Data mining methods for longitudinal data

Gilbert Ritschard, Dept of Econometrics, University of Geneva

Table of Content

http://mephisto.unige.ch

1 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 (ECD).

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

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

1.1 Kind of searched knowledge

Characterizing and discriminating classes

(Which attributes and which values best characterize and discriminate classes?)

Prediction and classification rules (supervised)

(How to best use predictors for predicting the outcome?)

Association Rules

(Which other books are ordered by a customer that buys a given book?)

Clustering (unsupervised)

(Which group emerge from the observed data?) ...

1.2 Main classes of methods

Supervised learning (discrimination, classification, prediction) The outcome variable is fixed at the learning stage.

Which predictors best discriminate the values (classes) of the outcome variable and how?

Ex: Distinguish countries according to age when leaving home, age at marriage, age when leaving education, ...

Mining association rules The predicate (outcome variable) of the rules is not necessarily fixed a priori.

Ex: Which event is most likely to follow the sequence (Ending a bachelor degree, Starting a love relation, Not finding a local job during 6 months)? Is it marriage, starting another formation, a higher level formation, moving abroad?

Unsupervised learning Clustering. No predefined outcome variable. Partition data into homogenous clusters.

Main supervised learning methods

- Induction Trees (Decision Trees, Classification Trees)
- k-Nearest Neighbors (KNN)
- Kernel Methods and Support Vector Machine (SVM)
- Bayesian Network
- ...

Here I will mainly discuss Induction Trees.

Characteristics of data mining methods

- Methods are mainly heuristics (non parametric, quasi optimal solutions)
- often very large data sets

  $\Rightarrow$ need for performance of algorithms
- heterogenous data (quantitative, categorial, symbolic, text,...)

  $\Rightarrow$ need for flexibility: should be able to handle many kinds of data (mixed data)

Breiman (2001) calls it the algorithmic culture and opposes it to the classical statistical culture based on stochastic data models.

2 Individual longitudinal data

Life course data

- Time stamped events

  Age when ending formation, age at marriage, age when first child, age at divorce, ...

  $\Rightarrow$ time to event, hazard (Event History Analysis)

- Sequences
  - of states

t	1	2	3	4	5	6	...
state	form	form	emp	emp	emp	unemp	...

  - of events

    first job $\rightarrow$ first union $\rightarrow$ first child $\rightarrow$ marriage $\rightarrow$ second child

  $\Rightarrow$ mobility analysis, optimal matching, frequent sequences

Mining longitudinal data: two approaches

1. Coding data to fit the input form of existing methods.

   This is what I will discuss here with two examples from the historical demography area

   - A three generation mobility analysis (with induction trees) (Ryczkowska and Ritschard, 2004; Ritschard and Oris, ming)
   - Detecting temporal changes in event sequences (mining frequent sequences)
     Blockeel et al. (2001)

2. Using (developing) dedicated tools (e.g. Survival Trees)

   I will here just briefly comment on an example from the literature
   De Rose and Pallara (1997)

3 Inducing a mobility tree

Geneva in the 19th century: historical background

- Eventful political, economic and demographic development
- City enclosed inside walls: lack of lands $\Rightarrow$ prevents development of agricultural sector.

  $\Rightarrow$ turns to trade and production of luxury items: textile ($\rightarrow$ beginning 19th) and clocks, jewelery, music boxes (Fabrique)
- Sector turned to exportation, hence sensitive to all the 19th political and economic crises.

  [1798-1816] French period (period of crises)

  [1816-1846] "Restauration" (annexation of the surrounding French parishes), economic boom during the 30's

  [1849- ...] Modernization of economic structure, destruction of the fortifications

Demographic evolution

- 1798: 21'327 inhabitants (larger than Bern 12000, Zurich, 10500 and Basel, 14000)

  Mainly natives (64%)

- French period: stagnation of population growth

- Positive growth by degrees after the 20's, boosted after the destruction of the walls (1850)

  1880: City 50'000, agglomeration 83'000

- High growth of immigrant population,

  lower growth of natives

  1860: 45% natives

  end of the century: 33% natives)

3.1 The data sources

Data collected by Ryczkowska (2003)

- City of Geneva, 1800-1880
- Marriage registration acts
- All individuals with a name beginning with letter B (socially neutral)

  $\Rightarrow$ 4865 acts
- Rebuild father - son histories by seeking the marriage act of the father for all marriages celebrated after 1829

  $\Rightarrow$ 3974 cases (1830-1880)

The social statuses

6 statuses build from the professions

unskilled : unskilled daily workmen, servants, labourer, ...

craftsmen : skilled workmen

clock makers : skilled persons working for the "Fabrique"

white collars : teachers, clerks, secretaries, apprentices, ...

petite et moyenne bourgeoisie : artists, coffee-house keepers, writers, students, merchants, dealers, ...

élites : stockholders, landlords, householders, businessmen, bankers, army high-ranking officers, ...

3.2 Two subpopulations: enrooted people and newcomers

enrooted population :

those for which the father of the groom or the bride also married in Geneva

newcomers :

