Using optimal matching to analyze the timing of daily life

Laurent Lesnard
Observatoire sociologique du changement (Sciences-po & CNRS)
Laboratoire de sociologie quantitative (Statistics France)

Beyond time-budgets

J. Gershuny and Oriel Sullivan (1998)

“The sociological uses of time-use diary analysis”

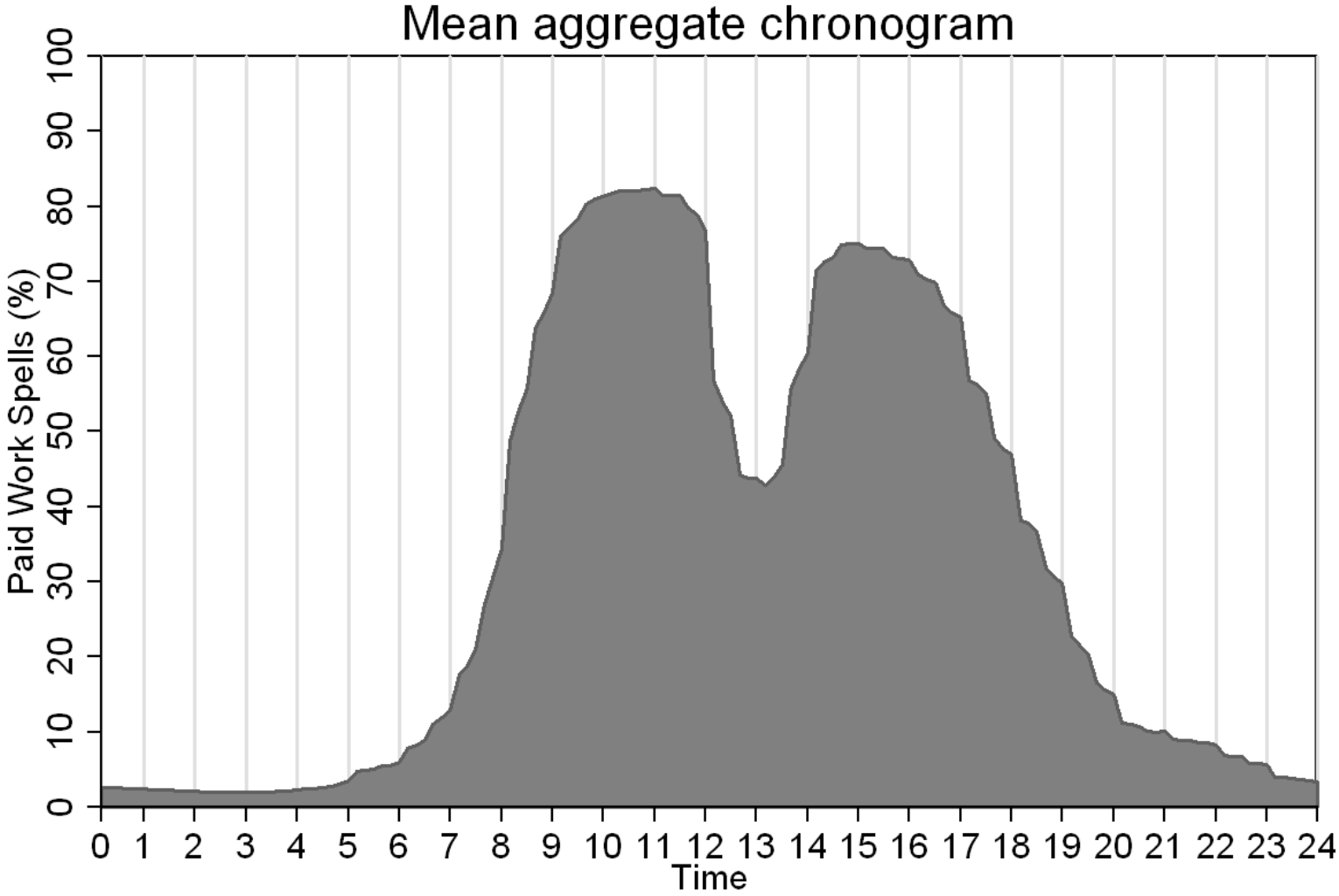
European Sociological Review

– Time-use analyses rely almost exclusively on time-budgets disregarding the sequential dimension of daily life

