EDSD 2024-2025 Sequence Analysis Assignment

Gilbert Habaasa

2025-07-05

#Instructor: Nicola Barban
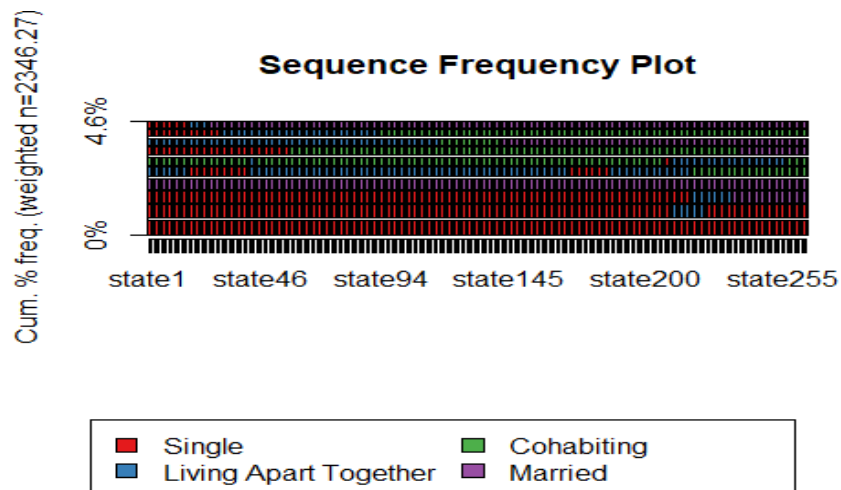
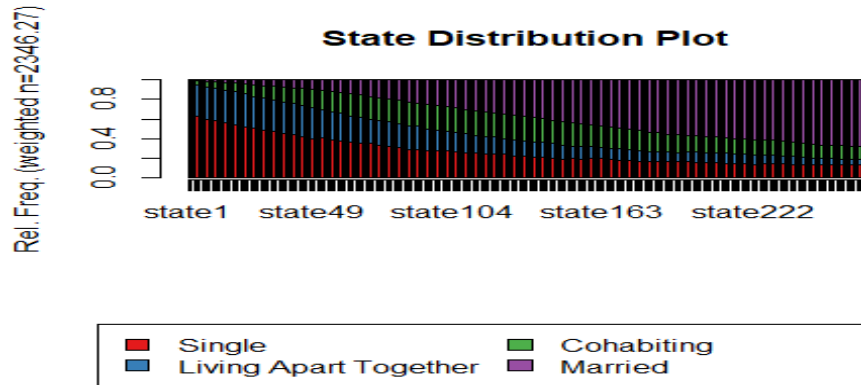##A. Import the data in R and describe the distribution of life states using relevant sequence

The German Family Panel (pairfam) data was imported from https://sa-book.github.io/. The detailed steps on importation from stata, recoding of life states, definition of state labels, codes,colours, and creation of sequence object are shown in the rmd file attached separately to the assignment. For question 2 and 3, detailed codes on mean duration in each state, plots,computation on life course dissimilarities in the data, as well as clustering methods to derive typical trajectories and describe groups are presented in rmd file indicated above.
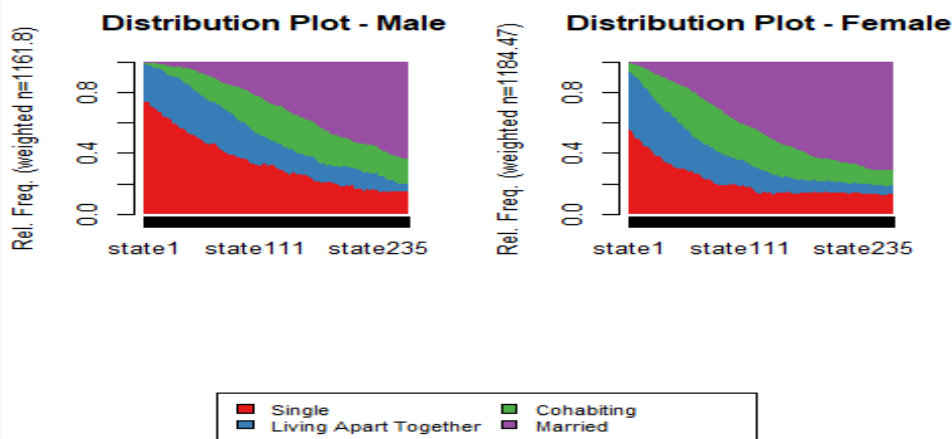
Step 1.1: Import Data from Stata
```
pairfam_raw <- read_dta("C:/Users/admin/OneDrive - London School of Hygiene and
Tropical Medicine/INED 2024/Sequence analysis/Assignment/PartnerBirthbio.dta")
```

