

Mining Sequence Data

Hazim Fitri

2024-12-22

Contents

ETL	3
EDA	3
Convert data into sequence format	5
Plot jujukan kekerapan	17
Plot taburan keadaan	18
Entropi	20
Transversal Entropy	21
sub jujukan	21

Categorical sequence data analysis. This type of data refers to observation of a particular individuals or entities over some period of time.

Objective :

- Analyze the behavior of sequence of states for a particular individuals or entities.
 - What are the characteristics of a sequence data?
 - What are the indicators that can be used to measure sequence data?
 - What are the appropriate plots to visualize a sequence data?
 - How can we compare the similarity between several sequence data?

Example :

- DNA sequence
- Life trajectory (occupation history, patient level history, cohabitation life course)
- Domain (biology, QC, text data, log-web data)

Sequential analysis technique

- Statistical summary indicators
 - Mean time spend in each state
 - Mean time spent in each state by groups
 - Number of transitions
 - Transition rates
 - Time varying transition states
- Visualization
 - Sequence index plot
 - Sequence frequency plot
 - State distribution plot

- Modal state plot
- Grouping
- Comparing sequences.

Types of sequence data :

- State-Sequence (STS)
- State-Permanence-Sequence (SPS) format
- Time-Stamped-Event (TSE) format
- SPELL format

Sequence characteristics by Entropy

- Visualization
 - Transversal Entropies
- Event sequence
- Categorizing patterns
 - State distribution
 - Sequence frequencies
 - Modal state
 - Discriminating transitions
- Sequence analysis
 - Other approaches
 - * Correspondence analysis of the states
 - * Markov modeling
 - * Event sequence analysis
 - * Survival analysis
 - * Longitudinal analysis
 - * Discrete panel data analysis

Case Study

-

Code	Example											
STS	Id	18	19	20	21	22	23	24	25	26	27	
	101	S	S	S	M	M	MC	MC	MC	MC	D	
	102	S	S	S	MC	MC	MC	MC	MC	MC	MC	
SPS (1)	Id	State 1		State 2		State 3		State 4		State 5		
	101	(S,3)		(M,2)		(MC,4)		(D,1)				
	102	(S,3)		(MC,7)								
SPS (2)	Id	State 1		State 2		State 3		State 4		State 5		
	101	S/3		M/2		MC/4		D/1				
	102	S/3		MC/7								
DSS	Id	State 1		State 2		State 3		State 4		State 5		
	101	S		M		MC		D				
	102	S		MC								
TSE	id	time		event								
	101	21		Marriage								
	101	23		Child								
	101	27		Divorce								
	102	21		Marriage								
	102	21		Child								
SPELL	id	index	from	to	status							
	101	1	18	20	Single							
	101	2	21	22	Married							
	101	3	23	26	Married w Children							
	101	4	27	..	Divorced							
	102	1	18	20	Single							
	102	2	21	27	Married w Children							

ETL

```
library(TraMineR)

## Warning: package 'TraMineR' was built under R version 4.4.2

##
## TraMineR stable version 2.2-11 (Built: 2025-02-20)

## Website: http://traminer.unige.ch

## Please type 'citation("TraMineR")' for citation information.

data("mvad")
class(mvad)

## [1] "data.frame"

class(mvad)

## [1] "data.frame"
```

EDA

```
str(mvad)

## 'data.frame':    712 obs. of  86 variables:
## $ id      : int  1 2 3 4 5 6 7 8 9 10 ...
## $ weight  : num  0.33 0.57 1.59 1.59 0.57 1.59 0.57 2.75 2 3.6 ...
## $ male    : Factor w/ 2 levels "no","yes": 1 1 2 1 2 2 2 2 1 1 ...
## $ catholic: Factor w/ 2 levels "no","yes": 1 1 2 1 1 2 2 2 1 1 ...
```

