sciendo

Swiss Journal of Sociology, 49 (2), 2023, 417–445

Back to the Features. Investigating the Relationship Between Educational Pathways and Income Using Sequence Analysis and Feature Extraction and Selection Approach

Leonhard Unterlerchner*, Matthias Studer*, and Andrés Gomensoro**

Abstract: This article compares two methods to study the link between educational pathways and income. Sequence analysis provides a holistic view but might fail to identify key trajectory characteristics. A new validation method overcoming this limit is proposed. Feature extraction and selection can directly identify these key characteristics. The conclusion summarizes the strengths and weaknesses of each method and provides guidelines on how to choose a method to study the relationship between a previous trajectory and a later-life outcome.

Keywords: Sequence analysis, cluster validation, feature extraction and selection, income, educational pathways

Back to the Features. Untersuchung des Zusammenhangs zwischen Bildungswegen und Einkommen mit Hilfe von Sequenzanalyse und Verfahren zur Auswahl und Extraktion von Merkmalen

Zusammenfassung: Der Artikel vergleicht zwei Methoden zur Untersuchung der Zusammenhänge zwischen Bildungswegen und Einkommen. Die Sequenzanalyse bietet eine ganzheitliche Sicht auf die Verläufe, aber sie kann die wichtigsten Verlaufsmerkmale, die das Einkommen erklären, ausser Acht lassen. Wir entwickeln eine neue Methode, die diese Einschränkung überwindet. Die Methode der Merkmalsextraktion und -auswahl identifiziert diese Schlüsselmerkmale direkt. Abschliessend werden die Vor- und Nachteile zusammengefasst und Leitlinien für die Methodenauswahl gegeben.

Schlüsselwörter: Sequenzanalyse, Merkmalsextraktion und -auswahl, Einkommen, Bildungswege

Back to the features. Étude du lien entre revenu et parcours de formation à l'aide de l'analyse de séquence et de techniques d'extraction et de sélection de caractéristiques

Résumé : L'article compare deux méthodes pour étudier les liens entre parcours de formation et revenu. L'analyse de séquences offre une vue globale, mais n'identifie pas de caractéristiques clefs. Une nouvelle méthode de validation dépassant cette limite est proposée. L'extraction et la sélection de caractéristiques permettent d'identifier directement les caractéristiques clefs. La conclusion résume les forces et faiblesses de chaque méthode et propose des recommandations pour l'étude des liens entre trajectoires et résultat postérieur.

Mots-clés : Analyse de séquences, validation de typologie, extraction et sélection de caractéristiques, revenu, parcours de formation

* Institute for Demography and Socioeconomics (IDESO), University of Geneva, CH-1211 Geneva 4, Center LIVES University of Geneva, CH-1211 Geneva 4, leonhard.unterlerchner@unige.ch, matthias.studer@unige.ch.

** University of Bern, CH-3012 Bern, andres.gomensoro@unibe.ch.

Seismo

DOI 10.2478/sjs-2023-0021

1 Introduction[1]

Several studies have been conducted on the occupational integration of the younger generation – a major political concern in Western societies (Arum and Shavit 1995; Brzinsky-Fay 2007; Gauthier and Gianettoni 2013). In this process, education is often considered a key resource for successful integration into the labor market (Bills 2003; Kramarz and Viarengo 2015). Studies have highlighted the role of education to explain status attainment and social reproduction (Meyer 2009), unemployment risk (Benda et al. 2019), job quality (Geier et al. 2013), employment prospects, and income (Korber and Oesch 2019).

The relationship between education and indicators of occupational integration, such as income, is often investigated by considering educational attainment (Gomensoro et al. 2017; Korber and Oesch 2019). Educational attainment, as an outcome of educational pathways, has been consistently shown to be positively associated with income (Falcon 2016; Gomensoro et al. 2017).

However, life course literature emphasizes the need to situate any outcome, such as education or income, within the constantly evolving trajectories of individuals (Bernardi et al. 2019). Several studies have examined how school-to-work trajectories affect later employment prospects from this perspective. Educational pathways might be interpreted by employers to evaluate the productivity of candidates (Spence 1973). In Switzerland, educational pathways, including bridging programs between lower and upper secondary levels and early unemployment after vocational education, were found to be associated with lower income and occupational status (Sacchi and Meyer 2016). More generally, atypical educational pathways involving, for instance, delayed education or repetitions of schooling years may also be interpreted as signals of lower productivity or motivation by employers. These developments call for a deeper understanding of how education, considered as a process, is linked with later-life employment prospects such as income (Gomensoro and Bolzman 2015; Zimmermann and Seiler 2019).

Methodologically, estimating how an entire trajectory is linked with later-life outcomes is not straightforward. Three approaches can be distinguished. The first approach involves using a proxy for the key dimension of the previous trajectory. In educational research, educational attainment or the number of years spent on education could be used as Previous Trajectory Proxy (PTP). These proxy indicators are often collected retrospectively and do not require a full longitudinal design. However, they generally focus on a specific aspect and do not consider the whole trajectory. The second approach relies upon sequence analysis to build a typology of educational pathways, which is later used in a regression to explain income (e. g.,

1 Matthias Studer and Leonhard Unterlerchner gracefully acknowledge for the financial and research support of the Swiss National Science Foundation (project "Strengthening Sequence Analysis", grant No.: 10001A_204740). The figures are available in color at https://centre-lives.ch/sites/default/files/figunterlerchner2023.pdf.

Gomensoro and Bolzman 2015; Zimmermann and Seiler 2019). However, creating a typology implies simplification of the data, which might lead to wrong conclusions (Studer 2013). The third approach, introduced by Bolano and Studer (2020), proposes a procedure based on feature extraction and selection, to understand the specific aspects of a previous trajectory linked with later life outcomes.

This article aims to review these three methodological approaches and discuss their applications to the study of the relationship between educational pathways and income. We further compare their results and highlight their connection to different research questions. We show that sequence analysis and PTP approaches lead to similar results, even if the latter approach is more efficient from a statistical point of view. By design, sequence analysis focuses on the identification of recurrent trajectories and, therefore, fails to capture atypical trajectories. Alternatively, by adapting the feature extraction and selection approach to account for educational attainment, several aspects of atypical trajectories are found to be significantly associated with income at age 30.

Additionally, we propose a new method to validate a sequence analysis typology to be subsequently used in regression. The method measures the impact of the data reduction of cluster analysis on the studied relationship. It does so by computing the share of the original relationship, measured without prior clustering, that is accounted for by a typology. As illustrated by our application, it can be used to guide the choice of the number of groups. This method is available using the *clustassoc* function of the *WeightedCluster R* library (Studer 2013).

This article is organized as follows. We begin by briefly presenting the Swiss education system and its expected educational pathways before introducing the data and the coding of the variable. We then present each method and the associated results. We conclude by comparing the results and interpretations before making recommendations regarding the choice of method.

2 Overview of the Swiss Education System

The Swiss educational system is considered to be highly selective and to reproduce social inequalities (Meyer 2009). This inequality is often attributed to the lower secondary tracking system in which pupils are oriented early toward different types of upper secondary education (Gomensoro et al. 2017). While the pre-baccalaureate or extended requirement tracks at the lower secondary level allow direct access to all types of upper secondary education, including general baccalaureates and school-based vocational education and training (VET), the attendance of basic requirements tracks only allows direct access to firm-based VET. Therefore, lower secondary education tracking strongly impacts upper secondary education (Hupka-Brunner et al. 2010; Falter 2012; Buchmann et al. 2016). Those oriented in high

requirement tracks almost exclusively attend general education or a vocational baccalaureate (89 %), while most in basic requirement tracks attend VET (69 %) (Gomensoro and Meyer 2021).

The transition between lower and upper secondary education is often considered to be direct. However, more than 20 % of the students undertake one or two years of transitional solutions after compulsory schooling and thus experience discontinuities in the transition to secondary education (Sacchi and Meyer 2016).

At the upper secondary level, VET is the norm in Switzerland and enjoys a higher degree of social prestige in comparison to other countries. With a duration of two to four years and covering more than 230 different professions with significant variance in intellectual requirements, VET is undertaken by nearly two-thirds of young people in Switzerland (Gomensoro and Meyer 2021). These trainings provide a rapid transition into the labor market (OFS 2018), and offer good employment and income prospects, albeit lower than general (Korber and Oesch 2019) and tertiary diplomas (Gomensoro et al. 2017). During or post VET, apprentices can undertake a vocational baccalaureate, a diploma acquired by 14 % of young people in Switzerland. Vocational baccalaureate allows transitions between VET and general tertiary education, and more than two-thirds of those with a vocational baccalaureate pursue general or VET tertiary education (OFS 2018). General upper secondary education (general baccalaureate and specialized schools), undertaken by about one-third of young people in Switzerland (Gomensoro and Meyer 2021), allows access to general tertiary education.