 # Step 1.5: Sequence Summary Plots: Frequency and distribution plots

State Distribution Plot

Rel. Freq. (weighted n=2346.27)

state1   state49   state104   state163   state222

Legend:
- Single (red)
- Living Apart Together (blue)
- Cohabiting (green)
- Married (purple)

```
# Plot distribution plots grouped by gender
```

Distribution Plot - Male

Rel. Freq. (weighted n=1161.8)

state1   state111   state235

Distribution Plot - Female

Rel. Freq. (weighted n=1184.47)

state1   state111   state235

Legend:
- Single (red)
- Living Apart Together (blue)
- Cohabiting (green)
- Married (purple)

Step 1.6: Statistics and Distribution of Life States
First 10 Most Common Sequences

```
##                                                  Freq Percent
## S/264                                            17.0    0.72
## S/210-LAT/12-S/42                                14.6    0.62
## S/217-LAT/17-MAR/30                              13.2    0.56
## MAR/264                                          11.8    0.50
## LAT/17-S/22-LAT/130-S/16-LAT/33-COH/46 10.2      0.43
## COH/39-LAT/3-COH/165-S/1-LAT/46-COH/10  9.1      0.39
## S/57-COH/180-MAR/27                               9.0    0.39
## LAT/116-COH/24-MAR/124                            8.1    0.34
## S/28-LAT/63-COH/173                               7.7    0.33
## S/15-LAT/8-MAR/241                                7.3    0.31


# Computed mean durations by gender
## # A tibble: 2 × 5
##   sex          S   LAT   COH   MAR
##   <fct>    <dbl> <dbl> <dbl> <dbl>
## 1 Male      80.1  45.7  49.8  88.4
## 2 Female    48.3  42.1  53.1 120.
```
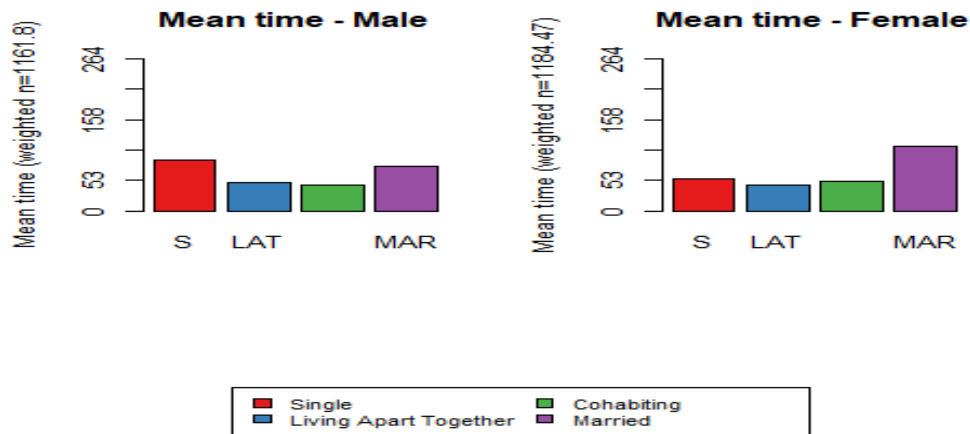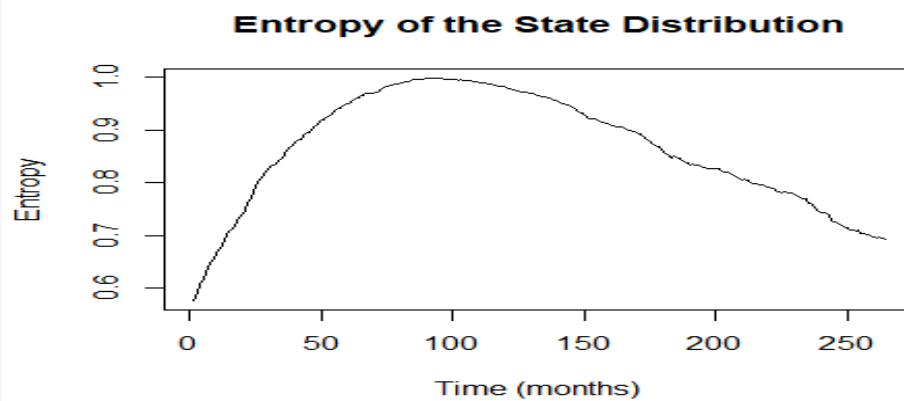
Mean time in each state by Gender
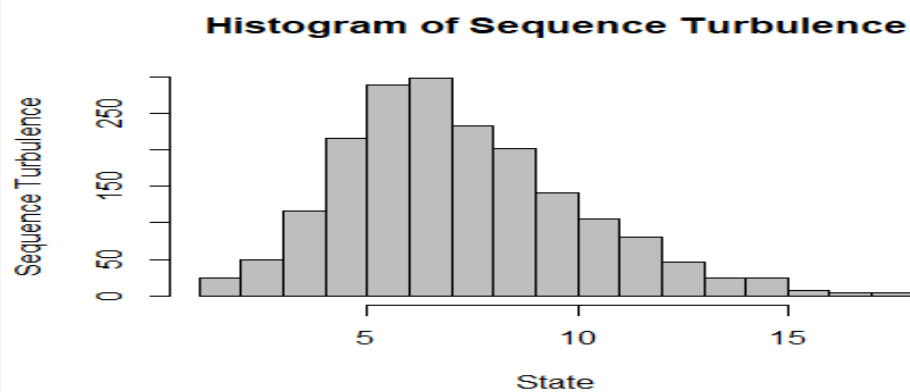

