2021-2022

M2 Expertise Sciences des Populations

ANALYSE DES DONNÉES LONGITUDINALES

Arno Muller

arno.muller@ined.fr

(Ined - Service Méthodes Statistiques)

Séance Précédente

- Format Long, Large, Semi-long
- Exercice sur R :
  - *Pivot_longer*
  - *Pivot_wider*

L’analyse de séquences

Concepts & Définition

- On analyse une «**trajectoire**» au cours du temps c'est à dire:
  - *Une liste **ordonnée** d'états dans le temps*
  - *Un **état** est défini comme une **modalité** d'une variable catégorielle (qualitative)*
  - *La valeur d'un état étant prise parmi un **ensemble fini de valeurs** possibles qu'on va par la suite appeler **alphabet***

- Une séquence de longueur k peut être vue comme une suite ordonnée de k éléments pris dans un alphabet préalablement déterminé.

Objectifs

- On cherche à comprendre la succession des évènements.
- Existe-t-il des «patterns» typiques d‘évènements ?
- Quels patterns dépendent d'un facteur donné?

Exemples de problématiques

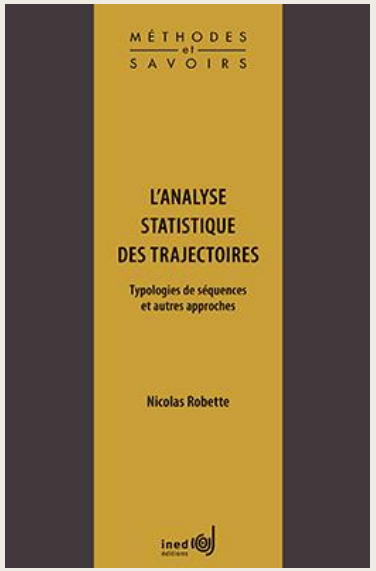
- Quelles sont les «trajectoires standards» ?
- Qui sont les personnes les plus à risque de suivre une trajectoire «chaotique» ?
- Quel(s) est/sont le(s) lien(s) entre les différents types de parcours et le profil social des individus ?

Exemples de séquences

- Statuts maritaux annuels de 30 à 50 ans
- Trajectoires horaires d'activités quotidiennes
  - *Lesnard Laurent, de Saint Pol Thibaut. Organisation du travail dans la semaine des individus et des couples actifs : le poids des déterminants économiques et sociaux. In: Economie et statistique, n°414, 2008. pp. 53-74.*
- Etude de l'insertion professionnelle après les études
- Trajectoires conjugales, passage à l'âge adulte
  - *Morand, E., et L. Toulemon. «Analyse des séquences et Optimal Matching : Le passage à l'âge adulte des Femmes et des Hommes en France». Paris: INSEE, 2009.*

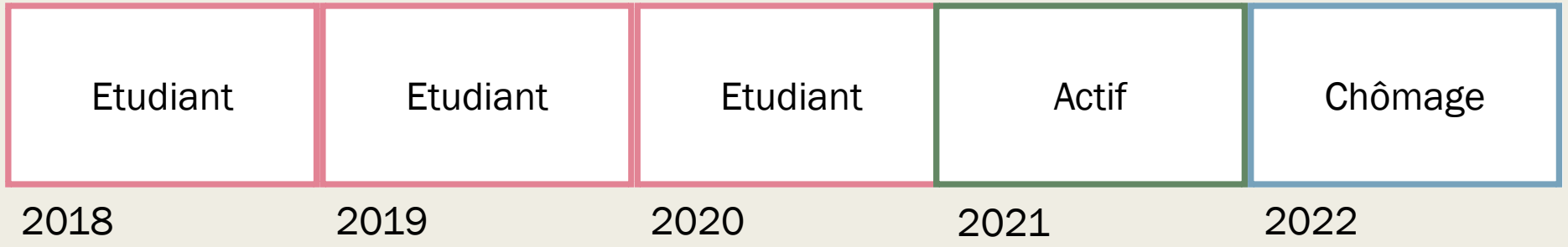
Bibliographie complémentaire

- Construire et analyser les trajectoires en démographie, coordonné par Philippe Cordazzo et Éva Lelièvre, INED Documents de travail, n°225, 2016, 81 pages
  - https://www.ined.fr/fichier/s_rubrique/25476/document_travail_2016_225_trajectoires_demographie.fr.pdf
- Le livre de Nicolas Robette



