Highlighting changes and differences in 20th century Swiss life trajectories with TraMineR

Gilbert Ritschard
Alexis Gabadinho, Nicolas S. Müller, Matthias Studer

Department of Econometrics and Laboratory of Demography
University of Geneva
http://mephisto.unige.ch/traminer

Swiss Statistical Meeting, Geneva, October 28-30, 2009

Outline

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES ÉCONOMIQUES ET SOCIALES
Département d'économétrie

Outline

Section outline

Objectives

- Illustrate some of the many exploratory features of TraMineR
- A package for Life Trajectory Mining in R
  - State sequences (education, full time, at home, part time, ...)
  - Event sequences (ending education, starting job, ...)
- Highlighting results about Swiss occupational trajectories
  - Differences between women and men
  - Evolution across birth cohorts

Using Data from the 2002 biographical retrospective survey carried on by the Swiss Household Panel

Objectives

- Illustrate some of the many exploratory features of TraMineR
- A package for Life Trajectory Mining in R
  - State sequences (education, full time, at home, part time, ...)
  - Event sequences (ending education, starting job, ...)
- Highlighting results about Swiss occupational trajectories
  - Differences between women and men
  - Evolution across birth cohorts

  Using Data from the 2002 biographical retrospective survey carried on by the Swiss Household Panel

Section outline

TraMineR's features

- Handling of longitudinal data and conversion between various sequence formats
- Plotting sequences (density plot, frequency plot, index plot and more)
- Centro-type and discrepancy measure of a set of sequences
- Individual longitudinal characteristics of sequences (length, time in each state, longitudinal entropy, turbulence and more)
- Sequence transversal characteristics by age point (transversal state distribution, transversal entropy, modal state)
- Other aggregated characteristics (transition rates, average duration in each state, sequence frequency)
- Dissimilarities between pairs of sequences (Optimal matching, longest common subsequence, Hamming, Dynamic Hamming, Multichannel and more)
- ANOVA-like analysis of sequences and tree structured ANOVA from dissimilarities
- Extracting frequent event subsequences
- Identifying most discriminating event subsequences
- Association rules between subsequences

Section outline

1. Introduction
   - Objectives
   - TraMineR
   - Data

The data

- Derived from 2002 biographical SHP survey
- Yearly data
- 1503 life trajectories between ages 20 and 45 (25 years length)
- Focus on
  - Occupational trajectories (8 states)
  - Cohabitational trajectories (10 states)

Outline

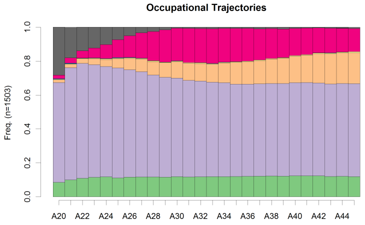
Section outline

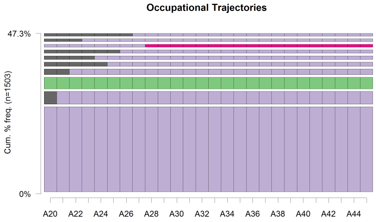
2 State sequences

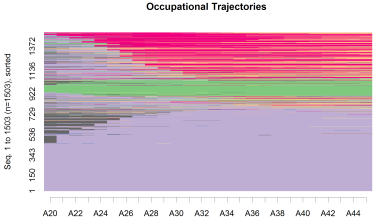
- Basic plots for state sequences
- Characterizing a set of sequences
- Individual longitudinal characteristics
- Computing and exploring pairwise dissimilarities
- Analysis of sequence discrepancy (ANOVA)
- Tree structured discrepancy analysis

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES ÉCONOMIQUES ET SOCIALES
Département d'économétrie

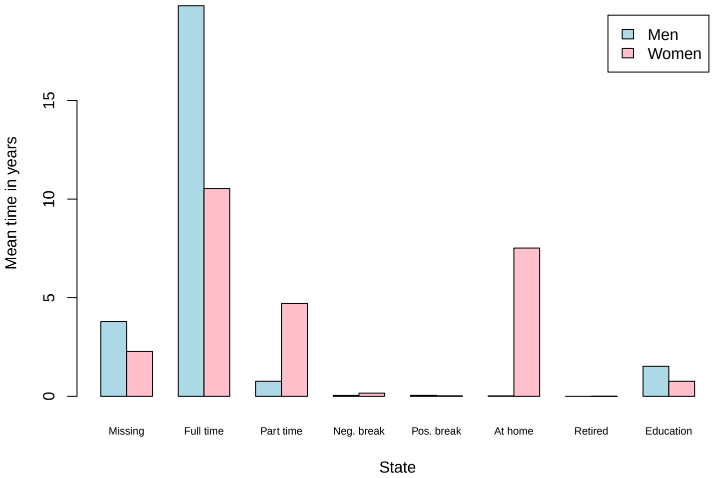
Rendering state sequences







Mean time in each state



Section outline

Characterizing a set of sequences

- Sequence of transversal measures (modal state, between entropy, ...)

id	t_1	t_2	t_3	$\cdots$
1	B	B	D	$\cdots$
2	A	B	C	$\cdots$
3	B	B	A	$\cdots$

- Summary of longitudinal measures (within entropy, transition rates, mean duration ...)

id	t_1	t_2	t_3	$\cdots$
1	B	B	D	$\cdots$
2	A	B	C	$\cdots$
3	B	B	A	$\cdots$

- Other global characteristics: Centro-type sequence, diversity of sequences, ...