– Statistical tools not adapted

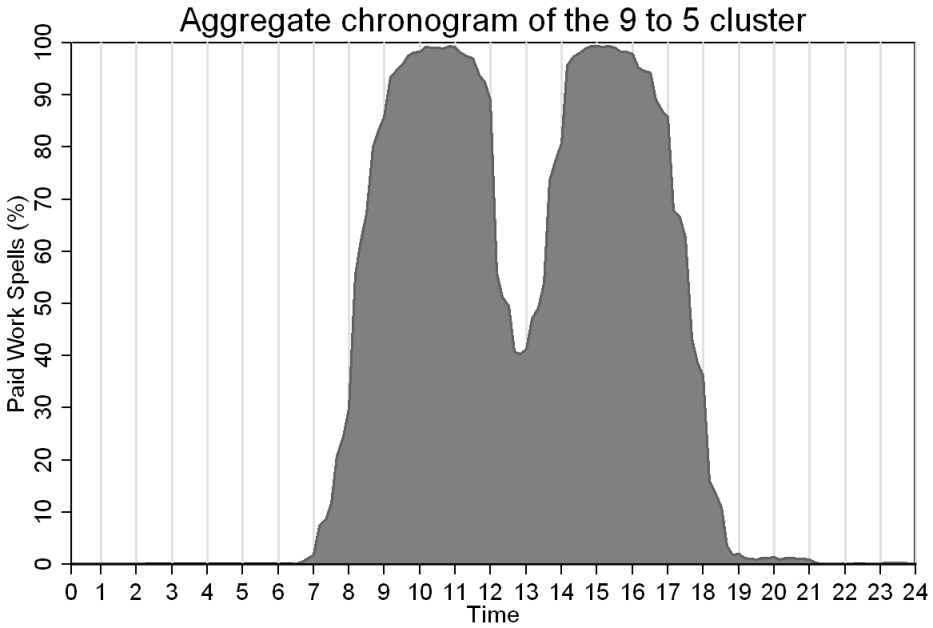
– New tools are needed

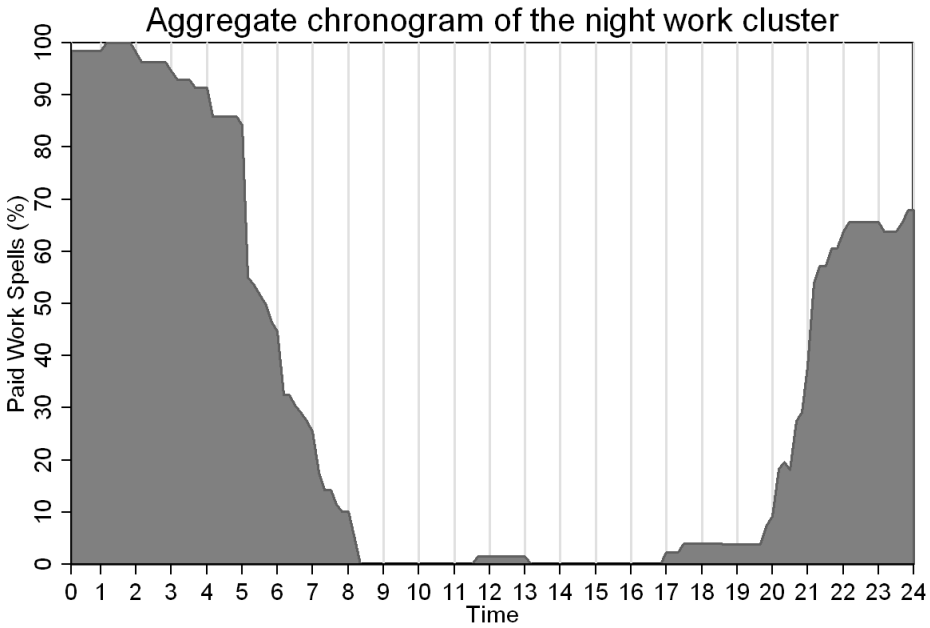
When the mean situation doesn't mean much...



Proportion of workers for each time slot of the 1999 French time-use survey

...while an empirical typology does





9 to 5 workers (33.9%)

Night workers (1.6%)

Proportion of workers for each time slot of the 1999 French time-use survey

Outline of the presentation

1. Optimal Matching (OM) in a (tiny) nutshell
2. How OM is used in biology
3. Why we cannot do the same and what we should do

1. Optimal Matching (OM) in a (tiny) nutshell

- Distance measure between sequences
- Methodology:
  1. For each pair of sequences, one sequence is transformed so that it matches the other one
  2. Three transformations are used: insertion, deletion, and substitution
  3. Each edit operation has a cost
  4. The distance between sequences is the minimal transformation cost
  5. The cluster analysis of the distance matrix gives the empirical typology

Example

A : X – Y – Y – Y

B : X – X – X – X – Y

- One solution:

A : X – X – X – X – X – X – Y

B : X – X – X – X – Y

- Another one:

A : X – X – Y – Y – Y

B : X – X – X – X – Y

Example

A : X – Y – Y – Y

B : X – X – X – X – Y

- One solution:

A : X – X – X – X – X – X – Y

B : X – X – X – X – Y

3 insertions

2 deletions

- Another one:

A : X – X – X – X – Y

B : X – X – X – X – Y

1 insertion

2 substitutions

2. How OM is used in biology

How OM is used in biology

- Aim: transfer information between known and unknown DNA or proteins
- Sequence analysis is used in biology as an approximation to avoid costly and lengthy experimentations
- The three edit operations are not reproducing any bio-chemical phenomena
- Costs are determined empirically and not theoretically

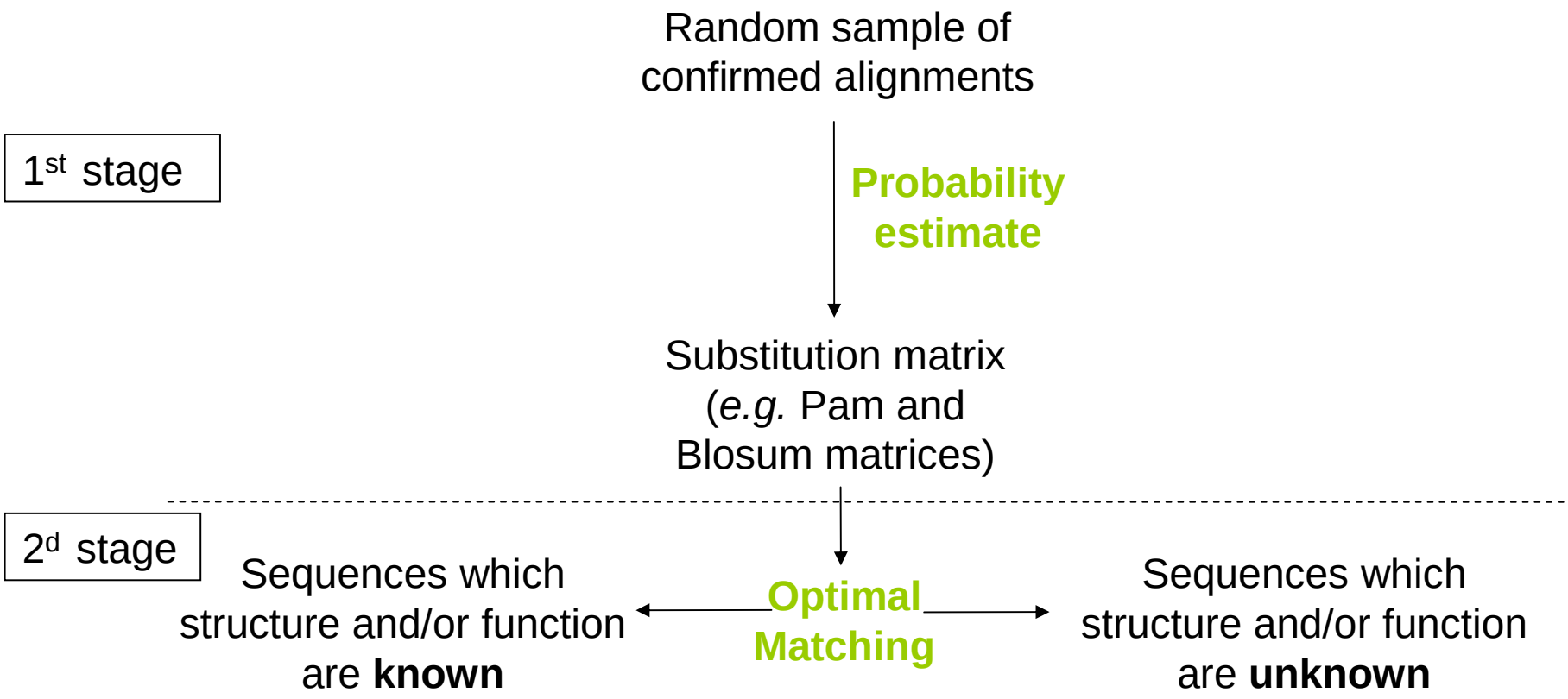
Insertion and deletion costs in biology

“In practice, people choose [insertion and deletion] costs empirically [read ‘not scientifically’] once they have chosen their substitution scores.”

