ATELIER FOUILLE VISUELLE DE DONNEES : METHODOLOGIE ET EVALUATION

Dans le cadre de la conférence Extraction et Gestion de Connaissances (EGC'2012)

Bordeaux, 31 janvier 2012

Hanene Azzag, Bénédicte Le Grand, Monique Noirhomme, Fabien Picarougne, François Poulet

Exploration graphique de données séquentielles

Reto Bürgin, Gilbert Ritschard, Emmanuel Rousseaux

Institut d'études démographiques et du parcours de vie, Université de Genève
et Pôle de recherche national LIVES 'Vulnérabilités et parcours de vie'
{reto.buergin, gilbert.ritschard, emmanuel.rousseaux}@unige.ch,
http://mephisto.unige.ch/

Résumé. Nous proposons une façon originale de représenter graphiquement des données longitudinales catégorielles. La visualisation proposée, inspirée des courbes de séries temporelles, se prête particulièrement bien à la description et à l'exploration de trajectoires individuelles décrites sous forme de séquences d'événements. Outre la description de la méthode de visualisation et des principes qui la président, l'article comprend des exemples d'application et une discussion des propriétés des graphiques produits. De plus, nous expliquons également quelques astuces de spécification permettant d'optimiser le rendu des données.

1 Introduction

L'analyse de données longitudinales a connu un essor important ces dernières années dans de nombreux domaines, par exemple pour suivre les comportements de consommateurs ou d'utilisateurs de services, suivre les développement cognitifs d'individus, contrôler le fonctionnement d'appareils, ou encore pour étudier les carrières professionnelles ou les trajectoires familiales. L'étendue et la portée des applications ne cesse de croître (Frees, 2004), et cette demande a engendré le développement de méthodes d'analyse spécifiques pour données longitudinales. Du côté de la statistique, l'effort a été mis, pour les séquences numériques sur les approches confirmatoires comme les modèles de régression multiniveaux (Hox, 2010), les modèles d'équations structurelles à croissance latente pour données répétées (McArdle, 2009), les modèles de survie, ou encore les modèles de transition de type markovien (Berchtold et Raftery, 2002) et, dans une optique plus exploratoire, sur la caractérisation de configurations typiques de séquences d'états. Du côté de la fouille de données, l'accent a porté sur la recherche de sous-séquences fréquentes et de règles d'association séquentielles (Agrawal et Srikant, 1995; Masseglia, 2002).

L'objectif premier de la présente contribution est de proposer des outils graphiques pour l'exploration de parcours de vie décrits sous forme de séquences d'événements (quitter ses parents, finir les études, se mettre en couple, déménager, etc.). Plus particulièrement, nous nous intéressons aux graphiques qui mettent en évidence les caractéristiques principales de l'ensemble tout en rendant compte des trajectoires individuelles et de leur diversité.

Les graphiques pour données longitudinales que l'on trouve dans la littérature concernent principalement le rendu de séquences d'états (Gabadinho et al., 2011) que l'on visualise par

exemple facilement sous forme de barres empilées dont la couleur des éléments représente l'état et la longueur le temps passé dans l'état. Pour les séquences numériques, on recourt en général aux graphiques superposant les courbes de séries temporelles (Tufte, 2001), parfois appelés 'spaghetti plots'. Parmi les autres formes de représentation de séquences individuelles, on peut citer encore les graphes reliant les états ou événements successifs de chaque séquence (Hébrail et Cadalen, 2000), ou les graphiques qui visualisent sur une ligne ou une barre distincte pour chaque séquence le calendrier d'occurrence des événements. Dans cette optique, les variantes proposées sous forme de 'Life Flow' dans Wongsuphasawat et al. (2011) sont particulièrement intéressantes.

Les séquences d'événements qui nous intéressent s'apparentent à celles que l'on considère dans la fouilles des séquences d'achats de consommateurs (Agrawal et Srikant, 1995). Ces séquences se distinguent des séquences d'états sur deux points qui rendent beaucoup plus difficile leur visualisation. D'une part, les événements interviennent à un moment donné et n'ont donc pas de durée, ce qui exclut l'utilisation de barres de longueur proportionnelle à la durée, et d'autre part, plusieurs événements peuvent intervenir simultanément, par exemple terminer ses études et se mettre en ménage le même mois ou acheter plusieurs produits lors d'une même commande. Ceci explique peut-être le peu de place accordée jusqu'ici à la visualisation de séquencements d'événements. Parmi les rares propositions que l'on trouve dans la littérature figurent les calendriers d'événements et leurs variantes (Wongsuphasawat et al., 2011) déjà cités, et des représentations plus synthétiques comme la superposition des courbes de survie jusqu'à l'occurrence des divers événements (Studer et al., 2010). Ces représentations s'avèrent inadaptées à nos objectifs qui sont :

– visualiser la diversité des trajectoires individuelles observées ;
– identifier les séquencements les plus communs ;
– représenter des séquences individuelles pouvant comprendre des événements simultanés.

