

La Boîte à Outils TraMineR pour l'Analyse de Trajectoires

De la visualisation à la recherche de séquences types

Gilbert Ritschard

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

Ateliers Ouvertures du CEREQ
Marseille, 8 décembre 2011

Plan

- 1 La boîte à outils TraMineR
- 2 Mesures descriptives transversales et longitudinales
- 3 Analyses fondées sur les dissimilarités deux-à-deux
- 4 Documentation et communauté d'utilisateurs
- 5 Conclusion

TraMineR, c'est quoi?

- TraMineR: **T**rajectory **M**iner in **R**
- Boîte à outils pour l'analyse de séquences.
 - Librement disponible sur le CRAN (Comprehensive R Archive Network) <http://cran.r-project.org/web/packages/TraMineR/>
 - Installation: `install.packages("TraMineR")`

TraMineR, c'est qui?

- Développé dans le cadre d'un projet de recherche FNS (Fonds national suisse de la recherche scientifique) 1/2007-1/2011
- sur la **fouille de données séquentielles en sciences sociales**
- Aujourd'hui, sous contrôle d'un comité scientifique:
 - Gilbert Ritschard (Statistique pour sciences sociales)
 - Alexis Gabadinho (Démographie)
 - Nicolas S. Müller (Sociologie, Système d'information)
 - Matthias Studer (Economie, Sociologie)
- ... développement se poursuit dans le cadre de l'IP 14 (module méthodologique) du PRN LIVES: Vulnérabilité, perspectives du parcours de vie.
 - Reto Bürgin (Statistique)
 - Emmanuel Rousseaux (Système d'information)

TraMineR, pourquoi?

- TraMineR a été conçu pour répondre à des questions de sciences sociales
- Les séquences (suites 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?

Ce qu'offre TraMineR pour répondre à ces questions

- une séries de graphiques et de mesures descriptives des séquences individuelles
- des outils pour calculer les dissimilarités entre séquences qui ouvrent la porte à toute une série d'outils statistiques
 - Clustering et analyse en coordonnées principales (MDS)
 - Analyse de la dispersion de séquences (ANOVA et arbre de régression)
 - Identification de séquences représentatives (trajectoires types)
 - ...

Séquences catégorielles

Données séquentielles catégorielles

Suites ordonnées de symboles tels que lettres, signaux, protéines, états, événements, ...

- au cœur de divers domaines
 - fouille de texte (séquences de lettres, mots, expressions, ...)
 - biologie (séquences de protéines, ADN, ...)
 - monitoring de l'activation d'appareils (ON/OFF),
 - étude des comportements temporels d'acheteurs ou utilisateurs (web logs),
 - étude de carrières et **parcours de vie**

Séquences d'états: exemples

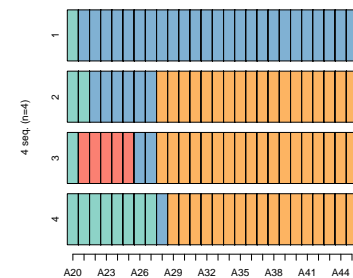
- Exemple de séquences cohabitationnelles (Données SHP)

2P = avec 2 parents, U = avec partenaire, C = enfant, A = seul, ...

- Sequence
- 1 2P-U
 - 2 2P-2P-U-U-U-U-U-U-UC
 - 3 2P-A-A-A-A-A-U-U-UC
 - 4 2P-2P-2P-2P-2P-2P-2P-2P-U-UC

- Présentation compacte

- Sequence
- 1 (2P, 1)-(U, 25)
 - 2 (2P, 2)-(U, 6)-(UC, 18)
 - 3 (2P, 1)-(A, 5)-(U, 2)-(UC, 18)
 - 4 (2P, 8)-(U, 1)-(UC, 17)



