Preprint de

Studer, M., A. Gabadinho, N.S. Müller et G. Ritschard (2008), Approches de type n-grammes pour l'analyse de parcours de vie familiaux, Revue des Nouvelles Technologies de l'Information, E-11 (vol. II), EGC'2008, 511-522.

Approches de type n-grammes pour l'analyse de parcours de vie familiaux

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

*Département d'économétrie et Laboratoire de démographie, Université de Genève
{matthias.studer, nicolas.muller, gilbert.ritschard}@metri.unige.ch,
alexis.gabadinho@ses.unige.ch
http://www.unige.ch/ses/metri/

Résumé. Cet article[1] porte sur l'analyse de parcours de vie représentés sous forme de séquences d'événements. Plus spécifiquement, on examine les possibilités d'exploiter des codages de type n-grammes de ces séquences pour en extraire des connaissances. En fait, compte tenu de la simultanéité de certains événements, une procédure stricte de n-grammes comme on peut par exemple l'appliquer sur des textes, n'est pas applicable ici. Nous discutons diverses alternatives qui s'avèrent finalement plus proches de la fouille de séquences fréquentes. Les concepts discutés sont illustrés sur des données de l'enquête biographique rétrospective réalisée par le Panel suisse de ménages en 2002. Enfin, on précisera sur quels aspects l'approche proposée peut apporter un éclairage complémentaire utile par rapport à d'autres techniques plus classiques d'analyse exploratoire de parcours de vie.

1 Introduction

Existe-t-il des séries typiques d'événements qui structurent la vie familiale ? Est-ce que certaines séquences d'événements sont typiques d'une partie de la population ou d'une sous-population ? Pour répondre à ces questions, les sciences sociales ont besoin de méthodes pour analyser les parcours de vie dans leur totalité. Mais comment décrire ou comparer des séquences d'événements ? Dans cet article, nous proposons de nous centrer sur les transitions dans les parcours de vie pour les décrire. Ainsi, l'approche proposée adopte un point de vue complémentaire à l'alignement de séquences, par exemple, qui se base sur des séquences d'états.

Les parcours de vie familiaux peuvent être compris comme des séries de transitions entre états de la vie familiale telles que fonder un nouveau foyer, l'arrivée d'un nouvel enfant ou le remariage d'un parent...[2] Ces transitions peuvent être caractérisées par plusieurs événements simultanés, par exemple, lorsqu'une personne fonde un foyer en quittant son domicile parental

[1] Etude soutenue financièrement par le Fonds national suisse de la recherche (FNS) FN-100012-113998, et réalisée avec les données collectées dans le cadre du projet « Vivre en Suisse 1999-2020 », piloté par le Panel suisse de ménages et supporté par le FNS, l'Office fédéral de la statistique et l'Université de Neuchâtel.

[2] Dans cet article, nous nous centrerons sur la vie familiale, mais nous pourrions inclure d'autres ensembles d'événements tels que ceux affectant la vie professionnelle.

ou encore lorsqu'elle se marie en même temps que la naissance d'un enfant. Nous nous intéresserons à caractériser les parcours de vie familiaux des individus vivant en suisse en prenant les données de l'enquête biographique rétrospective réalisée par le Panel suisse de ménages en 2002.

L'idée de départ de cet article est d'explorer l'applicabilité d'une approche de type n-grammes pour caractériser les parcours de vie et en extraire des connaissances. Les n-grammes ont été utilisés dans des contextes multiples pour caractériser des textes (Damashek, 1995; Mayfield et McNamee, 1998). Cette méthode se base sur le découpage en courtes sous-séquences de caractères d'un texte afin de pouvoir le caractériser et le comparer. Dès lors, il peut sembler logique d'utiliser cette méthode pour caractériser les parcours de vie représentés sous forme de séquences d'événements. Cependant, la méthode n'est pas directement applicable à cause des différences dans les types de données. En effet, les n-grammes sont construits sur une séquence continue qui ne connaît ni la simultanéité des événements (les caractères forment une séquence stricte) ni les trous (absence de caractère pendant une période indéfinie). De plus, il importe ici de tenir compte de la variabilité des durées qui séparent des événements consécutifs. Nous proposons d'utiliser la recherche de sous-séquences fréquente pour pallier ce manque.