En effet, les graphiques de type calendrier se prêtent par exemple mal à la gestion d'événements simultanés et au cas de grands nombres de séquences, tandis que les représentations sous forme de graphes ainsi que les représentation agrégées comme les courbes de survie ne rendent pas compte des parcours individuels observés.

La représentation que nous proposons s'inspire des 'spaghetti plots' de séries temporelles numériques que nous adaptons pour satisfaire les trois critères ci-dessus. Bien que conçu pour visualiser des séquencements d'événements, le graphique proposé s'avère toutefois également utile pour explorer d'autres types de données longitudinales. Nous le verrons en particulier avec l'un des deux jeux de données illustratives considérés qui concerne des séquences d'états.

Un graphique se doit d'être aisément lisible et interprétable. Dans ce but, nous nous sommes attachés à suivre les indications de Diggle et al. (2002) pour une visualisation efficace de données longitudinales, à savoir :

1. montrer les données brutes pertinentes plutôt que des résumés synthétiques ;
2. mettre en évidence les modèles de configuration potentiellement intéressants ;
3. identifier les caractéristiques transversales et longitudinales de l'ensemble ;
4. faciliter l'identification des individus et observations atypiques.

L'article est organisé comme suit. Nous commençons par introduire deux jeux de données qui nous servirons d'illustration. Nous explicitons ensuite les principes de notre méthode et illustrons leur usage. Finalement, nous discutons les avantages et inconvénients de notre proposition avant d'évoquer en conclusion quelques pistes de développements futurs.

2 Données illustratives

Nous considérons deux jeux de données réels. Dans le premier cas, les éléments des séquences sont les états mutuellement exclusifs (pas de simultanéité) d'une variable ordinale. Dans le second exemple, les éléments sont des événements de vie qui peuvent être simultanés.

Anxiété après un séjour aux soins intensifs. 149 patients ayant séjournés dans l'unité de soins intensifs (ICU) de l'Hôpital universitaire de Berne ont été interrogés à trois reprises après leur séjour (Jeitziner et al., 2011). Ils devaient indiquer l'intensité de leur anxiété sur une échelle numérique 0-10 (0 = aucune anxiété, 10 = pire anxiété). Les interviews ont respectivement eu lieu $t_1 = 1$ semaine, $t_2 = 6$ mois et $t_3 = 12$ mois après leur séjour à l'ICU. 117 patients ont répondus les trois fois, et les autres une ou deux fois.

Des analyses préliminaires ont montré qu'il était difficile de voir les trajectoires effectives et d'identifier les séquences typiques avec une représentation sous forme de courbes de séries temporelles, les courbes étant à la fois trop diverses et se masquant les unes les autres. Pour notre illustration, nous discrétisons l'échelle des valeurs en 3 classes : $[0, 2]$, $(2, 6]$ and $(6, 10]$.

id	trajectoire
1	$[0,2]^{t_1} \rightarrow (6,10]^{t_2} \rightarrow (2,6]^{t_3}$
2	$[0,2]^{t_1} \rightarrow [0,2]^{t_2} \rightarrow [0,2]^{t_3}$
3	$[0,2]^{t_1} \rightarrow [0,2]^{t_2} \rightarrow [0,2]^{t_3}$

TAB. 1 – *Extrait de séquences de niveaux d'anxiété, patients 1, 2 and 3.*

Le tableau 1 présente les trajectoires de niveaux d'anxiété de trois patients avec les notations de Studer et al. (2010), où les flèches séparent les éléments successifs et l'indice supérieur reflète la date d'observation, ou, lorsque celle-ci n'est pas disponible, simplement la position dans la séquence. Le premier patient, par exemple, déclare une anxiété de niveau $[0, 2]$ une semaine après son séjour au soins intensifs, $(6, 10]$ six mois plus tard et de niveau $(2, 6]$ après un an.

Evénements de vie familiale. Il s'agit de données tirées de l'enquête biographique réalisée en 2002 par le Panel Suisse des Ménages (`http://www.swisspanel.ch`). Plus précisément, nous utilisons ici les données relatives à 2601 enquêtés qui avaient atteint l'âge de 45 ans au moment de l'enquête, c'est-à-dire ceux né en 1957 ou avant. Les événements retenus sont le départ du domicile parental (Depart), la première mise en couple (Couple), le premier mariage, le premier enfant et le premier divorce. On connaît l'année d'occurrence de ces événements, mais on ne retient ici que leur ordre en considérant comme simultanés les événements survenant une même année. On dispose en tout de 9021 événements pour l'ensemble des 2601 cas retenus.

