A synthetic longitudinal individual generator

Candice Baud, Michel Bierlaire

TRANSP-OR laboratory, EPFL

May 2026

EPFL

Outline

Introduction

Motivation

- Agent-based models need **micro-level agents** (persons, households, firms).
- Real microdata is often **unavailable** (privacy) or **incomplete** (coverage).

Goals

- Generate a population that **matches observed aggregates** while remaining **plausible**.
- Contribution : generate **panel data** of individuals to track them over time

Existing simulation methods

[1] **Dynamic microsimulation :** Lomax, N., Smith, A. P. (2017). An introduction to microsimulation for demography. Australian Population Studies, 1(1), 73-85.

[2] **Projection and reweighting :** Kukic, M., Bierlaire, M. (2025). Adaptive synthetic generation using one-step gibbs sampler

[3] **Panel data generation :** Kukic, M., Ilinov, P., Bierlaire, M. (2025). Simulation framework for generating synthetic panel data (Tech. Rep. No. 251013).

$\longrightarrow$ **No general framework for panel data**
$\longrightarrow$ **Limited real data integration**

Outline

Key idea : Time independent individuals

- Individuals are usually described by **time-dependent variables**:

$$\text{Age}_t, \quad \text{Income}_t, \quad \text{DrivingLicense}_t$$

- Instead, we describe each individual by a **time-independent life-course vector** X.
- Time-dependent states are recovered through a deterministic mapping:

$$Y_t = T(X, t)$$

Example :

- Knowing the date of birth is sufficient to recover age at any time:

$$X = \text{Date of birth}, \qquad \text{Age}_t = T(X, t) = t - X$$

Key idea : Life dimensionality

- An individual life unfolds along multiple **dimensions = independent axes** : work, education, residence, driving license, ...
- Each dimension is composed of possible **events**: mandatory education, secondary education, tertiary education, ...
- A dimension trajectory is the collection of **realized events**, organized by dimension.
- Along each dimension, event durations must satisfy **structural constraints**:
  - **Span constraint**: the sum of event durations equals the lifespan
  - **Non-overlap constraint**: events cannot overlap in time

Life trajectory representation

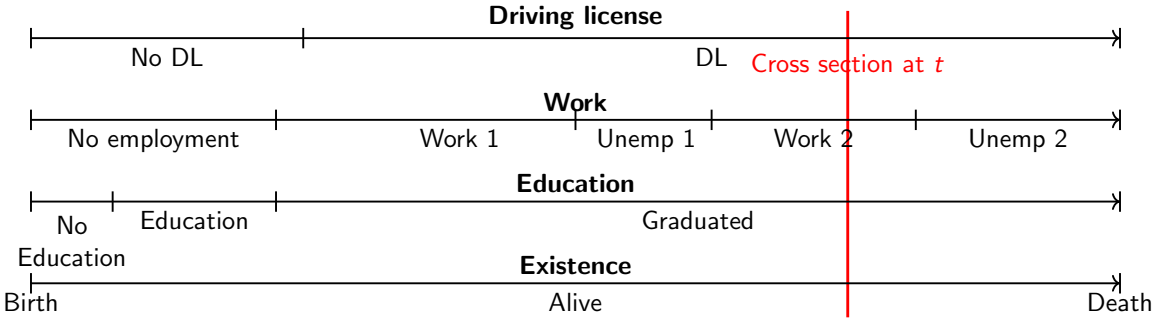


Figure: Complete life trajectory and cross-section at t

Additional constraints

- Span and non-overlap constraints define the structure of life trajectories, but are not sufficient to produce meaningful individuals

  - **Biological and legal constraints:**
    - Biological constraint (e.g. no individual lives beyond 150 years)
    - Legal age requirements (e.g. driving license only after the legal age), duration constraints
  - **Inter-event constraints:** relationships between event durations across dimensions
    - Example: individuals relocate only when changing job, implying equal durations for work and residence events

Good news : All constraints can be expressed as linear constraints !

Outline

Need to sample

$$(\tau_D)_D, \quad \text{where} \quad \tau_D = \big([\, i_D^{(e_{1,D})},\ s(e_{1,D}),\ d(e_{1,D}),\ a_{1,D}(e_{1,D}),\ a_{2,D}(e_{1,D}), \dots],$$
$$\dots$$
$$[\, i_D^{(e_{n_{e,D},D})},\ s(e_{n_{e,D},D}),\ d(e_{n_{e,D},D}),\ a_{1,D}(e_{n_{e,D},D}), \dots,\ a_{n_{a,D},D}(e_{n_{e,D},D})]\big)$$

$i(e)$ corresponds to indicator of the event happening

$s(e)$ corresponds to the starting date of event e

$d(e)$ corresponds to the duration of event e

$a_i(e)$ corresponds to attribute i of the event e

Strategy for sampling from the priors

- Prior distributions are specified using the literature and domain knowledge.
- Sampling is performed sequentially using a **Gibbs** scheme and **Metropolis–Hastings** steps:
  - Date of birth and lifespan: **Hit-and-Run MH**
  - Event occurrence indicators: **Gibbs sampler**
  - Event durations and attributes: **Gibbs Hit-And-Run MH on a convex set**

NB : the prior must be evaluable on the **time-independent variables**.

