Beyond the search of ideal typical sequences

Analyzing, interpreting and visualizing relationships between sequences and explanatory factors using discrepancy analysis

Matthias Studer & Reto Bürgin

Institute for demographic and life course studies
University Geneva
http://mephisto.unige.ch/traminer

Lausanne Conference on Sequence Analysis (LaCOSA),
Lausanne, 6th of June 2012

Outline

1. Introduction
   - Objectives
   - Example problematic
2. Cluster-based strategy
   - Clustering
   - Testing differences between groups of trajectories
3. Discrepancy analysis
   - Introduction
4. Interpreting the differences
   - Interpreting differences
   - Implicative Statistics
   - Residual gains
5. Extensions
   - Regression trees
6. Conclusion
   - Conclusion

Objectives

- Studying the relationships between trajectories and explanatory factors.
  - Sex and academic carrier.
  - Cohort and familial trajectory.
  - Social origin and sequence of school to work transitions.
- Multifactor approach.
  - Sex, horizontal segregation and academic carrier.

Objectives of the presentation:

- The usual cluster-based strategy.
- The discrepancy analysis framework.
- Present new tools to interpret the results.

Objectives

- Studying the relationships between trajectories and explanatory factors.
  - Sex and academic carrier.
  - Cohort and familial trajectory.
  - Social origin and sequence of school to work transitions.
- Multifactor approach.
  - Sex, horizontal segregation and academic carrier.

Objectives of the presentation:

- The usual cluster-based strategy.
- The discrepancy analysis framework.
- Present new tools to interpret the results.

Example problematic

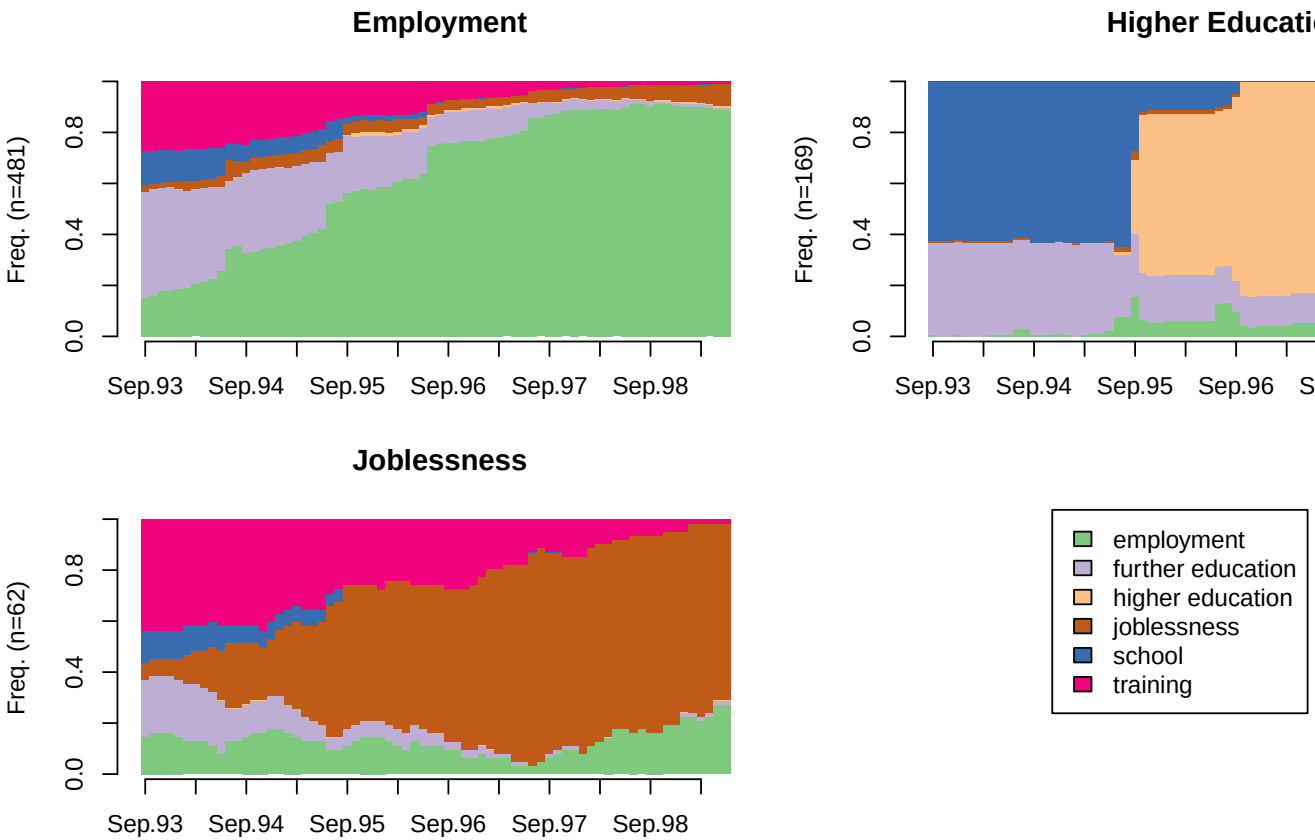
- Study of sequences of school-to-work transitions in North Irland by McVicar and Anyadike-Danes (2002).
- Focus on father unemployment: do we observe some kind of joblessness transmission?
- Sequences begin at the end of compulsory school.
- Length: 70 months.
- Alphabet (states): EM (Employment), FE (Further Education), HE (Higher Education), JL (Joblessness), SC (School), TR (Training).

