Les Méthodes d'Appariement Optimal

De nouveaux outils pour l'analyse de séquences

Laurent Lesnard

Université Paris 10 – Nanterre et CREST (INSEE)

Chercheur invité à l’Université de Essex (ISER)

Les processus en sciences sociales

- Théories sociales dynamiques
- Données souvent statiques
- Outils souvent statiques
- Exception: modèles d'analyse statistique des biographies (modèles de durée, etc.)
- Modélisation sans description
- Régression avec pour hypothèse centrale la causalité

Plan de la présentation

1. Les méthodes d’appariement optimal
    1. Principe
    2. La question des coûts

2. Application : les journées de travail
    1. Une méthode d’appariement non optimal
    2. Résultats

1.1. Principe des M.A.O.

- Objectif des méthodes d'analyse de séquence : déterminer la similarité de séquences (data reduction)
- Méthodes d'Appariement Optimal : une méthode particulière d'analyse de séquences
- Similarité : nombre minimal d'opérations nécessaires pour rendre identiques deux séquences
- Origine : biologie, phylogénétique
- M.A.O. utilisées pour la 1re fois en sciences sociales par Andrew Abbott

1.1. Principe des M.A.O.

- 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 »

1.1. Principe des M.A.O.

- Exemple :

  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

.

1.1. Principe des M.A.O.

- Exemple :

  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

.

1.1. Principe des M.A.O.

- Exemple :

  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

1.1. Principe des M.A.O.

- Exemple :

  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

1.1. Principe des M.A.O.

- Exemple :

  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

1.1. Principe des M.A.O.

- Exemple :

  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

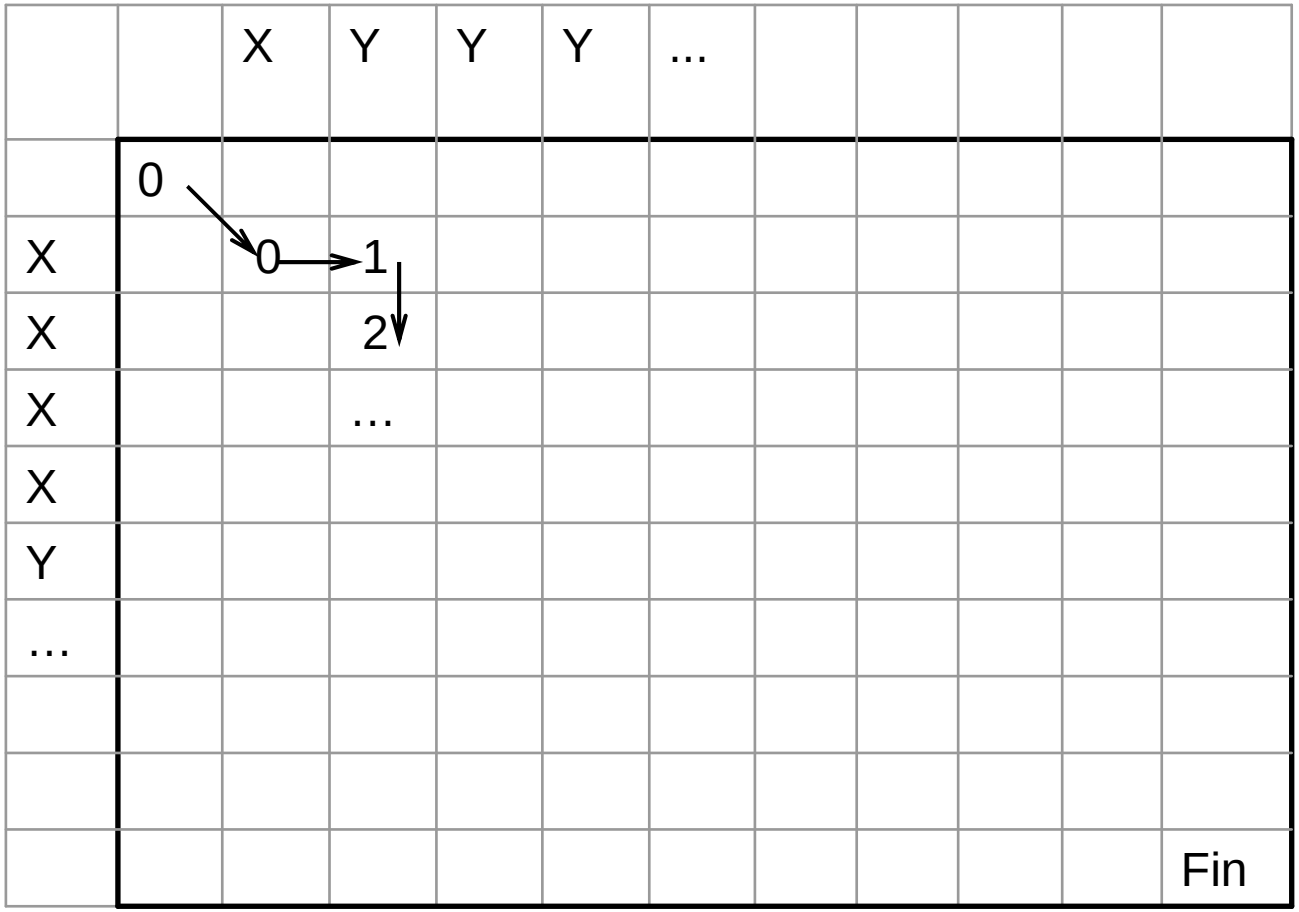
3 insertions
2 suppressions
Coût total=7

- Autre possibilité :

  A : X – X – X – X – Y

  B : X – X – X – X – Y

1 insertion
2 substitutions
Coût total=5


X
Y
Y
Y
...
0
X
0
1
X
2
X
...
X
Y
...
Fin

1.2. Détermination des coûts

- Biologie :
  - inspirés de propriétés électrochimiques et mécaniques
  - hypothèses dérivés d'un modèle théorique
- Sciences sociales :
  - Abbott (1995) : choix des coûts doivent être si possible relié à des considérations théoriques
  - Abbott (2000) : M.A.O. purement descriptives ; paramètres permettent de varier le type de régularité recherché

