Revue de psychoéducation

Introduction à l'analyse de séquence et illustration de son application en sciences sociales à partir de patrons de transitions de l'école au travail

Introduction to sequence analysis and its application in the social sciences using school-to-work transition patterns

Éliane Thouin, Clémentine Courdi, Elizabeth Olivier, Véronique Dupéré, Anne-Sophie Denault et Éric Lacourse

Volume 51, numéro 2, 2022

URI : https://id.erudit.org/iderudit/1093470ar
DOI : https://doi.org/10.7202/1093470ar

Aller au sommaire du numéro

Éditeur(s)

Revue de Psychoéducation

ISSN

1713-1782 (imprimé)
2371-6053 (numérique)

Découvrir la revue

Citer cet article

Thouin, É., Courdi, C., Olivier, E., Dupéré, V., Denault, A.-S. & Lacourse, É. (2022). Introduction à l'analyse de séquence et illustration de son application en sciences sociales à partir de patrons de transitions de l'école au travail. *Revue de psychoéducation*, *51*(2), 427–449. https://doi.org/10.7202/1093470ar

Résumé de l'article

Plusieurs champs de recherche en psychoéducation, en psychologie développementale et en sociologie visent à examiner les changements individuels et au sein de groupes dans la population. On cherche, par exemple, à identifier les parcours typiques lors de périodes développementales précises, comme la délinquance à l'adolescence et la transition de l'école au travail au début de l'âge adulte, afin de comprendre celles qui paraissent plus ou moins adaptatives ou optimales. Cet article a comme objectif de présenter une approche algorithmique permettant de tracer de tels parcours sous forme de séquences à partir de variables catégorielles/nominales, comme dont des statuts (p. ex. : occupationnels, maritaux), des états (p. ex. : de santé) ou la présence de comportements (p. ex. : consommation ou pas). Cette approche, l'analyse de séquences, est abondamment utilisée par les chercheurs européens, mais demeure peu connue en Amérique du Nord. L'article présente les fondements et l'application de cette approche analytique, en décrivant chacune des étapes de l'analyse à partir d'un exemple fictif tiré de banques de données portant sur le passage de l'adolescence à l'âge adulte. L'article conclut par une discussion rapportant les forces et les limites de l'analyse de séquences dite algorithmiques en sciences sociales. Le script utilisé pour réaliser les analyses de cet article est également fourni en ligne pour les lecteurs intéressés par cette technique analytique.

Revue de psychoéducation
Volume 51, numéro 2, 2022, 427-449

Mesure et évaluation

Introduction à l'analyse de séquence et illustration de son application en sciences sociales à partir de patrons de transitions de l'école au travail

Introduction to sequence analysis and its application in the social sciences using school-to-work transition patterns

É. Thouin[1]
C. Courdi[2]
E. Olivier[3]
V. Dupéré[1]
A.-S. Denault[4]
É. Lacourse[2]

1 École de psychoéducation, Université de Montréal
2 Département de sociologie, Université de Montréal
3 Département de psychopédagogie et d'andragogie, Université de Montréal
4 Département des fondements et des pratiques en éducation, Université Laval

Correspondance :

Éliane Thouin
École de psychoéducation, Université de Montréal
C.P. 6128, Succ. Centre-Ville
Montréal (Québec) H3C 3J7

eliane.thouin@umontreal.ca

Résumé

Plusieurs champs de recherche en psychoéducation, en psychologie développementale et en sociologie visent à examiner les changements individuels et au sein de groupes dans la population. On cherche, par exemple, à identifier les parcours typiques lors de périodes développementales précises, comme la délinquance à l'adolescence et la transition de l'école au travail au début de l'âge adulte, afin de comprendre celles qui paraissent plus ou moins adaptatives ou optimales. Cet article a comme objectif de présenter une approche algorithmique permettant de tracer de tels parcours sous forme de séquences à partir de variables catégorielles/nominales, comme dont des statuts (p. ex. : occupationnels, maritaux), des états (p. ex. : de santé) ou la présence de comportements (p. ex. : consommation ou pas). Cette approche, l'analyse de séquences, est abondamment utilisée par les chercheurs européens, mais demeure peu connue en Amérique du Nord. L'article présente les fondements et l'application de cette approche analytique, en décrivant chacune des étapes de l'analyse à partir d'un exemple fictif tiré de banques de données portant sur le passage de l'adolescence à l'âge adulte. L'article conclut par une discussion rapportant les forces et les limites de l'analyse de séquences dite algorithmiques en sciences sociales. Le script utilisé pour réaliser les analyses de cet article est également fourni en ligne pour les lecteurs intéressés par cette technique analytique.

Mots-clés : analyse de séquences; appariement optimal; trajectoires et parcours longitudinaux; approche algorithmique.

Abstract

Several areas of research in psychoeducation, developmental psychology, and sociology are aimed at examining individual and group changes in the population. For example, they seek to identify typical pathways during

specific developmental periods, such as delinquency in adolescence and the school to work transition in early adulthood, in order to understand those that appear maladaptive or optimal. The purpose of this paper is to present an algorithmic approach for tracing such pathways as sequences based on categorical/nominal variables, like statuses (e.g., occupational, marital), states (e.g., health), or behaviors (e.g., drinking habits). This approach, named sequence analysis, is widely used by European researchers, but remains little known in North America. This article presents the foundations and application of this analytical approach, describing each step of the analysis using a fictitious example drawn from data banks on the transition from adolescence to adulthood. The article concludes with a discussion of the strengths and limitations of sequence analysis in the social sciences. The script used to perform the analyses in this article is also provided online for readers interested in this analytical technique.

Keywords: sequence analysis; optimal matching; longitudinal pathways and trajectories; algorithmic approach

Que ce soit à l'enfance, à l'adolescence ou à l'âge adulte, les parcours de vie des individus sont marqués par des périodes de stabilités et de changements qui peuvent prendre différentes formes tout au long du développement. De manière générale, on peut classer les périodes de changement en deux grandes catégories : les changements continus et les changements catégoriels, également nommés changements par stade (Elder et Shanahan, 2006). Les changements continus se définissent par une augmentation ou une diminution des comportements, affects ou cognitions au fil du temps, comme la croissance physique à l'enfance ou la dégradation des habiletés de mémorisation chez les adultes d'âge avancé. Les changements catégoriels (ou par stade) concernent plutôt une transformation significative et substantielle entre deux états, par exemple, passer du statut de célibataire à conjoint à l'âge adulte, ou encore changer de mode de consommation de substances psychoactives à l'adolescence, avec des périodes marquées par l'usage de cannabis et d'autres par l'usage de drogues de synthèse.

Dans les travaux de recherche s'inscrivant dans une perspective de psychologie développementale, le type de changement à l'étude doit être pris en compte puisqu'il influencera le choix des méthodes statistiques à privilégier. Pour les changements continus, des techniques basées sur des modèles probabilistes (*model-based)*, telles que les analyses de courbes de croissances (*latent growth curve model*), des analyses de cheminements (*path analysis*) et des analyses de régressions à décalage croisé (*autoregressive cross-lagged*) sont souvent mises de l'avant (Dupéré et al., 2007; Morin et Litalien, 2019). Pour les changements catégoriels, les analyses disponibles peuvent paraître plus limitées, et moins souvent appliquées en recherche en sciences sociales, plus particulièrement en psychoéducation. En effet, même si les ouvrages de référence en statistiques font mention d'approches analytiques dont les modèles de Markov (voir p.ex. : Laditka et Laditka, 2015) et les analyses de transitions latentes ou à mesures répétées (voir p.ex. : Maas et al., 2019), elles sont plus rarement utilisées dans les recherches empiriques longitudinales. Pourtant, de nombreux phénomènes psychosociaux changent de façon catégorielle à travers le temps, comme les statuts occupationnels entre l'adolescence à l'âge adulte (p.ex. : étudiant → travailleur), les multiples rôles joués au sein des conflits d'intimidation pendant l'enfance (p.ex. : intimidateur →

victime → témoin) ou encore l'évolution des placements des jeunes sous la loi de la protection de la jeunesse (p.ex. : milieu familial → centre d'hébergement → foyer d'accueil).

Le présent article a pour objectif de présenter une approche algorithmique (une épistémologie analytique différente de celle basée sur des modèles probabilistes) permettant de capter ce type de changement à travers le temps. Il s'agit de l'analyse de séquences (AS), qui est plus couramment utilisée en sciences sociales en Europe (ex. : anthropologie et sociologie), mais encore peu connues en Amérique du Nord et ce, plus particulièrement en psychoéducation et en psychologie au Québec. Le texte débute avec une brève présentation des fondements des AS, puis se poursuit avec d'une explication détaillée de la démarche à suivre pour les appliquer à partir du logiciel gratuit et libre d'accès R et de la bibliothèque de code TraMineR. En conclusion, une synthèse est proposée, résumant les forces et les limites des AS en sciences sociales.