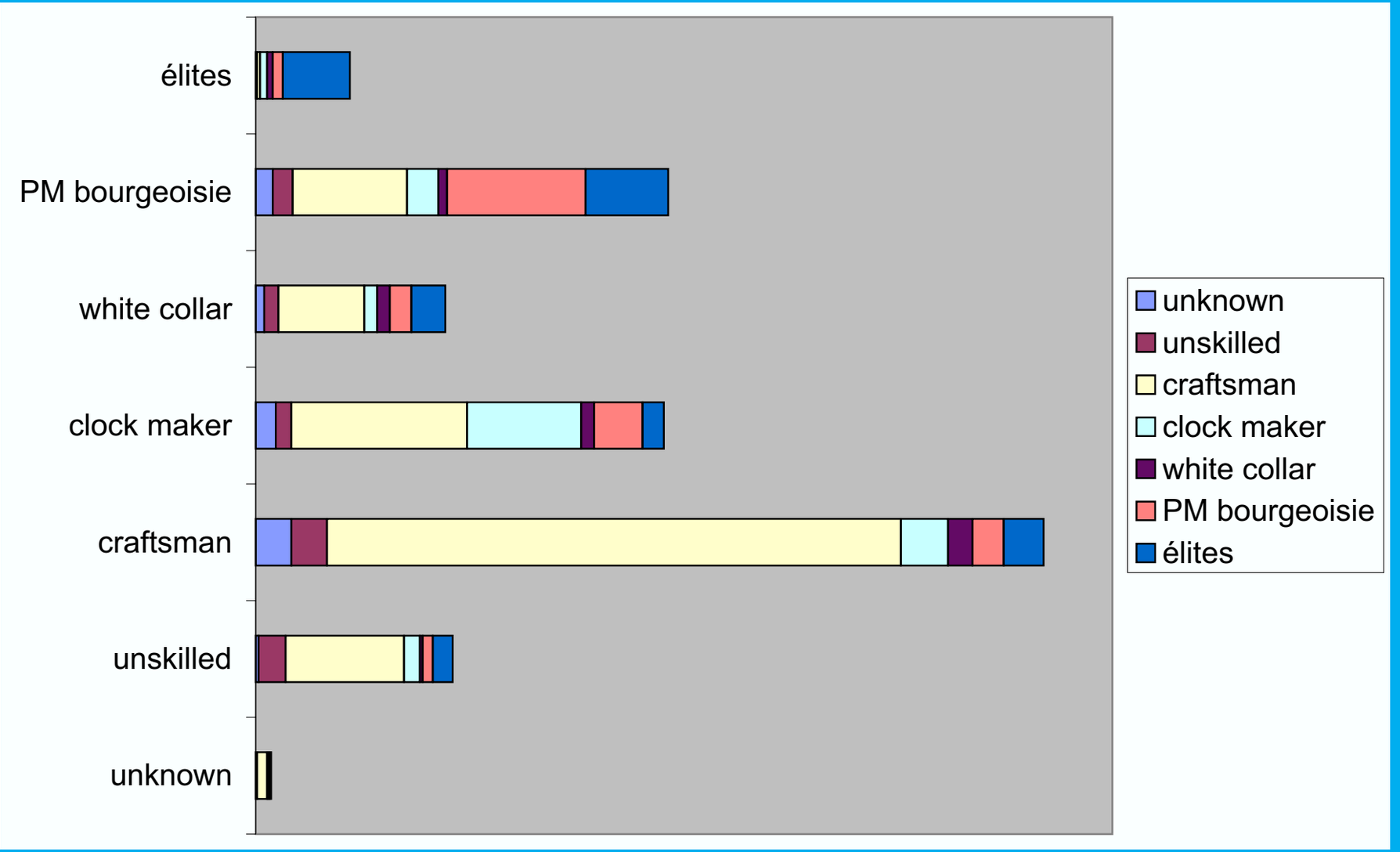
all others

Age at first marriage

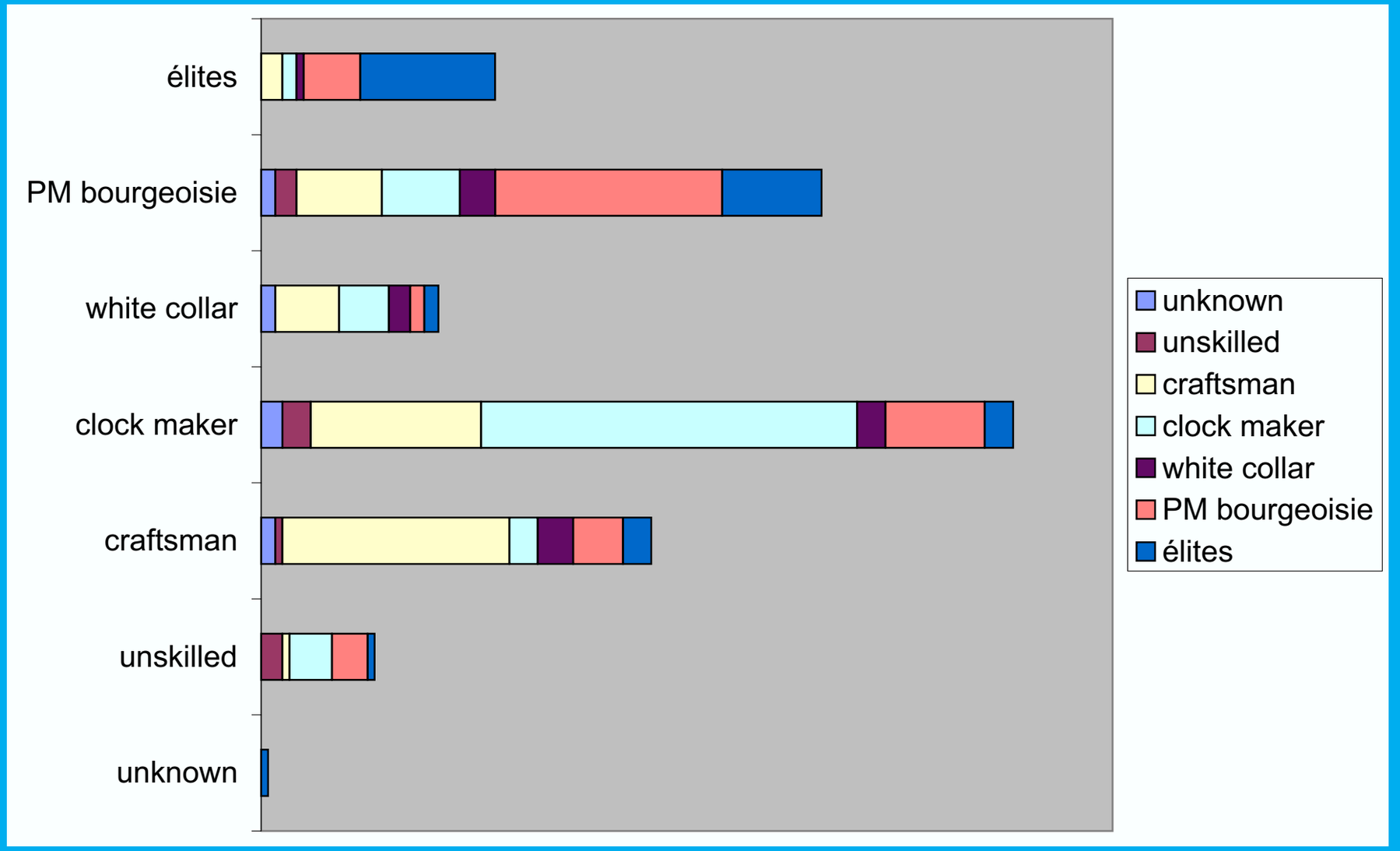
	enrooted		newcomers		
	mean age	n	mean age	n	deviation (stdev)
men	28.9	572	31.9	3402	3 (.32)
women	25.1	572	28.5	3402	3.4 (.27)

3.3 One generation social transitions

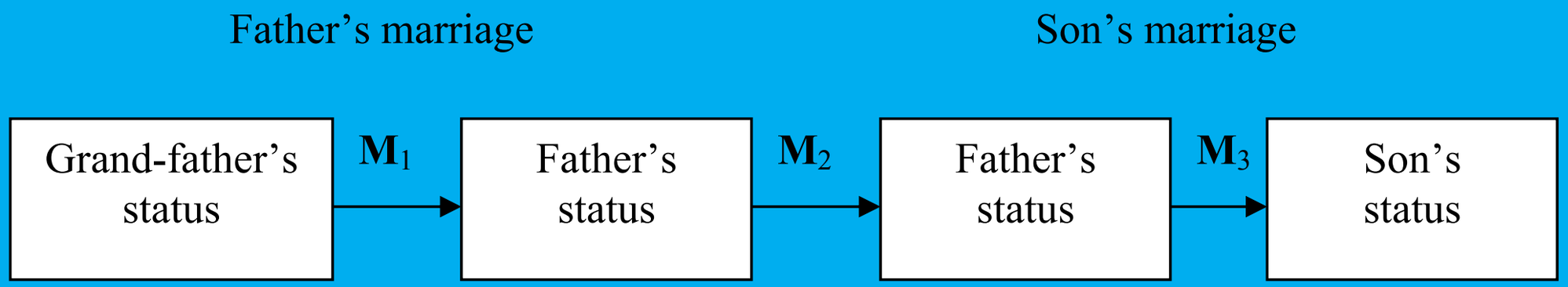
Newcomers (3402 cases), social origin, without deceased fathers



