'utilisation de l'analyse factorielle pour la description de séquences pose également, et peut-être même surtout, la question du statut du temps dans l'analyse. Par définition, l'ordre des variables n'intervient pas dans l'analyse factorielle[12]. Les épisodes qui forment les séquences perdent leur caractère ordinal dès lors qu'ils sont introduits dans une analyse factorielle. La mesure de la proximité entre séquences met sur le même plan des transitions entre deux épisodes consécutifs ou très éloignés dans le temps. Cette éviction du temps permet d'identifier d'éventuelles relations qui mettent en jeu des épisodes espacés dans le temps, au risque que cela ne soit qu'en vertu de transitions vers des états peu fréquentés. Toutefois, cela rend également l'interprétation d'autant plus délicate que les axes factoriels qui sont utilisés ensuite dans la procédure de classification ne sont que très rarement tous interprétés. L'objectif étant de produire une classification, ce n'est qu'au niveau de l'interprétation à l'issue de la classification ascendante hiérarchique que le temps resurgit au moment de représenter graphiquement les séquences. L'absence de prise en compte du temps semble toutefois poser problème dans la mesure où s'il peut être souhaitable de mettre en regard des épisodes très éloignés dans le temps, les proximités qui semblent les plus intéressantes (et les plus interprétées) sont celles qui se situent au voisinage des événements considérés. L'intérêt d'une typologie de séquences est de voir comment se déroulent les trajectoires et non pas d'identifier des corrélations entre états éloignés dans le temps. Cet éventuel lien se fait lors de l'interprétation et de l'examen visuel de l'ensemble des trajectoires et non sur les relations entre un petit nombre d'états qui ont de surcroît de grandes chances de rester hors du champ de l'analyse s'ils n'apparaissent pas sur les tout premiers facteurs.

3. Conclusion

Les deux familles de méthodes qui viennent d'être présentées permettent de décrire des données séquentielles selon des modalités très différentes qui s'inscrivent dans deux courants de recherche distincts. Issues de travaux en informatique et en théorie du signal, les méthodes d'appariement optimal appartiennent aux procédures algorithmique qui permettent d'automatiser les opérations que l'on fait intuitivement pour comparer des séquences. Au contraire, les analyses factorielles sont fondées sur l'algèbre linéaire et notamment sur le théorème d'analyse spectrale. Les méthodes d'appariement optimal ont été spécifiquement conçues pour mesurer le degré de ressemblance de séquences et reposent sur la comparaison de toutes les combinaisons d'accélération et de décélération du temps et de substitutions possibles. C'est le coût relatif de ces deux opérations qui permet de définir si c'est la contemporanéité des événements qui importe ou bien au contraire le rapprochement d'événements identiques plus lointains. L'utilisation de l'analyse factorielle pour établir des typologies de trajectoires semble moins assurée puisque cette méthode ne prend pas en compte l'ordre des variables dans le temps. La prise en compte du temps n'intervient qu'au moment de l'interprétation, elle est donc surajoutée à l'analyse. Cela permet de mettre en évidence des relations entre des transitions éloignées dans le temps, au prix d'une interprétation délicate qui passe par l'analyse de l'ensemble des facteurs sur lesquels la classification est construite. Une autre différence essentielle entre analyses factorielles appliquées aux séquences et méthodes d'appariement optimal est leur conception de la similiarité entre séquences. Pour l'analyse factorielle, ce sont les probabilités empiriques de transition qui permettent de rapprocher ou de différencier les séquences alors que pour les méthodes d'appariement optimal, le choix est vaste puisqu'il est possible d'utiliser les transitions, moyennes ou pour l'ensemble des épisodes contigus [25,26], mais également des coûts déterminés théoriquement ou selon d'autres critères empiriques. Cette flexibilité offre l'avantage de pouvoir adapter la mesure de similarité aux données et aux types de régularité recherchées mais exige en contrepartie que soit explicité et justifié le choix des coûts utilisés. Pour terminer, ajoutons que l'accroissement régulier de la puissance des ordinateurs et la diffusion de plus en plus large des programmes qui permettent de mettre en œuvre les méthodes d'appariement optimal (librairie TraMineR pour le logiciel R ; plugin sq pour le logiciel Stata) rend possible l'utilisation de méthodes intensives en calcul spécialement conçues pour l'analyse de séquence qui étaient jusqu'à très récemment hors de portée.