Le tableau 2 donne à titre d'exemple les trajectoires de vie familiale de trois individus. On y lit notamment que l'individu 2600 a vécu tout d'abord le départ du domicile de ses parents avant de connaître plus tard trois événements au cours d'une même année, à savoir la première mise en couple, le mariage et la naissance du premier enfant.

id	trajectoire
2599	(Depart, Couple, Mariage, Enfant)^1
2600	(Depart)^1 $\rightarrow$ (Couple, Mariage, Enfant)^2
2601	(Depart, Couple, Mariage, Enfant)^1

TAB. 2 – *Trajectoires de vie familiale de trois individus.*

3 Méthode

Le point de départ de la représentation graphique proposée est une simple transposition du principe des 'spaghetti plots' pour séries temporelles numériques au cas de données catégorielles. L'axe des x reflète la date ou position dans la séquence, et l'axe des y la valeur de la variable, soit dans notre cas l'événement ou l'état. On reporte ainsi sur l'axe des ordonnées les divers éléments de l'alphabet catégoriel considéré, dans un ordre arbitraire si l'alphabet est nominal et dans l'ordre établi s'il est ordinal. On peut ainsi visualiser les trajectoires en reportant en regard de chaque position un point à la hauteur de l'événement (ou état) concerné et en reliant ces points-événements successifs par des lignes. En cas d'événements simultanés, on connecte les points du bas vers le haut par un trait vertical. Ceci est arbitraire et une alternative serait par exemple de les connecter de haut en bas.

La Figure 1 illustre le graphique ainsi obtenu pour les données des événements de vie familiale. L'ordre des événements est arbitraire et nous les avons ici ordonnés selon ce qui est encore l'ordre standard en Suisse. Le résultat est un graphe où toutes les paires observées

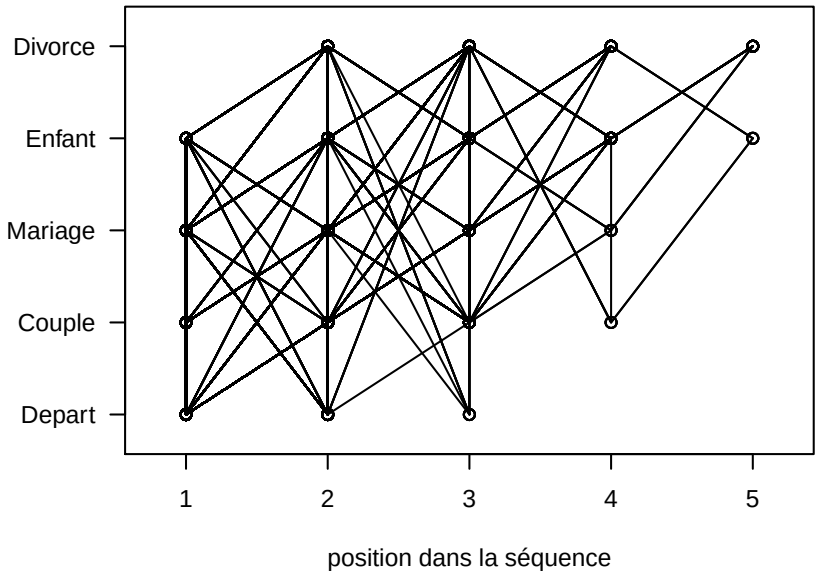

FIG. 1 – *Graphique de type 'séries temporelles' des séquences d'événements de vie familiale.*

d'événements consécutifs ou simultanés sont reliés. Si ce premier graphique permet de rendre compte de la simultanéité d'événements, il reste insatisfaisant car, en raison du caractère discret des événements, les trajectoires tendent à se masquer partiellement ou totalement les unes les autres. Ainsi, il ne permet en général ni de repérer les trajectoires fréquentes, ni a fortiori les trajectoires atypiques, ni de retracer les trajectoires individuelles. Il peut même être trompeur en laissant penser que toute séquence d'événements successivement connectés comme $(\text{first child})^1 \rightarrow (\text{divorce})^2 \rightarrow (\text{first child})^3 \rightarrow (\text{divorce})^4 \rightarrow (\text{first child})^5$ qui est clairement impossible, aurait été observée.

Décalement des courbes. Une première solution pour éviter le recouvrement de séquences consiste à décaler légèrement chaque séquence représentée. Pratiquement, on applique la même translation à chaque point-événement d'une trajectoire.

Afin de faciliter le suivi d'une trajectoire, nous représentons en arrière plan la *zone de translation* sous forme de carrés gris clair (voir figure 2). Le point central de la zone de translation indique la coordonnée d'origine commune des points visibles dans la zone. Tous les points d'une même trajectoire occupent la même position dans leur zone respective de déplacement. Il est ainsi aisé de repérer, par exemple, où se termine une trajectoire donnée quand on connaît la position de son point de départ.

La taille de la zone de translation est paramétrable. Il convient de laisser un espace entre les zones que l'on suggère de fixer à au moins la demi-largeur, respectivement la demi-longueur des zones.

Le seul décalage des courbes reste cependant insuffisant, le graphique devenant rapidement illisible en présence d'un grand nombre de trajectoires.

