WZB

Wissenschaftszentrum Berlin
für Sozialforschung

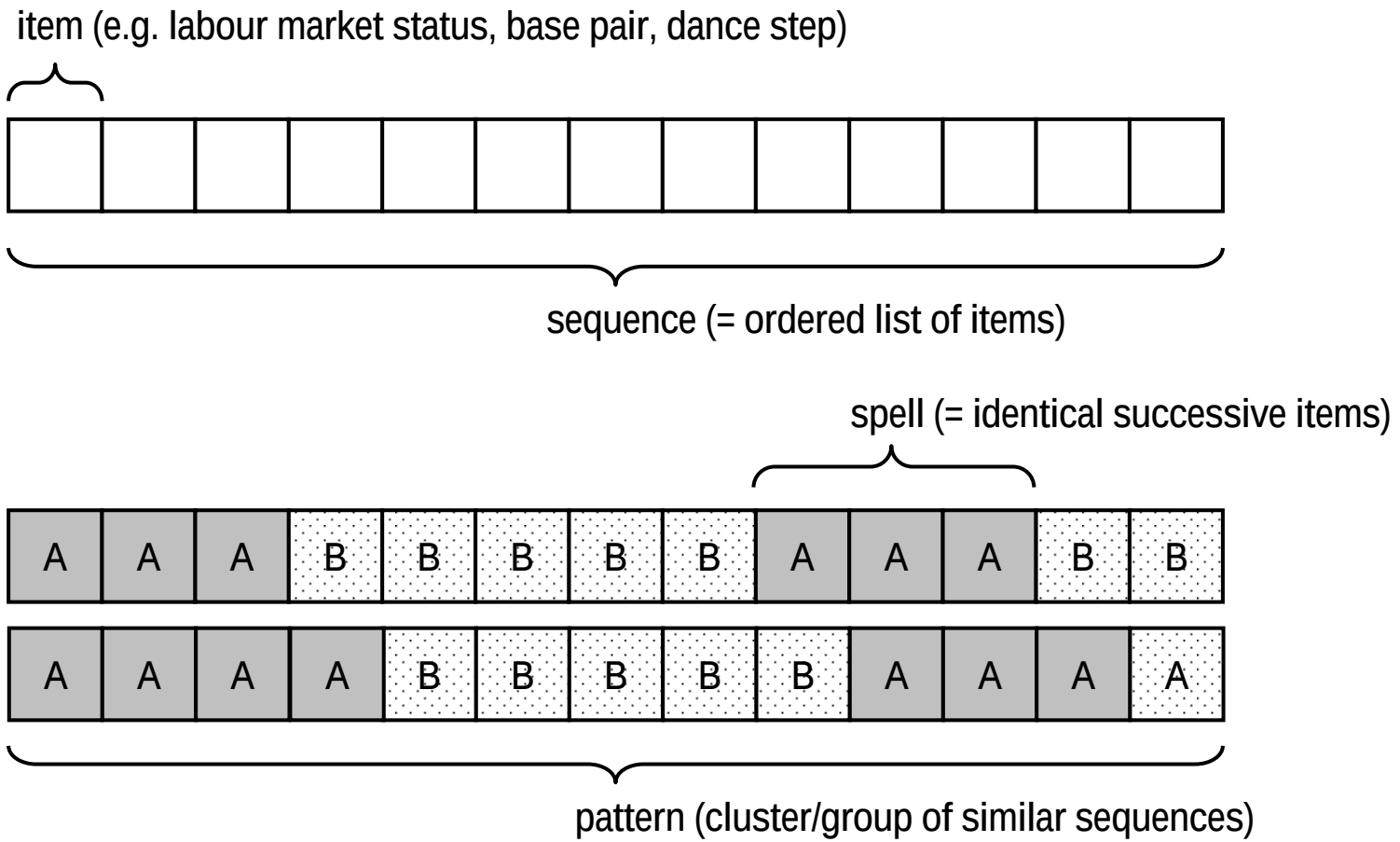
Sequence Analysis Using STATA

The Optimal Matching Technique and an Application Example in Social Science

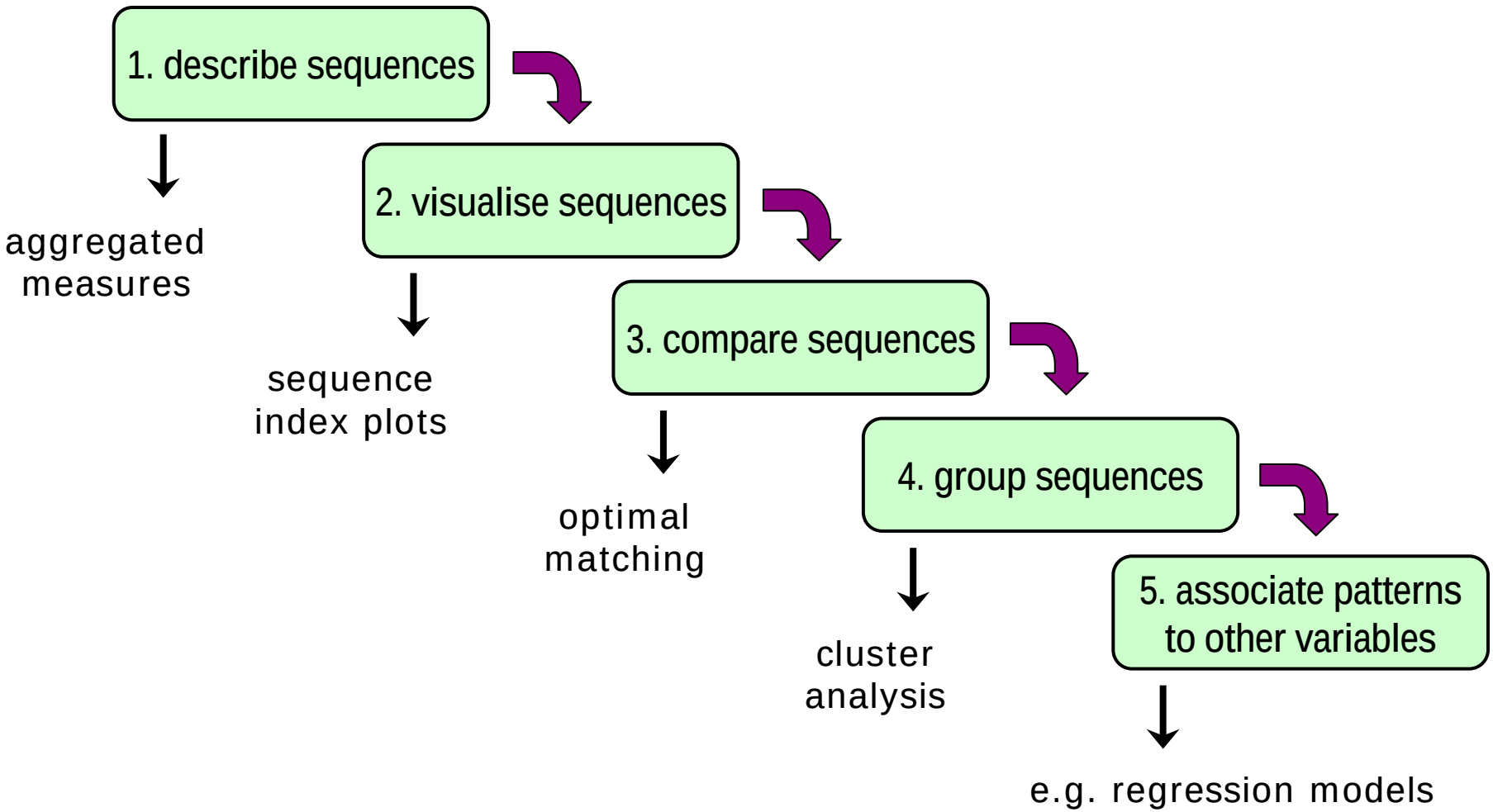
Christian Brzinsky-Fay

Mannheim, 31st March, 2006

concepts


item (e.g. labour market status, base pair, dance step)
sequence (= ordered list of items)
spell (= identical successive items)
A A A B B B B B A A A B B
A A A A B B B B B A A A A
pattern (cluster/group of similar sequences)

