

From Priors to Data: A Flexible Bayesian Framework for Panel Synthetic Population Generation

Candice Baud, Michel Bierlaire

TRANSP-OR laboratory, EPFL

30th September 2026



Outline

- 1 Context and motivation
- 2 Prior Framework
- 3 Posterior sampling
- 4 Case study
- 5 Conclusion

Context : cross-sections vs panel data

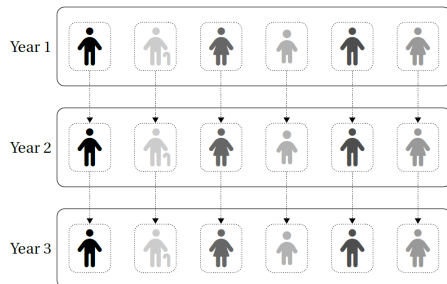
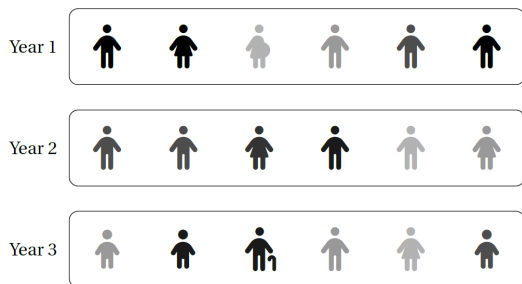


Figure: Cross-section vs panel data

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 use information in cross-sectional data to enrich the generation ?

To what extent multiple datasets improve the generation more than one ?

Outline

- 1 Context and motivation
- 2 Prior Framework**
- 3 Posterior sampling
- 4 Case study
- 5 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

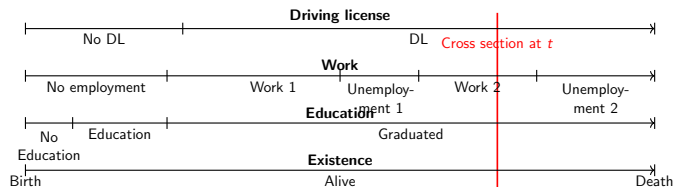


Figure: Complete life trajectory and cross-section at t

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*.

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 Prior Framework
- 3 Posterior sampling**
- 4 Case study
- 5 Conclusion

- ➡ 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 data information

Bayesian update

Posterior :

$$f(\mathbf{X} \mid \tilde{Y}_t) \propto f(\tilde{Y}_t \mid \mathbf{X}) \cdot \pi_X(\mathbf{X}), \quad \tilde{Y}_{t_1} \mid X \perp \tilde{Y}_{t_2} \mid X$$

Population-mixture likelihood : an observed individual is compared simultaneously with every alive synthetic individual.

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

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

$$\mathcal{I}_t(\mathbf{X}) = \{i, \text{Alive}_{t,i} = 1\}, \quad M_t(\mathbf{X}) = |\mathcal{I}_t(\mathbf{X})|, \quad Y_{t,i} = T_t(X_i)$$

Small example

	Date of birth	Lifespan	Life status in 2000
Lucy	3.2 million years BC	?	Dead
Cleopatra	69 BC	39	Dead
Michael Jackson	1958	50	Alive
Michel Bierlaire	1967	> 59	Alive
Candice Baud	2001	> 24	Not born yet

Table: Renowned individuals and their life statuses in 2000

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:

$$-\frac{1}{2\sigma^2} \int_{t_{\min}}^{t_{\max}} \left(\tilde{M}_t(\mathbf{X}) - M^*(\mathbf{X}) \right)^2 dt \rightarrow -\frac{1}{2\sigma^2} \sum_k w_k \left(\tilde{M}_{t_k}(\mathbf{X}) - M_{t_k}^*(\mathbf{X}) \right)$$

Outline

- 1 Context and motivation
- 2 Prior Framework
- 3 Posterior sampling
- 4 Case study**
- 5 Conclusion

Switzerland case study

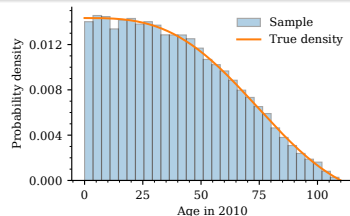
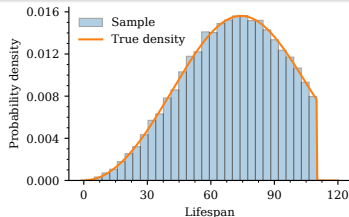
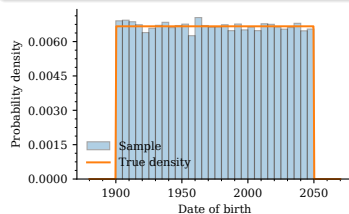
Table: Event and dimensions considered in the model

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	$\emptyset$ $\emptyset$ $\emptyset$ $\emptyset$
Work	No employment Employment spells	$\emptyset$ Income at start of work spell Place of Work
Driving license ownership	No driving license Driving license ownership	$\emptyset$ $\emptyset$

