Journées de Méthodologie Statistique – 24 mars 2009

Décrire des données séquentielles en sciences sociales : panorama des méthodes existantes

Laurent Lesnard

Sciences Po, Centre de données socio-politiques
Crest, Laboratoire de sociologie quantitative

Thibaut de Saint Pol

Insee, Conditions de vie des ménages
Crest, Laboratoire de sociologie quantitative

Introduction

- Données séquentielles :
  - trajectoires d'insertion sur le marché du travail
  - carrières professionnelles
  - emplois du temps
- Développement récent de nouvelles techniques, parallèlement aux possibilités informatiques
- Quels outils pour quelles données et quelles questions ?

Plan de la présentation

1. Les méthodes d'appariement optimal
   1. Principe de la comparaison des séquences
   2. Regrouper les séquences voisines
   3. La question des coûts
2. Analyse factorielle et données séquentielles
   1. L'analyse harmonique qualitative
   2. L'analyse factorielle de séquences

1. Les Méthodes d'appariement optimal

- en anglais *Optimal Matching Analysis* (OMA)
- Issues de recherche en informatique dans les années 1950 et 1960 où elles sont connues sous le nom de distance de Hamming et de Levenshtein
- Biologie : séquençage du génome
- Sciences sociales : travaux de Andrew Abbott
- Objectif des Méthodes d'appariement optimal : comparer et regrouper les séquences

1re étape : la minimisation

- Se donner une distance entre séquences : nombre minimal d'opérations nécessaires pour rendre identiques deux séquences
- Trois opérations sont possibles : insertion, suppression et substitution
- Chaque opération a un coût
- Le coût total minimal pour rendre identique deux séquences fournit une mesure de leur « distance »

Exemple : les engagements successifs de deux militants A et B dans les associations X et Y

- Deux séquences à comparer :

  A : X – Y – Y – Y

  B : X – X – X – X – Y

Exemple : les engagements successifs de deux militants A et B dans les associations X et Y

- Deux séquences à comparer :

  A : X – Y – Y – Y

  B : X – X – X – X – Y

- Une transformation possible de A en B :

  A : X – X – X – X – Y – ~~Y~~ – ~~Y~~

  B : X – X – X – X – Y

Exemple : les engagements successifs de deux militants A et B dans les associations X et Y

- Deux séquences à comparer :

  A : X – Y – Y – Y

  B : X – X – X – X – Y

- Une transformation possible de A en B :

  A : X – X – X – X – Y – ~~Y~~ – ~~Y~~

  B : X – X – X – X – Y

- Autre possibilité :

  A : X – X – Y – Y – Y

  B : X – X – X – X – Y

Exemple : les engagements successifs de deux militants A et B dans les associations X et Y

- Deux séquences à comparer :

  A : X – Y – Y – Y

  B : X – X – X – X – Y

- Une transformation possible de A en B :

  A : X – X – X – X – Y – ~~Y~~ – ~~Y~~

  B : X – X – X – X – Y

- Autre possibilité :

  A : X – X – X – X – Y

  B : X – X – X – X – Y

Exemple : les engagements successifs de deux militants A et B dans les associations X et Y

- Deux séquences à comparer :

  A : X – Y – Y – Y

  B : X – X – X – X – Y

- Une transformation possible de A en B :

  A : X – X – X – X – Y – ~~Y~~ – ~~Y~~

  B : X – X – X – X – Y

  3 insertions
  2 suppressions

- Autre possibilité :

  A : X – X – X – X – Y

  B : X – X – X – X – Y

  1 insertion
  2 substitutions

Exemple : les engagements successifs de deux militants A et B dans les associations X et Y

- Deux séquences à comparer :

  A : X – Y – Y – Y

  B : X – X – X – X – Y

Coûts « classiques »

Insertion et suppression=1

Substitution=2

- Une transformation possible de A en B :

  A : X – X – X – X – Y ~~– Y – Y~~

  B : X – X – X – X – Y

3 insertions
2 suppressions
Coût total=7

- Autre possibilité :

