

Congrès AFSP 2009

Section thématique 13

Variables, Individus, Contextes. Comment observer et analyser leurs interactions ?

Axe 2

Laurent Lesnard (Sciences Po, OSC)

laurent.lesnard@sciences-po.fr

Thibaut de Saint Pol (INSEE, CREST)

thibaut.de-saint-pol@insee.fr

Décrire les données séquentielles en sciences sociales avec les méthodes d'appariement optimal

Que l'objectif soit de décrire les trajectoires d'insertion sur le marché du travail, les emplois du temps ou les comportements électoraux, disposer d'outils adaptés pour décrire les données séquentielles ou longitudinales est essentiel pour les chercheurs en sciences sociales. Cette communication a pour objectif de présenter les Méthodes d'Appariement Optimal (en anglais *Optimal matching analysis*) technique qui s'impose comme la méthode de référence pour dresser des typologies empiriques de séquences.

Issues des travaux en théorie du signal dans les années 1950 et 1960, les Méthodes d'Appariement Optimal reposent sur un principe simple et l'automatisation des opérations que l'on fait intuitivement pour comparer des séquences entre elles. Elles 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. Au travers d'exemples, ce texte a pour objectif de montrer comment la flexibilité de cette méthode permet de l'adapter avec pertinence à des données et des questions très diverses.

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 (Levenshtein 1966), Hamming (Hamming 1950), ou encore *edit distance* (Sankoff et Kruskal 1983), les Méthodes d'Appariement Optimal (M.A.O), traduction que nous avons proposée pour *Optimal Matching Analysis* (Lesnard et Saint Pol 2006), sont plus connues en biologie où elles ont contribué au séquençage du génome¹. 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, (Abbott et Forrest 1986; Abbott et Hrycak 1990) 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 (Abbott 1995; Abbott et Tsay 2000).

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

¹ Ce texte a été repris et adapté de l'article d'introduction aux Méthodes d'Appariement Optimal (Lesnard et Saint Pol 2006) : <http://bms.revues.org/index638.html>

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 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).

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 1^{re} position dans la séquence X et de transformer le D alors en 3^e 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 1^{re} 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.

		y_1	y_2	y_3	y_4	...							y_n
	0												
x_1													
x_2													
x_3			...										
x_4													
...													
x_m													Fin

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 $X = (x_1, \dots, x_m)$ et $Y = (y_1, \dots, y_n)$, 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. À 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.

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.

		y_1	y_2	y_3	y_4	...					y_n
	0										
x_1											
x_2											
x_3											
x_4											
...											
x_m											

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

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

² Ce graphique et le suivant sont inspirés de Chan (Chan 2002).

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'accrochent mal *a priori* de la mesure d'agrégation de CAH euclidienne (la méthode de Ward) et ce d'autant plus que certaines configurations de coûts de substitution peuvent produire des dissimilarités qui ne respectent pas l'inégalité triangulaire³. 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 (Belbin, Faith et Milligan 1992; Milligan 1980; Milligan 1981). 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 10 de Stata.

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 *composables* 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

³ Par conséquent, la mesure de dissimilarité induite par les méthodes d'appariement optimal n'est pas toujours une distance au sens mathématique strict.

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 au-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'autre⁴. Au final, les M.A.O. sont donc une combinaison d'accélération, de ralentissements et d'écoulements normaux⁵ 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.

Tableau 3 – Distances de Hamming et de Levenshtein

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

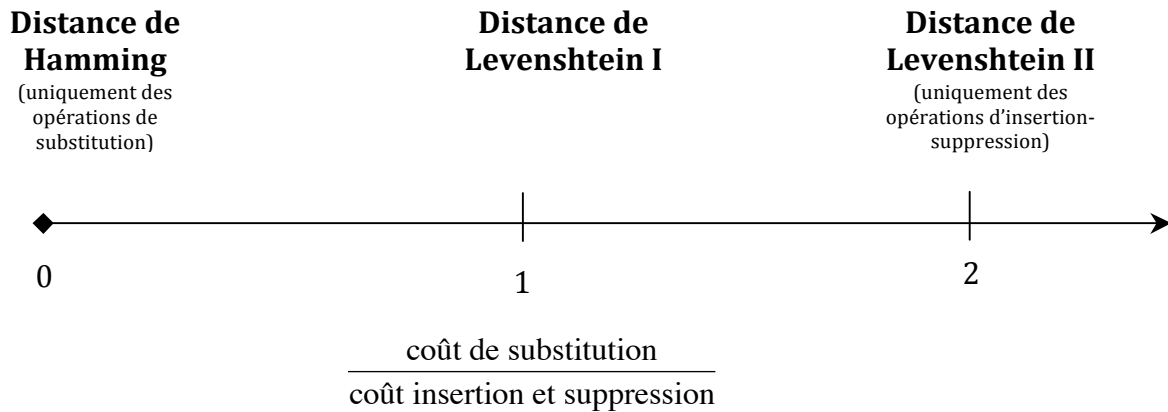
Du choix des coûts associés aux trois opérations des M.A.O. dépendent en effet l'équilibre entre les insertions-suppressions et les substitutions mais également le degré de simplification que ces opérations induisent. C'est pourquoi nous avons choisi de parler « des » Méthodes d'Appariement Optimal, 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 communes⁶.

