

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

Baud Candice<sup>\*1</sup> and Bierlaire Michel<sup>2</sup>

<sup>1</sup>PhD student, Transport and Mobility Laboratory, EPFL, Switzerland

<sup>2</sup>Professor, Transport and Mobility Laboratory, EPFL, Switzerland

## SHORT SUMMARY

Most methods for generating synthetic populations rely on cross-sectional snapshots or pseudo-panels, which do not track individuals consistently over time. This paper proposes a general framework for constructing synthetic populations whose panel structure is specified by design. Individuals are represented through life-based trajectories defined independently of calendar time, and a deterministic mapping recovers their state at any time  $t$ , allowing the reconstruction of panel data and population distributions at arbitrary points in time. The framework is model-agnostic and enforces internal consistency through structural constraints embedded in the life representation. We further introduce a Bayesian updating mechanism that incorporates information from observed cross-sectional datasets. When data are available, the synthetic population is sampled from the posterior distribution, combining prior knowledge with the evidence contained in the observations. This allows cross-sectional information, such as census data, to inform the generation of coherent longitudinal populations.

**Keywords:** Activity-based models, Bayesian Statistics, Simulation, Synthetic Population Generation

## 1 INTRODUCTION

Activity-based models (ABMs) rely on synthetic populations to represent individuals whose activities and travel behavior are simulated Rezvani et al. (2024); Müller & Axhausen (2010). In practice, these populations are often constructed from cross-sectional data, which provide large and accessible snapshots of individuals at a given time Borysov & Rich (2021). While informative, such data lack a temporal dimension and therefore do not capture behavioral dynamics, persistence, or causal effects Wooldridge (2010); Maier et al. (2023); Bhat & Koppelman (1999).

Panel data partially address this limitation by observing individuals over time, enabling the identification of dynamic effects and heterogeneity Frees (2004); Hsiao (2014). However, they are costly, subject to attrition, and typically limited in scope Deschaintres et al. (2022). As a result, most synthetic population methods reproduce marginal distributions at different time points using fitting or simulation approaches Farooq et al. (2013); Kukic & Bierlaire (2025); Kukic et al. (2024), without ensuring that individuals correspond to consistent entities across time. When panel structures are needed, projection models or pseudo-panel techniques are typically used Deaton (1985); Verbeek & Nijman (1998), but these approaches infer dynamics rather than reconstruct coherent individual trajectories. The fundamental limitation is therefore the absence of a time-consistent generative representation of individuals.

In this paper, we propose a general framework to generate synthetic panel populations by construction. Individuals are defined through time-independent variables that encode their full life trajectories, following the time-independent principle of Kukic et al. (2025). The state of an individual at time  $t$  is obtained through a deterministic mapping, ensuring temporal consistency over the entire time horizon. The framework is model-agnostic, fully generative, and guarantees micro-level coherence across periods.

The second contribution is the integration of information from observed data through a Bayesian framework that updates the distribution of life trajectories using cross-sectional observations. Starting from a prior distribution over time-independent individual representations, a likelihood function is defined using the observed cross-sectional data. The posterior distribution is obtained using MCMC methods. This approach allows cross-sectional data to inform the full life-course distribution, despite the absence of longitudinal observations.

## 2 METHODOLOGY

### *Framework definition*

This paper builds on the time-independent paradigm of Kukic et al. (2025). Instead of modeling variables at a fixed time  $t$ , individuals are described by **time-independent variables** encoding their full life trajectory. Time-dependent states are then obtained through a deterministic mapping. For instance, age is not modeled directly but derived from the date of birth  $b$  and lifespan  $L$ , ensuring temporal consistency by construction.

Each individual is represented by an existence interval  $[b, b+L]$ , which is decomposed into sequences of events organized into dimensions (e.g., employment, education). A dimension defines a set of admissible event types and a common attribute structure. For each individual, it generates an ordered sequence of events, each characterized by an occurrence indicator, temporal attributes, and optional non-temporal attributes.

