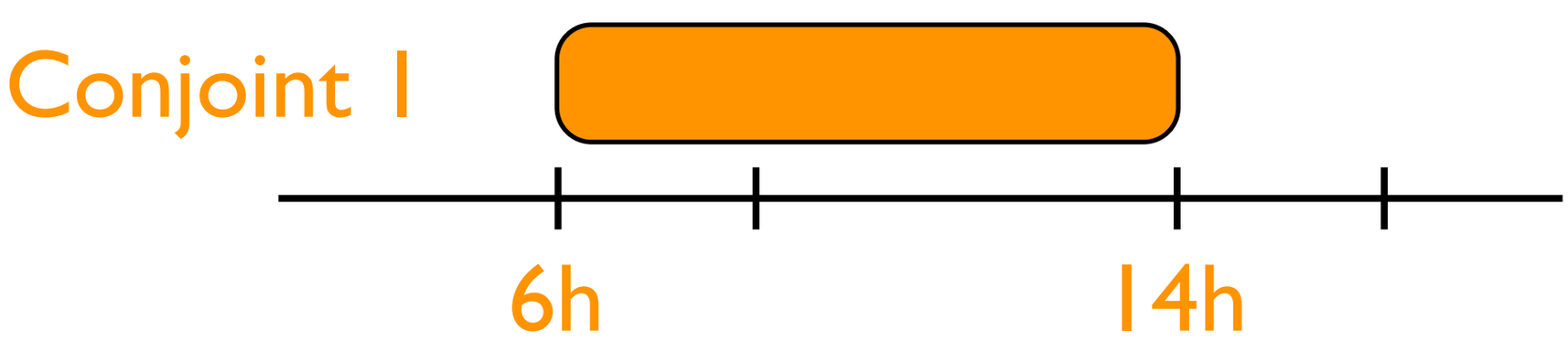
Méthodes d'appariement optimal et sciences sociales

Laurent Lesnard
Centre de données socio-politiques et
Observatoire sociologique du changement (CNRS - Sciences Po)
Laboratoire de sociologie quantitative du CREST (Insee)