Séquences d'événements

- Au lieu de s'intéresser aux états:
2P-2P-U-U-U-UC-UC-UC-UC
- on peut s'intéresser aux **événements datés**:
(Quitter parents, 22), (Mise en union, 22), (Naissance enfant, 25)
- Les deux premiers événements définissent la **transition 2P→U**
- C'est une représentation alternative de trajectoires qui **nécessite d'autres outils** que les séquences d'états.
- Leur **visualisation est plus difficile** (pas de durée)
- TraMineR propose aussi des solutions pour séquences d'événements (non discutées ici)

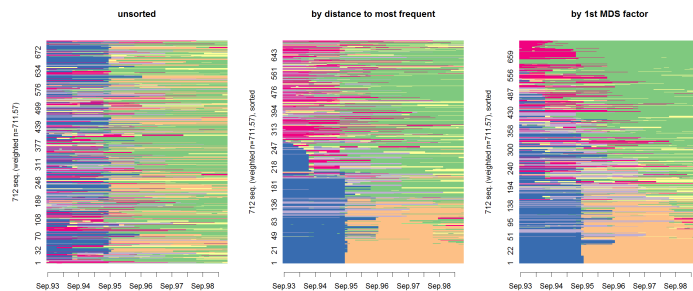
Présentations graphiques: Exemples

Séquences d'états, données McVicar and Anyadike-Danes (2002)



i-plot, et ordre des séquences

- Si grand nombre de séquences, ordonner pour améliorer la lisibilité



Aperçu des possibilités de TraMineR

Création d'une typologie par "Optimal matching"

