IPUC, Neuchâtel, February 23-24, 2007

# Innovative Data Mining based approaches for life course analysis

Gilbert Ritschard
Alexis Gabadinho, Nicolas Müller, Matthias Studer

University of Geneva, Switzerland

## Outline



http://mephisto.unige.ch

# 1 Aim of the research project

Just started February 1, 2007 FNS project on
"Mining event histories: Towards new insight on personal Swiss life courses"

Methodological concern Explore and develop data mining approaches for individual longitudinal data

- Methods for time to event analysis
- Methods for sequence data analysis

Socio-demographic concern Using mainly SHP data, but also other sources, gain original insight on

- How familial, professional and other socio-demographic events are entwined,
- Typical characteristics of Swiss life trajectories,
- Changes in these characteristics over time.

## What is data mining?

> "Data Mining is the process of finding new and potentially useful knowledge from data"
> Gregory Piatetsky-Shapiro editor of http://www.kdnuggets.com

> "Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner"
> (Hand et al., 2001)

Also called Knowledge Discovery in Databases, KDD.

Origin: IJCAI Workshop, 1989, Piatetsky-Shapiro (1989)

Textbooks : Han and Kamber (2001), Hand et al. (2001)

# What is data mining? (2)

**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<br>- Frequencies of typical patterns<br>- Discovering typical patterns |
| causality | - Hazard regression models<br>- Survival trees | - Markov models, Mobility trees<br>- Association rules between subsequences |

# 2 Our first results

- Mobility trees
- Survival trees
- Characteristic sequences

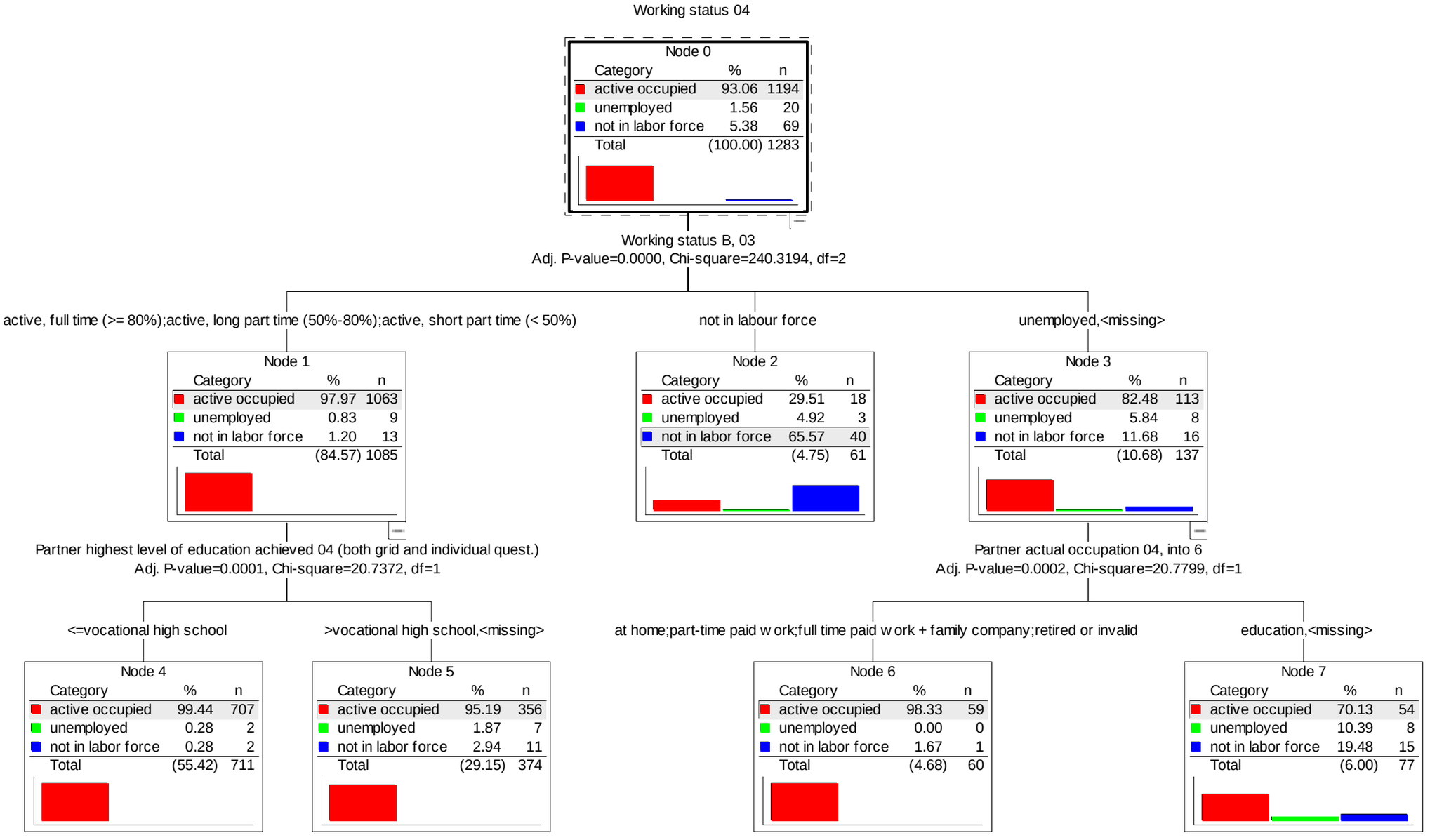
## 2.1 Mobility trees

- (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, ...)

