

Supervised and non-supervised analysis of the diversity of categorical sequences

Gilbert Ritschard

Department of Econometrics and Institute of Demography
University of Geneva

<http://mephisto.unige.ch/traminer>

Séminaire de recherche, Labo ERIC, Lyon, 9 décembre 2010

Outline

- 1 Objectives
- 2 Pairwise dissimilarities between sequences
- 3 Unsupervised clustering
- 4 Supervised Classification
- 5 Conclusion

Outline

- 1 Objectives
- 2 Pairwise dissimilarities between sequences
- 3 Unsupervised clustering
- 4 Supervised Classification
- 5 Conclusion

Aim of the presentation

- Using Swiss cohabitational and occupational data from the SHP biographical survey
- Demonstrate the results that can be extracted from **state sequences**
- Unsupervised learning
 - Clustering
 - Representative sequences
- Supervised learning
 - Discrepancy analysis (ANOVA)
 - Regression trees of state sequence data

Typical questions in social sciences

- In the field of **Life course analysis**
 - Are there standard of lives, ideal-types?
 - What are those standards, those ideal-tyes?
 - How are those standards linked to covariates (sex, birth cohort, ...)?
- Can answer to these questions by seeking typologies and studying groupings in terms of covariates

Typical questions in social sciences

- In the field of **Life course analysis**
 - Are there standard of lives, ideal-types?
 - What are those standards, those ideal-types?
 - How are those standards linked to covariates (sex, birth cohort, ...)?
- Can answer to these questions by seeking typologies and studying groupings in terms of covariates

Typical questions in social sciences

- In the field of **Life course analysis**
 - Are there standard of lives, ideal-types?
 - What are those standards, those ideal-types?
 - How are those standards linked to covariates (sex, birth cohort, ...)?
- Can answer to these questions by seeking typologies and studying groupings in terms of covariates

Prerequisite: pairwise dissimilarities

- Common point between all addressed methods:
- ... all are based on **pairwise dissimilarities** between sequences

Outline

- 1 Objectives
- 2 Pairwise dissimilarities between sequences**
- 3 Unsupervised clustering
- 4 Supervised Classification
- 5 Conclusion

Dissimilarity measure – 1

Based on count of matching attributes

- Measures Based on **count of matching attributes** $A(x, y)$ (proximity)

$$d(x, y) = A(x, x) + A(y, y) - 2A(x, y)$$

available in TraMineR

- **LCP** $A(x, y)$ = length of longest common prefix
- **RLCP** $A(x, y)$ = length of longest common suffix
- **LCS** $A(x, y)$ = length of longest common subsequence
- **HAM simple** $A(x, y)$ = half of number of matching elements

Dissimilarity measure – 2

Edit distances

- **Edit distance**: (minimal) cost of transforming x into y available in R
 - **OM** Optimal matching of **state séquences** (Levenshtein, 1966)
 - indel cost (insertion/deletion)
 - pairwise substitution costs
 - Generalized **HAM**, Hamming = OM without indel
 - **DHD**, Dynamic Hamming Distance, position-varying substitution cost (Lesnard, 2006)

Discrepancy of a set of sequences

- From a dissimilarity matrix, we can define the **discrepancy** of a set of sequences
- Sum of squares **SS** can be expressed in terms of pairwise distances

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

- Replacing $d_{e,ij}^2$ with the dissimilarity OM, LCP, LCS ... (or its square), we get a **pseudo SS**.

Discrepancy of a set of sequences

- From a dissimilarity matrix, we can define the **discrepancy** of a set of sequences
- Sum of squares **SS** can be expressed in terms of pairwise distances

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

- Replacing $d_{e,ij}^2$ with the dissimilarity OM, LCP, LCS ... (or its square), we get a **pseudo SS**.

Dissimilarity based analysis

- If we know how to compute a dissimilarity,
- we can apply any analysis based on dissimilarities or variances
 - Clustering (agglomerative, divisive, partitioning, ...) (Kaufman and Rousseeuw, 2005)
 - Principal coordinate analysis (PCO, MDS) (Gower, 1966)
 - Representative sequences (Gabadinho et al., 2009b)
 - ANOVA (Studer et al., 2010)
 - ...

Outline

- 1 Objectives
- 2 Pairwise dissimilarities between sequences
- 3 Unsupervised clustering**
- 4 Supervised Classification
- 5 Conclusion

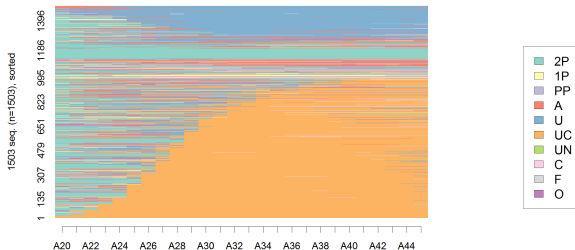
Section outline

- 3 Unsupervised clustering
 - Typology
 - Representative sequences

Building a typology

- To illustrate, hierarchical clustering with Ward criterion
- Data: Cohabital trajectories,
 - 1503 sequences from the 2002 biographic survey of the SHP
 - alphabet of 10 states
 - yearly data, from age 20 to 45 years (length 26)

Cohabital trajectories



Clustering from dissimilarities

- Compute the dissimilarity matrix, for instance with TraMineR


```
om.coh <- seqdist(seqs.coh, method="OM", sm="TRATE", indel=1)
```
- `om.dist.coh` is a 1503×1503 matrix that can be passed to any clustering method that accepts a distance matrix as input
- In R, we can use the `cluster` library (Maechler et al., 2005) which proposes among others
 - `agnes()` an agglomerative method
 - `diana()` a divisive method
 - `pam()` partitioning around medoids
- Illustration: agglomerative method with **Ward**
- We use the `agnes()` function


```
clw.coh <- agnes(om.coh, diss=T, method="ward")
```
- and retain the partition into 5 clusters


```
cutree(clw.coh, k=5)
```

