Bulletin de méthodologie sociologique
Bulletin of sociological methodology

90 | 2006
April

Introduction aux méthodes d'appariement optimal (Optimal Matching Analysis)

Laurent Lesnard et Thibaut de Saint Pol

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/bms/638
ISSN : 2070-2779

Éditeur
Association internationale de méthodologie sociologique

Édition imprimée
Date de publication : 1 avril 2006
Pagination : 5-25
ISSN : 0759-1063

Référence électronique
Laurent Lesnard et Thibaut de Saint Pol, « Introduction aux méthodes d'appariement optimal (Optimal Matching Analysis) », *Bulletin de méthodologie sociologique* [En ligne], 90 | 2006, mis en ligne le 01 avril 2009, consulté le 19 avril 2019. URL : http://journals.openedition.org/bms/638

Introduction aux méthodes d'appariement optimal (Optimal Matching Analysis)

Laurent Lesnard et Thibaut de Saint Pol

Introduction

1 Si l'analyse du changement ou des processus occupe une position centrale dans le discours théorique des sciences sociales, il a fallu attendre les années 1980 pour que sa transposition empirique soit possible. La première revue de littérature consacrée à l'analyse statistique dynamique de Nancy Tuma, Michael Hannan et Lyle Groeneveld (1979) invitait ainsi les sociologues à s'approprier les outils théoriques que la diffusion de l'informatique rendait enfin accessibles. Beaucoup de chemin a été parcouru depuis cet appel : si elles ne sont pas aussi couramment utilisées que les modèles de régression classiques, les analyses démographiques des biographies par exemple font désormais partie des outils classiques de la statistique sociale. Autrement dit, il est désormais possible de modéliser finement des processus en sciences sociales. Cependant, en l'absence de toute description préliminaire ces modélisations sont en quelque sorte aveugles : ce n'est qu'au travers de modèles que sont approchés les processus. En effet, la statistique « statique » n'est pas seulement inférentielle : aux côtés des modèles statistiques et économétriques se trouvent également des procédures descriptives fondées sur des décompositions géométriques (les analyses factorielles) ou algorithmiques (classifications) des données, et, plus simplement, sur des moyennes ou d'autres abstractions statistiques élémentaires.

2 De manière générale, ces méthodes descriptives n'ont pas réussi aussi bien la transition dynamique que leurs consœurs inférentielles : la moyenne est par nature statique, les mesures de similarité des classifications n'ont pas été prévues pour les données séquentielles tandis que les capacités dynamiques des analyses factorielles restent pour

l'instant limitées. On en arrive à une situation paradoxale : *en pratique il est actuellement plus facile de modéliser des processus que de les décrire*. Si la complexité des modèles dynamiques est souvent supérieure à celle des outils statiques, leur mise en œuvre reste généralement plus aisée qu'une description des processus. Le principe d'un modèle est en effet de simplifier la réalité, de la ramener à un petit nombre de principes élémentaires censés expliquer l'essentiel des phénomènes observés en fonction d'hypothèses, du choix de privilégier certains aspects sur d'autres.

3 Un bon modèle est par conséquent un jeu d'hypothèses qui offre une représentation simplifiée mais relativement fidèle de la réalité, c'est-à-dire qui permet de distinguer l'information principale du « bruit », de trier ce qui est important de ce qui l'est moins. La description des phénomènes étudiés est donc un préalable indispensable à toute modélisation :

> Ne nous lassons donc pas de répéter que, pour avoir chance de ne pas se prendre à des représentations inexactes et par suite à des coïncidences fortuites ou trompeuses, notre expérimentation statistique doit toujours s'appliquer à saisir, d'abord, dans son allure propre le fait étudié, à le saisir dans la *succession de ses phases*, dans la *décomposition de ses parties* si c'est le cas ; et si elle en simplifie ensuite l'expression, comme il est peut-être utile ou nécessaire pour la recherche même, si elle en laisse tomber telles ou telles particularités pour n'en retenir que certaines autres, elle doit savoir qu'elle fait cette élimination et pourquoi et avec quelles conséquences possibles sur les résultats ultérieurs. (Simiand, 1922, p. 48, italiques par nous).

4 La phase descriptive permet de prendre la mesure d'un phénomène, de confronter une première fois la théorie aux faits. En raison de son potentiel simplificateur, la modélisation statistique autorise l'étude des processus sans qu'il soit nécessaire de les décrire : l'analyse statistique des processus manque actuellement de procédures de description des séquences. L'objectif de cet article est de présenter une nouvelle méthode qui permet précisément de décrire les processus : les Méthodes d'Appariement Optimal (MAO)[1].

5 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 et de Hamming[2], les M.A.O sont plus connues en biologie où elles ont contribué au séquençage du génome. De manière plus générale, les MAO 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, (Abbott et Forrest, 1986 ; Abbott et Hrycak 1990) qui se trouve principalement à l'origine de l'introduction des MAO en sciences sociales au travers de l'étude de processus historiques. Principes que Andrew Abbott a ensuite approfondis dans deux articles (Abbott, 1995 ; Abbott et Tsay, 2000 ; Abbott, 2001).

6 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 MAO permettent de les regrouper et de dégager des idéaux-types. La première étape de cette procédure consiste à calculer une distance 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* (Halpin et Chan, 1998). Après avoir

présenté le principe et la mise en œuvre de cette technique, nous présenterons deux exemples d'utilisation de cette méthode aux emplois du temps des Français.

Comparer des séquences avec les méthodes d'appariement optimal

7 Dans cette première étape, il s'agit d'arriver à comparer des séquences qui peuvent être de longueurs différentes et contenir des éléments divers[3]. 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 MAO 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.

8 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.

