Generating Synthetic Panel Populations from Life Trajectories: A Model-Based Framework

Candice Baud Michel Bierlaire *

August 07, 2026

Report TRANSP-OR 260807
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, {michel.bierlaire, candice.baud}@epfl.ch

Abstract

This paper proposes a general framework for constructing synthetic populations whose panel structure is specified by design. The full life trajectories of individuals are generated, and a deterministic mapping recovers their state at any time t, allowing the reconstruction of population distributions at arbitrary points in time for the same set of individuals. The framework can accommodate different domain-specific probabilistic models and enforces structural feasibility and temporal consistency through constraints embedded in the life-trajectory representation.

We further introduce a Bayesian updating mechanism that incorporates information from observed cross-sectional datasets. When data are available, the synthetic population is sampled from the posterior distribution, combining prior knowledge with the evidence contained in the observations. This allows cross-sectional information, such as census or survey microdata, to update the distribution of coherent synthetic life trajectories conditional on the prior model and structural assumptions.

1 Introduction

Activity-based models (ABMs) rely on synthetic populations to represent the socio-economic characteristics of individuals whose activities and travel behavior are simulated (Rezvany et al., 2024; Müller and Axhausen, 2010). In practice, most applications rely on cross-sectional data capturing individuals at a single point in time due to the accessibility and relatively low cost of such data. However, individual habits and preferences evolve over time, and modelling these changes across heterogeneous populations remains difficult (Gärling and Axhausen, 2003; Borysov and Rich, 2021). Individual preferences shift as life events unfold and socio-economic status changes, but habits are also shaped by public policies and incentives. By construction, cross-sectional data provide only a static snapshot of the population. Repeated cross-sectional surveys still sample a different set of individuals at each wave. As a result, they do not allow the observation of behavioral evolution over time, the identification of dynamic effects, or the establishment of causal relationships (Wooldridge, 2010; Maier et al., 2023; Bhat and Koppelman, 1999; Baltagi, 2005).

Panel data offer a natural way to overcome these limitations by tracking the same individuals over time, enabling the identification of persistent heterogeneity and dynamic behavioral patterns (Frees, 2004; Hsiao, 2014). By observing the same units repeatedly, they provide a stronger basis for policy analysis (Deschaintres et al., 2022). However, panel surveys remain rare and costly, are subject to attrition, panel fatigue, and conditioning effects, and typically cover only limited time horizons (Deschaintres et al., 2022). As a result, available data remain predominantly cross-sectional, and most synthetic population methods have accordingly focused on generating cross-sectional data, typically reproducing marginal distributions at a given time through fitting procedures or simulation-based approaches (Farooq et al., 2013; Kukic and Bierlaire, 2025; Kukic et al., 2024). When panel analyses are required, researchers rely on projection models (Kukic et al., 2023; Kukic and Bierlaire, 2025) or pseudo-panel techniques (Deaton, 1985; Verbeek and Nijman, 1998) that infer dynamics from repeated cross-sections but do not reconstruct consistent individual life trajectories. In the transport field, for instance, pseudo-panel techniques have been used to model car ownership dynamics (Dargay and Vythoulkas, 1999). However, generating populations independently at different times or for different cohorts does not guarantee that individuals correspond to stable underlying entities. The core limitation is therefore not only the scarcity of panel data, but the absence of a time-consistent generative representation of individuals.

Longitudinal data are needed in the transport field to analyze behavioral dynamics and the long-term effects of transportation policies, yet the data that are actually available are mostly cross-sectional. This article addresses the follow-

ing question: can longitudinal synthetic populations be generated using cross-sectional data? Our contributions are as follows:

- We introduce a general, structurally feasible and temporally coherent framework for generating complete synthetic life trajectories from prior probabilistic models. The event-based representation and the associated structural constraints ensure temporal and logical consistency, while probabilistic sampling preserves population heterogeneity and reproduces the distributions implied by the input models.
- We define a deterministic mapping that recovers the state of each synthetic individual at any time of interest. This makes it possible to construct cross-sectional datasets at arbitrary dates from the same underlying population. The longitudinal structure is inherent to the representation: individual identities are preserved over time.
- We propose a Bayesian updating mechanism that incorporates information from one or several cross-sectional datasets into the generation of life trajectories. The observed data update the prior distribution of the underlying trajectory variables, while the structural constraints continue to guarantee coherent and feasible individual histories. Because the update is performed on the complete trajectory representation, the information introduced at specific observation dates can propagate in the long term.

The paper is structured as follows. Section 2 reviews the related literature. Section 3 presents the proposed methodology. Section 4 applies the approach to a case study on Switzerland. Finally, Section 5 concludes by discussing the limitations of the method and outlining directions for future research.

2 Literature review

Most current methods generate cross-sectional synthetic populations using iterative proportional fitting (IPF), which focuses on reproducing marginal distributions of socio-demographic variables (Beckman et al., 1996). Such approaches are inherently limited to aggregate consistency: they do not capture the full joint distribution of attributes and may introduce unrealistic independence assumptions between variables. To address this limitation, simulation-based approaches have been proposed, notably relying on Markov chain Monte Carlo (MCMC) frameworks to generate populations consistent with observed marginal and, where available, joint distributions (Farooq et al., 2013). Despite this methodological progress, most synthetic population approaches remain fundamentally cross-sectional in scope. Even methods designed to generate populations without microdata focus on static consistency rather than temporal coherence (Barthélemy and Toint, 2013), and approaches that improve the representation of joint distributions (Ye et al., 2009) do not address individual-level dynamics. Consequently, when synthetic populations are generated independently for different time periods, nothing guarantees that individuals in one period correspond to the same underlying entities as in another.

To move beyond static representations, alternative approaches based on repeated cross-sections have been developed, including pseudo-panel methods (Deaton, 1985; Verbeek and Nijman, 1998) and projection models (Kukic et al., 2023). Projection methods generate populations at multiple time points but do not preserve persistent individual identities across them. Pseudo-panel approaches instead reconstruct cohort-level dynamics by tracking synthetic cohorts rather than individual life trajectories. Both strategies therefore fall short of producing true panel data: one sacrifices identity persistence, the other operates at a coarser level of aggregation than the individual.

In demography, Oshanreh et al. (2025) propose a dynamic microsimulator that uses dynamic Bayesian networks to project individual life-course transitions forward in time; however, this approach requires genuine panel data as input to estimate transition probabilities and it projects an already-sampled population rather than generating a new synthetic panel sample.

The closest attempt to bridge the literature gaps is Kukic et al. (2025), which generates panel data as a proof-of-concept case study on Switzerland. However, this work does not propose a formal, general framework for producing temporally coherent longitudinal populations from cross-sectional data alone; the result is demonstrated for a specific setting rather than derived as a generalizable method. Our paper builds on this example, generalizing it into a coherent framework capable of constructing longitudinal synthetic populations that can be enriched using cross-sectional measurements.

3 Methodology

The methodology builds on the general principle of dynamic microsimulation, where population dynamics are represented at the individual level by simulating life trajectories as sequences of events. Following the idea of Orcutt (1957) and Oshanreh et al. (2025), socio-economic systems are described through individual units rather than aggregate relationships. Individuals are followed over time and transitions between life states are typically governed by probabilistic rules, often derived from hazard, transition, or multi-state models (van Imhoff and Post, 1998; Willekens, 2014). Although the present methodology does not rely directly on such models, it shares the same fundamental idea: generating complete individual life trajectories.

The time-independent principle introduced in Kukic et al. (2025) can be embedded in this broader life-trajectory concept. Instead of constructing a population independently at each point in time, the method first generates complete trajectories for individuals. The state of an individual at any time $\mathtt{t}$ is then not simulated separately, but derived from the pre-generated trajectory. This makes it possible to recover the population at any date as a cross-sectional snapshot of the same underlying longitudinal structure.

3.1 Unified life trajectory framework

This section introduces a general framework for generating longitudinal life trajectories that are structurally feasible and temporally coherent by design. The framework takes as input a probabilistic model describing the occurrence, timing, and attributes of individual life events. The resulting output is a synthetic population composed of individuals with complete life histories. A mapping mechanism is then used to recover each individual's state at any time $\mathtt{t}$, allowing the generated trajectories to be represented as longitudinal or panel data.

Each individual is represented by a finite existence interval $[\mathtt{b}, \mathtt{b} + \mathtt{L}]$, where $\mathtt{b}$ denotes the date of birth and $\mathtt{L}$ the lifespan. This interval defines the temporal support of the individual's life, over which multiple life axes evolve in parallel. These axes, referred to as *dimensions*, represent homogeneous aspects of life such as education, employment, place of residence, or driving license ownership. At every instant of the existence interval, each dimension is associated with exactly one active event describing the individual's state in that domain. Within each dimension, the existence interval is partitioned into a finite sequence of consecutive events. Thus, while different dimensions evolve concurrently, events belonging to the same dimension occur sequentially and together form a complete, non-overlapping decomposition of the individual's lifespan. A dimension also speci-

fies the set of admissible event types and the common attribute structure shared by all events belonging to that dimension.

The objective is therefore to sample, for each individual, a trajectory in each dimension. We define a random vector X containing the variables required to describe these trajectories: event occurrence indicators, temporal characteristics, and event-specific attributes. We denote by X_i the trajectory-defining variables of individual i, and $\mathbf{X} = (X_1, \ldots, X_N)$ the complete synthetic population. Occurrence indicators determine whether a given event is realized in the life of the individual; temporal characteristics determine when the event starts and ends; and attributes describe additional properties associated with the event. Individuals are sampled independently from this random vector, thereby generating a population with heterogeneous life trajectories.

For a given individual and dimension D, the realized trajectory is represented as a finite ordered sequence

$$\tau_D = (e_{1,D}, \ldots, e_{n_D,D}), \quad n_D > 0$$

where each event is associated with an occurrence indicator, temporal attributes, and, when relevant, non-temporal attributes. The set of dimensions and possible event types is common to all individuals, whereas the realized trajectories are individual-specific and sampled from the probabilistic model.

The *Existence* dimension differs from the other dimensions because it defines the temporal bounds of the individual's life. It contains a single event, referred to as the *Birth* event,

$$\tau_{\text{Existence}} = (e_{\text{birth}}),$$

which records the individual's date of birth, lifespan, and attributes determined at birth that remain constant throughout their life, such as sex.

Temporal characteristics are expressed through durations. Each realized event is characterized by a *main duration*, corresponding to the substantive state, and optionally by a *gap duration*, allowing flexible spacing between successive substantive events. Once the durations are known, the temporal position of every event is uniquely determined. Attributes, such as income, workplace location, or occupation type, remain constant throughout the corresponding event they characterize. Figure 1 presents a *Work* dimension trajectory for an individual. The individual undergoes 5 work spells comprising a main part, corresponding here to being employed and a gap part, corresponding here to an interruption of employment. For each spell, the Log Income and Place of Work attributes are defined.

Figure 1: Work dimension example

The event-based representation imposes structural consistency. Within each dimension, realized events must form a complete and non-overlapping partition of the lifespan. In particular, events in the same dimension cannot overlap, their total duration must cover the existence interval, and occurring events must have positive main durations. Therefore, at any time $t \in [b, b + L]$, a cross-section of the trajectories identifies one active event per dimension. The individual's time-dependent state is then obtained by applying the deterministic mapping to this cross-section.

Additional logical, biological, legal, and institutional constraints restrict the set of admissible trajectories. These constraints encode, for example, minimum or maximum starting ages, duration bounds, and event relations across dimensions. These restrictions can all be expressed as linear constraints on the event durations, which defines a feasible set of trajectories that can be represented in linear-algebraic form. Denoting x the individual vector of durations, the constraints are represented using a matrix A and a vector b such that the feasible set of durations lies in

$$\mathcal{F} = \{x \in \mathbb{R}^p : Ax \leq b\}.$$

The constraint types currently implemented and their algebraic representation are available in the Appendix 7.1.

The synthetic trajectories $\tau = (\tau_D)_D$ are generated through a structured Markov Chain Monte Carlo procedure. First, the *Existence* dimension, containing date of birth and lifespan is sampled because it defines the temporal support of all other events. The joint density of the durations, and the exact conditionals for the attributes must be provided. Second, conditional on *Existence*, event indicators are sampled using a Gibbs sampler subject to eligibility and logical constraints. This step requires as input the exact conditionals of the indicators. Third, conditional on the indicators, temporal and non-temporal attributes are sampled using

a Gibbs–Metropolis–Hastings scheme (Metropolis et al., 1953; Hastings, 1970; Gelfand and Smith, 1990). For this, we alternate between two steps: sampling event durations conditional on the attributes, and sampling attributes conditional on the durations. Conditional on the current attribute values, the duration vector is updated using a Hit-and-Run algorithm (Smith, 1984; Lovász and Vempala, 2006) targeting its joint conditional distribution over the feasible set. Before sampling, we isolate any constraints that are equalities in disguise and collect them into $A_{eq} \cdot x = b_{eq}$. These directions carry no uncertainty and therefore do not need to be explored by the sampler. Conditional on the sampled durations, the attributes are then updated through a Metropolis-Hastings step. If the exact conditionals are known, this step reduces to a Gibbs update with an acceptance ratio of 1. Once the sampling procedure is complete, event starting dates are recovered deterministically from the sampled durations.

The pseudo-algorithm in Algorithm 1 makes explicit the sampling strategy for one individual. The algorithm applies to both the sampling of the *Existence* dimension and that of the other dimensions. The key distinction is that, for the latter, the constraints depend on the *Existence* dimension sampled upstream, whereas the constraints for the *Existence* dimension are self-contained and do not rely on other dimensions. Consequently, when sampling the other dimensions, the realization of the *Existence* dimension is taken as input to construct the corresponding constraints. Moreover, the input model ($\mathcal{M}$) encodes all probabilistic specifications required to sample each quantity of interest. It defines the conditional distributions from which the event indicators are sampled as well as the set of constraints and the target prior log-density from which samples are drawn. The accept-reject step is based on this joint target density. In the algorithm, the instance I denotes the representation of the sampled individual trajectory X_i.

The framework fulfills its purposes. First, it ensures temporal consistency by construction. Instead of generating independent cross-sectional states at different points in time, the framework samples complete individual life trajectories. The evolution of an individual is therefore represented as a coherent sequence of events for each dimension. The structural rules and constraints prevent impossible or unrealistic temporal configurations. Temporal consistency is therefore not imposed ex post, but follows directly from the event-based representation and the feasible set of trajectories.

Second, the framework preserves individual identity across an infinite time horizon. Individuals are not defined relative to a given observation year. Instead, each synthetic individual is generated with a date of birth and a life trajectory, which together define the individual independently. At any time t, the deterministic mapping projects this underlying trajectory onto the individual's observable

Algorithm 1 Generic sampling algorithm for one individual

Require: model $\mathcal{M}$, burn-in B
Ensure: sampled instance I

Initialization
Initialize the instance I containing the life trajectory of the individual
Sample indicators using Gibbs sampling
Initialize attributes
Build constraints $Ax \leq b$ based on the indicators
Extract A_{eq}, b_{eq} and compute nullspace basis $N(A_{eq})$
Find a feasible x_0 and fill it in I
Compute the initial log-density $\pi(I)$ defined by $\mathcal{M}$

MCMC iterations
for $m = 1$ to B **do**
    Attribute update (MH-within-Gibbs)
    Propose new attributes a' and accept/reject using the MH ratio on $\pi(I')$

    Duration update (Hit-and-Run)
    Sample direction $u \in \text{Null}(A_{eq})$
    Compute the feasible segment $[t_{min}, t_{max}]$
    Sample $t \sim \mathcal{U}[t_{min}, t_{max}]$
    Propose $x' = x + tu$ and accept/reject using MH ratio on $\pi(I')$
end for

return final instance I

state at that date. Depending on t, the same individual may be unborn, alive with a given set of characteristics and events, or deceased. Thus, although the individual is represented over an arbitrarily broad time horizon, their lifetime -and therefore the period during which life events can occur- is bounded by their dates of birth and death. In practice, the framework is applied over a finite calendar interval of interest, which can be chosen as broadly as required. However, using a single time-invariant model over a very long period would generally be unrealistic, since demographic processes and individual behavior evolve over calendar time. The probabilistic models used to generate or constrain the different dimensions may therefore depend on t, reflecting the historical period to which their calibration data correspond.

Third, the framework is naturally designed to integrate multiple cross-sectional datasets. Each cross-sectional dataset provides information on the population observed at a particular point in time. Since the deterministic mapping can project the synthetic longitudinal population to any observation date, the simulated population can be compared with the corresponding observed data at that date. Cross-sectional datasets act as temporal reference points for the broader trajectory generation mechanism.

The methodology makes it possible to generate complete population trajectories using probabilistic specifications for event occurrences, durations, and attributes, that can be extracted from literature as an input. The open-source code is available on GitHub[1] and Zenodo (Baud, 2026).

3.2 Integration of cross-sectional data

The second part of the methodology describes how a probabilistic model derived from the literature or other prior knowledge can be enriched using external information about the target population. Before introducing this approach, it is important to distinguish it from a direct modification of the prior model itself. The assumptions and parameters of the prior model may be adjusted to represent specific temporal, spatial, or policy contexts, after which the resulting population can be compared with that generated under the baseline specification. For example, the survival probabilities of young men could be reduced to represent the demographic consequences of a war, while mortality rates among older individuals could be increased to represent a pandemic. Such interventions do not necessarily require observed microdata; they may instead rely on assumptions, expert knowledge, demographic projections, or other scenario-specific evidence. This type of intervention remains within the methodology presented in the previous section:

[1]Github repository : candicebaud/longitudinal_synthetic_population

external knowledge is used to modify the specification of the probabilistic model, but no data are directly incorporated into the population-generation process.

The present section addresses a different setting, in which external information is explicitly introduced in the form of cross-sectional, disaggregate data. These data may consist of actual observations, demographic forecasts, policy projections, or counterfactual data generated for a particular scenario. The framework therefore takes as inputs both an initial model of individual life trajectories and an external dataset describing the target population. The incorporated information then propagates through the longitudinal model. For example, a decline in births affects not only the number of children generated in the short term, but also the future age structure of the population. The resulting output is a synthetic population composed of individual life trajectories that are more closely aligned with the characteristics of the target population.