Results : Existence, priors

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

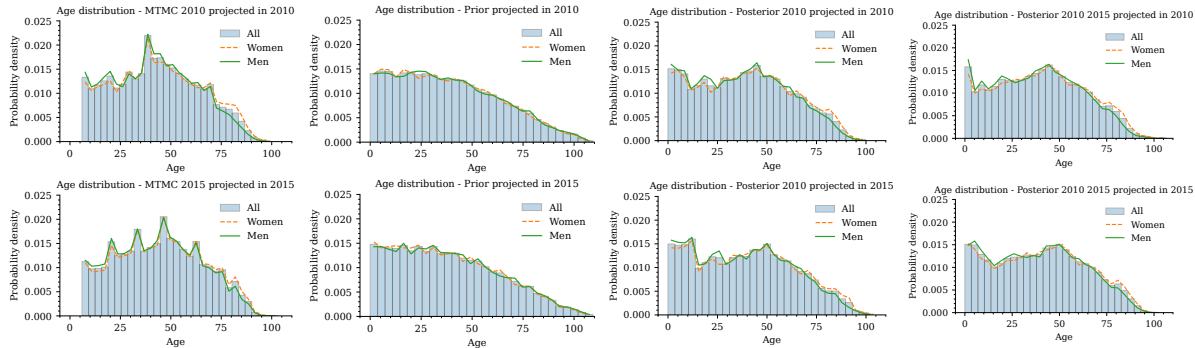
Constraints : $0 \leq \text{Lifespan} \leq 110$, $1900 \leq \text{Date of birth} \leq 2050$



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 : Existence, posterior



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

Results: Improvement by adding dataframes

Table: 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]

➡ Kernel's smoothing imposes a resolution limit of under 1 year on any in-sample comparison, independent of remaining model misfit. All update procedures evaluated in-sample fall below this bound suggesting the calibration is near the resolution allowed by the likelihood specification itself.

Results : Number of jobs per Lifespan bin, prior

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

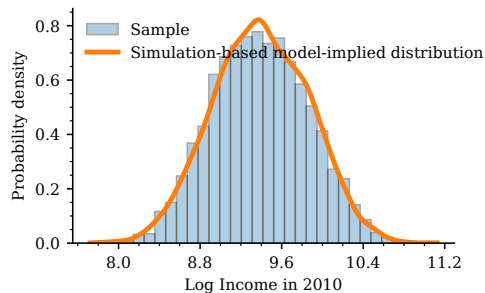
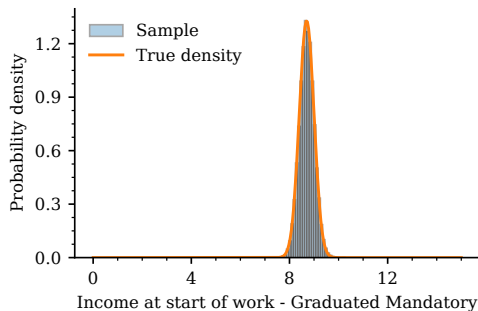
Lifespan Bin	[0, 15]	]15, 20]	]20, 25]	]25, 30]	]30, 35]	]35, 40]	]40, 45]	]45, 50]	]50, 55]	]55, 60]	]60, 64]	]64, 110]
Empirical Mean	0.00	1.55	2.77	3.62	4.17	4.72	5.19	5.65	5.91	6.35	6.58	6.67
Theoretical Mean	0.00	1.58	2.74	3.53	4.17	4.72	5.19	5.60	5.97	6.29	6.56	6.67

Results: Income at start of work, prior

Variables : The income distribution associated to each work spell is

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

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

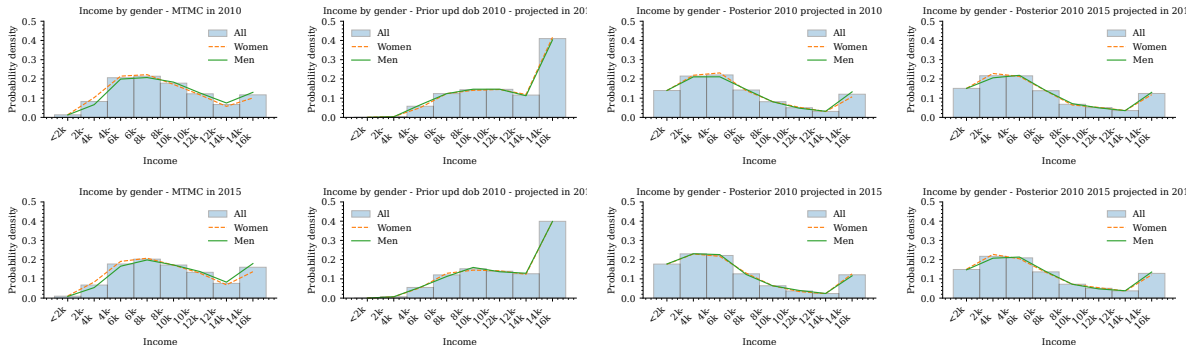


Results: Employment status, posterior

Table: 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

Results: Income, posterior



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

Outline

- 1 Context and motivation
- 2 Prior Framework
- 3 Posterior sampling
- 4 Case study
- 5 Conclusion**

Conclusion

Contributions :

- ➡ Generate **life trajectories** from any probabilistic model ensuring consistency and feasibility that can be mapped to recover the state of each individual at any time t , hence creating longitudinal data following the same entities
- ➡ **Update** the prior model with observed cross-sectional data which **improves** the sampling

Future work

- ➡ Relax assumption of individuals independence → **Households and social networks**
- ➡ Relax assumption of conditionally independent observed data-frames → **Use longitudinal data to improve the generation of the trajectories**

Github QR



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