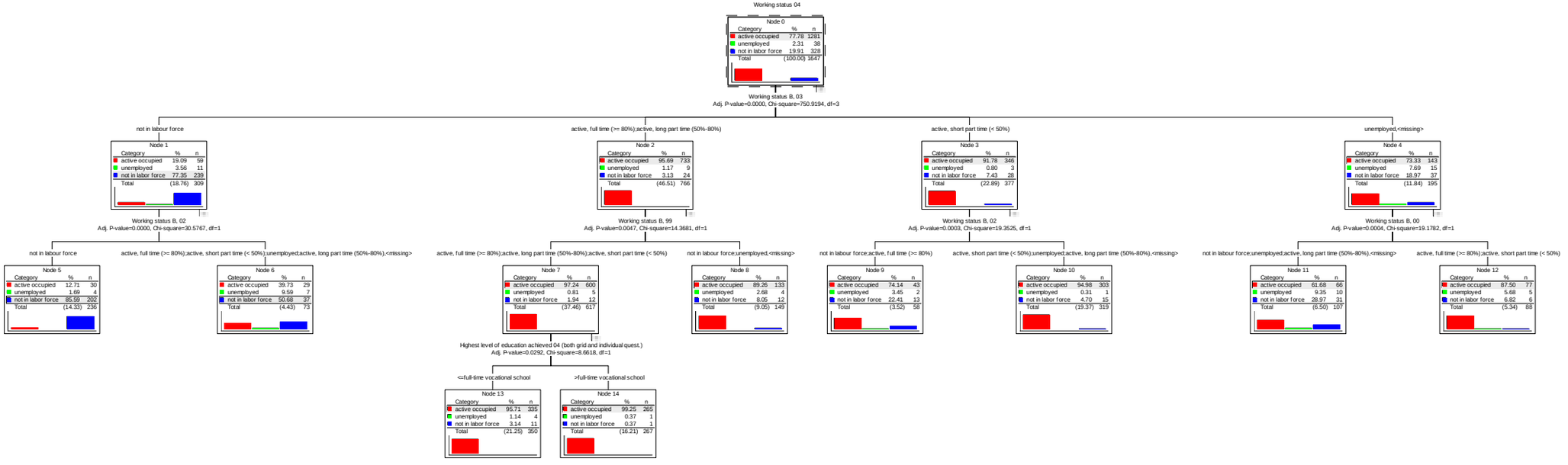
  and what are main interaction effects?

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

# Mobility tree, Men

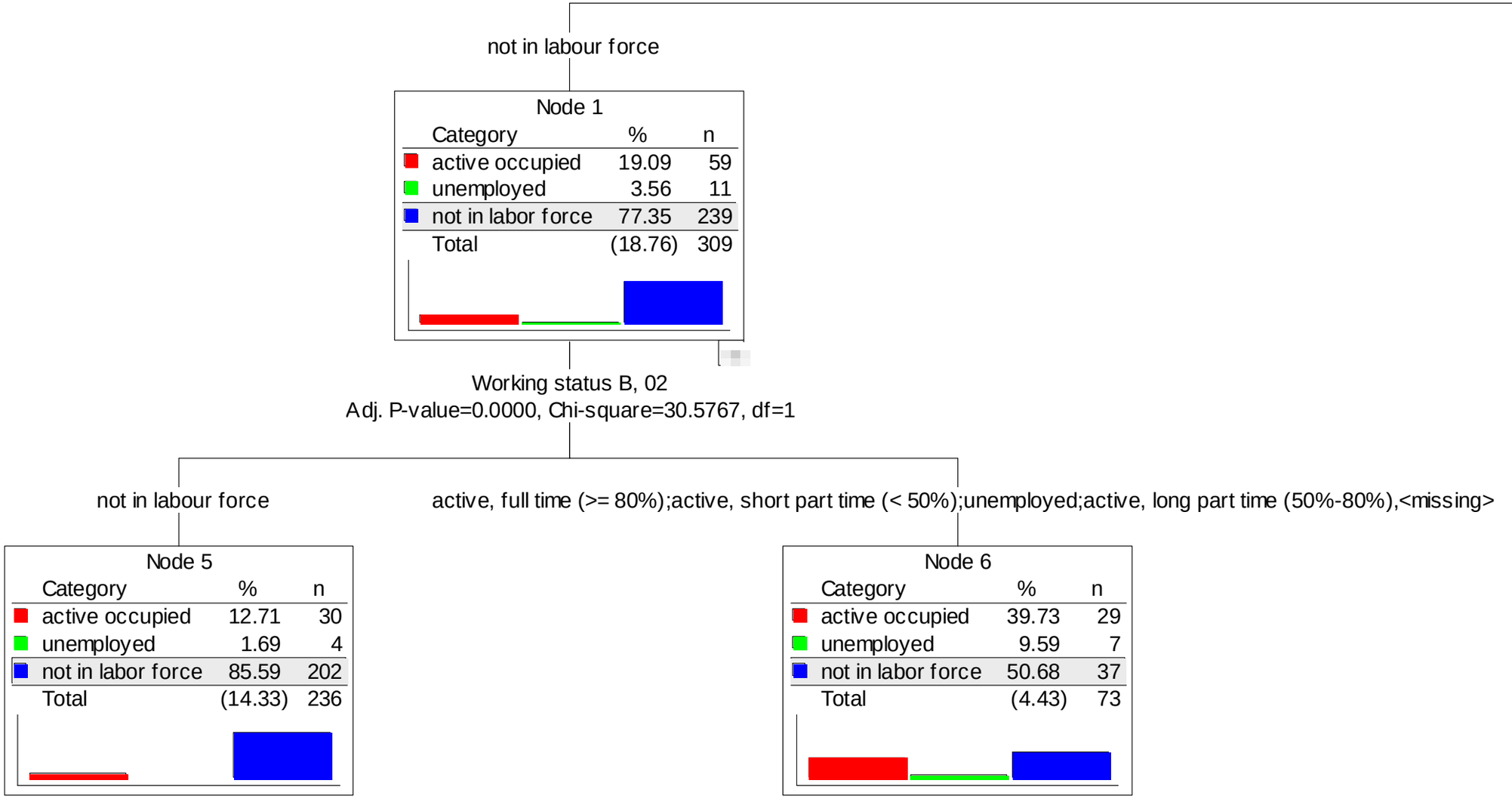


# Mobility tree, Women




Working status B (full time, long part time, short part time, unemployed, inactive) in 2003 used for first split

