Une introduction à l'analyse de séquences et quelques-unes de ses utilisations en sciences sociales

Matthias Studer
Gilbert Ritschard

LIVES NCCR, Université de Genève
Institut de Démographie et Socioéconomie, Université de Genève
http://mephisto.unige.ch/traminer

Déjeuner R, Université de Genève, 6 décembre 2016

Analyse de séquences

- Cadre méthodologique pour l'analyse de trajectoires et processus (Abbott and Tsay, 2000) :
  - Carrières (criminelle, professionnelle...)
  - Développement d'organisation, entreprise, nations.
  - Transitions de l'école à l'emploi...
  - Utilisation journalière (ou hebdomadaire) du temps...
  - ...
- But : analyser les processus dans leur globalité, développement entre plusieurs statuts.

Question sur des processus

- Centré sur un événement ou état.
  - Événement : Trouver un emploi...
  - État : Être employé...
- Questions :
  - Occurrence (récurrence ?)
  - Ordonnancement.
  - Temporalité.
  - Fréquence.
- Méthodes :
  - Analyse de biographie.
  - Analyse de données de panel.
  - ...

Question sur les processus II

- Vue holistique sur les processus :
  - Développement entre plusieurs états.
  - Transition qui prend du temps.
  - Synthétiser une période de vie.
    - Transition de l'école à l'emploi.
    - Réinsertion professionnelle.
- Dimensions clefs :
  - Ordonnancement.
  - Temporalités.
  - Durées.
- Méthodes : Analyse de séquences.

Questions de l'analyse de séquences

- Description de trajectoires.
- Identification de trajectoires typiques :
  - Existe-t-il des trajectoires typiques (fréquentes) ?
  - Quelles sont ces trajectoires ?
  - Quelle dispersion autour de ces trajectoires types ?
- Expliquer les trajectoires :
  - Quels facteurs influent sur la trajectoire suivie par un individu ?
- Impact d'une trajectoire passée.
  - Comment un résultat est-il influencé par une trajectoire passée ?

Méthodes de l'analyse de séquences

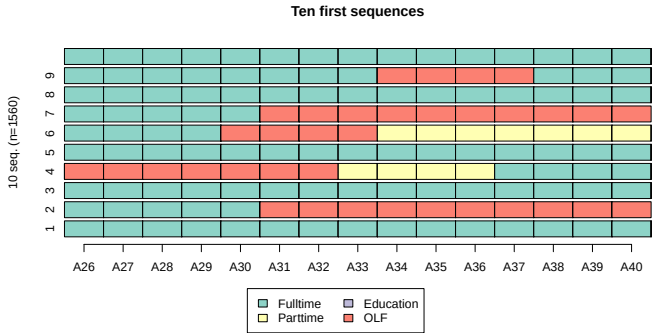
- Analyse descriptive de trajectoires.
  - Visualisation.
  - Statistiques descriptives.
- Typologie de trajectoires (Abbott and Forrest, 1986).
  - Utilisation de la typologie dans d'autres analyses.
- Analyse de dispersion pour séquences d'états (Studer et al., 2011).
  - Analyse des relations entre facteurs explicatifs et trajectoires.

Exemple

- Trajectoires de taux d'occupation en Suisse.
- Étude de Levy et al. (2006).
- But :
  - Identifier les changements de trajectoires au 20e siècle.
  - Parcours de vie genré.
  - Identifier les patterns d'aller-retour des femmes sur le marché du travail en Suisse.
- Données :
  - 1560 individus entre 26 et 40 ans.

Les trajectoires comme séquences

- Analyse de séquence : Trajectoires codées comme séquences d'états.
  - Séquence : liste ordonnée des états tirés de l'alphabet.
  - Alphabet : plein temps, temps partiel, inactifs (ou chômeur), éducation.



Séquences d'états ou d'événements

- Séquences chronologiques :
  - La position dans la séquences indique le temps.
  - Et la précédence.
- Distinction importante
  séquences d'états et séquences d'événements
  - Un état dure pour toute la durée de l'unité de temps.
  - Un événement, tel que "enménager" ou "terminer les études", ne dure pas, mais provoque un changement d'état, potentiellement en interaction avec d'autres événements.

Séquences d'états ou d'événements : exemple

Time stamped events (TSE)

Sandra	Ending education in 1980	Start working in 1980
Jack	Ending education in 1981	Start working in 1982

- Les événements peuvent être simultanés (cf. Sandra).
- Eléments à la position ne se produise pas nécessairement au même moment.