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

Substitution costs in biology

Aim: transfer information between known and unknown DNA or proteins



3. Why we cannot do the same and what we should do

What we can learn from how OM is used in biology

- The three edit operations have no theoretical meaning
- The success of OM in biology is the result of ingenious substitution costs
- Substitution costs are
  - interpreted as the probability that sequences are evolutionary related
  - empirically determined

Why we cannot do the same (use the same software)

- Different “matter”

  sequences = events + time

- Different purpose

  Identify collective rhythms

What we should do

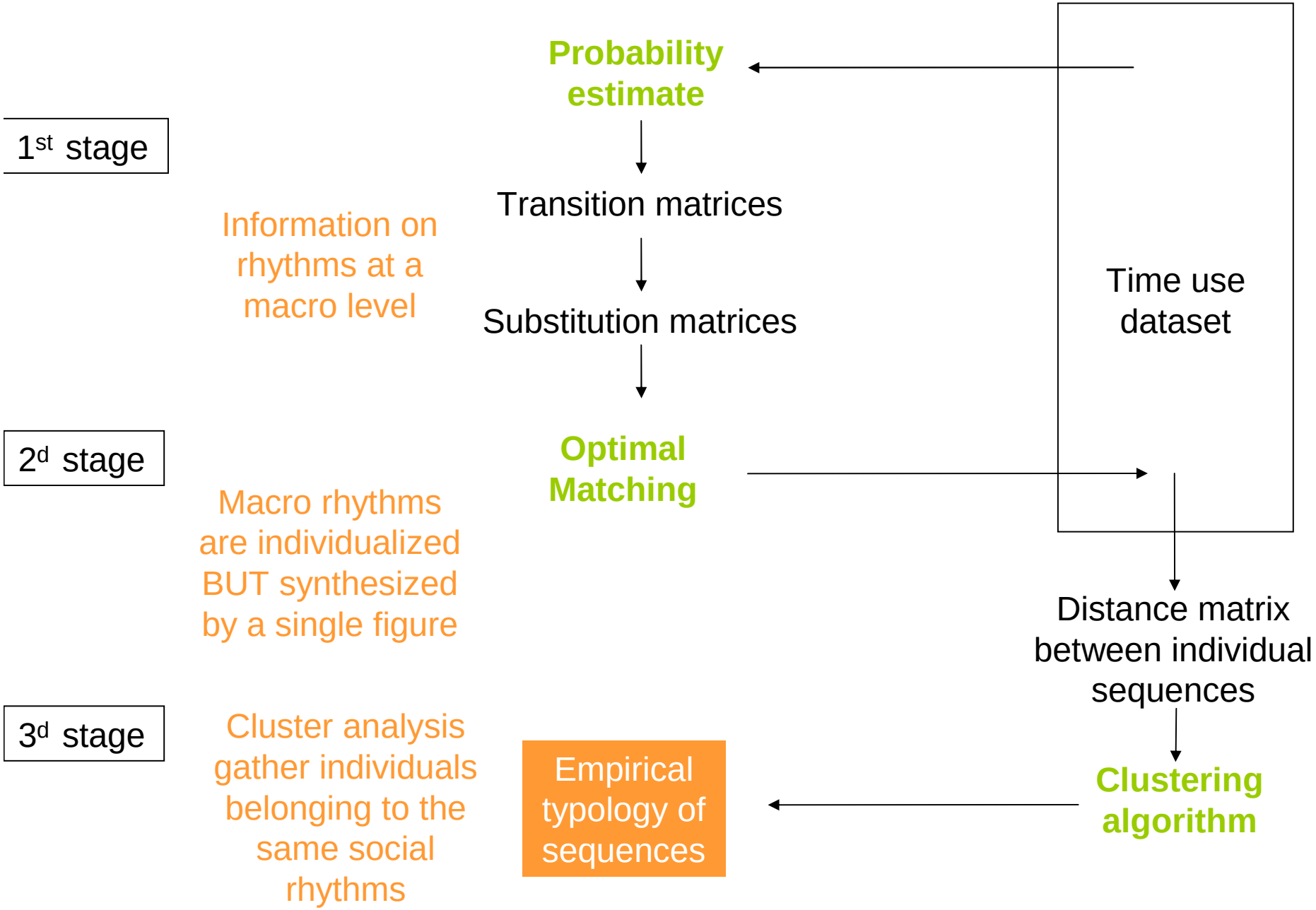
- Insertion and deletion operations should not be used since they distort the timing of sequences
- Substitution operations preserve the timing of sequences
- Substitution costs must capture the probability that two events belong to the same collective rhythm

Collective rhythm = Transition matrix

		07:30 AM		
		Working	Non working	
07:20 AM	Working	92.3 %	7.7%	100,0%
	Not working	1.0%	99.0%	100,0%