Conjoint 1
6h
14h

Conjoint 2

Conjoint 1

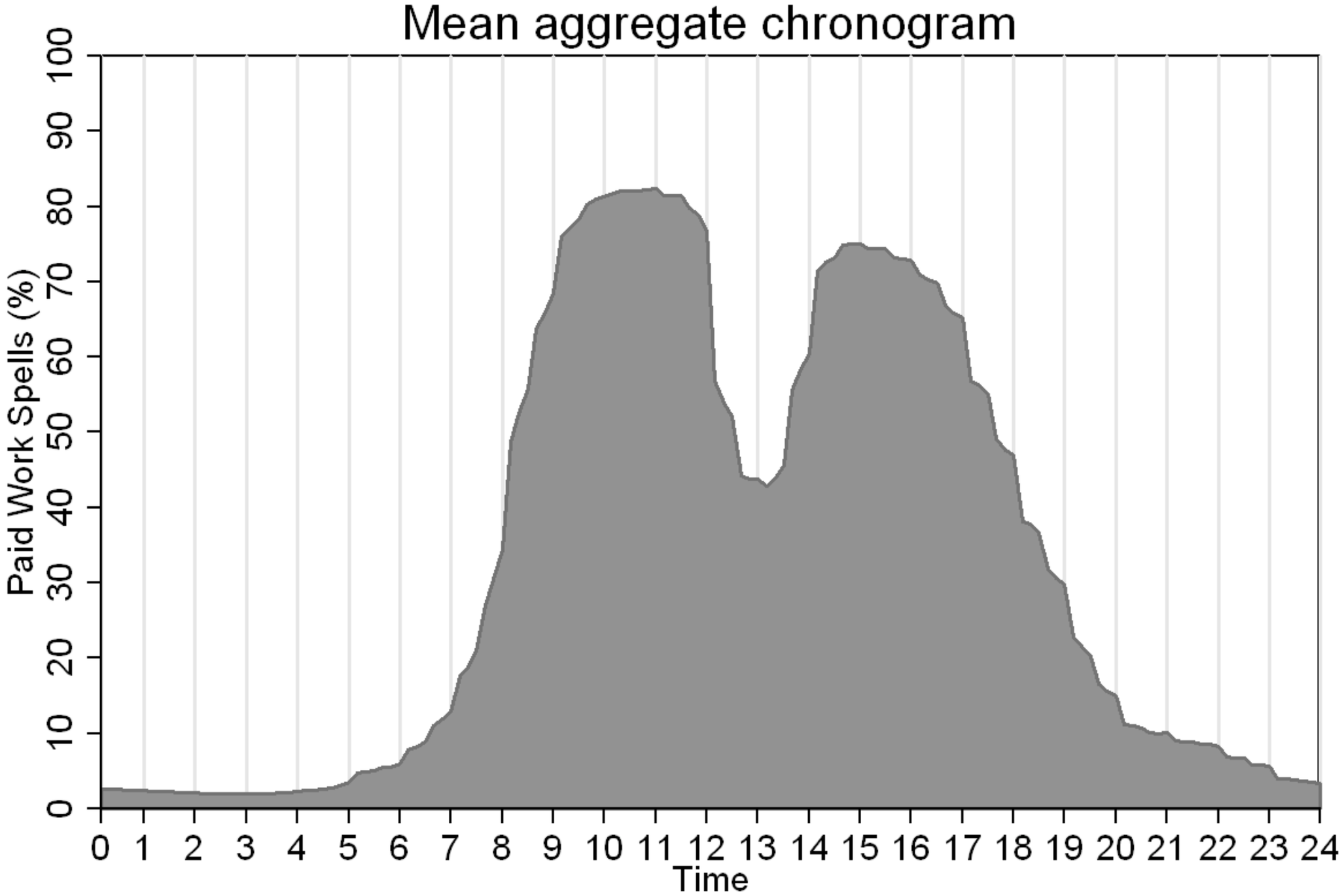
6h 9h 14h 17h

Désynchronisation

Conjoint 2

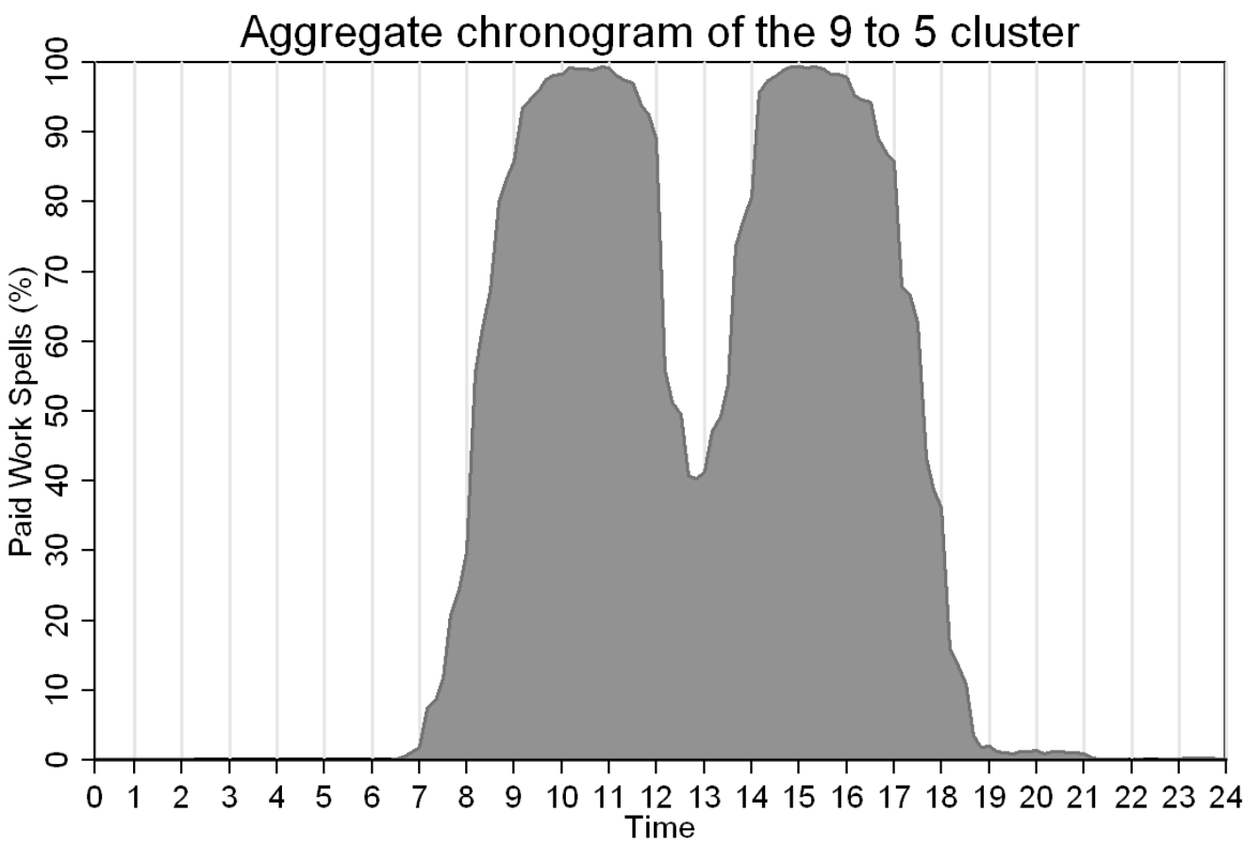
Conjoint 1

6h 9h 14h 17h

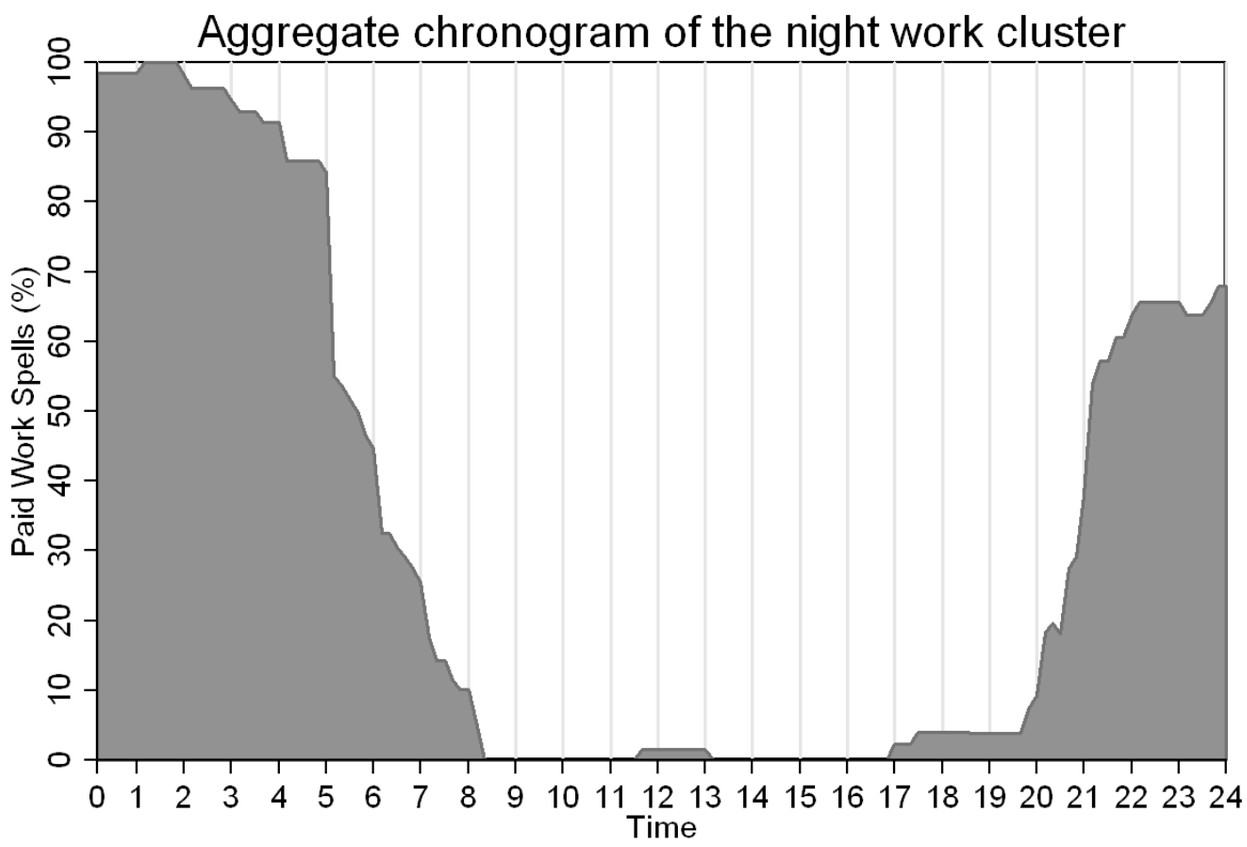


Journée de travail moyenne en France en 1999

(Insee, enquête emploi du temps 1999)



Journée de travail standard (33,9%)



Travail de nuit (1,6%)

Application des méthodes d'appariement optimal

- Principales applications : analyse de carrières
  - Abbott et Hrycak (1990) : analyse des carrières des musiciens
  - Halpin et Chan (1998) : analyse de la mobilité intra-générationnelle
  - Blair-Loy (1999) : analyse des carrières des femmes cadres dans la finance
  - Han et Moen (1999) : analyse des fins de carrières (retraite)
- Autres :
  - Analyse historique : Abbott et Forrest (1986)
  - Transports : Wilson (1998)
  - Analyse de la structure rhétorique des articles en sociologie : Abbott et Barman (1997)
  - Emplois du temps : Lesnard (2006a, 2006b, 2004, à paraître), Saint Pol (2006) pour analyser les différents modes d'inscriptions du dîner dans la soirée

Plan de la présentation

1. Une brève présentation des méthodes d'appariement optimal (MAO)
2. L'utilisation des MAO en biologie
3. Comment utiliser les MAO en sciences sociales ?
4. Exemples d'utilisation des MAO
5. En pratique

1. Une brève présentation des MAO

- 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

- *Optimal Matching Analysis* ou Méthodes d'Appariement

Principe des MAO

- Objectif : comparer et regrouper les séquences

- Deux étapes : construction de la distance et classification

- La spécificité de ces méthodes tient à la première étape

L'algorithme

- Utilisation de trois opérations élémentaires :
  - Insertion
  - suppression
  - Substitution

- Considérer tous les manières de passer d'une séquence à l'autre au moyen de ces opérations

A : X – Y – Y – Y

B : X – X – X – X – Y

L'algorithme

- Utilisation de trois opérations élémentaires :
  - Insertion
  - suppression
  - Substitution

- Considérer tous les manières de passer d'une séquence à l'autre au moyen de ces opérations

A : X – X – X – X – Y – Y – Y

B : X – X – X – X – Y

L'algorithme

- Utilisation de trois opérations élémentaires :
  - Insertion
  - suppression
  - Substitution

- Considérer tous les manières de passer d'une séquence à l'autre au moyen de ces opérations

A : X – X – X – X – Y – Y – Y

B : X – X – X – X – Y

L'algorithme

- Utilisation de trois opérations élémentaires :
  - Insertion
  - suppression
  - Substitution

- Considérer tous les manières de passer d'une séquence à l'autre au moyen de ces opérations

A : X – Y – Y – Y

B : X – X – X – X – Y

L'algorithme

- Utilisation de trois opérations élémentaires :
  - Insertion
  - suppression
  - Substitution

- Considérer tous les manières de passer d'une séquence à l'autre au moyen de ces opérations

A : X – X – Y – Y – Y

B : X – X – X – X – Y

L'algorithme

- Utilisation de trois opérations élémentaires :
  - Insertion
  - suppression
  - Substitution

- Considérer tous les manières de passer d'une séquence à l'autre au moyen de ces opérations

A : X – X – X – X – Y

B : X – X – X – X – Y

- 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

Représentation sous forme matricielle

		B_1	B_2	B_3	B_4	...					B_n
	0 →										
A_1											
A_2											
A_3			...								
A_4											
...											
A_m											Fin