We assume that individual-level cross-sectional data $\tilde{Y}_t$ of a population are observed at time t. Let X_i denote the primary time-independent variables of synthetic individual i, let $\mathbf{X} = (X_1, ..., X_N)$ denote the full synthetic population of size N. Given the observed sample $\tilde{Y}_t = \{\tilde{Y}_{t,j}\}_{j=1}^{n}$, the objective is to sample $\mathbf{X}$ from the posterior :

$$f(\mathbf{X} \mid \tilde{Y}_t) \propto f(\tilde{Y}_t \mid \mathbf{X}) \cdot \pi_X(\mathbf{X}),$$

which combines the prior distribution $\pi_X(\mathbf{X})$ with a likelihood $f(\tilde{Y}_t \mid \mathbf{X})$ measuring the consistency between simulated and observed data. A draw from this posterior is a single candidate population $\mathbf{X} = (X_1, ...X_N)$ of size N, generated jointly, and not a collection of N independent draws nor a set of n draws paired one-to-one with the observations.

The state of a synthetic individual i at time t is obtained by applying the time-specific mapping T_t to their trajectory:

$$Y_{t,i} = T_t(X_i)$$

The mapping recovers the alive status of an individual as

$$\text{Alive}_{t,i} = \begin{cases} 1 & \text{if } b_i \leq t < b_i + L_i \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

If the individual is alive, $Y_{t,j}$ also contains the observable characteristics of individual i at time t such as age, employment status, license status, etc. Let

$$\mathcal{I}_t(\mathbf{X}) = \{i,\ \text{Alive}_{t,i} = 1\}$$

the set of alive individuals at time t, and $M_t(\mathbf{X}) = |\mathcal{I}_t(\mathbf{X})|$ its size.

Since $\tilde{Y}_t$ is obtained from surveys or censuses, it only contains individuals alive at time t, whereas the simulated population also includes individuals not yet born or already deceased. $\tilde{Y}_t$ is only informative about the characteristics of the subset of synthetic individuals alive at t. For individuals not alive at t, the observations contain no direct information about their attributes; they may influence birth-date and lifespan distributions only through population-level constraints and the shared generative model.

The empirical cross-sectional distribution induced by this eligible subpopulation is

$$\mu_t^{\mathbf{X}} = \frac{1}{M_t(\mathbf{X})} \sum_{i \in \mathcal{I}_t(\mathbf{X})} \delta_{Y_{t,i}}$$

Let $K_t(\tilde{y}, y)$ denote a an observation kernel quantifying the plausibility that an observed value $\tilde{y}$ is consistent with a synthetic trajectory value y at time t. Its specification reflects assumptions about the relationship between the observed data and the model's projection, and must be adapted to the nature of the variable. For continuous variables, we use continuous kernels. For discrete variables, we rely on categorical kernels representing mismatch probabilities between categories, interpretable as misclassification models that allow probabilistic discrepancies between simulated and observed categorical outcomes (Li and Racine, 2007; Hausman et al., 1998; Aitchison and Aitken, 1976). Misclassification probabilities are specified following Li and Racine (2007) and depend on both data quality and the variable considered - for instance, sex is typically measured with less error than education. Exact kernel definitions used are given in Appendix 7.3. The parameters governing K_t - the bandwidth of the continuous kernels, the mismatch probabilities of the categorical kernels - should not be interpreted as literal estimates of survey or census reporting error. They act as a discrepancy, or smoothing, tolerance: they determine how much difference between an observed individual and a simulated population state is treated as consistent with the model, and thus how much mass the eligible synthetic population must place near $\tilde{y}$ for it to contribute to the likelihood.

Because no observed individual is matched to a specific synthetic individual, the likelihood of each observation is obtained by averaging the observation kernel over the eligible synthetic population. The likelihood of a single observed individual is then defined directly as the integral of the kernel against μ_t^X, i.e. its average value over the eligible synthetic population:

$$L_{t,j} = P(\tilde{Y}_{t,j} \mid \mathbf{X}) = \int K_t(\tilde{Y}_{t,j}, y) d\mu_t^X(y) = \frac{1}{M_t(\mathbf{X})} \sum_{i \in \mathcal{I}_t(\mathbf{X})} K_t(\tilde{Y}_{t,j}, Y_{t,i})$$

This is a population-mixture likelihood, not a paired-measurement model: the observed individual is compared simultaneously with every alive synthetic indi-

vidual, each contributing with equal weight $\frac{1}{M_t(\mathbf{X})}$. Assuming observed individuals are independent draws given $\mathbf{X}$, the likelihood of the population cross-section $\tilde{Y}_t = (\tilde{Y}_{t,j})_j$ is:

$$L_t = P(\tilde{Y}_t \mid \mathbf{X}) = \prod_{j=1}^{n} \frac{1}{M_t(\mathbf{X})} \sum_{i \in \mathcal{I}_t(\mathbf{X})} K_t\big(\tilde{Y}_{t,j}, Y_{t,i}\big).$$

When survey designs over- or under-sample certain groups, individuals do not contribute equally to the population, which is why the observed data are weighted. Each observation is assigned a weight reflecting the number of population units it represents. Ignoring these weights would implicitly treat the sample as if it were drawn uniformly from the population, which can lead to biased inference. To account for this, the likelihood is modified by weighting each observation according to its importance in the population. This leads to the following weighted likelihood:

$$L_t = P(\tilde{Y}_t \mid \mathbf{X}) = \prod_{j=1}^{n} \left[\frac{1}{M_t(\mathbf{X})} \sum_{i \in \mathcal{I}_t(\mathbf{X})} K_t\big(\tilde{Y}_{t,j}, Y_{t,i}\big) \right]^{w_j}.$$

The weights w_j can be interpreted as frequency or expansion factors: an observation with weight w_j contributes as if it were replicated w_j times in the sample. This formulation approximates the likelihood that would be obtained from a representative population sample. Raw survey weights are typically expansion factors calibrated so that $\sum_{j=1}^{n} w_j$ equals the size of the target population, not the size of the observed sample. Used directly as exponents, such weights would inflate the total evidence mass of the likelihood to the scale of the population rather than the scale of the observed data, producing an artificially peaked likelihood. To avoid this, we normalize the weights so that they sum to the sample size : $w_j^* = n \cdot \frac{w_j}{\sum_{k=1}^{n} w_k}$, and use w_j^* in place of w_j in the weighted likelihood.

Additionally, some categories might be absent from the survey by design. If this sampling rule is ignored, the posterior distribution may under-represent such categories, since the absence of such observations would be interpreted as evidence that they are absent from the population. To avoid this bias, the likelihood must account for the fact that the observed dataset is censored by design. Let S be a binary indicator equal to one if an individual is sampled in the time-dependent dataset, and equal to zero otherwise. The posterior distribution is therefore written as

$$P(\mathbf{X} \mid \tilde{Y}, S = 1) \propto P(\tilde{Y} \mid \mathbf{X}, S = 1) P(\mathbf{X}),$$

where the likelihood is conditional on the sampling rule. We define

$$q(y) = P(S = 1 \mid Y_t = y),$$

where y denotes the time-dependent state of an individual. If the sampling probability is deterministic and depends only on the state, this becomes :

$$q(y) = \begin{cases} 0 & \text{if } y \notin \mathcal{E}, \\ 1 & \text{if } y \in \mathcal{E} \end{cases}$$

where $\mathcal{E}$ denotes the set of eligible states to be observed. For a given simulated population $\mathbf{X}$, let

$$\mathcal{S}_t(\mathbf{X}) = \{i \in \mathcal{I}_t(\mathbf{X}) : q(Y_{i,t}) = 1\}$$

denote the set of sample-eligible individuals at time t. Here, $\mathcal{I}_t(\mathbf{X})$ is the set of individuals alive at time t, and $Y_{i,t}$ is the time-dependent state obtained by mapping individual i from the time-independent representation to time t.

The likelihood of the observed data is then computed only over the sample-eligible simulated individuals:

$$L_t = P(\tilde{Y}_t \mid \mathbf{X}, S = 1) = \prod_{j=1}^{n} \left[\frac{1}{|\mathcal{S}_t(\mathbf{X})|} \sum_{i \in \mathcal{S}_t(\mathbf{X})} K_t\left(\tilde{Y}_{t,j}, Y_{i,t}\right) \right]^{w_j}.$$

Therefore, the absence of such categories in the data is not treated as evidence that such categories are absent from the population. Instead, it is treated as a consequence of the sampling design. In practice, this means that the likelihood contribution of each observed individual is evaluated against the subset of simulated individuals that could have been sampled. A short derivation is provided in the Appendix, property 7.3.

Finally, when sampling from the prior, the population size at each time point is jointly determined by the input distributions of dates of birth and lifespans. During posterior inference, both distributions are sampled simultaneously from the observed population sizes, which gives rise to a non-identifiability problem: different combinations of birth-date and lifespan distributions may generate the same population size trajectory. Moreover, individuals who are not alive at any observation time do not contribute to the likelihood. Their dates of birth and lifespans may therefore take arbitrary values, provided that these values remain consistent with their absence from the observed population. To address this issue, we incorporate information on population size dynamics during the posterior sampling of the Existence dimension. Let $M_t^\star$ denote a target population size at time t, and let $\tilde{M}_t(\mathbf{X})$ denote the population size at time t implied by a simulated population $\mathbf{X}$. A penalty or regularization term is introduced to measure the discrepancy between these two functions:

$$L = \exp\left(-\frac{1}{2\sigma^2} \int_{t_{min}}^{t_{max}} (\tilde{M}_t(\mathbf{X}) - M_t^\star)^2 dt \right)$$

This expression is written as a continuous-time integral, as both $\tilde{M}_t(\mathbf{X})$ and $M_t^\star$ are defined over a continuous time horizon. In practice, the integral is approximated on a discrete grid t_k:

$$L = \int_{t_{min}}^{t_{max}} (\tilde{M}_t(\mathbf{X}) - M_t^\star)^2 dt \approx \sum_k w_k (\tilde{M}_{t_k}(\mathbf{X}) - M_{t_k}^\star)^2, \quad w_k = \frac{\Delta t_k}{t_{max} - t_{min}}$$

To capture relative changes instead of levels, the discrepancy is defined in logarithmic scale:

$$\ell = -\frac{1}{2\sigma^2} \sum w_k \left(\log(\tilde{M}_{t_k}(\mathbf{X})) - \log(M_{t_k}^\star) \right)^2,$$

This formulation emphasizes proportional differences and ensures that the penalty focuses on matching growth trends over time, while still relying on the same piecewise-constant representation of the population size trajectory. The function $\tilde{M}_t(\mathbf{X})$ is computed exactly from the simulated trajectories using the birth dates and lifespans of individuals. More precisely, it is a piecewise-constant (staircase) function of time: starting from zero, it increases by one at each birth event and decreases by one at each death event. Between two consecutive events (birth or death of one individual), the function remains constant. As a result, the entire trajectory of $\tilde{M}_t(\mathbf{X})$ is fully characterized by the ordered sequence of event times, and can be evaluated without discretization error. In this application, the objective is not to match the absolute population size as we simulate a sample, but rather its temporal evolution. The log likelihood used in the acceptance ratio is obtained as the sum of $\log(L_t)$ –denoted by ℓ_t– and ℓ. Summing the two log-terms is equivalent to multiplying the cross-sectional kernel likelihood and the population-size penalty together, treating agreement with individual-level observed profiles and agreement with the aggregate population trajectory as two independent sources of evidence about the same underlying population.

In order to perform the posterior sampling, we do the following procedure. The *Existence* dimension is sampled first at the population level and the rest of the life trajectory is sampled conditional on the lifespan and birth characteristics. To sample the *Existence* dimension, the likelihood compares observed ages with simulated ages of alive individuals, while the population constraint ensures identification through the evolution of $M_t(\mathbf{X})$. The update is performed at the population level, as the population size constraint couples all individuals. Candidate populations are generated by modifying a subset of individuals (e.g., birth dates and lifespans), using proposals such as Hit-and-Run in the constrained space.

The remaining dimensions are then sampled conditionally on the *Existence* trajectories at the individual level. For each individual, event indicators are generated given the *Existence* trajectory. Cross-sectional data are not informative

about lifetime event occurrences and thus the indicators are sampled from the prior distribution. For instance, observing employment at time t does not inform the number of past or future employment spells. More generally, cross-sectional measurements capture prevalence conditional on age and survival, and are subject to censoring and cohort effects. Durations and attributes are updated using MCMC targeting the posterior defined above.

If several conditionally independent cross-sectional datasets $\{\tilde{Y}_{t_k}\}_k$ are available, the posterior becomes

$$P(\mathbf{X} \mid \{\tilde{Y}_{t_k}\}_k) \propto \left(\prod_k P(\tilde{Y}_{t_k} \mid \mathbf{X}) \right) P(\mathbf{X}),$$

The algorithm used to sample from the posterior follows Algorithm 1, extended to incorporate both the target log-density and the likelihood function. For each observed dataset and corresponding observation time, a likelihood term is defined based on the available variables. The accept–reject step is then performed using the log-posterior density, which combines the prior and likelihood contributions. Unlike the original formulation, which evaluates a single individual trajectory, the extended algorithm operates at the population level. The acceptance ratio therefore accounts for the joint posterior contribution of all individuals in the population. For the *Existence* dimension, the penalization term is also included in the acceptance ratio.

Overall, the methodology provides a flexible way to combine prior knowledge with individual-level observations when generating synthetic life trajectories. Disaggregate cross-sectional data are used to align the resulting population with observed characteristics while preserving the information encoded in the initial model. The approach can also accommodate incomplete or non-representative samples, broadening the range of data sources that can contribute to the generation process.

4 Results

In this section, we show how the methodology can be used to generate a synthetic population for Switzerland. Because the proposed framework relies on simulation-based techniques, the quality of the generated population depends primarily on the validity of the input models and on the ability of the sampling procedure to reproduce their implied distributions. Validation focuses on whether the generated population satisfies the objectives of the framework. More specifically, we address two main questions:

- **Does the implementation generate structurally feasible longitudinal populations consistent with the specified prior model?** We verify that the generated life trajectories are consistent with the probabilistic models provided as inputs and that the sampling procedure correctly reproduces the longitudinal dynamics implied by these models. We verify that trajectories satisfy the specified structural, biological, legal, and institutional constraints.
- **Does the inclusion of cross-sectional information improve population generation?** We assess whether heterogeneous cross-sectional data sources can effectively update and balance the prior longitudinal model in light of observed population characteristics. In particular, we examine whether a limited number of cross-sectional reference points is sufficient to adjust the simulation and produce coherent complete life trajectories. We also analyze to what extent using several datasets is more beneficial than a single one.

4.1 Model and data

4.1.1 Prior model

We consider a prior model to describe the Swiss population including life events that are relevant to the transport field as presented in Table 1. We model individuals' education, employment, driving license and residential spells trajectories using the following structural rules. A driving license, once obtained, is permanent. Education follows a sequential path: mandatory schooling must be completed by all individuals who survive to the required age; once studies are interrupted, they cannot be resumed; no repetition or skipping is allowed. Secondary education must last at least three years to grant access to tertiary education. Individuals may combine work and study. Residential moves occur only upon job changes, and transport mode availability is assumed to adjust with relocation, in line with the literature on mobility and life events (Beige and Axhausen, 2008; Scheiner, 2007).

The probabilistic specifications and constraints are based on Kukic et al. (2025), Baud and Bierlaire (2026) and are presented in Appendix 7.2.

Table 1: Model structure

Dimension	Associated Main Events	Attributes
Existence	Birth	Sex
Residence	Residential spells	Canton of residence
		Urban/rural
		Car availability
		GA availability[2]
Education	No education	$\emptyset$
	Mandatory education	$\emptyset$
	Secondary education	$\emptyset$
	Tertiary education	$\emptyset$
Work	No employment	$\emptyset$
	Employment spells	Income at start of work spell
		Place of Work
Driving license ownership	No driving license	$\emptyset$
	Driving license ownership	$\emptyset$

4.1.2 Data integration

We use cross-sectional data from the Swiss Mobility and Transport Microcensus from the 2010 and 2015 waves. The Mobility and Transport Microcensus (MTMC), conducted every five years by the Federal Office for Spatial Development and the Federal Statistical Office, is the largest national survey on travel behavior in Switzerland, with a sample exceeding 55,000 individuals. It provides a description of mobility patterns by combining information on socioeconomic characteristics of households and individuals, ownership and access to mobility tools such as private vehicles and public transport subscriptions.

To map our life trajectories, we identify among the alive individuals the active event at time t for each individual and extract the associated attributes. For each individual and each dimension, the end dates of the main and gap events are computed as the sum of their respective start dates and durations. An event is considered active at time t if t lies within its time interval, that is, if the start date is less than or equal to t and t is strictly smaller than the corresponding end date. We additionally derive the age of respondents as

$$\text{Age}_t = t - b, \quad b \text{ denoting the date of birth} \tag{2}$$

[2]GA stands for General Abonnement which is a Swiss subscription from SBB-CFF-FFS the Swiss operator giving access to the complete public transport network in all Switzerland.

An important feature of the MTMC dataset is that it does not contain observations for children younger than 7 years old. We therefore define the sampling rule as:

$$q(y) = \begin{cases} 0, & \text{if } y(\text{age}) < 7, \\ 1, & \text{otherwise.} \end{cases}$$

We compute the likelihood of the observed data only over the sample-eligible simulated individuals as described in the methodology. The absence of children younger than 7 in the MTMC dataset is not treated as evidence that such children are absent from the population. Instead, it is treated as a consequence of the sampling design.

4.2 Results

4.2.1 Valid longitudinal population

This section focuses on showing that the framework generates a valid longitudinal population. In order to do so, we sample 50,000 individuals from the prior model and show coherence at both the disaggregate (at the individual level) and aggregate levels (at the population level). Generating the 50,000 individuals – with a burn-in of 1,500 iterations for the *Existence* dimension and 6,500 iterations for the other dimensions – takes approximately 5 hours on a high-performance computing cluster using 55 parallel workers, corresponding to roughly 15 seconds per individual. Convergence diagnostics are available in the Appendix 7.5.

At the disaggregate level, Figure 2 shows Individual 1 trajectory. This individual's lifespan is slightly more than 80 years. They obtain a driving license at around 23, and graduate from secondary education at exactly 18. They start working at almost 16, and relocate five times: starting in Bern, then living in Zurich across three consecutive spells, before finally moving to Valais and then Geneva.