Tertiary education can be categorized into vocational and general. Vocational tertiary education is accessible to students with work experience who obtain a secondary VET diploma. General tertiary education is accessible to students who obtain a general or specialized (and under certain conditions a vocational) baccalaureate (Gomensoro et al. 2017; Gomensoro and Bolzman 2019). General tertiary education can be undertaken either in universities or in universities of applied science/pedagogy. The latter is more oriented toward rapid labor market integration, and their curriculums tend to be shorter.

Previous findings on the link between education and income, demonstrated an income gradient across educational attainment levels. Holders of tertiary level credentials – either academic or vocational – earned the highest income. Among holders of secondary level diplomas, VET credentials are associated with the highest income at the beginning of the career, but this tendency is then reversed around the age of 30 (Gomensoro et al. 2017; Korber and Oesch 2019; Zimmermann and Seiler 2019).

3 Data

We rely on the first cohort of the TREE (TRansitions from Education to Employment) survey. This is a longitudinal follow-up of a nationally representative sample of compulsory school leavers. It includes comprehensive data measured at different points in time and covers a span of 14 years, based on nine survey waves (TREE 2016). The TREE data allows us to relate the trajectory between ages 15 and 30 with the wage perceived at age 30.

Educational pathways are captured using monthly episodic information on educational status from September 2000 to December 2014, i. e., 172 monthly records.[2] These educational pathways are described using six states; the first two states depict upper secondary education. Secondary II Vocational Education and Training (SECII VET) includes all types of two-to-four-year VET programs and vocational baccalaureate. Secondary II General Education (SECII Gen.) includes upper secondary general baccalaureates and specialized schools. The next two states characterize tertiary education. Tertiary General Education (TER Gen.) regroups universities and universities of applied sciences/pedagogy. Tertiary VET includes all types of vocational tertiary education such as higher specialized schools and higher VET. The fifth state is Transitional Solutions (TS) which groups bridging programs, pre-apprenticeships, internships, and non-awarding education. The final state is out of education or training (OET) which groups all the situations not involving education such as work, joblessness, or inactivity. We grouped these situations for two reasons. First, we aim to focus on educational trajectories, which do not include working status per se. Second, low level of complexity is required to comprehensively compare methodological approaches. This implies focusing only on one dimension of the life course: education.

Educational attainment is measured using the highest achieved qualification. Secondary VET qualifications are the most frequent and cover a wide range of occupations. Income expectations may differ dramatically depending on the economic sector, such as the banking or retail sectors. To account for this diversity, we distinguish between secondary VET levels using the Stalder and Nägele (2011) classification. This six-level ordinal scale depicts the requirement level and the prestige of secondary VET qualifications. We grouped levels one to three and levels four to six, along with commercial school diplomas.

The employment outcome is operationalized using full-time equivalent (FTE) income measured in 2014. This is a composite variable computed from all working episodes observed in 2014. Since we use full-time equivalent income, the results are not impacted by differences in employment rates (see Gomensoro 2022 for computation details).

2 Episodic data on education are currently not included in the TREE1 data release. They can be obtained from TREE on request.

Several control variables are considered to explain income. To avoid anticipatory analysis, these variables are taken from the baseline survey in 2000 at the end of compulsory school. We control for the sex of the respondent as educational choices are highly segregated (Gauthier and Gianettoni 2013; Imdorf and Hupka-Brunner 2015), and this helps to account for the gender wage differences (Bertschy et al. 2014; Korber and Oesch 2019). The lower secondary school track is of key importance and is implicitly linked with unmeasured skills (Meyer 2009). We also control for the linguistic regions as they favor different educational pathways (Scharenberg et al. 2017), and the labor markets and wages are different (Gomensoro et al. 2017). Finally, we rely on parental household wealth with a continuous variable as a proxy to control for social origin (Samuel et al. 2013).

Finally, we retained all cases without missing information, resulting in the final sample of 2 230 observations. Sampling weights are used throughout the analysis to compensate for attrition and deletion of missing data.

All the statistical analyses were performed with the *R* statistical software (R Core and Team 2020) using *Boruta* (Kursa and Rudnicki 2010), *TraMineR* (Gabadinho et al. 2011), and *WeightedCluster* (Studer 2013) packages.

4 Methodological Strategies

The aim of the article is to investigate the differences, complementarities, strengths, and weaknesses of the three methodological approaches to examine the relationship between a previous trajectory and a future-life outcome. Accordingly, we applied them to the study of the relationship between educational pathways and later-life income.

Each subsection presents one method before presenting the associated results. We further highlight the main required decisions and how to specify the typology in this setting. We use the feature extraction and selection (FES) approach and discuss its parametrization to understand the relationship between income and relevant atypical patterns observed in educational pathways. Finally, we compare the results and discuss their expected respective contributions.

4.1 Educational Attainment as Previous Trajectory Proxy

The relationship between education and income is often studied using educational attainment. Such a strategy implicitly summarizes the educational pathway followed by an individual using a single indicator, educational attainment. The highest educational attainment is a commonly accepted measurement of human capital and institutionalized cultural capital (Bourdieu 1979; Becker 1993). This information is even directly used by employers and employees to bargain wages. Therefore, it is expected to have a direct impact on income. This is a perfect example of the previous trajectory proxy approach as the whole educational process is operational using one

indicator. However, since it only relies on completed education, it does not capture bridge education programs (or any non-award study), incomplete, off-time, or repeated education spells. Any information on the trajectory itself is, therefore, omitted.

To study this relationship, we rely on linear regressions with income as a dependent variable. As informed in the data section, we control for linguistic regions, parental household wealth, sex, and lower secondary school track. The results of this linear regression are presented in the first column of Table 1.

These results are in line with previous findings (Gomensoro et al. 2017). Tertiary general degrees are associated with the highest income levels, followed by tertiary VET. No significant differences are observed between secondary educational levels, such as between secondary VET levels. Finally, individuals who have not achieved any qualification earn the lowest income. Interestingly, we do not observe any significant differences between secondary VET and general education levels.

In conclusion, we observe an income gradient following educational attainment. The PTP approach takes education into account by considering the cumulated acquired educational capital as described by Becker (1993). Furthermore, it does not require a longitudinal follow-up, and does not consider the path leading to this educational attainment. The use of Sequence Analysis (SA) precisely aims to focus on the path itself.

4.2 Sequence Analysis

Several authors have relied on SA to take educational pathways into account (Laganà et al. 2014; Pollien and Bonoli 2018; Zimmermann and Seiler 2019). Such an approach aims to consider education as a process, rather than only as an outcome. Correspondingly, it aims to understand how the pathways followed by individuals might affect their later-life income. Aside from educational attainment, the off-track educational pathways, for instance, those characterized by detours, reorientations, or longer than expected durations, may cause different school-to-work transitions and career paths (Brzinsky-Fay 2007; Achatz et al. 2022).

SA aims to provide a holistic view of trajectories or processes by creating a typology of the trajectories. This typology is expected to describe the different (ideal-) types of trajectories observed in the data. They can then be used in regression models to study the link between the type of previous educational pathway and later-life income. Here, we start by describing the typology creation, highlighting the key decisions to be made, and the recent developments in SA to answer common criticisms. We also propose a new method to validate the typology to be used in subsequent regression. We then use the typology in the same regression model as the educational attainment model before discussing the advantages of each approach.

Table 1 Regression Models of Full-Time Equivalent Income Depending on Educational Pathways Using a Previous Trajectory Proxy (PTP), Sequence Analysis (SA) and Feature Extraction and Selection (FES)

	PTP		SA		PTP & SA		PTP & FES	
Educational attainment (ref.: higher level of SECII VET)								
Lower level of SECII VET	−953.26	(442.04)*			−849.82	(442.22)	−778.64	(440.14)
Level of SECII VET missing	−324.27	(272.92)			−280.12	(274.35)	−320.30	(272.13)
Lower secondary	700.27	(578.57)			740.76	(577.38)	711.66	(570.47)
Secondary general	−484.55	(261.54)			−402.63	(280.05)	−322.97	(270.12)
Tertiary VET	1 346.14	(224.38)***			1 354.30	(256.54)***	970.91	(274.52)***
Tertiary general	1 595.09	(221.66)***			1 800.44	(325.72)***	1 809.31	(324.62)***
Educational pathways (ref.: secondary VET)								
Secondary VET and tertiary general			1 589.15	(231.41)***	384.81	(302.07)		
Secondary VET and tertiary VET			1 145.12	(213.56)***	203.54	(251.51)		
Short tertiary general			1 284.91	(239.83)***	−165.93	(330.62)		
Long tertiary general			621.48	(258.67)*	−900.84	(352.31)*		
Secondary general and tertiary VET			305.33	(270.41)	−326.71	(295.28)		
Sequence features								
TS → SECII VET							−791.78	(182.21)***
SECII VET → TS							550.82	(161.06)***
SECII VET → TER VET							686.06	(235.50)**
Start of SECII Gen in 2nd year							−1 833.93	(570.59)**
Start of SECII VET in 7th year							−1 147.63	(779.80)
Time spent out of education (ref.: under 3 years)								
Between 3 and 9 years							1 284.11	(306.12)***
More than 9 years							1 476.83	(372.34)***
Time spent in TER Gen.							2.74	(4.50)
Lower secondary track (ref.: extended academic requirements)								
Basic academic requirements	−437.38	(197.08)*	−579.83	(194.05)**	−446.41	(196.75)*	−414.27	(195.20)*
Pre-gymnasial	−140.51	(176.17)	72.99	(182.93)	−6.72	(179.45)	−4.75	(177.93)
No formal tracking	−609.84	(514.28)	−560.34	(522.65)	−532.50	(512.88)	−603.70	(507.81)
Sex (ref.: female)	1 108.98	(139.05)***	961.11	(148.87)***	963.90	(145.79)***	975.31	(139.87)***

Continuation of Table 1 on the next page.