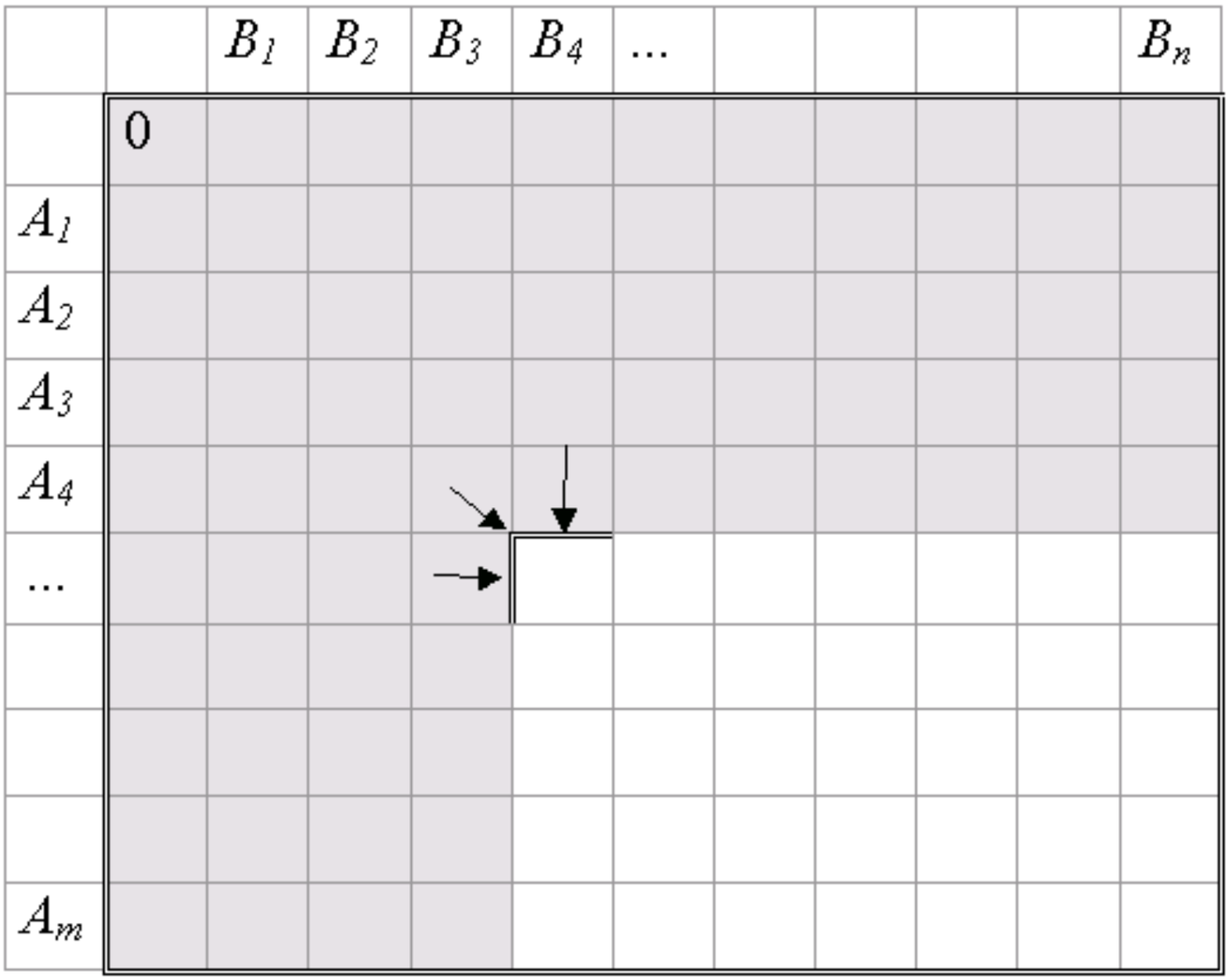
Ici à titre d'exemple :

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.

Seconde étape

- Phase de classification : passage d'une distance entre individus à une distance entre groupes

- Choix d'une méthode parmi toutes les méthodes de classification qui existent

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

Article	Coûts
Abbott et Forrest 1986 Optimal matching methods for historical sequences	Substitutions = nombre de différences dans la hiérarchie des pas de danse divisé par le nombre de pas total (5) Indel = 1 (max subs)
Stovel et al. 1996 “Ascription into achievement: models of career systems at Lloyds bank”	Substitutions = somme de deux matrices de transition entre les professions et les filiales inspirées par l’examen des deux matrices de transition entre 1890 et 1970 Indel = 6.5 (max subs) Dissimilarités standardisées par la longeur de la séquence
Halpin et Chan 1998	Substitutions déterminées théoriquement (varient entre 1 et 4) Indel = 3
Abbott et Hrycak 1999 Measuring Resemblance in Sequence Data: An Optimal Matching Analysis of Musicians' Careers	Substitutions dérivéEs de la matrice de transition entre tous les états pour l’ensemble des dates (0,47 - 1) Indel (1,088)
Blair-Loy 1999 “Career patterns of executive women in finance”	Coûts de substitution déterminés théoriquement (varient entre 0 et 1) Coûts d’insertion-suppression fixés à 0.48 après plusieurs essais
Han et Moen 1999 “Clocking Out Temporal Patterning of Retirement”	Non spécifiés. Coûts de substitution semblent dépendre des transitions entre les différents états
Stovel 2001 “Local sequential patterns the structures of lynching in the deep south”	Sij = |i - j|/(j+1)
Clark et al. 2003 “Housing Careers in the United States”	Indel = 1 Subs ϵ {0.1,0.8]
Stovel et Bolan 2004 “Residential trajectories”	Coûts de substitution inspirés par l’examen de la matrice de transition agrégée pour les 14 épisodes Indel = 2.7 lorsque les séquences comparées sont de même longueur et = 0.45 sinon Argument : fixed indel cost is drawing sequences together on the basis of similar lengths rather than similar patterns
Brzinsky-Fay 2007 “Lost in transition”	Indel = 1 Subs = 2
Pollock 2007 “Holistic trajectories: a study of combined employment, housing, and family careers by using multiple-sequence analysis”	Somme de 4 matrices de coût de substitution inspirées par l’observation des transitions pour chaque dimension varient entre 0,5 et 2 Indel = 1

"The assignment of transformation
costs haunts all
optimal matching analyses"

Katherine Stovel, Mike Savage and Peter Bearman, 1996, "Ascription into Achievement: Models of Career Systems at Lloyds Bank, 1890-1970", *American Journal of Sociology*, 102, p. 394

Deux critiques des MAO

1. En biologie, l'insertion, la suppression et la substitution d'éléments ont (auraient) un sens. Quelle est l'interprétation sociologique des MAO ?

    - "The analogy between DNA and careers is not obvious. Sociological structures may evolve, in a specific sense that remains to be defined, but they do not have genes" (Levine 2000)

    - "Replacements and indels [are] closer to what goes on with DNA" (Wu 2000)

    - "In biochemistry [...] making an insert, delete, or substitution of a token or even a subsequence of considerable length is justified by arguments rooted in a theory about the electrochemical, mechanical, or functional properties of what is deleted, inserted, or substituted for" (Elzinga 2003)

Deux critiques des MAO

2. Quels sont les effets de différentes combinaisons de coûts ?

    - “in the worst case — that is, if results are sensitive to alternative choices of costs — then findings obtained using sequence analysis could be driven solely by one’s choice in setting the various cost matrices” (Wu 2000)
    - “If two different setups of the sequence-matching algorithms provide essentially the same result, it means that neither one is very close to the data, or else that the data are not sufficiently detailed [...]. Ultimately, if a model were to prove so robust that it survived equally well under a great range of changes [...], then we would have to say that it is not falsifiable, which means it is not a model.” (Levine 2000)

2. L’utilisation des MAO en biologie

L’utilisation des MAO en biologie

- Objectif : transférer de l’information entre des protéines ou de l’ADN
- Appariement Optimal remplace les expériences qui sont coûteuses et longues

L’utilisation des MAO en biologie

Objectif : transférer de l’information entre des protéines ou de l’ADN

Séquences dont la structure et/ou les fonctions sont **connues** ← **Appariement Optimal** → Séquences dont la structure et/ou les fonctions sont **inconnues**

L'utilisation des MAO en biologie

- Les trois transformations n'ont pas de signification théorique
- Les trois transformations ne reproduisent pas des phénomènes biochimiques
- Les coûts de substitutions
  - Interprétés comme la probabilité qu'ont deux séquences d'être liées phylogénétiquement
  - Estimés empiriquement à partir d'un échantillon de séquences de référence

$$s(a,b) = \frac{p_{ab}}{q_a q_b}$$

Les coûts d'insertion et de suppression en biologie

"In practice, people choose [insertion and deletion] costs empirically once they have chosen their substitution scores."