# Mobility tree, Women: Details for women inactive in 2003



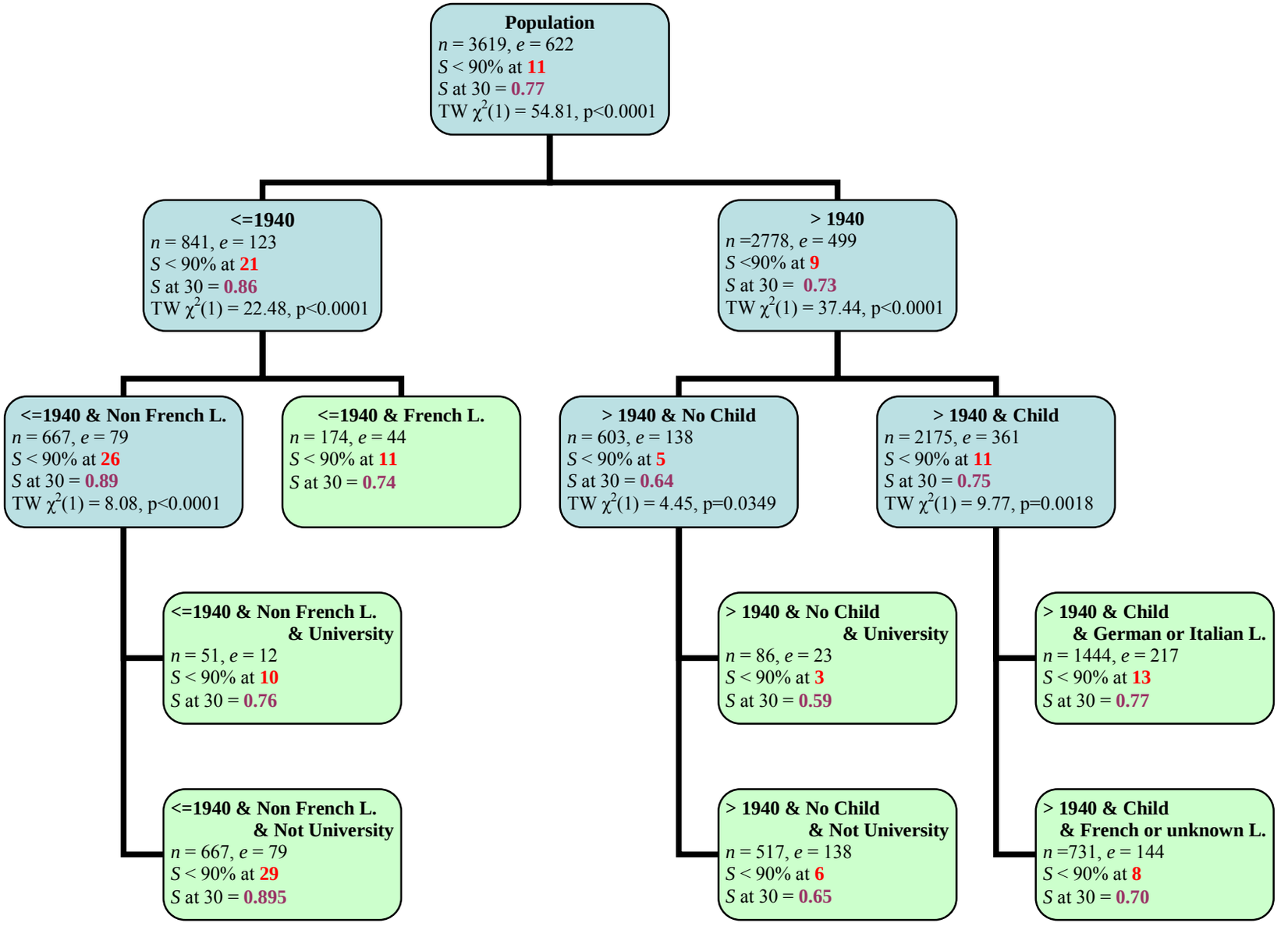
## 2.2 Survival trees

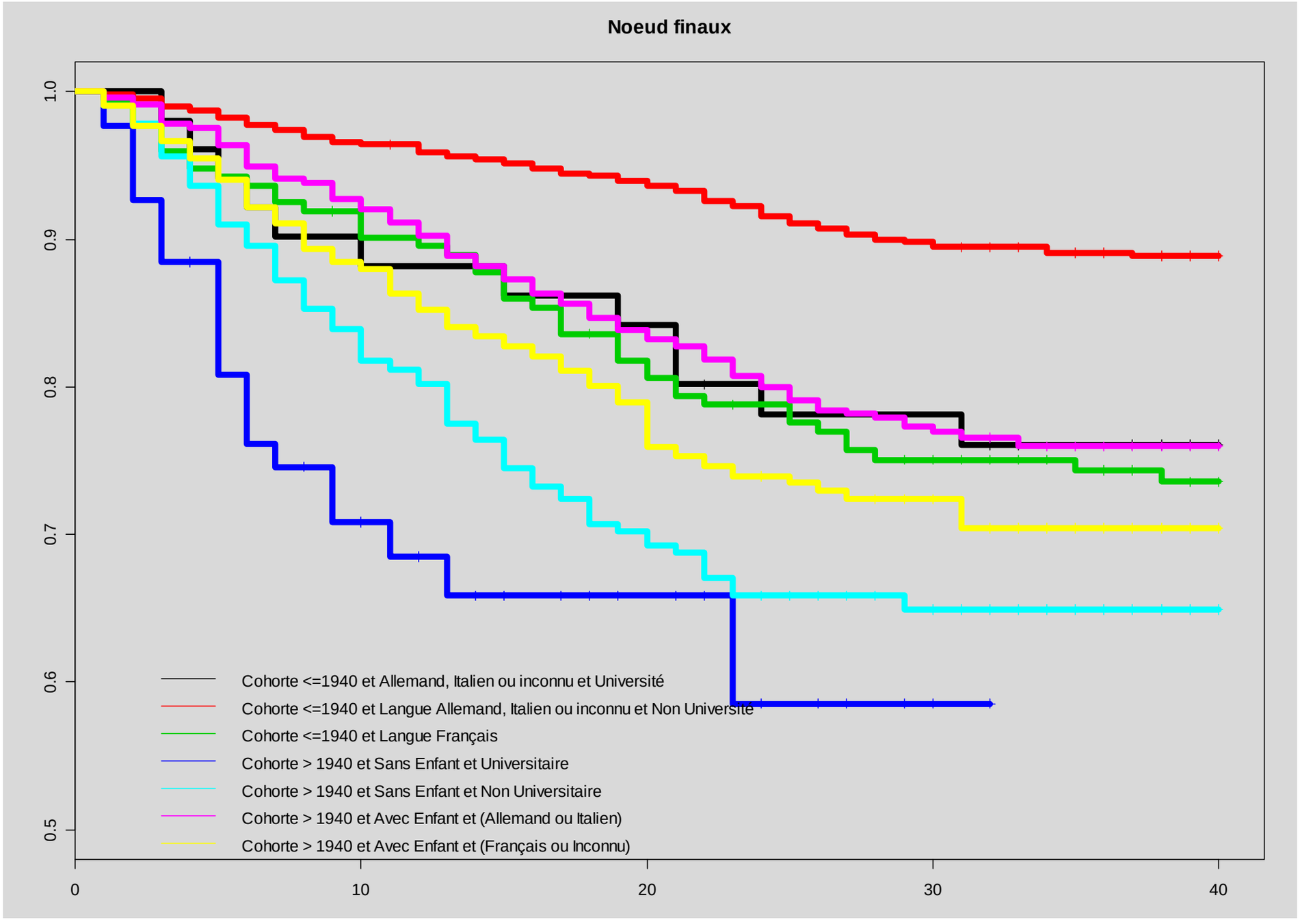
- (SHP 2002 biographical data, 2002 Wave data for some potential explanatory factors)
- Which are the most discriminating factors for marriage duration until divorce/separation?