Characterizing a set of sequences

- Sequence of transversal measures (modal state, between entropy, ...)

id	t_1	t_2	t_3	$\cdots$
1	B	B	D	$\cdots$
2	A	B	C	$\cdots$
3	B	B	A	$\cdots$

- Summary of longitudinal measures (within entropy, transition rates, mean duration ...)

id	t_1	t_2	t_3	$\cdots$
1	B	B	D	$\cdots$
2	A	B	C	$\cdots$
3	B	B	A	$\cdots$

- Other global characteristics: Centro-type sequence, diversity of sequences, ...

Characterizing a set of sequences

- Sequence of transversal measures (modal state, between entropy, ...)

id	t_1	t_2	t_3	$\cdots$
1	B	B	D	$\cdots$
2	A	B	C	$\cdots$
3	B	B	A	$\cdots$

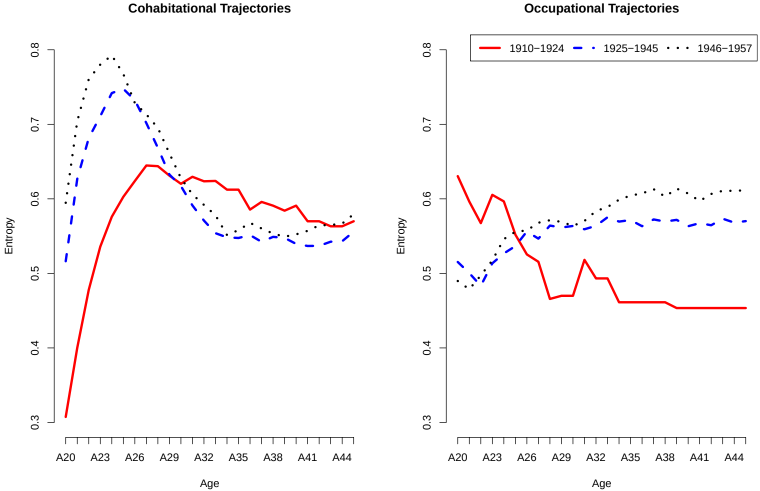
- Summary of longitudinal measures (within entropy, transition rates, mean duration ...)

id	t_1	t_2	t_3	$\cdots$
1	B	B	D	$\cdots$
2	A	B	C	$\cdots$
3	B	B	A	$\cdots$

- Other global characteristics: Centro-type sequence, diversity of sequences, ...

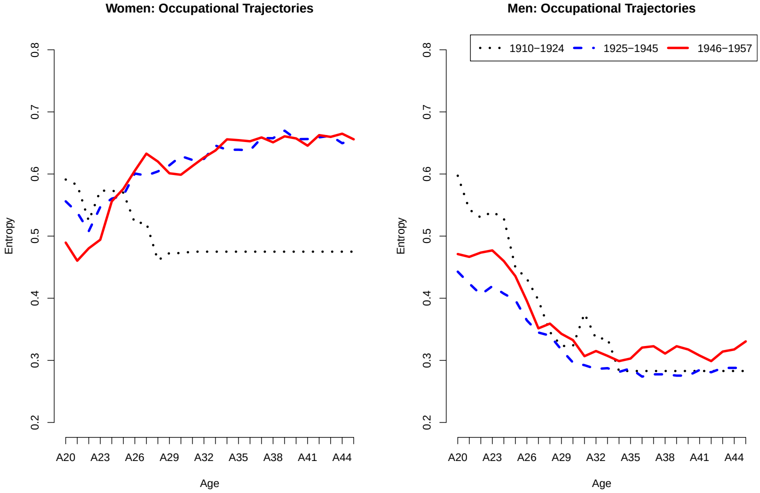
Heterogeneity: Sequence of transversal entropies

Cohabitational vs Occupational



Heterogeneity: Sequence of transversal entropies

Occupational, Women vs Men



Section outline

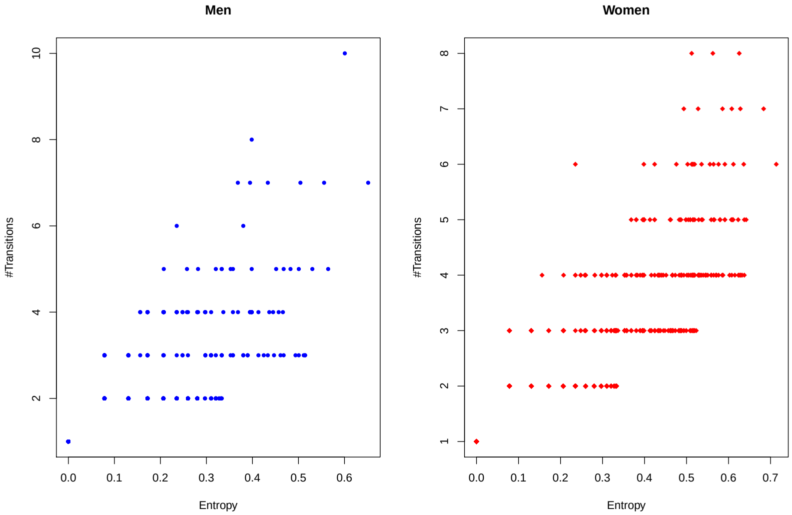
2 State sequences

- Basic plots for state sequences
- Characterizing a set of sequences
- Individual longitudinal characteristics
- Computing and exploring pairwise dissimilarities
- Analysis of sequence discrepancy (ANOVA)
- Tree structured discrepancy analysis

Longitudinal entropy

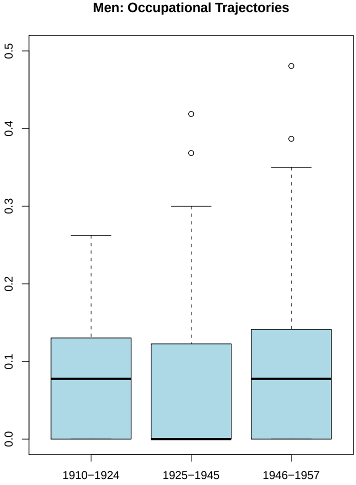
Number of distinct successive states (i.e. transitions)

