

A Flexible and Realistic Synthetic Panel Population Generation

Candice Baud, Michel Bierlaire

TRANSP-OR laboratory, EPFL

16 September 2026



Outline

- 1 Context and motivation
- 2 Framework
- 3 Case study
- 4 Conclusion

Context : cross-sections vs panel data

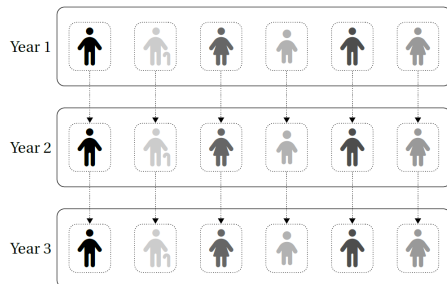
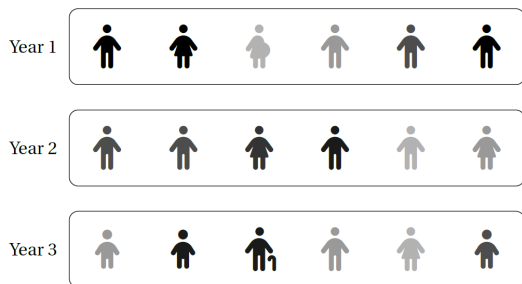


Figure: Cross-section vs panel data

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

Motivation

Data availability

Cross sectional data : widely available, but no identification of causal effects

Longitudinal data : rare and costly; subject to attrition, panel fatigue, and conditioning effects, often cover only limited time horizons

Transportation systems

Activity-based models rely on synthetic populations to represent the socio-economic characteristics of individuals whose activities and travel behavior are simulated

Research questions addressed

How to generate longitudinal individuals ?

How to ensure practical feasibility of the individuals ?

How to define a framework general enough to generate such populations at arbitrary temporal and spatial scales ?

Outline

- 1 Context and motivation
- 2 Framework**
- 3 Case study
- 4 Conclusion

Inherent longitudinal structure

Generation of full individuals life trajectories

Definition of a mapping mechanism at any time t , recovering the state of individuals hence creating **longitudinal** data

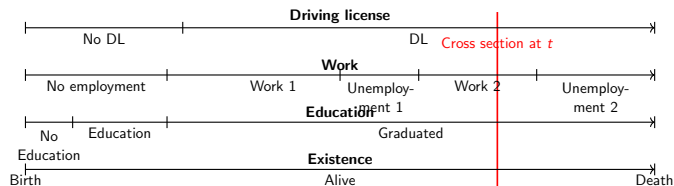


Figure: Complete life trajectory and cross-section at t

Event-based life trajectories

Decomposition and representation of life trajectories along **dimensions**, in which **events** form a **partition** of the individual lifespan

Event = building block, decomposed in a "main" phase and a "gap" phase

To each event are attached **attributes** characterizing it

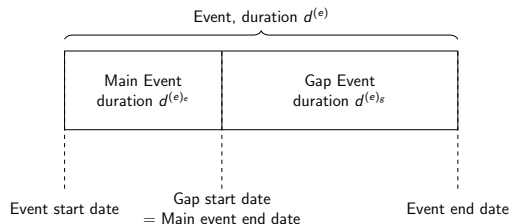


Figure: Event representation

Structural feasibility of the trajectories

Partition of the lifespan

Mutual exclusivity:

$$I_{\tau_D,i} \cap I_{\tau_D,j} = \emptyset \quad \text{for } i \neq j.$$

Full coverage :

$$\cup_{k=1}^{n_D} I_{\tau_D,k} = [b, b + L]$$

Positivity:

$$i_D^{(e_k)} = 1 \implies d^{(e_k,D)_e} > 0.$$

Additional constraints

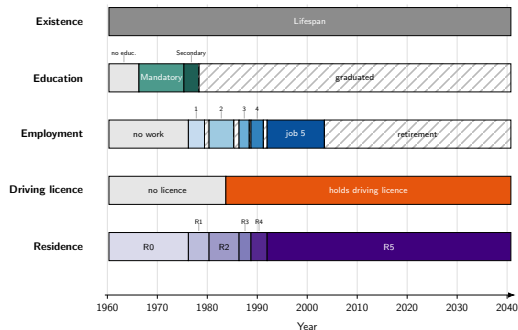
Start age minimum and maximum, durations minimum and maximum

Relations with other events of other dimensions

→ All constraints can be expressed using linear algebra

Life trajectory representation

Example life trajectory



Work dimension with attributes



At each time, one can make a cross-section of the state of the individual in each dimension

Vector to sample

Trajectories

Individual life trajectories for each dimension $T = (\tau_D)_{D=1}^N \stackrel{\text{iid}}{\sim} P_T$ follow the same law;
each individual sample is different realization

Population is a collection of independent draws $= \{T_l\}_{l=1}^n$,

Event-vector:

$$v(e) := (i(e), s(e), d(e), a_1(e), \dots, a_{n_a}(e))$$

i : indicator s : start d : duration a_k : attribute k

Trajectory:

$$\tau_D = (v(e_{1,D}), \dots, v(e_{n_e,D}, D))$$

Sampling and mapping

Sampling

Existence : Date of birth, lifespan using Hit-and-Run. Attributes are sampled using Metropolis Hastings.

Indicators : Gibbs sampler, conditionnally on *Existence* and indicators of other events.

Other dimensions : For each group of independent dimensions, sample the durations by Hit-and-Run. Attributes are sampled using Metropolis Hastings.

