Analyse de dispersion de séquences d'états: analyser les liens entre trajectoires et variables explicatives

Matthias Studer

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

Rencontres méthodes et recherche
FORS, Lausanne, 24 janvier 2012

Plan

- Analyse de séquences.
- Clustering de séquences.
- Tester l'association entre séquences et variables explicatives avec les clusters.
- Analyse de dispersion des séquences.
- Approches multifacteurs.
- Interprétation et test de différences de dispersion.
- Mise en œuvre avec `TraMineR`.

Objectifs de la présentation :

- Présentation des méthodes offertes par l'analyse de dispersion.
- Comparaison avec les méthodes usuellement utilisées.

A quoi ça sert ?

- Mesure des liens entre trajectoires (séquences) et facteurs explicatifs.
  - Sexe et carrière académique.
  - Cohorte et trajectoire familiale.
  - Origine sociale et insertion professionnelle.
- Approches multifacteurs.
  - Sexe, ségrégation horizontale et carrière académique.
- Interprétation de la dispersion et test d'égalité des dispersions.

Pour les détails :

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

Analyse de séquences en sciences sociales

- Étude des trajectoires de vies ou, plus généralement, des processus sociaux.
  - Carrières professionnelles.
  - Trajectoires de cohabitations.
  - Histoires d'organisations.
- Particularité :
  - Analyse de l'ordonnancement des séquences (Patterns).
  - Prise en compte de la multiplicité des états possibles.
  - Temporalités (âge aux différentes moments de la trajectoires).
- Vue holistique sur les trajectoires.
- Analyse des états et des transitions dans le contexte du processus dans son ensemble (Billari, 2001).

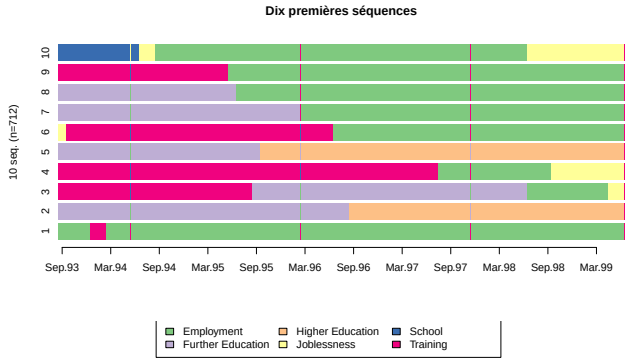
Problématique

- Étude de McVicar and Anyadike-Danes (2002) sur les transitions à l'emploi en Irlande du Nord.
- But : Identifier les profils de jeunes irlandais qui ont le plus de risques de connaître des transitions à l'emploi difficiles.
- Description des données :
  - Séquences commencent à la fin de l'école obligatoire.
  - Durée de 70 mois.
  - États possibles : EM (Employment), FE (Further Education), HE (Higher Education), JL (Joblessness), SC (School), TR (Training).
  - 712 individus.
  - Données incluses dans `TraMineR` (mvad).

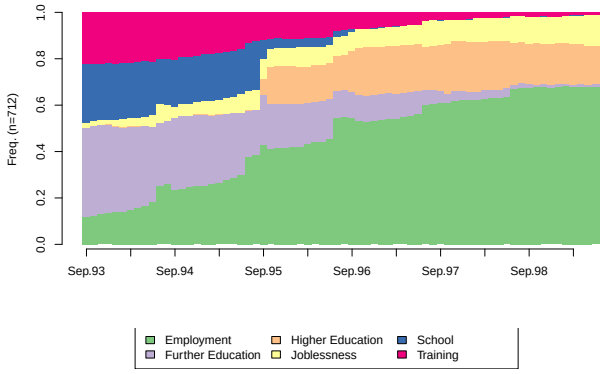
Définition d'une séquence d'états

Définition

- Alphabet A : ensemble des états possibles.
- Séquence de longueur k : liste ordonnée de k éléments appartenant à A.



Chronogramme



Analyse en cluster

- But : création de groupes de séquences :
  - les plus homogènes possible.
  - Les plus différents possibles les uns des autres.
- Mise en évidence des patterns récurrents des trajectoires.
- Point de vue descriptif et exploratoire.