Events are represented as intervals composed of a main duration and an optional gap. Within each dimension, these intervals form a partition of the lifespan satisfying:

$$\begin{aligned} \textbf{Mutual exclusivity: } & I_{\tau_D, i} \cap I_{\tau_D, j} = \emptyset \quad (i \neq j), \\ \textbf{Full coverage: } & \bigcup_{k=1}^{n_D} I_{\tau_D, k} = [b, b+L], \quad \sum_{k=1}^{n_D} d^{(e_k, D)} = L, \\ \textbf{Positivity: } & i_D^{(e_k)} = 1 \Rightarrow d^{(e_k, D)} > 0. \end{aligned}$$

To ensure well-defined trajectories, each dimension includes a default “no-event” state that is active from birth whenever no substantive event occurs. Two dimensions are exceptions: *Existence*, which contains the Birth event and defines the bounds  $b$  and  $L$  of the individual’s life (a temporal axis), and *Residence*, which represents a spatial axis. In these dimensions, individuals must exist and be located from birth onward. Therefore, trajectories start directly with a substantive event, ensuring full coverage of the lifespan without introducing an artificial placeholder.

Admissibility is enforced through structural constraints. Events must satisfy precedence relations (e.g., education before employment) and temporal bounds on starting ages and durations reflecting biological, legal, and institutional limits. Additional inter-dimensional constraints can link events across domains. All constraints are linear, so the set of feasible trajectories forms a convex region.

Synthetic trajectories are generated using a structured MCMC procedure combining Gibbs sampling, Metropolis–Hastings updates Metropolis et al. (1953); Hastings (1970); Gelfand & Smith (1990), and Hit-and-Run sampling Smith (1984); Lovász & Vempala (2006). First, the existence dimension (birth and lifespan) is sampled, defining the temporal support of all events. Conditional on this, event indicators are drawn using Gibbs updates under structural constraints (e.g., age bounds and precedence relations). Given the indicators, event attributes are sampled using a block Gibbs–Metropolis–Hastings scheme. Durations are sampled within the convex feasible region defined by the linear constraints. Finally, event starting times are recovered deterministically from the sampled durations. The complete methodology is available in Bierlaire & Baud (2026).

### *Bayesian update*

We assume that individual-level cross-sectional data  $\tilde{Y}_t$  of a population are observed at time  $t$ . Let  $X_i$  denote the primary time-independent variables of synthetic individual  $i$ , let  $\mathbf{X} = (X_1, \dots, X_N)$  denote the full synthetic population of size  $N$ . Given the observed sample  $\tilde{Y}_t = \{\tilde{Y}_{t,j}\}_{j=1}^n$ , the objective is to sample  $\mathbf{X}$  from the posterior:  $f(\mathbf{X} | \tilde{Y}_t) \propto f(\tilde{Y}_t | \mathbf{X}) \cdot \pi_X(\mathbf{X})$ , which combines the prior distribution  $\pi_X(\mathbf{X})$  with a likelihood  $f(\tilde{Y}_t | \mathbf{X})$ . The state of a synthetic individual  $i$  at time  $t$  is obtained by applying the time-specific mapping  $T_t$  to their trajectory:  $Y_{t,i} = T_t(X_i)$ . The empirical cross-sectional distribution induced by this eligible subpopulation is  $\mu_t^X = \frac{1}{M_t(\mathbf{X})} \sum_{i \in \mathcal{I}_t(\mathbf{X})} \delta_{Y_{t,i}}$ . Let  $K_t(\tilde{y}, y)$  denote an observation kernel quantifying the plausibility that an observed value  $\tilde{y}$  is consistent with a synthetic trajectory value  $y$  at time  $t$ . The likelihood of a single observed individual is then defined directly as the integral of the kernel against  $\mu_t^X$ , i.e. its average value over the eligible synthetic population:

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

Assuming observed individuals are independent draws given  $\mathbf{X}$ , the likelihood of the population cross-section  $\tilde{Y}_t = (\tilde{Y}_{t,j})_j$  is:  $L_t = P(\tilde{Y}_t | \mathbf{X}) = \prod_{j=1}^n \frac{1}{M_t(\mathbf{X})} \sum_{i \in \mathcal{I}_t(\mathbf{X})} K_t(\tilde{Y}_{t,j}, Y_{t,i})$ . Cross-sectional data inform only the distribution of  $Y_t$  among individuals alive at time  $t$ , and provide no direct information on individuals not alive at  $t$ . As a result, the posterior is not identified without additional constraints, since non-alive individuals do not enter the likelihood. To address this issue, we incorporate information on population size dynamics during the sampling of the Existence dimension. Let  $M_t^*$  denote a target population size at time  $t$ . We introduce the likelihood on the population size

$$\exp \left( -\frac{1}{2\sigma^2} \sum_k w_k (\log M_{t_k}(\mathbf{X}) - \log M_{t_k}^*)^2 \right),$$

which enforces consistency between simulated and target population sizes. The function  $M_t(\mathbf{X})$  is computed exactly from the simulated trajectories, using the birth dates and lifespans of individuals. It defines a piecewise-constant function over time, with increments of +1 at birth events and -1 at death events.