[illegible]

```
## $ Jan.98 : Factor w/ 6 levels "school","FE",...: 3 6 2 3 6 3 3 3 3 3 ...
## $ Feb.98 : Factor w/ 6 levels "school","FE",...: 3 6 2 3 6 3 3 3 3 3 ...
## $ Mar.98 : Factor w/ 6 levels "school","FE",...: 3 6 2 3 6 3 3 3 3 3 ...
## $ Apr.98 : Factor w/ 6 levels "school","FE",...: 3 6 2 3 6 3 3 3 3 3 ...
## $ May.98 : Factor w/ 6 levels "school","FE",...: 3 6 2 3 6 3 3 3 3 3 ...
## $ Jun.98 : Factor w/ 6 levels "school","FE",...: 3 6 2 3 6 3 3 3 3 3 ...
## $ Jul.98 : Factor w/ 6 levels "school","FE",...: 3 6 3 3 6 3 3 3 3 5 ...
## $ Aug.98 : Factor w/ 6 levels "school","FE",...: 3 6 3 3 6 3 3 3 3 5 ...
## $ Sep.98 : Factor w/ 6 levels "school","FE",...: 3 6 3 3 6 3 3 3 3 5 ...
## $ Oct.98 : Factor w/ 6 levels "school","FE",...: 3 6 3 5 6 3 3 3 3 5 ...
## $ Nov.98 : Factor w/ 6 levels "school","FE",...: 3 6 3 5 6 3 3 3 3 5 ...
## $ Dec.98 : Factor w/ 6 levels "school","FE",...: 3 6 3 5 6 3 3 3 3 5 ...
## $ Jan.99 : Factor w/ 6 levels "school","FE",...: 3 6 3 5 6 3 3 3 3 5 ...
## $ Feb.99 : Factor w/ 6 levels "school","FE",...: 3 6 3 5 6 3 3 3 3 5 ...
## $ Mar.99 : Factor w/ 6 levels "school","FE",...: 3 6 3 5 6 3 3 3 3 5 ...
## $ Apr.99 : Factor w/ 6 levels "school","FE",...: 3 6 3 5 6 3 3 3 3 5 ...
## $ May.99 : Factor w/ 6 levels "school","FE",...: 3 6 5 5 6 3 3 3 3 5 ...
## $ Jun.99 : Factor w/ 6 levels "school","FE",...: 3 6 5 5 6 3 3 3 3 5 ...
```

Column 1 - 14 is demography information and not sequence data

Sequence data is column 15 - 86

Convert data into sequence format

Define label and code for each state

```
unique(mvad$Jul.93)
```

```
## [1] training    joblessness employment school      FE
## Levels: school FE employment training joblessness HE
```

```
mvad.labels=c('bekerja', 'sambung belajar', 'pengjian tinggi', 'penganggur',
              'sekolah', 'latihan')
mvad.scode = c('EM', 'FE', 'HE', 'JL', 'SC', 'TR')
mvad.seq = seqdef(mvad, 15:86, states=mvad.scode, labels = mvad.labels, xstep = 6)
```

```
## [>] state coding:
```

```
##      [alphabet] [label] [long label]

##      1  employment  EM      bekerja

##      2  FE          FE      sambung belajar

##      3  HE          HE      pengjian tinggi

##      4  joblessness JL      penganggur

##      5  school      SC      sekolah

##      6  training    TR      latihan
```

```
## [>] 712 sequences in the data set
```

```
## [>] min/max sequence length: 72/72
```