Stable population (572 cases), social origin, without deceased fathers



3.4 Three generations social transitions



First Order Transition Matrix

	t								half confidence interval
t-1	unknown	unskilled	craft	clock	wcolar	PMB	elite	deceased	
unknown			30.30%	15.15%	6.06%	24.24%	6.06%	18.18%	19.65%
unskilled	1.79%	10.71%	7.14%	19.64%	1.79%	21.43%	3.57%	33.93%	15.08%
craft	0.89%	3.25%	37.87%	17.75%	4.73%	9.47%	2.96%	23.08%	6.14%
clock	0.57%	2.83%	8.50%	46.46%	5.95%	13.60%	2.55%	19.55%	6.01%
wcolar		4.62%	21.54%	13.85%	15.38%	10.77%	6.15%	27.69%	14.00%
PMB	1.48%	4.44%	10.74%	14.81%	3.33%	33.70%	10.00%	21.48%	6.87%
elite	1.04%	2.08%	6.25%	12.50%	3.13%	26.04%	39.58%	9.38%	11.52%
deceased	1.78%	7.13%	21.58%	31.09%	11.09%	20.99%	6.34%		5.02%

Principle of tree induction

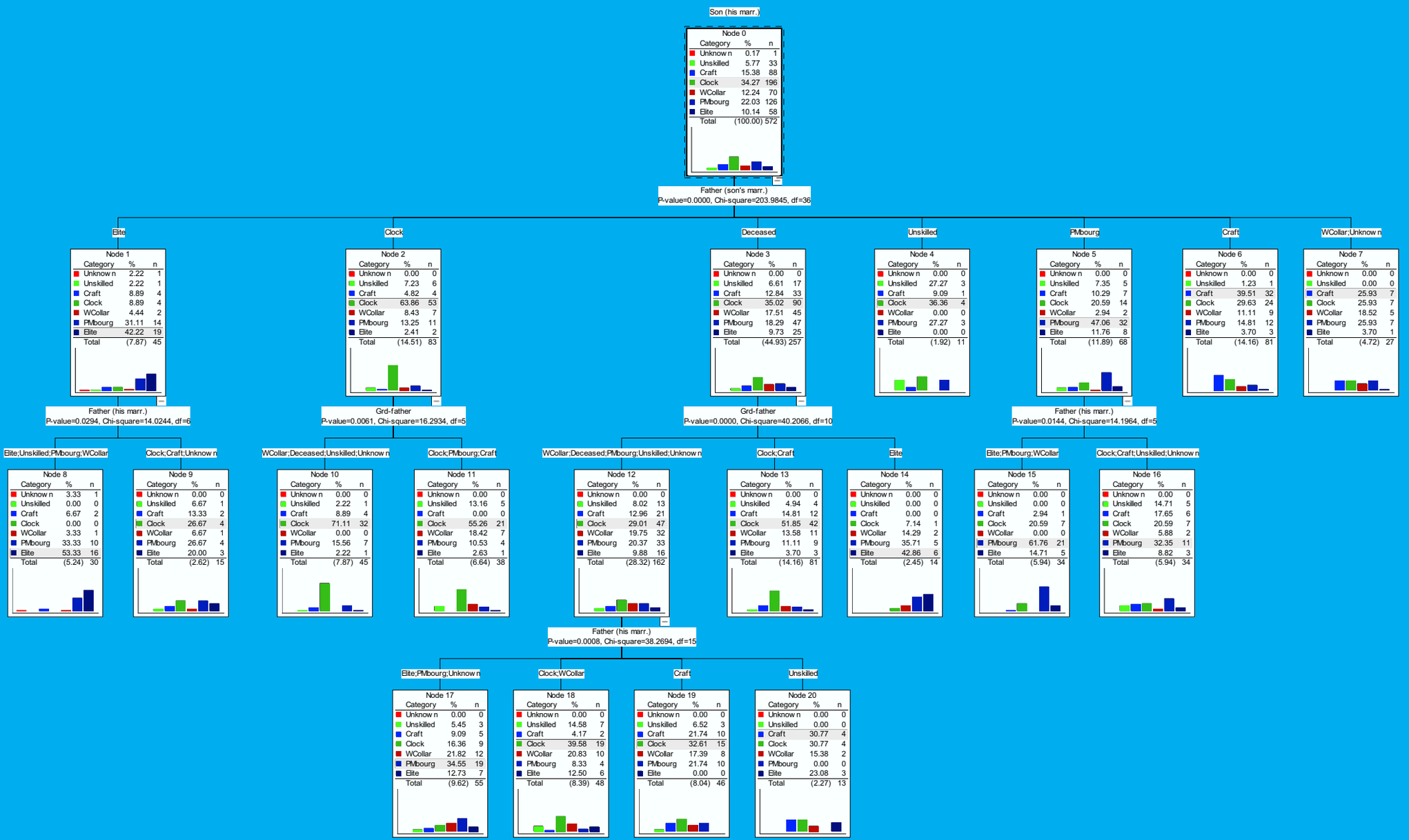
Goal: Find a partition of data such that the distribution of the outcome variable differs as much as possible from one leaf to the other.