Premier exemple

- On étudie la mobilité professionnelle :
- Au cours de la vie on passe par une succession d'état vis-à-vis de l'emploi.
- Il est nécessaire de définir un alphabet:
  - *Par exemple {Actif, Chômage, Inactif, Retraité, Etudiant}*

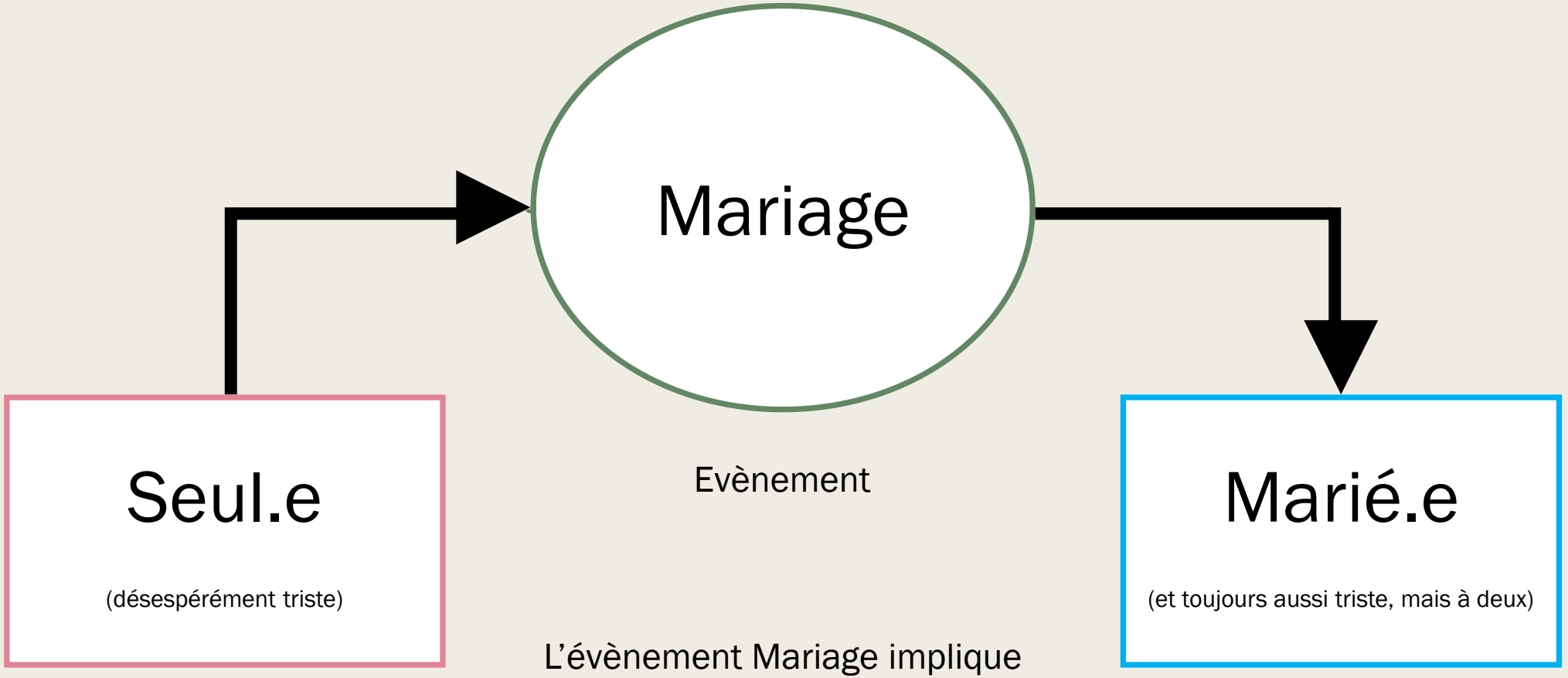


Evénement

- On désigne par événement le moment où un individu passe d'un état E1 à un état E2

- Exemple d'événement:
  - *Perdre son emploi : Passer de l'état «Actif» à l'état «Chômage»*
  - *Se marier : Passer de « Célibataire » à « Marié »*
  - *Venir en cours à 9h : Passer de l'état « Heureux » à l'état « Sérieusement-qu'est-ce-que-je-fais-là »*

Evénement, transitions, états



Evènement

Etat initial

L'évènement Mariage implique un changement d'état : une transition de l'état Seul.e à Marié.e.