Clustering from dissimilarities

- Compute the dissimilarity matrix, for instance with TraMineR

```
om.coh <- seqdist(seqs.coh, method="OM", sm="TRATE", indel=1)
```
- `om.dist.coh` is a 1503×1503 matrix that can be passed to any clustering method that accepts a distance matrix as input
- In R, we can use the `cluster` library (Maechler et al., 2005) which proposes among others
 - `agnes()` an agglomerative method
 - `diana()` a divisive method
 - `pam()` partitioning around medoids
- Illustration: agglomerative method with `Ward`
- We use the `agnes()` function
 - ```
clw.coh <- agnes(om.coh, diss=T, method="ward")
```
- and retain the partition into 5 clusters
  - ```
cutree(clw.coh, k=5)
```

Clustering from dissimilarities

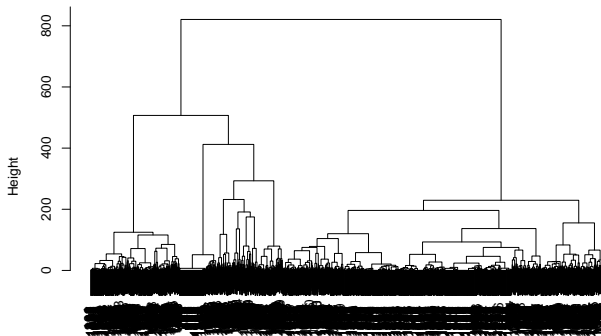
- Compute the dissimilarity matrix, for instance with TraMineR

```
om.coh <- seqdist(seqs.coh, method="OM", sm="TRATE", indel=1)
```
- `om.dist.coh` is a 1503×1503 matrix that can be passed to any clustering method that accepts a distance matrix as input
- In R, we can use the `cluster` library (Maechler et al., 2005) which proposes among others
 - `agnes()` an agglomerative method
 - `diana()` a divisive method
 - `pam()` partitioning around medoids
- Illustration: agglomerative method with **Ward**
- We use the `agnes()` function
 - ```
clw.coh <- agnes(om.coh, diss=T, method="ward")
```
- and retain the partition into 5 clusters
  - ```
cutree(clw.coh, k=5)
```

Hierarchical clustering, Ward

Dendrogram

Cohabitational trajectories



om.coh
Agglomerative Coefficient = 1

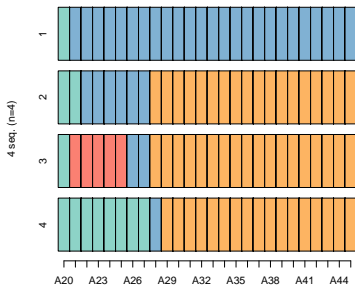
Rendering the clusters

- State sequences can easily be visualized

- Example: **i-plot**

Sequence

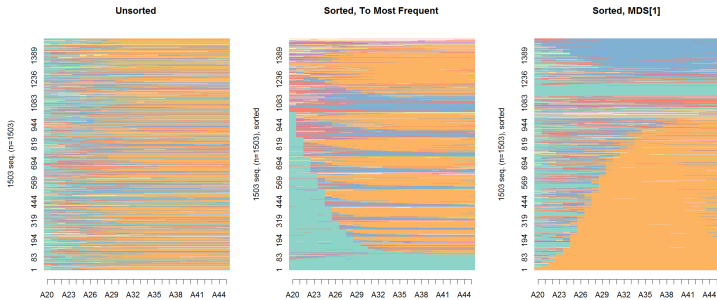
- (2P, 1) - (U, 25)
- (2P, 2) - (U, 6) - (UC, 18)
- (2P, 1) - (A, 5) - (U, 2) - (UC, 18)
- (2P, 8) - (U, 1) - (UC, 17)



- (Horizontal) stacking of rectangles, with color representing the state and length its duration.
- The vertical alignment informs about the distribution at each position

i-plot, and order of the sequences

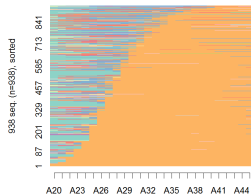
- When number of sequences is high, sorting sequences helps readability



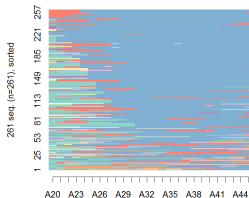
Typology of cohabitational state sequences

i-plot, sorted with MDS[1]

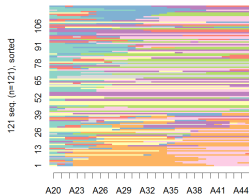
Type 1 : Parental Trajectories (62 %)



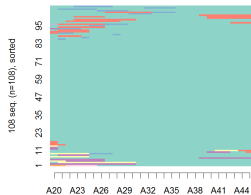
Type 2 : Conjugal Trajectories (17 %)



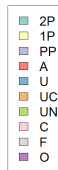
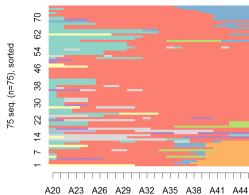
Type 3 : Mixed Cohabitation Trajectories (8 %)



Type 4 : Parental Home Trajectories (7 %)



Type 5 : Solo Trajectories (5 %)