The structural constraints are satisfied by design: within each dimension, the sequence of events spans the full timeline without overlap. The institutional constraints are also respected - for instance, the individual only obtains a driving license after turning 18, and only starts working after turning 15. The sampled life trajectory of this individual is therefore coherent with both sets of constraints.

Moreover, at the aggregate level, the life trajectories follow the probabilistic specification, which validates the implementation. Figure 3 compares sampled values with their corresponding target densities for key variables, including date of birth, lifespan, age at labor market entry (tertiary graduates), and age at driving license acquisition. Sampled values respect the imposed constraints and follow the target densities. Again, no one gets a driving license before 18 or starts to work before 15, and no one lives more than 110 years.

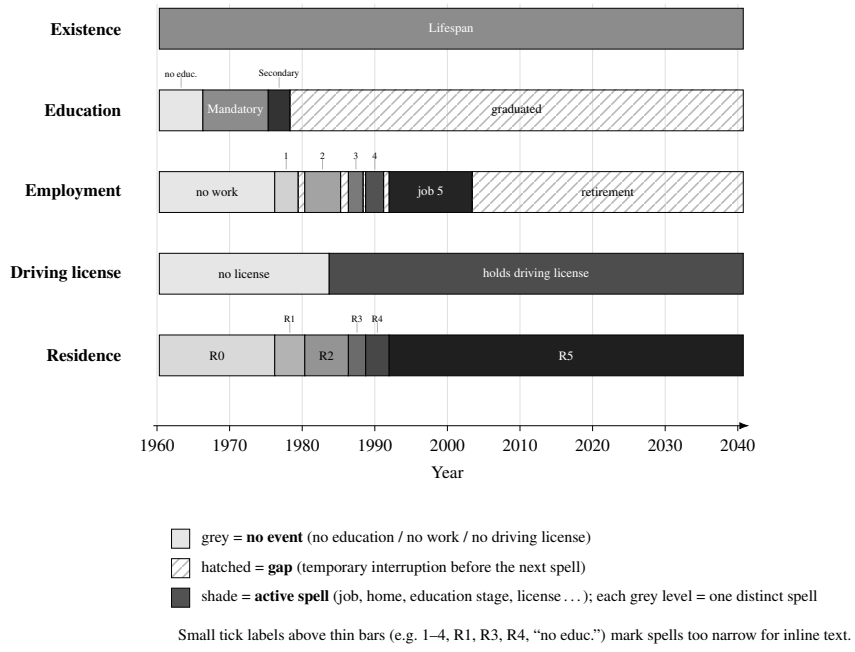


Figure 2: Life timeline of Individual 1

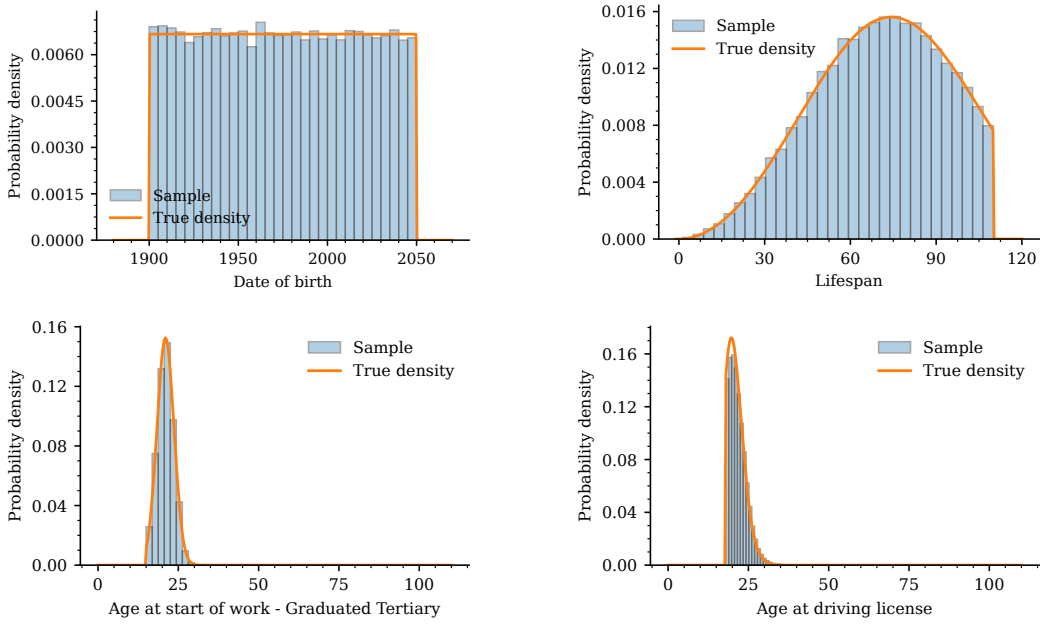


Figure 3: Empirical samples and theoretical densities for date of birth, lifespan, age at labor market entry (tertiary graduates), and age at driving license acquisition.

From these sampled life-trajectory variables, we can derive time-dependent cross-sectional datasets for any time t, tracking the same individuals. Taking individual 1 once more as an example, we can map their state in 2010. The person is alive and around 50 years old. They hold a driving license, have completed their education, and have retired from work. They now live in an urban area of Geneva and do not have access to a car. Such cross-sectional snapshot can be constructed for any sampled individual at any point in time.

At the population level, the implied age distribution at different dates matches the theoretical distributions obtained in closed form from the underlying birth and lifespan laws. Similarly, the distribution of place of residence and of car availability in 2010 are consistent with the specified density as shown in Figure 4. The income and car availability theoretical distributions are not tractable and are estimated using simulation. Indeed, such distributions depend on the distributions of other variables.. Therefore, once the other variables are sampled, we can estimate the theoretical density that should be observed in this sample.

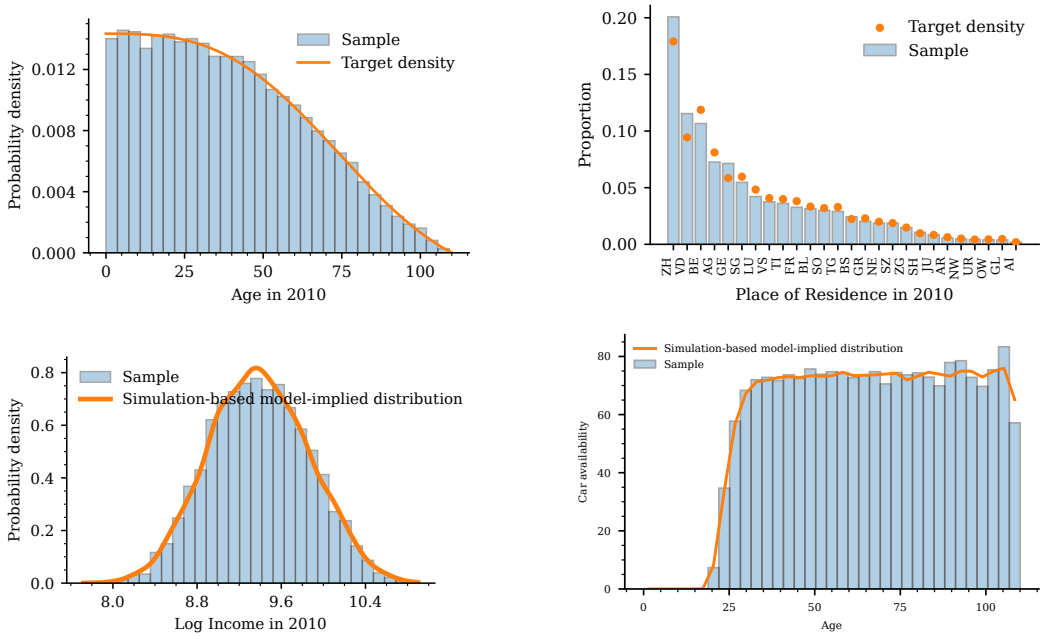


Figure 4: Age, canton of residence, log-income and car availability distributions in 2010

The prior sampling experiment confirms the internal coherence of the proposed framework along three points. First, the generated individual trajectories satisfy the structural constraints imposed by the prior models. Second, these trajectories reproduce the model-implied distributions of interest - event occurrences, durations, and associated attributes. Third, because individuals persist longitudi-

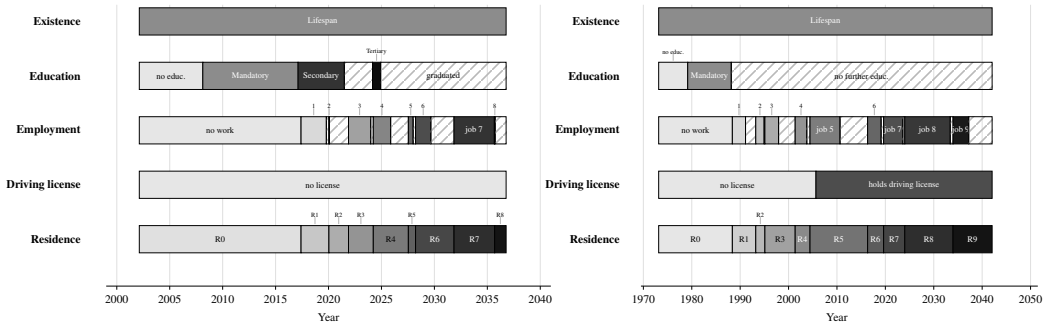
nally and can be located at any time t, the framework deterministically recovers the expected cross-sectional distributions of the population alive at t through the mapping mechanism.

4.2.2 Data integration improvements

We use the observed cross-sectional data and focus in this section on showing that the cross-sectional data effectively enable to infer longitudinal trajectories and that using multiple datasets brings more information than one. To do so, we use the cross-sectional data from the MTMC of 2010 and 2015 and compare the improvement brought by one and then the two datasets. The objective of the proposed framework is not to maximize predictive accuracy, but to provide a coherent generative representation of life trajectories that can integrate heterogeneous sources of cross-sectional information. Consequently, the framework is evaluated through its structural properties and its ability to recover coherent longitudinal populations, rather than through conventional out-of-sample prediction metrics.

The computational cost of sampling from the posterior of the *Existence* dimension is 45 minutes for 70,000 iterations for one dataset. This cost roughly doubles for two datasets. The computational cost of sampling the other dimensions from the posterior with a single dataset is approximately one second per chain iteration at the population level, on a high-performance computing cluster. This cost roughly doubles when two datasets are used, since the likelihood is computed separately for each dataset. The posterior chains for the other dimensions are run for 750,000 iterations.

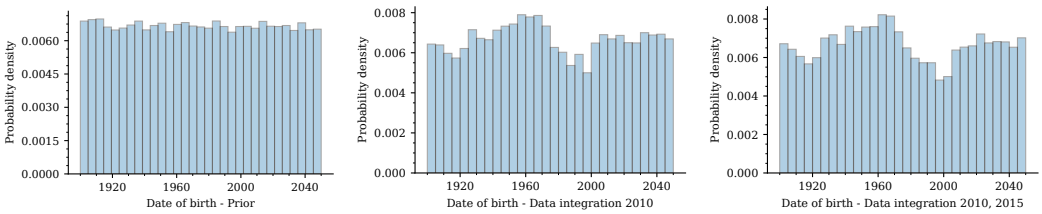
First, data integration makes it possible to generate complete synthetic life trajectories while using external population-level information to balance the prior model. The resulting individuals are therefore genuinely longitudinal: their entire life course is generated. The generated trajectories also satisfy the structural rules and constraints imposed by the model, such as a maximum lifespan of 110 years or the requirement that a driving license can only be obtained after the age of 18. Figure 5 illustrates trajectories resampled using 2010 data and using both 2010 and 2015 data. These trajectories should not, however, be compared as two versions of the same individual. Resampling is performed at the population level so that the characteristics of the synthetic population match the target distributions at the observation dates. The label "Individual 1" is therefore only an identifier within each generated population and does not imply any correspondence between the two trajectories. The distributional properties at the population level are presented hereafter.



(a) Using 2010 data (b) Using 2010–2015 data

Figure 5: Sampled life trajectories of Individual 1 using 2010 data; and using 2010 and 2015 data. The legend is the same as figure 2.

The Bayesian update is performed on the time-independent variables, yielding a posterior distribution in this space. Such modifications of the time-independent variables directly propagate in the projections at any time t. Figure 6 presents the date of birth distributions. It compares the prior distribution with the posterior distributions obtained using the 2010 data alone and using both the 2010 and 2015 datasets. The posterior distributions clearly depart from the uniform prior with births drops and peaks.

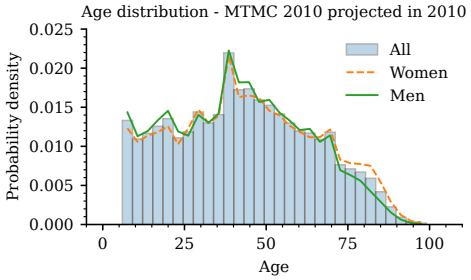


(a) Prior distribution (b) Posterior, 2010 (c) Posterior, 2010-2015

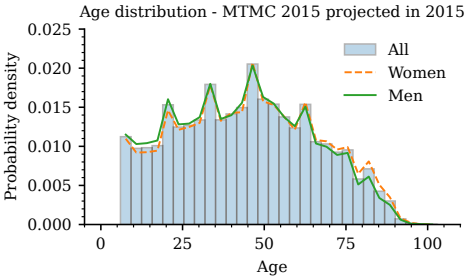
Figure 6: Date of birth distributions. The left panel shows the prior distribution. The middle and right panels show posterior updates, based on 2010 MTMC dataset only and on 2010 and 2015 MTMC datasets, respectively.

Since age is defined as age $=$ t $-$ date of birth, any modification in the distribution of dates of birth directly translates into a corresponding transformation of the age distribution at time t. The two distributions are therefore in one-to-one correspondence. The posterior over dates of birth induces a mirrored structure in the age distributions. This relationship is illustrated in Figure 7.

Figure 7 compares the observed and simulated age distributions in 2010 and 2015. The first row reports the observed distributions, the second row the prior model, the third row the posterior obtained using only the 2010 data, and the

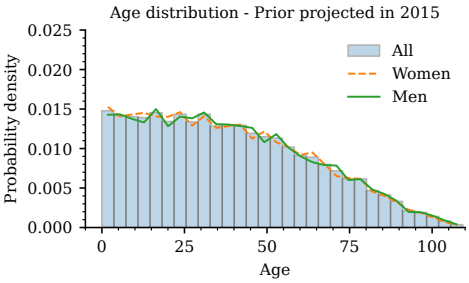


(a) 2010

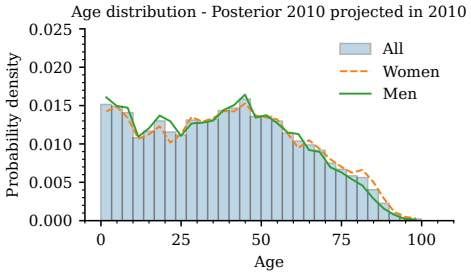


(b) 2015

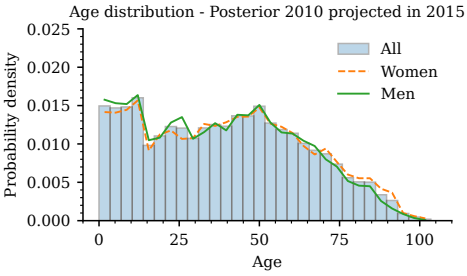
(c) Prior (2010)



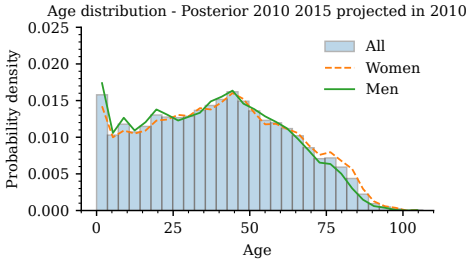
(d) Prior (2015)



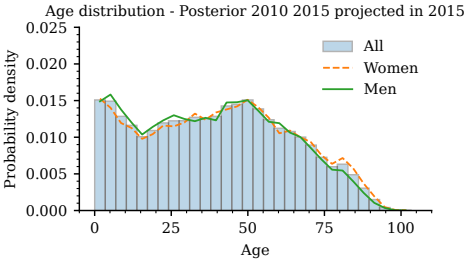
(e) Update with 2010 data (2010)



(f) Update with 2010 data (2015)



(g) Update with 2010+2015 data (2010)



(h) Update with 2010+2015 data (2015)

Figure 7: Age distributions. Columns correspond to years (2010 left, 2015 right). Rows correspond to observed data, prior, update with 2010 data, and update with 2010 and 2015 data. Update taking into account the sampling rule.

last row the posterior obtained using both datasets. When the sampling rule is accounted for, the posterior reproduces the observed age distribution for the population covered by the data, while preserving a coherent structure for unobserved age groups. In particular, children below age 7 remain present in the posterior population and retain a prior-like behavior, since they are not represented in the observations. Similarly, future births are not constrained by present observations and therefore continue to follow the prior. The update mainly affects the part of the trajectory that is informative for the observed cross-sections, which is then reflected in the age distributions through the relation age $= \mathrm{t} -$ date of birth. When the sampling rule is ignored, the absence of children aged 0 to 7 in the observations is incorrectly interpreted as evidence that these ages should have low probability in the population. This produces an artificial gap in the posterior age distribution, as displayed in Figure 14 in the Appendix. The *Existence* dimension posterior also captures correlations between variables such as Sex and Date of birth that were not present in the prior model, hence improving the generation of the population, as can be seen in Figure 7. Indeed, women are more represented than men in the older years, which is a trend observed in the data.

Table 2: Wasserstein distance (in years) between generated and observed age distributions (age > 7), with 95% bootstrap confidence intervals. Each generation procedure was mapped to a target survey year and compared against the observed age distribution for that year.

