Les Méthodes d'Appariement Optimal en sociologie

Congrès de l'AFS - RT n°20 - 7 septembre 2006

Laurent Lesnard et Thibaut de Saint Pol[1]

Observatoire sociologique du changement (Sciences-po, CNRS)

Laboratoire de sociologie quantitative (Crest, Insee)

Introduction

- Méthodes d'Appariement Optimal (*en anglais* Optimal Matching Analysis)
- Origine : informatique (puis biologie)
- M.A.O. utilisées pour la 1re fois en sciences sociales par Andrew Abbott.
- Objectif des méthodes d'Appariement Optimal : comparer et regrouper les séquences.

Plan de la présentation

- Les méthodes d'appariement optimal
  - Principe
  - La question des coûts
- Applications aux enquêtes Emploi du temps
  - Le dîner des Français
  - Le temps de travail

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 ».

[1] Courriels : laurent.lesnard@sciences-po.fr et thibaut.desaintpol@sciences-po.org

Exemple : les engagements successifs de deux militants X et Y 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

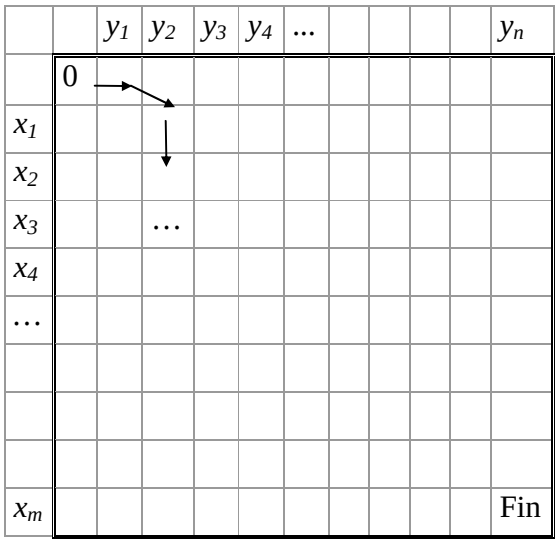
- Autre possibilité :

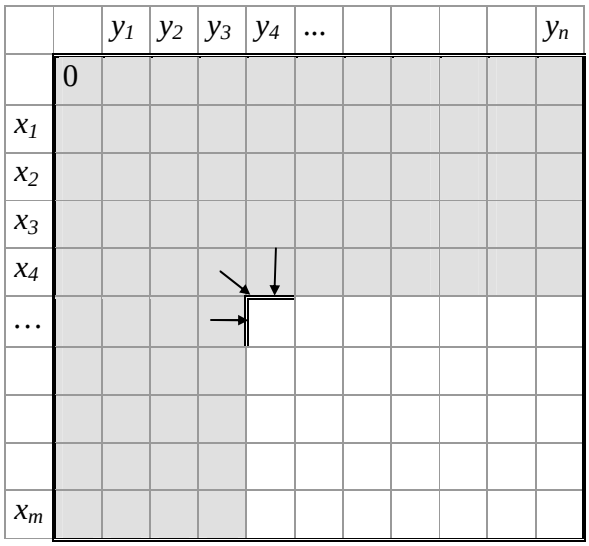
A : X – X – X – X – Y

B : X – X – X – X – Y

1 insertion
2 substitutions
Coût total=5

La représentation matricielle des différents chemins :





Ici à titre d'exemple :
Insertion de Y1 / Transformation de X1 en Y2 / Suppression de X2

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.

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

---

La détermination des coûts

- 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

Deux applications

- 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)
- 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

En pratique

- TDA (freeware) : il existe un module M.A.O. mais les coûts ne peuvent pas varier avec l'échelle de temps
- Programmation dans le logiciel SAS (module de calcul matriciel + langage macro)

+

- Classification Ascendante Hiérarchique
  (ici méthode Beta-Flexible ; méthode de Ward non recommandée)

Premier exemple: le dîner des Français

- Traitement des séquences d'activités sur la période 18h50-21h30 pendant laquelle se concentrent les prises alimentaires.
- Analyse du carnet journalier de l'enquête Emploi du Temps 1998 dont les activités ont été regroupées en 25 catégories.
- Séquences de 16 éléments (16 plages horaires de 10 minutes) pouvant prendre 25 valeurs différentes.