Typology of cohabitational state sequences

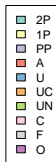
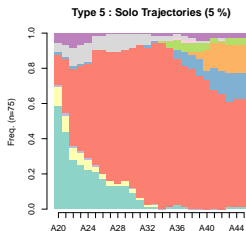
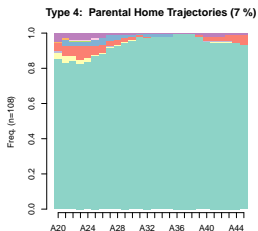
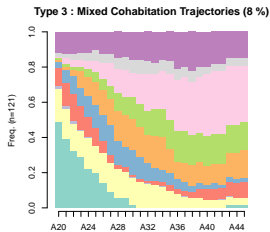
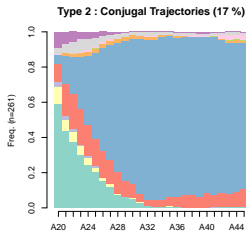
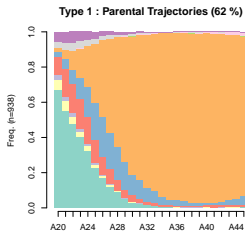
Discrepancies

- Discrepancy (pseudo variance) $\frac{1}{2n^2} \sum_i \sum_j d(i, j)$

	Count	Percent	Discrepancy
Parental	938	62.4	7.819
Conjugal	261	17.4	8.209
Mixed	121	8.1	19.842
Parental Home	108	7.2	3.002
Solo	75	5.0	9.185
Total	1503	100.0	15.526

Typology of cohabitational state sequences

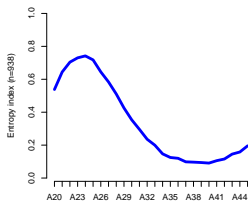
d-plot, transversal distributions



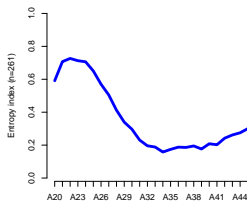
Typology of cohabitational state sequences

Ht-plot, transversal entropies

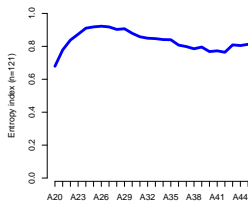
Type 1 : Parental Trajectories (62 %)



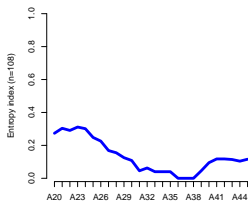
Type 2 : Conjugal Trajectories (17 %)



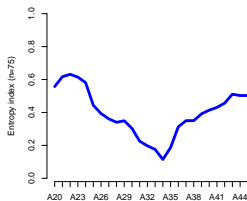
Type 3 : Mixed Cohabitation Trajectories (8 %)



Type 4 : Parental Home Trajectories (7 %)



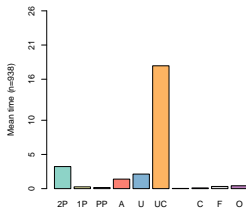
Type 5 : Solo Trajectories (5 %)



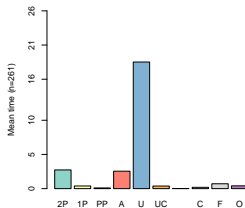
Typology of cohabitational state sequences

mt-plot, mean time in each state

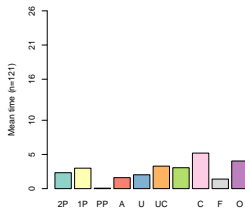
Type 1 : Parental Trajectories (62 %)



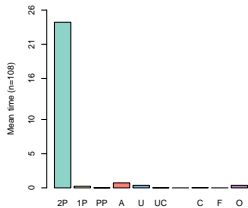
Type 2 : Conjugal Trajectories (17 %)



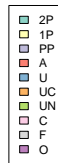
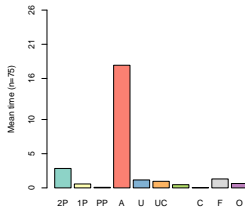
Type 3 : Mixed Cohabitation Trajectories (8 %)



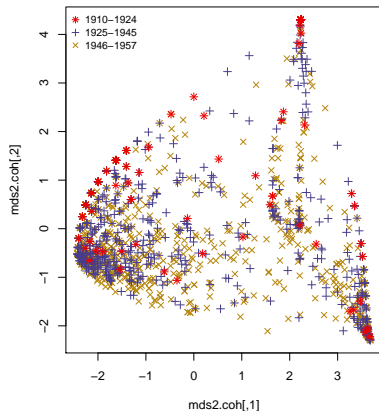
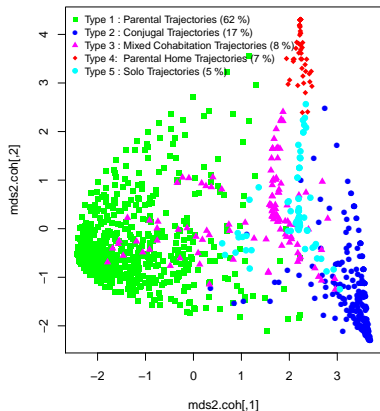
Type 4 : Parental Home Trajectories (7 %)



Type 5 : Solo Trajectories (5 %)



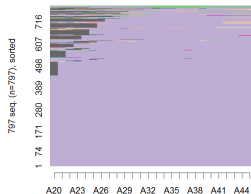
MDS: Cloud of points



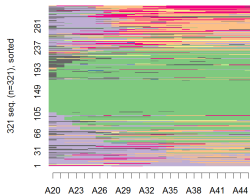
Typology of occupational state sequences

i-plot

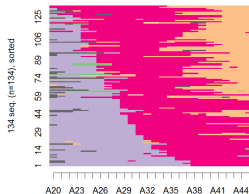
Type 1: Full Time Trajectories (52 %)