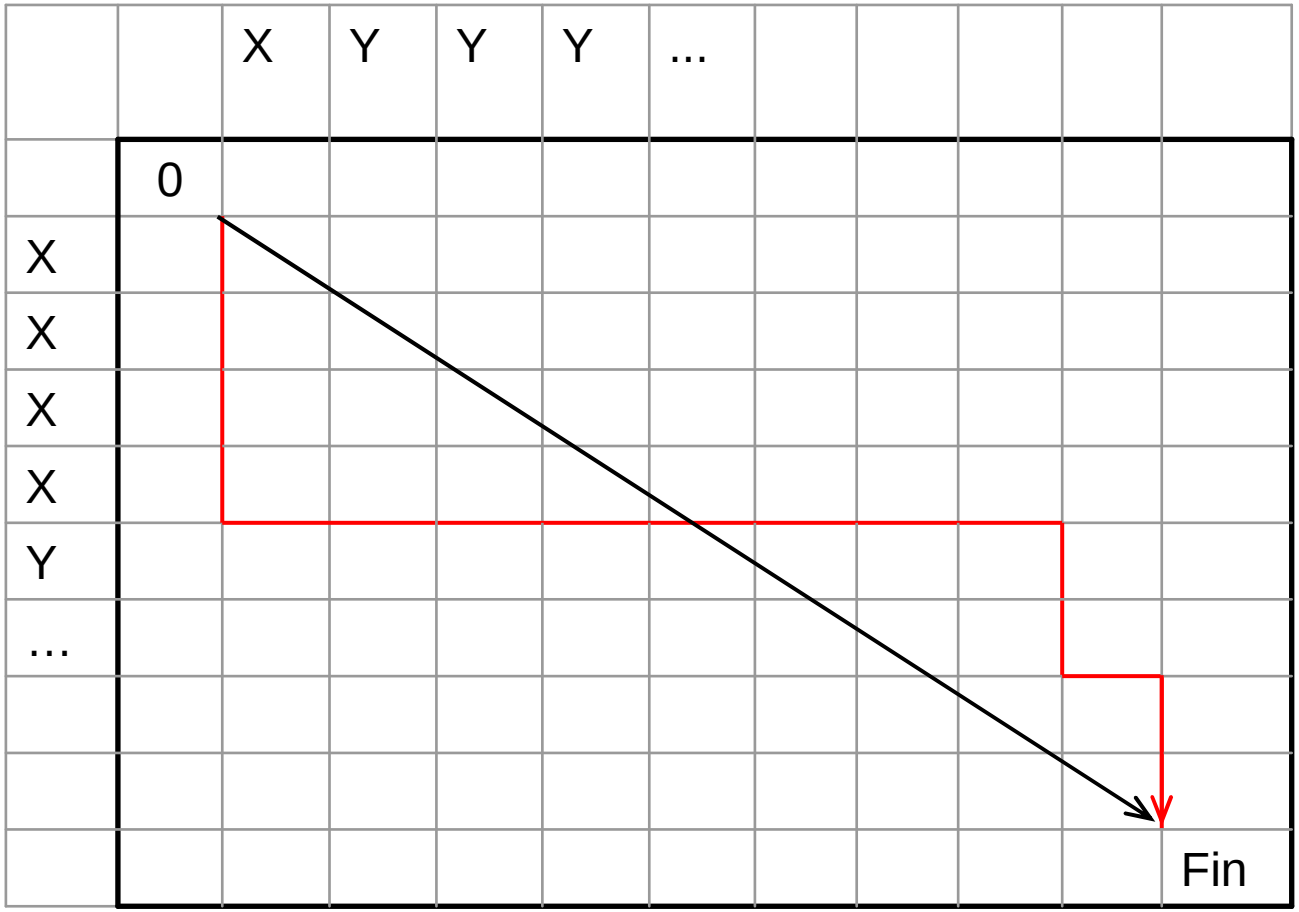
1.2. Détermination des coûts

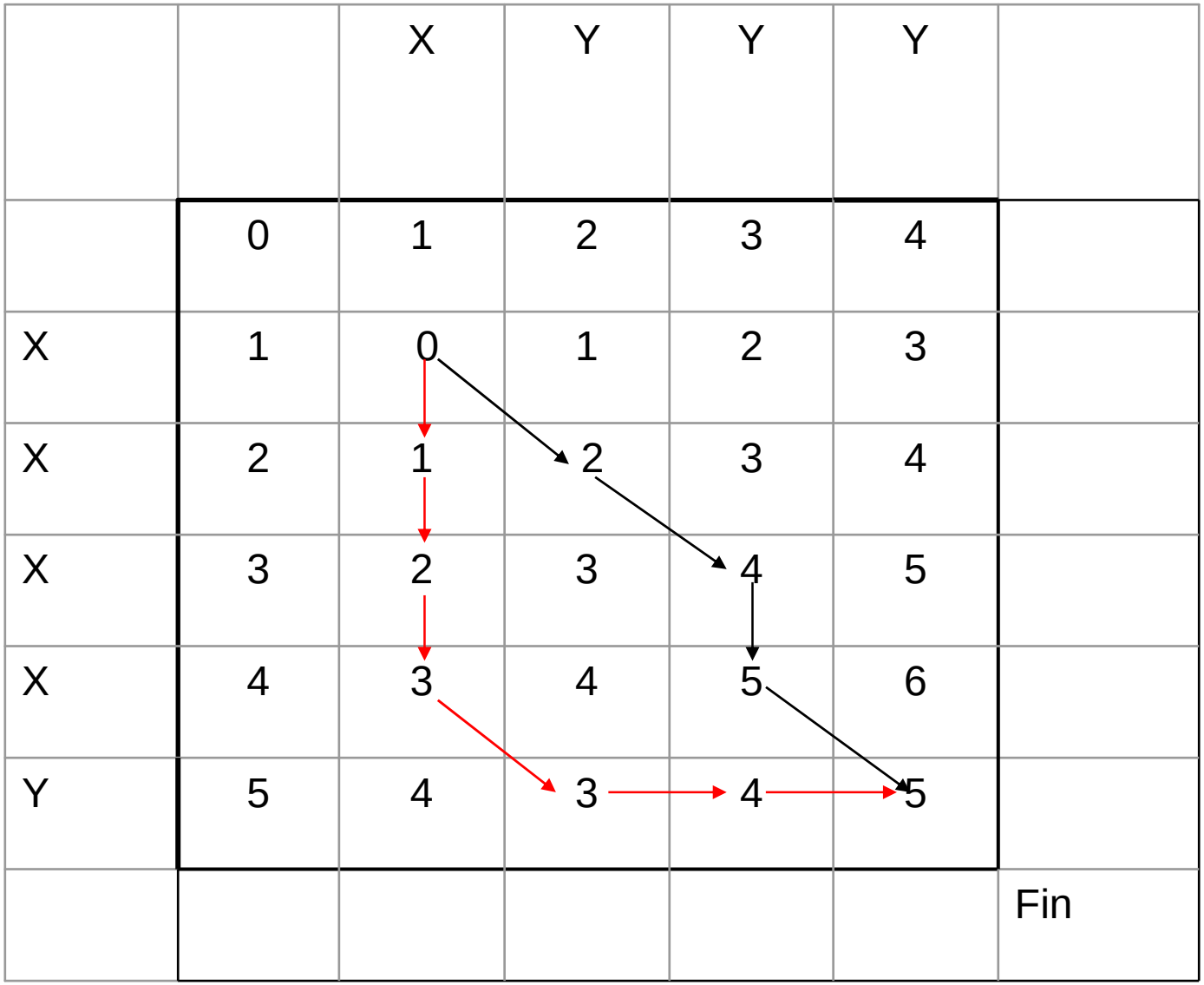
- Sciences sociales (suite) :
  - Levine (2000), Wu (2000), Elzinga (2003) : il faut absolument un modèle théorique pour déterminer les coûts
  - Abbott (2000) : ce n’est pas le cas en biologie donc les M.A.O. n’ont pas à mimer un hypothétique processus de production de séquence