Descriptif des dix classes

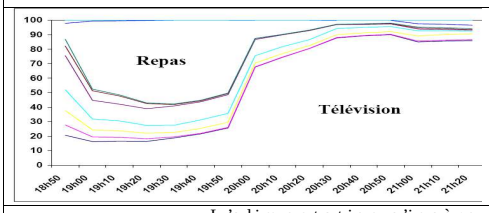
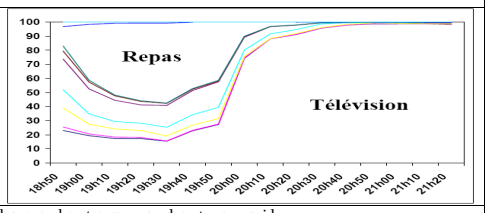
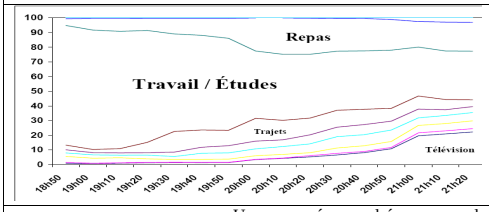
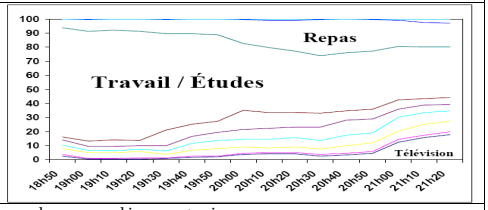
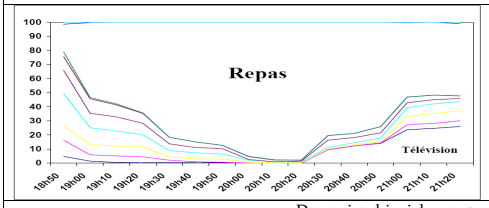
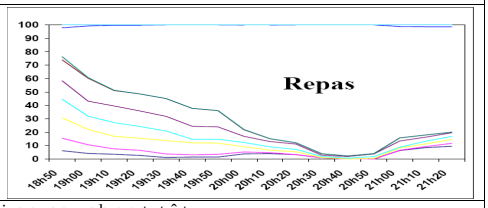
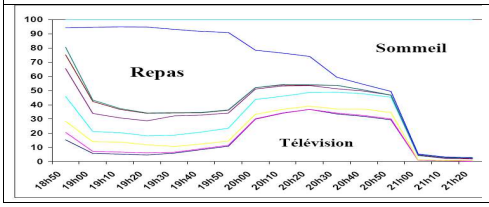
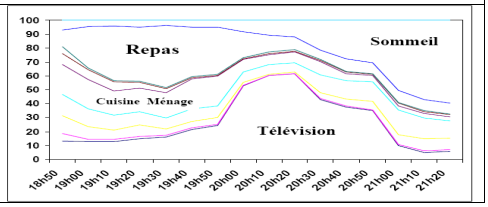
Classe	Effectif (%)	Description sommaire
1	5,8	Un repas encadré de tâches ménagères
2	12	Un repas pris tôt suivi par une multitude d'activités
3	7,5	Un repas pris tard et encadré de tâches ménagères
4	20,8	Une soirée découpée entre le repas et la télévision
5	5,4	Une soirée entièrement consacrée à la télévision
6	9,4	L'alimentation s'insère dans le temps de travail
7	6	Une prépondérance des plages alimentaires
8	4,9	Un repas qui s'insère dans des pratiques de loisirs à l'extérieur
9	24,5	Un repas pris tard, après une multitude d'activités qui se prolongent jusqu'à 20h00.
10	3,7	Des individus qui se couchent tôt