Fondements des AS en sciences sociales

Initialement dérivées d'études en bio-informatique ayant comme objectif de mesurer la correspondance entre des chaines d'ADN (Kruskal, 1983), les AS ont été introduites une première fois en sciences sociales dans le cadre de travaux de recherche sociologiques évaluant les degrés de similitude entre des rituels de danse (Abbott et Forrest, 1986) et des carrières musicales (Abbott et Hrycak, 1990). Plus de trente ans plus tard, elles font aujourd'hui partie des approches les plus couramment utilisées en Europe pour étudier les parcours de vie (Han et al., 2017). Elles ont été maintes fois utilisées pour examiner la transition de l'école au travail (Lorentzen et al., 2019; Ng-Knight et Schoon, 2017; Ranasinghe et al., 2019), la formation des couples (Pollock, 2007; Struffolino et al., 2016) ou encore l'atteinte des marqueurs importants du début de l'âge adulte (Schoon, 2015).

Les AS se réalisent à partir de séquences de variables catégorielles conceptualisées en tant que liste d'éléments suivant un ordre chronologique. Cet ordre fait toujours référence au temps qui passe, sans toutefois inclure explicitement une variable représentant une mesure du temps (Brzinsky-Fay et al., 2006). Autrement dit, contrairement aux analyses de trajectoires latentes, par exemple, qui intègrent généralement une variable pour tenir compte du temps, le temps est implicitement reconnu comme l'ordre de la liste d'éléments dans les analyses de séquences. Les éléments, quant à eux, représentent soit des statuts (p. ex. : statut d'emploi, statut marital), des états (p. ex. : état de santé) ou des comportements (p. ex. : mouvements de danse, type d'activité criminelle) pouvant prendre différentes formes au fil du temps. Ces éléments sont positionnés à un endroit précis sur les séquences (p. ex. : au 16e mois ou à la 80e minute) et organisés dans un ordre spécifique considéré non-interchangeable. Au moment de faire les analyses, les séquences de chaque individu sont ainsi incorporées à titre d'entités uniques et fixées sur la liste, ce qui permet de préserver l'ordre dans lequel se succèdent les éléments, de même que leur durée et leur moment d'apparition (*timing*). Les calculs visant à comparer les séquences tiennent donc compte de ces caractéristiques importantes pour mesurer à quel point les séquences sont similaires ou dissimilaires l'une par rapport à l'autre.

Tel que mentionné en introduction, sur le plan opérationnel, les AS s'appuient sur une approche algorithmique et heuristique pour dériver des trajectoires. Elles utilisent donc une épistémologie statistique différente de celle généralement retrouvée dans les méthodes longitudinales plus communément appliquées en sciences sociales (p.ex. : analyses de trajectoires ou de transition latente, régressions à décalage croisé), soit l'approche probabiliste qui mise sur les paramètres de distribution et de covariations de variables continues pour estimer l'hétérogénéité non-observée au sein des populations et faire des inférences. Les AS se déploient plutôt en trois grandes étapes analytiques : (1) organiser les éléments en séquences longitudinales, (2) comparer les séquences entre elles pour établir leur degré de similarité/dissimilarité et (3) regrouper les séquences similaires par analyses d'appariement (*clusters*) dans des groupes qui seront mutuellement exclusifs (Abbott, 1995; Ritschard et Studer, 2018). L'objectif final est de créer une typologie de séquences représentant des parcours longitudinaux et estimer leurs proportions. Ces groupes de séquences peuvent ensuite être soumis à des analyses complémentaires visant à examiner l'hétérogénéité de ces parcours et leurs liens avec des antécédents ou des conséquences pouvant survenir plus tard dans le développement.

Application des AS en sciences sociales

Chaque grande étape analytique des AS sera expliquée en utilisant, à titre d'exemple, des données tirées d'un projet de recherche portant sur la transition de l'école au travail auprès de jeunes en situation de vulnérabilité (Dupéré et al., 2018; Thouin, 2022). Ce projet comprend un échantillon de 545 participants initialement interviewés en personne une première fois à l'adolescence alors qu'ils étaient en fin de scolarité secondaire ($M_{âge}$ = 16 ans). Ils ont ensuite été recontactés quatre ans plus tard pour une seconde entrevue ($M_{âge}$ = 20 ans), dans laquelle ils étaient invités à énumérer tous les emplois et les programmes scolaires occupés depuis leur entrevue initiale au secondaire. Au total, 386 jeunes ont participé à cette seconde phase d'entrevue, représentant un taux de rétention de 70 % (pour plus de détails sur ce projet, voir Dupéré et al., 2018; Thouin 2022).

Tel que mentionné en introduction, les instructions pour réaliser des AS correspondront à celles offertes sur le logiciel R[1], à partir des bibliothèques de code TraMineR (Gabadinho et al., 2011) et WeightedCluster (Studer, 2013). Bien qu'il soit également possible d'effectuer des AS en utilisant le logiciel Stata avec les bibliothèques de code SQ-ADOs (Brzinsky-Fay et al., 2006) ou SQ (Kohler et al., 2020), la méthode de TraMineR a été retenue puisqu'elle offre davantage de flexibilité et de fonctions, en plus d'être accessible gratuitement sur le web par le biais de la communauté R. Chaque composante du script utilisé sera présentée au fur et à mesure dans le texte. Le script est également disponible dans son ensemble avec des explications sur la plateforme Github (https://anonymous.4open.science/r/Analyse-sequence-416E/README.md).

[1] Disponible gratuitement sur le web au : https://www.r-project.org/. Pour une explication du fonctionnement du logiciel, voir : https://catalogue.edulib.org/fr/cours/UMontreal-ISDS/

Première étape de la stratégie analytique : organisation des données sous forme de séquences

La première étape pour réaliser des AS est de s'assurer que les données soient organisées de façon séquentielle (Brzinsky-Fay et al., 2006; Ritschard et Studer, 2018). Pour ce faire, il faut d'abord identifier l'élément initial avec lequel débuteront les séquences, qui peut être un âge (p. ex. : 16 ans), un événement (p. ex. : décrochage scolaire, premier mariage) ou une date (p. ex. : 1990). La longueur des séquences doit ensuite être établie en comptabilisant le nombre d'unités temporelles (p. ex. : diurnes, mensuelles ou annuelles) suivant l'élément initial. Généralement, une longueur minimale de 25 unités temporelles est conseillée pour favoriser la stabilité et la reproductibilité des résultats (Dlouhy et Biemann, 2015). Avec les données utilisées pour l'illustration, le point de départ sélectionné est celui de la date de l'entrevue initiale à l'adolescence. La longueur des séquences est de 48 unités temporelles, ce qui correspond au nombre total de statuts mensuels couvrant la période de quatre ans entre l'entrevue à l'adolescence et celle réalisée au début de l'âge adulte.

Une fois la longueur établie, la prochaine étape consiste à élaborer une liste des valeurs qui composeront les éléments à l'intérieur des séquences, nommée *alphabet* en AS (Brzinsky-Fay et al., 2006). Il s'agit plus concrètement des différents statuts qui s'appliquent à chaque individu à chaque temps de mesure. En règle générale, plus l'*alphabet* comporte un grand nombre de statuts, plus l'interprétation des résultats risque d'être complexe une fois les analyses complétées. Il est alors recommandé d'éviter de dépasser dix – ou préférablement huit – valeurs, en priorisant les distinctions les plus importantes et en éliminant les valeurs rarement observées auprès des participants de l'échantillon (Biemann et Datta, 2014). Dans le cadre de cet article, l'*alphabet* comprend quatre valeurs qui représentent l'occupation principale en emploi ou en éducation au cours du mois : 1) ni au travail ni en éducation (N), 2) au travail (T), 3) en éducation secondaire (ES) et 4) en éducation postsecondaire (EP). Il faut donc indiquer au logiciel de traiter les variables visées comme des données d'une même séquence.

```
(df.alph <- seqstatl(df[, 1:48]))
df.seq <- seqdef(df[, 1:48], alphabet = df.alph,
                 labels = df.lab, states = df.shortlab,
                 xtstep = 4)
```

Enfin, un dernier aspect à considérer lors de la préparation des banques de données en AS est de s'assurer qu'elles soient organisées de sorte que les unités temporelles soient chronologiquement alignées. Le Tableau 1 illustre le modèle à suivre pour 12 mois consécutifs.