[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

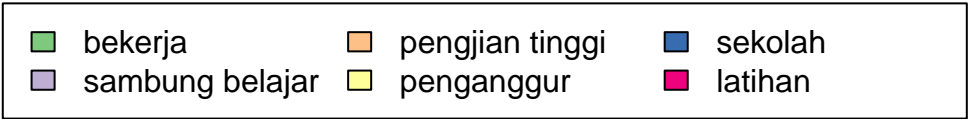
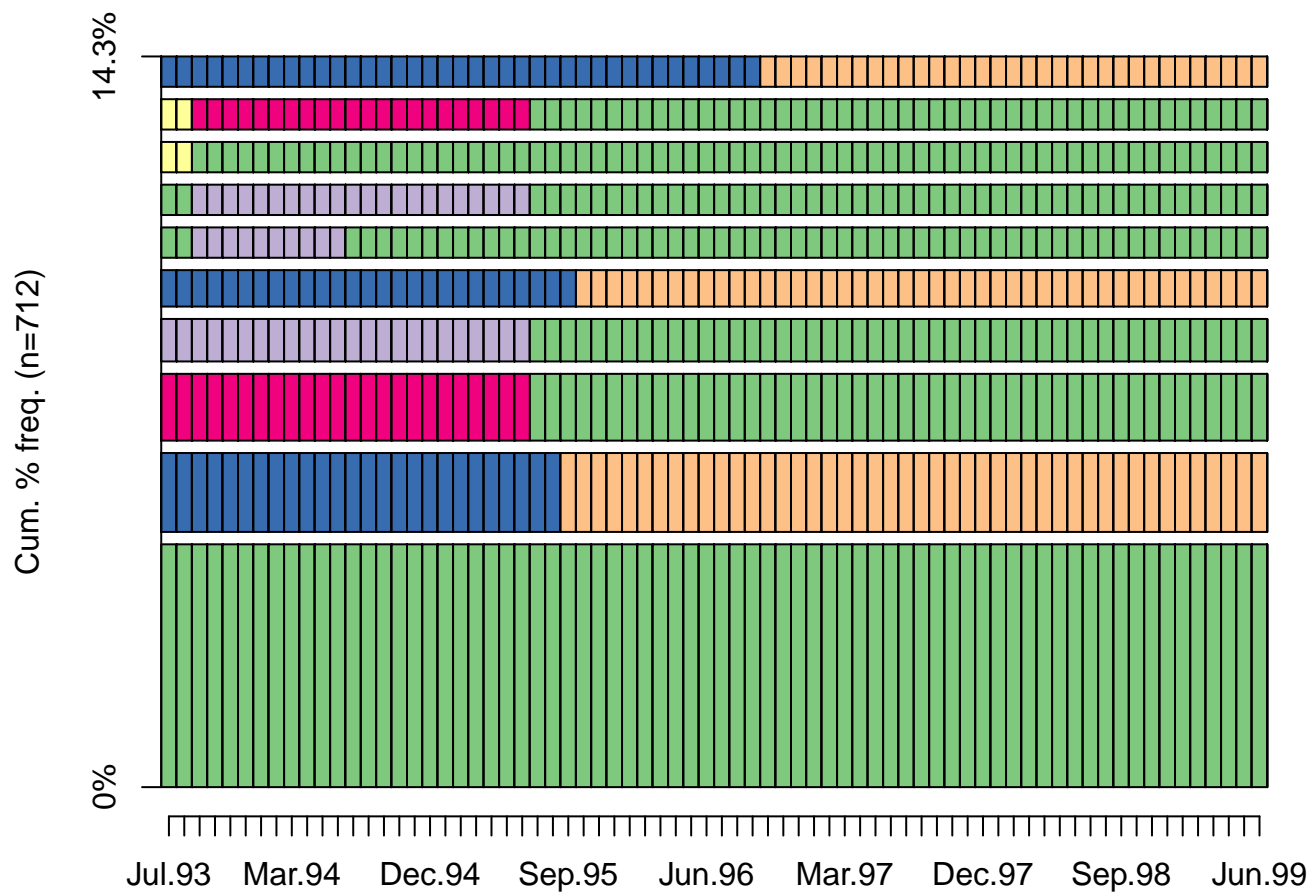
[illegible]


```
# mean time spent on a state
seqmeant(mvad.seq)
```

