Classer, discriminer et visualiser des séquences d'événements

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

*Institut d'études démographiques et des parcours de vie, Université de Genève
{matthias.studer, nicolas.muller, gilbert.ritschard, alexis.gabadinho}@unige.ch,
http://www.unige.ch/ses/metri/

Résumé. Cet article [1] présente un ensemble d'outils destiné à analyser des séquences d'événements en sciences sociales et à visualiser les résultats obtenus. Nous commençons par formaliser la notion de séquence d'événements avant de définir une mesure de dissimilarité entre ces séquences afin de construire des typologies et de tester les liens entre ces séquences et d'autres variables d'intérêts. Initialement définie par Moen (2000), cette mesure se base sur la notion de distance d'édition entre séquences et permet d'identifier les différences d'ordonnancement et de temporalité des événements. Nous proposons une extension de celle-ci afin de pouvoir prendre en compte la simultanéité des événements ainsi qu'une méthode de normalisation qui garantit le respect de l'inégalité triangulaire. Dans un deuxième temps, nous présentons un ensemble d'outils destinés à interpréter les résultats. Nous proposons ainsi deux méthodes de visualisation d'un ensemble de séquences et nous introduisons la notion de sous-séquence discriminante qui permet d'identifier les différences d'ordonnancement des événements les plus significatives entre groupes. L'ensemble des outils présentés est disponible au sein de la librairie R TraMineR.

1 Introduction

L'analyse de séquences en sciences sociales se centre essentiellement sur deux types de questions. Premièrement, dans une optique descriptive, on cherche à construire une typologie des séquences, ce que l'on réalise généralement à l'aide d'une classification automatique. Deuxièmement, on s'intéresse à comprendre les liens entre ces séquences et d'autres variables d'intérêt afin de mettre en relation ces séquences avec leur contexte. Il existe déjà plusieurs méthodes pour répondre à ces questions dans le cas de séquences d'états. Dans cet article, nous développons un ensemble de méthodes destinées à analyser des séquences d'événements. Toutes ces méthodes sont disponibles au sein de la librairie R TraMineR (Gabadinho et al., 2009).

La méthode généralement utilisée en sciences sociales consiste à recoder les séquences d'événements sous forme de séquence d'états puis d'utiliser une mesure de dissimilarité dérivée des méthodes d'alignement de séquences (Abbott, 1990). Cette méthode a deux désavantages. Premièrement, les méthodes d'alignement de séquences ne prennent qu'indirectement en

1. Etude réalisée avec le soutien financier du Fonds national suisse (FNS), subside 100015-122230.

compte la structure temporelle des séquences (Wu, 2000; Levine, 2000) et ne permettent pas, par exemple, l'opération consistant à déplacer dans le temps une transition entre deux états. Deuxièmement, le nombre d'états nécessaire pour représenter sans perte d'information une séquence d'événements est grand (2^n) ce qui rend vite l'interprétation des résultats difficile.

Nous commençons par introduire une extension de la dissimilarité entre séquences d'événements conçue par Moen (2000). Cette mesure est basée sur la notion de distance d'édition et peut être définie comme le coût minimum nécessaire pour transformer une séquence en une autre à l'aide de deux opérations : l'insertion d'événements et le déplacement de groupes d'événements simultanés. De plus, notre extension nous permettra de proposer une normalisation des dissimilarités — nécessaire puisque les séquences contiennent généralement un nombre d'événements variable — qui respecte l'inégalité triangulaire. Cette dernière propriété nous permettra d'utiliser ces dissimilarités dans une procédure de classification automatique ou encore pour mesurer les liens avec d'autres variables d'intérêt (Studer et al., 2009).

Nous introduisons ensuite un ensemble de méthodes destiné à caractériser les différences entre les groupes obtenus. Cette étape doit également permettre de donner une interprétation aux liens statistiquement significatifs avec d'autres variables. Pour ce faire, nous introduisons la notion de sous-séquence discriminante définie comme les sous-séquences fréquentes dont la présence permet de discriminer l'appartenance à un groupe spécifié. Nous introduisons également une méthode de visualisation d'un ensemble de séquences d'événements.

L'article est organisé de la manière suivante. Nous commençons par donner une courte introduction à l'étude des trajectoires familiales qui nous servira d'exemple d'application tout au long de cet article. Nous formalisons ensuite la notion de séquence d'événements en reprenant les termes et les concepts utilisés en sciences sociales, notre principal domaine d'application. Sur cette base, nous présentons ensuite la méthode de calcul de dissimilarité entre séquences d'événements ainsi que les extensions disponibles dans TraMineR. Finalement, nous introduisons les différentes méthodes de visualisation utilisées pour qualifier les différences entre groupes de séquences.