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

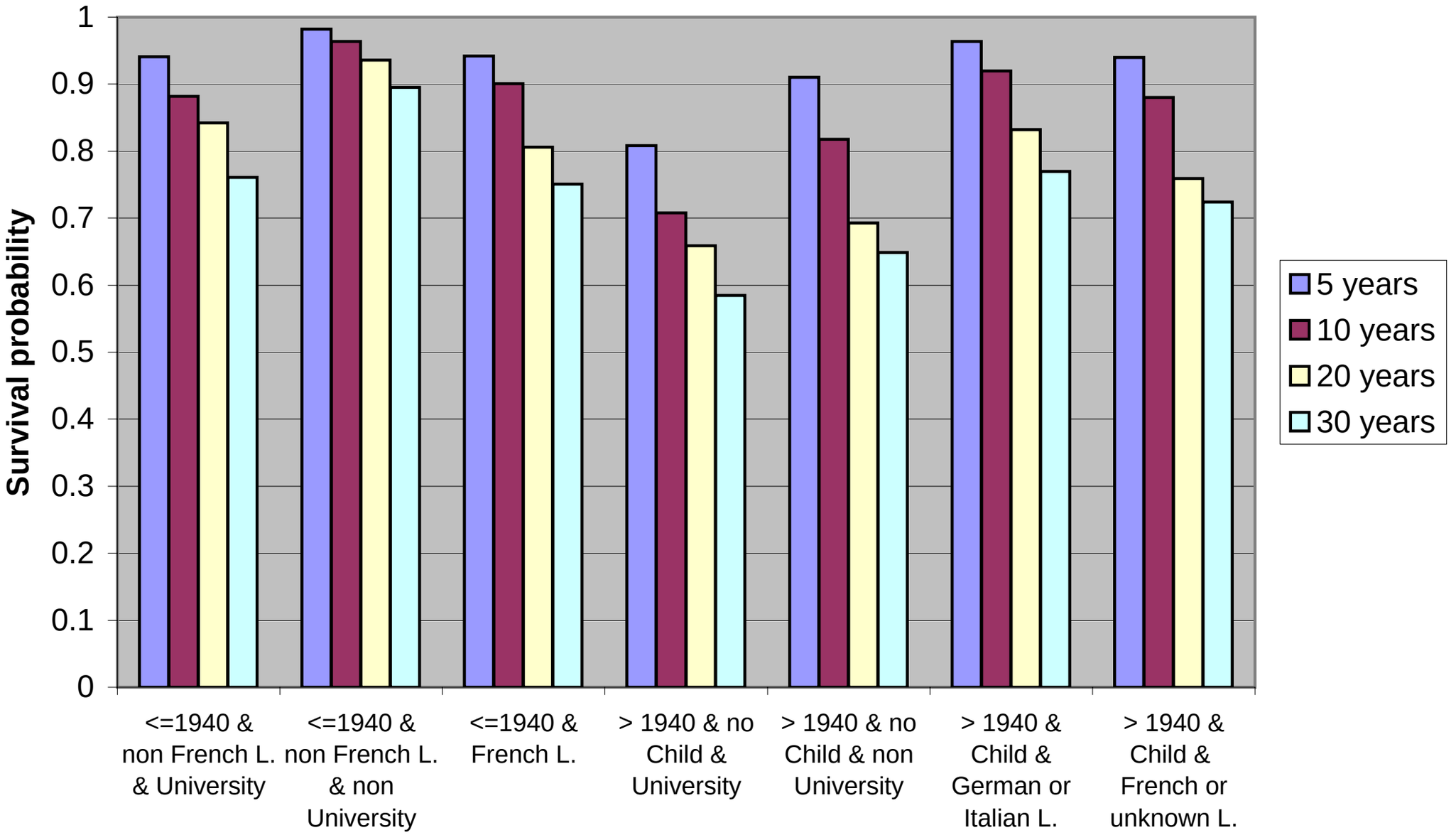
- Tried two methods
  - Maximize differences in KM survival curves using Tarone-Ware (T-W) $p$-value (Segal, 1988).
  - Cox regression tree: maximize differences in proportionality factors among groups (Leblanc and Crowley, 1992; Therneau and Atkinson, 1997)