Outline

1. Introduction
   - Objectives
   - Example problematic
2. Cluster-based strategy
   - Clustering
   - Testing differences between groups of trajectories
3. Discrepancy analysis
   - Introduction
4. Interpreting the differences
   - Interpreting differences
   - Implicative Statistics
   - Residual gains
5. Extensions
   - Regression trees
6. Conclusion
   - Conclusion

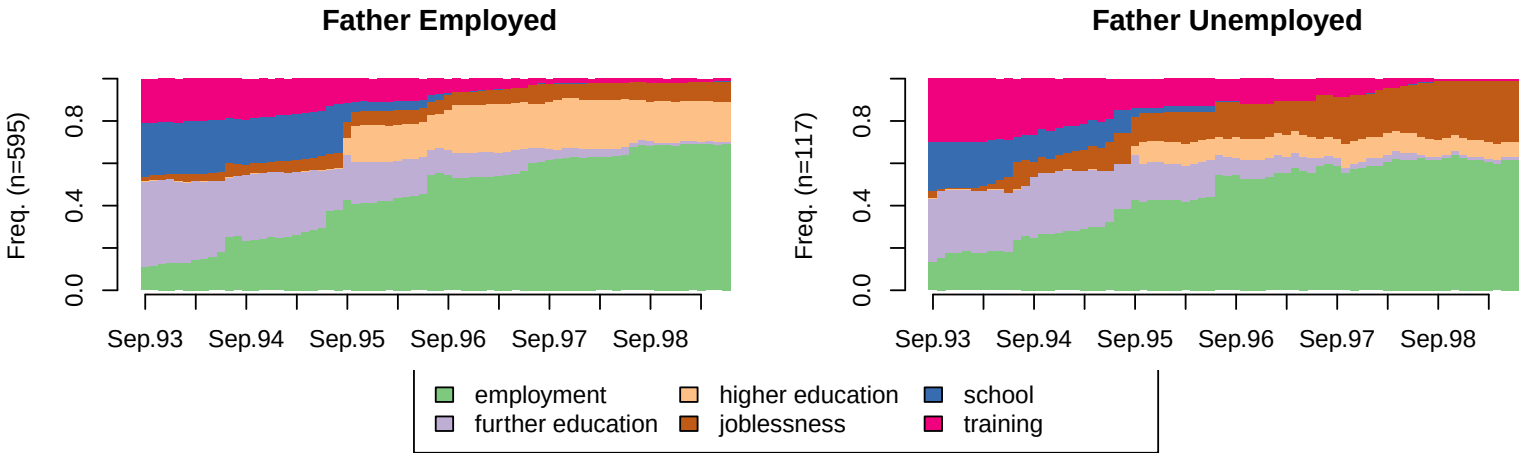
Usual cluster-based strategy

- Start by building a typology of the sequences using cluster analysis.
- Keep three groups (best Average Silhouette Width=0.41).

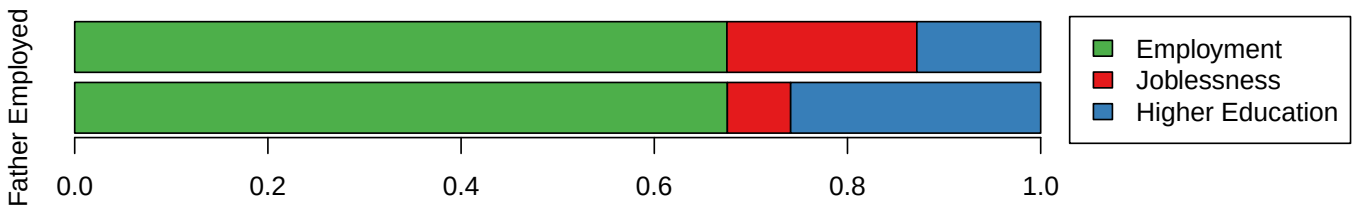


Comparing groups of sequences: father unemployment

- To test if the trajectories differ according to father unemployment status...



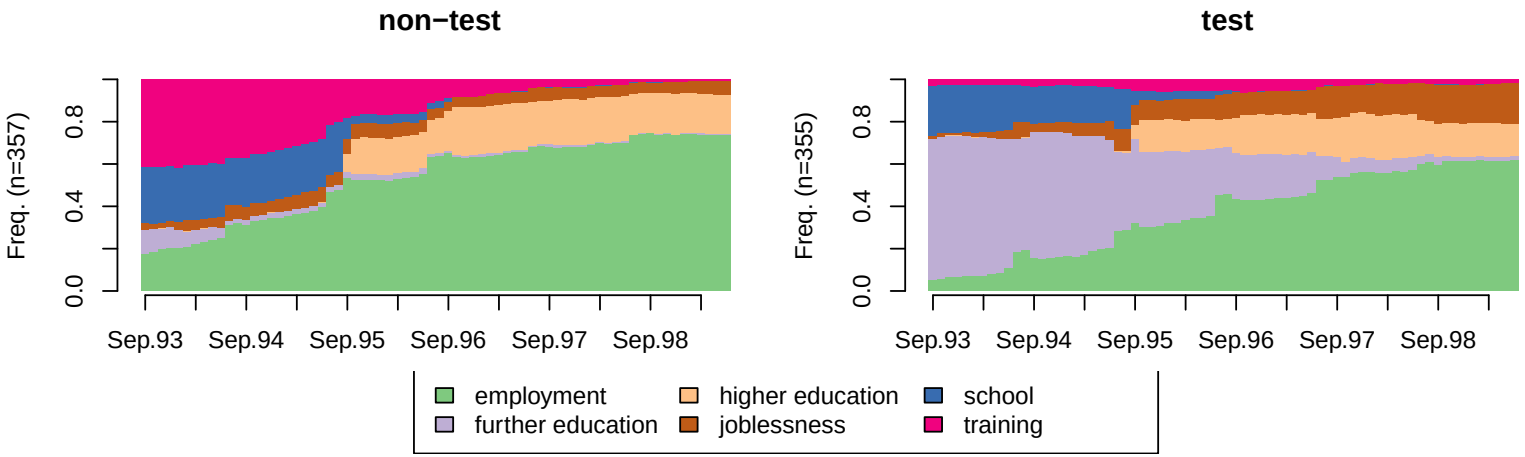
- We test this:



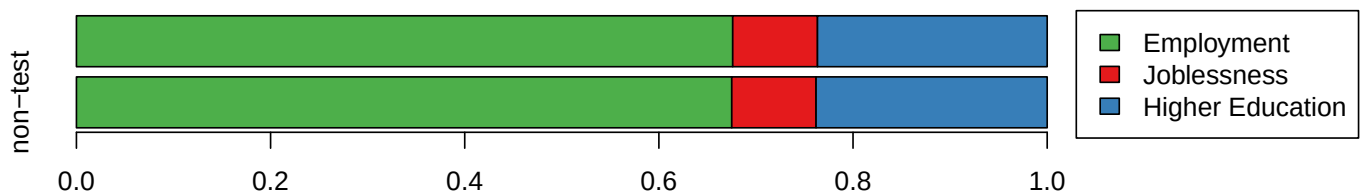
- Chi-square p-value: 1.939×10^{-6}, Cramer's v: 0.192

No relationships with the test factor ?

- To test if the trajectories differ according to the test factor...
- The test factor was artificially built using an MDS.



- We test this:



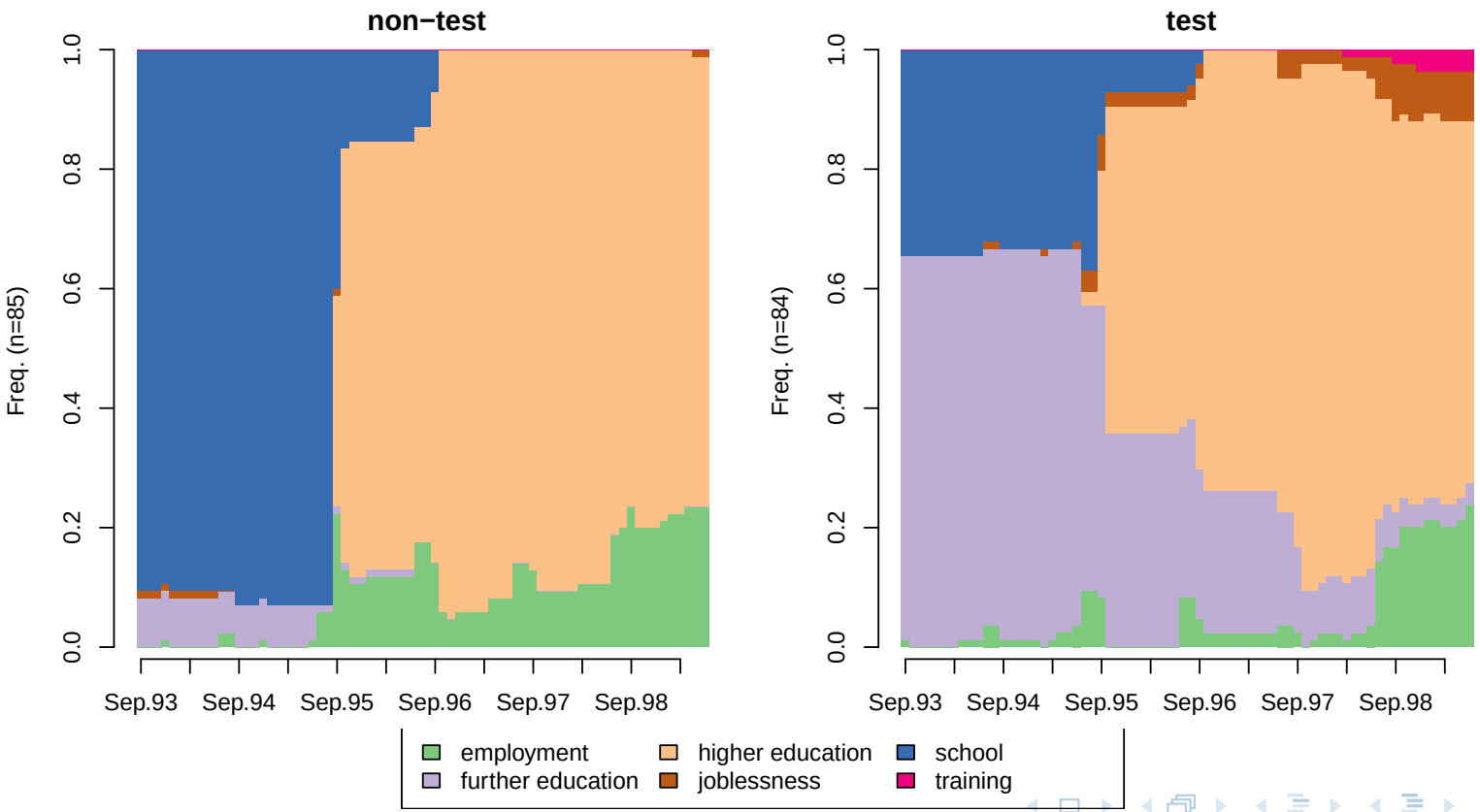
- Chi-square p-value: 0.999, Cramer's v: 0.002

What's the problem?