  A : X – X – X – X – Y

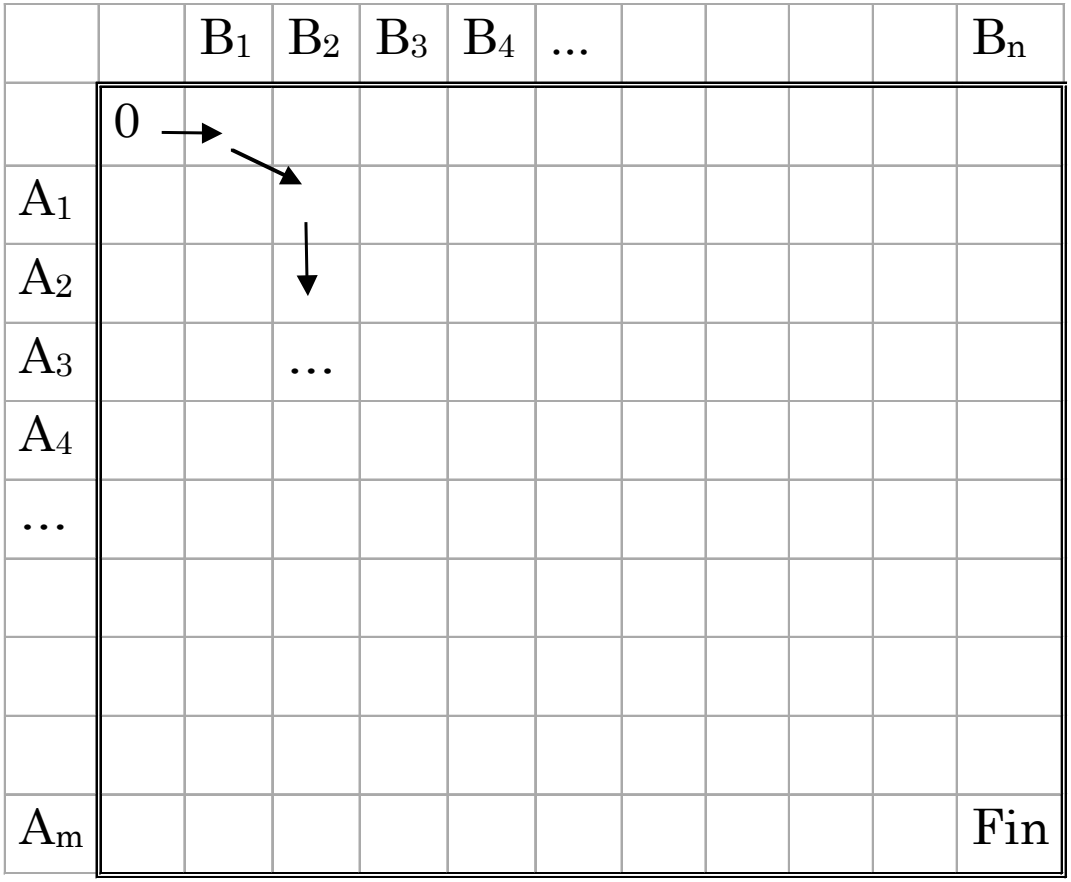
  B : X – X – X – X – Y

1 insertion
2 substitutions
Coût total=5

La distance

- Le coût du passage de A en B est donc :
  - cas 1 : coût de l'insertion de 3 X et la suppression de 2 Y
  - cas 2 : coût de l'insertion d'1 X et de la transformation de 2 Y en X
- Considérer toutes les manières de passer d'une séquence à une autre
- La distance entre les deux séquences sera le coût minimal pour passer de l'une à l'autre au moyen des 3 opérations

Représentation sous forme matricielle :