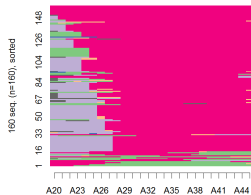
Type 2: Mixed Occupational Trajectories (22 %)



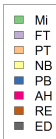
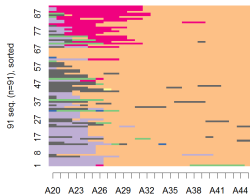
Type 3: Return Trajectories (11 %)



Type 4: At Home Trajectories (9.5 %)

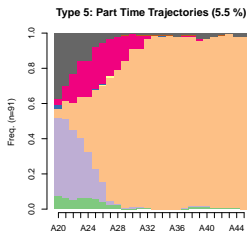
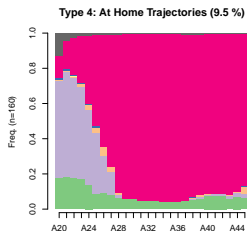
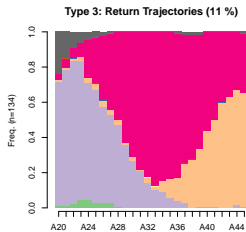
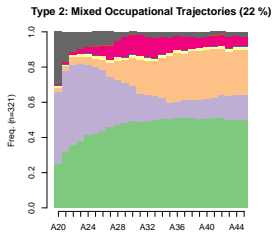
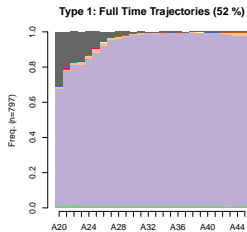


Type 5: Part Time Trajectories (5.5 %)



Typology of cohabitational state sequences

d-plot



Section outline

- 3 Unsupervised clustering
 - Typology
 - Representative sequences

Representative sequences

- Aim: summarize a set of sequences
- Find a small set of sequences, such that
 - non redundant
 - cover a minimal percentage of the set
- Redundance and coverage defined in terms of neighborhood
 - x and y non redundant if $d(x, y) > \delta_{tsim}$
 - coverage: % of sequences that have at least one representative r in their neighborhood ($d(x, r) < \delta_{tsim}$)

Representative sequences

- Aim: summarize a set of sequences
- Find a small set of sequences, such that
 - non redundant
 - cover a minimal percentage of the set
- Redundance and coverage defined in terms of neighborhood
 - x and y non redundant if $d(x, y) > \delta_{tsim}$
 - coverage: % of sequences that have at least one representative r in their neighborhood ($d(x, r) < \delta_{tsim}$)

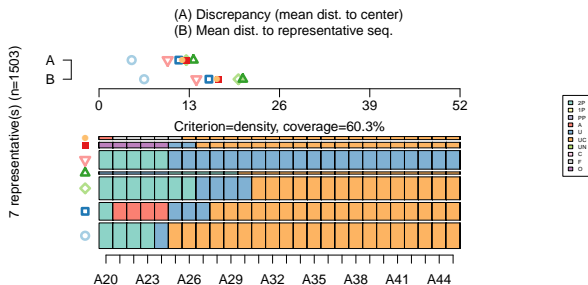
Representative sequences

Heuristic

- 1 **Sort sequences** according to a representativeness criterion
 - **density** number of sequences in its neighborhood
 - **centrality** sum of distances to all other sequences
 - others: frequency, mean of its state frequencies, likelihood, ...
- 2 **Suppress redundancy**
 - Compute coverage of the sequence with highest score
 - Then, for the next ones
 - drop out if redundant with sequences already retained
 - else, compute coverage of the new set of representatives
 - Stop when the wanted coverage is reached.

Representative sequences: Example

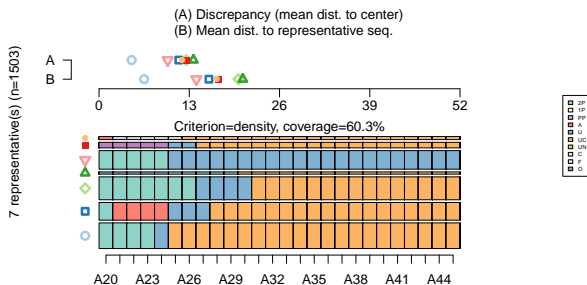
Cohabital trajectories (tsim=.2, trep=.6)



- With representative sequences, we miss small groups (Tanguy)

Representative sequences: Example

Cohabital trajectories (tsim=.2, trep=.6)

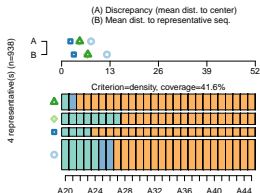


- With representative sequences, we miss small groups (Tanguy)

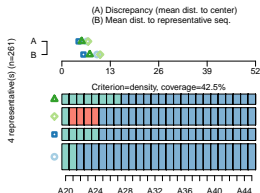
Representative sequences by cluster

Cohabital trajectories (tsim=.1, trep=.4)

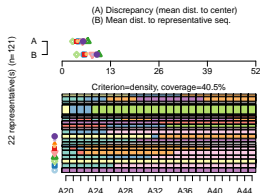
Type 1 : Parental Trajectories (62 %)



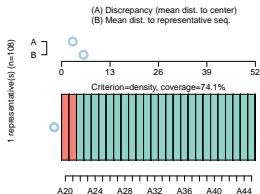
Type 2 : Conjugal Trajectories (17 %)



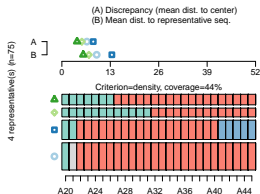
Type 3 : Mixed Cohabitation Trajectories (8 %)



