Synthetic Population Generation

Generating panel data using Bayesian methods

Candice Baud, Michel Bierlaire

Transport and Mobility Laboratory
School of Architecture, Civil and Environmental Engineering
Ecole Polytechnique Fédérale de Lausanne

February 2, 3 - 2026

Outline

Why synthetic populations?

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

Acknowledgments

Marija Kukic and Michel Bierlaire (June 2025). "Gibbs Sampler for Generating Longitudinal Synthetic Populations". In: *Proceedings of the 12th Triennial Symposium on Transportation Analysis (TRISTAN XII)*. Japan

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 life 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

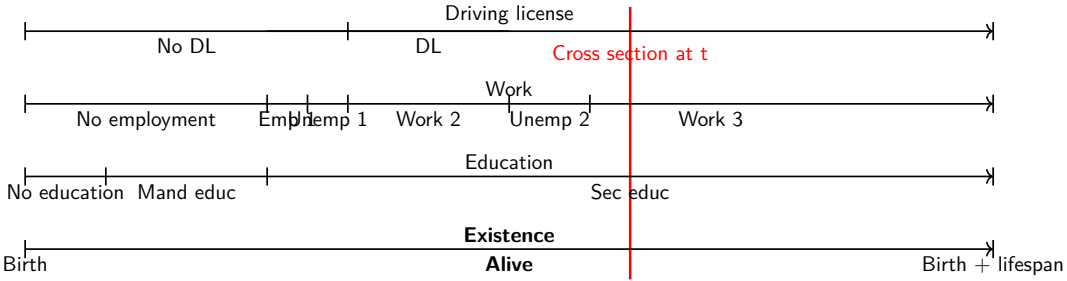

Figure: Example geometrical representation

Building block : Event

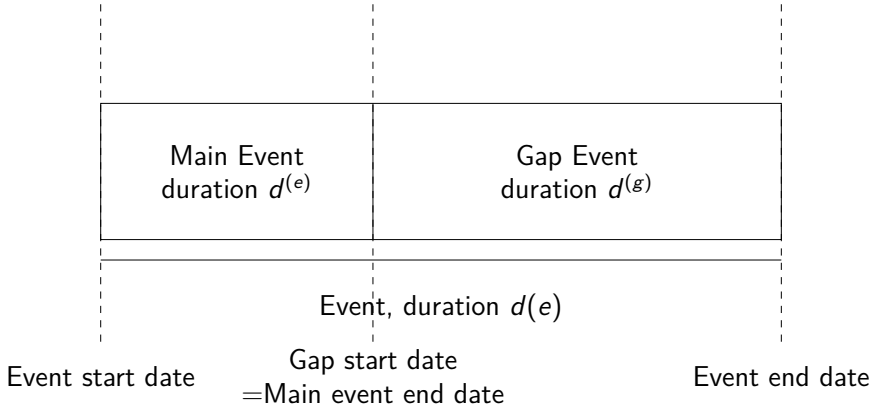

Figure: Event representation

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

Sampling process

Need to sample

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

$i(e)$ 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 with embedded **Metropolis–Hastings** steps:
  - Date of birth and lifespan: **Metropolis–Hastings**
  - Event occurrence indicators: **Gibbs sampler**
  - Event durations and attributes: **Metropolis–Hastings on a convex set**

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

Metropolis Hastings in convex space

General idea

- All the constraints are linear : defines a convex polytope
- Any point in the convex space can be generated by sampling

$$\lambda = (\lambda_1, \lambda_2, ...\lambda_3), \quad \lambda_i > 0, \quad \sum_k \lambda_k = 1, \quad x = \sum_k \lambda_k \cdot V_k$$

where V are the vertices

- Find the vertices to generate a point in the convex space that respects all constraints and avoid domain rejection
- Improves speed of convergence and rejection rate of the chain

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

Motivation

- 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

Simplifying assumption

Assumption :

- Cross-sectional measurements are independent

$$\mathcal{L}((\tilde{Y}_t)_t|X) = \prod_t \mathcal{L}(\tilde{Y}_t|X)$$

- That means, for each simulated time-independent individual, the likelihood is

$$\mathcal{L}((\tilde{Y}_t)_t|X_i) = \prod_{t,\text{alive}(X_i,t)=1} \mathcal{L}(\tilde{Y}_t|X_i)$$

Challenge

- At the individual level, the likelihood is deterministic

Example :

- **Observed individual (2025):**

$$\text{Age} = 24, \quad \text{Driving license} = 0, \quad \text{Sex} = \text{Female}$$

- **Simulated time-independent individual:**

$$\text{Date of birth} = 06/11/2001, \quad \text{Lifespan} = 110$$

Event	Start date	Duration	Attribute
Sex	06/11/2001	110	Male
No driving license	06/11/2001	110	–

➨ The likelihood of the time-dependent individual conditional on the time-independent one is **zero**. Because the mapping of the time-independent individual in 2025 produces a mismatch (Sex).

2 way of calculating the likelihood

IPF-like likelihood :

Likelihood evaluator at the Population level with "counts" : similar to IPF methods

$$\mathcal{L}(\tilde{Y}_{t_1}|X_i) = \frac{1}{n}\sum_{k=1}^{n_Y} P(\tilde{Y}_{t_1,k}|X_i) = \frac{1}{n}\sum_{k=1}^{n_Y} \mathbf{1}\left(T(X_i,t) = \tilde{Y}_{t_1,k}\right)$$

Annealing-like likelihood :

Likelihood evaluator at the Individual level using a noise assumption

$$\tilde{Y}_{t_1} = Y_{t_1} + \varepsilon = T(X,t_1) + \varepsilon$$

$$P(\tilde{Y}_{t_1,k}|X_i) = f_\varepsilon(T(X_i,t_1) - \tilde{Y}_{t_1,k}), \quad f_\varepsilon = \mathcal{N}_{\mu=0,\ \sigma}$$

Sampling from the posterior

Algorithm :

- Metropolis Hastings algorithm targeting the posterior distribution
- Decreasing noise of the normal distribution for Annealing method

Key components :

- A mapping of the time-independent individuals at any time t
- Priors from the literature that we can evaluate
- Observed cross-sectional data and associated times of observation

Outline

Contributions and future work

Contributions :

- Generate **panel individuals and populations** from the time-independent framework
- Generation based only on the model (sample-free) $\longrightarrow$ **Prior sampling**
- Generation taking into account to readjust the model $\longrightarrow$ **IPF and Annealing**

Future work :

- Get results with **real data** (Swiss micro-census)
- Relax assumption of individuals independence $\longrightarrow$ **Households**
- Relax assumption of independence of the observed cross-sectional data-frames $\longrightarrow$ **Integrate panel data observations**