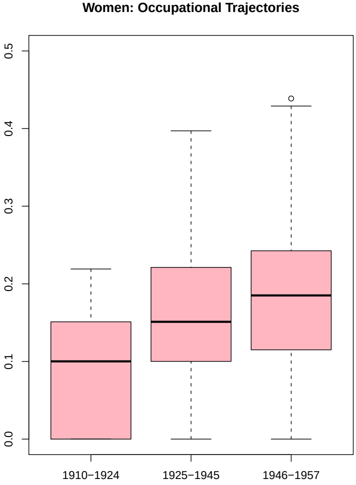
Entropy versus Number of transitions



Sequence complexity

Combines longitudinal entropy and number of transitions





Section outline

2. State sequences
   - Basic plots for state sequences
   - Characterizing a set of sequences
   - Individual longitudinal characteristics
   - Computing and exploring pairwise dissimilarities
   - Analysis of sequence discrepancy (ANOVA)
   - Tree structured discrepancy analysis

Pairwise dissimilarities between sequences

- Distance between sequences
  - Different metrics (LCP, LCS, OM, HAM, DHD, ...)
- Once we have pairwise dissimilarities, we can
  - Determine a central sequence (centro-type)
  - Measure the discrepancy between sequences
  - Cluster a set of sequences
  - MDS scatterplot representation of sequences
  - Discrepancy analysis of a set of sequences (ANOVA)
  - Tree Structured Discrepancy Analysis (Induction trees)

Pairwise dissimilarities between sequences

- Distance between sequences
  - Different metrics (LCP, LCS, OM, HAM, DHD, ...)
- Once we have pairwise dissimilarities, we can
  - Determine a central sequence (centro-type)
  - Measure the discrepancy between sequences
  - Cluster a set of sequences
  - MDS scatterplot representation of sequences
  - Discrepancy analysis of a set of sequences (ANOVA)
  - Tree Structured Discrepancy Analysis (Induction trees)

Pairwise dissimilarities between sequences

- Distance between sequences
  - Different metrics (LCP, LCS, OM, HAM, DHD, ...)
- Once we have pairwise dissimilarities, we can
  - Determine a central sequence (centro-type)
  - Measure the discrepancy between sequences
  - Cluster a set of sequences
  - MDS scatterplot representation of sequences
  - Discrepancy analysis of a set of sequences (ANOVA)
  - Tree Structured Discrepancy Analysis (Induction trees)

Pairwise dissimilarities between sequences

- Distance between sequences
  - Different metrics (LCP, LCS, OM, HAM, DHD, ...)
- Once we have pairwise dissimilarities, we can
  - Determine a central sequence (centro-type)
  - Measure the discrepancy between sequences
  - Cluster a set of sequences
  - MDS scatterplot representation of sequences
  - Discrepancy analysis of a set of sequences (ANOVA)
  - Tree Structured Discrepancy Analysis (Induction trees)

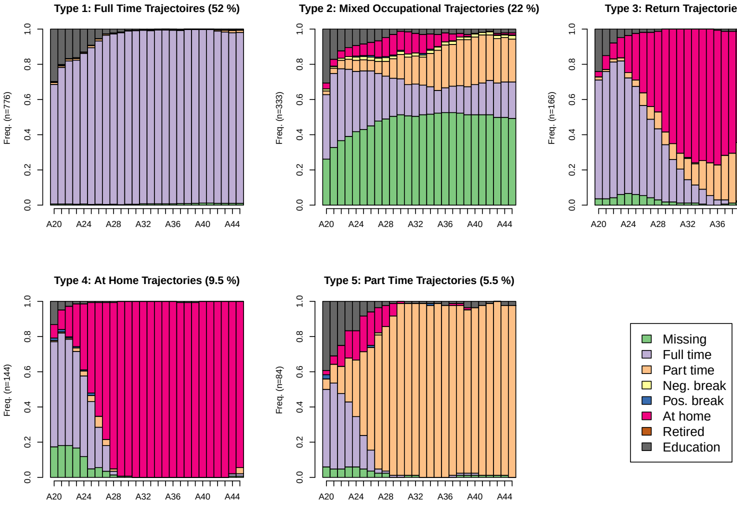
Deriving clusters from pairwise dissimilarities

- For each of the two sets of sequences: cohabitational and occupational
- Compute Pairwise dissimilarities (a 1503×1503 matrix)
- Here, we used Optimal Matching (OM)
  - For each pair $\{x, y\}$ of sequences, OM is the minimal cost of transforming one sequence into the other
    - insert/deletion (indel) cost $= 1$
    - substitution cost $c_{i,j} = c_{j,i} = 2 - p(i_t \mid j_{t-1}) - p(j_t \mid i_{t-1})$
- Cluster by plugging obtained dissimilarity matrix in any cluster algorithm
- We used an agglomerative hierarchical method with Ward's criteria
- and retained partition into 5 clusters