- The clustering is a simplification of the sequences.
- By using the typology in the chi-square test, implicit assumptions are made:
  - The typology carries enough information to describe all the differences between sequences.
  - I.e. The sequences in each groups are all equals.
  - The groups are equally different.
- But the "test" covariate explains the differences of sequences inside each groups.
- In this case, the simplification is abusive.

Clustering and test factor

- Example: difference inside cluster “Higher Education” according to test factor.
- All sequences are not equals in this cluster!
- The same applies to the other clusters.



What's the problem?

- These assumptions may be correct if:
  - If the clusters are very homogeneous.
  - **And** if the clusters are clearly separated from each other.
- Otherwise, this simplification may hide or create an association with a covariate.

Sociological assumptions

- We assign to each individual sequence an ideal-type of sequences.
- The difference between the individual sequence and the ideal-type is ignored.
- Justification:
  - The sequence are realization of well defined ideal-types (common patterns) of sequences (Abbott and Forrest, 1986).
  - I.e. The sequence were generated following a well defined model (common pattern) of trajectories.
  - The small differences between individual sequence and common pattern can be ignored because:
  - These common patterns (models) were correctly identified by the cluster analysis.
  - Intra-cluster variability is a kind of uninformative error term.
- These are rather strong assumptions!

Sociological assumptions

Link to complexity of sequential models (Abbott, 1992).

- Natural histories.
  - Sequences are strongly structured.
  - Example: sequence of compulsory schooling.
  - Cluster based strategy may be ok.
  - But what if small accidents (repeating a grade) only happen to those from a peculiar social origin?
- More complexes trajectories.
  - Lot of different trajectories.
  - Some sequences are often between two ideal-types.
  - These assumptions are too strong.

Why using cluster analysis?

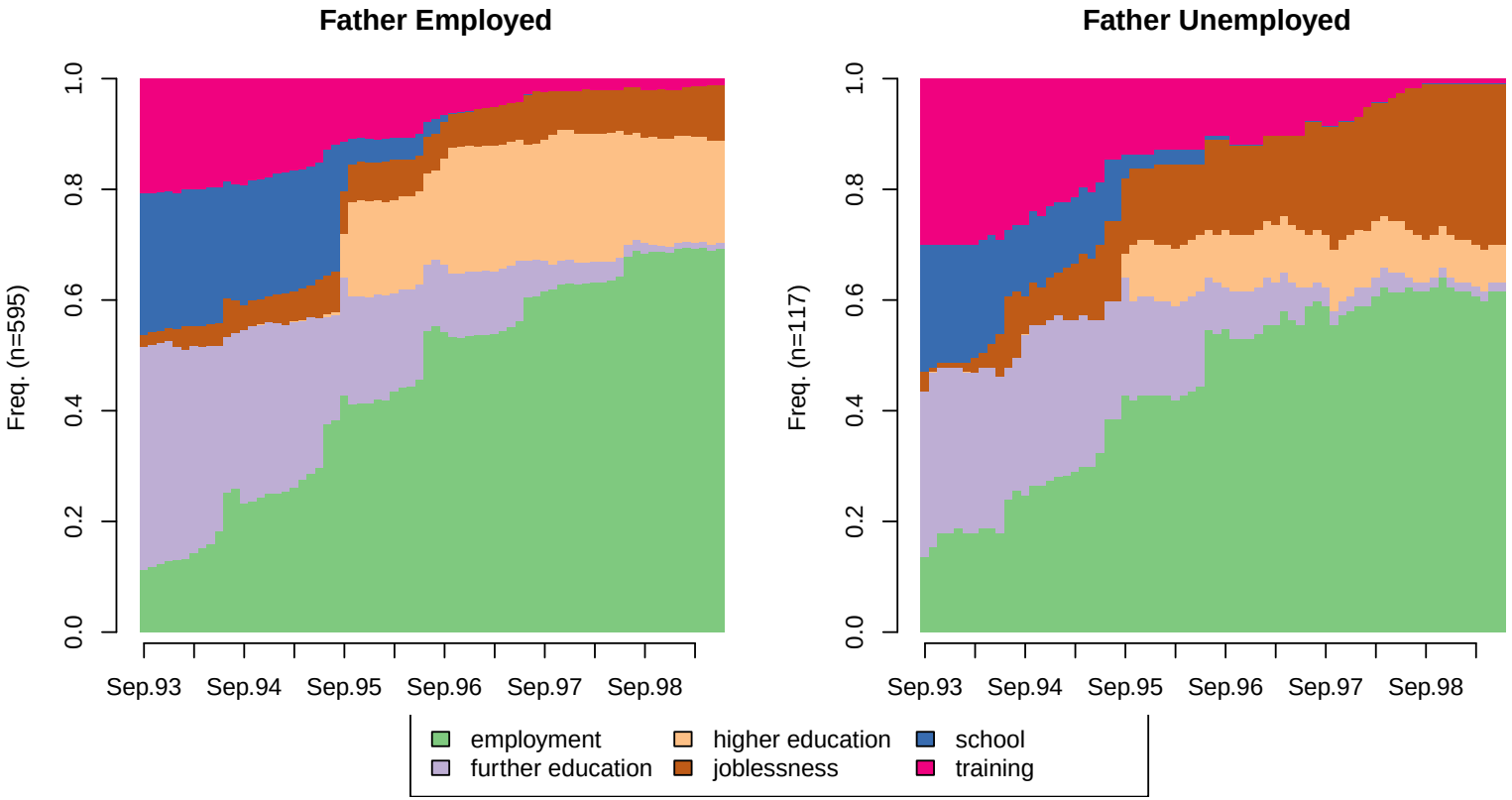
- Cluster analysis is a *descriptive* method.
  - Find some common patterns that may results from social, legal or economical constraints on trajectories.
  - Identify the main common patterns that may act as models of trajectories (Abbott and Hrycak, 1990).
  - Common patterns may highlight dependencies between different moment in the trajectories.
- These kind of interpretations are powerful and can be done using cluster analysis.
- The problematic assumptions are made when the typology is used in an explanatory analysis.