State sequence view

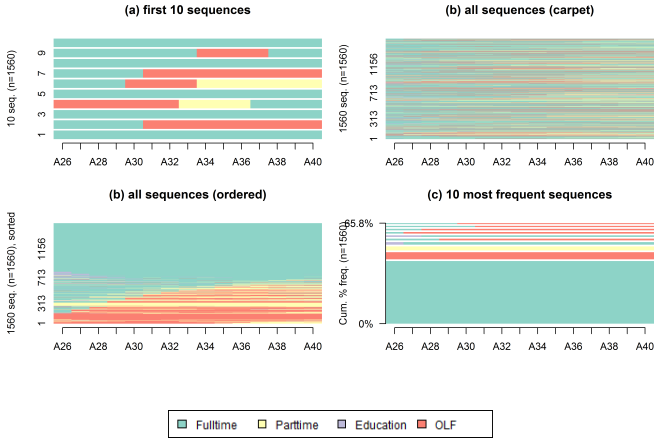
year	1979	1980	1981	1982	1983
Sandra	Education	Education	Employed	Employed	Employed
Jack	Education	Education	Education	Unemployed	Employed

- Un état observé uniquement pour chaque unité de temps.
- La position indique la temporalité : tous les états à la position 2 indique l'état en 1980.

Considérations pratiques

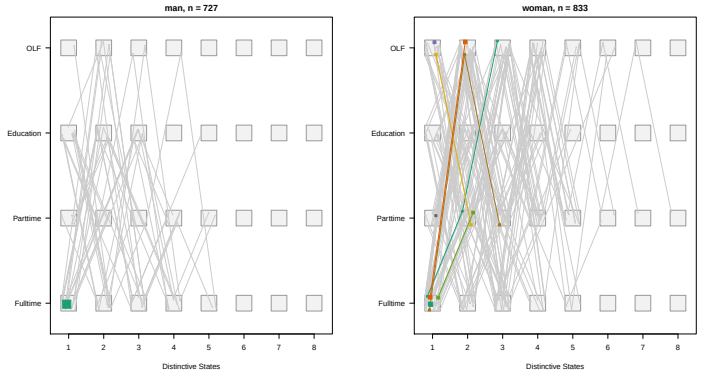
- Comment coder un processus ?
- Points clefs :
  - Définition du début du processus.
    - Un événement particulier (terminer les études, ...).
    - Un âge (15 ans).
    - Une année précise (2000).
  - Définition de l'alphabet :
    - Basé sur la question de recherche.
    - Uniquement les distinctions les plus importantes.
    - Grand versus petit alphabet.

Visualisation basée sur les séquences



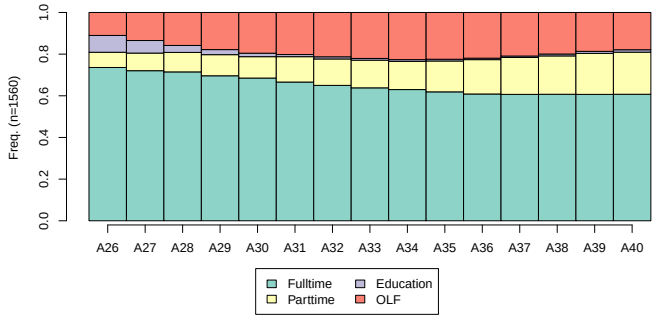
Visualisation de l'ordonnancement

- Parallel coordinates plot (Bürgin and Ritschard, 2014).
- Patterns fréquents (75% des cas).



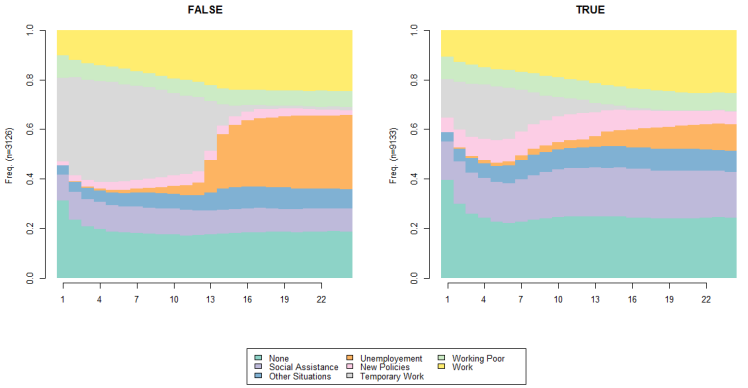
Chronogramme

- Distribution entre état à chaque position.