Etat final

Ce qu'il faut pour faire une analyse de séquences

- Un objet d'étude
  - *exemple: trajectoire maritale, de travail, …*
- Une période claire d'étude: début , fin et horloge utilisée
  - *exemple: entre 20 et 40 ans, chaque année*
- Un ensemble de modalités pour l'objet étudié (Alphabet)
  - *Des modalités très précises donc très nombreuses vs*
  - *Un nombre de modalités limité (moins de précision mais moins de dilution de l'information)*
- Une séquence d'états ou d'événements

Représentation d'une séquence individuelle

- On peut représenter une séquence par:
  - *Étude -Emploi -Emploi -Sans Emploi –Emploi -Étude –Emploi*
    - L'alphabet est ici l'ensemble des modalités d'états possibles
      - *{Emploi, Etude, Sans Emploi}*

- Autre mode de représentation :
  - *(Etude ,1) -(Emploi,2) -(Sans Emploi,1) -(Emploi,1) -(Étude ,1) -(Emploi,1)*

Le package TraMineR

- Site dédié : http://traminer.unige.ch/

- Ressource :

Gabadinho Alexis, Ritschard Gilbert, Studer Matthias, Müller Nicolas. 2011. Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software, 40(4), p.1-37*

Utilisation du logiciel

- Définition des données de séquence
- Eventuellement, changement de format des données
- Fonctions de visualisation graphique
- Construction de la matrice de coûts
- Etape de CAH, construction des classes
- Caractérisation des classes

Etape préliminaire

- Définition d'un objet de type séquence
  - *seqdef(df, states=, labels=, xtstep)*

    - df: Base de Donnée
    - states : Labels courts des éléments de l'alphabet
    - labels : Labels longs des éléments de l'alphabet
    - xtstep: Pas pour affichage des labels sur les graphes

Jul.93	Aug.93	Sep.93	Oct.93	Nov.93	Dec.93	Jan.94	Feb.94	Mar.94	Apr.94
training	training	employment	employment	employment	employment	training	training	employment	employment
joblessness	joblessness	FE	FE	FE	FE	FE	FE	FE	FE
joblessness	joblessness	training	training	training	training	training	training	training	training
training	training	training	training	training	training	training	training	training	training
joblessness	joblessness	FE	FE	FE	FE	FE	FE	FE	FE
joblessness	joblessness	joblessness	training	training	training	training	training	training	training
joblessness	joblessness	FE	FE	FE	FE	FE	FE	FE	FE
employment	employment	FE	FE	FE	FE	FE	FE	FE	FE
joblessness	joblessness	training	training	training	training	training	training	training	training
employment	employment	school	school	school	school	school	school	school	school
joblessness	employment	FE	FE	FE	FE	FE	FE	FE	FE

```
> mvad.seq=seqdef(mvad, 17:86,
+                 alphabet = mvad.alphabet,
+                 states = mvad.scodes,
+                 labels = mvad.labels,
+                 xtstep = 6)
 [>] state coding:
       [alphabet]  [label]  [long label]
     1  employment  EM       employment
     2  FE          FE       further education
     3  HE          HE       higher education
     4  joblessness JL       joblessness
     5  school      SC       school
     6  training    TR       training
 [>] 712 sequences in the data set
 [>] min/max sequence length: 70/70
```

Indicateurs individuels

Traitements exploratoires

- Nombre de transitions par individus
- Nombre d'états par individu
- Temps moyen passé dans chaque état

Nombre de transitions

- Pour chaque séquences on regarde le nombre de transition, de passages entre deux états distincts.