9 **Figure 1 : Deux séquences à comparer**

10 X : C – A – B – D – D

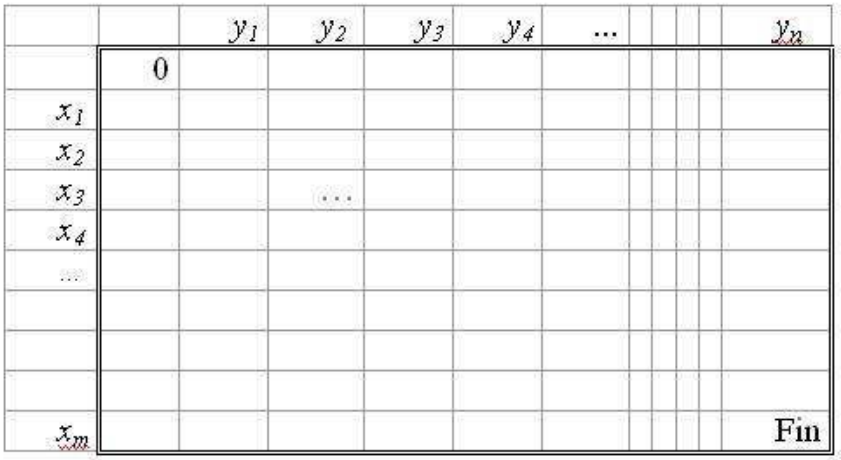
11 Y : A – B – C – D

12 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.

13 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.

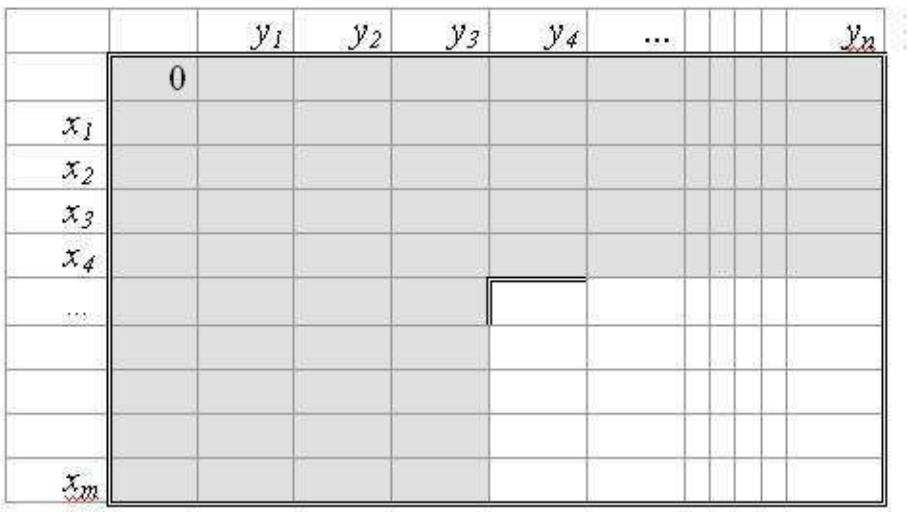
14 Si on généralise ce processus à deux séquences de taille m et n, on peut représenter cette procédure sous la forme d'une matrice de taille m,n. Ainsi, si on compare les séquences X= $(x_1,...,x_m)$ et Y= $(y_1,...y_n)$, on obtient la matrice représentée ci dessous. Passer de X à Y, c'est passer de 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[4].

Figure 2 : Représentation matricielle de la comparaison de deux séquences par les MAO



15 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 MAO



16 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.

Regrouper les séquences voisines

17 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.

18 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 classifications 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.

19 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 sociologue 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 MAO a donné ce critère. Il suffit de retenir une méthode et de regrouper les séquences[5].

Le problème des coûts

20 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 MAO, 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 MAO publiées de 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.

21 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 semblables 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 de la Méthode d'Appariement Optimal

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

22 Le modèle de comparaison de séquences proposé par les MAO 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 identiques des séquences tandis que les opérations de substitution distordent les événements pour mieux comparer leur dimension temporelle. Les MAO alternent donc ces deux types de simplifications que permet de visualiser la représentation matricielle du processus (voir Figure 2 supra) : 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[6]. Au final, les MAO sont donc une combinaison d'accélérations, de ralentissements et d'écoulements normaux[7] du temps qui permettent de comparer 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.

23 Du choix des coûts associés aux trois opérations des MAO 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. Prenons un exemple de deux séquences largement semblables mais dont le calendrier est décalé (voir Figure 4). Avec le système de coût traditionnel dans lequel une insertion-suppression coûte une unité contre deux pour toute substitution, l'appariement optimal est obtenu 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[8].

Figure 4 : Deux séquences décalées

24 X : A – A – A – A – B – B – B – B

25 Y : C – C – A – A – A – A – B – B

26 Plus précisément, les éléments qui apparaissent communs dépendent de l'ordre des événements dans chacune des séquences[9], 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.

27 Au contraire, préserver toute l'échelle de temps de l'action requiert des coûts d'insertion-suppression très élevés[10] mais pose la question de la distance entre événements, question que la stratégie classique supprime littéralement 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, tout est envisageable, depuis l'affectation de coûts guidés théoriquement jusqu'à l'application de critères empiriques. Il est en effet 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. Une contrainte tout de même : afin d'éviter le recours systématique aux opérations d'insertion-suppression, il est indispensable de fixer le coût de ces opérations au moins au niveau du coût de substitution le plus élevé, éventuellement augmenté de la différence qui existe entre les deux coûts de substitution les plus élevés (Abbott et Hrycak, 1990).

28 La comparaison de séquences par la technique de l'Appariement Optimal nécessite donc de simplifier 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 (2000) 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 à la sociologie. Cette critique minimise la portée sociologique des choix qui se trouvent derrière la modélisation statistique : discriminer un signal du bruit, c'est, au travers des hypothèses sur lesquelles s'appuie tout modèle 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, 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 MAO sont moins mystérieuses : 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 la majorité des méthodes statistiques classiques, les Méthodes d'Appariement Optimal rendent visible les enjeux sociologiques de la statistique : elles permettent de véritablement réfléchir et de choisir ce qui convient théoriquement le mieux, comme nous allons le voir avec les deux exemples que nous allons maintenant présenter.