Type 4 : Parental Home Trajectories (7 %)



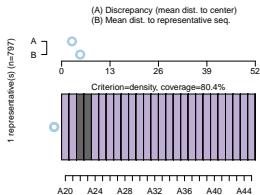
Type 5 : Solo Trajectories (5 %)



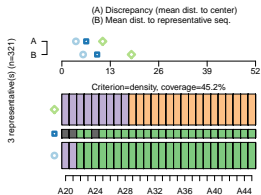
Representative sequences by cluster

Occupational trajectories (tsim=.1, trep=.4)

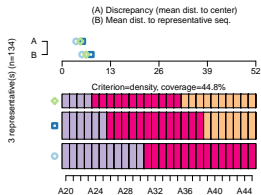
Type 1: Full Time Trajectories (52 %)



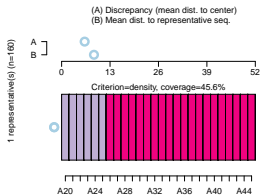
Type 2: Mixed Occupational Trajectories (22 %)



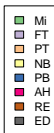
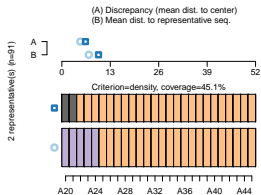
Type 3: Return Trajectories (11 %)



Type 4: At Home Trajectories (9.5 %)



Type 5: Part Time Trajectories (5.5 %)



Outline

- 1 Objectives
- 2 Pairwise dissimilarities between sequences
- 3 Unsupervised clustering
- 4 Supervised Classification**
- 5 Conclusion

Section outline

- 4 Supervised Classification
 - ANOVA on sequential data
 - Regression trees for categorical sequences

Association between sequences and a covariate

- For supervised clustering (according the value of a covariate)
- ... must be able to measure **association between sequential data and a covariate**
- Since we know how to determine the discrepancy

$$SS = \frac{1}{n} \sum_{i=1}^n \sum_{j=1+1}^n d_{ij}$$

- We can compute pseudo R^2 's and pseudo F 's

Association between sequences and a covariate

- For supervised clustering (according the value of a covariate)
- ... must be able to measure **association between sequential data and a covariate**
- Since we know how to determine the discrepancy

$$SS = \frac{1}{n} \sum_{i=1}^n \sum_{j=1+1}^n d_{ij}$$

- We can compute pseudo R^2 's and pseudo F 's

Analysis of sequence discrepancy

(Studer et al., 2009, 2010)

- 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

$$SS_T = \frac{1}{n} \sum_{i=1}^n \sum_{j=i+1}^n d_{ij}$$

$$SS_W = \sum_g \left(\frac{1}{n_g} \sum_{i=1}^{n_g} \sum_{j=i+1}^{n_g} d_{ij,g} \right)$$

$$SS_B = SS_T - SS_W$$

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 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 defensible in this setting.
- F cannot be compared with an F distribution.
- The significance is assessed 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 defensible in this setting.
- F cannot be compared with an F distribution.
- The significance is assessed 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

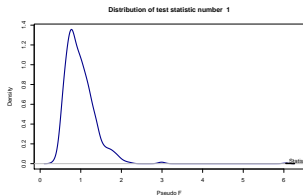
Cohabital trajectories with birth cohort

	SS	df	MSE
Exp	186.63	2	93.32
Res	23149.63	1500	15.43
Total	23336.26	1502	15.54

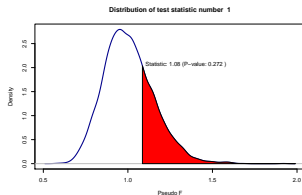
	t0	p.value
Pseudo F	5.057	0.001
Pseudo Fbf	5.851	0.001
Pseudo R2	0.007	0.001
Bartlett	8.731	0.001
Levene	14.122	0.001

Empirical F distribution

Cohabital trajectories with
birth cohort



Cohabital trajectories with
month of birth



Simple ANOVA, cohabitational trajectories

Simple ANOVA, cohabitational trajectories

	categories	R^2	F	Sig.
Birth Cohort	3	0.008	6.0	0.001
Sex	2	0.004	6.6	0.001
Education Level	4	0.007	3.3	0.001
Birth Month	12	0.009	1.1	0.272

Simple ANOVA, occupational trajectories

Simple ANOVA, occupational trajectories

	categories	R^2	F	Sig.
Birth Cohort	3	0.007	5.1	0.001
Sex	2	0.183	336.6	0.001
Education Level	4	0.065	34.8	0.001

Homogeneity of within group discrepancy

- Is discrepancy the same in all groups?
- Contribution to inertia $d_{x\tilde{g}} = \frac{1}{n} \left(\sum_i d_{xi} - SS \right)$
- Letting z_i be the dissimilarity between sequence i and its group center
- **Levene:** F test (ANOVA) on the z_i 's

Homogeneity of within group discrepancy

- Is discrepancy the same in all groups?
- Contribution to inertia $d_{x\tilde{g}} = \frac{1}{n} \left(\sum_i d_{xi} - SS \right)$
- Letting z_i be the dissimilarity between sequence i and its group center
- **Levene:** F test (ANOVA) on the z_i 's

Homogeneity of within group discrepancy

Levene test, Cohabital Trajectories

	categories	<i>L</i>	Sig.
Birth Cohort	3	0.1	0.915
Sex	2	9.2	0.002
Education Level	4	2.4	0.063
Birth Month	12	1.3	0.238

Homogeneity of within group discrepancy

Levene test, Occupational Trajectories

	categories	<i>L</i>	Sig.
Birth Cohort	3	14.1	0.001
Sex	2	912.8	0.001
Education Level	4	15.2	0.001

Multi-factor ANOVA

- Generalization to multi-factor case voir (Studer et al., 2010)
- Here, we consider Type II effects
- Measures contribution added by each factor v when we control for all the others.
- The F statistic is

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

where SS_{B_c} and SS_{W_c} are the explained and residual SS of the full model. SS_{B_v} the explained part of the model after deletion of v , and p the number of indicator or contrasts used for coding variable v .

