Simulation framework for generating synthetic panel data

Marija Kukic * Pavel Ilinov * Michel Bierlaire *

October 13, 2025

Report TRANSP-OR 251013
Transport and Mobility Laboratory
School of Architecture, Civil and Environmental Engineering
Ecole Polytechnique Fédérale de Lausanne
transp-or.epfl.ch

*École Polytechnique Fédérale de Lausanne (EPFL), School of Architecture, Civil and Environmental Engineering (ENAC), Transport and Mobility Laboratory, Switzerland, {marija.kukic, pavel.ilinov, michel.bierlaire}@epfl.ch

Abstract

Most existing synthetic data generation methods produce cross-sectional datasets that replicate only aggregated population characteristics, limiting their ability to capture individual-level dynamics over time. This paper introduces a simulation framework for generating synthetic panel data that consistently tracks the same individuals across years. The contribution of this work is threefold: (i) it defines a universal set of time-independent variables representing life trajectories through parametric models informed by demographic literature (ii) it establishes mapping rules to translate these universal variables into time-dependent attributes for any observation year and (iii) it updates model parameters via maximum likelihood estimation using one or more cross-sectional datasets, assessing their impact on time-dependent outcomes. Using data from the Swiss Mobility and Transport Microcensus, we compare data-free and data-integrated implementations of the framework. Results show that the approach produces consistent individual trajectories and that data integration enhances the alignment of synthetic samples with observed aggregates. The proposed framework provides a flexible basis for constructing realistic longitudinal datasets that evolve with new data sources, enabling temporally consistent population modeling and supporting long-term behavioral and policy analyses in the absence of real panel data.

1 Introduction

Transportation research often relies on cross-sectional data, which capture individual characteristics (e.g., age, income, vehicle ownership) and behaviors (e.g., trips taken, modes used) at a single point in time (Borysov and Rich, 2021). This type of data is easy to collect and widely used in transport and social sciences due to its accessibility, relatively low cost, and typically large sample sizes. While cross-sectional data provide a useful snapshot of the population, they lack a temporal dimension, making it impossible to observe changes, establish causal relationships, or capture the dynamic evolution of preferences and habits (Maier et al., 2023). Instead, there is an implicit assumption that variation across individuals reflects the same dynamics that would be observed over time (Wooldridge, 2010).

However, understanding transportation behavior dynamics necessitates a methodology capable of capturing temporal processes (Bhat and Koppelman, 1999) as travel demand is not static, i.e., it evolves in response to both predictable events and unforeseen disruptions (Vagni and Cornwell, 2018). Understanding these dynamics is essential for effective transportation planning and demand forecasting (Haghighi and Miller, 2025).

In contrast, panel data, also known as longitudinal data, are designed to capture the temporal dimension of behavior since they track the same units of observation (e.g., individuals, households) over an extended period. This allows the ability to study dynamic relationships and to model differences, or heterogeneity among the subjects (Frees, 2004). Panel data enable the modeling of state dependence, where past behavior influences future choices, and unobserved heterogeneity, which accounts for persistent, time-invariant differences between individuals. This ability to track change over time provides a much stronger basis for causal inference (Hsiao, 2014).

Although the panel approach has several theoretical advantages over the cross-sectional approach, it is often faced with practical challenges related to data collection. For example, respondents may drop out of the survey altogether (e.g., attrition), provide lower-quality answers as they become tired of repeated participation (e.g., panel fatigue), or even change their behavior and responses because of the repeated questioning itself (e.g., panel conditioning). Also, true panel surveys are rarely available due to high costs, and when they are, they often cover only short periods (Deschaintres et al., 2022). The absence of rich panel data limits the development of transportation models that depend on them, forcing a focus on single-period rather than multi-period analyses (Kukic, Rezvany and Bierlaire, 2024).

The pseudo-panel data are created as a compromise between cross-sectional and panel data. Rather than tracking the same individuals over time, it constructs

pseudo-panel data from repeated independent cross-sections by forming groups of individuals that can be followed across periods. These groups, or cohorts, are defined by stable socio-demographic characteristics such as birth year, gender, or region. Each cohort is then assumed to behave like a "representative individual" (Deaton, 1985). The visualization of these three data types is presented in Figure 1, and their strengths and limitations are summarized in Table 1.

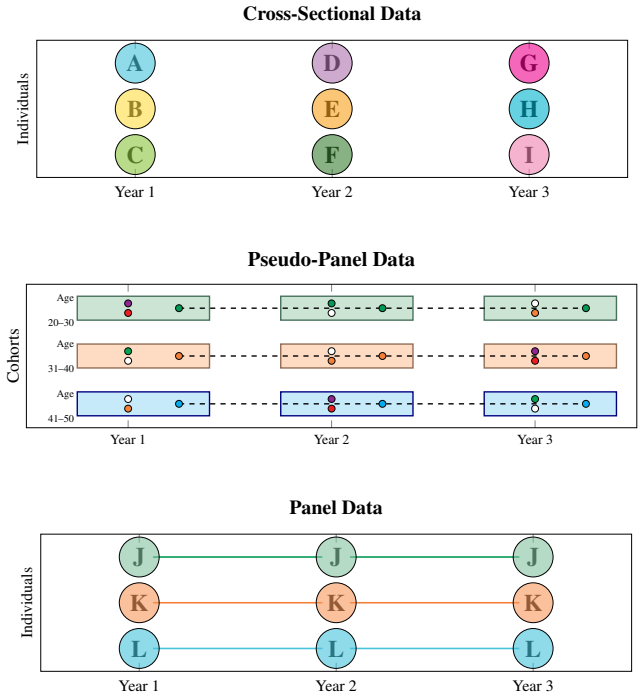


Figure 1: Comparison of cross-sectional, pseudo-panel, and panel approaches

In recent years, interest in generating synthetic data has grown as a way to address the limitations of restricted, incomplete, or sensitive real datasets. Typically, synthetic data are produced from available datasets or from combinations of aggregated and disaggregated sources, with the aim of replicating their statistical properties (Müller and Axhausen, 2010). Since cross-sectional data are relatively easy to obtain, most efforts in synthetic data generation have focused on replicating them, typically producing datasets for a specific region at a single point in time. In contrast, much less attention has been given to synthetic approaches that capture temporal dynamics. An important step in this direction is the study by Borysov and Rich, 2021, who demonstrate how repeated cross-sectional surveys

	Cross-sectional (Goulias and Kitamura, 1992)	**Pseudo-panel** (Deaton, 1985)	**Panel** (Goulias and Kitamura, 1992)
Strengths	Low cost Easy to collect Accessible Large samples	Affordable Avoids attrition Averaging reduces noise	Tracks behavior dynamics Captures hidden differences Enables causal insights
Limitations	No temporal aspect Misleading conclusions Assumption: *Variation = Dynamics*	No individual tracking Aggregation bias Reduced heterogeneity Assumption: *Cohorts are stable*	Expensive to collect Attrition Panel fatigue Panel conditioning

Table 1: Strengths and limitations of different data types

can be combined with generative models to construct synthetic pseudo-panels. However, to the best of our knowledge, there is no standardized approach to generating synthetic panel data.

To bridge this gap, we introduce the concept of universal time-independent variables that characterize an individual's life and propose a novel method for generating panel data based solely on them. Defined in line with the life-event variables commonly discussed in the demographic literature (Nurul Habib, 2018; Hush et al., 2021; Simmons, 2023; Wilmoth et al., 2025), these universal variables enable a data-free process for generating synthetic panel data at any point in time. By construction, any information embedded in the universal variables is automatically carried over into the derived synthetic panels, allowing them to capture typical demographic trends even when real data are unavailable. However, this model-based approach alone cannot reproduce the heterogeneity and context-specific dynamics of real populations. To address this limitation, we develop a mechanism that integrates cross-sectional data into the universal variables and updates the synthetic panels accordingly. As a result, we provide a flexible framework that combines model-based assumptions with data-driven updates, allowing the fusion of diverse information sources (e.g., cross-sectional surveys, published models, census data). In this way, we allow synthetic panels to follow general demographic patterns while also being tailored to specific needs (e.g., testing scenarios, assessing the potential impacts of new policies, integrating constraints).

Our contributions are as follows:

- We introduce a framework based on a universal set of variables with parameterized distributions derived from established literature, ensuring that life events occur in the correct order and adhere to basic demographic rules. In addition, we define the corresponding time-dependent variables, which

enable the construction of synthetic panel data over time, and specify rules for deriving them from the universal variables.

- We design a mechanism to integrate information from potentially multiple cross-sectional datasets (e.g., Swiss Mobility and Transport Microcensus (MTMC) (Swiss Federal Office of Statistics, 2012; 2018; 2023)) into the model using specifically tailored maximum likelihood estimation (MLE).
- We demonstrate the advantage of using estimated parameters to generate the universal variables and examine their impact on the derived synthetic panels, compared to generation with parameters assumed in the literature.

The remainder of this chapter is organized as follows. In the following section, we review the relevant literature. Section 3 introduces the proposed methodology, including the universal and time-dependent variables, estimation, and generation procedures. In Section 4, we first present the outcomes of the generation procedure using both fixed and estimated parameters at a given point in time. We then assess the adequacy of the assumed distributions by comparing the generated samples with real observations from the same period. Finally, Section 5 concludes with a discussion of the findings and potential directions for future research.

2 Literature review

To capture how individual and household mobility evolves over time, several panel data collections have been established. In Switzerland, the Swiss Mobility Panel (SMP, 2020) surveys about 9,500 residents every two years, while the MOBIS GPS Panel combined daily GPS tracking with online surveys and continued during the COVID-19 pandemic (Molloy et al., 2023). In Germany, the German Mobility Panel has surveyed around 1,500 households annually since 1994 (Chlond et al., 2024), and in the Netherlands the Mobility Panel has collected travel diaries since 2013 from about 2,000 households (MPN, 2024). Denmark provides a unique case where registry data cover the entire population with daily records on health, education, employment, and residence (Lynge et al., 2011).

In transportation literature, such data have proven essential for understanding how life events shape travel behavior. Beige and Axhausen, 2017, for instance, show with Swiss panel data that choices of residence, employment, and commuting modes evolve jointly over time. Beige, 2008 applies event-history and duration models to retrospective panels to analyze long- and mid-term mobility decisions. Similarly, Ahmed and Moeckel, 2023 find that while travel behavior is generally stable, life events such as job changes or relocations trigger gradual rather than abrupt shifts, highlighting the value of panel data for distinguishing stable

from changing attributes. These studies emphasize that travel demand evolves on a yearly rather than daily basis, with habitual choices and demographic shifts playing a central role (Gärling and Axhausen, 2003). Panel data thus provide a valuable means of reconstructing life-course sequences to identify the events that most strongly shape travel behavior.

Beyond transportation, panel data have also been applied in other domains. Savcisens et al., 2024 use them in a machine learning framework to predict life outcomes such as early mortality and personality traits from sequences of life events, showing the importance of temporal patterns for accurate prediction. In demography, Oshanreh et al., 2024 develop a dynamic microsimulator based on panel surveys and Bayesian networks to model life-course transitions such as employment, household formation, and mortality.

For our purposes, these approaches are important not because of their specific models, but because they highlight that life trajectories in panel data are best represented as sequences of events with durations. This perspective provides a simple and coherent way to structure universal variables in a panel context. While prior studies use such representations to estimate transition risks or predict outcomes, we employ them as a generative structure for creating synthetic panel datasets in situations where real panel data are unavailable.

These studies demonstrate the value of panel data for capturing life-course dynamics, but they also underline the practical difficulties of collecting such information. In addition, restricted access limits both availability and reproducibility. Taken together, these challenges provide a strong motivation for developing synthetic panel data approaches.

As shown in Table 2, in the field of synthetic population generation, numerous techniques have mostly been focused on cross-sectional data. Following (Yaméogo et al., 2021), these techniques can be broadly classified into three categories: (i) statistical reconstruction (Beckman et al., 1996; Ye et al., 2009), (ii) simulation-based methods (Farooq et al., 2013; Casati et al., 2015; Kukic, Li and Bierlaire, 2024), and (iii) machine learning approaches (Xu and Veeramachaneni, 2018; Borysov et al., 2019; Aemmer and MacKenzie, 2022; Lederrey et al., 2022; Qian et al., 2024). Although these methods employ different strategies for data generation, they all follow a common procedure in which they rely on real data, either aggregated or disaggregated, as input and aim to replicate its statistical properties. The availability of real cross-sectional samples may partly explain why synthetic data generators have paid less attention to the creation of pseudo-panel or panel data.

An alternative approach is proposed by Borysov and Rich, 2021, who construct synthetic pseudo-panels by fixing a set of individuals with given sociodemographics and assigning their travel preferences for each survey year using repeated cross-sectional surveys with a conditional variational autoencoder (CVAE).

	Cross-sectional	Pseudo-panel	Panel
Statistical reconstruction	Beckman et al., 1996 Ye et al., 2009	✗	✗
Simulation methods	Farooq et al., 2013 Casati et al., 2015 Kukic, Li and Bierlaire, 2024	✗	**This paper**
Machine learning	Xu and Veeramachaneni, 2018 Borysov et al., 2019 Aemmer and MacKenzie, 2022 Lederrey et al., 2022 Qian et al., 2024	Borysov and Rich, 2021	✗

Table 2: Comparison of synthetic data generation methods across different data types

These are called pseudo-panels because socio-demographics are fixed while preferences evolve across survey years, with cross-sectional data directly populating the panel year by year. Consequently, without a survey sample, no data can be produced for that year. By contrast, in our framework, the use of cross-sectional data serves a different purpose. Data integration is optional, since the model can operate in a data-free mode. When available, one or more data sources can be integrated to calibrate the parameters of a generative model of universal variables. The goal of this calibration is to shape the distributions in the synthetic panel so that they resemble the patterns observed in the chosen data sources. Importantly, even if cross-sectional data come from a single year, the calibrated model still generates full synthetic life trajectories, allowing panel datasets to be constructed for any year.

As shown in Table 2, to the best of our knowledge, no existing method can generate synthetic panel data without access to real panel data. Our aim is therefore to develop a framework that combines the strengths of model-based and data-driven mechanisms for panel data generation. While Borysov and Rich, 2021 integrate repeated cross-sections by directly filling in pseudo-panels year by year, no published work has introduced a method for incorporating new cross-sectional data into a generative framework that produces full synthetic life trajectories. Our approach allows such integration, enabling synthetic panels to be continuously updated and enriched as new data sources become available. This gap motivates the methodological framework presented in the following section.

3 Methodology

In this section, we formally introduce the components of the framework shown in Figure 2.

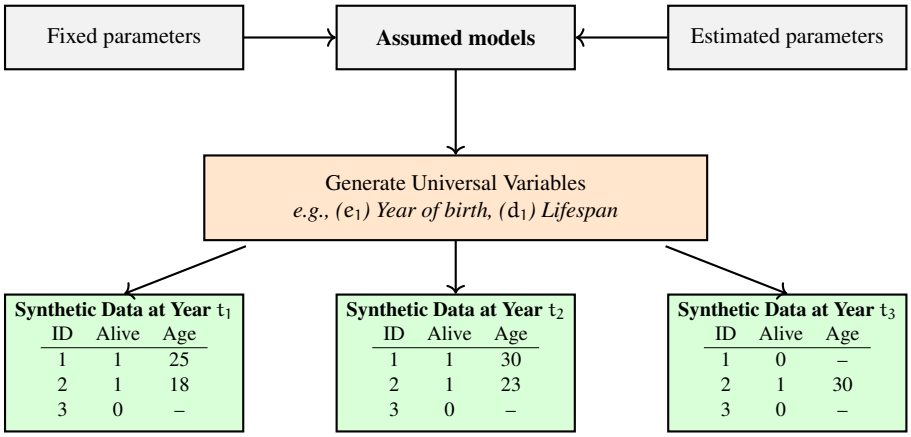


Figure 2: Overview of the framework

The central methodological idea of this work is to characterize the synthetic population using universal time-independent descriptors rather than varaibles that evolve over time. These time-independent variables define individual life trajectories from birth to death. Formally, for each individual i, they are represented by a set of life events, each defined by its starting year e and the corresponding duration d. In contrast, time-dependent variables such as age, income, or driving license ownership change with time. Instead of updating these variables dynamically, our approach generates a fixed set of time-independent descriptors, collectively referred to as the universal dataset, only once at the beginning and then uses them to reconstruct time-dependent variables at any point in time. These descriptors should enforce consistency during the reconstruction phase, and any changes made to them will automatically propagate to all derived datasets.

The selection of universal variables focuses on capturing socio-demographic attributes that usually appear in synthetic populations (e.g., age, driving license, education, employment) (Hradec et al., 2022). Thus, for these attributes, we define their corresponding universal counterparts, allowing us to reconstruct their value at any year of observation. The choice of variables that is considered in our framework is motivated by their relevance in the context of activity-based models (Ahmed and Moeckel, 2023) and aligns with our previous efforts in this domain (Kukic, Li and Bierlaire, 2024; Kukic and Bierlaire, 2025).

As an illustration of the relationship between universal and time-dependent variables, consider the process of generating the age variable, shown in Figure 2.

Age clearly varies with time and is therefore time-dependent. However, in the context of our universal dataset, we can replace it with two time-independent variables: the individual's year of birth e_1 and her lifespan d_1, instead of treating it directly. These two descriptors do not change over time, yet they allow us to deduce both whether the individual is alive at a given point in time and what her exact age is. For all other time-dependent variables, we can establish a similar event-duration model.

Generally speaking, the universal variables are assumed to be unobserved, meaning that we do not have access to real panel data describing their joint distribution. Therefore, these variables must be generated either from existing models with parameters reported in the literature or by extracting information from cross-sectional datasets. Building on this idea, in this work we use parametrized models as the foundation and propose a method to integrate real data through parameter estimation. The complete procedure consists of:

1. **Design of universal variables:** Constructing a fully data-free model by extracting and adapting relevant parametric models from the demographic literature to describe universal variables with available models, and assigning noninformative models to those without.

2. **Specification of time-dependent variables:** Defining the relationship between universal variables and the corresponding time-dependent variables.

3. **Cross-sectional data integration:** Integrating information from multiple cross-sectional datasets into initial distributions of universal variables via maximum likelihood estimation (MLE).

In what follows, we formally introduce the building blocks of our framework.

3.1 Design of universal variables

In our framework, we adopt one year as the unit of time, following evidence that key life events shaping demographic and mobility patterns typically evolve at this temporal scale (Ahmed and Moeckel, 2023). We further define a universal time horizon spanning from y_{min} to y_{max}, representing the predefined lower and upper bounds of individuals' birth years.

Within this temporal setting, an individual's life course is represented as a sequence of universal time-independent variables, each defined by an event and its duration, capturing key milestones from birth to death. The term time-independent reflects the idea that all individuals conceptually exist throughout the entire universal time horizon, and all information about their lives is encoded within it.

However, whether an individual is alive, deceased, or not yet born becomes meaningful only relative to a specific observation year $\mathtt{t}$. To illustrate, consider a universal dataset $\{(1920, 60), (1950, 60), (2010, 70)\}$ representing three individuals characterized by their universal variables: year of birth and lifespan, as shown in Figure 3. In this example, if we observe the population in 1950, individuals 1 and 2 are alive, while individual 3 has not yet been born. By contrast, in 2010, individuals 1 and 2 are no longer alive, whereas individual 3 still is.

