L'analyse de séquences dans R avec la librairie TraMineR

Alexis Gabadinho, Gilbert Ritschard
Nicolas S. Müller, Matthias Studer

LIVES et Institut d'études démographiques et du parcours de vie
Université de Genève
http://mephisto.unige.ch/traminer

Rencontre de statistique appliquée
Palette d'applications sous R, 28 avril 2011, Paris

Outline

TraMineR

- TraMineR : Trajectory Miner in R

  (Accessoirement inspiré par notre goût pour le Gewürztraminer)
- Librement disponible sur le CRAN (Comprehensive R Archive Network)

  `http://cran.r-project.org/web/packages/TraMineR/`
- Installation : `install.packages("TraMineR")`
- TraMineR peut être simplement et directement combiné avec les autres librairies de R
  - Par exemple, les dissimilarités entre séquences obtenues avec TraMineR peuvent être utilisées avec les procédures déjà optimisées de clustering, de MDS, les outils de régression linéaire et non-linéaire, ...

Le projet de recherche

TraMineR est le fruit d'un projet FNS

- Mining event histories : Towards new insights on personal Swiss life courses
- Project FN-113998 et FN-122230 de février 2007 à janvier 2011

- Gilbert Ritschard, prof. de statistique, requérant principal,
- Eric Widmer, prof. de sociologie, co-requérant
- Alexis Gabadinho, démographie
- Nicolas S. Müller, sociologie, systèmes d'information
- Matthias Studer, économie, sociologie

- Aujourd'hui, le développement se poursuit dans le cadre de l'IP14 de LIVES

Séquences en sciences sociales

- TraMineR a été conçu pour répondre à des questions de sciences sociales
- Les séquences (suite d'états ou d'événements) décrivent des trajectoires de vie
- Types de questions :
  - Les parcours de vie obéissent-ils à une norme sociale ?
    - Quelles sont les types de trajectoires standards ?
    - Quels écarts observe-t-on par rapport à ces normes ?
  - Pourquoi certaines personnes suivent-elles des trajectoires plus chaotiques que d'autres ?
  - Comment les trajectoires de vie sont-elles liées au sexe, à l'origine sociale et à d'autres facteurs

Logiciels pour l'analyse de séquences

Etat des lieux au début du projet

- Optimize le logiciel d'Abbott (Abbott, 1997)
  - Calcul des distances d'optimal matching
  - Plus maintenu
- TDA (Rohwer and Pötter, 2002)
  - logiciel statistique gratuit, calcul des distances d'optimal matching
- SQ-ados, macros Stata, (Brzinsky-Fay et al., 2006)
  - gratuit si on a une licence Stata
  - distances optimal matching, visualisation
- CHESA logiciel gratuit de Elzinga (2007)
  - Nombreuses métriques, dont plusieurs non fondées sur l'alignement
  - Turbulence
- MARCH (Berchtold and Berchtold, 2004)
  - Modèles des transitions, chaînes de Markov cachées, ...

Objectifs initiaux

- Développer une plate-forme pour chercheurs en sciences sociales
  - rassemblant les fonctionnalités offertes séparément par les programmes existants ;
  - proposant de nouvelles méthodes pour l'analyse de trajectoires de vie.
- Développer une plate-forme accessible au plus grand nombre
  - gratuite ;
  - documentée avec un manuel de l'utilisateur détaillé.

Ce que TraMineR permet de faire

- Prise en charge et conversion de différents types de données longitudinales
- Gestion des poids et des données manquantes
- Visualisation d'un ensemble de séquences (index plot, séquences fréquentes, distributions transversales, et plus...)
- Caractéristiques longitudinales de séquences individuelles (complexité, durées de séjour dans chaque état, entropie longitudinale, turbulence, et plus ...)
- Séquence de caractéristiques transversales (distribution des états, entropie transversale, état modal)
- Autres caractéristiques agrégées (taux de transition, durées moyennes de séjour dans chaque état)
- Dissimilarités entre paires de séquences (Optimal matching, Longest Common Subsequence, Hamming, Dynamic Hamming, Multichannel et plus ...)
- Mesure de dispersion d'un ensemble de séquences
- Séquences représentatives
- ANOVA et arbres de régression à partir de matrices de dissimilarités
- Extraction de séquences d'événements fréquents
- Identification de séquences d'événements discriminantes