2 Présentation des données

Nous utilisons les données de l'enquête biographique rétrospective réalisée par le Panel suisse de ménages [2] en 2002. Dans le cadre de cette étude, nous retenons cinq événements constitutifs de la vie familiale : le départ du foyer parental, la première cohabitation en couple, le premier mariage, le premier divorce et la naissance du premier enfant. Afin de comparer des séquences similaires, nous n'avons retenu que les répondants ayant 45 ans lors de l'enquête, soit 2601 individus. Ce choix nous permettra de comparer les résultats de l'analyse directe des séquences d'événements avec ceux obtenus dans (Müller et al., 2008) en passant par le recodage sous forme de séquences d'états. L'objectif de l'étude est de mettre en évidences les différents types de parcours familiaux et les changements que ceux-ci ont connus au cours du $20^{\text{ème}}$ siècle, notamment concernant le mariage.

2. http ://www.swisspanel.ch

3 Formalisation des séquences d'événements

Dans cette section, nous présentons une formalisation de la notion de séquence qui permet de prendre compte la temporalité des événements. Cette formalisation se base sur une adaptation à la terminologie sociologique de la formalisation généralement utilisée en data-mining et présentée notamment par Agrawal et Srikant (1995) et Zaki (2001).

Notre optique consiste à analyser les trajectoires comme une succession ordonnée de changements, de transitions et d'événements. Dès lors, nous définissons une *séquence* comme une liste *ordonnée* de *transitions*. Une séquence X est notée $(\tau_1 \rightarrow \tau_2 \rightarrow \ldots \rightarrow \tau_q)$, où chaque τ_i désigne une transition. Les transitions peuvent être formalisées comme une liste *non ordonnée* d'*événements* distincts. Autrement dit, une transition correspond à la réalisation d'un ou de plusieurs événements simultanés, chaque événement ne pouvant se produire qu'une seule fois dans une même transition. Une transition τ_i est notée $(\epsilon_1, \epsilon_2, \ldots, \epsilon_r)$, ou chaque ϵ_i désigne un événement distinct. La distinction entre transition et événement permet de rendre compte de la simultanéité de certains événements. L'ensemble des événements pouvant se réaliser constitue l'alphabet Σ^* des séquences d'événements. Prenons un exemple.

$$(\text{Départ, Couple}) \rightarrow (\text{Mariage}) \rightarrow (\text{Enfant}) \tag{1}$$

Cette séquence décrit le parcours d'une personne qui quitterait ses parents pour se mettre en couple avant de se marier puis d'avoir un premier enfant. Selon notre formalisation, "Départ", "Couple", "Mariage" et "Enfant" sont des événements alors que (Départ, Couple) dénote une *transition*, contenant deux événements distincts, entre les états "Encore chez les parents et sans partenaire" et "Parti et avec un partenaire".

Afin de garantir l'unicité de la représentation d'une transition, nous répertorions les événements qui la composent dans un ordre constant prédéfini. Sans ce point, la séquence (1) pourrait également s'écrire (Couple, Départ) $\rightarrow$ (Mariage) $\rightarrow$ (Enfant).

Cette formalisation laisse cependant de côté la temporalité des événements, alors qu'il s'agit d'une notion centrale à l'étude des séquences d'événements en sciences sociales. Nous proposons de la compléter en datant les transitions et en mesurant les durées qui les séparent. Nous utilisons un exposant pour indiquer l'âge auquel chaque transition se produit. Ainsi, une transition τ_1 se produisant à l'instant t s'écrira τ_1^t. A titre d'exemple, la séquence 1 pourrait être enrichie de la manière suivante :

$$(\text{Départ, Couple})^{23} \rightarrow (\text{Mariage})^{29} \rightarrow (\text{Enfant})^{30} \tag{2}$$

Dans certaines situations, on préférera des *écarts* relatifs entre événements et transitions plutôt que des âges absolus. Dans notre exemple, on peut ainsi noter que le premier enfant survient à trente ans ou qu'il survient un an après le mariage. Nous introduisons une notation basée sur les écarts à l'aide d'un exposant au dessus de la flèche $\rightarrow$. Afin de disposer de l'ensemble des informations, il est nécessaire d'inclure l'écart à l'origine (âge 0) pour la première transition. Notons que les deux notations, sous forme d'âge absolu ou d'écarts, comportent la même information et que l'une ou l'autre doit être utilisée en fonction de la perspective selon laquelle on entend aborder une trajectoire. Ainsi, nous pourrions réécrire la séquence 2 de la manière suivante.

$$\xrightarrow{23} (\text{Départ, Couple}) \xrightarrow{6} (\text{Mariage}) \xrightarrow{1} (\text{Enfant}) \xrightarrow{15} \tag{3}$$

De manière similaire, nous pouvons inclure l'écart jusqu'à la fin de l'observation de la séquence. Ce formalisme permet de prendre en compte les données censurées. En effet, en sciences sociales, il est fréquent que les durées d'observation de chaque individu diffèrent. Notons qu'en interne TraMineR utilise une représentation basée sur les écarts.