⁴ 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..

⁵ Par « écoulement normal du temps » il faut entendre « conformément au rythme de l'échelle de temps des séquences ».

⁶ 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.

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 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 substitution⁷.

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 séquences⁸, 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 élevés⁹ 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

⁷ Lorsque les séquences comparées sont de même longueur, l'utilisation des seules opérations de substitution revient à appliquer la distance de Hamming.

⁸ Ce n'est donc pas un simple dénombrement des éléments communs de chaque séquence.

⁹ Voire de réduire l'analyse aux seules opérations de substitution, solution qui est présentée dans le second exemple présenté plus bas.

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 séquences¹⁰.

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 (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 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 enjeux sociologiques de la statistique : elles permettent de véritablement réfléchir et de choisir ce qui convient théoriquement le mieux comme les deux exemples vont le montrer.

Deux applications sur les mêmes données

Les deux applications qui seront présentées dans ce texte 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.

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 M.A.O. 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.

La mise en œuvre informatique

C'est dans le cadre de la bio-informatique que la plupart des logiciels d'appariement de séquences ont été développés et continuent de l'être. La vigueur de la recherche appliquée dans cette discipline est telle qu'il est impossible de proposer un inventaire exhaustif de logiciels pour la plupart très spécialisés. Citons un des programmes les plus anciens *Clustal*, qui semble avoir suivi le rythme vertigineux des améliorations

¹⁰ 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.

de ce champ, et un plus récent, *EMBOSS* (European Molecular Biology Open Software Suite). Promoteur de ces méthodes en sciences sociales, Andrew Abbott a supervisé le développement d'un logiciel, *Optimize*, qui n'a cependant plus évolué depuis 1997. Un ensemble de programmes est disponible pour le logiciel *Stata* (SQ) et *R* (TraMineR). Le logiciel de statistique libre *TDA* proposé par Goetz Rohwer et Ulrich Pötter de l'université de Bochum comporte un module d'analyse de séquence pourvu de quelques fonctionnalités d'appariement optimal. C'est ce logiciel qui a été utilisé pour le premier exemple, complété par le logiciel SAS pour les manipulations des données et la classification des distances issues de TDA. Le second exemple a été entièrement mené à bien sous SAS, les simplifications supplémentaires facilitant encore davantage la mise en œuvre informatique de l'appariement.

L'étude du repas dans le cadre de la soirée

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.

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.

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 (Saint Pol 2006). Les activités présentes dans les carnets journaliers de l'enquête Emploi du Temps 1998 ont été regroupées en 25 catégories¹¹ 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. À 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.

Mais les M.A.O. 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

¹¹ Ces catégories sont celles proposées par Alain Chenu et Nicolas Herpin (Chenu et Herpin 2002).

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 faible¹². 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.

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 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.

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 4¹³. Mis à part les individus de la première classe qui consacrent 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 M.A.O. Ainsi, au sein de ce qui n'était auparavant qu'un groupe assez difforme d'individus aux temps moyens alimentaires similaires, il est désormais possible de distinguer différents types de pratiques.

¹² En termes mathématiques, ces coûts sont les inverses des probabilités de transition entre deux activités sur l'ensemble des séquences de notre échantillon.

¹³ Pour plus de précision, voir Sankoff et Kruskal (1983).

Tableau 4 – Descriptif des dix classes.