Continuation of Table 1.

	PTP		SA		PTP & SA		PTP & FES	
Linguistic region								
French	–467.88	(184.76)*	–576.48	(187.54)**	–375.91	(185.57)*	–377.99	(182.68)*
Italian	–902.14	(326.96)**	–705.12	(338.43)*	–639.20	(332.00)	–720.74	(326.64)*
Parental household wealth (continuous)	268.42	(87.08)**	297.27	(88.38)***	264.86	(86.78)**	259.04	(85.85)**
Constant	4 981.06	(195.06)***	5 027.46	(158.02)***	4,950.90	(201.10)***	3 530.31	(399.95)***
Observations	2 230		2 230		2 230		2 230	
R^2	0.12		0.08		0.13		0.15	
BIC	43229.28		43301.85		43245.16		43217.28	

Note: *p < 0.05; **p < 0.01; ***p < 0.001. SECII Gen: Secondary General Education; SECII VET: Secondary VET; TER VET: Higher Vocational Education and Training; TER Gen.: Tertiary General Education; TS: Transitional Solutions; OET: Out of Education or Trainin.
Source: TREE wave 1, author's calculation

Typology

The typology of sequences is created in three steps. First, all individual sequences are compared to one another using a distance measure. Second, based on the previous information, similar sequences are grouped using cluster analysis. Finally, the typology is validated using one of the available methods. We discuss their relative strength and propose a new method.

In order to group similar trajectories, one needs to compare them. This step is technically achieved by using a distance measure. From a substantive point of view, this comparison is carried out by relying on a criterion. The life-course perspective emphasizes several key aspects of trajectories to be taken into account and could serve as criteria to compare sequences (Settersten and Mayer 1997; Billari et al. 2006). These aspects can be grouped into timing, duration, and sequencing aspects of trajectories (Elzinga and Studer 2015; Studer and Ritschard 2016). Each of these aspects can be theoretically linked with later-life income.

The *sequencing* refers to ordering of the various states of educational pathways, and codes the dynamics of the trajectory. *Sequencing* also captures the *quantum,* i. e., the presence or absence of specific steps in trajectories. Differences in sequencing are often considered to have strong later-life consequences. For instance, Sacchi and Meyer (2016) showed that following a bridging program might hinder the careers of those experiencing it and therefore result in lower income. Similarly, pathways characterized by many back-and-forth movements in education are expected to result in lower income, as educational attainment and occupational experience would typically remain low. The *timing* aspect refers to when states and transitions are experienced within trajectories. Differences in *timing* are also expected to have later-life consequences, an idea framed under the *critical or sensitive period model* in epidemiology (Kuh et al. 2003). For instance, delayed start of tertiary education has been found to result in lower income in the United States (Yu 2021). Finally, the *duration* aspect corresponds to the time spent in each state. Longer tertiary education has been found to be associated with lower income prospects in Italy (Aina and Casalone 2011). The time spent in education can, for instance, be lengthened by reorientations or year repetitions

Several distance measures, which differ according to their sensitivity to the above-mentioned aspects, are available (Studer and Ritschard 2016). The choice among them should be made according to the aspects regarded to be theoretically relevant by the researchers. In our application, all three aspects are important, but a choice is required. We used the optimal matching distance with standard parameters, which is shown to be sensitive to the sequencing and duration aspects (Studer and Ritschard 2016). Sequencing is a key aspect to uncovering the various (potentially atypical) successions of education spells and coding the path leading to tertiary education for instance. The duration aspect is central for capturing the duration of tertiary education or repeated years.

Once the sequences are compared, cluster analysis can be used to create a typology by grouping similar trajectories and assigning dissimilar ones to different groups. Here, we use the Partitioning Around Medoids (PAM) algorithm, which aims to minimize a global criterion (Han et. al. 2017).

Cluster analysis always produces a typology, which may or may not reveal a structure found in the data (Levine 2000). Specifically, the typology might be a statistical artifact. Therefore, it is imperative to carefully evaluate its quality, which is also a key step in choosing the number of groups or the clustering algorithm. In this article, we first evaluate the statistical quality of the typology using cluster quality indices (CQI), as usually recommended (Piccarreta and Studer 2018). We then propose a new method specifically designed to assess the quality of a typology to be used in a subsequent regression. Further, we show how it overcomes a common issue with the CQI's approach.

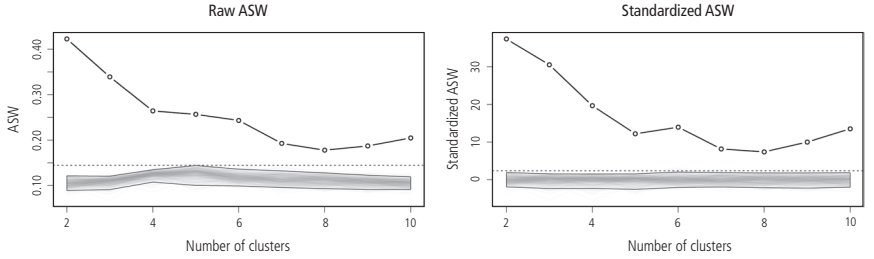
The statistical quality of a clustering can be measured using several CQIs (see Studer 2013 for a review). These indices typically take into account the homogeneity of the types and their separation. Cluster analysis simplifies the data by reducing the differences between all the individual sequences to the differences between a few types. This is a necessary step to understanding the diversity of the trajectories, but there is an *oversimplification* risk. If the types are homogeneous, this risk is low. Conversely, cluster separation ensures that we are not creating unnecessary distinctions between trajectories. The Average Silhouette Width (ASW) is the most used index that balances these two aspects.

The left-hand side plot of Figure 1 presents the value of the ASW for different groups. The best clustering solutions are then found for two groups, which show the highest ASW.

Since CQIs lack clear interpretation thresholds, Studer (2021) proposed comparing the obtained CQI to the ones obtained by clustering randomly generated *similar* but *non-clustered* data, i. e., the CQI values obtained when we should not cluster the data. These CQI values, obtained by clustering non-clustered data, are represented using thin gray lines in Figure 1. Using these CQI values, a more formal statistical test for the presence of a clustering structure in the data can be derived. The threshold value of this test, which accounts for multiple testing, is represented using a dotted horizontal line. Since all CQI values are above this line, we can conclude that a "significant" structure is found in the data for all number of groups. These CQI values can also be used to standardize the observed ASW values, making them more comparable across different groups. The standardized values are represented in the right-hand side plot of Figure 1. These standardized ASW values still favor clustering with two or three groups, but a local maximum is found for six groups.

When the typology is subsequently used in regressions, any within-cluster variation of the trajectories is ignored. Indeed, all trajectories clustered in the same type are described by a single value, i.e., the type itself. As shown by Studer (2013),

Figure 1 Average Silhouette Width (ASW) Values for Varying Number of Groups and ASW Values Obtained by Clustering Randomly Generated Similar but Non-Clustered Data Using the Combined Null Model



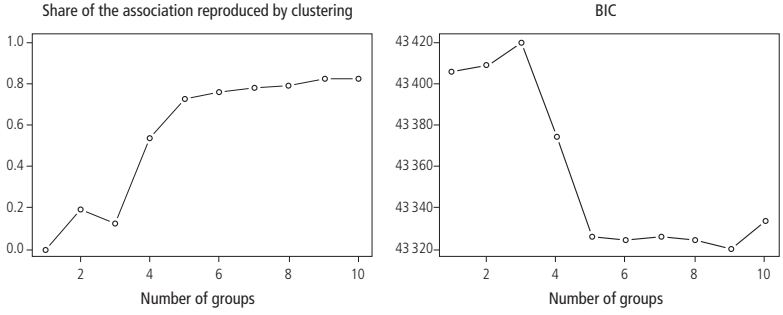
Note: see Studer, 2021. Figure available in color at https://centre-lives.ch/sites/default/files/figunterlerchner2023.pdf#page=1.
Source: TREE wave 1, author's calculation.

the remaining within-cluster heterogeneity can lead to wrong conclusions, if it is linked with the outcome of interest, i. e., income in our case. Here, the clustering in two groups mostly distinguishes secondary VET without further educational curriculum from other pathways (See Figure 5 in the appendix). However, if the distinction between secondary general education pathways followed or not followed by tertiary education is important to explain future income, we would not capture it with this typology. Since all these pathways are now described using the same type, we would wrongly conclude that the educational pathway is not relevant to explain income. In this article, we propose a new method to ensure that we are not simplifying relevant variations of the trajectories. This method is made available through the *clustassoc* function of the *WeightedCluster R* library.