Generation procedure	2010 data	2015 data
Prior	2.69 [2.50, 2.91]	3.39 [3.14, 3.65]
Update (2010 only)	0.50 [0.36, 0.83]	1.60 [1.33, 1.90]
Update (2010–2015)	0.44 [0.35, 0.76]	0.65 [0.46, 0.96]

Table 2 reports the Wasserstein distance between generated and observed age distributions for each generation procedure, with 95% bootstrap confidence intervals. The prior distributions are further from the observed data than either update procedure, for both survey years, confirming that the update step meaningfully improves representativeness. Comparing the two update procedures, incorporating both 2010 and 2015 data yields the closest fit to observed data in both years. The improvement is asymmetric: for 2010, the confidence intervals of the two update procedures overlap (0.50 [0.36, 0.83] vs. 0.44 [0.35, 0.76]), indicating no statistically distinguishable gain from adding 2015 data when evaluating against 2010. For 2015, however, the joint procedure produces a distance roughly 2.5 times smaller than the 2010-only procedure, with non-overlapping confidence intervals (1.60 [1.33, 1.90] vs. 0.65 [0.46, 0.96]), indicating a robust improvement.

This means that combining multiple years of data primarily benefits generalization to years not directly used for that specific fit, rather than uniformly improving performance everywhere. Additionally, since the age likelihood kernel (more details in Appendix 7.3) is $\mathcal{N}(\tilde{a}; a, 1)$, observed ages are modeled as draws from the synthetic age distribution $\mu_t^{\mathbf{X}}$ convolved with $\mathcal{N}(0, 1)$. Under this specification, a coupling argument gives (Villani, 2009)

$$W_1(\mu_t^{\mathbf{X}}, \mu_t^{\mathbf{X}} * \varphi) \leq \sqrt{\frac{2}{\pi}} \approx 0.8.$$

That means the kernel's own smoothing imposes an expected resolution limit of under one year on any in-sample comparison, independent of remaining model misfit. All update procedures evaluated in-sample (Table 2) fall below this bound suggesting the calibration is near the resolution allowed by the likelihood specification itself. The single out-of-sample comparison (2010-only update evaluated against 2015 data) is not subject to this bound and shows a substantially larger distance (1.60 years).

Once the *Existence* dimension is sampled, we sample the other dimensions. Two sub-populations emerge: the ones that are alive at the time of observation and the ones that are not (either already dead, or not born yet).

Individuals not alive at the observation dates For individuals who are not alive at the time of observation, their characteristics are sampled from the prior model, as no data are available to inform these attributes. Indeed, only the characteristics of alive individuals-such as employment or education status-are observed, and therefore the data provide no information on these variables for individuals outside the observation period. Consequently, the sampled characteristics of unobserved individuals follow the distributions specified by the prior.

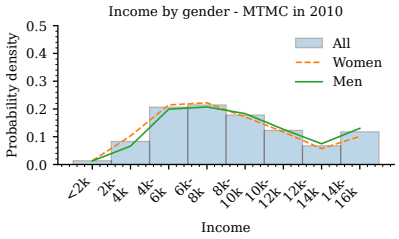
Alive population For the alive individuals at the time of observation, the dimension-specific characteristics are sampled from the posterior. In Table 3, we present the employment statuses for different simulations. *Observed data* corresponds to the MTMC data, while *Mapped year* indicates the observation year. For all simulated models, the column *Existence update* specifies which updated dates of birth and lifespans were used to generate the population. For example, the *Prior* model with an *Existence update* from 2010 uses the dates of birth and lifespans inferred from the 2010 update, while employment trajectories are generated entirely from the prior model. This design ensures that all models are compared on the same underlying demographic structure. Without updating the existence process, differences in the age composition and survival of the population would propagate to employment outcomes, making it harder to distinguish errors due to demographic

mismatches from those due to the employment model itself. The prior model underestimates the proportion of employed individuals across all periods, which is partly corrected by the posterior sampling. Since the demographic structure is fixed, some categories have limited flexibility. In particular, the *Retired* category can only be adjusted within the constraints imposed by the age structure, while the *Age* $<$ *15* category is entirely determined by age and therefore remains unchanged. The Appendix also presents the same Table differentiated by Sex (Section 7.6.9) which shows that the posterior also captures the different employment-status dynamics between men and women, notably higher employment levels among men and higher proportions of retired women.

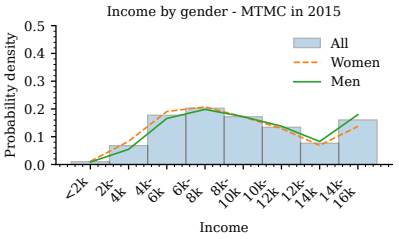
Table 3: Comparison of employment-status distributions across simulation settings, restricted to individuals aged > 7, in %

Model	**Existence update**	**Data update**	**Mapped year**	**Employed**	**Unemployed**	**Retired**	**Age** <15
Observed data	–	–	2010	60.36	13.38	18.23	8.02
Prior	2010	–	2010	32.67	38.34	18.22	10.77
Posterior	2010	2010	2010	44.48	26.67	18.07	10.77
Prior	2010–2015	–	2010	33.12	38.91	17.87	10.10
Posterior	2010–2015	2010–2015	2010	45.05	27.24	17.60	10.10
Observed data	–	–	2015	60.32	12.48	19.54	7.60
Prior	2010	–	2015	31.25	35.97	19.97	12.82
Posterior	2010	2010	2015	36.96	31.05	19.17	12.82
Prior	2010–2015	–	2015	31.54	37.98	19.88	10.60
Posterior	2010–2015	2010–2015	2015	43.18	26.99	19.23	10.60

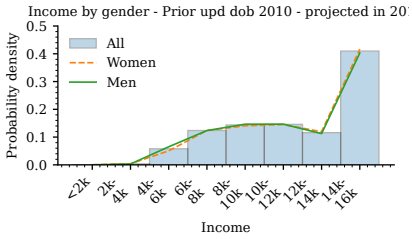
In addition to employment status, Figure 8 presents the income distribution by sex. The prior model appears poorly calibrated, as it produces a pronounced peak at high income levels that is not observed in the data. Although income in the prior model does not directly depend on sex, the resulting distributions show women earning slightly more than men. This effect is mainly driven by differences in the age structure: women are less represented among younger individuals and more represented among older individuals. Since the prior income model depends on work experience (years spent working), older individuals tend to receive higher salaries on average. As a result, women are under-represented at lower income levels in the prior simulation. This contrasts with the observed data, where women earn less than men on average, being more represented at low income levels and less represented at high income levels. The posterior distributions based on the 2010 data, and on both the 2010 and 2015 data, mitigate these issues by reducing the over-representation of high salaries. It also better captures the observed association between sex and income, with women slightly more concentrated in the lower income categories and men slightly more concentrated in the higher income categories. Income per age category figures are also available in the Appendix (Figure 16) that show how the posterior corrects the income distribution not only by sex but also by age group.



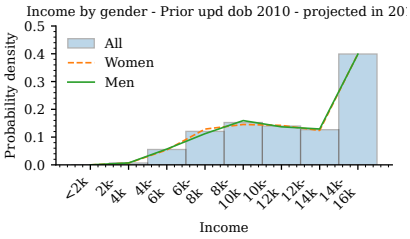
(a) 2010



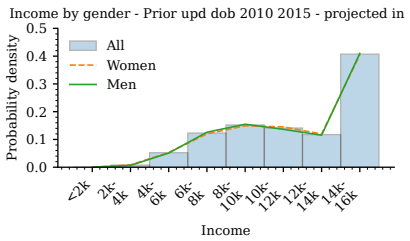
(b) 2015



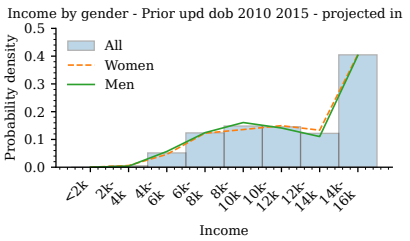
(c) Prior with 2010 data (2010)



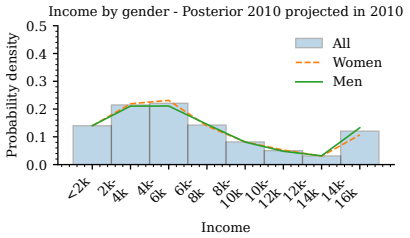
(d) Prior with 2010 data (2015)



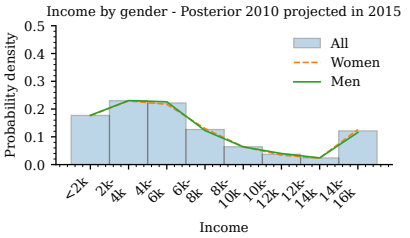
(e) Prior with 2015 data (2010)



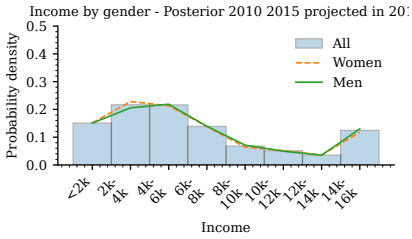
(f) Prior with 2015 data (2015)



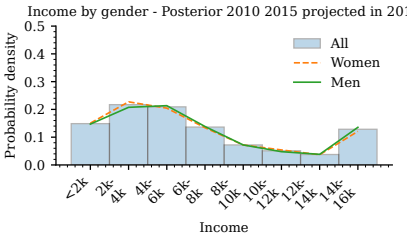
(g) Update with 2010 data (2010)



(h) Update with 2010 data (2015)



(i) Update with 2010+2015 data (2010)



(j) Update with 2010+2015 data (2015)

Figure 8: Income category shares by gender. Columns correspond to years (2010 left, 2015 right). Rows correspond to observed data, priors with updated demographic structure, full update with 2010 data, and full update with 2010 and 2015 data.

Additional results are available in the Appendix for other variables and dimensions and show that the integration of cross-sectional data improves the generated population. However, the driving-license dimension provides an informative illustration of the scope and limitations of the proposed framework. Table 4 reports the share of individuals holding a driving license in the total population, as well as separately among women and men. The prior underestimates license ownership overall. It also produces an unrealistic sex pattern: the share of women holding a license is higher than the share of men holding a license, whereas the opposite is observed in the data. This discrepancy can again be related to the age structure of the simulated population. Since women are, on average, older in the prior population, they are more likely to belong to age groups in which the license-acquisition process has already occurred. The posterior remains very close to the prior and fails to recover the observed gender gap. Rather than reproducing the higher prevalence among men, it mainly acts to equalize the shares between women and men. These results are expected. As described previously, the occurrence of lifetime events is sampled from the prior distribution because cross-sectional data only inform the prevalence of a characteristic at a given time, not the proportion of individuals who will ever experience the event. In our experiments, we assumed that 85% of individuals eventually obtain a driving license, independently of gender or birth cohort. While this simplifying assumption is sufficient to illustrate the methodology, it does not reflect the historical evolution of driving license acquisition, which differs across generations and between men and women.

Table 4: Comparison of driving-license ownership across simulation settings, among individuals aged > 18, in %

Model	Existence update	Data update	Mapped year	License, all	License, women	License, men
Observed data	–	–	2010	81.89	74.72	89.49
Prior	2010	–	2010	79.46	80.14	78.74
Posterior	2010	2010	2010	77.93	77.97	77.88
Prior	2010–2015	–	2010	79.73	80.38	79.04
Posterior	2010–2015	2010–2015	2010	78.54	78.60	78.47
Observed data	–	–	2015	83.25	77.13	89.63
Prior	2010	–	2015	79.70	80.15	79.22
Posterior	2010	2010	2015	78.20	78.26	78.14
Prior	2010–2015	–	2015	80.11	80.49	79.72
Posterior	2010–2015	2010–2015	2015	79.02	79.38	78.64

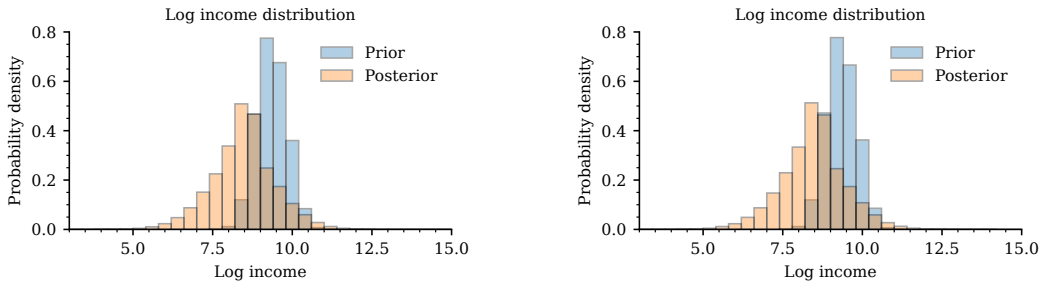
This misspecification has direct consequences for the posterior distribution. The driving license dimension consists of only two states—without a driving license and with a driving license—and, for individuals who eventually obtain one, the sampled acquisition age entirely determines their status at every point in time. In the observed data, older women in 2010 and 2015 are much less likely to hold a driving license than younger cohorts. Under our prior, however, 85% of these

individuals are assumed to obtain a license during their lifetime. Since the model also imposes the structural constraint that licenses must be acquired before the age of 70, these individuals cannot simply postpone the acquisition to match the cross-sectional observations. Consequently, the posterior is constrained by the prior assumptions, despite the evidence contained in the likelihood. Additional results in the Appendix (Figures 17 and 18) show this age mismatch effect.

This result illustrates a limitation of the framework: the posterior cannot recover characteristics that are excluded or inadequately represented by the prior and structural model. Attention should be paid to the temporal validity of the prior model. In this application, a more realistic specification would allow the probability of ever obtaining a driving license to depend on birth cohort and sex, accounting for historical changes in driving license ownership.

Time-independent updated population For the Bayesian update, the population was split into two groups. The first group consists of individuals who cannot be observed in the data because, at the time of observation, they have either already passed away or have not yet been born. The second group consists of individuals who are alive at the time of observation and can therefore be observed in the data. For the latter group, the variables are sampled from the posterior distribution, which combines the prior with the likelihood. For the former group, the variables are sampled from the prior only, since the data are uninformative for these individuals. As a result, when all individuals are aggregated, two distinct trends can appear in the time-independent variables: one corresponding to individuals sampled from the prior, and one corresponding to individuals sampled from the posterior.

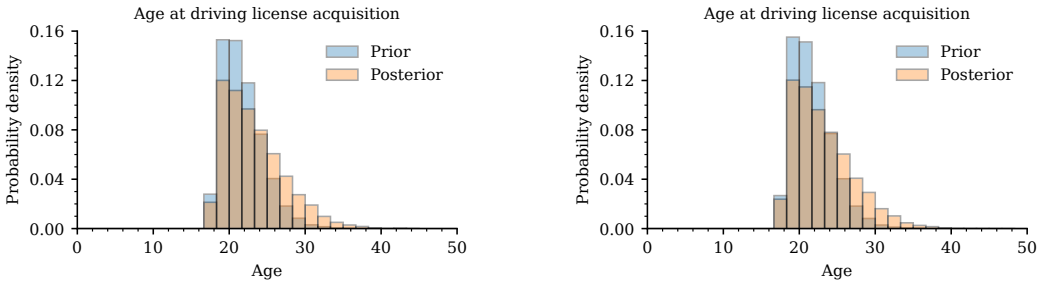
Figure 9 shows the log income distribution in the resulting longitudinal population. Two clear trends can be observed between the prior and posterior distributions. The posterior is shifted to the left, indicating overall lower incomes, and is slightly more dispersed, suggesting greater variability than in the prior.



(a) Using 2010 data

(b) Using 2010 and 2015 data

Figure 9: Log Income in the updated population.



(a) Using 2010 data

(b) Using 2010 and 2015 data

Figure 10: Age at driving license acquisition in the updated population.

Similarly, Figure 10 shows two trends in the age at driving license acquisition, with the posterior displaying a longer tail.

Overall, incorporating cross-sectional data substantially improves the generation of the population. These data provide reference points for the observed population at different moments in time. The primitive variables governing individual life trajectories are resampled so that the simulated population better matches the observed trends, and these modifications propagate to all future years in which the individuals are present. Individual trajectories are adjusted both in terms of when events occur and how they occur, for example by postponing or advancing the start of an event, or by modifying its attributes.

Using multiple cross-sectional datasets is particularly desirable because each dataset provides information at a different point in time and therefore contributes to the readjustment of the trajectories. Each observation reshapes part of the past, and these changes also propagate into the future. However, since each dataset only informs a limited temporal window, combining several cross-sectional sources helps reshape longer time horizons. It also leads to a better trade-off between the information contained in the different datasets, rather than relying too strongly on a single source.

5 Discussion

This paper shows that it is possible to generate longitudinal synthetic populations from cross-sectional data by fundamentally changing the representation of individuals: instead of representing them through their observed states at a given point in time, we represent them through their complete life trajectories. The approach is based on a unified representation of individual life courses, decomposed into dimensions along which we can structure constraints to ensure structural feasibility and temporal coherence of the generated trajectories. The population state at any point in time is obtained through a deterministic mapping. This representation provides a coherent way to describe the same synthetic population over time, without reconstructing individuals independently at each time period; and therefore creates true panel data. Demographic evolution emerges endogenously from individual histories, rather than being imposed separately at each time period.

The approach is also general and model-agnostic. Any process that can be expressed within a life-trajectory representation can be incorporated. The same formal structure can be adapted to different temporal resolutions, spatial scales, and application domains. Cross-sectional data are used to update the distribution of synthetic trajectory variables conditional on the prior model. It can incorporate multiple datasets observed at different points in time to enrich the prior model with empirical evidence. At the same time, the structural constraints of the life trajectories are preserved, so that the updated population remains feasible and temporally coherent.

As with any general modeling framework, some aspects define the current scope of the approach and open directions for future extensions. First, individuals are currently modeled as independent entities. The individual is therefore the highest-level unit of the population, which means that correlations across individuals, household formation, and intra-household interactions are not explicitly represented. Extending the framework to include household-level structures is a natural and important direction for future work. Second, the framework requires detailed input information, since it aims to describe complete life trajectories rather than isolated cross-sectional observations. The framework is not tied to a specific observation date, hence the term "time-independent", used to refer to a higher-level generative mechanism. In practice, however, the prior models available for demographic, econometric, or transport-related processes are usually calibrated on a given time interval and geographical area. Therefore, the practical implementation of the framework inherits the temporal and spatial validity of these input models. For population generation over long time horizons, using several sub-models, each valid over a specific time range, is recommended instead of relying on a single model for the entire period. Finally, a natural next step would

be to use longitudinal data to improve the generation of the life trajectories. By having access to longitudinal data, we could improve the updating power of the Bayesian step as we would have much richer information.