Méthode traditionnellement utilisée :

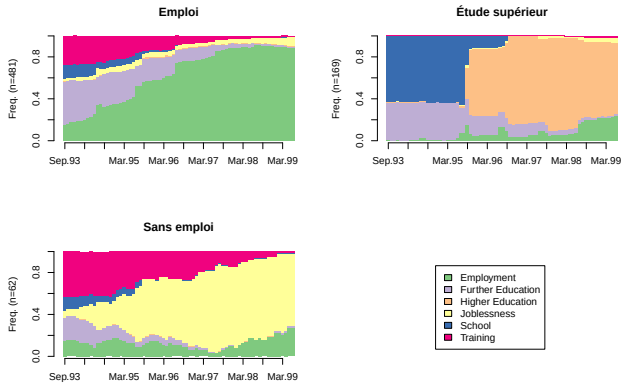
- Calculer des dissimilarités entre séquences.
- Regroupement des séquences à l'aide de clustering.
- Mesure des liens entre la solution du clustering et les variables explicatives.
- Régressions logistiques, test du chi-carré, ...

Dissimilarité

- Une mesure de dissimilarité est une quantification de l'éloignement de deux objets.
- Par exemple, pour deux revenus x et y :
  - $d(x, y) = |x - y|$
  - $d(x, y) = (x - y)^2$ (ANOVA).
- Grand nombre de dissimilarités pour séquences (Optimal Matching, Hamming, NMS, ...)
- Ces distances comparent :
  - L'ordre des états et des transitions dans les séquences.
  - La temporalité des transitions.
  - Temps total passé dans chaque état.
- Ici, optimal matching avec les coûts définis dans l'article original.

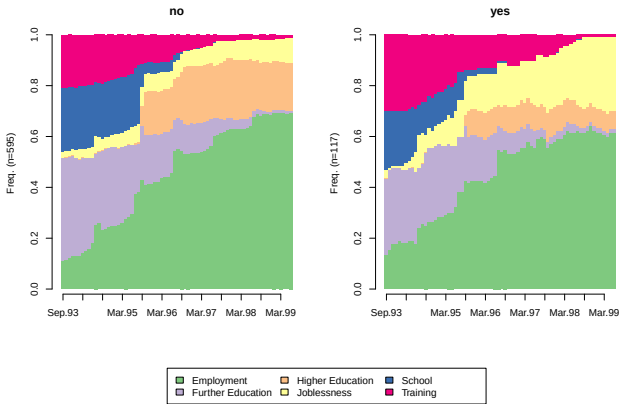
Analyse en cluster

- Retient trois groupes (meilleure silhouette=0.41).



Comparer des groupes des séquences

- Est-ce que les trajectoires diffèrent selon le statut professionnel du père (sans emploi) ?
- Ces différences sont-elles significatives ?

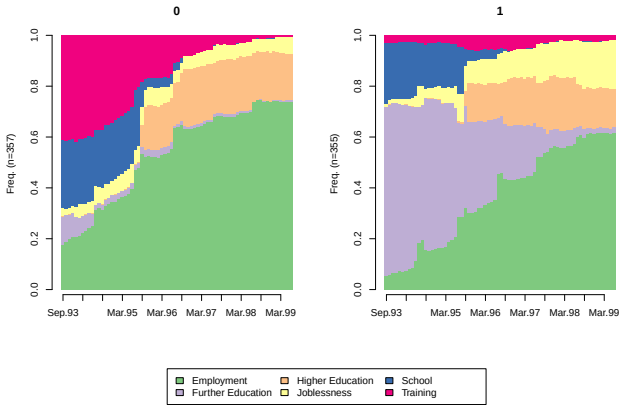


Lien avec les variables explicatives

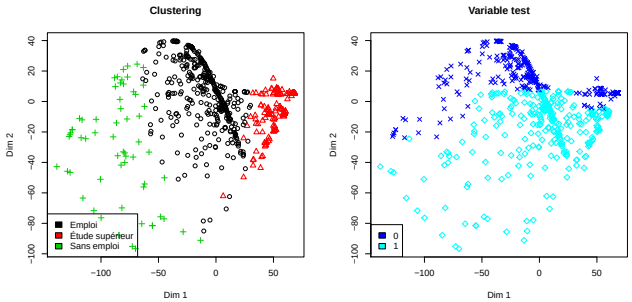
- Test du chi-carré et v de Cramer.