The existence dimension is sampled first and at the population level. The likelihood compares observed ages with simulated ages of alive individuals, while the population constraint ensures identification through the evolution of  $M_t(\mathbf{X})$ . The update is performed at the population level, as the population size constraint couples all individuals. Candidate populations are generated by modifying a subset of individuals (e.g., birth dates and lifespans), using proposals such as Hit-and-Run in the constrained space. The remaining dimensions are then sampled conditionally on the existence trajectories, and at the individual level. For each individual, event indicators and attributes are generated given the already sampled existence path. Note that cross-sectional data are not informative about lifetime event occurrences. For instance, observing employment at time  $t$  does not inform the number of past or future employment spells. More generally, cross-sectional measurements capture prevalence conditional on age and survival, and are subject to censoring and cohort effects. Therefore, event indicators are sampled from the prior, while durations and attributes are updated using MCMC targeting the posterior defined above. If several independent cross-sectional datasets  $\{\tilde{Y}_{t_k}\}_k$  are available, the posterior becomes

$$P(\mathbf{X} | \{\tilde{Y}_{t_k}\}_k) \propto \left( \prod_k P(\tilde{Y}_{t_k} | \mathbf{X}) \right) P(\mathbf{X}),$$

### 3 RESULTS AND DISCUSSION

#### *Probabilistic model definition*

Let us consider a synthetic life-course model based on the dimensions and events presented in Table 1. For selected events, attributes are defined: residence spells are associated with a canton

Table 1: Dimensions Considered

| Dimension                        | Associated Main Events      |
|----------------------------------|-----------------------------|
| <b>Existence</b>                 | Birth                       |
| <b>Residence</b>                 | Residential spells          |
| <b>Education</b>                 | No education                |
|                                  | Mandatory education         |
|                                  | Secondary education         |
|                                  | Tertiary education          |
| <b>Work</b>                      | No employment               |
|                                  | Employment spells           |
| <b>Driving license ownership</b> | No driving license          |
|                                  | Driving license acquisition |

of residence, while employment spells are characterized by a canton of work and an income level.

The model is governed by the following structural rules. The driving license, once obtained, is irrevocable. Education is sequential: mandatory education must be completed, no repetition or skipping is allowed, and once studies are left they cannot be resumed. Secondary education must last at least three years to allow access to tertiary education. Individuals may work and study simultaneously, thus defining two dimensions. We assume that residential moves occur only when employment changes.

Time constraints are defined as follows. Birth occurs in [1900, 2050] and lifespan is bounded above by 110 years. Mandatory education starts at age 6 and lasts 9 years. Secondary education starts at age 15 and lasts between 3 and 6 years, while tertiary education may start between ages 18 and 25 and lasts at most 15 years. Employment may start from age 15; the last employment spell must start before age 64 and end before age 70. A driving license may be obtained between ages 18 and 70.

The probabilistic specification follows Kukic et al. (2025). The date of birth is uniformly distributed over the considered period, while lifespan follows a Weibull distribution with parameters  $\lambda = 85$  and  $k = 3$ . Mandatory education is universal; secondary and tertiary education are undertaken with probabilities 0.95 and 0.60 (conditional on eligibility), respectively. The duration of secondary education follows a truncated normal distribution  $\mathcal{N}(4, 0.5)$  on  $[3, 6]$ , while tertiary education duration follows a Weibull distribution ( $\lambda = 4.5, k = 3.5$ ), with an exponential gap (mean 1) between secondary and tertiary education.