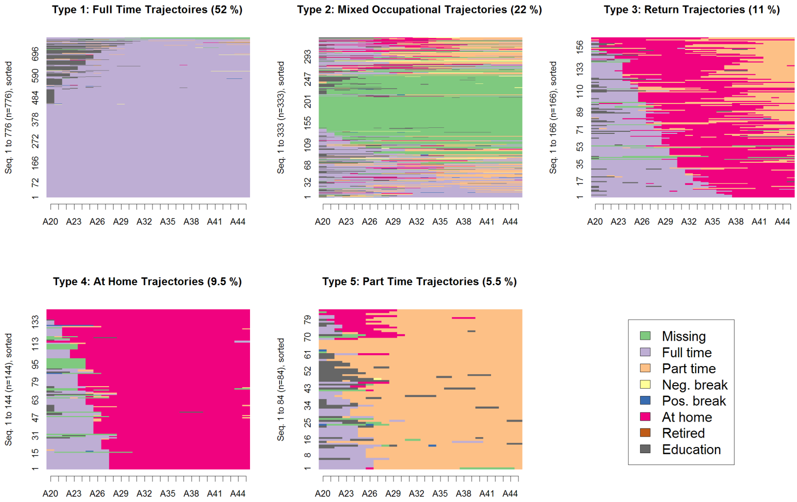
Deriving clusters from pairwise dissimilarities

- For each of the two sets of sequences: cohabitational and occupational
- Compute Pairwise dissimilarities (a 1503×1503 matrix)
- Here, we used Optimal Matching (OM)
  - For each pair $\{x, y\}$ of sequences, OM is the minimal cost of transforming one sequence into the other
    - insert/deletion (indel) cost $= 1$
    - substitution cost $c_{i,j} = c_{j,i} = 2 - p(i_t \mid j_{t-1}) - p(j_t \mid i_{t-1})$
- Cluster by plugging obtained dissimilarity matrix in any cluster algorithm
- We used an agglomerative hierarchical method with Ward's criteria
- and retained partition into 5 clusters

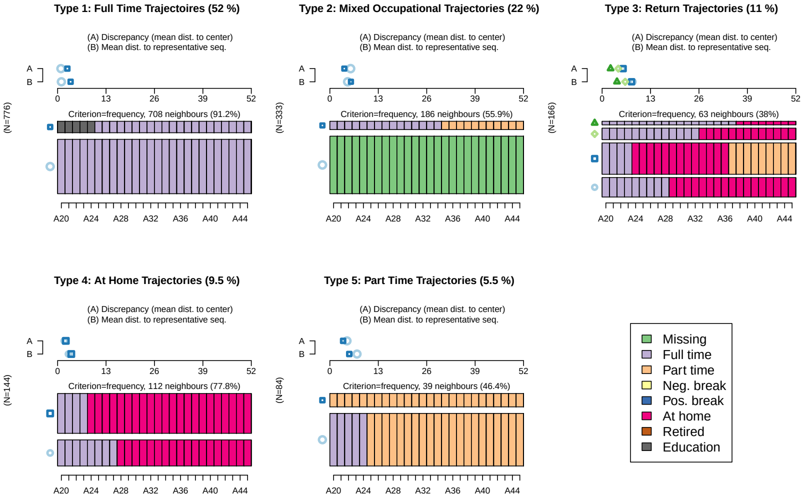
Cluster analysis: determining typologies



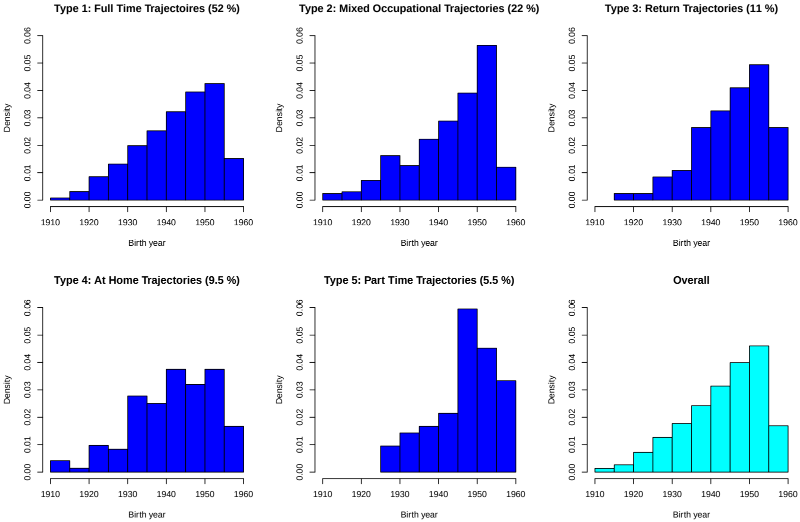
Cluster analysis: i-plots (sorted by 1st MDS factor)



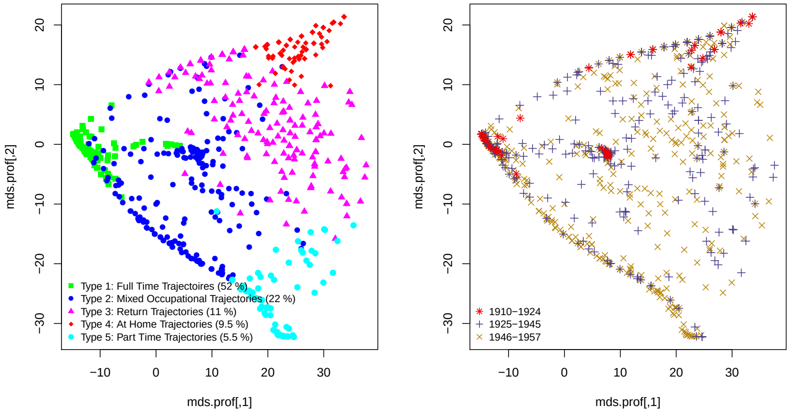
Cluster analysis: representative sequences



Birth year distribution by cluster



MDS: Scatterplot view of sequences



Section outline

2 State sequences

- Basic plots for state sequences
- Characterizing a set of sequences
- Individual longitudinal characteristics
- Computing and exploring pairwise dissimilarities
- Analysis of sequence discrepancy (ANOVA)
- Tree structured discrepancy analysis

Dispersion of the set of sequences

- From the distance matrix, we get the pseudo-variance of the set of sequences.
- Sum of squares SS can be expressed in terms of distances between pairs

$$SS = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n}(y_i - y_j)^2$$
$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n} d_{ij}$$

- Setting d_{ij} equal to OM, LCP, LCS ... distance, we get SS.
- Can apply ANOVA principle (Studer et al., 2009).