4 Distance entre séquences d'événements

La définition d'une mesure de dissimilarité permet de construire une typologie des séquences à l'aide d'une procédure de classification hiérarchique. Elle permet aussi de tester le lien avec d'autres variables d'intérêts. Cette mesure de dissimilarité doit permettre de comparer les séquences en fonction de l'ordonnancement des événements et de leur temporalité. Dans cette section, nous construisons, sur la base de la mesure établie par Pirjo Moen (Moen, 2000), une distance entre séquences d'événements.

Dans sa thèse de doctorat, Moen (2000) propose d'établir une mesure de dissimilarité entre séquences de transitions [3] en se basant sur la notion de distance d'édition. La dissimilarité entre deux objets est alors définie comme le coût minimum nécessaire à la transformation d'une séquence en une autre en considérant deux types d'opérations : l'insertion d'une transition et son déplacement temporel. Cette mesure prend ainsi deux paramètres : un tableau spécifiant le coût d'insertion de chaque type de transition et le coût (constant) associé au déplacement d'une transition d'une unité de temps. L'approche ressemble ainsi fortement à l'alignement de séquences, procédure souvent utilisée en sciences sociales pour l'analyse de séquences d'états. Elle permet toutefois d'inclure des opérations propres aux séquences d'événements, tels que le déplacement temporel d'une transition.

Cette dissimilarité présente plusieurs inconvénients. Premièrement, elle ne permet pas de prendre en compte les événements simultanés, alors qu'ils sont courants en sciences sociales où les processus sont souvent mesurés à l'échelle de l'année. Deuxièmement, la normalisation des distances en fonction du coût maximal peut remettre en cause l'inégalité triangulaire.

L'extension que nous présentons permet de se baser sur la notion d'événement plutôt que celle de transition. Deux opérations sont considérées : l'insertion d'événements et le déplacement temporel de transitions. Dès lors, deux transitions comportant des événements en commun pourront être considérées comme similaires. Notre formulation est générale et en l'absence d'événements simultanés, correspond à l'algorithme de Moen (2000). De plus, notre formulation permet de proposer une normalisation qui respecte l'inégalité triangulaire.

4.1 Définition

La distance entre deux séquences d'événements est définie comme le coût minimum nécessaire pour transformer une séquence en une autre à l'aide de deux opérations : le déplacement temporel d'une transition et l'insertion d'un événement dans une séquence. La métrique dépend ainsi de deux paramètres : δ le coût associé au déplacement d'une transition d'une unité de temps et $\omega(\epsilon)$ le coût associé à l'insertion-suppression d'un événement de type ϵ. Moen (2000) propose de faire dépendre ce dernier coût de la fréquence d'apparition de l'événement ϵ.

3. Pirjo Moen parle de séquence d'événements. Toutefois, sa définition de l'événement suppose la non-simultanéité des événements et sa proposition de considérer les combinaisons d'événements simultanés comme des événements distincts correspond à la définition de transition que nous avons introduite dans notre formalisation.

Le calcul de cette distance peut être réalisé en redéfinissant les coûts associés aux deux opérations de base définies dans la distance de Levenshtein généralisée (Yujian et Bo, 2007) : l'insertion-suppression et la substitution d'une transition. Soit λ l'élément nul que nous ajoutons à l'alphabet Σ^* et $\Omega(A)$ la somme des coûts ω d'insertion de l'ensemble des événements appartenant à A, A étant une transition ou une séquence :

$$\Omega(A) = \sum_{\epsilon_i \in A} \omega(\epsilon_i) \tag{4}$$

Nous notons $\gamma(\lambda \rightarrow \tau^t)$ le coût d'insertion d'une transition τ à l'instant t et $\gamma(\tau^t \rightarrow \lambda)$ celui d'une suppression. Le coût de cette opération dépend des événements inclus dans la transition et est défini comme la somme des coûts de leur insertion-suppression, soit :

$$\gamma(\lambda \rightarrow \tau^t) = \Omega(\tau^t) = \sum_{\epsilon_i \in \tau^t} \omega(\epsilon_i) \tag{5}$$

Le coût de la substitution d'une transition $\tau_1^{t_1}$ par $\tau_2^{t_2}$ est notée $\gamma(\tau_1^{t_1} \rightarrow \tau_2^{t_2})$. Soit $\Delta(\tau_1^{t_1} \rightarrow \tau_2^{t_2})$, le coût associé au déplacement de la transition de t_1 à t_2 plus le coût nécessaire pour transformer τ_1 en τ_2 :

$$\Delta(\tau_1^{t_1} \rightarrow \tau_2^{t_2}) = \delta|t_1 - t_2| + \Omega(\tau_1) + \Omega(\tau_2) - 2 \cdot \Omega(\tau_1 \cap \tau_2) \tag{6}$$

$$\gamma(\tau_1^{t_1} \rightarrow \tau_2^{t_2}) = \min\{\Delta(\tau_1^{t_1} \rightarrow \tau_2^{t_2}); \gamma(\tau_1^{t_1} \rightarrow \lambda) + \gamma(\lambda \rightarrow \tau_2^{t_2})\} \tag{7}$$