Trajectoires distinctes. Pour réduire le nombre de courbes, on propose de ne visualiser que les configurations distinctes et de jouer sur l'épaisseur du trait et des points-événements pour rendre compte de la fréquence d'observation des trajectoires représentées. Considérons à titre d'exemple les quatre trajectoires du tableau 3. Comme les séquences 1 et 3 sont identiques, on représente les quatre séquences avec trois courbes (partie gauche de la figure 2). L'épaisseur des traits et points de la trajectoire $(A)^1 \rightarrow (B)^2$ indique qu'elle est deux fois plus fréquente.

Le risque est que les trajectoires les plus fréquentes, qui donnent lieu aux traits les plus épais, cachent les moins fréquentes. On peut éviter en partie cet effet, en variant les couleurs des trajectoires, éventuellement en jouant avec des effets de transparence, et en dessinant en premier les traits les plus épais de sorte à placer les traits plus fins par dessus. Une option permet de fixer le support minimal des séquences à représenter et de visualiser les autres en gris clair pour rendre compte de la diversité.

id	trajectoire
1	$(A)^1 \rightarrow (B)^2$
2	$(A)^1 \rightarrow (A, B)^2$
3	$(A)^1 \rightarrow (B)^2$
4	$(A)^1 \rightarrow (B)^2 \rightarrow (C)^3$

TAB. 3 – *Exemple de trajectoires.*

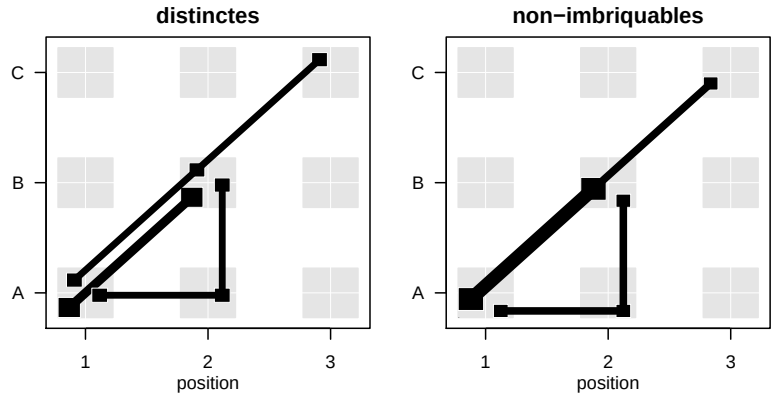

FIG. 2 – *Trajectoires distinctes et trajectoires non-imbriquables, données du tableau 3.*

Trajectoires non-imbriquables. Afin de réduire plus encore le nombre de courbes nécessaires pour rendre compte des séquences individuelles, on propose de ne visualiser que les trajectoires non imbriquées dans une séquence plus longue (voir figure 2 graphique de droite).

Définition. *Une séquence* $s1$ *débutant à la position* t_{s1} *est dite* imbriquée *dans une séquence* $s2$ *si est seulement si* $s2$ *contient de façon contiguë et alignée aux mêmes positions toute la séquence* $s1$*, c'est-à-dire, si* $s2$ *est de la forme* $s2 = sp \rightarrow s1 \rightarrow ss$*, avec des préfixes* sp *et suffixes* ss *éventuellement vides et le début de* $s1$ *en position* t_{s1}.

Par exemple, la séquence $(A)^1 \rightarrow (B)^2$ est imbriquée dans la séquence $(A)^1 \rightarrow (B)^2 \rightarrow (C)^3$, mais n'est pas imbriquée dans $(A)^1 \rightarrow (A, B)^2$. La notion de séquence imbriquée s'apparente à celle de sous-chaîne de caractères (*substring*), mais est plus générale puisqu'ici les caractères peuvent être des transactions ou transitions définies par des ensembles d'événements simultanés. Dans le cas où, comme dans l'exemple ci-dessus, les débuts de séquences sont tous alignés sur la 1ère position, il ne peut évidemment pas y avoir de préfixe sp.

Pour visualiser la présence de séquences imbriquées, on adapte l'épaisseur des points et segments communs indépendamment selon leur fréquence. Ainsi, dans le graphique de droite de la figure 2, on voit clairement qu'il existe des séquences $(A)^1 \rightarrow (B)^2$ imbriquées dans $(A)^1 \rightarrow (B)^2 \rightarrow (C)^3$.

Une difficulté est qu'une même séquence peut parfois s'imbriquer dans plusieurs séquences. La solution retenue est d'affecter son poids entièrement à une seule des séquences non imbriquables concernées, la plus fréquente. Alternativement, on pourrait répartir son poids de façon uniforme entre les diverses séquences non imbriquables concernées.

Si toutes les séquences sont de même longueur (même nombre de transactions dans le cas de séquences d'événements), les trajectoires distinctes et les trajectoires non-imbriquables sont les mêmes.