The number of employment spells follows a Poisson distribution with parameter 6. Entry into the labor market depends on education level: the duration of the initial no-employment spell is normally distributed with mean 16, 18, or 21 (and variances 2, 4, and 7) for mandatory, secondary, and tertiary education, respectively. The duration of an employment spell starting at age  $a_k$  follows an exponential distribution with mean  $\lambda(a_k)$  defined as

$$\lambda(a_k) = \begin{cases} 1.5 & a_k < 25, \\ 2.5 & 25 \leq a_k < 35, \\ 3.5 & 35 \leq a_k < 45, \\ 4.5 & 45 \leq a_k < 55, \\ 9 & a_k \geq 55. \end{cases}$$

Income during employment spell  $m$  follows

$$\ln(\text{inc}_m) = \beta_0 + \beta_1 \cdot \text{work}_m + \beta_2 \cdot \text{work}_m^2 + \rho \cdot \text{edu}_m + \varepsilon_m, \quad \varepsilon_m \sim \mathcal{N}(0, 0.3),$$

with parameters  $\beta_0 = 7.8$ ,  $\beta_1 = 0.025$ ,  $\beta_2 = -0.0004$ , and  $\rho = 0.1$ .

Driving license acquisition follows a log-normal distribution  $\log\mathcal{N}(\ln(20.5), 0.15)$ , and occurs for 85% of individuals.

Spatial attributes are assigned as follows. The canton of birth and work are sampled proportionally to population size, with a multiplicative factor of 1.5 for urban cantons when income is high. The residence canton coincides with the work canton with probability 0.95, and is otherwise drawn from neighboring cantons.

We sample 50,000 individuals from the proposed model. The Markov chains exhibit convergence toward the stationary distribution, and the empirical distributions closely match the theoretical specifications. Figure 1 compares sampled values with their corresponding densities for key variables, including date of birth, lifespan, age at labor market entry (tertiary graduates), and age at driving license acquisition. In all cases, the empirical distributions closely reproduce the theoretical targets.

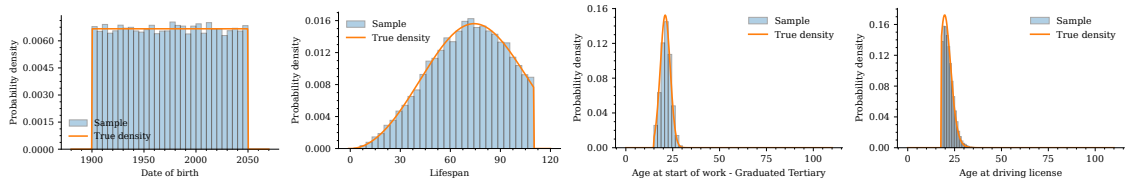


Figure 1: Empirical samples and theoretical densities for key variables: date of birth, lifespan, age at labor market entry (tertiary graduates), and age at driving license acquisition.

Event frequencies are likewise consistent with the imposed probabilities, conditional on survival constraints. Among individuals whose lifespan exceeds 18 years, 84.6% obtain a driving license.

All individuals with lifespan greater than 6 undergo mandatory education. Among those living beyond 15 years, 94% complete secondary education, and among those living beyond 18 years, 57% complete tertiary education. These proportions are in line with the model specification. The number of job spells also follows the intended truncated Poisson(6) distribution, truncated at 11 spells. Table 2 reports the empirical and theoretical survival probabilities over the support  $\{0, \dots, 11\}$ . The close correspondence between empirical and theoretical values confirms that the truncation and sampling procedure correctly reproduces the target distribution.

Table 2: Empirical vs. Truncated Poisson(6) Survival Probabilities (%)

| $x$                      | 0      | 1     | 2     | 3     | 4     | 5     | 6     | 7     | 8     | 9     | 10   | 11   |
|--------------------------|--------|-------|-------|-------|-------|-------|-------|-------|-------|-------|------|------|
| <b>Empirical</b>         | 100.00 | 99.27 | 97.82 | 93.41 | 85.01 | 71.16 | 56.09 | 40.03 | 25.57 | 16.02 | 8.9  | 5.3  |
| <b>Trunc. Poisson(6)</b> | 100.00 | 99.75 | 98.24 | 93.71 | 84.65 | 71.07 | 54.77 | 38.47 | 24.48 | 13.99 | 7.00 | 3.21 |

From these time-independent sampled variables, time-dependent cross-sectional datasets can be derived for any time  $t$ . The implied age distributions at different dates match the theoretical distributions obtained in closed form from the underlying birth and lifespan laws. Similarly, the distribution of place of residence in 2010 is consistent with the specified density. Figure 2 illustrates the agreement between empirical and theoretical distributions for both age and canton of residence in 2010.