Les séquences moyennes des dix classes

Classe	18h50	19h00	19h10	19h20	19h30	19h40	19h50	20h00	20h10	20h20	20h30	20h40	20h50	21h00	21h10	21h20
1	Ménager	Ménager	Ménager	Ménager	Repas	Repas	Repas	Repas	Repas	Repas	Télévision	Télévision	Télévision	Télévision	Télévision	Télévision
2	Ménager	Repas	Repas	Repas	Repas	Repas	Repas	Ménager	Ménager	Ménager	Enfants	Enfants	Télévision	Télévision	Télévision	Télévision
3	Ménager	Ménager	Ménager	Ménager	Ménager	Ménager	Ménager	Repas	Repas	Repas	Repas	Repas	Repas	Ménager	Ménager	Ménager
4	Ménager	Repas	Repas	Repas	Repas	Repas	Repas	Télévision	Télévision	Télévision	Télévision	Télévision	Télévision	Télévision	Télévision	Télévision
5	Télévision	Télévision	Télévision	Télévision	Télévision	Télévision	Télévision	Repas	Repas	Repas	Repas	Repas	Repas	Télévision	Télévision	Télévision
6	Travail	Travail	Travail	Travail	Travail	Travail	Travail	Travail	Travail	Travail	Travail	Travail	Travail	Travail	Travail	Travail
7	Repas	Repas	Repas	Repas	Repas	Repas	Repas	Repas	Repas	Repas	Repas	Repas	Repas	Repas	Repas	Repas
8	Rencontres	Rencontres	Rencontres	Rencontres	Rencontres	Rencontres	Rencontres	Rencontres	Rencontres	Rencontres	Rencontres	Rencontres	Rencontres	Rencontres	Rencontres	Rencontres
9	Brico.	Brico.	Brico.	Brico.	Repas	Repas	Repas	Repas	Repas	Repas	Repas	Repas	Repas	Télévision	Télévision	Télévision
10	Repas	Repas	Repas	Repas	Repas	Repas	Repas	Télévision	Télévision	Télévision	Sommeil	Sommeil	Sommeil	Sommeil	Sommeil	Sommeil

Classe	1	2	3	4	5	6	7	8	9	10
Temps moyen consacré à l'alimentation (min)	48	42	40	37	**43**	26	113	29	36	**43**

Comparaison des enquêtes 1985-1998.

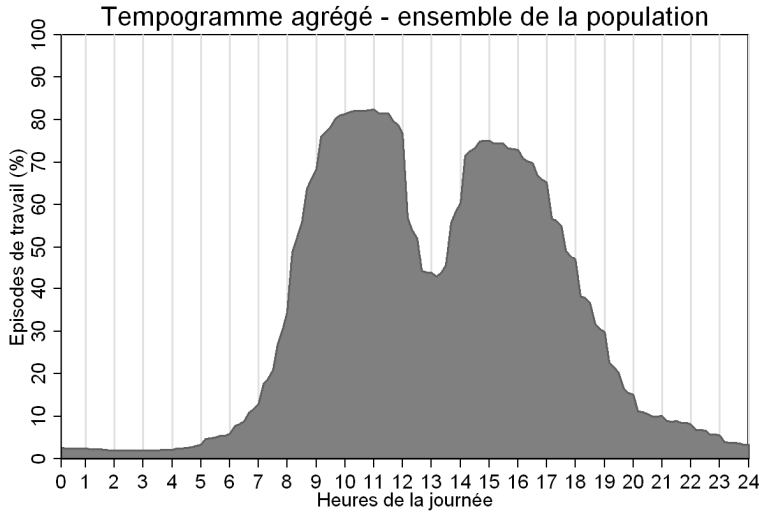
1998-1999	1985-1986
Une soirée découpée entre le repas et la télévision	
	