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

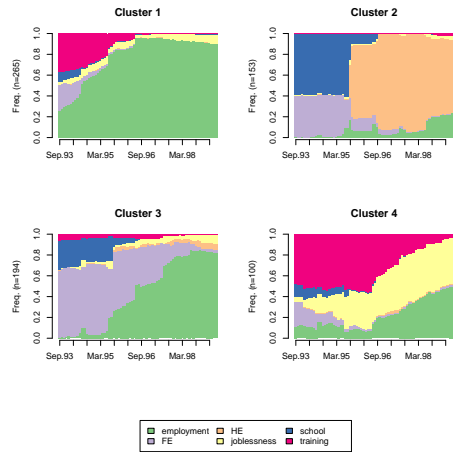

```
R> library(TraMineR)
R> data(mvad)
R> mvad.seq <- seqdef(mvad, 17:86, xtstep = 6)
```
- Calcul des dissimilarités OM entre paires de séquences avec un coût d'indel de 1 et des coûts de substitutions déduits des taux de transitions


```
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)

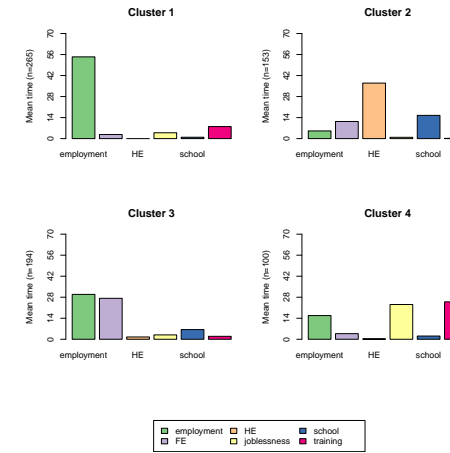
- Visualisation des classes: distributions transversales des états
`R> seqplot(mvad.seq, group = c14.lab, border = NA)`



7/12/2011gr 17/59

Aperçu des possibilités de TraMineR (suite 2)

- Temps moyen dans les états par classe
`R> seqmplot(mvad.seq, group = c14.lab)`



7/12/2011gr 18/59

Caractéristiques transversales versus longitudinales

- Distribution **transversale** à chaque position

id	t_1	t_2	t_3	...
1	B	B	D	...
2	A	B	C	...
3	B	B	A	...

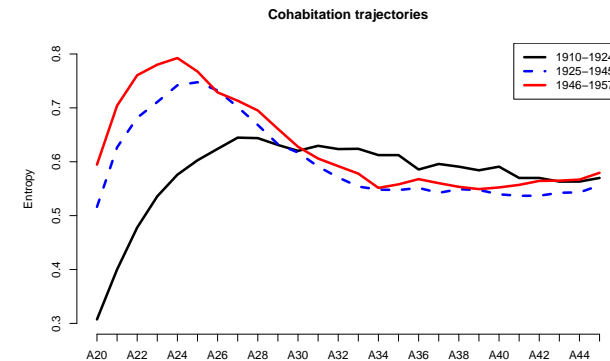
- Distribution **longitudinale** de chaque trajectoire

id	t_1	t_2	t_3	...
1	B	B	D	...
2	A	B	C	...
3	B	B	A	...

7/12/2011gr 21/59

Séquences de caractéristiques transversales Evolution de l'entropie transversale par cohortes (Widmer and Ritschard, 2009)

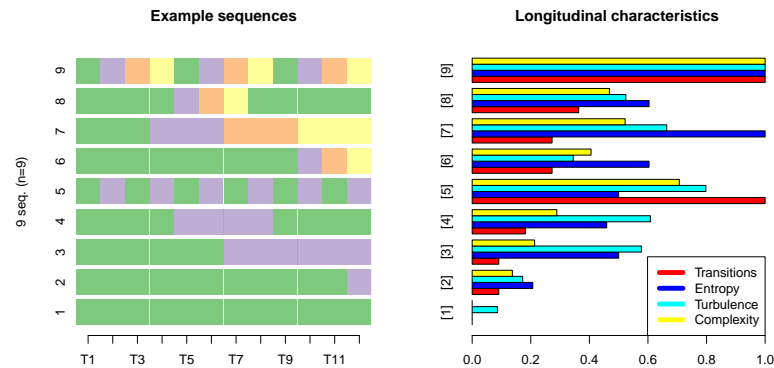
Données de l'enquête biographique du Panel suisse des ménages (2002)
 $n = 1503$, $a = 10$



7/12/2011gr 22/59

Caractéristiques longitudinales individuelles: complexité

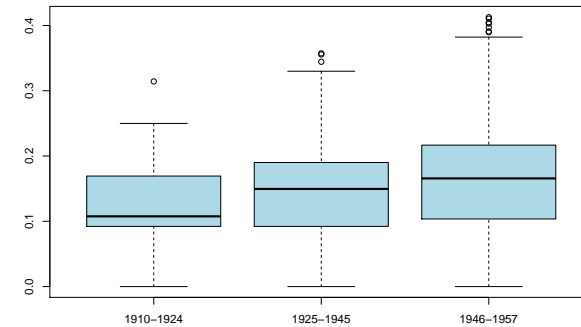
Séquences exemple ($\ell = 12$, $a = 4$) et valeur normalisée de mesures de complexité.



Distribution de la complexité

Complexité des trajectoires cohabitationnelles par cohortes

Données de l'enquête biographique du Panel suisse des ménages (2002)

 $n = 1503$, $a = 10$ 

Dissimilarités

Distance	Method	Position-wise	Additional arguments
<i>Count of common attributes</i>			
Simple Hamming	HAM	Yes	