Le jeu de données mvad

- Etude de McVicar and Anyadike-Danes (2002) sur la transition entre formation et emploi en Irlande du Nord.
  - Jeu de données distribué avec la librairie TraMineR.
  - Provient d'une enquête auprès 712 jeunes irlandais.
  - Les séquences représentent leur suivi pendant les 6 années suivant la fin de la scolarité obligatoire (16 ans) et sont constituées des 70 variables indiquant les états mensuels successifs de chaque individu entre septembre 1993 et juin 1999.
  - Les états sont :

EM	en emploi
FE	formation secondaire
HE	formation supérieure
JL	au chômage
SC	école
TR	en stage ou apprentissage.

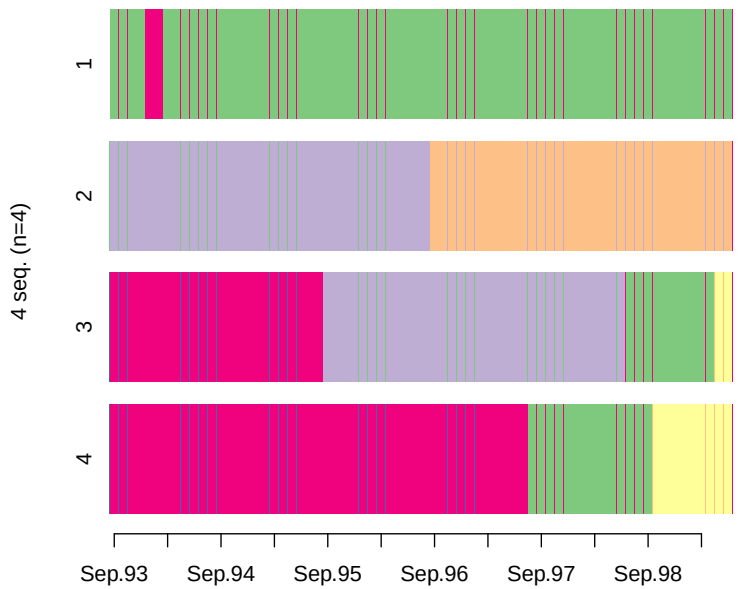
Séquences d'états - Jeu de données mvad

- Premières séquences du jeu de données (20 premiers mois)

```
  Sequence
1 EM-EM-EM-EM-TR-TR-EM-EM-EM-EM-EM-EM-EM-EM-EM-EM-EM-EM-EM-EM
2 FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE
3 TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR
4 TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR
```

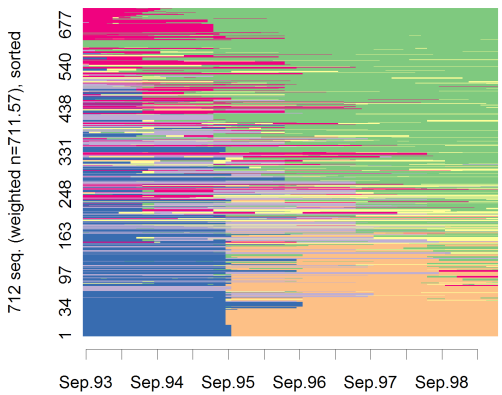
- Représentation compacte (format SPS)

```
  Sequence
[1] (EM,4)-(TR,2)-(EM,64)
[2] (FE,36)-(HE,34)
[3] (TR,24)-(FE,34)-(EM,10)-(JL,2)
[4] (TR,47)-(EM,14)-(JL,9)
```

