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...

Mean aggregate chronogram

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 : \textcolor{red}{X} – \textcolor{red}{X} – \textcolor{red}{X} – X – \textcolor{red}{X} – \textcolor{red}{X} – Y
  B : X – X – X – X – Y

• Another one:
  A : \textcolor{green}{X} – X – Y – Y – Y
  B : X – X – X – X – X – Y
Example

A : X – Y – Y – Y
B : X – X – X – X – Y

• One solution:
  B : X – X – X – X – X – Y

• Another one:
  A : X – X – X – X – Y
  B : X – X – X – X – X – Y

3 insertions 2 deletions
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

Random sample of confirmed alignments

Probability estimate

Substitution matrix (e.g. Pam and Blosum matrices)

1st stage

2nd stage

Sequences which structure and/or function are known

Optimal Matching

Sequences which structure and/or function are unknown
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

<table>
<thead>
<tr>
<th></th>
<th>Working</th>
<th>Non working</th>
</tr>
</thead>
<tbody>
<tr>
<td>Working</td>
<td>92.3%</td>
<td>7.7%</td>
</tr>
<tr>
<td>Not working</td>
<td>1.0%</td>
<td>99.0%</td>
</tr>
</tbody>
</table>

07:20 AM

07:30 AM

1999 French Time Use Survey
Identifying social rhythms

1st stage

- Information on rhythms at a macro level

2nd stage

- Macro rhythms are individualized BUT synthesized by a single figure

3rd stage

- Cluster analysis gather individuals belonging to the same social rhythms

Time use dataset

- Transition matrices
- Substitution matrices

Optimal Matching

- Probability estimate

Distance matrix between individual sequences

Empirical typology of sequences

Clustering algorithm
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
Aggregate 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.lesnardin.free.fr
Appendices

Quality
Stata plugin
Stata plugin output
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
  ```stata
  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
## Distance list

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>21.461657</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>10.937172</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>10.524483</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>19.102806</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>40.564461</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>30.039978</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>10.893203</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>10.568453</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>7.303132</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>29.996008</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>38.929619</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>39.605457</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>35.339550</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>50.682537</td>
</tr>
</tbody>
</table>

## 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 |

**Transition matrix for 45 -> 46**

|   | 0.995668 | 0.004332 | 0.002153 | 0.997846 |

**Substitution costs for 45**

|   | 0.000000 | 3.977141 |