# T-W Survival Tree: Marriage until Divorce/Separation

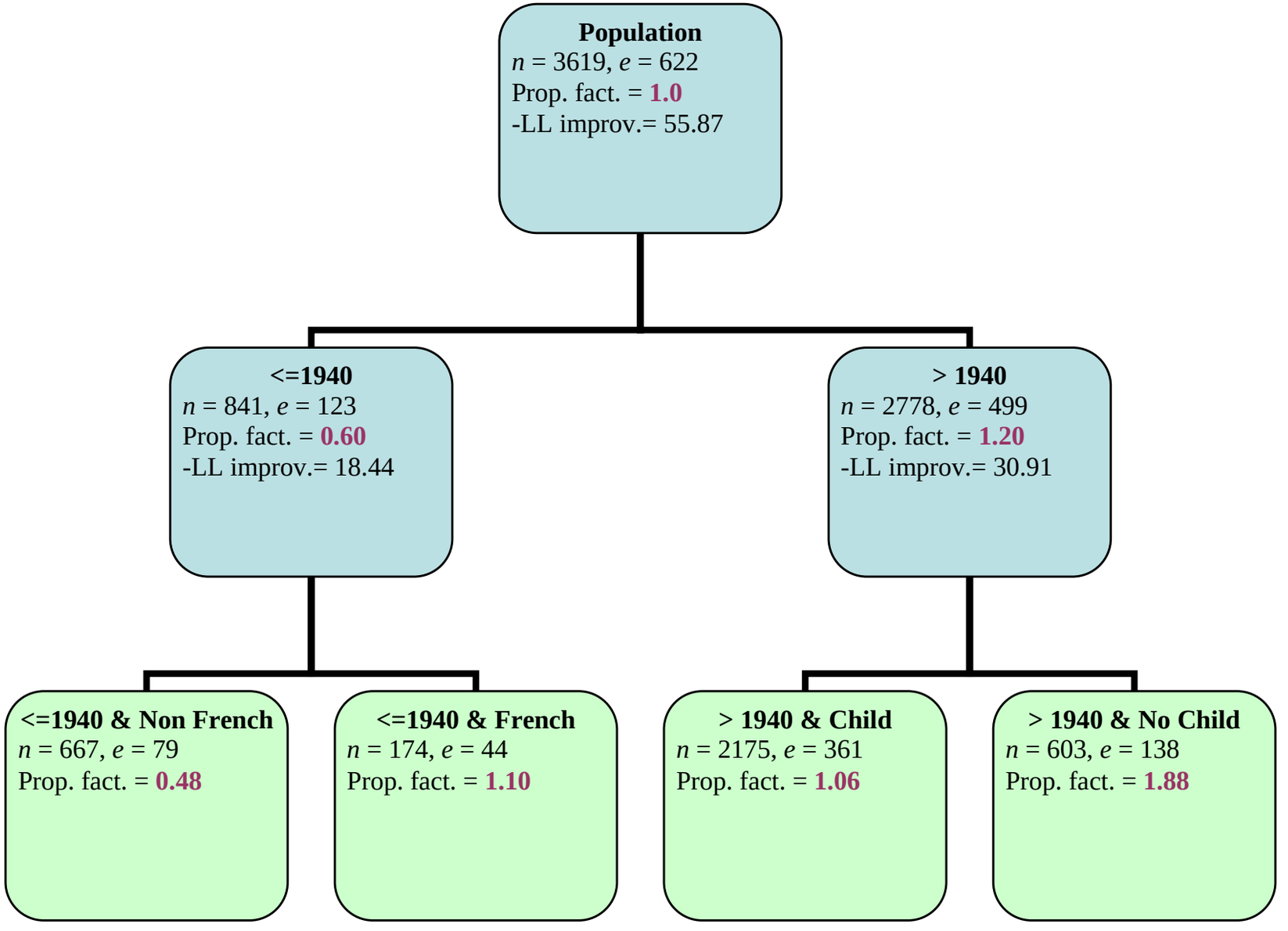


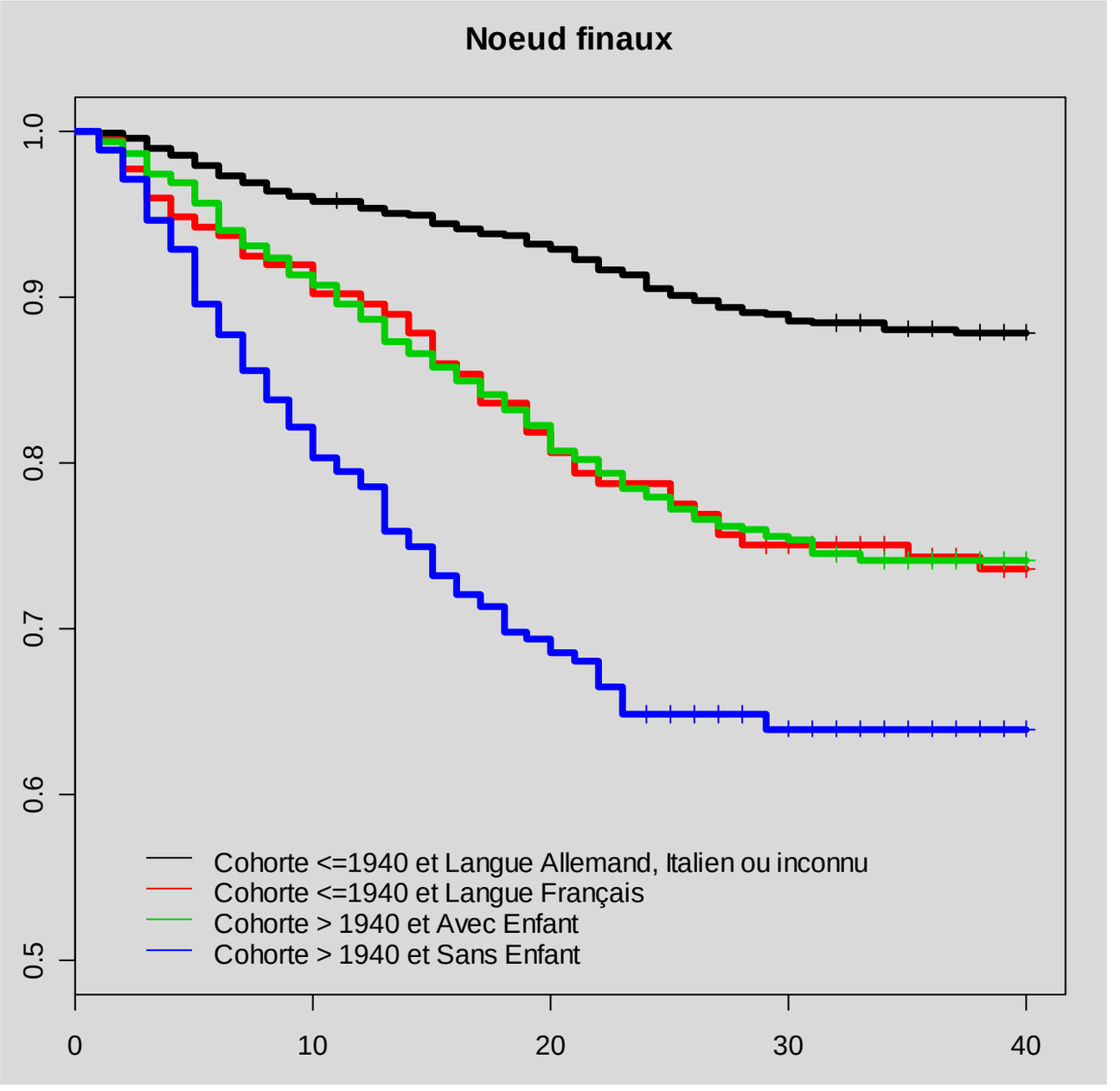

Noeud finaux
1.0
0.9
0.8
0.7
0.6
0.5
0
10
20
30
40
Cohorte <=1940 et Allemand, Italien ou inconnu et Université
Cohorte <=1940 et Langue Allemand, Italien ou inconnu et Non Université
Cohorte <=1940 et Langue Français
Cohorte > 1940 et Sans Enfant et Universitaire
Cohorte > 1940 et Sans Enfant et Non Universitaire
Cohorte > 1940 et Avec Enfant et (Allemand ou Italien)
Cohorte > 1940 et Avec Enfant et (Français ou Inconnu)