Tableau 1. Segments de séquences extraites de la banque de données pour les 12 premiers mois

Participants	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
ID162	ES	ES	ES	ES	ES	EP	EP	T	T	EP	EP	EP
ID163	ES	ES	ES	ES	ES	EP	EP	EP	EP	EP	EP	EP
ID164	ES	ES	ES	ES	ES	ES	ES	T	T	T	T	T
ID165	ES	ES	ES	ES	ES	T	T	T	T	T	T	T
ID166	ES	ES	ES	ES	ES	N	N	N	N	N	N	N

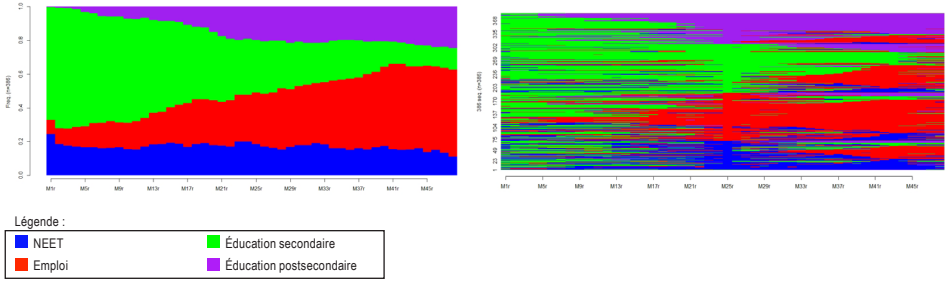
Note. M = Mois; N = Ni au travail et ni en éducation; T = Travail; ES = Éducation secondaire; EP = Éducation postsecondaire.

Deuxième étape de la stratégie analytique : comparaison et dissimilarité des séquences

Analyses préliminaires. Après avoir organisé adéquatement les données, il est suggéré d'examiner globalement la composition de l'échantillon à l'étude à l'aide d'analyses descriptives préliminaires (Brzinsky-Fay et al., 2006). Ces analyses permettent d'avoir un aperçu rapide de la nature des séquences retrouvées dans l'échantillon, avant de passer aux opérations plus complexes de comparaison et de regroupement. Pour faciliter l'interprétation, une couleur est attribuée à chaque statut. Un graphique (nommé chronogramme) est ensuite créé représentant les séquences individuelles, qui sont superposées en suivant l'ordre des observations dans la base de données (voir figure 1). L'axe des X indique le temps de mesure (de 1 à 48) et l'axe des Y, le numéro de l'observation dont la séquence est illustrée. Il est également possible de regrouper le nombre d'observations attribuées à chaque statut en fonction des mois, ce qui permet de générer un graphique plus aisément interprétable.

```
#Determiner les couleurs
cpal(df.seq)<- c("blue", "red", "green", "purple")
#Faire graphique de toutes les séquences de l'échantillon
seqIplot (df.seq, border = NA)
#par mois, le nombre de personnes dans chaque statut (plus lisible)
seqdplot(df.seq,border = NA)
```

Figure 1. Chronogrammes des séquences individuelles sans et avec regroupement par statut



Sans toutes les présenter par souci de parcimonie, la fonction seqiplot de la bibliothèque de code TraMineR sur R permet notamment d'examiner la fréquence des séquences les plus souvent retrouvées dans l'échantillon. Par exemple, avec les données utilisées dans le cadre de cet article, la séquence la plus fréquente (représentant 2.3 % des 386 participants de l'étude) est celle affichant le statut de l'éducation secondaire tout au long de la période examinée. Il est également possible d'obtenir des informations sur le nombre moyen d'unités temporelles occupées par statut. Par exemple, le statut de l'éducation secondaire présente une durée moyenne de 18 mois sur la période de quatre ans. Le statut occupé le moins souvent est celui de l'éducation postsecondaire, avec une durée moyenne de sept mois. Enfin, une autre mesure descriptive à explorer avec la bibliothèque de code TraMineR permet de tracer les taux de transition entre les statuts (voir Figure 1), indiquant le nombre moyen de transitions entre un statut en particulier (p. ex. : travail) à un autre (p. ex. : éducation secondaire). Par exemple, dans la matrice illustrée à la Figure 1, en s'attardant aux lignes horizontales, ion constate qu'en moyenne, les jeunes qui affichent le statut N (ni au travail ni en éducation) à un mois donné transitionnent dans une proportion de 87 % vers le même statut (N) le mois suivant, de 8 % vers le travail et de 4 % vers l'éducation secondaire.

```
#Moyenne de temps passé dans chacun des statuts (en nombre de mois)
seqmeant(df.seq)
#En graphique
seqmtplot(df.seq, ylim = c(0, 20))
#Taux de transition
round(seqtrate(df.seq), digits = 2)
#Séquence la plus commune de l'échantillon
seqtab(df.seq, idxs = 1:10)
#Donnees individuelles
mean(seqtransn(df.seq)) ##Nombre de transition
```

Figure 1. Matrice de transition obtenue avec les données utilisées dans le cadre de l'article.

	[-> N]	[-> T]	[-> ES]	[-> EP]
[N ->]	0.87	0.08	0.04	0.00
[T ->]	0.04	0.93	0.02	0.01
[ES ->]	0.02	0.03	0.93	0.02
[EP ->]	0.00	0.02	0.00	0.97

Note. N= Ni au travail ni en éducation, T = Travail, ES = Éducation secondaire; EP = Éducation postsecondaire. Les taux de transition se retrouvent sur les lignes horizontales du tableau.

Mesures de dissimilarité Une fois les caractéristiques générales des séquences mieux connues, la procédure de comparaison entre les séquences peut être entamée. En AS, ces comparaisons s'effectuent en mesurant la distance qui sépare chaque paire de séquences retrouvée dans l'échantillon, nommée mesure de dissimilarité. Les mesures de dissimilarité peuvent être dérivées à partir de différentes techniques, telles les méthodes d'appariement optimal (*optimal matching*), les distances de Levenshtein, la distance de khi-deux ou encore la méthode du nombre de sous-séquences communes (*Number of Matching Subsequences*) (Studer et Ritschard, 2016). La méthode la plus communément utilisée en sciences sociales est de loin celle de l'appariement optimal. Les prochains paragraphes se concentrent sur celle-ci (pour un aperçu complet des différentes techniques, voir Studer et Ritschard, 2016).

Expliquée simplement, la méthode d'appariement optimal produit les mesures de dissimilarité en calculant le nombre d'opérations nécessaires pour rendre deux séquences strictement identiques par l'entremise de l'algorithme d'optimisation de Needlman-Wuncsh (Abbott et Forrest, 1986; Dlouhy et Biemann, 2015). Un petit nombre d'opérations génère une petite mesure de dissimilarité, ce qui en retour indique un haut niveau de similarité entre deux séquences. Cette procédure est répétée systématiquement auprès de toutes les paires de séquences retrouvées dans l'échantillon, afin de générer ultimement une matrice de dissimilarité à partir de laquelle les typologies de séquences peuvent être dérivées.

Avec la méthode d'appariement optimal, deux types d'opérations élémentaires sont possibles : la substitution ou l'*insertion-deletion*, nommée « indel ». L'application de ces deux types d'opérations varie selon la définition des coûts. La substitution est l'opération qui permet de transformer directement un élément en un autre, sans modifier la longueur de la séquence. Il s'agit simplement de substituer un statut pour un autre, tel qu'illustré au tableau 2 qui reprend les six premiers mois des séquences ID164 et ID165 du Tableau 1. On peut y constater que le statut T (Travail) de l'ID164 au sixième mois (M6) a été à substitué par le statut Éducation Secondaire (ES), afin de rendre sa séquence équivalente à celle du ID165. La longueur des séquences n'a pas été modifiée.

Tableau 2. Opération de substitution

	M1	M2	M3	M4	M5	M6
ID164	ES	ES	ES	ES	ES	T
ID165	ES	ES	ES	ES	ES	ES
Présence de dissimilarités	≠	≠	≠	≠	≠	X
ID164 transformé en ID165	ES	ES	ES	ES	ES	∓ → ES
						Substitution

De son côté, l'opération « indel » consiste à insérer et à supprimer un élément dans la séquence. Cette opération peut altérer la longueur des séquences, mais permet toutefois de conserver l'ordre dans lequel les éléments se succèdent. Cette opération est illustrée au Tableau 3, en prenant cette fois-ci en exemple les six derniers mois des séquences ID162 et ID163. Les statuts de l'ID163 au huitième et neuvième mois (M8-M9) ont été supprimés (*deletion*), afin de rajouter (*insertion*) les deux statuts manquants à la fin de sa séquence pour la rendre équivalente à celle du ID164. La longueur de la séquence se retrouve alors changée.

Tableau 3. Opération d'indel

Sujets	**M6**	**M7**	**M8**	**M9**	**M10**	**M11**	**M12**		
ID162	EP	EP	T	T	EP	EP	EP		
ID163	EP	EP	EP	EP	EP	EP	EP		
Présence de dissimilarités	≠	≠	X	X	≠	≠	≠		
ID162 transformé en ID163	EP	EP	T	T	EP	EP	EP	EP	EP
			Del	Del				In	In

Note. Del = « deletion »; In= « insertion ».

