Comparing and classifying personal life courses
From time to event methods to sequence analysis

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

Dpt of Econometrics and LaboDemo, University of Geneva, Switzerland

Outline

http://mephisto.unige.ch

1 Aim of the presentation

Well-being relies on the dynamics of life courses

$\Rightarrow$ We have to analyse life courses.

Survey of possible approaches, with focus on new data-mining-based ones.

Survival Methods (Event History Analysis, Blossfeld and Rohwer (2002))

- Focus on a specific event (marriage, childbirth, starting new job, ...).
- How does the hazard of experiencing the event evolve with time and other personal characteristics?

Sequence Analysis (Holistic approach, Billari (2005))

- Focus on whole sequences of family, professional, education ... events.
- Clustering sequences (optimal matching),
- Sequencing, Characteristic subsequences and their relationships with personal characteristics.

What is data mining?

Concerned with characterization of interesting patterns

- per se (unsupervised learning)
  - Clustering
  - Frequent itemsets
  - Association rules
- for classification or prediction purposes (supervised learning)
  - Decision trees
  - Bayesian networks
  - SVM and Kernel Methods
  - CBR (case based reasoning), K-NN (k nearest neighbors)

Proceeds mainly heuristically.

Unlike statistical modeling, makes no assumptions about process generating the data.

Typology of methods for individual longitudinal data

questions	nature of data: time stamped event	nature of data: state/event sequences
descriptive	- Survival curves: Parametric (Weibull, Gompertz) and non parametric (Kaplan-Meier, Nelson-Aalen) estimators	- Optimal matching clustering - Frequencies of typical patterns - Discovering typical patterns
causality	- Hazard regression models - Survival trees	- Markov models, Mobility trees - Association rules between subsequences