6 Acknowledgements

This research is funded by ERC Project: 101264336 – COBRA: Combinatorial Optimization for Behavioral Response Analysis

7 Appendix

7.1 Constraints and linear algebra representation

7.1.1 Constraints

We define here the **Bound Constraints** to ensure trajectories remain admissible with respect to biological and legal rules. Notice that since ages and start dates satisfy

$$\text{Date at age} = \text{Date of birth} + \text{Age},$$

it is equivalent to writing the constraints in terms of age. For clarity, all start-date constraints are expressed in age units.

Notation conventions are the following:

e_k is the notation for event k

e_{e_k} is the main part of event k

g_{e_k} is the gap part of event k

$d^{(e_{e_k})}$ is the duration of the main part of the event k

$d^{(g_{e_k})}$ is the duration of the gap part of the event k

$\tilde{i}^{(e_k)}$ is the sampled indicator of the event k occurring or not

$d(e_k)$ is the total duration of event k, meaning the sum of the main duration and the gap duration

$a_{\_*}$ refers to ages

The framework supports several classes of temporal constraints. Starting-age constraints determine when a state may begin, duration constraints restrict how long a state may persist, and saturation constraints impose a minimum duration only when progression to a subsequent event is observed. The latter distinction is important because an event may be interrupted by the end of the individual's lifespan and should therefore not necessarily be required to reach its usual minimum duration.

The **Bound** constraints that can be specified are the following:

1. **Main event minimal starting age:** This constraint prevents the main part of an event from starting before a prescribed age. It can represent biological,

legal, or institutional eligibility conditions. For example, mandatory education cannot begin before the minimum school-entry age, and employment cannot begin before the minimum legal working age.

The starting age of the main part of event e_{n_e} is obtained by summing the durations of the events that precede it. The corresponding lower-bound constraint is

$$\sum_{k=1}^{n_e} \tilde{i}^{(e_k)} \cdot (d^{(e_{e_{k-1}})} + d^{(g_{e_{k-1}})}) \geq a_{\text{min, main}},$$

The occurrence indicators ensure that only realized events contribute to the starting age. Because the events form an ordered sequence, the occurrence of event e_k implies that all preceding events have also occurred. This sequential structure may equivalently be represented as

$$\tilde{i}^{(e_k)} = \prod_{l=1}^{k} \tilde{i}^{(e_l)}.$$

2. **Main event maximal starting age:** This constraint prevents the main part of an event from beginning after a prescribed age. It is useful when entry into a state is only meaningful or institutionally possible within a limited period of life. For example, it may be desirable to exclude trajectories in which an individual enters mandatory education at an implausibly late age.

   The constraint is written as

$$\sum_{k=1}^{n_e} \tilde{i}^{(e_k)} \cdot (d^{(e_{e_{k-1}})} + d^{(g_{e_{k-1}})}) \leq a_{\text{max, main}}$$

   Together, the minimum and maximum starting-age constraints define the admissible age interval within which the main part of an event may begin.

3. **Main event minimal saturation duration:** A strictly positive minimum duration cannot generally be imposed on every realized event. An individual may die before completing the event, in which case the final event of the trajectory is truncated by the lifespan constraint. For example, an individual may begin an education stage but die before its normal completion.

   Nevertheless, progression to a subsequent event may imply that the preceding event has been completed to a sufficient extent. For instance, entry into secondary education requires that a minimum amount of time has been spent in mandatory education. The minimum duration is therefore imposed only when both the current event e_k and the subsequent event e_{k+1} occur.

This conditional minimum-duration requirement is expressed as: [3]

$$\tilde{i}^{(e_k)} \cdot \tilde{i}^{(e_{k+1})} d^{(e_k)} \geq \tilde{i}^{(e)} \cdot \tilde{i}^{(e_{k+1})} \cdot d^{(e_k)}_{\text{min, sat}}$$

If the subsequent event does not occur, the constraint becomes inactive. This allows the current event to be truncated by death or by the end of the individual's existence interval.

4. **Main event maximal duration:** This constraint limits the amount of time that may be spent in the main part of an event. It can be used to exclude implausibly long states or to represent institutional limits. For example, the duration of a particular education stage may be bounded above.

   The constraint is

   $$\tilde{i}^{(e)} \cdot d^{(e)} \leq \tilde{i}^{(e)} \cdot d_{\text{max, main}}$$

   When event e_k occurs, its main duration is bounded by $d^{(e_k)}_{\text{max, main}}$. When it does not occur, its duration is constrained to zero by the structural constraints of the framework.

5. **Gap event minimal starting age:** In some applications, it is necessary to constrain not only the beginning of the substantive state, but also the beginning of the gap that follows it. A gap may represent an interruption, inactivity period, or terminal state. For example, retirement may be represented by the final gap of the employment dimension. Imposing a minimum starting age on this gap ensures that retirement cannot begin before a prescribed age.

   The starting age of the gap associated with event e_{n_e} includes the initial *No event* duration, the durations of the preceding events, and the main duration of the current event. The corresponding constraint is

   $$d^{(e_0)} + \sum_{k=1}^{n_e} \tilde{i}^{(e_k)} \cdot (d^{(g_{e_{k-1}})} + d^{(e_{e_k})}) \geq a_{\text{min, gap}}$$

6. **Gap event maximal starting age:** This constraint imposes an upper bound on the age at which a gap may begin. It is relevant when the transition out of a substantive state must occur before a given age. For example, it may be used to require the final employment spell to end, and hence retirement to begin, before a maximum retirement age.

[3] Note that we can also write it as $d^{(e_k)} \geq \tilde{i}^{(e_k)} \cdot \tilde{i}^{(e_{k+1})} \cdot d^{(e_k)}_{\text{min, sat}}$, when either indicator is zero, the right-hand side is zero and the constraint reduces to the non-negativity condition already imposed by the framework.

The constraint is

$$d^{(e_0)} + \sum_{k=1}^{n_e} \tilde{i}^{(e_k)} \cdot (d^{(g_{e_{k-1}})} + d^{(e_{e_k})}) \leq a_{\max,\,\text{gap}}$$

The minimum and maximum gap-starting ages jointly define the admissible interval within which the gap may begin.

7. **Gap event minimal saturation duration:** As for the main part of an event, a minimum gap duration should not necessarily be imposed whenever the gap exists, since the individual's lifespan may end before the gap is completed. A minimum duration becomes necessary, however, when the individual subsequently enters another event.

   This constraint is useful for representing required waiting periods between consecutive states. For example, a biological minimum interval may be imposed between two pregnancies, or an institutional waiting period may be required before an individual can enter a subsequent state.

   The minimum duration is activated only when both event e_k and the subsequent event e_{k+1} occur[4]:

$$\tilde{i}^{(e_k)} \cdot \tilde{i}^{(e_{k+1})} d^{(g_k)} \geq \tilde{i}^{(e_k)} \cdot \tilde{i}^{(e_{k+1})} \cdot d^{(g_k)}_{\min,\,\text{sat}}$$

   If no subsequent event occurs, the condition is inactive, allowing the final gap to be truncated by the end of the lifespan.

8. **Gap event maximal duration:** This constraint limits the duration of the gap associated with an event. It can be used to prevent unrealistically long interruptions or to represent a maximum admissible waiting period between consecutive states. For example, a model may restrict the maximum duration of unemployment between two employment spells.

   The constraint is

$$\tilde{i}^{(e_k)} \cdot d^{(g_k)} \leq \tilde{i}^{(e_k)} \cdot d_{\max,\text{gap}}$$

   When event e_k occurs, the duration of its associated gap is bounded above. Otherwise, the corresponding duration is zero.

[4]Note that we can also write it as $d^{(g_k)} \geq \tilde{i}^{(e_k)} \cdot \tilde{i}^{(e_{k+1})} \cdot d^{(g_k)}_{\min,\,\text{sat}}$, when either event does not occur, the right-hand side is zero and the constraint reduces to the non-negativity condition already imposed by the framework.

9. **Total event maximal duration:** In some cases, the relevant quantity is not the duration of the main part or the gap separately, but their combined duration. This constraint limits the total time associated with an event, including both the substantive state and the gap that follows it.

   Denoting the total duration of event e_k by

   $$d(e_k) = d^{(e_k)} + d^{(g_k)},$$

   the constraint is

   $$\tilde{\imath}^{(e_k)} \cdot d(e_k) \leq \tilde{\imath}^{(e_k)} \cdot d_{\max}$$

   This formulation is useful when flexibility is allowed in the allocation of time between the main and gap components, while the entire event must remain within a prescribed maximum duration.

We also define **interdependence constraints**, which express temporal relationships between events belonging to the same dimension or to different dimensions. Unlike the preceding constraints, which restrict a single event in isolation, interdependence constraints link several event durations and therefore allow dependencies between different components of an individual's life trajectory to be represented.

For example, the duration of the initial residence spell may be constrained by the duration of the initial period without education or employment. Such a constraint can represent the assumption that an individual leaves their initial household no later than the beginning of their first education or employment episode. The initial residence spell may nevertheless end earlier, so the constraint does not require the residential move to coincide exactly with either transition.

Let d_i denote the duration to be constrained. Depending on the application, d_i may represent a main-event duration, a gap duration, or a total event duration. Let $(d_k)_k$ denote the durations on which d_i depends. The framework supports the following affine interdependence constraints:

1. **Linear transformation equality.**

   An equality constraint imposes an exact affine relationship between the duration d_i and a collection of other durations:

   $$d_i = \sum_k a_k d_k + b.$$

   This formulation is appropriate when the timing or duration of one event is fully determined by other components of the trajectory. For example,

it can be used to require two events in different dimensions to end at the same time, or to impose a fixed temporal offset between them. The coefficients a_k determine how each constraining duration contributes to d_i, while b represents a constant temporal offset.

2. **Linear transformation inequality.**

   An inequality constraint imposes an upper bound on d_i as an affine function of other durations:

$$d_i \leq \sum_k a_k d_k + b.$$

   This formulation is appropriate when the duration or timing of one event must not exceed that implied by another part of the trajectory, while earlier transitions remain admissible. In the residence example, such a constraint ensures that the initial residence spell ends no later than the relevant education or employment transition, without forcing the move to occur exactly at that time.

7.1.2 Linear algebra representation

All the constraints above translate in matrix representations since there are only linear constraints. Let us denote by

$$d = \begin{pmatrix} d^{(e_{\text{no event}})} \\ d^{(g_{\text{no event}})} \\ d^{(e_1)} \\ d^{(e_2)} \\ \vdots \\ d^{(e_n)} \end{pmatrix} \quad \text{the vector of durations.}$$

Equality constraints There are two design equality constraints, the gap of the no event is equal to zero, and the sum of all durations is equal to the lifespan to ensure that in each dimension, we have a partition of the lifespan by the intervals of event durations.

$$A_{eq} = \begin{pmatrix} 0 & \tilde{i}^{(g_{\text{no event}})} & 0 & .. & 0 \\ \tilde{i}^{(e_{\text{no event}})} & \tilde{i}^{(g_{\text{no event}})} & \tilde{i}^{(e_1)} & .. & \tilde{i}^{(g_n)} \end{pmatrix}, \quad b_{eq} = \begin{pmatrix} 0 \\ \text{lifespan} \end{pmatrix}, \quad A_{eq} \cdot d = b_{eq}$$

Remark: In the second line of the vector b, there is "lifespan", but rigorously, it should be $\max_k(\tilde{i}^{(e_k)}) \cdot$ lifespan. If there are no events, the maximum of the indicators is zero, and therefore we have the constraint $0 \leq 0$, thus the constraint is

inactive. In this framework, the "No event" is mandatory and has an indicator one, therefore, the maximum is always one, which is why for notational simplicity the max is omitted.

This can be rewritten using inequality matrix

$$A_{ineq} = \begin{pmatrix} 0 & \tilde{i}^{(g_{\text{no event}})} & 0 & \cdots & 0 \\ 0 & -\tilde{i}^{(g_{\text{no event}})} & 0 & \cdots & 0 \\ \tilde{i}^{(e_{\text{no event}})} & \tilde{i}^{(g_{\text{no event}})} & \tilde{i}^{(e_1)} & \dots & \tilde{i}^{(g_n)} \\ -\tilde{i}^{(e_{\text{no event}})} & -\tilde{i}^{(g_{\text{no event}})} & -\tilde{i}^{(e_1)} & \dots & -\tilde{i}^{(g_n)} \end{pmatrix},$$

$$b_{ineq} = \begin{pmatrix} 0 \\ 0 \\ \text{lifespan} \\ -\text{lifespan} \end{pmatrix}, \quad A_{ineq} \cdot d \leq b_{eq}$$

Large inequality constraints There are inequalities inherited from the structural constraints, and other from the specified bound constraints. Let's start with the structural ones: all the durations must be greater or equal than zero.

$$A_{\text{ineq, gaps}} = \begin{pmatrix} -\tilde{i}^{(g_{\text{no event}})} & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -\tilde{i}^{(g_{\text{no event}})} & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & -\tilde{i}^{(e_1)} & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & -\tilde{i}^{(g_1)} & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\tilde{i}^{(g_n)} \end{pmatrix},$$

$$b_{\text{ineq,gaps}} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad A_{\text{ineq, gaps}} \cdot d \leq b_{\text{ineq, gaps}}$$

All durations must be lower or equal than the lifespan.

$$A_{\text{ineq, lifespan bound}} = \begin{pmatrix} \tilde{i}^{(e_{\text{no event}})} & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \tilde{i}^{(e_{\text{no event}})} & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \tilde{i}^{(e_1)} & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \tilde{i}^{(e_1)} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & \tilde{i}^{(e_n)} \end{pmatrix},$$

$$b_{eq} = \begin{pmatrix} \text{lifespan} \\ \text{lifespan} \\ \text{lifespan} \\ \text{lifespan} \\ \vdots \\ \text{lifespan} \end{pmatrix}, \quad A_{\text{ineq, lifespan bound}} \cdot d \leq b_{eq}$$

which is equivalent to adding the following row in the inequality matrix for each duration

$$row = \begin{pmatrix} 0 & 0 & \underbrace{\tilde{i}^{(e)}}_{\text{index of the duration concerned}} & 0 & \cdots & 0 \end{pmatrix}, \quad value = \text{lifespan}, \quad row \cdot d \leq value$$

Then, by design, the gap event of the "No Event" must have a duration equal to zero, which imposes the additional following constraint (written only for the row):

$$row = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \end{pmatrix}, \quad value = 0, \quad row \cdot d \leq value$$

By design, the events for which the indicator is equal to zero have durations for the main and for the gap that are constrained to zero. Since those events don't occur in the person's life, their duration is zero. This implies the following constraint

$$row = \begin{pmatrix} 0 & 0 & \cdots & (1 - \tilde{i}^{(e)}) & 0 & \cdots 0 \end{pmatrix}, \quad value = 0, \quad row \cdot d \leq value$$

$$row = \begin{pmatrix} 0 & 0 & \cdots & 0 & (1 - \tilde{i}^{(g)}) & 0 & \cdots 0 \end{pmatrix}, \quad value = 0, \quad row \cdot d \leq value$$

In the case where the indicator is zero, (1 - i) is equal to one, and therefore the constraint is active, meaning the upper bound for the duration is zero. Since the lower bound is also zero, this constrains the duration to be equal to zero. In the case where the indicator is one, (1-i) is equal to zero, giving the inequality $0 \leq 0$, which is an inactive constraint.

For the bound constraints, we adopt the following convention. The first event of the chain is the No Event, but is considered "event 0", meaning an index of 0 for the main duration of the No Event. Therefore, the k-th event is event of index k. The 0-th event is the No event, the 1st event is the first event after the No event etc.

First, a minimal starting age constraint on the event k can be translated by the following row:

$$row = \begin{pmatrix} \underbrace{-\tilde{i}^{(e_1)} & -\tilde{i}^{(e_1)} & -\tilde{i}^{(e_2)} & -\tilde{i}^{(e_2)} & \cdots}_{2k} & 0 & 0 \end{pmatrix},$$

$$value = -a_{\text{min, main}}, \quad row \cdot d \leq value$$

There are 2k values of -1 because we have to start after both the main durations and the gap durations for each preceding event. Similarly, for a maximal starting age constraint, we have

$$row = \left(\underbrace{\tilde{i}^{(e_1)} \quad \tilde{i}^{(e_1)} \quad \tilde{i}^{(e_2)} \quad \tilde{i}^{(e_2)} \quad \cdots}_{2k} \quad 0 \quad 0 \right)$$

$$value = a_{\text{max, main}}, \quad row \cdot d \leq value$$

A minimal saturating duration on the main event translates as follows

$$row = \left(0 \quad 0 \quad \underbrace{-\tilde{i}^{(e_k)} \cdot \tilde{i}^{(e_{k+1})}}_{\text{index } 2k} \quad 0 \quad 0 \quad \cdots \quad 0 \right),$$

$$value = -\tilde{i}^{(e_k)} \cdot \tilde{i}^{(e_{k+1})} \cdot d^{(e_k)}_{\text{main, min}}, \quad row \cdot d \leq value$$

A maximal duration on the main event translates as follows

$$row = \left(0 \quad 0 \quad \underbrace{\tilde{i}^{(e_k)}}_{\text{index } 2k} \quad 0 \quad 0 \quad \cdots \quad 0 \right),$$

$$value = \tilde{i}^{(e_k)} \cdot d_{\text{main, max}}, \quad row \cdot d \leq value$$

There are similar constraints on the gaps. For the k-th event, a minimal starting age of the gap can be rewritten as follows

$$row = \left(\underbrace{-1 \quad -\tilde{i}^{(e_1)} \quad -\tilde{i}^{(e_1)} \quad -\tilde{i}^{(e_2)} \quad -\tilde{i}^{(e_2)} \quad -\tilde{i}^{(e_3)} \quad \cdots}_{2k+1} \quad 0 \, 0 \right),$$

$$value = -a_{\text{min, gap}}, \quad row \cdot d \leq value$$

and a maximal starting age for the gap of k-th event can be rewritten as follows

$$row = \left(\underbrace{1 \quad \tilde{i}^{(e_1)} \quad \tilde{i}^{(e_1)} \quad \tilde{i}^{(e_2)} \quad \tilde{i}^{(e_2)} \quad \tilde{i}^{(e_3)} \quad \cdots}_{2k+1} \quad 0 \quad 0 \right),$$

$$value = a_{\text{max, gap}}, \quad row \cdot d \leq value$$

A minimal saturating duration on the gap event translates as follows

$$row = \left(0 \quad 0 \quad \underbrace{-\tilde{i}^{(e_k)} \cdot \tilde{i}^{(e_{k+1})}}_{\text{index } 2k+1} \quad 0 \quad 0 \quad \cdots \quad 0 \right),$$

$$value = -\tilde{i}^{(e_k)} \cdot \tilde{i}^{(e_{k+1})} \cdot d^{(g_k)}_{\text{main, min}}, \quad row \cdot d \leq value$$

A gap event maximal duration includes the following row in the matrix (for event k)

$$\mathrm{row} = \begin{pmatrix} 0 & 0 & \underbrace{\tilde{i}^{(e_k)}}_{\text{index } 2k+1} & 0 & 0 & \cdots & 0 \end{pmatrix},$$

$$\mathrm{value} = \tilde{i}^{(e_k)} \cdot d_{\text{gap, max}}, \quad \mathrm{row} \cdot d \leq \mathrm{value}$$

Finally, the total event duration can also have an upper limit which gives the additional row considering event k

$$\mathrm{row} = \begin{pmatrix} 0 & 0 & \cdots & \underbrace{\tilde{i}^{(e_k)} \; \tilde{i}^{(e_k)}}_{\text{indices } 2k \text{ and } 2k+1} & 0 & \cdots & 0 \end{pmatrix},$$

$$\mathrm{value} = \tilde{i}^{(e_k)} \cdot d_{\text{max}}, \quad \mathrm{row} \cdot d \leq \mathrm{value}$$

There are also the linear interdependence inequality constraints that can be written in the matrix form by adding the following row:

$$\mathrm{row} = \begin{pmatrix} a_0^{(e)} & a_0^{(g)} & a_1^{(e)} & a_1^{(g)} & \ldots & a_n^{(g)} \end{pmatrix},$$

$$\mathrm{value} = b, \mathrm{row} \cdot d \leq \mathrm{value}$$

with the a coefficients that can be equal to zero, but not all of them equal to zero, and the b coefficient that can be equal to zero. To specify the equality, one needs to add the opposite of quantities, that means

$$-\mathrm{row} \cdot d \leq -b, \quad \text{with the same row and value as previously}$$

7.2 Swiss prior model

We consider the model described in 1 and define first constraints and then the probabilistic distributions.

