# Measuring labour market transitions using a life-course perspective in selected developed and developing countries: an inventory of existing panel data and methods of analysis

Vaccaro, Giannina; Orsholits, Dan; Steinmetz, Stephanie; Studer, Matthias; Vandecasteele, Leen

International Labour Organization

European Commission

Background paper n°4

February 2022

# Measuring labour market transitions using a life-course perspective in selected developed and developing countries

## An inventory of existing panel data and methods of analysis


- Dr. Giannina Vaccaro (University of Lausanne)
- Dr. Dan Orsholits (University of Geneva)
- Prof. Stephanie Steinmetz (University of Lausanne)
- Prof. Matthias Studer (University of Geneva)
- Prof. Leen Vandecasteele (University of Lausanne)



Background Paper Series of the Joint EU-ILO Project
"Building Partnerships on the Future of Work"

# Abstract

The study of labour market transitions is an important aspect not only for researchers but also for policy makers. Such transitions, including retirement, extended working life, unemployment or school-to-work transitions etc., can be challenging and are often identified as turning points in the (re)production and accumulation of social inequalities. Therefore, they are also receiving an increasing attention for the design of new social policies.

However, to improve our understanding of how social inequalities evolve over the life course, labour transitions cannot be completely understood by only focusing on one particular point in time. This requires a more holistic approach through the implementation of a 'life course perspective'. For this however, it is essential that those transitions are situated in their wider trajectories, i.e. linked with previous and future situations, in order to understand their medium and long-term consequences. Focusing on a particular direct transition might be misleading. However, such a more holistic approach requires specific methods and data.

Against this background, this report aims to provide interested researchers and relevant stakeholders with an overview of available methods and data sets to analyse these labour market transitions in a life course perspective. More concretely, it provides an overview and evaluation (discussion of the strengths and weaknesses) of the methodological approaches to perform basic as well as advanced life-course analysis for two core phases of working lives: a) school-to-work and b) work-to-retirement transitions. Based on the description and evaluation of 32 datasets from 14 high-, middle- and low-income countries (Bolivia, Brazil, China, Egypt, Ethiopia, India, Indonesia, Japan, México, Peru, South Africa, South Korea, United States and Vietnam) it provides an overview of selected data sets as well as an evaluation of their suitability to study those selected transitions in a life course perspective with the discussed methodological approaches.

Keywords: labour market transitions, life course perspective, panel data, school to work, old age transition

JEL: C23, C32, C33, J11, J46, J62

# 1. Introduction[1]

The study of labour market transitions is getting increasing attention in high-income economies. These transitions, including retirement, extended working life, unemployment or school-to-work transitions to name a few, can be challenging and are often identified as turning points in the (re)production of social inequalities. For instance, the school-to-work transition is regularly identified as a key transition in professional trajectories. Experiencing a smooth transition between education and work might secure a future career, while atypical trajectories or those marked by unemployment or insecure job positions might have long-standing effects on future employment (e.g. Brzinsky-Fay, 2014). For instance, unemployment might leave scars suspected to affect both future wages and job quality (Buchs et al., 2017). For all these reasons, labour market transitions are also receiving an increasing attention for the design of new social policies.

The life course perspective offers a theoretical framework increasingly used in the social sciences. This perspective insists on several key aspects in the study of these transitions. First, transitions should be situated in their wider trajectories, i.e. linked with previous and future situations, in order to understand their medium and long-term consequences. Many labour transitions cannot be really understood by only focusing on a particular point in time, but should rather be studied in a more holistic perspective. A holistic perspective doesn't focus on just a specific transition but situates it in context by changing the scale of analysis to trajectories which include periods before and/or after a transition. For instance, school-to-work transitions can be complex, as they can involve preparatory courses or breaks. For these reasons, focusing on a particular direct transition might be misleading. One should rather take into account the whole trajectory to measure nonlinear trajectories, complex back-and-forth dynamics, and recurrent events (Shanahan, 2000). This more holistic perspective requires specific methods and data.

This report aims to provide interested researchers and relevant stakeholders with an overview of available methods and data to analyse these labour market transitions in a life course perspective. More concretely, it has two objectives. First, it provides an overview and evaluation (discussion of the strengths and weaknesses) of the methodological approaches to perform basic as well as advanced life-course analysis for two core phases of working lives: a) school-to-work and b) work-to-retirement transitions. Second, it provides an overview of selected data sets as well as an evaluation of their suitability to study those transitions in a life course perspective with the discussed methodological approaches.

Before moving on, some key concepts and classical terminology used in life course research should be introduced. Generally speaking, the life course approach focuses on how processes unfold over individuals' lives as they grow older and go through events that mark their personal trajectories. This is one of the first aspects of the life course approach: understanding how an individual's trajectory changes over time as they age. Also referred to as the age effect, this is the individual-specific aspect of the life course focusing on which stages an individual goes through and during which age bracket certain transitions typically happen in their lives. For instance, age effects are particularly important in school-to-work transitions, where some situations (such as unemployment or internships), are typically thought to be harmless if they occur at the beginning of the transitions, but harmful if they are experienced later on.

At the same time, these changes and trajectories over time must be contextualized in the larger historical and country context in which the individual lives. Important events affecting the whole population, such as the COVID-19 pandemic or an important economic recession, can also alter people's life courses and their trajectories. These are referred to as period effects in the life course literature. The main interest, when studying period effects, is to understand how these events affect individuals. Again, the COVID-19 will most probably profoundly affect all labour market transitions experienced by individuals.

Next to age and period effects, we also need to account for cohort effects. Cohorts, in the life course approach, are groups of individuals who were born in the same period, or who experienced another important event, like leaving school, at the same time. Cohorts are distinct groups of individuals whose life

[1] The authors would like to thank the ILO, in particular Guillaume Delautre, Drew Gardiner, Sher Verick, and Dorothea Schmidt-Klau for the very for the constructive and helpful suggestions and comments on the report. In addition, the authors are very grateful for the support and input we received from Diahhadi Setyonaluri with regard to the Indonesian data.

courses unfolded in similar historical contexts. A life course analysis would typically compare different cohorts who may have not experienced the same events or who experienced them at a different time. Recently, there has been a growing concern that young people starting their school-to-work transition during the COVID-19 pandemic might experience particularly difficult transitions having long-term consequences. This would be a cohort effect.

Life course processes unfold over time and therefore requires longitudinal data and methods. Not all methods are suited to investigate age, period, and cohort effects. As it will be presented, some methods are better suited to investigate age effects or period effects, while others can potentially be used to study all three. Similarly, only some methods offer a more holistic view on these processes. Thus, the research question of interest, in addition to data availability, plays an important role in determining which method should be used.

We review two broad groups of methods: those focusing on within-individual differences and methods focusing on between-individual differences. In the case of within-individual differences, the main goal is to understand how change in an individual's situation affects a later outcome for them. In the case of between-individual differences, the approach is more descriptive as the main goal is to compare individuals and how they differ in terms of change over time.

This report also aims to review available datasets to study labour market transition in a life course perspective for a selection of countries. This selection was carefully made to represent the different continents and to include low- and middle-income countries, where life-course studies are rather infrequent. As a result, the following countries were included: Bolivia, Brazil, China, Egypt, Ethiopia, India, Indonesia, Japan, México, Peru, South Africa, South Korea, the United States and Vietnam.

The report reviews datasets that are representative of the labour-active population in these countries. It focuses on a panel data, i.e. where the same individuals are followed over time. Indeed, this statistical instrument allows for a more causal analysis of transitions between states, as well as studying people's trajectories by examining sequences of transitions over time. If panel data is not available (or only to a limited extent), retrospective data is also considered.

# 2. Analytical Approaches to study life courses

The main analytical approaches presented in the following are (1) within-individual and (2) between individual models. In within-individual analysis, the main focus is on estimating how change occurs for individuals and the effects of events such as job loss, birth of a child etc. on later labour market outcomes. On the other hand, between-individual models mainly focus on describing differences in trajectories between individuals.

The analytical approaches presented in this section are designed to be used with individual longitudinal data, where the same individuals are followed over time. However, when such data are not unavailable, one can use pseudo-panels, which allow an over-time follow-up at the group level.

Further details about pseudo-panel data sets can be found in Box 1. In the remainder of the text, we focus on analytical approaches for individual longitudinal data. However, once a pseudo-panel is constructed, longitudinal analysis techniques can be applied to them.

**▶ Box 1 – Pseudo-panels**

Pseudo-panels are data sets that are derived from repeated cross-sectional surveys and are designed to approximate the structure of a panel data set. The aim is to create pseudo-panels, i.e. groups based on information collected from repeated representative cross-sectional studies such as birth year, gender, socio-economic position. This allows the use of methods designed for individual panel data, such as those presented in this section, even when only cross-sectional data is available.

In a pseudo-panel, the same individuals are not observed over time, but groups are created based on constant socio-demographic characteristics. These same groups are constructed for every repeated cross-sectional dataset, and hence change over time can be examined at the group level. For instance, a pseudo-panel could group people together on the basis of their year of birth, education level and gender and then examine change in unemployment rate across the years of a recession.

Pseudo-panels are created by grouping together individuals who have the same time-constant characteristics. Among the most common characteristics used to make these groups are sex, birth cohort or highest educational level achieved. The variables used to construct these groups must be measured for a representative sample of the population at every one of the repeated cross-sectional surveys. Once these groups are established, a longitudinal data set is constructed by taking the averages or percentages for the variables of interest. The models, such as fixed-effects models or dynamic panel models, are then estimated using these pseudo-panels.

Oftentimes, pseudo-panel data are a good solution to examine the effect of labour market transitions on indicators that are scarcely available in individual longitudinal data, such as measures of social attitudes. An example are the studies by Reeskens and Vandecasteele (2017, 2021), whereby they examine the effect of unemployment and economic insecurity on well-being and social attitudes with a pseudo-panel and a fixed-effects approach. They also present a cohort analysis of these over-time changes, which allows them to study these mechanisms in a life-course framework (Reeskens & Vandecasteele, 2021).

Obviously, pseudo-panels are not a perfect replacement for individual longitudinal panel data. First, there are some specific issues related to how to treat unobserved heterogeneity when using pseudo-panels in conjunction with within-individual difference estimators such as fixed-effects models. Second, they can lead to finding effects, which may actually be due to compositional changes in the sampled individuals for each cross-sectional measurement rather than an actual change over time.

## 2.1. Within-Individual Differences

The main focus of within-individual difference methods is to analyse what changes in an outcome are due to a change in a (time-varying) explanatory variable at the individual level. With these types of models, longitudinal data is typically used to control for unobserved (time-constant) variables by "controlling" for a person's unmeasured characteristics. Models focusing on within-individual differences are especially useful in trying to understand what time-varying elements can lead to a change in a specific individual's life course. The most commonly used method for investigating within-individual differences are the so-called *fixed-effects estimators*. They can be used for continuous outcomes (such as salaries) or non-continuous outcomes (for instance employment status).

Another group of models focusing on within-individual differences are the *family of dynamic panel models*. They go further than models using fixed-effects estimators by allowing the past values of an outcome to be included as explanatory variables. In the next section, we will present some suggestions for models to investigate within-individual differences as well as the advantages and disadvantages of these methods.

### 2.1.1. Fixed-effects Models

**Which questions can be answered?** — The goal of fixed-effects models is essentially to understand how a change in a certain explanatory variable affects a specific outcome at the individual level by removing any unobserved time-constant differences that exist between different individuals. In other words, these models let us understand how a change in a person's life can affect a specific outcome considering any person-specific characteristics that are stable over time. This makes it most suitable for understanding age- and/or period-related changes in an individual life course but not cohort differences.

**Estimation —** Fixed-effects models in the case of continuous outcomes are not difficult to estimate or interpret. They are among the most commonly used statistical methods. Various estimators for these models exist with the most common being the "within-transformation" where the overall mean for each variable is calculated and then subtracted from each observation.

**Data requirements —** Fixed-effects models are flexible as they can be estimated with few waves of longitudinal data (in fact even two waves are sufficient). Therefore, short panels such as rotating panels from labour force surveys or panel components are enough for these types of models. They also do not require evenly spaced data or complete data. This means that there these models can be used even if respondents don't respond to every survey wave or if they join after the start of the survey, as well as if the data is collected at irregular intervals. This type of data is also referred to as unbalanced panel data.

**Inclusion of covariates —** There is one major disadvantage associated with traditional fixed-effects models using the within transformation: time-constant explanatory variables cannot be included. If we take the case

of the transition from education to employment, this means that fixed-effects models estimated with the standard within transformation would not allow us to estimate, for instance, how salaries vary between a people who completed an upper secondary level of education and others who completed a tertiary-level degree program. Another disadvantage of the standard fixed-effects approach is that it when considering processes over time, the assumption is made that the process unfolds the same way for all individuals. This assumption can be relaxed using the within transformation in combination with random-effects/multilevel/mixed-effects models (sometimes termed the "hybrid model" or the Mundlak method) (Allison, 2009). This combines all the advantages of fixed-effects models, namely controlling for time-constant unobserved characteristics, with the additional flexibility of the estimators used for mixed-effects models allowing the inclusion of time-constant variables as well as random coefficients, or multiple levels of observation.

**Categorical outcomes —** Estimating fixed-effects models for non-continuous outcomes is slightly more complicated. Firstly, fixed-effects models with binary or multinomial outcomes can *only* use the logit link. Secondly, the conditional logit model, the most common estimator, produces unbiased parameter estimates by removing any individuals who experience no change in an outcome. This means that in the case where there are small sample sizes, this could potentially lead to a large reduction in the data available. Thirdly, the hybrid-model specification using a mixed-effects estimator does not provide the same estimates for time-varying variables as standard fixed-effects estimators in the case of non-continuous outcomes. Consequently, a compromise needs to be made in the case of non-continuous outcomes when time-constant variables are to be included in the analysis. An additional issue is that fixed-effects estimators for multinomial outcomes are not as easily or widely implemented in software. This means that there might be a loss of information especially in the case where the research question aims to distinguish between many meaningfully different states.

**Conclusion —** To summarize, fixed-effects models are designed to be used with variables that vary over time and are thus not useful when the goal is to compare between individuals. For example, in the case of the school-to-work transition, fixed-effects models are not necessarily the best suited — hybrid models notwithstanding — to comparing individuals' salaries growth depending on their level of education because education level does not change for most working people. However, if the interest is in understanding if changing jobs or moving to a city leads to an increase in wages for a *specific* individual, then fixed-effects models are a good choice. These models let us understand how a change in a person's employment career can affect a specific outcome whereby any person-specific characteristics that are stable over time are controlled. Thus, this is an approach that centres on the individual and consequences of changes *for a specific* individual rather than comparing different groups or subgroups of people. As the minimum number of time points is two for fixed effects models they can work well with very short panels.

### 2.1.2. Dynamic Panel Models

**Which questions can be answered?** — Dynamic panel models can be thought of as an extension of the fixed effects model in that it is still an estimation technique the focuses on within-individual differences, but it goes beyond standard fixed-effects approaches by also including previous, or lagged, values of the main dependent variable as an additional independent variable. This introduces a more dynamic approach to within-individual differences by incorporating past information thus taking into account previous individual trajectories. In other words, these models mainly aim to study contemporaneous relationships once past information is also considered rather than the contribution of each past measurement. Nevertheless, dynamic panel models can also be used, especially in the case of discrete outcomes, to test the degree to which past states continue to influence the present. In the case of the school-to-work transition for instance, this could allow the assessment of the degree to which an individual's immediate labour market status after leaving school continues to influence their future employment prospects.

**Data requirements —** One advantage dynamic panel models share with fixed-effects models is that they do not necessarily need many observations as estimators have been developed to overcome bias related to short periods of observation. They do however require at least three measurements rather than two as at least one measurement is lost through the inclusion of the lagged dependent variable as an explanatory variable. One disadvantage of using these models for data covering long periods is that they can become difficult to interpret as there can be many lagged values.

**Inclusion of covariates —** A major disadvantage in the case of continuous outcomes is that the standard estimators used by dynamic panel models cannot include between-individual characteristics (i.e., those that do not vary over time) as explanatory variables. However, recent developments showing that dynamic panel models can be estimated in the structural equation modelling framework can solve this limitation (Moral-Benito et al., 2018). As for discrete or categorical outcomes, these models are typically restricted to binary variables meaning they are not suited to comparing multiple outcomes simultaneously.

**Conclusion —** Dynamic panel models build on fixed-effects models by further incorporating past information on the dependent variable as additional covariates thus taking into account the dynamic nature of processes. This provides a more encompassing view of processes as the present is not thought to be independent of the past which is especially relevant for labour market statuses and processes. Dynamic panel models, however, are not always easy to estimate and there can be limits when they are used with categorical outcomes and very long periods of observation.

### 2.1.3. Other "Within" Models

**Which questions can be answered? —** There are two other commonly used and important methods focusing on within-individual differences: difference-in-difference and regression discontinuity. Both of these methods are principally used to investigate the impact of a policy change on an outcome by comparing a treatment group to a control group usually between two points in time. The main difference is that the treatment or policy change in a difference-in-difference design is often considered to be at a level higher than the individual level, while in regression discontinuity designs, treatment is determined by a cut-off for eligibility.