	Cramer's v	p-value
gcse5eq	0.51	0.000
Grammar	0.29	0.000
region	0.21	0.000
funemp	0.19	0.000
fmpr	0.19	0.000
religion	0.14	0.001
sex	0.12	0.005
livboth	0.10	0.031
test	0.00	0.999

Absence de lien avec la variable test ?



Quel est le problème ?



Quel est le problème ?

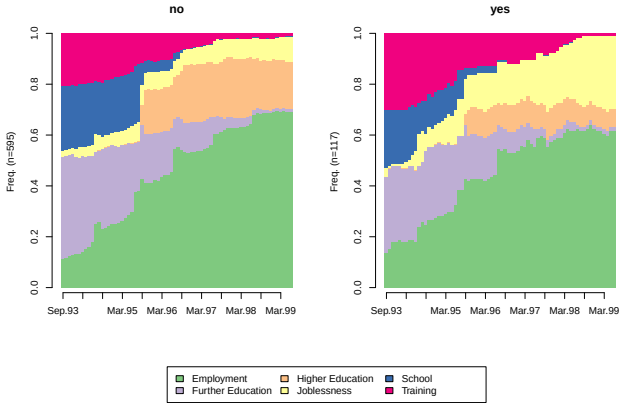
- Le clustering est une simplification de l'ensemble des séquences.
- On associe chaque séquence à un modèle (ou type).
- Chaque séquence est considérée comme une réalisation de ce modèle.
- L'écart au « vrai » modèle est considéré comme un "terme d'erreur" négligeable.
  - La diversité des séquences dans chaque type est supprimée.
  - Les différences de proximités entre types sont ignorées.

Quel est le problème ?

- Cette simplification peut être justifiée si les groupes sont très homogènes et clairement séparés les uns des autres.
- Dans les cas contraires, la simplification peut créer ou cacher des associations.
- Par exemple, si la variable explicative explique les différences de trajectoire au sein d'un type.

But de l'analyse de dispersion

- Étudier les liens entre des séquences d'états et des variables explicatives.
- Analyse directe, sans clustering préalable.
- Est-ce que les trajectoires diffèrent selon le statut professionnelle du père (sans emploi) ?
- Ces différences sont-elles significatives ?



Principes généraux

- Définition d'une mesure de la dispersion des séquences sur la base de la matrice des distances.
- Mesure la force de l'association à l'aide de la part de cette dispersion expliquée par un facteur explicatif.
- Atteste la significativité de l'association à l'aide de tests de permutation.
- Cette méthode est une généralisation l'ANOVA.
- Elle peut être étendue pour inclure plusieurs facteurs explicatifs
- ou pour construire des arbres de régressions.

Origine et domaines d'application de la méthode

- Premiers développements par Mielke and Berry (1983)
- Méthodes présentées ici : Anderson (2001); McArdle and Anderson (2001)
- Domaines d'applications :
  - L'écologie.
  - La génétique.
  - L'analyse de courbe de Lorenz.
  - Images cérébrales.
  - Quand il y a plus de variables dépendantes que d'individus.

Principes

- Dans le cas euclidien :

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

- En remplaçant d_{ij} avec une autre distance (OM, ...), on définit une mesure de dispersion d'un ensemble de séquences.