2 Three examples

- Mobility trees
- Survival trees
- Characteristic sequences

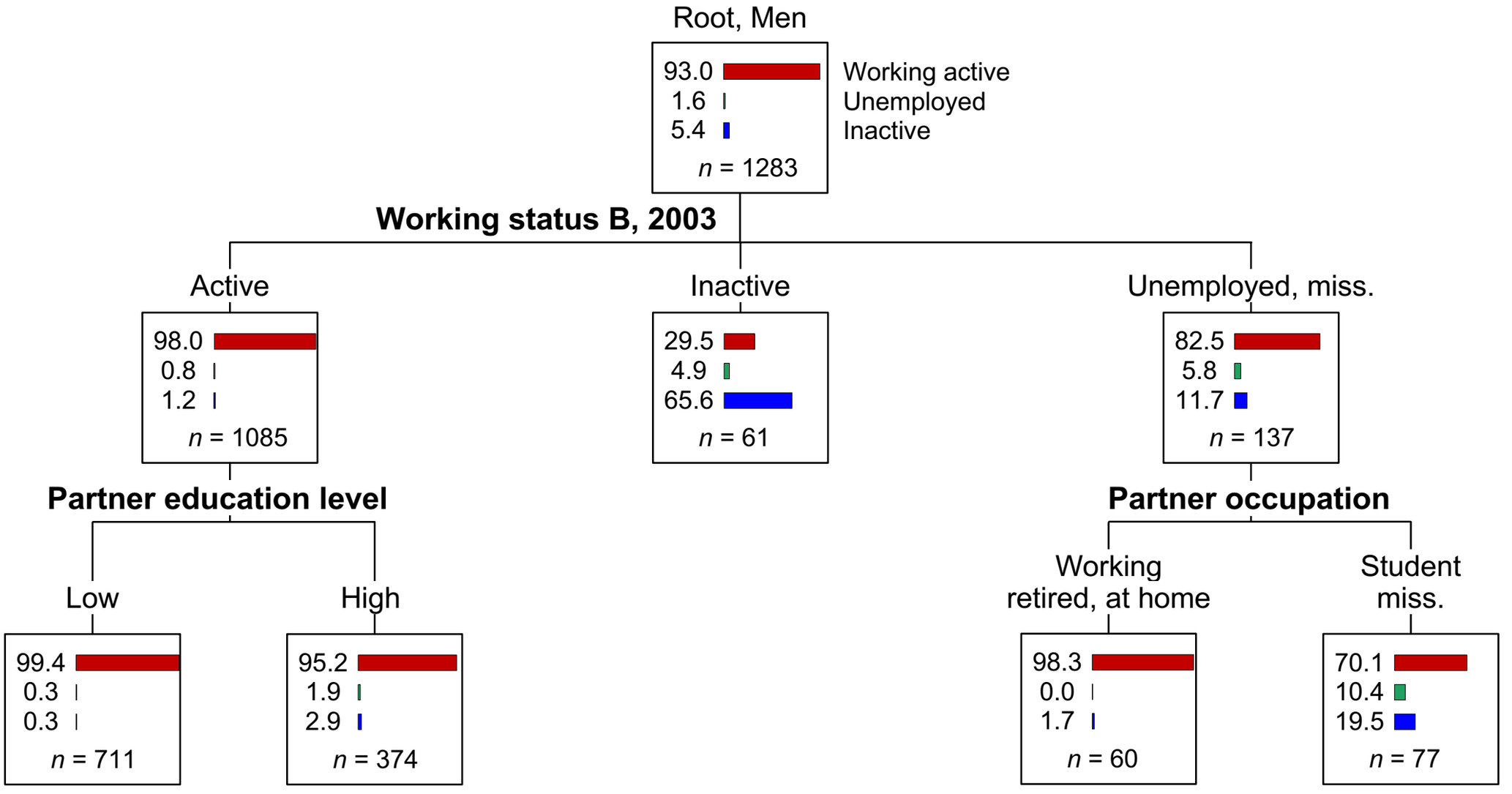
2.1 Mobility trees

- (SHP Data, Waves 1 to 6 (1999-2004), aged between 20 and 64 in 2004.)
- How does working status (occupied active, unemployed, inactive) in 2004 depend on
  - working status in previous year (1999 to 2003)