1.2. Détermination des coûts

Séquence=événements + échelle de temps

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

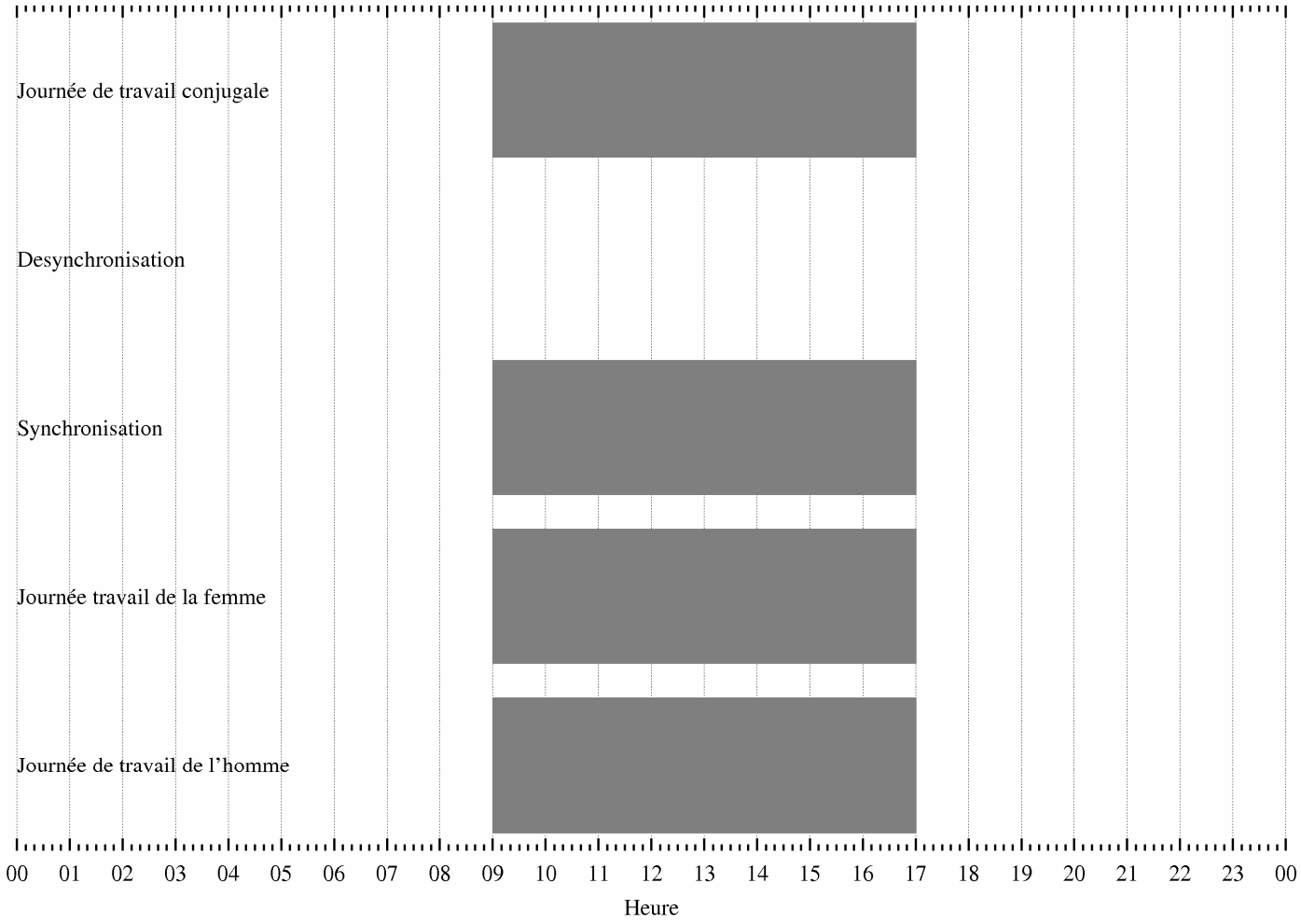

X
Y
Y
Y
...
0
X
X
X
X
Y
...
Fin


X
Y
Y
Y
0
1
2
3
4
X
1
0
1
2
3
X
2
1
2
3
4
X
3
2
3
4
5
X
4
3
4
5
6
Y
5
4
3
4
5
Fin

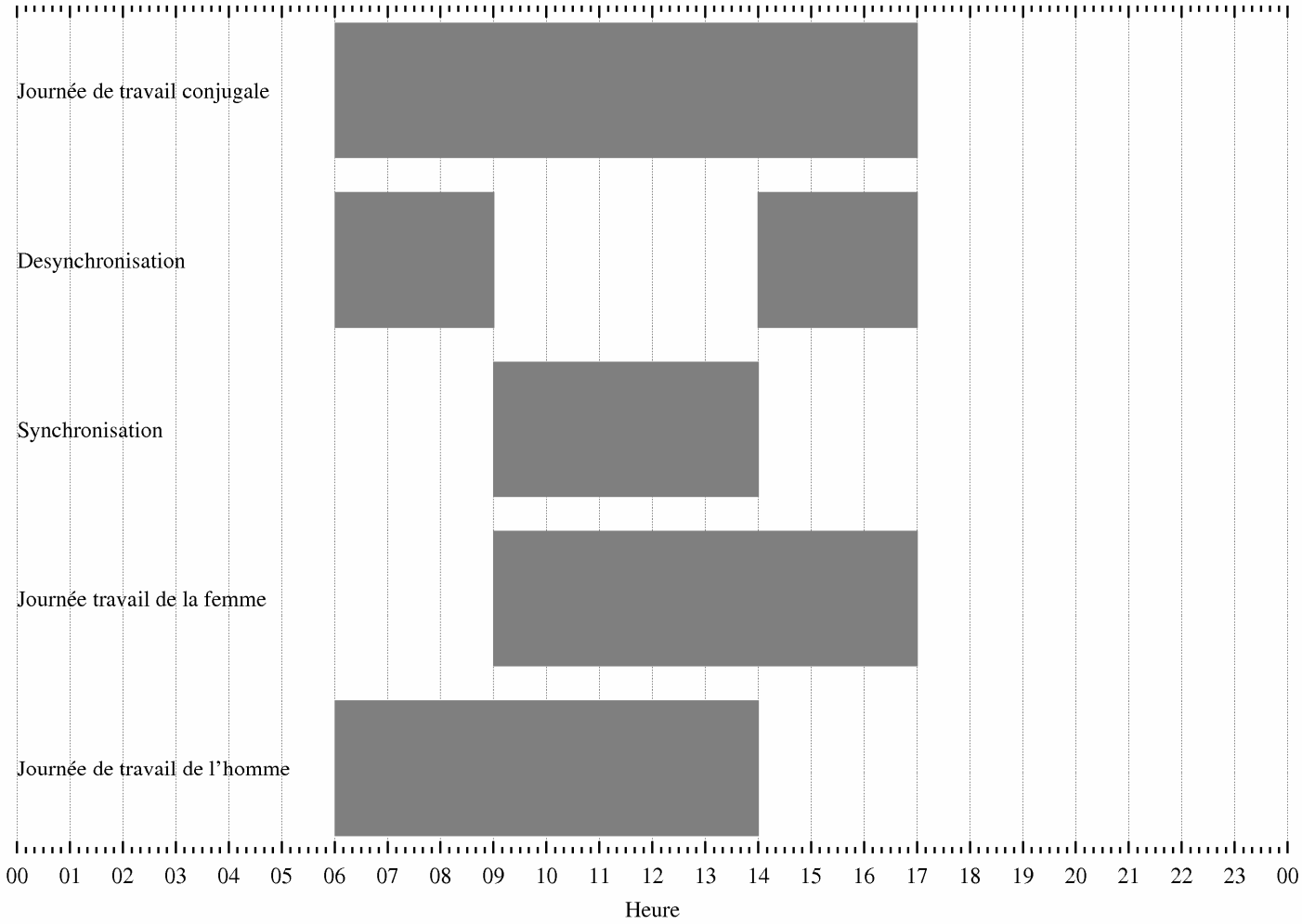
2. Application : les journées de travail