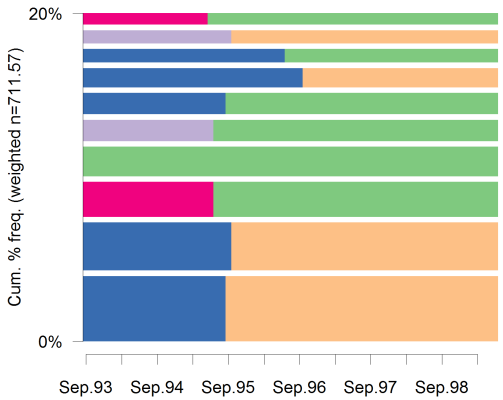


Présentations graphiques : Exemples

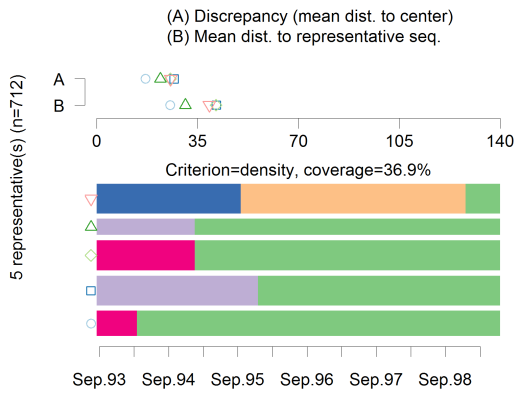
I-plot, tapis des sequences individuelles



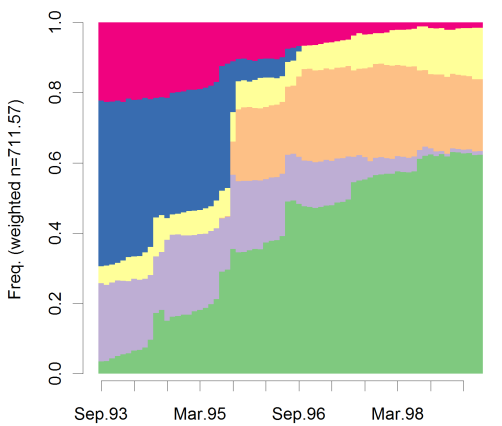
f-plot, sequences les plus frequentes



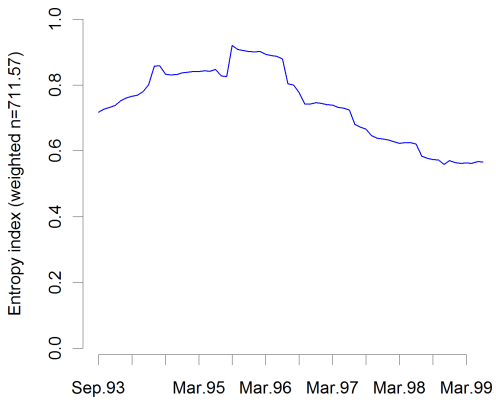
r-plot, sequences representatives



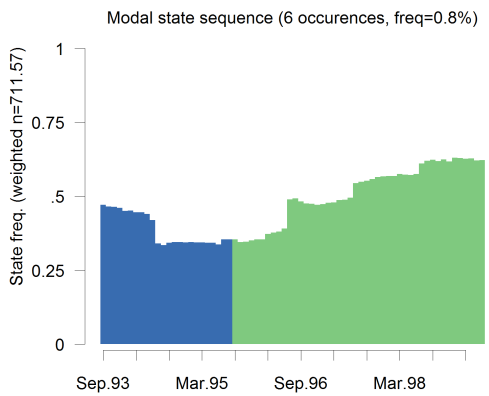
d-plot, distributions transversales des etats



Ht-plot, entropies transversales



ms-plot, sequence des etats modaux



Aperçu des possibilités de TraMineR

- Charger TraMineR et créer un objet 'séquences d'états'

```
R> library(TraMineR)
R> data(mvad)
R> mvad.seq <- seqdef(mvad, 17:86, xtstep = 6)
```

- Calcul des dissimilarités OM entre paires de séquences avec un coût d'indel de 1 et des coûts de substitutions déduits des taux de transitions

```
R> mvad.om <- seqdist(mvad.seq, method = "OM", indel = 1, sm = "TRATE")
```