  - other factors (attained education level, partner working status, partner education level, ...)

  and what are main interaction effects?

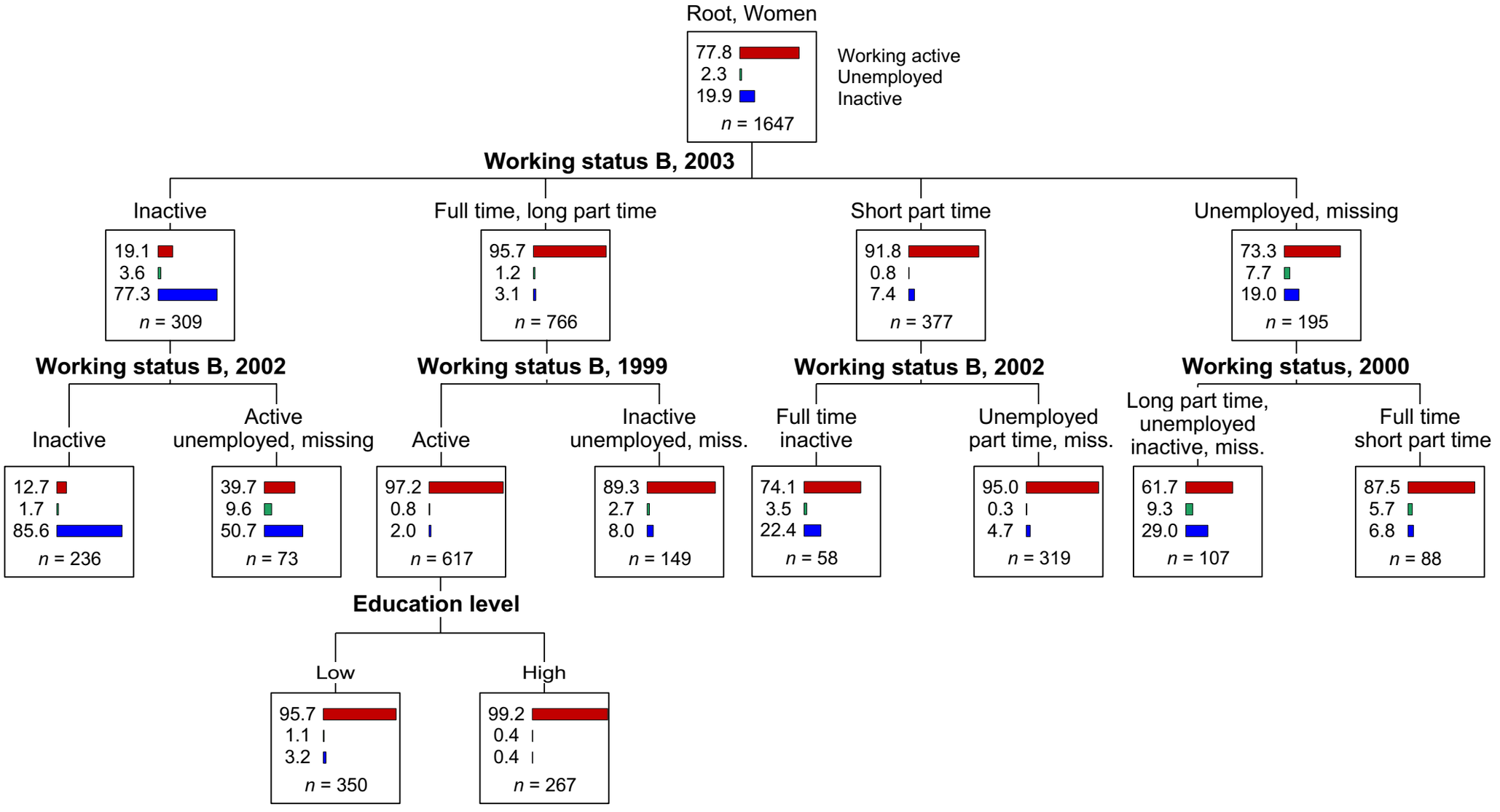
- Mobility trees are alternative to Markovian transition models.
- Growing separate classification trees for women and men highlights gender differences.