Le reste de l'article est organisé de la manière suivante. Nous commençons par décrire plus précisément les données de type « parcours de vie » que nous avons à notre disposition. Nous passons ensuite en revue les principes de l'analyse n-grammes et discutons des limites de leurs applications au parcours de vie. Nous présentons ensuite la recherche de sous-séquences fréquentes avant de discuter de son adaptation pour caractériser les parcours de vie. Finalement, nous appliquons la méthode décrite avant de discuter de ses apports par rapport à d'autres méthodes plus classiques d'analyse des parcours de vie.

2 Présentation des données

Nous utilisons les données de l'enquête biographique rétrospective réalisée par le Panel suisse de ménages[3] en 2002. Nous utilisons les résultats du questionnaire sur les parcours de cohabitation. Pour chaque année, nous connaissons les personnes avec lesquelles habitaient le répondant. Ainsi, nous disposons, pour chaque individu, d'une histoire de sa vie familiale. Nous avons choisi de centrer notre analyse sur la transition vers l'âge adulte en prenant les parcours de vie depuis la naissance jusqu'à trente ans. Afin de comparer des séquences similaires, nous n'avons retenu que les répondants ayant trente ans lors de l'enquête, soit 3557 individus.

Notre base de données se présente sous la forme d'une liste « individu-transition » où chaque transition est décrite à l'aide d'un ensemble d'événements. Les événements considérés sont les arrivées et les départs de personnes qui compose le ménage dans lequel vit un individu donné. Le tableau 1 donne la liste des événements que nous avons utilisés dans notre analyse. Le codage des départs nous permet de capter les transitions en terme d'un état vers un autre. En effet, la transition n'est pas seulement caractérisée par son état de destination, mais également par son origine.

Nous présentons dans le tableau 2 un extrait de notre base de données. Dans cet exemple, nous avons représenté deux individus. Le premier n'a connu que deux transitions : il naît en 1973 (« apparition » des parents dans le parcours) et quitte ses parents (événements L et

[3] http ://www.swisspanel.ch

Vit avec	Arrivée	Départ
Ami(s)	A	B
Autre(s)	C	D
Conjoint marié	E	F
Enfant	G	H
Enfant du conjoint	I	J
Mère	K	L
Conjoint non marié	M	N
Conjoint d'un parent	O	P
Père	Q	R
Seul	S	T

TAB. 1 – *Liste des événements*

R) pour se mettre en union (événement M) en 1993, soit à vingt ans. Le deuxième individu connaît un parcours marqué par un plus grand nombre de transitions. Il naît en 1965 et quitte ses parents pour se mettre en union en 1989 ; il se marie en 1990 et a son premier enfant en 1991.

ID individu	Date	Evénements
1	1973	KQ
1	1993	LMR
2	1965	KQ
2	1989	LMR
2	1990	EN
2	1991	G

TAB. 2 – *Extrait de la base de données*

Cette représentation des données — considérée notamment par Agrawal et Srikant (1995) — n'est pas spécifique des données de type parcours de vie, mais peut représenter plusieurs types de séquence qui doivent prendre en considération la simultanéité des événements. Il en va de même pour le codage des événements que nous avons réalisé. En effet, dans une optique plus générale, on peut penser les événements comme l'apparition ou la disparition d'un attribut au cours de la vie d'un individu.

Après avoir présenté les données et leur particularité, nous présentons les principes des n-grammes et nous discutons des limites de leurs applications à ces données.

3 Principe des n-grammes

Nous reprenons ici la présentation des n-grammes de Damashek (1995). Un texte peut être représenté en utilisant un vecteur composé des fréquences relatives de chaque n-gramme distinct qui le compose. La liste exhaustive des n-grammes correspond à l'ensemble des séquences

de n caractères obtenus en faisant coulisser une fenêtre de n caractères le long du texte, un caractère à la fois. Ainsi, si $n = 3$, le texte « abcde » sera décrit à l'aide des tri-grammes « abc », « bcd » et « cde ».

Le poids assigné à chaque n-gramme est égal à sa fréquence relative. Ainsi, si un texte comprend j n-grammes distincts, le poids x_i du i-ième éléments est égal à :

$$x_i = \frac{m_i}{\sum_{j=1}^{j} m_j} \tag{1}$$

avec m_j, le nombre d'occurrences du j-ième n-gramme. Par construction, on a $\sum_{j=1}^{j} x_j = 1$.