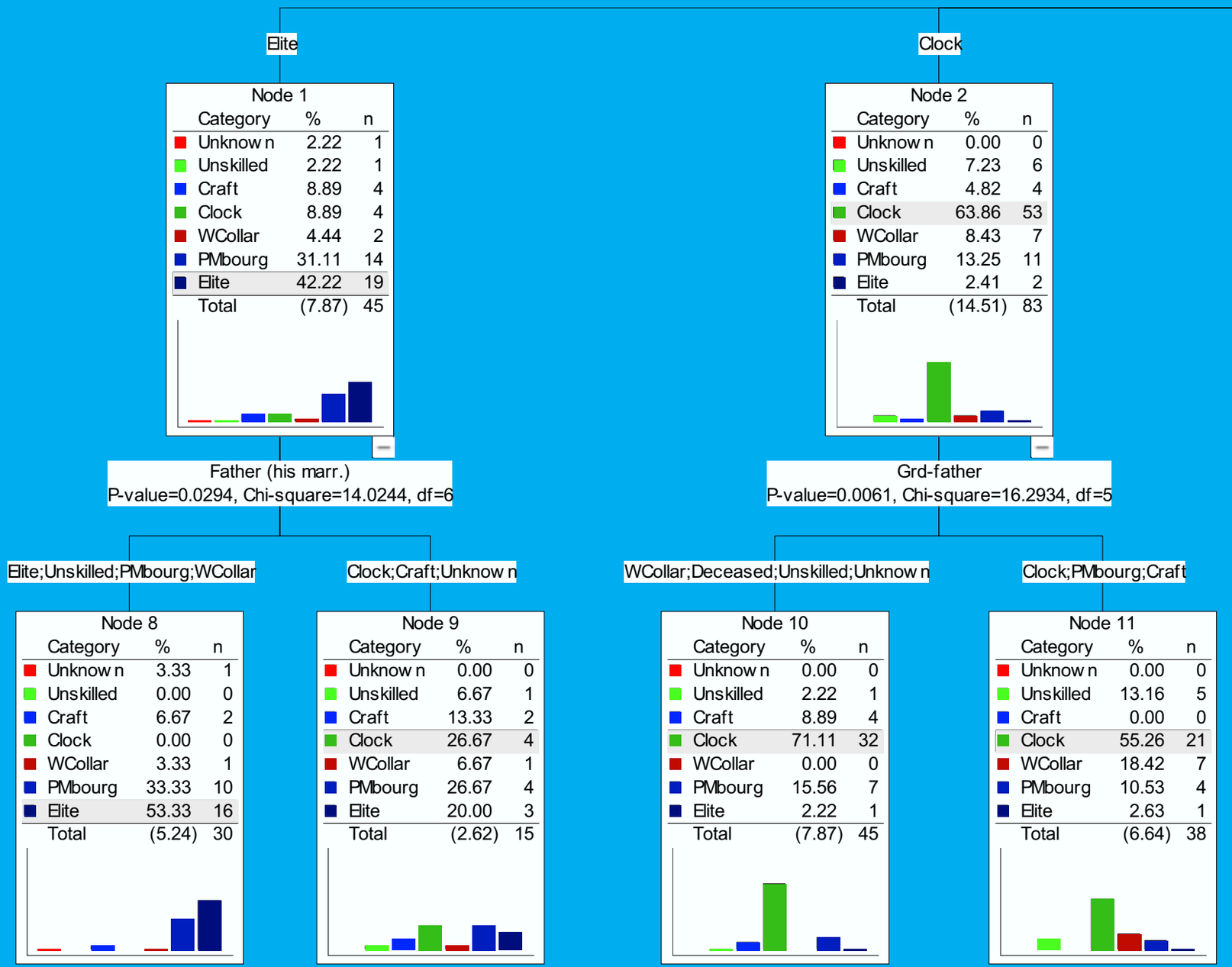
How: Determine the partition by successively splitting nodes. Starting with the root node, seek the attribute that generates the best split according to a given criterion. This operation is then repeated at each new node until some stopping criterion, a minimal node size for instance, is met.

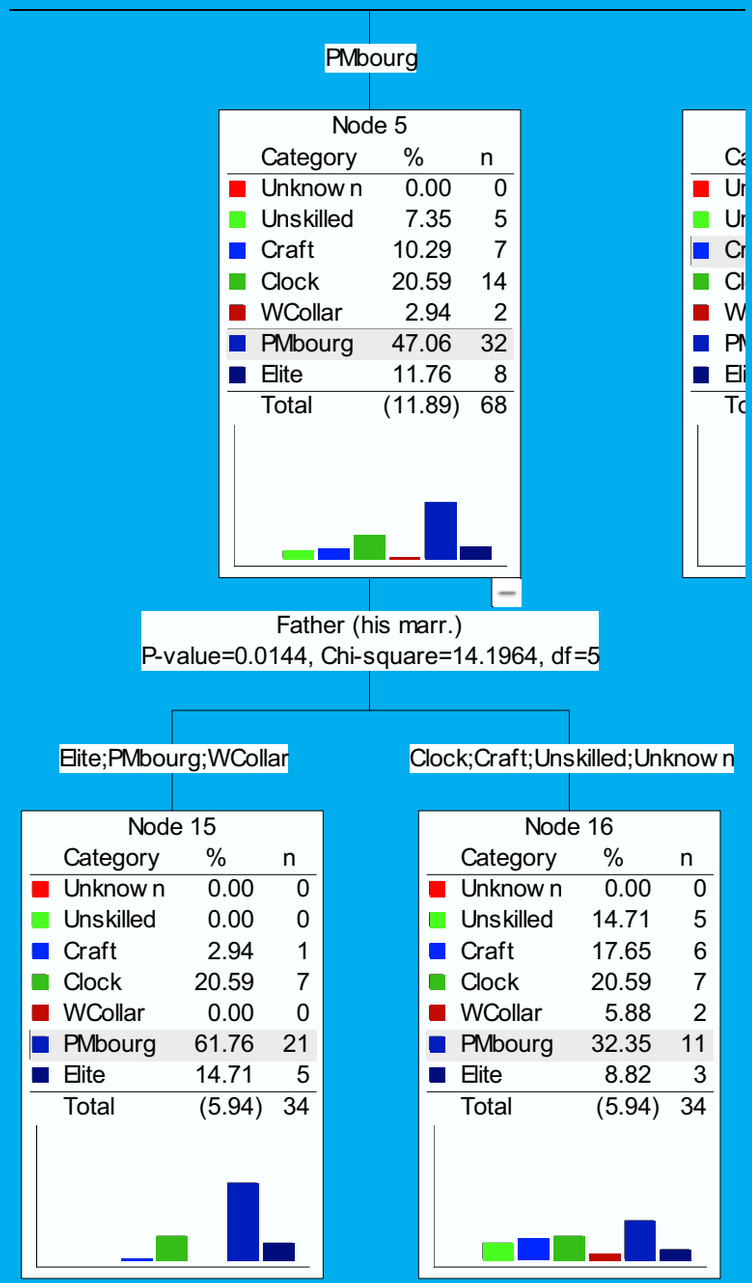
Main algorithms:

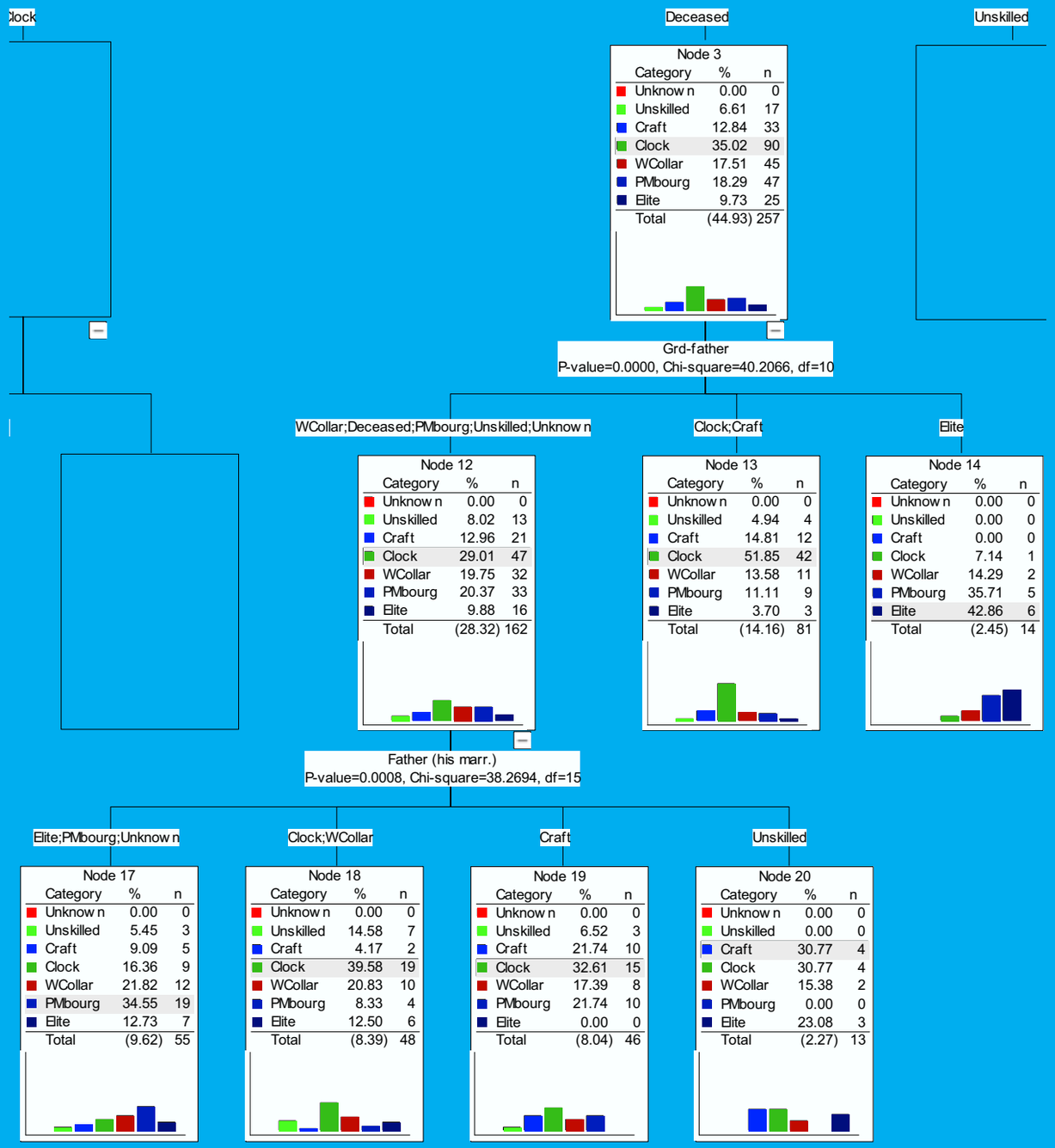
CHAID (Kass, 1980), significance of Chi-2
CART (Breiman et al., 1984), Gini index, binary trees
C4.5 (Quinlan, 1993), gain ratio