Plot jujukan kekerapan

20 jujukan yang paling kerap berlaku

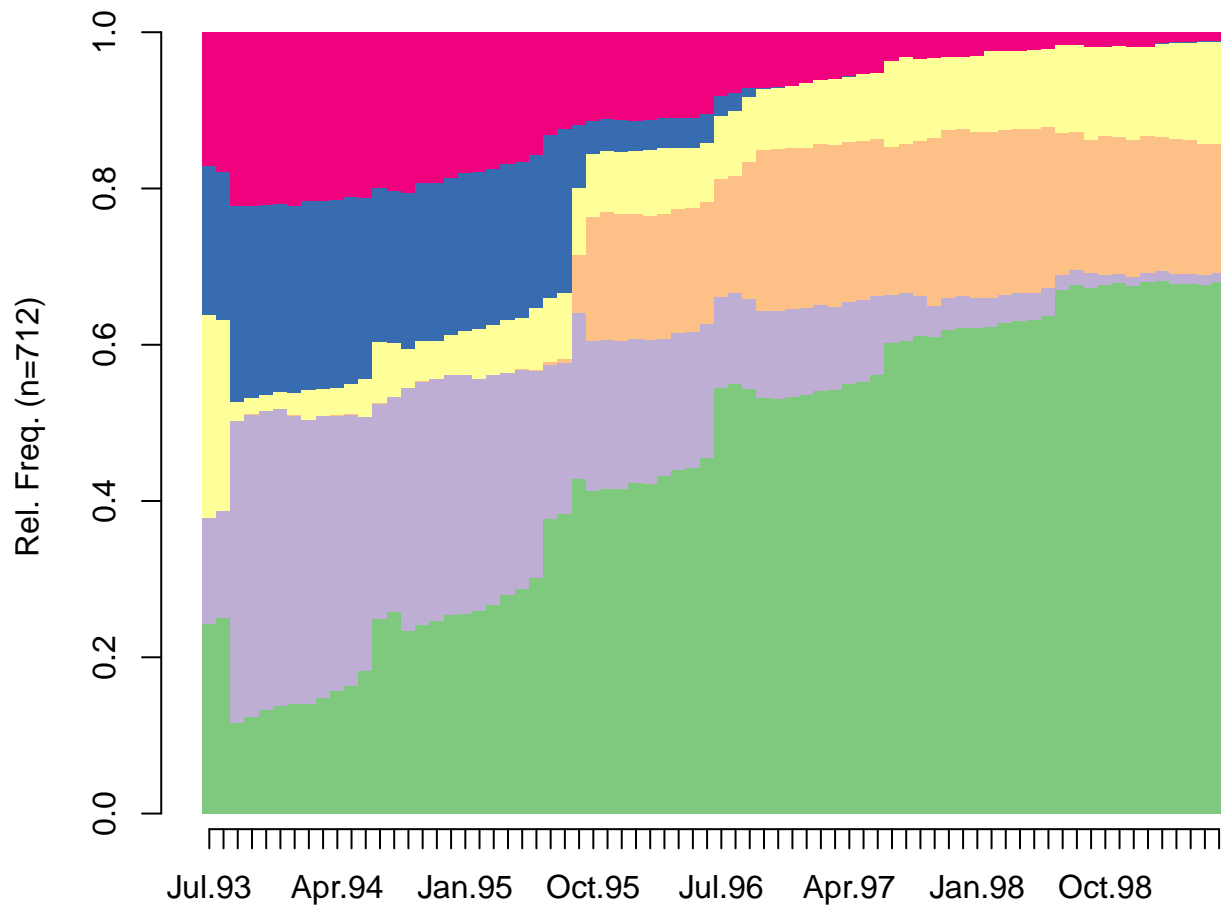
```
seqfplot(mvad.seq, idxs=1:10)
```



Plot taburan keadaan

```
# sequence density plot
seqdplot(mvad.seq, border=NA, main='Plot taburan keadaan')
```

Plot taburan keadaan

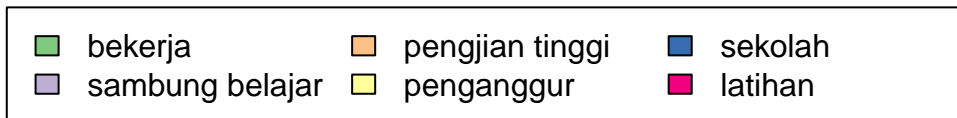
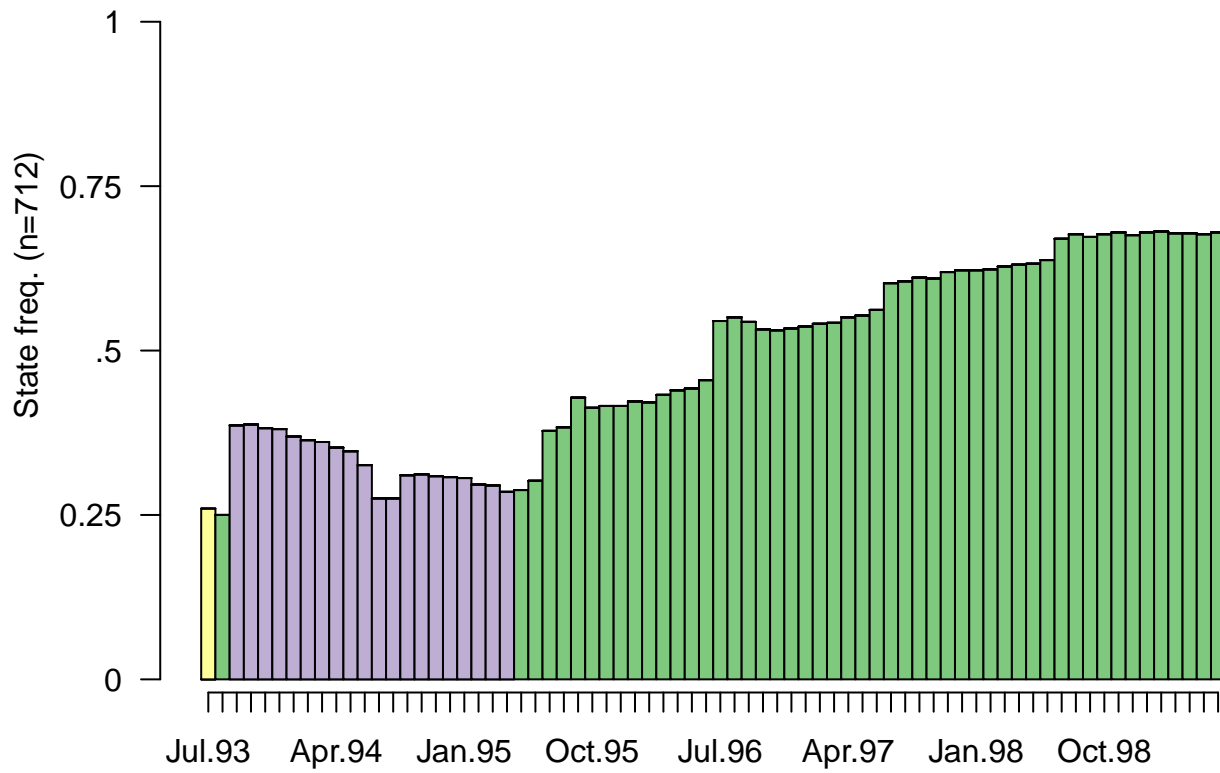