Transition Matrix
```
##            [-> S] [-> LAT] [-> COH] [-> MAR]
## [S   ->]    0.98     0.02     0.00     0.00
## [LAT ->]    0.02     0.96     0.02     0.00
## [COH ->]    0.00     0.00     0.98     0.01
## [MAR ->]    0.00     0.00     0.00     1.00
```
Entropy Over Time

Sequence Turbulence

The most common sequence is Singlehood state and is followed by sequences that start with singlehood and transition to LAT, cohabitation and marriage. On average, males spend 80.1 months in single state while females spend 48.3 months in single state. However, in marriage state, the males spend less time than females (88.4 months vs 120.0 months).On transitions,persons single have 98% chance of staying single. The 98% people stay in LAT state while 100% remain married.On theother hand, the 98% cohabiting are likely to keep in same state.

##2. Calculate life course dissimilarities in the data.

```
#Step 2.6: Comparison of Dissimilarity matrices using Pairwise Correlation
##              OM_constant  OM_trate OM_seqcost       LCS       DHD
## OM_constant   1.0000000 0.9999721  0.9999736 1.0000000 0.9089094
## OM_trate      0.9999721 1.0000000  0.9999999 0.9999721 0.9098161
## OM_seqcost    0.9999736 0.9999999  1.0000000 0.9999736 0.9097426
## LCS           1.0000000 0.9999721  0.9999736 1.0000000 0.9089094
## DHD           0.9089094 0.9098161  0.9097426 0.9089094 1.0000000
```
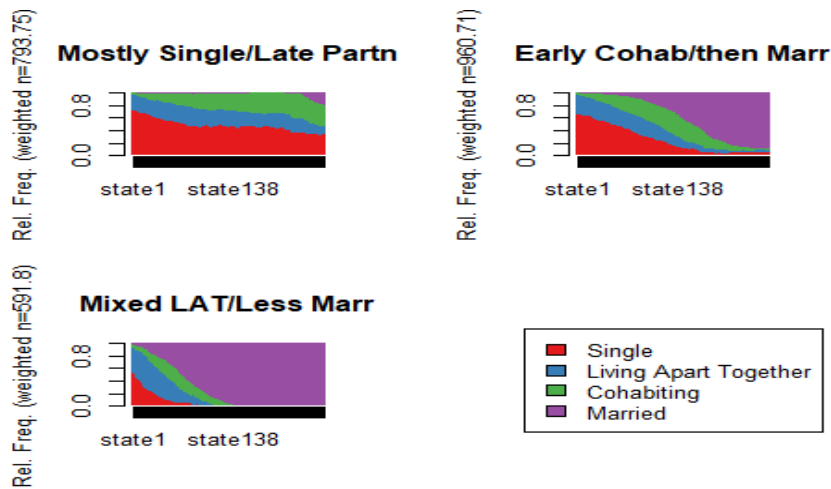
There is high correlation between OM,LCS and DHD. This implies similar results despite the method used.

##3. Use a clustering methods to derive typical trajectories and describe groups
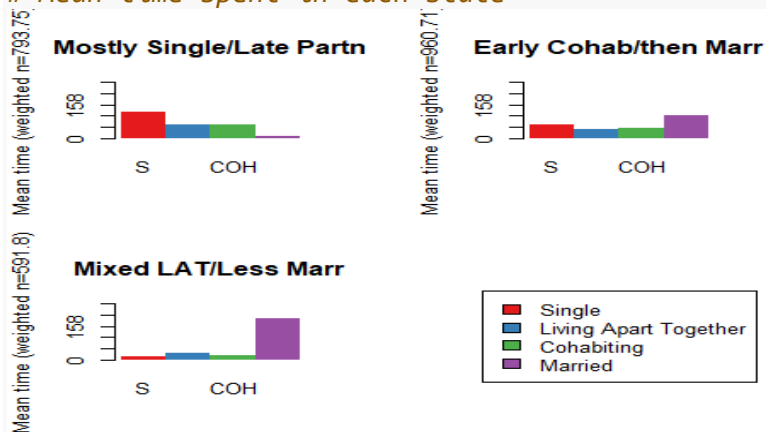
```
# Step 3.1: Clustering using Ward's method
```


Dendrogram of agnes

pairfam.dist3
Agglomerative Coefficient = 0.99
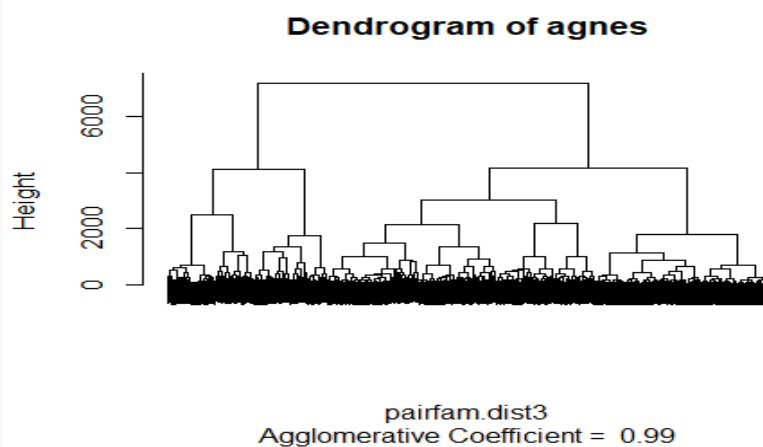
Based on 3 clusters, a diverse life course trajectory in partnership formation is observed: stable singlehood or delayed partnership; early cohabitation leading to marriage; mixed living apart together and less marriage. The first cluster is dominated by singlehood sequence. The second cluster is largely dominated by early cohabitation and marriage in the longrun. Both early cohabiting and mixed LAT report partners staying long in marriage in the long run.