Cette pondération donne un poids identique à chaque caractère de la séquence — chacun apparaissant dans n n-grammes[4]. D'autres types de pondération existent, tel que l'utilisation du TF/IDF (Mayfield et McNamee, 1998).

Afin de juger de la similarité de deux textes, Damashek (1995) propose d'utiliser le cosinus de l'angle entre deux vecteurs. Il note qu'il peut être intéressant de mesurer cette différence par rapport au centroïde. En effet, ceci permet de mesurer les dissimilarités à partir d'un point commun (qui pourrait former l'origine « réelle ») plutôt que par rapport à l'origine mathématique, nécessairement arbitraire.

A première vue, l'application des principes des n-grammes aux parcours de vie peut sembler directe. Il suffirait de coder les parcours de vie sous la forme de séquences de caractères. Cependant, ce codage n'est pas possible pour deux raisons. Premièrement, les textes sont caractérisés par de très longues séquences alors que les parcours de vie ne contiennent que peu de caractères (ou d'événements) en comparaison. Ainsi, l'information apportée par un caractère dans un texte est bien moindre que celle d'un événement dans un parcours de vie.

Deuxièmement, la notion de succession des caractères est clairement établie dans le cas des textes, où chacun est précédé et suivi d'un autre (hormis pour la fin du texte). Il n'en va pas de même pour les parcours de vie où les événements peuvent être simultanés. De plus, la notion de succession est également différente dans une séquence d'événements de vie. En effet, celle-ci n'est pas complètement définie : est-ce qu'un événement qui se produit vingt ans après un autre lui succède de la même manière que si l'événement suivant se produit 3 ans après ?

Pour pallier ces problèmes, nous proposons de nous baser sur la notion de k-séquence plutôt que celle de n-grammes pour décrire les séquences à l'aide de leurs sous-séquences. Cette notion a été introduite par Agrawal et Srikant (1995) dans le cadre de la recherche de sous-séquences fréquentes d'événements. Elle nous permet d'aborder les notions de succession et de simultanéité des événements et semble ainsi mieux adaptée aux données de parcours de vie.

4 K-séquence

Le concept de k-séquence et sa formalisation sont introduits par Agrawal et Srikant (1995). Nous utilisons ici la notation proposée par Zaki (2001) dans la présentation de son algorithme SPADE de recherche de sous-séquences fréquentes et de règles d'association entre celles-ci.

[4]Pour autant que l'on omette les $n-1$ premiers et derniers caractères. Cette différence de pondération est cependant négligeable dans le cas d'un texte en raison de la longueur de celui-ci.

Nous reprenons ici sa formulation en adaptant les termes à la terminologie que nous avons déjà introduite.

Une *séquence* peut être comprise comme une liste *ordonnée* de *transitions*. Une séquence α est notée $(\alpha_1 \rightarrow \alpha_2 \rightarrow ... \rightarrow \alpha_q)$, où chaque α_i désigne une transition. Les transitions sont des listes *non ordonnées* d'*événements* distincts (c'est-à-dire qu'un événement ne peut apparaître deux fois dans la même transition). Une séquence composée de k événements est appelée une *k-séquence*.

Agrawal et Srikant définissent alors la sous-séquence de la manière suivante : α est une sous-séquence de β si chaque transition de α est un sous-ensemble de β et que l'ordre des transitions est conservé. On le note $\alpha \preceq \beta$. Ainsi, par exemple, $(B \rightarrow AC)$ est une sous-séquence de $(AB \rightarrow E \rightarrow ACD)$ puisque $B \subseteq (AB)$ — l'événement B est contenu dans la transition (AB) — et que $(AC) \subseteq (ACD)$ — les événements A et C sont contenus dans la transition (ACD) .

Une sous-séquence est dite *fréquente* si on l'observe dans un nombre de séquences supérieures à un support minimum[5] défini au préalable. Une sous-séquence est dite *maximale* si elle n'est incluse dans aucune autre sous-séquence fréquente.

Dans un but d'interprétation des résultats, nous avons jugé utile de spécifier une fenêtre de temps maximale pour la recherche de sous-séquences (spécification qui existe également dans SPADE). Ainsi, il est possible de ne rechercher que les sous-séquences qui se déroulent dans un laps de temps donné. Cette spécification s'avère très utile dans le cas de longue séquence avec peu d'événements, car il est plus difficile d'assumer une liaison entre des événements très éloignés dans le temps.