**Data requirements —** At the individual level the difference-in-difference can be estimated with at least two measurements of the same person i.e., two waves of panel data or with cross-sectional data measured before and after the intervention of interest. At for instance a country level, cross-sectional data can also be used in combination with a difference-in-difference design if the goal is to compare the effect of a labour market policy, for instance, between two different countries in two different periods marked by the introduction of a policy change. Regression discontinuity designs are designed around cross-sectional data and therefore do not require panel data.

### 2.1.4. Examples of research

Noelke & Horn (2014), use a difference-in-difference design to assess the impact of a policy reform in Hungary concerning vocational education. During the 1990s, vocational training in Hungary shifted from a model where training was principally done by employers to one where training occurred mainly in schools. They find that this shift made labour market entry more difficult as the unemployment rate for males. Kracke et al. (2018) use fixed-effects models to study the consequences of overqualification on wages at labour market entry in Germany. They find that even when individual heterogeneity is taken into account, there is still a wage penalty associated with not being able to find employment which matches individuals' skills and training be it for people with university degrees or vocational training.

For the work-to-retirement transition, Chan & Stevens (2004) use panel data in combination with fixed effects to study what influences individuals' decision to retire in the United States based on their financial resources. The authors find that pension plans that continue to accrue benefits the longer an individual works discourage retirement and lead to delay in entering retirement. They also find that wealth contributes to explaining retirement decisions as individuals with greater wealth are more likely to retire earlier. Another study using fixed-effects models by Lux & Scherger (2018) investigates changes in self-rated health for individual who returned to employment after the statutory retirement age in the United Kingdom and Germany. Interestingly they find that taking up work after retirement does not contribute to worse health even if the job is not necessarily a high quality one.

### 2.1.5. Suggested readings

For a deeper look into models studying within-individual change, we suggest:

Allison, P. D. (2009). *Fixed Effects Regression Models*. Sage Publications.

Andreß, HJ., Golsch, K. & Schmidt, A.W. (2013). *Applied Panel Data Analysis for Economic and Social Surveys*. Springer.

Angrist, J. D. & Pischke, J.S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data* (2nd ed.). MIT Press.

## 2.2. Between-Individual Differences

Contrary to approaches focused on investigating within-individual differences which primarily aim to establish causal relationships, methods that are used to analyse between-individual differences are of a more descriptive nature (Brüderl et al., 2019). Nevertheless, these methods are more useful when investigating longer-term trajectories, especially in the case of multiple simultaneous outcomes or multidimensional measures. It is especially the second aspect, the study of complete trajectories (Piccarreta & Studer, 2019), that make these methods suitable for a holistic analysis of life courses; which is difficult to achieve using methods focused on within-individual change. In addition, these methods allow incorporating time-constant variables, which makes them more suitable to studying period or cohort effects in relation to individual life courses.

Each method presented here will be accompanied by examples of the types of research questions they can be used to answer. We will also propose situations where combining methods can be useful in providing insights concerning labour market transitions. To begin, we present two holistic methods that can be used to analyse long life phases or even entire life courses: growth curve models and sequence analysis. We then present other methods, survival analysis and latent transition analysis, which principally focus on transitions to different statuses.

### 2.2.1. Growth Curve Models

**Which questions can be answered? —** Growth curve models were historically developed to chart development, or *growth* over time while accounting for inter-individual variability in these trajectories. They are useful for studying the evolution of continuous outcomes over time. From a more substantive point of view, growth curve models are especially useful when studying topics such as the evolution of labour income over time or also subjective measures of job quality measured using continuous scales. In the simplest models, without any explanatory variables, we would then model the change in income or job quality as a function of time.

In the larger life course perspective, growth curve models were initially developed to study change over time, which makes them particularly well suited for studying age effects. They can also be used to study period effects and to test group or cohort effects through the inclusion of covariates or through testing equality constraints in the structural equation modelling framework.

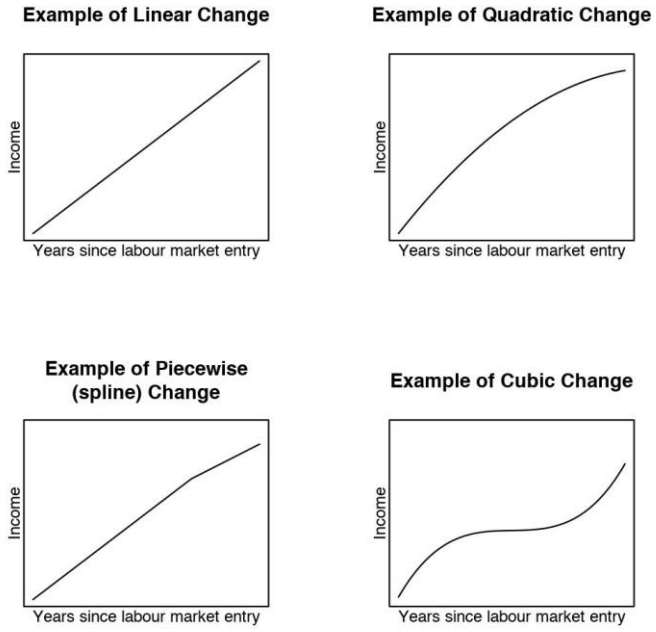
**Estimation —** Growth curve models were initially used in the psychological sciences. They were estimated within the structural equation modelling (SEM) framework, where they were shown to be a special case of confirmatory factor analysis (Grimm et al., 2017). However, the development of hierarchical linear modelling and subsequently multilevel random- and mixed-effects estimators has allowed these models to be estimated in another statistical framework, which offers several advantages over the structural equation modelling framework. They are better at incorporating multiple levels of data (for instance measurements nested in individuals who are nested in households) and are more widely implemented in statistical analysis software.

Besides the choice of statistical framework, the researcher also has to specify how an outcome is expected to change as a function of time. The simplest function is a linear one, which supposes that the outcome can only increase or decrease over time. It does not allow for complex forms of change, which would allow for an increase followed by a decrease, nor for instance, reaching a plateau, which would be more common when studying income. More complex forms of change over time can nevertheless be specified. A quadratic relationship, including time and time squared, for instance would allow for modelling more complex relationships and things such as plateaus. Figure 1 shows some examples of what the types of trajectories could be chosen and estimated with growth curve models in the case of studying income trajectories after

labour market entry. Moreover, presenting the mean trajectories is often a useful tool for investigating the results especially when more complex forms of change over time are used.

Even more complex relationships can be specified such as cubic, or quartic, but the more complex the function of time, the harder it becomes to interpret the results especially once change over time is related to additional explanatory variables. Another option is the use of a spline function, which allows specifying a change as a function of two (or more) linear functions and a point from which the switch occurs. An additional possibility in the SEM framework is the "latent basis" model where a linear change is estimated but for each time point there is a multiplicative factor, which introduces non-linearity. In other words, an overall mean value for change over time is estimated using a time-specific multiplicative factor deviating from this overall mean. This allows for modelling non-linear change over time without actually estimating a non-linear model of change.

**▶ Figure 1: Examples of trajectories of change that can be estimated with growth curve models**




**Data requirements —** Growth curve models do not require data that has been measured at identical intervals over time. However, the more complex the function of time, the more longitudinal time points are required. In the case of a linear growth curve model, the rule of thumb is that, on average, each respondent should provide at least three waves of responses. For a quadratic function four waves are needed and with a cubic or spline function five waves of data is recommended (Bollen & Curran, 2006). A latent basis model, by virtue of it being a special case of a linear change function, shares the same requirements as regular linear model. These recommendations apply to both the SEM and multilevel frameworks.

**Inclusion of covariates —** Growth curve models can be further extended by incorporating variables that can explain differences in trajectories between individuals. The simplest cases are time-invariant or (mainly) constant individual-specific characteristics such as sex or birth cohort, which can be included as explanatory variables directly affecting the parameters describing the trajectory of change over time. time-varying explanatory variables, like employment status or salary only directly affect the measured outcome a specific point in time. Consequently, they do not explain complete individual trajectories but rather differences between individuals in the observed outcome at a specific time point.

**Extensions —** Growth curve models offer the possibility for numerous extensions but some of them are dependent on the framework chosen for estimation. They can be extended to model changes in discrete (binary or ordinal) outcomes over time in both the SEM and multilevel framework. This however makes the assumption that we are measuring an underlying continuous construct – for example in the case of unemployment a continuous propensity or probability of an individual being unemployed over time. The

method is however not suitable to study multiple movements between different statuses, for example between education, unemployment and employment, simultaneously as this would normally require the use of multinomial outcomes.

Another extension available in the SEM framework is modelling change over time in multiple indicators simultaneously allowing for multi-dimensional approaches and analyses. This can be achieved in two ways: creating a composite latent variable using factor analysis or simultaneously estimating growth processes for multiple outcomes (Mund & Nestler, 2019). The first option is well-suited to using multiple indicators to measure a concept such as employment security with different indicators corresponding to different dimensions (contract type, subjective evaluations, etc.). The second option is interesting if the goal is to link changes in one outcome over time to changes in another outcome over time. For example, one could relate changes in an individual's labour income over time to a measure of life satisfaction to see how they evolve together.

A final extension, also within the SEM framework, is the growth mixture model. This model simultaneously estimates a growth curve while clustering or grouping individuals according to their individual deviation from the mean estimate trajectories. Consequently, it would be possible to have groups of individuals whose incomes are characterized by slower growth, quicker growth, or stagnation for instance. There are nevertheless criticisms of this method notably in relation to its potential for over-interpretation of results that could be due to the violation of the statistical assumptions these models make in relation to dependent variable. Thus, results from growth mixture models should be interpreted with caution as the groups are not necessarily representative of true differences in growth trajectories.

**Examples of research —** Growth curve models have been used to study topics such as career progression and the earnings trajectories after labour market entry. Manzoni et al. (2014) used growth curve models to analyse trajectories of career progression, measured using occupational prestige, for eight birth cohorts in Germany in the first 15 years of. Using retrospective data, they find that career progression takes places in the earliest phases of working life with career progression mainly happening in the first 10 years of labour market participation.

Another example of research using growth curve models is the work by Gabay-Egozi & Yaish (2021) which compares earning trajectories between individuals who followed a vocational education track and those who followed an academic one after the school-to-work transition in Israel. They find that it is predominantly the highest level of education that matters the most rather than the academic track even if, for an equal level of education, following a vocational track can provide a short-term income advantage in early stages of one's working life.

**Conclusion —** Growth curve models offer the possibility of modelling changes in outcomes over time under the assumption that they are continuous variables. The function of change over time can be specified flexibly with potentially complex forms being chosen. There is however a trade-off with more complex specifications of change over time requiring more data and which is harder to interpret. Despite this limitation, growth curve models offer many possibilities for analysing employment outcomes over the entire life course but are most useful when the outcomes being investigated can be operationalized using continuous scales.

**Suggested reading —** For a deeper look into growth curve models and their estimation in the SEM or multilevel frameworks we suggest:

Bollen, K. A., & Curran, P. J. (2006). *Latent Curve Models: A Structural Equation Perspective*. Wiley.

Grimm, K. J., Ram, N., & Estabrook, R. (2017). *Growth Modeling: Structural Equation and Multilevel Modeling Approaches*. The Guilford Press.

### 2.2.2. Sequence Analysis and Longitudinal Latent Class Analysis

**Which questions can be answered? —** Sequence analysis in the social sciences is often used to describe individuals' movements through stages of the life course. It encompasses a broad set of methods aiming to analyse trajectories from a holistic perspective, ranging from visualization to explanatory methods. It provides a comprehensive overview of the observed trajectories taken as a whole. This overview can then be used for various purposes, such as describing a set of entire trajectories, identifying relevant or atypical regularities, or contrasting these trajectories according to other key aspects such as gender or cohort. It is a holistic method in that it can capture long or medium life course phases in a single analysis. In this

framework, the trajectories are coded as a sequence of successive statuses occupied by individuals over time. This allows for studying age effects and, through the use of covariates, cohort effects. However, because the main object of analysis is the complete sequence, and not its components, it is not a method that is suited to studying period effects.

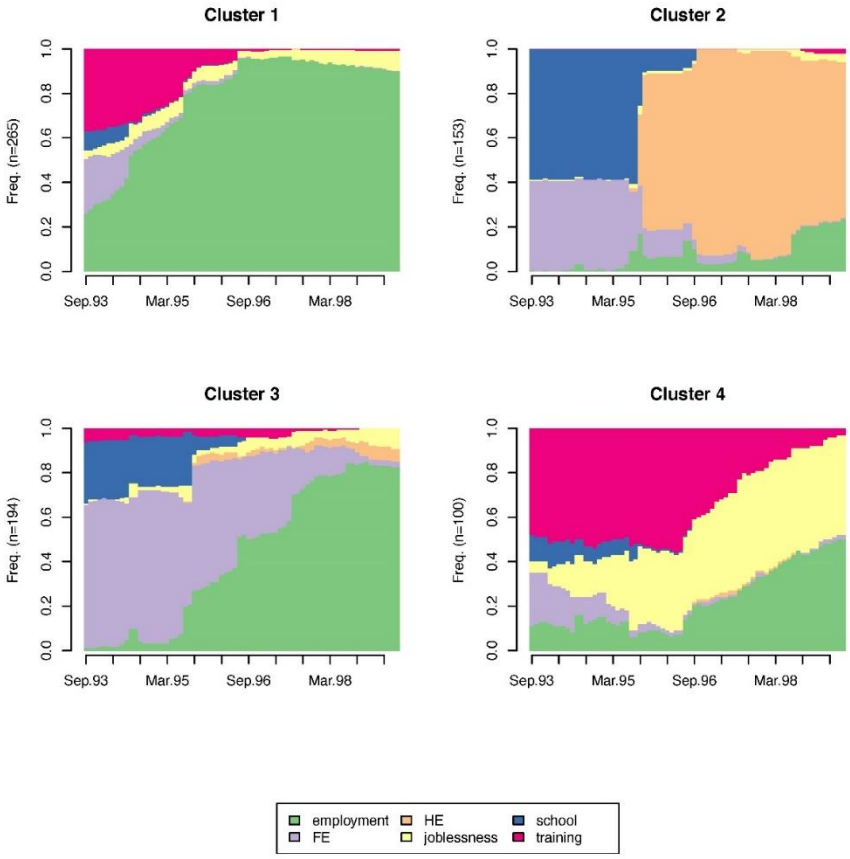
Sequence analysis can be used to focus on a single domain of the life course such as education and employment, but it can also be used to investigate multiple domains at once such as family trajectories in conjunction with employment trajectories. It is important to note that sequence analysis assumes that the statuses we are considering are measured using discrete categorical variables meaning that is can be thought of as the categorical counterpart to growth curve models.

**Estimation —** The standard use of SA revolves around a "core program" involving four typical steps. First, the trajectories are coded as states' sequences. This implies specifying the situation occupied by an observation at each time point to describe its trajectory over time. In the second step, the trajectories are compared to one another using a dissimilarity measure.

Third, the distances between each pair of sequences are used to create a typology of the trajectories with cluster analysis. This typology describes the various kinds of patterns observed in the data without making any assumptions on the data generation mechanisms. This exploratory approach can capture complex and potentially unexpected dynamics in trajectories, which is particularly suited to understanding the many interdependencies of the life course. Depending on the research question, such a typology might highlight regularities in the timing of situations, ordering of states, or time spent in each state.

Visual inspection of the sequences and the cluster assignments if often used to designate the typologies. For example, Figure 2 shows four clusters describing school-to-work transitions in Northern Ireland (McVicar & Anyadike-Danes, 2002). For each cluster, a sequence density plot shows the proportion of individuals in a given state (employment, higher education, school, further education, joblessness, or training) for each point in time. For the first cluster, we see that it mainly comprises individuals who immediately entered the labour market. Clusters two and three are characterized by individuals who continued their education before entering employment while the final cluster is composed of people who had difficulty entering the labour market.

**▶ Figure 2: Examples of school-to-work trajectories grouped using a combination of sequence analysis and clustering**




Source: Gabadinho et al. (2011)

Most of the time, the typology is then used in subsequent analyses. Technically, it can be included in any statistical method handling categorical data. Two main uses can be distinguished. First, the typology might be used as a dependent variable in a multinomial (or similar) regression model. The aim is then to understand how the type of trajectory is associated with variables of key interest. For instance, McVicar & Anyadike-Danes (2002) used this strategy to identify the profiles of young individuals "at-risk" of following a school-to-work trajectory marked by joblessness. Second, the typology might be used as an explanatory variable in a subsequent regression. In this case, the aim is to understand how a previous trajectory is linked with a later-life outcome. For instance, Brzinsky-Fay & Solga (2016) looked at how school-to-work transition patterns are linked with occupational attainment at 30 years old, aiming to understand the consequence of nonlinear pathways on later-life careers.

This general procedure involves choosing a dissimilarity measure. While most studies use optimal matching, many other distance measures were developed and are briefly presented below. A more comprehensive and detailed review of these developments and their usefulness for life-course research is available in Studer & Ristchard (2016).

