EXPLORING THE LIFE COURSE:

a theoretical and empirical comparison of methods

Nicolas Robette, Università Bocconi

Life course quantitative studies

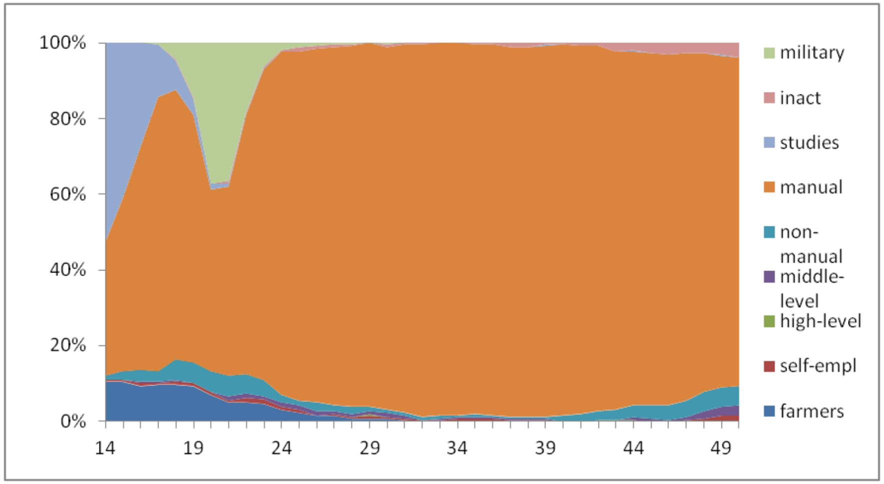
- ***Atomistic approach:***
  - Unit of analysis = event
  - Modelisation of transition likelihoods / durations
  - Stochastic, explanatory

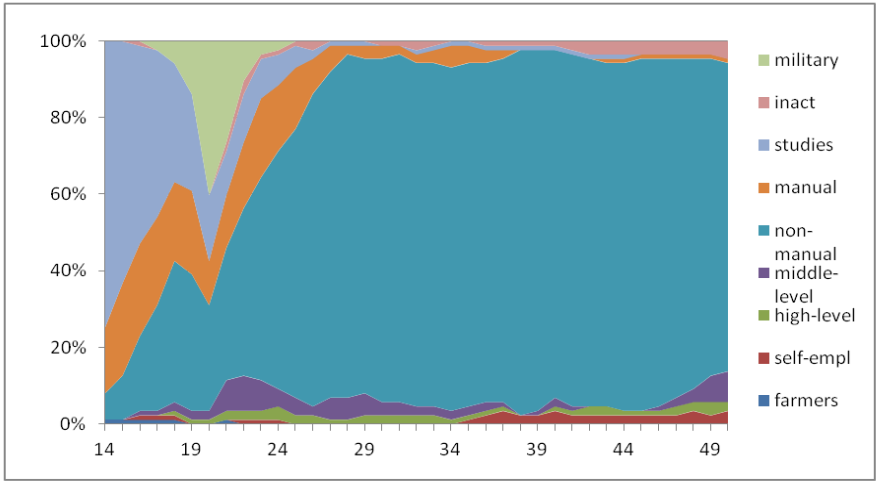
- ***Holistic approach:***
  - Unit of analysis = trajectory « as a whole »
  - Identification of ideal-types, patterns
  - Exploratory, descriptive