In addition to the temporal dimension, universal variables can also be associated with a spatial component. In the context of synthetic individuals, this spatial dimension typically corresponds to a specific country divided into administrative units. In this paper, we focus on Switzerland, which is partitioned into 26 cantons. However, the level of granularity (for both space and time) can be adapted to suit the specific context.

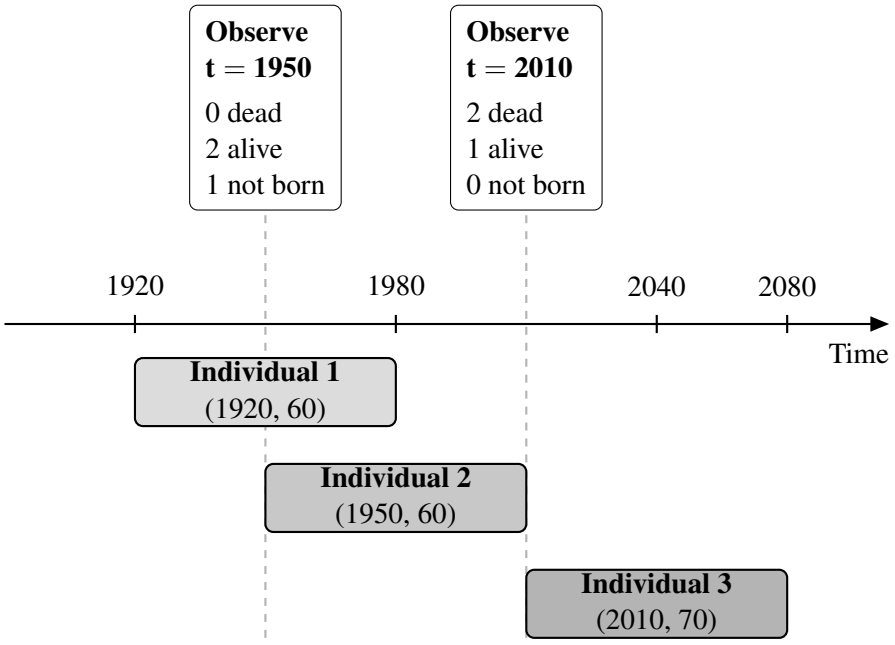


Figure 3: Modeling context

Universal variables	Name
$u_1 = (e_1, d_1)$	Year of birth Lifespan
$u_2 = (e_2, d_2)$	Licence acquisition year Validity
$u_3 = \{(e_3, d_3), (e_4, d_4)\}$	Educational phase Duration of a phase
$u_4 = \{(e_m, d_m, inc_m)\}_m$ $m \in \{5, \cdots, M+4\}$	Employment phase Duration of a phase Initial income
$u_5 = (e_{M+5}, d_{M+5}, c_0^h)$	$e_{M+5} = e_1$ Duration of residence Birthplace
$u_6 = \{(e_n, d_n, c_n^h)\}_n$ $n \in \{M+6, \cdots, M+N+5\}$	Moving year Duration of a phase Location
$u_7 = \{(e_q, d_q, c_q^w)\}_q$ $q \in \{M+N+6, \cdots, M+N+J+5\}$	Starting year of work Duration of a phase Location

Table 3: Set of universal variables

Formally speaking, let the set of universal variables $\mathbf{u} = \{u_j\}_{j=1}^7$ be given by tuples summarized in Table 3. For consistency, within each tuple, the variable e denotes the starting year of a life-course event, d represents its duration, and additional elements may also appear depending on the variable. Unlike u_1, u_2, and u_5 which consist of a single tuple, the other variables represent sets of sequential periods corresponding to distinct phases of life. These periods are guaranteed to be non-overlapping by construction, as they are generated sequentially according to predefined rules that enforce continuity. For example, the number of education phases in u_3 corresponds to different education levels (secondary and tertiary), the number of employment phases in u_4 to the total number of jobs held over a lifetime, and the number of home and work phases, u_6 and u_7, to the number of relocations or workplace changes.

An example of an individual's life sequence is shown in Figure 4. We illustrate the case of an individual born in 1995 who lived for 30 years and obtained their driving licence in 2015. After completing secondary and tertiary education in 2010 and 2014, respectively, they became employed in 2018. After spending 24 years in their birth canton, they relocated to another canton for work.

We define a set of parametric models for each universal variable to represent individual life-course trajectories. These models serve to illustrate the proposed framework through concrete and interpretable examples rather than to reproduce highly detailed, context-specific formulations. Drawing on simple distributions from demographic research, they capture the essential regularities of population

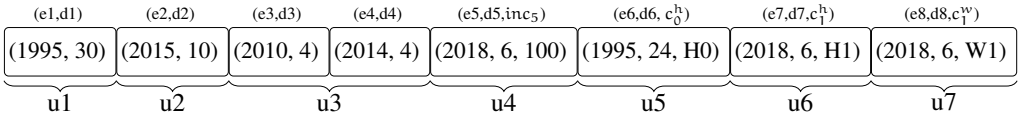


Figure 4: One row of the universal dataset

dynamics while keeping the approach general, transparent, and easily adaptable across regions and time periods through parameter estimation from available data.

In the following subsections, we describe how we model each universal variable introduced in Table 3. For each variable, a probabilistic model is introduced, followed by a specification of its parameter values and a related discussion.

3.1.1 Year of birth and lifespan

Probabilistic model. The year of birth e_1 is treated as a discrete random variable drawn from a uniform distribution over the possible birth years:

$$e_1 \sim \text{Uniform}\{y_{\min}, y_{\min}+1, \ldots, y_{\max}\},$$

with the probability mass function

$$P(e_1 = l) = \frac{1}{y_{\max} - y_{\min} + 1}, \quad y_{\min} \le l \le y_{\max}, \tag{1}$$

where $y_{\min}$ and $y_{\max}$ denote the minimum and maximum possible birth years.

Standard demographic survival models assume a continuous two-parameter Weibull distribution for lifespan d_1:

$$f_p(d_1) = \begin{cases} \frac{k}{\lambda}\left(\frac{d_1}{\lambda}\right)^{k-1} \exp\left[-\left(\frac{d_1}{\lambda}\right)^k\right], & d_1 \ge 0, \\ 0, & d_1 < 0. \end{cases} \tag{2}$$

The shape parameter $k > 0$ controls how mortality risk changes with age, and the scale parameter $\lambda > 0$ determines the characteristic lifespan. Given that we model the year of birth as a discrete random variable and adopt a yearly time granularity, we adapt the existing continuous lifespan model from the literature into a discrete version $\hat{d}_1$. To this end, we define the lifespan probability mass function as:

$$P(\hat{d}_1 = l) = F_{d_1}(l + 0.5) - F_{d_1}(l - 0.5),$$

where F_{d_1} is the CDF of $f_p(d_1)$ introduced in (2). Consequently, we let $F_{\hat{d}_i}$ denote the CDF of $\hat{d}_1$.

Parameters. The parameters with their default values are based on Weon, 2004 and given in Table 4.

Parameter	Meaning	Default value
y_{min}	Lower bound of possible birth years	1920
y_{max}	Upper bound of possible birth years	2050
k	Weibull shape parameter	3
λ	Weibull scale parameter	85 years

Table 4: Parameter definitions and default values for year of birth and lifespan.

Discussion. The year of birth e_1 and lifespan d_1 define the temporal span of an individual's life and constitute the foundation for all subsequent universal variables. The uniform specification of e_1 represents a non-informative assumption, assigning equal probabilities to all potential birth years within the range (Ciganda and Todd, 2024; Wilmoth et al., 2025). We adopt this as a starting point due to the absence of information about the global distribution of birth years. In practice, however, real populations exhibit non-uniform birth year distributions shaped by fertility, mortality, survivorship, historical events, and the sampling frame. Consequently, a uniform prior may differ from empirical data in terms of realism. The Weibull specification for d_1 is in line with (Mahevahaja and Josoa Michel, 2023) and captures realistic survival behavior observed in human populations, ensuring internally consistent demographic structure while keeping the model simple (Weon, 2004). Larger values of λ imply longer expected lifetimes, while $k > 1$ aligns with biological aging, where mortality increases with age. Finally, official population projections (e.g., (Eurostat, 2025)) could provide a valuable foundation for developing more sophisticated models, making this a promising direction for future research.

3.1.2 Licence acquisition year and validity

Probabilistic model. The year of licence acquisition, denoted by e_2, can be expressed as

$$e_2 = e_1 + a_{dl},$$

where the additional universal variable a_{dl} represents the age at which an individual obtains their driving licence. The support of a_{dl} is given by

$$a_{dl} \in [A_{min}, \infty) \cup \{\infty\}$$

where $A_{\min}$ is the legal minimum driving age and $a_{dl} = \infty$ corresponds to individuals who never obtain a driving licence. As a result, we can adopt the following mixed distribution:

$$a_{dl} \sim \begin{cases} \infty, & d_1 < A_{\min}, \\ \infty, & d_1 \geq A_{\min} \quad \text{with probability } \pi, \\ \log \mathcal{N}(\mu, \sigma^2) \text{ truncated to } [A_{\min}, d_1], & d_1 \geq A_{\min} \quad \text{with probability } 1-\pi. \end{cases}$$

The first case corresponds to individuals who die before reaching the legal minimum age $A_{\min}$, the second to those who live longer but never obtain a licence, and the third to individuals who do. For the third case, the corresponding truncated log-normal component has the probability density function:

$$f_\Delta(a_{dl}; d_1) = \begin{cases} \dfrac{1}{a_{dl}\sigma\sqrt{2\pi}} \exp\left(-\dfrac{(\ln a_{dl} - \mu)^2}{2\sigma^2}\right) \\ \hline \Phi\left(\frac{\ln d_1 - \mu}{\sigma}\right) - \Phi\left(\frac{\ln A_{\min} - \mu}{\sigma}\right), \end{cases} \quad A_{\min} \leq a_{dl} \leq d_1 \\ 0, \qquad \text{otherwise}, \end{cases} \tag{3}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function and (3) is rescaled to ensure that it integrates to 1. Analogous to lifespan, we define the discrete version of $\hat{a}_{dl}$ via

$$P(\hat{a}_{dl} = \infty \mid \hat{d}_1) = \begin{cases} 1, & \hat{d}_1 \leq A_{\min} \\ \pi, & \hat{d}_1 \geq A_{\min}, \end{cases} \tag{4}$$

$$P(\hat{a}_{dl} = l \mid \hat{d}_1) = \begin{cases} 0, & \hat{d}_1 \leq A_{\min}, \\ (1-\pi)\left[F_\Delta(l+0.5; \hat{d}_1) - F_\Delta(l-0.5; \hat{d}_1)\right], & \hat{d}_1 \geq A_{\min}, \end{cases} \tag{5}$$

where $l \in \{A_{\min}, \cdots, \hat{d}_1\}$ and $F_\Delta(a_{dl}; d_1)$ is the CDF of the truncated log normal introduced in (3).

Since licences are assumed to be irrevocable, the duration of holding is defined as $d_2 = \hat{d}_1 - \hat{a}_{dl}$, while it is set to zero for non-holders.

Parameters. All parameters with their default values are based on Federal Statistical Office (FSO), 2017 and given in Table 5.

Parameter	Meaning	Default value
μ	Log-location parameter	$\ln(20.5)$
σ	Log-scale parameter	0.15
$A_{\min}$	Minimum legal age	18 years
π	Probability of never obtaining a licence	0.15

Table 5: Parameter definitions and default values for age of driving licence acquisition based on Federal Statistical Office (FSO), 2017.

Discussion. Inspired by empirical distributions shown in Nurul Habib, 2018, we introduce a_{dl} as an additional universal variable that accounts for both individuals who acquire a licence during their lifetime and those who never do. This variable is defined conditional on lifespan d_1, since a person cannot obtain a driving licence before the legal minimum age or after death. The model differentiates between two population groups: a point mass at $a_{dl} = \infty$, representing individuals who never obtain a driving licence, and a truncated log-normal distribution describing the age of licence acquisition among holders. The parameter π denotes the probability of never obtaining a licence.

3.1.3 Education phases

Probabilistic model. We represent secondary and tertiary education as event-duration pairs, (e_3, d_3) and (e_4, d_4), respectively, and assume that individuals pursue each level with a certain probability. To fully model the level of education individuals achieve during their life, we introduce the following variables:

$Z_i^L \in \{0, 1\}$ – education level attendance indicator for individual i,

$a_{end}^S = a_{start}^S + d_3$ – age at the end of secondary education,

g^T – non-negative gap between secondary and tertiary education,

where $L \in \{S, T\}$, $L = S$ refers to secondary, $L = T$ to tertiary education, and a_{start}^S is the starting age of secondary education.

Indicators Z_i^L are defined via following distributions:

$$Z_i^S \sim \text{Bernoulli}(p^S), \qquad Z_i^T \sim \begin{cases} 0 & \text{, if } Z_i^S = 0, \\ \text{Bernoulli}(p^T) & \text{, if } Z_i^S = 1, \end{cases}$$

where p^S is the probability of attending secondary education, and p^T is the probability of attending tertiary education conditional on completing secondary education. If $Z_i^S = 0$, we have $e_3 = d_3 = e_4 = d_4 = 0$; if $Z_i^T = 0$, we have $e_4 = d_4 = 0$. The models of all other variables are presented in what follows.

Duration of secondary education:

$$d_3 \sim \mathcal{N}\big(\mu_d^S,\ \sigma_d^S\big) \text{ truncated to } [d_{min}^S,\ d_{max}^S], \tag{6}$$

where μ_d^S and σ_d^S represent the mean and standard deviation of the secondary education duration, while $[d_{min}^S, d_{max}^S]$ specifies the plausible range of program lengths across different types of secondary schooling. The corresponding discretized version $\hat{d}_3$ is given by:

$$P(\hat{d}_3 = l) = F_{d_3}(l + 0.5) - F_{d_3}(l - 0.5),$$

where $F_{d_3}(d_3)$ is the CDF of the truncated normal distribution in (6).

Gap between secondary and tertiary education:

$$g^T \sim \text{Exponential}\big(\mu_g^T\big), \tag{7}$$

where μ_g^T is the expected length of the gap before entering tertiary education. The corresponding discrete version $\hat{g}_T$ is given by:

$$P(\hat{g}^T = l) = F_{g^T}(l + 0.5) - F_{g^T}(l - 0.5),$$

where F_{g^T} denotes the CDF of the exponential distribution in (7).

Duration of tertiary education:

$$d_4 \sim \text{Weibull}\big(k^T, \lambda^T\big), \tag{8}$$

where k^T and λ^T are the shape and scale parameters defining the duration of tertiary studies. Its discrete counterpart $\hat{d}_4$ is given by:

$$P(\hat{d}_4 = l) = F_{d_4}(l + 0.5) - F_{d_4}(l - 0.5),$$

where F_{d_4} denotes the CDF of the Weibull distribution in (8).

The starting years of the education phases are then given by

$$e_3 = e_1 + a_{start}^S, \qquad e_4 = e_1 + a_{end}^S + \hat{g}^T.$$

Parameters. All introduced parameters with their assumed values are described in Table 6.

Parameter	Meaning	Default value
p^S	Probability of attending secondary education	0.95
p^T	Conditional probability of attending tertiary education	0.60
a^S_{start}	Starting age of secondary education	15
μ^S_d	Mean duration of secondary education	4 years
σ^S_d	Standard deviation of secondary duration	0.5 years
d^S_{min}	Min. duration of secondary education	3 years
d^S_{max}	Max. duration of secondary education	5 years
μ^T_g	Mean gap between secondary and tertiary education	1 years
k^T	Weibull shape parameter	3.5
λ^T	Weibull scale parameter	4.5 year

Table 6: Parameter definitions and default values for education phases based on Organisation for Economic Co-operation and Development, 2021a; Organisation for Economic Co-operation and Development, 2021b

Discussion. Secondary education follows compulsory schooling and typically includes high school or vocational programmes, while tertiary education encompasses studies beyond the secondary level, such as university or other higher education institutions, and usually spans several years. These definitions are consistent with the international standard classification of education (OECD, 2021). Both phases are optional, and individuals typically attend them with certain country-specific probabilities (Härkönen and Sirniö, 2020). Moreover, both phases may be truncated by lifespan d_1, meaning that if death occurs before or during schooling, the phase is truncated accordingly. In this formulation, the duration of secondary education d_3 follows the parametric formulation proposed by Mills, 2011. The variable $\hat{g}^T$ may take the value 0, corresponding to an immediate transition between education levels without any gap, or a positive value, corresponding to a delayed entry into tertiary education.

3.1.4 Employment phases

Probabilistic model. We model an individual's employment history as a sequence of phases, each defined by a start year e_m and a duration d_m, for $m \in \{5, \ldots, M+4\}$, where M is the maximum number of jobs a person can have. Each job phase m is constrained by lifespan and retirement such that:

$$e_m + d_m \leq \min(e_1 + d_1,\ a_{ret}),$$

where $e_1 + d_1$ is the year of death and a_{ret} is the retirement year. To fully model an individual's employment phases, we introduce the following variables

g - transitional gap between education and first employment,

d_m - age dependent job durations,

b_m - breaks between consecutive jobs.

Entry into the labor market:

The start of the first employment phase is dictated by the completion of the individual's highest level of education and a transitional gap, resulting in the starting age of the first employment

$$a_1 = a_{\text{edu,end}} + g, \tag{9}$$

where $a_{\text{edu,end}}$ is the age at the end of the highest completed education and g is the exponentially distributed transitional gap with mean μ_g

$$g \sim \text{Exponential}(\mu_g), \tag{10}$$

representing the expected waiting time before the first job. Given the yearly time granularity, we discretize g into $\hat{g} \in \{0, 1, 2, \dots\}$ and define:

$$P(\hat{g} = l) = F_g(l + 0.5) - F_g(l - 0.5),$$

where F_g denotes the CDF of the exponential distribution in (10).

If no secondary education is attended, entry into the labor market is allowed from the legal minimum working age, i.e., $a_{\text{edu,end}} = a_{\text{min}}^{\text{work}}$. The initial employment start year is then

$$e_5 = e_1 + a_{\text{edu,end}} + \hat{g},$$

where e_1 is the year of birth.

Age dependent job durations:

$$d_m \sim \text{Exponential}(\lambda(a_m)), \tag{11}$$

where $a_m = e_m - e_1$ is the individual's age at the start of job m, and $\lambda(a_m)$ is a piecewise age dependent mean parameter:

$$\lambda(a_m) = \begin{cases} 1.5, & a_m < 25, \\ 2.5, & 25 \leq a_m < 35, \\ 3.5, & 35 \leq a_m < 45, \\ 4.5, & 45 \leq a_m < 55, \\ 9.0, & a_m \geq 55. \end{cases}$$

The corresponding discrete exponential $\hat{d}_m$ is given by

$$P(\hat{d}_m = l) = F_{d_m}(l + 0.5) - F_{d_m}(l - 0.5),$$

where $F_{d_m}(d_m)$ is the CDF of the exponential distribution in (11).