typical research design



optimal matching – example

substitution costs: 1
indel costs: 1

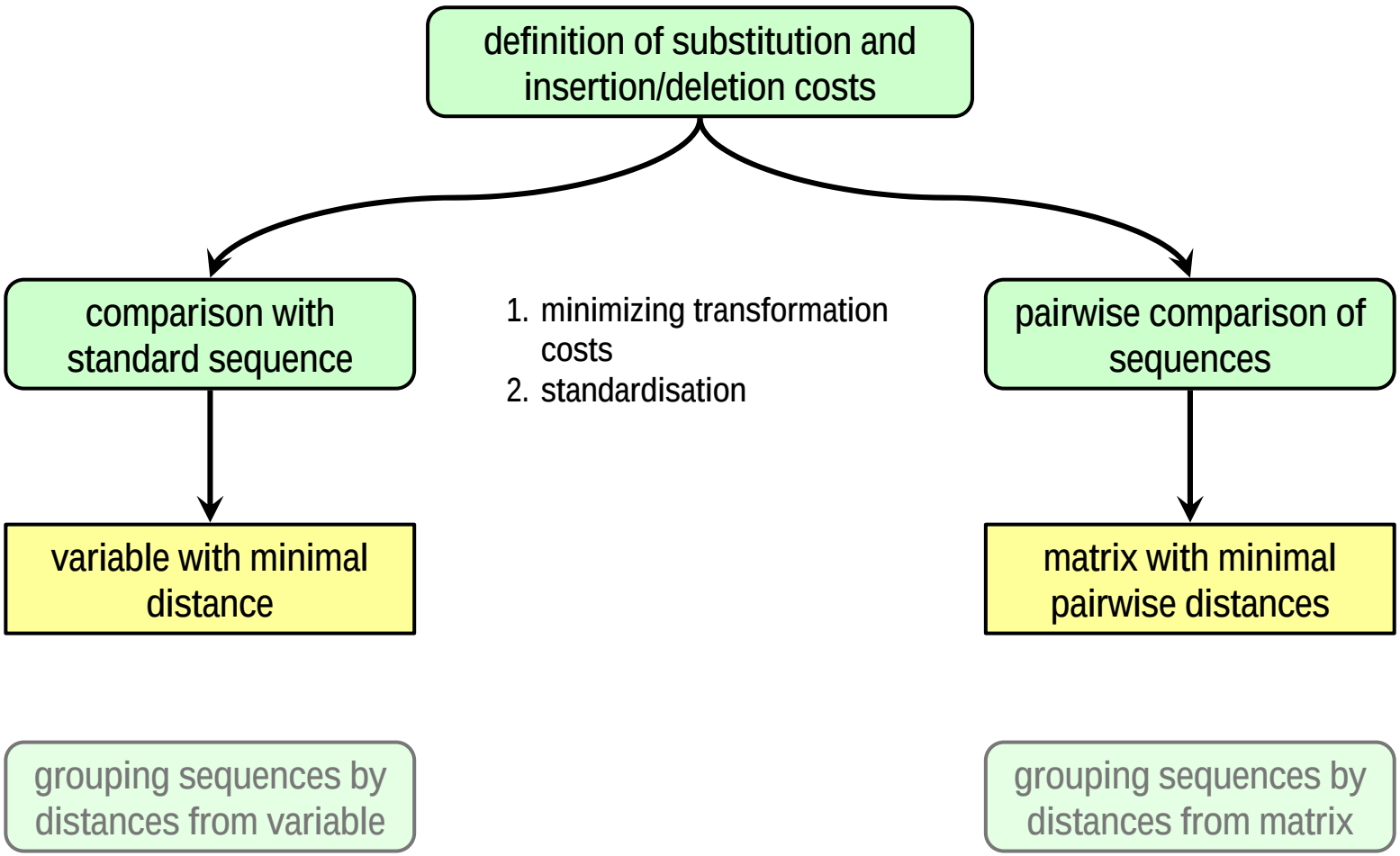
```
B E A T L E S
S T O N E S -
-------------
1 1 1 1 1 1 1 = 7
                7/7 = 1
```

```
B E A T L E S
K I N K S - -
-------------
1 1 1 1 1 1 1 = 7
                7/7 = 1
```

```
B E A T L E S
- S T O N E S
---------------
1 1 1 1 1 0 0 = 5
                5/7 ≈ 0.71
```

```
B E A T L E S
- - K I N K S
-------------
1 1 1 1 1 1 0 = 6
                6/7 ≈ 0.86
```