Longest Common Prefix	LCP	Yes	
Longest Common Suffix	RLCP	Yes	
Longest Common Subsequence	LCS	No	
<i>Edit distances</i>			
Optimal Matching	OM	No	Insertion/deletion costs (<i>indel</i>) and substitution costs matrix (<i>sm</i>)
Hamming	HAM	Yes	substitution costs matrix (<i>sm</i>)
Dynamic Hamming	DHD	Yes	substitution costs matrix (<i>sm</i>)

Plusieurs autres mesures seront prochainement disponibles dans TraMineR

Dispersion de séquences

- A partir de la matrice des dissimilarités, on peut définir la **dispersion** d'un ensemble de séquences.
- Somme des carrés **SC** peut être exprimée en terme des distances deux-à-deux

$$SC = \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n} \sum_{i=1}^n \sum_{j=i+1}^n (y_i - y_j)^2$$

$$= \frac{1}{n} \sum_{i=1}^n \sum_{j=i+1}^n d_{e,ij}^2$$

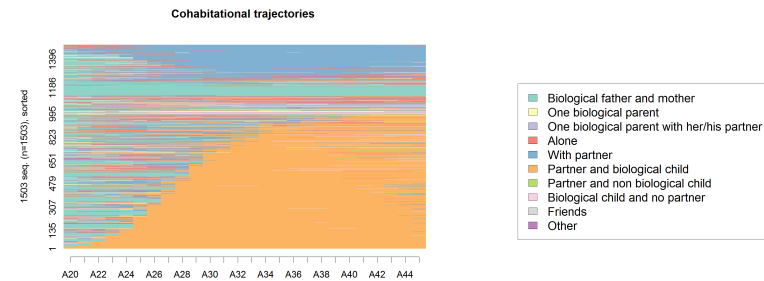
- En remplaçant $d_{e,ij}^2$ par la dissimilarité OM, LCP, LCS ... (ou son carré), on obtient une **pseudo SC**.

Analyses fondées sur dissimilarités

- Dès qu'on sait calculer des dissimilarités, on a accès
- à toute analyse fondée sur dissimilarités ou variance
 - Classification agglomérative, divisive, partitionnante, ... (Kaufman and Rousseeuw, 2005)
 - Cartes de Kohonen (Rousset and Giret, 2009)
 - Analyse en coordonnées principales (PCO, MDS) (Gower, 1966)
 - Séquences représentatives (Gabadinho et al., 2009b)
 - Analyse de variance (Studer et al., 2011)
 - Arbre de régression (Studer et al., 2011)

Construction d'une typologie

- Pour illustrer, classification hiérarchique de Ward
- Données: Trajectoires cohabitationnelles en Suisse,
 - 1503 séquences tirées de l'enquête biographique 2002 du PMS
 - alphabet de 10 états
 - données annuelles, de 20 à 45 ans (longueur 26)

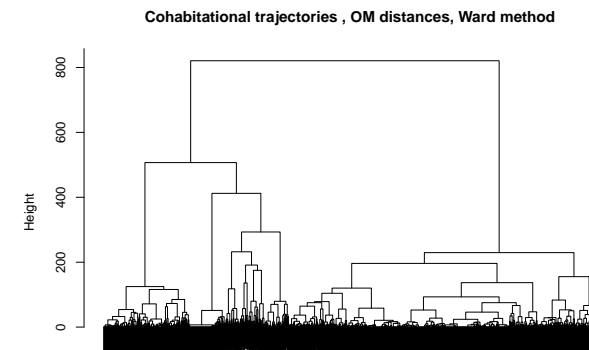


Classification à partir des dissimilarités

- Calculer la matrice des distances, par exemple avec TraMineR