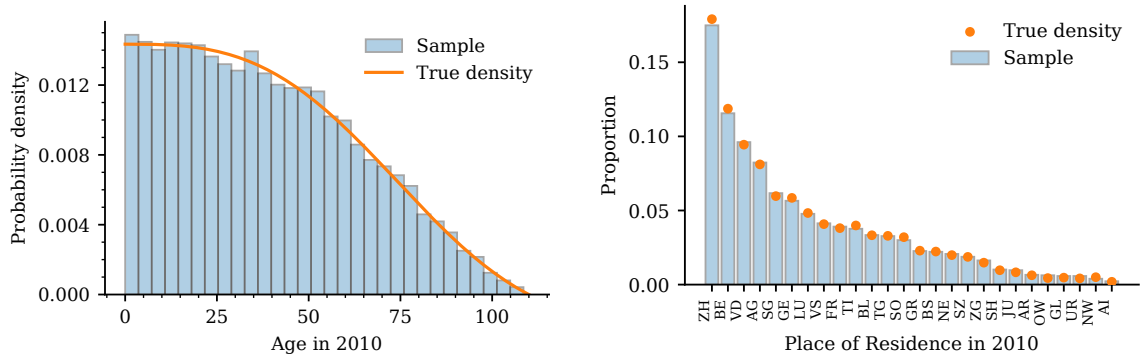


Figure 2: Age and canton distribution in 2010

Overall, the empirical results validate the framework at two levels. The marginal distributions of the time-independent variables are correctly reproduced by the sampling procedure. Moreover, the mapping operator that transforms life trajectories into time-specific observations preserves the theoretical structure of the model, as the resulting cross-sectional distributions coincide with the analytically derived ones.

### Data integration

We use the Swiss Mobility and Transport Microcensus data to illustrate the data integration process and posterior sampling. We first consider the 2010 dataset, and then jointly use the 2010 and 2015 datasets to assess the impact of multiple observations. The observation noise  $\varepsilon$  is assumed Gaussian with standard deviation 5. The parameter  $\sigma^2$ , controlling the alive population size, is fixed to 1, allowing for a deviation of about 10%. The population size curve is specified constant. Figure 3 reports the age distributions. The first row shows the data for 2010 and 2015. The second row shows the prior model. The third row presents the posterior obtained using only the 2010 data. The last row presents the posterior obtained using both datasets. The Bayesian approach combines prior information and data in a consistent way.

The Microcensus excludes children aged 0 to 7, which directly affects the posterior distributions. This induces a gap in the posterior distributions, visible in subfigure (f). Births after 2010 are not informed by the data and therefore follow the prior. This creates the appearance of a sharp discontinuity, which is in fact a consequence of the data collection process rather than a structural feature of the model. Such artifacts should be mitigated by explicitly accounting for the sampling design.

Future births are not affected by present observations and follow the prior. In contrast, the past is partially reshaped. Individuals observed in 2010 must have been born between 1900 and 2010, assuming a maximum lifespan of 110 years. As a result, the posterior distribution of dates of birth

over this period differs from the prior uniform distribution. Since  $\text{age} = t - \text{date of birth}$ , the distribution of dates of birth mirrors the age distributions.