The relationship between trajectories and covariates can be studied directly using discrepancy analysis (see Studer et al. 2011). This method evaluates the strength of the relationship with a Pseudo-R^2, measuring the share of the variation of the trajectories explained by a covariate and the statistical significance of the relationship. The method works without prior clustering, and therefore, without data simplification. However, discrepancy analysis has a strong limitation. There is no indication of how the trajectories differ according to the included covariates. Consequently, most studies continue to rely on cluster analysis.

In our case, income is found to be significant using discrepancy analysis, and the strength of the association is measured with a Pseudo-R^2 of 0.71 %. It should be noted that low Pseudo-R^2 values are expected in SA because the diversity of the trajectories is generally very large (Studer et al. 2011; Liao and Fasang 2021). However, there is no indication of how the pathways vary according to income.

Figure 2 Evolution of Cluster Quality Measures for a Typology Used in Subsequent Regression



Source: TREE wave 1, author's calculation.

Multifactor discrepancy analysis extends the previous method and allows measuring a relationship while controlling for other covariates. Here, we propose to measure the relationship between trajectories and income while controlling for the typology. If the income's Pseudo-R^2 remains at the same level, it means that the typology does not capture the relationship between income and trajectories. In other words, the typology simplifies all the relevant information to capture this relationship. Conversely, if the income Pseudo-R^2 is much lower, it means that the typology reproduces the key information to understand the income–pathway relationship. Using this strategy, we can compute the share of the original Pseudo-R^2 that is taken into account by our clustering.

Figure 2 shows the evolution of the share of the original association reproduced by the clustering for a varying number of groups. A minimum of five groups is required to reproduce most of the association (approximately 80 %), and nine groups would give the best results. However, this means that 20 % of the variation cannot be reproduced by a typology. By comparison, with only two groups, as recommended by the ASW, one would only reproduce 20 % of the association and overlook 80 % of the relationship between educational pathways and income.

Han et al. (2017) proposed a similar strategy to evaluate typologies. Their procedure evaluates the efficiency of a typology to explain a key variable, such as income, using the Bayesian Information Criterion (BIC), which measures the explanatory power of a typology while accounting for complexity. As a recall, the BIC should be minimized. The right-hand side plot of Figure 2 presents the evolution of the BIC of a linear model explaining income according to a typology in different number of clusters. The general trend follows our previous approaches and the results lead to the same conclusion. However, the BIC approach faces two issues. First, the data reduction of the clustering can lead to the creation of a statistical relationship

(see Studer 2013), and this would not be captured using the BIC. Conversely, the proposed approach would identify it, as the association would not be significant from the beginning. Second, the BIC approach provides no information on the extent to which the data reduction of clustering affects the considered relationship.

Summarizing, the ASW favors clusterings with few groups with a local maximum for six groups and any number of groups is found to be "significant." Meanwhile, the newly developed method shows that at least five groups are required to describe the relationship between educational pathways and income. Therefore, we retain six groups, which are presented in Figure 3. The newly developed method further highlights that 20 % of the association with income cannot be reproduced using a typology with fewer than ten groups.

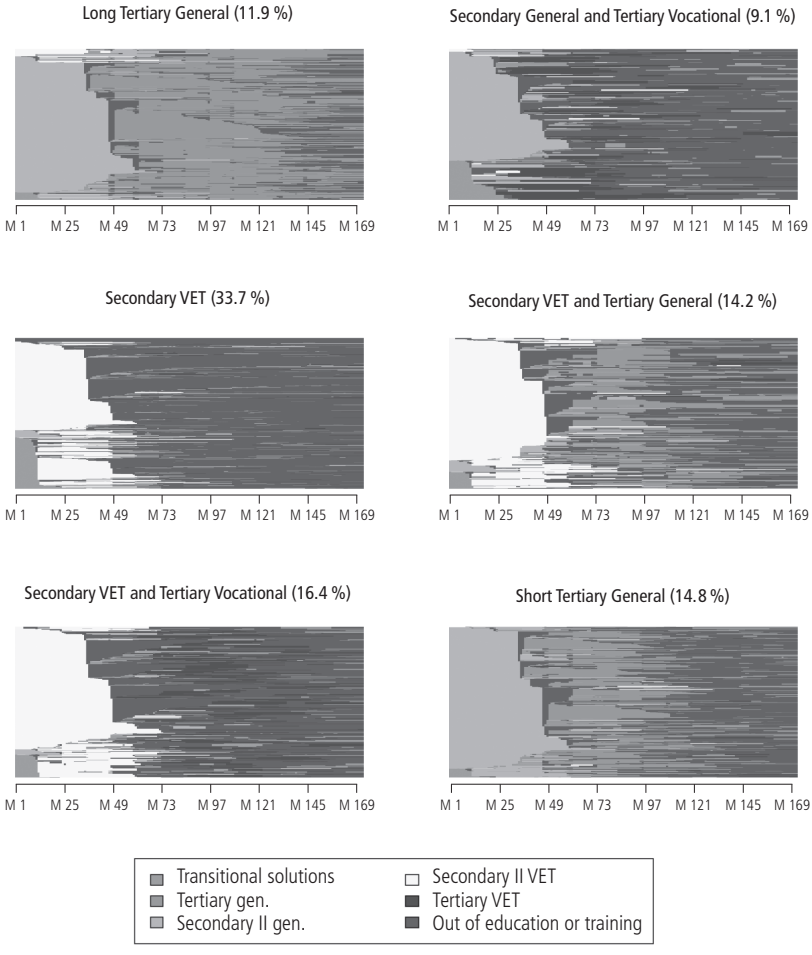
The typology, though obtained using explorative methods, is in line with the Swiss education system. It clearly distinguishes educational pathways according to the upper secondary track followed by an individual. There are three types of pathways that start with a spell in secondary VET education.

The first type is *Secondary VET,* characterized by a direct transition out of education (usually meaning employment). It corresponds to the expected VET curriculum of the Swiss education system. About 80 % of these pathways end with a VET qualification, either from the secondary or tertiary level. Interestingly, approximately half of these pathways start with a spell in transitional solutions, such as bridge education programs. Therefore, the typology cannot distinguish pathways with or without transitional solutions. The second type identifies the pattern of *Secondary VET* and *Tertiary General* education, where 65 % of the pathways end up with a tertiary general qualification. Finally, the last type describes the patterns of secondary VET followed by tertiary VET education, where 68 % of the students obtain a tertiary VET qualification. Collectively, these three groups account for 64 % of the observations in the sample.

The second set of pathways starts with a general education spell. A first type, named *Short Tertiary General,* groups sequences characterized by general secondary education followed by higher tertiary education, with a median time of 48 months spent in tertiary education. In this group, 81 % of the individuals hold a tertiary general degree. The *Long Tertiary General* group is similar to the previous one, but the median time spent in higher tertiary education is 86 months. Approximately 90 % of the individuals in this group obtain a tertiary general degree. Finally, a third type identifies those pursuing secondary general education followed by tertiary VET education. More than 50 % of the individuals in this group obtained a tertiary VET qualification, but 31 % ended up with a secondary general qualification. This type regroups relatively diverse pathways, including those starting with transitional solutions.

The typology seems to effectively describe the expected pathways in the Swiss education system. Consequently, the typology is strongly associated with the high-

Figure 3 Typology of Educational Pathways in Six Groups, Sequences Ordered From the Starting State



Note: Individual educational pathways, in months. Figure available in color at https://centre-lives.ch/sites/default/files/figunterlerchner2023.pdf#page=3.
Source: TREE wave 1, author's calculation.

est educational attainment (Cramér's V = 0.46). However, the SA approach makes further distinctions between pathways ending with the same qualifications but differing in spell lengths. This is the case for the *Short Tertiary General* and *Long Tertiary General* groups. It also distinguishes the steps leading to tertiary education. The typology reveals a group following secondary VET before entering tertiary VET and another group following secondary general before entering tertiary VET. However,

the typology also fails to identify the pattern starting with transitional solutions or the pathways ending without qualifications. Those patterns are included in broader types. Furthermore, the secondary general followed by tertiary education type seems to regroup relatively diverse pathways.

Regression Model

Once the typology of the educational pathway is created, it can be used in regression models as any categorical variable. Correspondingly, we can investigate the relationship between income and education, based on the pathways instead of the highest educational attainment. To obtain comparable results, we used the same control variables. The results are presented in Table 1.

The results of the two approaches are similar. In both cases, tertiary education is associated with higher wages followed by secondary education. However, the SA typology makes further distinctions within tertiary education. Tertiary general education following secondary VET is associated with the highest wages, followed by the tertiary vocational and short tertiary general types.

The type *Long Tertiary General* is associated with a lower wage. Individuals in this group might not have the time to complete their occupational integration in the studied age range (until age 30). They are still at the beginning of their occupational careers and have limited occupational experience. Measuring income at an older age would probably lead to a different conclusion.