Notons que s'il est aisé de retracer les séquences individuelles à partir du graphique des séquences distinctes, la démarche demande un peu plus d'efforts dans le cas du graphique des séquences non-imbriquées. L'utilisateur doit dans ce cas repérer et interpréter les variations

d'épaisseurs des divers segments des trajectoires représentées, ce qui reste très intuitif. C'est là le prix à payer pour avoir un graphique plus synthétique sans perte d'information.

4 Illustrations

Anxiété après un séjour aux soins intensifs. La figure 3 présente les 26 trajectoires distinctes parmi les 149 observées. Les couleurs aident à les distinguer et l'épaisseur des traits reflète la fréquence d'observation des trajectoires, si bien qu'il apparaît immédiatement que, et heureusement, le cas le plus fréquent est celui d'individus éprouvant peu ou pas d'anxiété aux trois termes considérés. Comme il s'agit ici de séquences d'états exclusifs, on n'a pas de simultanéités et donc pas de traits verticaux. Par ailleurs, les états sont ordonnés avec la situation la moins désirable au haut de l'échelle. Comme les courbes tendent plutôt à descendre qu'à monter, cela traduit une baisse de l'anxiété au cours du temps. En particulier, on peut observer une fréquence relativement importante dans le carré nord-ouest correspondant aux patients fortement anxieux après une semaine. En suivant les trajectoires qui partent de ce carré, on voit que la plupart de ces individus se déclarent moins anxieux 6 et 12 mois plus tard, que rares sont ceux qui restent fortement anxieux 6 mois plus tard, et qu'aucun ne se déclare encore fortement anxieux après 12 mois.

Outre les séquences sans anxiété, les trajectoires les plus fréquentes identifiables par l'épaisseur des traits sont [(forte)^{t_1} $\rightarrow$ (faible)^{t_2} $\rightarrow$ faible)^{t_3}], [(moyenne)^{t_1} $\rightarrow$ (faible)^{t_2} $\rightarrow$ (faible)^{t_3}],

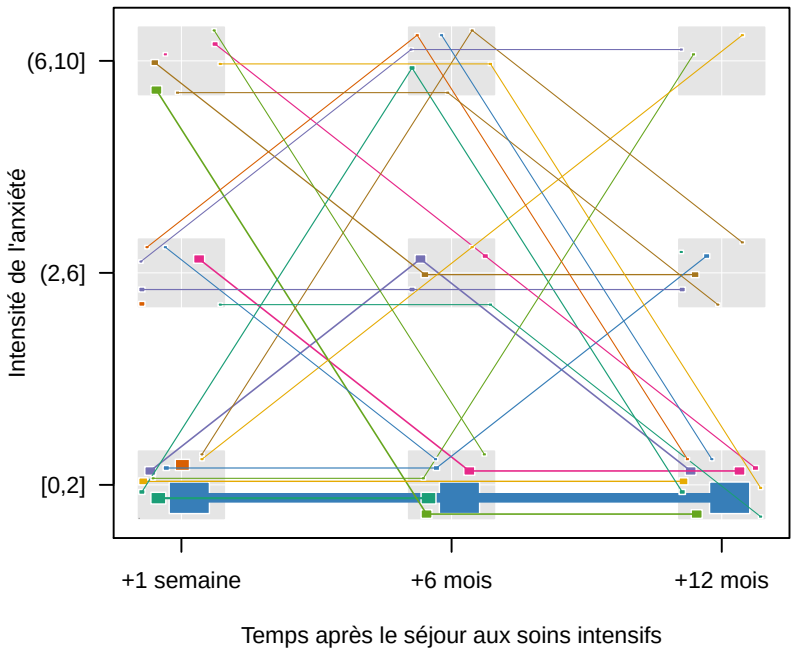

FIG. 3 – *Trajectoires distinctes du niveau d'anxiété de 149 patients.*

et [(faible)^{t_1} $\rightarrow$ (moyenne)^{t_2} $\rightarrow$ (faible)^{t_3}]. Enfin, avec un peu d'attention on peut voir des points-états isolés qui correspondent à des individus qui n'ont répondu qu'une fois, ainsi que des trajectoires de longueur deux, notamment la séquence fréquente [(faible)^{t_1} $\rightarrow$ (faible)^{t_2}], mais aussi une trajectoire plus atypique [(forte)^{t_2} $\rightarrow$ (faible)^{t_3}]. Cette dernière est imbriquée dans la séquence [(forte)^{t_1} $\rightarrow$ (forte)^{t_2} $\rightarrow$ (faible)^{t_3}], et n'apparaîtrait donc pas sous forme distincte dans un graphique des séquences non-imbriquables.