Breaks between jobs:

$$b_m \sim \text{Exponential}(\mu_b), \tag{12}$$

where μ_b is the mean expected break between the jobs. The correspondig discretized variable $\hat{b}_m \in \{0, 1, ...\}$ is given by

$$P(\hat{b}_m = l) = F_{b_m}(l + 0.5) - F_{b_m}(l - 0.5),$$

where F_{b_m} denotes the CDF of the exponential distribution in (12). The start year of the next phase is then

$$e_{m+1} = e_m + \hat{d}_m + \hat{b}_m.$$

Parameters. The list of all introduced parameters with their corresponding values is shown in Table 7. All defaults are based on empirical evidence from Bureau of Labor Statistics, 2019.

Parameter	Meaning	Default value
M	Maximum number of jobs	12
a_{ret}	Retirement age	65 years
$a_{\text{min}}^{\text{work}}$	Minimum legal working age	15 years
μ_g	Mean of exp. dist. for gap before first job	1 year
μ_b	Mean of exp. dist. for break between jobs	1 year

Table 7: Parameter definitions and default values used in the employment history model based on Bureau of Labor Statistics, 2019.

Discussion. We assume a person can have at most M jobs in their lifetime (Bureau of Labor Statistics, 2019). We define the retirement age a_{ret} such that no new employment phase can begin after this threshold. If a phase extends beyond retirement or death, its duration is truncated accordingly. If an individual reaches M jobs before retirement, the final job continues until the earlier of retirement or death. The transitional gap g^T is modeled based on Edemealem, 2022. In this formulation, $\hat{g}$ may take the value 0, corresponding to an immediate transition to the labor market without any gap, or a positive value, corresponding to a

delayed entry into employment. Since research from labor economics shows that job tenure increases with age, i.e., young adults often change jobs while older individuals stay in positions longer (Employee Benefit Research Institute, 2025), we modeled the job duration d_m as a sample drawn from an exponential distribution with an age-dependent mean. Moreover, after each job phase ends, the individual may experience a gap b_m before the next job (Simmons, 2023). As before, $\hat{b}_m$ may take the value 0, representing an immediate transition between jobs without breaks, or a positive value, corresponding to a period of unemployment.

3.1.5 Initial income per employment phase

Probabilistic model. To model the initial income per employment phase, i.e., income at the time of recruitment, we first introduce:

$$\text{edu} = \hat{d}_3 + \hat{d}_4, \quad \text{work}_m = \sum_{5 \le r < m} \hat{d}_r, \tag{13}$$

where edu is the total number of schooling years, work_m is the total accumulated work experience up to the beginning of the employment phase indexed by m, and r indexes all preceding phases. Then, we can adopt a Mincer-type specification (Polachek, 2007):

$$\ln(\text{inc}_m) = \beta_0 + \rho\,\text{edu} + \beta_1\,\text{work}_m + \beta_2\,\text{work}_m^2 + \varepsilon_m,$$

where:

β_0 represents the baseline log-income for an individual with no schooling and no experience,

ρ denotes the marginal return to education,

β_1 and β_2 determine the shape of the experience–earnings profile,

$\varepsilon_m \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ is a normally distributed error term capturing unobserved heterogeneity in income across individuals.

This gives the income at the beginning of the employment phase m:

$$\text{inc}_m = \exp\left(\beta_0 + \rho\,\text{edu} + \beta_1\,\text{work}_m + \beta_2\,\text{work}_m^2 + \varepsilon_m\right). \tag{14}$$

Parameters. Table 8 lists the parameters of the Mincer specification with their default values calibrated to the Swiss case.

Parameter	Meaning	Default value
β_0	Intercept	7.8
ρ	Return to schooling	0.10
β_1	Linear effect of experience on income	0.025
β_2	Quadratic effect of experience on income	-0.0004
σ_ε^2	Standard deviation of the error term	0.3

Table 8: Parameter definitions and default values based on Patrinos, 2016.

Discussion. Note that income is calculated only for employed individuals. Education is assumed to be completed before the onset of the first employment period, and the total number of schooling years, i.e., edu, is deterministically derived from the duration of education. Moreover, we set $\text{work}_5 = 0$ for the first employment phase (indexed by $m = 5$), since no experience is accumulated beforehand. In our framework, both edu and work_m can be derived from the previously defined universal variables that describe periods of education and employment. Once education and work experience have been defined, the initial income for each employment phase, inc_m, can be computed using a Mincer-type specification as in Polachek, 2007. In this model, β_0 corresponds to a baseline monthly income of approximately 2,400 CHF for an unskilled, inexperienced worker. The initial income does not explicitly depend on earnings from the previous job but is indirectly influenced by accumulated experience. Parameter ρ dictates the increase in income per additional year of education, whereas β_1 and β_2 generate a concave earning profile reflecting increasing and then diminishing returns to work experience, while ε_m is the stochastic term used to produce realistic income dispersion. In the future, this framework could also be extended by explicitly incorporating job type as a universal variable (for example, using ISCO groups reported in Swiss labor statistics), which would allow income to depend directly on factors such as education and age rather than solely through accumulated work experience.

3.1.6 Birthplace

Probabilistic model. We model the birthplace of each individual by setting the starting year e_{M+5} and the duration of residence d_{M+5} as

$$e_{M+5} = e_1, \quad d_{M+5} = \max(a_{fm}, a_1), \tag{15}$$

where a_1 is the starting age of the first job defined in (9), and a_{fm} is a parameter representing the typical age at which individuals leave home for the first time.

Locations are modeled at the canton level, with $C = \{1, \ldots, 26\}$ denoting the set of Swiss cantons with population sizes N_c. Birthplaces c_0^h are sampled proportional to N_c, i.e.,

$$P(c_0^h = c) = \frac{N_c}{\sum_{c' \in C} N_{c'}}. \tag{16}$$

Parameters. Parameter a_{fm} is set to 18.

Discussion. The birthplace variable c_0^h is introduced to ensure that each individual is assigned a place of residence prior to their first employment. It is sampled from empirical population data (Moralti et al., 2023), with probabilities proportional to N_c. For simplicity, we assume that individuals who never enter employment remain in their birth canton throughout their lives.

3.1.7 Home and work location phase

Probabilistic model. Home phases are defined via tuples:

$$(e_n, d_n, c_n^h), \quad n = M + 6, \ldots, M + N + 5,$$

where e_n and d_n denote the start year and duration of each subsequent residence c_n^h after c_0^h, and N denotes the total number of moves. Similarly, work location phases are defined as:

$$(e_q, d_q, c_q^w), \quad q = M + N + 6, \ldots, M + N + J + 5,$$

where e_q, d_q, and c_q^w denote the start time, duration, and work canton, respectively. We assume that individuals relocate only for employment purposes. With a slight abuse of notation, this implies that the timing of employment periods and residential phases is coupled as follows:

$$(e_m, d_m) = (e_n, d_n) = (e_q, d_q).$$

Each residential c_n^h and workplace c_q^w phase associated with employment phase m is assigned probabilistically as:

$$P(c_n^h = c) = \frac{w_c(m)}{\sum_{c' \in C} w_{c'}(m)},$$

with employment-specific weights $w_c(m)$ given by

$$w_c(m) = \begin{cases} N_c \cdot \delta, & \text{if } c \in U \text{ and } \text{inc}_m > y^*, \\ N_c, & \text{otherwise,} \end{cases} \qquad \delta > 1,$$

where inc_m is the expected income at the start of employment phase m given by (14), y^* is the income threshold above which individuals are considered high earners, and $U \subset C$ is the set of urban cantons.

Once c_n^h is drawn, the work canton c_q^w is assigned to match the home canton with probability p_h or to a neighbouring canton, based on a predefined adjacency list, with probability p_w.

Parameters. All parameters and their assumed default values are given in Table 9.

Parameter	Meaning	Default value
y^*	Income threshold	6000 CHF
U	Urban cantons	{ZH, GE, BS, VD}
δ	Urban boost factor	1.5
p_h	Prob. work equals home canton	0.95
p_w	Prob. work in neighbouring canton	0.05

Table 9: Parameter definitions and values based on Moralti et al., 2023. Urban cantons: ZH (Zürich), GE (Geneva), BS (Basel-Stadt), VD (Vaud).

Discussion. Home and work phases describe the spatial dimension of an individual's life course. To ensure temporal consistency, home and work phases replicate the start and duration of each employment phase m. Hence, the corresponding triplets could be written compactly as (e_m, d_m, c_m^h) and (e_m, d_m, c_m^w). The construction of $P(c_n^h = c)$ is inspired by gravity-based migration models (Reina et al., 2024). Here, each canton is weighted by its population, with urban cantons receiving an additional boost for high-income individuals. As a result, this rule captures both population size effects and the observed tendency of higher-income individuals to cluster in urban cantons. Finally, the assignment of work locations c_q^w captures the empirical tendency for most individuals to live and work in the same canton, while still allowing for cross-border commuting. During periods without employment (i.e., before the first job, between jobs, or after retirement), work location remains undefined.

3.2 Specification of time-dependent variables

We construct eight time-dependent variables to describe each individual i at time t, as shown in Figure 5: life indicator $x_1(t)$, age $y_1(t)$, driving licence $y_2(t)$, education $y_3(t)$, employment $y_4(t)$, income $y_5(t)$, and two spatial variables $c^h(t)$

and $c^w(t)$ representing home and work cantons. Indicator $x_1(t)$ specifies whether the individual is alive at time t. The following section outlines the mapping rules used to derive these variables from the universal ones. With a slight abuse of notation, we introduce an additional subscript to universal variables to indicate their association with individual i.

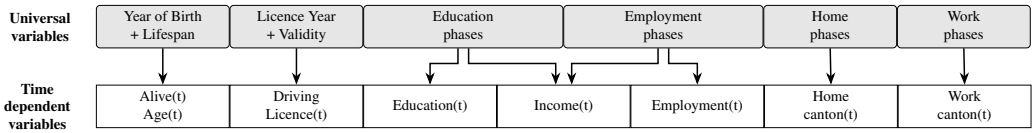


Figure 5: The generated sequence of time-dependent variables

Life-indicator and age: If $e_{1,i}$ and $d_{1,i}$ denote the year of birth and lifespan, we first define the indicator $x_{1,i}(t)$ for an individual i, as follows:

$$x_{1,i}(t) = \begin{cases} 1, & \text{if } e_{1,i} \leq t < e_{1,i} + d_{1,i}, \\ 0, & \text{otherwise.} \end{cases}$$

The rest of the variables are defined only for individuals that are alive at the moment t (i.e., $x_{1,i}(t) = 1$). Consequentially, the age of an individual i that is alive at time t is $y_{1,i}(t) = t - e_{1,i}$.

Driving license status: If $e_{2,i}$ is the year of license acquisition and $d_{2,i}$ is the length of its validity, and an indivdiual i is alive at moment t (i.e., $x_{1,i} = 1$), then the driving license status of individual is defined as:

$$y_{2,i}(t) = \begin{cases} 1, & \text{if } e_{2,i} \leq t < e_{2,i} + d_{2,i} \text{ and } e_{2,i} < \infty \\ 0, & \text{otherwise.} \end{cases}$$

Education status: Education status is defined only for individuals who are alive at time t (i.e., $x_{1,i}(t) = 1$). For such individuals, let $e_{3,i}$ and $d_{3,i}$ denote the start year and duration of secondary education, and $e_{4,i}$ and $d_{4,i}$ the start year and duration of tertiary education. Then, the education status of individual i at time t is a categorical variable defined as:

$$y_{3,i}(t) = \begin{cases} \texttt{Compulsory}, & \text{if } t < e_{3,i}, \\ \texttt{No secondary}, & \text{if } e_{3,i} = 0, \\ \texttt{Secondary}, & \text{if } e_{3,i} \leq t < e_{3,i} + d_{3,i}, \\ & \quad \text{or } (t \geq e_{3,i} + d_{3,i} \text{ and } (e_{4,i} = 0 \text{ or } t < e_{4,i})), \\ \texttt{Tertiary}, & \text{if } e_{4,i} \leq t \text{ and } e_{4,i} > 0, \end{cases}$$

Note that the category "Under 15" comprises individuals below the typical age for secondary education, whereas "No secondary" refers to those old enough to have attended but who did not. These categories are in line with the available categories in the real dataset.

Employment status: Employment status is defined only for individuals alive at time t. Based on this, we can define the following auxiliary variable:

$$\text{Employed}_i(t) = \mathbf{1}\{\exists m \in \{5, \cdots, M+4\} \text{ s.t. } e_{m,i} \leq t < e_{m,i} + d_{m,i}\}$$

Then, for individual i alive at t, employment status is a categorical variable:

$$y_{4,i}(t) = \begin{cases} \texttt{Under 15}, & \text{if } t - e_{1,i} < 15, \\ \texttt{Retired}, & \text{if } t - e_{1,i} \geq a_{\text{ret}}, \\ \texttt{In education}, & (e_{3,i} > 0 \text{ and } d_{3,i} > 0 \text{ and } e_{3,i} \leq t < e_{3,i} + d_{3,i}) \\ & \text{or } (e_{4,i} > 0 \text{ and } d_{4,i} > 0 \text{ and } e_{4,i} \leq t < e_{4,i} + d_{4,i}), \\ \texttt{Employed}, & \text{Employed}_i(t) = 1, \\ \texttt{Unemployed}, & \text{otherwise.} \end{cases}$$

These categories are in line with the available categories in the real dataset.

Income: At the time-dependent level, income is evaluated deterministically for each year t when the individual is alive and employed. To do so, we calculate how much the income of individual i has increased up to time t, relative to their initial income in the current employment phase. Let $m_i^*(t)$ denote the index of the employment phase at time t, defined as

$$m_i^*(t) = \underset{m \in \{5,\ldots,M+4\}, t - e_{m,i} \geq 0}{\arg\min} \; t - e_{m,i}.$$

Additionally, we calculate $\text{work}_i(t)$, which represents the number of years since the beginning of the first employment period, up to time t, as follows:

$$\text{work}_i(t) = \sum_{m=1}^{M_i} \min\big(d_{m,i}, \max(0, t - e_{m,i})\big).$$

Following the same formulation of the effect of work experience as in Mincer's equation (14), and using the previously defined $m_i^*(t)$ and $\text{work}_i(t)$, we compute the income at time t as:

$$\ln(y_{5,i}(t)) = \ln(\text{inc}_{m_i^*(t)}) + \beta_1 \Delta\text{work}_i(t) + \beta_2 \Delta\text{work}_i(t)^2,$$

where $\Delta\text{work}_i(t) = \text{work}_i(t) - \text{work}_{m_i^*(t)}$ and $\text{work}_{m_i^*(t)}$ is given in (13).

In other words, the income $y_{5,i}(t)$ value expressed in CHF:

$$y_{5,i}(t) = \begin{cases} e^{\ln(\text{inc}_{m_i^*(t)}) + \beta_1 \Delta\text{work}_i(t) + \beta_2 \Delta\text{work}_i(t)^2} & \text{if } x_{1,i}(t) = 1 \text{ and } \text{Employed}_i(t) = 1, \\ 0, & \text{otherwise,} \end{cases}$$

can be evaluated for any year within an active employment phase, based on the cumulative work experience up to that moment, with the parameters defined in the same way as in Section 3.1.5.

Although age does not appear explicitly in the equation, income is age-dependent because education and work experience accumulate over time. Moreover, the stochasticity comes implicitly from $\varepsilon_{m_i^*(t)}$ that is embedded in $\text{inc}_{m^*(t)}$.

Spatial variables: At the time-dependent level, home and work locations are obtained from the corresponding spatial phases. For each individual, the home canton at time t is defined as

$$c_i^h(t) = c_{n,i}^h \quad \text{if } e_{n,i} \le t < e_{n,i} + d_{n,i},$$

and the work canton as

$$c_i^w(t) = \begin{cases} c_{q,i}^w, & \text{if } e_{q,i} \le t < e_{q,i} + d_{q,i}, \\ \text{undefined}, & \text{otherwise.} \end{cases}$$

Before the first employment, the home canton equals the birthplace c_0^h, while after the last employment, it remains fixed at the final home canton.

3.3 Sampling scheme

In order to generate synthetic panel data, we need to sample the joint distribution of the universal variables and map to different time steps. The sampling process is guided by the observation that life-course events unfold sequentially, with each stage depending on the previous one. By following this natural order (e.g., birth precedes schooling, which in turn precedes employment), we ensure that the generated data remain consistent and realistic.

Figure 6 illustrates our model as a directed graph, where each variable is represented by a node and incoming edges indicate the variables from which it is derived. The graph distinguishes between two types of nodes: random and derived, with edges labeled either as deterministic or conditional. In line with Section 3.1, random nodes without parent variables are sampled independently from

their parameterized distributions, while those with parents are sampled conditionally, once the values of their parent variables are known. On the other hand, derived nodes are not sampled directly. Instead, they are computed as simple functions (i.e., sums or differences) of their parent variables, once those have been determined. As a result of our simplifying assumptions, random variables are sampled in a sequential order which, together with the introduced deterministic links, yields a tractable model that preserves the temporal structure of life events and generates internally consistent life trajectories.

Because this factorized structure is explicitly defined, synthetic data can be generated by sequentially drawing and computing variables in the order defined by the graph. This process corresponds to forward sampling (Koller and Friedman, 2009), where each step uses the outputs of the previous ones. Starting with independent random nodes (e.g., birth and lifespan), the procedure sequentially samples variables describing subsequent life phases, resulting in a synthetic population where each individual is defined by a vector of time-independent variables (see Figure 4). Once this so-called universal dataset is constructed, it can be mapped to a synthetic panel population at any point in time using the rules outlined in Section 3.2.

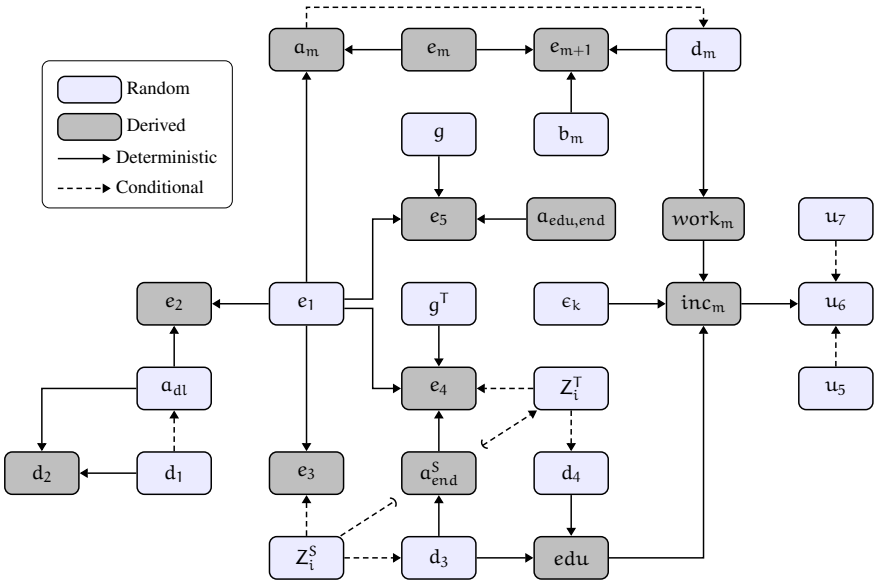


Figure 6: Sampling scheme

3.4 Cross-sectional data integration

Generally speaking, Sections 3.1 and 3.2 provide a complete, data-free framework for generating synthetic panel data at any point in time. However, because

general parametric models for certain universal variables are lacking in the literature, we must rely on uninformative assumptions such as uniform distributions (e.g., for the year of birth). Consequently, this generation process may yield aggregated properties that deviate from those observed in real cross-sectional data. Our objective, therefore, is to develop a mechanism that incorporates information from potentially multiple cross-sectional datasets into the parameters governing universal variables. Specifically, we propose integrating information from cross-sectional datasets using Maximum Likelihood Estimation (MLE). To do so, we first define a model that captures the probability of an individual being sampled in a survey at a given time t. This model establishes a mapping from the universal dataset to the cross-sectional data, enabling us to construct the likelihood of the observed variables by marginalizing over the unobserved components of the universal dataset.