bekerja	pengajian tinggi	sekolah
sambung belajar	penganggur	latihan

Plot ini memaparkan keadaan dalam rentas masa

```
# mode of sequence based on certain time period
# sequence modal state plot
seqmsplot(mvad.seq)
```

Modal state sequence (0 occurrences, freq=0%)

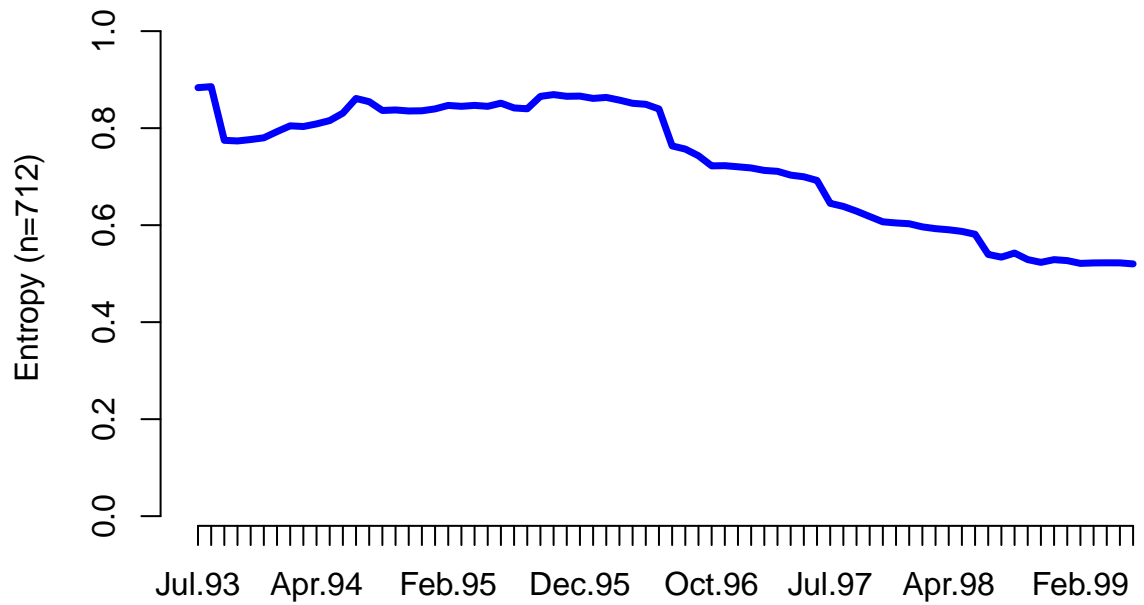


Entropi

$$h(p_1, \dots, p_a) = - \sum_{i=1}^a p_i \log(p_i)$$

```
# rate of change
seqHtplot(mvad.seq, main='entropi rentas lintang')
```