Outline

1. Introduction
   - Objectives
   - Example problematic
2. Cluster-based strategy
   - Clustering
   - Testing differences between groups of trajectories
3. Discrepancy analysis
   - Introduction
4. Interpreting the differences
   - Interpreting differences
   - Implicative Statistics
   - Residual gains
5. Extensions
   - Regression trees
6. Conclusion
   - Conclusion

Discrepancy Analysis

- Aim: study the relationships between states sequences and explanatory variables without prior clustering.



General principles

- Define a measure of the discrepancy of the sequences based on the distance matrix.

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

- Measure the strength of the relationship using the share of the total discrepancy that is explained by this covariate.
- Attest the significance of the relationship using permutation tests.
- This method is a generalization of the ANOVA.
- The method can be extended to
  - Include several covariates at the same time (control for the effect of other covariates).
  - Build regression trees.

Discrepancy Analysis Results

	F	R^2	p-value
test	71.56	0.092	0.000
funemp	9.51	0.013	0.000

With discrepancy analysis, we can measure:

- The strength of the relationship with the R^2.
- The statistical significance using permutation tests.

Underlying assumptions

- Main sequence analysis are preserved because of the use of distance.
- Holistic view (Abbott, 1992)
  - Highlight the endogenous dynamic (logic) of trajectories (enchainment).
  - Analyze the order of sequences (order).
- Relationships between sequences and exploratory factors.
  - Effect of exogenous condition on endogenous logic.
  - Trajectories are shaped by social and/or historic context (Elder, 1999).
- Focus on discrepancy:
  - Preserve the notion of inter-individual variability of the individual within these context.
  - Inside that context individual are able to make choices (Elder, 1999).

When use discrepancy analysis?

- If there is an interest in the effect of exogenous factors.
- More complex sequential models.
- For instance, we may think that:
  - Some contexts influence the beginning of the trajectories.
  - Others influence the end.
  - People are inserted in many different contexts that influence, each on its own way, the trajectories.
  - This leads to a large amount of individual trajectories.
- Using discrepancy analysis, less assumptions are made!

Outline

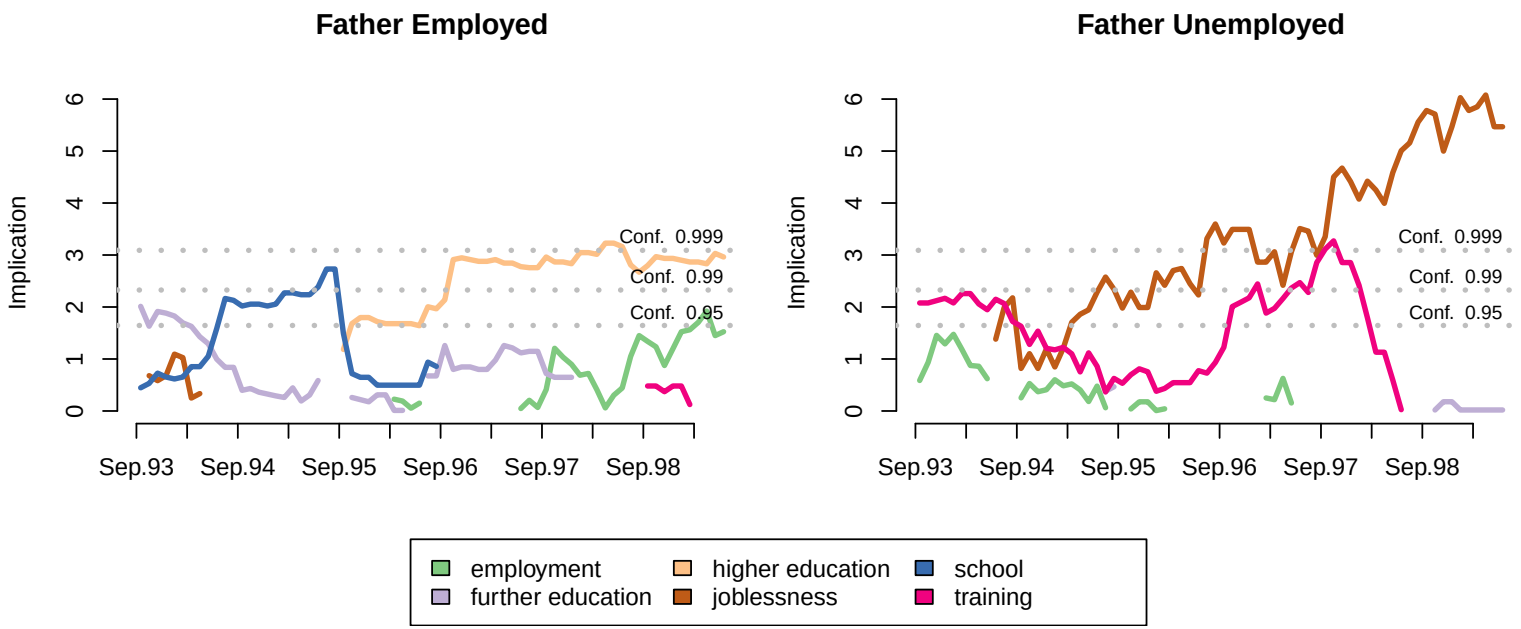
1. Introduction
   - Objectives
   - Example problematic
2. Cluster-based strategy
   - Clustering
   - Testing differences between groups of trajectories
3. Discrepancy analysis
   - Introduction
4. Interpreting the differences
   - Interpreting differences
   - Implicative Statistics
   - Residual gains
5. Extensions
   - Regression trees
6. Conclusion
   - Conclusion