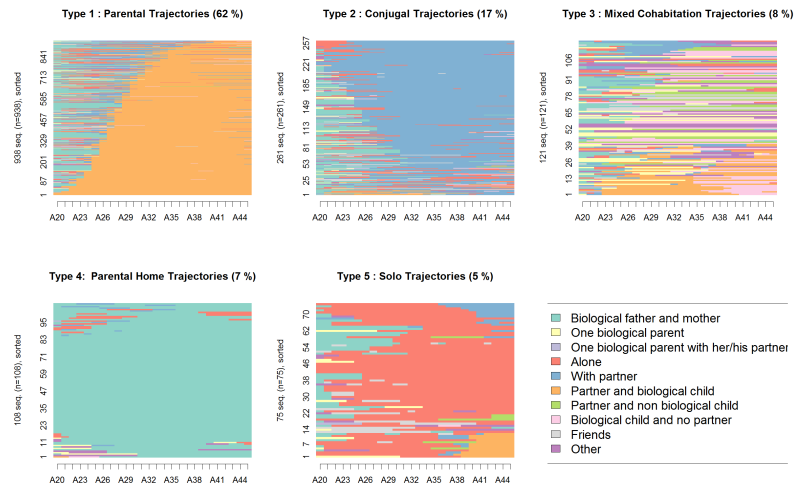
```
om.coh <- seqdist(seqs.coh, method="OM", sm="TRATE", indel=1)
```
- `om.dist.coh` matrice 1503×1503 que l'on peut passer à tout algorithme acceptant une matrice de dissimilarités en entrée
- Dans R, on peut par exemple utiliser la librairie `cluster` (Maechler et al., 2005) qui propose notamment
 - `agnes()` méthode agglomérative
 - `diana()` méthode divisive
 - `pam()` partitionnement autour de médoïdes
- Illustration: méthode agglomérative avec critère de Ward
- On utilise la fonction `agnes()`
 - ```
clw.coh <- agnes(om.coh, diss=T, method="ward")
```
- et retenons la partition en 5 classes
  - ```
cutree(clw.coh, k=5)
```

Classification hiérarchique, Ward Dendrogramme



Typologie des trajectoires cohabitationnelles

i-plot, ordre selon MDS[1]



Typologie des trajectoires cohabitationnelles

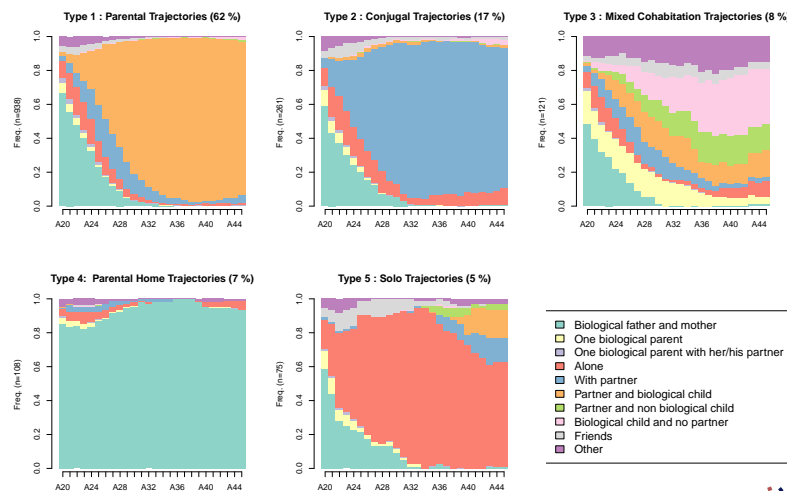
Dispersions

• Dispersion (pseudo variance) $\frac{1}{2n^2} \sum_i \sum_j d(i, j)$

	Count	Percent	Discrepancy
Parental	938	62.4	7.819
Conjugal	261	17.4	8.209
Mixed	121	8.1	19.842
Parental Home	108	7.2	3.002
Solo	75	5.0	9.185
Total	1503	100.0	15.526

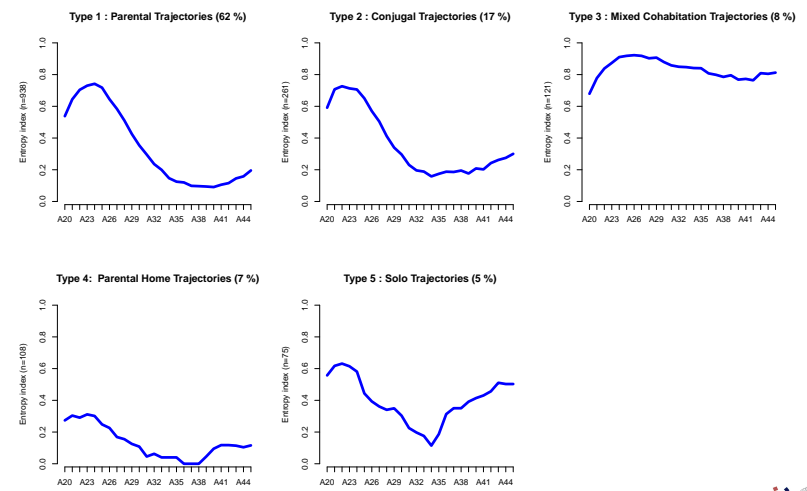
Typologie des trajectoires cohabitationnelles

d-plot, distributions transversales



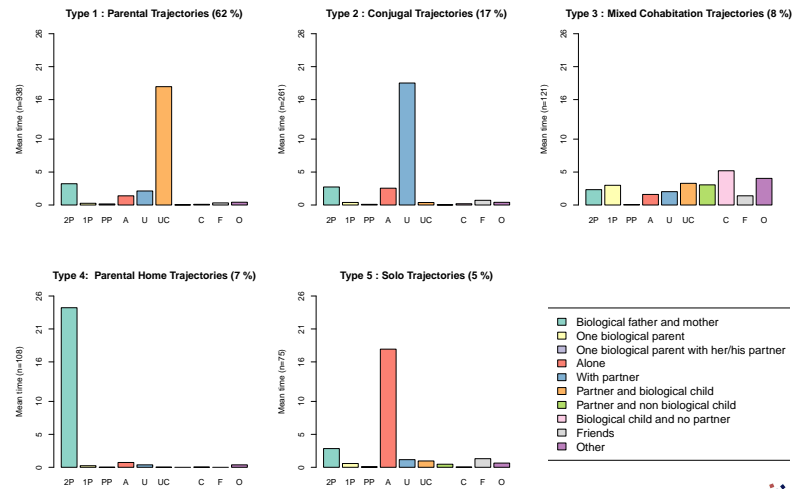
Typologie des trajectoires cohabitationnelles

Ht-plot, entropies transversales



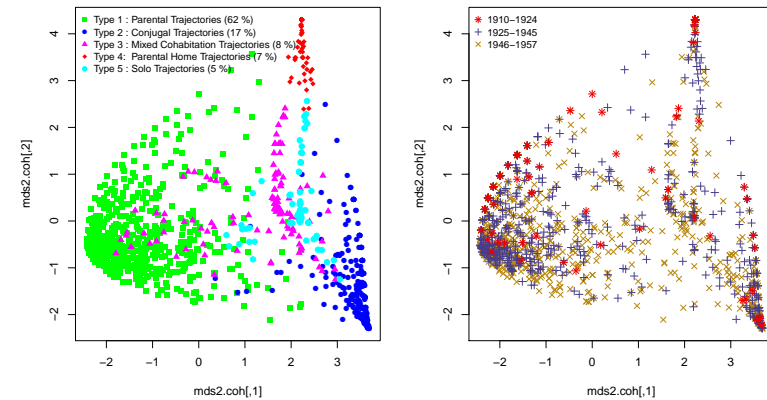
Typologie des trajectoires cohabitationnelles

mt-plot, temps moyen par état



7/12/2011gr 38/59

MDS: Nuage de points



7/12/2011gr 39/59

Séquences représentatives

- Objectif: synthétiser un ensemble de séquences
- Trouver le plus petit ensemble de séquences
 - non redondantes
 - assurant une couverture donnée de l'ensemble
- Redondance et couverture définies en termes de **voisinage**