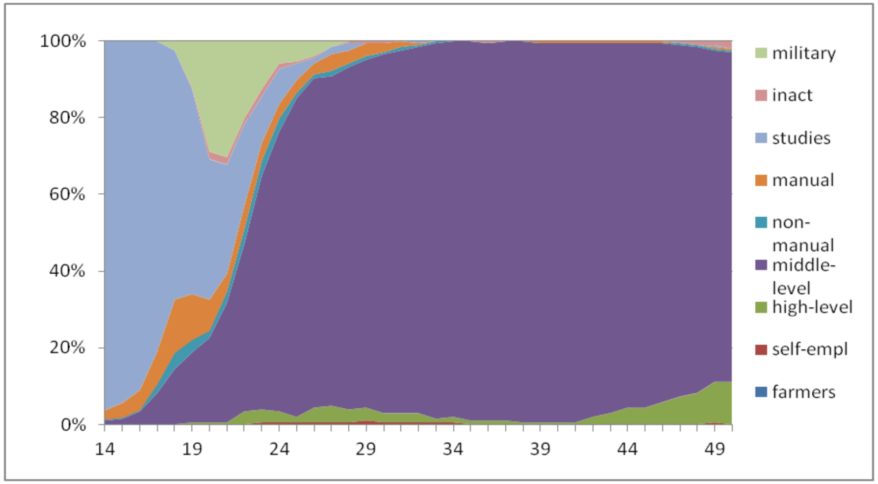
A short example: data

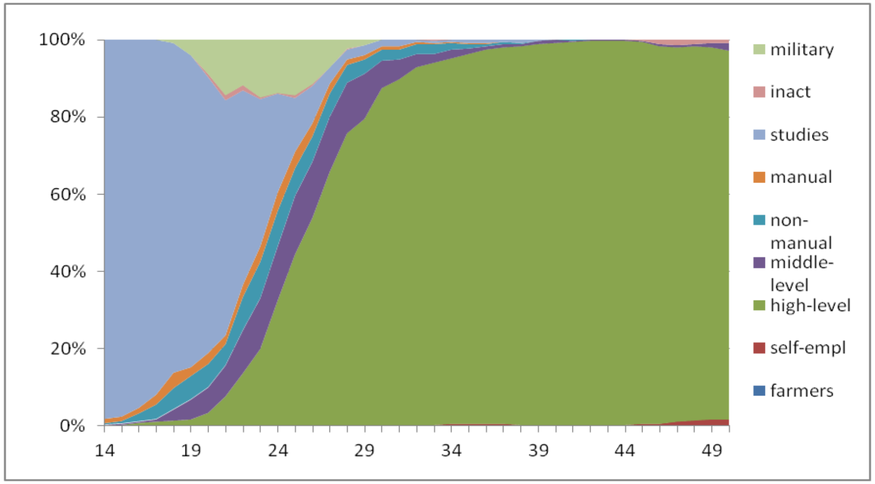
- Event-history survey *Biographies et entourage* (INED, 2001)
- 1421 men's occupational trajectories
- 37 years, from 14 to 50
- 9 states:
  - *farmers, self-employed, higher-level intellectual occupations, intermediate occupations, clerical and sales workers, manual workers,*
  - *students,*
  - *military conscripts,*
  - *other inactive*

A short example: patterns (1)