Définition des coûts. Ces deux exemples mettent en évidence qu'en AS, plus d'une stratégie de remplacement peut être utilisée pour évaluer la dissimilarité de deux séquences (Aisenbrey et Fasang, 2010). Comment déterminer la meilleure stratégie à prioriser? Avec la méthode d'appariement optimal, l'algorithme de Needleman-Wunsh choisit systématiquement l'option qui requiert le moindre coût, c'est-à-dire le plus petit nombre d'opérations possible. La sélection des coûts est donc une étape cruciale, considérant que ceux-ci influenceront les procédures d'appariements et par conséquent, les mesures de dissimilarité (Lesnard, 2010; Ritschard et Studer, 2018).

De manière générale, il est conseillé de commencer cette procédure analytique en choisissant les coûts de substitutions (Lesnard, 2010). Il faut d'abord décider si ces opérations auront un coût fixe ou un coût variable. Un coût de substitution fixe signifie que toutes les transformations seront considérées

équivalentes. Par exemple, avec les données illustratives, si cette option est choisie, la substitution du statut « Éducation Secondaire » par le statut « Ni au travail ni en éducation » a le même coût qu'une substitution par le statut « Éducation Postsecondaire ». Cependant, en réalité, ces transitions ne sont souvent pas considérées équivalentes. En effet, passer de l'école secondaire à un statut d'inactivité (ni au travail ni en éducation) est associé à différents risques sur les plans social, financier et de la santé, alors que transiter vers l'éducation postsecondaire est vu comme un moyen de prévenir ces risques (Marshall et Symonds, 2021).

Pour ces raisons, les études en sciences sociales privilégient généralement d'attribuer des coûts variables aux opérations de substitution (Aisenbrey et Fasang, 2010; Bison, 2009; Studer, 2019). L'une des méthodes les plus couramment employées pour établir ces coûts est de se référer à la matrice de transition, illustrée à la Figure 1 (Studer et Ritschard, 2016). La matrice de transition retrace les probabilités qu'un statut en particulier soit suivi d'un autre, en se basant sur le nombre moyen de transitions observées au sein de l'échantillon pour toutes les combinaisons d'éléments possibles (pour connaitre plus de méthodes d'attribution des coûts, voir Gauthier et al., 2009). À partir de ces informations, des coûts différenciés sont obtenus pour chaque type de substitution. C'est cette méthode qui a été choisie dans le présent article.

Une fois les coûts de substitution choisis, la prochaine étape consiste à déterminer les coûts de l'opération d'indel, qui, traditionnellement, sont fixés au coût de substitutions le plus élevé, qui se situent généralement à 1 (Aisenbrey et Fasang, 2010; Bison, 2009; Studer, 2019). Des coûts d'indel avoisinant 1 optimisent la méthode d'appariement, en permettant à l'algorithme de prioriser les similitudes observées sur le plan de l'ordre lorsqu'elles sont dominantes, mais sans toutefois négliger les opérations de substitution (privilégiant le *timing* des transitions) lorsqu'elles s'avèrent plus optimales.

Enfin, il faut souligner que différentes configurations de coûts peuvent être testées pour examiner celles qui conviennent le mieux aux données. Il faut toutefois garder en tête que ces décisions auront une incidence majeure sur les mesures de dissimilarité (et conséquemment sur les typologies créées) et qu'à ce jour, aucune méthode de validation n'est disponible pour vérifier si la configuration de coût choisie est effectivement la plus adéquate en fonction des caractéristiques de l'échantillon et de la structure de données utilisées (Bison, 2009; Ritschard et Studer, 2018). À défaut d'une telle méthode, il est recommandé d'utiliser la procédure usuelle décrite ci-haut (c.-à.-d., taux de transition pour les coûts de substitution, 1 pour les coûts d'indel), et de la modifier au besoin[2].

[2] Pour une revue plus exhaustive des différentes techniques visant à définir les coûts en AS, les ouvrages de Studer et Ritschard (2016) et de Gauthier et al. (2009) peuvent être consultés

```
#Calculer des mesures de dissimilarité par appariement optimal (avec coûts de substitution correspondant aux taux de transition et ceux des indel établis à 1)
OMtrate <- seqdist(df.seq, method = "OM", indel = 1, sm = "TRATE")
hc.ward <- hclust(as.dist(OMtrate), method = "ward.D")
#Clustering hiérarchique avec test de modèles de 1 à 8 classes
df.clust <- as.clustrange(hc.ward, diss = OMtrate, ncluster = 8)
```

Troisième étape de la stratégie analytique : regroupement (*clustering*) des séquences

Analyses de regroupement. Après avoir déterminé les coûts des opérations et procédé aux analyses en calculant les mesures de dissimilarité entre les séquences, la dernière étape des AS peut être entamée. Elle consiste à regrouper les séquences similaires dans des catégories, à partir de la matrice de dissimilarité créée aux étapes précédentes. Tout comme pour les calculs des distances entre les séquences, différentes stratégies peuvent être déployées pour réaliser cet objectif, se basant tantôt sur les modèles théoriques, tantôt sur des méthodes empiriques. Sur le plan des méthodes empiriques, il est important de mentionner que la majorité de ces techniques proviennent de la famille des analyses algorithmiques de type regroupement/classification (*cluster*). Elles comprennent notamment la méthode de classification ascendante hiérarchique (*hierarchical clustering*), les algorithmes de k-médoïdes (*partitioning around medoids*; PAM) et les *single linkage clusterings*. Dans le cadre de cet article, la classification ascendante hiérarchique a été retenue et appliquée aux données, puisqu'il s'agit d'une méthode largement répandue et qui génère souvent les meilleurs indices d'ajustement aux tests de robustesse post-hoc (Dlouhy et Biemann, 2015; Han et al., 2017).

En bref, la classification ascendante hiérarchique amorce ses analyses en assignant chaque séquence dans une seule catégorie (une séquence = une catégorie), pour progressivement regrouper les séquences dans des catégories de plus en plus grandes (c.-à-d., regroupant de plus en plus de séquences). Autrement dit, l'algorithme procède au départ en regroupant les séquences démontrant de très faibles mesures de dissimilarité, résultant en un grand nombre de catégories comprenant un petit nombre de séquences (p. ex. : 50 sous-groupes regroupant en moyenne 6 individus). Progressivement, la condition exigeant de faibles mesures de dissimilarité pour regrouper des séquences s'assouplit, ce qui fait en sorte que de plus en plus de séquences peuvent être jumelées ensemble dans des catégories recoupant un plus grand nombre de séquences (p. ex. : 6 sous-groupes regroupant en moyenne 50 individus).

Cette procédure d'agglomération graduelle est reproduite jusqu'à ce que le critère de fin (c'est-à-dire le nombre de catégories déterminé par le chercheur) soit atteint. Typiquement, les chercheurs débutent en assignant un petit critère de fin aux analyses, pour progressivement l'augmenter, allant de 2 à 10 configurations par exemple (Navarro et al., 1997). La prochaine étape consiste à déterminer la configuration qui semble la plus adéquate, à partir de trois principales stratégies : celles des tests statistiques, de l'examen visuel et de la pertinence théorique.

Sélection du nombre de catégories. L'utilisation des indices d'ajustement permet de comparer les configurations obtenues à l'aide de critères d'information permettant de garantir la robustesse du modèle sélectionné (Studer, 2013, 2019). Parmi ces indicateurs, certains utilisent des mesures de corrélation pour examiner la capacité des configurations à reproduire la matrice de dissimilarité originale. Parmi ceux-ci, le point biserial de corrélation (PBC) utilise la corrélation de Pearson, alors que le HG (*Hubert's Gamma*, dans sa version originale) se fie plutôt à des corrélations non paramétriques basées notamment sur les distances moyennes des séquences à l'intérieur des regroupements et entre ceux-ci (Newell et al., 2013). Pour leur part, les critères de ASW et de ASWw (*Average Silhouette Width; Average Silhouette Width weighted*), quant à eux, examinent le degré d'homogénéité des catégories obtenues à l'intérieur des configurations et si elles parviennent à se distinguer significativement les unes des autres. Enfin, l'index de Hubert (*Hubert's C*; HC) indique l'écart entre la classification testée et la meilleure classification théoriquement possible si celle-ci était effectuée avec le même nombre de catégories et les mêmes propriétés de séquences. Pour tous les indicateurs mentionnés à l'exception du HC, les valeurs obtenues peuvent se situer entre -1 et 1, et plus elles sont élevées, plus la configuration est considérée robuste. Dans le cas du HC (qui peut varier entre 0 et 1), c'est l'inverse : plus la valeur obtenue est faible, meilleure la configuration est considérée.