 - x et y **redondant** si $d(x, y) \leq \delta_{tsim}$
 - **couverture**: % de séquences dans le voisinage d'au moins une des séquences représentatives r

7/12/2011gr 41/59

Séquences représentatives

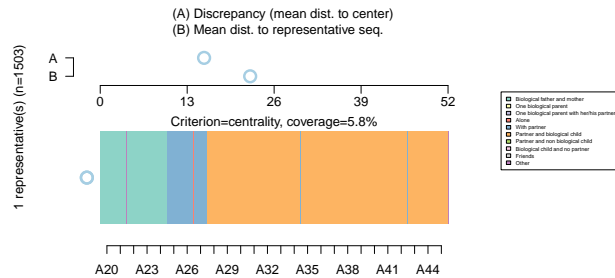
Heuristique

- 1 **Ordonner les séquences** selon un critère de représentativité
 - **densité** nombre de séquences dans son voisinage
 - **centralité** somme des distances à toutes les autres séquences
 - autres: fréquence, moyenne des fréquences des états qui la composent, vraisemblance, ...
- 2 **Supprimer la redondance**
 - Calculer la couverture de celle qui a le meilleur score
 - Puis, pour les suivantes
 - supprimer si redondante avec représentants retenus
 - sinon, calculer couverture du nouvel ensemble de représentants
 - Arrêt lorsque la couverture voulue est atteinte.

7/12/2011gr 42/59

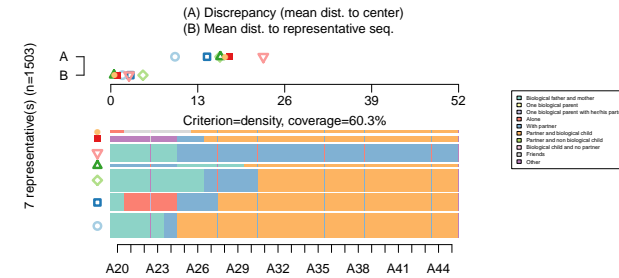
Séquences représentatives: Exemple Médoïde des trajectoires cohabitationnelles

- Critère: centralité
- taille ensemble représentatif = 1



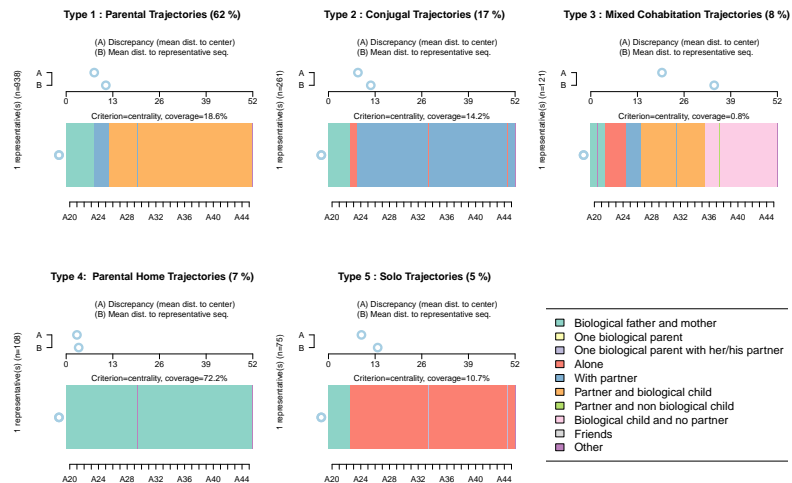
Séquences représentatives: Exemple densité Trajectoires cohabitationnelles (tsim=.2, trep=.6)

- Critère: densité, diamètre voisinage = 20% de la distance maximale
- couverture minimale = 60%

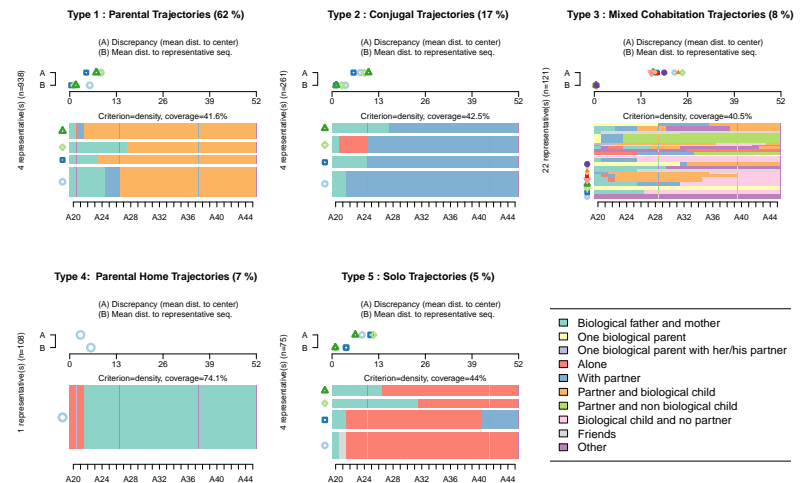


- Avec les séquences représentatives, on ne repère pas les petits groupes (Tanguy)

Séquences représentatives par classe Médoïdes des trajectoires cohabitationnelles



Séquences représentatives par classe Trajectoires cohabitationnelles (tsim=.1, trep=.4)



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

7/12/2011gr 48/59

7/12/2011gr 49/59

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.

7/12/2011gr 50/59

Conclusion 1: Sur l'analyse de séquences

- Analyser trajectoires jusqu'à 45 ans, => ignorer les **générations récentes**