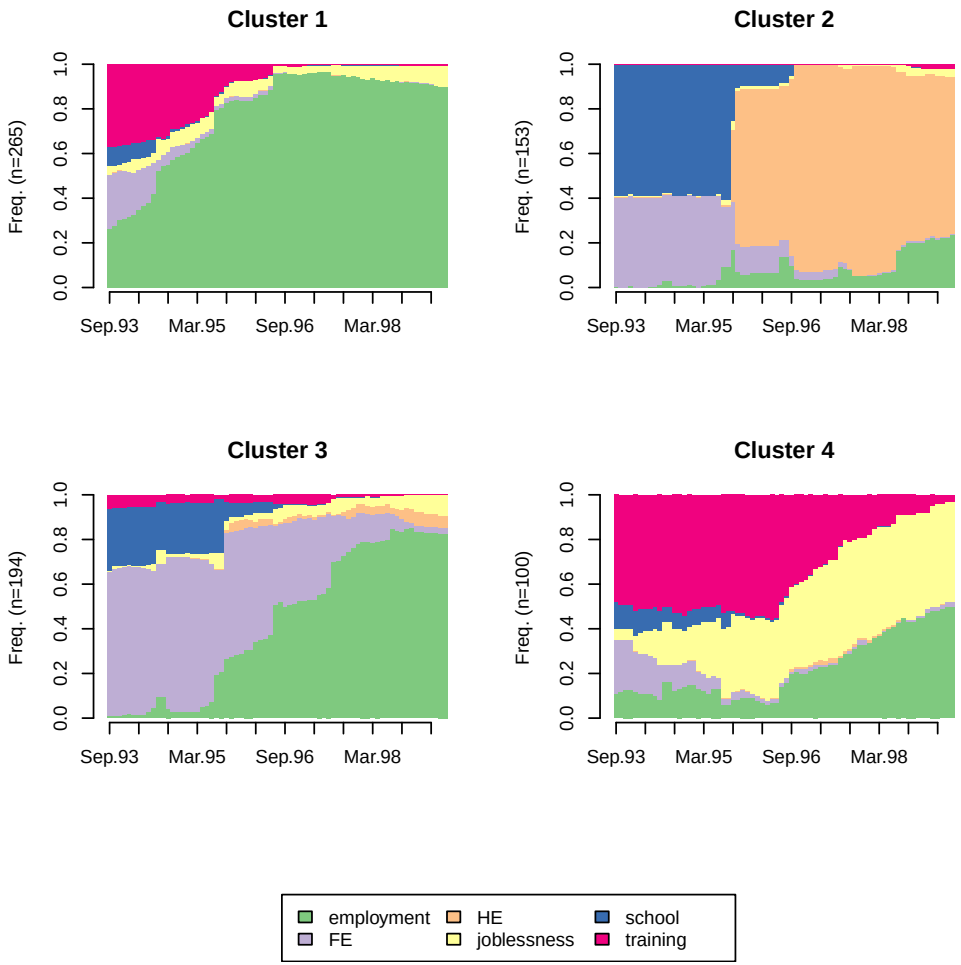
- Classification en 4 groupes par une procédure agglomérative avec critère de Ward

```
R> library(cluster)
R> clusterward <- agnes(mvad.om, diss = TRUE, method = "ward")
R> mvad.cl4 <- cutree(clusterward, k = 4)
R> cl4.lab <- factor(mvad.cl4, labels = paste("Cluster", 1:4))
```

Aperçu des possibilités de TraMineR (suite 1)