Namely, let $\mathcal{T}$ denote the set of years for which cross-sectional datasets $\mathcal{D}_t$ are available, where each $\mathcal{D}_t$ represents a sample of the population at time $t \in \mathcal{T}$. At each t, we assume the observed individuals $i \in \mathcal{D}_t$ are described by a subset $z_{i,t}$ of all time-dependent variables, i.e.,

$$z_{i,t} \subseteq \{x_{1,i}(t), y_{1,i}(t), y_{2,i}(t), y_{3,i}(t), y_{4,i}(t), y_{5,i}(t), c_i^h(t), c_i^w(t)\}.$$

To integrate information from cross-sectional datasets $\mathcal{D}_t$, we aim to optimize all or a selected subset of the parameters collected in the vector θ, which govern the probability distributions describing the universal variables.

Since cross-sectional datasets typically do not track the same individuals over time, we treat observations from different datasets as independent. Under this assumption, the combined likelihood of multiple datasets simplifies to the product of their individual likelihoods. Moreover, to account for the fact that cross-sectional datasets include only living individuals who were sampled according to specific survey protocols, we introduce a simplified surrogate sampling model that maps universal variables to a particular point in time. Based on these considerations, we define the likelihood $L(\theta)$ of observing all cross-sectional datasets as:

$$L(\theta) = \prod_{t \in \mathcal{T}} \prod_{i \in \mathcal{D}_t} P(s_{i,t} = 1, z_{i,t} \mid \theta) = \prod_{t \in \mathcal{T}} \prod_{i \in \mathcal{D}_t} \sum_{\mathbf{u}_\theta} P(s_{i,t} = 1, z_{i,t}, \mathbf{u}_\theta \mid \theta), \quad (17)$$

where $P(s_{i,t} = 1, z_{i,t} \mid \theta) = \sum_{\mathbf{u}_\theta} P(s_{i,t} = 1, z_{i,t}, \mathbf{u}_\theta \mid \theta)$ is an individual's contribution to the joint likelihood, $\mathbf{u}_\theta$ is a subset of relevant universal variables parametrized by θ, and $s_{i,t} = 1$ indicates that the individual i was sampled for survey at time t. In the following subsections, we first formally introduce the adopted sampling model and then show how the likelihood is derived for the joint distribution of age and driving license status, i.e., $z_{i,t} = \{y_{1,i}(t), y_{2,i}(t)\}$.

3.4.1 Mapping universal dataset to a specific cross-sectional via sampling mechanism

To introduce the sampling model and illustrate its relationship to the actual population and the cross-sectional data at time t, we consider a motivating example shown in Figure 7. In this example, the real population consists of three individuals whose driving license status evolves over time. At time steps t', t'', and t''', two of these individuals are included in the cross-sectional datasets $\mathcal{D}_{t'}$, $\mathcal{D}_{t''}$, and $\mathcal{D}_{t'''}$ according to a specified sampling protocol. As illustrated in Figure 7, at time steps t' and t''', the distribution of driving license status in the cross-sectional data matches that of the true population, while at time t'' a discrepancy arises between the two. To correctly update the parameters of the universal dataset based on the cross-sectional data at time t'', it is necessary to ensure that information is propagated consistently with the sampling protocol and that the parameters are adjusted accordingly.

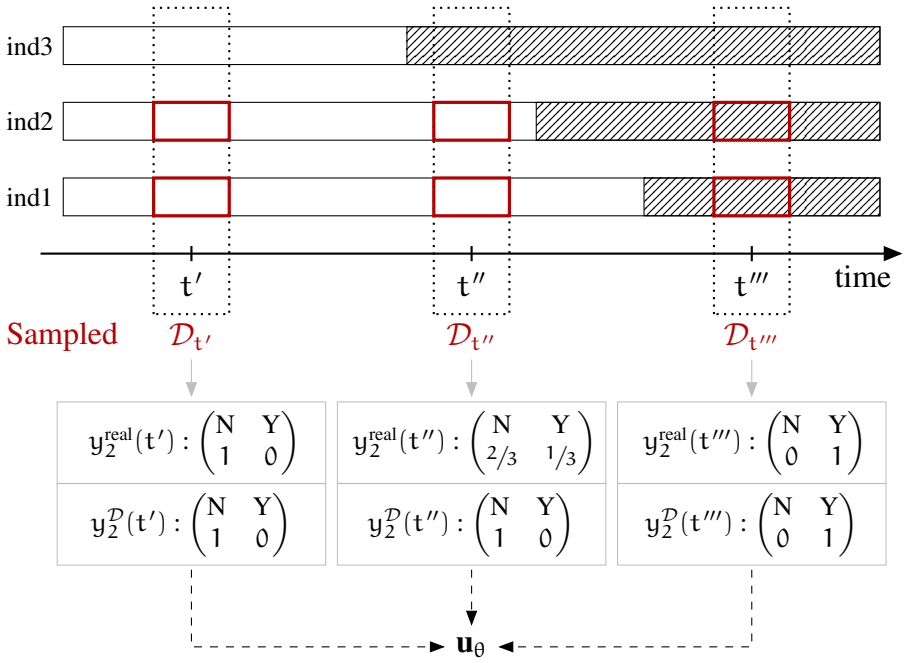


Figure 7: Illustration of sampling probabilities

To model the probability that an individual is sampled for a cross-sectional dataset at moment t, we propose a simple surogate function that reflects empirical sampling bias across different age groups. We assume that adults, especially those of working age, are more frequently sampled, while children and the elderly tend to be underrepresented. To capture this trend, we divide the population into three

smoothly transitioning age groups: young individuals, adults, and old individuals, and introduce two threshold parameters:

τ_1: the age boundary between young and adult individuals,

τ_2: the boundary between adult and old individuals,

with $\tau_1 < \tau_2$. Thus, the final sampling probability is defined as follows:

$$P(s_t = 1 \mid e_1) = \underbrace{\alpha_y \cdot s_1(t - e_1)}_{\text{young contribution}} + \underbrace{0.5 \cdot \alpha_a \cdot (2 - s_1(t - e_1)) - s_2(t - e_1))}_{\text{adult contribution}} + \\ + \underbrace{\alpha_o \cdot s_2(t - e_1)}_{\text{old contribution}} \tag{18}$$

where:

$$s_1(t - e_1) = \frac{1}{1 + \exp(-\gamma(\tau_1 - (t - e_1)))}$$
$$s_2(t - e_1) = \frac{1}{1 + \exp(-\gamma((t - e_1) - \tau_2))}.$$

The sigmoid functions $s_1(t - e_1)$ and $s_2(t - e_1)$ define smooth transitions across age groups, allowing individuals near age thresholds to partially contribute to multiple categories. The α parameters then control how much each group contributes to the overall sampling probability, reflecting real differences in how often different age groups are sampled. In other words, the parameters α_y, α_a, and α_o represent the mixing coefficients for individuals in the young, adult, and old age groups, respectively. These parameters are constrained such that $\alpha_y, \alpha_a, \alpha_o \geq 0$, and $\alpha_y + \alpha_a + \alpha_o = 1$. The parameter $\gamma > 0$ controls the sharpness of the sigmoid transition, with larger values producing steeper transitions and smaller values yielding smoother, more gradual changes around the threshold age. Note that $s_1(t - e_1)$ and $s_2(t - e_1)$ are sigmoid functions and therefore always lie within the interval $[0, 1]$, which makes the expression $0.5 \cdot (2 - s_1(t - e_1) - s_2(t - e_1))$ also bounded between 0 and 1. Since all three terms in the expression for $P(s_t = 1 \mid e_1)$ are multiplied by non-negative weights $\alpha_y, \alpha_a, \alpha_o$ that sum to 1, the total is always greater than or equal to 0 and cannot exceed 1, which makes the model valid.

We initially considered a step-function model, assigning fixed weights to discrete age bins. However, this approach produced unnatural discontinuities and sharp jumps in the distribution at the group boundaries. To make these transitions smoother, we decided to use a sigmoid formulation instead. In Figure 8, we illustrate the differences between the sampling probability distributions generated using a step function and a sigmoid-based formulation.

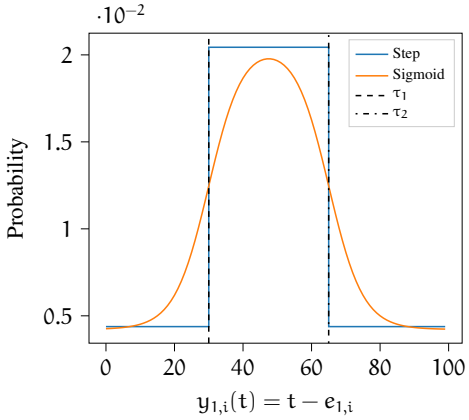


Figure 8: Comparison of sampling probability using step and sigmoids

Apart from enabling us to link universal variables with cross-sectional data observed at different points in time, the parameterized sampling model also adds flexibility by allowing us to correct irregularities that may arise from initially assumed uninformative distributions of certain universal variables. In Section 4.1, we show that this is particularly the case for the uniform distribution of the year of birth variable.

In the following section, we demonstrate how the likelihood in (17) for the joint observation of age and licence status in cross-sectional data can be computed, with the goal of calibrating the parameters of the corresponding universal variables that govern these time-dependent variables.

3.4.2 Estimating parameters based on observed age and licence status in cross-sectional data

With the sampling procedure defined, we can now express individual i's contribution to the likelihood in (17) for the case of jointly observed age and driving license status in cross-sectional dataset $\mathcal{D}_t$, i.e., for $z_{i,t} = \{y_{1,i}(t), y_{2,i}(t)\}$:

$$P(s_{i,t} = 1, z_{i,t} \mid \theta) = P(s_{i,t} = 1, y_{1,i}(t), y_{2,i}(t) \mid \theta).$$

This likelihood enables a direct calibration of the parameters of the universal variables $\mathbf{u}_\theta = \{e_1, \hat{d}_1, \hat{a}_{dl}\}$ that are used to derive $z_{i,t}$, i.e.,

$$\theta = (k, \lambda, \alpha_y, \alpha_a, \alpha_o, \tau_1, \tau_2, \pi, \mu, \sigma).$$

For notational simplicity, in the following derivation of the likelihood, we omit explicitly indicating the dependence of $y_{1,i}$ and $y_{2,i}$ on time t, as well as the fact that all universal variables refer to individual i, when this is clear from the context. We start by writing:

$$P(s_{i,t} = 1, z_{i,t} \mid \theta) = \sum_{\mathbf{u}_\theta} P(s_{i,t} = 1, y_{1,i}, y_{2,i} \mid e_1, \hat{d}_1, \hat{a}_{dl}; \theta) P(e_1, \hat{d}_1, \hat{a}_{dl} \mid \theta). \tag{19}$$

Next, based on the Bayes formula, we can write (19) as follows:

$$\sum_{e_1, \hat{d}_1, \hat{a}_{dl}} P(y_{1,i}, y_{2,i} \mid s_{i,t} = 1, e_1, \hat{d}_1, \hat{a}_{dl}; \theta) P(s_{i,t} = 1 \mid e_1, \hat{d}_1, \hat{a}_{dl}; \theta) P(e_1, \hat{d}_1, \hat{a}_{dl} \mid \theta) \tag{20}$$

Under the assumption that e_1 is independent of both $\hat{d}_1$ and $\hat{a}_{dl}$, and based on our sampling model (18), the expression (20) can be further simplified:

$$\begin{aligned} &\sum_{e_1, \hat{d}_1, \hat{a}_{dl}} P(y_{1,i}, y_{2,i} \mid s_{i,t} = 1, e_1, \hat{d}_1, \hat{a}_{dl}; \theta) P(s_{i,t} = 1 \mid e_1; \theta) P(e_1) P(\hat{d}_1, \hat{a}_{dl} \mid \theta) = \\ &\sum_{e_1, \hat{d}_1, \hat{a}_{dl}} P(y_{1,i}, y_{2,i}, s_{i,t} = 1, e_1, \hat{d}_1, \hat{a}_{dl}; \theta) P(s_{i,t} = 1 \mid e_1; \theta) P(e_1) P(\hat{a}_{dl} \mid \hat{d}_1; \theta) P(\hat{d}_1 \mid \theta) \end{aligned} \tag{21}$$

As $e_1 = t - y_{1,i}$ is fixed, (21) simplifies to:

$$\begin{aligned} &\sum_{\hat{d}_1, \hat{a}_{dl}} P(y_{1,i}, y_{2,i} \mid s_{i,t} = 1, e_1 = t - y_{1,i}, \hat{d}_1, \hat{a}_{dl}; \theta) \\ &\qquad \underbrace{P(s_{i,t} = 1 \mid e_1 = t - y_{1,i}; \theta)}_{\text{sampling model}} \underbrace{P(e_1 = t - y_{1,i})}_{\text{constant}} P(\hat{a}_{dl} \mid \hat{d}_1; \theta) P(\hat{d}_1 \mid \theta) \\ &= P(s_{i,t} = 1 \mid e_1 = t - y_{1,i}; \theta) P(e_1 = t - y_{1,i}) \\ &\qquad \sum_{\hat{d}_1, \hat{a}_{dl}} P(y_{1,i}, y_{2,i} \mid s_{i,t} = 1, e_1 = t - y_{1,i}, \hat{d}_1, \hat{a}_{dl}; \theta) P(\hat{a}_{dl} \mid \hat{d}_1; \theta) P(\hat{d}_1 \mid \theta) \end{aligned} \tag{22}$$

Observe that the second term represents the previously defined sampling model from (18), while $P(e_1 = t - y_{1,i})$ is fixed to $\frac{1}{y_{max} - y_{min} + 1} = \frac{1}{100}$ based on (1). We can distinguish between two cases based on the value of $y_{2,i}$, one for $y_{2,i} = 0$ and another for $y_{2,i} = 1$.

Case $y_{2,i} = 0$:

$$P(y_{1,i}, y_{2,i} = 0 \mid s_{i,t} = 1, e_1 = t - y_{1,i}, \hat{d}_1, \hat{a}_{dl}; \theta) = \begin{cases} 1, & \hat{d}_1 \geq y_{1,i} \text{ and } \hat{a}_{dl} > y_{1,i}, \\ 0, & \text{otherwise.} \end{cases}$$

Then, after splitting the sum over $\hat{a}_{dl}$ in two parts, equation (22) becomes:

$$\begin{aligned} \sum_{\substack{\hat{d}_1 \geq y_{1,i}, \\ \hat{a}_{dl} > y_{1,i}}} P(\hat{a}_{dl} \mid \hat{d}_1; \theta) P(\hat{d}_1 \mid \theta) = &\sum_{\substack{\hat{d}_1 \geq y_{1,i}, \\ \hat{a}_{dl} = \infty}} P(\hat{a}_{dl} \mid \hat{d}_1; \theta) P(\hat{d}_1 \mid \theta) \\ &+ \sum_{\substack{\hat{d}_1 \geq y_{1,i}, \\ y_{1,i} < \hat{a}_{dl} < \infty}} P(\hat{a}_{dl} \mid \hat{d}_1; \theta) P(\hat{d}_1 \mid \theta) \end{aligned}$$

$$= \underbrace{P(\hat{a}_{dl} = \infty \mid \hat{d}_1; \theta) \sum_{\hat{d}_1 \geq y_{1,i}} P(\hat{d}_1)}_{\pi(1 - F_{\hat{d}_1}(y_{1,i} - 1))} + \sum_{\substack{\hat{d}_1 \geq y_{1,i} \\ y_{1,i} < \hat{a}_{dl} \leq \hat{d}_1}} P(\hat{a}_{dl} | \hat{d}_1; \theta) P(\hat{d}_1 \mid \theta)$$

where, in the first summation, $P(\hat{a}_{dl} \mid \hat{d}_1; \theta) = \pi$ is constant based on (4), and the summation over $\hat{d}_1$ can be reformulated using the CDF of $\hat{d}_1$. To calculate the second term, we consider two cases based on the ordering of $y_{1,i}$ and A_{min}:

1. If $A_{min} < y_{1,i} \leq \hat{d}_1$ then, $\sum_{\substack{\hat{d}_1 \geq y_{1,i} \\ y_{1,i} < \hat{a}_{dl} \leq \hat{d}_1}} P(\hat{a}_{dl} | \hat{d}_1; \theta) P(\hat{d}_1 \mid \theta)$ simplifies to:

$$\begin{aligned}
&\sum_{\substack{\hat{d}_1 \geq y_{1,i} \\ y_{1,i} < \hat{a}_{dl} \leq \hat{d}_1}} (1 - \pi) \left[F_\Delta(\hat{a}_{dl} + 0.5; \hat{d}_1) - F_\Delta(\hat{a}_{dl} - 0.5; \hat{d}_1) \right] P(\hat{d}_1 \mid \theta) \\
&= (1 - \pi) \sum_{\hat{d}_1 \geq y_{1,i}} P(\hat{d}_1 \mid \theta) \sum_{y_{1,i} < \hat{a}_{dl} \leq \hat{d}_1} \left[F_\Delta(\hat{a}_{dl} + 0.5; \hat{d}_1) - F_\Delta(\hat{a}_{dl} - 0.5; \hat{d}_1) \right] \\
&= (1 - \pi) \sum_{\hat{d}_1 \geq y_{1,i}} P(\hat{d}_1 \mid \theta) \left[F_\Delta(\hat{d}_1 + 0.5; \hat{d}_1) - F_\Delta(y_{1,i} + 0.5; \hat{d}_1) \right] \qquad (23)
\end{aligned}$$

   where, in the first line, we substitute $P(\hat{a}_{dl} \mid \hat{d}_1; \theta)$ from (5), and the last line follows from simplifying the telescoping sum over $\hat{a}_{dl}$.

2. If $y_{1,i} \leq A_{min} \leq \hat{d}_1$ then, $\sum_{\substack{\hat{d}_1 \geq y_{1,i} \\ y_{1,i} < \hat{a}_{dl} \leq \hat{d}_1}} P(\hat{a}_{dl} \mid \hat{d}_1; \theta) P(\hat{d}_1 \mid \theta)$ simplifies to:

$$\begin{aligned}
&\underbrace{\sum_{\substack{\hat{d}_1 \geq y_{1,i} \\ y_{1,i} < \hat{a}_{dl} \leq A_{min}}} P(\hat{a}_{dl} \mid \hat{d}_1; \theta) P(\hat{d}_1 \mid \theta)}_{:=0} + \sum_{\substack{\hat{d}_1 \geq y_{1,i} \\ A_{min} \leq \hat{a}_{dl} \leq \hat{d}_1}} P(\hat{a}_{dl} \mid \hat{d}_1; \theta) P(\hat{d}_1 \mid \theta) \\
&= \sum_{\hat{d}_1 \geq y_{1,i}} P(\hat{d}_1 \mid \theta) \underbrace{\sum_{A_{min} \leq \hat{a}_{dl} \leq \hat{d}_1} P(\hat{a}_{dl} \mid \hat{d}_1; \theta)}_{:=1} = (1 - \pi)(1 - F_{\hat{d}_1}(y_{1,i} - 1))
\end{aligned}$$

   where, in the first sum, we have $P(\hat{a}_{dl} \mid \hat{d}_1; \theta) = 0$ based on (5), and the sum over $\hat{a}_{dl}$ is equal to 1 by definition of $\hat{a}_{dl}$'s support.