For our mobility tree, we used CHAID as implemented in Answer Tree 3.1 (SPSS, 2001)


Son (his marr.)
Node 0
Category % n
Unknown 0.17 1
Unskilled 5.77 33
Craft 15.38 88
Clock 34.27 196
WCollar 12.24 70
PMbourg 22.03 126
Elite 10.14 58
Total (100.00) 572
Father (son's marr.)
P-value=0.0000, Chi-square=203.9845, df=36
Elite
Clock
Deceased
Unskilled
PMbourg
Craft
WCollar;Unknown
Node 1
Category % n
Unknown 2.22 1
Unskilled 2.22 1
Craft 8.89 4
Clock 8.89 4
WCollar 4.44 2
PMbourg 31.11 14
Elite 42.22 19
Total (7.87) 45
Node 2
Category % n
Unknown 0.00 0
Unskilled 7.23 6
Craft 4.82 4
Clock 63.86 53
WCollar 8.43 7
PMbourg 13.25 11
Elite 2.41 2
Total (14.51) 83
Node 3
Category % n
Unknown 0.00 0
Unskilled 6.61 17
Craft 12.84 33
Clock 35.02 90
WCollar 17.51 45
PMbourg 18.29 47
Elite 9.73 25
Total (44.93) 257
Node 4
Category % n
Unknown 0.00 0
Unskilled 27.27 3
Craft 9.09 1
Clock 36.36 4
WCollar 0.00 0
PMbourg 27.27 3
Elite 0.00 0
Total (1.92) 11
Node 5
Category % n
Unknown 0.00 0
Unskilled 7.35 5
Craft 10.29 7
Clock 20.59 14
WCollar 2.94 2
PMbourg 47.06 32
Elite 11.76 8
Total (11.89) 68
Node 6
Category % n
Unknown 0.00 0
Unskilled 1.23 1
Craft 39.51 32
Clock 29.63 24
WCollar 11.11 9
PMbourg 14.81 12
Elite 3.70 3
Total (14.16) 81
Node 7
Category % n
Unknown 0.00 0
Unskilled 0.00 0
Craft 25.93 7
Clock 25.93 7
WCollar 18.52 5
PMbourg 25.93 7
Elite 3.70 1
Total (4.72) 27
Father (his marr.)
P-value=0.0294, Chi-square=14.0244, df=6
Grd-father
P-value=0.0061, Chi-square=16.2934, df=5
Grd-father
P-value=0.0000, Chi-square=40.2066, df=10
Father (his marr.)
P-value=0.0144, Chi-square=14.1964, df=5
Elite;Unskilled;PMbourg;WCollar
Clock;Craft;Unknow
WCollar;Deceased;Unskilled;Unknow
Clock;PMbourg;Craft
WCollar;Deceased;PMbourg;Unskilled;Unknow
Clock;Craft
Elite
Elite;PMbourg;WCollar
Clock;Craft;Unskilled;Unknow
Node 8
Category % n
Unknown 3.33 1
Unskilled 0.00 0
Craft 6.67 2
Clock 0.00 0
WCollar 3.33 1
PMbourg 33.33 10
Elite 53.33 16
Total (5.24) 30
Node 9
Category % n
Unknown 0.00 0
Unskilled 6.67 1
Craft 13.33 2
Clock 26.67 4
WCollar 6.67 1
PMbourg 26.67 4
Elite 20.00 3
Total (2.62) 15
Node 10
Category % n
Unknown 0.00 0
Unskilled 2.22 1
Craft 8.89 4
Clock 71.11 32
WCollar 0.00 0
PMbourg 15.56 7
Elite 2.22 1
Total (7.87) 45
Node 11
Category % n
Unknown 0.00 0
Unskilled 13.16 5
Craft 0.00 0
Clock 55.26 21
WCollar 18.42 7
PMbourg 10.53 4
Elite 2.63 1
Total (6.64) 38
Node 12
Category % n
Unknown 0.00 0
Unskilled 8.02 13
Craft 12.96 21
Clock 29.01 47
WCollar 19.75 32
PMbourg 20.37 33
Elite 9.88 16
Total (28.32) 162
Node 13
Category % n
Unknown 0.00 0
Unskilled 4.94 4
Craft 14.81 12
Clock 51.85 42
WCollar 13.58 11
PMbourg 11.11 9
Elite 3.70 3
Total (14.16) 81
Node 14
Category % n
Unknown 0.00 0
Unskilled 0.00 0
Craft 0.00 0
Clock 7.14 1
WCollar 14.29 2
PMbourg 35.71 5
Elite 42.86 6
Total (2.45) 14
Node 15
Category % n
Unknown 0.00 0
Unskilled 0.00 0
Craft 2.94 1
Clock 20.59 7
WCollar 0.00 0
PMbourg 61.76 21
Elite 14.71 5
Total (5.94) 34
Node 16
Category % n
Unknown 0.00 0
Unskilled 14.71 5
Craft 17.65 6
Clock 20.59 7
WCollar 5.88 2
PMbourg 32.35 11
Elite 8.82 3
Total (5.94) 34
Father (his marr.)
P-value=0.0008, Chi-square=38.2694, df=15
Elite;PMbourg;Unknow
Clock;WCollar
Craft
Unskilled
Node 17
Category % n
Unknown 0.00 0
Unskilled 5.45 3
Craft 9.09 5
Clock 16.36 9
WCollar 21.82 12
PMbourg 34.55 19
Elite 12.73 7
Total (9.62) 55
Node 18
Category % n
Unknown 0.00 0
Unskilled 14.58 7
Craft 4.17 2
Clock 39.58 19
WCollar 20.83 10
PMbourg 8.33 4
Elite 12.50 6
Total (8.39) 48
Node 19
Category % n
Unknown 0.00 0
Unskilled 6.52 3
Craft 21.74 10
Clock 32.61 15
WCollar 17.39 8
PMbourg 21.74 10
Elite 0.00 0
Total (8.04) 46
Node 20
Category % n
Unknown 0.00 0
Unskilled 0.00 0
Craft 30.77 4
Clock 30.77 4
WCollar 15.38 2
PMbourg 0.00 0
Elite 23.08 3
Total (2.27) 13


