# Longitudinal Synthetic Population Generation : A Unified Framework

Candice Baud, Michel Bierlaire

TRANSP-OR laboratory, EPFL

2 July 2026


EPFL

# Outline

# Introduction

## Previous work

- Cross sectional synthetic population generation at fixed time t
- Projection and reweighting
- Pseudo panel

→ **No general method to create panel data**

## Contributions

- A general model based framework to sample **panel data** of individuals that are coherent individually and match observed aggregates
- A Bayesian update mechanism enabling to use real data to correct the prior model

# Introduction


Figure: Cross-section vs panel data

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

# 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 : Full life trajectory representation

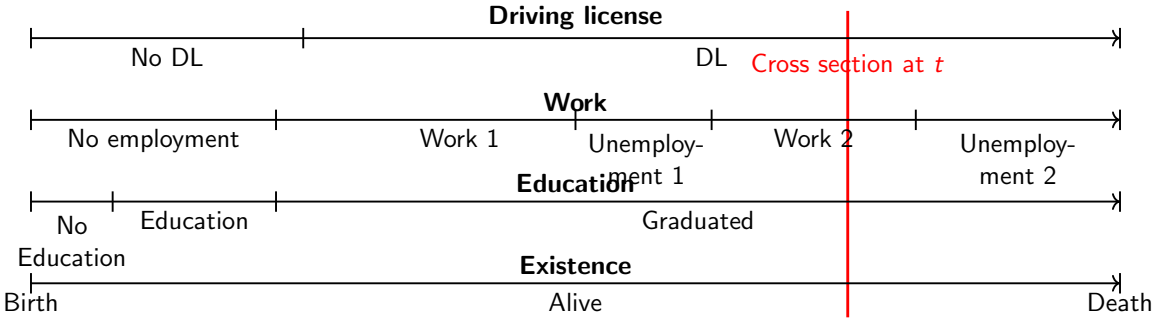



Figure: Complete life trajectory and cross-section at $t$

# Framework summary

- Each individual life is decomposed along multiple **axes** (called dimensions) which form a **partition** of their respective lifespan
- Each individual can undergo the same events but they all have different **realized trajectories** which are sampled probabilistically
- **Constraints** enable to create plausible life trajectories:
  - *Span constraints :* the sum of event durations equals the lifespan
  - *Non-overlap constraints :* events in the same dimension cannot overlap in time
  - *Biological/legal constraints*
  - *Other constraints* : linking events of the same or different dimension

# 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 literature
- Sampling is performed sequentially using a **Gibbs** scheme and **Metropolis–Hastings** steps enforcing the constraints and exploring efficiently the space:
  - 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

**Current 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: Renowned individuals and their life statuses in 2000

# Likelihood calculation

**Key idea : Mapping**

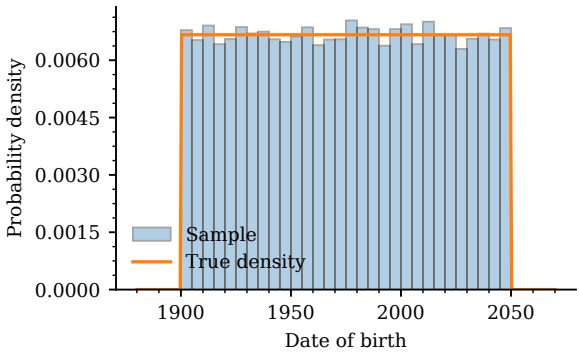
- Each individual $X_i$ in the time-independent framework can be mapped in the time-dependent framework $Y_{t,i} = T(X_i, t)$
- Observed individual $\tilde{Y}_{t,j} = Y_{t,j} + \varepsilon = T(X, t) + \varepsilon$
- With $\varepsilon$ distributed,