Mobility tree, Men



Working status B (full time, long part time, short part time, unemployed, inactive)

Mobility tree, Women



Working status B (full time, long part time, short part time, unemployed, inactive)

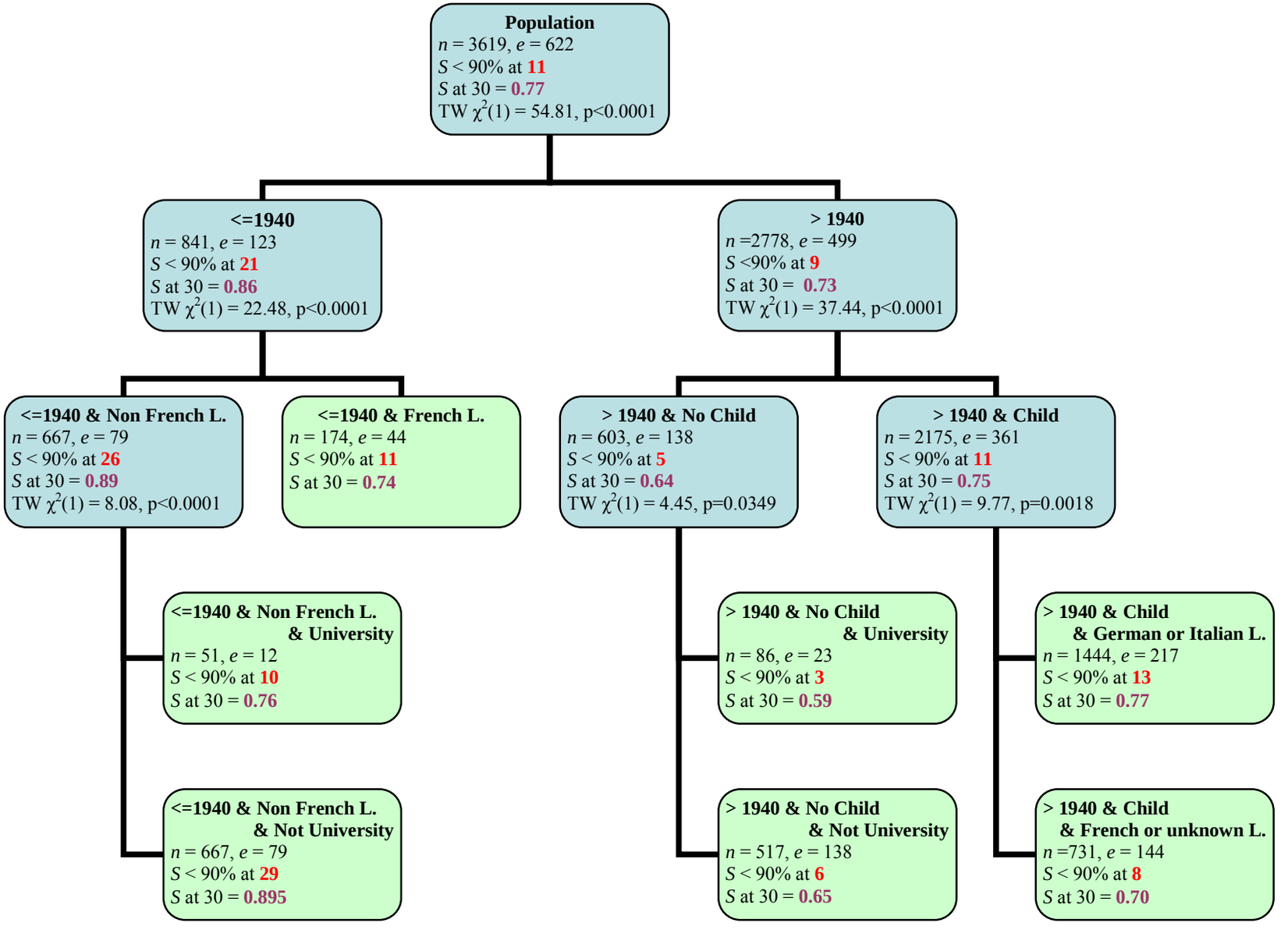