We observe lower wages if tertiary VET follows general education, as no significant difference with *Secondary VET* type is found. However, as already discussed, this group is relatively heterogeneous. It is, therefore, difficult to draw any clear conclusion. Finally, the SA typology cannot capture individuals without any qualifications.

From a statistical perspective, the highest educational attainment model provides a better performance than the SA model, as it explains a higher share of the income variation ($R^2 = 11.7\%$ vs. $R2 = 8.5\%$). It is also a more parsimonious model according to the BIC. This can probably be explained by the fact that the SA typology only imperfectly accounts for educational attainment, which remains a vital piece of information to explain later-life income. Further distinctions made by the SA typology on the pattern leading to tertiary education or on time spent in tertiary general education do not fully compensate for this loss. It also fails to account for the difference in wages of those without qualifications. The similarity of the conclusions is confirmed by looking at the model including the PTP and SA approaches, where the coefficients of the typology becomes non-significant, except for the distinction between short and long tertiary general education. This is further confirmed by the comparison of the BICs, which shows that the increase in explanatory power is not compensated by the loss of parsimony. Logically, the recurrent pathways are those leading to the main educational attainments; therefore,

both approaches convey similar statistical information. As pointed out by Labussière et al. (2021), the focus of SA on common patterns makes it difficult to shed light on atypical paths. However, it provides a useful holistic view on common trajectories, in line with the life-course perspective (Bernardi et al. 2019).

4.3 Feature Extraction and Selection

Bolano and Studer (2020) proposed using the feature extraction and selection (FES) procedure to study the link between a previous trajectory and later-life outcome. This method works in two steps. First, potentially interesting characteristics are extracted from the trajectories. Second, the characteristics that are statistically relevant in explaining the outcome are selected. Overall, the method aims to identify the key characteristics of trajectories to explain income.

In this section, we start by presenting the method and its application to study the relationship between educational pathways and income. In the meantime, we discuss its use to understand how the critical characteristics of educational pathways are linked with later-life income.

Feature Extraction

The first step is to identify characteristics of the trajectories that might be relevant to identify later-life income. Bolano and Studer (2020) proposed a framework to automatically extract a set of these characteristics, called features. Their procedure aims to capture key aspects of trajectories according to their duration, timing, and sequencing. These three aspects are potentially relevant as they can each be associated with common life-course models linking previous trajectories and later-life outcomes (Bolano and Studer 2020). However, the authors insist on the need to adapt it to the research objectives. In this section, we briefly describe the extracted features, their relevance for our research, and their required specifications.

The *duration* aspect of the trajectories refers to the time spent in each state. This aspect covers the potential impact of exposure to a given situation (Kuh et al. 2003; Bolano and Studer 2020). For instance, the overall exposure to tertiary education is typically considered to have later-life consequences. This aspect also aims to capture the potential effect of deviation from the expected spell duration (sometimes referred to as "spacing," see Settersten and Mayer 1997). This is a key aspect of educational trajectories, where educational spells have expected durations. Deviations from these durations, resulting (for instance) from repeating a year, could be linked with later-life income (Aina and Casalone 2011). In the framework developed by Bolano and Studer (2020), the *duration* aspect is measured by creating one variable storing the overall time spent in each of the states. We, therefore, end up with eight duration features.

The *timing* aspect refers to when state and transitions occur in trajectories. Deviation from expected timings is also expected to have consequences on later-life

income (Aina and Casalone 2011). Typically, starting or ending an education spell at an unexpected age could be interpreted as a lack of commitment by employers. Bolano and Studer (2020) propose to capture this aspect by considering the timing of key transitions and events. In our case, we considered the timing of the start and end of each education spell. Technically, the timing of events is captured by measuring whether each event occurred in predefined age ranges. The dummy variable resulting from it takes the value 1 if the transition occurs in the given age range, and 0 otherwise.

In our application, we used the same age ranges for every transition, by considering periods of 12 months. Transitions are measured yearly, which should not lead to a significant loss of information since the start and end of education spells are expected to take place once a year. However, it could also be interesting to consider theoretically driven age ranges. For instance, one might code whether the transition is occurring at before, or after the expected time. The procedure resulted in the automatic creation of 106 features to measure the timing of transitions in the trajectories.

Finally, the *sequencing* aspect refers to the ordering of the state and the presence of specific spells in trajectories. Typically, following a bridge education spell before secondary VET or experiencing a direct transition might be interpreted differently. Bolano and Studer (2020) propose to rely on frequent subsequence mining to identify frequent successions of states within trajectories. This method automatically finds recurrent subsequences of varying lengths and creates dummy variables storing the presence or absence of each pattern. For instance, the pattern "SECII VET" stores the presence of a secondary VET spell in trajectories. The feature "SECII VET → TER VET" stores the presence of a secondary VET spell followed, directly or later on, by a tertiary VET spell. Due to the first pattern being embedded in the second one, the two features are overlapping. All patterns of a maximum length of three occurring in at least 5 % of the trajectories were considered, resulting in 170 sequencing features.

The feature extraction step aims to automatically identify potentially relevant aspects of the trajectories to explain later-life income. Overall, we extracted 282 features of pathways. However, these features are most probably not all relevant to explain later-life income.

Features Selection

The second step is to select the relevant features to explain later-life income. This selection is automatically made using a machine-learning algorithm. Following Bolano and Studer (2020), we relied on Boruta (Kursa and Rudnicki 2010). This algorithm aims to select all the features that are relevant to explain the outcome, even if they are redundant.

Technically, Boruta is an extension of random forest, an ensemble method based on regression trees. The random forest works by estimating many regression trees (500 in our cases) to explain the outcomes. These trees are estimated on random subsamples of the data and features. All these trees are then combined (i. e., averaged) to predict the outcome. Random forest has several key features. First, it can capture non-linear relationships and interaction effects between features. Second, it can be used to identify the features that are recurrently important in explaining the outcome by computing an "importance" score. However, this score lacks well-defined threshold values.

In order to select a subset of significant features, multiple testing should be taken into account. This is a key issue when considering 282 potential features. With the usual 5 % threshold, one could expect $0.05 \cdot 282 = 14.1$ features to be flagged as significant if all 282 features were independent of the outcome. The Boruta algorithm provides threshold values for the importance scores, which allows the selection of "significantly important" features. Internally, this is achieved using permutation tests taking into account multiple testing.

Bolano and Studer (2020) discuss the controlling of confounders in the selection process. This is performed by residualization of the outcome variable, i.e., income. The method allows avoiding selecting features that are indirect measurements of key dimensions, such as gender or unmeasured skills. To keep our presentation simple, we controlled for sex (as a gender pay gap is expected), and lower secondary school track. In this article, we propose to further control for the highest educational attainment. Accordingly, we aim to identify the features of the pathways explaining income, net of the effect of educational attainment. We, therefore, look specifically at the characteristics of the pathways that are not taken into account by educational attainment, such as incomplete or delayed education.

Selected Features

Table 2 presents the 25 features selected by Boruta and their "importance" score: four are related to durations, seven are related to timing aspect, and 14 are related to sequencing.

To simplify the interpretation of the results, we propose relying on correlation plots and feature clustering. Figure 4 presents a graphical representation of the Pearson correlation between the selected features, hatched circles represent negative correlations. This plot highlights the overlaps in the information carried by the features. The features are ordered according to their similarities, and the clustering of the features in nine groups is represented using black squares. Combined with Boruta's importance scores, we can use this clustering to guide our interpretation.

The time spent in secondary VET is the most important feature, but it is highly correlated with several other features, making it difficult to interpret. Looking more closely at the features involving secondary VET, several patterns can be identified.

Table 2 Features Selected by the Boruta Algorithm

Features selected			
Sequencing		Duration (time spent)	
TS → SECII VET → OET	8.4	SECII VET	10.2
TS → SECII VET	8.0	SECII Gen	7.4
SECII VET → OET → OET	6.3	OET	4.3
TS → OET → SECII VET	6.2	TER GEN.	4.0
TER GEN. → OET → OET	5.0	Timing of transitions	
TER GEN. → OET → TER GEN.	4.5	Start SECII Gen in 2nd year	5.8
OET → SECII VET	4.0	Start TER Gen. in 3rd year	4.5
SECII VET → TS	4.0	Start SECII VET in 7th year	4.4
SECII VET → OET → TER VET	3.9	End TS in 2nd year	4.1
OET → TER GEN. → TER GEN.	3.8	End TER Gen. in 5th year	3.9
SECII Gen → OET	3.5	End SECII VET in 5th year	3.6
SECII VET → TER VET	3.5	End TS in 1st year	3.4
SECII VET → TS → OET	3.3		
SECII Gen	3.2		