Interpreting differences

- The sequences are significantly different according to father unemployment.
  - What does this mean?
  - What are the differences?

Implicative Statistic

- Aim: Measure the relevance of the rule "C implies A_t" (Gras, 1979).
- Let C be one of the category of the explanatory variable.
- Let A_t denote being in state A at time t
- In each plot C, we represent the evolution of the relevance of the rule "C implies A_t".

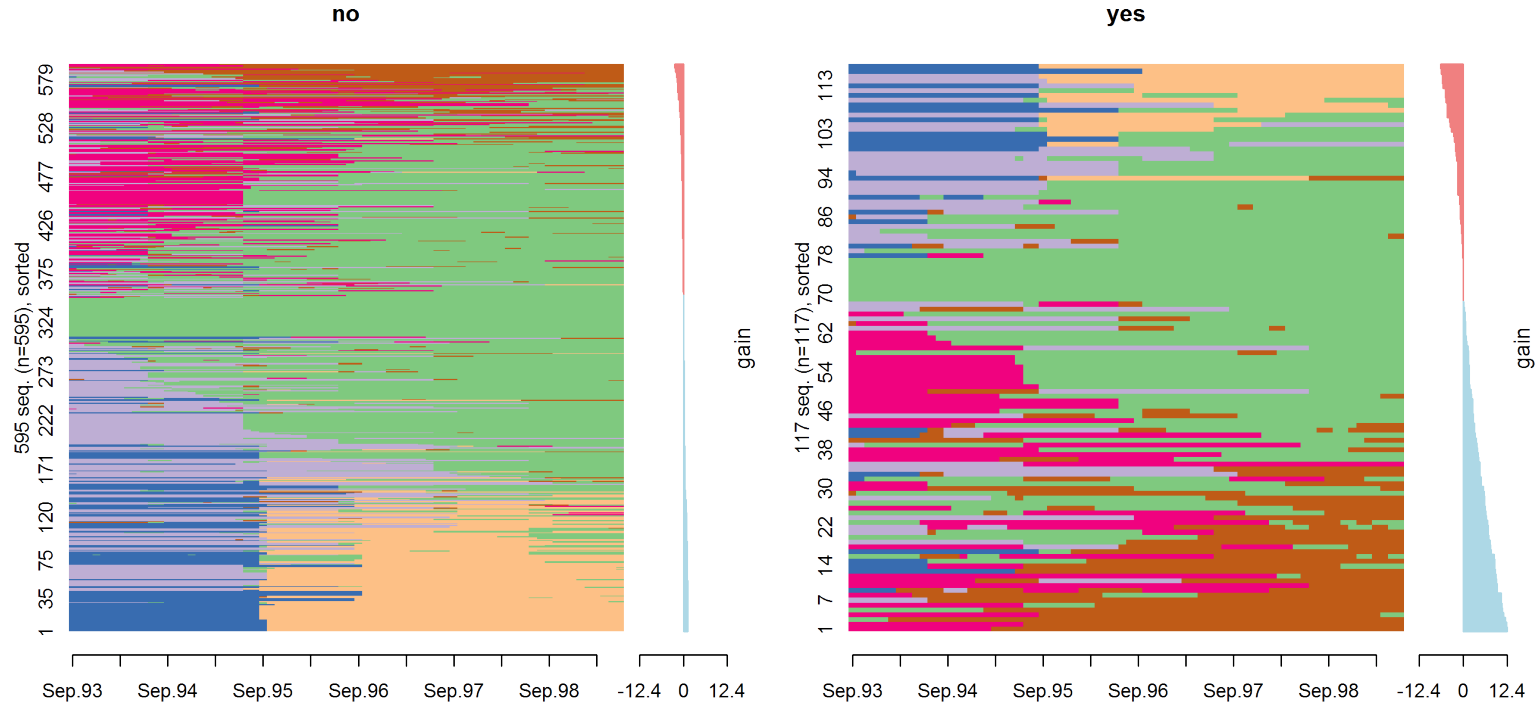


Implicative Statistic

- This plot makes the interpretation easier.
- Highlights the most important differences in occupied states.
- Similar to distribution plot (chronogram), it is a sequence of transversal characteristics.
- We loose all longitudinal information!
- Are there patterns that differ according to father unemployment?