optimal matching – technique



WZB

Wissenschaftszentrum Berlin
für Sozialforschung

Example for Sequence Analysis

Labour Market Entry Patterns of School Leavers in Europe

graphs can be downloaded from the author's website:

www.wz-berlin.de/~brzinsky-fay

overview

- population: people < 25 years
- period of labour market integration begins after leaving school & ends five years later
- items of labour market status:
  1. education (further education, university etc.)
  2. apprenticeship
  3. employment
  4. unemployment
  5. inactivity (household, military/civil service)
- time unit: months

description 1

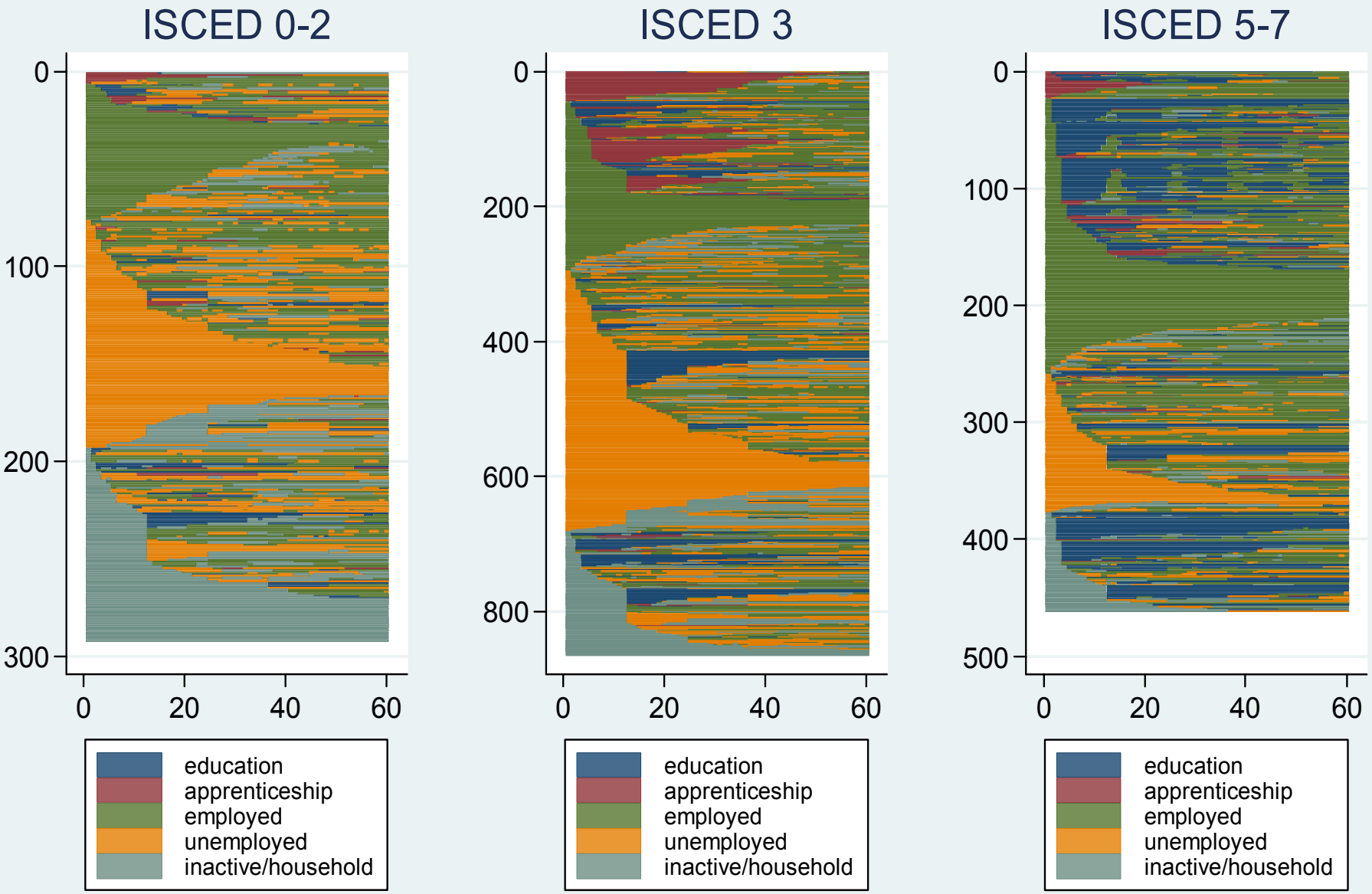
		DK	BE	FR	IE	IT	GR	ES	PT	DE	UK	Total
average duration in...	education	19.7	26.2	12.6	17.8	11.4	7.1	14.0	12.7	5.5	5.5	11.5