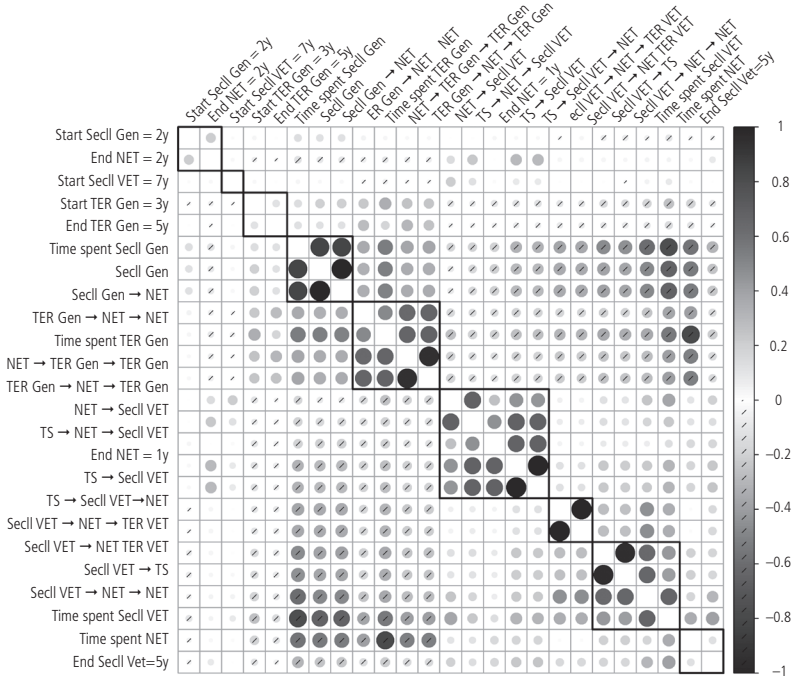
Note: SECII Gen: Secondary General Education; SECII VET: Secondary VET; TER VET: Higher Vocational Education and Training; TER Gen.: Tertiary General Education; TS: Transitional Solutions; OET: Out of Education or Training.
Source: TREE wave 1, author's calculation.

First, and most importantly, several features capture the pattern of transitional solutions or OET before secondary VET (*TS → SECII VET → OET*; *TS → SECII VET*; *TS → OET → SECII VET*; *OET → SECIIVET*; *End TS in 2nd year*).
Second, different patterns of further education are found to be relevant, including tertiary or transitional solutions spells (*SECII VET → TS*; *SECII VET → OET → TER VET*; *SECII VET → TER VET*; *SECII VET → TS → OET*). The feature "*SEC VET → OET → OET*" regroups both situations because it implies an education spell between the two spells of OET. Interestingly, starting secondary VET education seven years later also seems to be linked with later-life income (feature *start SECII VET in 7th year*). Ending secondary VET later is also selected but its interpretation is difficult, as the feature is linked with longer VET education and bridge formation.

Conversely, secondary general education is also listed among the most important features, either through its presence in the patterns (*SECII Gen. → OET and SECII Gen.*) or the overall time spent in it. Interestingly, the delayed start of secondary general education is also flagged as important (*start SECII Gen. in 2nd year*).

Several features linked to tertiary education are selected, either following vocational or general secondary education spells. The overall time spent in tertiary general education is strongly linked with several patterns including the timing of

Figure 4 Correlations and Clustering of the Features Selected by the Boruta Algorithm



Note: Figure available in color at https://centre-lives.ch/sites/default/files/figunterlerchner2023.pdf#page=4.
Source: TREE wave 1, author's calculation.

higher education spells (*End TER Gen.in 5th year*; *Begin TER Gen. in 3rd year*), and back-and-forth movement in tertiary general education (*TER Gen.* → *OET* → *OET*; *TER Gen.* → *OET* → *TER Gen.*; *OET* → *TER Gen.* → *TER Gen.*).

Finally, the time spent not in education nor training is strongly negatively correlated with secondary general and tertiary education spells. Its interpretation is not straightforward as it mostly regroups working spells and is therefore linked with working experience. Looking at the partial dependence plot (see Figure 6 in the appendix), a non-linear relationship and interaction effect with other features can be identified. There is a substantial average income increase for those with more than 36 months in OET, followed by another increase after 9 years in OET. However, the latter increase is found only for some observations (depending on the value of the other features) and a stable evolution, or even a decrease, is expected for others.

Regression Model

Boruta aims to identify all the features linked with an outcome. Based on random forest, it can capture non-linear relationships and interaction effects between features. However, even with the use of partial dependence plots, the interpretation of how the features are related to the outcome remains difficult. To understand these relationships, Bolano and Studer (2020) proposed to include the features in a regression model. This also makes the results comparable with the other approaches presented so far. However, the features are overlapping and strongly correlated, and therefore, including all of them in the same model would raise multicollinearity issues. Hence, they rely on a carefully and comprehensively chosen subset of the features.

Following our discussion, we selected seven features. We selected the patterns *TS* → *SECII VET* to capture bridge formation and *SECII VET* → *TER VET* and *SECII VET* → *TS* to measure the two types of further educational spells following secondary VET. We included the time spent in tertiary education and the one in OET in three categories as identified before. Finally, the features related to delayed entry into secondary education were also added.

We used the same controls as in the previous models to make them comparable. Our specification of the selection procedure controlled for the highest educational attainment, to emphasize the specific effects of the features. Therefore, we also included it in the model. This model is presented in the fourth column of Table 1.

Compared with the previous models, the coefficients of the highest educational attainment are similar in their direction, but not in their size. The coefficients of no qualification and tertiary VET qualification are lower compared to the previous model and the one for tertiary general education is greater. This is directly related to the features added to the model.

Delayed entry into general secondary education and transitional solutions before secondary VET is associated with lower income. These results are in line with the one from Sacchi and Meyer (2016), who found lower income and occupational status attainment for those following bridge programs after compulsory education.

Continuing education after a secondary VET qualification is positively associated with income, as shown by the coefficient of the patterns "*SECII VET* → *TS*" (which includes non-awarding education) and "*SECII VET* → *TER VET*." The relationship with tertiary VET is even stronger if we consider that higher educational attainment is also taken into account and is significant. It shows that the path leading to tertiary VET is relevant, probably because it leads to different types of institutions and, therefore, qualifications. The Federal Office of Statistics (OFS 2021) reached similar conclusions. They found that further education spells following secondary VET are frequent and associated with upward mobility and higher income.

The time in OET is highly significant and shows a non-linear relationship. Those who spent less than three years in OET have lower incomes. This could be related to a lower work experience or a shortened occupational integration process.

The relationship is non-linear and bounded, as a longer time OET is only linked to a lower increase.

The time spent in higher general education is not significant on top of tertiary general qualification. Further investigation shows that this is directly linked with the time in OET, as the two are strongly negatively correlated. In this sense, the results presented here are fully compatible with the one from SA, where we found lower income for those experiencing long tertiary education spells. However, the "feature" approach provides a clearer interpretation, by emphasizing the role of work experience.

From a statistical point of view, the "feature" model is better than the SA or PTP models; it explains a higher share of income variation and is more parsimonious according to the BIC, even if some of the included features remain non-significant. These results were expected, as our parametrization of the FES approach combines information from the highest educational attainment and the pathways, and therefore relies on more information. The key conclusions of the PTP and SA models are also confirmed by the FES approach, which further highlights the role of bridge education. Generally, these results highlight that aside from educational attainment, the paths leading to it are also important. Atypical paths or delayed timing might well be interpreted as signals (Spence 1973).

5 Conclusion and Discussion

In this article, we compared three methods to study the relationship between educational pathways and later-life income. We started with educational attainment as an illustration of the Previous Trajectory Proxy (PTP) approach. Here, we primarily aimed to offer a contrast to the other methods and highlight their relative benefits. Unsurprisingly, the approach mostly underlined the income gradient following educational attainment level and confirmed previous research (Gomensoro et al. 2017; Korber and Oesch 2019). However, by design, it does not consider the path followed by individuals and does not require longitudinal data.

The Sequence Analysis (SA) approach focuses on educational pathways described as a sequence of educational situations. It then provides a typology of recurrent or typical paths observed in the data. While it requires longitudinal data, it provides a holistic view of trajectories in line with the life-course perspective. The obtained typology distinguishes educational pathways according to the upper secondary track followed by individuals, and then by tertiary education or the end of education spells. The typology is strongly associated with educational attainment, but still makes further distinctions according to the spell lengths in some states, or the steps leading to tertiary education. However, atypical patterns, such as those starting with transitional solutions or ending without qualifications, are not identified (for

instance, in our research one cluster regrouped quite diverse pathways). Labussière et al. (2021) explored alternative coding of trajectories to overcome this limitation.

When used in a regression, the SA typology also emphasizes the income gradient following the education level. It further highlights that the path leading to tertiary education is relevant and the lowest wage is earned by those following long tertiary education pathways, due to their lower occupational experience. Similar results were reported by Zimmermann and Seiler (2019), both on the typology obtained and on its relationship with income.

The educational attainment model showed better performance from a statistical perspective than the SA model. The SA typology only imperfectly accounts for educational attainment, which provides key information to explain later-life income. This is expected, as no information on graduation was included in the coding of the pathways. However, the typology remains highly associated with educational attainment and, therefore, conveys similar information.