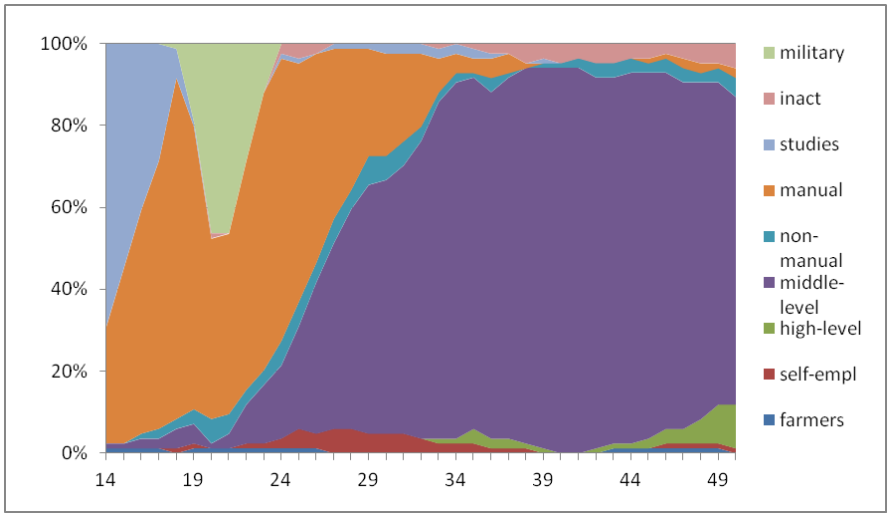


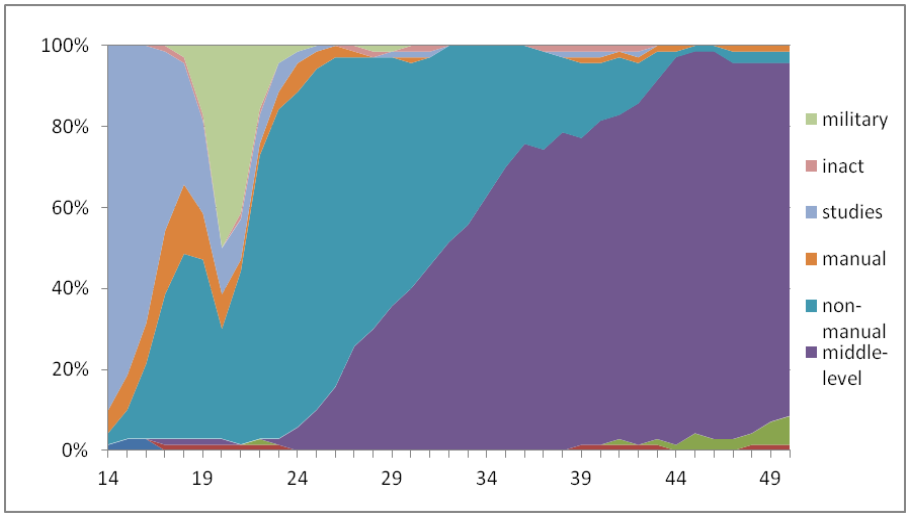


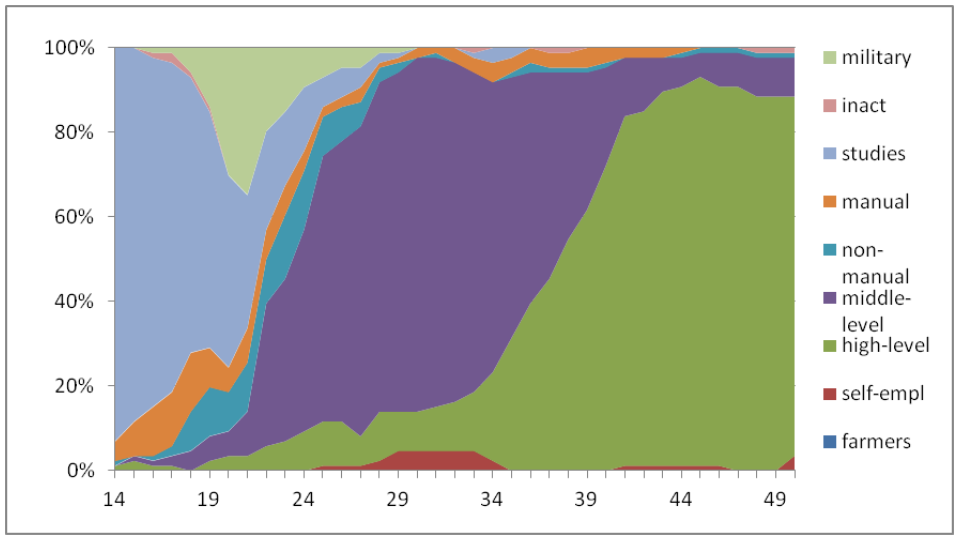




A short example: patterns (2)







Holistic approach

- ***CA / geometric methods:***
  - Choice of the career's coding
  - Correspondence analysis <-> type of distance
  - Clustering (or MDS…)

- ***Sequential / algorithmic methods:***
  - Career as a sequence of states
  - Pairwise dissimilarities metrics
  - Clustering (or MDS…)

Example of trajectory

School-to-work transition:

S = studies

U = unemployment

J = job

18	19	20	21	22	23	24	25
S	S	S	U	J	J	J	J

CA methods(1)

18	19	20	21	22	23	24	25
S	S	S	U	J	J	J	J

- **Disjonctif coding**

18S	18U	18J	…	25S	25U	25J
1	0	0	…	0	0	1

→ PCA: euclidian distance

→ CA: khi-2 distance

CA methods(2)

18	19	20	21	22	23	24	25
S	S	S	U	J	J	J	J

- **Summarized calendar**

  = *Qualitative Harmonic Analysis*

18-20 S	18-20 U	18-20 J	21-25 S	21-25 U	21-25 J
1	0	0	0	0,2	0,8

→ CA (khi-2 distance)

CA methods(3)

18	19	20	21	22	23	24	25
S	S	S	U	J	J	J	J

- **Indicators**
  - Durations

S	U	J
3	1	4

  - Transitions

SS	SU	SJ	US	UU	UJ	JS	JU	JJ
2	1	0	0	0	1	0	0	3

  - Number of spells, duration before first spell in a specific state...