Évaluation d'une politique publique

- Réforme de la politique de réinsertion à Genève.
- Trajectoires des chômeurs en fin de droits.
  - Transfert entre institutions.
  - Emplois discontinus.
  - État « Aucun ».

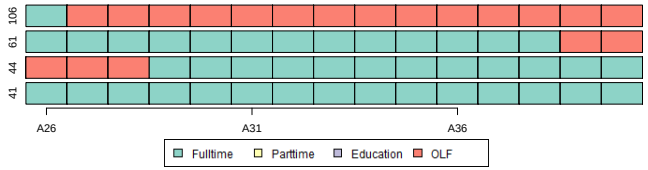


Typologie de séquences en sciences sociales

- Analyse de séquences :
  - Existe-t-il des séquences typiques (ou récurrentes) ?
  - Quelles sont ces trajectoires ?
- Analyse exploratoire
- Analyse en cluster.
  - But : regrouper les séquences similaires dans des types en ignorant les petites différences.
  - Etapes :
    - Calcul des dissimilarités entre séquences.
    - Clustering.

Dissimilarité entre séquences

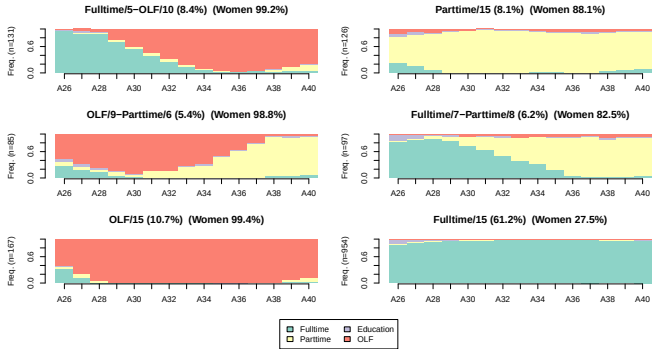
- Quantification des différences entre deux objets.



- Optimal Matching.
- Mais beaucoup d'autres qui diffèrent selon (Studer and Ritschard, 2016) :
  - Ordonnancement des états et transitions.
  - Temporalité des états et transitions.
  - Temps passé dans chaque état.

Analyse en cluster

- Six types de séquences.



Interprétation de la typologie

- Identifications de patterns liés à :
  - Logiques de développement différentes.
  - Interdépendances entre moments de la trajectoire.
  - Contraintes légales, économiques ou sociales.
  - Norme sociale sur les parcours de vie.
- Modèles pour les acteurs ?
- Trajectoires attendues ?

Expliquer les trajectoires

- La typologie est une variable catégorielle.
- Peut être utilisé comme variable dépendante.
- But : Comprendre les facteurs qui influencent les chances de suivre un type de trajectoires.
- Modèle multinomial (référence : plein-temps).

	Full-OLF	Part	OLF-Part	Full-Part	OLF
Constante	-6.61^{***}	-4.82^{***}	-7.01^{***}	-4.81^{***}	-6.31^{***}
	(1.01)	(0.36)	(1.03)	(0.38)	(1.01)
Femme	5.84^{***}	2.96^{***}	5.39^{***}	2.50^{***}	6.09^{***}
	(1.01)	(0.29)	(1.01)	(0.28)	(1.01)
Né après 1945	0.11	1.29^{***}	0.67^{*}	1.42^{***}	-0.39^{*}
	(0.23)	(0.29)	(0.29)	(0.34)	(0.20)

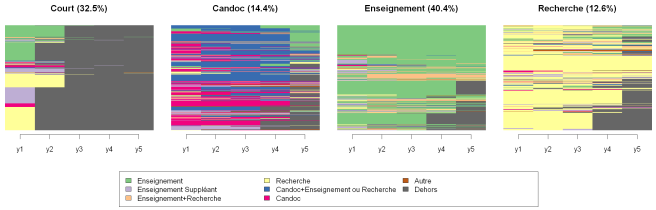
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Effet d'une trajectoire passée

- La typologie comme variable indépendante.
- But : Comprendre l'effet d'un type de trajectoire passée sur la situation présente ou future.
- Exemple :
  - Effet de la trajectoire professionnelle sur la pauvreté chez les personnes âgées (Gabriel et al., 2015).
  - Condition de travail passée et réussite au doctorat chez les assistants-doctorants à Genève (Studer, 2012).