Deux applications à l'emploi du temps des français

29 Les Méthodes d'Appariement Optimal peuvent s'appliquer à bien des objets dès lors que l'étude porte sur une succession d'états ou d'actions. Ce peut être le cas à l'échelle d'une vie ou d'une portion de vie. On peut par exemple étudier les différents métiers qu'ont exercés des individus ou les différents lieux sur lesquels ils ont habité. Mais on peut également raisonner sur des périodes plus courtes : une année, un mois ou une semaine. Les deux applications présentées ici portent sur une durée encore plus courte : elles s'intéressent aux activités réalisées pendant une journée à partir des deux dernières enquêtes Emploi du Temps de l'INSEE. L'emploi du temps est en effet un bon exemple de matériel séquentiel : tout au long de la journée, les individus enchaînent des activités plus ou moins longues. Calculer les moyennes des temps consacrés à chaque activité, comme on le fait fréquemment, ne suffit cependant pas à rendre compte de la diversité des organisations quotidiennes. Travailler huit heures en continu n'est pas pareil que de consacrer quatre plages de deux heures au travail à divers moments de la journée.

30 Les Méthodes d'Appariement Optimal permettent justement de tirer parti de la spécificité de telles données. Les études développées ci-dessous proposent deux utilisations différentes de cette méthode à partir du même matériau. La présentation de chacun de ces exemples mériterait lui-même beaucoup plus d'espace : on se concentrera ici sur les apports empiriques des MAO et on ne fera qu'esquisser les éléments théoriques. Le premier exemple peut être considéré comme une application typique des techniques d'appariement optimal : les deux ensembles d'opérations sont ici utilisés de manière à faire émerger la manière dont le dîner s'inscrit dans la soirée des Français. Le second, au contraire, illustre la souplesse de la technique d'appariement de séquences et propose

d'identifier les différents types d'horaires de travail à l'aide des seules opérations de substitutions, de manière à préserver au maximum leurs dimensions temporelles.

L'étude du repas dans le cadre de la soirée

31 L'étude des comportements alimentaires s'effectue généralement indépendamment de l'analyse des autres temps quotidiens. On analyse fréquemment les durées moyennes que les femmes dédient à l'alimentation par rapport à celles des hommes, ou encore celles des plus jeunes par rapport aux plus vieux, sans s'intéresser aux activités qui encadrent les repas. Or il existe une forte interdépendance en termes d'horaires, de lieu, ou même de compagnie entre les différentes activités qui composent notre emploi du temps. Si le repas du midi s'effectue souvent à l'extérieur du domicile et dure moins longtemps en moyenne que le dîner, c'est parce qu'il s'inscrit généralement au milieu d'activités professionnelles qui influent directement sur la manière dont va se dérouler cette prise alimentaire. Le repas du soir quant à lui se déroule entre des activités plus diversifiées qui vont des travaux domestiques aux loisirs les plus divers, en passant par le travail professionnel ou même le sommeil.

32 Comprendre les logiques qui président à la réalisation du dîner, ce n'est pas seulement étudier sa durée, l'heure de son commencement ou encore calculer le temps moyen que lui consacre telle ou telle catégorie de la population. On ne peut également se contenter d'une analyse du type que celles que permettent les régressions. Expliquer les pratiques alimentaires en fonction de paramètres tels que l'âge, le sexe ou encore le lieu de résidence comporte un intérêt certain. Mais ces démarches négligent la possibilité d'une causalité multiple inscrite dans la temporalité : le fait que tel individu dîne sur telle plage horaire est le résultat d'un processus auquel ont contribué, par exemple, le fait qu'il ait quitté tard le bureau, qu'il soit ensuite resté une heure dans les embouteillages, mais qu'il ne veuille en aucun cas manquer le film qui débute à 20h50. Ce sont ces contraintes imposées par les activités qui encadrent le repas qui permettent de comprendre la manière dont se déroule le dîner et qui participent par leur récurrence à la construction d'habitudes alimentaires. Une régression ferait par exemple apparaître la corrélation entre les repas les plus courts et le fait d'être actif. Mais elle oublierait que l'explication de cette relation se situe dans l'enchaînement des différentes activités.

33 Afin de décrire le contexte dans lequel s'effectue le dîner, cette étude s'est intéressée à la période 18h50-21h30 pendant laquelle se concentrent les prises alimentaires de la soirée. Les activités présentes dans les carnets journaliers de l'enquête Emploi du Temps 1998 ont été regroupées en 25 catégories (qui vont du travail aux rencontres en passant par la travail ménager) qui sont autant d'éléments possibles constitutifs de nos séquences. L'Appariement Optimal porte sur des séquences de même taille, c'est-à-dire de 16 éléments qui correspondent aux 16 plages horaires de dix minutes de la période étudiée. A aucun moment de ce traitement séquentiel, les plages alimentaires ne seront privilégiées dans le regroupement des séquences. Ce point qui pourrait paraître anodin est en fait fondamental. Trop de classifications font apparaître des dissemblances qui découlent directement de ce choix d'agrégation et qui biaisent l'analyse sociologique qui s'y rapporte. Ici, le regard du sociologue n'intervient pas dans le processus de regroupement des séquences.