- Significance is again evaluated through permutation tests.

Multi-factor ANOVA

- Generalization to multi-factor case voir (Studer et al., 2010)
- Here, we consider Type II effects
- Measures contribution added by each factor v when we control for all the others.
- The F statistic is

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

where SS_{B_c} and SS_{W_c} are the explained and residual SS of the full model. SS_{B_v} the explained part of the model after deletion of v , and p the number of indicator or contrasts used for coding variable v .

- Significance is again evaluated through permutation tests.

Multi-factor, derivation of the results

- Consider the linear model $\mathbf{Y} = \mathbf{X}\mathbf{B}$
- Its 'Hat' matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, such that $\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$
- Gower's matrix $\mathbf{G} = -\frac{1}{2}\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)\mathbf{D}\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)$
with \mathbf{D} matrix of squared Euclidean distances
- We have

$$SS_T = tr(\mathbf{G})$$

$$SS_B = tr(\mathbf{H}\mathbf{G})$$

$$SS_W = tr((\mathbf{I} - \mathbf{H})\mathbf{G})$$

- Generalization by substituting \mathbf{D} with the matrix of dissimilarities.

Multi-factor analysis

sex, birth cohort, level of education

Multi-factor analysis, Occupational trajectories

Variable	PseudoF	PseudoR2	p_value
sex	497.039	0.226	0.0000
cohort3b	5.281	0.005	0.0010
edu_lev	34.353	0.047	0.0000
Total	116.800	0.319	0.0000

Evolution of the differences

- How do differences vary over time?
- At which age do trajectories most differ between cohorts?
- Compute R^2 on small **sliding windows** (length 2)
- => **Series of R^2** , and we plot their evolution
- Likewise we can plot the series of
 - the total residual discrepancies (SS_W)
 - the residual discrepancy of each group (SS_G)

Evolution of the differences

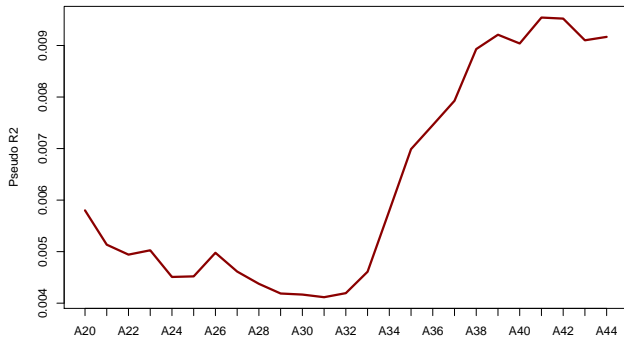
- How do differences vary over time?
- At which age do trajectories most differ between cohorts?
- Compute R^2 on small **sliding windows** (length 2)
- => **Series of R^2** , and we plot their evolution
- Likewise we can plot the series of
 - the total residual discrepancies (SS_W)
 - the residual discrepancy of each group (SS_G)

Evolution of the differences

- How do differences vary over time?
- At which age do trajectories most differ between cohorts?
- Compute R^2 on small **sliding windows** (length 2)
- => **Series of R^2** , and we plot their evolution
- Likewise we can plot the series of
 - the total residual discrepancies (SS_W)
 - the residual discrepancy of each group (SS_G)

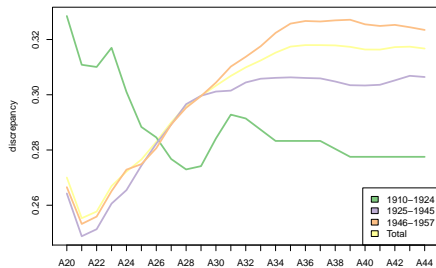
Series of R-squares

R^2 , Occupational Trajectories, Birth cohort



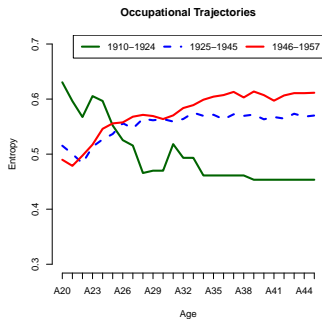
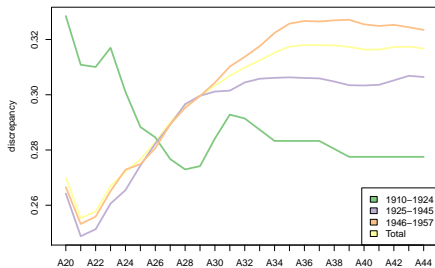
Series of residual discrepancies

Birth cohort



Series of residual discrepancies

Birth cohort



Section outline

- 4 Supervised Classification
 - ANOVA on sequential data
 - Regression trees for categorical sequences

Tree structured analysis of sequence data

- Aim: Find out most important predictors and their interactions.
- Iteratively segment cases using covariate values
- Form as homogeneous groups.

- at each step, select covariate and split that generates the highest R^2 .
- Split significance tested with permutation F .
- Stop when the selected split is not significant.

Tree structured analysis of sequence data

- Aim: Find out most important predictors and their interactions.
- Iteratively segment cases using covariate values
- Form as homogeneous groups.

- at each step, select covariate and split that generates the highest R^2 .
- Split significance tested with permutation F .
- Stop when the selected split is not significant.