Analyse de dispersion des séquences

- La définition de la dispersion permet d'utiliser le cadre d'analyse ANOVA.
- Basé sur la décomposition de la somme des carrés (SC) (ou inertie)

$$SC_{tot} = SC_{inter} + SC_{intra}$$

- On peut ensuite calculer un Pseudo-R^2 pour mesurer la force de l'association.
- R^2 est la part de la dispersion totale expliquée par une variable.

$$R^2 = \frac{SC_{inter}}{SC_{tot}}$$

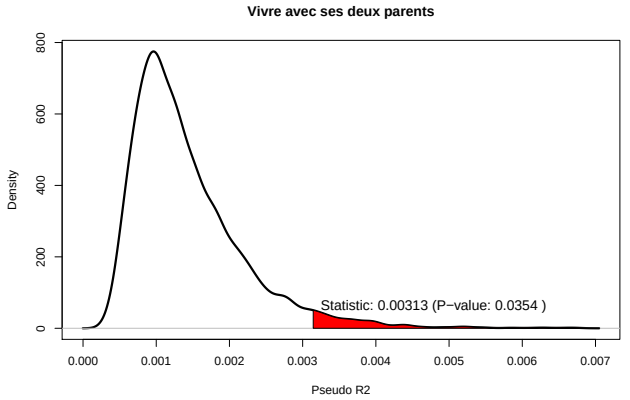
- Pseudo F

$$F = \frac{SC_{inter}/(m-1)}{SC_{intra}/(n-m)}$$

Significativité

- Estimation de la significativité à l'aide de test de permutations.
- Tests de permutations :
  - Estimation de la probabilité qu'une partition aléatoire explique une plus grande part de la dispersion.
  - Calcul P (grand) valeurs du R^2 pour des partitions aléatoires R^2_{perm}.
  - Estimation de la p-valeur : $p(R^2_{obs} > R^2_{perm})$.
  - Valeurs de P (Manly, 2007) :
    - 5000 permutations pour un seuil de 1% .
    - 1000 pour un seuil de 5%.

Distribution nulle du R^2



Analyse de dispersion

- Association bivariée avec chaque variable explicative.

	F	R^2	p-value
test	71.56	0.092	0.000
gcse5eq	67.69	0.087	0.000
Grammar	23.16	0.032	0.000
region	5.50	0.030	0.000
funemp	9.51	0.013	0.000
fmpr	8.76	0.012	0.000
sex	6.84	0.010	0.000
religion	2.75	0.004	0.013
livboth	2.23	0.003	0.035

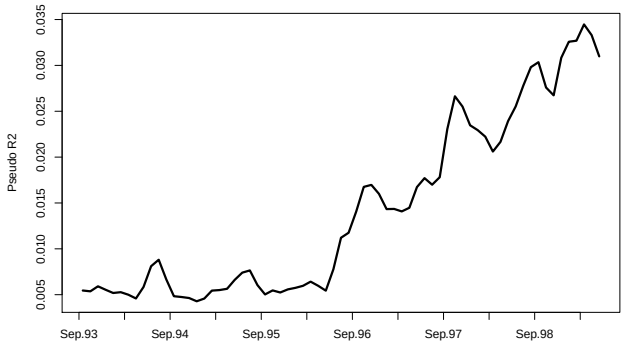
L'analyse de dispersion permet ainsi de mesurer :

- La force de la relation à l'aide du R^2.
- La significativité de la relation avec des tests de permutations.

Évolution de l'association

- Comment est-ce que les différences entre groupes varient au cours du temps ?
- A quel moment les trajectoires diffèrent le plus ?
- Calcul du R^2 pour une période de temps.
- Ici de longueur 6.
- Nous obtenons une séquence de R^2 qui peut être représentée graphiquement.

Évolution de l'association



Analyse de dispersion multifacteur

- Généralisation de l'approche précédente.
- Inclusion de plusieurs facteurs explicatifs en même temps.
- L'apport d'une variable est calculé en considérant la contribution additionnelle de chaque covariable lorsqu'on a déjà pris en compte l'effet de toutes les autres (effet de Type III).

Mise en oeuvre d'une analyse multifacteur.