Soit $T_{X,Y}$ un ensemble d'opérations T_i permettant de transformer la séquence X en Y, la *distance* $d(X, Y)$ est définie comme :

$$d(X, Y) = \min\{\sum_{T_I \in T_{X,Y}} \gamma(T_i)\} \tag{8}$$

Cette distance peut être calculée à l'aide d'un algorithme de programmation dynamique classique. Remarquons qu'il s'agit d'une distance au sens mathématique du terme. En effet, la matrice des opérations de base γ ainsi définie respecte les conditions mathématiques de la distance, ce qui est une condition suffisante pour assurer que l'algorithme produise une distance (Yujian et Bo, 2007). Il est toutefois nécessaire que les paramètres respectent les conditions suivantes : $\omega(\epsilon) > 0$ et $\delta \geq 0$.

A titre d'exemple, prenons les deux séquences X et Y suivantes :

$$\begin{aligned} X &= \text{(Naissance)} \xrightarrow{20} \text{(Couple, Depart)} \xrightarrow{4} \text{(Mariage)} \xrightarrow{21} \\ Y &= \text{(Naissance)} \xrightarrow{19} \text{(Couple, Depart, Mariage)} \xrightarrow{26} \end{aligned}$$

La série de transformation associée au coût minimal avec un coût d'insertion constant $\omega = 1$ et un coût de déplacement $\delta = 0.2$ est :

$$\begin{aligned} d(X, Y) &= \gamma((\text{Naissance})^0 \rightarrow (\text{Naissance})^0); \\ &\quad \gamma((\text{Couple, Depart})^{20} \rightarrow (\text{Couple, Depart, Mariage})^{19}); \\ &\quad \gamma((\text{Mariage})^{24} \rightarrow \lambda) \\ &= 0 + 1.2 + 1 = 2.2 \end{aligned}$$

4.2 Normalisation

En sciences sociales, les séquences d'événements sont souvent de longueurs différentes, même pour des temps d'observations identiques. Certaines personnes connaissent des trajectoires plus turbulentes que d'autres. Pour pallier ce problème qui affecte divers domaines, il est nécessaire de normaliser les distances. Sans normalisation, une insertion dans la comparaison de deux séquences comportant mille événements conduira à la même distance que si les séquences en comportent deux. Moen (2000) propose de diviser chaque distance par le coût maximum, mais cette normalisation peut entraîner le non-respect de l'inégalité triangulaire. Ceci est notamment problématique pour calculer l'association des séquences avec d'autres variables sur la base des dissimilarités (Studer et al., 2009).

Yujian et Bo (2007) proposent une méthode de normalisation qui maintient l'inégalité triangulaire pour la distance de Levenshtein généralisée à coût d'insertion constant. Nous proposons une adaptation de celle-ci à notre métrique basée sur des coûts d'insertion variables (mais sans substitution). Yujian et Bo (2007) montrent qu'une telle normalisation peut être obtenue en définissant une mesure de similarité s qui satisfait les conditions suivantes :

$$s(Y,Y) + s(X,Z) \geq s(X,Y) + s(Y,Z) \tag{9}$$

$$0 \leq s(X,Y) \leq \min\{s(X,X), s(Y,Y)\} \tag{10}$$

Sur la base de la distance introduite, nous définissons la mesure de similarité suivante.

$$s(X,Y) = \frac{\Omega(X) + \Omega(Y) - d(X,Y)}{2} \tag{11}$$

Cette mesure de similarité satisfait la relation (9), ce qui peut être montré de la manière suivante en notant que $s(Y,Y) = \Omega(Y)$:

$$\begin{aligned} d(X,Z) &\leq d(X,Y) + d(Y,Z) \\ \Omega(X) + \Omega(Z) - 2s(X,Z) &\leq \Omega(X) + \Omega(Y) - 2s(X,Y) + \Omega(Y) + \Omega(Z) - 2s(Y,Z) \\ s(X,Y) + s(Y,Z) &\leq \Omega(Y) + s(X,Z) \end{aligned}$$

La relation (10) est également respectée. En effet, $d(X,Y)$ est nécessairement supérieur ou égal à la somme des insertions et suppressions nécessaires pour transformer X en Y sans considération d'ordre ni de temporalité (qui ne peuvent qu'augmenter la distance), ce qui est égal à $\Omega(X) + \Omega(Y) - 2 \cdot \Omega(X \cap Y)$. Ainsi, $s(X,Y) \leq \Omega(X \cap Y) \leq \min\{\Omega(X), \Omega(Y)\} = \min\{s(X,X), s(Y,Y)\}$. De plus, comme $d(X,Y) \leq \Omega(X) + \Omega(Y)$, on a $s(X,Y) \geq 0$.