Durbin *et al.*, 1998, p. 44

Coûts de substitution en biologie

Objectif : transférer de l'information entre des protéines ou de l'ADN

Échantillon aléatoire de séquences phylogénétiquement liées

1re étape

Estimation probabiliste

Matrice de substitution (par ex. les matrices Pam ou Blosum)

2e étape

Séquences dont la structure et/ou les fonctions sont **connues**

Appariement Optimal

Séquences dont la structure et/ou les fonctions sont **inconnues**

Leçons à tirer de la biologie

1. Nature des séquences, objectifs et utilisation des MAO en biologie et en sciences sociales n'ont rien en commun

➡ Il n'est pas possible d'utiliser les mêmes outils

2. Les trois transformations n'ont pas de signification théorique et ne reproduisent pas de phénomènes biochimiques

3. Les coûts sont déterminés empiriquement

4. La clé du succès des MAO en biologie tient aux coûts de substitution, à leur interprétation

3. Comment utiliser les MAO en sciences sociales ?

Différences avec la biologie

- « Matière » différente

  séquences = événements + temps

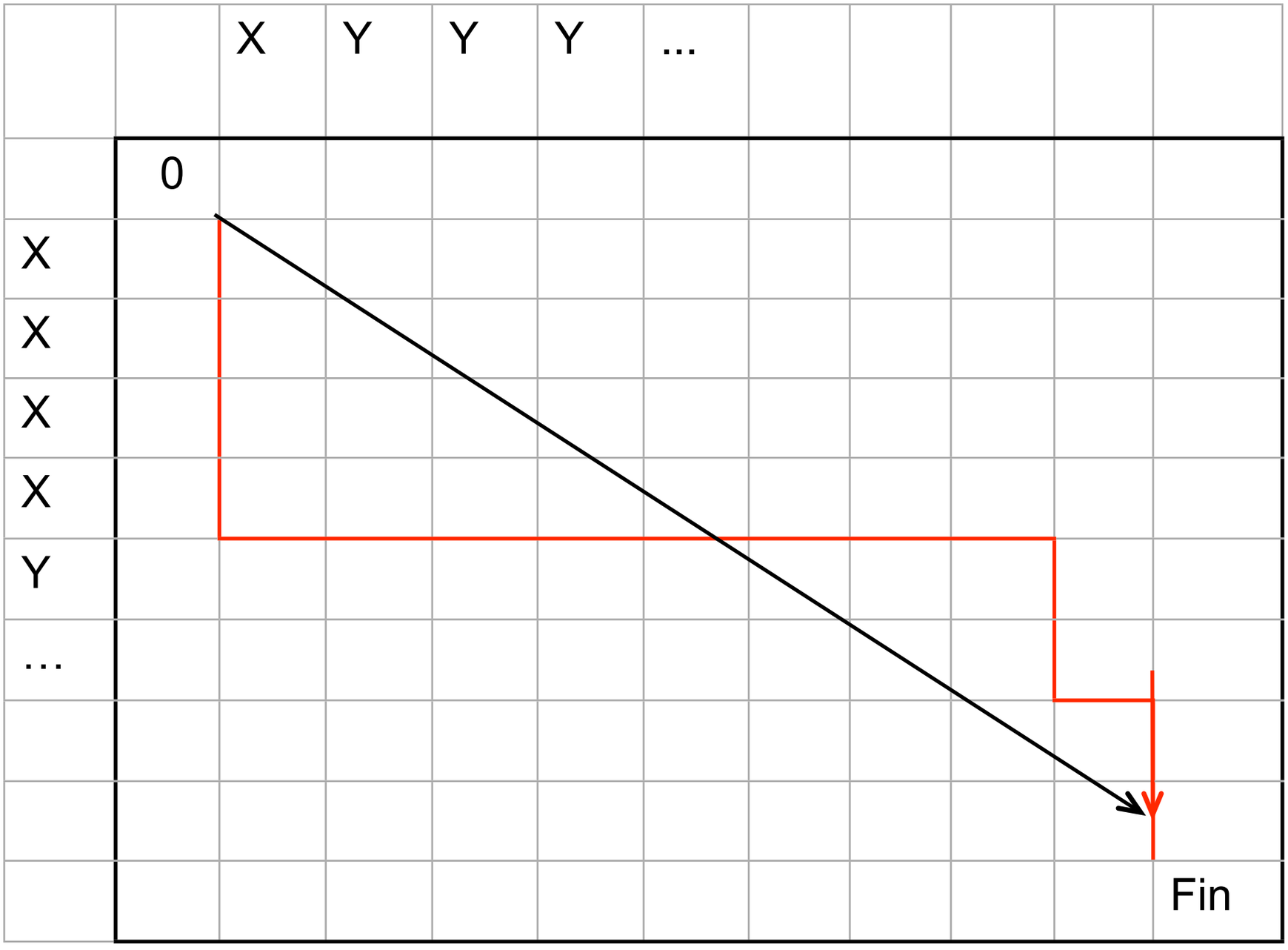
- Objectif différent

  Identifier la structuration sociale des séquences

Comment adapter les MAO aux sciences sociales ?

- États : création d'un espace sociologiquement pertinent
- Temps : quelle origine ? Existe-t-il une échelle de temps commune ? Quel type de régularité temporelle est recherché ?
- Opérations et leurs coûts

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


X
Y
Y
Y
...
0
X
X
X
X
Y
...
Fin

Quelles opérations ? Quels coûts ?

- Les opérations d'insertion et de suppression distordent le timing des séquences pour rapprocher des événements "identiques" éloignés
- Au contraire les opérations de substitution préservent le timing des séquences
- Aucune interprétation des opérations n'est nécessaire
- Quelle interprétation des coûts ?
  - coûts de substitution doivent capturer la probabilité que deux événements différents appartiennent en fait à une même trajectoire
  - coûts d'insertion-suppression doivent être déterminés en fonction de l'importance du timing des événements

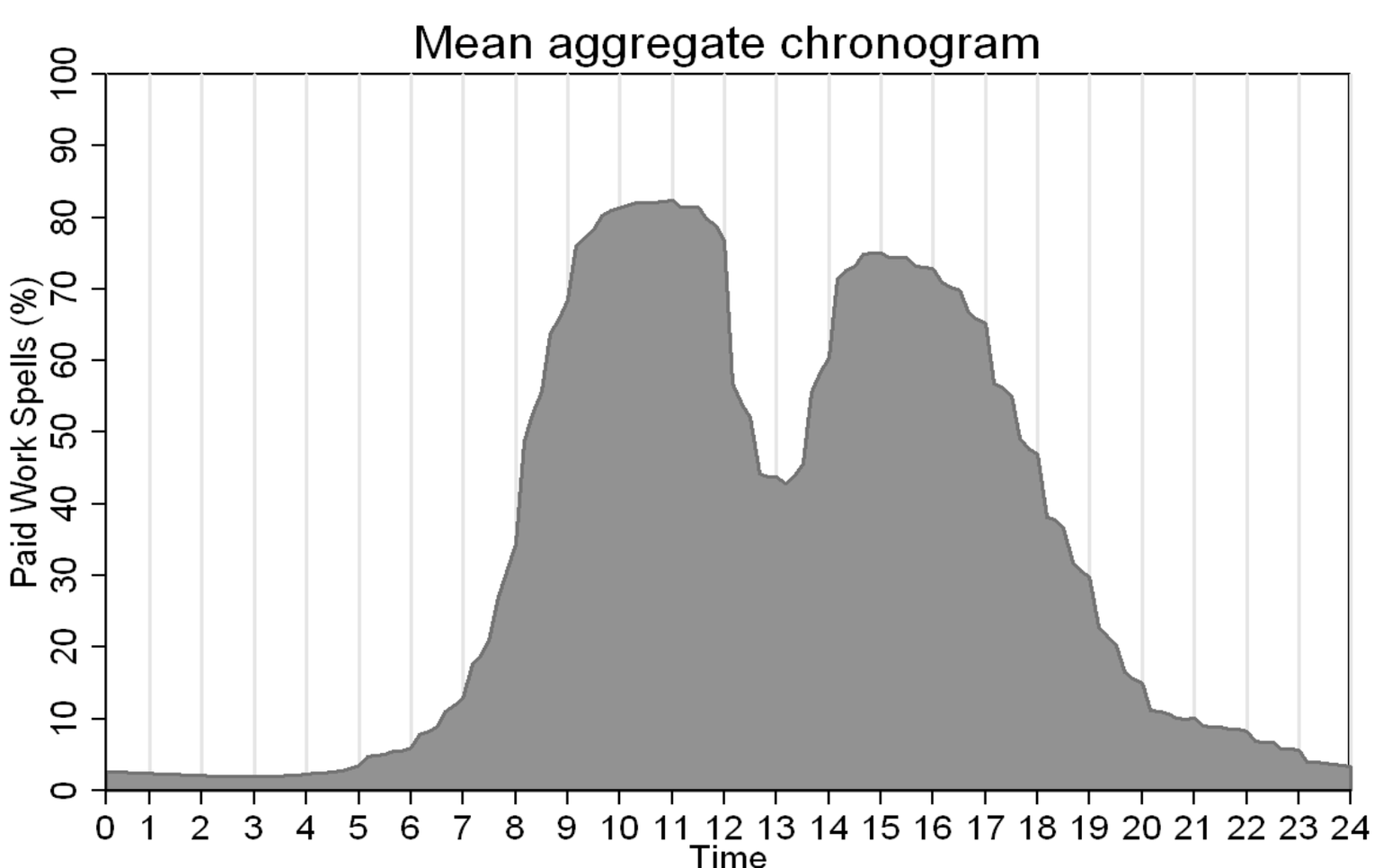
Distance de Hamming			Distance de Levenshtein II
Uniquement des opérations de substitution	coût de substitution < 2 × coût d'insertion suppression	coût de substitution ≥ 2 × coût d'insertion suppression	Uniquement des opérations d'insertion-suppression
Nombre d'événements contemporains identiques communs			Longueur des sous-séquences identiques