2.2 Survival trees

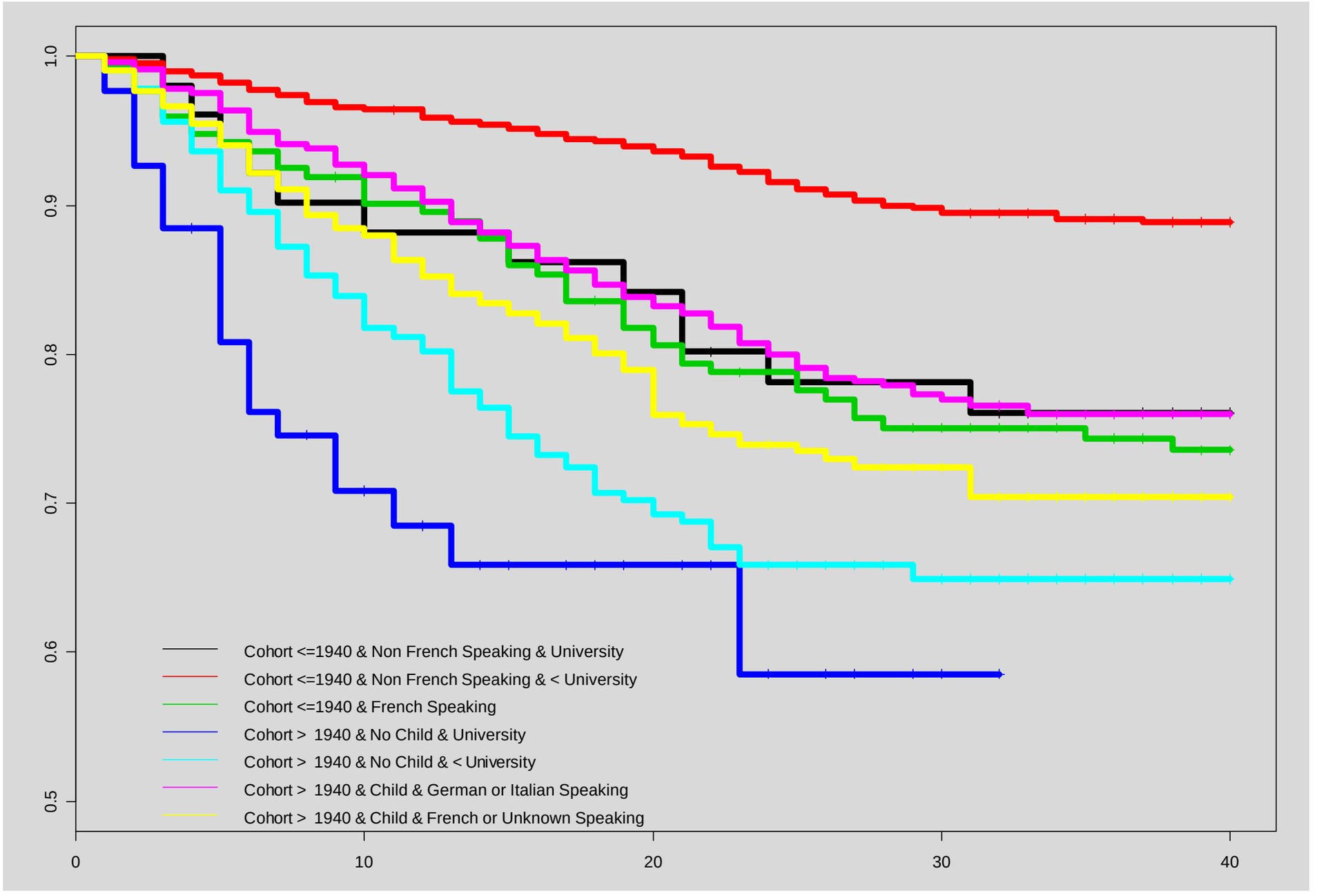
- (SHP 2002 biographical data, 2002 Wave data for some potential explanatory factors)
- Which are the most discriminating factors for marriage duration until divorce/separation?