Birth occurs in $[1900, 2050]$ and lifespan (denoted by L) is bounded above by 110 years. Mandatory education starts at age 6 and lasts 9 years. Secondary education starts at age 15 and lasts between 3 and 6 years. Tertiary education may start between ages 18 and 25 and lasts at most 15 years. Employment may start from age 15. The last employment spell must start before age 64 and end before age 70. A driving license may be obtained between ages 18 and 70.

The probabilistic specifications are based on (Kukic et al., 2025; Baud and Bierlaire, 2026) and described hereafter for each dimension.

Existence Date of birth (denoted by b) follows a uniform distribution over the considered period. This is equivalent to assuming a constant birth rate which defines a homogeneous Poisson process. Such process has the property that the expected number of births in any interval only depends on the interval length. Lifespan follows a Weibull distribution with parameters $\lambda = 85$ and $k = 3$. As a result, the expected size of the alive population at any time t is constant. The Sex attribute is sampled as a Bernoulli of parameter 0.5.

Residence The canton of birth is sampled proportionally to population size. Work cantons are sampled proportionally to population size, with a multiplicative factor of 1.5 for urban cantons when income is high. Residence canton coincides with the work canton with probability 0.95, and is drawn from neighboring cantons with probability 0.05.

For the Urban residence attribute, we rely on the canton typology provided by the Swiss Federal Statistical Office (FSO), as well as the share of individuals living in urban and rural areas within each canton, reported in Table 7. The shares are rounded to the nearest 5%. At the aggregate level, the proportion of individuals living in urban areas is 84.9%, which is consistent with FSO estimates.

Car availability is modeled as an attribute associated with residence spells. The literature shows that individuals tend to adjust their mobility behavior at key life events such as residential relocation, childbirth, or job transitions (Scheiner, 2007; Beige and Axhausen, 2008; Gu et al., 2021). These events disrupt existing habits and often lead to changes in mobility resources, including car ownership and public transport subscriptions. In the proposed model, car availability is therefore resampled at each residence spell. However, it is constrained to be zero prior to the acquisition of a driving license. After this point, and conditional on holding a driving license, access to a car is modeled as a Bernoulli random variable with parameter 0.76, in line with the empirical findings reported in Danalet and Mathys (2018). Using the same source, the availability of a GA (General Abonnement) at residence spell m is 10% among the individuals aged 16 and above. This assumption is extended to individuals aged between 6 and 16, as the GA is available from age 6. Accordingly, GA availability is modeled as a Bernoulli random variable with parameter 0.1.

Education Mandatory education is universal. Secondary education is attended with probability 0.95, and tertiary education with probability 0.60 (conditional on eligibility). The duration of secondary education follows $\mathcal{N}(\mu = 4,\ \sigma^2 = 0.5)$, truncated to $[3, 6]$. If tertiary education occurs, the gap between secondary and tertiary education follows an exponential distribution with mean 1. The duration of tertiary education follows a Weibull distribution with parameters $\lambda = 4.5$ and

$k = 3.5$.

Work The number of employment spells follows a Poisson($\gamma(L)$) distribution, truncated to 11, where

$$\gamma(L) = \sqrt{\max(0, \min(L, 64) - 15)}$$

where 15 represents the minimal age to enter the job market, and 64 the retirement age. The square root shape accounts for the multiple changes of jobs in younger ages and lower rate of change of job in older ages (Topel and Ward, 1992; Swiss Federal Statistical Office, 2026b). For a person working until 64 which is the Swiss retirement age, this specification defines an average of slightly less than 7 jobs; which is consistent with Swiss estimates (Swiss Federal Statistical Office, 2026a). Individuals retire before 70, and can not start a new employment spell after 64.

Entry into the labor market is centered around graduation. The duration of the initial no-employment spell for each graduation level follows:

$$\begin{cases} \mathcal{N}(\mu = 16,\ \sigma^2 = 2) & \text{mandatory} \\ \mathcal{N}(\mu = 18,\ \sigma^2 = 4) & \text{secondary} \\ \mathcal{N}(\mu = 21,\ \sigma^2 = 7) & \text{tertiary} \end{cases}$$

The duration of an employment spell starting at age a_k follows an exponential distribution with age-dependent mean

$$\lambda(a_k) = \begin{cases} 2.5 & a_k < 25 \\ 3.5 & 25 \leq a_k < 35 \\ 4.5 & 35 \leq a_k < 45 \\ 5.5 & 45 \leq a_k < 55 \\ 9 & 55 \leq a_k \end{cases}$$

And the associated gap durations between employment spells follow an exponential distribution of mean 1.

Income during employment spell m follows the log-linear model

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

where work_m denotes the accumulated work experience, edu_m the total schooling years at the beginning of spell m and where

$$\varepsilon_m \sim \mathcal{N}(\mu = 0, \sigma^2 = 0.3)$$

$$\beta_0 = 7.8,\ \rho = 0.1,\ \beta_1 = 0.025,\ \beta_2 = -0.0004$$

Driving license ownership The duration of the "No driving license" spell (which is equivalent to the age of acquisition) follows a log-normal distribution $\log\mathcal{N}(\mu = \ln(20.5), \sigma^2 = 0.2)$ and 85% of individuals obtain a license during their lifetime.

7.3 Likelihood definition

For each observed variable $\tilde{Y}_t$ and its simulated counterpart $Y_t = T(X, t)$, a likelihood $P(\tilde{Y}_t \mid Y_t)$ is specified through variable-specific kernels. Continuous variables are modeled using Gaussian kernels, while categorical variables rely on misclassification probabilities. In all cases, the highest probability is assigned to exact matches.

Age is derived from the date of birth and lifespan. Let $\tilde{a}$ and a denote the observed and simulated ages. The likelihood is defined using a Gaussian kernel,

$$P(\tilde{a} \mid a) = \mathcal{N}(\tilde{a}; a, 1),$$

which reflects small measurement or reporting errors around the true age.

For unordered categorical variables, a simple misclassification model is adopted. Let $\tilde{x}$ and x denote the observed and simulated values. The likelihood is defined as

$$P(\tilde{x} \mid x) = \begin{cases} p & \text{if } \tilde{x} = x, \\ 1 - p & \text{otherwise,} \end{cases}$$

where the parameter p depends on the variable. This specification is used for sex ($p = 0.95$), urban status ($p = 0.7$), education status ($p = 0.8$), and GA status ($p = 0.9$).

For spatial variables, namely canton of residence and canton of work, the likelihood accounts for geographical proximity. Let $\mathcal{C}$ denote the set of cantons and N_c the set of neighboring cantons of c. The likelihood is defined as

$$P(\tilde{c} \mid c) = \begin{cases} 0.8 & \text{if } \tilde{c} = c, \\ \frac{0.2}{|N_c|} & \text{if } \tilde{c} \in N_c, \\ \frac{0.05}{|\mathcal{C} \setminus (N_c \cup \{c\})|} & \text{otherwise,} \end{cases}$$

so that misclassification is more likely to occur between neighboring cantons.

Let $\tilde{e}$ denote the observed employment status and e the simulated employment status. Since employment status is a categorical variable with no natural metric structure, the likelihood is defined through a misclassification matrix. For the categories

$$1: \text{ employed}, \qquad 2: \text{ unemployed}, \qquad 4: \text{ age } \leq 15, \qquad 5: \text{ retired},$$

we define

$$\big(P(\tilde{e} \mid e)\big) = \begin{array}{c|cccc} & e = 1 & e = 2 & e = 4 & e = 5 \\ \hline \tilde{e} = 1 & 0.90 & 0.08 & 0 & 0.02 \\ \tilde{e} = 2 & 0.08 & 0.90 & 0 & 0.02 \\ \tilde{e} = 4 & 0 & 0 & 1 & 0 \\ \tilde{e} = 5 & 0.02 & 0.02 & 0 & 0.96 \end{array}$$

so that exact matches are the most likely, small discrepancies between employed and unemployed are allowed, and the category corresponding to individuals aged at most 15 is treated as deterministic.

Income is an ordered categorical variable. Let $\tilde{y}, y \in \{1, \dots, 9\}$ denote the observed and simulated categories. The likelihood is defined through a distance-based kernel,

$$P(\tilde{y} \mid y) = \frac{\exp\left(-\frac{(\tilde{y}-y)^2}{2\sigma_{\text{inc}}^2}\right)}{\sum_{k=1}^{9} \exp\left(-\frac{(k-y)^2}{2\sigma_{\text{inc}}^2}\right)},$$

which assigns decreasing probability to categories that are farther from the simulated value.

Finally, driving license status and car availability are modeled jointly to capture their dependence. Let $z = (l, ca)$ and $\tilde{z} = (\tilde{l}, \tilde{ca})$. The likelihood is defined by the misclassification matrix

$$\big(P(\tilde{z} \mid z)\big) = \begin{array}{c|cccc} & (0,0) & (0,1) & (1,0) & (1,1) \\ \hline (0,0) & 0.85 & 0.30 & 0.10 & 0.03 \\ (0,1) & 0.05 & 0.60 & 0.02 & 0.02 \\ (1,0) & 0.08 & 0.05 & 0.78 & 0.10 \\ (1,1) & 0.02 & 0.05 & 0.10 & 0.85 \end{array},$$

which captures asymmetric misclassification patterns and avoids degeneracy by assigning strictly positive probabilities to all outcomes.

The parameter σ^2, controlling the alive population size, is fixed to 0.1, allowing for a deviation of about 1%. The population size curve is specified constant.

Property 7.1: Likelihood under deterministic sampling rule

Let $q(y) = P(S = 1 \mid Y = y)$. Then,

$$P(\tilde{Y}_{t,j} \mid X, S = 1) = \frac{\sum_{i \in \mathcal{I}_t(X)} q(Y_{t,i}) K_t(\tilde{Y}_{t,j}, Y_{t,i})}{\sum_{i \in \mathcal{I}_t(X)} q(Y_{t,i})}. \tag{3}$$

Since $q(Y_{t,i}) \in \{0, 1\}$, the expression simplifies to

$$P(\tilde{Y}_{t,j} \mid X, S = 1) = \frac{\sum_{i \in \mathcal{S}_t(X)} K_t(\tilde{Y}_{t,j}, Y_{t,i})}{\sum_{i \in \mathcal{S}_t(X)} 1} = \frac{1}{|\mathcal{S}_t(X)|} \sum_{i \in \mathcal{S}_t(X)} K_t(\tilde{Y}_{t,j}, Y_{t,i}). \tag{4}$$

For the complete observed data $\tilde{Y}_t$, the likelihood factorizes as

$$P(\tilde{Y}_t \mid X, S = 1) = \prod_{j=1}^{n} \left[\frac{1}{|\mathcal{S}_t(X)|} \sum_{i \in \mathcal{S}_t(X)} K_t(\tilde{Y}_{t,j}, Y_{t,i}) \right]^{w_j}. \tag{5}$$

where w_j corresponds to individual weights of the observed individuals in the data.

7.4 Data information

Rank	Canton (Eurostat NUTS 3)	Population	Share (%)
1	Zurich	1605508	17.91
2	Canton of Bern	1063533	11.87
3	Canton of Vaud	845870	9.44
4	Canton of Aargau	726894	8.11
5	Canton of St. Gallen	535114	5.97
6	Canton of Geneva	524410	5.85
7	Canton of Lucerne	432744	4.83
8	Canton of Valais	365844	4.08
9	Ticino	357720	3.99
10	Canton of Fribourg	341537	3.81
11	Canton of Basel-Landschaft	298837	3.33
12	Canton of Thurgau	295220	3.29
13	Canton of Solothurn	286844	3.20
14	Canton of Graubunden	204888	2.29
15	Canton of Basel-Stadt	200031	2.23
16	Canton of Neuchatel	178291	1.99
17	Canton of Schwyz	167403	1.87
18	Canton of Zug	132556	1.48
19	Canton of Schaffhausen	87111	0.97
20	Canton of Jura	74548	0.83
21	Appenzell Ausserrhoden	56495	0.63
22	Canton of Nidwalden	45016	0.50
23	Canton of Glarus	42056	0.47
24	Canton of Obwalden	39272	0.44
25	Canton of Uri	37931	0.42
26	Appenzell Innerrhoden	16585	0.19
Total (sum of 26 cantons)		**8962258**	**100.00**

Table 5: Population by Swiss canton (Eurostat NUTS 3, via DataCommons) and share of the total Swiss population (total computed as the sum of the 26 cantons).

Table 6: neighboring Swiss cantons

Canton	**neighboring cantons**
Zurich	Aargau, Schaffhausen, Thurgau, St. Gallen, Schwyz, Zug
Bern	Fribourg, Vaud, Neuchatel, Jura, Solothurn, Aargau, Lucerne, Obwalden
Lucerne	Aargau, Zug, Schwyz, Obwalden, Nidwalden, Bern
Uri	Schwyz, Nidwalden, Obwalden, Ticino, Graubunden
Schwyz	Zurich, Zug, Lucerne, Nidwalden, Uri, Glarus, St. Gallen
Obwalden	Lucerne, Nidwalden, Uri, Bern
Nidwalden	Lucerne, Obwalden, Uri, Schwyz
Glarus	Schwyz, St. Gallen, Graubunden
Zug	Zurich, Aargau, Lucerne, Schwyz
Fribourg	Vaud, Bern
Solothurn	Basel-Stadt, Basel-Landschaft, Jura, Bern, Aargau
Basel-Stadt	Basel-Landschaft
Basel-Landschaft	Basel-Stadt, Aargau, Solothurn, Jura
Schaffhausen	Zurich, Thurgau
Appenzell Ausserrhoden	St. Gallen
Appenzell Innerrhoden	St. Gallen
St. Gallen	Zurich, Thurgau, Appenzell Ausserrhoden, Appenzell Innerrhoden, Schwyz, Glarus, Graubunden
Graubunden	St. Gallen, Glarus, Uri, Ticino
Aargau	Zurich, Zug, Lucerne, Bern, Solothurn, Basel-Landschaft
Thurgau	Zurich, St. Gallen, Schaffhausen
Ticino	Uri, Graubunden, Valais
Vaud	Geneva, Valais, Fribourg, Bern, Neuchatel
Valais	Vaud, Bern, Uri, Ticino
Neuchatel	Vaud, Bern, Jura
Geneva	Vaud
Jura	Basel-Landschaft, Solothurn, Bern, Neuchatel

Rank	Canton	Population	Urban (%)	Rural (%)
1	Zurich	1605508	100	0
2	Canton of Bern	1063533	75	25
3	Canton of Vaud	845870	90	10
4	Canton of Aargau	726894	85	15
5	Canton of St. Gallen	535114	85	15
6	Canton of Geneva	524410	100	0
7	Canton of Lucerne	432744	65	35
8	Canton of Valais	365844	75	25
9	Ticino	357720	90	10
10	Canton of Fribourg	341537	75	25
11	Canton of Basel-Landschaft	298837	95	5
12	Canton of Thurgau	295220	70	30
13	Canton of Solothurn	286844	85	15
14	Canton of Graubunden	204888	45	55
15	Canton of Basel-Stadt	200031	100	0
16	Canton of Neuchatel	178291	90	10
17	Canton of Schwyz	167403	80	20
18	Canton of Zug	132556	100	0
19	Canton of Schaffhausen	87111	90	10
20	Canton of Jura	74548	50	50
21	Appenzell Ausserrhoden	56495	75	25
22	Canton of Nidwalden	45016	50	50
23	Canton of Glarus	42056	75	25
24	Canton of Obwalden	39272	30	70
25	Canton of Uri	37931	85	15
26	Appenzell Innerrhoden	16585	0	100
Urban population		**7611409**	**84.93 %**	
Rural population		**1350849**	**15.07 %**	
Total		**8962258**	**100.00 %**	

Table 7: Approximate share of population living in urban vs rural areas by Swiss canton (based on FSO 2012 typology of areas with urban character).