Mapping

Alive status

$$\text{Alive}_t = \begin{cases} 1 & \text{if } b \leq t < b + L \\ 0 & \text{otherwise} \end{cases}$$

Find the unique active event for each dimension, such that t lies between its start date and end date

Outline

- 1 Context and motivation
- 2 Framework
- 3 Case study**
- 4 Conclusion

Switzerland case study

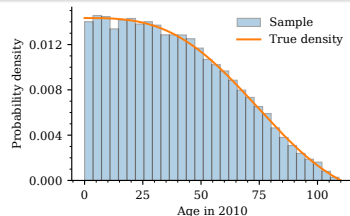
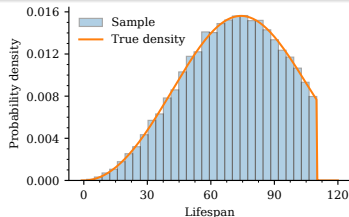
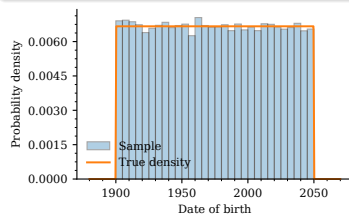
Table: Event and dimensions considered in the model

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

Results : Existence

Sampled variables : Date of birth b follows a Uniform distribution and Lifespan L a Weibull distribution of parameters $\lambda = 85$ and $k = 3$

Constraints : $0 \leq \text{Lifespan} \leq 110$, $1900 \leq \text{Date of birth} \leq 2050$



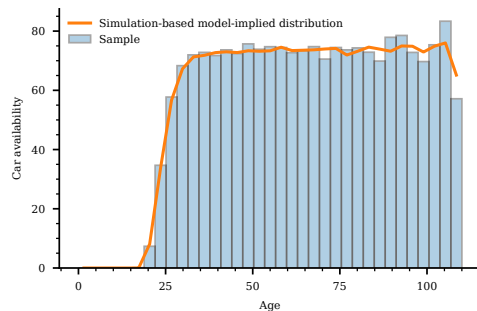
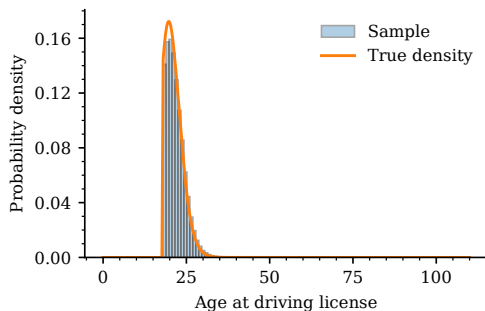
Date of birth and Lifespan sampled and true densities on the left. Age projected in 2010 on the right.

Age is derived from Date of birth and Lifespan : $\text{Age}_t = t - b$ if $b \leq t < b + L$

Results: Driving license

Variables : Age at driving license follows $\log\mathcal{N}(\mu = \ln(20.5), \sigma^2 = 0.15)$. Car availability follows a Bernoulli distribution of parameter 0.76 conditional on holding a license

Constraints : A driving license can be obtained only after 18 years old

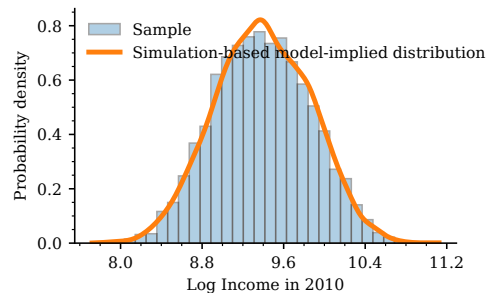
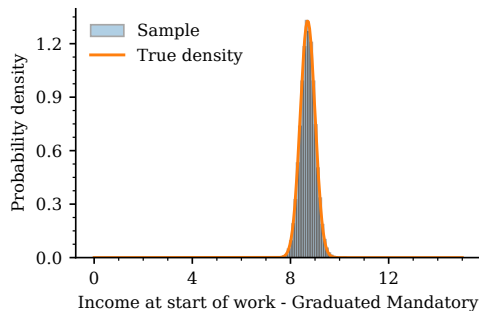


Results: Income at start of work

Variables : The income distribution associated to each work spell is

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

$$\varepsilon_m \sim \mathcal{N}(\mu = 0, \sigma^2 = 0.3), \quad \beta_0 = 7.8, \quad \rho = 0.1, \quad \beta_1 = 0.025, \quad \beta_2 = -0.0004$$



Results : Number of jobs per Lifespan bin

The number of employment spells follows a $\text{Poisson}(\gamma(L))$ 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.

Table: 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

Outline

- 1 Context and motivation
- 2 Framework
- 3 Case study
- 4 Conclusion**

Conclusion

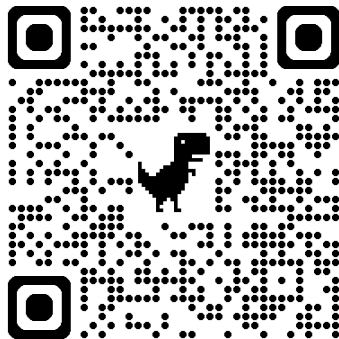
Contributions :

- ➡ Generate **life trajectories** from any probabilistic model ensuring consistency and feasibility
- ➡ The mapping mechanism recovers the state of each individual at any time t , hence creating longitudinal data following the same entities

Future work

- ➡ Relax assumption of individuals independence → **Households and social networks**
- ➡ Use external data (cross-section or longitudinal data) to improve the generation of the trajectories → **Bayesian posterior sampling**

Github QR



Zenodo DOI : [10.5281/zenodo.21293241](https://doi.org/10.5281/zenodo.21293241)