  Used same variables as for discrete time logistic model in Ritschard and Sauvain-Dugerdil (2007)

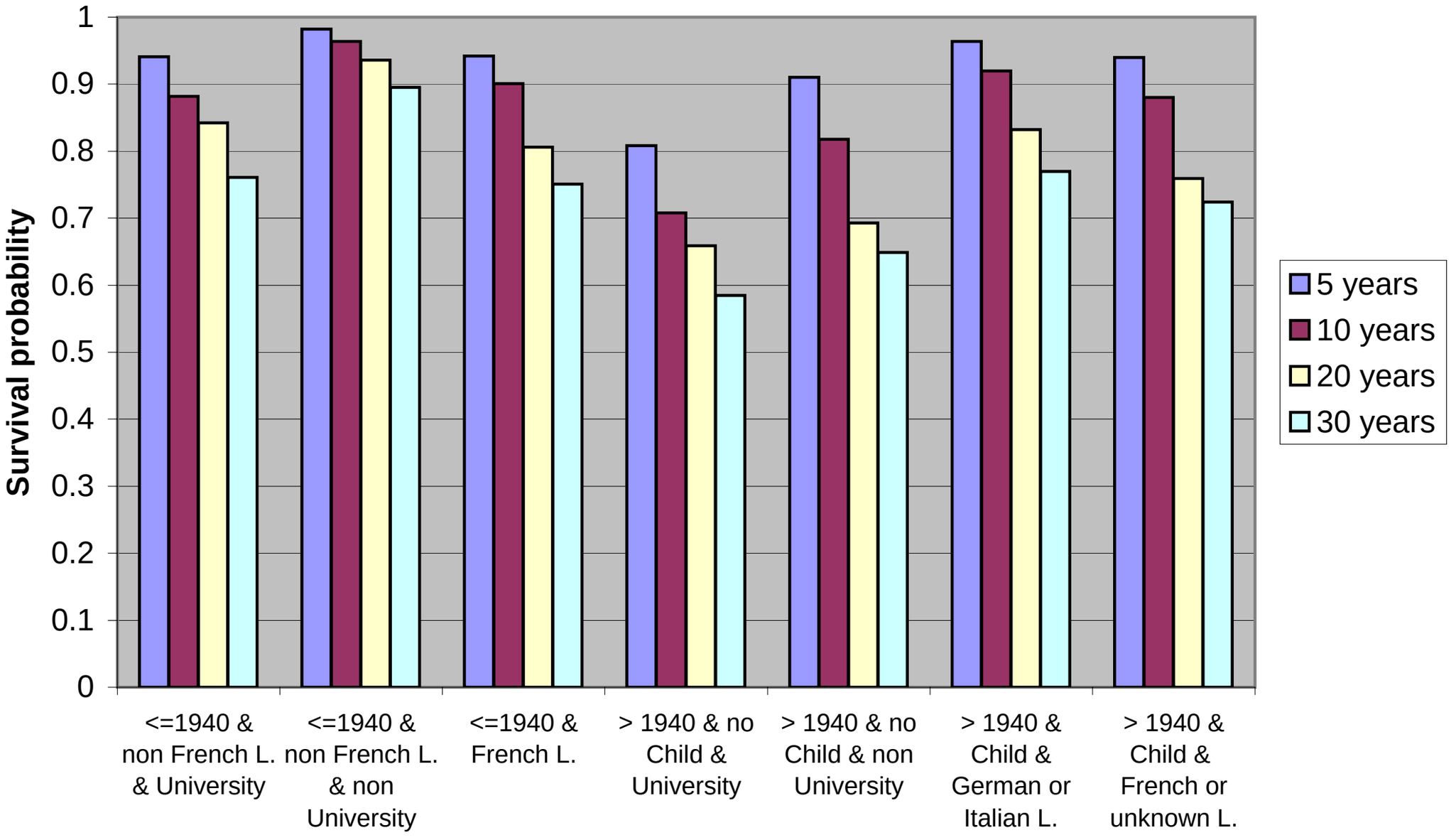
- Tried two methods
  - Maximize differences in KM survival curves using Tarone-Ware (T-W) p-value (Segal, 1988).
  - Cox regression tree: maximize differences in proportionality factors among groups (Leblanc and Crowley, 1992; Therneau and Atkinson, 1997)