Case $y_{2,i} = 1$:

$$P(y_{1,i}, y_{2,i} = 1 \mid s_{i,t} = 1,\, e_1 = t - y_{1,i}, \hat{d}_1, \hat{a}_{dl}; \theta) = \begin{cases} 1, & \hat{d}_1 \geq y_{1,i} \text{ and } \hat{a}_{dl} \leq y_{1,i}, \\ 0, & \text{otherwise.} \end{cases}$$

Then, the sum in equation (22) becomes:

$$\sum_{\substack{\hat{d}_1 \geq y_{1,i} \\ \hat{a}_{dl} \leq y_{1,i}}} P(\hat{a}_{dl} \mid \hat{d}_1; \theta) P(\hat{d}_1 \mid \theta)$$

which simplifies to:

$$\begin{aligned}
&\sum_{\hat{d}_1 \geq y_{1,i}} P(\hat{d}_1 \mid \theta) \sum_{A_{\min} \leq \hat{a}_{dl} \leq y_{1,i}} (1-\pi) \left[F_\Delta(\hat{a}_{dl} + 0.5; \hat{d}_1) - F_\Delta(\hat{a}_{dl} - 0.5; \hat{d}_1) \right] \\
&= (1-\pi) \sum_{\hat{d}_1 \geq y_{1,i}} P(\hat{d}_1 \mid \theta) \sum_{A_{\min} \leq \hat{a}_{dl} \leq y_{1,i}} \left[F_\Delta(\hat{a}_{dl} + 0.5; \hat{d}_1) - F_\Delta(\hat{a}_{dl} - 0.5; \hat{d}_1) \right] \\
&= (1-\pi) \sum_{\hat{d}_1 \geq y_{1,i}} P(\hat{d}_1 \mid \theta) [F_\Delta(y_{1,i} + 0.5; \hat{d}_1) - F_\Delta(A_{\min} - 0.5; \hat{d}_1)] \qquad (24)
\end{aligned}$$

where, in the first line, we substitute $P(\hat{a}_{dl} \mid \hat{d}_1; \theta)$ from (5), and the last line is again obtained by canceling out terms in the telescoping sum over $\hat{a}_{dl}$.

The parameters θ are estimated using Maximum Likelihood Estimation (MLE). To evaluate the infinite sums in (23) and (24), we approximate them by truncating at $\hat{d}_1 = 130$, as $P(\hat{d}_1 \mid \theta)$ becomes negligible beyond this age. To identify the optimal parameter values, we employ differential evolution, a population-based global optimization algorithm well suited to non-convex problems. This procedure yields the parameter set that maximizes the log-likelihood given the empirical data.

4 Results

In this section, we compare data-free and data-integrated generation processes. To this end, we generate a universal dataset of 100,000 individual life sequences following the sampling procedure described in Section 3.3. In the data-free approach, we use the assumed distributions and parameter values defined in Section 3.1. In contrast, the data-integrated approach estimates the parameters via MLE using one or multiple datasets and then constructs the corresponding distributions based on the estimated values.

With that in mind, in Section 4.1, we first examine how different models, i.e., a data-free model, a model with parameters estimated using a single dataset, and a model with parameters estimated using multiple datasets, affect the generation of synthetic panels at both aggregate and individual levels. In these experiments, we focus specifically on generating year of birth and lifespan, and on how these variables shape the time-dependent age distribution. Subsequently, in Sections 4.2 and 4.3, we extend the analysis by comparing the data-free model with the model whose parameters are estimated using all cross-sectional datasets, focusing on their impact on the generation of the universal dataset as well as on the resulting aggregate and disaggregate properties of the complete set of time-dependent panel distributions. In Section 4.4, we demonstrate the framework's capability to test hypothetical scenarios by designing a pandemic case study and showing how changes in universal variables automatically propagate to the derived, time-dependent datasets. Finally, as an illustrative example, Appendix A presents the step-by-step development of a model for generating year of birth and lifespan, along with its impact on the resulting age distribution.

4.1 Comparison of data-free and data-integrated generation processes

In this section, we demonstrate both the data-free generation of synthetic panel data and the capability of the proposed framework for data integration. By data integration, we refer to the use of real cross-sectional datasets to estimate the parameters of the universal variables, which are then employed to derive corresponding synthetic time-dependent panel data. To evaluate the effect of integrating different amounts of information, we compare four models that differ in their level of data integration:

- **Model 1 (data-free baseline):** Parameters are fixed to values reported in the literature, without using empirical data for calibration.
- **Model 2 (one-dataset integration):** Parameters are estimated using the

2010 cross-sectional dataset, while the 2015 and 2021 datasets are used for validation.

- **Model 3 (two-dataset integration):** Parameters are estimated jointly using the 2010 and 2015 datasets, with 2021 kept for validation.
- **Model 4 (all-dataset integration):** Parameters are estimated using all three datasets (2010, 2015, and 2021), leaving no data for external validation.

Through this comparison, we aim to assess how the progressive integration of additional data sources affects both the estimated parameters of the universal variables and the resulting synthetic panel distributions. For the likelihood derived in Section 3.4.2, we estimate the parameter vector

$$\theta = (k, \lambda, \alpha_y, \alpha_a, \alpha_o, \tau_1, \tau_2, \pi, \mu, \sigma),$$

which governs the generation of the year of birth, lifespan, and the age of driving licence acquisition. The parameters are updated through maximum likelihood estimation (MLE) when data are integrated, while the remaining model components retain their literature-based default values, following the procedure described in Section 3.4.

Parameter	Meaning	Model 1	Model 2	Model 3	Model 4
k	Shape (lifespan)	3.00	2.84	3.44	3.37
λ	Scale (lifespan)	85.00	69.28	77.36	74.41
α_y	Weight of young group	–	0.02	0.16	0.06
α_a	Weight of adult group	–	0.95	0.79	0.84
α_o	Weight of old group	–	0.04	0.05	0.10
τ_1	Young–adult threshold	–	33.98	38.06	37.73
τ_2	Adult–old threshold	–	84.89	82.50	87.76
π	Prob. no licence	0.15	0.17	0.19	0.21
μ	Log-mean (licence age)	3.02	2.89	2.87	2.90
σ	Log-sd (licence age)	0.15	0.13	0.11	0.11

Table 10: Parameter values and meaning for the four models

Table 10 summarizes the parameter values obtained for the four tested models. Model 1 does not include any parameters related to the sampling function, as it represents the data-free baseline. In this case, the year of birth is drawn uniformly from $y_{min} = 1920$ to $y_{max} = 2050$, and lifespans are generated from a Weibull distribution. The resulting derived age distribution is therefore determined solely by the survival model and the uniform distribution of birth years. In

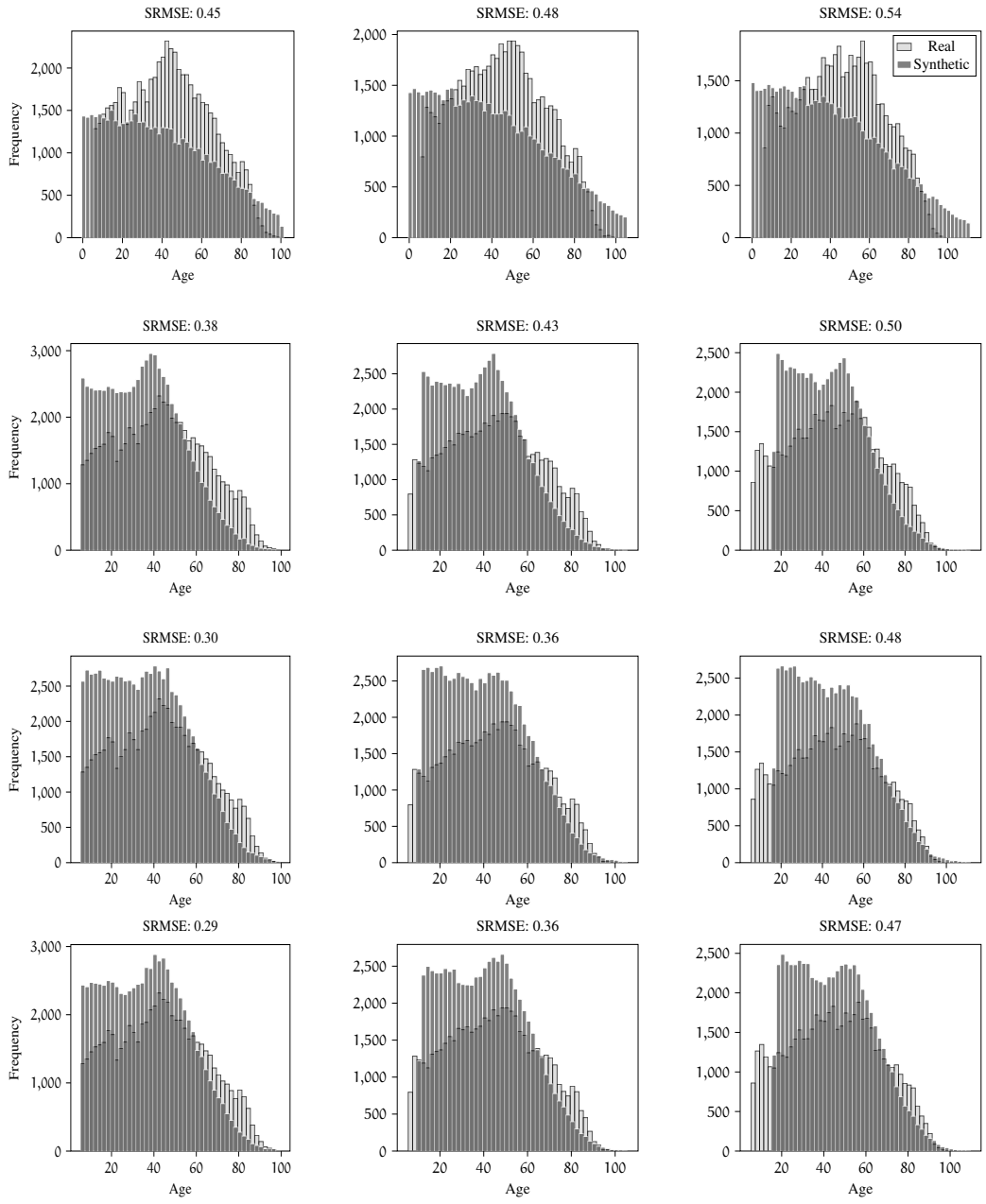
contrast, in Models 2–4, the birth-year distribution is additionally influenced by the sampling mechanism introduced in Section 3.4.1. If we were interested only in a single point in time, we could simply sample from the empirical age distribution. However, to enable the universal birth-year distribution to capture multiple observed age distributions, we have to account for its mapping to cross-sectional data through the sampling mechanism.

The parameter estimates across Models 2–4 appear to differ only slightly, likely because the datasets used for estimation share a similar structure and describe the same underlying population. This consistency demonstrates that data integration is robust across datasets and yields parameter values close to those reported in the demographic literature. In contexts where demographic parameters are well documented, such as Switzerland, literature values can be used directly. Where such information is unavailable, the results suggest that the proposed procedure can infer these parameters from any available data.

Four universal datasets, one for each model, were generated, from which we derived synthetic panel data for $t = 2010, 2015$, and 2021, as shown in Figure 9. When the universal dataset is generated under a uniform birth-year distribution, no constraint is imposed on individuals' presence at a specific point in time. As a result, the dataset may include individuals who are alive, already deceased, or not yet born in a given reference year. As we can see in Figure 9, for the data-free model, newborns appear in each snapshot, reflecting the continuous inflow of new individuals over time. When the year of birth is generated from a uniform distribution, sampling probabilities are not incorporated, leading to a differently shaped age distribution compared to other models.

On the other hand, when the universal distribution is calibrated using cross-sectional data, the sampling mechanism better aligns the birth-year distribution with the observed age structure at time t. This means that only birth years feasible for that period are generated, and no individuals are born after the reference year. However, since lifespans are sampled independently according to the assumed survival model, some of these individuals may already be deceased by time t. As a result, the calibrated universal dataset contains only individuals who could have existed at that moment, without newborns appearing in later snapshots.

For Models 2–4, when the panel is projected forward to 2015 and 2021, the same individuals are retained and simply age over time. Consequently, the shape of the resulting distribution reflects the model formulation: the Weibull survival function emphasizes younger ages, while the sampling function redistributes probability across young, adult, and old groups. Over time, this distribution shape changes due to mortality.



(a) 2010 (b) 2015 (c) 2021

Figure 9: Comparison between the derived age panel distribution and the observed age distribution over time.

To evaluate the quality of the aggregated properties of the synthetic panels, we compare the real and synthetic age distributions based on their normalized

frequencies (i.e., probabilities across age bins). Let $p^{(t)}_{real,j}$ and $p^{(t)}_{synth,j}$ denote the relative frequencies of age bin j in year t for the real and synthetic datasets.

To obtain an overall measure of fit across multiple periods, we concatenate the normalized yearly distributions into a single vector and compute the joint standardized root mean squared error (joint SRMSE):

$$\text{SRMSE}_{\text{joint}} = \frac{\sqrt{\frac{1}{N}\sum_{t\in\mathcal{T}}\sum_{j=1}^{n_t}\left(p^{(t)}_{synth,j} - p^{(t)}_{real,j}\right)^2}}{\overline{p}_{real,all}},$$

where n_t is the number of bins in year t, $N = \sum_{t\in T} n_t$, and $\overline{p}_{real,all}$ denotes the average bin probability across all years (equal to $1/n_t$ when all distributions are normalized and share the same bin structure).

When the evaluation concerns a single period ($|\mathcal{T}| = 1$), the joint SRMSE reduces to the independent SRMSE:

$$\text{SRMSE}_t = \frac{\sqrt{\frac{1}{n_t}\sum_{j=1}^{n_t}\left(p^{(t)}_{synth,j} - p^{(t)}_{real,j}\right)^2}}{\overline{p}^{(t)}_{real}}.$$

The independent SRMSE therefore measures the fit for a single year, while the joint SRMSE generalizes this metric to capture the overall agreement between real and synthetic distributions across multiple years.

Model	Calibration data	Unseen data		
		2010	2015	2021
Model 1	x	**0.45**	**0.48**	**0.54**
Model 2	0.38	(0.38)	**0.43**	**0.50**
Model 3	0.40	(0.30)	(0.36)	**0.48**
Model 4	0.39	(0.47)	(0.39)	(0.39)

Table 11: Model evaluation through SRMSE

Table 11 summarizes the results of generating time-dependent age distributions under different model specifications, evaluated using the joint SRMSE on the estimated data and the independent SRMSE on unseen data. SRMSE values reported in parentheses correspond to the independent SRMSE calculated for the calibration year, providing a reference for model fit within the estimation period. From these results, several observations can be made:

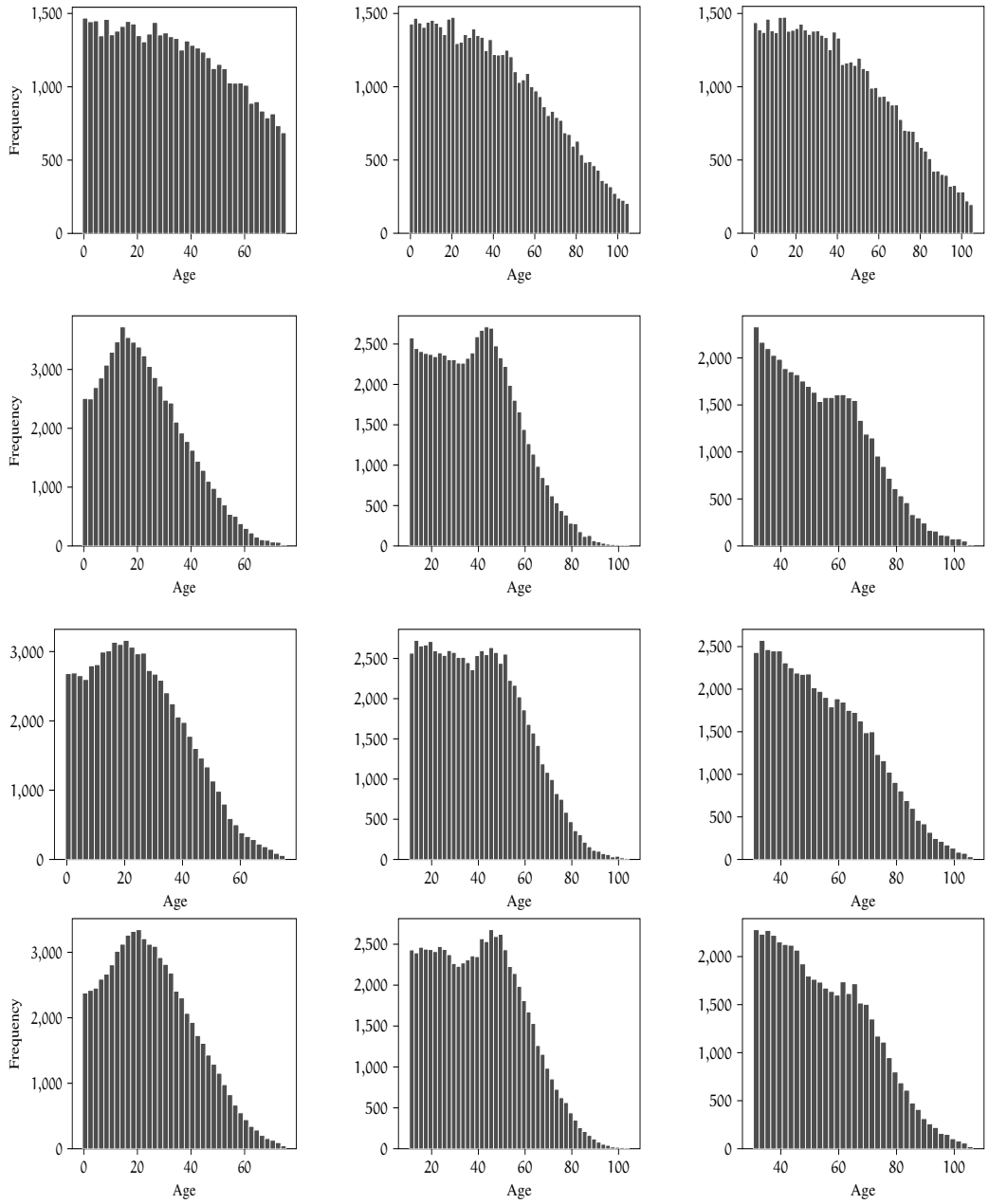
- Based on the performance on the calibration data, Model 2 appears to achieve the best fit. However, its low error likely indicates overfitting, as

its performance on the unseen datasets is worse than that of the other models. When looking at Models 3 and 4, we observe that incorporating more data generally improves the fit.