Classe	Effectif (en %)	Temps moyen consacré au de à	REPAS 18h50 21h30	Nom de la classe	Séquence-type ¹⁴				
					19h00- 19h30	19h30- 20h00	20h00- 20h30	20h30- 21h00	21h00- 21h30
1	6,0	113 min		<i>Les Mangeurs</i>	Repas	Repas	Repas	Repas	Repas
2	3,7	43 min		<i>Les Couche-tôt</i>	Repas	Repas	Télévision	Sommeil	Sommeil
3	20,8	37 min		<i>Soirée télé</i>	Repas	Repas	Télévision	Télévision	Télévision
4	12,0	42 min		<i>Les Dîne-tôt</i>	Repas	Repas	Télévision	Télévision	Télévision
5	5,8	48 min		<i>Cuisine et Ménage</i>	Repas	Repas	Ménager	Enfants	Télévision
6	24,5	36 min		<i>Les Dîne-tard</i>	Bricolage	Repas	Repas	Repas	Télévision
7	7,5	40 min		<i>Deuxième journée</i>	Ménager	Ménager	Repas	Repas	Télévision
8	5,4	43 min		<i>Les Téléphages</i>	Télévision	Télévision	Repas	Repas	Télévision
9	9,4	26 min		<i>Les Travailleurs</i>	Travail	Travail	Travail	Travail	Travail
10	4,9	29 min		<i>Sorties</i>	Rencontres	Rencontres	Rencontres	Rencontres	Rencontres

Source : Enquêtes Emploi du Temps de l'Insee de 1998.

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.

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 suspecte¹⁵. L'utilisation de M.A.O. 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 M.A.O.. 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 objets de recherche.

¹⁴ La séquence-type reproduite ici est celle qui se situe au centre de la classe : elle minimise la distance à toutes les autres séquences du groupe. Dans une visée illustrative, on a enlevé le premier élément qui correspond à la plage 18h50-19h00 et qui est identique à l'élément suivant pour toutes les séquences, mis à part pour la 3^e et la 4^e où il s'agit de travaux ménagers, et on a regroupé les éléments trois par trois.

¹⁵ Il faut donc interpréter avec prudence les résultats de G. Larmet qui conclut, uniquement en termes de durée, à l'accroissement de la sociabilité alimentaire entre 1985 et 1998 (Larmet 2002).

Le temps du travail

L'analyse scientifique du temps de travail est généralement réduite dans les enquêtes Emploi du temps à de simples durées¹⁶, 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*¹⁷ (Schor 1991) 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 (Chenu 2002; Gershuny 2000) 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.

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.

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.

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*¹⁸.

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 (Durkheim 1912; Lesnard 2004; Lesnard 2006c), le temps est un système symbolique qui, parce qu'il cristallise le rythme de l'activité collective, permet d'anticiper les régularités 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.

¹⁶ 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.

¹⁷ Terme anglo-saxon qui désigne les travailleurs compulsifs.

¹⁸ En effet, sans opérations d'insertion-suppression, un seul chemin est possible : celui situé sur la diagonale de la matrice d'appariement.

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 état¹⁹. 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 dynamique²⁰. 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.

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 5²¹).

Tableau 5 – 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 durée moyenne est de huit heures et quart.

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

¹⁹ En termes statistiques, ces flux sont mesurés par les matrices de transition entre les différents états. Pour plus de détails, voir (Lesnard 2004; Lesnard 2006c).

²⁰ Cette version des M.A.O. 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; Lesnard 2006c). Cette méthode est disponible sur Internet sous la forme d'une extension Stata (voir <http://laurent.lesnard.free.fr>).

²¹ Seuls les résultats pour 1985-86 sont présentés ici. Pour plus de détails, voir (Lesnard 2006a; Lesnard 2006b).

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 ».

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 nuit²² 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 atypiques²³. 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.

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 Pendaries 1994; Prunier-Poulmaire 2000).

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.

²² 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.

²³ 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.

Conclusion

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.

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.

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. 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 quand il souligne le caractère intégrateur de cette approche qui tient à la fois de la démarche analytique et de la démarche narrative (Fabiani 2003). Les méthodes d'appariement optimal constituent un important pour l'analyse de données séquentielles et méritent d'être intégrées à la boîte à outil des chercheurs en sciences sociales pour être mobilisées quand les besoins théoriques l'exigent.