	apprenticeship	8.4	3.1	0.1	3.5	1.5	0.3	1.3	0.1	19.7	1.4	3.6
	employment	23.3	18.3	27.6	31.0	13.8	19.2	19.8	26.6	24.3	37.8	23.6
	unemployment	4.4	9.9	9.9	5.4	25.8	20.0	17.6	10.2	5.1	8.2	14.0
	inactivity	4.2	2.5	9.8	2.2	7.5	13.5	7.3	10.3	5.4	7.2	7.3
average number of different statuses		3.3	2.5	2.6	2.4	2.4	2.6	2.7	2.4	2.8	2.2	2.5
average number of episodes (total)		5.8	3.7	4.8	4.9	3.7	3.7	5.0	3.8	4.2	4.0	4.3
average number of episodes in...	education	1.2	0.8	0.7	1.5	0.6	0.4	0.9	0.6	0.3	0.4	0.7
	apprenticeship	0.6	0.3	0.0	0.2	0.2	0.1	0.2	0.0	0.9	0.2	0.2
	employment	2.2	1.1	1.9	2.2	0.9	1.1	1.8	1.3	1.8	1.9	1.6
	unemployment	0.9	0.8	1.3	0.7	1.4	1.3	1.5	0.9	0.8	1.2	1.2
	inactivity	0.9	0.8	0.9	0.3	0.6	0.9	0.7	0.9	0.4	0.4	0.6
volatility indicator		2.1	1.4	1.2	4.0	0.8	0.7	1.3	1.1	2.6	1.5	1.4
N		52	48	160	164	361	156	206	101	177	208	1633