Evénements de vie familiale. Pour notre second exemple, le nombre de trajectoires est beaucoup plus conséquent, c'est pourquoi nous optons pour la représentation des séquences non-imbriquées. La figure 4 montre les 55 trajectoires non-imbriquées de vie familiale identifiables parmi les 2601 trajectoires considérées. A titre de comparaison, le nombre de trajectoires distinctes est de 130 et leur visualisation donnerait un graphique sensiblement plus confus. Le carré noir dans le coin sud-ouest reflète les 52 individus qui n'ont connu aucun événement. Pour faciliter l'identification des séquences les plus communes, les trajectoires dont un segment au moins a un support minimal donné (5% dans notre exemple) sont rendues en couleur, tandis que les trajectoires moins fréquentes apparaissent en gris clair en arrière-plan.

L'examen du graphique révèle que les trajectoires tendent à croître de façon monotone avec la position dans la séquence, ce qui démontre que l'ordre des événements retenu pour l'axe des y est en Suisse la norme, du moins pour les générations nées avant 1958. Deux configurations sont clairement les plus fréquentes : (en bleu) quitter le domicile parental, se mettre en ménage et se marier durant une même année, puis avoir un premier enfant une ou plusieurs années plus tard, et (en vert) quitter d'abord le domicile parental et quelques années plus tard se mettre en ménage et se marier une même année, puis, encore plus tard, avoir le premier enfant. Deux trajectoires un peu moins fréquentes sont de connaître ces quatre événements la même année, ou de les vivre l'un après l'autre. On remarque aussi, que parmi ceux qui ont suivi l'une des quatre trajectoires précédentes, le risque de divorcer est le plus important lorsque l'on expérimente les quatre événements la même année.

De façon non surprenante puisque le divorce doit obligatoirement être précédé d'un mariage, on voit qu'aucune trajectoire ne débute avec le divorce. Plus intéressant est le fait qu'un certain nombre d'individus se marient et/ou ont un premier enfant avant de quitter le domicile parental. On le voit, cette visualisation des séquences permet de rendre compte des normes sociales dans le séquencement des événements de vie tout en informant sur la diversité des situations observées.

5 Discussion

Les illustrations précédentes ont démontré l'utilité de notre représentation graphique pour l'exploration de séquences d'états comme de séquences d'événements. L'intérêt du graphique réside dans sa capacité à visualiser toutes les configurations de trajectoires observées et à faire ressortir clairement les plus courantes d'entre elles.

Nous nous proposons dans cette section, d'une part, d'examiner dans quelle mesure notre représentation graphique répond aux critères de Diggle et al. (2002) que nous avons rappelés en introduction, et d'autre part, de discuter les possibilités d'adapter le graphique aux spécificités des données pour en optimiser le rendu.