Variable	Description	Type
Demographics		
age	Age of the individual	numeric
gender	Sex of the individual	binary
canton_residence	Canton of residence	categorical
urban	Urban residence indicator	binary
Household characteristics		
hh_size	Household size	numeric
hh_type	Household type	categorical
hh_income	Household income	categorical
Socio-economic status		
employment_status	Employment status	categorical
education_status	Education status	binary
Mobility and transport		
has_license	Driving license ownership	binary
car_available	Access to a car	binary
has_ga	GA travelcard ownership	binary
canton_work	Canton of work	categorical
Weights		
WP	Individual population weight	numeric

Table 8: Recoded and extracted time-dependent variables from the MTMC dataset

7.5 Convergence diagnostics

In this section, we present diagnostics of the chains convergence when sampling from the prior. The sampling is done in three steps as described above:

- A first Hit-And-Run coupled with a Gibbs sampler for the attributes samples the date of birth, lifespan, and attributes associated with the *Existence* dimension.
- A Gibbs sampler for the event occurrences indicators.
- A Hit-And-Run sampler for durations couples with a Gibbs sampler for the attributes for the other dimensions.

Therefore, there is not one single MCMC procedure that generates the samples. There are several ones succeeding each other. Therefore, we present here

the diagnostics for the chains targeting the *Existence* dimension and the other dimensions characteristics.

Existence We sample 10,000 realizations of date of birth and lifespan across four chains with a burn-in of 1500 iterations. Table 9 presents the ESS and $\hat{R}$ values obtained for date of birth, lifespan using the Hit-and-Run, and the attribute "Sex" using the Gibbs sampler. All $\hat{R}$ values are below 1.01, which means that

	$\hat{R}$	ESS
Date of birth	1.0009	406.57
Lifespan	1.0002	552.35
Sex	0.9999	742.77

Table 9: Diagnostics for the *Existence* dimension

all chains are statistically indistinguishable from each other. The ESS for 10,000 realizations is over 400 for all quantities sampled which is also reassuring.

Other dimensions For the other dimensions, we can not directly compare chains for trajectories where individuals don't have the same event indicators. Indeed, having or not an event in their life trajectories constrains the durations of the other events of each dimension. Therefore, we generate *Existence* and the event indicators, and then run 4 chains with a burn-in of 6500 for the durations and attributes. We report in Table 10 the $\hat{R}$ and ESS measures of the relevant variables. For the driving license dimension, as there are only two durations sampled whose sum equals the lifespan, they both have the same diagnostics as each one is a linear transformation of the other. We only report the Driving license event main duration. No Education and Mandatory Education are constrained to fixed durations, their durations are deterministic, hence we don't report diagnostics for them. Residence spells are superposed with Work spells in our model, therefore, we restrict the results reporting to Work spells. We also report diagnostics for the log Income attribute.

Table 10 shows that the vast majority of the $\hat{R}$ values are below the conventional threshold of 1.1, with a substantial proportion even below 1.01. Some of the $\hat{R}$ values associated with the main durations of Work spells are, however, noticeably higher. This can be explained by the specification of these durations: although they are all exponentially distributed, the rate parameter λ depends on the age at entry into the spell and, consequently, on the cumulative duration of the preceding spells. As a result, later Work spells may be generated from exponential distributions with substantially different rate parameters (as the sum of previ-

	$\hat{R}$	ESS
Driving license main duration	1.04	24.15
Secondary Education main duration	< 1.01	103.57
Secondary Education gap duration	1.02	56.27
Tertiary Education main duration	< 1.01	51.38
No Work main duration	1.09	32.45
Work 1 main duration	< 1.01	45.49
Work 1 gap duration	1.03	63.04
Work 1 log Income	1.04	135.24
Work 2 main duration	1.32	39.78
Work 2 gap duration	< 1.01	71.63
Work 2 log Income	1.04	119.43
Work 3 main duration	1.24	31.92
Work 3 gap duration	1.02	64.04
Work 3 log Income	1.09	96.78
Work 4 main duration	1.03	42.36
Work 4 gap duration	< 1.01	66.97
Work 4 log Income	1.09	124.95
Work 5 main duration	1.21	22.92
Work 5 gap duration	< 1.01	81.19
Work 5 log Income	1.05	167.54
Work 6 main duration	1.24	28.16
Work 6 gap duration	< 1.01	79.57
Work 6 log Income	1.02	222.83
Work 7 main duration	1.26	28.98
Work 7 gap duration	1.02	64.33
Work 7 log Income	1.02	231.74
Work 8 main duration	1.16	25.56
Work 8 gap duration	< 1.01	50.59
Work 8 log Income	1.02	250.30
Work 9 main duration	1.23	24.95
Work 9 gap duration	< 1.01	65.87
Work 9 log Income	< 1.01	267.39
Work 10 main duration	1.16	22.31
Work 10 gap duration	< 1.01	59.42
Work 10 log Income	< 1.01	260.63
Work 11 main duration	1.07	20.7
Work 11 log Income	< 1.01	267.46

Table 10: Diagnostics for the other dimensions

ous durations varies more and more), leading by design to greater heterogeneity across sampled trajectories. Their durations are therefore not expected to follow an identical marginal distribution as their λ parameter differ. By contrast, all gap durations are generated from the same exponential distribution with a common parameter, which explains their consistently good convergence diagnostics. The effective sample size (ESS) values are relatively low compared with commonly recommended levels. This is not unexpected, since both the *Existence* dimension and the event indicators are fixed and impose strong structural constraints on the set of feasible trajectories. Sampling is therefore restricted to a constrained feasible space, which can induce stronger dependence between successive draws and slower mixing of the Markov chains.

7.6 Additional results

7.6.1 Age at start of work - prior model

(a) Graduated from Mandatory education

(b) Graduated from Secondary education

(c) Graduated from Tertiary education

Figure 11: Age at start of work

7.6.2 Event frequencies - prior

Event frequencies are consistent with the imposed probabilities, conditional on survival constraints. Among individuals whose lifespan exceeds 18 years, 85% obtain a driving license. All individuals with lifespan greater than 6 undergo mandatory education. Among those living beyond 15 years, 94.9% complete secondary education, and among those living beyond 18 years, and have done secondary education, 60% complete tertiary education. These proportions are in line with the model specification.
The number of job spells also follows the intended truncated Poisson$(\lambda(L))$ at 11 spells, and where L denotes the lifespan, as Table 11 displays. Note that individuals whose lifespan is lower than 15 have no job spell in their life.

Table 11: Empirical vs. Theoretical Number of jobs per Lifespan bin

Lifespan Bin	[0, 15]	]15, 20]	]20, 25]	]25, 30]	]30, 35]	]35, 40]	]40, 45]	]45, 50]	]50, 55]	]55, 60]	]60, 64]	]64, 110]
Empirical Mean	0.00	1.55	2.77	3.62	4.17	4.72	5.19	5.65	5.91	6.35	6.58	6.67
Theoretical Mean	0.00	1.58	2.74	3.53	4.17	4.72	5.19	5.60	5.97	6.29	6.56	6.67

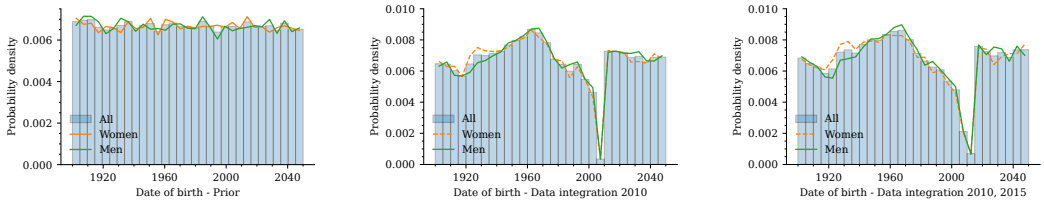
7.6.3 Attributes - prior

The sampled attributes are consistent with the specified distributions. In particular, the income sampled at the start of each work spell matches the intended distribution. Table 12 reports the theoretical and empirical average log-incomes, differentiated by work spell and education level.

	Mandatory		Secondary		Tertiary	
Work spell	Sampled	Theoretical	Sampled	Theoretical	Sampled	Theoretical
1	8.71	8.70	9.06	9.07	9.40	9.42
2	8.74	8.74	9.10	9.10	9.54	9.55
3	8.81	8.80	9.13	9.13	9.62	9.63
4	8.85	8.86	9.17	9.17	9.69	9.69
5	8.91	8.91	9.21	9.21	9.74	9.74
6	8.97	8.96	9.26	9.26	9.78	9.79
7	8.99	9.00	9.30	9.30	9.82	9.82
8	9.03	9.03	9.32	9.33	9.85	9.85
9	9.03	9.05	9.35	9.35	9.87	9.87
10	9.05	9.06	9.36	9.36	9.89	9.88
11	9.00	9.07	9.38	9.37	9.90	9.89

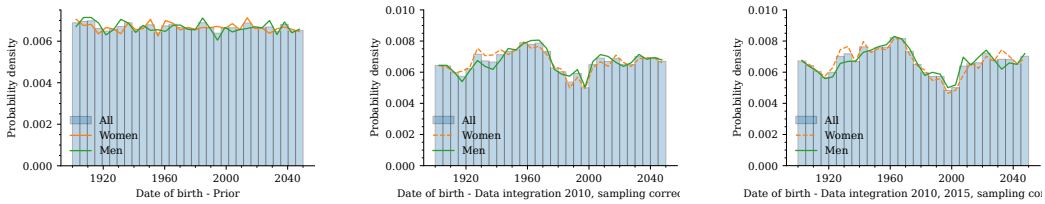
Table 12: Sampled and theoretical expected log-income by work spell and education level

7.6.4 Date of birth - posterior



(a) Prior distribution (b) Posterior 2010 (c) Posterior 2010, 2015

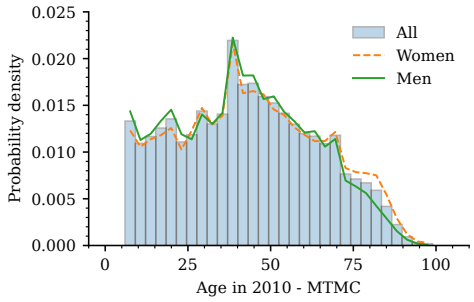
Figure 12: Date of birth distributions: prior and posterior updates, differentiated by gender. Sampling rule not taken into account.



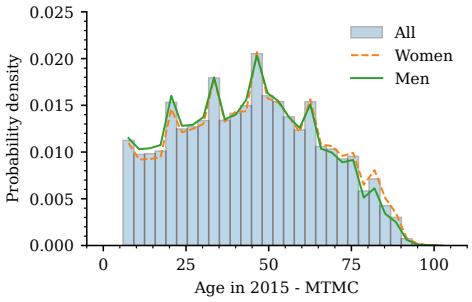
(a) Prior distribution (b) Posterior 2010 (c) Posterior 2010, 2015

Figure 13: Date of birth distributions: prior and posterior updates, differentiated by gender. Sampling rule taken into account.

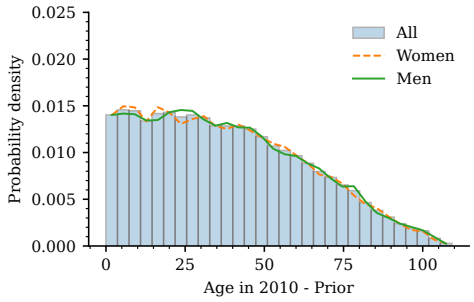
7.6.5 Age distribution without taking into account the sampling rule - posterior



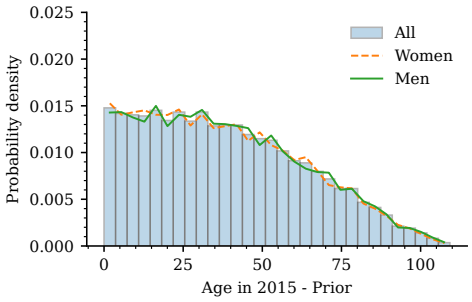
(a) 2010



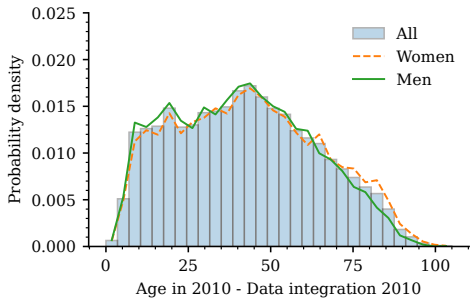
(b) 2015



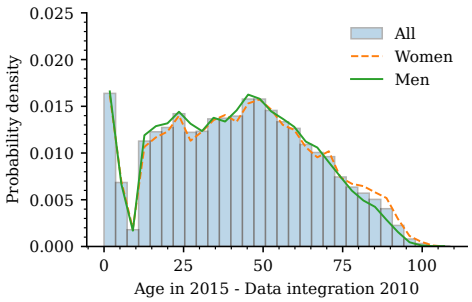
(c) Prior (2010)



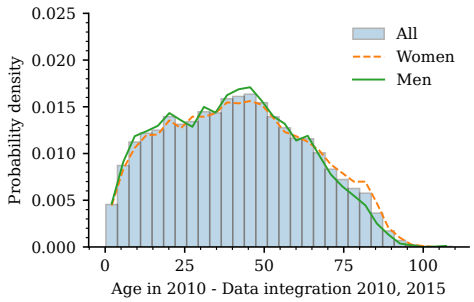
(d) Prior (2015)



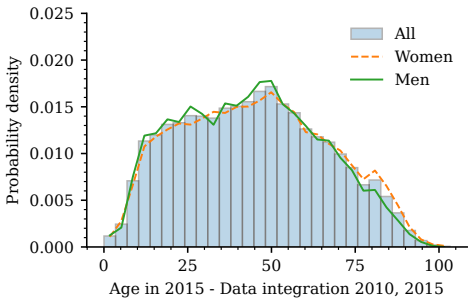
(e) Update with 2010 data (2010)



(f) Update with 2010 data (2015)



(g) Update with 2010+2015 data (2010)



(h) Update with 2010+2015 data (2015)

Figure 14: Age distributions. Columns correspond to years (2010 left, 2015 right). Rows correspond to observed data, prior, update with 2010 data, and update with 2010 and 2015 data. Update without taking into account the sampling rule.

7.6.6 Age and Sex - posterior

The marginal distributions of age and sex are improved in the posterior distribution, but more importantly, the results also show that the correlation between age and sex is captured. The prior population does not exhibit any age difference across genders. After the update, the model reproduces the empirical pattern in which women are, on average, older than men, as shown in Figure 7 and Table 13. Without taking into account the sampling rule however, the Table indicates noticeable over-estimation of the age as it doesn't consider the missing children by design. By under-representing children, the adults age is shifted towards higher values as an attempt for the posterior to get close to the censored empirical distribution.

Table 13: Average age among those aged > 7, by population, year, and gender

Population	**Year**	**All**	**Men**	**Women**
Prior	2010	42.54	42.67	42.42
Prior	2015	42.57	42.75	42.39
Data	2010	43.19	42.07	44.26
Data	2015	44.43	43.60	45.25
Updated (2010 $\rightarrow$ 2010)	2010	43.07	41.94	44.15
Updated (2010 $\rightarrow$ 2015)	2015	46.09	45.12	47.03
Updated (2010 $\rightarrow$ 2010), sampling	2010	42.89	41.93	43.82
Updated (2010 $\rightarrow$ 2015), sampling	2015	43.50	42.60	44.37
Updated (2010, 2015 $\rightarrow$ 2010)	2010	43.05	42.10	43.97
Updated (2010, 2015 $\rightarrow$ 2015)	2015	45.04	44.14	45.92
Updated (2010, 2015 $\rightarrow$ 2010), sampling	2010	42.95	41.88	43.98
Updated (2010, 2015 $\rightarrow$ 2015), sampling	2015	44.05	43.16	44.91

7.6.7 Event occurrences in the posterior

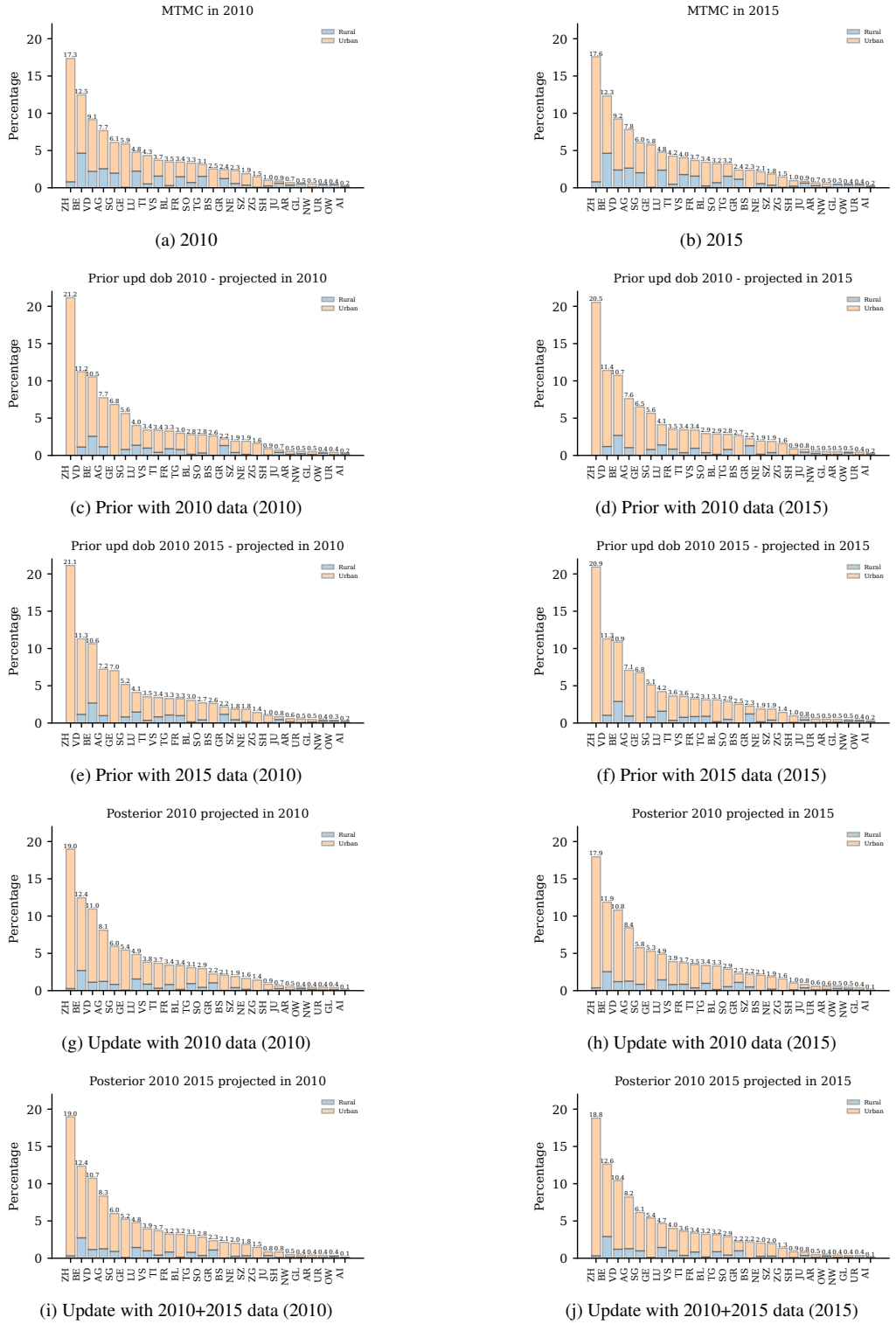
Only the event occurrences are sampled from the prior, as the observation of cross-sectional data does not inform about the number of spells the individual experiences. Table 14 displays the number of jobs for each lifespan bin, sampled in the prior model and for the different updates. One can see that the event occurrences are indeed sampled according to the prior model. Similarly for other dimensions, among individuals whose lifespan is greater than 18, 85 % acquire a driving license in their life. Among the individuals whose lifespan is greater than 6, all of them undergo Mandatory Education. Among those surviving up to 15 and having undergone Mandatory Education, 95 % of them pursue Secondary Education, and among those, 60 % pursue Tertiary Education, as specified in the prior distribution.