34 Mais les MAO ont un autre intérêt : c'est une technique particulièrement flexible qui s'adapte très aisément aux contraintes imposées par le matériau utilisé et à la théorie au

travers du choix des coûts des différentes opérations. Ici, nous avons calculé les coûts de chacune des trois opérations en termes de fréquences des différents éléments constitutifs des séquences. Ainsi, le coût de substitution d'un épisode de travail par un épisode de sommeil, situation peu courante dans nos emplois du temps, sera élevé. Au contraire, le coût de la substitution d'un repas par un épisode de télévision sera plus faible[11]. Ces coûts sont donc calculés sur l'ensemble des 16 épisodes : ils ne tiennent pas compte de la tranche horaire considérée. Il pourrait en effet être plus probable que la télévision suive le repas à 21h00 qu'à 19h00. C'est pourquoi nous l'avons introduite dans les coûts d'insertion et de suppression. Il est important pour notre étude des rythmes que l'algorithme puisse prendre en compte cette dimension temporelle. Dans les emplois du temps, une opération n'a pas toujours le même coût quelle que soit l'heure à laquelle l'activité se déroule.

35 Or ce sont les insertions et les suppressions qui, en décalant les activités de dix minutes à chaque fois, posent le problème de différence des tranches horaires. L'emploi de ces deux opérations est ce qui fait la singularité et l'originalité de cette analyse séquentielle. Mais une trop grande utilisation de ces mouvements revient à décaler totalement les éléments de nos séquences et à perdre de ce fait les particularités de chaque tranche horaire. Il nous fallait donc autoriser le recours à ces deux opérations, tout en en empêchant une utilisation abusive. Nous avons choisi de rehausser légèrement les coûts d'insertion et de suppression par rapport aux substitutions pour que ces dernières soient privilégiées, suivant en cela les recommandations d'Abbott et Hrycak (1990). L'étude comparative des résultats avec ou sans cette modification montre qu'elle augmente la variance inter classe et diminue celle intra classe : limiter le recours à l'insertion et à la suppression a permis d'éviter des classements abusifs et a amélioré l'homogénéité de nos classes. Ce jeu sur le coût des opérations est moins complexe qu'il ne peut paraître de prime abord. C'est un des intérêts de cette méthode d'analyse. Si l'objet d'étude n'intervient pas directement dans la construction des classes, il est néanmoins possible d'adapter simplement l'algorithme de la méthode aux particularités temporelles de cet objet afin de coller au plus près à la réalité sociale que traduit la séquence.

36 Ces choix ont conduit à regrouper dans notre exemple les séquences en dix classes que l'on peut résumer très imparfaitement à partir du tableau 2[12]. Mis à part la première classe qui consacre beaucoup de temps au repas et les deux dernières où le temps alimentaire est faible, le recours par exemple aux moyennes voile un certain nombre de pratiques fort diverses que le recours à une analyse séquentielle permet de mettre en lumière. Ainsi, les temps moyens consacrés au repas pour les quatrième et huitième classes sont très proches. Pour autant, le repas est pris en début de période pour le premier groupe et en seconde partie pour le deuxième groupe. Cette observation simple suffit à éclairer les possibilités nouvelles qu'offrent les MAO Ainsi, au sein de ce qui n'était auparavant qu'un groupe assez difforme d'individus au temps moyens alimentaires similaires, il est désormais possible de distinguer différents types de pratiques.

Tableau 2 : Descriptif des dix classes

Classe	Effectif (en %)	Temps moyen consacré au REPAS de 18h50 à 21h30	*Nom de la classe*
1	6,0	113 min	*Les Mangeurs*
2	3,7	43 min	*Les Couche-tôt*
3	20,8	37 min	*Soirée télé*
4	12,0	42 min	*Les Dîne-tôt*
5	5,8	48 min	*Cuisine et Ménage*
6	24,5	36 min	*Les Dîne-tard*
7	7,5	40 min	*Deuxième journée*
8	5,4	43 min	*Les Téléphages*
9	9,4	26 min	*Les Travailleurs*
10	4,9	29 min	*Sorties*

Séquence-type[13]				
19h00-19h30	19h30-20h00	20h00-20h30	20h30-21h00	21h00-21h30
Repas	Repas	Repas	Repas	Repas
Repas	Repas	Télévision	Sommeil	Sommeil
Repas	Repas	Télévision	Télévision	Télévision
Repas	Repas	Télévision	Télévision	Télévision
Repas	Repas	Ménager	Enfants	Télévision
Bricolage	Repas	Repas	Repas	Télévision
Ménager	Ménager	Repas	Repas	Télévision
Télévision	Télévision	Repas	Repas	Télévision
Travail	Travail	Travail	Travail	Travail
Rencontres	Rencontres	Rencontres	Rencontres	Rencontres

Source : Enquêtes Emploi du Temps de l'INSEE de 1998

37 Chacune de ces soirées-type fait apparaître une logique d'insertion du dîner dans la soirée et montre l'importance du contexte dans lequel se déroule cette prise alimentaire. Cette approche nouvelle est d'autant plus intéressante que l'on peut caractériser ces séquences

au moyen des caractéristiques des individus auxquelles elles appartiennent. On va ainsi pouvoir opposer par exemple des séquences d'activités féminines, marquées par le poids du travail ménager comme la classe 7 par exemple, s'opposant à des séquences plus masculines comme la classe 8, où la télévision occupe une grande place.

38 Par ailleurs, le recours à une méthode d'Appariement Optimal permet de dépasser certaines limites propres aux analyses classiques. Dans le cas de notre exemple, le passage entre l'enquête Emploi du Temps de 1985 et celle de 1998 de l'interligne du carnet journalier rempli par les personnes interrogées de cinq à dix minutes conduit à un biais méthodologique qui empêche de mener des comparaisons satisfaisantes des durées des activités entre les deux enquêtes. En effet, les activités les plus courtes, comme mettre ou débarrasser la table ou plus généralement le travail domestique qui est très fractionné, ont été souvent intégrées par les enquêtés dans des activités plus longues. Ainsi, on observe en moyenne entre 1985 et 1998 une augmentation de près de dix minutes du temps consacré aux repas, qui est plus que suspecte[13]. L'utilisation de MAO rend possible le dépassement de ce biais méthodologique en ne considérant pour 1985 qu'une ligne pour deux ; ce qui a pour conséquence directe de faire disparaître la moitié des activités de cinq minutes et de prendre en compte leur relative disparition en 1998. L'application du même protocole réalisé pour les soirées des Français en 1998 pour les données de 1985 amène ainsi à la construction d'une typologie extraordinairement proche de celle présentée ci-dessus. Ce qui offre d'ailleurs une possibilité de contrôle de la grande robustesse des classes obtenues au moyen de MAO. Ce bref aperçu de l'approche novatrice autorisée par l'utilisation de cette technique pour la compréhension des pratiques alimentaires milite pour son adoption et sa mise en pratique à d'autres champs de la sociologie.