- Année naissance la plus récente **1957 (2002 – 45)**
- Problèmes:
 - **Granularité**: année, mois, jour, ...
 - **Définition des états**: faut-il distinguer {séparé, divorcé, veuf} ou considérer comme un seul état? travaux de Raffaella Piccaretta

7/12/2011gr 52/59

Conclusion 2: Données manquantes et pondérations

- **Données manquantes** dans les séquences: problème capital
- TraMineR permet des traitements différenciés pour les données manquant à droite, gauche et à l'intérieur de la séquence
 - considérer comme un état propre
 - supprimer (glissement à gauche des états subséquents)
 - imputer, mais comment?
- **Pondération des cas**
 - Prise en compte dans le rendu des séquences (pondération des caractéristiques transversales)
 - Solutions également pour ANOVA et test de permutation
 - Pas pertinent pour calcul des dissimilarités et des caractéristiques longitudinales

Conclusion 3: Extension de l'analyse

- Comme TraMineR est une librairie R, ses sorties peuvent facilement être combinées dans même script avec d'autres procédures R
- Nous avons vu: l'analyse en clusters, MDS, ...
- In Widmer and Ritschard (2009),
 - Relation entre trajectoires **occupationnelles** and **cohabitationnelles** par des régressions des entropies longitudinales de chacune d'entre-elles sur les types occupationnels and cohabitationnels en contrôlant pour les cohorte de naissance et le sexe.
 - Etude aussi de **l'appartenance au type** par des régressions logistiques.

Conclusion 4: Application à d'autre type de données

- Les techniques discutées pour les séquences
- ... s'appliquent à toutes données non mesurables caractérisées par leur dissimilarités deux-à-deux.
- Seul aspect propre aux séquences d'états: rendu visuel.

Conclusion 4: A propos de TraMineR

- **TraMineR** est un outil unique pour analyse de séquences discrètes