Growing the tree

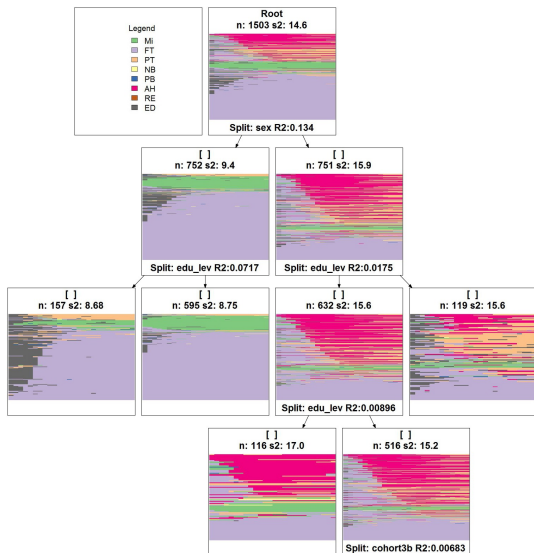
```

|-- Root (n: 1503 disc: 15)
  |--> sex 0.13
    |-- [ ] (n: 752 disc: 9.4)
      |--> edu_lev 0.072
        |-- [ ] (n: 157 disc: 8.7) [(ED,6)-(FT,20)] *
        |-- [ ] (n: 595 disc: 8.7) [(FT,26)] *
      |-- [ ] (n: 751 disc: 16)
        |--> edu_lev 0.017
          |-- [ ] (n: 632 disc: 16)
            |--> edu_lev 0.009
              |-- [ ] (n: 116 disc: 17) [(FT,9)-(AH,17)] *
              |-- [ ] (n: 516 disc: 15)
                |--> cohort3b 0.0068
                  |-- [ ] (n: 280 disc: 15) [(FT,10)-(AH,10)-(PT,6)] *
                  |-- [ ] (n: 236 disc: 15) [(FT,12)-(AH,14)] *
                |-- [ ] (n: 119 disc: 16) [(ED,1)-(FT,14)-(PT,11)] *

```

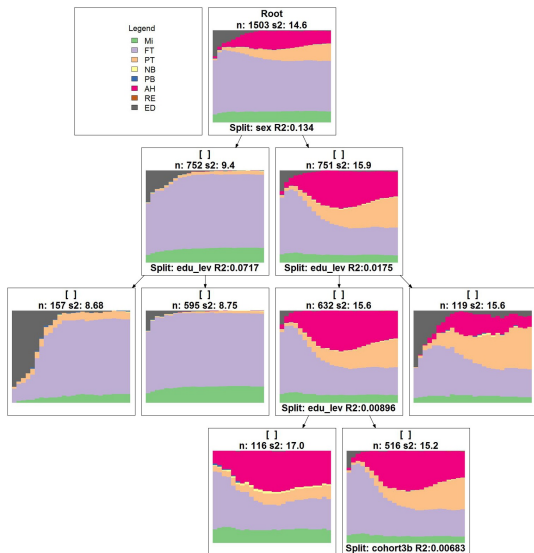

Rendering the tree

Occupational trajectories



Rendering the tree

Occupational trajectories



Quality of the tree

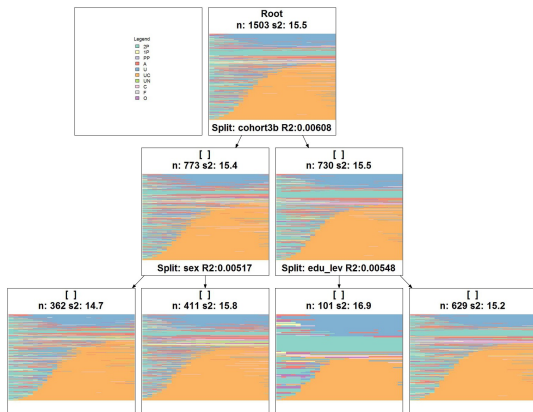
Occupational trajectories

ANOVA for the leaves of the tree,

	t0	p.value
Pseudo F	62.49	0.000
Pseudo Fbf	55.66	0.000
Pseudo R2	0.17	0.000
Bartlett	60.60	0.000
Levene	43.66	0.000

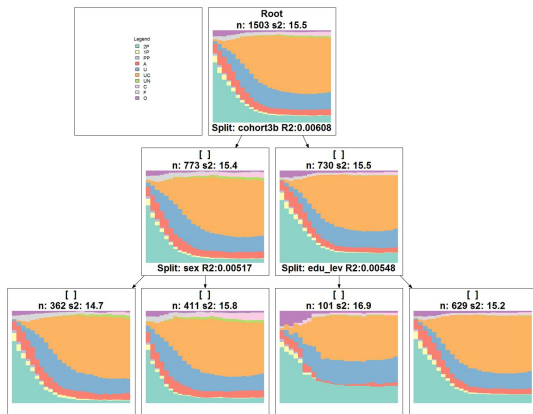
Rendering the tree

Cohabital trajectories



Rendering the tree

Cohabital trajectories



Quality of the tree

Cohabital trajectories

ANOVA for leaves of the tree,

	t0	p.value
Pseudo F	5.74	0.000
Pseudo Fbf	5.62	0.000
Pseudo R2	0.01	0.000
Bartlett	0.96	0.049
Levene	2.73	0.041

Outline

- 1 Objectives
- 2 Pairwise dissimilarities between sequences
- 3 Unsupervised clustering
- 4 Supervised Classification
- 5 Conclusion**