- The independent SRMSE on unseen data provides additional insight. Examining the results across rows, Model 1 consistently exhibits the weakest fit in all years, as expected, since it is not calibrated using observed data. Looking down the columns, SRMSE values generally decrease as more data are incorporated into model estimation, indicating that additional information improves generalization. Across all models, however, the fit deteriorates as the projection horizon extends further into the future.

- Overall, the results show a clear incremental development: starting from a data-free baseline with a weak fit (Model 1), through single-dataset calibration (Model 2), to joint calibration on multiple datasets (Model 3), each step reduces error and increases robustness, confirming that integrating more information leads to a better fit across time.

Additionally, to demonstrate that the method can generate synthetic panels for years outside of the observed range, Figure 10 displays age distributions of living individuals for 1985 and 2035, alongside the reference year 2015, across Models 1–4. As expected, the backward projection to 1985 yields more young individuals, while the forward projection to 2035 shifts the distribution toward older ages, with its shape evolving over time due to aging and rising mortality.



(a) 1985 (b) 2015 (c) 2035

Figure 10: Synthetic panel 1985, 2015, 2035

4.2 Derivation of time-dependent panel synthetic samples from universal variables: aggregated level

In this section, we generate a universal dataset using two approaches: (i) a data-free approach with fixed parameters and (ii) a data-integrated approach with parameters estimated from three cross-sectional datasets. From these universal datasets, we derive time-dependent data for $\mathrm{t} = 2010$ and compare it with real data from the same period.

The objectives of this experiment are threefold: (i) to assess whether the model can reproduce the marginal and conditional distributions of the real data (i.e., aggregated properties) at a specific point in time, (ii) to evaluate the validity of the assumed universal variable models, and (iii) to examine how simplifications at the universal level affect the resulting time-dependent distributions. Our analysis focuses on the discrete distributions of age, driving licence ownership, employment status, and education level.

First, using the fixed parameters introduced in Section 3.1, we generate the universal dataset and derive the age, driving licence, employment, and education status for $\mathrm{t} = 2010$, as shown in Figure 11. In our model, most universal variables are derived from the year of birth. Consequently, any discrepancy in the age distribution propagates to other variables. In this case, the uniform prior for the year of birth results in an unrealistically flat age distribution. As a result, for example, there is an excessive share of individuals under 15 in both education and employment categories.

Next, we generate a universal dataset using the estimated parameters for Model 4 in Table 10. We start by generating the year of birth and driving licence using the previously obtained estimates, while all other universal variables (i.e., education and employment phases) remain based on the assumed parameters. As shown in Figure 12, the synthetic age distribution now follows the empirical shape more closely. However, the model still underestimates the number of older individuals and overestimates younger ones. This could stem from simulating lifespans using a Weibull distribution, which assigns higher survival probabilities to younger ages. Moreover, since the lifespan and year of birth are generated independently, the synthetic population includes fewer elderly individuals than in the observed data. The improvement in the age distribution, particularly the more realistic share of adults, directly translates into better alignment of the education and employment structures with the observed data. Nevertheless, some discrepancies remain: the underrepresentation of older adults results in fewer individuals classified as retired, while the excess of youth increases the shares in the "under 15" and "in education" categories. In contrast, the "employed" and "unemployed" groups are generated directly from the employment model and align well with the observed data, suggesting that the employment-phase generator performs as in-

tended. Overall, these results show that even with simple models, it is possible to reproduce the main demographic trends. However, achieving a closer fit would require more advanced demographic formulations.

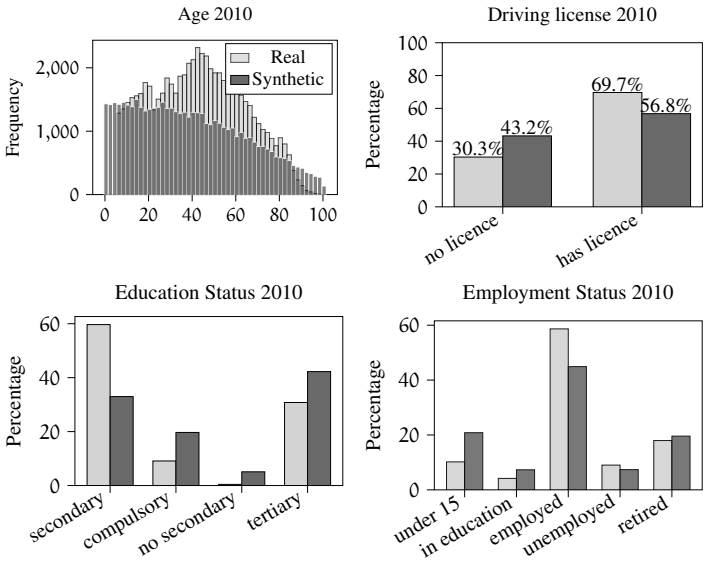


Figure 11: Comparison of marginal distributions of synthetic time-dependent variables derived from simple models and real data in 2010 - data-free approach

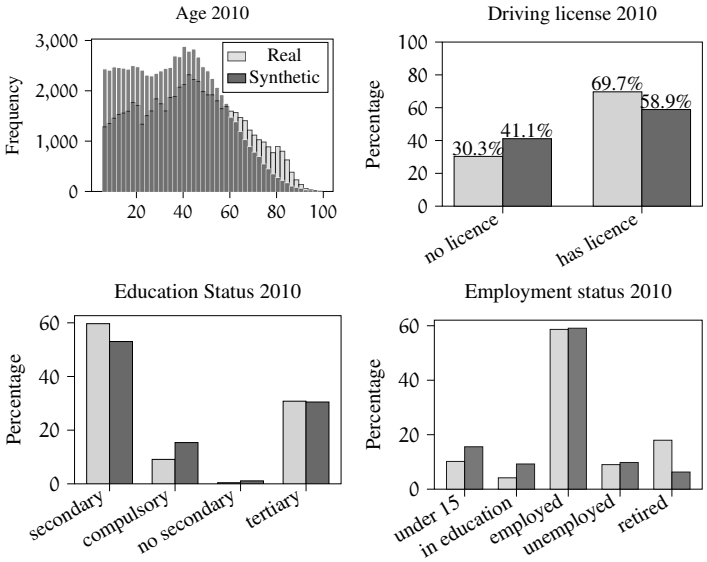
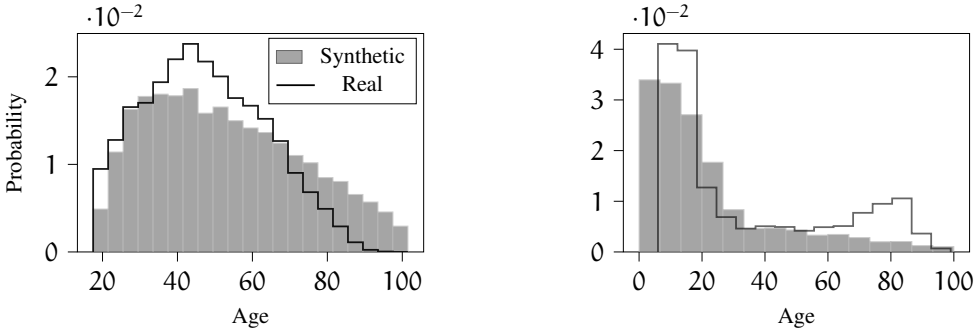


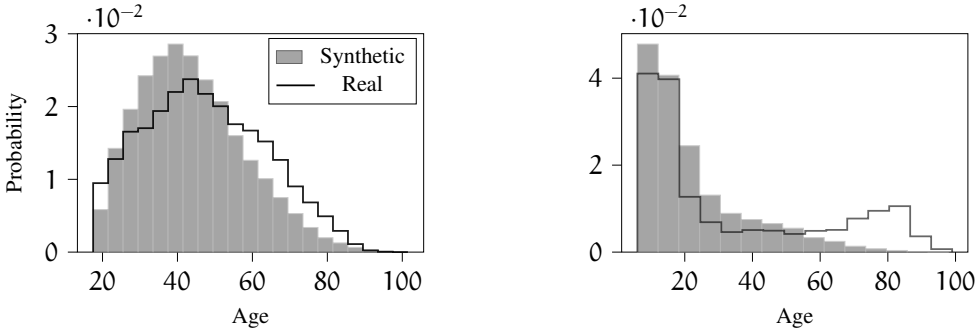
Figure 12: Comparison of marginal distributions of synthetic time-dependent variables derived from simple models and real data in 2010 - data-integrated approach

Since we also estimate parameters related to driving licence ownership, we further examine the performance in this category. Figures 13 and 14 show the conditional probability of age given licence possession for both data-free and data-integrated approaches. These plots provide a clearer view of how distortions in the age distribution propagate to licence status. For instance, in Figure 14, we observe that the sample includes too many individuals without a licence and too few with one. This discrepancy arises because licence status is generated conditional on the simulated age, which already underestimates the proportion of older individuals and overestimates that of younger ones.



(a) People having a driving license (b) People without a driving license

Figure 13: Comparison of real and synthetic age distribution of people having and not having a driving licence in 2010 - data-free approach



(a) People having a driving license (b) People without a driving license

Figure 14: Comparison of real and synthetic age distribution of people having and not having a driving licence in 2010 - data-integrated approach

However, this behavior may also partly stem from the simplified model used to generate e_2 and d_2 in Section 3.1.2. As we see in Figure 14, after the legal threshold (i.e., $A_{\min} = 18$), the model produces a decline in the share of non-holders. This is because it relies solely on the parameter π, which determines the overall fraction of individuals who never obtain a licence but does not account for age or socio-demographic factors. Moreover, the assumption that licences are irrevocable might prevent the model from generating additional non-holders at older ages, further reinforcing the misfit among the elderly.

To further investigate this, we additionally calculate the license possession rate by age group, as shown in Figure 15. The license possession rate is defined as the proportion of individuals with a driving license within each age group, relative to the total number of individuals in that group. This measure is important because it normalizes license holding by age, allowing us to assess how likely it is for someone of a given age to possess a license. Consistent with the previous results, the synthetic data shows too few licence holders among young and middle-aged adults (i.e., an excess of non-holders in Figure 14) and too many licence holders among the elderly (i.e., too few elderly non-holders in Figure 14), confirming the same deviations observed in the separate "with" and "without license" distributions.

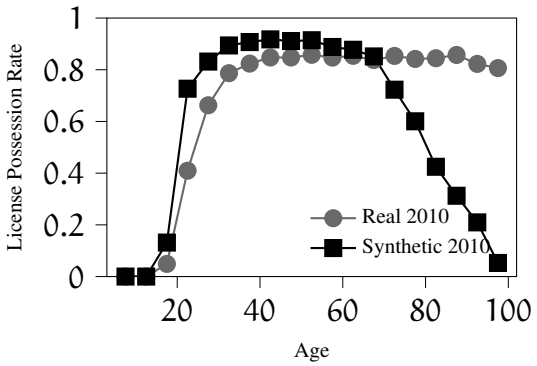
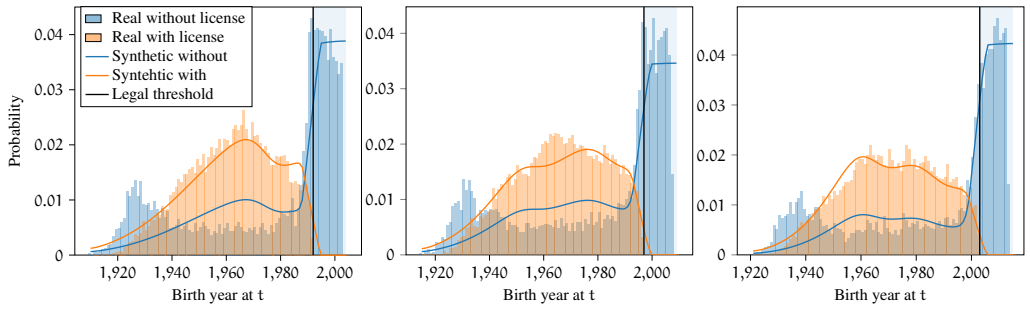


Figure 15: License possession rate by age group

To assess whether this behaviour persists using different datasets, Figure 16 presents the conditional distribution of birth year at time t, given driving licence possession, for 2010, 2015, and 2021. These distributions are generated independently using the estimated parameters for the corresponding year. The figure shows that the systematic deviations observed for 2010 reappear in later years. In other words, the synthetic population consistently underrepresents older individuals without a licence and overrepresents younger cohorts, indicating that the mismatch stems from structural limitations of the model rather than from year-specific effects. These results suggest that a model relying solely on literature values can approximate the overall distributions reasonably well. However, its



(a) 2010 (b) 2015 (c) 2021

Figure 16: Conditional distribution of birth year at time t, given driving licence possession, generated for 2010, 2015, and 2021 using estimated parameters

simplifying assumptions lead to noticeable discrepancies in the marginal distributions of certain variables, indicating the need for more refined demographic models to better capture the underlying relationships.

4.3 Derivation of time-dependent panel synthetic samples from universal variables: disaggregated level

In this section, we illustrate (i) the panel effect of the synthetic time-dependent dataset and its evolution over time, and (ii) the relevance of the correlations integrated into the model (e.g., between income and education). To this end, we select individuals born in 1985 from the universal dataset and derive time-dependent data for 2005, 2015, 2025, and 2035. To highlight the panel effect, we focus on income and spatial evolution, meaning that all figures track the same individuals over time. Note that we do not have access to real panel data, therefore, comparison with observed data is not possible in this case.

Figure 17 shows how income evolves for employed individuals and how it varies across education levels. The simulated incomes follow a log-normal distribution that is strictly positive, right-skewed, and age-dependent. To avoid unrealistic outliers, monthly incomes above CHF 20,000 are truncated.

In 2005 (age 20), most individuals with tertiary education were still studying and not yet active in the labor market, while employed individuals were predominantly secondary graduates with entry-level earnings. By 2015, as tertiary-educated individuals entered employment, their earnings surpassed those of people with secondary or no education, shifting the overall income distribution upward. At age 40, incomes continued to rise, and the gap between education groups widened, with tertiary graduates concentrated at the upper end of the distribution. By age 50, some individuals had left the workforce due to retirement or mortality,

the mean income stabilized, and education-related differences remained evident.

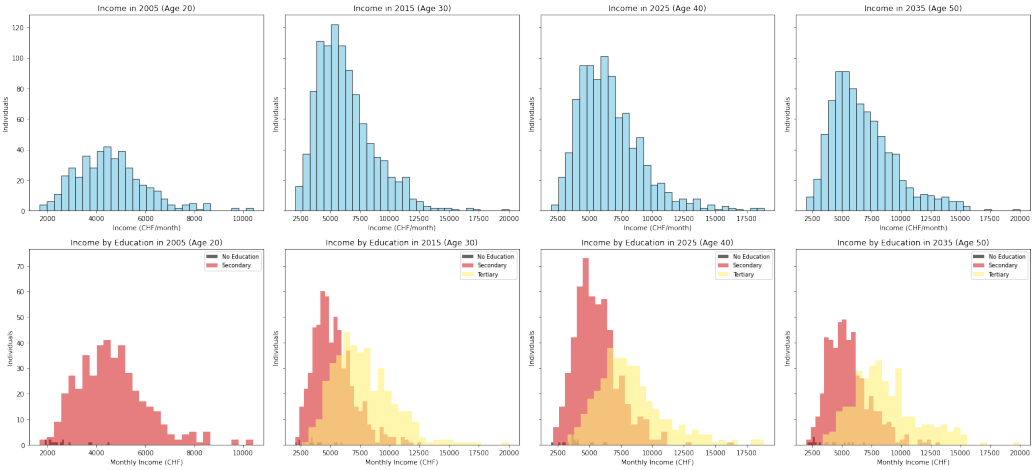


Figure 17: Tracking the income and education distributions of individuals born in 1985 over the years 2005, 2015, 2025, and 2035

To illustrate the dynamics of spatial mobility, we track the home distribution of the same 1985 birth cohort at four points in time: 2005, 2015, 2025, and 2035. At birth, the spatial distribution of this cohort reflects only birthplace sampling, which follows cantonal population shares (Figure 18). We then derive the spatial shares of these individuals for each subsequent year, as shown in Figure 19.

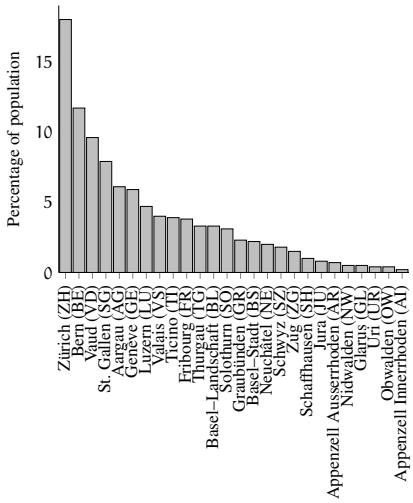


Figure 18: The cantonal home distribution of people born in 1985

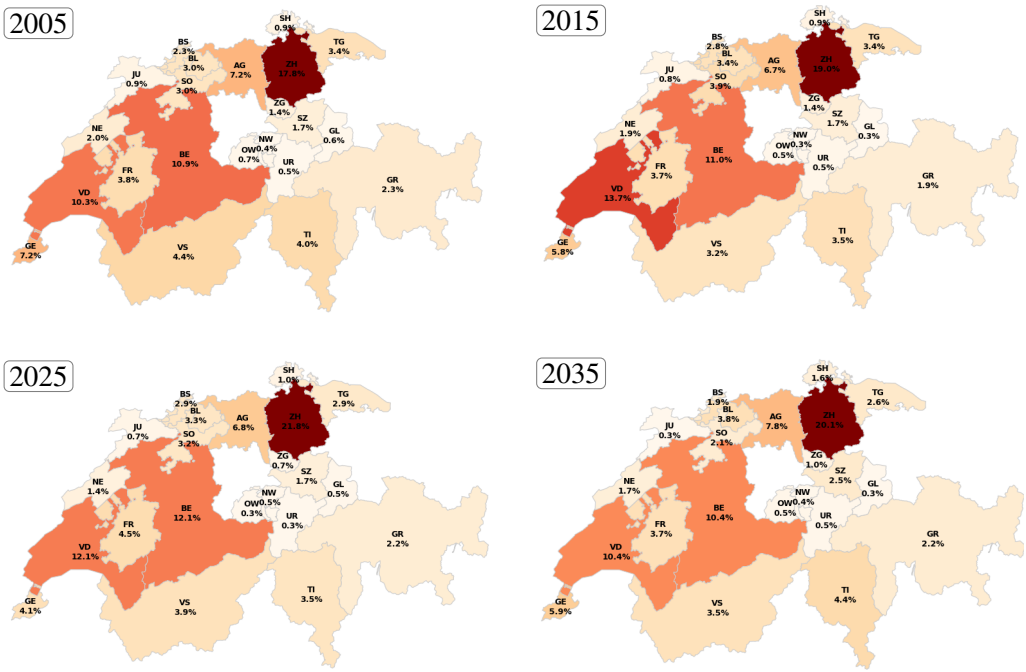
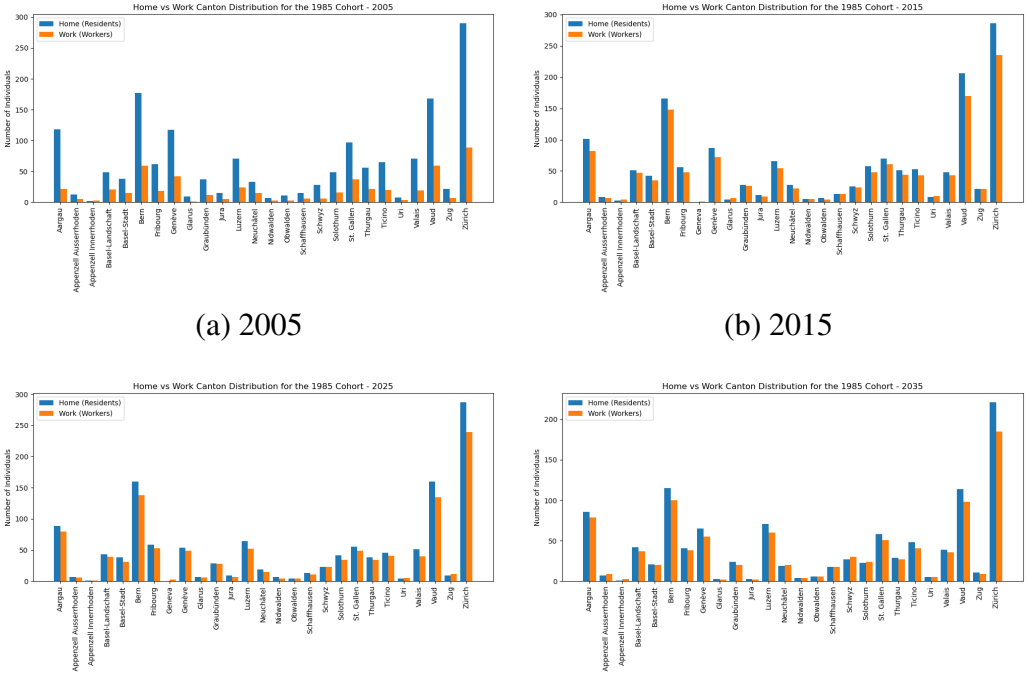


Figure 19: Swiss canton shares across years

In 2005 (age 20), most individuals still live in their birth canton. By 2015, as more people entered the labor market, the first signs of concentration appeared in metropolitan cantons, particularly Zürich. This pattern reflects the income-dependent urban boost, which increases the likelihood that higher earners relocate to major economic centers. By 2025, clustering in Zürich (over 21% of the cohort) and Bern (around 12%) becomes more pronounced, revealing the cumulative effect of repeated employment-related moves. Other cantons, such as Aargau, continue to retain notable shares, while the relative weight of smaller cantons declines. By 2035, mobility slows as individuals approach the later stages of their careers. These results confirm that the model effectively captures the income boost mechanism, whereby higher earners increasingly concentrate in urban cantons over time, while the rest of the population remains distributed according to baseline demographic patterns.