Studer and Ritschard (2016) consider three key aspects of interest when studying trajectories in a life-course perspective. Timing and duration focus on the time aspect of states in sequences with timing being related to when a *state or a transition occurs* while duration measures the amount of time spent in the same state. The timing aspect is a key element in school-to-work trajectories for instance, as experiencing unemployment at the beginning or the end of the process is typically thought to have very different consequences on later life outcomes, such as income. The duration aspect is also important, as the total amount of time spent in unemployment is also thought to have long-lasting consequences through scaring

effects for instance. Sequencing is the third aspects of interest in trajectories. It refers to the order in which the different states appear in a sequence, and it is a key aspect to understand its *dynamics*. Typically, experiencing the "employed-unemployed" or the "unemployed-employed" ordering reveals different dynamics in school-to-work transitions. Studer & Ritschard (2016) compare the distance measures based on their sensitivity to these three aspects.

Optimal matching (OM) is the most commonly used approach for calculating distances in SA. OM is derived from work in the fields of information theory and computer science used to measure the differences between two pieces of information, for instance lines of text, by calculating the "cost" of transforming one into another. Three possible operations are considered: insertions, deletions, and substitutions. One of the major criticisms of the OM is that the costs for each operation need to be chosen in advance. This choice can then influence how the distance between sequences is calculated and which aspects of sequences affect the calculation of dissimilarities. Several approaches are available, but the most common one is to use constant costs, which assumes that all states describing the sequences are equally different from one another.

According to Studer & Ritschard (2016), OM is mostly sensitive to duration and sequencing aspects of trajectories. However, it can be made less sensitive to timing and more to sequencing by increasing the costs of insertions and deletions relative to substitutions while decreasing those costs relative to substitutions makes the distance measure more sensitive to timing.

While OM is the most popular and common approach to measuring sequence dissimilarity, there are approaches to measuring sequence dissimilarity which are not derived from the cost of transforming one sequence into another, but on shared characteristics. Measures of dissimilarity derived from differences in state distribution (such as Euclidean distances or Chi-Square) compare sequences based on the amount of time spent in each possible state in each sequence. Typically, these measures are insensitive to the timing of states and their sequencing, but they can be augmented to take into account non-matching positions and thus include timing in the measure of dissimilarity.

The simple Hamming distance compares sequences at each position and the number of mismatches is a metric for how similar or dissimilar sequences are. It is a good choice when the main interest is comparing the timing of the trajectories. However, the main disadvantage is that all the sequences being compared must be of the same length.

Finally, another group of distance measures are based on shared characteristics, such as shared subsequence's or time spent in a state. The SVR-spell measure uses the number of matching sub-sequences in total in a pair of sequences weighted by the length of each subsequence. This distance measure is sensitive to sequencing and duration.

As a rule of thumb, if the main interest is explaining differences in sequencing (the order of states), the best dissimilarity measures are OM-based variants accounting for transitions or spell durations, or SVR-spell, which is a subsequence-based measure. If the focus is on timing, the Hamming distance is the best choice followed by timing-sensitive state distribution measures. Finally, if the interest is in the duration or time spent in spells, state distribution measures are the best suited dissimilarity measures followed by various OM-based measures (classic OM, OM of spells).

An alternative to sequence analysis based on dissimilarity measures is longitudinal latent class analysis (Barban & Billari, 2012). This approach is conceptually simpler in that no choices need to be made relative to dissimilarity measures or clustering algorithms, but it makes certain statistical assumptions that are dependent on the model itself. Among others, the conditional state independence assumption states that, given group memberships, states measured at different times are independent from one another. This is particularly unlikely for longitudinal data. Furthermore, it is not possible to choose among different dissimilarity measures and therefore on which specific characteristics of the sequences should be analysed. It works with the same type of data – categorical variables – and often produces similar results to sequence analysis conducted with a combination of OM and clustering. It is computationally close to the use of the Hamming distance. However, it does have some advantages when it comes to missing data as all available information can be included in the model without having to undertake any missing data imputation or consider missing data as a specific state.

After choosing the method of clustering the sequences, there is one final choice researchers need to make: the number of clusters or groups to keep. In the case of dissimilarity-based approaches combined with clustering, there are numerous measures of cluster quality that can be used to aid in determining the

number of clusters to retain. In the case of longitudinal latent class analysis, there are model-based measures of relative goodness-of-fit (AIB, BIC, likelihood ratio tests, etc.) to test which number of latent classes or groups' best fits the observed data. These model-based measures rely on a conditional state independence assumption, which might not be relevant for longitudinal data.

**Data requirements —** In sequence analysis a minimum of four to five time points are required. This makes it hard to use these models with shorter-duration panels such as those employed in rotating panels as part of labour force surveys. In addition, some dissimilarity measures require that all the sequences that are being compared be of the same length which for panel studies with drop-out may not always be realistic. Furthermore, sequence analysis requires (at least approximately) evenly spaced data. Therefore, additional retrospective information to fill in the gaps is necessary when the data are collected at irregular time intervals.

**Inclusion of covariates —** It is possible to incorporate time-constant explanatory variables when using sequence analysis or latent class analysis. With sequence analysis using dissimilarity measures and clustering, the analysis needs to be done in two steps: assigning individuals to clusters and then re-using the cluster assignments as a dependent variable in a regression model such as multinomial logistic regression. In the case of longitudinal latent class analysis, the covariates can be directly included in the estimated model if the aim is to study the association between certain individual characteristics and the likelihood of being in a group or cluster. However, the interpretation becomes cumbersome. It is also possible to use the classification as a time-constant explanatory variable. For instance, after classifying educational trajectories, they can then be used as an explanatory variable in the analysis of changes in income over time.

**Extensions —** Both sequence analysis and longitudinal latent class analysis can be extended to simultaneously analyse sequences in multiple domains jointly through the use of multichannel sequence analysis while latent class analysis can simply incorporate additional items of measurement directly into the model. This can be used to investigate the parallel evolution of working life and family life for instance in midlife, or to look at parallel education and employment sequences in the early phases of labour market entry. The groups established in the case of multichannel analyses become composite typologies in the vein of "working and married" or "in education with part-time employment."

Recently, extensions such as SAMM or CTA have been proposed to handle time-varying covariates. This allows for instance the inclusion (at least to some extent) of period effects in the analysis.

**Examples of research —** In the case of school-to-work transitions, the work of Brzinsky-Fay & Solga (2016) studies how occupational attainment at age 30 relates to different school-to-work trajectories in West Germany. They find that more non-linear school-to-work transitions are not necessarily associated with worse occupational attainment except for younger women.

For the work-to-retirement transition, work by Riekhoff (2016) analysed work-to-retirement trajectories in the Netherlands for individuals aged between 56 and 66 with sequence analysis using monthly employment data. The results of this study showed that the self-employed or individuals working in smaller companies were more likely to work for longer and retire later. Women were more likely to experience a significant period of inactivity (i.e., to have no income) prior to reaching the statutory retirement age than men.

**Conclusion —** In summary, sequence analysis and longitudinal latent class analysis can be thought of as the categorical counterpart to growth curve models in that they allow the analysis of trajectories in a holistic manner when trajectories are measured using a series of categorical variables as opposed to continuous ones. Sequence analysis offers more potential for the researcher to choose which aspects of sequences should be highlighted in the subsequent typologies being created while latent class analysis doesn't offer as much flexibility. However, latent class analysis might better handle missing data. As for data requirements, longer panels are generally required, and the information should be available at a regular time interval.

**Suggested readings —** For a deeper look into sequence analysis and related methods we suggest the following:

General introductions and critical review papers:

Abbott, A., & Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology, Review and prospect. *Sociological Methods and Research, 29*(1), 3–33. https://doi.org/10.1177/0049124100029001001

Aisenbrey, S., & 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 and Research, 38*(3), 430–462. https://doi.org/10.1177/0049124109357532

Piccarreta, R., & Studer, M. (2019). Holistic analysis of the life course: Methodological challenges and new perspectives. *Advances in Life Course Research*, *41*. https://doi.org/10.1016/j.alcr.2018.10.004

Technical presentations and manuals:

Gabadinho, A., Ritschard, G., Müller, N. S., & Studer, M. (2011). Analyzing and Visualizing State Sequences in R with TraMineR. *Journal of Statistical Software*, 40(4), 1–37. http://www.jstatsoft.org/v40/i04

Studer, M., & 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. (2013). WeightedCluster Library Manual: A practical guide to creating typologies of trajectories in the social sciences with R (LIVES Working Papers No.24; Number 24). NCCR LIVES. https://doi.org/10.12682/lives.2296-1658.2013.24

### 2.2.3. Transition- and Event-Focused Models

In this section we present two transitions and event-focused models: survival analysis and latent transition analysis. The former is mainly of use when the aim is to investigate whether an event or transition will occur at some point in the future while the latter is more suited to investigating a series of transitions in detail.

Survival Analysis

**Which questions can be answered? —** Survival analysis, or event history analysis, is a collection of methods aimed at understanding how, after a certain amount of time, certain individuals experience an event while others don't. Therefore, the focus of these models is the study of the *occurrence* of an event. For example, in the case of employment transitions, survival analysis can be used to investigate what makes certain groups of individuals more likely to enter employment after completing education or in the case of retirement transitions, to understand what makes certain groups more likely to retire earlier or later. Because survival analysis simultaneously models time and the occurrence of an event, it is well suited for studying age effects or period effects, and, through the inclusion of covariates, cohort effects as well.

**Estimation —** While there are many methods used in survival analysis, the most common approaches are discrete-time survival analysis and, to a lesser extent, the Cox proportional hazards model. The proportional hazards estimator has an advantage over many other methods in that it doesn't require the specification of a survival function i.e. the researcher doesn't have to specify how the risk of an event occurring is expected to change over time. The proportional hazards approach is especially suited for comparing differences in the likelihood of an event occurring between different groups. In the case of discrete-time survival analysis, the researcher has to specify a time function with the most flexible option being a stepwise function using dummy variables for each time period.

**Data requirements —** Survival analysis does not require much data for model estimation. Generally, three waves of longitudinal data are considered to be the minimum for these models. Survival analysis can also accommodate unevenly spaced data and gaps between measurements. However, one major concern with these methods is when the period of observation begins. We need to observe individuals from the moment they become likely to experience the event of interest in order to avoid biased estimates. Similarly, while estimation with unevenly spaced data is not a problem, gaps can also introduce risks of bias as an event can occur during an unobserved period.

**Inclusion of covariates —** Time-constant and time-varying covariates can easily be included in either discrete-time or proportional hazards to assess which characteristics make individuals more or less likely to experience an event.

**Extensions —** Survival analysis can also be extended to include competing risks that is to analyse the potential occurrence of more than one type of event rather than the occurrence or non-occurrence of a single event. In the case of the school-to-work transition, this can be useful in understanding whether certain characteristics are associated with entering the labour force immediately after completing mandatory education compared to pursuing further education or entering a period of unemployment while searching

for a job, or in the case of reaching the statutory retirement age whether individuals choose to continue in some form of formal employment, enter informal employment or leave the labour market completely. Another extension to survival models allows for repeated events. This, for example, could be used to investigate whether experiencing unemployment during one's employment career multiple times increase the chances of experiencing it again.

Survival analysis can also be combined with sequence analysis in order to take into account the past and incorporate individual trajectories into the analysis of the occurrence of events. In the case of retirement, we could investigate, for instance, if individuals with more stable employment trajectories were more or less likely to retire early than those with less stable trajectories characterized by a mix of part-time employment, unemployment, or being out of the labour market. This gives a more dynamic and holistic emphasis to what is a method essentially focused on a single aspect of labour market trajectories.

**Examples of research —** One use of survival analysis is the study of the duration of a transition. The work by Pastore et al. (2021) compares the duration of the school-to-work transition in 14 European countries. They find that higher levels of education are associated with a shorter transition, in other words the time necessary to find a job is reduced. Moreover, they find that in countries with a more developed vocational education, the school-to-work transition is shorter than in countries where these types of programs are less common.

For work to retirement, a study by Platts et al. (2019) uses survival analysis to investigate the likelihood of retired individuals returning to employment in the United Kingdom. They find that it is predominantly well-educated and recently retired individuals that are the most likely to return to work. Moreover, there find that there was no link between an individual's financial situation and the risk of returning to work showing that the "undoing" of this transition is not done with the aim of improving income.

**Conclusion —** To summarize, survival analysis is a method which focuses on the occurrence of events and understanding differences between individuals in the chances of an event occurring. It is thus not a holistic method that aims to analyse entire trajectories but rather focuses on the occurrence of punctual events. Nevertheless, it is a flexible method which doesn't require many consecutive observation waves.

**Suggested reading —** For a more detailed look at survival analysis we suggest:

Allison, P. D. (2014). *Event History and Survival Analysis* (2nd ed.). Sage Publications.

Singer, J., & Willett, J. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford University Press.

Latent Transition Analysis

**Which questions can be answered? —** Latent transition analysis (LTA) is another longitudinal extension of latent class analysis. The difference is that rather than modelling entire trajectories as in the case of longitudinal latent class analysis, it is primarily used to study movements, or transitions, between different states, measured by categorical variables, while taking into account past states (Collins & Lanza, 2010). Like latent class analysis, it can be used to analyse multiple indicators simultaneously for instance combining contract type, level of wages, part-time work, and limited duration contract to study transitions to and from precarious employment. These models are sometimes described as holistic in that they can be used to jointly investigate multiple domains of the life course. However, the entire trajectory is not analysed as once. Typically, latent transition models are used to estimate movements between states conditionally on the previous state i.e. the likelihood of being in a specific state at T2 given the state a person was in at T1. This makes latent transition models similar to dynamic panel models as past information contributes to explaining what we observe. Therefore, they are well suited for the study of age-related transitions, but also potentially, period effects. With the inclusion covariates, there is also the potential to compare differences between cohorts.

**Estimation —** While these models are flexible in that they can include covariates, multiple indicators to create composite states, and included past information, they are very difficult models to estimate computationally as the estimation time increases substantially for each additional wave of data used. In applied research, these models are generally used with few waves of data, typically between two and four waves, as it might be unfeasible to estimate a model with more waves of data even if it is available. Finally, latent transition analysis doesn't focus on overall trajectories and therefore only offers a short-term, almost instantaneous, view of individuals labour market position.

**Data requirements —** Latent transition analysis requires a minimum of two waves of panel data in order to estimate a model. This makes it useful for analysing short-duration panel data. However, LTA is also very difficult to estimate with many waves of data and generally three or four waves is quite often the maximum feasible number of data waves that can be used with LTA. The data also does not need to be evenly spaced but the interpretation of the results needs to take into account any uneven spacing.

**Inclusion of covariates —** LTA models can include covariates which affect the probability of transitioning between states over time. The covariates can be time-inhomogeneous meaning that the effect of the covariates varies for each time point (this is more or less equivalent to specifying an interaction between a variable and time). However, the more waves of data are included, the more difficult it becomes to interpret the effects of explanatory variables if the time inhomogeneous formulation is employed. The covariates can also be time-homogeneous by constraining the association of the covariates to be the same over time and thus estimate an effect that does not vary over time.

**Examples of research —** Tang & Burr (2015) use latent transition analysis to describe transitions to retirement over time, in the United States, using a complex typology that distinguishes between multiple states (Full-time worker; Disabled Partial retiree/part-time worker; Full retiree; Partial retiree/full-time worker) for men and for women (Disabled; Full retiree; Partial retiree; Full-time worker; Home- maker). The results showed that, in general, once individuals retired, they were unlikely to leave retirement. Full-time workers became more likely to transition to full retirement over time compared to any form of partial retirement and, for women, homemakers also became more likely to transition to full retirement over time.

**Conclusion —** LTA is flexible method that allows investigating movements between different states over time. It also allows for incorporating multidimensional constructs directly in the models. However, LTA is not suitable for analysing long duration processes or adopting a holistic approach. This makes it a useful model if the goal is to study short-term changes especially for complex concepts, but not to establish a long-term perspective in relation to labour market outcomes.

**Suggested reading —** For a more detailed discussion of LTA we suggest:

Collins, L. M., & Lanza, S. T. (2010). *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral and Health Sciences*. Wiley.

Lanza, S. T., Bray, B. C., & Collins, L. M. (2013). An introduction to latent class and latent transition analysis. In J. A. Schinka, W. F. Velicer, & I. B. Weiner(Eds.), *Handbook of psychology* (2nd ed.,Vol. 2, pp. 691-716). Wiley.

## 2.3. General Conclusion

Methods which focus on within-individual differences are very commonly used to study the labour market in many disciplines, especially economics, to establish causal effects of changes over time. These methods can be applied to short-duration panels (2 or 3 waves of data) and to continuous or discrete outcomes. Additionally, many "within" methods are not particularly difficult to implement and the results, generally, are easy to interpret. However, they do not offer a holistic view of labour market transitions as they are not designed to simultaneously investigate multiple outcomes in multiple domains nor study entire trajectories. In addition, their focus on within-individual differences means that any time-constant differences, such as cohorts, cannot be studied with these models without certain adjustments to estimation techniques.