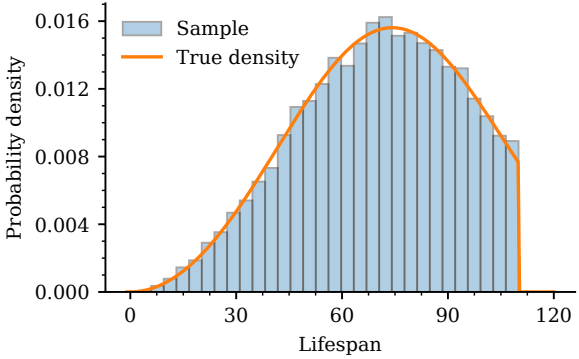
$$P(\tilde{Y}_t \mid X) = \prod_{j=1}^{n} \frac{1}{M_t(X)} \sum_{i \in \mathcal{I}_t(X)} p_\varepsilon\big(\tilde{Y}_{t,j} - Y_{t,i}\big).$$

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

# Outline

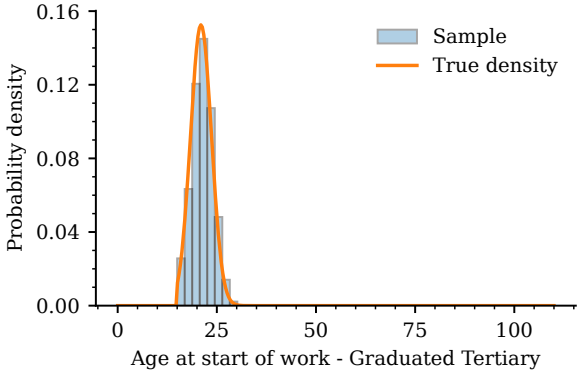
# Priors : Existence

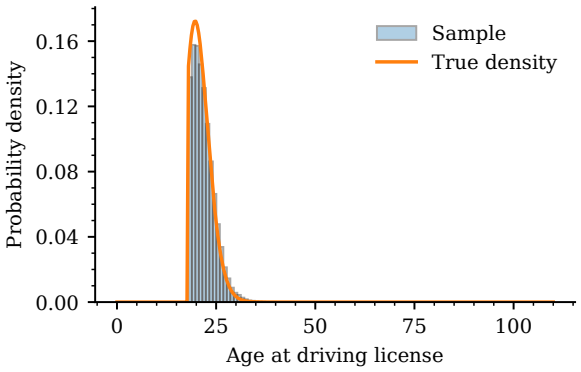



Date of birth and Lifespan sampled and true densities

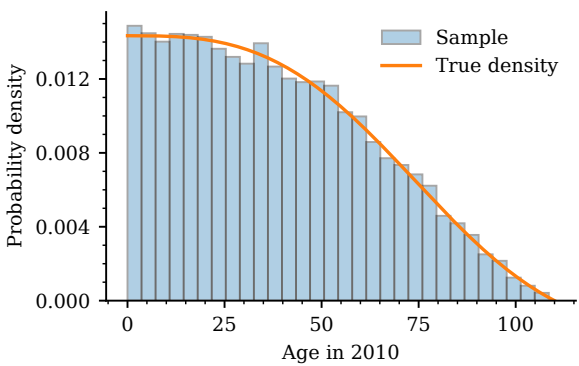