- Enquêtes Emploi du Temps
- Analyse séparée de la durée du travail et de la régularité
  - Budgets-temps
  - Indicateurs (travail de nuit, journées de travail de plus de 10h, etc.)
- Absence de description des journées de travail individuelles et conjugales

Journées de travail des couples



Journées de travail des couples



M.A.O. et horaires de travail

- Le « présent » peut dépendre du « futur »
- L’échelle de temps doit être impérativement respectée
- Utiliser seulement des opérations de substitutions
- Coûts de substitution doivent refléter l’éloignement d’événements différents qui ont lieu au même moment

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 » Les formes élémentaires de la vie religieuse
- Le temps est socialement différencié
- C'est le rythme de la vie collective qui différencie le temps

Une méthode d'appariement non optimal

- Les transitions entre les différents états fournissent une mesure de la similarité entre les séquences
- Des transitions élevées entre deux états indiquent un changement de rythme collectif
- Des transitions faibles entre deux états indiquent deux sous-rythmes très différents
- Les coûts de substitutions dépendent donc de l'échelle de temps
- Exemple :
  - Un actif qui travaille à 9h et un autre qui ne travaille pas à la même heure seront jugés similaires
  - La même situation à minuit conduit à une faible similarité
- Interprétation probabiliste possible

Données et codage

- Enquêtes Emploi du Temps 1985-86 et 1998-99
- Toute journée avec au moins 10 minutes de travail rémunéré est considérée comme travaillée
- Codage :
  - Journées individuelles, deux états :
    - Travail
    - Non travail
  - Journées conjugales, quatre états :
    - Les deux conjoints travaillent
    - Seul l'homme travaille
    - Seul la femme travaille
    - Les deux conjoints ne travaillent pas

En pratique

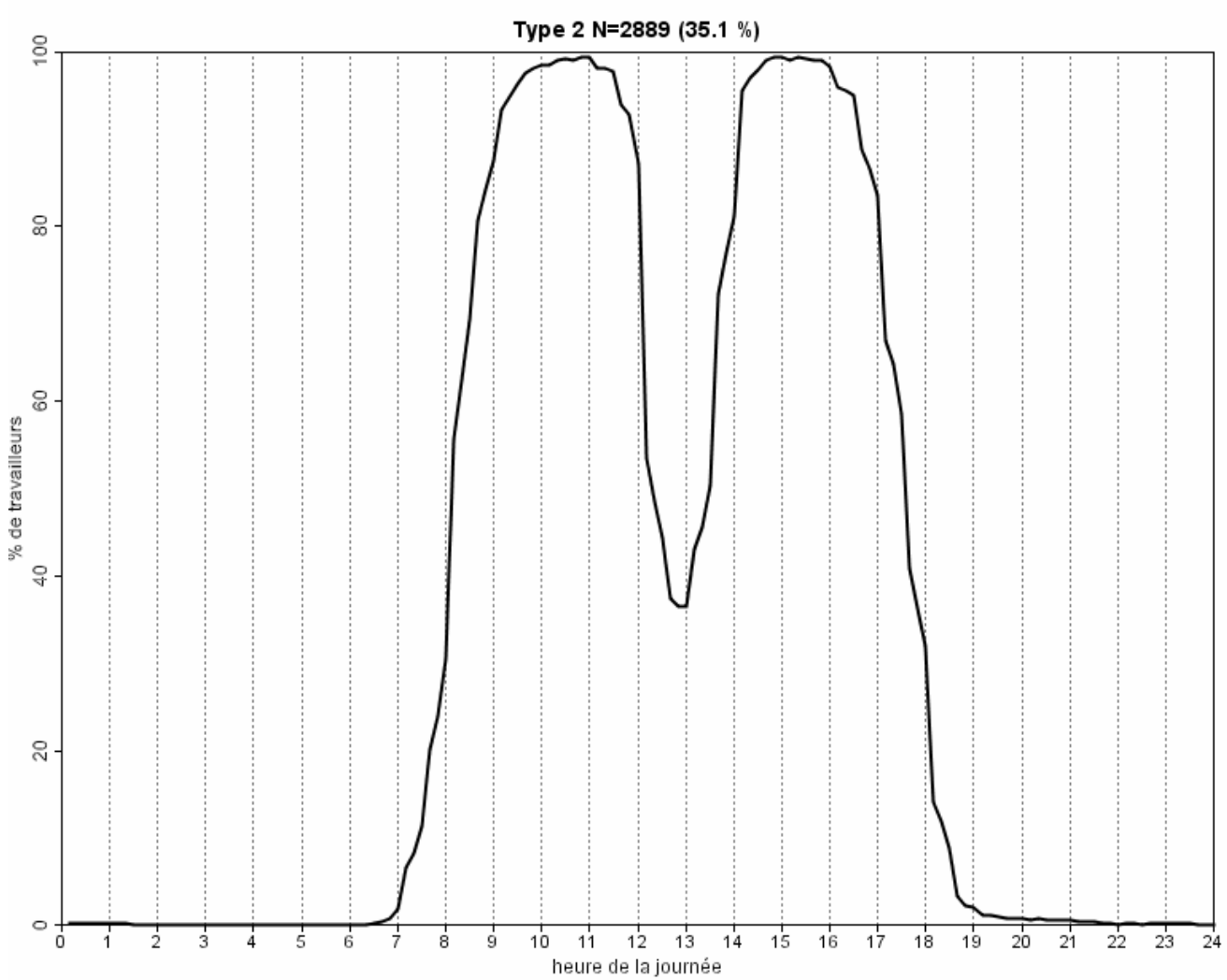
- TDA (freeware) : il existe un module M.A.O. mais les coûts ne peuvent pas varier avec l'échelle de temps
- Programmation dans le logiciel SAS (module de calcul matriciel + langage macro)