# Marriage survival probabilities until Divorce/Separation, by leaves



# Cox Survival Tree: Marriage until Divorce/Separation

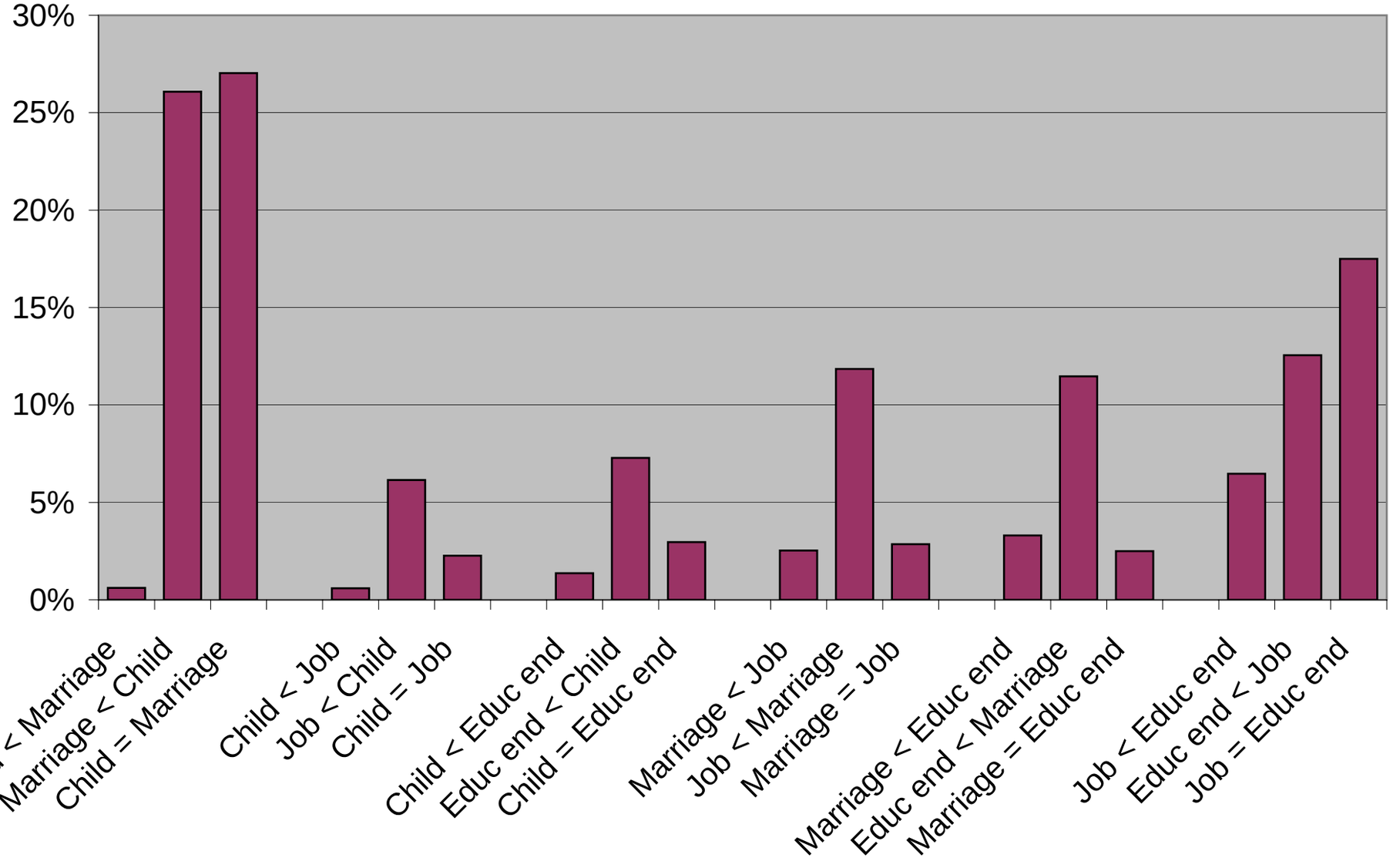



Noeud finaux
1.0
0.9
0.8
0.7
0.6
0.5
0
10
20
30
40
Cohorte <=1940 et Langue Allemand, Italien ou inconnu
Cohorte <=1940 et Langue Français
Cohorte > 1940 et Avec Enfant
Cohorte > 1940 et Sans Enfant