Dispersion of the set of sequences

- From the distance matrix, we get the pseudo-variance of the set of sequences.
- Sum of squares SS can be expressed in terms of distances between pairs

$$SS = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n}(y_i - y_j)^2$$
$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n} d_{ij}$$

- Setting d_{ij} equal to OM, LCP, LCS ... distance, we get SS.
- Can apply ANOVA principle (Studer et al., 2009).

Dispersion of the set of sequences

- From the distance matrix, we get the pseudo-variance of the set of sequences.
- Sum of squares SS can be expressed in terms of distances between pairs

$$
\begin{aligned}
SS &= \sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n}(y_i - y_j)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n} d_{ij}
\end{aligned}
$$

- Setting d_{ij} equal to OM, LCP, LCS ... distance, we get SS.
- Can apply ANOVA principle (Studer et al., 2009).

Analysis of sequence discrepancy

- ANOVA like analysis based on pairwise dissimilarities
- We decompose the SS (Sum of squares equivalent)

$$SS_T = SS_B + SS_W$$

- Here, with the formula shown earlier

$$\begin{aligned} SS_T &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n} d_{ij} \\ SS_W &= \sum_{g}\left(\frac{1}{n_g}\sum_{i=1}^{n_g}\sum_{j=i+1}^{n_g} d_{ij,g}\right) \\ SS_B &= SS_T - SS_W \end{aligned}$$

Pseudo R-square and ANOVA Table

- ANOVA table for m groups

	Discrepancy	df	Mean Discr.	F
Between	SS_B	$df_B = m - 1$	$\frac{SS_B}{df_B}$	$\frac{SS_B}{SS_W}\frac{df_W}{df_B}$
Within	SS_W	$df_W = \sum_g n_g - m$	$\frac{SS_W}{df_W}$	
Total	SS_T	$df_T = n - 1$		

- Pseudo R^2

$$R^2 = \frac{SS_B}{SS_T}$$

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES ÉCONOMIQUES ET SOCIALES
Département d'économétrie

Pseudo R-square and ANOVA Table

- ANOVA table for m groups

	Discrepancy	df	Mean Discr.	F
Between	SS_B	$df_B = m - 1$	$\frac{SS_B}{df_B}$	$\frac{SS_B}{SS_W}\frac{df_W}{df_B}$
Within	SS_W	$df_W = \sum_g n_g - m$	$\frac{SS_W}{df_W}$	
Total	SS_T	$df_T = n - 1$		

- Pseudo R^2

$$R^2 = \frac{SS_B}{SS_T}$$

Pseudo F

- Pseudo F

$$F = \frac{SS_B/(m-1)}{SS_W/(n-m)}$$

- Normality is not defendable in this setting.
- F cannot be compared with an F distribution.
- The significance is assesses through a permutation test
- Permutation test: iteratively randomly reassign each covariate profile to one of the observed sequence and recompute the F.
- Empirical distribution of F under independence.

Pseudo F

- Pseudo F

$$F = \frac{SS_B/(m-1)}{SS_W/(n-m)}$$

- Normality is not defendable in this setting.
- F cannot be compared with an F distribution.
- The significance is assesses through a permutation test
- Permutation test: iteratively randomly reassign each covariate profile to one of the observed sequence and recompute the F.
- Empirical distribution of F under independence.

Analysis of sequence discrepancy

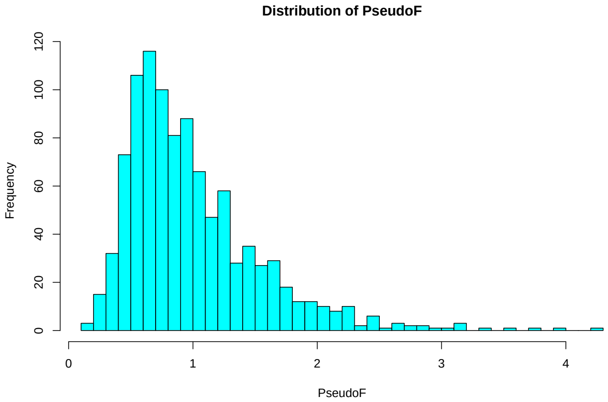
- Running an ANOVA like analysis for `cohort3b`

```
Pseudo ANOVA table:
              SS   df      MSE
Exp     106.4437    2 53.22183
Res   15645.8712 1500 10.43058
Total 15752.3148 1502 10.48756

Test values  (p-values based on 999 permutation):
 PseudoF    PseudoR2 PseudoF_Pval  PseudoT PseudoT_Pval
 5.10248 0.006757335            0 7.361347            0

Variance per level:
             n  variance
1910-1924   71  7.713761
1925-1945  659  9.651546
1946-1957  773 11.303784
Total     1503 10.480582
```

Distribution of pseudo F



Multiple factor analysis

- Generalize previous approach for multiple covariates.
- Here, we consider Type III effects
- Measure the additional contribution of each covariate v when we accounted for all other covariates.
- The F statistics reads

$$F_v = \frac{(SS_{B_c} - SS_{B_v})/p}{SS_{W_c}/(n-m-1)}$$

where the SS_{B_c} and SS_{W_c} are the explained and residual sums of squares of the full model, SS_{B_v} the explained sum of squares of the model after removing variable v, and p the number of indicators or contrasts used to encode the covariate v.

- Significance is assessed again through permutation tests.