La formalisation d'Agrawal et Srikant permet donc de distinguer l'ordre d'apparition des événements, mais également leurs simultanéités, en considérant les transitions comme des listes non ordonnées d'événements. Les k-séquences nous permettent de caractériser les parcours de vie de la même manière que les n-grammes permettaient de caractériser des textes. Les parcours de vie peuvent être considérés comme des séquences caractérisées par des sous-séquences. C'est l'aspect que nous présentons à présent.

5 Caractérisation de parcours de vie à l'aide de k-séquences

En suivant le principe des n-grammes, on peut représenter chaque séquence à l'aide d'un vecteur composé des fréquences relatives de chacune des sous-séquences. Pour cela, nous devons encore définir un critère pour choisir les sous-séquences que nous utiliserons. Nous devons également choisir une méthode de comptage des sous-séquences afin de calculer les fréquences relatives.

5.1 Choix des sous-séquences

En suivant la méthode des n-grammes, on devrait utiliser un k fixé pour le calcul des sous-séquences. Cependant, ceci n'est pas pertinent pour les parcours de vie. En effet, certaines dynamiques ne peuvent être captées. Dans notre cas, un individu qui resterait chez ses parents jusqu'à trente ans ne pourrait être décrit à l'aide d'une 3-séquence, car il n'aurait connu que

[5] nombre minimum de séquences qui contiennent la sous-séquence.

deux événements. Il en va de même pour une personne ayant un enfant après dix ans de mariage, si notre fenêtre de temps est plus restreinte. Dès lors, il faudrait prendre des séquences d'ordre 1. Cette solution enlève toutes les notions de successions des événements et donc de séquence. Deux méthodes sont envisageables pour choisir les sous-séquences :

1. On retient l'ensemble des sous-séquences fréquentes qui sont applicables à une séquence donnée. Ainsi, si l'on considère la séquence $(A \rightarrow B \rightarrow C)$, on utilisera les sous-séquences $(A \rightarrow C)$ $(A \rightarrow B)$ et $(B \rightarrow C)$ ainsi que les « sous-séquences événements » (A),(B) et (C) pour autant qu'elles satisfassent le support minimum.

2. On ne retient que les sous-séquences maximales d'une séquence donnée. Ainsi, la solution dépend du support minimum. Si l'on considère la séquence $(A \rightarrow B)$, on ne retiendra que $(A \rightarrow B)$ si cette sous-séquence est fréquente ou les « sous-séquences événements » (A) et (B), pour autant qu'elles soient fréquentes, sinon. Cette dernière méthode implique de recalculer plusieurs fois les sous-séquences fréquentes (puisque les fréquences dépendent de la sélection effectuée). Elle a cependant l'avantage de ne pas multiplier le comptage de chaque événement entre les sous-séquences.

5.2 Méthode de comptage et calcul des poids

Afin de calculer les poids de chaque sous-séquence, il est nécessaire de choisir une méthode de comptage. Il existe plusieurs méthodes pour compter le nombre de fois qu'une sous-séquence apparaît dans un parcours de vie (Joshi et al., 2001). Nous proposons de compter le nombre d'occurrences d'une sous-séquence dans chaque séquence de vie, c'est-à-dire le nombre de fois où l'on observe la sous-séquence dans chaque parcours de vie. Par exemple, dans la séquence $(A \rightarrow B \rightarrow B)$, on compte deux fois la sous-séquence $(A \rightarrow B)$, une fois en liant (A) au premier (B) et une deuxième fois en le liant au second.

Si l'on s'en tient aux fréquences relatives des sous-séquences, chaque événement et chaque transition peuvent avoir des poids différents en fonction de la séquence. En effet, un événement peut être décrit à l'aide de plus ou moins de sous-séquences. Ce problème n'apparaît pas explicitement dans le cas des n-grammes. En effet, on considère une fenêtre coulissante, ce qui implique que chaque lettre est comptée n fois. Il est vrai que les $n-1$ premiers et derniers caractères sont compté moins souvent. Cependant, cette différence de poids est compensée par la longueur du texte. Ainsi, l'utilisation de la fréquence relative dans les n-grammes assure un poids plus ou moins équivalent à chaque caractère.

Plusieurs solutions peuvent être adoptées :

- Un poids identique à chaque sous-séquence.
- Un poids établi de manière à ce que chaque événement ait un poids identique dans l'ensemble.
- Un poids établi de manière à ce que chaque transition ait un poids identique dans l'ensemble.