+

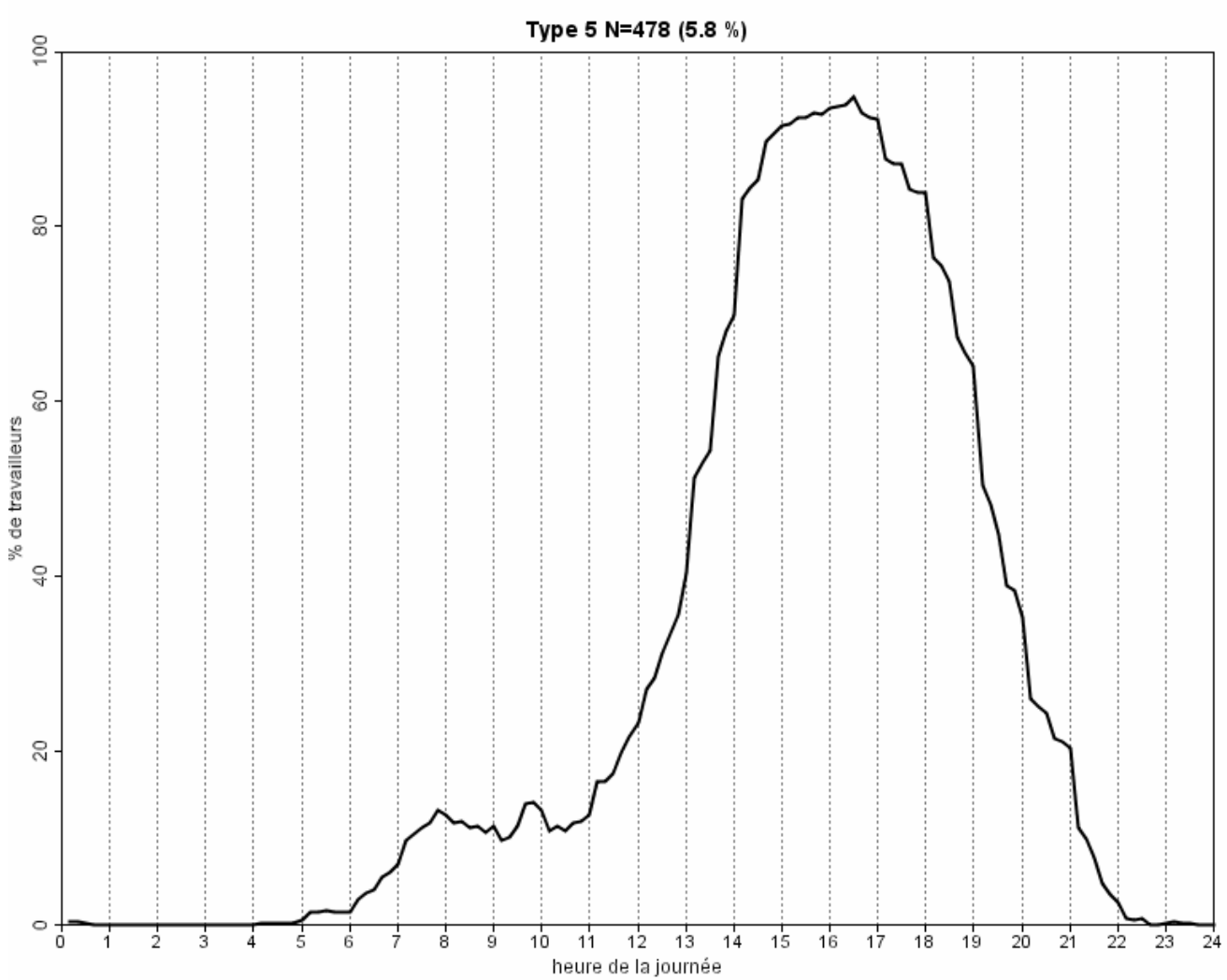
- Classification Ascendante Hiérarchique

(ici méthode Beta-Flexible : méthode de Ward non recommandée)

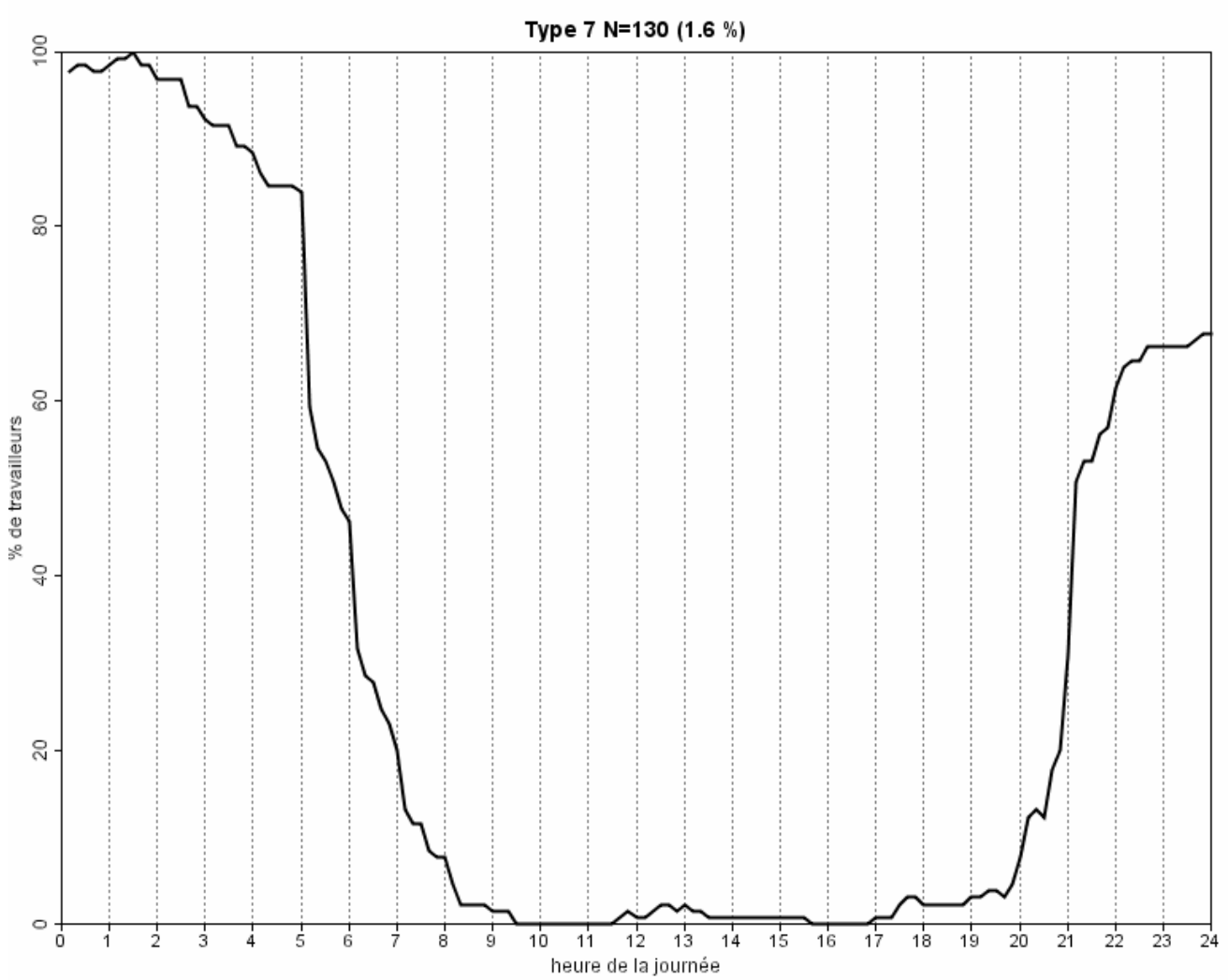
Journée 9 à 5

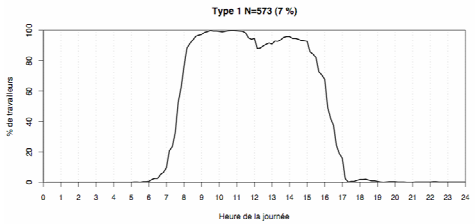

Type 2 N=2889 (35.1 %)
% de travailleurs
0
20
40
60
80
100
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
heure de la journée