Elite
Clock
Node 1
Category % n
Unknown 2.22 1
Unskilled 2.22 1
Craft 8.89 4
Clock 8.89 4
WCollar 4.44 2
PMbourg 31.11 14
Elite 42.22 19
Total (7.87) 45
Node 2
Category % n
Unknown 0.00 0
Unskilled 7.23 6
Craft 4.82 4
Clock 63.86 53
WCollar 8.43 7
PMbourg 13.25 11
Elite 2.41 2
Total (14.51) 83
Father (his marr.)
P-value=0.0294, Chi-square=14.0244, df=6
Grd-father
P-value=0.0061, Chi-square=16.2934, df=5
Elite;Unskilled;PMbourg;WCollar
Clock;Craft;Unknown
WCollar;Deceased;Unskilled;Unknown
Clock;PMbourg;Craft
Node 8
Category % n
Unknown 3.33 1
Unskilled 0.00 0
Craft 6.67 2
Clock 0.00 0
WCollar 3.33 1
PMbourg 33.33 10
Elite 53.33 16
Total (5.24) 30
Node 9
Category % n
Unknown 0.00 0
Unskilled 6.67 1
Craft 13.33 2
Clock 26.67 4
WCollar 6.67 1
PMbourg 26.67 4
Elite 20.00 3
Total (2.62) 15
Node 10
Category % n
Unknown 0.00 0
Unskilled 2.22 1
Craft 8.89 4
Clock 71.11 32
WCollar 0.00 0
PMbourg 15.56 7
Elite 2.22 1
Total (7.87) 45
Node 11
Category % n
Unknown 0.00 0
Unskilled 13.16 5
Craft 0.00 0
Clock 55.26 21
WCollar 18.42 7
PMbourg 10.53 4
Elite 2.63 1
Total (6.64) 38


PMbourg
Node 5
Category % n
Unknown 0.00 0
Unskilled 7.35 5
Craft 10.29 7
Clock 20.59 14
WCollar 2.94 2
PMbourg 47.06 32
Elite 11.76 8
Total (11.89) 68
Father (his marr.)
P-value=0.0144, Chi-square=14.1964, df=5
Elite;PMbourg;WCollar
Clock;Craft;Unskilled;Unknown
Node 15
Category % n
Unknown 0.00 0
Unskilled 0.00 0
Craft 2.94 1
Clock 20.59 7
WCollar 0.00 0
PMbourg 61.76 21
Elite 14.71 5
Total (5.94) 34
Node 16
Category % n
Unknown 0.00 0
Unskilled 14.71 5
Craft 17.65 6
Clock 20.59 7
WCollar 5.88 2
PMbourg 32.35 11
Elite 8.82 3
Total (5.94) 34


lock
Deceased
Unskilled
Node 3
Category % n
Unknown 0.00 0
Unskilled 6.61 17
Craft 12.84 33
Clock 35.02 90
WCollar 17.51 45
PMbourg 18.29 47
Elite 9.73 25
Total (44.93) 257
Grd-father
P-value=0.0000, Chi-square=40.2066, df=10
WCollar;Deceased;PMbourg;Unskilled;Unknown
Clock;Craft
Elite
Node 12
Category % n
Unknown 0.00 0
Unskilled 8.02 13
Craft 12.96 21
Clock 29.01 47
WCollar 19.75 32
PMbourg 20.37 33
Elite 9.88 16
Total (28.32) 162
Node 13
Category % n
Unknown 0.00 0
Unskilled 4.94 4
Craft 14.81 12
Clock 51.85 42
WCollar 13.58 11
PMbourg 11.11 9
Elite 3.70 3
Total (14.16) 81
Node 14
Category % n
Unknown 0.00 0
Unskilled 0.00 0
Craft 0.00 0
Clock 7.14 1
WCollar 14.29 2
PMbourg 35.71 5
Elite 42.86 6
Total (2.45) 14
Father (his marr.)
P-value=0.0008, Chi-square=38.2694, df=15
Elite;PMbourg;Unknown
Clock;WCollar
Craft
Unskilled
Node 17
Category % n
Unknown 0.00 0
Unskilled 5.45 3
Craft 9.09 5
Clock 16.36 9
WCollar 21.82 12
PMbourg 34.55 19
Elite 12.73 7
Total (9.62) 55
Node 18
Category % n
Unknown 0.00 0
Unskilled 14.58 7
Craft 4.17 2
Clock 39.58 19
WCollar 20.83 10
PMbourg 8.33 4
Elite 12.50 6
Total (8.39) 48
Node 19
Category % n
Unknown 0.00 0
Unskilled 6.52 3
Craft 21.74 10
Clock 32.61 15
WCollar 17.39 8
PMbourg 21.74 10
Elite 0.00 0
Total (8.04) 46
Node 20
Category % n
Unknown 0.00 0
Unskilled 0.00 0
Craft 30.77 4
Clock 30.77 4
WCollar 15.38 2
PMbourg 0.00 0
Elite 23.08 3
Total (2.27) 13

Tree quality

- Error rate: 55.7%, i.e. 15% reduction of the classification error rate of the initial node which is 65%. Indeed: $(65 - 55.7)/65 = 15\%$
- Goodness-of-fit. See Ritschard and Zighed (2003)

Tree	Variation of the LR Chi-square level 1	level 2	level 3	saturated	pseudo R^2
indep.	173.01 (36 *df*)	263.96 (66 *df*)	309.51 (84 *df*)	791.73 (852 *df*)	0
level 1		90.95 (30 *df*)	136.49 (48 *df*)	618.72 (816 *df*)	.18
level 2			45.55 (18 *df*)	527.77 (786 *df*)	.28
level 3				482.22 (768 *df*)	.32