Table 14: Prior vs. Updated Number of jobs per Lifespan bin

Lifespan Bin	[0, 15]	]15, 20]	]20, 25]	]25, 30]	]30, 35]	]35, 40]	]40, 45]	]45, 50]	]50, 55]	]55, 60]	]60, 64]	]64, 110]
Prior	0.00	1.55	2.77	3.62	4.17	4.72	5.19	5.65	5.91	6.35	6.58	6.67
2010 - Not alive in 2010	0.00	1.63	2.68	3.51	4.20	4.73	5.21	5.63	5.95	6.23	6.54	6.65
2010, 2015 - Not alive in 2010 and 2015	0.00	1.61	2.76	3.68	4.16	4.71	5.17	5.63	5.99	6.27	6.50	6.67
2010 - Alive in 2010	0.00	1.49	2.75	3.38	4.0	4.86	5.19	5.37	6.01	6.33	6.49	6.69
2010, 2015 - Alive in 2010 or 2015	0.00	1.81	2.86	3.53	4.39	4.67	5.32	5.61	6.09	6.33	6.57	6.68

7.6.8 Residence dimension - posterior

The posterior residence dimension is also improved compared to the prior distribution, as shown in Figure 15. The prior model is slightly misaligned with the data and did not accurately capture the urban distribution within each canton. Indeed, the model is calibrated using rough estimates (table 7), which led to some cantons having their entire population assigned to urban areas, a pattern that is not supported by the observed data. The posterior distribution mitigates these limitations and produces an overall better distribution of the population across cantons, as well as a more accurate representation of urban shares within each canton.



(a) 2010

(b) 2015

(c) Prior with 2010 data (2010)

(d) Prior with 2010 data (2015)

(e) Prior with 2015 data (2010)

(f) Prior with 2015 data (2015)

(g) Update with 2010 data (2010)

(h) Update with 2010 data (2015)

(i) Update with 2010+2015 data (2010)

(j) Update with 2010+2015 data (2015)

Figure 15: Residence distributions. Columns correspond to years (2010 left, 2015 right). Rows correspond to observed data, priors based on the 2010 and 2015 data, the update with 2010 data, and the update with both 2010 and 2015 data.

7.6.9 Employment status by Sex

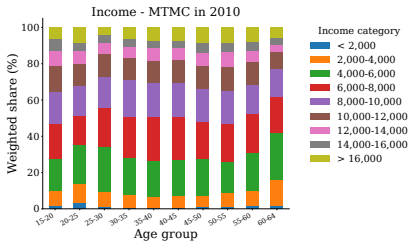
Table 15: Comparison of employment-status distributions across simulation settings by sex, among those aged > 7, in %

(a) Women

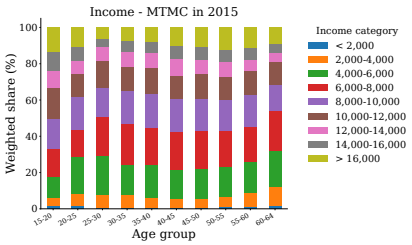
Model	Existence update	Data update	Mapped year	Employed	Unemployed	Retired	Age < 15
Observed data	–	–	2010	54.38	16.89	21.02	7.71
Prior	2010	–	2010	32.09	37.44	20.08	10.38
Posterior	2010	2010	2010	41.65	27.90	20.06	10.38
Prior	2010–2015	–	2010	32.80	37.75	19.90	9.54
Posterior	2010–2015	2010–2015	2010	42.19	28.48	19.79	9.54
Observed data	–	–	2015	55.72	15.08	21.90	7.21
Prior	2010	–	2015	30.66	35.75	21.20	12.40
Posterior	2010	2010	2015	36.89	30.16	20.54	12.40
Prior	2010–2015	–	2015	31.17	37.29	21.46	10.08
Posterior	2010–2015	2010–2015	2015	41.43	27.51	20.98	10.08

(b) Men

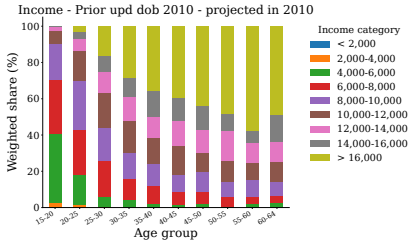
Model	Existence update	Data update	Mapped year	Employed	Unemployed	Retired	Age < 15
Observed data	–	–	2010	66.61	9.73	15.32	8.35
Prior	2010	–	2010	32.28	39.28	16.28	11.17
Posterior	2010	2010	2010	47.42	25.40	16.01	11.17
Prior	2010–2015	–	2010	33.44	40.11	15.76	10.68
Posterior	2010–2015	2010–2015	2010	48.03	25.95	15.33	10.68
Observed data	–	–	2015	65.05	9.80	17.11	7.98
Prior	2010	–	2015	31.85	36.20	18.69	13.25
Posterior	2010	2010	2015	37.03	31.97	17.74	13.25
Prior	2010–2015	–	2015	31.92	38.69	18.26	11.12
Posterior	2010–2015	2010–2015	2015	44.98	26.46	17.44	11.12



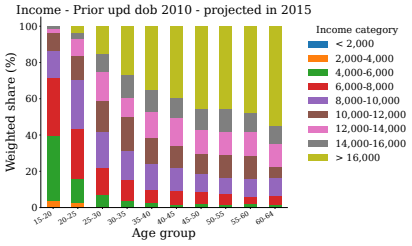
(a) 2010



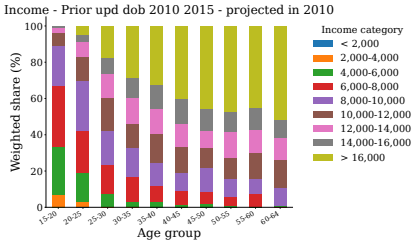
(b) 2015



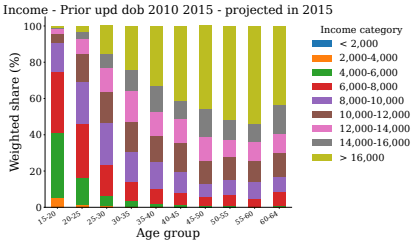
(c) Prior with 2010 data (2010)



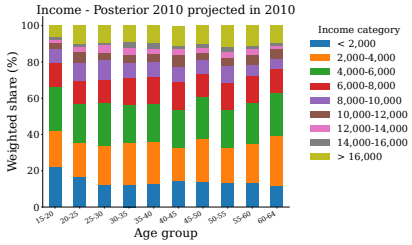
(d) Prior with 2010 data (2015)



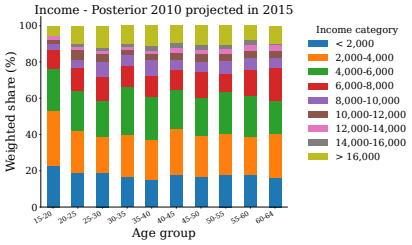
(e) Prior with 2015 data (2010)



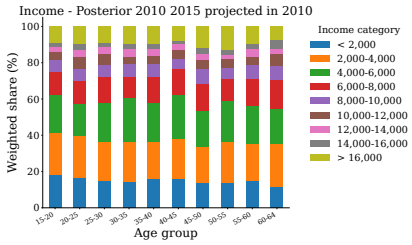
(f) Prior with 2015 data (2015)



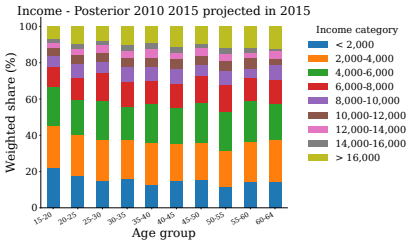
(g) Update with 2010 data (2010)



(h) Update with 2010 data (2015)



(i) Update with 2010+2015 data (2010)



(j) Update with 2010+2015 data (2015)

Figure 16: Income distributions by age. Columns correspond to years (2010 left, 2015 right). Rows correspond to observed data, priors based on the 2010 and 2015 data, the update with 2010 data, and the update with both 2010 and 2015 data.

7.6.10 Income by age category - posterior

7.6.11 Driving license by age and gender - posterior

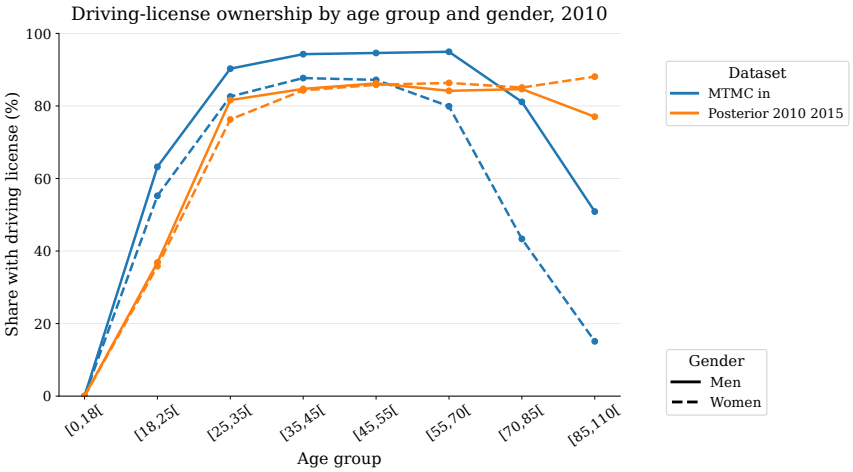


Figure 17: Driving license by age and gender, 2010

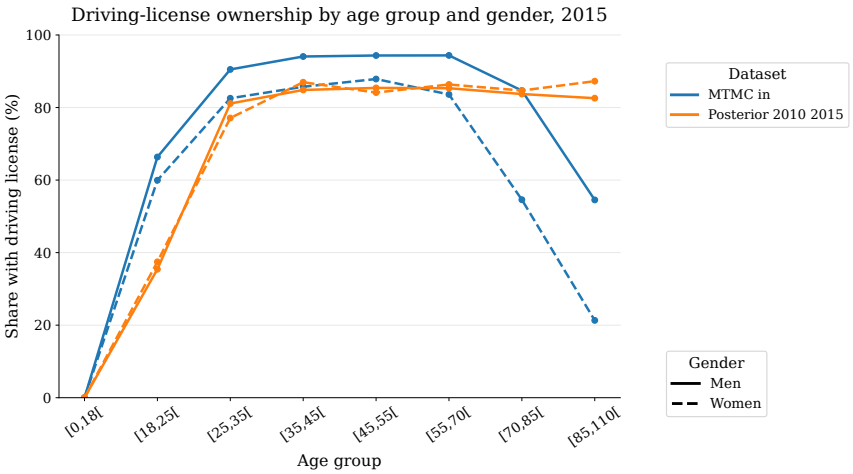


Figure 18: Driving license by age and gender, 2015

References

Aitchison, J. and Aitken, C. G. G. (1976). Multivariate binary discrimination by the kernel method, *Biometrika* **63**(3): 413–420.

Baltagi, B. H. (2005). *Econometric Analysis of Panel Data*, 3 edn, John Wiley & Sons, Chichester.

Barthélemy, J. and Toint, P. L. (2013). Synthetic population generation without a sample, *Transportation Science* **47**(2): 266–279.

Baud, C. (2026). Longitudinal synthetic population framework. Computer software.
URL: *https://doi.org/10.5281/zenodo.21293242*

Baud, C. and Bierlaire, M. (2026). A flexible and realistic synthetic panel population generation, *Proceedings of the 2026 IEEE 29th International Conference on Intelligent Transportation Systems*. In press.

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

Beige, S. and Axhausen, K. W. (2008). Long-term and mid-term mobility decisions during the life course: Experiences with a retrospective survey, *IATSS Research* **32**(2): 16–33.

Bhat, C. R. and Koppelman, F. S. (1999). Activity-based modeling of travel demand, *in* R. W. Hall (ed.), *Handbook of Transportation Science*, Springer, Boston, MA, pp. 35–61.

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

Danalet, A. and Mathys, N. (2018). Mobility resources in switzerland in 2015, *Proceedings of the 18th Swiss Transport Research Conference (STRC)*, Federal Office for Spatial Development (ARE), Monte Verita / Ascona, Switzerland. May 16–18, 2018.

Dargay, J. M. and Vythoulkas, P. C. (1999). Estimation of a dynamic car ownership model: A pseudo-panel approach, *Journal of Transport Economics and Policy* **33**(3): 287–301.

Deaton, A. (1985). Panel data from time series of cross-sections, *Journal of Econometrics* **30**(1–2): 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.

Farooq, B., Bierlaire, M., Hurtubia, L. and Flötteröd, G. (2013). Simulation-based population synthesis, *Transportation Research Part B* **58**: 243–263.

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

Gärling, T. and Axhausen, K. W. (2003). Introduction: Habitual travel choice, *Transportation* **30**(1): 1–11.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association* **85**(410): 398–409.

Gu, G., Feng, T., Yang, D. and Timmermans, H. (2021). Modeling dynamics in household car ownership over life courses: a latent class competing risks model, *Transportation* **48**: 809–829.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications, *Biometrika* **57**(1): 97–109.

Hausman, J. A., Abrevaya, J. and Scott-Morton, F. M. (1998). Misclassification of the dependent variable in a discrete-response setting, *Journal of Econometrics* **87**(2): 239–269.

Hsiao, C. (2014). *Analysis of Panel Data*, 3 edn, Cambridge University Press, Cambridge, U.K.

Kukic, M., Benchekroun, S. and Bierlaire, M. (2023). Hybrid simulator for capturing dynamics of synthetic populations, *Proceedings of the 26th IEEE International Conference on Intelligent Transportation Systems (ITSC)*, Bilbao, Spain.

Kukic, M. and Bierlaire, M. (2025). Adaptive synthetic generation using one-step Gibbs sampler, *Transportation Research Interdisciplinary Perspectives* **33**: 101597.

Kukic, M., Ilinov, P. and Bierlaire, M. (2025). Simulation framework for generating synthetic panel data, *Technical Report 251013*, Transport and Mobility Laboratory (TRANSP-OR), EPFL.

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.

Li, Q. and Racine, J. S. (2007). *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

Lovász, L. and Vempala, S. (2006). Hit-and-run from a corner, *SIAM Journal on Computing* **35**(4): 985–1005.

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.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equation of state calculations by fast computing machines, *Journal of Chemical Physics* **21**(6): 1087–1092.

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.

Orcutt, G. H. (1957). A new type of socio-economic system, *The Review of Economics and Statistics* **39**(2): 116–123.

Oshanreh, M. M., Khan, N. A. and MacKenzie, D. (2025). Propagating synthetic populations with dynamic bayesian networks: A framework for long-horizon demographic forecasting, *SSRN Electronic Journal* . Preprint, not peer reviewed.
URL: *https://ssrn.com/abstract=5281670*

Rezvany, N., Kukic, M. and Bierlaire, M. (2024). A review of activity-based disaggregate travel demand models, *Findings* .

Scheiner, J. (2007). Mobility biographies: Elements of a biographical theory of travel demand (mobilitätsbiographien: Bausteine zu einer biographischen theorie der verkehrsnachfrage), *Erdkunde* **61**(2): 161–173.
URL: *https://www.jstor.org/stable/25647982*

Smith, R. L. (1984). Efficient monte carlo procedures for generating points uniformly distributed over bounded regions, *Operations Research* **32**(6): 1296–1308.

Swiss Federal Statistical Office (2026a). Swiss labour force survey (slfs). Accessed: 2026-04-07.
URL: *https://www.bfs.admin.ch/bfs/en/home/statistics/work-income/surveys/slfs.html*

Swiss Federal Statistical Office (2026b). Work and income statistics. Accessed: 2026-04-07.
URL: *https://www.bfs.admin.ch/bfs/en/home/statistics/work-income.html*

Topel, R. H. and Ward, M. P. (1992). Job mobility and the careers of young men, *Quarterly Journal of Economics* **107**(2): 439–479.

van Imhoff, E. and Post, W. (1998). Microsimulation methods for population projection, *Population: An English Selection* **10**(1): 97–138.

Verbeek, M. and Nijman, T. (1998). Can cohort data be treated as genuine panel data?, *Empirical Economics* **23**: 9–23.

Villani, C. (2009). *Optimal Transport: Old and New*, Vol. 338 of *Grundlehren der mathematischen Wissenschaften*, Springer, Berlin, Heidelberg.

Willekens, F. (2014). *Multistate Analysis of Life Histories with R*, Springer.

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

Ye, X., Konduri, K. C., Pendyala, R. M., Sana, B. and Waddell, P. (2009). Methodology to match distributions of both household and person attributes in generation of synthetic populations.
URL: *https://api.semanticscholar.org/CorpusID:127155690*