By contrast, methods focused on between-individual differences are well suited to adopting a holistic approach. They can be used to estimate entire trajectories measured using continuous or categorical measures. This makes them useful tools for investigating transitions that can unfold over long periods of time, which can vary substantially between individuals, and which can be very complex. Moreover, they are suitable to analyse multiple outcomes or domains simultaneously in the case of sequence analysis and growth curve models. This means that no indices need to be calculated in advance and instead multiple outcomes can be directly included in the models and studied in parallel or jointly. These approaches are also very flexible. In the case of growth curve models, many different time specifications can be tested for describing trajectories which can also be grouped together directly using model-based approaches. In the case of sequence analysis, different dissimilarity measures can be chosen depending on which aspect of trajectories, or sequences, should be focused on.

However, these methods require more data, with three waves being the minimum for the simplest growth curve models and four or five waves being needed for more complex specifications. Sequence analysis also

usually requires four or five waves of data. However, beyond the number of observations required, this data requirement is relative. For example, in the case of school-to-work transitions, four or five waves of data measured from the transition itself would be required. The data requirements for event- and transition-focused approaches (survival analysis and latent transition analysis) are lower with two or three waves being sufficient. Another potential disadvantage is the extreme flexibility of these methods requiring researchers to make choices and test them prior to finalizing an analysis.

Table 1 provides a summary of the methods proposed in this section. For each method which type of data is necessary is specified. In the case of longitudinal data this includes panel studies in general, rotating panels, and panel components of cross-sectional studies. The minimum number of waves is also specified, however, this should be interpreted as a very rough guideline as the effective number of waves that is available from a study depends on the event or transition of interest. For example, in a study on school-to-work transitions using data focused on youth, the chances of having sufficient waves to analyse trajectories are higher than in a panel study with a sample mainly composed of individuals in the middle of their careers. Spacing refers to the regularity of the measurements for longitudinal data. Certain methods, most notably sequence analysis, cannot be used with data that has irregular measurements or gaps (spacing) while this is not a problem for other methods.

The “Aspect of life course” column refers to which of the three main aspects of the life course – age, period, and cohort – can be studied with a given method. The column for types of covariates indicates whether the method can incorporate explanatory variables that change over time (time-varying covariates) and/or explanatory variables that are constant (time-invariant covariates). Certain methods cannot incorporate both, but we have noted situations where non-standard estimation methods can overcome limitations. Finally, we specify if a specific method can directly incorporate multidimensional outcomes (rather than having to create an aggregated outcome measurement beforehand) and whether the method takes a holistic approach i.e. focuses on entire trajectories rather than changes in an outcome.

In the next section, we will discuss which datasets are available for investigating labour market trajectories and transitions and before making recommendations on which methods and data sources can be used together to potentially investigate labour market transitions and trajectories.

▶ **Table 1 — Summary of methods, data requirements, and possible life course aspects they can be used to study**

| Methods | Data type | Min. No. of Longitudinal Time points | Regularly spaced data required | Aspect of life course | Types of covariates | Multi-dimensional Approach | Holistic Approach (trajectories) |
|---|---|---|---|---|---|---|---|
| Fixed-Effects | Longitudinal = panel data, rotating panel, panel component, pseudo-panel | 2 | No | Age, period | Time-varying; time-invariant* | No | No |
| Dynamic panel models | Longitudinal = panel data, rotating panel, panel component, pseudo-panel | 3 | No | Age, period | Time-varying; time-invariant* | No | No |
| Difference-in-difference | Longitudinal **or** cross-sectional | 2 | No | Period | Time-varying and time-invariant | No | No |
| Regression Discontinuity | Longitudinal **or** cross-sectional | 2 | No | Period | Time-varying and time-invariant | No | No |
| Growth Curve Models | Longitudinal = panel data, rotating panel, panel component | 3 | No | Age, period, cohort | Time-varying and time-invariant | Yes | Yes |
| Sequence Analysis | Longitudinal = panel data, rotating panel, panel component | 4 to 5 | Requires regularly spaced data (constant measurement interval) | Age, cohort | Time-invariant; time-varying* | Yes | Yes |
| Event history analysis | Longitudinal = panel data, rotating panel, panel component | 3 | No | Age, period, cohort | Time-invariant and time-varying | No | No |
| Latent transition analysis | Longitudinal = panel data, rotating panel, panel component | 2 | No | Age, period, cohort | Time-invariant and time-varying | Yes | No |

* = non-standard estimation methods necessary

# 3. Data description for 14 selected high-, middle- and low-income countries

This section aims to provide examples of datasets for analysing school-to-work and work-to-retirement transitions. More concretely, it describes 32 datasets from 14 high-, middle- and low-income countries (Bolivia, Brazil, China, Egypt, Ethiopia, India, Indonesia, Japan, México, Peru, South Africa, South Korea, United States and Vietnam).

The data is organized in four main groups considering their main structure: (1) Panel data, (2) Cross-sectional data and two additional categories when the data includes both elements (3) Rotating panels, as well as (4) Datasets with combined designs, which contain datasets with panel and cross-sectional components. For each of the data sets we will provide a brief overview about the principal characteristics of these data, including the sample design, the number of observations, the frequency, as well as the availability of main variables to study the school-to-work and work-to-retirement transitions (including socio-demographic characteristics, labour market indicators, information about retirement, etc.).[2] In addition, if applicable, some comments for the use of the data, as well as details regarding the availability of complementary data are included as they are important supplements for the data presented here. In case information allows, we also mention the availability of harmonized data so researchers can use the here presented data in a comparative framework (for example, the Luxembourg Income Study Database when describing the NIDS). To further guide the work of potential interested users, it also highlights relevant information to take into account when using the data. Finally, modalities on how to access the data are described at the end of each dataset description. The concrete links[3] to the different data sets are provided in the annex (Table A1).

The section ends with a summary highlighting the main features of the data presented here. In addition, it also evaluates whether it is possible to investigate school-to-work and work-to-retirement transitions and provides a suggestion which method(s) detailed in section 2 could be applied for the analyses.

## 3.1 Panel data

### 3.1.1. The Keio Household Panel Survey (KHPS) and Japan Household Panel Survey (JHPS)

The KHPS is an annual household panel survey continuously implemented since 2004. It provides nationwide information on 4,000 households and 7,000 individuals aged between 20 to 69 years. The attrition rate is about 7%.[4] To compensate for dropouts, two refreshment surveys of about 1,400 and 1,000 households were added in 2007 and 2012. To complement the KHPS, the Panel Data Research Center at Keio University, established the JHPS parallel to the KHPS. It collects information of around 4,000 individuals aged 20 years and older.

Both surveys include questions on topics such as household structure, individual attributes, family composition, academic background, labour market participation, etc. However, while the KHPS covers a wide range of topics including employment behaviour, poverty trends, etc., the JHPS, in addition to these topics, focuses on education and healthcare.

The Panel Data Research Center at Keio University provides microdata for non-profit and academic purposes only. Researchers, undergraduate and graduate students can request access to the data online at no cost. Questionnaires are freely available online up to the 16$^{th}$ wave (KHPS 2019) and the 11$^{th}$ wave (JHPS 2019).

*Comments:* Given that the KHPS and JHPS focus mainly on people at working age including early retirement age, they may allow to investigate education-labour and labour-retirement transitions.[5]

[2] The data is complemented with an excel spreadsheet in which more detailed information about each data set is provided.

[3] The information regarding the access to the data sets was made with documentation available during autumn 2021.

[4] This attrition rate has been computed using information provided from the most recent available i.e. wave 15$^{th}$ (KHPS2018), which details 93% as collection rate.

[5] Japanese workers are allowed to draw their pension at the age of 62, but most people work until almost 70 years (69.1 years for women, 70.8 years for men). Source: Tokyo Shoko Research.

Useful information: The Panel Data Research Center at Keio University also provides other panel data such as the *Japan Child Panel Survey (JCPS);* and the *Japanese Panel Survey of Consumers (JPSC);* and other surveys such as the *Great East Japan Earthquake Special Survey (GEES);* the *Cross-National Equivalent File (CNEF)*, which provides harmonized data from general population household-based panel surveys from Australia, Canada, Germany, Great Britain, Japan, Korea, Russia, Switzerland and the United States.

### 3.1.2. (South African) National Income Dynamics Study (NIDS)

NIDS is the first national household panel in South Africa. Between 2008 and 2017, information has been collected every two to three years for over 28,000 individuals in 7,300 households across the country (wave 1 for 2008-2009, wave 2 for 2010-2011, wave 3 for 2012, wave 4 for 2014-2015, and wave 5 for 2017). In 2020, a special follow up, called the *National Income Dynamic Study: Coronavirus Rapid Mobile Survey* (*NIDS-CRAM*), with a subsample of adults from households' interview by NIDS in 2017 was conducted. The response rates vary for each wave. Over the combined field work period of NIDS the cumulative response rate equates to 69%.

NIDS provides information about how households cope with positive (i.e. unemployed relative obtaining a job) and negative shocks (i.e. death of a family member) over time, focusing on the provision of information about changes in a range of topics such as poverty, well-being and inequality, education, fertility and mortality, health, household composition and structure, human capital formation, labour market participation and economic activity, migration, and vulnerability and social capital.

The NIDS microdata can be freely downloadable for analysis, and additionally NIDS secure data is hosted at DataFirst's Secure Research Data Centre.

*Comments:* The NIDS data is also included in the harmonized Luxembourg Income Study Database (LIS). This represents a potential data source for comparative analysis on life-course. See details in https://www.lisdatacenter.org/our-data/lis-database/

### 3.1.3. (American) National Longitudinal Surveys of Youth (NLSY)

The NLSY is a longitudinal survey that follows two cohorts of nationally representative samples of American youth born between 1957-64 (NLSY79) and 1980-84 (NLSY97). The NLSY79 included originally 12,686 respondents aged 14-22 when first interviewed in 1979; after two subsamples were dropped, 9,964 respondents remain in the eligible samples. The NLSY79 is available from Round 1 (1979) to Round 28 (2018). The NLSY97 interviews 9,984 respondents aged 12-17 years when first interviewed in 1997. This ongoing cohort has been surveyed 18 times to date and is now interviewed biennially. The NLSY97 is available from Round 1 (1997-98) to Round 18 (2017-18). Interviews for the NLYS79 were collected annually from 1979 to 1994 and on a biennial basis thereafter. Interviews of the NLSY97 were conducted annually from 1997 to 2011 and biennially since then. During the years since that first survey, the participants in these cohorts typically have finished their schooling, moved out of their parents' homes, made decisions to continuing education and training, entered the labour market, served in the military, married, started families of their own, and thought about their retirement expectations. Data collected from the NLSY79 and NLSY97 respondents chronicle these changes and provide researchers the opportunity to study the life-course transitions, including school-to-work transitions or other labour market transitions, of Americans men and women.

Access to NLSY data for each cohort is available at no cost via the NLS Investigator, an online search and extraction site that allow to review and obtain the NLS variables to create customized data sets. It is not necessary to get an account to browse data, but an account is necessary to save datasets online.

*Comments:* In addition to the NLSY surveys, the US Bureau of Labour Statistics collects the *National Longitudinal Survey of Youth 1979 Child and Young Adult (NLSCYA)* that follows about 10,000 biological child of women in the NLSY79. Also, other currently discontinued longitudinal surveys were collected such as the *National Longitudinal Survey of Mature and Young Women (NLSW)* and the *National Longitudinal Survey of Older and Young Men (NLSM)*.

### 3.1.4. (American) Panel Study on Income Dynamics (PSID)

The PSID is the longest running longitudinal household survey in the world. Since 1968 it collects national representative data of over 18,000 individuals living in 5,000 households in the United States (until 1996 annually, since 1997 biennial).[6] The PSID provides information on life course and multigenerational economic conditions, including includes social, demographic, health, geospatial, and psychological data in a long-term panel representative of the full U.S. population. PSID follows sample members when they change households. A family member who moves out of a PSID family unit is eligible for interviewing as a separate family unit if they are a sample member and living in a different, independent household. The PSID started with two distinct samples: (i) The SRC Sample, a nationally representative sample, designed by the Survey Research Center at the University of Michigan, and (ii) The Survey of Economic Opportunity (SEO) Sample, a non-nationally representative sample of individuals drawn from lower income levels. This allows to perform multiple analysis to study changes in family behaviour and composition in the US. Sample sizes and number of individuals and households vary across waves.[7]

Access to the data is free of charge. Before downloading data for the first time, users must register by completing a short registration form that allows them to access the public use data archive.

*Comments:* The PSID has different complementary modules. The *Child Development Supplement (CDS).* The original CDS included up to two children per household who were 0 to 12 years old in 1997, and followed those children over three waves, ending in 2007-08. Beginning with CDS-2014, the new steady state design of CDS includes all eligible children in PSID households born since 1997. The CDS provides extensive data on children and their extended families with which to study initial live transitions including human and social capital formation. The *Transition into Adulthood Supplement (TAS),* collected from 2005 to 2015 and relaunched in 2017, follow children from the original CDS cohort into adulthood. The *Disability and Use of Time Supplement (DUST),* collected between 2009 and 2013, aimed to investigate the connections between disability, time use, and well-being for older adults.

### 3.1.5. The Egypt Labour Market Panel Survey (ELMPS)

The ELMPS is a nationally representative household panel survey (LFS) implemented during the years 1988 (special LFS), 1998, 2005, 2012 and 2018. Over this time, the ELMPS has become the main source of information for labour market and human development research in Egypt, being the first and most comprehensive source of publicly available microdata on these topics. It includes modules with retrospective labour market information, mainly histories from the past three months.

It surveys a national sample of households and households' members aged 6+, in addition to enterprises operated by the household. In every wave, a refreshment sample of 2,000 to 3,000 households is added to keep the representativeness of the overall sample. The final sample of 2018 included 15,746 households and 61,231 individuals. Of these households, 13,793 households were included also in 2012 (10,041 panel and 3,751 split households). Among individuals, 53,040 were in households that included at least one individual interviewed in 2012 (i.e., either panel or split households). Of the 49,186 individuals included in the 2012 sample, 39,153 (about 80%) were successfully re-interviewed in 2018.

The ELMPS covers a range of topics including sociodemographic characteristics (place of birth, residence, and parental background), education, housing, access to services, residential mobility, migration and remittances, time use, marriage patterns and costs, fertility, job dynamics, and usual employment statistics typical in LFS. In addition to the survey's panel design, the ELMPS also contains several retrospective questions about the timing of major life events such as education, residential mobility, jobs, marriage, and fertility.

To access the microdata, researchers are required to register to the Economic Research Forum (ERF) website and comply with the data access agreement. The data can only be used for scholarly, research or educational purposes.

*Comments:* The ELMPS is particularly rich to study individual's schooling development. It provides very detail information about school and college attendance at various stages of an individual's trajectory, allowing the individual records to be linked to individual sociodemographic information and to school characteristics.

[6] The sample size has grown from 4,800 in 1968 to more than 9,569 families in 2019. Last 2019 has interviewed more than 75,000 individuals.

[7] See details in Table 1 in PSID User Guide 2019 https://psidonline.isr.umich.edu/data/Documentation/UserGuide2019.pdf

### 3.1.6. Young Lives Panel Data (Young Lives)

*Young Lives Study* started in 2002 as a panel study to follow 12,000 children in four developing countries: Ethiopia, India (Andhra Pradesh and Telangana), Peru, and Vietnam. It has the objective to investigate causes and consequences in childhood poverty. Therefore, it is not intended to be nationally representative, but covers children from over-represented poor areas in the country. To focus on poor households, these are randomized within a study site, while the sites themselves were chosen on the phases of predetermined criteria.

Young Lives collects information on two cohorts: (1) Children born in 1994 (with 7 to 8 years), and (2) Children born in 2001 (between 6 to 17 months). Each country sample follows 3,000 children: 2,000 children born in 2001-02, and 1,000 children born in 1994-95. So far, five rounds (2002, 2006, 2009, 2013, 2016) of an in-person household survey, and on by phone in the context of COVID-19 in 2020 are available. These surveys were interspersed with four waves of further in-depth, qualitative interviews. In addition, another phone interview has been planned for 2021, to research the mid-term effect of the pandemic.

Young Lives focuses on five main topics such as education, growth and nutrition, poverty and inequality and youth transitions. It includes questions about life aspirations for their future, as well as details about children's social and environmental conditions. As longitudinal study, it aims to investigate individual changes over time and the impact of earlier circumstances on children's later outcomes, with a special focus on differences between age, ethnicity, gender, location, and income/wealth percentile. Due to the focus of the sample, it allows to study transitions to adulthood including school-to-work transitions, but not labour-to-retirement. Overall attrition from the first to the fifth wave is about 6.5% in general, 4.5% for the youngest cohort and 10.7% for the oldest cohort.[8]

The Young Lives datasets are publicly available from the UK Data Archive. For accessing the data individuals are required to register and apply for a password with the UK Data confidentiality agreement. Access to data details is detailed in Table 2. Detailed documentation is accessible in https://www.younglives.org.uk/content/sampling-and-attrition.