L'alimentation s'insère dans le temps de travail	
	
Une prépondérance des plages alimentaires	
	
Des individus qui se couchent tôt	
	

Second exemple: le temps de travail

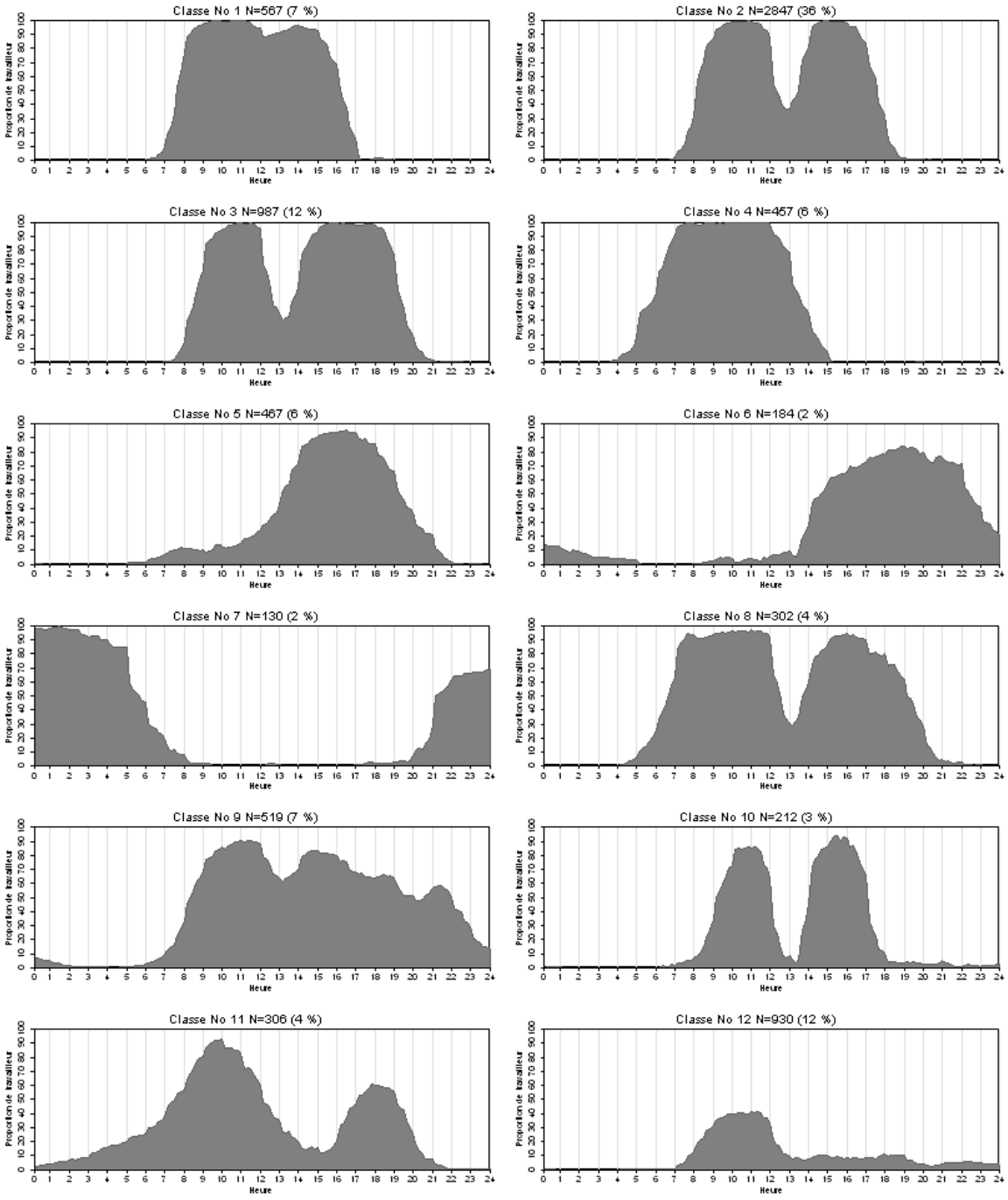
- Deux états : travail et non-travail
- Seule la position du travail dans la journée nous importe
- Utilisation de la seule opération de substitution
- Coûts de substitution dynamiques dérivés des probabilités de transition