Avec les données utilisées dans le cadre de cet article, des configurations de 3 à 8 catégories de parcours de transition de l'école au travail ont été testées. Le Tableau 4 affiche les indices d'ajustement correspondant à chacune de ces configurations. Selon ces résultats, la solution à cinq groupes semble la plus adéquate. En effet, elle présente les valeurs les plus élevées aux tests de PB, d'ASW et d'ASWw. Pour le HG, bien que la valeur de la configuration à cinq classes ne soit pas le plus élevé, l'augmentation est moins abrupte une fois la configuration à cinq groupes atteinte (de 0,75 à 0,77), ce qui signifie que les différences sont plutôt minimes avec les configurations suivantes. La même observation peut être faite avec l'indice HC, qui doit être le plus bas possible : une fois que la configuration à cinq groupes est atteinte, la diminution de l'indice se fait très progressivement (de 0,12 à 0,10), indiquant peu de différences entre la configuration à cinq groupes et celles à huit groupes.

```
#Indices d'ajustement
df.clust
#Visualisation des indices d'ajustement sous forme de graphique
plot(df.clust, stat = 'all', norm = 'zscore', lwd = 2)
```

Tableau 4. Indices d'ajustement des configurations testées avec les données

	PBC	HG	ASW	ASWw	HC
3 catégories	0,48	0,57	0,28	0,29	0,20
4 catégories	0,51	0,64	0,27	0,28	0,17
5 catégories	0,56	0,75	0,29	0,30	0,12
6 catégories	0,55	0,76	0,26	0,27	0,12
7 catégories	0,53	0,79	0,24	0,25	0,10
8 catégories	0,50	0,77	0,21	0,22	0,10

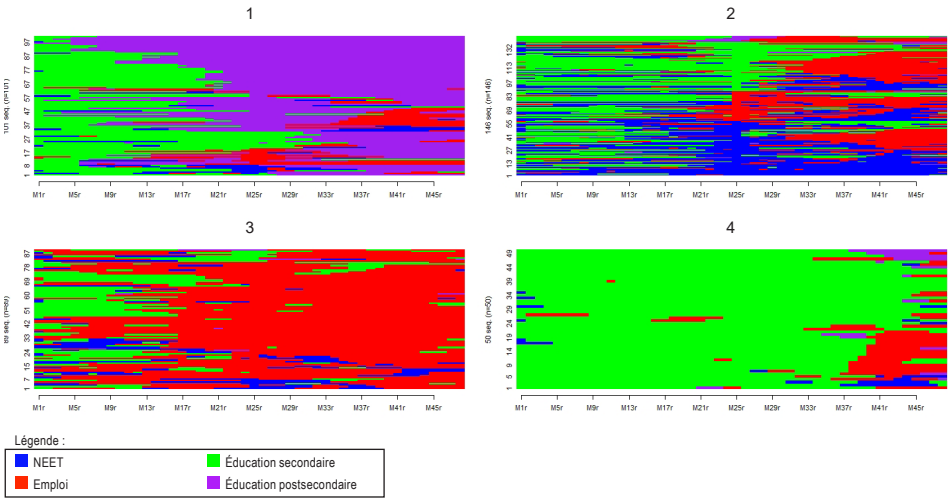
Note. PB = Point Biseral, HG= Hubert's Gamma, ASW = Average Silhouette Width, ASWw = Average Silhouette Width weighted, HC = Hubert's.

La deuxième stratégie utilisée en AS pour décider du nombre de catégories à retenir est celle de l'examen visuel. Elle consiste à apprécier visuellement les séquences types créées par la bibliothèque de code TraMineR, nommées chronogrammes (Gabadinho et al., 2011). Les chronogrammes (voir Figure 2 et 3) sont des graphiques où sont superposées chacune des séquences contenues dans une catégorie et où chaque statut est représenté par une couleur. L'axe horizontal représente le temps selon l'unité temporelle choisie (48 mois dans le cas du présent article) et l'axe vertical représente les identifiants auxquels appartiennent les séquences. Une bonne classification des séquences devrait générer des chronogrammes qui s'interprètent facilement, c'est-à-dire qui permettent en un coup d'œil d'identifier les types de parcours suivis par les individus regroupés dans les catégories (Brzinsky-Fay et al., 2006). À l'inverse, lorsque les chronogrammes sont diffus et qu'aucune tendance n'est perceptible, la configuration n'est pas optimale.

À titre d'exemple, deux différentes configurations obtenues à partir des données utilisées dans le cadre de cet article sont présentées aux Figures 2 et 3. D'abord, la configuration illustrée à la Figure 2 contient quatre groupes. Le premier groupe de parcours (#1; 26 % de l'échantillon) s'interprète facilement : il est constitué de jeunes ayant transité de l'éducation secondaire à l'éducation postsecondaire, en majorité vers la moitié de la période examinée. Les groupes #3 (23 %) et #4 (13 %) sont aussi faciles à interpréter ou à identifier et paraissent assez homogènes. Le groupe #3 représente un parcours de jeunes s'insérant rapidement sur le marché du travail et le groupe #4 illustre un parcours de jeunes demeurant principalement en éducation secondaire tout au long de la période, avec quelques jeunes passant à l'emploi vers la fin de la période. Le deuxième groupe (#2; 38 %) est toutefois plus problématique. Il contient différents sous-groupes de parcours, avec des jeunes ni au travail ni en éducation tout au long de la période, jumelés à d'autres passant de l'école secondaire au travail à différents moments de la période. Il est donc potentiellement trop hétérogène pour représenter un réel sous-groupe de jeunes dans leur transition de l'école au travail.

```
#Enregistrer la solution à 4 clusters
clusterH4 <- df.clust$clustering$cluster4
#graphique des séquences individuelles par classes
seqdplot(df.seq, group = df.clust$clustering$cluster4, border = NA)
#graphique de la sequence Moyenne/typique de chaque classe
icenter <- disscenter(OMtrate, factor(clusterH4), medoids.index="first")
seqiplot (df.seq[icenter,])
```

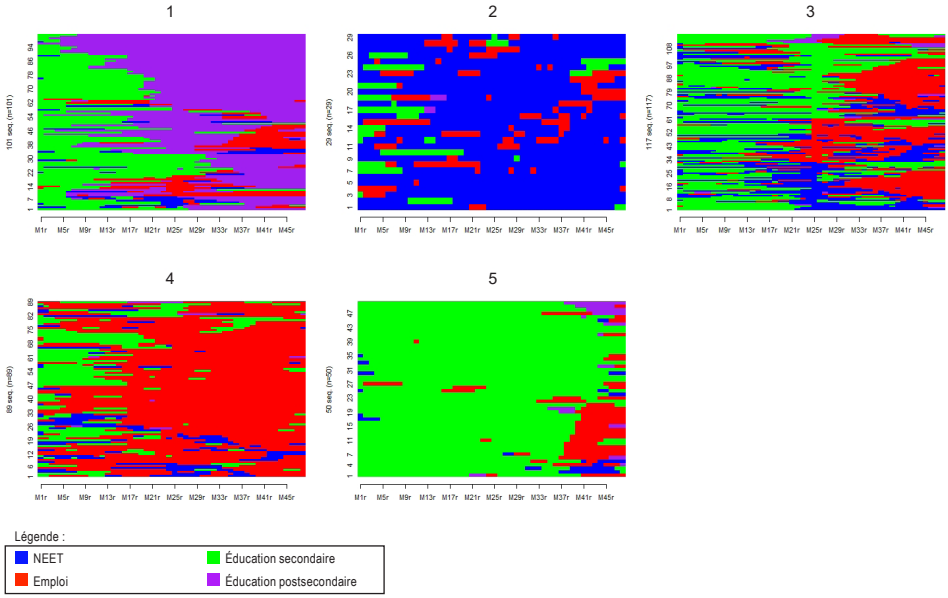
Figure 2. Chronogrammes avec une classification à quatre groupes de transition de l'école au travail



Concernant la configuration à cinq groupes qui est illustrée à la Figure 3, on constate que les groupes de parcours les plus homogènes de la configuration précédente sont reproduits. Il est question ici du parcours de transition vers l'école secondaire (#1 → #1), du parcours vers le travail précoce (#3 → #4) et du parcours d'études secondaires prolongées (#3 → #4). Le groupe #2 s'est quant à lui divisé en deux différents parcours, soit le #2 et le #3. Le #2 (8 %) représente un parcours de jeunes principalement ni au travail ni en éducation tout au long de la période. Le #3 (30 %) demeure hétérogène, mais il est désormais plus facile de dégager un parcours général du graphique. En effet, les jeunes suivant ce parcours semblent commencer la période en éducation secondaire, pour ultimement s'orienter vers le travail. Le moment précis (timing) de la transition varie d'un individu à l'autre, mais survient généralement après la moitié de la période examinée.

```
#Enregistrer la solution à 5 clusters
clusterH5 <- df.clust$clustering$cluster5
#graphique des séquences individuelles par classes
seqdplot(df.seq, group = df.clust$clustering$cluster5, border=NA, cols=3)
#graphique de la sequence Moyenne/typique de chaque classe
icenter <- disscenter(OMtrate, factor(clusterH5), medoids.index="first")
seqiplot (df.seq[icenter,])
```

Figure 3. Chronogrammes avec une classification à cinq groupes de transition de l'école au travail



Ainsi, selon la stratégie de l'examen visuel des deux configurations, celle comprenant cinq groupes paraît plus optimale que celle à quatre groupes. D'une part, parce qu'elle sépare un groupe diffus et hétérogène en deux groupes plus distincts et d'autre part, car elle parvient à cibler un sous-groupe hautement à risque : les jeunes ni au travail ni en éducation sur une période prolongée. Ce parcours est d'un grand intérêt pour la recherche et l'intervention, puisque l'on souhaite réduire son occurrence et diminuer les conséquences qui y sont associées. Il est donc préférable de préserver ce groupe et d'approfondir les analyses pour explorer des facteurs collatéraux qui y sont liés (p. ex. : données sociodémographiques, de la santé et scolaires).