Despite it is a multinational study with similar characteristics that allow cross-country comparisons, samples of each of the included countries have their own characteristics.:

Young Lives Ethiopia

The Ethiopian sample follows households in both rural (60 per cent) and urban (40 per cent) areas located in 20 sentinel sites in the five major regions of Ethiopia (Amhara, Oromia, SNNPR and Tigray, plus the capital city Addis Ababa). An in-depth look at particular aspects of children's lives with a smaller sample of 100 children and their caregivers in five communities from each region. The sample also reflects an equal number of older and younger age groups (cohorts) and boys and girls. This involves group-based activities, individual interviews, and observational work exploring their local environment. Additionally, a school survey was introduced to Young Lives Ethiopia in 2010. This was followed by two survey rounds in 2013 (Grades 4 and 5), and 2016-17 (Grades 7 and 8).

Young Lives India

Young Lives in India follows 3,000 children in two States, Andhra Pradesh and Telangana. Alongside the household and child surveys, Young Lives India developed a qualitative longitudinal stream of research following a sub-set of 200 children over a seven-year period. The data from qualitative research is not shared nor put into a public archive because of concerns about confidentiality, but it can be accessed by contacting lead qualitative researchers. Attrition in India is about 4.1% (3% for the younger cohort, and 6,4% for the older cohort).

Young Lives Peru (Niños del Milenio)

Young Lives Peru, commonly called *Niños del Milenio,* consists of two rounds of surveyed households: 2,000 children were one year old when their parents were first interviewed in 2002, and an older cohort of 700 children interviewed when they were eight years old. Children were selected in 20 sentinel sites from the poorest areas of the country. Importantly, while the sampling of clusters in the other countries was semi-random, in Peru sampling of clusters was random, and district level was used as a sample frame. Departments covered in Peru were Tumbes, Piura, Amazonas, San Martin, Cajamarca, La Libertad, Ancash, Huánuco, Lima, Junín, Ayacucho, Apurimac, Arequipa and Puno. Sample attrition in Peru is about 11.2% (8.2% for the younger cohort, and 14.1% for the older cohort).

[8] See details about sampling and attrition in https://www.younglives.org.uk/content/sampling-and-attrition.

Young Lives Vietnam

The sample of children in Young Lives Vietnam was selected in 2001 using a semi-purpose sampling strategy. A sentinel site was defined as commune-based (20 sentinel sites and 31 communes were included in the study sample. Among these communes, 15 were from the poor group (48%), 9 from the average group (29%), and 7 from the above-average group (23%). Provinces included in the sample are Phu Yen, Ben Tre, Lao Cai, Hung Yen and Da Nang. Due to the non-random sampling of poor sites, the sample is not nationally representative. However, the Young Lives Vietnam sample represents the share of different ethnic groups and gender, allowing to study ethnic diversity within the children sample. The Young Lives of Vietnam includes 5% (156 children) of the second biggest ethnic group (H'Mong), and 64 Dao children.[9] Sample attrition in the Vietnam sample is 2.5% for the younger cohort and 8.6% for the older cohort.

### 3.1.7. The Greater Jakarta Transition to Adulthood Longitudinal Survey (GJTAS)

The Greater Jakarta Transition to Adulthood Longitudinal Survey is a survey conducted by the Demography Program, Australian National University (ANU), focusing on young adults' life transitions including their political and religious affiliations, their voting behaviors, gender aspiration, health and wellbeing and risky health behavior in Greater Jakarta areas (Province of Jakarta, Bogor, Depok, Tangerang, and Bekasi cities and municipalities). The study interviewed young adult Indonesians at three points in time to find their economic and social outcomes changes across time or to see early experience impact on longer-term outcomes. It collected retrospection information about education history, school to work transitions, employment over the lifecycle (first jobs, unemployment, re-employment), marital and fertility histories, and migration. This study has a broad aim which to examine how changes in young Indonesians' lives affect their progression to becoming independent, secure adults and in what ways this progress differs for those who experienced their schooling in the different eras and in the transition period. It also examines whether the outcomes differ by gender and by socio-economic strata. This study is using longitudinal analysis to examine whether young adults in Jakarta can change their circumstances across time or whether early poor outcomes cannot be reversed.

The GJTAS samples were collected through a multi-stage PPS random sampling method in Jakarta, Bekasi, and Tangerang. The survey was fielded in three waves (2010, N=3,006; 2014, N=1,816; and 2018, N=1,120) of a panel survey of young adults (ages 20-34 years old). The survey was also accompanied by qualitative case studies with smaller sample and target of respondent (N=80; ages 20-34 years old) in Greater Jakarta. The qualitative study examined the life courses of young people.

Description of the study and list of publications can be accessed at: https://demography.cass.anu.edu.au/greater-jakarta-transition-adulthood-longitudinal-survey. There is no publicly available data, and researchers interested to work with this data are advised to contact the School of Demography Program at Australian National University (ANU) at https://demography.cass.anu.edu.au/contact-us

### 3.1.8. The Indonesia Family Life Survey (IFLS)

The IFLS started in 1993 to collect information on socioeconomic and demographic characteristics, including fertility, health, education, migration, and employment at the individual-, and household[10]-level. Its household survey collects retrospective information for most topics (fertility, marriage, education, employment, migration) in the survey, providing researchers the possibility to study the effects of changes over time in government programs and household decisions. Besides what is commonly provided in household surveys (household and individual information), the IFLS links household-level data to community-level data on public services and economic infrastructure, which allows researchers to potentially understand how surrounding conditions affect family behaviour and investigate the effect of community policies. The original sample comprises 7,224 households and over 22,000 individuals across 13 provinces on the islands of Java (DKI Jakarta, West Java, Central Java, DI Yogyakarta, and East Java), Sumatra (North Sumatra, West Sumatra, South Sumatra and Lampung), Bali, West Nusa Tenggara, Kalimantan (South Kalimantan), and Sulawesi (South Sulawesi)[11]. It was

[9] Nationally, neither H'Mong nor the Dao is among the five largest ethnic groups.

[10] IFLS uses National Statistics Office (BPS)'s definition of household. A household is defined as a person or a group of people who inhabit part or all of a building and usually live together and eat from the same kitchen. What is meant by eating from the same kitchen is if the management of daily needs is managed together into one.

[11] In 1993, there were 26 provinces in Indonesia.

designed to comprise approximately 83% of the Indonesian population and much of its socioeconomic and culture heterogeneity. The sample of Wave 1 was selected from the nationally representative sampling frame used in the 1993 SUSENAS or National Socioeconomic Survey. Its sampling scheme randomly selects households stratified across provinces. Provinces are selected to maximize representation for the population. IFLS sampling scheme had selected 321 enumeration areas (EAs), with oversampling of urban areas and smaller provinces to facilitate urban-rural and Javanese–non-Javanese comparisons. For each selected household, a representative member provides household-level demographic and economic information. IFLS interviewed with the following household members: household head and his/her spouses, two randomly selected children of the head and spouse age 0-14, one randomly selected individual age 50 or older and her/his spouse, and one individual age 15-49 and his/her spouse-randomly selected from remaining members.

The initial objective in 1993's survey was to interview both women of reproductive age and older individuals in the same survey. The subsequent surveys applied a broader objective to include the life-cycle transition of men and women allowing for studying life-cycle transitions such as school-to-work, changes in employment status, and work-to-retirement. Also, within-households sampling rules include interviews for never married men and women. In addition to the cross-sectional sample, the IFLS designed as a longitudinal study to follow households and individuals over time. The baseline panel was collected 1993 in-home face-to-face information of the household head, spouse, and a sample of their children.

The survey has currently 6 waves of data collection since then in 1997 (IFLS 2) and 1998 (IFLS 2+), 2000 (IFLS 3), 2007 (IFLS 4) and 2014 (IFLS 5). In each wave after 1993, IFLS re-contacted respondents from the main households in 1993 as long as they still resided in any of 13 IFLS' provinces. The target respondent in IFLS re-contact protocols includes the main respondents in 1993 or those who were born before 1968. Individuals splitting off or moving out from the original households in 1993 and their spouses and biological children were also tracked and interviewed. IFLS2+ was conducted in 1998 to capture the impact of Asian economic crisis in 1997 and its sample covered 25% of the sample in IFLS 2. From 2000 and onwards, IFLS expanded the criteria of the targeted respondents. Besides tracking the 1993 main respondents and IFLS 1 household members who were born after 1968, the survey also tracked individuals born since 1993 and after 1988 in the original households, and origin household members born between 1968-1988 if they were interviewed or not interviewed in 1997 (or 2000 and 2007 in the subsequent waves). The last survey or IFLS wave 5 (2014) has around 50,000 individuals from 16,000 households in the sample, with around 14,000 individuals are the household members from the original 1993's survey.

IFLS data updates, notes as well as tips and frequently asked questions are provided here: https://www.rand.org/well-being/social-and-behavioral-policy/data/FLS/IFLS/datanotes.html#IFLS1. IFLS datasets are publicly available and can be downloaded by registering at: https://www.rand.org/well-being/social-and-behavioral-policy/data/FLS/IFLS/access.html.

Comments: The IFLS is part of the set of Family Life Surveys that contained detailed household and community surveys of developing countries conducted by RAND Corporation. The currently available country surveys cover Malaysia (The *Malaysian Family Life Survey - MFLS,* 1976-77, 1988-89), Indonesia (The *Indonesian Family Life Survey – IFLS,* 1993, 1997, 2000, 2007, 2014), Guatemala (The *Guatemalan Survey for Family Health – EGSF,* 1995), and Bangladesh (The *Matlab Health and Socio-Economic Survey MHSS-1,* 1996), available at https://www.rand.org/well-being/social-and-behavioral-policy/data/FLS.html

### 3.1.9. The (Chinese) Rural Urban Migration Data (RUMiC)

The RUMiC is a longitudinal annual survey covering a four-year time span, aiming to understand institutional barriers that limit rural-to-urban migration to improve living conditions of rural migrants. It tracks respondents so long as they remain in the surveyed cities and villages. It collects data about migrant's health, education, employment, social networks, household income and expenditure, housing conditions, and place of origin.

It consists of 3 independent surveys: the Urban Household Survey (UHS with 5,000 households), the Rural Household Survey (RHS, of about 8,000 households), and the Migrant Household Survey (MHS, of about 500 households).[12] Since 2008, four waves of the UHS and RHS as well as five waves of the MHS have been collected. Each of the surveys include comprehensive information on household and personal characteristics. The MHS additionally includes questions related

[12] Data collection started in 2006 by a group of researchers at the Australian National University, the University of Queensland and the Beijing Normal University and was supported by the Institute for Labor (IZA). Since 2017 it is run independently by the Survey Data Center at Jinan University.

to the migration history. For the RHS and UHS, individuals are tracked using their permanent address. The MHS tracks individuals through exploiting individuals' work and home address information, as well as other contact details in both cities and home villages. Also, lottery incentives were implemented to improve tracking of migrants.

The RHS and UHS were conducted using random samples from the annual household income and expenditure surveys carried out in cities and rural villages. The RHS was conducted in villages across 9 provinces, while UHS and MHS were carried out in 19 and 15 cities respectively. Between the first and second waves, the attrition rate for the RHS was 1% and the UHS about 5,7%. The attrition rates for these two samples increased between the second and the third waves because a change in survey conductor, but they remain in a low range (http://idsc.iza.org/rumic).

The RUMiC complements existing surveys such as the recently collected Chinese Household Income Project (CHIP) survey which collects data on rural and urban surveys, and the China Health and Retirement Longitudinal Study (CHARLS), a biennial survey focused on individuals aged at 45 and older.[13]

The scientific use files for the first two waves publicly available, and data application forms can be downloaded from the International Data Service Center of the IZA website and should be completed with a description of the research project and submitted to idsc@iza.org.

### 3.1.10. China Family Panel Studies (CFPS)

The CFPS is a nationally representative, longitudinal survey of Chinese communities, families, and individuals. It started in 2010 by the Institute of Social Science Survey (ISSS) of the Peking University, China, with three waves of full sample follow-up surveys in 2012, 2014, and 2016. In addition, a small-scale sample maintenance survey was conducted in 2011. The CFPS covers 25 provinces/municipalities/autonomous regions, representing 95% of the Chinese population. The baseline survey of 2010 includes 15,000 households and almost 30,000 individuals with an approx. 79% response rate.

The CFPS collects information at three levels: (i) individual- (from 9 years or older), (ii) family-, and (iii) community-level longitudinal data in China. The original target sample size was 16,000 households. Half of the sample (8,000) was generated by oversampling with five independent sampling frames (called -large provinces) of Shanghai, Liaoning, Henan, Gansu, and Guangdong. Each of the sub-samples had 1,600 households. The other 8,000 households were from an independent sampling frame composed of 20 provinces (called -small provinces). Each sub-sample in the CFPS study is drawn through three stages: county (or equivalent), then village (or equivalent), then household.

Access to this data is provided by the Institute of Social Science Survey. For accessing the public data, new users must register. Also, the CFPS provides access to county-level restricted data which includes county ids that can potentially be linked to the CFPS public-use datasets. To have access to the county-level restricted data, users must fill in an application form (http://www.isss.pku.edu.cn/cfps/docs/20190813210906969141.pdf?CSRFT=YHED-K0LD-IUFN-5AAR-P3EN-7Q4V-6YMH-RX05)

### 3.1.11. The Indian Human Development Survey (IHDS)

The IHDS is a collaborative project from the University of Maryland, College Park; the National Council of Applied Economic Research (NCAER) in Delhi; Indiana University, and the University of Michigan. In 2005, it collected information of 41,554 households and 215,754 individuals across India (from 33 states, 384 districts, 1,503 villages, 971 urban blocks located in 276 towns and cities) The IHDS is designed as two main panels: IHDS1 (2004-2005) and IHDS2 (2011-2012). IHDS 2 re-interviewed about 83% of the IHDS1 households in addition to any split households that resided in the same community. In case of attrition or lost households, a replacement household was randomly selected in the same neighbourhood to refresh the sample. The IHDS2 sample contains 2,134 new households. The IHDS3 is currently in the field using CAPI and it is planned to be released in 2023. Linking information from the two rounds is possible by using the linking files which are provided after online registration and downloading the files (https://ihds.umd.edu/data/data-download).

The IHDS profits from a rich survey history in India conducted by the National Council of Applied Economic Research (NACER). Questionnaire's design and the IHDS sample were borrowed from important Indian surveys including *the National Sample Surveys*, the *National Family and Health Surveys* and the 1993-1994 *Human Development Profile of India*

[13] A scientific paper describing in detail the content of the survey can be found here https://ftp.iza.org/dp7860.pdf

*(HDPI)*, a survey conducted from 1993-994. The documentation states that, with some caveats, it is feasible to identify few individual households in the HDPI so that can be linked to the IHDS.

Access to the data is free of charge after following online registration.

### 3.1.12. The Korean Labour & Income Panel Study (KLIPS)

The KLIPS is the only labour-related panel survey in South Korea that comprises relevant cross-sectional and time-series data. It is administrated annually since 1998 to a sample of 5,000 urban households, and about 14,000 household members age 15+. Currently, 23 waves are available with the last wave completed in 2020. The sample household retention rate is 87,6% in wave 2 (1999), 80,9% in wave 3 (2000), 77,3% in wave 4 (2001), and stabilized after that at a decrease rate of about 1% per year, eventually reaching 66,2% in wave 21 (2018). Attempting to re-engage the reserve sample resulted in a in 4,334 surveyed households in 2017, and 5,004 surveyed households in 2018. To our knowledge, no more recent information regarding response rates is available.

The KLIPS has two main sections: (i) The Household Dataset derived from the Household Questionnaire, and (ii) The Individual Dataset compiled from the Individuals questionnaires. It provides extensive information about family and individual characteristics such as household income and expenditures, financial status, education, employment (hours worked, occupations), personal and life satisfaction, job-seeking activities, labour market mobility, etc.

The Supplemental Survey (SP) focuses on different topics such as school-to-labour market and work-to-retirement transitions. For instance: Wave3-SP was administered to young persons between ages of 15 and 30. Wave4-SP focused on health and retirement targeting individuals aged 45 and old. Wave6-SP was administered to those aged 50 and old.

To access to the data, researchers must register online by agreeing to the Terms of Use and acknowledge Notice on Collection and Usage of Personal Information.

## 3.2. Cross-sectional data

### 3.2.1. The (Japanese) Employment Status Survey (ESS)

The ESS is a household survey, conducted every three years from 1956 to1982 and every five years since 1982. It collects information of individuals aged 15+ of approximately 490,000 households. It is representative at the national level. The survey is collected in person by trained interviewers. Items include sociodemographic information, labour market characteristics related to main and second jobs of persons engaged in work, details of childbearing, childcare and housekeeping activities. Questions for persons not engaged in work include human capital characteristics (education, experience, tenure, etc.) and reasons for wishing to work. The most recent available dataset is from 2017.

The survey reports are freely available online, and access to meta- and microdata should be requested to the Statistics Bureau of Japan.

*Comments:* Given that surveyed population collects information of people with at least 15 years old, it allows to investigate education-labour and labour-retirement transitions.