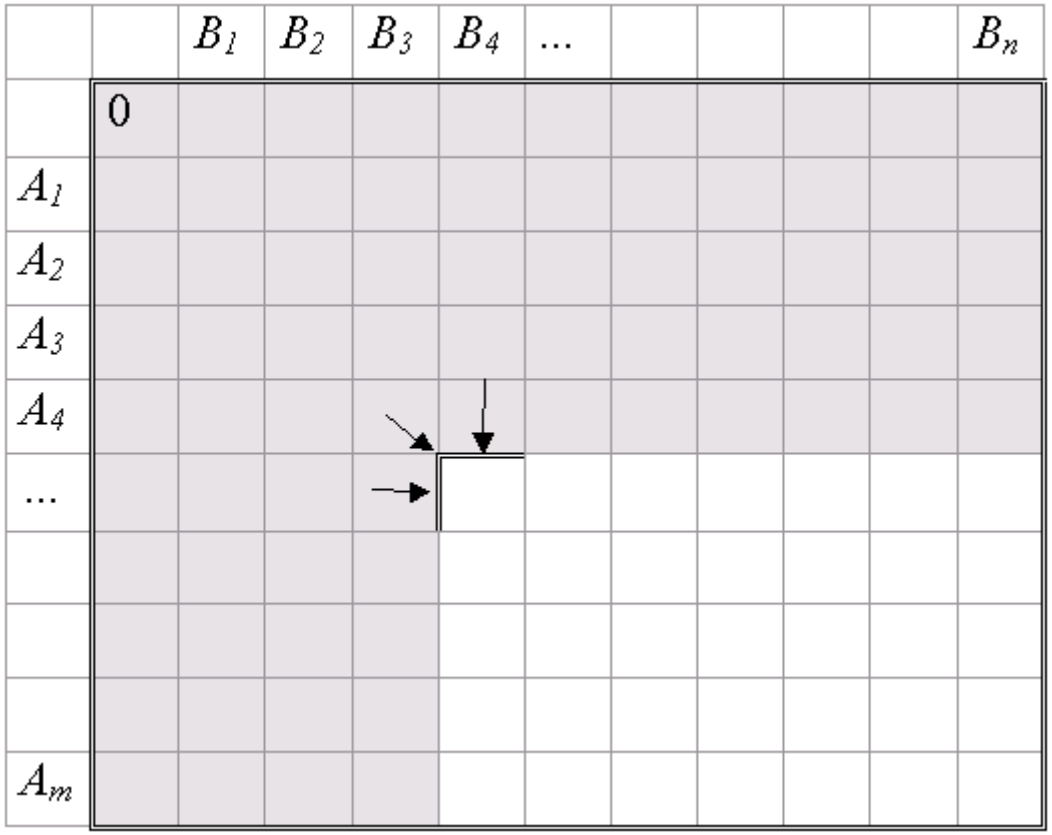
<u>Ici à titre d'exemple :</u>

Insertion de B1

Transformation de A1 en B2

Suppression de A2

Représentation matricielle du processus de minimisation



- Seulement 3 façons d'arriver sur une case
- 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

2e étape : la classification

- Passage d'une distance entre séquences à une distance entre groupes (plusieurs méthodes possibles)
- Choix d'une méthode parmi toutes celles qui existent
- Construction de groupes de séquences

Intérêt de la méthode

- Bâtir des groupes à partir de milliers de séquences
- Compare les séquences sans avantager aucun élément
- Prend en compte toutes les dimensions de la séquence (verticale et horizontale)
- Rendre compte des choix théoriques au travers des coûts des opérations

En pratique

- SAS
  - TDA (freeware) : il existe un module M.A.O.
  - Programmation dans le logiciel SAS (module de calcul matriciel + langage macro)
  - Classification Ascendante Hiérarchique (ici méthode Beta-Flexible ; méthode de Ward non recommandée)
  - Macro SAS : Dynamic Hamming matching http://laurent.lesnard.free.fr
- Stata : module sq
- R : librairie TraMineR

Deux applications aux emplois du temps

1. L'insertion du repas dans la soirée
   - Plusieurs états : les activités de la soirée
   - Identification de séquences d'activités typiques
   - Utilisation des trois opérations : indel=1 et substitution=f(matrice de transition)

2. Le temps du travail
   - Deux états : travail et non-travail
   - Seule la position du travail dans la journée nous importe
   - Utilisation de la seule opération de substitution

Contribution associée à la session 9 :

L. LESNARD, Th. de SAINT POL, Décrire des données séquentielles en sciences sociales : mise en pratique des Méthodes d'Appariement Optimal.