description 2

	sequence order	frequency	percent	cumulated	sequence type
1.	UE - EM	95	5.8	5.8	detour
2.	EM	84	5.1	10.9	rupture
3.	UE	54	3.3	14.2	rupture
4.	IN - ED	48	2.9	17.1	return
5.	UE - EM - UE - EM	40	2.4	19.6	detour
6.	EM - AP	39	2.4	21.9	fusion
7.	UE - ED	35	2.1	24.1	return
8.	IN	32	2.0	26.0	rupture
9.	EM - ED	29	1.8	27.8	interruption
10.	AP - EM	26	1.6	29.4	bridge
11.	UE - IN	26	1.6	31.0	change
12.	IN - EM	25	1.5	32.5	detour
13.	EM - UE - EM	24	1.5	33.9	detour
14.	EM - ED - EM	23	1.4	35.3	detour
15.	EM - IN - EM	22	1.3	36.7	detour

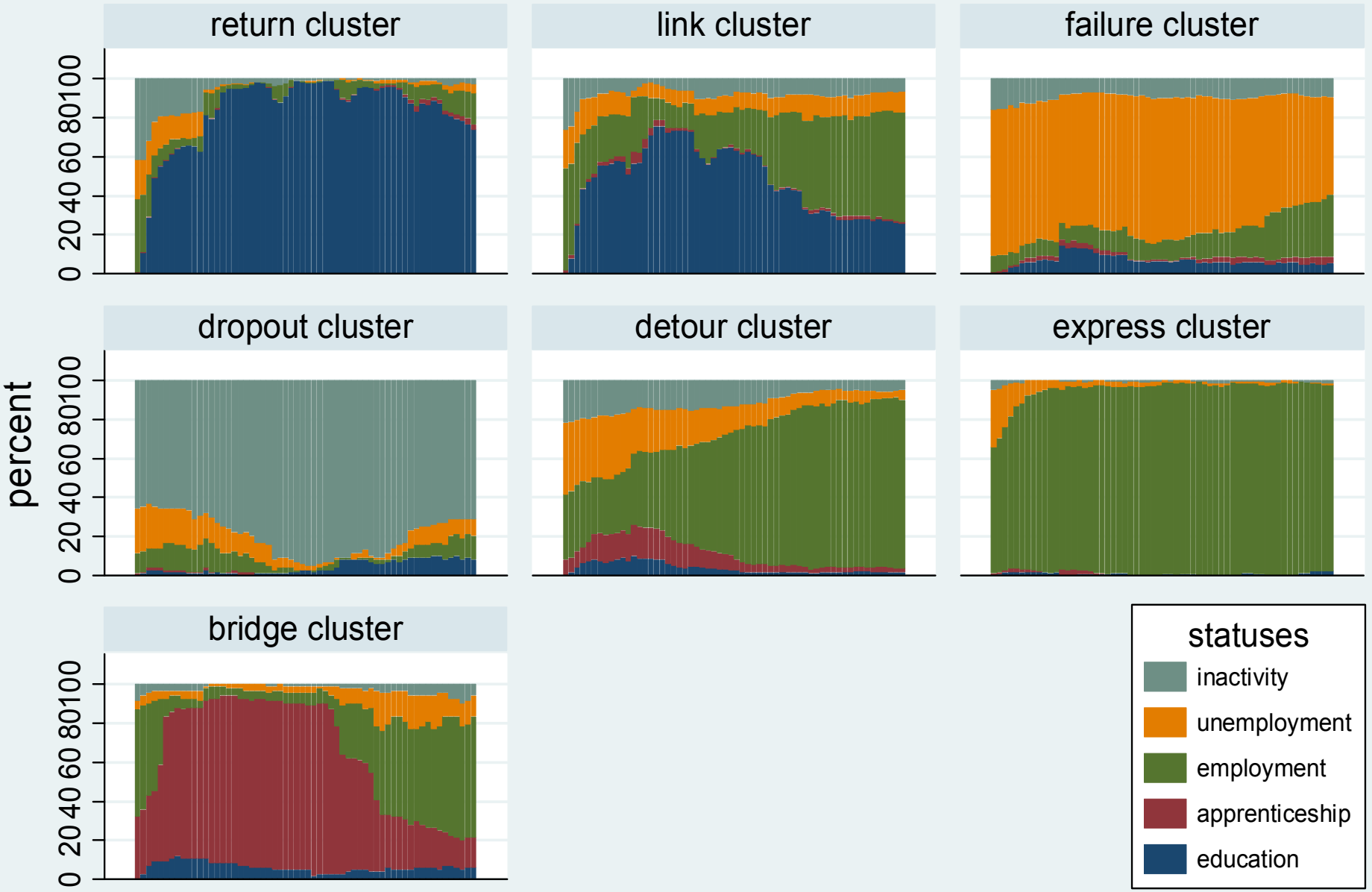
sequence index plots by level of education

source: ECHP, own calculations
(c) Christian Brzinsky-Fay (2006)



monthly proportion of statuses by cluster

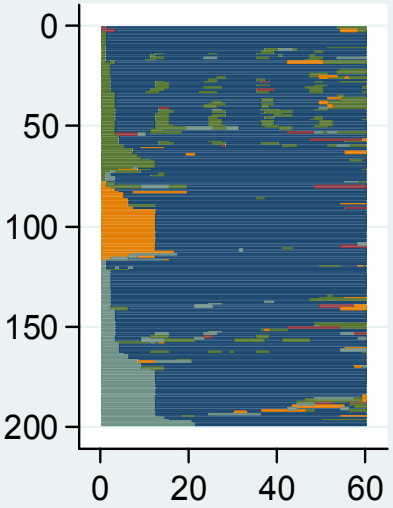
(c) Christian Brzinsky-Fay (2006)
source: ECHP, own calculations



