Indice de complexité pour le tri et la comparaison de séquences catégorielles

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

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

Résumé. Cet article [1] propose un nouvel indice de la complexité de séquences catégorielles. Bien que conçu pour des séquences représentant des trajectoires biographiques telles que celles rencontrées dans les sciences sociales, il s'applique à tous types de listes ordonnées d'états. L'indice prend en compte deux aspects distincts, soit la complexité induite par l'ordonnancement des états successifs qui est mesurée par le nombre de transitions (changements d'état) et la complexité liée à la distribution des états dont rend compte l'entropie.

1 Introduction

Dans les sciences sociales, les séquences catégorielles sont des listes ordonnées d'états ou d'événements décrivant typiquement des trajectoires ou parcours de vie, familiale ou professionnelle par exemple. Dans ce contexte, il importe de pouvoir distinguer les trajectoires selon leur complexité. A cette fin, deux mesures ont été considérées dans la littérature, d'une part l'*entropie* de la distribution des divers états dans une séquence (Gabadinho et al., 2009; Widmer et Ritschard, 2009) et d'autre part la *turbulence* (Elzinga et Liefbroer, 2007). La première mesure présente l'inconvénient de ne pas tenir compte du séquencement des états. La seconde, sensible à cet ordonnancement, reste difficile à interpréter car s'appuyant sur le concept peu intuitif du nombre de sous-séquences distinctes qui peuvent être extraites de la séquence et sur la variance des durées de séjour dans chacun des états. Nous proposons ici un nouvel indice qui combine l'entropie avec un indicateur de la complexité de l'ordonnancement des états. Bien que conçu pour des séquences décrivant des parcours de vie, l'indice caractérise utilement tout type de séquence d'états.

2 Données et définitions

Une séquence de longueur ℓ est une liste ordonnée de ℓ éléments choisis successivement dans un ensemble fini A de taille $|A|$ appelé *alphabet*. Pour illustrer notre propos et comparer le comportement de l'indice proposé nous considérons un jeu de séquences type, 3 jeux de

1. Travail réalisé avec le soutien financier du Fonds national suisse, subside FN-100015-122230.

Index	Sequence	nd	nt	ϕ	v	h	T	C
1	A-A-A-A-A-A-A-A-A-A-A-A	1	0	2	0.00	-0.00	1.00	-0.00
2	A-A-A-A-A-A-A-A-A-A-A-B	2	1	4	50.00	0.21	2.00	0.14
3	A-A-A-A-A-A-B-B-B-B-B-B	2	1	4	0.00	0.50	6.70	0.21
4	A-A-A-A-A-B-B-A-A-A-A-A	2	2	7	3.00	0.33	5.47	0.24
5	A-B-B-B-B-B-B-B-B-B-B-A	2	2	7	27.00	0.33	2.81	0.24
6	A-B-A-A-A-A-A-A-A-A-B-A	2	4	20	9.80	0.33	4.32	0.34
7	A-B-A-B-A-B-A-B-A-A-A-A	2	8	143	1.00	0.46	7.16	0.58
8	A-B-A-B-A-B-A-B-A-B-A-B	2	11	609	0.00	0.50	9.25	0.71
9	A-A-A-A-A-A-A-A-A-A-B-C	3	2	8	27.00	0.41	3.00	0.27
10	A-A-A-A-B-B-B-B-C-C-C-C	3	2	8	0.00	0.79	7.25	0.38
11	A-A-A-A-A-B-C-A-A-A-A-A	3	3	15	5.33	0.41	5.29	0.33
12	A-B-B-B-B-B-C-C-C-C-C-A	3	3	15	5.33	0.74	5.29	0.45
13	A-B-C-A-A-A-A-A-A-C-B-A	3	6	92	3.57	0.63	6.52	0.58
14	A-B-C-A-B-C-A-B-A-A-A-A	3	8	302	1.00	0.69	8.24	0.71
15	A-B-C-B-A-C-A-C-B-C-B-A	3	11	1572	0.00	0.79	10.62	0.89
16	A-B-C-A-B-C-A-B-C-A-B-C	3	11	2031	0.00	0.79	10.99	0.89
17	A-A-A-A-A-A-A-A-A-B-C-D	4	3	16	16.00	0.60	4.00	0.41
18	A-A-A-B-B-B-C-C-C-D-D-D	4	3	16	0.00	1.00	7.70	0.52
19	A-A-A-A-B-C-D-A-A-A-A-A	4	4	31	3.80	0.60	6.08	0.47
20	A-B-B-B-C-C-C-C-D-D-D-A	4	4	31	1.80	0.98	6.81	0.60
21	A-B-C-D-A-A-A-A-D-C-B-A	4	8	396	1.00	0.90	8.63	0.81
22	A-B-C-D-A-B-C-D-A-A-A-A	4	8	432	1.00	0.90	8.75	0.81
23	A-B-C-D-B-A-C-D-A-C-B-D	4	11	2898	0.00	1.00	11.50	1.00
24	A-B-C-D-A-B-C-D-A-B-C-D	4	11	3096	0.00	1.00	11.60	1.00