1.2 La détermination des coûts

- Jouer sur les coûts pour adapter la méthode à l'objet traité
- Opérations d'insertion-suppression : déformer le temps pour rapprocher les événements identiques
- Opérations de substitution : distorsion des événements pour mieux comparer leur dimension temporelle
- Déterminer les coûts, c'est donc déterminer les modalités de la comparaison des séquences

Evénements et temps

- Séquence = événements + temps
- Comparer des séquences, c'est simplifier l'une ou l'autre dimension
- Lien entre ces deux dimensions et les opérations des M.A.O. :

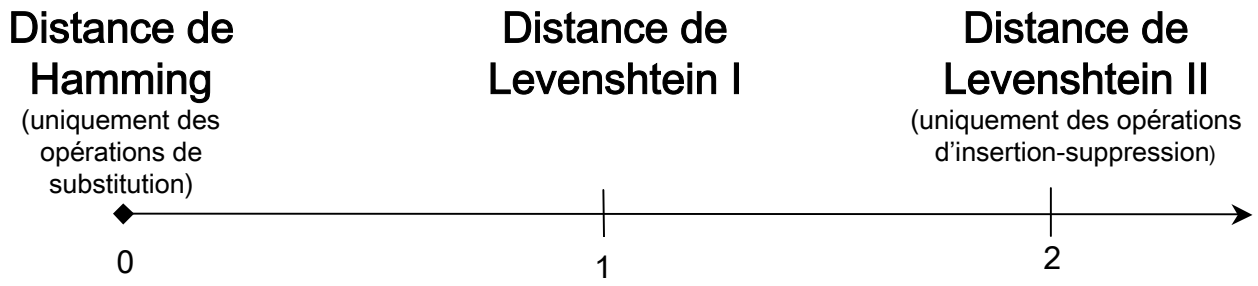
	Insertion-Suppression	Substitution
Ce qui est préservé	Événements	Temps
Ce qui est simplifié	Temps	Événements

Coûts et type de régularité statistique privilégié

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

Coûts et type de régularité statistique privilégié

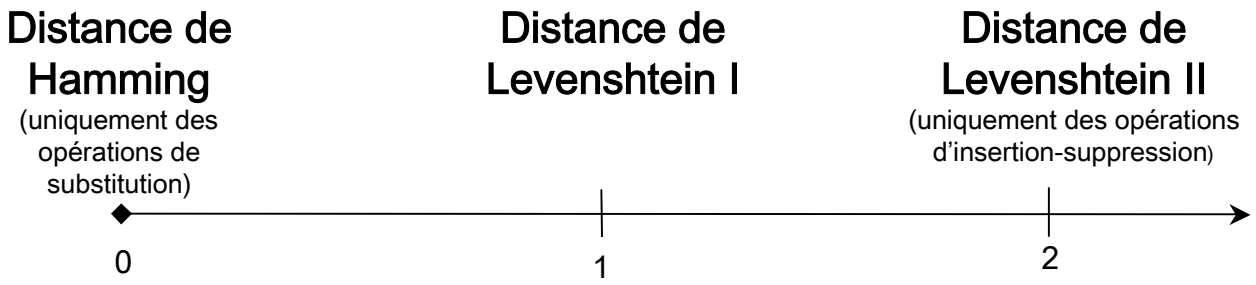
	Opérations utilisées	
	Substitution	Insertion et suppression
Hamming	Oui (coût = 1)	Non
Levenshtein I	Oui (coût = 1)	Oui (coût = 1)
Levenshtein II	Non	Oui (coût = 1)



$$\frac{\text{Coût d'insertion et suppression}}{\text{Coût de substitution}}$$

Coûts et type de régularité statistique privilégié

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



Nombre d'événements contemporains identiques

$$\frac{\text{Coût d'insertion et suppression}}{\text{Coût de substitution}}$$

Sous-séquence identique de longueur maximale

2. Analyse factorielle et données séquentielles

- Deux approches :
  - Analyse harmonique qualitative (AHQ)
  - Analyse des correspondances multiples
- Principes communs :
  - Recherche de la structure des séquences (analyse factorielle)
  - Ré-injection de cette structure au niveau individuel (classification sur les facteurs de l'analyse factorielle)
  - La première phase est essentielle du point de vue de la mesure de similarité entre séquences