- Exemple :

  A-A-A-A-A-A-B-B-B-A-A-C-C-C

  - *Seq = ABAC, NT = 3*

- Syntaxe R :

  - *seqtransn()*

Durée passée dans chaque état

- La durée passée dans chaque états comptabilise le nombre d'unités de temps passés dans chacun des états
  - *Exemple : A-A-A-A-A-A-B-B-B-A-A-C-C-C*
  - $D_A=8$; $D_B=3$; $D_C=3$

- Syntaxe :
  - *seqistatd()*

	EM	FE	HE	JL	SC	TR
1	68	0	0	0	0	2
2	0	36	34	0	0	0
3	10	34	0	2	0	24
4	14	0	0	9	0	47
5	0	25	45	0	0	0
6	36	0	0	1	0	33
7	40	30	0	0	0	0
8	48	22	0	0	0	0
9	49	0	0	0	0	21
10	46	0	0	14	10	0

Indicateurs globaux

Séquences les plus fréquentes

- Quelles sont, au sein de mon ensemble d'individus, la séquence qui est la plus fréquente ?

- Attention. Ne pas confondre:
  - *Séquence modale :*
    - Séquence la plus fréquente
    - seqmodst()
  - *Séquence des états modaux :*
    - Séquence constituée de l'état modal à chaque unité de temps.
    - seqtab(sequence, tlim, format)

Durée moyenne passée dans chaque état

- Calcul du temps passé par l' ensemble des individus dans chaque état défini dans l'alphabet

```
> apply(seqistatd(mvad.seq),2,mean)
 [>] computing state distribution for 712 sequences ...
       EM        FE        HE        JL        SC        TR
31.721910 11.426966  8.398876  5.674157  5.723315  7.054775
```

Taux de transition entre états

- Probabilité de passer d'un état à un autre état (Total de 100% pour chacun des états)
  - *Représentation matricielle*
  - *Calculé sur l'ensemble des couples d'états*
- Syntaxe R:
  - *seqtrate()*

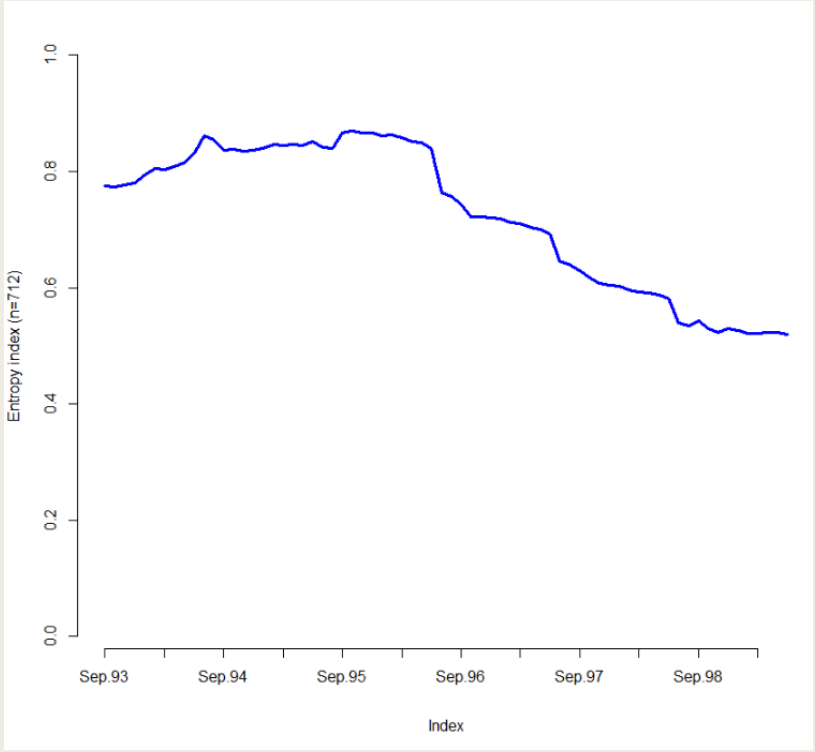
```
> transition <- seqtrate(mvad.seq)
 [>] computing transition probabilities for states EM/FE/HE/JL/SC/TR ...
> round(transition,2)
        [-> EM] [-> FE] [-> HE] [-> JL] [-> SC] [-> TR]
[EM ->]    0.99    0.00    0.00    0.01    0.00    0.00
[FE ->]    0.03    0.95    0.01    0.01    0.00    0.00
[HE ->]    0.01    0.00    0.99    0.00    0.00    0.00
[JL ->]    0.04    0.01    0.00    0.94    0.00    0.01
[SC ->]    0.01    0.01    0.02    0.01    0.95    0.00
[TR ->]    0.04    0.00    0.00    0.01    0.00    0.94
```