TAB. 1 – *Séquences type et caractéristiques : nombre d'états distincts (nd), nombre de transitions (nt), nombre de sous-séquences dans la DSS (ϕ), variance des durées de séjour (v), entropie normalisée (h), turbulence (T), indice de complexité (C).*

séquences générées artificiellement et un jeu de données réelles. Les séquences type sont celles du tableau 1. Elles sont construites à partir d'un alphabet de 4 états (A,B,C et D) et contiennent chacune 12 états, mais diffèrent selon le nombre d'états distincts, l'ordre et les fréquences de ces états. Les trois jeux artificiels comptent chacun 1000 séquences générées à partir du même alphabet. Les deux premiers, A1 et A2, résultent du tirage aléatoire de chaque état de la séquence [2] et le troisième (A3) d'un modèle de Markov d'ordre 1 avec des probabilités élevées (>.9) de transition d'un état vers lui-même.

Les données réelles (*mvad*) sont celles de l'étude de McVicar et Anyadike-Danes (2002) sur la transition entre formation et emploi en Irlande du Nord. Les séquences représentent le suivi mensuel de 712 jeunes à partir de la fin de la scolarité obligatoire et sont constituées de 70 états mensuels. L'alphabet contient les états EM (en emploi), FE (formation complémentaire), HE (formation supérieure), JL (au chômage), SC (école), TR (en stage).

2. le premier (A1) avec des probabilités identiques pour chacun des états, le second (A2) avec p(A)=0.42, p(B)=0.16, p(C)=0.10, p(D)=0.32

3 Complexité de séquences d'états

La mesure de la complexité de séquences catégorielles est abordée dans de nombreux domaines tels que la biologie ou l'informatique. Nous souhaitons ici définir et quantifier plus particulièrement la *complexité* de séquences catégorielles représentant des biographies.

Etats distincts successifs La premier aspect que nous retenons pour caractériser la complexité est le séquencement, c'est-à-dire l'ordre d'apparition des états. Soit $s = x_1 x_2 ... x_\ell$ une séquence d'états de longueur ℓ. La *séquence des états distincts successifs* (DSS pour '*Distinct Sucessive States*') est obtenue en ne retenant qu'une seule des t_x occurrences consécutives d'un même état x. Si les séquences sont définies sur une échelle temporelle, t_x est la *durée de séjour*[3] dans l'état x. Une séquence d'états peut ainsi être représentée par un ou plusieurs couples (x, t_x) où x est un état et t_x la durée de séjour associée. Les séquences type 1 et 7 deviennent ainsi :

```
[1] (A,12)
[7] (A,1)-(B,1)-(A,1)-(B,1)-(A,1)-(B,1)-(A,1)-(B,1)-(A,4)
```

La première séquence peut être représentée à l'aide d'un seul couple (x, t_x) alors qu'il en faut 9 pour représenter la seconde. Comme pour la complexité algorithmique de Kolmogorov nous considérons ici la longueur de la description d'un objet comme indicateur de sa complexité. Le nombre de couples (x, t_x) nécessaires pour représenter une séquence est équivalent à la longueur ℓ_d de la séquence des états distincts successifs (DSS), avec $\ell_d \leq \ell$. Ce nombre est lié au nombre $nt = \ell_d - 1$ de transitions — passages d'un état à un autre — contenues dans la séquence (tableau 1), avec $0 \leq nt \leq \ell - 1$.

Fréquence des états Le nombre d'occurrences des états dans une séquence est un autre aspect de la complexité. Les séquences type 2 et 3 sont de complexité égale selon le précédent critère, mais la distribution des états de l'alphabet dans la séquence (les durées totales passées dans chacun des états) est différente :

```
     A B C D
[2] 11 1 0 0
[3]  6 6 0 0
```

L'entropie de Shannon permet de capturer cet aspect de la complexité. Elle est définie par $h(s) = -\sum_{i=1}^{|A|} \pi_i \log \pi_i$ où π_i est la proportion d'occurrence du $i$ème état de l'alphabet dans la séquence s. La séquence 1 composée d'un seul état, présente une entropie minimale (tableau 1). L'entropie est maximale lorsque le nombre d'occurrences de chacun des états est égal (séquences 18 et 24). Ce maximum, utilisé pour normaliser l'entropie, vaut $h_{max} = -\log(1/|A|)$ et est donc indépendant de ℓ.