La distance normalisée est alors définie comme :

$$d_N(X,Y) = \frac{s(X,X) + s(Y,Y) - 2s(X,Y)}{s(X,X) + s(Y,Y) - s(X,Y)} = \frac{2d(X,Y)}{\Omega(X) + \Omega(Y) + d(X,Y)} \tag{12}$$

4.3 Prise en compte du temps

La définition du temps auquel les transitions ont lieu n'est pas triviale. Pirjo Moen propose de prendre l'âge absolu comme référence temporelle. Toutefois, comme nous l'avons vu dans notre formalisation, il est également possible de raisonner en terme d'écarts à la transition

précédente. Dans le premier cas, un décalage au début peut se répercuter sur l'ensemble des événements suivants si les écarts sont les mêmes, ce qui est évité avec la deuxième solution. Le calcul sur la base des âges mettra en avant la temporalité absolue des transitions alors que les écarts permettent de prendre en compte la dynamique. Il est également possible d'utiliser l'écart à la transition suivante ou à la fin de l'observation de la séquence pour la dernière transition. Dans ce dernier cas, on mettra en avant le fait qu'une transition influence une période de longueur déterminée.

Toutes ces manières de prendre en considération la temporalité des séquences d'événements entrent dans le cadre de la mesure de distance que nous avons présentée. Le choix de l'une ou l'autre des solutions devrait être effectué en fonction de la problématique considérée.

4.4 Application

Nous proposons à présent de mettre en oeuvre cette mesure de dissimilarité sur l'exemple que nous avons présenté. Nous avons utilisé un coût δ de 0.1 et un coût d'insertion-délétion constant égal à 0.5. Cette définition implique qu'un déplacement de dix ans correspond à l'insertion et la suppression d'un événement. Dans notre logique, à dix ans d'intervalle, la signification sociale d'un événement change tellement qu'il est nécessaire de le considérer comme différent. Nous avons choisi de calculer les distances en fonction des âges absolus.

A l'aide des distances obtenues, on peut procéder à une classification automatique en utilisant diverses procédures. Étant donné que TraMineR est un package R, l'ensemble des procédures disponibles dans R peut être utilisé. Nous avons ici créé trois groupes à l'aide de l'algorithme PAM (Partitioning Around Medoids) (Maechler et al., 2005), une extension des nuées dynamiques à tout type de dissimilarités. Une fois les groupes construits, il est possible de se faire une idée de leurs contenus à l'aide de leur centroïde. Le tableau1 présente le nombre d'individus, la dispersion calculée sur la base des dissimilarités (Studer et al., 2009) ainsi que le centroide de chaque groupe obtenu.

Groupe	N	Dispersion	Centroïde
1	1256	0.236	(Naissance) $\xrightarrow{24}$ (Couple, Depart, Mariage) $\xrightarrow{21}$
2	973	0.275	(Naissance) $\xrightarrow{20}$ (Depart) $\xrightarrow{6}$ (Couple, Mariage) $\xrightarrow{2}$ (Enfant) $\xrightarrow{17}$
3	372	0.262	(Naissance) $\xrightarrow{24}$ (Mariage) $\xrightarrow{21}$
Total	2601	0.311	

TAB. 1 – *Groupes obtenus*

A l'aide de ces dissimilarités, il est également possible de mesurer le lien entre les séquences d'événements représentées par les dissimilarités et d'autres variables d'intérêts. Nous pouvons ainsi affirmer que la relation entre les séquences et la cohorte de naissance est statistiquement significative. Autrement dit, les modèles de construction de la vie familiale de ceux qui sont nés après 1949 sont différents de ceux nés avant.

Toutefois, en l'état, la procédure de classification automatique reste d'un intérêt limité : il nous manque des méthodes pour décrire ou visualiser les groupes obtenus. L'utilisation des centroides est trop réductrice pour pouvoir interpréter les groupes. Il en va de même pour la mesure d'association. Il est nécessaire de décrire en quoi les cohortes diffèrent pour donner du

sens à la relation. Dans le prochain chapitre, nous abordons des méthodes destinées à visualiser un ensemble de séquences et à identifier les différences entre groupes de séquences.

5 Visualisation et caractérisation de groupes de séquences

Dans cette section, nous décrivons de nouvelles méthodes pour visualiser des ensembles de séquences d'événements. Nous abordons également, dans un deuxième temps, la localisation de différences entre groupes de séquences. Ces méthodes nous permettront d'interpréter les résultats obtenus précédemment.