1999 French Time Use Survey

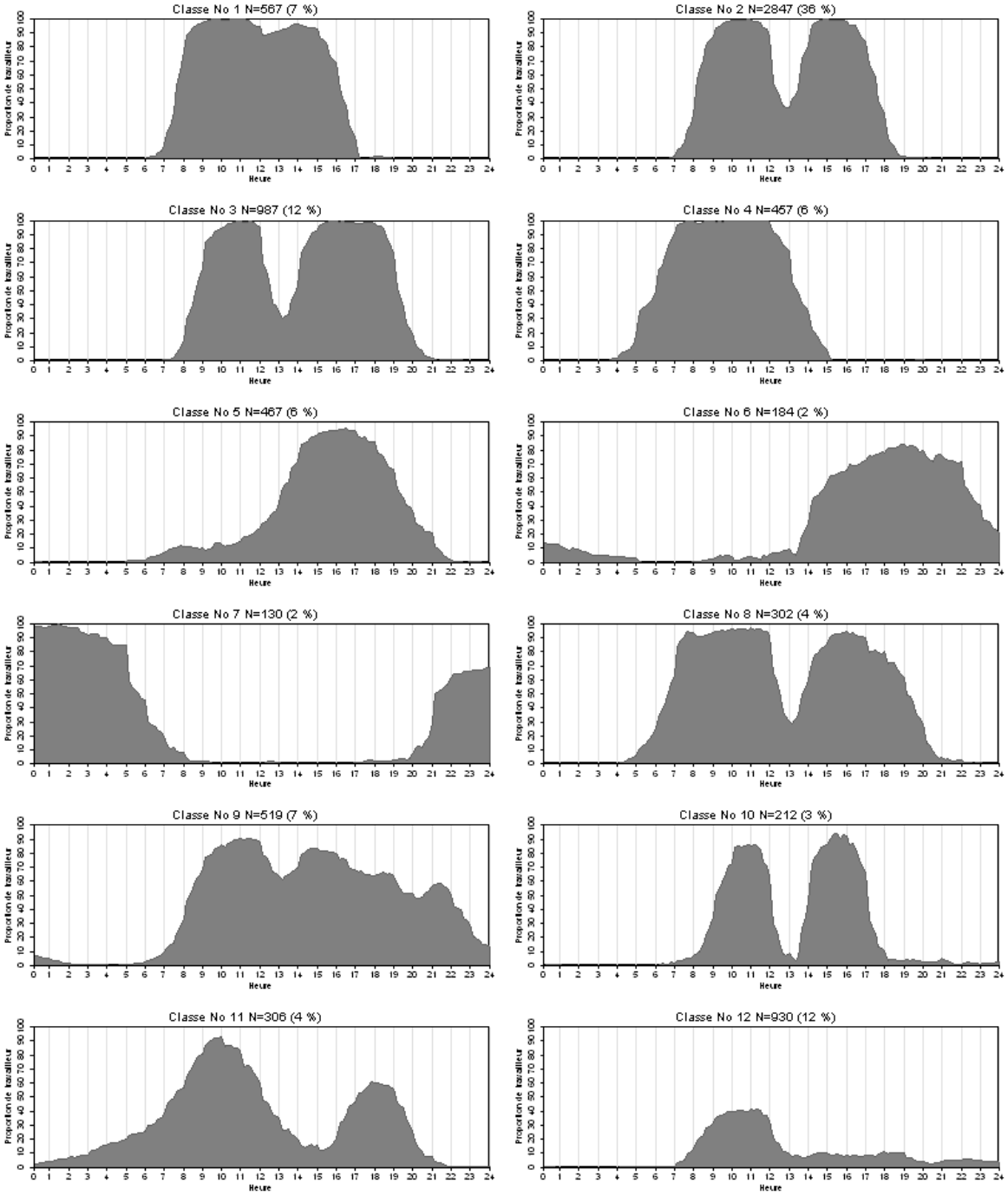
Identifying social rhythms



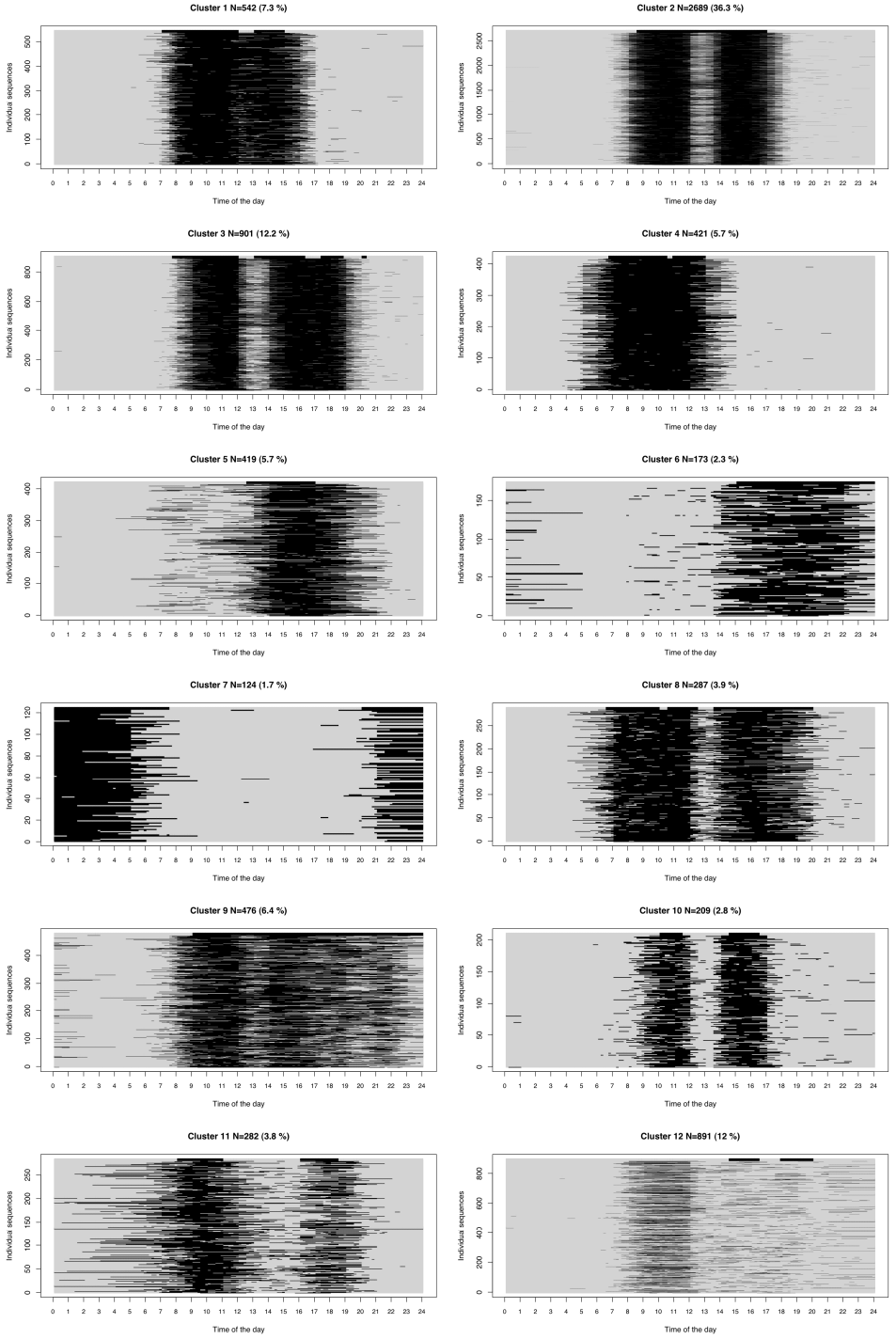
Example: The timing of paid work

- 1986 and 1999 French time use surveys
- Two states: working vs. not working
- Substitution costs
  - Are high when the probability of changing states are low (low transition rate): e.g. at midnight, the transition from work to non work or from non work to work is low, indicating two distinct rhythms
  - Are low when the probability of changing states are high (high transition rate): e.g. at 9 pm. There is a high transition rate between work and non work, indicating that at that time these two events may belong to the same rhythm
- 12 clusters

Aggregage chronogram of the twelve types of work days



Individual chronogram of the twelve types of work days



Software

- Sas macro
- Stata plugin (10 times faster…)
- Clustering: flexible WPGMA or, better, flexible UPGMA
  - Stata
    - Version 9: old algorithms
    - Version 10 (next version): both
  - Sas & Clustan Graphics: only flexible WPGMA
  - R: both