4 Mesure composite de la complexité

Le nombre de transitions et l'entropie d'une séquence mesurent des aspects distincts (voir figure 1-a). L'entropie ne rend pas compte du séquencement des états : les séquences type 10 et

3. Ce terme est utilisé par la suite pour désigner t_x.

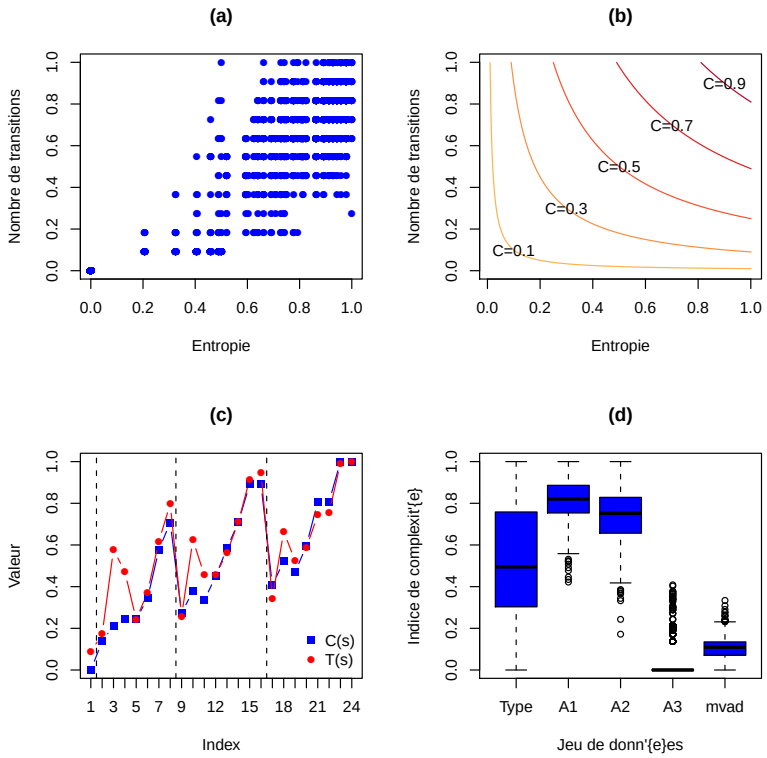

FIG. 1 – *(a) Relation entre les deux composantes de C(s) pour les jeux "type", A1, A2 et A3 ; (b) Valeurs théoriques des deux composantes de $C(s)$; (c) Valeurs de C(s) et de la turbulence les séquences type ; (d) Distribution de C(s).*

16 ont par exemple une entropie identique. Inversement, un nombre de transitions donné peut correspondre à des valeurs d'entropie variées (séquences 8 et 24). Nous proposons une mesure composite qui rend compte simultanément de ces deux aspects.

Définition L'*indice de complexité* $C(s)$ est la moyenne géométrique de l'entropie longitudinale $h(s)$ et du nombre $nt(s) = \ell_d(s) - 1$ de transitions dans la séquence s, chacun de ces termes étant normalisé, soit

$$C(s) = \sqrt{\frac{nt(s)}{(\ell(s)-1)} \frac{h(s)}{h_{max}}}$$

où h_{max} est le maximum théorique de l'entropie et $\ell(s)$ est la longueur de la séquence.

Propriétés Par construction on a $0 \leq C(s) \leq 1$. $C(s) = 0$ pour une séquence composée d'un seul état distinct, car alors $h(s) = 0$ et $nt(s) = 0$. $C(s) = 1$ si (i) le nombre de transitions dans la séquence est égal au maximum $\ell - 1$ et si (ii) chacun des états apparaît $\ell/|A|$ fois dans la séquence. Notre indice est plus facilement interprétable que la *turbulence* $T(s)$ (Elzinga

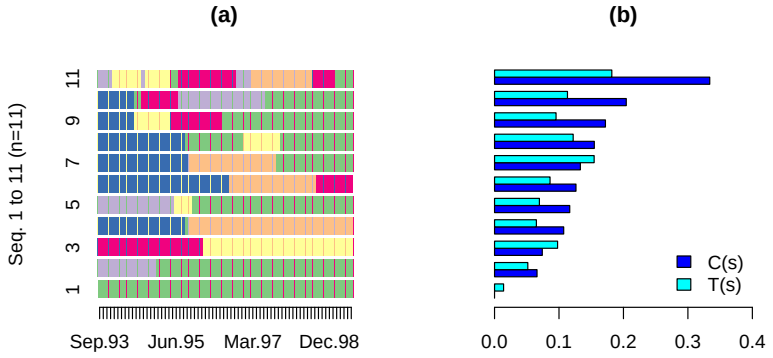

FIG. 2 – *Données* mvad *: (a) Séquences sélectionnées et triées selon la valeur de l'indice de complexité ; (b) valeurs correspondantes de la turbulence (normalisée) et de l'indice de complexité.*

et Liefbroer, 2007), autre mesure composite qui combine le nombre $\phi(s)$ de sous-séquences distinctes de la DSS et la variance $v_t(s)$ des durées de séjour dans les états successifs (tableau 1). Il est aussi beaucoup plus simple et donc plus rapide à calculer.