Bibliographie

- Abbott, Andrew. 1995. "Sequence analysis: new methods for old ideas." *Annual Review of Sociology* 21:93-113.
- Abbott, Andrew et John Forrest. 1986. "Optimal matching methods for historical sequences." *Journal of Interdisciplinary History* 16:471-494.
- Abbott, Andrew et Alexandra Hrycak. 1990. "Measuring resemblance in sequence analysis: an optimal matching analysis of musicians careers." *American Journal of Sociology* 96:144-185.
- Abbott, Andrew et Angela Tsay. 2000. "Sequence analysis and optimal matching methods in sociology." *Sociological Methods and Research* 29:3-33.
- Belbin, Lee, Dan Faith et Glenn W. Milligan. 1992. "A Comparison of Two Approaches to Beta-Flexible Clustering." *Multivariate Behavioral Research* 27:417-433.
- Bouffartigue, Paul et Jean-René Pendaries. 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* 36:337-359.
- Chan, Tak Wing. 2002. "Optimal Matching Analysis." *Social Research Update* 24.
- Chenu, Alain. 2002. "Les horaires et l'organisation du temps de travail." *Économie et Statistique* 352-353:151-167.
- Chenu, Alain et Nicolas Herpin. 2002. "Une pause dans la marche vers la civilisation des loisirs ?" *Économie et Statistique*:15-37.
- Durkheim, Émile. 1912. *Les formes élémentaires de la vie religieuse*. Paris: Alcan.

- Fabiani, Jean-Louis. 2003. "Pour en finir avec la réalité unilinéaire. Le parcours méthodologique de Andrew Abbott." *Annales HSS* 3:549-565.
- Gershuny, Jonathan. 2000. *Changing Times: Work and Leisure in Postindustrial Society*. Oxford: Oxford University Press.
- Godard, Francis. 2003. "Les temps du quotidien." Pp. 15-22 in *Le(s) public(s) de la culture*, edited by O. Donnat et P. Tolila. Paris: Presses de Sciences Po.
- Halpin, Brandan et Tak Wing Chan. 1998. "Class careers as sequences: an optimal matching analysis of work-life histories." *European Sociological Review* 14:111-130.
- Hamming, Richard W. 1950. "Error-detecting and error-correcting codes." *Bell System Technical Journal* 29:147-160.
- Larmet, Gwenaël. 2002. "La sociabilité alimentaire s'accroît." *Économie et Statistique* 352-353:191-211.
- Lesnard, Laurent. 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 of Time Use Research* 1:63-88.
- . 2006a. "Flexibilité des horaires de travail et inégalités sociales." Pp. 371-378 in *Données Sociales - La société française*, edited by Insee. Paris: Insee.
- . 2006b. "Flexibilité et concordance des horaires de travail dans le couple." Pp. 379-384 in *Données Sociales - La société française*, edited by Insee. Paris: Insee.
- . 2006c. "Optimal matching and the social sciences." in *Document de travail du CREST*. Paris: Centre de Recherche en Économie et Statistique - Insee.
- Lesnard, Laurent et Thibaut de Saint Pol. 2006. "Introduction aux méthodes d'appariement optimal (optimal matching analysis)." *Bulletin de Méthodologie Sociologique* 90:5-25.
- Levenshtein, Vladimir I. 1966. "Binary codes capable of correcting deletions, insertions, and reversals." *Soviet Physics Doklady* 10:707-710.
- Levine, Joel H. 2000. "But what have you done for us lately?: Commentary on Abbot and Tsay." *Sociological Methods and Research* 29:34-40.
- Milligan, Glenn W. 1980. "An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms." *Psychometrika* 45:325-342.
- . 1981. "A Monte Carlo Study of Thirty Internal Criterion Measures for Cluster Analysis." *Psychometrika* 46:187-199.
- Prunier-Poulmaire, Sophie. 2000. "Flexibilité assistée par ordinateur. Les caissières d'hypermarché." *Actes de la recherche en sciences sociales* 134:29-65.
- Robinson, John P. et Geoffrey Godbey. 1999. *Time For Life. The Surprising Ways Americans Use Their Time*. University Park: Pennsylvania State University Press.
- Saint Pol, Thibaut de. 2006. "Le dîner des Français : un synchronisme alimentaire qui se maintient." *Économie et Statistique* 400:45-69.
- Sankoff, David et Joseph B. Kruskal (dir.). 1983. *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison*. Reading, MA: Addison-Wesley.
- Schor, Juliet. 1991. *The Overworked American: the unexpected decline of leisure*. New York: Basic Books.