Entropie transversale

- Variabilité des états à chaque temps

$$h = -\sum_{i=1}^{s} p_i \log(p_i)$$

- Avec s nombre d'états et pi part de la modalité i pour l'ensemble des individus à un instant t
- H vaut 0 quand tous les individus sont dans le même état
- H est maximale quand toutes les modalités sont identiquement représentées (et correspond donc à des pi égaux).

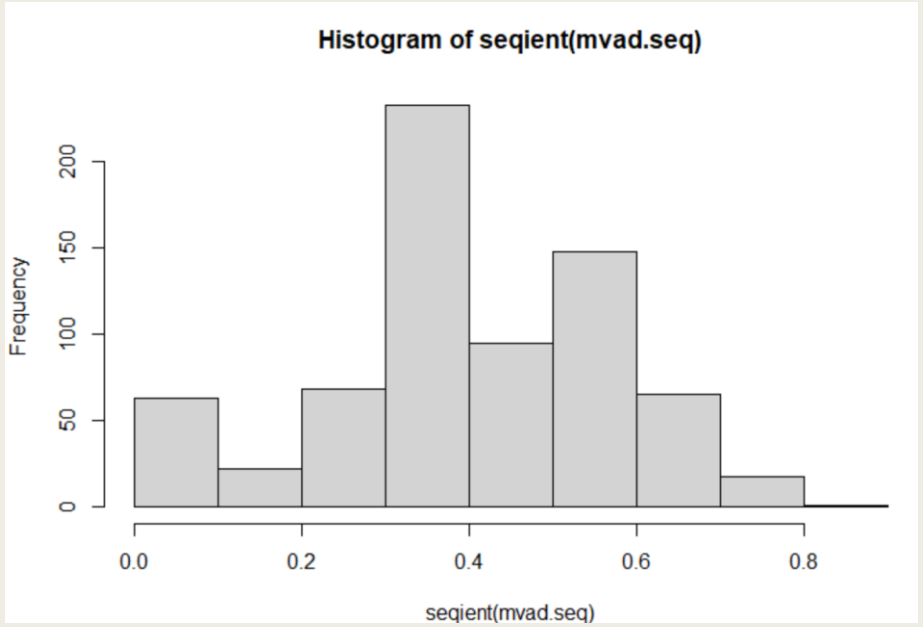


Entropie longitudinale

- Variabilité des états pour chaque individu

$$h = -\sum_{i=1}^{s} p_i \log(p_i)$$

- Avec s nombre d'états et pi part de la modalité i pour un individu sur toute la période d'analyse
- H vaut 0 quand l'individu reste dans le même état
- H est maximale quand toutes les modalités sont identiquement représentées pour un individu (et correspond donc à des pi égaux).
- Syntaxe R : seqient()



Visualisation graphiques

Deux principaux :