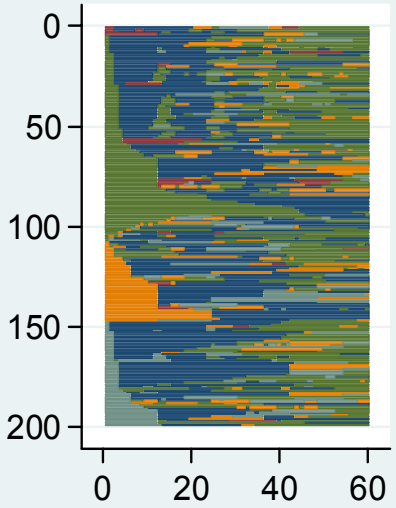
sequence index plots by cluster

(c) Christian Brzinsky-Fay (2006)
source: ECHP, own calculations

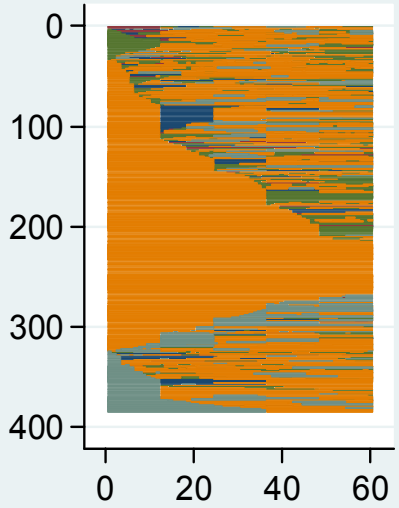
return cluster



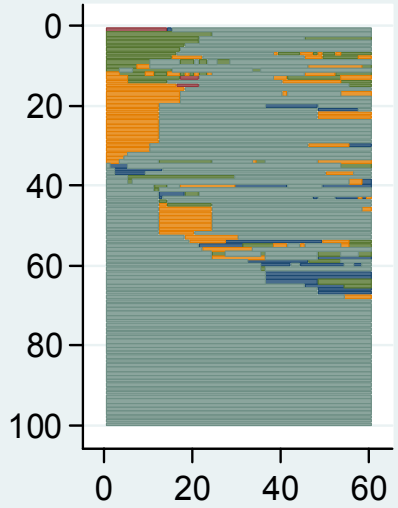
link cluster



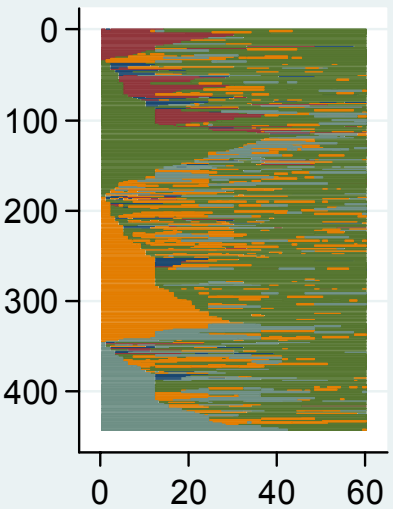
failure cluster



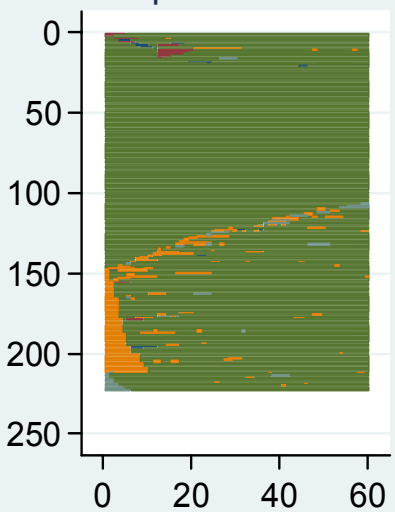
dropout cluster



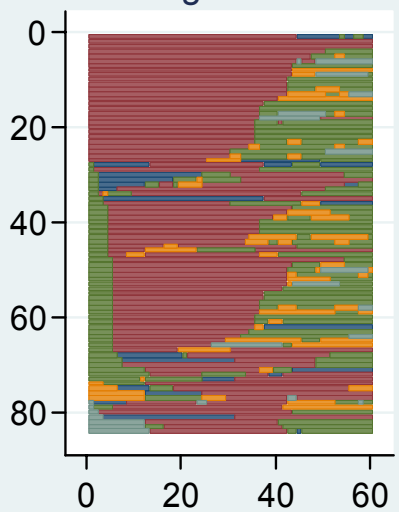
detour cluster



express cluster



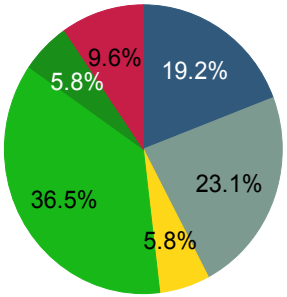
bridge cluster



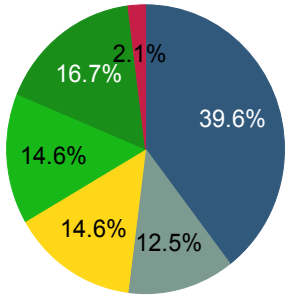
appearance of clusters in countries

(c) Christian Brzinsky-Fay (2006)
source: ECHP, own calculations

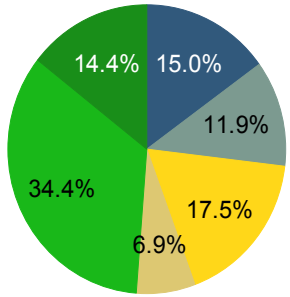
Denmark



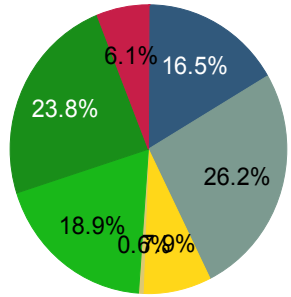