5.1 Visualisation et description d'un ensemble de séquence

Il est possible de décrire un ensemble de séquences à l'aide des sous-séquences les plus fréquentes dans cet ensemble (Agrawal et Srikant, 1995; Zaki, 2001). Une sous-séquence d'une séquence donnée x peut être définie comme une séquence d'événements composée d'un sous-ensemble des événements de x et qui respecte l'ordre des événements de x. Par exemple, (Départ, Couple) $\rightarrow$ (Enfant) est une sous-séquence de la séquence (1) puisque l'ordre des transitions et des événements est respecté. Le support d'une sous-séquence correspond au nombre de séquences dans lesquelles cette sous-séquence apparaît. Une sous-séquence est dite *fréquente* si son support dépasse un certain seuil appelé le support minimum.

TraMineR permet de rechercher les sous-séquences fréquentes à l'aide d'un algorithme simple, basé sur les arbres de préfixes (Masseglia, 2002). Cette recherche peut être contrôlée à l'aide de contraintes temporelles. Ainsi, il est possible de spécifier une fenêtre maximum (temps maximal pendant lequel une sous-séquence peut se produire) et le temps maximum séparant deux transitions. Un âge minimum et maximum peuvent également être spécifié afin de se concentrer sur l'étude de tranches d'âges spécifiques. La figure 1 présente les vingt sous-séquences les plus fréquentes dans nos séquences sans aucune contrainte temporelle.

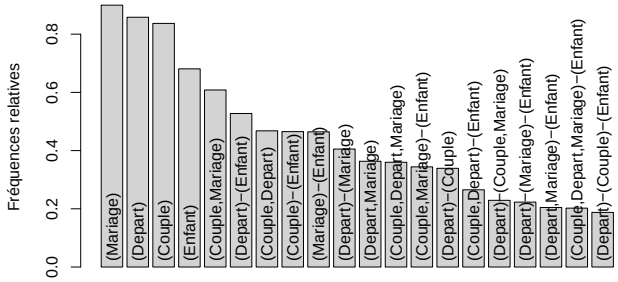

FIG. 1 – *Les vingt sous-séquences les plus fréquentes.*

Cette représentation permet d'observer les enchaînements les plus fréquents, mais sa lecture peut s'avérer difficile. Ce n'est pas parce qu'une sous-séquence est fréquente qu'elle est intéressante et le nombre de sous-séquences à afficher est difficile à déterminer. De plus, certaines sous-séquences peuvent être redondantes. Finalement, cette description ne prend pas en compte la temporalité des événements, hormis si l'on spécifie des contraintes temporelles.

La figure 2 propose deux représentations d'un ensemble de séquences. Le premier graphique montre l'évolution du nombre moyen d'événements par individu pour chaque type d'événement (l'événement "Naissance" a été omis pour plus de clarté). Ce graphique permet de mettre en lumière les régularités temporelles de certains événements. Ainsi, le départ connaît un pic aux alentours de vingt ans alors que l'arrivée du premier enfant se concentre entre 25 et 30 ans. Cette représentation tient compte des apparitions multiples des mêmes événements dans une même séquence.

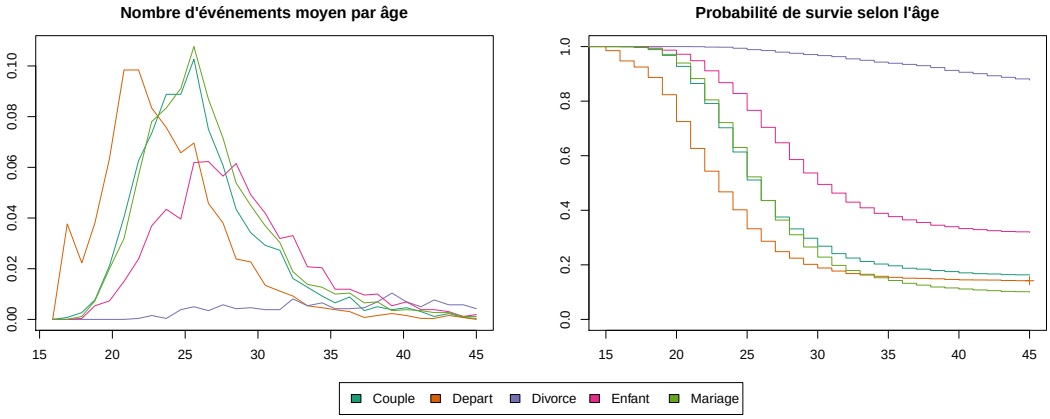

FIG. 2 – *Représentation d'un ensemble de séquences.*

Le graphique de droite affiche les courbes de survie $S(t)$ jusqu'à la première occurrence de chaque événement. $S(t) = P(T \geq t)$ donne la probabilité que la durée d'attente T jusqu'à l'événement soit au moins t. Ce graphique à l'avantage de montrer la part de la population qui connaît chaque événement. Toutefois, il n'est pas possible d'inclure les événements récurrents.

5.2 Différencier des groupes de séquences

Les méthodes présentées dans la section précédente permettent de représenter un ensemble de séquences. Elles s'avèrent notamment utiles pour explorer les différences entre des groupes de séquences. La figure 3 présente les courbes de survie dans chaque groupe obtenu par la procédure de classification automatique.