entropi rentas lintang



Transversal Entropy

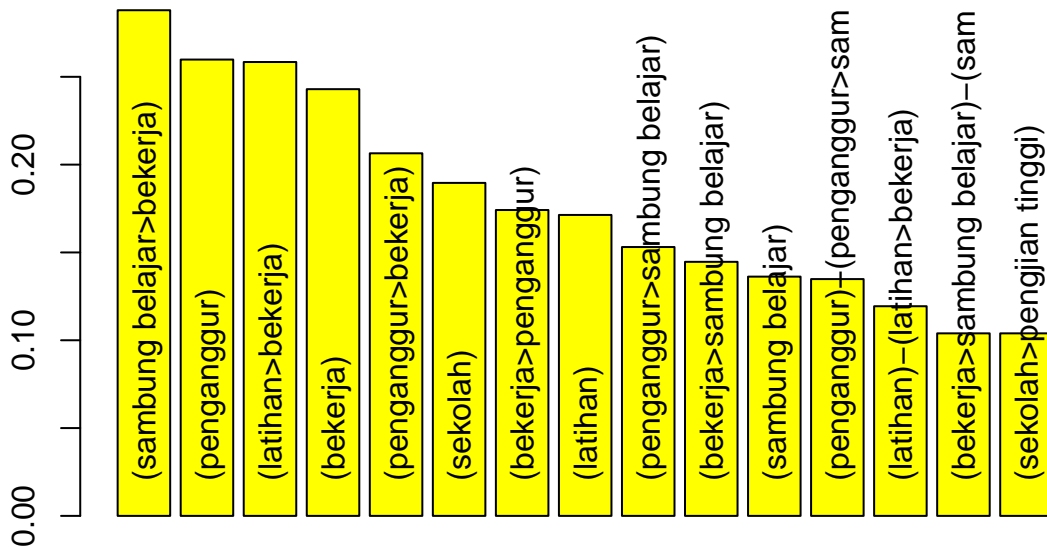
```
# mean time spent on a certain sequence  
mvad.seqe = seqecreate(mvad.seq)
```

sub jujukan

```
fsubseq = seqefsub(mvad.seqe, pmin.support=0.05)
```

15 sub jujukan paling kerap berlaku

```
plot(fsubseq[1:15], col='yellow')
```



Clustering

```
library(cluster)
# sequence substitution cost matrix
submat = seqsubm(mvad.seq, method='TRATE')
```

```
## [>] creating substitution-cost matrix using transition rates ...
```

```
## [>] computing transition probabilities for states EM/FE/HE/JL/SC/TR ...
```

```
# sequence distance computation
dist.om = seqdist(mvad.seq, method='OM', sm=submat)
```

```
## [>] 712 sequences with 6 distinct states
```

```
## [>] checking 'sm' (size and triangle inequality)
```

```
## [>] 557 distinct sequences
```

```
## [>] min/max sequence lengths: 72/72
```

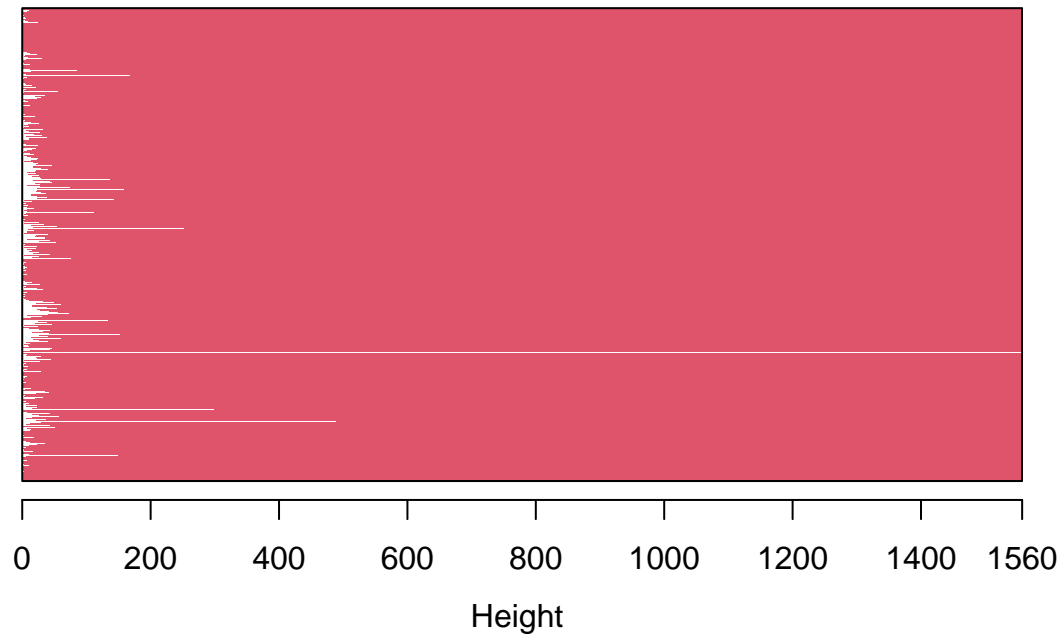
```
## [>] computing distances using the OM metric
```

```
## [>] elapsed time: 2.12 secs
```