T-W Survival Tree: Marriage until Divorce/Separation




1.0
0.9
0.8
0.7
0.6
0.5
0
10
20
30
40
Cohort <=1940 & Non French Speaking & University
Cohort <=1940 & Non French Speaking & < University
Cohort <=1940 & French Speaking
Cohort > 1940 & No Child & University
Cohort > 1940 & No Child & < University
Cohort > 1940 & Child & German or Italian Speaking
Cohort > 1940 & Child & French or Unknown Speaking

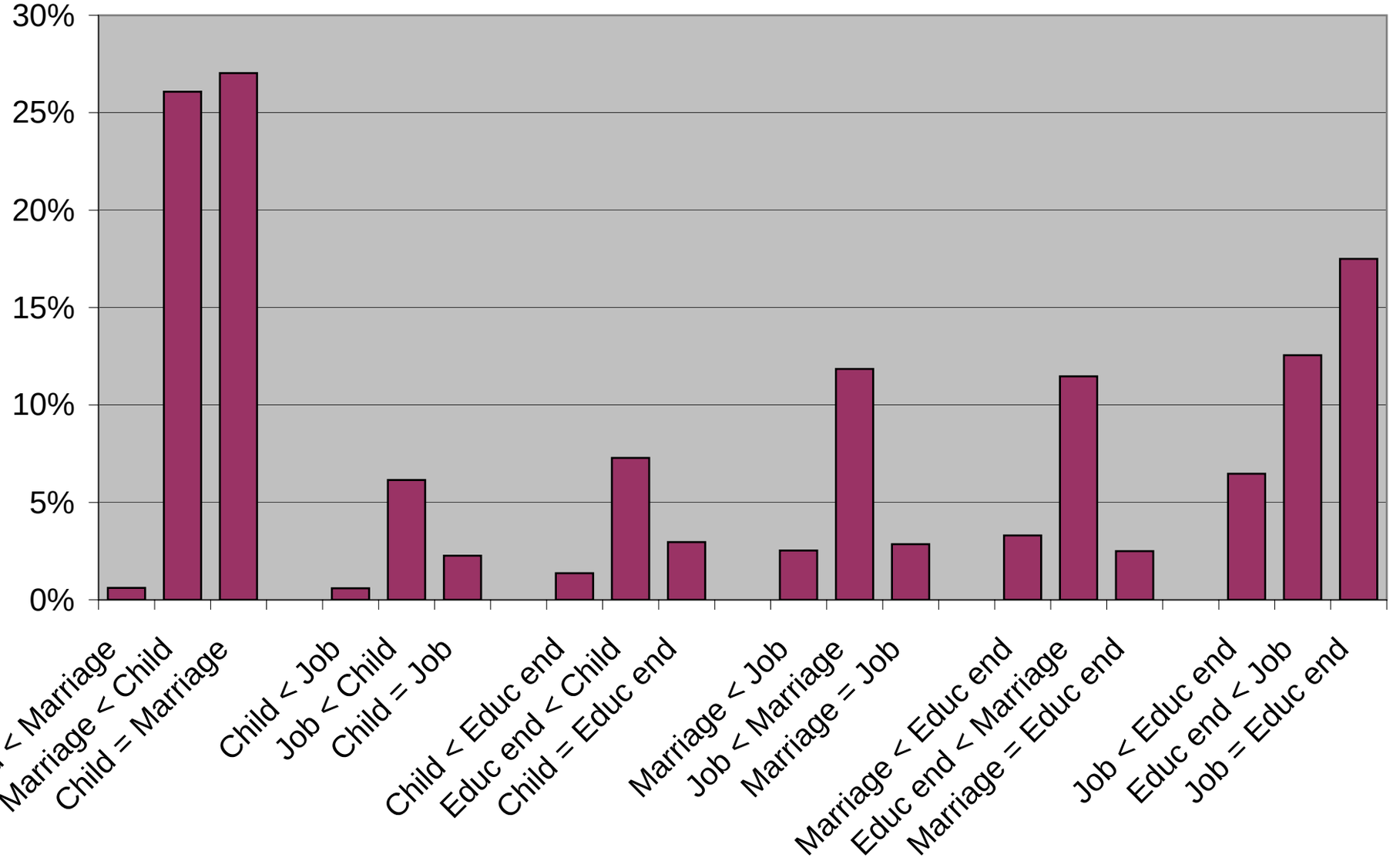
Marriage survival probabilities until Divorce/Separation, by leaves