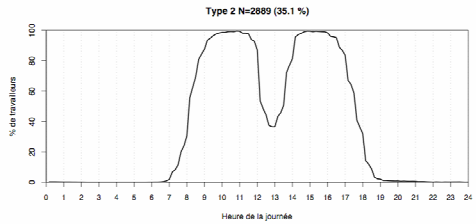
Journée décalée l’après-midi

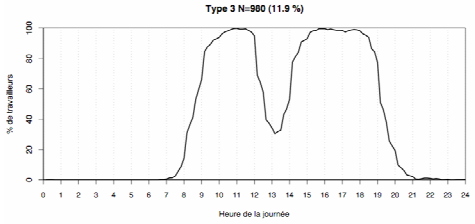


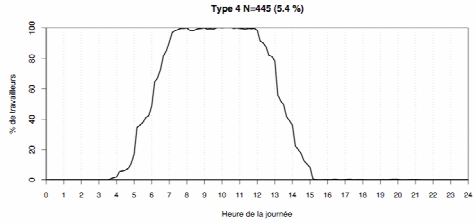
Deux mi-journées de travail de nuit

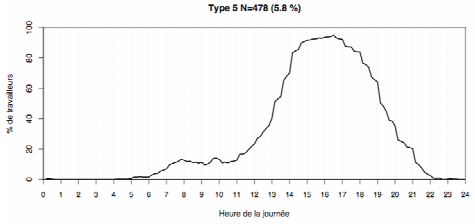


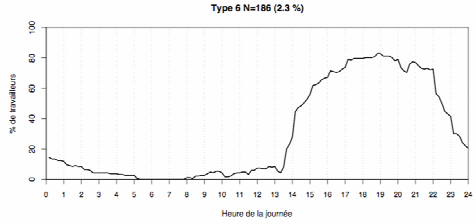

Type 1 N=573 (7 %)
% de travailleurs
Heure de la journée

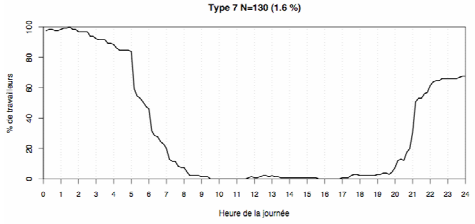

Type 2 N=2889 (35.1 %)
% de travailleurs
Heure de la journée

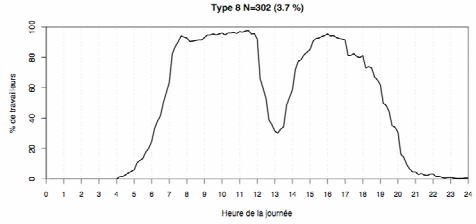

Type 3 N=980 (11.9 %)
% de travailleurs
Heure de la journée

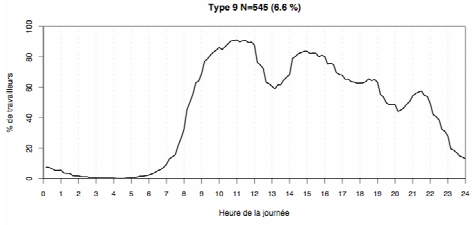

Type 4 N=445 (5.4 %)
% de travailleurs
Heure de la journée

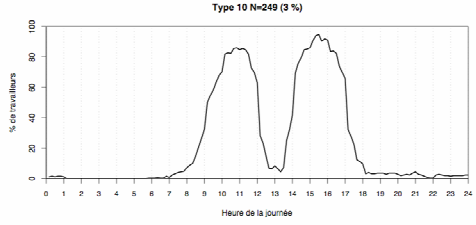

Type 5 N=478 (5.8 %)
% de travailleurs
Heure de la journée

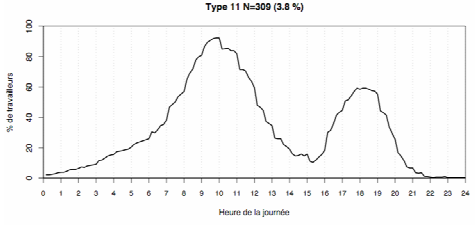

Type 6 N=186 (2.3 %)
% de travailleurs
Heure de la journée

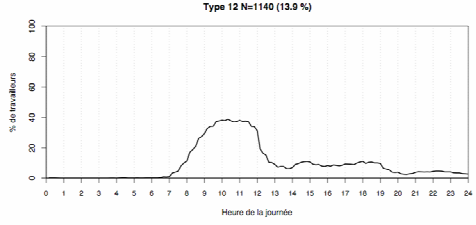

Type 7 N=130 (1.6 %)
% de travailleurs
Heure de la journée


Type 8 N=302 (3.7 %)
% de travailleurs
Heure de la journée

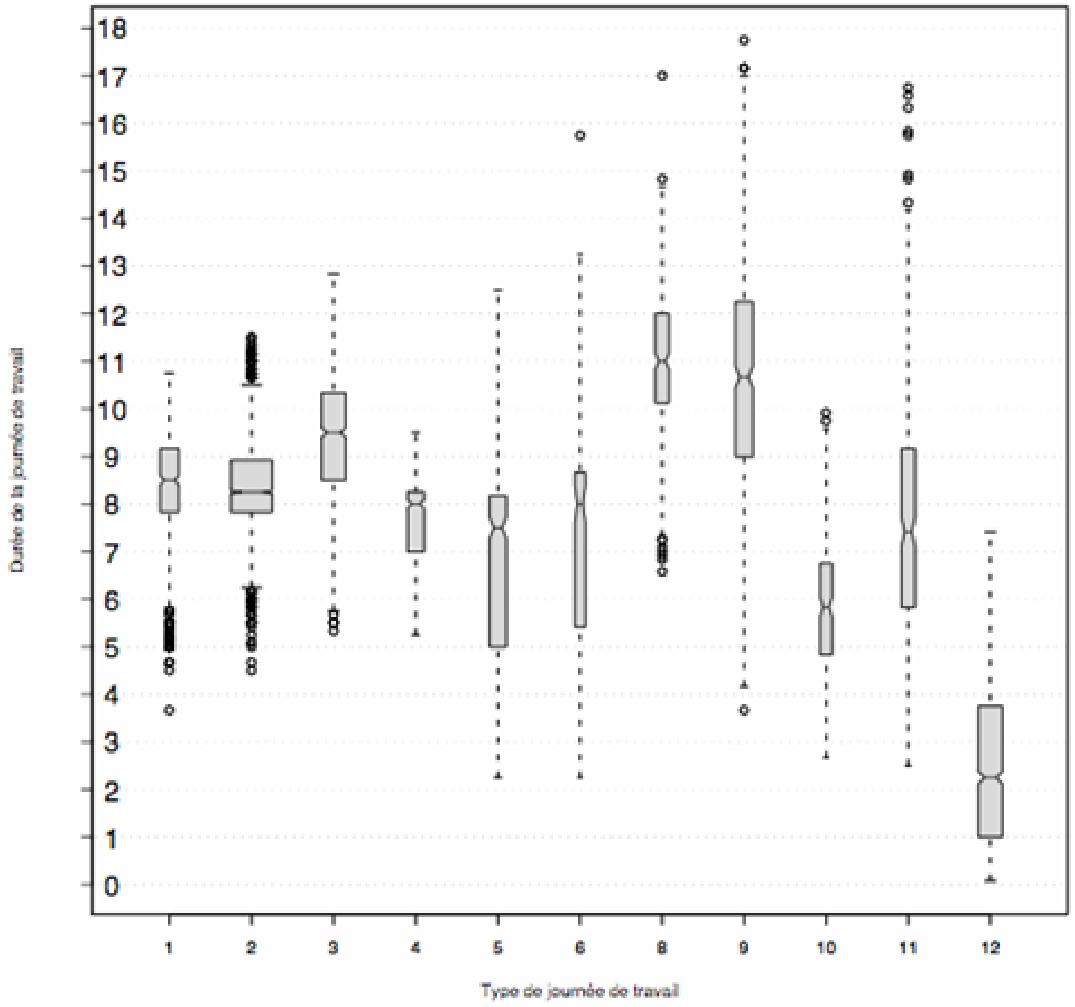

Type 9 N=545 (6.6 %)
% de travailleurs
Heure de la journée


Type 10 N=249 (3 %)
% de travailleurs
Heure de la journée

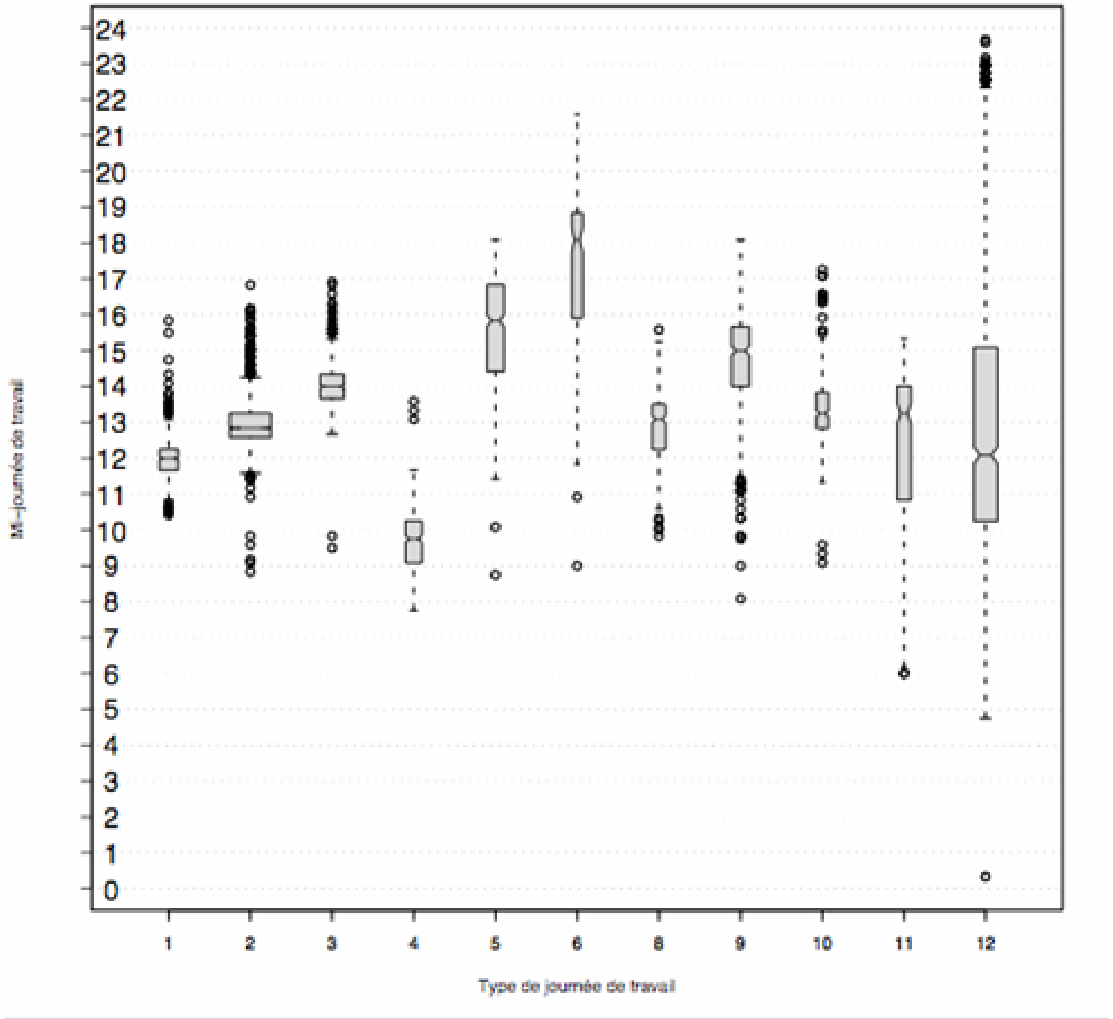

Type 11 N=309 (3.8 %)
% de travailleurs
Heure de la journée


Type 12 N=1140 (13.9 %)
% de travailleurs
Heure de la journée

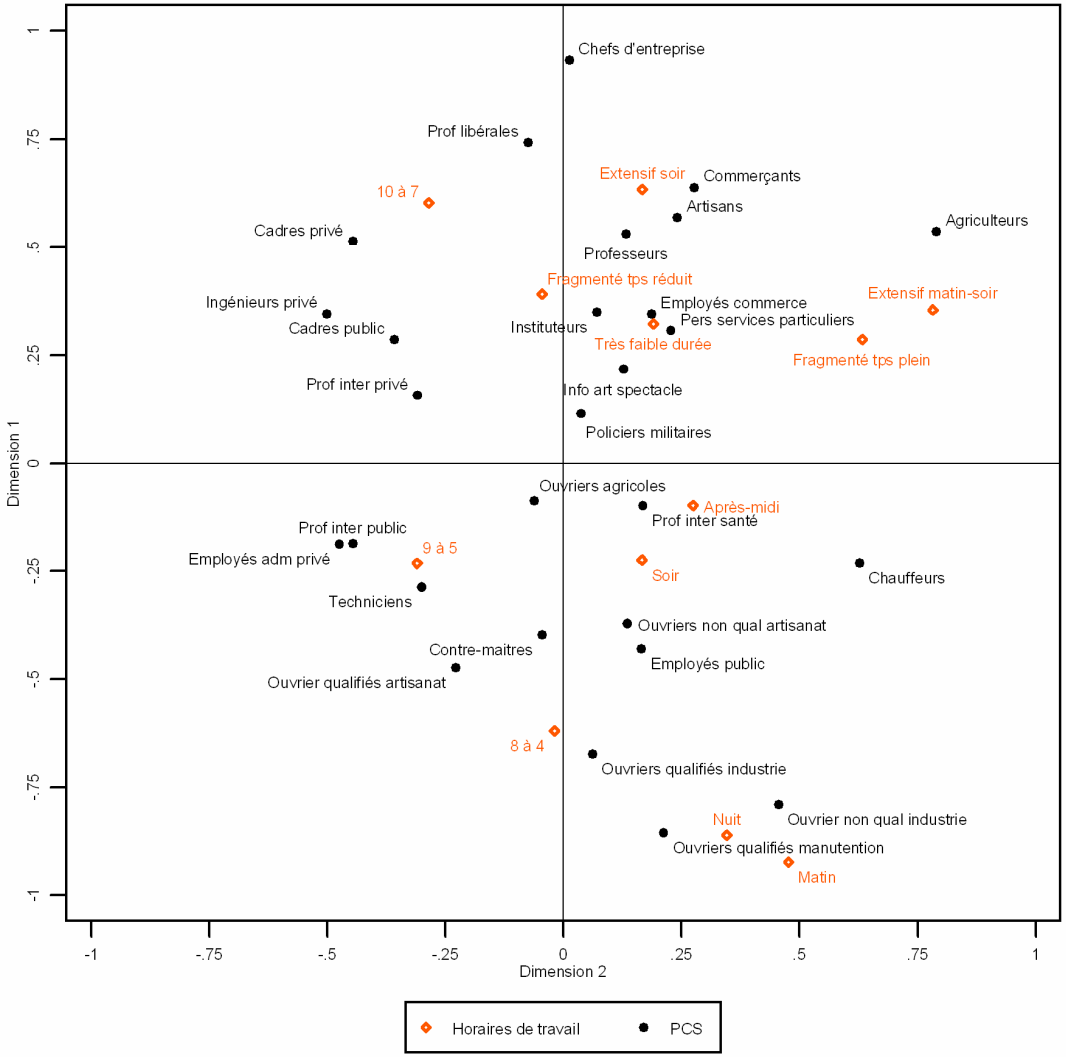