Additionally, to illustrate the relationship between home and work locations, Figure 20 presents the cantonal distributions of residents and workers over time for the same cohort. The observed evolution aligns with the expected modeling dynamics. At the age of 20, most individuals still reside in their birth canton, as many are not yet employed, resulting in a lower frequency of work locations, particularly in smaller cantons. By the age of 30, employment becomes more widespread, especially in larger cantons, and most individuals live where they

work. By the age of 40, the gap between home and work locations narrows further, while urban cantons account for a larger share of both residence and employment. Conversely, some smaller cantons (e.g., Jura) lose relative weight, reflecting mid-career migration toward economic centers. By the age of 50, the home and work distributions remain consistent, reflecting the model assumption that 95% of individuals live and work in the same canton, while the remaining 5% are employed in neighbouring cantons.



(a) 2005 (b) 2015

(c) 2025 (d) 2035

Figure 20: Home and work canton distribution for the 1985 cohort over time

4.4 Scenario testing

In this section, we want to demonstrate how changes can be applied to the universal dataset and reflected in all derived datasets, which opens the door for testing the impact of hypothetical scenarios in both short-and long-term simulations. In Figure 21, we illustrate the steps of the performed case study.

First, we generate a universal dataset using Model 4, introduced in Section 4.1. Next, we simulate a hypothetical 2025 pandemic on this dataset, targeting older individuals. Specifically, we randomly select people aged over 50 and apply a 70% mortality rate. To assess the impact, we compare two scenarios: (i) normal,

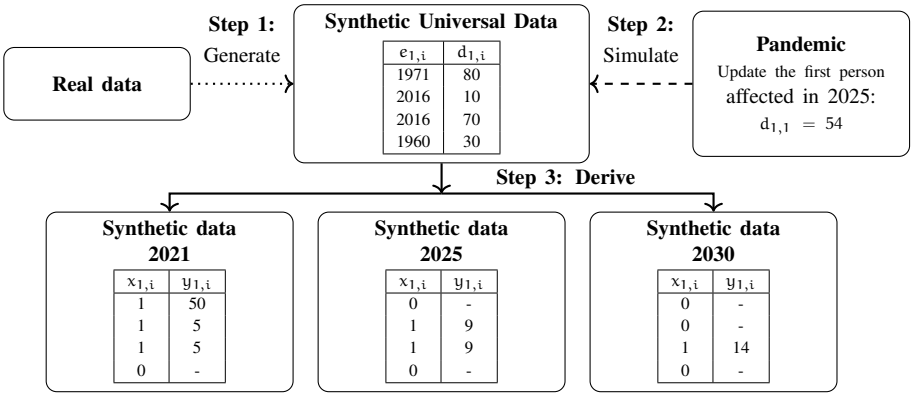
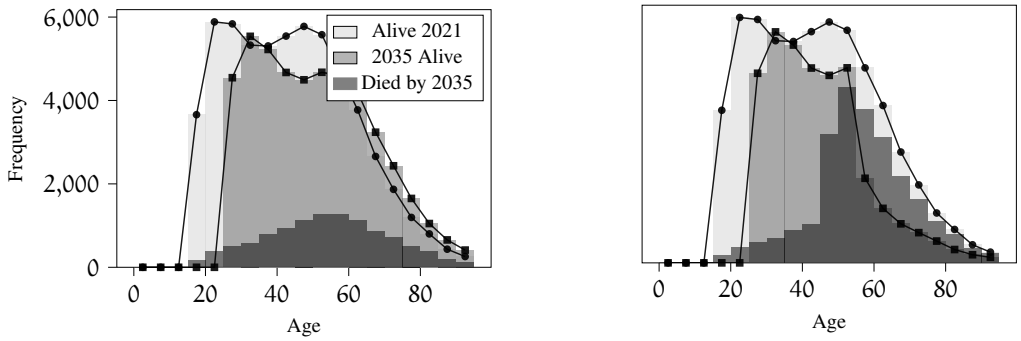


Figure 21: Hypothetical scenario setting

where we derive synthetic panels for $t = 2021$ and $t = 2035$ directly from the original universal dataset, and (ii) hypothetical, where the same panels are derived after applying the pandemic simulation. By contrasting these two scenarios, we aim to evaluate whether we can capture a disruptive demographic event between 2021 and 2035. In Figure 22, we compare the normal and disaster scenarios. The



(a) Before pandemic

(b) After pandemic

Figure 22: Age distribution of synthetic panels from 2021 and 2030

normal scenario shows the expected age distribution shift due to natural aging and mortality. In the disaster scenario, the 2021 sample remains unchanged, while by 2035 a larger share of older individuals has died as a result of the simulated pandemic. Comparing these two time-dependent datasets allows us to observe the impact of the demographic event that occurred between them.

Figure 23 illustrates the setup for testing how the choice of time step s affects the visibility of pandemic effects relative to the year t in which the pandemic occurred. By comparing death rates at $t - s$ and $t + s$ (see Table 12), we analyze

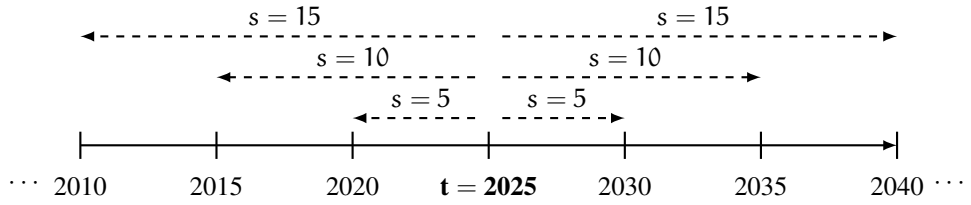


Figure 23: The effect of time step s on identifying pandemic impacts around t

the extent to which the disaster's impact can be identified over varying temporal distances. We calculate the death rate for both scenarios as the difference between the death percentage at $t + s$ and $t - s$, divided by the time step s. Since no pandemic has occurred before t, the death percentage at $t - s$ is the same for both scenarios. The disaster becomes evident through the highest number in the death rate for smaller time steps (e.g., $s = 5$), with the death rate in the disaster scenario being 2.06 times higher than in the normal scenario. For larger steps (e.g., $s \geq 25$), the natural rise in deaths hides short-term effects, making the disaster harder to detect.

Time Step s	**Death % at** $t - s$	**Death % at** $t + s$ **Normal**	**Death % at** $t + s$ **Disaster**	**Death Rate Normal** (DR_n)	**Death Rate Disaster** (DR_p)	$\frac{DR_p}{DR_n}$
5	38.99	51.09	63.93	2.42	4.99	2.06
10	33.12	57.24	67.55	2.41	3.44	1.43
15	27.59	63.15	71.22	2.37	2.91	1.23
20	22.73	68.82	75.00	2.30	2.61	1.13
25	18.30	74.05	78.66	2.23	2.41	1.08
30	14.36	78.84	82.17	2.15	2.26	1.05
35	11.01	83.13	85.45	2.06	2.13	1.03
40	8.28	86.81	88.38	1.96	2.00	1.02
45	6.02	89.94	90.95	1.86	1.89	1.01

Table 12: Comparison of cumulative death percentages and death rates for $t = 2025$ for different time steps in normal and disaster scenarios.

This example serves as a foundation for extending the framework to other types of simulations, such as changes in legislation, wars, or migration crises.

5 Conclusion

This paper introduced a framework for generating synthetic panel data based on the event–duration approach in the absence of real panel data. The central idea is to generate a single set of universal variables from which time-dependent variables can be derived at any point without recalibration. This provides (i) disaggregated longitudinal information about the same individual, offering richer insights than aggregated sociodemographic marginals alone, (ii) internal consistency across time by relying on one set of universal variables, (iii) flexibility, as changes to the universal dataset are automatically reflected in all derived datasets, and (iv) efficiency, since time-dependent data are derived directly rather than regenerated.

This work introduces a framework that combines a data-free generative approach with the adaptation of synthetic panel data to observed cross-sectional information. The proposed methodology enables the generation of synthetic panel datasets that exploit available cross-sectional observations while remaining adaptable to diverse application contexts. The version presented here is deliberately simple and illustrated through examples designed to clarify the essential building blocks of the approach.

Future developments may refine the methodology by extending the framework from individuals to households, or by adopting Bayesian approaches that allow continuous updating as new data become available. Potential applications include scenario testing through the integration of multiple datasets, improvements in demographic modeling, additional parameterizations of behavioral models, and the use of alternative estimation procedures. These avenues highlight the flexibility and extensibility of the framework, and its potential to evolve in step with advances in data availability, statistical methods, and application needs.

A From Survival Probability to Sampling-Based Likelihood Integration

We initially explored whether the empirical age distribution observed in the cross-sectional data could be reproduced using only the survival component of the model, i.e., by combining the distributions of birth year e_1 and lifespan d_1. The idea was that, if lifespans follow a realistic distribution and births are uniformly distributed across years, then survival probability alone might capture the overall age structure.

Formally, we define the probability that an individual is alive at time $t = 2010$, given their year of birth $e_1 = y$, as

$$P(x_{1,2010} = 1 \mid e_1 = y) = P(d_1 \geq y_{1,2010}) = 1 - F_{d_1}(y_{1,2010}), \tag{25}$$

where $x_{1,2010} = 1$ denotes that the individual is alive in 2010, d_1 represents the lifespan, and $y_{1,2010} = 2010 - e_1$ is the age of the individual in 2010. The function $F_{d_1}(\cdot)$ is the cumulative distribution function (CDF) of the Weibull distribution assumed for lifespan:

$$F_{d_1}(y_{1,2010}) = 1 - \exp\left[-\left(\frac{y_{1,2010}}{\lambda}\right)^k\right], \tag{26}$$

where $k > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter of the Weibull distribution.

We then performed a sensitivity analysis by testing various combinations of the parameters k and λ to identify those that best fit the real data. To evaluate the fit, we use two metrics: Mean Squared Error (MSE) and Kullback-Leibler (KL) divergence. MSE measures the average squared difference between the real and synthetic distributions, capturing overall differences in shape and magnitude. KL divergence quantifies the information loss when the synthetic distribution approximates the real one, placing greater weight on regions with high probability mass in the empirical data. Since KL divergence is not symmetric, that is, $D_{KL}(P \,\|\, Q) \neq D_{KL}(Q \,\|\, P)$, it is essential to define the direction of comparison. When P is the real distribution (i.e., observed data) and Q the synthetic one (i.e., given by the model), we use the following expression:

$$D_{KL}(P \,\|\, Q) = \sum_i p_i \cdot \log\left(\frac{p_i}{q_i}\right)$$

We tested these metrics with a random set of parameters k and λ for the Weibull distribution using datasets from 2010, 2015, 2021. The results of the comparison are presented in Table 13 and Figure 24.

Year	Shape (k)	Scale (λ)	MSE	KL Divergence
2010	3.0	80	7.4×10^{-5}	9.76×10^{-2}
2010	3.0	85	7.3×10^{-5}	1.02×10^{-1}
2010	3.0	90	7.0×10^{-5}	1.06×10^{-1}
2015	3.0	80	3.9×10^{-5}	6.38×10^{-2}
2015	3.0	85	3.3×10^{-5}	6.63×10^{-2}
2015	3.0	90	3.4×10^{-5}	7.61×10^{-2}
2021	3.0	80	4.5×10^{-5}	7.50×10^{-2}
2021	3.0	85	4.1×10^{-5}	8.18×10^{-2}
2021	3.0	90	4.2×10^{-5}	9.55×10^{-2}

Table 13: Top 3 Weibull parameter combinations per year, selected based on MSE and KL divergence.

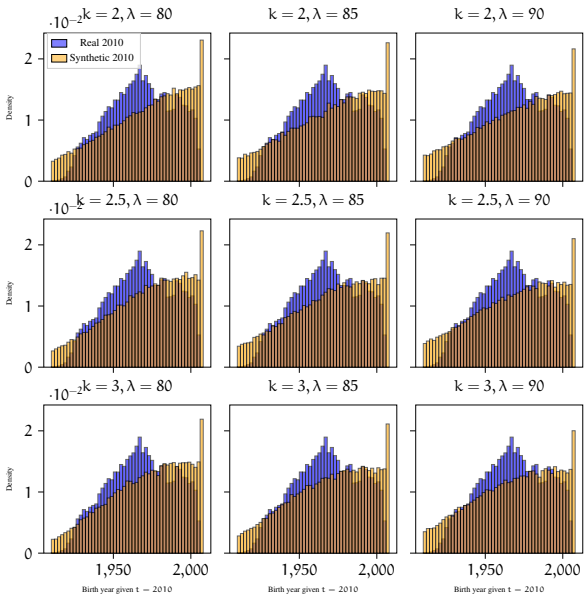


Figure 24: Sensitivity analysis of birth year given $t = 2010$ generated using Weilbull distribution with different k and λ compared to real sample

The results show that, although different parameter combinations produce varying degrees of fit, none of them are able to reproduce the shape of the empirical distribution. The generated data systematically overrepresented younger individuals and underrepresented older individuals, resulting in a flatter and unrealistic age profile.

While KL divergence and MSE serve as distance measures between distributions, they are not suitable for parameter estimation. MLE instead provides

a systematic approach by choosing the parameters k, λ that maximize the likelihood of the observed data under the model. We choose KL divergence as the primary evaluation metric because, in theory, maximum likelihood estimation MLE minimizes the KL divergence between the empirical distribution and the model (Murphy, 2012). Accordingly, we expect the parameter set obtained via MLE to achieve the lowest KL divergence, and potentially also the lowest mean squared error, when compared to alternative parameter sets. Thus, we estimate parameters k and λ using MLE for data from 2010, 2015, and 2021 and obtain the results presented in Table 14 and Figure 25.

Year	**k**	λ
2010	2.16	48.74
2015	2.24	50.16
2021	2.23	51.12

Table 14: Weibull parameter estimates by year

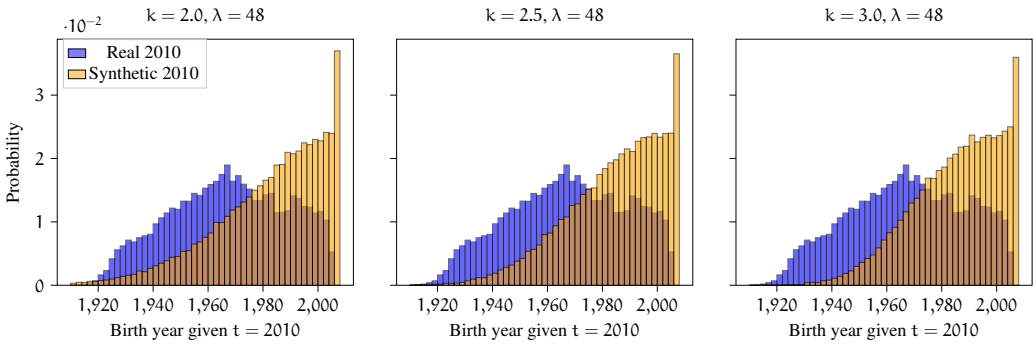


Figure 25: Comparison of real data from 2010 and generated data using estimated k and λ of the Weilbull distribution.

Parameter estimates from KL divergence and MLE differ significantly, with MLE yielding an unrealistic lifespan ($\lambda = 48$), suggesting a poor model fit. Although MLE theoretically minimizes KL divergence under ideal conditions, this discrepancy indicates that the Weibull distribution is insufficient to fully capture the true structure of the birth year distribution.

The results from both the sensitivity analysis and the MLE estimation confirm that a survival-only model cannot reproduce the empirical birth year structure at moment t. To address this limitation, we extend the model by introducing an explicit age-dependent sampling probability, as described in Section 3.4. Building on this extension, we now present the estimation results obtained using the maximum likelihood framework, which integrates empirical information into the

assumed distributions and calibrates their parameters to better reflect the observed data.

Two likelihood formulations are considered. First, we estimate a simplified version of the likelihood that focuses solely on calibrating the parameters governing the year of birth (e_1) and lifespan (d_1). This specification captures the joint effect of survival and sampling and yields the optimal parameter set

$$\theta_1^* = (k, \lambda, \alpha_y, \alpha_a, \alpha_o, \tau_1, \tau_2),$$

representing the combination of survival and sampling parameters that best explain the observed birth year structure for a given year t.