Le temps du travail

39 L'analyse scientifique du temps detravail est généralement réduite dans les enquêtes emploi du temps à de simples durées[14], ce qui occulte nombre de variations (Godard, 2003). Ainsi, des positions *a priori* opposées comme celles de la diminution du temps consacré au travail (Robinson et Godbey, 1999) et de l'extension du *workaholism*[15] (Schor, 1993) peuvent-elles être réconciliées dès lors que les moyennes nationales sont décomposées selon la position sociale ou le niveau d'éducation : la thèse du renversement du gradient du niveau d'éducation-travail (Gershuny, 2000 ; Chenu, 2002) permet à cet effet de réconcilier ces deux théories en soulignant les changements des rapports entretenus entre position dans la hiérarchie sociale et temps de travail.

40 Toutefois, cette décomposition de moyenne ne permet pas de relier les évolutions des heures moyennes travaillées avec un autre thème majeur, celui de la *flexibilité,* des horaires de travail notamment. La moyenne ne permet donc pas d'appréhender le travail dans son déroulement, de connaître la répartition des heures travaillées dans la journée. De la même manière, les indicateurs de flexibilité apparaissent sensibles à la durée du travail et à la répartition du travail dans la journée : puisqu'ils sont construits *a priori*, les indicateurs de flexibilité sont bien souvent hétérogènes. On peut citer l'exemple du travail de nuit des Enquêtes Emploi : une personne travaille de nuit si sa période d'activité se situe, même partiellement, entre minuit et cinq heures du matin. Les journées de travail qui commencent à cinq heures (horaires décalés le matin), celles qui

se terminent à minuit (horaires décalés le soir) se trouvent ainsi mélangées au véritable travail de nuit.

41 Seule une classification peut conjuguer régularité statistique et diversité et dépasser ainsi l'antagonisme de la moyenne et des indicateurs *a priori*. Pour mesurer la dissimilarité des journées de travail en termes de durée mais également de répartition des heures travaillées dans la journée, l'approche séquentielle des Méthodes d'Appariement Optimal semble idéale. Le recodage binaire (travail/non-travail) des carnets des enquêtes Emploi du Temps de 1985 et 1998 associé à un algorithme d'appariement optimal devrait donc permettre de construire une typologie des horaires de travail en France.

42 Toutefois, comme il ne s'agit pas ici d'identifier des enchaînements typiques, les journées étant stylisées à l'extrême à l'opposé de l'exemple précédent, mais au contraire d'identifier les *décalages* temporels du travail, les opérations de substitutions doivent être fortement privilégiées au détriment des opérations d'insertion-suppression (qui brouilleraient les décalages des horaires de travail). Mieux, les opérations d'insertion-suppression peuvent être bannies du processus d'appariement puisque seule la dimension temporelle du travail nous intéresse ici. Cet exemple illustre donc particulièrement bien la souplesse de l'analyse d'appariement qui, dans ce cas très particulier, n'est plus *optimal* [16].

43 Reste à déterminer les différents coûts de substitution entre les deux états travail et non-travail. Si la théorie sociologique ne semble pas ici en mesure de déterminer directement de tels coûts, elle peut cependant guider leur construction : puisque, comme l'a montré Durkheim (1912), le temps est un système symbolique qui, parce qu'il cristallise le rythme de l'activité collective, permet d'anticiper les régularité sociales, c'est le rythme collectif qui va fournir le moyen de différencier les différents emplois du temps de travail. En effet, la traduction du postulat durkheimien de la différenciation sociale du temps, autrement dit la différenciation du flux incessant d'événements par l'activité collective, en des termes plus opérationnels nous donne un moyen de détermination des coûts de substitution : c'est la position relative des emplois du temps individuels par rapport au rythme collectif qui va nous donner une mesure de la similarité des emplois du temps.

44 Le rythme de l'activité collective qui nous intéresse ici est le rythme du travail et peut être approché simplement par les « flux » entre les deux états « travail » et « non-travail » : un flux élevé entre ces deux états signifie qu'un changement de rythme est en cours donc qu'un travailleur et un inactif sont assez proches puisqu'ils risquent de partager le même état[17]. Au contraire, une faible circulation entre ces états est signe d'un certain hermétisme (les deux rythmes coexistent), ce qui fait qu'un travailleur et un inactif seront alors éloignés. Par exemple, la transition entre travail et non-travail a de bonnes chances d'être élevée vers 9h, ce qui va limiter la distance entre un travailleur et un inactif. En revanche, vers 3h ou 15h, cette même transition sera très vraisemblablement plus faible, ce qui accentuera la différence entre un travailleur et un inactif à de telles heures. L'appariement des horaires proposé s'accorde donc avec le sens commun qui voit la différence comme un écart à la norme : des horaires ne deviennent atypiques qu'en relation à une norme collective de rythme de travail. Plutôt que de la fixer arbitrairement, la norme émerge ici des régularités observées : c'est le rythme collectif qui va déterminer le degré de différence entre deux horaires de travail : la mesure de dissimilarité proposée est donc à la fois endogène et dynamique[18]. Au final, la distance entre deux emplois du temps individuels est obtenue par la somme de leurs

différences instantanées, i.e. par la suite de leurs positions relatives par rapport aux rythmes temporels du champ considéré.