4. Exemples

- Les journées de travail (Dynamic Hamming Matching)
- Les semaines de travail (Two-stage sequence analysis)
- Multiple sequence analysis

Les journées de travail

- Enquêtes emploi du temps 1986 et 1999
- Population : journées avec au moins 5 ou 10 minutes de travail
- Deux états : travail et non travail



Journée de travail moyenne en France en 1999

Distance de Hamming			Distance de Levenshtein II
Uniquement des opérations de substitution	coût de substitution < 2 × coût d'insertion suppression	coût de substitution ≥ 2 × coût d'insertion suppression	Uniquement des opérations d'insertion-suppression
Nombre d'événements contemporains identiques communs			Longueur des sous-séquences identiques

Théorie sociologique du temps

- Durkheim : « Un calendrier exprime le rythme de l'activité collective en même temps qu'il a pour fonction d'en assurer la régularité »

<u>Les formes élémentaires de la vie religieuse</u>

- Le temps est socialement différencié
- C'est le rythme de la vie collective qui différencie le temps

Rythme collectif = matrice de transition

8:00	Non travail	Travail	
Non travail	0,78	0,22	1
Travail	0,04	0,96	1
		8:10	

Enquête emploi du temps 1999

Rythme collectif = matrice de transition

8:10

8:00	Non travail	Travail	
Non travail	0,78	0,22	1
Travail	0,04	0,96	1

Enquête emploi du temps 1999

Rythme collectif = matrice de transition

10:40 \ 10:50	Non travail	Travail	
Non travail	0,97	0,03	1
Travail	0,01	0,99	1

Enquête emploi du temps 1999

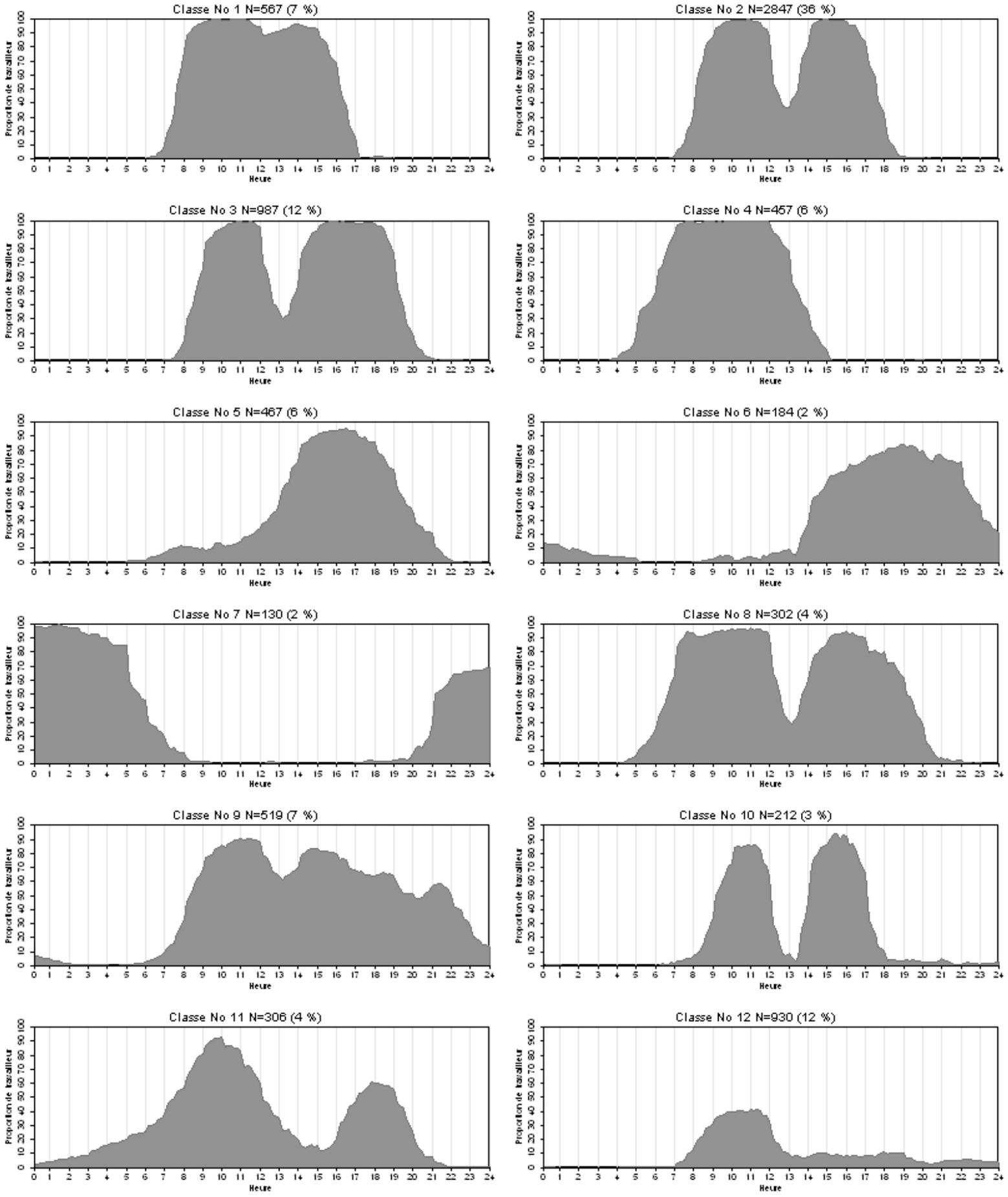
Quand le timing est crucial

- Utilisation des seules opérations de substitution (Hamming)
- Mais plusieurs matrices de coûts de substitution
- Coûts de substitution interprétés comme la probabilité d'appartenir au même rythme collectif
- Dynamic Hamming Matching

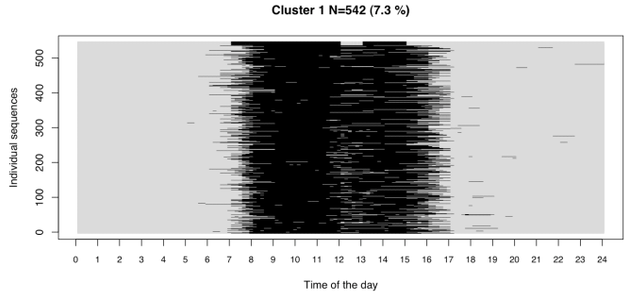
$$s_t(a,b) = \begin{cases} 4 - \left[p\left(X_t = a \middle| X_{t-1} = b\right) + p\left(X_t = b \middle| X_{t-1} = a\right) + p\left(X_{t+1} = a \middle| X_t = b\right) + p\left(X_{t+1} = b \middle| X_t = a\right) \right] & \text{if } a \neq b \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

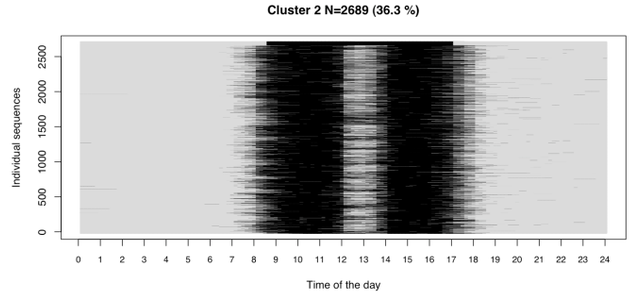
- Plus les transitions entre $t-1$ et $t+1$ et entre t et $t+1$ sont élevées entre les états A et B
  Plus le coût de substitution entre les événements A et B est faible (A et B appartiennent au même rythme collectif)

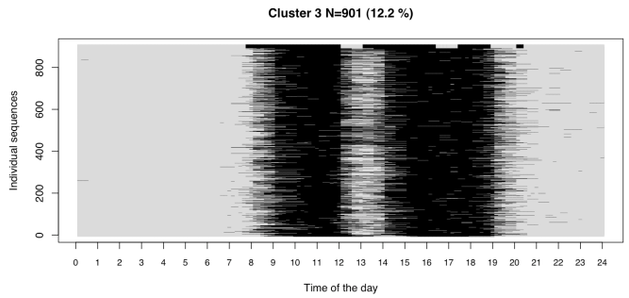
Tempogramme agrégé des douze types de journées de travail