Qualité de la typologie

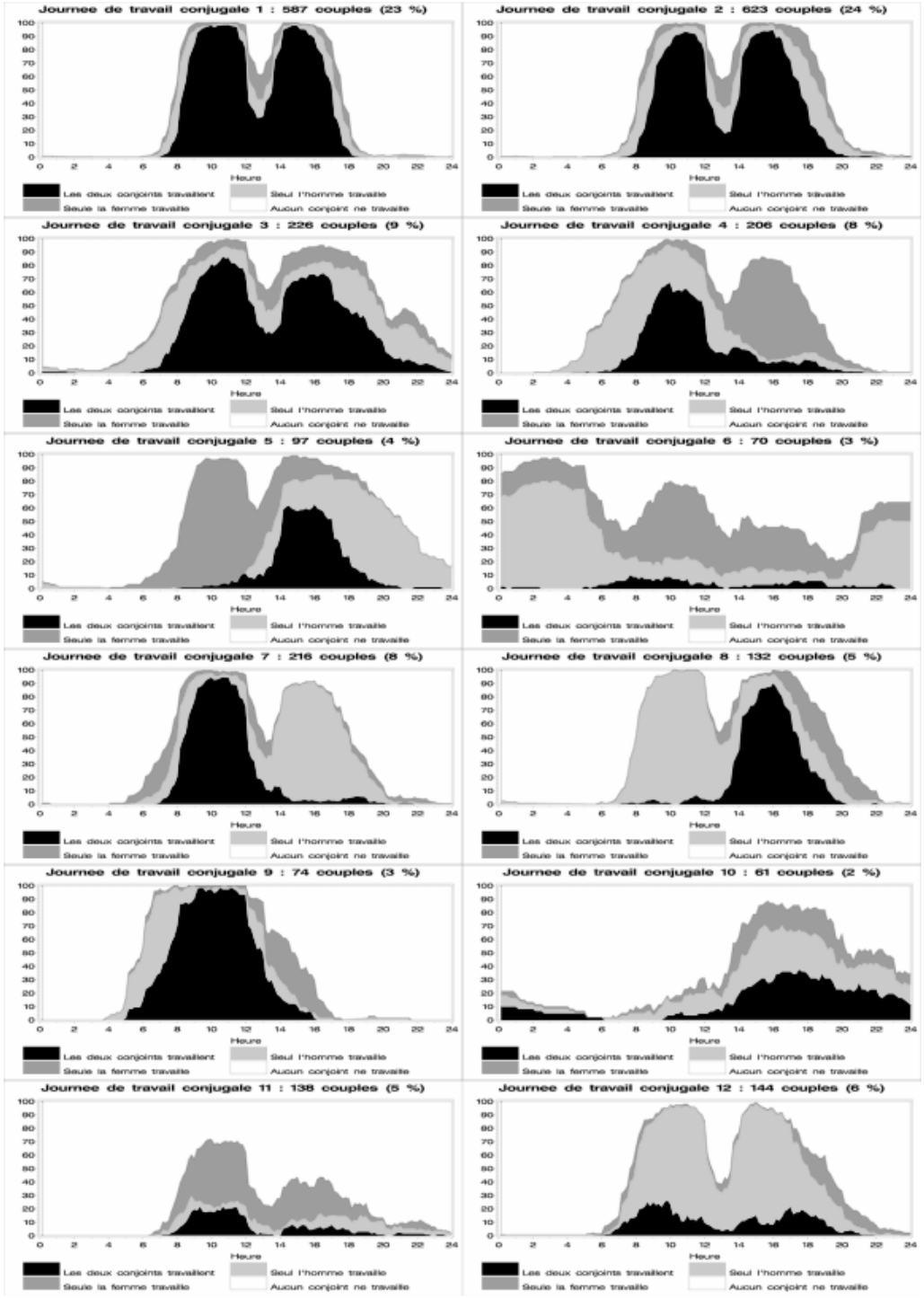


Qualité de la typologie



Utilisation de la typologie




Journee de travail conjugale 1 : 587 couples (23 %)
Journee de travail conjugale 2 : 623 couples (24 %)
Journee de travail conjugale 3 : 226 couples (9 %)
Journee de travail conjugale 4 : 206 couples (8 %)
Journee de travail conjugale 5 : 97 couples (4 %)
Journee de travail conjugale 6 : 70 couples (3 %)
Journee de travail conjugale 7 : 216 couples (8 %)
Journee de travail conjugale 8 : 132 couples (5 %)
Journee de travail conjugale 9 : 74 couples (3 %)
Journee de travail conjugale 10 : 81 couples (2 %)
Journee de travail conjugale 11 : 138 couples (5 %)
Journee de travail conjugale 12 : 144 couples (6 %)
Heure
Les deux conjoints travaillent
Seule la femme travaille
Seul l'homme travaille
Aucun conjoint ne travaille

Résumé des journées de travail conjugales

Type de journée de travail conjugale		1985				1998			
		Taille	Durée trav. Hommes	Durée trav. Femmes	Synchronie (en %)	Taille	Durée trav. Hommes	Durée trav. Femmes	Synchronie (en %)
Standard	Double journée standard	49	08:36	07:54	72.8	44	09:02	08:20	70.1
Atypique	Double journée extensive	8	11:04	08:51	57.7	10	11:15	09:36	57.8
	Double journée décalée	14	07:05	07:07	23.9	15	07:06	07:16	23.3
	Journée partiellement travaillée par les femmes	12	08:54	04:49	36.9	16	09:07	05:18	36.5