Residual gain

- Residual gain: Gain for each sequences to take an explanatory factor into account.
- Residual gain are computed from discrepancy analysis.
- Plot of the sequences sorted according to residual gains.
- Sequences at the bottom are the most typical of each profile (highest gain).



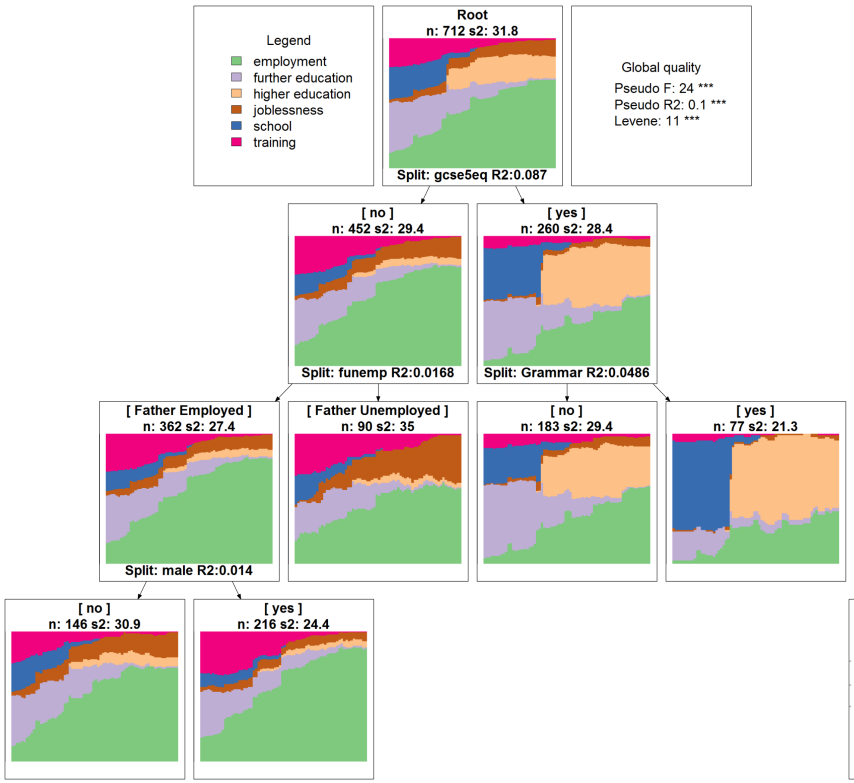
Residual gain

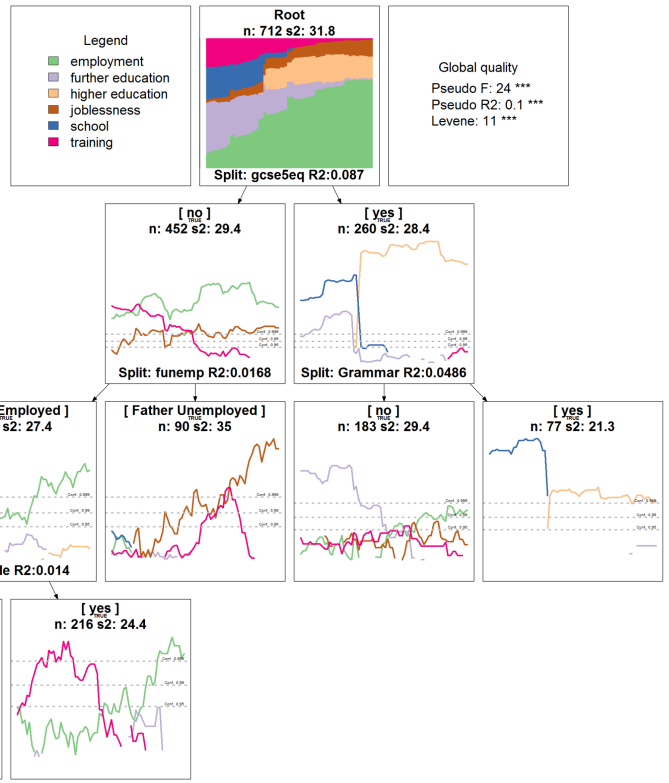
- Interpretation: tendencies to follow the kind of sequences that have the highest gains.
- Possible extensions:
  - Multifactor analysis: regression like interpretation of the effect (proper effect).
  - Tree structured analysis.
- Residual gain allows to visualize the relationships between explanatory factors and sequences.

Outline

1. Introduction
   - Objectives
   - Example problematic
2. Cluster-based strategy
   - Clustering
   - Testing differences between groups of trajectories
3. Discrepancy analysis
   - Introduction
4. Interpreting the differences
   - Interpreting differences
   - Implicative Statistics
   - Residual gains
5. Extensions
   - Regression trees
6. Conclusion
   - Conclusion

Regression trees





Outline

1. Introduction
   - Objectives
   - Example problematic
2. Cluster-based strategy
   - Clustering
   - Testing differences between groups of trajectories
3. Discrepancy analysis
   - Introduction
4. Interpreting the differences
   - Interpreting differences
   - Implicative Statistics
   - Residual gains
5. Extensions
   - Regression trees
6. Conclusion
   - Conclusion

Conclusion I

- Cluster analysis.
  - Descriptive analysis.
  - Problematic assumptions in explanatory analysis.
- With discrepancy analysis, we can estimate:
  - The strength of the relationship with the R^2.
  - The statistical significance using permutation tests.
- The test are much more powerful than the one computed using clusters.
- Analyze directly the links between the trajectories and the covariates.
- Brings an explanatory framework in sequences analysis.
- Preserve the inter-individual variability of the trajectories while studying the relationships between the trajectories and their contexts.