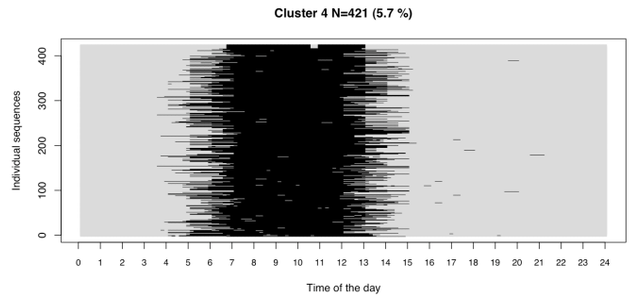


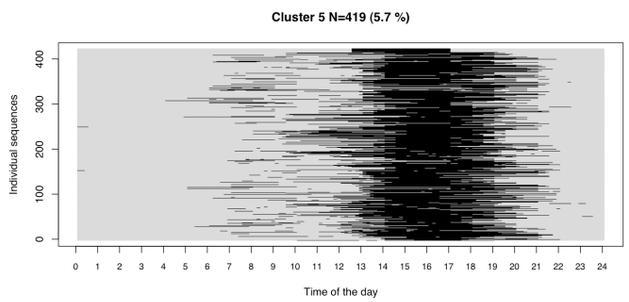
Tempogramme individuel des douze types de journées de travail

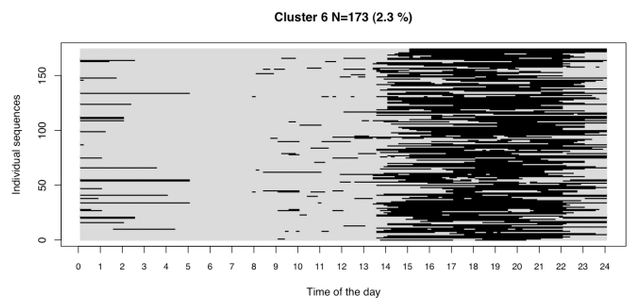


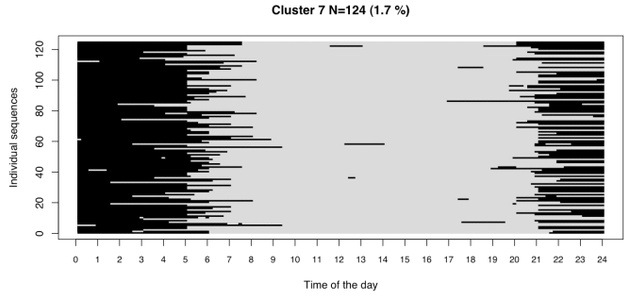


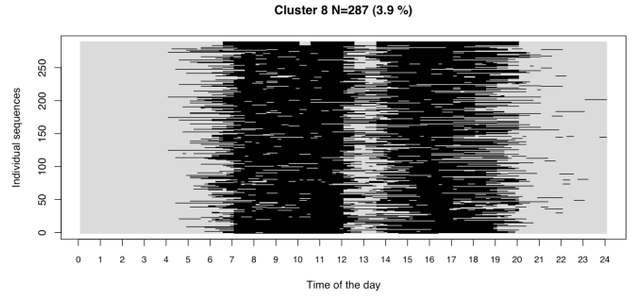


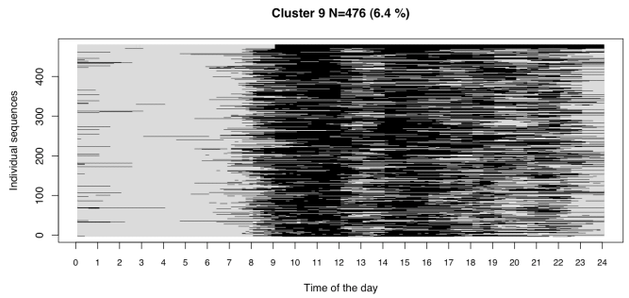


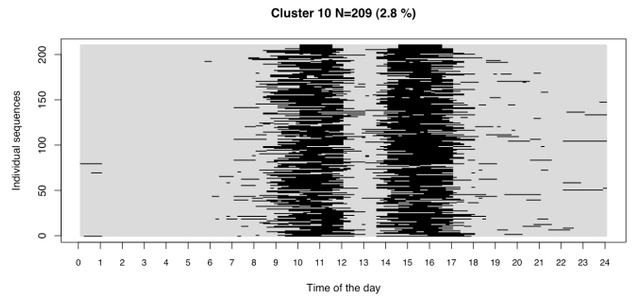


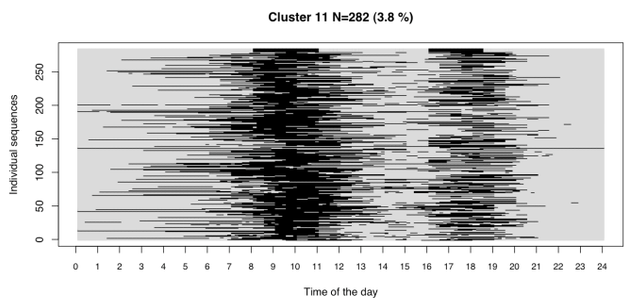


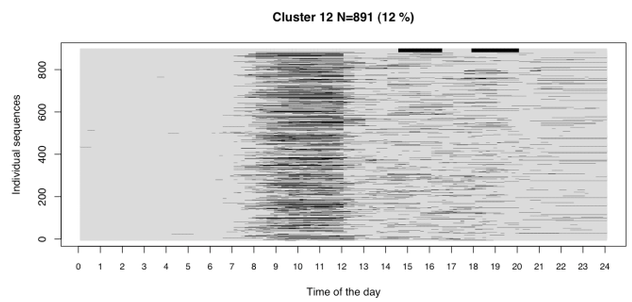












	Type d'horaire de travail	Effectifs (en %) 1985-86	1998-99
Standard	**Standard moyenne**	**56,5**	**54,7**
	Journée 8 à 4	7,6	6,8
	Journée 9 à 5	38,2	33,9
	Journée 10 à 7	10,7	14,0
Atypique	**Décalé**	**14,4**	**16,6**
	Matin	5,3	6,1
	Après-midi	5,4	6,4
	Soir	2,1	2,5
	Nuit	1,7	1,6
	Extensif	**9,1**	**11,6**
	Matin et après-midi	3,5	4,1
	Soir	5,6	7,5
	Irrégulier	**20,0**	**17,1**
	Fragmenté temps réduit	3,2	2,4
	Fragmenté temps plein	3,5	4,2
	Très faible durée	13,3	10,5
	Total	100,0	100,0

Les semaines de travail

- Enquête emploi du temps 1999
- Semainier : horaires de travail pendant sept jours avec une précision de 15 min.
- Utilisation des MAO en deux temps
- Typologie de journées de travail pour simplifier les semaines de travail
- Typologie des semaines de travail à partir des semaines de travail simplifiées

1re étape : typologie des journées de travail

4 920 semaines (2 états et 672 épisodes)