Running a Multiple factor analysis

	Variable	PseudoF	PseudoR2	p_value
1	sex	486.157573	0.222836269	0.000000000
2	cohort3b	5.297978	0.004856786	0.000999001
3	edu_lev	33.998319	0.046750636	0.000000000
4	Total	114.523325	0.314748465	0.000000000

Differences over time

- How do differences between groups vary over time?
- At which age do trajectories most differ across birth cohorts?
- Compute R^2 for short sliding windows (length 2)
- We get thus a sequence of R^2, which can be plotted
- Similarly, we can plot series of
  - total residual discrepancy (SS_W)
  - residual discrepancy of each group (SS_G)

Differences over time

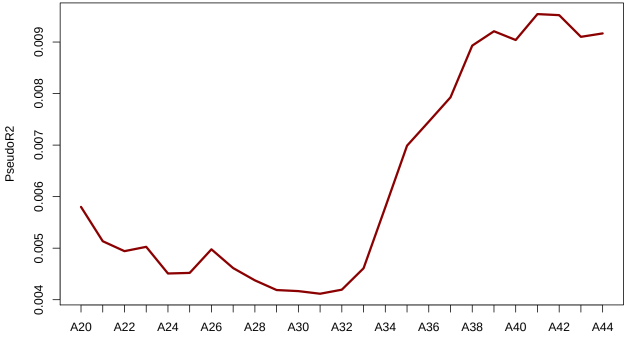
- How do differences between groups vary over time?
- At which age do trajectories most differ across birth cohorts?
- Compute R^2 for short sliding windows (length 2)
- We get thus a sequence of R^2, which can be plotted
- Similarly, we can plot series of
  - total residual discrepancy (SS_W)
  - residual discrepancy of each group (SS_G)

Differences over time

- How do differences between groups vary over time?
- At which age do trajectories most differ across birth cohorts?
- Compute R^2 for short sliding windows (length 2)
- We get thus a sequence of R^2, which can be plotted
- Similarly, we can plot series of
  - total residual discrepancy (SS_W)
  - residual discrepancy of each group (SS_G)

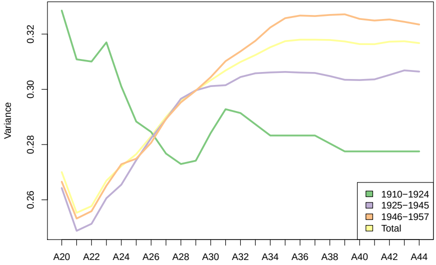
Plotting R-squares over time

Birth cohorts



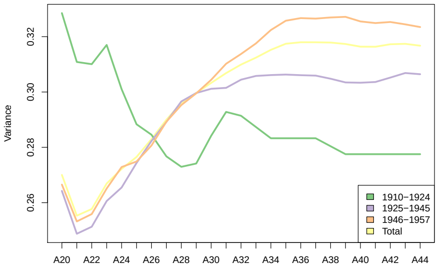
Plotting residual discrepancy over time

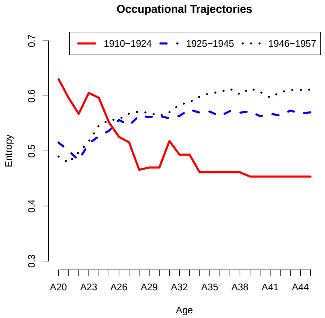
Birth cohorts



Plotting residual discrepancy over time

Birth cohorts





Section outline

2 State sequences

- Basic plots for state sequences
- Characterizing a set of sequences
- Individual longitudinal characteristics
- Computing and exploring pairwise dissimilarities
- Analysis of sequence discrepancy (ANOVA)
- Tree structured discrepancy analysis

Tree structured discrepancy analysis

- Objective: Find the most important predictors and their interactions.
- Iteratively segment the cases using values of covariates (predictors)
- Such that groups be as homogenous as possible.

- At each step, we select the covariate and split with highest R^2.
- Significance of split is assessed through a permutation F test.
- Growing stops when the selected split is not significant.

Tree structured discrepancy analysis

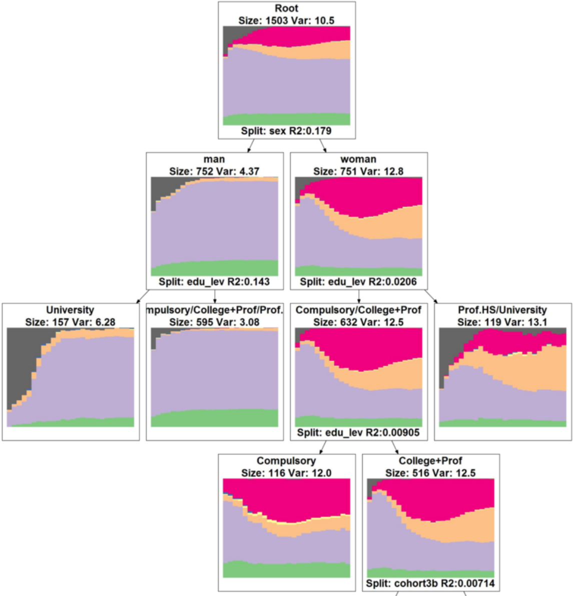
- Objective: Find the most important predictors and their interactions.
- Iteratively segment the cases using values of covariates (predictors)
- Such that groups be as homogenous as possible.

- At each step, we select the covariate and split with highest R^2.
- Significance of split is assessed through a permutation F test.
- Growing stops when the selected split is not significant.