45 La mise en œuvre qui vient d'être décrite, associée à l'algorithme WPGMA flexible de classification ascendante hiérarchique nous permet d'identifier douze horaires de travail typiques. Ces types peuvent être décrits à l'aide de deux indicateurs : la mi-journée de travail et la durée de cette journée de travail, autrement dit par un indicateur de position centrale et un autre indiquant la dispersion autour de cette tendance. L'interprétation de la plupart des douze classes est aisée (voir Tableau 2 *supra*[19]).

Tableau 3 : Principales caractéristiques des douze types d'horaires de travail

No. classe	Type d'horaire de travail	Effectifs (% de la pop. tot.)	Mi-journée de travail	Durée de travail
	Standard	**56,5%**	**12:59**	**8:26**
1	7-16	7,6%	12:00	8:14
2	8-18	38,2%	12:53	8:17
3	9-19	10,7%	14:01	9:09
	Décalé	**14,4%**		**7:16**
4	Matin	5,3%	9:44	7:39
5	Après-midi	5,4%	15:32	6:46
6	Soir	2,1%	17:02	7:20
7	Nuit	1,7%		7:38
	Extensif	**9,1%**	**13:57**	**10:29**
8	Régulier	3,5%	12:54	10:47
9	Irrégulier	5,6%	14:38	10:18
	Irrégulier	**20,0%**	**12:50**	**3:45**
10	Fragmenté	3,2%	13:21	3:50
11	Étalé	3,5%	12:15	8:06
12	Faible durée	13,3%	12:52	2:14

Source : Enquêtes Emploi du Temps de l'INSEE de 1985-86

Lecture : la première classe (No. 1) appartient au sous-groupe des horaires standards et représente 7,6% des journées travaillées. La mi-journée de travail de ce type d'horaire se situe en moyenne à midi alors que sa duré moyenne est de huit heures et quart.

46 Les trois premiers types constituent des horaires standard correspondant à une journée de travail de huit heures avec des horaires de bureau centrés autour de la mi-journée (13h) et regroupent un peu plus de la moitié des journées travaillées. Ces trois types d'horaires de travail apparaissent conformes à ce que l'on considère comme une « journée de travail normale » : la technique d'appariement optimal permet donc d'isoler les horaires de travail les plus courants, dont les caractéristiques apparaissent conformes à la norme tacite des horaires de travail « normaux ».

47 La déviance la plus conséquente à cette norme de journée de travail repose essentiellement sur une divergence temporelle considérable de la mi-journée de travail par rapport à la mi-journée « normale » qui se situe ici aux alentours de 13h. Ces types d'horaires peuvent être considérés comme atypiques et contiennent notamment le travail de nuit[20] qui ne correspond pas du tout aux définitions classiques retenues usuellement. Dans les enquêtes Emploi de l'INSEE, est considéré comme travail de nuit toute période de travail située, même partiellement, entre minuit et cinq heures du matin : l'analyse proposée ici permet en quelque sorte d'affiner cette catégorie avec laquelle elle se superpose en partie, mais surtout, parce qu'elle ne repose pas sur des règles strictes fixées *a priori*, elle augmente significativement le nombre des horaires atypiques[21]. La durée moyenne inférieure à huit heures indique que ces horaires de travail contiennent une proportion importante de journées partiellement travaillées, autrement dit que la réduction du temps de travail s'accompagne d'une marginalisation de la répartition de ces heures travaillées dans la journée.

48 Mais les horaires décalés ne sont pas la seule source de déviance par rapport aux horaires de travail « normaux » : les longues journées de travail peuvent également être légitimement considérées comme atypiques. Deux classes d'horaires de travail présentent ainsi une durée de travail supérieure à dix heures, situation qui représente près de 10 % des journées travaillées. De même, les petites journées de travail apparaissent non-standard, anormalité de durée parfois redoublée par une fragmentation de ce travail au cours de la journée. Ces types d'horaires sont la conséquence du processus de sélection des journées travaillées (une journée est considérée comme travaillée dès lors qu'elle présente au moins une déclaration de travail dans le carnet d'emploi du temps) et contiennent un nombre non négligeable de séances de travail le week-end de cadres ou d'enseignants de même que d'horaires de travail fragmentés de certaines catégories d'employés comme les caissières qui peuvent enchaîner deux séances de travail 10-13h et 16-20h dans une journée (Bouffartigue et Pendariès, 1994 ; Prunier-Poulmaire, 2000).

49 Ainsi, contrairement à l'image véhiculée par les indicateurs construits à partir de règles rigides, les horaires atypiques, loin d'être minoritaires, représentent une part presque équivalente à la journée de travail « normale ». Parce qu'ils sont trop synthétiques, indicateurs et moyennes fragmentent et figent le travail, autrement dit offrent une vision partielle des transformations du travail. Seule une approche en termes de séquence permet de lier les changements de la durée de la journée de travail avec la flexibilité des horaires et d'apercevoir ainsi que la réduction de la durée de travail s'accompagne souvent d'une répartition non-standard de ces horaires.

Conclusion

50 Les Méthodes d'Appariement Optimal, en mettant au premier plan la séquence au sein de l'analyse sociologique, permettent non seulement de décrire autrement les phénomènes sociaux, mais remettent en lumière la dimension temporelle de la causalité. Penser en séquences, c'est saisir l'action au travers de sa durée et de ses bornes, comme on le fait habituellement lorsque par exemple on calcule des moyennes ou que l'on fait des régressions. Mais c'est aussi considérer l'action parmi un enchaînement d'autres actions qui ont elles aussi une durée et des bornes qui influent sur les éléments qui les précèdent ou qui les suivent.