- Visualisation des classes : distributions transversales des états

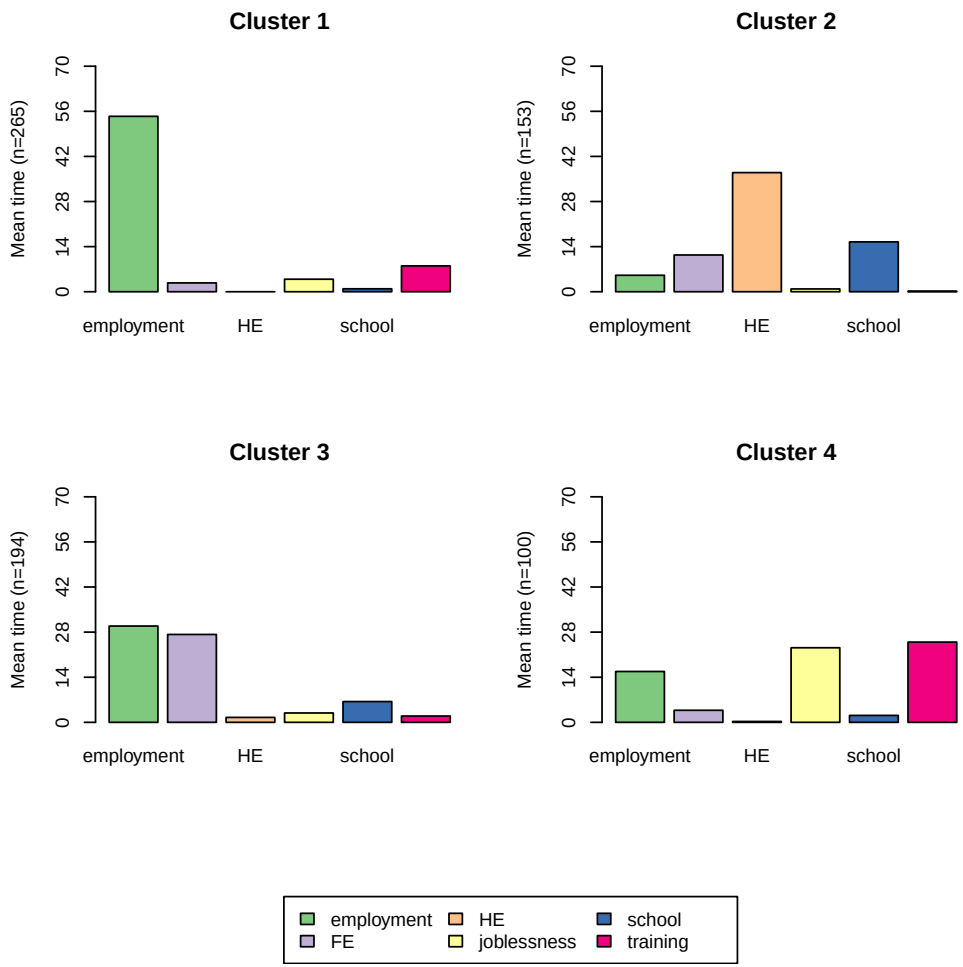
```
R> seqdplot(mvad.seq, group = cl4.lab, border = NA)
```



Aperçu des possibilités de TraMineR (suite 2)

- Temps moyen dans les états par classe

```
R> seqmtplot(mvad.seq, group = cl4.lab)
```