Growing the tree

```
Dissimilarity tree
Global R2: 0.229
  |-- Root [ 1503 ] var: 10.5
      |-> sex  R2: 0.179
        |-- man [ 752 ] var: 4.37
            |-> edu_lev  R2: 0.143
              |-- University [ 157 ] var: 6.28
              |-- Compulsory/College+Prof/Prof.HS [ 595 ] var: 3.08
        |-- woman [ 751 ] var: 12.8
            |-> edu_lev  R2: 0.0206
              |-- Compulsory/College+Prof [ 632 ] var: 12.5
                  |-> edu_lev  R2: 0.00905
                    |-- Compulsory [ 116 ] var: 12.0
                    |-- College+Prof [ 516 ] var: 12.5
                        |-> cohort3b  R2: 0.00714
                          |-- 1946-1957 [ 280 ] var: 12.5
                          |-- 1910-1924/1925-1945 [ 236 ] var: 12.2
              |-- Prof.HS/University [ 119 ] var: 13.1
```

Graphical Tree



Outline

Event sequences

- Time stamped events

  (end education, 21) (start full time job, 21) (at home, 28) (start part time, 29)

- Which are the most typical sequencings?
- Which are the most typical events that occur after the sub-sequence (leaving home, ending education)?
- Which sequencings do most differ among groups?
- ...

- Unlike state sequences, event sequences are hard to visualize

Event sequences

- Time stamped events

  (end education, 21) (start full time job, 21) (at home, 28) (start part time, 29)

- Which are the most typical sequencings?
- Which are the most typical events that occur after the sub-sequence (leaving home, ending education)?
- Which sequencings do most differ among groups?
- ...

- Unlike state sequences, event sequences are hard to visualize

Event sequences

- Time stamped events

  (end education, 21) (start full time job, 21) (at home, 28) (start part time, 29)

- Which are the most typical sequencings?
- Which are the most typical events that occur after the sub-sequence (leaving home, ending education)?
- Which sequencings do most differ among groups?
- ...

- Unlike state sequences, event sequences are hard to visualize

Events, Transitions and States

- An event occurs at a given time
  (leaving home, starting job, ...)
- Transition: set of events occurring simultaneously
- A transition corresponds to a state change
- Easy to transform between state and transition sequences
- Converting to and from events requires additional information

- To illustrate, we consider hereafter the events defined by state changes in our previous trajectories

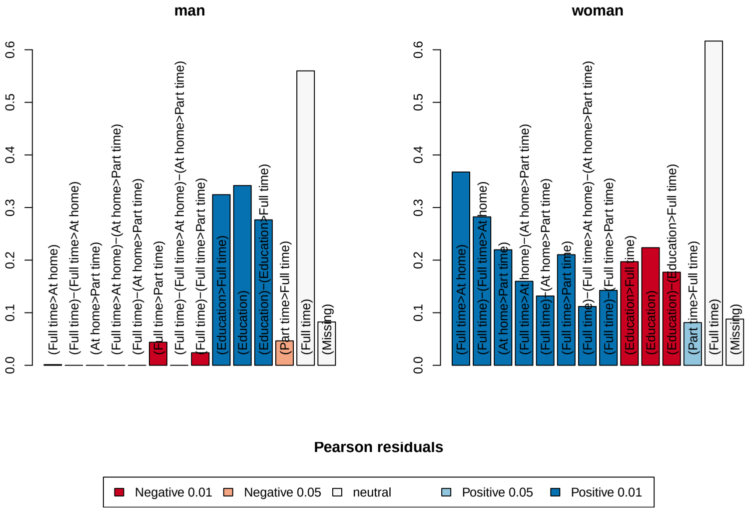
Events, Transitions and States

- An event occurs at a given time
  (leaving home, starting job, ...)
- Transition: set of events occurring simultaneously
- A transition corresponds to a state change
- Easy to transform between state and transition sequences
- Converting to and from events requires additional information

- To illustrate, we consider hereafter the events defined by state changes in our previous trajectories

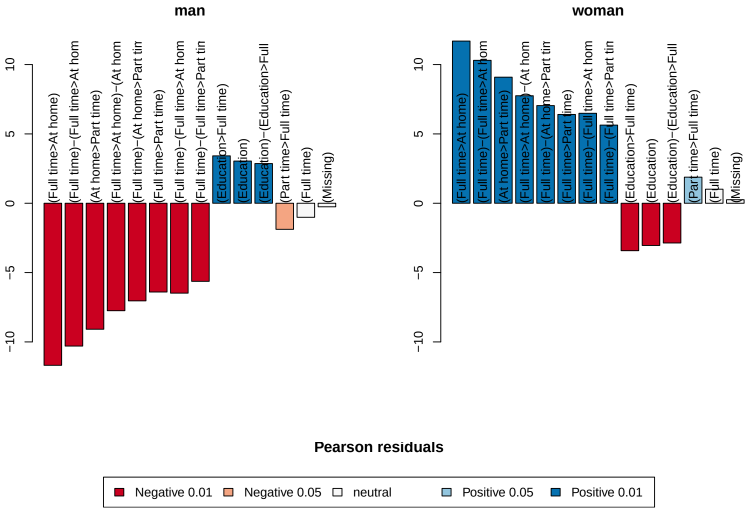
Event sequences: discriminating sub-sequences