1.1 L'analyse harmonique qualitative (AHQ)

- Travaux de Jean-Claude Deville dans les années 1970
- En pratique :
  1. Découpage de la période d'observation en sous-périodes
  2. Calcul des temps de séjour dans tous les états
  3. Analyse factorielle du tableau individus-temps de séjour (en %)
  4. Classification ascendante hiérarchique sur les premiers facteurs

Matrice harmonique

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

Nombre et bornes des sous-périodes

- Nombre de sous-périodes
  - Trop -> risque de durées nulles
  - Pas assez -> disparition des variations temporelles
- Bornes des sous-périodes
  - Intervalles peuvent être de longueurs différentes
  - Distribution des changements d'états
  - Choix guidé par la théorie : par exemple, intervalles courts et nombreux pour la période jugée la plus importante pour l'analyse

Analyse harmonique qualitative et temps

- Dans chaque sous-période
  - Séquences réduites à des temps de passage dans les différents états
  - Disparition du timing
  - Parcimonie
- Analyse factorielle de la matrice harmonique
  - Ordre des sous-périodes non pris en compte
  - Avantage : prise en compte des liaisons indépendamment de leur éloignement dans les séquences
  - Inconvénient : dimension ordonnée du temps ignorée

1.2 L'analyse factorielle de séquences (AFS)

- ACM sur les variables qui décrivent les trajectoires
  1. Séquences mises sous forme d'un tableau disjonctif complet
  2. AFC
  3. Classification ascendante hiérarchique sur les premiers facteurs

- Différences avec l'AHQ
  - Pas de regroupement en sous-périodes
  - Conséquence directe : durée non prise en compte

Comment est évaluée la proximité entre séquences ?

- Équivalence avec l'analyse factorielle du tableau des profils de Burt
- Tableau des profils de Burt proportionnel aux probabilités de transitions pour toutes les combinaisons d'états et de dates
- Deux états seront d'autant plus éloignés que les transitions vers les autres états divergeront
- Et ce d'autant plus que ces dates et états sont peu fréquentés

Analyse factorielle de séquences et temps

- Comme pour l'Analyse Harmonique Qualitative
  - Ordre des événements non pris en compte
  - Avantage : prise en compte des liaisons indépendamment de leur éloignement dans les séquences
  - Inconvénient : dimension ordonnée du temps ignorée
- Contrairement à l'AHQ
  - Le temps n'est à aucun moment pris en compte dans l'analyse
  - Il est réintroduit au moment de l'interprétation de la classification
  - Peu parcimonieux mais préserve la totalité des événements

Conclusion (1)

- Deux familles de méthodes qui s'inscrivent dans des paradigmes très différents
  - Algorithmique pour les méthodes d'appariement optimal (MAO)
  - Géométrique pour l'analyse factorielle (AHQ et AFS)

- Différences de mesure de similarité
  - Distance fondée sur la structure, appréhendée au travers des probabilités de transition pour l'AFS
  - Distance peut dépendre d'agrégats macro (transitions moyennes ou toutes les transitions entre périodes consécutives pour le Dynamic Hamming Matching), de coûts déterminés théoriquement, de coûts neutres (Levenshtein I), etc.

Conclusion (2)

- Statut du temps
  - MAO : jeu sur les coûts pour privilégier les ressemblances locales ou la recherche de sous-séquences identiques
  - AHQ : prise en compte de la durée, au prix d'une simplification des séquences
  - AFS : aucun statut particulier

- En pratique pour choisir une méthode pour construire une typologies de séquences
  - Type de régularité temporelle recherché
  - Statut du temps
  - Interprétation et désirabilité des éventuelles relations entre des événements très éloignés dans les séquences ?