Finally, we used the feature extraction and selection (FES) approach. By adapting the selection procedure to account for educational attainment, we identified 25 educational pathway features linked with later-life income. A comprehensive selection was required before specifying a regression model. Our use of correlation plots and hierarchical clustering allowed the identification of several key aspects, some of which were previously highlighted by the SA typology. Interestingly, the approach highlighted the importance of delayed entry and transitional solutions before upper secondary education. This confirms the results from Sacchi and Meyer (2016) on the negative relationship between income and bridging solutions. It also emphasized the role of awarding and non-awarding educational spells following secondary VET. Finally, it showed that a minimal occupational experience of three years is important to explain income at age 30. This shed new light on the results of long tertiary education spells identified with SA.

From a statistical point of view, the FES approach leads to the best regression model. This is expected, since it combines information on educational attainment with information on the pathway followed by individuals. Our results showed that the path itself is important to understand later-life outcomes. Although causal interpretation is not warranted, critical patterns such as bridge years or continuing education as well as the path followed until a given type of credentials are associated with later life income. Negative and positive signals, as well as skill acquisition, might play a role in the sociological understanding of these findings. However, educational attainment remains a key characteristic to explain later life income.

This methodological comparison illustrates the strengths of each approach. The PTP approach can lead to clear results if it is well-defined and captures the key aspects of the previous trajectory. However, it often may fail to capture the relationship with more peculiar aspects of the previous trajectory. SA provides a descriptive and holistic perspective on recurrent trajectories. However, it might fail to capture

the relationship between trajectories and an outcome. The proposed methodology can help choose the number of groups and lower the risk of misleading conclusions. Nevertheless, we do not recommend a simple application of SA if atypical trajectories are of interest. Alternative coding of trajectories or other distance measures might provide better results in such cases (Labussière et al. 2021). Finally, FES allows a detailed understanding of the characteristics of a previous trajectory that are linked with the outcome. The selection criteria obtained through residualization can be further specified to consider confounders and / or PTP, such as educational attainment. However, the selected features are redundant, and their number might be overwhelming, even with the proposed use of correlation plots and hierarchical clustering. By doing so, FES can refine the results obtained with PTP and sharpen them by identifying key dimensions lacking in the proxy.

We also made several methodological contributions. First, we discussed the specification of the selection procedure to account for educational attainment. Generally, such specification allows us to understand the specificities of the path taken, aside from the current situation, to explain a later-life outcome. This has many applications in life-course research.

Second, we developed a new validation technique for SA typologies to be used subsequently in regression. The method works by looking at the share of the direct relationship between sequences and the outcome that is accounted for by the typology. It ensures that the simplification carried out by the clustering step of SA does not eliminate key information to explain an outcome. In our analysis, the method emphasized the need to use at least five groups to understand the relationship between recurrent trajectories and income, instead of only two groups for the usual cluster quality index (CQI) approach. It further highlighted that, even with five groups, part of the relationship was not captured by the typology. This method will have many uses, as most SA applications subsequently rely on regressions. It is available in the *clustassoc* function of the *WeightedCluster R* library (Studer 2013).

Our analysis has several limitations. First, we considered school-to-work transitions until age 30. However, it might be too early to fully capture the consequences of educational pathways. As shown by the FES analysis, some respondents might not have completed their transition to employment. The next releases of the first cohort of the TREE survey should allow this. Second, we only considered educational pathways, even though the TREE data contains information about employment/unemployment. Finally, we used an automatic procedure to extract features regarding the timing of transitions, using constant and predefined time ranges (every 12 months). However, it might be interesting to directly distinguish between "on-time" and "off-time" transitions. This would reduce the number of potential features – a strategy generally recommended by the data mining literature (see Bolano and Studer 2020) – and ease the interpretation of the results.

6 References

Achatz, Juliane, Kerstin Jahn, and Brigitte Schels. 2022. On the Non-Standard Routes: Vocational Training Measures in the School-to-Work Transitions of Lower-Qualified Youth in Germany. *Journal of Vocational Education & Training* 74(2): 289–310.

Aina, Carmen, and Giorgia Casalone. 2011. Does Time-to-Degree Matter? The Effect of Delayed Graduation on Employment and Wages. *AlmaLaurea Working Papers* 38.

Arum, Richard, and Yossi Shavit. 1995. Secondary Vocational Education and the Transition from School to Work. *Sociology of Education* 68(3): 187–204.

Becker, Gary Stanley. 1993. *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. Third edition. Chicago: Univof Chicago Press.

Benda, Luc, Ferry Koster, and Romke J. van der Veen. 2019. Levelling the Playing Field? Active Labour Market Policies, Educational Attainment and Unemployment. *International Journal of Sociology and Social Policy* 39(3/4): 276–95.

Bernardi, Laura, Johannes Huinink, and Richard A. Settersten. 2019. *The Life Course Cube: A Tool for Studying Lives. Advances in Life Course Research*, Theoretical and Methodological Frontiers in Life Course Research, 41: 100258.

Bertschy, Kathrin, Philipp Walker, Annick Baeriswyl, and Michael Marti. 2014. Lohndiskriminierung beim Berufseinstieg. Eine quantitative Analyse für die Schweiz. *Schweizerische Zeitschrift für Soziologie/Swiss journal of sociology/Revue suisse de sociologie* 40(2): 279–305.

Billari, Francesco C., Johannes Fürnkranz, and Alexia Prskawetz. 2006. Timing, Sequencing, and Quantum of Life Course Events: A Machine Learning Approach. *European Journal of Population/Revue Européenne de Démographie* 22(1): 37–65.

Bills, David B. 2003. Credentials, Signals, and Screens: Explaining the Relationship between Schooling and Job Assignment. *Review of Educational Research* 73(4): 441–69.

Bolano, Danilo, and Matthias Studer. 2020. The Link Between Previous Life Trajectories and a Later Life Outcome: A Feature Selection Approach, *LIVES working paper* 82. https://doi.org/10.12682/lives.2296-1658.2020.82.

Bourdieu, Pierre. 1979. Les trois états du capital culturel. *Actes de la Recherche en Sciences Sociales* 30(1): 3–6.

Brzinsky-Fay, Christian. 2007. Lost in Transition? Labour Market Entry Sequences of School Leavers in Europe. *European Sociological Review* 23(4): 409–22.

Buchmann, Marlis, Irene Kriesi, Maarten Koomen, Christian Imdorf, and Ariane Basler. 2016. Differentiation in Secondary Education and Inequality in Educational Opportunities: The Case of Switzerland. *Models of Secondary Education and Social Inequality*.

Elzinga, Cees H., and Matthias Studer. 2015. Spell Sequences, State Proximities, and Distance Metrics. *Sociological Methods & Research* 44(1): 3–47.

Falcon, Julie. 2016. Les limites du culte de la formation professionnelle : comment le système éducatif suisse reproduit les inégalités sociales. *Formation emploi. Revue française de sciences sociales*, no. 133(April) : 35–53.

Falter, Jean-Marc. 2012. Parental Background, Upper Secondary Transitions and Schooling Inequality in Switzerland. *Swiss Journal of Sociology* 38(2): 201–222.

Gabadinho, Alexis, Gilbert Ritschard, Nicolas S. Müller, and Matthias Studer. 2011. Analyzing and Visualizing State Sequences in R with TraMineR. *Journal of Statistical Software* 40(1): 1–37.

Gauthier, Jacques-Antoine, and Lavinia Gianettoni. 2013. Socialisation Séquentielle et Identité de Genre Liées à La Transition de La Formation Professionnelle à l'emploi. *Swiss Journal of Sociology* 39(1): 33–55.

Geier, Boris, Sandra Hupka-Brunner, and Nora Gaupp. 2013. Les Trajectoires d'insertion Des Jeunes Peu Qualifiés En Suisse et En Allemagne. *Cahiers de La Recherche Sur l'éducation et Les Savoirs* 4 : 149–166.

Gomensoro, Andrés. 2022. Construction of Standardised Variables on Income from Gainful Occupation for the TREE1 Cohort. *TREE Technical Paper Series*, May.

Gomensoro, Andres, and Claudio Bolzman. 2015. The Effect of the Socioeconomic Status of Ethnic Groups on Educational Inequalities in Switzerland: Which "Hidden" Mechanisms? *Italian Journal of Sociology of Education* 7(2): 70–98.

Gomensoro, Andrés; Bolzman, Claudio (2019). When children of immigrants come of age. A longitudinal perspective on labour market outcomes in Switzerland. *TREE Working Paper Series* No. 2. Bern: TREE. https://doi.org/10.7892/boris.131250.

Gomensoro, Andrés, and Thomas Meyer. 2021. TREE2 Results: The First Two Years, Bern: TREE.

Gomensoro, Andrés, Thomas Meyer, Sandra Hupka-Brunner, Ben Jann, Barbara Müller, Dominique Fabienne Krebs-Oesch, Melania Rudin, and Katja Scharenberg. 2017. Employment Situation at Age Thirty. Results Update of the Swiss Panel Survey, Bern: TREE.