	F	R^2	p-value
male	3.55	0.004	0.002
Grammar	11.07	0.014	0.001
funemp	5.33	0.007	0.001
gcse5eq	46.00	0.058	0.001
fmpr	1.90	0.002	0.042
livboth	1.50	0.002	0.131
Total	15.76	0.118	0.001

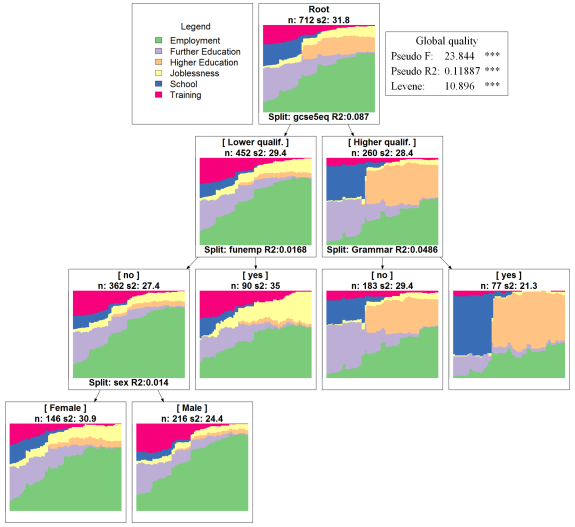
Conclusion de l'analyse multifacteur

- L'ensemble du modèle est statistiquement significatif
- Il permet d'expliquer 12% de la dispersion totale.
- La variable livboth n'est plus statistiquement significative dès lors que l'on contrôle pour les autres variables.
- La variable gcse5eq permet d'expliquer 5.8% de la dispersion une fois l'effet des autres variables pris en considération.

Analyse par arbre de régression

- But : Identifier les variables les plus importantes et leurs interactions.
- Représentation graphique.
- fonctionnement :
  - Segmente récursivement les individus en utilisant les valeurs des variables
  - de manière à ce que les groupes soient aussi homogènes que possible.
  - A chaque pas, on choisit la partition avec le R^2 le plus élevé.
  - La significativité d'un éclatement est attestée à l'aide de tests de permutations.
  - On arrête de grandir l'arbre lorsque l'association n'est plus significative.

Arbre de régression



Analyse de dispersion

- L'analyse de dispersion permet d'estimer :
  - La force de la relation à l'aide d'un R^2.
  - La significativité de la relation avec des tests de permutations.
- Les tests sont plus puissants qu'en passant par les clusters.
- Analyse direct des trajectoires sans hypothèses sur les « modèles de trajectoires ».
- Apporte une approche explicative à l'analyse de séquences.
- Prise en compte de la disparité des séquences (donc la variabilité inter-individuelle) tout en étudiant la relation des trajectoires avec leurs contextes.

Homogénéité des dispersions

- Le test d'homogénéité des dispersions permet de tester l'égalité des dispersions intra-groupes.
- Est-ce que la dispersion diffère significativement d'un groupe à l'autre ?
- Basé sur une généralisation du test de Levene.
- La significativité est attestée à l'aide de tests de permutation.
- Exemple d'utilisation : standardisation des trajectoires.

Dispersion d'un ensemble de séquences : interprétation

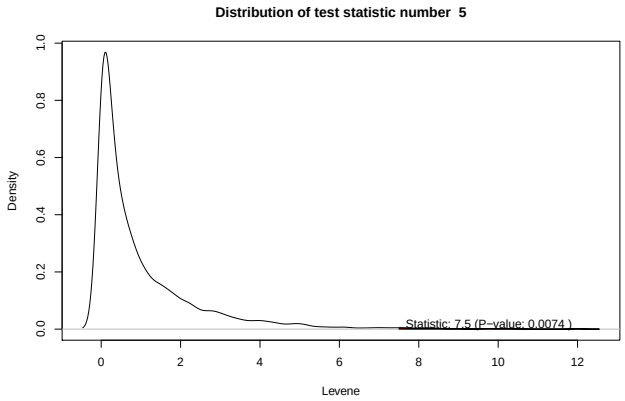
- Interprétation de la dispersion.
  - Mesure de la variabilité inter-individuelle des trajectoires.
  - Peut s'interpréter comme une mesure de l'incertitude quant à la trajectoire suivie et peut donc refléter
  - une forme de précarité
  - ou une multiplicité de choix auxquels l'individu doit faire face.
  - Selon si l'incertitude est "subie" ou "agie".
  - Ce n'est pas un concept nécessairement négatif !
- La dispersion est différente des mesures de complexité :
  - La complexité mesure la variabilité intra-trajectoire.
  - La dispersion mesure la variabilité inter-trajectoire.