Il est ainsi possible de donner une interprétation plus précise des groupes obtenus. Les groupes reprennent trois logiques connues des parcours de cohabitation suisses. Le premier correspond à l'ancien modèle qui consiste à quitter le foyer parental quand on se marie. Le deuxième groupe reprend le modèle moderne de construction de la vie familiale qui se distingue par l'existence d'une période "seule". Finalement, le dernier groupe reprend les individus qui n'ont jamais quitté leurs parents (éventuellement en se mariant).

La méthode traditionnellement employée, qui exploite un recodage sous forme de séquences d'états (Müller et al., 2008), favorise un regroupement selon le temps passé au sein de chaque état (défini par la combinaison des événements vécus). La dissimilarité entre séquences d'événements proposée ici donne un poids plus grand aux différences de séquencement, même

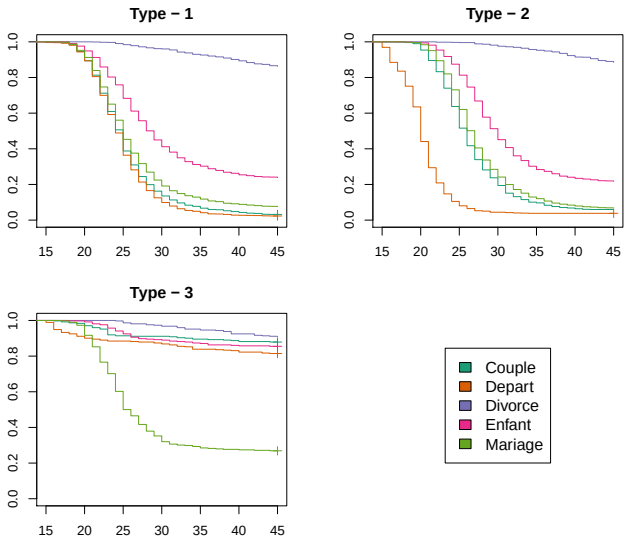

FIG. 3 – *Courbes de survie par groupe obtenu.*

si les écarts entre événements sont petits. Avec la première solution, les individus qui passent un grand nombre d'années dans l'état marié avec enfant sont regroupés, peu importe le processus qui les a conduits à connaître ou à quitter cet état. La première méthode se centre sur les états occupés, alors que celle que nous proposons se centre sur les événements et correspond ainsi mieux à l'étude des processus représentés sous la forme de succession d'événements.

Il serait possible de décrire les groupes à l'aide des sous-séquences les plus fréquentes. Toutefois, un tel graphique n'apporte généralement que peu d'information. En effet, une sous-séquence très fréquente peut l'être dans chacun des groupes comparés. Autrement dit, les groupes peuvent se distinguer sur des sous-séquences relativement peu fréquentes.

Afin de pallier ce problème, nous proposons d'ordonner les sous-séquences en fonction de leurs capacités à discriminer l'appartenance à un groupe plutôt qu'à un autre. Nous mesurons cette capacité de chaque sous-séquence à l'aide de la valeur du khi-carré de Pearson calculé sur le tableau croisé entre l'appartenance au groupe et la présence de la sous-séquence en question. Notons que les valeurs du khi-carré sont comparables puisque les degrés de liberté sont les mêmes pour l'ensemble des sous-séquences. Nous sélectionnons ensuite les sous-séquences pour lesquelles le test du khi-carré est statistiquement significatif en appliquant une correction de Bonferroni pour prendre en compte la multiplicité des tests effectués.

La figure 4 présente les fréquences d'apparition des 10 sous-séquences permettant de discriminer le plus ceux nés avant ou après 1950. Les sous-séquences sont ordonnées, de gauche à droite, selon leur pouvoir discriminant. Les barres sont coloriées en fonction de la significativité du résidu de Pearson correspondant au croisement entre l'appartenance au groupe et la présence de la sous-séquence. Un résidu négatif (resp. positif) nous montre que la sous-séquence apparaît significativement moins (resp. plus) souvent qu'attendu en cas d'indépendance.