Enfin, avant de décider du nombre optimal de groupes à retenir, il est également important de s'assurer que la configuration choisie ait un sens sur les plans théorique et pratique (Brzinsky-Fay et al., 2006), ce qui constitue la troisième et dernière stratégie de sélection. En effet, même si les AS suivent une démarche

exploratoire plutôt que confirmatoire, il est attendu que la configuration obtenue reflète des caractéristiques déjà examinées antérieurement dans le cadre d'études empiriques, ou fasse partie des postulats de modèles théoriques notoires (Ritschard et Studer, 2018). Dans le cas de la configuration à cinq groupes du présent article, ces critères sont majoritairement respectés. Les études antérieures sur la transition de l'école au travail ont déjà documenté des parcours similaires (p. ex. : Lorentzen et coll., 2019; Ng-Knight, et Schoon, 2017) tout comme des modèles théoriques (p. ex. : Modèle des Parcours de vie, voir Elder et al., 2015). On pourrait toutefois questionner la pertinence d'avoir deux groupes de parcours comprenant une transition vers le travail : une ayant lieu de manière précoce (catégorie #4) et l'autre se déroulant vers la moitié de la période (catégorie #3). Or, comme la création de la catégorie de la transition plus tardive vers le travail ne survient qu'avec celle de la catégorie des jeunes ni au travail ni en éducation, il est préférable de la conserver, De surcroît, le moment précis (*timing*) de la transition vers le travail pourrait aussi se révéler un facteur important associé à des conséquences distinctes sur le développement.

En somme, il est fréquent qu'une configuration ne corresponde pas parfaitement aux résultats empiriques des travaux antérieurs. Il en est de même pour les indices d'ajustement et l'examen visuel des chronogrammes. Le choix doit donc se faire en priorisant la configuration satisfaisant le plus grand nombre de critères possible à partir des trois stratégies de sélection décrites et qui s'arrime adéquatement avec les questions et le domaine de recherche. La configuration à sélectionner avec les données utilisées dans le cadre de cet article serait donc fort probablement celle à cinq catégories de parcours, puisqu'elle répond à la fois aux conditions statistiques, visuelles et théoriques.

Une fois la configuration choisie, le terme « typologie » peut maintenant lui être attribuée. De manière générale, les chercheurs soumettent ensuite la typologie à des analyses de régression (p. ex. : logistique, linéaire) visant à examiner si des antécédents sont associés à l'appartenance aux parcours et si les parcours sont associés à des conséquences ultérieures dans le développement. Ces analyses agissent également à titre de validation supplémentaire de la typologie sélectionnée (Ritschard et Studer, 2018). En effet, une bonne typologie devrait permettre de prédire des phénomènes psychosociaux pertinents au sujet à l'étude, montrant qu'elle revêt une valeur heuristique. Par exemple, dans la typologie de parcours de transition de l'école au travail créée dans le cadre de cet article, il serait attendu que les jeunes du parcours ni au travail ni à l'emploi (#2) proviennent de familles moins socioéconomiquement avantagées que ceux du sous-groupe poursuivant des études postsecondaires (#1), tel que révélé par des études antérieures (p. ex. : Lorentzen et al., 2019; Ranasinghe et al., 2019; Schoon et Lyons-Amos, 2017). Comme cet article se concentre sur les AS, les analyses de régressions qui pourraient être faites ne sont pas présentées dans le texte (pour un exemple, voir : Thouin et al., en révision) .

Discussion

Cet article avait comme objectif d'introduire la méthode des AS et leur application pour identifier des parcours de vie en sciences sociales. Il s'agit d'une approche algorithmique utile pour étudier des comportements, des affects ou des cognitions qui se transforment de façon qualitative au fil du temps, c'est-à-dire qui ne se développent pas de façon systématiquement croissante ou décroissante. Initialement surtout utilisée en Europe, la popularité des AS augmente en Amérique du Nord, grâce en partie à des articles récemment publiés dans des revues en sciences développementales introduisant cette méthode (Canales Sánchez et al., 2020; Johnston et al., 2020; Lu et al., 2021). En plus de permettre d'examiner la transition de l'école au travail comme démontré dans cet article, les AS sont utiles pour examiner une multitude d'autres phénomènes caractérisés par une évolution catégorielle entre différents statuts, par exemple l'évolution des comportements délinquants à l'adolescence. Les AS mettent en lumière des changements entre des statuts distincts et complexes à opérationnaliser en tenant compte de la durée, du moment d'apparition (*timing*) et de l'ordre dans lequel ceux-ci s'enchainent. Il s'agit de notions considérées conceptuellement importantes en sciences du développement, notamment car elles peuvent avoir une incidence sur l'adaptation, la santé et le bien-être subséquents (Blossfeld, 2009; Elder, 1998; Elder et Giele, 2009; Elder Jr et al., 2015). Par exemple, dans la transition de l'école au travail lors du passage à l'âge adulte, des études ont démontré que les jeunes qui demeuraient longtemps déconnectés du marché de l'emploi et de l'éducation éprouvaient davantage de difficultés financières et sociales comparativement à ceux y restant sur une courte *durée* (Mascherini et Ledermaier, 2016; Van de Velde, 2016). D'autres exemples peuvent également être trouvés dans le domaine de la toxicomanie, avec des études démontrant que l'âge (*timing*) auquel les individus consomment pour une première fois une substance psychoactive aura une influence sur leur parcours de consommation future, avec une plus grande probabilité de consommer différents types de substances, et de développer des problématiques d'abus et de dépendance (Collins et al., 1997; Richmond-Rakerd et al., 2017).

Un autre avantage de l'utilisation des AS est la qualité des représentations graphiques produites (Ritschard et Studer, 2018). Ces dernières permettent une interprétation facile et rapide des résultats contribuant ainsi à mieux vulgariser les typologies obtenues auprès des publics moins familiers avec ces approches typologiques, comme les décideurs politiques et les praticiens (voir Figure 1 et 2). Également, les AS ont peu de contraintes concernant les tailles d'échantillon, et peuvent être techniquement réalisées à partir d'un petit nombre de sujets sans que les résultats en soit gravement affectés (Dlouhy et Biemann, 2015). Enfin, un dernier avantage : les AS sont grandement flexibles (Han et al., 2017). Les chercheurs peuvent effectivement moduler la démarche analytique en fonction de leurs questions de recherche, que ce soit à l'étape de la sélection du type de mesure de dissimilarité, à celle de la définition des coûts ou encore au moment de choisir la stratégie d'agglomération des séquences en sous-groupes (Aisenbrey et Fasang, 2010).

Toutefois, cette flexibilité amène aussi son lot de limites, surtout en ce qui concerne l'étape de la définition des coûts. En effet, des auteurs soulignent que

la grande latitude laissée aux chercheurs pour établir les coûts des opérations peut devenir problématique, puisque, comme mentionné plus tôt, les coûts ont un rôle crucial dans l'élaboration des typologies (Bison, 2009; Wu, 2000). Ainsi, un chercheur pourrait tester arbitrairement différentes configurations de coûts jusqu'à l'obtention de la typologie souhaitée, allant à l'encontre d'une démarche scientifique rigoureuse et objective (Chalmers, 1977). Bien que certains outils de validation empirique des coûts aient été créées au courant des dernières années, aucune d'entre elles ne fait consensus actuellement au sein de la communauté des chercheurs utilisant les AS (Ritschard et Studer, 2018). Il est donc conseillé d'avoir recours à plusieurs méthodes de vérification des coûts pour éviter les biais, mais il demeure encore difficile aujourd'hui de déterminer ce qui se rapproche le plus du « vrai » coût des opérations (Gauthier et al., 2009; Ritschard et Studer, 2018).