### 3.2.2. The (Japanese) Labour Force Survey (J-LFS)

The J-LFS is a monthly household survey among members aged 15+ years usually residing in Japan. It is collected on a nationwide scale since July 1947, and currently includes approximately 100,000 individuals of 40,000 households from the whole country. It provides information about the labour force (weekly hours work, type of employment, etc.), as well as reasons for taking non-regular employment, unemployment duration, etc. From 2001, the J-LFS includes a *Special Survey* which aims to investigate the details on employment and unemployment status, to supplement the monthly LFS.

Questionnaires are freely available online. Historical statistics are publicly available at https://www.stat.go.jp/english/data/roudou/lngindex.html. Access to the meta and microdata should be requested from the Statistics Bureau of Japan.

### 3.2.3. (Mexican) National Household Survey (Encuesta Nacional de los Hogares - ENH)

Since 2014, the (Mexican) National Institute of Statistics and Geography (Instituto Nacional de Estadística y Geografía, INEGI) surveyed about 64,000 households in Mexico. The ENH is a cross-sectional annual survey which is nationally representative including urban and rural areas. It collects sociodemographic information, education, health, household characteristics, occupations, and information. It is currently available until 2017.

The metadata is freely available and can be downloaded online.

### 3.2.4. (Mexican) Telephonic Survey of Occupation and Employment 2020 (Encuesta Telefónica de Ocupación y Empleo - ETOE)

ETOE is a household survey which was collected monthly between April and June 2020 to monitor the national employment situation of people age 15+ during the COVID-19 pandemic. It has a maximum expected non-response rate of 35%. It uses the statistical designed of ENOE, which was not collected during this time. It provides information of about 45,547 households (i) April: 14,294, ii) May: 13,884, iii) June: 17,369).

Information and metadata can be obtained at no cost.

### 3.2.5. (South African) General Household Survey (GHS)

The GHS is an omnibus household survey collected annually since 2002. It replaced the October Household Survey (OHS) conducted from 1993 to 1999. It surveys a nationally representative sample of private households (non-institutionalized and non-military) in all nine provinces of South Africa including residents in worker's hostel. The Master Sample is designed to be representative at provincial level and within provinces at metro/non-metro levels. It includes a total of 20,000 households, with a response rate of around 87%.[14] The GHS aims to determine the progress of development in the country and investigate the extent of service delivery and the quality of services in key economic sectors including education, health disability, social security, energy, access to water and sanitation, environment, telecommunications, transport, refuse removal, household income, access to food, and agriculture.

Public files of microdata are publicly accessible by filling a register form provided by DataFirst.

### 3.2.6. (Bolivian) Household Surveys (Encuesta de Hogares: BOL-EH)

Since 2005, BOL-EH has been collected annually to obtain information about the living conditions of households in Bolivia, representative at the national level from the 9 national provinces in urban and rural areas. It also provides disaggregated information at the national, rural, urban and departmental level with exception of Beni and Pando for which estimates are provided together.

It covers multiple topics including household characteristics, such as migration, health, education, employment, income, expenses, etc. The 2020 wave follows the Sample design from 2021 that considers structural households' characteristics such as provision of basic services, highest level of education of chief of the household, etc. In 2020, the BOL-EH collected information of 11,292 households. The non-response rate for the BOL-EH in 2020 is about 2%.

Methodological details can be found in http://anda.ine.gob.bo/index.php/catalog/88. Data files can be downloaded freely.

*Comments:* The BOL-EH belongs to the group of multiple surveys that the Bolivian National Institute of Statistics started to collect from 1978. Among other collected surveys are the *Encuesta Permanente de Hogares (EPH)* between 1980-1988, the *Encuesta Integrada de Hogares (EIH)* between 1989-1995, the *Encuesta Nacional de Empleo (ENE) between* 1996-1998, and the *Encuesta Continua de Hogares del Programa MECOVI between* 1999-2004.

### 3.2.7. The Indonesian Family Life Survey (IFLS) East

The IFLS East is a replication of the IFLS Indonesia conducted at eastern provinces in 2012. The survey was conducted to close the gap in data availability to promote development to the eastern part of Indonesia. Similar to its national-scale survey, the IFLS East collects individuals, households and communities' data, including health and education facilities in

[14] Details available in https://www.statssa.gov.za/publications/P0318/P03182019.pdf

the enumeration areas. The sample comprises of around 10,000 individuals, living in 2,500 households in 99 EAs in seven provinces: Nusa Tenggara Timur, Kalimantan Timur, Sulawesi Tenggara, Maluku, Maluku Utara, Papua Barat, and Papua.

IFLS East sampling method was done in three stages: in the first stage, one province was randomly selected from each Kalimantan and Sulawesi while the rest of provinces in the survey were purposively selected. In the second stage, in each province, the survey selected 14 villages randomly from the nationally representative sampling frame used in SUSENAS 2010. Unlike the national-scale survey, the IFLS East used village rather than EAs given the difficulties in identifying the border of EAs in Eastern Indonesia's areas. In the third stage, the survey created a listing of smallest local area within village, based on the identification made in each village. From this list, the survey randomly selected smallest local area. In the final stage, the survey developed a list of all households in selected smallest local areas, and it applied a simple random sampling to select 20-30 households from each area. In IFLS East, all members of selected households were interviewed.

All the questions in the IFLS East were replicated from the national IFLS, including retrospective information on employment, fertility, marriage, education, and migration. To date, there has not been any follow up surveys since 2012.

Documentations, questionnaires, and datasets of IFLS East 2012 are available to download from: https://surveymeter.org/en/data_sakertim.

## 3.3. Cross-sectional data with rotating panels

### 3.3.1. (Egypt) Labour Force Survey (E-LFS)

The E-LFS is a quarterly survey collected by the Central Agency for Public Mobilization and Statistics (CAPMAS). It started in 1957 (November) and it is conducted in successive rounds (quarterly, bi-annually, or annually) till now. The survey covers a national sample (urban and rural) of households and all individuals permanently residing in surveyed households. Overall, each quarter of the E-LFS covers 22,896 households each quarter (January-March, April-June, July-September, and October-December) and about 85,000 households and 340,000 individuals annually.[15] The response rate is about 90% at the national level (85% urban, and 98% rural). Following international recommendations, from 2012, the E-LFS conducts two different questionnaires: (i) A short version collected in January-March, April-June and October-December, and (ii) A long version, collected from July-September, which includes more information about housing conditions and immigration. From 2018, the sample design was developed such that sample size was withdrawn 50% of the (panel households) visited in the same quarter last year and 50% of the sample size (new households) visited for the first time. The E-LFS has the objective to provide quarterly labour force statistics (size of manpower, employed and unemployed labour force) and their geographic characteristics.

It is worth mentioning that with the support of CAPMAS, the Economic Research Forum (ERF) has successfully conducted the *Egyptian Labour Market Panel Survey (ELMPS)* described in the subsection section 3.1.5 of this report.[16]

Similarly, to the EMPLS, access to the microdata is granted to researchers after registering on the ERF website and complying with the data access agreement. The data is only available for scholarly research or educational purposes. Microdata can be accessed after registration to the National Data Drive website.

*Comments:* Since 2009, the ERF has cleaned and harmonized the E-LFS to enable the analyses of this data together with other LFS of Arab countries. It would be interesting to use this data for cross-country comparisons of labour market transitions in the region. Moreover, since the E-LFS interviews all people in the household, it maybe be possible to investigate retirement conditions (although it has not more focused on active labour market transitions).

### 3.3.2. The National Labour Force Survey from Indonesia (SAKERNAS)

The SAKERNAS started in 1976 and is a national representative household survey with the objective to obtain labour market information of all working age individuals within sampled households. Up to now, SAKERNAS has undergone various changes: From 1986 to 1993 SAKERNAS was collected on a quarterly basis, and from 1994 to 2001 on an annual

[15] This number is approximated using number of cases interviewed in wave 2017: 82,902 HH and 335,396 individuals (Source: http://www.erfdataportal.com/index.php/catalog/149/data-dictionary)

[16] See details in Table 6 for the ELMPS in https://erf.org.eg/app/uploads/2019/10/1360.pdf

basis every August, from 2002 to 2004, apart from being on an annual basis, it was also carried out on a quarterly basis. Meanwhile, from 2005 to 2010 SAKERNAS was conducted semi-annually (February and August). Between 2011-2014, it was carried out on a quarterly basis (February (Quarter I), May (Quarter II), August (Quarter III), and November (Quarter IV)).

The SAKERNAS is designed to monitor the Indonesian labour market indicators, and cover information up to the provincial level. The latest SAKERNAS conducted in August 2020 and February 2021 included questions about the impact of COVID-19 pandemic on individuals employment status, work hours, and income. Since 1976, SAKERNAS has undergone various changes in terms of the frequency of the survey: From 1986 to 1993 SAKERNAS was collected on a quarterly basis, and from 1994 to 2001 on an annual basis every August, from 2002 to 2004, apart from being on an annual basis, it was also carried out on a quarterly basis. Meanwhile, from 2005 to 2010 SAKERNAS was conducted semi-annually (February and August). Between 2011-2014, it was carried out on a quarterly basis (February (Quarter I), May (Quarter II), August (Quarter III), and November (Quarter IV)). Starting in 2015, SAKERNAS was again held on a semi-annual basis. The result of the quarterly data collection at provincial level considered samples sizes of about 65,400 households (in 1976), and it reduced over time to 50,000 households approx. The result of data collection for the third quarter was published to regency/municipality level, and it included a bigger sample of about 200,000 households (50,000 households from the quarterly sample, and 150,000 households added to the sample package). In 2009, the sample interviewed about 300,000 households. For the period (2011-2014), in addition to the quarterly sample, a sample which is intended to obtain annual figures as an estimate for the presentation of data to the district/city level was collected. For the bi-annual surveys the sample size for February's SAKERNAS is smaller compared to the August survey, which limits the estimates to represent up to the provincial level for February's survey. The February's survey covers 75,000 households, while August' survey covers 300,000 households. Since 2011, SAKERNAS has been using a rotating panel sampling design. About 75% of the households remain in the sample for two consecutive survey rounds. A maximum of 4 times an ultimate sampling unit is interviewed.[17]

Data can be acquired via the Data Dissemination Division of the BPS https://www.bps.go.id/.

### 3.3.3. The (Mexican) National Survey of Occupation and Employment (Encuesta Nacional de Ocupación y Empleo - ENOE)

The ENOE is a nationally representative household survey that furnishes the main source of information about the urban Mexican labour market. Since 2005, it provides quarterly information of 120,260 households about socioeconomic and labour force characteristics (occupation, informal employment, sub-occupation) of the urban population. While socioeconomic characteristics are captured for youth of 12 years and older, until 2014 labour market information was collected only for people of 14 years or older. From the last quarter of 2014, the latter is only collected for people with 15 years old or more.[18] In addition, the ENOE has a rotation rate of 20%; in other words, 80% of the sample remains each quarter so that each household remains in the sample for about 5 quarters.[19]

It consolidates information from individuals age 12 and older from the National Survey of Urban Employment (Encuesta Nacional de Empleo Urbano - ENEU) and the National Employment Survey (Encuesta Nacional de Empleo - ENE).

Information and metadata can be downloaded freely.

### 3.3.4. (Indian) Periodic Labour Force Survey (PLFS)

The PLFS is a quarterly survey that covers whole India except the villages in Adaman and Nicobar Islands. It stared during July 2011-June 2013 with a Pilot survey. The PLFS has a rotating panel sampling scheme in urban areas of two-years. During this time, each selected household in urban areas is visited four times (one with first visit schedule and other three with revisit schedule). This scheme rotation ensures that 75% of the first-stage sampling units (FSUs) are matched between two

[17] Information based on BPS-Statistics Indonesia. 2019. Pedoman Pencacah Survei Angkatan Kerja Nasional 2019. Jakarta: BPS.http://sirusa.bps.go.id/webadmin/pedoman/2019_5_ped_Pedoman%20Pengawas%20Survei%20Angkatan%20Kerja%20Nasional%20Agustus%202019.pdf

[18] Changes in the age of collection of labor market indicators are explained due to changes in the Mexican Constitution that increased the minimum working age from 14 to 15 years.

[19] See sample design (p. 52) in http://www.erfdataportal.com/index.php/catalog/167/study-description.

consecutive visits. After the completion of every two-year period, the sample frame is updated to incorporate potential changes.

The most recent quarterly report for January – March 21 states that at all-India level in urban areas, the PLFS surveys 5,601 FSU blocks, with about 44,000 households and 172,484 individuals. For rural areas, samples for all the 8 quarters will be selected before the survey starts for each two-year period, and the frame remains for this duration (Documentation available on http://mospi.nic.in/sites/default/files/press_release/2.Press_note_QB10_30112021.pdf)

The PLFS has two primary objectives: first, measuring the dynamics of the labour force participation and employment status in the short time of three months for urban areas. Second, estimate the employment and unemployment indicators for both rural and urban areas annually.

Access to the data can be obtained from the Ministry of Statistics (adg.dsdd@mospi.gov.in).

### 3.3.5. (Bolivian) Labour Continuous Survey (Encuesta Continua de Empleo – ECE)

Since 2015, the ECE, a national household survey, has been collected monthly by the (Bolivian) National Institute of Statistics (INE). The ECE surveys members of the household age 14 and older from rural and urban areas in the country. The survey is designed to replace 25% of the surveyed households each quarter. In such a way each household is interviewed 4 times a year.

In 2017 the ECE covered 17,784 households per quarter. It collects monthly information on the employment situation, and job rural-urban mobility. It covers topics related to occupations, education, income, and household expenses.

Access to the data is provided by the INE (www.ine.gob.bo) by contacting ceninf@ine.gob.bo.

### 3.3.6. (Peruvian) Permanent Employment Survey (Encuesta Permanente de Empleo (EPE))

The (Peruvian) National Institute of Statistics (INEI) started to collect the EPE in March 2001 with the objective to provide monthly information about the situation and dynamics of the labour market of Metropolitan Lima and Callao (43 districts of Lima and 6 districts of Callao). It surveys households' members 14 years or older, about sociodemographic characteristics of each member and labour market outcomes. The sample includes about 5,000 households.[20]

The results correspond to the situation observed in cumulative periods of three months (Moving Quarter). The survey was originally conceived as a fixed panel with a quarterly return to households. After a year of implementation and given the increase in the non-response rate, modifications were introduced to improve the sample design to a panel with household rotation: approximately 1/6 of the sample is rotated every quarter, so only 17% of the sample was replaced. In other words, from 2001 to 2002, the survey had a mixed structure that include a rotating panel and mobile quarter. And from 2002 onwards, the survey has kept a mobile quarter structure which interviews one household for three consecutive quarters. The rate of non-response is about 20%.

Microdata is publicly and freely accessible.

### 3.3.7. (South African) Quarterly Labour Force Survey (QLFS)

QLFS is a quarterly rotated household panel collected by Statistics South Africa (StatsSA) from 2008 onwards. It replaced the South African Labour Force Survey and collects data on labour market activities of individuals between 15 to 64 years old (including informal sector, private households, agriculture and small businesses). The QLFS uses a "Master Sample frame" which has been developed with a general-purpose household frame that can be used in all other StatsSA. This Master Sample is designed to be representative at the provincial level, and within provinces at metro/non-metro levels where three geographical types are used: Urban, Tribal, and Farms. Within the metropolitan area, the QLFS is representative of the different geography types that may exist within that metro. The quarterly sample is approximately 30,000 dwellings in which household reside.

[20] In the mobile quarter of April-May-June 2018, about 4,800 households were programmed and 4,445 were interviewed.

The survey is divided equally into four subgroups. 25% of the sampled dwellings are rotated out of the sample. Thus, sample dwellings are expected to remain in the sample for four consecutive quarters.[21] The QLFS is usually collected face-to-face. However, due to the COVID-19 pandemic, the survey mode has switched to CATI since 19 March 2020.

Access to the data can be requested directly from StatsSA.

### 3.3.8. (Brazilian) Monthly Employment Survey (PME)

PME is a rotated household panel between 1980 and March 2016. It was completely revised in 1982, and partially revised in 1988 and 1993. The PME surveyed households for two periods of 4 consecutive months, eight months apart from each other. It interviewed households with members aged 10 years or older, and covered six Metropolitan Areas (Recife, Salvador, Bello Horizonte, Rio de Janeiro, Sao Paulo and Porto Alegre). In March 2014, PME's sample consisted of 33,809 households with 95,122 individuals. While it is possible to follow households over time, it is not possible to follow single individuals because PME does not assign the same ID number to each individual in the household across interviews. There are two versions of PME: The PME-Antiga (the original survey collected until the end of 2002) and the PME-Nova for which major changes to the design and questionnaires were introduced.

From March 2016, the *Continuous PNAD* replaced the PME covering the whole country. Both PM and the Continuous PNAD (see 3.3.8.) have coexisted between 2012 and 2016. Along 36 years, PME has become the main sources to follow short-term labour market outcomes, and income information from the population. It has been mainly used to compute the main unemployment index in the country.

Data can be freely downloaded.