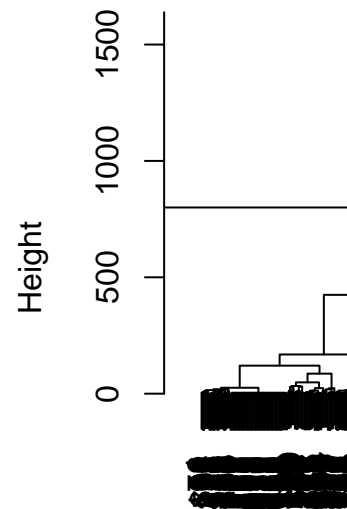
dapatkan kelompok dalam data

```
clusterward = agnes(dist.om, diss='T', method='ward')
plot(clusterward); abline(h=800)
```

Banner of `agnes(x = dist.om, diss = "T", method = "ward")`



Dendrogram of



Agglomerative Coefficient = 0.99

misalkan k=4 kelompok adalah signifikan

```
cl.4 = cutree(clusterward, k=4)
cl.4fac = factor(cl.4, labels=paste('Kumpulan', 1:4))
head(cl.4fac)
```

```
## [1] Kumpulan 1 Kumpulan 2 Kumpulan 3 Kumpulan 4 Kumpulan 2 Kumpulan 4
## Levels: Kumpulan 1 Kumpulan 2 Kumpulan 3 Kumpulan 4
```

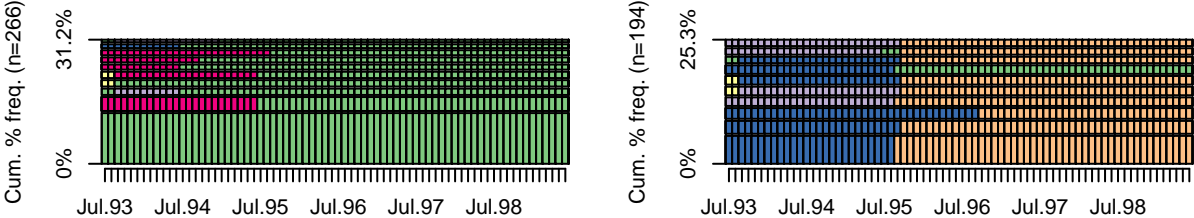
Jalankan analisis lanjutan terhadap setiap group.

Setiap individu dalam kumpulan yang sama akan mempunyai ciri yang hampir sama.

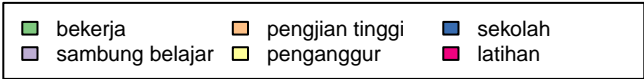
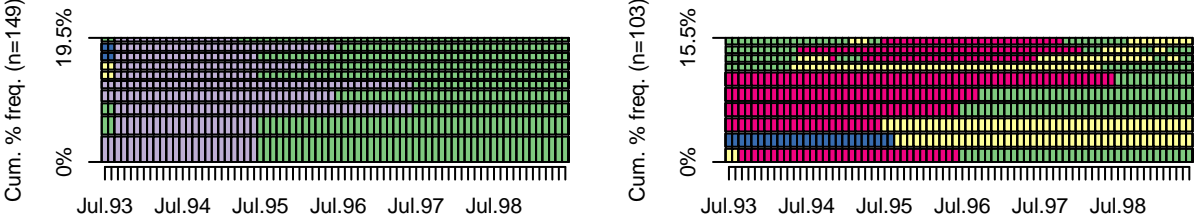
Plot kekerapan jujukan, 10 jujukan yang paling kerap berlaku.

```
seqfplot(mvad.seq, group=cl.4fac, main='10 Jujukan yang paling kerap berlaku',
         idxs=1:10)
```

10 Jujukan yang paling kerap berlaku – Kumpula 10 Jujukan yang paling kerap berlaku – Kumpula



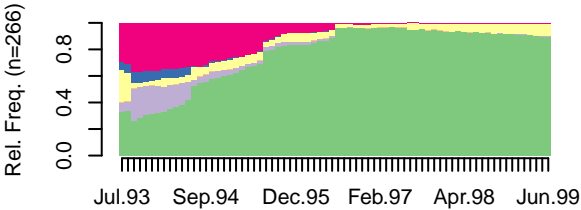
10 Jujukan yang paling kerap berlaku – Kumpula 10 Jujukan yang paling kerap berlaku – Kumpula



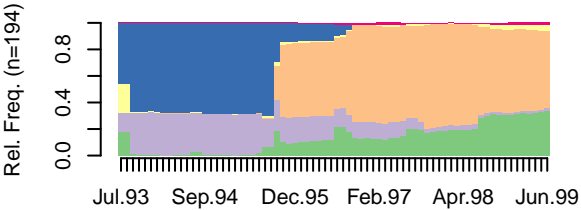
Plot taburan keadaan

```
seqdplot(mvad.seq, group=cl.4fac, border=NA, main='Plot taburan keadaan')
```