Il est nécessaire de fixer la base de calcul des poids accordés à chaque sous-séquence. Ils peuvent être calculés, par exemple, sur la base des événements concernés et du nombre de sous-séquences qui les décrivent. Ainsi si l'on prend un ensemble de parcours de vie composé de K événements distincts, on évaluera le poids w_s associé à une sous-séquence s, composée

de E événements en utilisant :

$$w_s = \sum_{e=1}^{E} (\frac{1}{K} \cdot \sum_{s_e \in S_e} \frac{q_s}{q_{s_e}}) \tag{2}$$

Où q_s désigne le nombre d'occurrences de la sous-séquence s, S_e l'ensemble des sous-séquences fréquentes contenant l'événement e et q_{s_e} le nombre d'occurrences de la séquence s_e qui exprime l'événement e. Ainsi, chaque événement repartit sa contribution (soit $\frac{1}{K}$) entre l'ensemble des sous-séquences qui le décrivent.

La caractérisation des séquences à l'aide de sous-séquences nécessite de spécifier les trois points suivants : la méthode de sélection des sous-séquences fréquentes, la méthode de comptage ainsi qu'une base permettant d'assigner des poids à chaque sous-séquence.

6 Application : Analyse en composantes principales

Nous avons caractérisé l'ensemble des séquences à l'aide des sous-séquences fréquentes en choisissant un support minimum de deux pour cent. Ce seuil nous fait retenir 326 sous-séquences différentes pour décrire les parcours de vie. Nous ne présentons pas ici les résultats obtenus avec la caractérisation à l'aide des sous-séquences maximales. En effet, les résultats sont difficiles à interpréter, sans-doute à cause du manque de redondance de l'information qui permet de calculer les proximités entre événements et sous-séquences. Pour cette première analyse, nous avons choisi d'accorder un poids identique à chaque sous-séquence en suivant les principes utilisés dans le cas des n-grammes.

A l'aide de ces données, nous avons effectué une analyse en composante principale pour offrir une visualisation de l'espace formé par les parcours de vie. Dans les graphiques qui suivent, les sous-séquences contribuant le plus significativement à la construction des axes sont représentées à l'aide d'un rond bleu. Nous avons également ajouté plusieurs variables supplémentaires : le sexe (triangle, pointe en haut, rouge), la langue d'interview que nous considérerons comme un proxy pour aborder la culture d'origine (triangle, pointe en bas, rouge clair), la cohorte de naissance — la période dans laquelle on est né — (carré vert) ainsi que la confrontation à des problèmes d'argent dans la jeunesse (losange bleu). Ces différentes variables nous permettront de caractériser le contexte dans lequel ces séquences apparaissent.

Nous avons également ajouté comme variable supplémentaire (représentée à l'aide d'un pentagone vert) la classification des parcours de vie décrite dans Müller et al. (2007) obtenue à l'aide de la méthode « optimal matching ». Cette méthode permet de classer les séquences d'états en respectant les états présents ainsi que leurs temporalités. Cette classification est construite sur la base de quatre événements soit le départ du domicile parental, le mariage, l'arrivée du premier enfant et le divorce. La méthode décrite ici permet de prendre en considération un plus grand nombre d'événements, tel que la mise en couple, la cohabitation avec des amis, etc... Nous pourrons ainsi comparer la méthode présentée avec celle de l'optimal matching.

La figure 1 nous montre les deux premières dimensions de l'espace formé par les séquences. La première dimension oppose ceux qui sont restés chez leurs parents (à gauche) aux autres. Ainsi, on retrouve à gauche les séquences K, Q et KQ qui dénotent l'arrivée des deux parents. De par notre système de pondération, ces sous-séquences ont un poids très élevé