Exemple :

- Différence de dispersion selon le statut professionnel du père.

	n	discrepancy
no	595.00	30.61
yes	117.00	35.27
Total	712.00	31.80

Distribution nulle de Levene

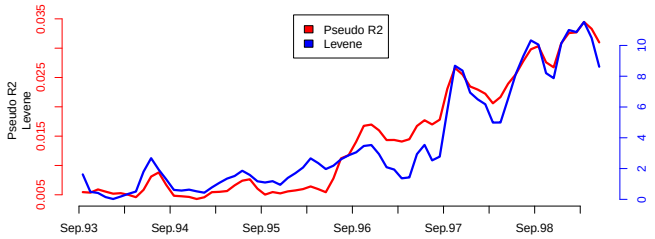


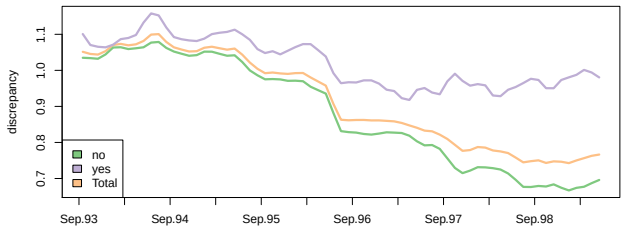
- La dispersion des trajectoires diffère significativement (Levene = 7.5 ; p-valeur=0.0074).

Différences au cours du temps

- Comme précédemment, on peut représenter l'évolution de
  - La statistique de Levene (homogénéité des dispersions).
  - La dispersion groupe totale.
  - La dispersion dans chaque groupe.

Évolution de la statistique de Levene





Analyse de dispersion dans `TraMineR`

Commande	Analyses
`dissassoc`	Tests bivariés, test d'homogénéité des dispersions, dispersion totale, dispersion par groupe
`dissmfacw`	Analyse multifacteur
`seqtree`	Analyse par arbre pour des séquences
`seqtreedisplay`	Représentation graphique de l'arbre. Nécessite l'installation de GraphViz.
`seqdiff`	Évolution de l'association.

Conclusion

- L'analyse en cluster offre un point de vue descriptif et exploratoire sur les séquences.
- Elle ne devrait pas être utilisée pour faire de l'inférence.
- L'analyse de dispersion permet d'estimer :
  - La force de la relation à l'aide d'un R^2.
  - La significativité de la relation avec des tests de permutations.
- Analyse direct des trajectoires sans hypothèses sur les « modèles de trajectoires ».
- Prise en compte de la disparité des séquences (donc la variabilité inter-individuelle) tout en étudiant la relation des trajectoires avec leurs contextes.
- Approches multifacteurs.
- Comparaison des dispersions.

References I

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology 26*, 32–46.

Billari, F. C. (2001). The analysis of early life courses : Complex description of the transition to adulthood. *Journal of Population Research 18*(2), 119–142.

Manly, B. F. J. (2007). *Randomization, Bootstrap and Monte Carlo Methods in Biology* (Third Edition ed.). New York : Chapman & Hall.

McArdle, B. H. and M. J. Anderson (2001). Fitting multivariate models to community data : A comment on distance-based redundancy analysis. *Ecology 82*(1), 290–297.

References II

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.

Mielke, P. W. and K. J. Berry (1983). Asymptotic clarifications, generalizations, and concerns regarding an extended class of matched pairs tests based on powers of ranks. *Psychometrika 48*(3), 483–485.

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

References III

Studer, M., G. Ritschard, A. Gabadinho, and N. S. Müller (2010). Discrepancy analysis of complex objects using dissimilarities. In F. Guillet, G. Ritschard, D. A. Zighed, and H. Briand (Eds.), *Advances in Knowledge Discovery and Management*, Volume 292 of *Studies in Computational Intelligence*, pp. 3–19. Berlin : Springer.

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