Han, Yu, Aart C. Liefbroer, and Cees H. Elzinga. 2017. Comparing Methods of Classifying Life Courses: Sequence Analysis and Latent Class Analysis. *Longitudinal and Life Course Studies* 8(4): 319–41.

Hupka-Brunner, Sandra, Stefan Sacchi, and Barbara Stalder. 2010. Social Origin and Access to Upper Secondary Education in Switzerland: A Comparison of Company-Based Apprenticeship and Exclusively School-Based Programmes. *Swiss Journal of Sociology* 36(1): 11–31.

Complete citation: Imdorf, Christian, and Sandra Hupka-Brunner. 2015. Gender Differences at Labor Market Entry in Switzerland. In *Gender, Education and Employment,* by Hans-Peter Blossfeld, Jan Skopek, Moris Triventi, and Sandra Buchholz, 267–86. Cheltenham: Edward Elgar Publishing.

Korber, Maïlys, and Daniel Oesch. 2019. Vocational versus General Education: Employment and Earnings over the Life Course in Switzerland. *Advances in Life Course Research* 40: 1–13.

Kramarz, Francis, and Martina Viarengo. 2015. Chapitre 2. Les systèmes d'éducation et de formation face au chômage des jeunes, In: F. Kramarz & M. Viarengo (Dir), *Ni en emploi, ni en formation: Des jeunes laissés pour compte* (pp. 57–84). Paris: Presses de Sciences: 57–84.

Kuh, D., Y. Ben-Shlomo, J. Lynch, J. Hallqvist, and C. Power. 2003. Life Course Epidemiology. *Journal of Epidemiology and Community Health* 57(10): 778–83.

Kursa, Miron B., and Witold R. Rudnicki. 2010. Feature Selection with the Boruta Package. *Journal of Statistical Software* 36(September): 1–13.

Labussière, Marie, Mark Levels, and Maarten Vink. 2021. Citizenship and Education Trajectories among Children of Immigrants: A Transition-Oriented Sequence Analysis. *Advances in Life Course Research* 50: 100433.

Laganà, Francesco, Julien Chevillard, and Jacques-Antoine Gauthier. 2014. Socio-Economic Background and Early Post-Compulsory Education Pathways: A Comparison between Natives and Second-Generation Immigrants in Switzerland. *European Sociological Review* 30(1): 18–34.

Levine, Joel H. 2000. But What Have You Done for Us Lately? Commentary on Abbott and Tsay. *Sociological Methods & Research* 29(1): 34–40.

Liao, Tim Futing, and Anette Eva Fasang. 2021. Comparing Groups of Life-Course Sequences Using the Bayesian Information Criterion and the Likelihood-Ratio Test. *Sociological Methodology* 51(1): 44–85.

Meyer, Thomas. 2009. On ne prête qu'aux riches: L'inégalité des chances devant le système de formation en Suisse. In *Rapport social 2008. La Suisse mesurée et comparée,* edited by Christian Suter, Silvia Perrenoud, René Levy, Ursina Kuhn, Dominique Joye, and Pascale Gazareth, 60–81, Zürich: Seismo.

OFS, Office fédéral de la statistique. 2018. Transitions après un titre du degré secondaire II et intégration sur le marché du travail. Neuchâtel. https://www.bfs.admin.ch/bfs/fr/home/statistiques/catalogues-banques-donnees/publications.assetdetail.5006700.html, accessed 14.03.2022.

OFS, Office fédéral de la statistique. 2021. Le revenu des certifiés de la formation professionnelle initiale – Analyses longitudinales dans le domaine de la formation – Évolution dans les cinq ans après le titre https://www.bfs.admin.ch/asset/fr/17544438, accessed 31.03.2022.

Piccarreta, Raffaella, and Matthias Studer. 2018. Holistic Analysis of the Life Course: Methodological Challenges and New Perspectives. *Advances in Life Course Research* 41.

Pollien, Alexandre, and Lorenzo Bonoli. 2018. Parcours de Formation: Analyse Des Trajectoires de Formation Des Personnes Résident En Suisse. https://forscenter.ch/working-papers/2012-00002/. Accessed 17.07.2023

R Core, and Team. 2020. *A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing.

Sacchi, Stefan, and Thomas Meyer. 2016. Übergangslösungen beim Eintritt in die Schweizer Berufsbildung: Brückenschlag oder Sackgasse? *Swiss Journal of Sociology* 42(1).

Samuel, Robin, Manfred Max Bergman, and Sandra Hupka-Brunner. 2013. Longitudinal Effects of Social Background on Educational and Occupational Pathways within Early and Strong School Tracking. *Longitudinal and Life Course Studies* 5(1): 1–18.

Scharenberg, Katja, Karin Wohlgemuth, and Sandra Hupka-Brunner. 2017. Does the Structural Organisation of Lower-Secondary Education in Switzerland Influence Students' Opportunities of Transition to Upper-Secondary Education? A Multilevel Analysis. *Swiss Journal of Sociology* 43(1): 63–88.

Settersten, Richard A., and Karl Ulrich Mayer. 1997. The Measurement of Age, Age Structuring, and the Life Course. *Annual Review of Sociology* 23: 233–61.

Spence, Michael. 1973. Job Market Signaling. *The Quarterly Journal of Economics* 87(3): 355–74.

Stalder, Barbara, and Christof Nägele. 2011. Vocational Education and Training in Switzerland: Organisation, Development and Challenges for the Future. *Youth in Transition in Switzerland: Results from the TREE Panel Study*, January, 18–39.

Studer, Matthias. 2013. WeightedCluster Library Manual: A Practical Guide to Creating Typologies of Trajectories in the Social Sciences with R. *LIVES Working Papers* 24: 1–32.

Studer, Matthias. 2021. Validating Sequence Analysis Typologies Using Parametric Bootstrap. *Sociological Methodology* 51(2): 290–318.

Studer, Matthias, and Gilbert Ritschard. 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.

Studer, Matthias, Gilbert Ritschard, Alexis Gabadinho, and Nicolas S. Müller. 2011. Discrepancy Analysis of State Sequences. *Sociological Methods & Research* 40(3): 471–510.

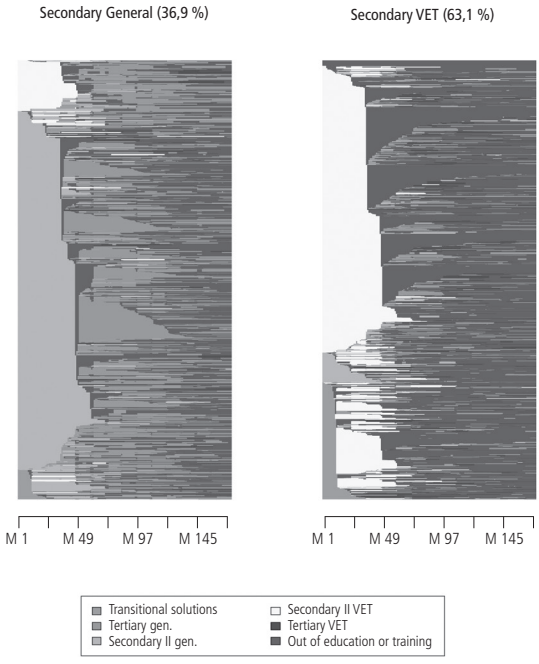
TREE. 2016. TRansition de l'École à l'Emploi, Documentation de La Première Cohorte de TREE (TREE1). 2000–2016. Bern: TREE

Yu, Xiao. 2021. Later Timing but Informed Decision? Delayed Postgraduate Attainment and U.S. College Graduates' Earnings. *Social Science Research* 98: 102583.

Zimmermann, Barbara, and Simon Seiler. 2019. The Relationship between Educational Pathways and Occupational Outcomes at the Intersection of Gender and Social Origin. *Social Inclusion* 7(3): 79–94.

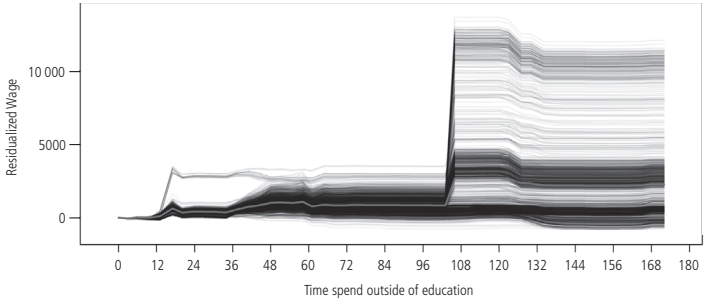
Appendix

Figure 5 Typology of Educational Pathways in Two Groups, Sequences Ordered From the Starting State



Note: Individual educational pathways, in months. Figure available in color at https://centre-lives.ch/sites/default/files/figunterlerchner2023.pdf#page=5.
Source: TREE wave 1, author's calculation.

Figure 6 Partial Dependence Plot of the Estimated Effect of Time Spent Not in Education nor Training on Residualized Income



Note: Monthly income of all working activities. Figure available in color at https://centre-lives.ch/sites/default/files/figunterlerchner2023.pdf#page=6.
Source: TREE wave 1, author's calculation.