[12] L'homologie des variables n'est en outre pas prise en considération [24, p. 182].

Bibliographie

[1] Levenshtein V.I., « Binary codes capable of correcting deletions, insertions, and reversals », *Soviet Physics Doklady*, Vol. 10, pp. 707-710, 1966 [1965].

[2] Hamming R.W., « Error-detecting and error-correcting codes », *Bell System Technical Journal*, Vol. 29, n° 2, pp. 147-160, 1950.

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

[4] Lesnard L., de Saint Pol Th., « *Introduction aux méthodes d'appariement optimal (optimal matching analysis)* », *Bulletin de Méthodologie Sociologique*, n° 90, pp. 5-25, 2006.

[5] Abbott A., Forrest J., « Optimal Matching Methods for Historical Sequences », *Journal of Interdisciplinary History,* Vol. 16, n° 3, pp. 471-494, 1986.

[6] Abbott A., Hrycak A., « Measuring ressemblance in sequence data : an optimal matching analysis of musicians' careers », *American journal of sociology*, vol. 96, n° 1, pp. 144-185, 1990.

[7] Abbott A., «Sequence analysis: new methods for old ideas », *Annual Review of Sociology*, Vol. 21, pp. 93-113, 1995.

[8] Abbott A., Tsay A., « Sequence analysis and optimal matching methods in sociology », *Sociological Methods and Research*, Vol. 29, n° 1, pp. 3-33, 2000.

[9] Halpin B., Chan T. W., « Class careers as sequences: an optimal matching analysis of work-life histories », *European Sociological Review*, vol. 14, n° 2, pp. 111-130, 1998.

[10] Milligan G.W., « An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms », *Psychometrika*, Vol. 45, n° 3, pp. 325-342, 1980.

[11] Milligan G.W., « A Monte Carlo Study of Thirty Internal Criterion Measures for Cluster Analysis », *Psychometrika*, Vol. 46, pp. 187-199, 1981.

[12] Belbin L., Faith D., Milligan G.W., «A Comparison of Two Approaches to Beta-Flexible Clustering », *Multivariate Behavioral Research*, Vol. 27, pp. 417-433, 1992.

[13] Chan T.W., « Optimal Matching Analysis », *Social Research Update*, Vol. 24, 2002.

[14] Levine J.H., « But what have you done for us lately?: Commentary on Abbot and Tsay », *Sociological Methods and Research*, Vol. 29, n° 1, pp. 34-40, 2000.

[15] Benzécri J.-P., *L'analyse des données tome 2 : l'analyse des correspondances*, Dunod, Paris,1973.

[16] Deville J-C., « Méthodes statistiques et numériques de l'analyse harmonique », *Annales de l'insee*, n°15, pp. 3-101, 1974.

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

[18] Degenne A., Lebeaux M.-O., Mounier L., « Typologies d'itinéraires comme instrument d'analyse du marché du travail », in A. Degenne, M. Mansuy, G. Podevin, P. Werquin (dir.), *Typologie des marchés du travail, suivi et parcours*, Marseille, Document du Céreq n°115, pp. 27-42, 1996.

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

[20] Robette N., Thibault N., « L'analyse exploratoire de trajectoire professionnelle : analyse harmonique qualitative ou appariement optimal ? », *Population*, Vol. 63, n° 4, 2008.

[21] Grelet Y., « Des typologies de parcours. Méthodes et usages », *Document Génération 92*, n°20, 47 p., 2002.

[22] Rouanet H. « Idées Forces », http://www.math-info.univ-paris5.fr/~rouanet/index.html, consulté le 24 février 2009.

[23] Lebart L., Piron M., Morineau A., *Statistique exploratoire multidimensionnelle : visualisation et inférence en fouilles de données*, Dunod, Paris,2006.

[24] Escofier B., Pagès J., *Analyses factorielles simples et multiples. Objectifs, méthodes et interprétation*, Dunod, Paris,1998.

[25] Lesnard L., « Schedules as sequences: a new method to analyze the use of time based on collective rhythm with an application to the work arrangements of French dual-earner couples », *Electronic International Journal of Time Use Research*, Vol. 1, n° 1, pp. 63-88, 2004.

[26] Lesnard L., « Optimal matching and the social sciences », *Document de travail du CREST*, n° 2006-01, Centre de Recherche en Économie et Statistique, Insee, Paris, 2006.