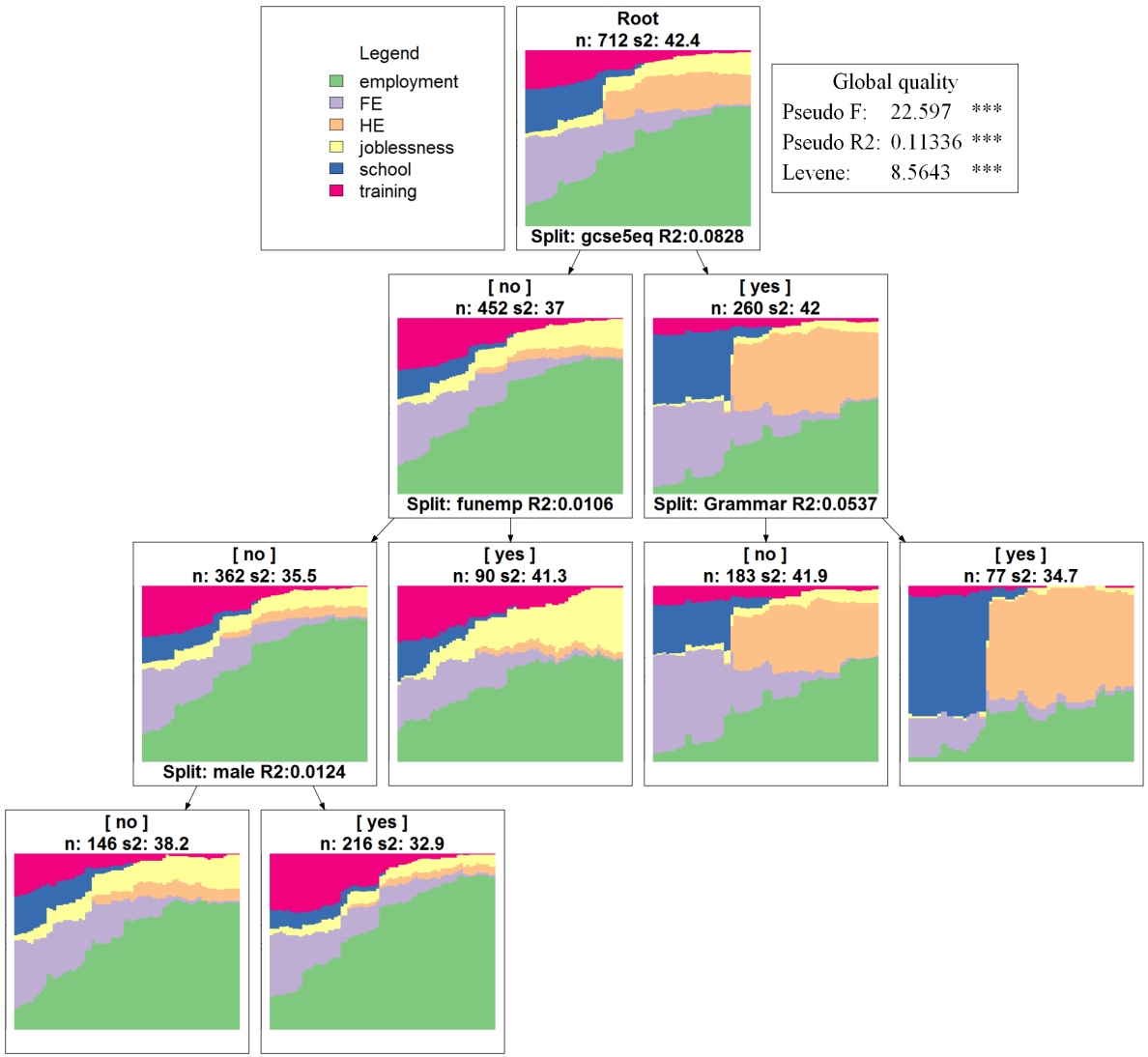
Arbre de régression

(Studer et al., 2011, 2009)

```
R> dt <- seqtree(mvad.seq ~ male + Grammar + funemp +
+       gcse5eq + fmpr + livboth, weighted = FALSE, data = mvad,
+       diss = mvad.om, R = 5000)
R> seqtreedisplay(dt, filename = "fg_mvadseqtree.png",
+       type = "d", border = NA, showtree = FALSE)
```

- La visualisation de l'arbre utilise Graphviz (http://www.graphviz.org/) qui doit être installé sur le système.

Arbre de régression



Documentation

- Le succès de TraMineR est largement dû à sa documentation.
- Site internet http://mephisto.unige.ch/traminer
  - dernières nouvelles
  - aperçu des possibilités
  - documentation :
    - manuel de l'utilisateur (env. 120 pages)
    - tutoriels
    - version en ligne (html) du manuel de référence
    - publications de l'équipe
    - publications d'utilisateurs de TraMineR
  - information sur les formations à TraMiner

TraMineR

Sequence analysis in R

[home] [doc] [training] [preview] [who uses it] [history] [install] [help & contact]

TraMineR mailing-list

If you have questions about using TraMineR and/or encounter problems, please write to the TraMineR mailing-list

Online help

You can see here a short preview of what TraMineR can do for you. Just have a look to get the flavour of TraMineR's main features and of how easy it is to put them at work.
Reference manual [html], [pdf]. See also the TraMineR page on the CRAN.

Documentation

- Mailing-list
- Online help
- User's Guide
- Citing TraMineR
- Trainings
- Publications
- Links

TraMineR User's Guide

The User's guide of TraMineR (pdf, ~3.6MB) describes the features and usage of TraMineR by means of many examples from the social sciences. It may also serve as an introduction to discrete sequential data analysis.