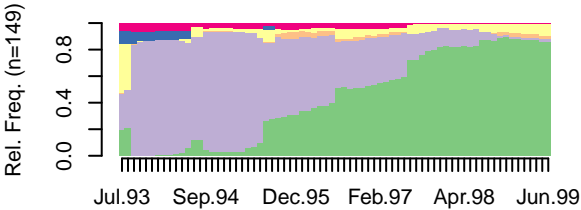
Plot taburan keadaan – Kumpulan 1



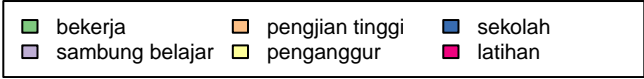
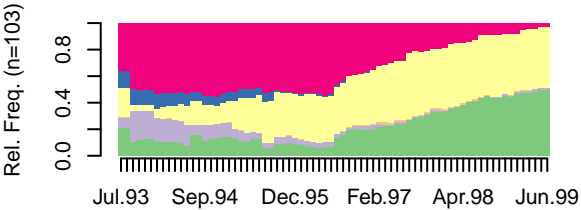
Plot taburan keadaan – Kumpulan 2



Plot taburan keadaan – Kumpulan 3

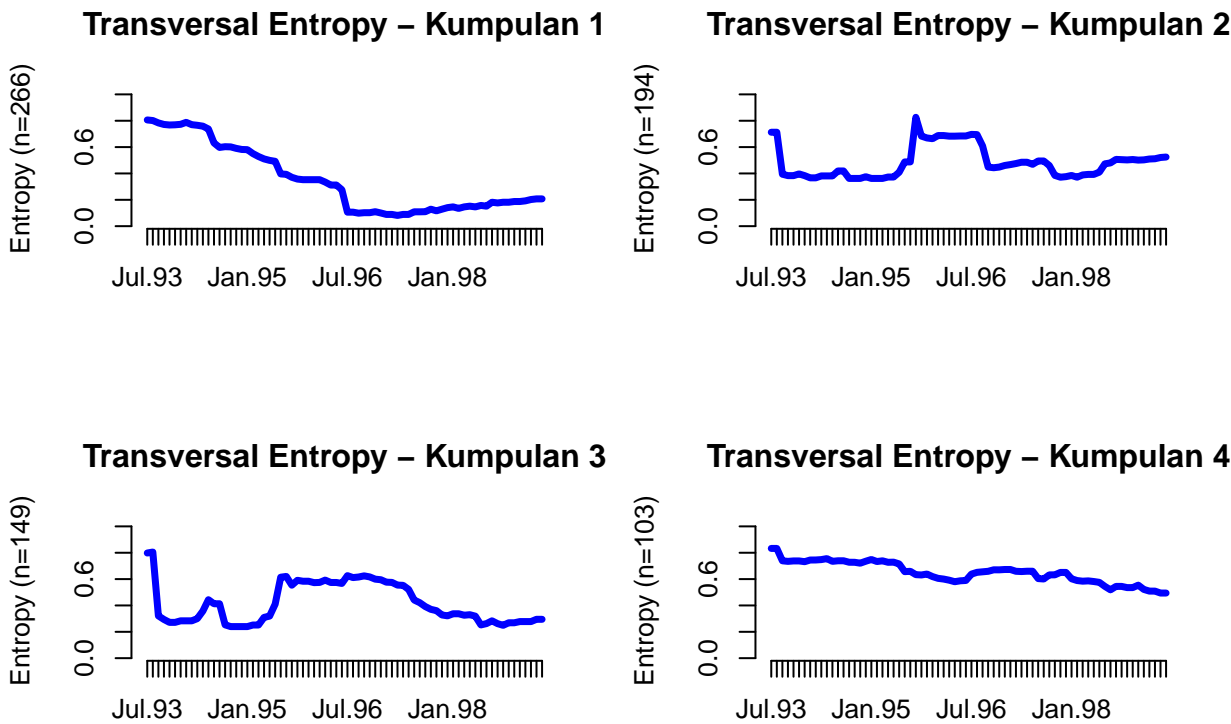


Plot taburan keadaan – Kumpulan 4



Plot rentas lintang


```
seqHtplot(mvad.seq, group=cl.4fac, main='Transversal Entropy')
```



Sub jujukan

```
disc = seqecmpgroup(fsubseq, group=cl.4fac)  
plot(disc[1:6])
```