Nous avions mis en évidence la présence d'un effet de génération à l'aide d'un test statis-

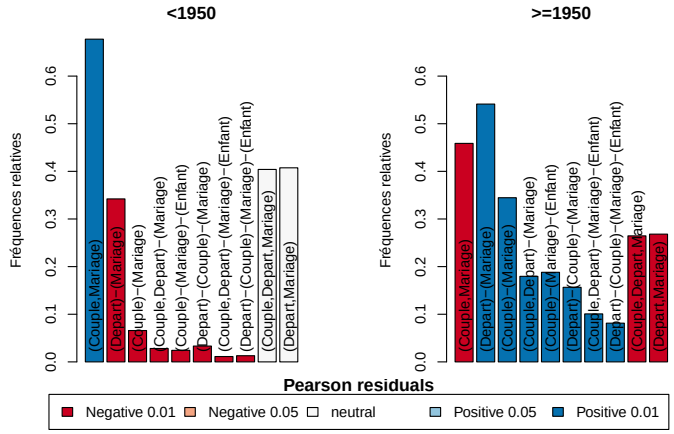

FIG. 4 – *Fréquences d'apparition des* 10 *sous-séquences les plus discriminantes.*

tique sur les dissimilarités. La figure 4 permet d'interpréter cet effet et notamment de mettre en évidence l'apparition d'un délai entre la cohabitation et le mariage pour les individus nés après 1949.

6 Conclusion

Dans cet article, nous avons présenté un ensemble de méthodes d'analyse de séquences d'événements visant à des questions communes en sciences sociales. Ces outils se basent sur la définition d'une mesure de dissimilarité entre séquences qui prend en compte l'ordonnancement et la temporalité des transitions. La mesure proposée prend en compte la simultanéité des événements ainsi que la longueur des séquences comparées tout en respectant l'inégalité triangulaire. Il est ainsi possible de construire des typologies et de mesurer les liens avec d'autres variables d'intérêts. Nous avons en particulier obtenu des résultats pertinents pour l'analyse des parcours de cohabitation en Suisse. La mesure doit être encore testée plus systématiquement, par exemple à l'aide de données générées artificiellement afin de mettre en lumière ses forces et ses limites.

Dans un deuxième temps, nous avons présenté des méthodes de visualisation qui permettent d'interpréter les résultats obtenus. L'utilisation de courbes de survie jusqu'à l'apparition de chacun des événements permet de comparer leur temporalité dans les séquences. Nous avons également introduit la notion de sous-séquence discriminante pour repérer les différences les plus significatives d'ordonnancement des événements entre groupes. Nous travaillons actuellement à l'extension de la recherche de séquences discriminantes à l'aide d'autres tests statistiques. Nous envisageons, par exemple, de comparer les durées ou encore l'attente jusqu'à l'apparition de chacune des sous-séquences.

Finalement, rappelons que l'ensemble des outils que nous avons présenté est disponible dans TraMineR. Les résultats peuvent ainsi être réutilisés en interaction avec d'autres méthodes d'analyses plus classiques disponible dans R.

Références

Abbott, A. (1990). A primer on sequence methods. *Organization Science 1*(4), 375–392.

Agrawal, R. et R. Srikant (1995). Mining sequential patterns. In P. S. Yu et A. L. P. Chen (Eds.), *Proceedings of the International Conference on Data Engeneering (ICDE), Taipei, Taiwan*, pp. 487–499. IEEE Computer Society.

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

Levine, J. (2000). But what have you done for us lately. *Sociological Methods & Research 29 (1)*, pp. 35–40.

Maechler, M., P. Rousseeuw, A. Struyf, et M. Hubert (2005). Package 'cluster': Cluster analysis basics and extensions. Refence manual, R-project, CRAN.

Masseglia, F. (2002). *Algorithmes et applications pour l'extraction de motifs séquentiels dans le domaine de la fouille de données : de l'incrémental au temps réel*. Ph. D. thesis, Université de Versailles Saint-Quentin en Yvelines.

Moen, P. (2000). *Attribute, Event Sequence, and Event Type Similarity Notions for Data Mining*. Ph. D. thesis, University of Helsinki.

Müller, N. S., S. Lespinats, G. Ritschard, M. Studer, et A. Gabadinho (2008). Visualisation et classification des parcours de vie. *Revue des nouvelles technologies de l'information RNTI E-11, II*, 499–510.

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.

Wu, L. L. (2000). Some comments on 'sequence analysis and optimal matching methods in sociology : Review and prospect'. *Sociological Methods Research 29*(1), 41–64.

Yujian, L. et L. Bo (2007). A normalized levenshtein distance metric. *IEEE Transactions On Pattern Analysis And Machine Intelligence 29*(6), 1091–1095.

Zaki, M. J. (2001). SPADE : An efficient algorithm for mining frequent sequences. *Machine Learning 42*(1/2), 31–60.

Summary

This article presents a set of tools to analyze event sequences in the social sciences and visualize the results. We begin by formalizing the notion of event sequence before defining a measure of dissimilarity between these sequences to cluster them and test the links between these sequences and other variables of interest. Initially defined by Moen (2000), this measure is based on the notion of edit distance between sequences and identifies the differences in sequencing and timing of events. We propose an extension of it in order to take into account the simultaneity of events and a normalization method that guarantees the respect of the triangle inequality. In a second step, we present a set of tools to interpret the results. We thus propose two methods of viewing a set of sequences and we introduce the concept of discriminant subsequence that identifies differences in sequencing that are the most significant between groups. All the tools presented are available in the TraMineR R library.