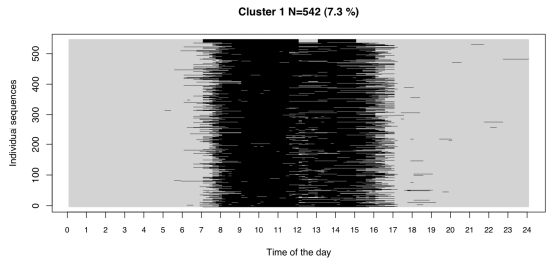
---

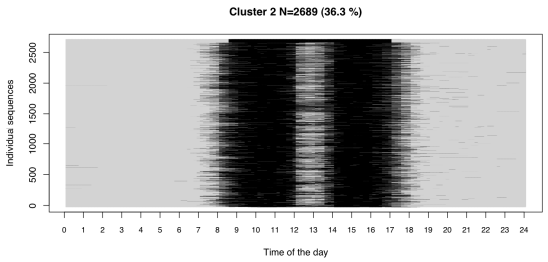
Douze types d'horaires de travail

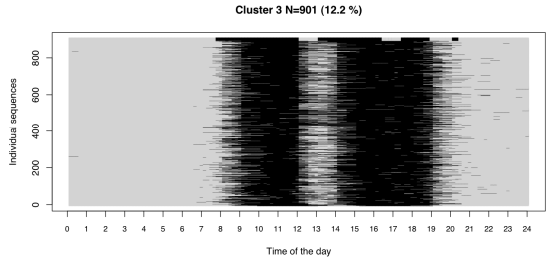


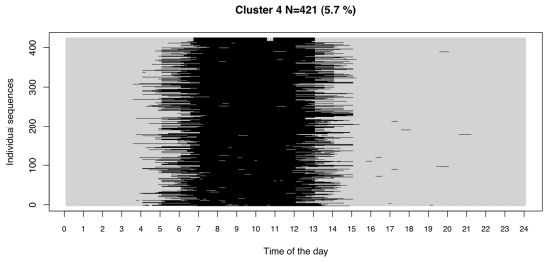
No. classe	Type d'horaire de travail	Effectifs (% de la pop. tot.)	Durée de travail
	Standard	**56,5%**	**8:26**
1	7-16	7,6%	8:14
2	8-18	38,2%	8:17
3	9-19	10,7%	9:09
	Décalé	**14,4%**	**7:16**
4	Matin	5,3%	7:39
5	Après-midi	5,4%	6:46
6	Soir	2,1%	7:20
7	Nuit	1,7%	7:38
	Extensif	**9,1%**	**10:29**
8	Régulier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

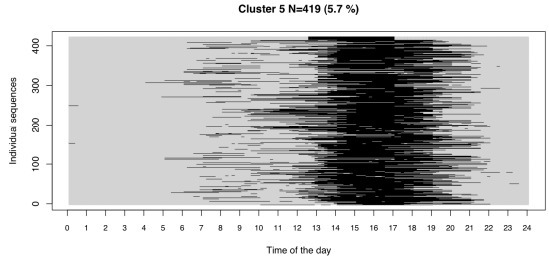

Classe No 1 N=567 (7 %)
Classe No 2 N=2847 (36 %)
Classe No 3 N=987 (12 %)
Classe No 4 N=457 (6 %)
Classe No 5 N=467 (6 %)
Classe No 6 N=184 (2 %)
Classe No 7 N=130 (2 %)
Classe No 8 N=302 (4 %)
Classe No 9 N=519 (7 %)
Classe No 10 N=212 (3 %)
Classe No 11 N=306 (4 %)
Classe No 12 N=930 (12 %)
Proportion de travailleur
Heure