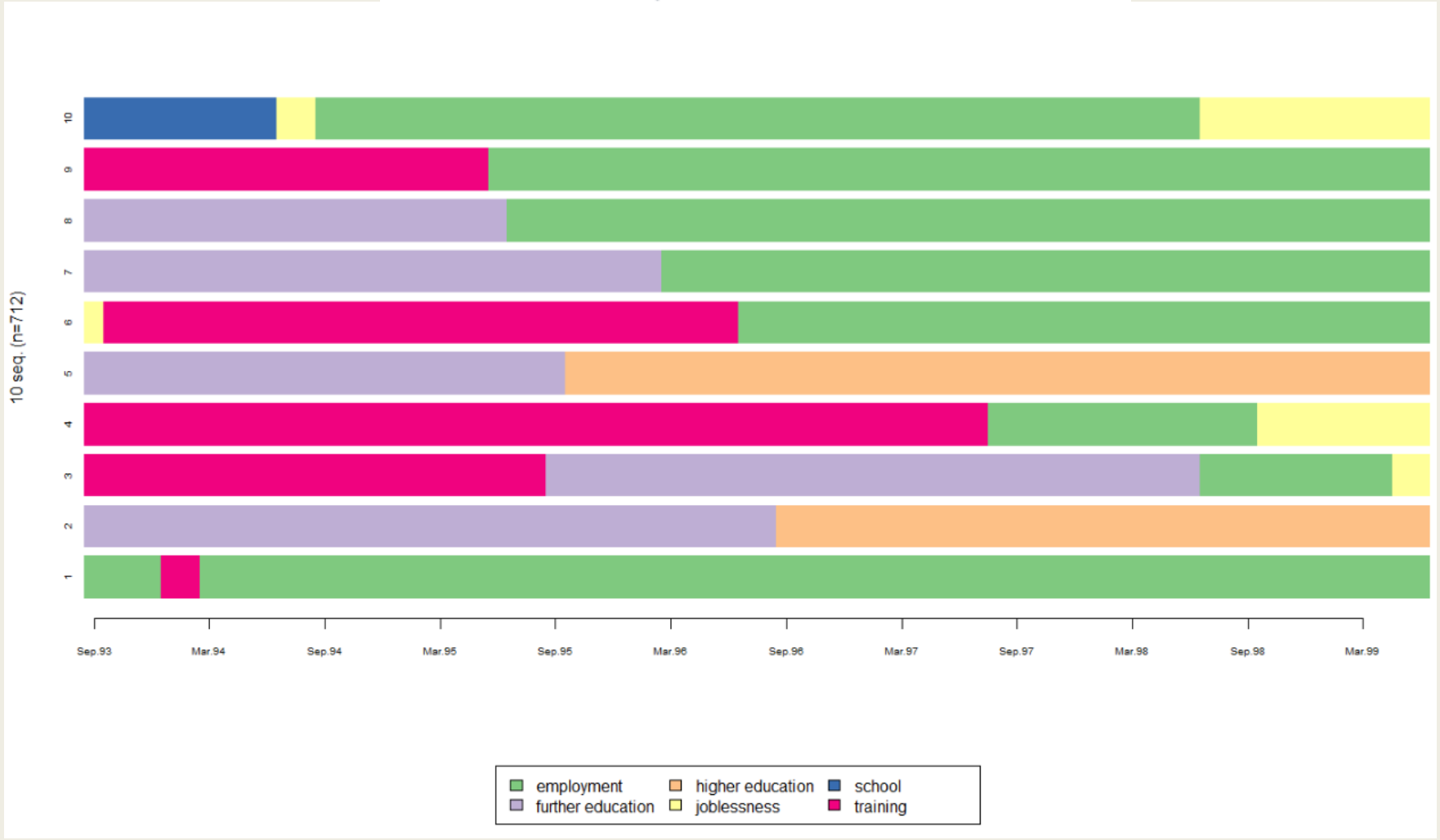
- Tapis :
  - *Visualisation des trajectoires individuelles*
- Chronogramme :
  - *Distribution des états à chaque âge*