Conclusion II

- Tools to interpret the relationship.
  - Implicative statistic plots.
  - Residual gain plots.

References I

Abbott, A. (1992). From causes to events: Notes on narrative positivism. *Sociological Method and Research 20*, 428.

Abbott, A. and J. Forrest (1986). Optimal matching methods for historical sequences. *Journal of Interdisciplinary History 16*, 471–494.

Abbott, A. and A. Hrycak (1990). Measuring resemblance in sequence data: An optimal matching analaysis of musician's carrers. *American Journal of Sociolgy 96*(1), 144–185.

Elder, G. H. (1999). *Children of the Great Depression*. Boulder: Westview Press.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika 53*(3/4), 325–338.

References II

Gower, J. C. (1982). Euclidean distance geometry. *Mathematical Scientist 7*, 1–14.

Gras, R. (1979). *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques*. Thèse d'état, Université de Rennes 1, France.

Gras, R., S. Ag Almouloud, M. Bailleul, A. Laher, M. Polo, H. Ratsimba-Rajohn, and A. Totohasina (1996). *L'implication statistique: Nouvelle méthode exploratoire de données*. Recherches en didactique des mathématiques. Grenoble: La pensée sauvage.

McArdle, B. H. and M. J. Anderson (2001). Fitting multivariate models to community data: A comment on distance-based redundancy analysis. *Ecology 82*(1), 290–297.

References III

McVicar, D. and M. Anyadike-Danes (2002). Predicting successful and unsuccessful transitions from school to work using sequence methods. *Journal of the Royal Statistical Society A 165*(2), 317–334.

Studer, M., G. Ritschard, A. Gabadinho, and 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, and N. S. Müller (2010). Discrepancy analysis of complex objects using dissimilarities. In F. Guillet, G. Ritschard, D. A. Zighed, and H. Briand (Eds.), *Advances in Knowledge Discovery and Management*, Volume 292 of *Studies in Computational Intelligence*, pp. 3–19. Berlin: Springer.

References IV

Studer, M., G. Ritschard, A. Gabadinho, and N. S. Müller (2011). Discrepancy analysis of state sequences. *Sociological Methods and Research 40*(3), 471–510.

Suzuki, E. and Y. Kodratoff (1998). Discovery of surprising exception rules based on intensity of implication. In J. M. Zytkow and M. Quafafou (Eds.), *Principles of Data Mining and Knowledge Discovery, Second European Symposium, PKDD '98, Nantes, France, September 23-26, Proceedings*, pp. 10–18. Berlin: Springer.

Mathematical developments

- Using MANOVA, we can write (McArdle and Anderson, 2001).
- $tr(\mathbf{Y}'\mathbf{Y}) = tr(\widehat{\mathbf{Y}}'\widehat{\mathbf{Y}}) + tr(\mathbf{R}'\mathbf{R})$
- Which can be rewritten (McArdle and Anderson, 2001):
- $tr(\mathbf{Y}\mathbf{Y}') = tr(\mathbf{H}\mathbf{Y}\mathbf{Y}') + tr\left[(\mathbf{I} - \mathbf{H})\mathbf{Y}\mathbf{Y}'\right]$
- Let $\mathbf{Y}$, a $n \times q$ matrix of q centered variables.
- $\mathbf{Y}\mathbf{Y}' = \mathbf{G}$ (Gower, 1966, 1982).
- $\mathbf{G} = -\frac{1}{2}(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{D}^2(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}')$
- By replacing $\mathbf{D}^2$ by another distance measure, we get the discrepancy analysis framework.

Mathematical developments

- Using MANOVA, we can write (McArdle and Anderson, 2001).
- $tr(\mathbf{Y}'\mathbf{Y}) = tr(\widehat{\mathbf{Y}}'\widehat{\mathbf{Y}}) + tr(\mathbf{R}'\mathbf{R})$
- Which can be rewritten (McArdle and Anderson, 2001):
- $tr(\mathbf{Y}\mathbf{Y}') = tr(\mathbf{H}\mathbf{Y}\mathbf{Y}') + tr\left[(\mathbf{I} - \mathbf{H})\mathbf{Y}\mathbf{Y}'\right]$
- Let $\mathbf{Y}$, a $n \times q$ matrix of q centered variables.
- $\mathbf{Y}\mathbf{Y}' = \mathbf{G}$ (Gower, 1966, 1982).
- $\mathbf{G} = -\frac{1}{2}(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{D}^{\mathbf{2}}(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}')$
- By replacing $\mathbf{D}^{\mathbf{2}}$ by another distance measure, we get the discrepancy analysis framework.

Mathematical developments II

$$SS_T = SS_B + SS_W \quad (1)$$

With:

$$SS_T = tr(\mathbf{G}) \quad (2)$$

$$SS_B = tr(\mathbf{HG}) \quad (3)$$

$$SS_W = tr\left[(\mathbf{I} - \mathbf{H})\mathbf{G}\right] \quad (4)$$

(5)

- Diagonal element of $\mathbf{G}$ are residuals of the null model.
- Diagonal element of $(\mathbf{I} - \mathbf{H})\mathbf{G}$ are residuals of the fitted model.
- Residual interpreted as distance to the center of class.
- Diagonal of $-\mathbf{HG}$ are the gains in term of residual if we use a given explanatory covariate.