The second formulation extends the likelihood to also account for driving licence acquisition, as presented in Section 3.4. In this case, licence acquisition is modeled jointly with survival and sampling, enabling the estimation to integrate information from one or multiple cross-sectional datasets simultaneously. The corresponding parameter vector is expanded to

$$\theta_2^* = (k, \lambda, \alpha_y, \alpha_a, \alpha_o, \tau_1, \tau_2, \mu, \sigma, \pi),$$

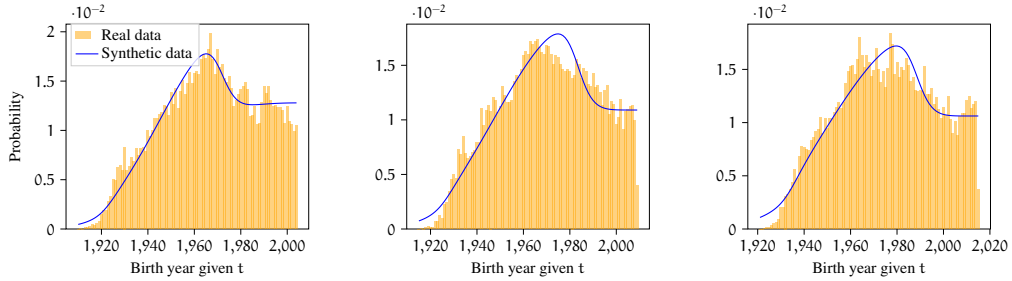
where (μ, σ, π) describe the parameters of the licence acquisition process.

The optimization returns the parameter values that maximize the log-likelihood on the empirical data. We perform the estimation independently for years $t = 2010, 2015, 2021$, and report the resulting optimal parameters and maximum log-likelihood values. Table 15 reports the estimation results obtained from the likelihood defined in Section 3.4.2. The distribution generated using these estimates is also visualized in Figure 26, alongside real data. We notice that we obtain realistic parameters (e.g., average life expectancy λ). This shows that modeling survival and sampling probabilities leads to more realistic results compared to modeling the survival component only.

Year	k	λ	α_y	α_a	α_o	τ_1	τ_2
2010	2.84	69.28	0.002	0.95	0.04	38.51	88.21
2015	3.43	74.25	0.019	0.933	0.048	31.75	89.21
2021	3.42	78.12	0.031	0.922	0.047	32.71	86.98

Table 15: Estimated model parameters θ_1^* using maximum likelihood

Additionally, we perform a sensitivity analysis by testing various combinations of the parameters to identify those that best fit the real data using MSE and KL divergence. To assess this, we generated a large number of random parameter sets drawn uniformly from plausible ranges. For each year, we evaluated the empirical fit of all random sets using both KL divergence and MSE. From these, we



(a) 2010 (b) 2015 (c) 2021

Figure 26: Birth year distributions given $t = 2010$, $t = 2015$, and $t = 2021$ generated using the estimated parameters from Table 15.

retained (i) the best-performing random set according to KL divergence and (ii) the best-performing random set according to MSE. Separately, we included the MLE-optimized parameter set (reported in Table 15), which was excluded from the random pool but evaluated in the same way, ensuring a fair comparison under the same model specification.

As shown in Table 16, the MLE solution consistently achieves the lowest KL divergence and, in most cases, also the lowest MSE across all three years. This supports the idea that the survival and sampling model not only maximizes likelihood, but also produces age distributions that closely align with the observed data under multiple evaluation criteria. Figure 27 further illustrates this by comparing, for each year, the distributions obtained from the best KL-divergence-based random solution with those from the MLE solution.

Year	**Metric**	**KL** 10^{-4}	**MSE** 10^{-4}	k	λ	α_y	α_a	α_o	τ_1	τ_2
2010	Random by KL	1.20	0.06	2.98	72.35	0.0044	0.9874	0.0082	32.83	81.80
2010	Random by MSE	1.20	0.06	2.98	72.35	0.0044	0.9874	0.0082	32.83	81.80
2010	MLE Solution	**0.84**	**0.05**	3.40	70.65	0.0452	0.9465	0.0084	38.51	88.21
2015	Random by KL	1.79	0.08	2.78	74.61	0.0282	0.9559	0.0159	33.88	78.46
2015	Random by MSE	1.89	0.08	3.09	79.24	0.1018	0.8605	0.0377	34.84	76.51
2015	MLE Solution	**1.32**	**0.08**	3.43	74.25	0.0192	0.9328	0.0480	31.76	89.21
2021	Random by KL	1.49	0.08	3.09	79.69	0.1078	0.8874	0.0048	31.44	85.19
2021	Random by MSE	2.08	0.07	3.38	75.68	0.0463	0.4891	0.4646	30.04	52.68
2021	MLE Solution	**1.35**	**0.07**	3.42	78.12	0.0307	0.9224	0.0469	32.71	86.98

Table 16: Comparison of KL divergence and MSE between best random samples and MLE solution for each year

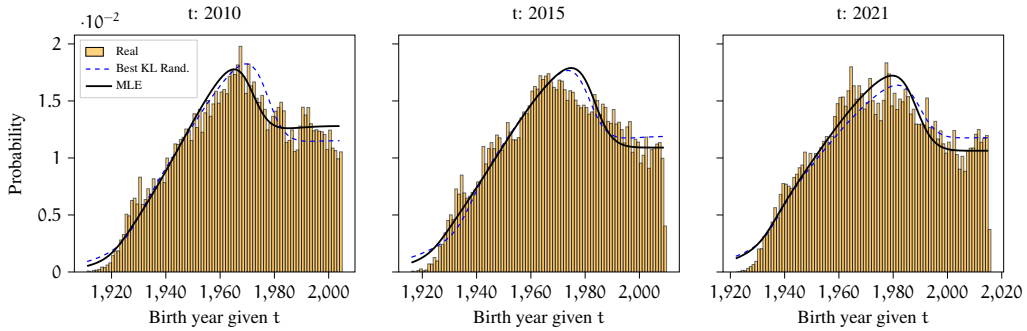


Figure 27: Comparison of birth year distributions given $t = 2010, 2015, 2021$ generated using the paramaters of best KL-divergence-based and the MLE-based solution

We now turn to the second estimation problem, introduced in Section 3.4. In this extended framework, the joint likelihood incorporates additional information, such as licence acquisition, and the parameter vector is augmented accordingly to include (μ, σ, π). The results of this extended estimation are presented in Table 17.

Year	k	λ	α_y	α_a	α_o	τ_1	τ_2	π	μ	σ	**Log-Likelihood**
2010	3.17	70.67	0.045	0.767	0.189	39.18	57.10	0.200	2.95	0.092	−265,776.00
2015	2.86	73.87	0.088	0.889	0.024	31.44	80.32	0.252	2.97	0.056	−242,626.21
2021	3.37	80.24	0.073	0.925	0.002	35.79	85.54	0.148	2.86	0.111	−234,659.54

Table 17: Estimated model parameters θ_2^* using the joint likelihood formulation.

From these results, several findings can be highlighted. First, the estimated survival and sampling parameters are consistent with those obtained in the previous procedure (see Table 15), and their values are interpretable and fall within realistic ranges. This indicates that the estimation procedure produces stable results, and the joint likelihood is well defined. Second, the similarity of the estimates across different years is expected. Although the cross-sectional samples are assumed to be independent (i.e., they do not track the same individuals), they are all drawn from the same underlying population (i.e., some dependencies may arise because surveys often stem from the same region or sampling frame), which naturally leads to consistent parameter values. Third, for the licence-related parameters (π, μ, σ), our estimates are close to the fixed values reported for Switzerland (see Section 3.1). This confirms the validity of our estimation procedure and highlights the advantage of the updating approach, since in countries where such reference values are not available, they can be directly inferred from the data.

References

Aemmer, Z. and MacKenzie, D. (2022). Generative population synthesis for joint household and individual characteristics, *Computers, Environment and Urban Systems* **96**: 101852.

Ahmed, U. and Moeckel, R. (2023). Impact of life events on incremental travel behavior change, *Transportation Research Record* **2677**(9): 594–605. Publisher Copyright: © National Academy of Sciences: Transportation Research Board 2023.

Beckman, R. J., Baggerly, K. A. and McKay, M. D. (1996). Creating synthetic baseline populations, *Transportation Research Part A: Policy and Practice* **30**(6): 415–429.

Beige, S. (2008). *Long-term and mid-term mobility decisions during the life course*, Dissertation, Institut für Verkehrsplanung und Transportsysteme (IVT), ETH Zürich, Zürich.

Beige, S. and Axhausen, K. W. (2017). The dynamics of commuting over the life course: Swiss experiences, *Transportation Research Part A: Policy and Practice* **104**: 179–194.

Bhat, C. R. and Koppelman, F. S. (1999). *Activity-Based Modeling of Travel Demand*, Springer US, Boston, MA, pp. 35–61.

Borysov, S. S. and Rich, J. (2021). Introducing synthetic pseudo panels: application to transport behaviour dynamics, *Transportation* **48**(5): 2493–2520.

Borysov, S. S., Rich, J. and Pereira, F. C. (2019). How to generate micro-agents? a deep generative modeling approach to population synthesis, *Transportation Research Part C: Emerging Technologies* **106**: 73–97.
URL: *http://dx.doi.org/10.1016/j.trc.2019.07.006*

Bureau of Labor Statistics (2019). Number of jobs, labor market experience, and earnings growth: Results from a national longitudinal survey, *News release*, U.S. Department of Labor, Washington, DC.
URL: *https://www.bls.gov/news.release/nlsoy.nr0.htm*

Casati, D., Müller, K., Fourie, P. J., Erath, A. and Axhausen, K. W. (2015). Synthetic population generation by combining a hierarchical, simulation-based approach with reweighting by generalized ranking, *Transportation Research Record* **2493**(1): 107–116.

Chlond, B., Ecke, L., Magdolen, M., Vallée, J. and Vortisch, P. (2024). The german mobility panel: – lessons learned from a longitudinal travel behavior survey over 30 years, *Transportation Research Record* **2678**(12): 1826–1842.

Ciganda, D. and Todd, N. (2024). Modelling the age pattern of fertility: an individual-level approach, *Royal Society Open Science* **11**(11).

Deaton, A. (1985). Panel data from time series of cross-sections, *Journal of Econometrics* **30**(1): 109–126.

Deschaintres, E., Morency, C. and Trépanier, M. (2022). Cross-analysis of the variability of travel behaviors using one-day trip diaries and longitudinal data, *Transportation Research Part A: Policy and Practice* **163**: 228–246.
URL: *https://www.sciencedirect.com/science/article/pii/S0965856422001823*

Edemealem, A. (2022). Modelling the transition process from higher education to employment: The case of undergraduates from debre markos university, *Education Research International* **2022**: 1119825.
URL: *https://doi.org/10.1155/2022/1119825*

Employee Benefit Research Institute (2025). Trends in employee tenure, 1983–2024. Accessed: 2025-08-28.
URL: *https://www.ebri.org/content/trends-in-employee-tenure–1983-2024*

Eurostat (2025). Population projections in the eu.
URL: `https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Population_projections_in_the_EU`

Farooq, B., Bierlaire, M., Hurtubia, R. and Flötteröd, G. (2013). Simulation based population synthesis, *Transportation Research Part B: Methodological* **58**.

Federal Statistical Office (FSO) (2017). Population's travel behaviour 2015, *Technical report*, Federal Statistical Office (FSO), Neuchâtel, Switzerland. Swiss Statistics, Mobility and Transport Microcensus (MTMC), FSO number 1697-1500.

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

Gärling, T. and Axhausen, K. W. (2003). Introduction: Habitual travel choice, *Transportation* **30**(1): 1–11.
URL: *https://doi.org/10.1023/A:1021230223001*

Goulias, K. G. and Kitamura, R. (1992). Travel demand forecasting with dynamic microsimulation, *Transportation Research Record* .
URL: *https://api.semanticscholar.org/CorpusID:152417751*

Haghighi, M. and Miller, E. J. (2025). Week-long activity-based modelling: a review of the existing models and datasets and a comprehensive conceptual framework, *Transport Reviews* **45**(1): 119–148.

Hradec, J., Craglia, M., Di Leo, M., S, D. N., N, O. and N, N. (2022). Multi-purpose synthetic population for policy applications, (KJ-NA-31116-EN-N (online)).

Hsiao, C. (2014). *Analysis of Panel Data*, Econometric Society Monographs, 3 edn, Cambridge University Press.

Hush, D. R., Ojha, T. and Al-Doroubi, W. (2021). The time to graduation problem: Survival analysis for education outcomes, *Technical report*, University of New Mexico, Department of Electrical and Computer Engineering.
URL: *https://digitalrepository.unm.edu/ece_rpts/54*

Härkönen, J. and Sirniö, O. (2020). Educational transitions and educational inequality: A multiple pathways sequential logit model analysis of finnish birth cohorts 1960–1985, *European Sociological Review* **36**(5): 700–719.
URL: *https://doi.org/10.1093/esr/jcaa019*

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*, MIT Press, Cambridge, MA.

Kukic, M. and Bierlaire, M. (2025). Adaptive synthetic generation using one-step gibbs sampler, *Transportation Research Interdisciplinary Perspectives* **33**: 101597.
URL: *https://www.sciencedirect.com/science/article/pii/S2590198225002763*

Kukic, M., Li, X. and Bierlaire, M. (2024). One-step gibbs sampling for the generation of synthetic households, *Transportation Research Part C: Emerging Technologies* **166**(104770).

Kukic, M., Rezvany, N. and Bierlaire, M. (2024). A Review of Activity-based Disaggregate Travel Demand Models, *Findings* .

Lederrey, G., Hillel, T. and Bierlaire, M. (2022). DATGAN: Integrating expert knowledge into deep learning for synthetic tabular data.

Lynge, E., Sandegaard, J. L. and Rebolj, M. (2011). The danish national patient register, *Scandinavian Journal of Public Health* **39**(7 Suppl): 30–33.

Mahevahaja, J. and Josoa Michel, T. (2023). Computation of human lifespan with a weibull distribution, *International Journal of Science and Research (IJSR)* **12**: 1927–1932.

Maier, C., Thatcher, J. B., Grover, V. and Dwivedi, Y. K. (2023). Cross-sectional research: A critical perspective, use cases, and recommendations for is research, *International Journal of Information Management* **70**: 102625.
URL: *https://www.sciencedirect.com/science/article/pii/S0268401223000063*

Mills, M. (2011). Parametric models, *Introducing Survival and Event History Analysis*, SAGE Publications Ltd, London, pp. 114–140.

Molloy, J., Castro Fernández, A., Götschi, T., Schoeman, B., Tchervenkov, C., Tomic, U., Hintermann, B. and Axhausen, K. W. (2023). The mobis dataset: a large gps dataset of mobility behaviour in switzerland, *Transportation* **50**(5): 1983–2007.
URL: *https://doi.org/10.1007/s11116-022-10299-4*

Moralti, J.-L., Maksim, H., Siegenthaler, C., Popovic, J., Balmer, M. and Danalet, A. (2023). Mobilitätsverhalten der bevölkerung. ergebnisse des mikrozensus mobilität und verkehr 2021, *Technical report*, Bundesamt für Statistik (BFS) Bundesamt für Statistik (BFS) Bundesamt für Statistik (BFS), Neuchâtel.

MPN (2024). The netherlands mobility panel. Accessed: 2025-09-10.
URL: *https://english.kimnet.nl/the-netherlands-mobility-panel*

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*, MIT Press, Cambridge, MA.

Müller, K. and Axhausen, K. W. (2010). Population synthesis for microsimulation: State of the art, *Proceedings of the 10th Swiss Transport Research Conference (STRC)*, Ascona, Switzerland.

Nurul Habib, K. (2018). Modelling the choice and timing of acquiring a driver's license: Revelations from a hazard model applied to the university students in toronto, *Transportation Research Part A: Policy and Practice* **118**: 374–386.

OECD (2021). *Education at a Glance 2021: OECD Indicators*, OECD Publishing, Paris.

Organisation for Economic Co-operation and Development (2021a). *Education at a Glance 2021: OECD Indicators*, OECD Publishing, Paris.
URL: *https://doi.org/10.1787/b35a14e5-en*

Organisation for Economic Co-operation and Development (2021b). *Labour Market Transitions across OECD Countries: Stylised Facts*, OECD Publishing, Paris.
URL: *https://www.oecd.org/employment/labour-market-transitions-across-oecd-countries-1d7f65fa-en.htm*

Oshanreh, M. M., Khan, N. A. and MacKenzie, D. (2024). Propagating synthetic populations with dynamic bayesian networks: A framework for long-horizon demographic forecasting. Available at SSRN: `https://ssrn.com/abstract=5281670` or `http://dx.doi.org/10.2139/ssrn.5281670`.

Patrinos, H. (2016). Estimating the return to schooling using the mincer equation, *IZA World of Labor* .

Polachek, S. (2007). Earnings over the life cycle: The mincer earnings function and its applications, *Foundations and Trends® in Microeconomics* **4**.

Qian, X., Gangwal, U., Dong, S. and Davidson, R. (2024). A deep generative framework for joint households and individuals population synthesis.
URL: *https://arxiv.org/abs/2407.01643*

Reina, J. C., Cuaresma, J. C., Fenz, K., Zellmann, J., Yankov, T. and Taha, A. (2024). Gravity models for global migration flows: A predictive evaluation, *Population Research and Policy Review* **43**(2): 29.
URL: *https://doi.org/10.1007/s11113-024-09867-6*

Savcisens, G., Eliassi-Rad, T., Hansen, L. K. et al. (2024). Using sequences of life-events to predict human lives, *Nature Computational Science* **4**: 43–56.
URL: *https://doi.org/10.1038/s43588-023-00573-5*

Simmons, M. (2023). Job to job transitions, job finding and the ins of unemployment, *Labour Economics* **80**: 102304.
URL: *https://www.sciencedirect.com/science/article/pii/S0927537122001944*

SMP (2020). Swiss mobility panel. Accessed: 2025-09-10.
URL: *https://istp.ethz.ch/research/swiss-mobility-panel.html*

Swiss Federal Office of Statistics (2012; 2018; 2023). *Comportement de la population en matière de mobilité*, Bundesamt für Statistik (BFS), Neuchâtel.

Vagni, G. and Cornwell, B. (2018). Patterns of everyday activities across social contexts, *Proceedings of the National Academy of Sciences* **115**(24): 6183–6188.
URL: *https://www.pnas.org/doi/abs/10.1073/pnas.1718020115*

Weon, B. M. (2004). Analysis of trends in human longevity by new model, *arXiv preprint q-bio/0402011* .
URL: *https://arxiv.org/abs/q-bio/0402011*

Wilmoth, J. R., Andreev, K., Jdanov, D., Glei, D. A., Riffe, T., Boe, C., Bubenheim, M., Philipov, D., Shkolnikov, V., Vachon, P., Winant, C. and Barbieri, M. (2025). Methods protocol for the human mortality database, *Technical report*, University of California, Berkeley and Max Planck Institute for Demographic Research. Version 6, Last revised August 5, 2025.
URL: *http://www.mortality.org*

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*, The MIT Press.
URL: *http://www.jstor.org/stable/j.ctt5hhcfr*

Xu, L. and Veeramachaneni, K. (2018). Synthesizing tabular data using generative adversarial networks, *arXiv:1811.11264 [cs, stat]* .

Yaméogo, B. F., Gastineau, P., Hankach, P. and Vandanjon, P.-O. (2021). Comparing methods for generating a two-layered synthetic population, *Transportation Research Record* **2675**(1): 136–147.

Ye, X., Konduri, K., Pendyala, R. M., Sana, B. and Waddell, P. (2009). A methodology to match distributions of both household and person attributes in the generation of synthetic populations, *88th Annual Meeting of the transportation research Board, Washington, DC*.