# Priors : Other dimensions

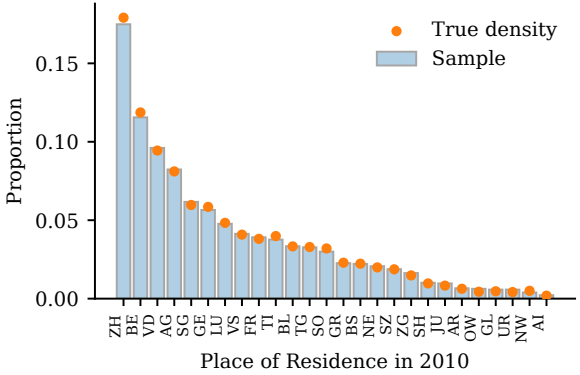



Age at labor entry (Graduated from Tertiary) and Driving license age

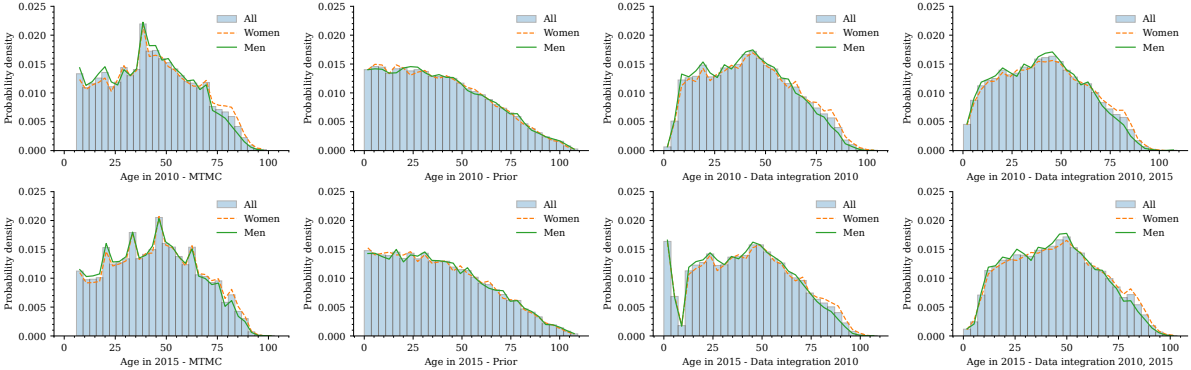
# Priors : Projection in 2010



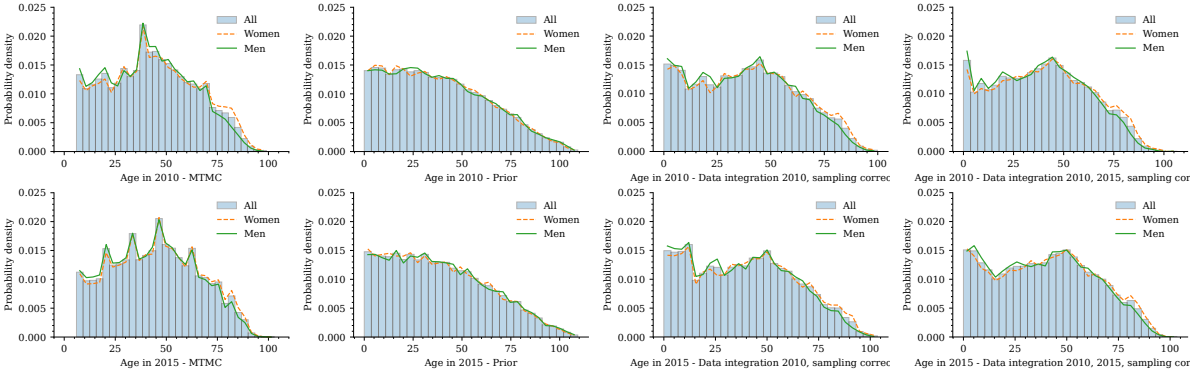

Age and canton of residence in 2010

# Existence naive posterior projected



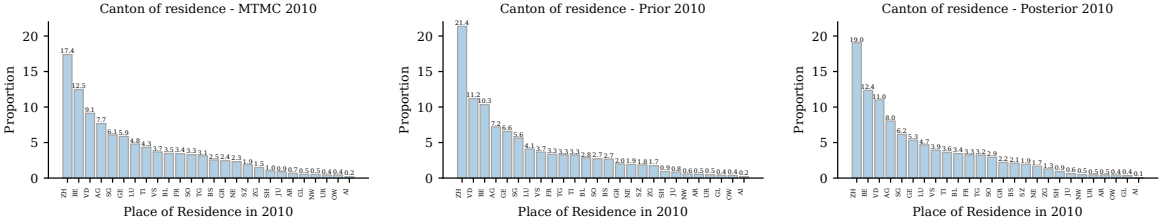

Observed data, prior, posterior with 2010, posterior with 2010 and 2015; in 2010 and 2015