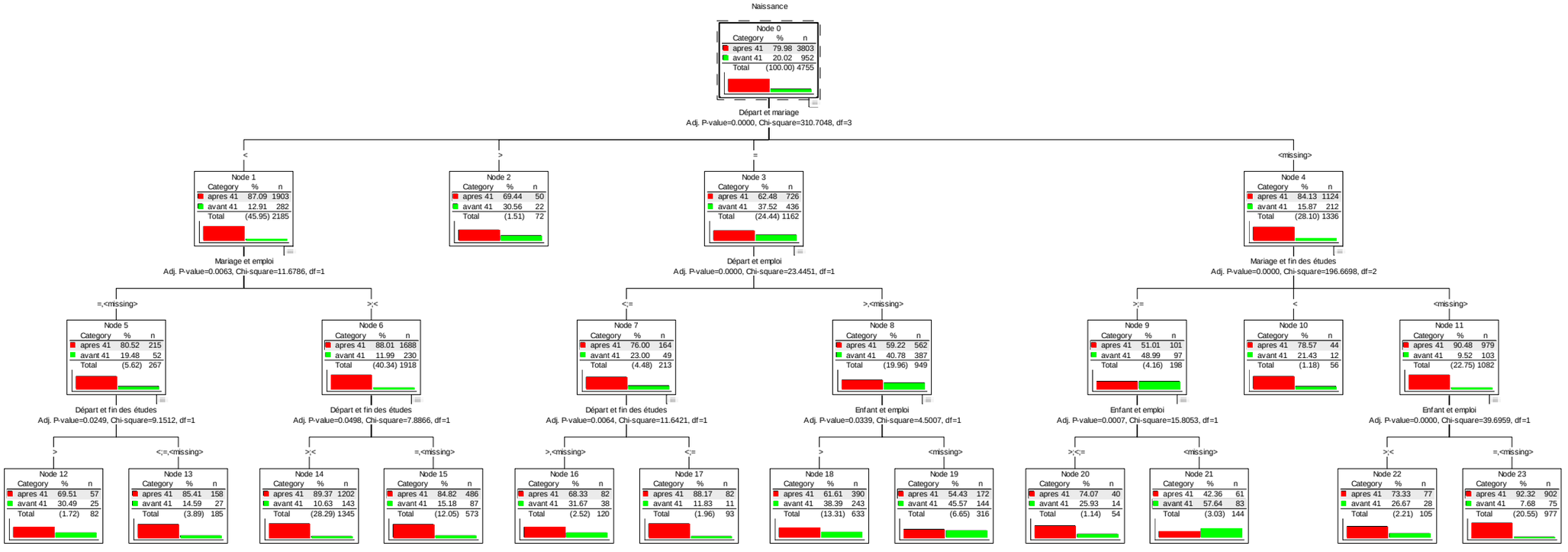
## 2.3 Characteristic sequences

- (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

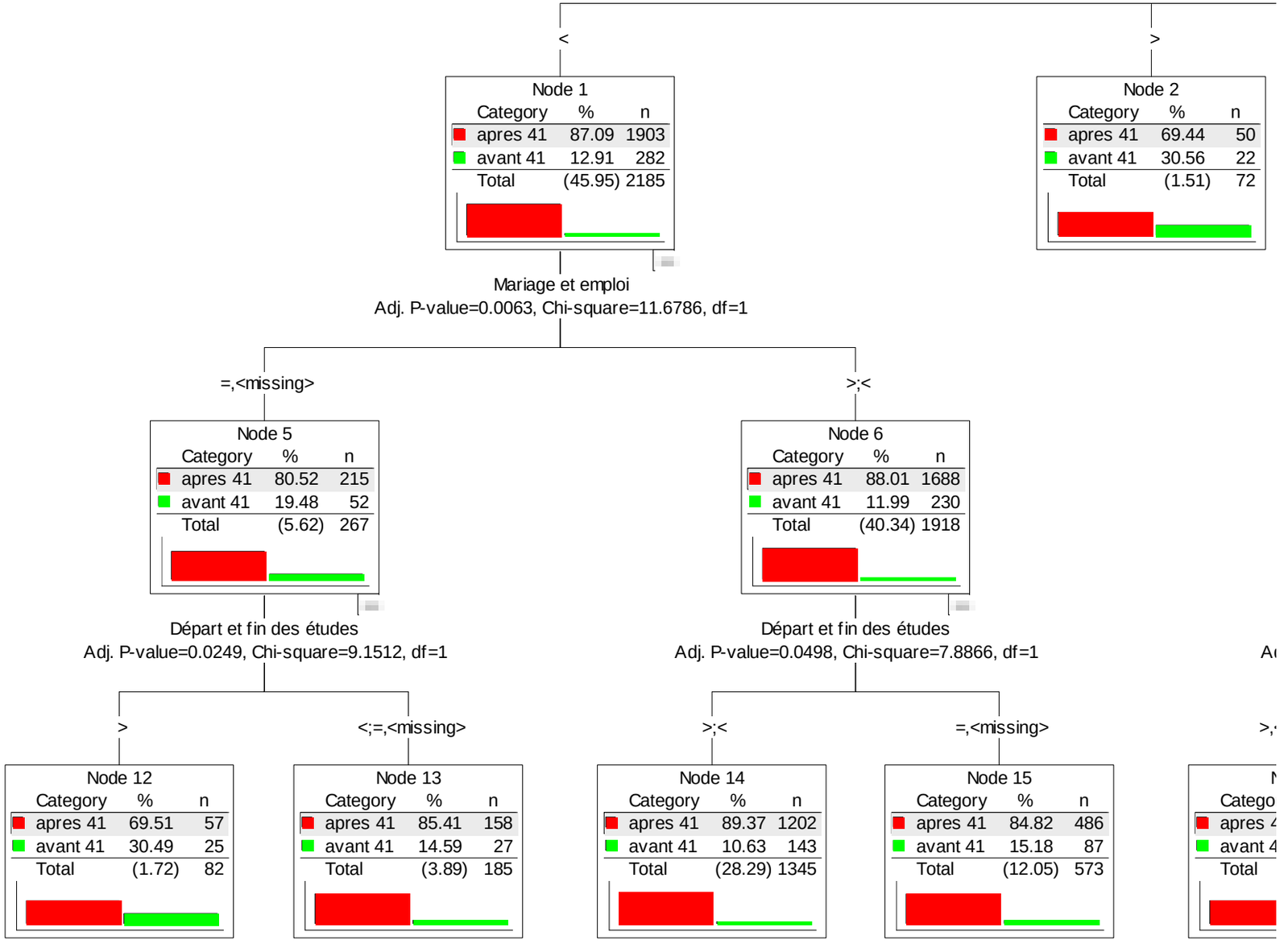


# Cohort discriminating 2-event sequences

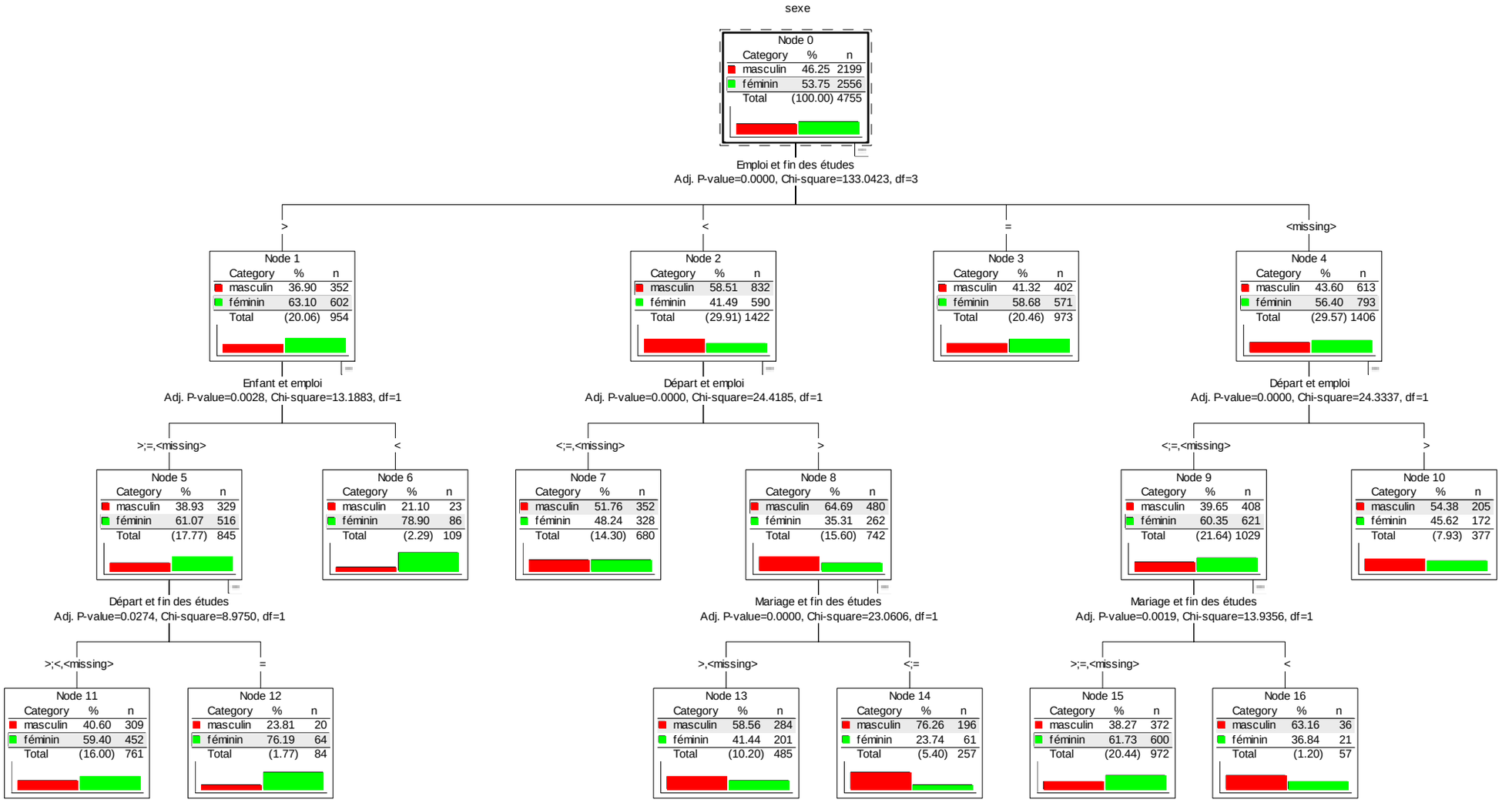



First split variable: {Marriage, Leaving Home}

# Cohort: details for Leaving Home before Marriage

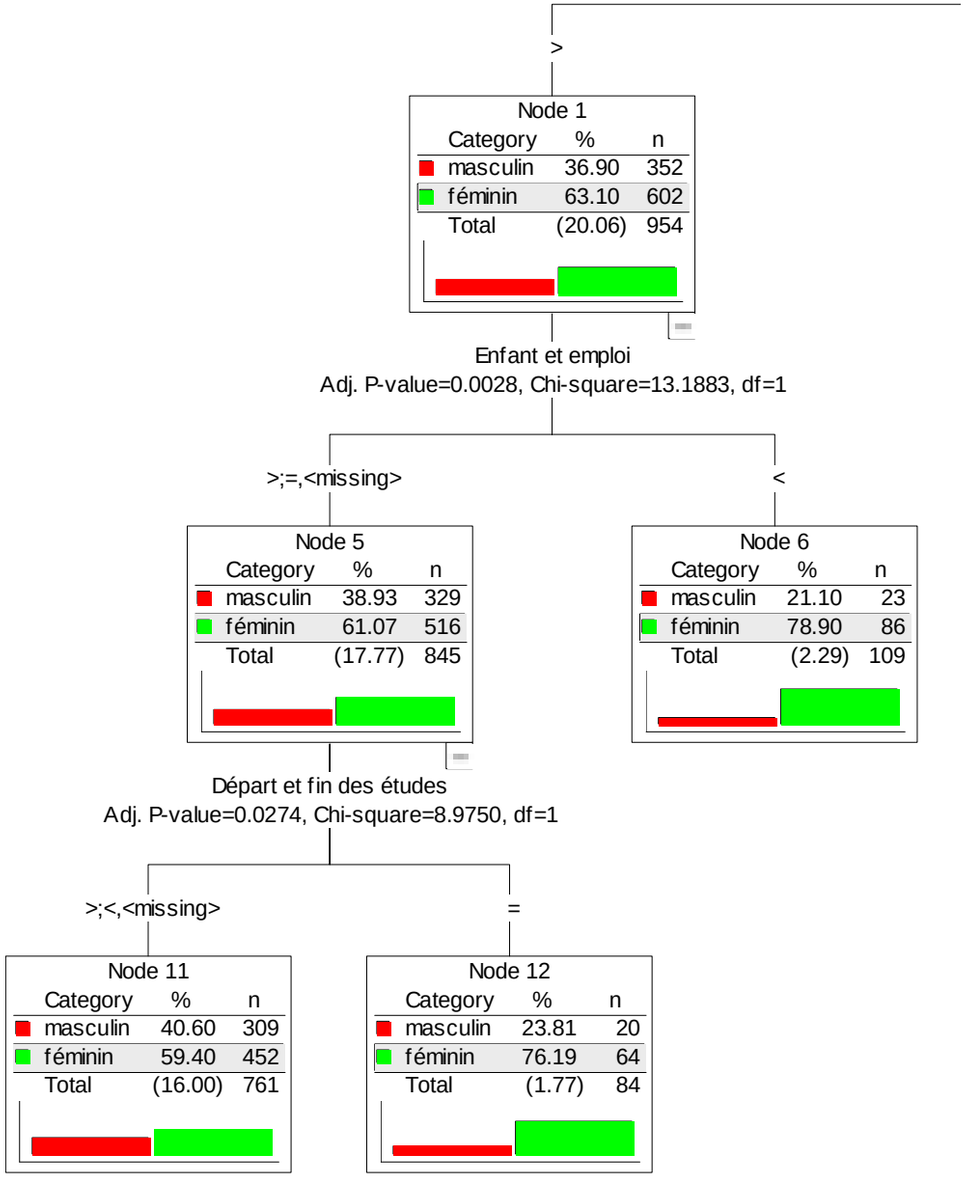


# Sex discriminating 2-event sequences




First split variable: {Job, Education End}

# Sex: details for Job after Education end



# 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
  - ...

# References


Han, J. and M. Kamber (2001). *Data Mining: Concept and Techniques*. San Francisco: Morgan Kaufmann.

Hand, D. J., H. Mannila, and P. Smyth (2001). *Principles of Data Mining*. Adaptive Computation and Machine Learning. Cambridge MA: MIT Press.

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

Piatetsky-Shapiro, G. (Ed.) (1989). *Notes of IJCAI'89 Workshop on Knowledge Discovery in Databases (KDD'89)*, Detroit, MI.

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.