	Irrégulière et/ou faible durée	17	05:47	04:15	27.0	15	06:45	04:46	31.9
Total		**100**	**08:09**	**06:53**	**52.4**	**100.0**	**08:39**	**07:18**	**50.5**

Autres applications

- Principale application : analyse de carrières
  - Abbott et Hrycak (1990) : analyse des carrières des musiciens
  - Halpin et Chan (1998) : analyse des trajectoires sociales
  - Blair-Loy (1999) : analyse des carrières des femmes cadres dans la finance
  - Han et Moen (1999) : analyse des fins de carrières (retraite)
- Autres applications :
  - 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 : Saint Pol (2003) pour analyser les différents modes d'inscriptions du dîner dans la soirée

Conclusion

- Mode d’emploi des M.A.O. :
  - Données séquentielles
  - Codage des séquences cohérent avec la problématique
  - Choix des coûts
  - Application informatique de l’analyse :
    - TDA et Optimize (Abbott)
    - Clustal et Emboss : coûts adaptés aux pratiques courantes de la biologie
  - Algorithme de classification ou tout autre méthode capable d’exploiter une matrice de distance (multidimensional scaling)

Conclusion

- Séquences en biologie et sciences sociales
- Coûts plutôt que les opérations
- Les coûts(re-)posent la question du temps pour les sciences sociales
- Le temps n’est pas qu’un support de classement
- Les coûts doivent pouvoir varier avec le temps
- Problème de la mise en œuvre