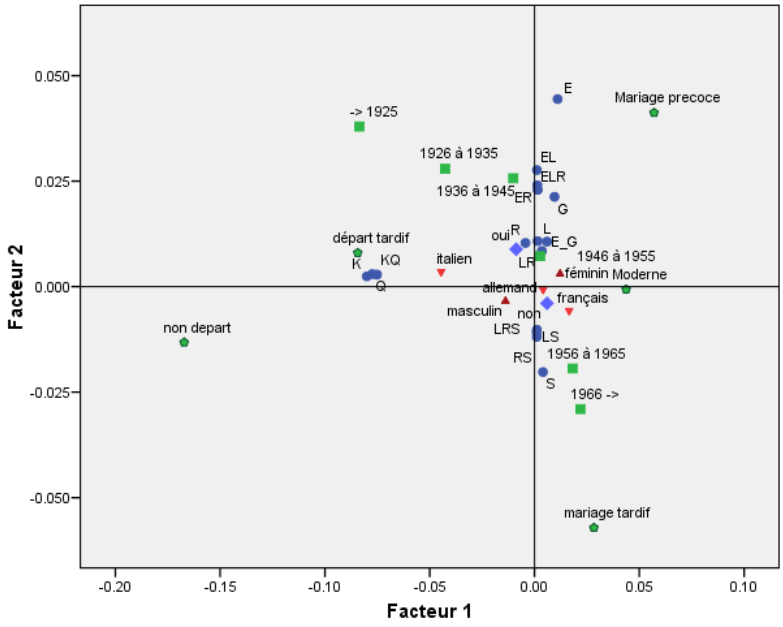

FIG. 1 – *Positionnement des séquences et des variables supplémentaires sur le plan des première (33.37% de la variance) et deuxième (11.27%) composantes principales.*

chez les individus qui ne sont pas partis à l'âge de trente ans ou qui ont connu peu d'événements ou de transitions. On retrouve ainsi une forte corrélation entre cette dimension et le nombre d'événements ($r = 0.61$) ou avec le nombre de transitions ($r = 0.55$). Le deuxième axe distingue les individus qui quittent leurs parents pour s'établir seuls (LRS) de ceux qui se marient (E) et ont un enfant (G) ou quittent leur parent pour se marier (ELR). Cet axe est négativement corrélé au nombre d'événements ($r = -0.37$)

Les positions des variables supplémentaires sont intéressantes. La cohorte nous montre que les individus nés avant 1945 semblent plus propices à rester chez leurs parents jusqu'à trente ans ou connaissent moins de transitions. Les individus nés après 1946 se distinguent essentiellement le long de l'axe vertical et on voit apparaître les départs pour vivre seul. Ceux qui ont répondu au questionnaire en italien (venant de la Suisse Italienne) sont plus propices à rester chez leur parent. Ces résultats confirment ceux trouvés par Schumacher et al. (2006).

La classification de l'optimal matching se conforme aux axes et à leur description. Ainsi, le premier axe oppose ceux qui ne partent pas ou tard de chez leurs parents aux autres, alors que le second distingue la temporalité du mariage. On remarque que cette temporalité est prise en compte à l'aide des sous-séquences du départ du domicile parental qui marque ainsi des transitions différentes pour le mariage tardif et le mariage précoce.

La troisième dimension (figure 2) oppose ceux qui ont un poids fort sur les événements de type départ de chez les parents aux autres. Ainsi, le point le plus à droite correspond à ceux qui ne partent pas selon la classification « optimal matching ». C'est le dernier axe à être corrélé

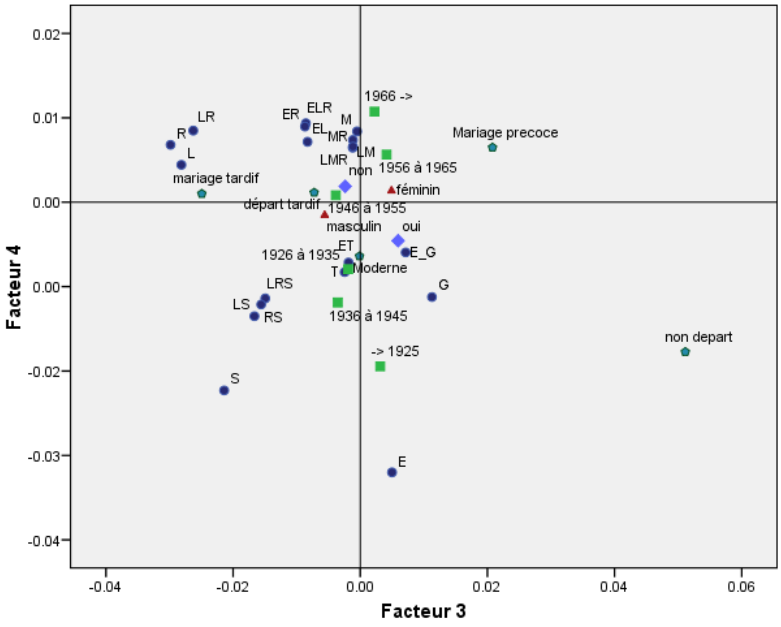

FIG. 2 – *Positionnement des séquences et des variables supplémentaires sur le plan des troisième (7.55%) et quatrième (5.45%) composantes principales.*

avec le nombre d'événements ou de transition ($r = 0.27$). La quatrième dimension distingue ceux qui partent pour vivre en couple ou pour se marier, de ceux qui partent vivre seuls. Ainsi, les regroupements de trajectoires menant directement au mariage se situent dans les valeurs positives.