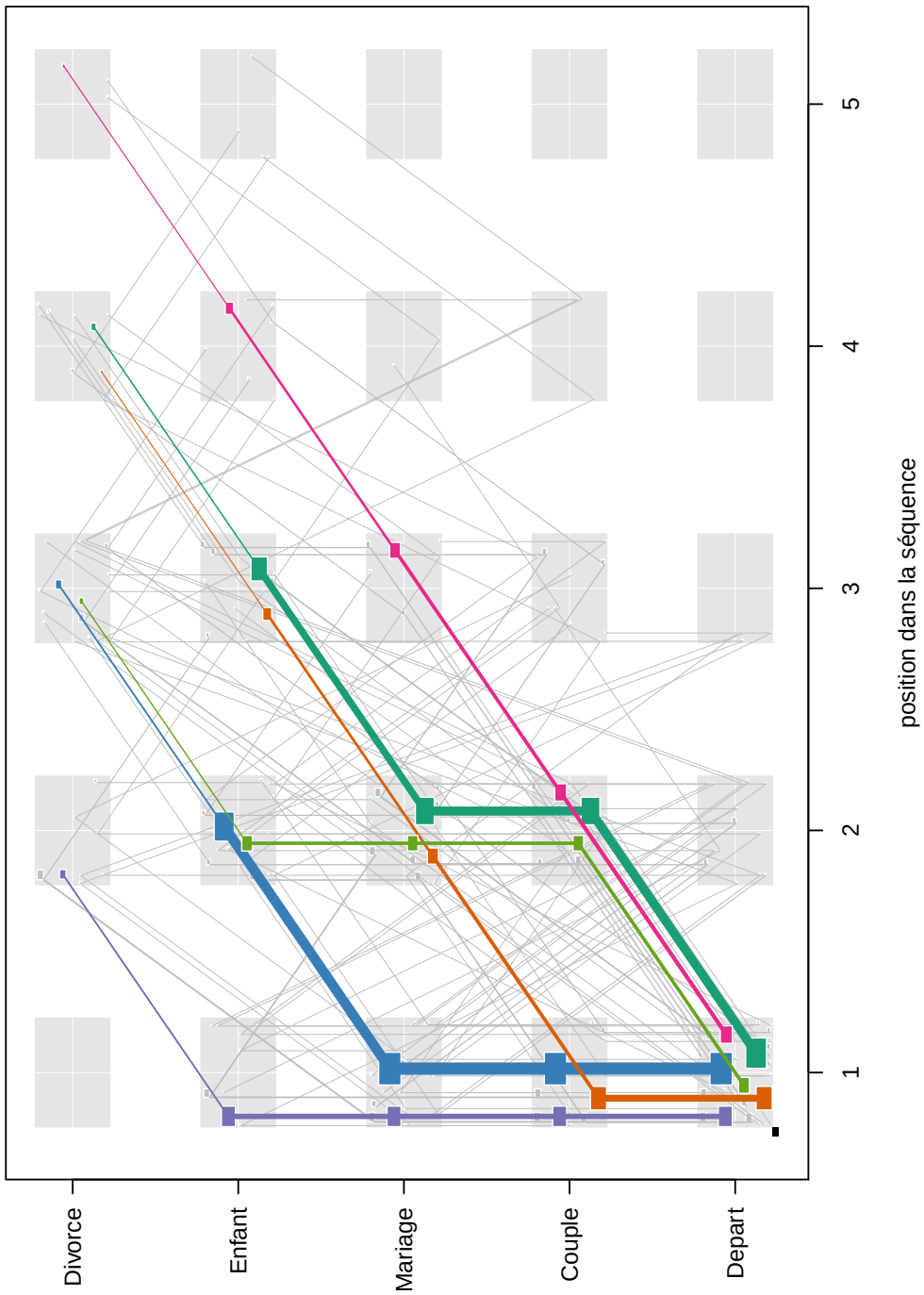

FIG. 4 – *Trajectoires non-imbriquables des 2601 parcours de vie familiale.*

5.1 Critères de Diggle et al. (2002)

Le chapitre 3 de l'ouvrage de Diggle et al. (2002) contient une discussion fort intéressante sur la visualisation de données longitudinales numériques. Les auteurs y énoncent quatre critères qu'un graphique de données longitudinales devrait satisfaire pour être efficient. Bien que notre proposition soit destinée à des données catégorielles plutôt que numériques, il est instructif de l'examiner à la lumière de ces quatre critères.

Montrer les données brutes pertinentes plutôt que des résumés synthétiques. La visualisation proposée conserve l'essentiel de l'information sur les trajectoires individuelles, puisque l'information perdue par l'agrégation des trajectoires identiques ou l'imbrication de trajectoires est restituée par l'épaisseur variable des traits et points correspondants. Il est vrai que telle que présentée la relation entre le graphique et les données n'est pas totalement biunivoque car la figure ne porte pas les identifiants des trajectoires et ne renseigne que de façon relative sur les fréquences des trajectoires distinctes ou segments de trajectoires imbriquées. Si nécessaire, on pourrait afficher ces informations en regard de quelques trajectoires d'intérêt.

Mettre en évidence les modèles de configuration potentiellement intéressants. L'épaisseur différenciée des points et traits permet précisément de distinguer les trajectoires les plus fréquentes. On pourrait évidemment envisager un autre critère d'intérêt que la fréquence, par exemple la centralité (somme des distances par rapport aux autres séquences), et utiliser l'épaisseur des traits pour le refléter.

Identifier les caractéristiques transversales et longitudinales de l'ensemble. Le graphique proposé vise prioritairement à rendre compte des caractéristiques longitudinales. Il visualise cependant aussi la diversité des trajectoires entre individus. De plus, bien que qu'un chronogramme puisse être plus approprié pour cela, il donne également une certaine idée de la distribution transversale des états ou événements en chaque position.

Faciliter l'identification des individus et observations atypiques. A nouveau, c'est ici la taille des points et l'épaisseur des traits qui permet de repérer les trajectoires peu fréquentes : ce sont celles qui sont reproduites avec les traits les plus fins. Comme pour l'identification des trajectoires d'intérêt potentiel, il serait aisé d'utiliser un autre critère que la fréquence pour juger de l'atypicité. Notons cependant que le repérage des lignes les plus fines demande plus d'efforts que l'identification des trajectoires les plus typiques. Notre implémentation prevoit une option pour ne colorier que les trajectoires les moins fréquentes et ainsi mieux les mettre en évidence. Quant à l'identification des individus, on pourrait imaginer d'afficher leur identificateur en regard des trajectoires les moins fréquentes, du moins tant que leur nombre reste limité. L'examen du nombre et de la taille des points représentant à chaque position les états ou événements observés permet d'identifier les états ou événements rares.

Au final, on peut donc affirmer que notre proposition est efficace au sens de Diggle et al. (2002).

5.2 Spécification du graphique et astuces pratiques

En expérimentant la visualisation proposée sur divers jeux de données, nous avons constaté qu'il reste difficile d'extraire des régularités lorsque le nombre de trajectoires distinctes est très

grand. Le résultat est en général d'autant plus brouillé que le nombre d'individus, la taille de l'alphabet des états ou événements, et la longueur possible des séquences sont grands. Plusieurs astuces ou ajustements relativement simples permettent d'améliorer sensiblement les représentations.