25 138 jours de travail

9 302 jours non travaillés

Optimal Matching

4 types de journées de travail

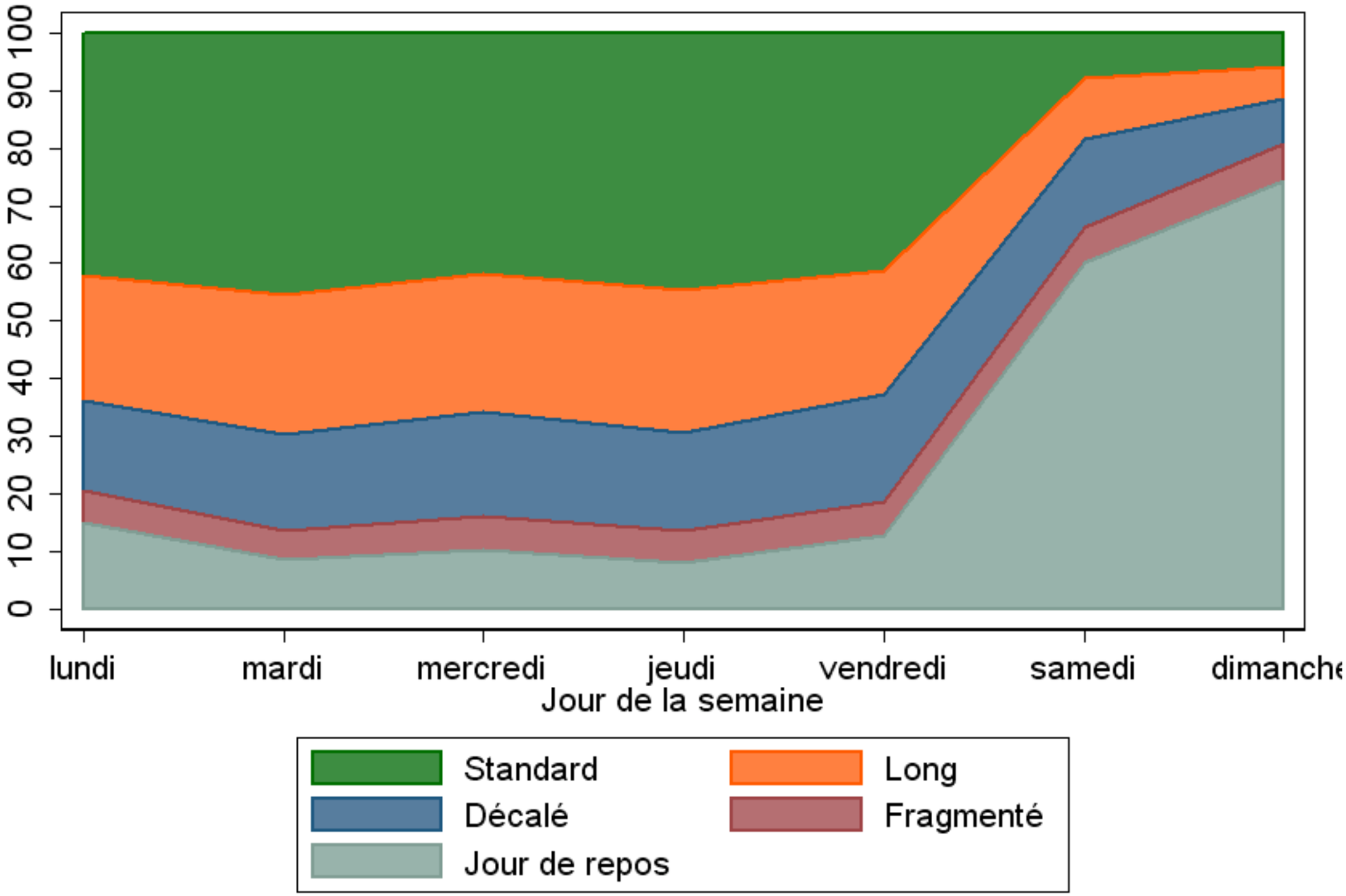
1 type de journée non travaillée

4 920 semaines simplifiées (5 états et 7 épisodes)

I re étape : typologie des journées de travail

Type d'horaires de travail	Taille (%)
9 to 5	45
Longs	26
Décalés	21
Fragmentés	8

Semaines simplifiées

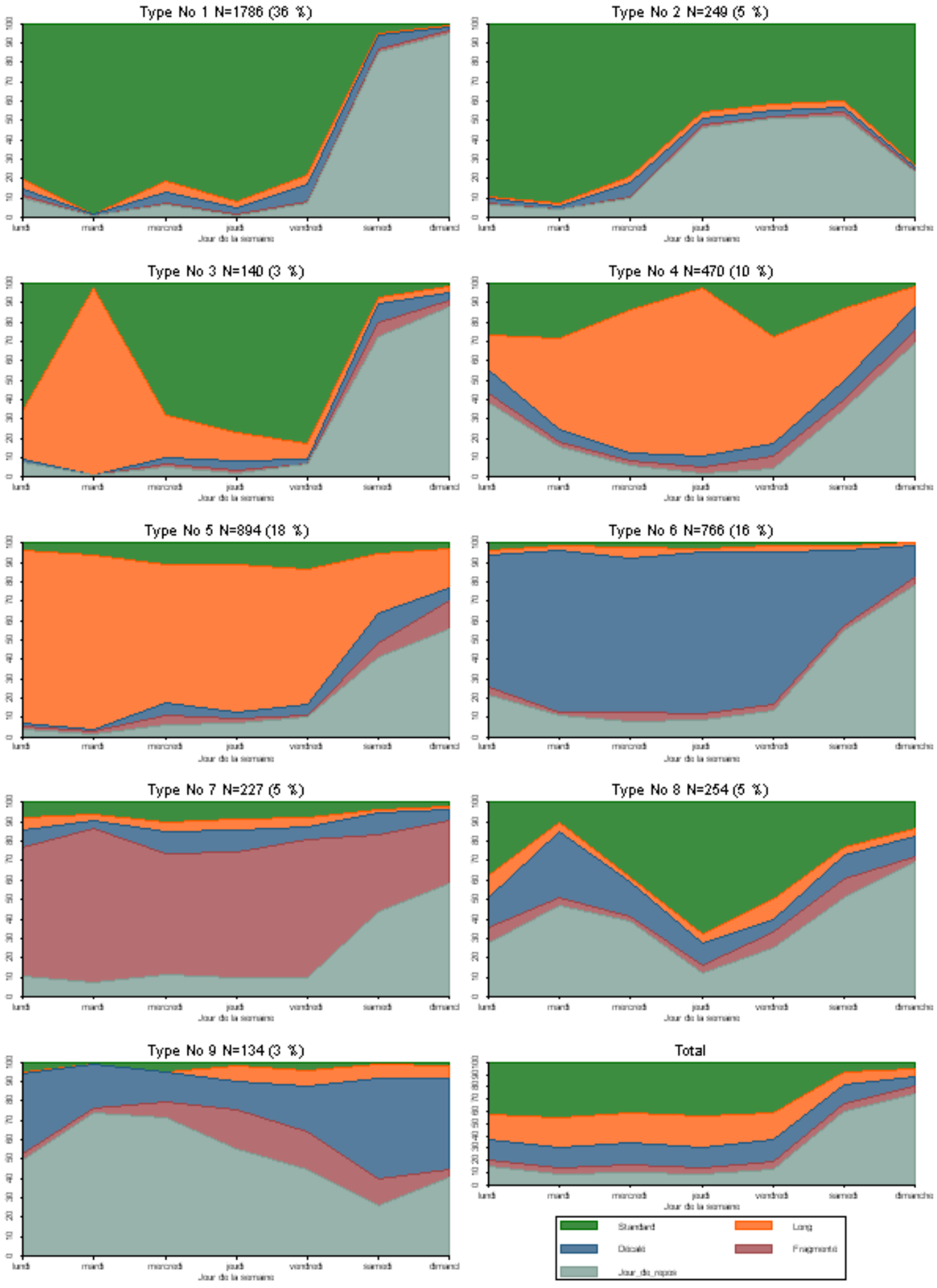


2e étape : typologie des semaines de travail

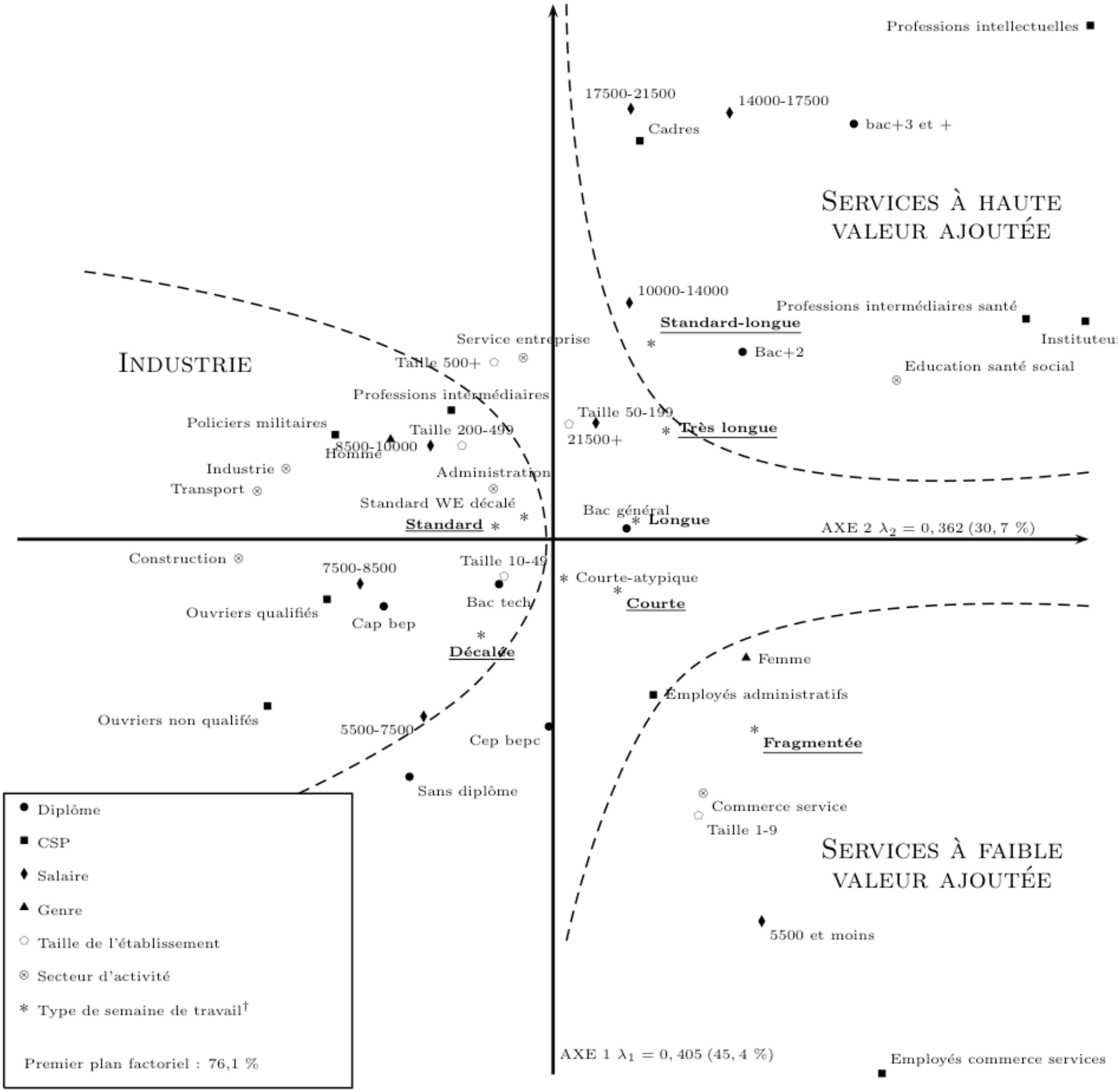
4 920 semaines simplifiées (5 états et 7 épisodes)