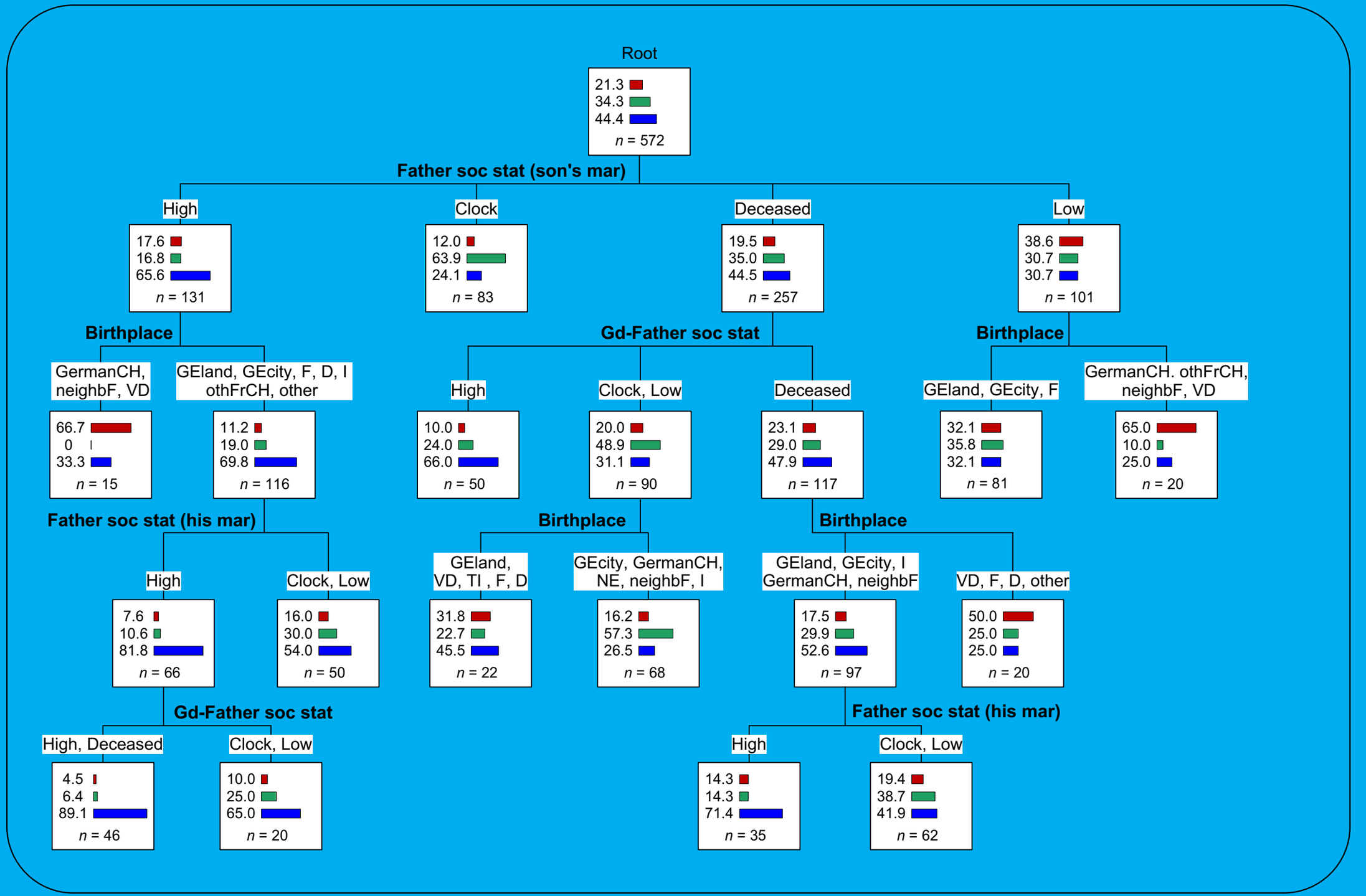
3.5 Social status and geographical origin

Statuses 3 categories

Low	unknown
	unskilled
	craft
Clock	clock
High	white collar
	PMB
	elite

Birth place 12 values:

GEcity	Geneva city
GEland	Geneva surrounding land
neighbF	neighboring France
VD	Vaud
NE	Neuchatel
otherFrCH	other French speaking Switzerland
GermanCH	German speaking Switzerland
TI	Italian speaking Switzerland
F	France
D	Germany
I	Italy
other	other


Root
21.3
34.3
44.4
n = 572
Father soc stat (son's mar)
High
17.6
16.8
65.6
n = 131
Clock
12.0
63.9
24.1
n = 83
Deceased
19.5
35.0
44.5
n = 257
Low
38.6
30.7
30.7
n = 101
Birthplace
GermanCH, neighbF, VD
66.7
0
33.3
n = 15
GEland, GEcity, F, D, I othFrCH, other
11.2
19.0
69.8
n = 116
Gd-Father soc stat
High
10.0
24.0
66.0
n = 50
Clock, Low
20.0
48.9
31.1
n = 90
Deceased
23.1
29.0
47.9
n = 117
Birthplace
GEland, GEcity, F
32.1
35.8
32.1
n = 81
GermanCH. othFrCH, neighbF, VD
65.0
10.0
25.0
n = 20
Father soc stat (his mar)
High
7.6
10.6
81.8
n = 66
Clock, Low
16.0
30.0
54.0
n = 50
Birthplace
GEland, VD, TI , F, D
31.8
22.7
45.5
n = 22
GEcity, GermanCH, NE, neighbF, I
16.2
57.3
26.5
n = 68
Birthplace
GEland, GEcity, I GermanCH, neighbF
17.5
29.9
52.6
n = 97
VD, F, D, other
50.0
25.0
25.0
n = 20
Gd-Father soc stat
High, Deceased
4.5
6.4
89.1
n = 46
Clock, Low
10.0
25.0
65.0
n = 20
Father soc stat (his mar)
High
14.3
14.3
71.4
n = 35
Clock, Low
19.4
38.7
41.9
n = 62

Tree quality

- Error rate: 42.4%, i.e. 24% reduction of the classification error rate of the initial node
- Goodness of fit

Tree	$G2$	df	sig	BIC	AIC	pseudo R^2
Indep	482.3	324	0.000	2319.6	812.3	0
Level 1	408.2	318	0.000	1493.9	750.2	0.14
Level 2	356.0	310	0.037	1492.5	714.0	0.23
Level 3	327.6	304	0.168	1502.2	697.6	0.28
Fitted	312.5	300	0.298	1512.5	690.5	0.30
Saturated	0	0	1	3104.7	978.0	1