Gabadinho, A., G. Ritschard, M. Studer and N. S. Müller, *Mining sequence data in R with the TraMineR package: A user's guide* University of Geneva, 2009. (http://mephisto.unige.ch/traminer)

Citing TraMineR

Thank you for citing the article below when presenting analyses realized with the help of TraMineR.

Gabadinho, A., Ritschard, G., Müller, N.S. & Studer, M. (2011), Analyzing and visualizing state sequences in R with TraMineR, *Journal of Statistical Software*. Vol. 40(4), pp. 1-37.

Tutorials and trainings

On our training page, you may find training materials from past course, workshops and tutorials.

Publications

QuickSearch: clear Number of matching entries: 19/19. Search Settings

2011

Gabadinho, A., Ritschard, G., Studer, M. & Müller, N.S. (2011), *"Extracting and Rendering Representative Sequences"*, In Fred, A., Dietz, J.L.G., Liu, K. & Filipe, J. (eds) Knowledge Discovery, Knowledge Engineering and Knowledge Management. Series: Communications in Computer and Information Science (CCIS). Volume 128, pp. 94-106. Springer-Verlag.
[Abstract] [BibTeX] [DOI] [Preprint (pdf)]

Gabadinho, A., Ritschard, G., Müller, N.S. & Studer, M. (2011), *"Analyzing and visualizing state sequences in R with TraMineR"*, Journal of Statistical Software. Vol. 40(4), pp. 1-37.

Site R-forge et communauté d'utilisateurs

Nous avons également créé

- une liste de discussion
- un site sur R-forge (https://r-forge.r-project.org/projects/traminer/) pour
  - mettre à disposition la version de développement,
  - permettre aux utilisateurs de reporter des bugs,
  - et de proposer des fonctionnalités.

- ... et prochainement : un blog

Merci!

References I

Abbott, A. (1997). Optimize. http://home.uchicago.edu/~aabbott/om.html.

Abbott, A. (2001). *Time Matters. On Theory and Methods*. Chicago: Chicago Press.

Berchtold, A. and A. Berchtold (2004). MARCH 2.02: Markovian model computation and analysis. User's guide.

Brzinsky-Fay, C., U. Kohler, and M. Luniak (2006). Sequence analysis with Stata. *The Stata Journal 6*(4), 435–460.

Elzinga, C. H. (2007). CHESA 2.1 User manual. User guide, Dept of Social Science Research Methods, Vrije Universiteit, Amsterdam.

Gabadinho, A., G. Ritschard, N. S. Müller, and M. Studer (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software 40*(4), 1–37.

Gabadinho, A., G. Ritschard, M. Studer, and N. S. Müller (2009). Mining sequence data in R with the TraMineR package: A user's guide. Technical report, Department of Econometrics and Laboratory of Demography, University of Geneva, Geneva.

References II

Gabadinho, A., G. Ritschard, M. Studer, and N. S. Müller (2011). Extracting and rendering representative sequences. In A. Fred, J. L. G. Dietz, K. Liu, and J. Filipe (Eds.), *Knowledge Discovery, Knowledge Engineering and Knowledge Management*, Volume 128 of *Communications in Computer and Information Science (CCIS)*, pp. 94–106. Springer-Verlag.

McVicar, D. and M. Anyadike-Danes (2002). Predicting successful and unsuccessful transitions from school to work using sequence methods. *Journal of the Royal Statistical Society A 165*(2), 317–334.

Ritschard, G., A. Gabadinho, N. S. Müller, and M. Studer (2008). Mining event histories: A social science perspective. *International Journal of Data Mining, Modelling and Management 1*(1), 68–90.

Rohwer, G. and U. Pötter (2002). TDA user's manual. Software, Ruhr-Universität Bochum, Fakultät für Sozialwissenschaften, Bochum.

Studer, M., G. Ritschard, A. Gabadinho, et N. S. Müller (2009). Analyse de dissimilarités par arbre d'induction. *Revue des nouvelles technologies de l'information RNTI E-15*, 7–18.

Studer, M., G. Ritschard, A. Gabadinho, et N. S. Müller (2011). Discrepancy analysis of state sequences. *Sociological Methods and Research 40*(3), 471–510.