La cinquième dimension (figure 3) oppose ceux qui partent de chez leurs parents pour se mettre en couple aux autres formes de départ. Cette dimension montre une forte association avec la cohorte. On remarque que le départ du domicile parental pour se mettre en union sans mariage immédiat (ce qui ne signifie pas sans mariage) est une dynamique relativement récente. Cette dimension oppose les répondants originaires de Suisse romande (francophone) aux autres. Finalement, la sixième dimension distingue ceux qui se marient la même année où ils ont leur premier enfant des autres formes de départ. Ce modèle est relativement plus présent en Suisse italienne et dans les cohortes plus anciennes.

Ici encore, on trouve un lien avec la classification de l'optimal matching. Cependant, les liens ne sont pas directs. Ainsi, la catégorie « Mariage tardif » est située tout à gauche montrant qu'on se marie tard, mais qu'on est déjà en union. Il en va de même pour les trajectoires « Modernes ». Les « Mariages précoces » se situent à l'opposé et montre qu'on se marie directement en quittant ses parents et qu'on tend également à avoir des enfants à ce moment.

Nous ne présentons ici que les six premières dimensions. Cependant, il est possible d'associer un sens aux dix premières dimensions. Ceci nous montre qu'un grand nombre de dimensions sont nécessaires à la compréhension de l'espace des séquences. Ce n'est pas une surprise

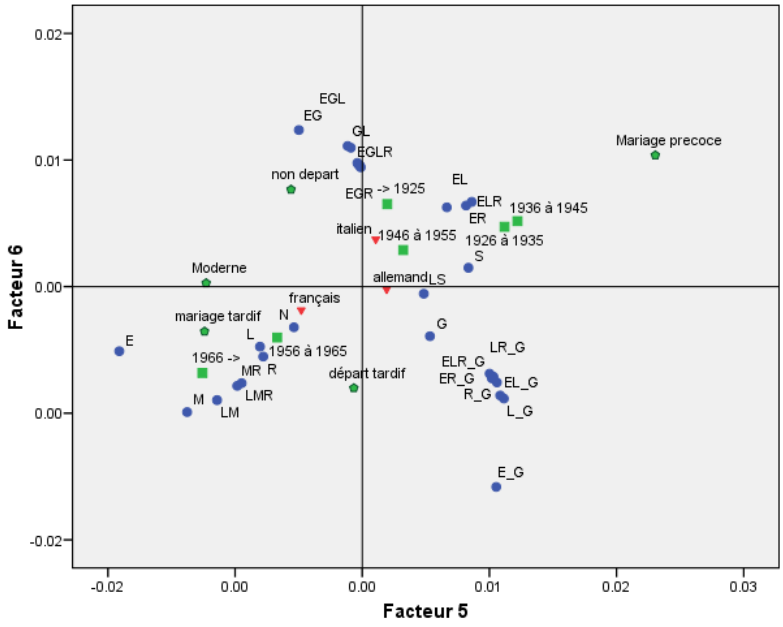

FIG. 3 – *Positionnement des séquences et des variables supplémentaires sur le plan des cinquième (4.17%) et sixième (3.5%) composantes principales.*

étant donné que nous avons regroupé un grand ensemble d'événements qui sont souvent analysés séparément, tel que le départ du foyer parental, la mise en union ou encore l'arrivée du premier enfant.