From individuals to population

Key Assumption

- A population is a collection of independent individuals

- Using the individual generation process, sample individuals independently
- Very restrictive assumption $\rightarrow$ will be relaxed in future work.

Outline

- What if the prior distributions do not reflect the true population?
- What if the priors are outdated or miss recent structural changes?
- What if a large-scale shock occurs (pandemic, war, policy change)?

Integrate cross-sectional measurement

Key idea :

- Use Bayesian statistics to recover the posterior distribution of X given the observed cross-sectional dataset(s)

$$X \sim f_{\text{prior}} \quad \text{without data measurement}$$

$$X \sim f(X|(\tilde{Y}_t)_t) \quad \text{when observing data}$$

- Formula

$$f(X|(\tilde{Y}_t)_t) \propto \mathcal{L}((\tilde{Y}_t)_t|X) \cdot f_{\text{prior}}(X)$$

- Not all time-independent individuals are concerned by the update, only the ones alive at the moment of the dataset

Small example

	Date of birth	Lifespan	Life status in 2000
Cleopatra	69 BC	39	Dead
Jesus Christ	0	33-39	Dead
Michael Jackson	1958	50	Alive
Michel Bierlaire	1967	> 59	Alive
Louise Lallemand	2000	> 26	Just born
Candice Baud	2001	> 24	Not born yet

Table: Example individuals life statuses

Likelihood calculation

Key idea : Mapping

- Each individual X_i in the time-independent framework can be mapped in the time-dependent framework $Y_{t,i} = T(X_i, t)$
- Observed individual $\tilde{Y}_{t,j} = Y_{t,j} + \varepsilon = T(X, t) + \varepsilon$
- With ε distributed,

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

For example : $f_\varepsilon = \mathcal{N}_{\mu=0,\ \sigma}$

Population size

Issues:

- The previous likelihood function only accounts for alive individuals.
- It is possible to generate an unlimited number of non-alive individuals without impacting the likelihood.
- The posterior distribution is not identifiable without incorporating an additional term.

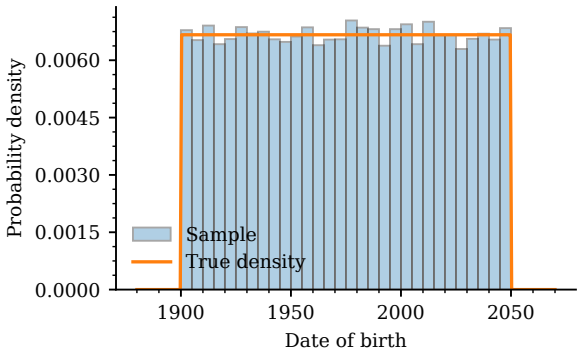
Key Idea:

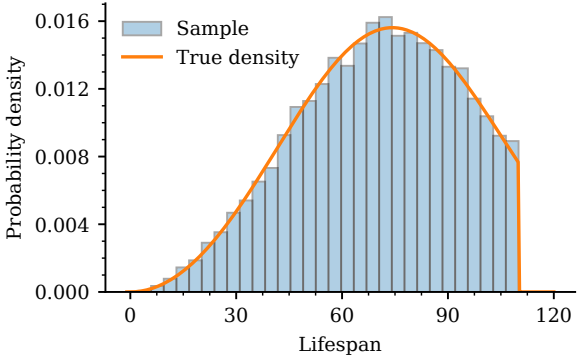
- Introduce a population size target (e.g., based on historical data, prior knowledge, or assumptions).
- Add a regularization term to the likelihood:

$$-\frac{1}{2\sigma^2}\int_{t_{\min}}^{t_{\max}}\left(\tilde{M}_t(G)-M^\star(G)\right)^2\,dt \quad\rightarrow\quad -\frac{1}{2\sigma^2}\sum_k w_k\left(\tilde{M}_{t_k}(G)-M^\star_{t_k}(G)\right)$$