Between sex, frequencies



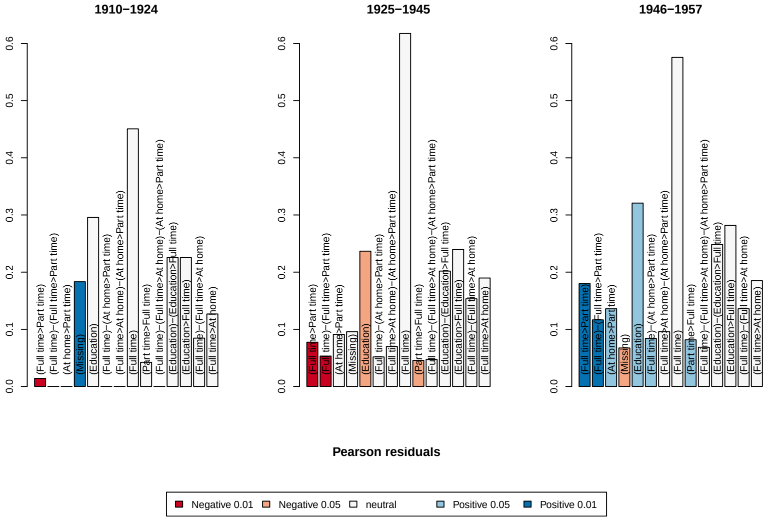
Event sequences: discriminating sub-sequences

Between sex, residuals



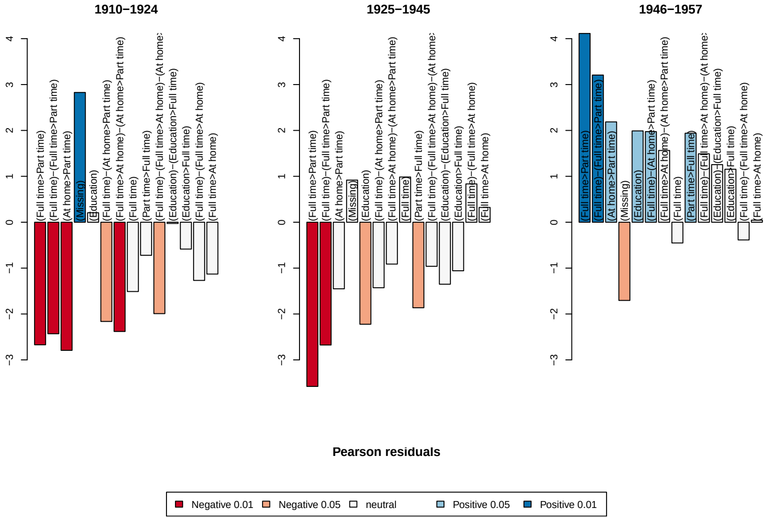
Event sequences: discriminating between birth cohorts

frequencies



Event sequences: discriminating between birth cohorts

residuals



Outline

Conclusion 1: about sequence analysis

- Analyzing trajectories until 45, implies ignoring recent generations
- Most recent birth year is 1957 ($2002 - 45$)
- Missing data in sequences is a crucial issue
- TraMineR permits different handling for left, right and in between missings
  - consider as a specific state
  - drop (shifts state sequences left)
  - impute, but how?
- Weights
  - Can be handled in sequence rendering (weighted transversal characteristics)
  - Not really an issue for computing dissimilarities and longitudinal charcateristics
  - We are working on a solution for permutation tests

Conclusion 1: about sequence analysis

- Analyzing trajectories until 45, implies ignoring recent generations
- Most recent birth year is 1957 (2002 – 45)
- Missing data in sequences is a crucial issue
- TraMineR permits different handling for left, right and in between missings
  - consider as a specific state
  - drop (shifts state sequences left)
  - impute, but how?
- Weights
  - Can be handled in sequence rendering (weighted transversal characteristics)
  - Not really an issue for computing dissimilarities and longitudinal charcateristics
  - We are working on a solution for permutation tests

Conclusion 1: about sequence analysis

- Analyzing trajectories until 45, implies ignoring recent generations
- Most recent birth year is 1957 (2002 – 45)
- Missing data in sequences is a crucial issue
- TraMineR permits different handling for left, right and in between missings
  - consider as a specific state
  - drop (shifts state sequences left)
  - impute, but how?
- Weights
  - Can be handled in sequence rendering (weighted transversal characteristics)
  - Not really an issue for computing dissimilarities and longitudinal charcateristics
  - We are working on a solution for permutation tests

Conclusion 2: extending analysis

- Since it runs in R, TraMineR's outcome can be easily combined in a same script with other R procedures
- We have shown: cluster analysis, MDS, ...
- In Widmer and Ritschard (2009), we studied
  - Relationship between occupational and cohabitational trajectories by regressing longitudinal entropies of each of them on both occupational and cohabitational clusters while controlling for birth cohorts and sex
  - Studied also cluster membership by means of logistic regressions.

Conclusion 3: about TraMineR

- TraMineR is a unique powerful tool for discrete sequences
- Can do much more than shown in this presentation, for instance
  - sequence data management
  - conversion between event and state sequences
  - multiple metrics, including multi-channel for parallel sequences
  - dissimilarities between event sequences
  - discovering association rules between event-subsequences
  - ...
- ... and, as R, it is available for free on the CRAN http://cran.r-project.org
- See also the package web page http://mephisto.unige.ch/traminer

Thank You!

References I

Gabadinho, A., G. Ritschard, M. Studer, and N. S. Müller (2008). Mining sequence data in R with TraMineR: A user's guide. Technical report, Department of Econometrics and Laboratory of Demography, University of Geneva, Geneva.

Ritschard, G., A. Gabadinho, N. S. Müller, and M. Studer (2008). Mining event histories: A social science perspective. *International Journal of Data Mining, Modelling and Management 1*(1), 68–90.

Studer, M., G. Ritschard, A. Gabadinho, and N. S. Müller (2009). Discrepancy analysis of complex objects using dissimilarities. In H. Briand, F. Guillet, G. Ritschard, and D. A. Zighed (Eds.), *Advances in Knowledge Discovery and Management*, Studies in Computational Intelligence. Berlin: Springer. (forthcoming).

Widmer, E. and G. Ritschard (2009). The de-standardization of the life course: Are men and women equal? *Advances in Life course Research 14*(1-2), 28–39.