Les résultats obtenus sont relativement similaires à ceux de l'optimal matching, mais apportent un éclairage complémentaire en se centrant sur les transitions. Ainsi, on a pu voir des différences entre le modèle de mariage tardif et précoce en identifiant certaines transitions intermédiaires (notamment la mise en union). La méthode proposée a cependant l'avantage de pouvoir intégrer un plus grand nombre d'événements, surtout s'ils peuvent être simultanés. Avec l'optimal matching, cette situation implique un trop grand nombre d'états et les résultats deviennent difficilement interprétables. Sans spécifier explicitement une temporalité, celle-ci se retrouve dans les axes. Cependant, ceci est dû aux séquences qui se centrent sur la période allant de la naissance à 30 ans ainsi qu'à la présence de l'événement départ du domicile parental que presque l'ensemble des individus connaît pendant cette période.

Ainsi, la caractérisation à l'aide des sous-séquences fréquentes permet d'offrir une visualisation de l'espace formé par les séquences et de dégager quelques dynamiques intéressantes sur les différences entre types de trajectoire, notamment avec l'aide des variables supplémentaires. Toutefois, nous voyons deux limites à la méthode présentée qui nécessite des développements futurs. Le système de pondération des sous-séquences rend parfois l'interprétation des axes difficile, puisqu'elle dépend de la présence d'autres événements dans la séquence. Il s'agit d'un poids relatif à la séquence. Il est ainsi nécessaire de développer d'autres méthodes qui

permettent une meilleure interprétation des poids.

Deuxièmement, l'analyse présentée — fondée sur l'ordonnancement des événements — ne tient que partiellement compte de la temporalité des événements, ce qui peut s'avérer crucial dans l'étude des parcours de vie : l'événement « devenir veuf » n'est pas le même si on le vit à trente ou à quatre-vingts ans, ses implications sur le reste du parcours ne sont pas les mêmes. Dès lors, il nous semble nécessaire d'ajouter cette dimension à l'analyse, par exemple en utilisant une approche similaire à celle de Deville et Saporta (1983).

7 Conclusion

En suivant les principes des n-grammes, nous avons proposé une méthode qui permet de caractériser les parcours de vie par des sous-séquences fréquentes. Cette caractérisation rend ensuite possible l'utilisation d'un vaste ensemble d'outils de l'analyse exploratoire, tel que l'analyse en composante principale. A notre avis, deux points au moins peuvent encore être améliorés. Premièrement, il est nécessaire de développer un système de pondération qui rende l'interprétation des coordonnées plus aisée. Deuxièmement, des méthodes doivent être intégrées pour incorporer la temporalité des sous-séquences si l'on veut aborder de plus longs parcours de vie.

Cette étude établit cependant clairement que la méthodologie proposée complète les analyses exploratoires traditionnelles en se concentrant sur l'ordonnancement des événements. En effet, nous avons montré que nos résultats complétaient utilement ceux fournis par la classification à l'aide de l'optimal matching. De plus, elle permet d'étudier un plus grand nombre d'événements distincts.

Références

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.

Damashek, M. (1995). Gauging similarity with ngrams: Language-independent categorization of text. *Science 267*, 843–848.

Deville, J. C. et G. Saporta (1983). Correspondence analysis with an extension towards nominal time series. *Journal of Econometrics 22*, 169–189.

Joshi, M. V., G. Karypis, et V. Kumar (2001). A universal formulation of sequential patterns. In *Proceedings of the KDD'2001 workshop on Temporal Data Mining, San Fransisco, August 2001*.

Mayfield, J. et P. McNamee (1998). Indexing using both n-grams and words. In *TREC*, pp. 361–365.

Müller, N. S., M. Studer, et G. Ritschard (2007). Classification de parcours de vie à l'aide de l'optimal matching. In *XIVe Rencontre de la Société francophone de classification (SFC 2007), Paris, 5 - 7 septembre 2007*, pp. 157–160.

Schumacher, R., T. Spoorenberg, et Y. Forney (2006). Déstandardisation, différenciation régionale et changements générationnels. départ du foyer parental et modes de vie en Suisse au XXe siècle. *European Journal of Population 22*, 153–177.

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

Summary

This article is concerned with life courses represented in the form of sequences of time stamped events. More specifically, we examine the possibility of extracting relevant knowledge through n-gram coding of such life course sequences. The possible simultaneity of life events prevents a straightforward application of n-gram-based procedures such as applied on texts for instance. We discuss various alternatives, which lead us to techniques more related to the mining of frequent subsequences than to n-grams. The discussed concepts are illustrated using data from the retrospective biographical survey carried out in 2002 by the Swiss Household Panel. Finally, we identify the nature of the additional knowledge brought by this approach when compared with more classical exploratory techniques used in life course analysis.