Conclusion

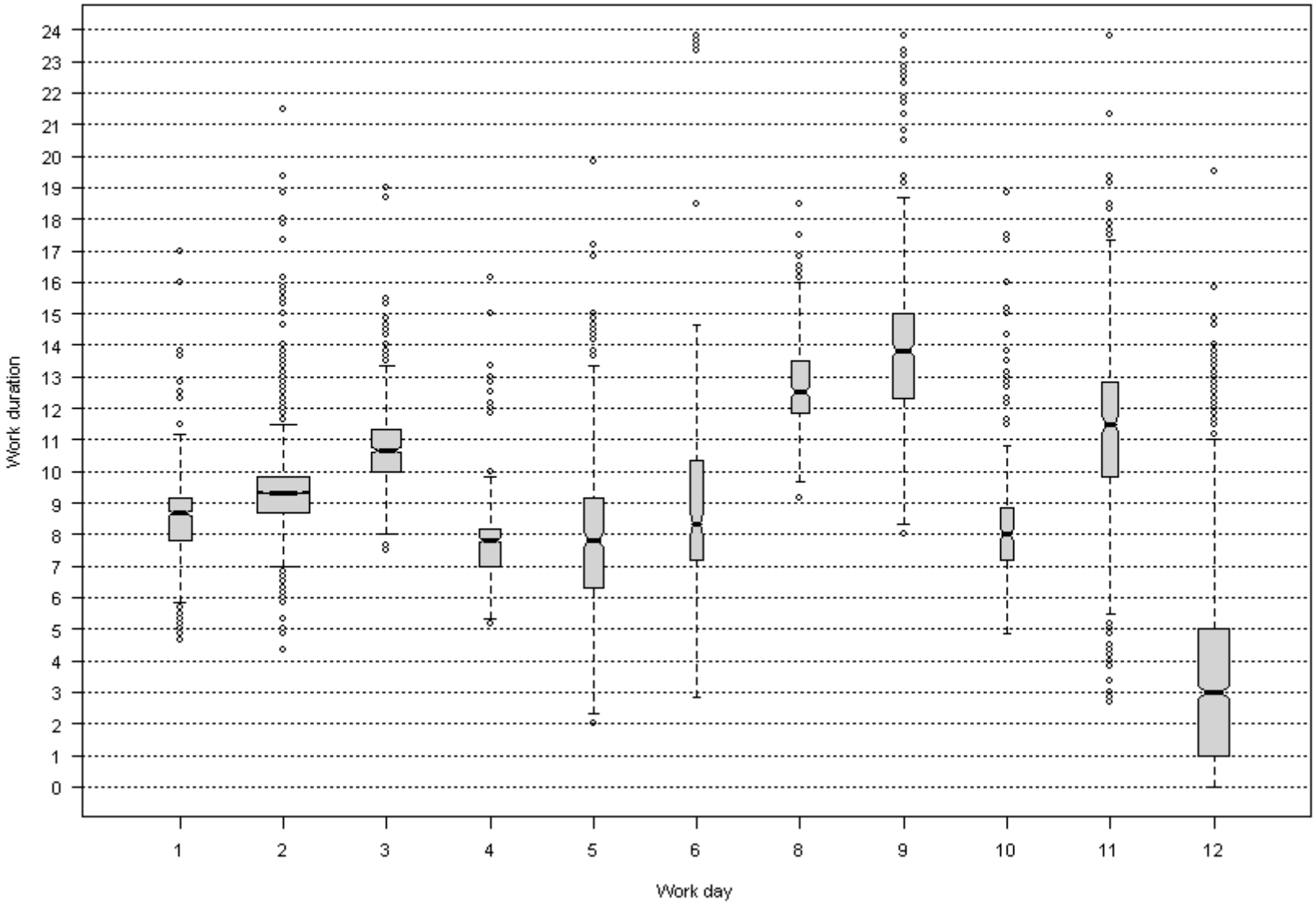
- Give optimal matching a try!
- Interested in an international comparison of work schedules?
- E-mail:

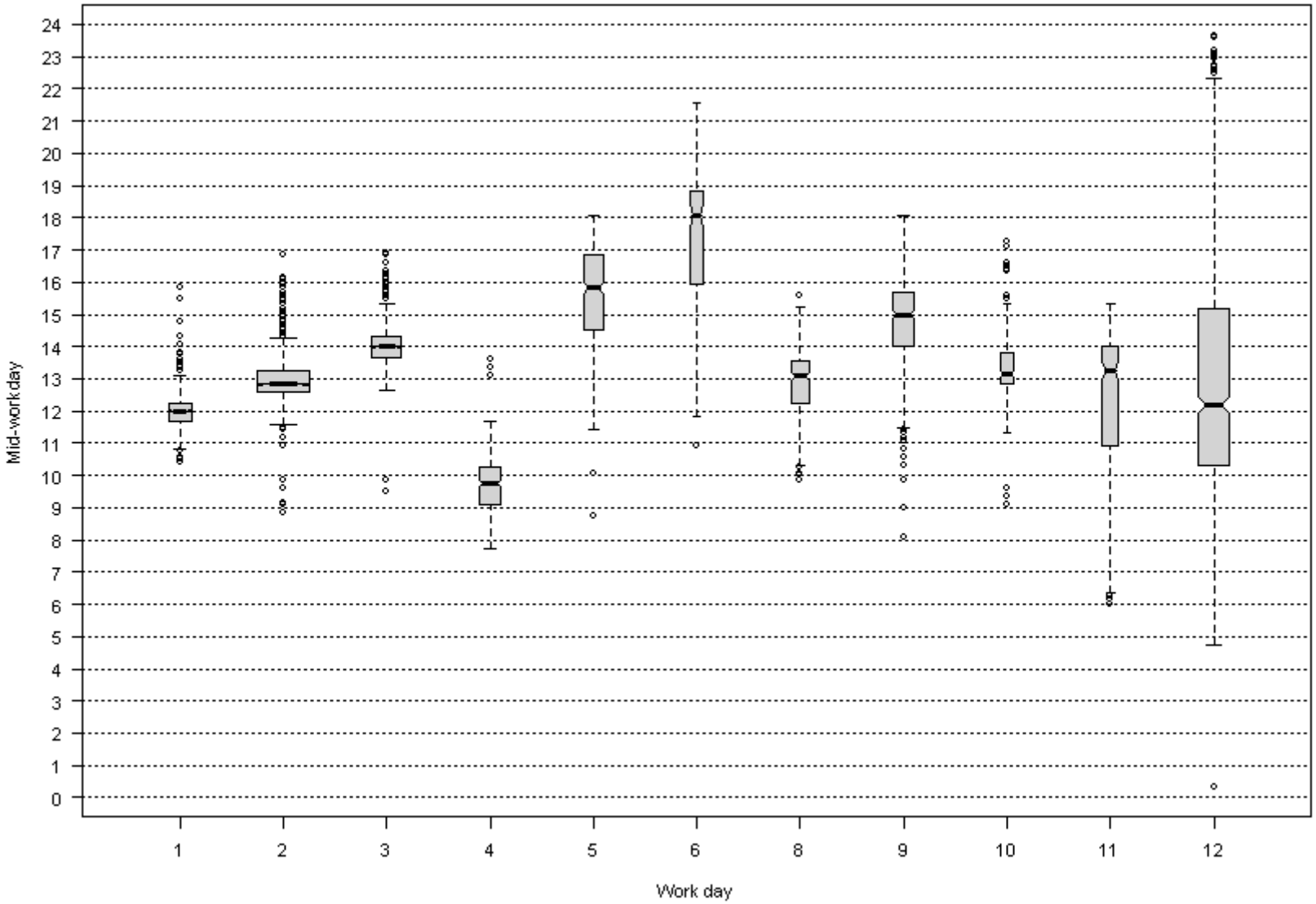
laurent.lesnard@sciences-po.fr

- Plugin & working paper:

http://laurent.lesnard.free.fr

Appendices


Work duration
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
1
2
3
4
5
6
8
9
10
11
12
Work day


Mid-work day
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
1
2
3
4
5
6
8
9
10
11
12
Work day

Stata plugin v0.6 (1)

- Capabilities:
  - Limitations (Stata's) if distance matrix is analyzed within Stata
  - None if Stata is just used as a frontend for the C program behind the plugin, and distance matrix is reduced with external cluster program
- Syntax

  seqcomp seq1-seq144 if subsample==1 [iw=weight] using "c:\temp", id(idseq)

Stata plugin v0.6 (2)

- Output : 3 files
  1. Distance list ready to be analyzed by Clustan Graphics
  2. File with the series of transition and substitution costs matrices
  3. File to match the internal id with the id of the dataset
- Future developments:
  - Further integration into Stata when the next version (10) is released (e.g. direct use of cluster analysis)
  - Individual and aggregate chronograms
  - Complementarity with geometric data analysis applied to sequences

Stata plugin output

Distance list

2 1 21.461657
3 1 10.937172
3 2 10.524483
4 1 19.102806
4 2 40.564461
4 3 30.039978
5 1 10.893203
5 2 10.568453
5 3 7.303132
5 4 29.996008
6 1 38.929619
6 2 39.605457
6 3 35.339550
6 4 50.682537

Transition and substitution matrices

Transition matrix for 44 -> 45
0.922689 0.077312
0.010072 0.989927

Substitution costs for 44
0.000000 3.974521
3.974521 0.000000

Transition matrix for 45 -> 46
0.995668 0.004332
0.002153 0.997846

Substitution costs for 45
0.000000 3.977141
3.977141 0.000000