De plus, le nombre d'unités temporelles retraçant les statuts, les états ou les comportements examinés doit être relativement élevé en AS. Comme évoqué précédemment, il est recommandé d'avoir au minimum 25 unités pour optimiser la démarche, ce qui requiert d'avoir recueilli des données sur des périodes couvrant 25 mois, 25 trimestres, 25 années ou autres (Dlouhy et Biemann, 2015). Obtenir autant d'informations peut être ardu pour les chercheurs, surtout considérant que l'attrition est un problème souvent inévitable en sciences sociales, y compris dans les études qui se penchent sur la transition à l'âge adulte (Gould et al., 1990; Nicholson et al., 2017). Sur ce plan, les AS sont aussi limitées puisqu'elles offrent peu d'options pour pallier les problèmes de données manquantes (Courgeau, 2018). Les deux principales stratégies consistent à 1) supprimer les segments des séquences avec des valeurs manquantes ou 2) attribuer une valeur désignée aux valeurs manquantes (p. ex. : T = travail, E = éducation, M = manquant), soit deux méthodes qui sont démontrées moins efficaces que d'autres alternatives (imputations multiples, estimations par maximum de vraisemblance) (Little, 2013). Dans les deux cas, différents problèmes peuvent survenir au moment de calculer les mesures de dissimilarité. Par exemple, avec la première stratégie, des comparaisons seront faites entre des séquences n'ayant pas la même longueur, ce qui peut causer des biais considérant que les périodes couvertes ne sont pas équivalentes et peuvent cacher des informations importantes. Avec la deuxième stratégie, des séquences discordantes, mais affichant plusieurs valeurs manquantes, pourraient être considérées semblables seulement parce que les deux possèdent des valeurs manquantes. Pour ces raisons, il est conseillé de supprimer les séquences incomplètes, ce qui peut résulter en une diminution significative du nombre de participants à analyser dans un échantillon et, dans certains cas, altérer la signification substantielle des trajectoires (p. ex. : données manquantes pour des individus ayant commis un suicide dans des études de trajectoires sur les états dépressifs) (Studer et al., 2018).

Tout comme la sélection des coûts des opérations, le choix du nombre de sous-groupes de séquences au cœur des typologies fait aussi l'objet de certaines critiques (Courgeau, 2018; Wu, 2000). Plus précisément, la robustesse dans la découverte d'une classification (cluster) optimale en AS présente aussi certaines limites, notamment parce qu'elle se base sur un seul paramètre (mesure de dissimilarité) pour établir les regroupements et que les tests statistiques visant à évaluer la qualité de la configuration élaborée sont considérés insuffisants et limités (Kline, 2016)

Conclusion

Dans cet article, nous nous sommes concentrés sur les AS dans leur forme la plus simplifiée. Or, ce champ analytique est en constante évolution et de nouvelles spécificités et variations des AS sont régulièrement publiées. Par exemple, une nouvelle version des AS (voir Rossignon et al., 2018), nommée analyses historiques des séquences (*sequence history analysis*), permet d'être jumelée à des analyses historiques des événements (*event history analysis*), alors qu'une autre version (voir Helske et al., 2018) combine l'algorithme des AS à celui du modèle de Markov caché (*hidden Markov model*). Les lecteurs intéressés à ces nouveaux développements peuvent consulter le site internet du regroupement international de la *Sequence Analysis Association* (https://sequenceanalysis.org/), qui offre des webinaires et des articles de blog présentant les plus récentes avancées et de nouvelles fonctionnalités des AS.

Références

Abbott, A. (1995). Sequence analysis: new methods for old ideas. *Annual Review of Sociology, 21*, 93-113.

Abbott, A. et Forrest, J. (1986). Optimal matching methods for historical sequences. *The Journal of Interdisciplinary History, 16*(3), 471-494. https://doi.org/10.2307/204500

Abbott, A. et Hrycak, A. (1990). Measuring resemblance in sequence data: An optimal matching analysis of musicians' careers. *American Journal of Sociology, 96*(1), 144-185. https://doi.org/10.1086/229495

Aisenbrey, S. et Fasang, A. E. (2010). New life for old ideas: The "second wave" of sequence analysis bringing the "course" back into the life course. *Sociological Methods & Research, 38*(3), 420-462. https://doi.org/10.1177/0049124109357532

Biemann, T. et Datta, D. K. (2014). Analyzing sequence data:Optimal matching in management research. *Organizational Research Methods, 17*(1), 51-76. https://doi.org/10.1177/1094428113499408

Bison, I. (2009). OM natters: The interaction effects between indel and substitution costs. *4*(2), 53-67. https://doi.org/10.1177/205979910900400205

Blossfeld, H.-P. (2009). Comparative Life Course Research. A cross-national and longitudinal perspective. Dans G. H. Elder et J. Z. Giele (dir.), *The Craft of Life Course Research* (p. 280-306). Guilford Press.

Brzinsky-Fay, C., Kohler, U. et Luniak, M. (2006). Sequence analysis with stata. *The Stata Journal 6*(4), 435-460. https://doi.org/10.1177/1536867x0600600401

Canales Sánchez, D., Bautista Godínez, T., Moreno Salinas, G., García-Minjares, M. et Sánchez-Mendiola, M. (2020). Curricular change in a medical school: a new method for analysis of students academic pathways. *medRxiv* https://doi.org/10.1101/2020.04.25.20079715

Chalmers, A. (2013). *What is this thing called science?* (4e ed.) Hackett Publishing.

Collins, L. M., Graham, J. W., Rousculp, S. S. et Hansen, W. B. (1997). Heavy caffeine use and the beginning of the substance use onset process: An illustration of latent transition analysis. Dans Bryant, K.J, Windle, M. et West, S.G. (dir.) *The science of prevention: Methodological advances from alcohol and substance abuse research*. (p. 79-99). American Psychological Association. https://doi.org/10.1037/10222-003

Courgeau, D. (2018). Do different approaches in population science lead to divergent or convergent models? Dans G. Ritschard et M. Studer (dir.), *Sequence Analysis and Related Approaches: Innovative Methods and Applications* (p. 15-33). Springer International Publishing. https://doi.org/10.1007/978-3-319-95420-2_2

Dlouhy, K. et Biemann, T. (2015). Optimal matching analysis in career research: A review and some best-practice recommendations. *Journal of Vocational Behavior, 90*, 163-173. https://doi.org/10.1016/j.jvb.2015.04.005

Dupéré, V., Lacourse, É., Vitaro, F. et Tremblay, R. E. (2007). Méthodes d'analyse du changement fondées sur les trajectoires de développement individuelle : Modèles de régression mixtes paramétriques et non paramétriques/Longitudinal methods based on Individual development trajectories – Parametric and non parametric mixed models. *Bulletin de méthodologie sociologique, 95*(1), 26-57.

Dupéré, V., Dion, E., Leventhal, T., Archambault, I., Crosnoe, R. et Janosz, M. (2018). High school dropout in proximal context: The triggering role of stressful life events. *Child Development, 89*(2), e107-e122. https://doi.org/10.1111/cdev.12792

Elder, G. H., Jr. (1998). The life course as developmental theory. *Child Development, 69*(1), 1-12. http://www.jstor.org/stable/1132065

Elder, G. H., Jr. et Giele, J. (2009). Life course studies: An evolving field. Dans Elder, G. H., Jr. et Giele, J. (dir.)*The craft of life course research* (p. 1-24). Guilford Press.

Elder, G. H., Jr. et Shanahan, M. J. (2006). The life course and human development. Dans W. Damon et R. Lerner (dir.), *Handbook of Child Psychology* (6the éd., vol. 1: Theoretical Models of Human Development, p. 665-715). Wiley. https://doi.org/10.1002/9780470147658.chpsy0112

Elder Jr, G. H., Shanahan, M. J. et Jennings, J. A. (2015). Human development in time and place. Dans R. Lerner, M. H. Bornstein et T. Leventhal (dir.), *Handbook of Child Psychology and Developmental Science* (p. 6-54). John Wiley & Sons. https://doi.org/10.1002/9781118963418.childpsy402

Gabadinho, A., Ritschard, G., Müller, N. S. et Studer, M. (2011, 2011-04-07). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software, 40*(4). https://doi.org/10.18637/jss.v040.i04

Gauthier, J.-A., Widmer, E. D., Bucher, P. et Notredame, C. (2009). How much does it dost?: Optimization of costs in sequence analysis of social science data. *Sociological Methods & Research, 38*(1), 197-231. https://doi.org/10.1177/0049124109342065

Gould, M. S., Shaffer, D. et Davies, M. (1990). Truncated pathways from childhood to adulthood: Attrition in follow-up studies due to death. Dans Robins, L. et Rutter, M. (dir.) *Straight and devious pathways from childhood to adulthood.* (p. 3-9). Cambridge University Press.