Optimal Matching

9 types de semaines de travail


Type No 1 N=1786 (36 %)
Type No 2 N=249 (5 %)
Type No 3 N=140 (3 %)
Type No 4 N=470 (10 %)
Type No 5 N=894 (18 %)
Type No 6 N=766 (16 %)
Type No 7 N=227 (5 %)
Type No 8 N=254 (5 %)
Type No 9 N=134 (3 %)
Total
Jour de la semaine
Standard
Long
Diésel
Fragmenté
Jour de repos

Type de semaines de travail	Taille (%)
Standard	36.4
Standard avec week-ends décalés	5.0
Standard-long	2.8
Long (travail le samedi)	9.1
Très long (travail samedi et dimanche)	18.3
Décalée	15.6
Fragmentée	4.8
Court	5.2
Court, décalée avec travail le week-end	2.8
Total	100.0


Professions intellectuelles
17500-21500
14000-17500
Cadres
bac+3 et +
Services à haute valeur ajoutée
10000-14000
Professions intermédiaires santé
Standard-longue
Service entreprise
Instituteu:
Bac+2
Industrie
Taille 500+
Education santé social
Professions intermédiaires
Taille 50-199
Policiers militaires
Taille 200-499
Très longue
21500+
8500-10000
Homme
Industrie
Administration
Transport
Standard WE décalé
Bac général
Standard
Longue
AXE 2 $\lambda_2 = 0,362$ (30,7 %)
Construction
7500-8500
Taille 10-49
Courte-atypique
Courte
Ouvriers qualifiés
Bac tech
Cap bep
Décalée
Femme
Employés administratifs
Ouvriers non qualifés
5500-7500
Cep bepc
Fragmentée
Sans diplôme
Commerce service
Taille 1-9
Services à faible valeur ajoutée
5500 et moins
Diplôme
CSP
Salaire
Genre
Taille de l'établissement
Secteur d'activité
Type de semaine de travail†
Premier plan factoriel : 76,1 %
AXE 1 $\lambda_1 = 0,405$ (45,4 %)
Employés commerce services

Multiple sequence analysis

- Décrire des trajectoires multidimensionnelles
- Exemple (Pollock 2007):
  - emploi
  - logement
  - situation matrimoniale
  - responsabilités d'au moins un enfant de moins de 16 ans
- Coûts
  - indel = 1
  - substitution = somme de quatre coûts de substitution

Pollock 2007

Table 2. Substitution cost matrices†

		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Employment status										
Self employed	1	0								
Employed	2	0.6	0							
Unemployed	3	1.4	0.8	0						
Retired	4	1.2	1.2	1.2	0					
Maternity leave	5	1.4	0.6	1.4	1.4	0				
Family care	6	1.2	0.8	1.2	0.8	1	0			
Full-time student	7	1.4	0.6	1	1.4	2	1.4	0		
Long term sick	8	1.4	1.4	1.2	0.8	2	1.2	1.4	0	
Government training	9	1.4	0.8	1	1.4	2	1.4	1.4	1.4	0
Housing tenure										
Own outright	1	0								
Own with mortgage	2	0.6	0							
Local authority rent	3	1.4	0.8	0						
Housing association rent	4	1.4	1	0.8	0					
Rent from employer	5	1.4	1	1.4	1.4	0				
Private rent (unfurnished)	6	1.4	1	1.4	1	0.8	0			
Private rent (furnished)	7	1.4	0.8	1.4	1.4	1.2	1	0		
Marital status										
Married	1	0								
Separated	2	1	0							
Divorced	3	0.5	0.6	0						
Widowed	4	0.5	1.4	2	0					
Never married	5	0.5	2	2	2	0				
Responsibility for children aged under 16 years										
Yes	1	0								
No	2	0.5	0							

†The cost of insertion or deletion is 1; hence a substitution cost of 2 will invoke an indel instead of a substitution. This means that there is no need to create a substitution cost which is greater than the sum of an insertion and a deletion.

Article	Coûts
Abbott et Forrest 1986 Optimal matching methods for historical sequences	Substitutions = nombre de différences dans la hiérarchie des pas de danse divisé par le nombre de pas total (5) Indel = 1 (max subs)
Stovel et al. 1996 “Ascription into achievement: models of career systems at Lloyds bank”	Substitutions = somme de deux matrices de transition entre les professions et les filiales inspirées par l’examen des deux matrices de transition entre 1890 et 1970 Indel = 6.5 (max subs) Dissimilarités standardisées par la longeur de la séquence
Halpin et Chan 1998	Substitutions déterminées théoriquement (varient entre 1 et 4) Indel = 3
Abbott et Hrycak 1999 Measuring Resemblance in Sequence Data: An Optimal Matching Analysis of Musicians' Careers	Substitutions dérivéEs de la matrice de transition entre tous les états pour l’ensemble des dates (0,47 - 1) Indel (1,088)
Blair-Loy 1999 “Career patterns of executive women in finance”	Coûts de substitution déterminés théoriquement (varient entre 0 et 1) Coûts d’insertion-suppression fixés à 0.48 après plusieurs essais
Han et Moen 1999 “Clocking Out Temporal Patterning of Retirement”	Non spécifiés. Coûts de substitution semblent dépendre des transitions entre les différents états
Stovel 2001 “Local sequential patterns the structures of lynching in the deep south”	Sij = |i - j|/(j+1)
Clark et al. 2003 “Housing Careers in the United States”	Indel = 1 Subs ϵ {0.1,0.8]
Stovel et Bolan 2004 “Residential trajectories”	Coûts de substitution inspirés par l’examen de la matrice de transition agrégée pour les 14 épisodes Indel = 2.7 lorsque les séquences comparées sont de même longueur et = 0.45 sinon Argument : fixed indel cost is drawing sequences together on the basis of similar lengths rather than similar patterns
Brzinsky-Fay 2007 “Lost in transition”	Indel = 1 Subs = 2
Pollock 2007 “Holistic trajectories: a study of combined employment, housing, and family careers by using multiple-sequence analysis”	Somme de 4 matrices de coût de substitution inspirées par l’observation des transitions pour chaque dimension varient entre 0,5 et 2 Indel = 1

5. En pratique

Logiciels d'analyse de séquences

- MAO
  - SAS : macro "distance" introuvable
  - TDA
    - Avantages : très utilisé, fait figure de référence, gratuit, libre
    - Inconvénients : très peu convivial (mais pas trop difficile à utiliser) et ne permet pas de faire d'exploiter la matrice de distance
  - Stata : sq (Ulrich Kohler et al.)
    - Avantages : outils pour représenter graphiquement les séquences, permet d'analyser la matrice de distance
    - Inconvénients : très lent (et il faut avoir Stata)
- Dynamic Hamming matching
  - Sas macro
  - Stata plugin (10 fois plus rapide)

Exploitation du fichier de distance

- Clustering
  - Ward : critère le plus utilisé mais OMA n'est pas une distance euclidienne et Ward n'est pas la plus performante empiriquement
  - WPGMA flexible (beta flexible) ou, mieux, UPGMA flexible :
    1. R: WPGMA flexible
    2. Sas & Clustan Graphics: WPGMA flexible
    3. Stata : aucune
    4. SPSS : aucune
- Partitionnement : cartes de Kohonen (cf. travaux de Patrick Rousset)
- Échelonnement multidimensionnel (multidimensional

Conclusion

- Strauss : ordre social ne peut être compris que dans une perspective diachronique
- position vs. trajectoire ou parcours
- Lien individu - société
- MAO : outil qui permet de d'identifier des trajectoires types vs. régression