4 Event sequences with most varying frequencies

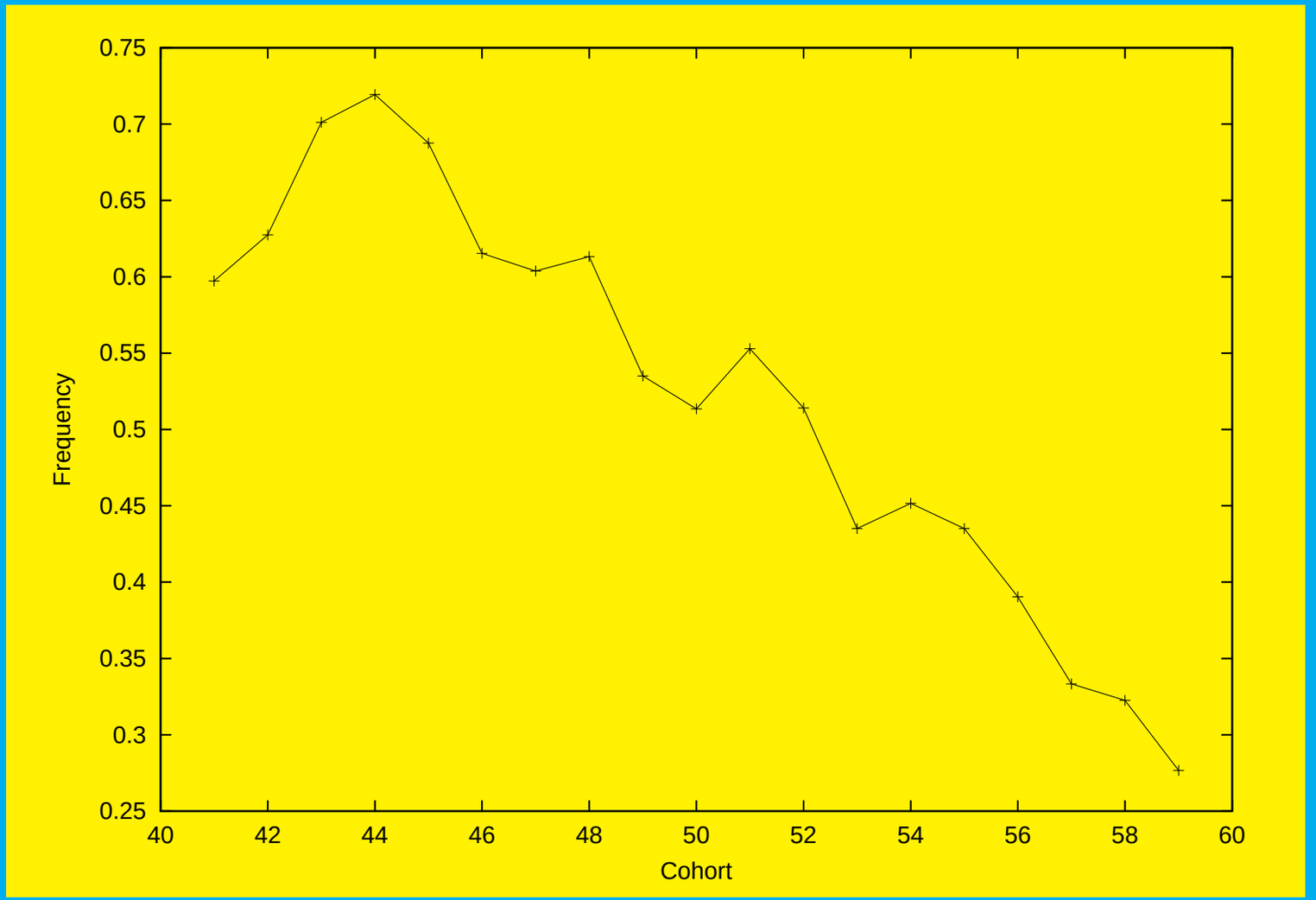
Algorithm for mining frequent sequences (Agrawal and Srikant, 1995; Mannila et al., 1997) are derived from those for mining frequent itemsets, essentially `apriori` (Agrawal and Srikant, 1994; Mannilia et al., 1994)

Blockeel et al. (2001) have experimented this approach for discovering frequent partnership and birth event patterns that mostly varied among (year) cohorts.

Data : 1995 Austrian Fertility and Family Survey (FFS).

Retrospective histories of 4,581 women and 1,539 men aged between 20 and 54 at the survey time $\Rightarrow$ cohorts $=$ 41 to 75.

Example of outcome:



Negative trend in the proportion of first unions starting at marriage

5 Other examples from the literature

De Rose and Pallara (1997) study the duration in years between 16th birthday and marriage on a sample of about 1500 Italian women.

They use survival trees, a method originated in biostatistics at the end of the 80's, (Segal, 1988; Ciampi et al., 1988)

A survival tree successively splits the data such that the survival curves estimated for each node are as different as possible.

Billari et al. (2000) use classification trees and induction of rule sets for discriminating Austrian and Italian behaviors in terms of time until leaving home, marriage, 1st child, end of formation and first job.

Propose a triple coding of the data in terms of quantum (does the event happen?), timing (when?) and sequencing.

References

Agrawal, R. and Srikant, R. (1994). Fast algorithm for mining association rules in large databases. In *Proceedings 1994 International Conference on Very Large Data Base (VLDB'94)*, pages 487–499, Santiago, Chile.

Agrawal, R. and Srikant, R. (1995). Mining sequential patterns. In *Proceedings of the International Conference on Data Engeneering (ICDE)*, pages 487–499, Taipei, Taiwan.

Billari, F. C., Fürnkranz, J., and Prskawetz, A. (2000). Timing, sequencing, and quantum of life course events: a machine learning approach. Working Paper 010, Max Planck Institute for Demographic Research, Rostock.

Blockeel, H., Fürnkranz, J., Prskawetz, A., and Billari, F. (2001). Detecting temporal change in event sequences: An application to demographic data. In De Raedt, L. and Siebes, A., editors, *Principles of Data Mining and Knowledge Discovery: 5th European Conference, PKDD 2001*, volume LNCS 2168, pages 29–41. Springer, Freiburg in Brisgau.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification And Regression Trees*. Chapman and Hall, New York.

Breiman, L. (2001). Satistical modeling: The two cultures (with discussion). *Statistical Science*, 16(3):199–231.

Ciampi, A., Hogg, S. A., McKinney, S., and Thiffault, J. (1988). RECPAM: a computer program for recursive partitioning and amalgamation for censored survival data and other

situations frequently occuring in biostatistics i. methods and program features. *Computer Methods and Programs in Biomedicine*, 26(3):239–256.

De Rose, A. and Pallara, A. (1997). Survival trees: An alternative non-parametric multivariate technique for life history analysis. *European Journal of Population*, 13:223–241.

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

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

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29(2):119–127.

Mannila, H., Toivonen, H., and Verkamo, A. I. (1997). Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery*, 1(3):259–289.

Mannilia, H., Toivonen, H., and Verkamo, A. I. (1994). Efficient algorithms for discovering association rules. In *Proceedings AAAI'94 Workshop Knowledge Discovery in Databases (KDD'94)*, pages 181–192, Seattle, WA.

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

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo.

Ritschard, G. and Oris, M. (forthcoming). Dealing with life course data in demography: Statistical and data mining approaches. In Levy, R., Widmer, E., Spini, D., Le Goff, J.-M., and Ghisletta, P., editors, *Advances in Interdisciplinary Life Course Research*, page 24. PaVie, Lausanne.

Ritschard, G. and Zighed, D. A. (2003). Goodness-of-fit measures for induction trees. In Zhong, N., Ras, Z., Tsumo, S., and Suzuki, E., editors, *Foundations of Intelligent Systems, ISMIS03*, volume LNAI 2871, pages 57–64. Springer, Berlin.

Ryczkowska, G. and Ritschard, G. (2004). Mobilités sociales et spatiales: Parcours intergénérationnels d'après les mariages genevois, 1830-1880. In *Fifth European Social Science History Conference ESSHC*, Berlin.

Ryczkowska, G. (2003). Accès au mariage et structure de l'alliance à genève, 1800-1880. Mémoire de dea, Déartement d'histoire économique, Université de Genève, Genève.

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

SPSS, editor (2001). *Answer Tree 3.0 User's Guide*. SPSS Inc., Chicago.