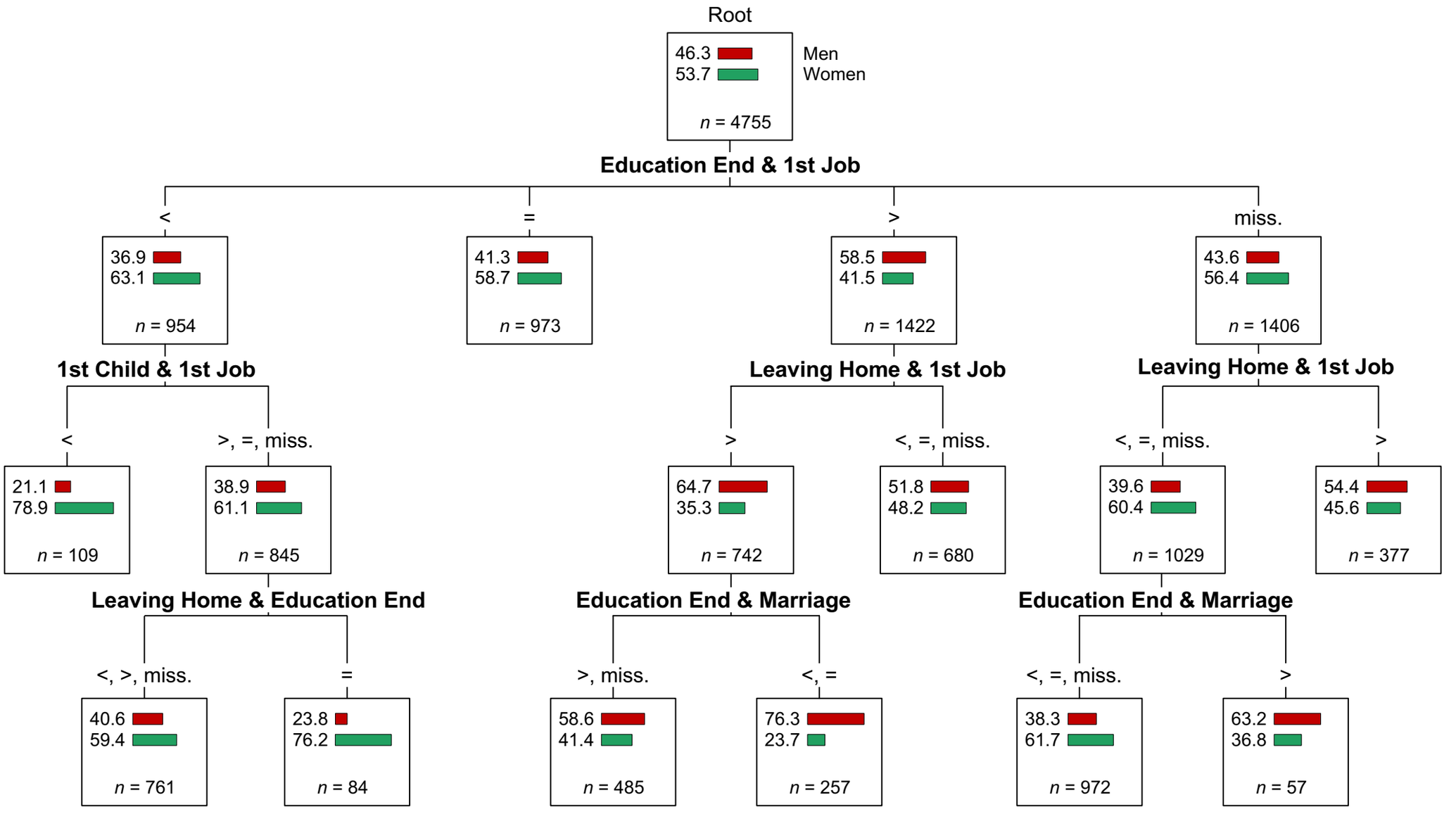
2.3 Characteristic sequences

- (SHP 2002 biographical data)
- Selection of pairs of events, e.g. marriage and first job.
- For each pair, order of sequence: $<$, $=$, $>$, missing
- Which are the most typical sequences?
- Most discriminating sequences between
  - sex
  - birth cohort (1940 and before, after 1940)

Frequencies of characteristic 2-event sequences



Discriminating sex with 2-event sequences



3 Foreseen Developments

- Extend tree approaches for
  - Time varying covariates
  - Multilevel contexts
- Mining typical sequence patterns and association rules
- Suitable validation criteria
- Friendly graphical interface for making methods easily accessible
- Analysis of Swiss life courses
  - Differential impact of various profiles of social insertion
  - Broken lives
  - ...

THANK YOU

References

Billari, F. C. (2005). Life course analysis: Two (complementary) cultures? Some reflections with examples from the analysis of transition to adulthood. In P. Ghisletta, J.-M. Le Goff, R. Levy, D. Spini, and E. Widmer (Eds.), *Towards an Interdisciplinary Perspective on the Life Course*, Advancements in Life Course Research, Vol. 10, pp. 267–288. Amsterdam: Elsevier.

Blossfeld, H.-P. and G. Rohwer (2002). *Techniques of Event History Modeling, New Approaches to Causal Analysis* (2nd ed.). Mahwah NJ: Lawrence Erlbaum.

Leblanc, M. and J. Crowley (1992). Relative risk trees for censored survival data. *Biometrics 48*, 411–425.

Ritschard, G. et C. Sauvain-Dugerdil (2007). L'enfant ciment du couple ou le couple comme ciment de la relation du père à l'enfant ? Quelques enseignements de l'enquête rétrospective du Panel Suisse de Ménages. In C. Burton-Jeangros, E. Widmer, et C. Lalive d'Epinay (Eds.), *Interactions familiales et constructions de l'intimité.*, coll. Questions sociologiques. Paris : L'Harmattan. (à paraître).

Segal, M. R. (1988). Regression trees for censored data. *Biometrics 44*, 35–47.

Therneau, T. M. and E. J. Atkinson (1997). An introduction to recursive partitioning using the rpart routines. Technical Report Series 61, Mayo Clinic, Section of Statistics, Rochester, Minnesota.