La figure 1-b présente les combinaisons de valeurs normalisées de $h(s)$ et $nt(s)$ qui produisent une valeur donnée de $C(s)$. Ces deux mesures ne sont cependant pas totalement indépendantes. Pour deux séquences $s_{nt=\ell-1}$ et $s_{nt=1}$ de longueur ℓ contenant respectivement $\ell-1$ et 1 transitions, et donc au moins deux états distincts, on a [4]

$$h(s_{nt=\ell-1}) \geq -(\frac{2}{|A|}\log\frac{1}{|A|}) \geq h(s_{nt=1}) \geq -\Big(\frac{1}{\ell}\log\frac{1}{\ell} + \frac{\ell-1}{\ell}\log\frac{\ell-1}{\ell}\Big)$$

Expérimentation On note les valeurs surévaluées de $T(s)$ lorsque la complexité de "séquencement" est faible et la variance des durées de séjour nulle (séquences 3, 10 et 18, [5] tableau 1 et figure [6] 1-c). La complexité est globalement la plus élevée pour le jeu de séquences aléatoires A1 généré avec des probabilités uniformes pour les états de l'alphabet (graphique 1-d), car l'entropie et le nombre moyen de transitions sont logiquement élevés. La "prévisibilité" augmente pour les séquences générées par le modèle de Markov (jeu A3) et en conséquence la complexité diminue fortement.

Pour les données réelles *mvad*, nous avons sélectionné et représenté 10 séquences correspondant aux déciles de leur indice de complexité $C(s)$ (figure 2). Nous pouvons voir entre autre comment l'augmentation de $C(s)$ est liée à l'augmentation du nombre d'états distincts dans la séquence. La séquence la moins complexe ne contient qu'un seul état de l'alphabet, alors que la plus complexe en contient 5 (sur 6). Une régression de l'indice sur le sexe et le niveau de réussite à la fin de scolarité obligatoire montre que la complexité est significative-

4. Pour $|A| = 4$ et $\ell = 12$, $h(s_{nt=11})/h_{max} \geq 0.5 \geq h(s_{nt=1})/h_{max} \geq 0.21$ (voir tableau 1)
5. L'entropie diffère pour ces trois séquences alors que la variance des durées de séjour est identique
6. La figure reporte la turbulence normalisée par son maximum théorique.

ment inférieure pour les hommes et supérieure pour les individus ayant obtenu les meilleures résultats en fin de scolarité obligatoire.

5 Conclusion

Nous avons identifié deux aspects principaux de la complexité de séquences d'états pertinents pour l'analyse de trajectoires biographiques : le séquencement et la distribution des états. Nous avons introduit un nouvel indice combinant ces aspects. Cet indice est d'interprétation plus claire que la mesure de turbulence proposée initialement, dont le comportement est par ailleurs parfois contre-intuitif. Il est également plus simple et donc plus rapide à calculer. L'indice s'avère en particulier utile pour ordonner des trajectoires biographiques et étudier la relation entre leur complexité et un ensemble de variables.

Références

Elzinga, C. et A. Liefbroer (2007). De-standardization of family-life trajectories of young adults : A cross-national comparison using sequence analysis. *European Journal of Population/Revue européenne de Démographie 23*(3), 225–250.

Gabadinho, A., G. Ritschard, M. Studer, et N. 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.

McVicar, D. et M. Anyadike-Danes (2002). Predicting successful and unsuccessful transitions from school to work by using sequence methods. *Journal of the Royal Statistical Society. Series A (Statistics in Society) 165*(2), 317–334.

Widmer, E. et G. Ritschard (2009). The de-standardization of the life course: Are men and women equal? *Advances in Life course Research 14*(1-2), 28–39.

Summary

This paper introduces a complexity index for categorical state sequences. Though, the index is more specifically intended for measuring the complexity of sequences describing biographical trajectories in social sciences, it applies to all kind of ordered lists of states. The measure accounts for two distinct aspects of complexity: the complexity of the sequencing of the states captured by the number of transitions and the diversity of states in the sequence measured with Shannon's entropy.