### 3.3.9. (Brazilian) Continuous National Household Sample Survey (Continuous PNAD)

The Continuous PNAD is a survey conducted by the Brazilian Institute of Geography and Statistics (IBGE) to continuously produce information on the labour market together with demographic and educational characteristics. It started in January 2012, replacing progressively information collected in Brazilian Monthly Employment Survey (PME) and the Brazilian National Household Sample Survey (PNAD).

The Continuous PNAD collects monthly, quarterly, and annual nationally representative information on sociodemographic characteristics and labour market outcomes in Brazil. Monthly information collects a restricted set of labour force indicators, quarterly for workforce indicators. The annual survey collects permanent topics of the supplementary survey and complementary indicators related to the workforce. The monthly data is representative only at national level and the rest are representative at the following geographical level: Brazil, Major Regions, Federative Units, 20 metropolitan areas that contain the Capital Municipalities, municipalities of the Capital Region and the Developed integrated region of Greater Teresina.

The Continuous PNAD has a rotating panel structure that surveys household units for five consecutive quarters. Each quarter, about 211,000 households are interviewed and they are visited every three months for five consecutive months; so that a panel data is generated. Progressively, Continuous PNAD replaced job statistics obtained from the Monthly Employment Survey, PME, and Brazilian National Household Sample Survey, PNAD.

Microdata of the Continuous PNAD can be accessed freely.

## 3.4. Datasets with combined (or mixed) designs

The datasets listed here contain two types of data design namely panel and cross-sectional components which follow a different systematic as cross-sectional surveys with a rotating panel. They either have a separate panel study with its own sampling and item characteristics in addition to the collection of the cross-sectional data (ENAHO), or they collect panel information only for particular modules while the rest of the survey follows a cross-sectional design (SUSENAS), or they have a complete different design than other surveys. For example, the CPS is a cross-sectional survey with an atypical

[21] The QLFS has as sampling unit the dwelling, and the unit of observation is the household. If a household moves out of a dwelling after being in the sample for two quarters (for example) and a new household moves in, the new household will be enumerated for the next two quarters. If no household moves into the sampled dwelling, the dwelling will be classified as vacant (or unoccupied).

design where households are followed for four consecutive months, than they are excluded from the sample for eight consecutive months, and after that they are included again in the sample for four consecutive months. In addition, it adds supplementary questionnaires to the monthly collected information on a variety of topics. Supplements are usually conducted annually or biannually. Furthermore, this data allows to mirror panel rotation patterns which is facilitated by IPSUM-CPS documentation for linking CPS monthly data across time spells.[22]

### 3.4.1. (Peruvian) National Household Survey about Poverty and Living Conditions (Programa de Mejoramiento de Encuestas y de la Medición de las Condiciones de Vida (MECOVI) and Encuesta Nacional de Hogares sobre condiciones de vida y pobreza – ENAHO)

MECOVI was collected from 1997 to 2002 by the (Peruvian) National Institute of Statistics (INEI) to obtain information about the employment and income of the population. The updated version of ENAHO, the most important national household survey, started to be continuously collected since 2003 to also include indicators about poverty in the country. The survey (MECOVI) is not collected anymore every quarter and using modules, but ENAHO is collected continuously using a unique questionnaire. ENAHO is collected at the national level, including urban and rural areas from the 24 national provinces including Callao.

The updated version of ENAHO includes two types of survey: (1) Cross-sectional data collected annually from 2002 to 2010, and quarterly from 2011, and (2) A panel data of 5 years, available from 2011.

Between 2002 and 2010, the panel sample was collected every 4 years. From 2002 and 2006, the sample included 6,123 households with no rotation. However, by 2006 the sample was reduced by about 30%. Since 2007, the panel sample was updated and the designed included a panel rotation of about 20% of the new panel sample every year. Also, households in the panel sample are only included for 5 consecutive years. After 2010, the size of the panel sample varied between periods: 7,770 households (2011 – 2015) and 12,700 households (2016-2020). Depending on the comparable years, the common panel sample varies reaching about 6,114 and 10,000 comparable households between 2010-2011 and 2019 - 2020 respectively.

The cross-sectional design of ENAHO is available from 1997-2012 with an old design (annually), and from 2004 to 2021 (quarterly). The size of the sample has been increasing continuously from 8,054 households in 1998 to 37,103 households in 2020. ENAHO includes 5 types of questionnaires: (1) About household characteristics to be filled by the chief of the households, (2) Individual questionnaire related to education, health, employment and income, (3) Individual questionnaire about perceptions of governability, democracy and transparency, (4) Individual questionnaire related to agriculture, (5) Individual questionnaire about labour conditions of independent workers (informal, etc.).

Data can be accessed at no cost.

### 3.4.2. The (Indonesian) National Socioeconomic Survey (SUSENAS)

SUSENAS constitutes the major source of data on social welfare of the Indonesian population. It was designed as a series of large-scale multi-purpose socioeconomic surveys initiated in 1963. The survey was conducted twice a year between 1963-1978 and at least once a year up to 1992. In the last two decades until 2010, SUSENAS was conducted semi-annually, and since then every year.

Between 1992-2005, in addition to a basic social and economic questionnaire (core), a more specialised questionnaire was introduced (module). The core questionnaire covers basic information about household and individual characteristics including health, death, education/literacy, employment, fertility and family planning, housing, and household expenditure. This questionnaire is asked annually with only slight modifications over the years. The module is intended to gather more detailed information on specific topics, especially on data that changes less frequently than a year. There are three modules of SUSENAS and each module is added in a three-year cycle, which covers 3 modules: (i) Household Consumption and Expenditure, (ii) Social, Culture and Education, and (iii) Housing and Health Module. During 1992-2005, the SUSENAS Core was targeted to cover 300,000 households (urban and rural) in the sample, while the module covered 68,000 households with national and provincial level estimates. Between 2004-2006, SUSENAS added a panel segment to the survey, specifically to gather consumption information at household level. This consumption panel survey was conducted every March with as sample around 10,000 households and it was used for national level estimations. Between

[22] See https://cps.ipums.org/cps/cps_linking_documentation.shtml

2007-2010, the SUSENAS consumption panel survey was still conducted every March, but with a larger sample of 68,800 households, allowing for national and provincial level estimations. Since 2011, the consumption module was conducted every quarterly with a sample of around 75,000 households for national and province level estimation. In SUSENAS 2020, the semi-annual panel households were selected from EAs framework of 2010's population census. The sample of first semester or March SUSENAS survey covered 7,5000 EAs or 75,000 households. These selected households became the panel households in the September's survey in 2020.

SUSENAS main sampling frame is taken from 180,000 Census Blocks or Enumeration Areas, or equivalent to 25% population. These EAs are drawn using Probability Proportional to Size (PPS) from the EAs of Population Census. In each selected EA, the Statistics Office updates the list of households before selecting 10 households randomly. The selection of EAs is done independently every survey period. Although it is not designed to capture the labour force statistics, SUSENAS collects information related to economic activities, from: main activities (working, school, doing domestic work, unable to do activities due to disabilities or old-age); working hours; industry, and employment status (self-employed, employees, etc.). These employment indicators and households' characteristics information available in SUSENAS are useful for analysis that examine the relationship between households' socioeconomic characteristics and employment's status of an individual in a particular year.

Access to data can be obtained from the BPS-Statistics Indonesia and can be done online, but only using an Indonesian VPN at https://internal.bps.go.id/. Users wishing to acquire this data should be contact the Data Dissemination Division of the BPS https://www.bps.go.id/.

### 3.4.3. (American) Current Population Survey (CPS)

It is one of the oldest, largest and most well-recognized monthly surveys in the United States (50 states and the District of Columbia). It started to be collected in the early 1950s as response to the economic depression of 1930s. The sample design follows the 4-8-4 sampling scheme which means that households are surveyed for 4 consecutive months, they are excluded for 8, and then included again for another 4 months before leaving the sample permanently.

The CPS survey uses a probability sample of about 60,000 occupied households, and it provides reliable estimates at the state level, and for 12 of the largest metropolitan statistical areas.[23]. It provides rich information about individuals including earnings, education and work characteristics. In addition to the monthly labour market information, the CPS adds supplementary questionnaires to the monthly collected information on a variety of topics, such as child support, health insurance coverage, school enrolment, etc. Supplements are usually conducted annually or biannually, but the frequency and recurrence of a supplement depend completely on what best meets the needs of the supplement's sponsor.

Tables and statistics obtained from the CPS can be obtained from the U.S Bureau of Labour Statistics (https://www.bls.gov/cps/tables.htm). Historical data files can be obtained from different institutions: The NBER (https://www.nber.org/research/data/current-population-survey-cps-data-nber), IPUMS USA (https://usa.ipums.org/usa/), and Inter-University Consortium For Political and Social Research (https://www.icpsr.umich.edu/web/pages/ICPSR/index.html).

# 4. Evaluating the possibility to study core labour market transitions for the selected data sets

As stated at the beginning of the report, one core aim of this report is to evaluate whether the selected data sets are suitable to study school-to-work as well as work-to-retirement transitions in a life course perspective using the discussed methodological approaches. While the choice of method should be primarily based on the research questions, it should also be recognised that practical aspects, such as data availability and whether the data fulfils the main requirements of the planned methodological approach, can constrain the research. Moreover, while some of the selected countries and

[23] The sample size does not allow proper estimates to be obtained at the county level, and in fact, data are not available for most counties sampled due to confidentiality laws.

data sets are more ideal than others to study school-to-work and work-to-retirement transitions, this is also heavily determined on whether enough cases are observed in the considered age range.

In this section we synthesize the information of section 2 and section 3 by summarizing the reviewed data characteristics alongside the data requirements of each method (see Table 2). We evaluate the data sets based on the type (see column 1). Starting with the classical panel data, which could be considered as the preferable type of data when adopting a life course perspective, this type of data is available in 9 of our 14 selected countries. It is predominantly available in high- and to some extent, in middle-income countries. For all those selected panel data sets school-to-work transitions can be analyzed, while work-to-retirement transitions can only be examined for 7 of those countries. The reason for that is that the oldest cohort in those surveys has not yet reached the retirement age, which however also implies that in the future this transition could be examined. In principle, for all these panel data sets the discussed methodological approaches can be used (in particular, the between-individual difference approaches). However, for some of them caution is required as larger measurement gaps (time between waves) might introduce biases and methodological challenges. While generally, the selected transitions can be measured well with short-time intervals, large time intervals usually lead to poorer transition measurement as many of them could have occurred without being observed, except when retrospective information (i.e., information concerning periods between waves) is available (see column 8 for a more detailed evaluation). Finally, for those users interested in comparative research, only three of the panels provide a comparative panel file delivering harmonized cross-national data (see column 6).

Besides panel data, datasets which combine designs – meaning they include a panel component – can be used for studying work-related transitions. Such data sets are available in 3 out of the 14 examined countries. However, due to the fact that one of the surveys includes a panel component to suitable to study transitions, we remain with the evaluation of two. For those countries, the analyses of school-to-work transitions can be implemented, work-to-retirement transitions can only be studied for one country. Regarding the methodological approaches, most of them can be applied to these data sets. While for them the risk of larger measurement gaps is low, the short time period of the repeated observations is however not ideal for employing sequence analysis. Although, it is technically possible, it is not advisable as it requires the observation of the full transition process from its beginning to its end (except when it is only used as a descriptive tool). For instance, when studying school-to-work transitions it would require the continuous observation of individuals between 15 and 25 years. This requirement also involves ensuring that a sufficient number of complete trajectories are observed. Finally, for none of these data sets a comparative panel file (see column 6) is available.

This brings us to datasets which are cross-sectional including a rotating panel. Half of our selected countries provide such data. While generally those datasets can be used to examine school-to-work and work-to-retirement transitions, the possibilities depend heavily on the structure of the survey, the availability of retrospective questions as well as enough information. For most of those data sets also the risk of large measurement gaps is low (besides for one which might require a careful examination of which method can be used), all discussed methods (besides sequence analyses, as this type of data does not include the observation of the full process) can be used.

Finally, for 5 of our 14 countries we have also identified cross-sectional data sources. While generally, cross-sectional data does not allow to study the selected work-related transitions with the discussed methodological approaches, the creation of pseudo-panels (see Box 1, p. 4-5) might offer some limited possibilities to simulate longitudinal data. However, this works only when the exact same information is collected in repeated time-intervals. With respect to the methodological approaches, to the best of our knowledge, pseudo-panels are almost exclusively used with fixed-effects or dynamic panel models. Methods, such as sequence analysis, has not yet been used with pseudo-panels and there are no trivial solutions to do so.

▶ **TABLE 2 (see excel file on the website)**

# 5. Concluding remarks and recommendations

This report provides researchers and relevant stakeholders interested in studying labour market transitions with an overview of available methods and data to analyse these labour market transitions in a life course perspective for a wider variety of countries. More concretely, first, it provides an overview and evaluation (discussion of the strengths and weaknesses) of the methodological approaches to perform basic as well as advanced life-course analysis for *two core phases of working lives*: a) school-to-work and b) work-to-retirement transitions. Furthermore, it gives an overview of 32

selected data sets as well as an evaluation of their suitability to study those transitions in a life course perspective with the discussed methodological approaches.

Overall, we can conclude that in particular for high-income but also to some extent for middle-income countries, the analyses of core labour market transitions have been done already for quite a time resulting in an advanced knowledge and understanding of them. This is reflected in the increasing amount of available scientific literature using the described data sets (for examples of research based on the describe data sets see Table A2 in the annex). For instance, in Latin America the availability of panel household data has allowed to studied school-to-work transitions (Fawcett, 2002), while work-to-retirement transitions are mainly based on household data complemented with data from the Retirement regulator (Calvo et al., 2010; Olivera and Bernal, 2020). However, for most of the middle- and low-income countries our knowledge on crucial life course transitions and more particular labour market transitions remains limited. And some world regions – although providing suitable data sets – remain understudied. Table A3 in the annex, provides for some of those understudied countries (China, Egypt, India and Indonesia, as well as studies using Young Lives) a few research examples related to the labour market transitions relevant for this report.

Based on our examination and evaluation, one core aspect for studying labour market transitions in a life course perspective is the availability of high-quality longitudinal data. While panel data would be the most adequate data type and such data collection should be promoted and financially supported globally, it is also one of the most expensive undertakings. In that respect, cross-sectional data including a rotating panel / or data set with a mixed design can be a suitable alternative This is reflected in the fact that it is a common practise for studying labour market transitions across a lot of countries. However, as described in section 4 those data come with some methodological challenges, related to large measurement gaps, the coverage of transitions and the provision of retrospective data. To overcome such weaknesses, data collectors and providers could consider addressing some of the aforementioned aspects through, for instance, the inclusion of retrospective questions (cheaper option) or the collection of data at shorter intervals (more cost-intensive).

The choice of method, notwithstanding the associated data requirements, should be guided by the substantive research question. If the main interest is to explore medium- or long-run dynamics, holistic methods such as sequence analysis or growth curve models are best suited for establishing and studying complete individual trajectories and the differences between them. In case the main research question concerns the chances of a certain type of event or transition occurring, then methods such as survival analysis or latent transition analysis are a better option. Finally, if there is interest in taking a policy evaluation perspective, or a semi-experimental approach in studying which changes at an individual level could potentially lead to better outcomes, then methods focusing on within-individual differences would be the best choice.

When preparing the report we noticed that one important aspect for the usage of the data is access and documentation. While most of the examined data sets are available for the scientific community at no (or low cost), the documentation is often more problematic. This relates on the one hand to the general accessibility and its quality (i.e. detailed information). On the other hand, to whether survey providers also have the intention to provide the data to a wider international community. In particular for middle- and low-income countries the data documentation is provided only in the national language which makes it difficult for international researchers to work with it. In that respect, although governments and data providers encourage the use of their data for research and store the related publications in their online domains, I might be advisable to invest in the translation of relevant documentation.

# References


Abbott, A., & Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology, Review and prospect. *Sociological Methods and Research*, 29(1), 3-33. https://doi.org/10.1177/0049124100029001001

Aisenbrey, S., & Fasang, A. (2010). New life for old ideas: The "second wave" of sequence analysis bringing the "course" back into the life course. *Sociological Methods and Research*, 38(3), 430-462. https://doi.org/10.1177/0049124109357532

Allison, P. (2009). *Fixed Effects Regression Models*. Sage Publications.

Barban, N., & Billari, F. (2012). Classifying life course trajectories: a comparison of latent class and sequence analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(5), 765-784. https://doi.org/10.1111/j.1467-9876.2012.01047.x

Bollen, K., & Curran, P. (2006). *Latent Curve Models: A Structural Equation Prespective*. Wiley.

BPS-Statistics Indonesia. 2019. *Pedoman Pencacah Survei Angkatan Kerja Nasional 2019*. Jakarta: BPS.

Brüderl, J., Kratz, F., & Bauer, G. (2019). Life course research with panel data: An analysis of the reproduction of social inequality. *Advances in Life Course Research*, 41. https://doi.org/10.1016/j.alcr.2018.09.003

Brzinsky-Fay, C. & Solga, H. (2016). Compressed, postponed, or disadvantaged? School-to-work-transition patterns and early occupational attainment in West Germany. *Research in Social Stratification and Mobility*, 46, 21-36. https://doi.org/10.1016/j.rssm.2016.01.004

Calvo, E., Bertranou, F., & Bertranou, E. (2010). Are old-age pension system reforms moving away from individual retirement accounts in Latin America? *Journal of social policy*. Vol.39 (2).