Le tapis

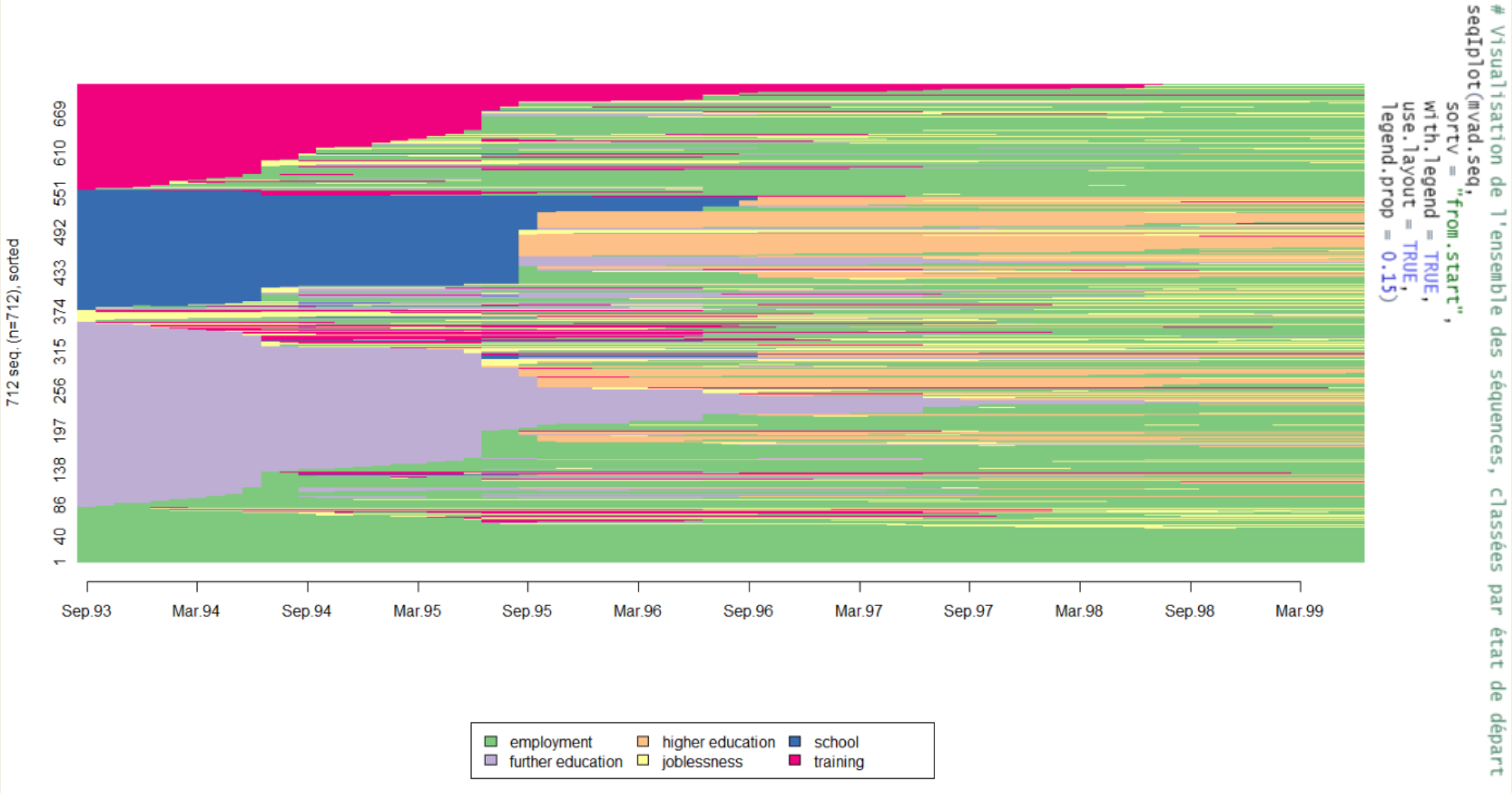
- Visualisation des trajectoires individuelles. Colorisation suivant l'état à un âge j.
- Remarques :
  - *Rapidement illisible*
  - *Possibilité de regroupement par valeur initiale (ou finale)*
  - *Possibilité de visualisation sur des sous-populations*

Le tapis

```
# Visualisation de la trajectoire des 10 premiers individus
seqiplot(mvad.seq,
         border = NA,
         use.layout = TRUE,
         with.legend = TRUE,
         legend.prop = 0.15,
         cex.axis = 0.6)
```



Le tapis, toutes les séquences



Le chronogramme

- Distribution des états à chaque âge pris en compte dans la distribution
- Attention à l'interprétation : la proportion de chaque état à chaque temps