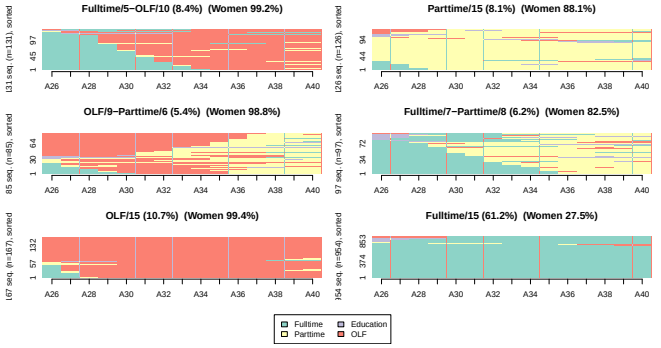
Trajectoires de doctorants

- Trajectoire de conditions de travail.
- Analyse des liens avec la réussite au doctorat.
- Basé sur les archives de l'Université de Genève (Studer, 2012).



Limites

- Simplification des séquences.
- Hypothèse : dispersion intra cluster peut être ignorée.
- Si non vérifiées, risque de fausses conclusions.
- Comment confirmer les liens observés ?



Analyse de dispersion des séquences

Analyse de dispersion (Studer et al., 2011) :

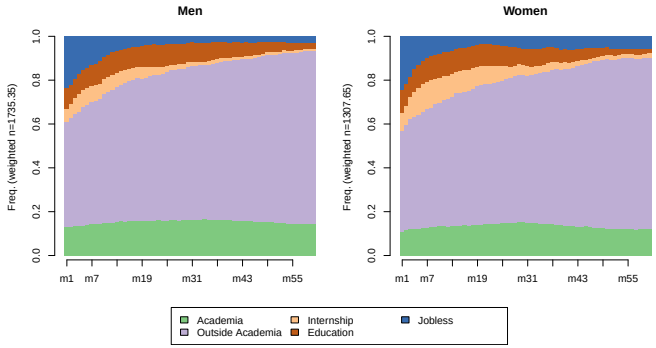
- Étude des liens entre séquences d'états et variables explicatives.
- Extension du cadre méthodologique de l'ANOVA.
- Mesure de la force du lien à l'aide de la part «expliquée» de la dispersion des séquences.
- Estimation de la significativité ("p-value") à l'aide de tests de permutation.

Problématique exemple

- Carrière professionnelle suite au Master en Suisse (Studer, 2012).
- But : mesure des inégalités hommes-femmes dans le monde académique.
- Données.
  - 3043 individus suivis pendant 5 ans après le Master.
  - Alphabet : Outside Academia, Academia, Internship, Jobless, Further Education.

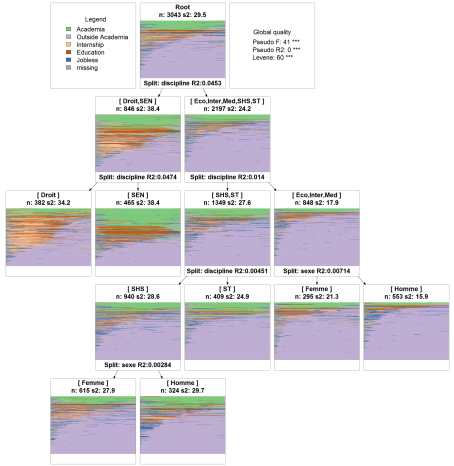
Comparaison de groupes de séquences

- Est-ce que les trajectoires diffèrent selon le sexe ?
- Pseudo-$R^2 = 0.002$ et $\hat{p}$-value $= 0.001$.



Arbre de régression

- Choix récursif de la « meilleure » variable explicative.
- Uniquement des séparations significatives.



Analyse de séquences

- Vue d'ensemble sur les trajectoires et les processus.
  - Analyse descriptive.
  - Identification de typologie.
  - Mesure des liens avec des co-variables.
- Vue globale en prenant en compte.
  - Ordonnancement.
  - Temporalité.
  - Temps passé dans chaque état.
- Approche complémentaire à d'autres.

Débuter : bibliographie choisie

- Introduction à l'analyse de séquences avec TraMineR :
  - 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.
- Création de typologie de séquences :
  - Studer, M. (2013).WeightedCluster library manual : A practical guide to creating typologies of trajectories in the social sciences with R. *LIVES Working Papers* 24, NCCR LIVES, Switzerland.
- Analyse de dispersion des séquences :
  - Studer, M., G. Ritschard, A. Gabadinho and N. S. Müller (2011). Discrepancy Analysis of State Sequences. *Sociological Methods and Research 40*(3), 471–510.