→ PCA (euclidian distance)

Sequence analysis

- Individual trajectories are built as sequences of states

- Computation of pairwise dissimilarities
  (algorithms = Optimal Matching Analysis,…)

  → Distance matrix

  → Clustering *(HCA…; or reduction by MDS)*

  → Typology of trajectories

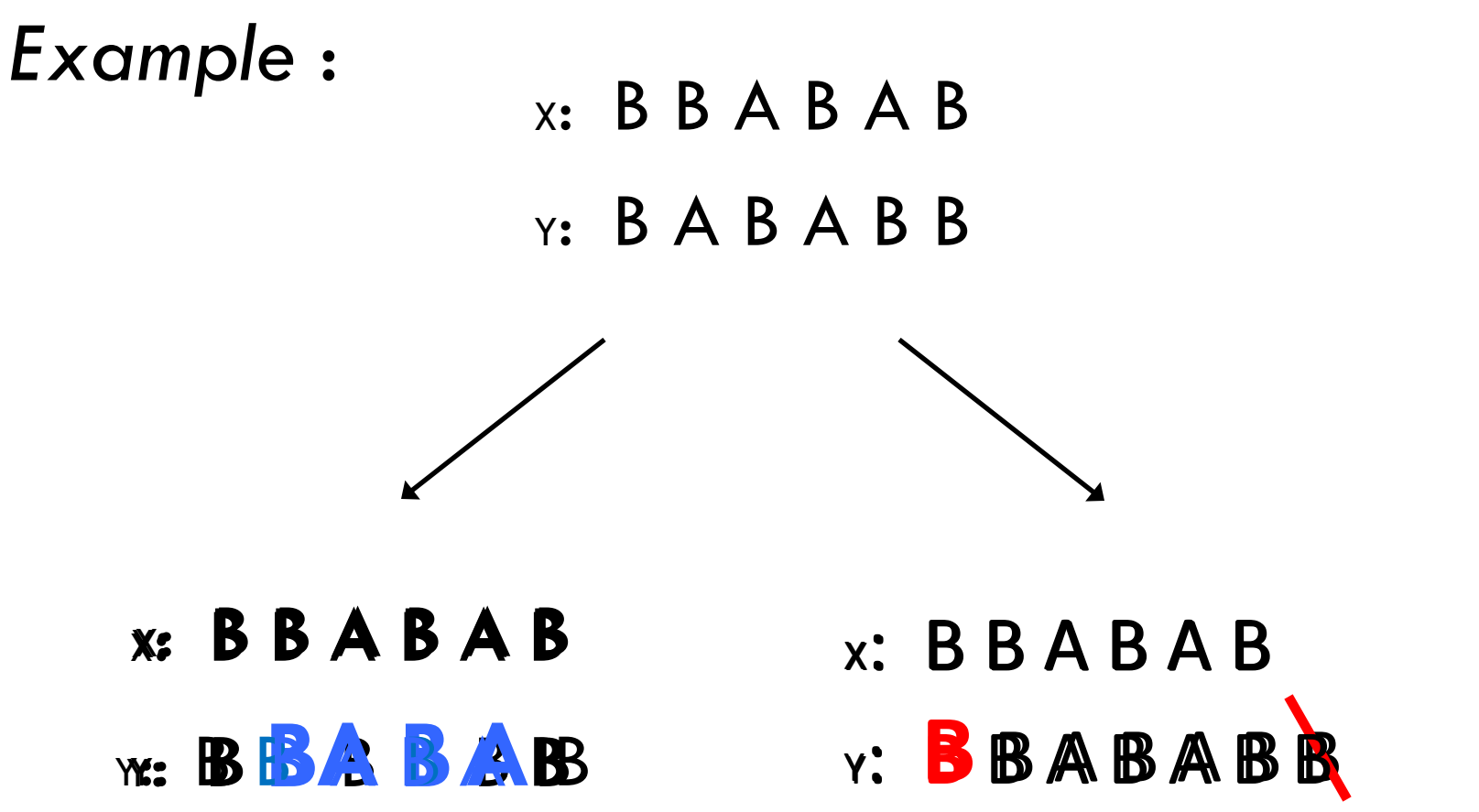
Optimal Matching Analysis (1)