Han, Y., Liefbroer, A. C. et Elzinga, C. H. (2017). Comparing methods of classifying life courses: sequence analysis and latent class analysis. Longitudinal and Life Course Studies; Vol 8, No 4 (2017): *Longitudinal and Life Course Studies*. https://doi.org/10.14301/llcs.v8i4.409

Helske, S., Helske, J. et Eerola, M. (2018). Combining Sequence Analysis and Hidden Markov Models in the Analysis of Complex Life Sequence Data. Dans G. Ritschard et M. Studer (dir.), *Sequence Analysis and Related Approaches: Innovative Methods and Applications* (p. 185-200). Springer International Publishing. https://doi.org/10.1007/978-3-319-95420-2_11

Johnston, C. A., Crosnoe, R., Mernitz, S. E. et Pollitt, A. M. (2020). Two methods for studying the developmental significance of family structure trajectories. *Journal of Marriage and Family 82*(3), 1110-1123. https://doi.org/10.1111/jomf.12639

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (Fourth edition.e éd.). The Guilford Press.

Kohler, U., Luniak, M. et Brzinsky-Fay, C. (2020). *SQ: Stata module for sequence analysis*. Boston College Department of Economics.

Kruskal, J. B. (1983). An overview of sequence comparison: Time warps, string edits, and macromolecules. *25*(2), 201-237. https://doi.org/10.1137/1025045

Laditka, J. N. et Laditka, S. B. (2015, 2016/12/01). Associations of educational attainment with disability and life expectancy by race and gender in the United States. *Journal of Aging and Health, 28*(8), 1403-1425. https://doi.org/10.1177/0898264315620590

Lesnard, L. (2010). Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns. *Sociological methods & research 38*(3), 389-419. https://doi.org/10.1177/0049124110362526

Little, T. D. (2013). Longitudinal structural equation modeling. Guilford press.

Lorentzen, T., Bäckman, O., Ilmakunnas, I. et Kauppinen, T. (2019). Pathways to adulthood: Sequences in the school-to-work transition in Finland, Norway and Sweden. *Social Indicators Research, 141*(3), 1285-1305. https://doi.org/10.1007/s11205-018-1877-4

Lu, Y., Zhang, R. et Du, H. (2021). Family structure, family instability, and child psychological well-being in the context of migration: Evidence from sequence analysis in China. *Child Development*. https://doi.org/https://doi.org/10.1111/cdev.13496

Maas, M. K., Bray, B. C. et Noll, J. G. (2019). Online sexual experiences predict subsequent sexual health and victimization outcomes among female adolescents: A latent class analysis. *Journal of Youth and Adolescence, 48*(5), 837-849. https://doi.org/10.1007/s10964-019-00995-3

Marshall, E. A. et Symonds, J. E. (2021). *Young adult development at the school-to-work transition: International pathways and processes*. Oxford University Press.

Mascherini, M. et Ledermaier, S. (2016). *Exploring the diversity of NEETs*. Publications Office of the European Union Luxembourg.

Morin, A. J. S. et Litalien, D. (2019). *Mixture modeling for lifespan developmental research*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190236557.013.364

Navarro, J. F., Frenk, C. S. et White, S. D. M. (1997, 1997/12). A Universal density profile from hierarchical clustering. *The Astrophysical Journal, 490*(2), 493-508. https://doi.org/10.1086/304888

Newell, M. A., Cook, D., Hofmann, H. et Jannink, J.-L. (2013). An algorithm for deciding the number of clusters and validation using simulated data with application to exploring crop population structure *The Annals of Applied Statistics, 7*(4), 1898-1916. http://www.jstor.org/stable/23566447

Ng-Knight, T. et Schoon, I. (2017). Can cocus of control compensate for socioeconomic adversity in the transition from school to work? *A Multidisciplinary Research Publication, 46*(10), 2114-2128. https://doi.org/10.1007/s10964-017-0720-6

Nicholson, J. S., Deboeck, P. R. et Howard, W. (2017). Attrition in developmental psychology: A review of modern missing data reporting and practices. International *Journal of Behavioral Development 41*(1), 143-153. https://doi.org/10.1177/0165025415618275

Pollock, G. (2007). Holistic trajectories: a study of combined employment, housing and family careers by using multiple-sequence analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 170*(1), 167-183. https://doi.org/10.1111/j.1467-985X.2006.00450.x

Ranasinghe, R., Chew, E., Knight, G. et Siekmann, G. (2019). School-to-Work Pathways. *National Centre for Vocational Education Research*, Adelaide, SA.

Richmond-Rakerd, L. S., Slutske, W. S. et Wood, P. K. (2017). Age of initiation and substance use progression: A multivariate latent growth analysis. *Psychology of Addictive Behaviors, 31*(6), 664-675. https://doi.org/10.1037/adb0000304

Ritschard, G. et Studer, M. (2018). *Sequence Analysis and Related Approaches: Innovative Methods and Applications*. Springer. https://doi.org/10.1007/978-3-319-95420-2

Rossignon, F., Studer, M., Gauthier, J.-A. et Goff, J.-M. L. (2018). Sequence history analysis (SHA): Estimating the effect of past trajectories on an upcoming event. Dans G. Ritschard et M. Studer (dir.), *Sequence Analysis and Related Approaches: Innovative Methods and Applications* (p. 83-100). Springer International Publishing. https://doi.org/10.1007/978-3-319-95420-2_6

Schoon, I. (2015). Diverse pathways: Rethinking the transition to adulthood. Dans Amato, P.R, Booth, A., McHale, S. et Van Hook, J. *Families in an Era of Increasing Inequality* (p. 115-136). Springer.

Schoon, I. et Lyons-Amos, M. (2017). A socio-ecological model of agency: The role of structure and agency in shaping education and employment transitions in England. *Journal of Longitudinal and Lifecourse Studies, 8*(1), 35-56.

Struffolino, E., Studer, M. et Fasang, A. E. (2016). Gender, education, and family life courses in East and West Germany: Insights from new sequence analysis techniques. *Advances in Life Course Research, 29*, 66-79. https://doi.org/10.1016/j.alcr.2015.12.001

Studer, M. (2013). WeightedCluster Library Manual: A practical guide to creating typologies of trajectories in the social sciences with R. *LIVES Working Papers, 2013*(24). https://doi.org/10.12682/lives.2296-1658.2013.24

Studer, M. (2019). *A general introduction to sequence analysis and its use in the social sciences [présentation d'un conférencier invitié].* Institut de démographie et socioéconomie, Université de Genève.

Studer, M. et Ritschard, G. (2016). What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society: Series A* (Statistics in Society) 179(2), 481-511. https://doi.org/10.1111/rssa.12125

Studer, M., Struffolino, E. et Fasang, A. E. (2018). Estimating the relationship between time-varying covariates and trajectories: The sequence analysis multistate model procedure. (Revue) *48*(1), 103-135. https://doi.org/10.1177/0081175017747122

Thouin, É. (2022). *La transition de l'école au travail chez les jeunes en situation de vulnérabilité scolaire ou sociale : examen des déterminants, des conséquences et des processus explicatifs* [thèse de doctorat, Université de Montréal]. Papyrus

Thouin, É., Dupéré, V., Denault, A. S. et Schoon, I. (en révision). Beyond college for all: Portrait of rapid and successful school-to-work Transitions among vulnerable youth.

Van de Velde, C. (2016). Les "NEETs": une déconstruction sociologique. *Bulletin d'information Observatoire jeunes et société,* (2), 18-19.

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. https://doi.org/10.1177/0049124100029001003

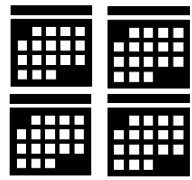
Annexe : résumé imagé de l'article

Collecte de données

Mesure du statut d'activité

Ni en emploi ni aux études (NEET)

Collecte mensuelle sur 48 mois



Statut vulnérable : ni en emploi, ni aux études (NEET)

Analyse de séquence

Étape 1 : Organisation des données en séquence

Temps : 1 2 3 4 5 6 7 8 ...
Statut : ▲ ▲ ...

Étape 2 : Comparaison des séquences

- Mesures de dissimilarité
  - Appariement optimal
  - Estimer les coûts de substitution et *indel*

Étape 3 : Regroupement des séquences

Classification ascendante hiérarchique

Classification en 5 sous-groupes sélectionnée

Validation du modèle sélectionné par l'analyse de séquence

Modèle favorisé : classification en 5 sous-groupes

- Groupe d'intérêt (NEET) bien distingué
- Corrélation des groupes d'appartenance avec les prédicteurs et les résultats (*outcomes*) d'intérêt
- Cohérence entre le modèle et la théorie

Applications possibles

Utiliser des prédicteurs (notes, scolarité parentale, revenu familial…) pour tenter de déterminer ce qui influence l'appartenance aux différents sous-groupes

Utiliser les sous-groupes pour tenter de prédire différents évènements (décrochage, évènement de santé mentale…)

Création d'interventions pour favoriser le maintien en emploi ou aux études