51 Une action ne se construit pas isolément dans une boîte noire en fonction de facteurs tel que l'âge, le sexe ou la profession du sujet. Elle est presque toujours déterminée par la suite d'actes dans laquelle elle s'inscrit. Ainsi, c'est parce qu'elle n'a pas un revenu suffisant que telle employée doit se passer des services d'une nourrice et qu'elle doit chaque matin conduire ses enfants à l'école. Et c'est parce que l'école de ses enfants ouvre à un horaire fixe que cette personne arrive fréquemment en retard au travail. Le processus de causalité s'établit dans la chronologie. C'est précisément cette chronologie que les Méthodes d'Appariement Optimal permettent de mettre en lumière.

52 L'intérêt de cette technique ne se limite pas à l'étude des emplois du temps. Elles s'appliquent à toutes les données dynamiques, notamment aux carrières (Abbott et Hrycak, 1990 ; Halpin et Chan, 1998). Si le principe de ces méthodes repose sur l'optimisation des opérations élémentaires engagées dans toute comparaison manuelle de séquences – insertion, suppression et substitution – l'automatisation de ce traitement exige que soient explicitées les règles de la comparaison au travers des coûts qui sont affectés à ces opérations. À cet égard, les Méthodes d'Appariement Optimal permettent de réconcilier les oppositions artificielles entre théorie et pratique, et entre traitement quantitatif et qualitatif des faits sociaux. C'est ce que note Jean-Louis Fabiani (2003) quand il souligne le caractère intégrateur de cette approche qui tient à la fois de la démarche analytique et de la démarche narrative. Les Méthodes d'Appariement Optimal renouvellent considérablement les perspectives quantitatives de la sociologie et méritent d'être intégrées à la boîte à outil du sociologue pour être utilisées quand les besoins théoriques l'exigent.

53 **6.** C'est la raison pour laquelle le même coût est attribué à ces opérations symétriques, symétrie qui apparaît clairement dans la représentation matricielle des MAO.

54 **7.** Par « écoulement normal du temps » il faut entendre « conformément au rythme de l'échelle de temps des séquences ».

BIBLIOGRAPHIE

Abbott A., 1995. - « Sequence Analysis: New Methods for Old Ideas », *Annual Review of Sociology*, Vol. 21, pp. 93-113.

Abbott A. et Forrest J., 1986. - « Optimal Matching Methods for Historical Sequences », *Journal of Interdisciplinary History*, Vol. 16, No. 3, pp. 471-494.

Abbott A. et Hrycak A., 1990. - « Measuring resemblance in sequence analysis: an optimal matching analysis of musicians careers », *American Journal of Sociology*, Vol. 96, No. 1, pp. 144-185.

Abbott A. et Tsay A., 2000. - « Sequence Analysis and Optimal Matching Methods in Sociology », *Sociological Methods and Research*, Vol. 29, No. 1, pp. 3-33.

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

Bouffartigue, P. et Pendaries, J.-R., 1994. - « Formes particulières d'emploi et gestion d'une main-d'œuvre féminine peu qualifiée : le cas des caissières d'un hypermarché », *Sociologie du travail*, Vol. 36, No. 3, pp. 337-359.

Chan T. W., 1999. - « Optimal Matching Analysis », *Social Research Update*, 24, University of Surrey.

Chenu, A., 2002. - « Les horaires et l'organisation du temps de travail », *Economie et Statistique*, No. 352-353, pp. 151-167.

Durbin R., Eddy S. R., Krogh A. et Mitchison G., 1998. - *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge (UK) et New York, Cambridge University Press.

Durkheim E., 1912. - *Les formes élémentaires de la vie religieuse*, Paris, Alcan. Rééd. Presses Universitaires de France, « Quadrige », Paris, 1965.

Fabiani J.-L., 2003. - « Pour en finir avec la réalité unilinéaire. Le parcours méthodologique de Andrew Abbott », *Annales HSS*, mai juin, No 3, pp. 549-565.

Gershuny J., 2000. - *Changing times: work and leisure in Postindustrial Society*, Oxford, Oxford University Press.

Godard, F., 2003. - « Les temps du quotidien », *in* O. Donnat and P. Tolila (dir.), *Le(s) public(s) de la culture*, Paris, Presses de Sciences Po.

Halpin B. et Chan T. W., 1998. - « Class careers as sequences: an optimal matching analysis of work-life histories », *European Sociological Review*, Vol. 14, No. 2, pp. 111-130.

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

Larmet G., 2002. - « La sociabilité alimentaire s'accroît », *Economie et Statistique*, No. 352-353, pp. 191-211.

Lesnard L., 2004. - « 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 or Time-Use Research*, Vol. 1, No. 1, pp. 60-84.

Lesnard L., 2006a. – « Optimal Matching and Social Sciences », *Document de travail du CREST*, No. 2006-01.

Lesnard L., 2006b. - « Flexibilité des horaires de travail et inégalités sociales », dans INSEE (dir.), *Données Sociales - La société française*, Paris, INSEE.

Lesnard, L., 2006c. - « Flexibilité et concordance des horaires de travail dans le couple », dans INSEE (dir.), *Données Sociales - La société française*, Paris, INSEE.

Levenshtein, V. I., 1966. - « Binary codes capable of correcting deletions, insertions, and reversals », *Soviet Physics Doklady*, Vol. 10, pp. 707-710. Traduction du Russe de l'article publié dans *Doklady Akademii Nauk SSSR*, Vol. 163, No. 4, pp. 845-848, 1965.

Levine J., 2000. - « But What Have You Done for Us Lately ? Commentary on Abbott and Tsay », *Sociological Methods and Research*, Vol. 29, No. 1, August.

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

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

Prunier-Poulmaire S., 2000. - « Flexibilité assistée par ordinateur. Les caissières d'hypermarché », *Actes de la recherche en sciences sociales*, No. 134, pp. 29-65.

Robinson J. et Godbey G., 1999. - *Time for life: the surprising ways Americans use their time*, Oxford, University Park, The Pennsylvania State University Press, 2e édition.