- Peut faire beaucoup plus que ce qui a été vu,
 - gestion de données séquentielles
 - conversion entre séquences d'événements et d'états
 - dissimilarité multi-canal pour séquences parallèles
 - analyse de dispersion, arbre de régression
 - séquences d'événements: sous-séquences fréquentes,
 - sous-séquences discriminantes
 - ...
- ... et, comme **R**, disponible gratuitement sur le **CRAN**
<http://cran.r-project.org>
- Voir aussi la page web
<http://mephisto.unige.ch/traminer>

Merci!

References I

- 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 (2009a). 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.
- Gabadinho, A., G. Ritschard, M. Studer, and N. S. Müller (2009b). Summarizing sets of categorical sequences. In *International Conference on Knowledge Discovery and Information Retrieval, Madeira, 6-8 October, 2009*, pp. 62–69. INSTICC. (Received the Best Paper Award).
- Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* 53(3/4), 325–338.
- Kaufman, L. and P. J. Rousseeuw (2005). *Finding Groups in Data*. Hoboken: John Wiley & Sons.

References II

- Maechler, M., P. Rousseeuw, A. Struyf, and M. Hubert (2005). Package 'cluster': Cluster analysis basics and extensions. Reference manual, R-project, CRAN.
- 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.
- Rousset, P. and J.-F. Giret (2009). A longitudinal analysis of labour market data with SOM. In J. R. Rabuñal, J. Dorado, and A. Pazos (Eds.), *Encyclopedia of Artificial Intelligence*, pp. 1029–1035. IGI Global.
- Studer, M., G. Ritschard, A. Gabadinho, and N. S. Müller (2011). Discrepancy analysis of state sequences. *Sociological Methods and Research* 40(3), 471–510.
- Widmer, E. and G. Ritschard (2009). The de-standardization of the life course: Are men and women equal? *Advances in Life Course Research* 14(1-2), 28–39.