# Existence corrected posterior projected

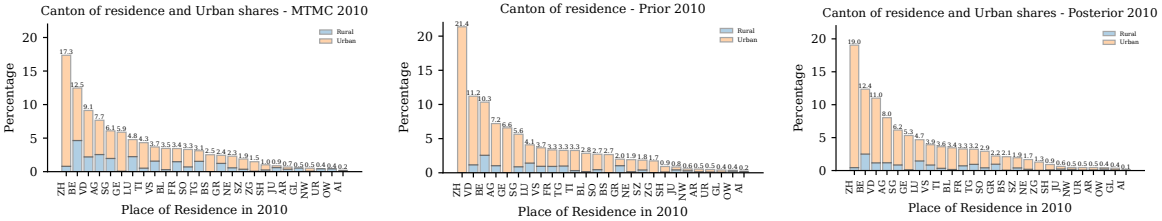



Observed data, prior, posterior with 2010, posterior with 2010 and 2015; in 2010 and 2015

# Other dimensions : Place of residence




Place of residence projected in 2010; data, prior, posterior



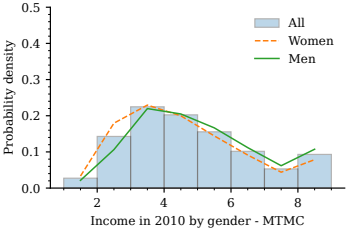

Place of residence projected in 2010 with urban shares; data, prior, posterior

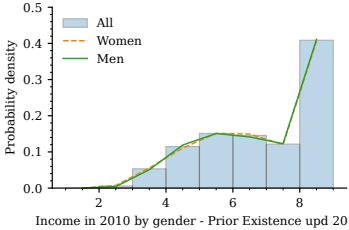
# Other dimensions : Employment status

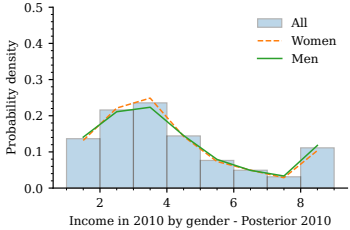
| Employment status | Prior | Posterior | Data |
|---|---|---|---|
| Unemployed / inactive | 39.19 | 27.85 | 13.10 |
| Employed | 32.56 | 43.42 | 59.11 |
| Retired | 18.17 | 17.95 | 17.85 |
| Under 15 | 10.08 | 10.08 | 9.92 |

Table: Employment status proportions in 2010.

# Other dimensions : Income

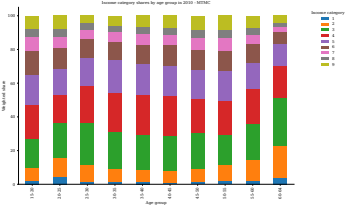


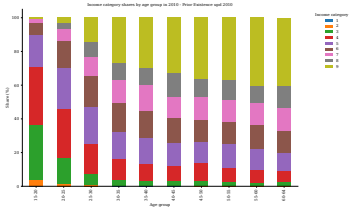



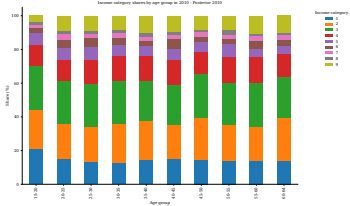




Posterior income distributions in 2010 : differentiated by age and gender

# Other dimensions : Income by age




➙ Over-representation of high incomes in the prior mitigated in the posterior; especially among older individuals

# 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 and social networks**
- Relax assumption of independence of the observed cross-sectional data-frames $\longrightarrow$ **Use panel data observations as input for the update**

# Questions ?

Scan the QR codes to access the Github repository !