TraMineR

- TraMineR est une librairie `R` pour l'analyse de séquences.
- Destiné aux sciences sociales (mais utilisé au-delà).
- TraMineR : Trajectory Miner in R (Aucun rapport avec le Gewürztraminer)
- Installation : `install.packages("TraMineR")`
- Pourquoi une librairie `R` ?
  - Importation de données (SPSS, Stata, ...).
  - Analyse des résultats produit par TraMineR à l'aide d'autres packages (Clustering, MDS, multi-level models, ...).
  - Utilisation des résultats d'autres packages dans TraMineR (régression, etc.).

Possibilités offertes par `TraMineR`

- Toutes les analyses présentées ici.
- Autre visualisation (librairie `TraMineRextras`).
- Prise en charge et conversion entre différents format de données longitudinal.
- Prise en charge des pondérations et données manquantes.
- Extraction et visualisation de séquences représentatives d'un ensemble de séquences.
- Autre statistiques descriptives (taux de transitions, ...).
- Calcul de caractéristiques longitudinal des séquences (indice de complexité, entropie longitudinale, turbulence, temps passé dans chaque état, ...)
- Test d'Homogénéité des dispersions
- Analyse de séquences d'événements.
  - Extraction de sous-séquences fréquentes.
  - Identification de sous-séquences discriminantes.

References I

Abbott, A. and J. Forrest (1986). Optimal matching methods for historical sequences. *Journal of Interdisciplinary History 16*, 471–494.

Abbott, A. and A. Tsay (2000). Sequence analysis and optimal matching methods in sociology, Review and prospect. *Sociological Methods and Research 29*(1), 3–33. (With discussion, pp 34-76).

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

Bürgin, R. and G. Ritschard (2014). A decorated parallel coordinate plot for categorical longitudinal data. *The American Statistician 68*(2), 98–103.

Elzinga, C. H. (2009). *CHESA 3.1 User Manual*. Amsterdam : Vrije Universiteit.

References II

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.

Gabriel, R., M. Oris, M. Studer, and M. Baeriswyl (2015). The persistence of social stratification ? a life course perspective on poverty in old-age in switzerland. *Swiss Journal of Sociology*.

Halpin, B. (2014). SADI : Sequence analysis tools for Stata. Department of Sociology Working Paper Series WP2014-03, University of Limerick.

Levy, R., J.-A. Gauthier, and E. Widmer (2006). Entre contraintes institutionnelle et domestique : les parcours de vie masculins et féminins en suisse. *Cahiers canadiens de sociologie 31*(4), 461–489.

References III

Studer, M. (2012). *Étude des inégalités de genre en début de carrière académique à l'aide de méthodes innovatrices d'analyse de données séquentielles*. Thèse de doctorat n° 777, Faculté des sciences économiques et sociales, Université de Genève.

Studer, M. and G. Ritschard (2016). What matters in differences between life trajectories : A comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society : Serie A 179*(2), 481–511.

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

Softwares

- **R** :
  - TraMineR (Gabadinho et al., 2011)
  - WeightedCluster (Studer, 2012)
- **Stata** :
  - SADI : Sequence Analysis Distance Measures (Halpin, 2014)
  - SQ (Brzinsky-Fay et al., 2006)
- Chesa (Elzinga, 2009)

Longitudinal data formats

Code	Example
STS	*Id* *a18* *a19* *a20* *a21* *a22* *a23* *a24* *a25* *a26* *a27* 101 S S S M M MC MC MC MC D 102 S S S MC MC MC MC MC MC MC
DSS	*Id* *s1* *s2* *s3* *s4* 101 S M MC D 102 S MC
SPS	*Id* *s1* *s2* *s3* *s4* 101 (S,3) (M,2) (MC,4) (D,1) 102 (S,3) (MC,7)
SPELL	*Id* *Index* *From* *To* *State* 101 1 18 20 Single (S) 101 2 21 22 Married (M) 101 3 23 26 Married w Children (MC) 101 4 27 27 Divorced (D) 102 1 18 20 Single (S) 102 2 21 27 Married w Children (MC)
TSE	*Id* *Time* *Event* 101 21 M (Marriage) 101 23 C (Childbirth) 101 26 C (Childbirth) 101 27 D (Divorce) 102 21 M (Marriage) 102 21 C (Childbirth)

Other methods in Sequence Analysis

- Extraction and visualization of representative sequences of a set of sequence.
- Analyze complexity/turbulance of sequences.
- Homogeneity of discrepancy
- Analysis of event sequences.
  - Extraction of frequent event subsequences.
  - Identification of discriminant subsequences.