Saint Pol, T. de, 2005. - « Le dîner des Français : étude séquentielle d'un emploi du temps », *Document de travail du CREST*, No. 2005-19.

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

Schor, J., 1991. - *The Overworked American : The Unexpected Decline of Leisure*, New York, Basic Books.

Simiand F., 1922. - *Statistique et expérience. Remarques de méthode*, Paris, Marcel Rivière, Bibliothèque des sciences économiques et sociales.

Tuma N. B., Hannan M. T. et Groeneveld L. P., 1979. - « Dynamic analysis of event histories », *The American Journal of Sociology*, Vol. 84, No. 4, pp. 820-854.

NOTES

1. Traduction de l'expression anglaise Optimal Matching Analysis (OMA).
2. Voir Levenshtein (1966) et Hamming (1950).
3. Pour plus de détails techniques, voir Sankoff et Kruskal (1983) et Durbin *et al.* (1998).
4. Ce graphique et le suivant sont inspirés de Chan (1999).
5. 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 MAO s'accommodent mal *a priori* de la mesure d'agrégation de CAH euclidienne (la méthode de Ward). 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 (Milligan 1980 et 1981 ; Belbin, Faith et Milligan, 1992). La méthode WPGMA flexible est disponible dans R, SAS et

ClustanGraphics mais reste indisponible dans la version 14 de SPSS et 9 de Stata. La méthode UPGMA flexible est disponible uniquement dans R.

8. Lorsque les séquences comparées sont de même longueur, l'utilisation des seules opérations de substitution revient à appliquer la distance de Hamming.

9. Ce n'est donc pas un simple dénombrement des éléments communs de chaque séquence.

10. Voire de réduire l'analyse aux seules opérations de substitution, solution qui est présentée plus loin dans le second exemple.

11. En termes mathématiques, ces coûts sont les inverses des probabilités de transition entre deux activités sur l'ensemble des séquences de notre échantillon.

12. Pour plus de précision, voir Saint Pol (2005).

13. Il faut donc interpréter avec prudence les résultats de G. Larmet (2002) qui conclut, uniquement en termes de durée, à l'accroissement de la sociabilité alimentaire entre 1985 et 1998.

14. La remarque s'applique également aux dernières enquêtes Emploi de l'INSEE dont la question sur l'heure de début et de fin du travail se transforme invariablement dans les exploitations en simple durée.

15. Terme anglo-saxon qui désigne les travailleurs compulsifs.

16. En effet, sans opérations d'insertion-suppression, un seul chemin est possible : celui situé sur la diagonale de la matrice d'appariement.

17. En termes statistiques, ces flux sont mesurés par les matrices de transition entre les différents états. Pour plus de détails, voir Lesnard (2006a).

18. Cette version des MAO peut être vue comme un cas particulier de la distance de Hamming pondérée par la série des matrices de transition entre épisodes. Pour une présentation plus complète et technique, voir Lesnard (2004 et 2006a). Cette méthode est disponible sur Internet sous la forme d'une extension Stata (voir http://laurent.lesnard.free.fr).

19. Seuls les résultats pour 1985-86 sont présentés ici. Pour plus de détails, voir Lesnard (2006b et 2006c).

20. Le travail de nuit est ici très particulier puisque deux « journées » de travail sont partiellement observées, le travail de nuit ne correspondant pas à la fenêtre d'une journée des enquêtes Emploi du Temps françaises.

21. À peu près 20 % des horaires non-standard identifiés par la classification entrent dans la définition du travail de nuit de l'enquête emploi. Par conséquent, identifier les horaires atypiques aux seules périodes de travail de nuit limite singulièrement l'appréciation de l'importance des horaires décalés. Si, par définition, les horaires de nuit sont complètement inclus dans le travail de nuit, seuls 10 % des horaires du matin et 30 % des horaires du soir entrent dans le champ du travail de nuit, sans parler de l'exclusion des horaires décalés dans l'après-midi.

RÉSUMÉS

Cet article vise à présenter les fondements d'une nouvelle technique statistique qui permet de décrire les séquences : les Méthodes d'Appariement Optimal (MAO). Empruntée à la biologie moléculaire, cette technique repose sur des principes assez simples et peut être adaptée aux exigences théoriques de l'analyse. Parce qu'elles permettent de comparer des séquences sans

présumer de relations de cause à effet, les MAO présentent également de nombreux atouts pour le sociologue puisqu'elles lui permettent de retemporaliser l'action en la saisissant en termes de processus et de proposer une nouvelle approche des faits sociaux. Deux applications de cette méthode aux emplois du temps des Français sont proposées pour illustrer le fonctionnement et la flexibilité mais également l'intérêt sociologique des Méthodes d'Appariement Optimal.

Introduction to Optimal Matching Analysis: This paper provides an introduction to a new statistical technique for describing sequences: Optimal Matching Analysis (OMA). Borrowed from biology, this technique is based on rather simple principles that can be adapted to suit the theoretical requirements of different analyses. Since OMA is not based on any causal assumptions, it is particularly well-adapted for sociologists who retemporalize action by analyzing it as a process. Two applications of Optimal Matching to time-use data are proposed to demonstrate its flexibility and also its sociological interest.

INDEX

Keywords : Dining, Epistemology, Evening Meal, Optimal Matching Analysis, Sequences, Work Schedules
Mots-clés : Dîner, Epistémologie, Horaires de travail, Méthodes d'appariement optimal, Repas du soir

AUTEURS

LAURENT LESNARD

Observatoire sociologique du changement - Sciences-po et CNRS ; Laboratoire de sociologie quantitative du Crest – INSEE, lesnard@laposte.net

THIBAUT DE SAINT POL

Observatoire sociologique du changement - Sciences-po et CNRS ; Laboratoire de sociologie quantitative du Crest - INSEE, thibaut.desaintpol@ensae.fr