En premier lieu, il est souvent très instructif de partitionner les données selon, par exemple, le point de départ ou d'arrivée de la séquence, c'est-à-dire selon le premier ou dernier élément de la séquence. Ceci permet non seulement de réduire le nombre de séquences par graphique, mais aussi la diversité des début ou fin de séquences, ce qui en facilite l'interprétation.

Il est par ailleurs plus facile de repérer des évolutions monotones que des trajectoires irrégulières. Lorsque l'alphabet est nominal, il peut être utile de l'ordonner de sorte à assurer cette monotonie. C'est ce que nous avons fait dans notre illustration avec les événements de vie. Dans le cas d'un alphabet ordinal comme dans notre première illustration, il convient évidemment de respecter l'ordre naturel de l'alphabet.

Des variantes sont possibles également au niveau de l'axe des x. Les options sont ici soit d'aligner les événements ou états sur la position qu'ils occupent dans la séquence, ou alors prendre en compte explicitement la date ou la durée du processus (âge) et aligner sur la date ou la durée. Dans notre seconde illustration, nous avons délibérément ignoré l'information sur les dates d'occurrence des événements en raison de notre intérêt sur leur séquencement plutôt que sur les moments de la vie où ils se réalisent. La prise en compte des âges produit un graphique nettement moins lisible.

Un dernier facteur d'ajustement est lié au décalage des diverses trajectoires ou si l'on veut au positionnement des points dans la zone de translation. Ces positions sont actuellement déterminées aléatoirement mais une idée serait de les choisir de façon à minimiser les croisements.

6 Conclusion

La procédure d'exploration visuelle de données séquentielles proposée dans cet article vient complémenter les nombreux outils développés ces dernières années dans le domaine de la fouille de séquences et de l'analyse de données longitudinales. L'intérêt de l'outil, qui s'applique aussi bien à des séquences d'événements qu'à des séquences d'états, réside dans l'aide qu'il fournit à l'interprétation et à sa capacité à rendre compte simultanément des tendances principales et de la diversité entre individus.

Les démographes et sociologues qui ont eu l'occasion d'expérimenter le graphique sur leurs données y ont vu un outil d'emploi facile, de lecture simple et ouvrant des perspectives nouvelles en particulier pour l'étude de la déstandardisation des parcours de vie. Le graphique sera prochainement disponible dans le cadre de la librairie R TraMineR (Gabadinho et al., 2011).

Les développements futurs devraient permettre d'automatiser l'optimisation du placement des diverses trajectoires dans la zone de translation ainsi que de l'ordre de l'alphabet, ou encore l'ordre de présentation d'événements simultanés.

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.

Berchtold, A. et A. E. Raftery (2002). The mixture transition distribution model for high-order Markov chains and non-gaussian time series. *Statistical Science 17*(3), 328–356.

Diggle, P. J., P. Heagerty, K.-Y. Liang, et S. L. Zeger (2002). *Analysis of Longitudinal Data* (2nd ed.). Oxford: Oxford University Press.

Frees, E. W. (2004). *Longitudinal and Panel Data: Analysis and Applications in the Social Sciences*. New York: Cambridge University Press.

Gabadinho, A., G. Ritschard, N. S. Müller, et M. Studer (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software 40*(4), 1–37.

Hébrail, G. et H. Cadalen (2000). Visualisation et classification automatique de parcours professionnels. In *Actes des XXXIIe Journées de statistique, Fès, Maroc*, pp. 458–462.

Hox, J. (2010). *Multilevel Analysis: Techniques and Applications* (2nd ed.). Abingdon UK: Routledge.

Jeitziner, M.-M., V. Hantikainen, A. Conca, et J. Hamers (2011). Long-term consequences of an intensive care unit stay in older critically ill patients: Design of a longitudinal study. *BMC Geriatrics 11*(52), 1–7.

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.

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology 60*, 577–605.

Studer, M., N. S. Müller, G. Ritschard, et A. Gabadinho (2010). Classer, discriminer et visualiser des séquences d'événements. *Revue des nouvelles technologies de l'information RNTI E-19*, 37–48.

Tufte, E. R. (2001). *The Visual Display of Quantitative Information* (2nd ed.). Cheshire CT: Graphics Press.

Wongsuphasawat, K., J. A. G. Gómez, C. Plaisant, T. D. Wang, M. Taieb-Maimon, et B. Shneiderman (2011). LifeFlow: Visualizing an overview of event sequences. In *Proceedings of the 2011 annual conference on Human Factors in Computing Systems (CHI), Vancouver, Canada, May 7-12, 2011*, pp. 1747–1756. New York: ACM.

Summary

The article introduces an original graphical display for categorical longitudinal data. The visualisation, inspired from the multiple time-series plot, particularly suits to descriptive and exploratory analyses of individual trajectories defined as event sequences. The article includes a description of the visualisation method and of its founding principles, application examples, and a discussion of the properties of the resulting plots. In addition, we explain fine-tuning specifications for optimally rendering given data.