Chan, S., & Stevens, A. (2004). Do changes in pension incentives affect retirement? A longitudinal study of subjective retirement expectations. *Journal of Public Economics*, 88(7-8), 1307-1333, https://doi.org/10.1016/S0047-2727(02)00223-2

Collins, L., & Lanza, S. (2010). *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral and Health Sciences*. Wiley.

Gabadinho, A., Ritschard, G., Müller, N., & Studer, M. (2011). Analyzing and Visualizing State Sequences in R with TraMineR. *Journal of Statistical Software*, 40(4), 1-37. http://www.jstatsoft.org/v40/i04

Gabay-Egozi, L., & Yaish, M. (2021). Short and long-term consequences of high-school tracks for earnings in Israel. *Acta Sociologica*, 64(3), 294-313. https://doi.org/10.1177/0001699320920919

Grimm, K., Ram, N., & Estabrook, R. (2017). *Growth Modeling: Structural Equation and Multilevel Modeling Approaches*. The Guilford Press.

Kracke, N., Reichelt, M., & Vicari, B. (2018). Wage Losses Due to Overqualification: The Role of Formal Degrees and Occupational Skills. *Social Indicators Research*, 139, 1085–1108. https://doi.org/10.1007/s11205-017-1744-8.

Lux, T., & Scherger, S. (2018). The Effects of Taking Up Employment After Pension Age on Self-Rated Health in Germany and the UK: Evidence Based on Fixed Effects Models. *Work, Aging and Retirement*, 4(3), 262–273. https://doi.org/10.1093/workar/way003

Manzoni, A., Härkönen, J., & Mayer K.-U. (2014), Moving On? A Growth-Curve Analysis of Occupational Attainment and Career Progression Patterns in West Germany. *Social Forces*, 92(4), 1285–1312. https://doi.org/10.1093/sf/sou002

McVicar, D. & Anyadike-Danes, M. (2002), Predicting successful and unsuccessful transitions from school to work by using sequence methods. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 165, 317-334. https://doi.org/10.1111/1467-985X.00641

Moral-Benito, E., Allison, P., & Williams, R. (2018). Dynamic panel data modelling using maximum likelihood: an alternative to Arellano-Bond. *Applied Economics*, 51(20), 2221-2232. https://doi.org/10.1080/00036846.2018.1540854

Mund, M., & Nestler, S. (2019). Beyond the Cross-Lagged Panel Model: Next-generation statistical tools for analyzing interdependencies across the life course. *Advances in Life Course Research*, 41. https://doi.org/10.1016/j.alcr.2018.10.002

Noelke, C., & Horn, D. (2014). Social Transformation and the Transition from Vocational Education to Work in Hungary: A Differences-in-differences Approach. *European Sociological Review*, 30(4), 431-443. https://doi.org/ 10.1093/esr/jcu048

Olivera, J., & Bernal, B. (2020). Choice of pension management fees and affects on pension wealth. *Journal of Economic Behavior & Organization*. Vol 176, 539-568.

Pastore, F., Quintano, C., & Rocca, A. (2021). Some young people have all the luck! The duration dependence of the school-to-work transition in Europe. *Labour Economics*, 70. https://doi.org/10.1016/j.labeco.2021.101982

Piccarreta, R., & Studer, M. (2019). Holistic analysis of the life course: Methodological challenges and new perspectives. *Advances in Life Course Research*, 41. https://doi.org/10.1016/j.alcr.2018.10.004

Platts, L., Corna, L., Worts, D., Mcdonough, P., Price, D., & Glaser, K. (2019). Returns to work after retirement: A prospective study of unretirement in the United Kingdom. *Ageing & Society*, 39(3), 439-464. https://doi.org/10.1017/S0144686X17000885

Reeskens, T. & Vandecasteele, L (2017). Hard Times and European Youth. The Effect of Economic Insecurity on Human Values, Social Attitudes and Well-being. *Journal of Psychology*. 52 (1), 19-27

Reeskens, T. & Vandecasteele, L. (2021). The Impact of Economic Insecurity on Social Capital and Well-Being: An Analysis Across Different Cohorts in Europe. In Almakaeva, A., Moreno, A., Wilkes, R. (Eds.), *Social Capital and Subjective Well-Being*. Cham: Springer

Riekhoff, A.-J. (2016). Institutional and socio-economic drivers of work-to-retirement trajectories in the Netherlands. *Ageing and Society*, 38(3), 568-593. https://doi.org/10.1017/S0144686X16001252

Studer, M., & 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*, 179(2): 481-511

Studer, M. (2013). Weighted Cluster Library Manual: A practical guide to creating typologies of trajectories in the social sciences with R (LIVES Working Papers No.24; Number 24). NCCR LIVES. https://doi.org/10.12682/lives.2296-1658.2013.24

Tang, F., & Burr, J. (2015). Revisiting the pathways to retirement: A latent structure model of the dynamics of transition from work to retirement. *Ageing & Society*, 35(8), 1739-1770. https://doi.org/10.1017/S0144686X14000634

# Annex

### ▶ Table A1. Data accessibility

| Country | Dataset | Link for accessibility* |
|---|---|---|
| Bolivia | ECE | https://www.inec.cr/encuestas/encuesta-continua-de-empleo |
| | BOL-EH | https://www.ine.gob.bo/index.php/censos-y-banco-de-datos/censos/bases-de-datos-encuestas-sociales/ |
| Brasil | PME | https://www.ibge.gov.br/en/statistics/social/labor/16897-monthly-employment-survey-old-methodology.html?=&t=downloads |
| | Continuos PNAD | https://www.ibge.gov.br/en/statistics/social/population/16833-monthly-dissemination-pnadc1.html?edicao=20780&t=microdados |
| China | RUMiC | http://idsc.iza.org/rumic |
| | CFPS | http://www.isss.pku.edu.cn/cfps/en/data/public/index.htm |
| Egypt | ELMPS | http://www.erfdataportal.com/index.php/auth/login/?destination=catalog/157/get-microdata |
| | E-LFS | https://erf.org.eg/erf-micro-data-catalogue-nada/; http://www.erfdataportal.com/index.php/auth/login/?destination=catalog/136/get-microdata |
| Ethiopia, Peru, Vietnam, and India | Young Lives | https://www.younglives.org.uk/data |
| India | IHDS (I and II) | https://www.icpsr.umich.edu/web/DSDR/studies/36151/datadocumentation |
| | PLFS | http://microdata.gov.in/nada43/index.php/catalog/central/about |
| Indonesia | IFLS | https://www.rand.org/well-being/social-and-behavioral-policy/data/FLS/IFLS/datanotes.html |
| | IFLS - East | https://surveymeter.org/en/data_sakertim |
| | SAKERNAS | https://www.bps.go.id/searchengine/# |
| | SUSENAS | https://internal.bps.go.id/ |
| | GJTAS | https://demography.cass.anu.edu.au/greater-jakarta-transition-adulthood-longitudinal-survey |
| Japan | KHPS and JHPS | https://www.pdrc.keio.ac.jp/en/paneldata/datasets/jhpskhps/ |
| | ESS | https://www.stat.go.jp/english/data/shugyou/2017/outline.html |
| | J-LFS | https://www.stat.go.jp/english/data/roudou/index.html |

| Country | Dataset | Link for accessibility* |
|---|---|---|
| Mexico | ENH | https://www.inegi.org.mx/programas/enh/2017/#Microdatos |
| | ENOE | https://www.inegi.org.mx/programas/enoe/15ymas/ |
| | ETOE | https://www.inegi.org.mx/investigacion/etoe/ |
| Peru | ENAHO | http://iinei.inei.gob.pe/microdatos/Consulta_por_Encuesta.asp |
| | EPE | http://iinei.inei.gob.pe/microdatos/Consulta_por_Encuesta.asp |
| South-Korea | KLIPS | https://www.kli.re.kr/klips_eng/index.do |
| South Africa | NIDS | http://www.nids.uct.ac.za/nids-data/data-access |
| | GHS | https://www.datafirst.uct.ac.za/dataportal/index.php/catalog |
| | QLFS | http://www.statssa.gov.za/?page_id=1854&PPN=P0211 |
| USA | NLSY | https://www.nlsinfo.org/content/access-data-investigator |
| | PSID | https://simba.isr.umich.edu/data/data.aspx |
| | CPS | https://www.census.gov/programs-surveys/cps/data/datasets.html |

Note: Links available at the time of the finalizing the report (Dec., 2021).

## ▶ Table A2: Examples of academic papers using data mentioned in this report

| Country | Data Sets | Research using data* |
|---|---|---|
| Bolivia | ECE, BOL-EH | https://www.ine.gob.bo/index.php/comunicacion/publicaciones/ |
| Brazil | PME | https://pubmed.ncbi.nlm.nih.gov/18769511/<br>https://publications.iadb.org/publications/english/document/Who_Suffers_During_Recessions_in_Brazil_en_en.pdf |
| | Continues PNAD | https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3658379<br>http://www.scielo.edu.uy/scielo.php?pid=S2301-15482019000200185&script=sci_arttext |
| Japan | KHPS, JHPS | https://www.pdrc.keio.ac.jp/en/publications/ |
| Mexico | ENH<br>ENOE<br>ETOE | http://www.scielo.org.mx/scielo.php?pid=S0301-70362020000200085&script=sci_arttext&tlng=en<br>http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0188-33802019000100145<br>https://www.jstor.org/stable/41237823?seq=1#metadata_info_tab_contents |
| Peru | ENAHO | https://renati.sunedu.gob.pe/simple-search?query=ENAHO |
| | EPE | https://renati.sunedu.gob.pe/simple-search?query=EPE |
| South Korea | KLIPS | http://www.jed.or.kr/full-text/22-2/Ahn.PDF<br>https://onlinelibrary.wiley.com/doi/abs/10.1111/obes.12404 |
| South Africa | GHS | https://equityhealthj.biomedcentral.com/track/pdf/10.1186/1475-9276-12-64.pdf |
| | QLFS | http://www.statssa.gov.za/publications/catalogue/Catalogue_of_products_and_publications_Latest.pdf |
| USA | NLSY | https://www.bls.gov/nls/additional-publications/ |
| | CPS | https://www.bls.gov/cps/publications.htm |

Note: Links available at the time of the finalizing the report (Dec., 2021).

▶ **Table A3 – Selected summary literature review for a collection of under-researched countries.**

| Dataset | Reference | Summary of the paper | Link to access the paper |
|---|---|---|---|
| **China** | | | |
| RUMiC | Sakellariou, C. & Fang, Z. (2016). Returns to schooling for urban and migrant workers in China: a detailed investigation. Applied Economics. 17. | It estimates the return to schooling for rural-to-urban migrants and urban residents | https://ideas.repec.org/a/taf/applec/v48y2016i8p684-700.html |
| RUMiC | Wang, C., Zhang, C. & Ni, J. Social network, intra-network education spillover effect and rural–urban migrants' wages: Evidence from China. *Economics Faculty Publications.* 33. | It estimates the determinants of rural-urban migrant wages | https://core.ac.uk/download/pdf/232778025.pdf |
| CFPS | Nauck, B., Gropler, N. & Yi, C. How kinship systems and welfare regimes shape leaving home: A comparative study of the United States, Germany, Taiwan, and China. *Demographic Research.* 2017. 40. | It explains societal differences in the event of leaving the parental home as part of the transition to adulthood | https://www.demographic-research.org/volumes/vol36/38/default.htm |
| **Egypt** | | | |
| ELMPS | Amer, M. Transition from education to work. ETF Sharing Expertise in Training. 2007. 65. | It analyzes the school-to-work transition | https://www.etf.europa.eu/sites/default/files/m/C12578310056925BC125744E004AF5DA_NOTE7ESHZN.pdf |
| ELMPS | Angel-Urdiola, D. F. & Semlali, A. Labor Markets and School-to-Work Transition in Egypt: Diagnostics, Constraints, and Policy Framework. Munich Personal RePEc Archive. 2010. 26. | It analyzes the school-to-work transition | https://mpra.ub.uni-muenchen.de/27674/1/MPRA_paper_27674.pdf |
| E-LFS | Assaad, R., Hendy, R., Lassassi, M. & Yassin, S. Explaining the MENA Paradox: Rising Educational Attainment, Yet Stagnant Female Labor Force Participation. *Demographic Research.* 2020. 34. | It studies female participation in different labor market states | https://pubmed.ncbi.nlm.nih.gov/34366710/ |

| Dataset | Reference | Summary of the paper | Link to access the paper |
|---|---|---|---|
| **India** | | | |
| IHDS<br>(I and II) | Dewan, S. & Khan, L. Breaking the Cycle of Vulnerability Education, Skills and Employability for Indian Youth. Just Jobs Network. 2019. 76. | It analyzes the school-to-work transition for vulnerable youth in India | https://www.unicef.org/rosa/media/3926/file/Breaking%20the%20Cycle%20of%20Vulnerability.pdf |
| IHDS (I and II) | Venumuddala, V. R. Determinants of occupational mobility within the social stratification structure in India. arXiv Cornell University. 2020. 9. | It identifies the relationship between education and social mobility of individuals | https://arxiv.org/abs/2005.06802 |
| PLFS | Vijay, A. S. Female Labour Force Participation in India: Insights Through Time Use Survey. *Review of Market Integration.* 2021. 41. | It aims to understand the factors affecting female labour force participation in rural and urban India | https://journals.sagepub.com/doi/abs/10.1177/09749292211031131 |
| PLFS | Bhatt, V., Bahl, S. & Sharma, A. COVID-19 Pandemic, Lockdown and the Indian Labour Market: Evidence from Periodic Labour Force Survey 2018–2019. *The Indian Economic Journal.* 2021. 20. | It analyzes the current labour market from the perspective of COVID-19 pandemic | https://journals.sagepub.com/doi/full/10.1177/00194662211013237 |
| **Indonesia** | | | |
| IFLS | Rizky, M., Surydarma, D. & Suryahadi, A. Effect of growing up poor on labor market outcomes: evidence from Indonesia. A*sian Development Bank Institute*. 2019. 22. | It analyzes the long-term effect of child poverty on labor market outcomes | https://www.voced.edu.au/content/ngv%3A85336 |
| IFLS-East | Cao, Junran & Anu Rammohan. Social capital and healthy aging in Indonesia. *BMC Public Health* 16. 2016.631 | It points out the importance of social capital measures for moderating the impact of poor health, particularly in daily life. | https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-016-3257-9 |
| SAKERNAS | Permata, M., Yanfitri, Y. & Prasuko, A. The labor shifting in Indonesian labor market. *Bulletin of Monetary Econommics and Banking.* 2010. 38. | It studies the direction of labor movement and the characteristics of the shifting labor | https://www.bmeb-bi.org/index.php/BEMP/article/view/373 |

| Dataset | Reference | Summary of the paper | Link to access the paper |
|---|---|---|---|
| **Indonesia** | | | |
| SUSENAS | Muhamad, R. & Firmana, V. Labor market development in Indonesia Has it been for all?. *Center for Economics and Development Studies.* 15. | It analyzes labor market outcomes over a long run period (1992-2012) | https://www.econbiz.de/Record/labor-market-development-in-indonesia-has-it-been-for-all-purnagunawan-muhamad/10010770419 |
| GTAS | Peter McDonald, Iwu Dwisetyani Utomo, Ariane Utomo, Anna Reimondos and Terence Hull. 2013. Migration and Transition to Adulthood: Education and Employment Outcomes among Young Migrants in Greater Jakarta. *Journal of Asian Population Studies* 9(1): 4-27. | It studies the importance of age and migration for schooling and employment patterns among joung people in Greater Jackarta in 2009/2011. | https://www.tandfonline.com/doi/full/10.1080/17441730.2012.736700 |
| **Ethiopia** | | | |
| Young lives | Tafere, Y. & Chuta, N. The Unrealised Promises of Education: The Challenges of School to Work Transition in Ethiopia. *Young Lives*. 2020. 34. | It aims to answer how young men and women make the transition from school to work and the problems they encounter in doing so | https://www.younglives.org.uk/sites/www.younglives.org.uk/files/YL-WP190-5.pdf |
| | Pankhurst, A. & Tafere, Y. Jobs, Businesses and Cooperatives: Young Men and Women's Transitions to Employment and Income Generation in Ethiopia. *Young Lives.* 2020. 50. | How qualifications are related to jobs | https://www.younglives.org.uk/sites/www.younglives.org.uk/files/YL-WP191-3.pdf |
| | Favara, M. Do Dreams Come True? Aspirations and Educational Attainments of Ethiopian Boys and Girls. *Journal of African Economies.* 2017. 23. | How parental and children's aspirations form and document the relation between early aspirations and educational attainment | https://academic.oup.com/jae/article-abstract/26/5/561/4096500?redirectedFrom=fulltext |

Source: Own elaboration

**ilo.org**

**International Labour Organization**

**Route des Morillons 4**

**European Commission**

Rue du Champ de Mars 21

1050 Brussels, Belgium