- Method used in molecular biology (DNA strings)
- Introduced in social sciences by Andrew Abbott in the 80's
- ***Principle***: measuring dissimilarity between pairs of sequences by calculating the cost of the transformation of one sequence into the other

See for example (Macindoe & Abbott, 2004)

Optimal Matching Analysis (2)

Example :



Optimal Matching Analysis (3)

- 3 elementary operations :
  - insertion
  - deletion
  - substitution
- Each operation is assigned a **cost**
- The distance between two sequences is equal to the minimal cost needed to transform one sequence into the other

The choice of costs (1)

- Crucial issue of OMA
- *Substitution*: retains the temporal structure (moment) but distorts events (order)

  Insertion/deletion: distorts time but retains the order of events
- A usual choice: indel=1, subst=2

The choice of costs (2)

- Substitution costs matrix:
  - According to theoritical assumptions:
    - ➔ *stratification*
  - Data driven:
    - ➔ *based on transition likelihoods*

The choice of costs (3)

- **in**sertion/**del**etion (indel) cost:
  - If order prevails:
    - ➔ *indel = 1/10 maximal substitution cost*
  - If the moment prevails:
    - ➔ *indel> 1/2 maximal substitution cost*

Elzinga's alternatives

- **Criticism:** OMA doesn't take order into account (substitution of A to B or B to A are equivalent)

- **Several metrics:**
  - Longest common prefix
  - Longest common subsequence
  - Number of common subsequences
  - Number of matching subsequences
  - …

Dynamic Hamming *(Lesnard, 2004)*

- **Criticism:** Transition likelihoods are time-dependant
- **Principle:**
  - no indel operations
  - substitution costs computed for each time point
- Applied to time-use diary data

To summarize…

- A large set of methods

- Each one has specificities, in the way it handles the dimensions of time:
  - moment
  - duration
  - transitions

→ Necessity of empirical comparisons

Previous comparisons

- A few articles mention the robustness of the results

- Comparison of OMA / summarized calendar (QHA), applied to occupational careers *(Robette,Thibault, 2008)*:
  - The main patterns are identical
  - Mobility patterns more homogeneous with OMA
  - Small clusters including rare states (farming, unemployment) appear with QHA

How to compare methods ?

- Data: event-histories on occupational trajectories
- Multidimensional Scaling
- Correlation between first components
- Observation of the dyads of sequences of which dissimilarities are the most different using two distinct methods

Methods tested

- CA methods: PCA, CA, summarized calendar
- OMA, with various cost schemes
- Elzinga's metrics
- Dynamic Hamming

First results (1)

- Several sets of very close methods
  - OMA + Dynamic Hamming
  - CA methods
  - Elzinga's metrics

First results (2)

- Main differences:
  - Transition *vs* duration (sequences with the same transition, occurring at a different moment)
  - Very chaotic and/or totally different sequences

Conclusion

- Exploration of complex trajectories, complementarity with stochastic approaches
- Robustness, flexibility
- To go further → attempt on simulated sequences ?

Bibliography

- Robette Nicolas, Thibault Nicolas, 2008, « Comparing qualitative harmonic analysis and optimal matching. An exploratory study of occupational trajectories », *Population-E*, 64(3), p. 533-556.
- Macindoe Heather, Abbott Andrew, 2004, « Sequence analysis and optimal matching techniques for social science data », in Hardy Melissa, Bryman Alan Handbook of Data Analysis, London, Sage, p. 387-406.
- Grelet Yvette, 2002, « Des typologies de parcours. Méthodes et usages », *Document Génération 92*, (20), 47 p.
- Elzinga Cees H., 2007, « Sequence analysis: metric representations of categorical time series », Sociological methods & research.
- Lesnard Laurent, 2006, « Optimal matching and social sciences », *Document de travail du CREST*, (01), 25 p.

Contact:

nicolas.robette@unibocconi.it

Presentations and papers here:

http://nicolas.robette.free.fr/Publis_eng.htm

A short handbook is forthcoming…