Conclusion 1: About sequence analysis

- Analyse trajectories until 45 years => **ignore recent generations**
- Most recent birth year is **1957** (2002 – 45)
- Issues:
 - **Granularity**: year, month, day, ...
 - **State definition**: should we distinguish {separated, divorced, widowed} or consider a single state? works by Raffaella Piccaretta

Conclusion 1: About sequence analysis

- Analyse trajectories until 45 years => **ignore recent generations**
- Most recent birth year is **1957** (2002 – 45)
- Issues:
 - **Granularity**: year, month, day, ...
 - **State definition**: should we distinguish {separated, divorced, widowed} or consider a single state? works by Raffaella Piccaretta

Conclusion 2: Missing data and weights

- **Missing data** in sequences
- TraMineR allows for differentiated handling of left, right and in-between missing values
 - consider 'missing' as a specific state
 - drop out (left shift of subsequent elements)
 - impute, but how?
- **Weighting cases**
 - Account for them in rendering of sequences (weighted transversal characteristics)
 - Implemented solutions for ANOVA and permutation test
 - Not relevant for dissimilarities and longitudinal characteristics

Conclusion 2: Missing data and weights

- **Missing data** in sequences
- TraMineR allows for differentiated handling of left, right and in-between missing values
 - consider 'missing' as a specific state
 - drop out (left shift of subsequent elements)
 - impute, but how?
- **Weighting cases**
 - Account for them in rendering of sequences (weighted transversal characteristics)
 - Implemented solutions for ANOVA and permutation test
 - Not relevant for dissimilarities and longitudinal characteristics

Conclusion 3: Extending the analysis

- Since TraMineR is an R library, its outcome can easily be combined in a same script with any other R process
- We have seen: cluster analysis, MDS, ...
- In Widmer and Ritschard (2009),
 - Relationship between **occupational** and **cohabitational** trajectories by regressing longitudinal entropies of each of them on occupational and cohabitational types while controlling for birth cohort and sex.
 - Studied also **cluster membership** with logistic regressions.

Conclusion 3: Extending the analysis

- Since TraMineR is an R library, its outcome can easily be combined in a same script with any other R process
- We have seen: cluster analysis, MDS, ...
- In Widmer and Ritschard (2009),
 - Relationship between **occupational** and **cohabitational** trajectories by regressing longitudinal entropies of each of them on occupational and cohabitational types while controlling for birth cohort and sex.
 - Studied also **cluster membership** with logistic regressions.

Conclusion 3: Extending the analysis

- Since TraMineR is an R library, its outcome can easily be combined in a same script with any other R process
- We have seen: cluster analysis, MDS, ...
- In Widmer and Ritschard (2009),
 - Relationship between **occupational** and **cohabitational** trajectories by regressing longitudinal entropies of each of them on occupational and cohabitational types while controlling for birth cohort and sex.
 - Studied also **cluster membership** with logistic regressions.

Conclusion 4: Application to other kind of data

- Discrepancy based analysis
- ... applies to any data that can be characterized by their pairwise dissimilarities.
- Only aspect specific to state sequences: their visual rendering.

Conclusion 4: About TraMineR

- **TraMineR** is a unique toolbox for discrete sequence analysis
- Can do much more than shown in this presentation,
 - handling of sequence data
 - conversion between states and events
 - multi-channel dissimilarity for parallel sequences
 - frequent and discriminant sub-sequences
 - extracting association rules between sub-sequences
 - ...
- ... and, like **R**, available for free on the **CRAN**
<http://cran.r-project.org>
- See also package's web site
<http://mephisto.unige.ch/traminer>

Conclusion 4: About TraMineR

- **TraMineR** is a unique toolbox for discrete sequence analysis
- Can do much more than shown in this presentation,
 - handling of sequence data
 - conversion between states and events
 - multi-channel dissimilarity for parallel sequences
 - frequent and discriminant sub-sequences
 - extracting association rules between sub-sequences
 - ...
- ... and, like **R**, available for free on the **CRAN**
<http://cran.r-project.org>
- See also package's web site
<http://mephisto.unige.ch/traminer>

Thank You!

References I

- Gabadinho, A., G. Ritschard, M. Studer, and N. S. Müller (2009a). Mining sequence data in R with the TraMineR package: A user's guide. Technical report, Department of Econometrics and Laboratory of Demography, University of Geneva, Geneva.
- Gabadinho, A., G. Ritschard, M. Studer, and N. S. Müller (2009b). Summarizing sets of categorical sequences. In *International Conference on Knowledge Discovery and Information Retrieval, Madeira, 6-8 October, 2009*, pp. 62–69. INSTICC. (Received the Best Paper Award).
- Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* 53(3/4), 325–338.
- Kaufman, L. and P. J. Rousseeuw (2005). *Finding Groups in Data*. Hoboken: Wiley.
- Lesnard, L. (2006). Optimal matching and social sciences. Série des Documents de Travail du CREST 2006-01, Institut National de la Statistique et des Etudes Economiques, Paris.
- Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10, 707–710.

References II

- Maechler, M., P. Rousseeuw, A. Struyf, and M. Hubert (2005). Package 'cluster': Cluster analysis basics and extensions. Reference manual, R-project, CRAN.
- 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, et N. S. Müller (2009). Analyse de dissimilarités par arbre d'induction. *Revue des nouvelles technologies de l'information RNTI E-15*, 7–18.
- Studer, M., G. Ritschard, A. Gabadinho, et N. S. Müller (2010). Discrepancy analysis of complex objects using dissimilarities. In F. Guillet, G. Ritschard, D. A. Zighed, et H. Briand (Eds.), *Advances in Knowledge Discovery and Management*, Volume 292 of *Studies in Computational Intelligence*, pp. 3–19. Berlin : Springer.
- 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.