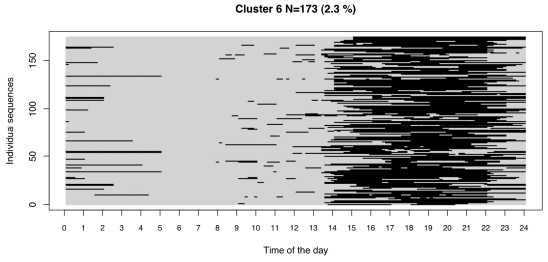

Cluster 1 N=542 (7.3 %)
Individua sequences
Time of the day

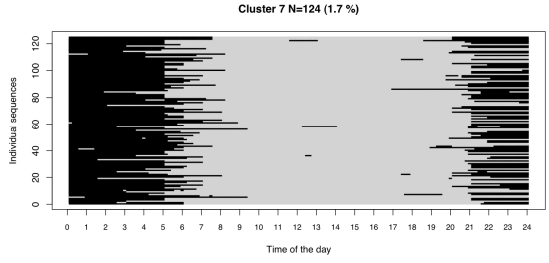

Cluster 2 N=2689 (36.3 %)
Individua sequences
Time of the day

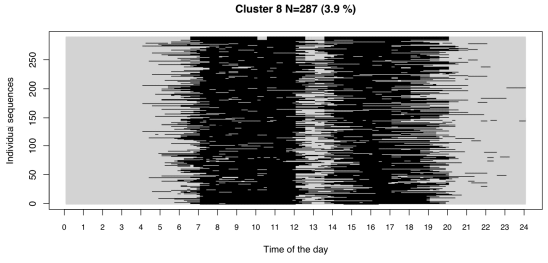

Cluster 3 N=901 (12.2 %)
Individua sequences
Time of the day

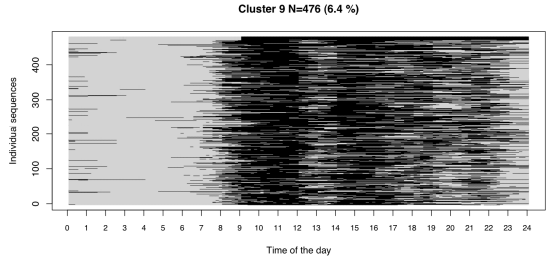

Cluster 4 N=421 (5.7 %)
Individua sequences
Time of the day

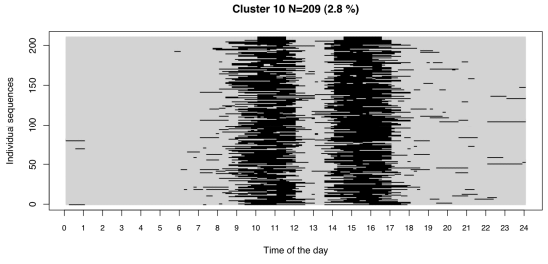

Cluster 5 N=419 (5.7 %)
Individua sequences
Time of the day

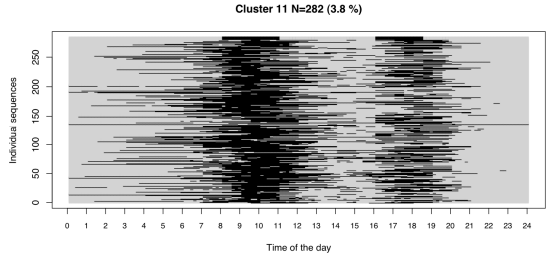

Cluster 6 N=173 (2.3 %)
Individua sequences
Time of the day

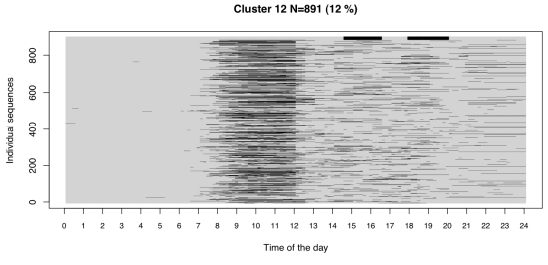

Cluster 7 N=124 (1.7 %)
Individua sequences
Time of the day


Cluster 8 N=287 (3.9 %)
Individua sequences
Time of the day


Cluster 9 N=476 (6.4 %)
Individua sequences
Time of the day


Cluster 10 N=209 (2.8 %)
Individua sequences
Time of the day


Cluster 11 N=282 (3.8 %)
Individua sequences
Time of the day


Cluster 12 N=891 (12 %)
Individua sequences
Time of the day

Conclusion: intérêt de la méthode

- Bâtir des groupes à partir de milliers de séquences.
- Compare les séquences sans avantager aucun élément, donc aucune activité particulière.
- Prend en compte toutes les dimensions de l’emploi du temps (verticale et horizontale).
- Permet de dépasser les limites de méthodes classiques (Comparaison des enquêtes 1985-1998).