Outline

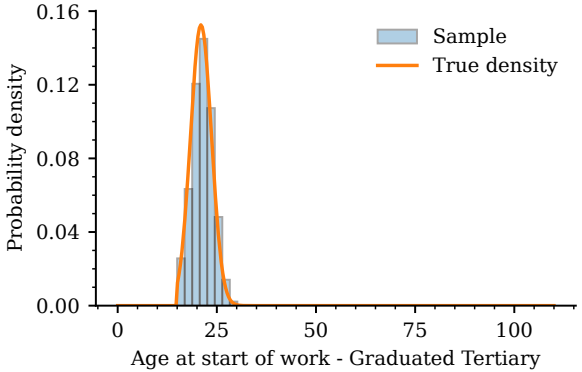
Priors : Existence

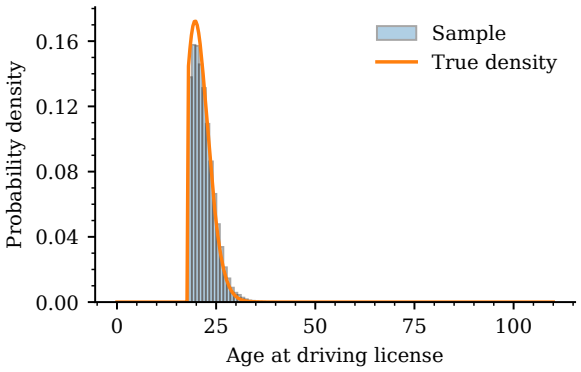




Date of birth and Lifespan sampled and true densities

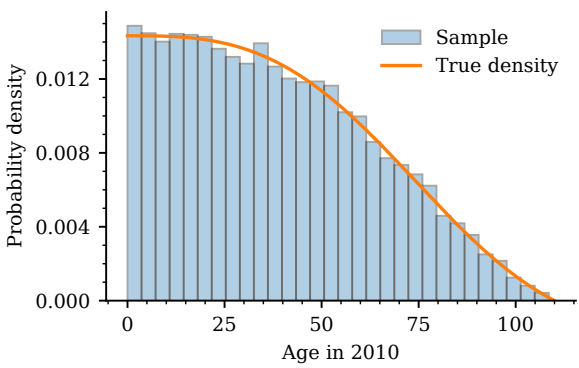
Priors : Other dimensions

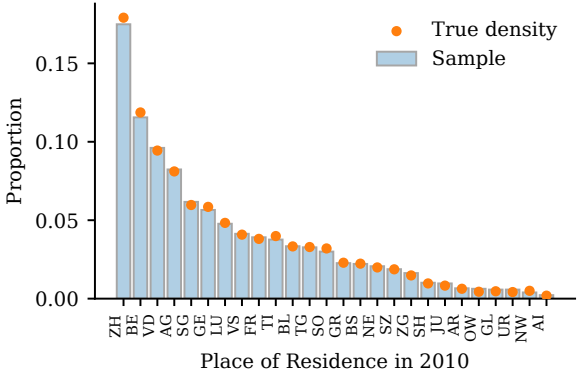




Age at labor entry (Graduated from Tertiary) and Driving license age

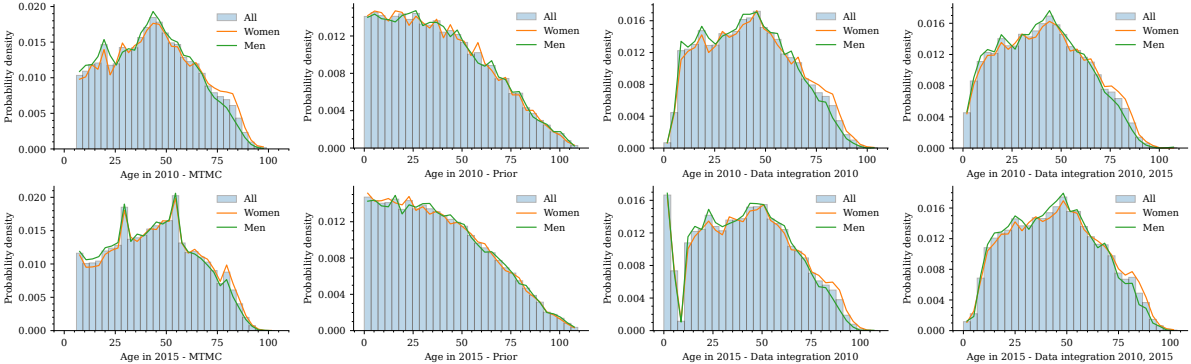
Priors : Projection in 2010





Age and canton of residence in 2010

Existence posterior projected



Observed data, prior, posterior with 2010, posterior with 2010 and 2015; in 2010 and 2015

Outline

Contributions and further steps

Contributions :

- Generate **panel individuals and populations** from any model ensuring consistency of individuals
- Data-free sampling $\longrightarrow$ **Prior sampling**
- Data and model sampling $\longrightarrow$ **Bayesian update**

Future work

- Relax assumption of individuals independence $\longrightarrow$ **Households**
- Relax assumption of independence of the observed cross-sectional data-frames $\longrightarrow$ **Integrate panel data observations**

Questions ?

Scan the QR code for to access the Github repository !