Belgium



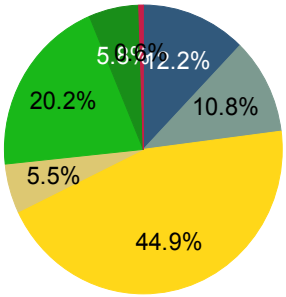
France



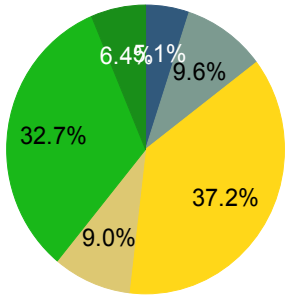
Ireland



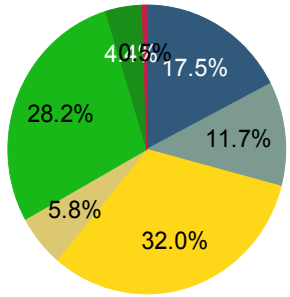
Italy



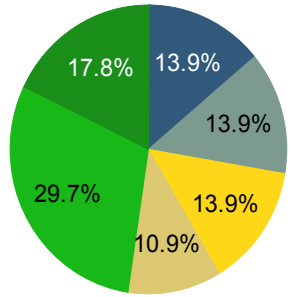
Greece



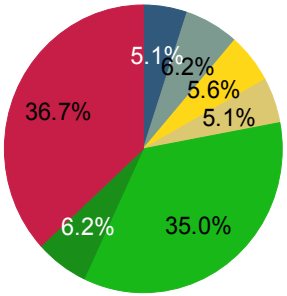
Spain



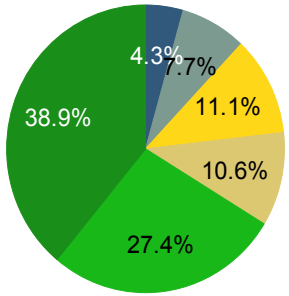
Portugal

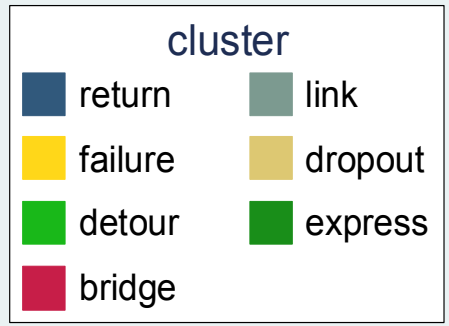


Germany



United Kingdom





notes on the software demonstration

1. extract the zip-archive to an arbitrary directory
2. from Stata change with **cd** to this directory
3. start software demonstration with view **`cover1.smcl`**
4. click on the links (blue) to perform the respective commands and
5. watch the results by switching to the Stata window