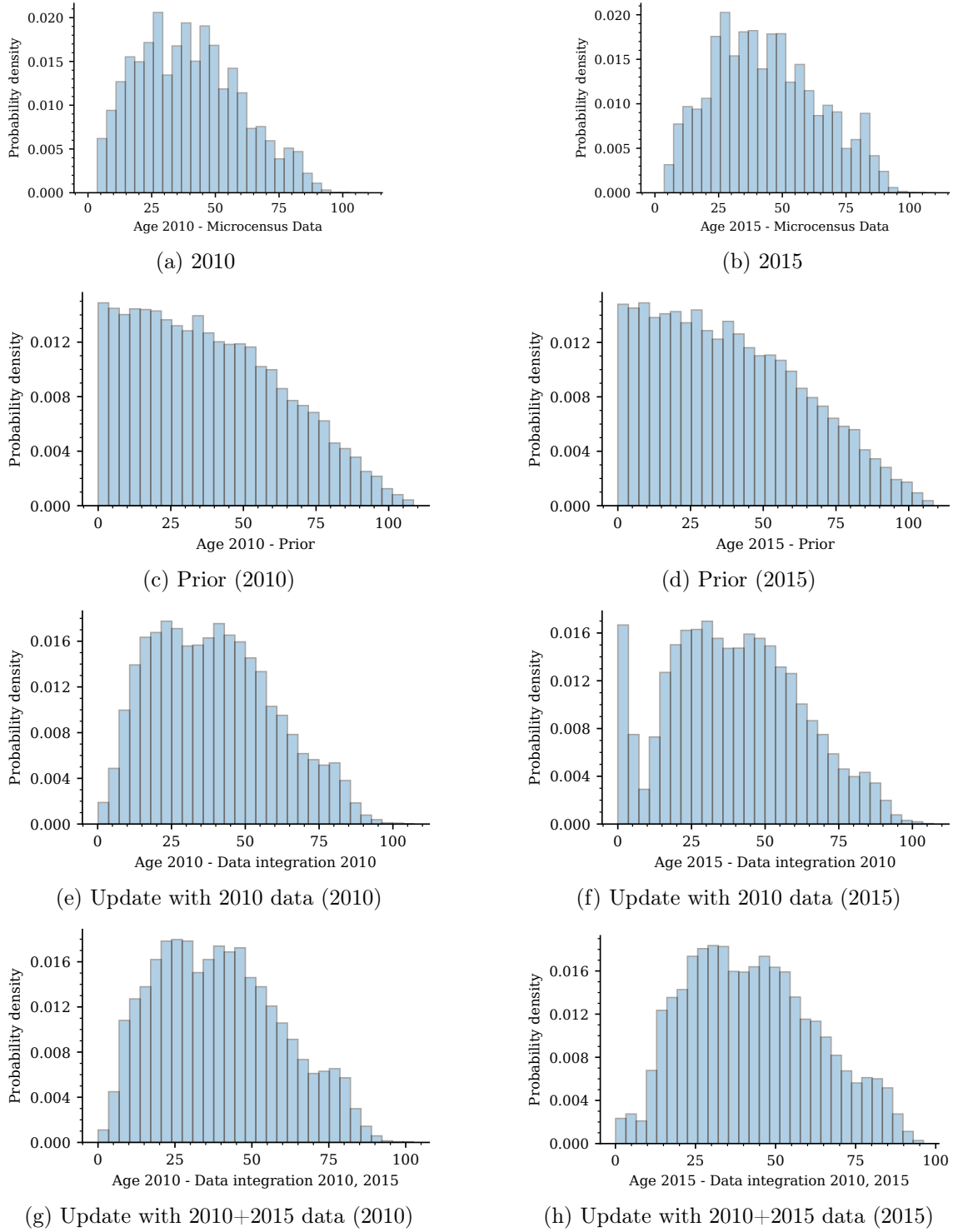


Figure 3: Age distributions. Columns correspond to years (2010 left, 2015 right). Rows correspond to observed data, prior, update with 2010 data, and update with 2010 and 2015 data.

Note that additional results on other dimensions will be presented during the conference.

## 4 CONCLUSIONS

In this paper, a general life-trajectory framework for synthetic panel population generation has been proposed. The approach relies on a unified and time-independent representation of individual life courses, from which states at any time are obtained through a deterministic mapping. The framework is realistic by construction. Biological, institutional, and logical constraints are embedded directly in the trajectory structure, preventing infeasible event sequences and ensuring plausible life-course dynamics. As a result, demographic evolution emerges endogenously from individual histories, rather than being imposed at each time period. The approach is also fully general and model-agnostic. Any event structure, attribute dynamic, or behavioral model that can be expressed within a life-trajectory representation can be incorporated. This flexibility allows the generation of complex demographic and behavioral systems at different temporal and spatial scales.

The contribution is completed by a Bayesian update mechanism that connects the generative model to observed data. The update operates through mappings between time-independent trajectories and time-dependent observations, and can accommodate multiple datasets observed at different time points. This provides a systematic way to enrich the prior model and to generate synthetic populations that are consistent with real data, while preserving the structural properties of the trajectories.

A key advantage of the panel structure is its ability to capture long-term compositional effects. Because individuals are generated with complete life trajectories, shocks propagate through time via the population structure. For example, a mortality shock affecting a cohort modifies the age distribution decades later. Such effects cannot be represented with purely cross-sectional approaches.

Future research will focus on relaxing two main assumptions : conditional independence of the observed data at different times, allowing for the integration of panel data for example; and independence of the individuals within a population, enabling to generate households.

## 5 RERITE CHALLENGE

The conference paper is eligible for the Reproducible Research Prize. The code and data associated with this work are available at [https://github.com/candicebaud/longitudinal\\_synthetic\\_population](https://github.com/candicebaud/longitudinal_synthetic_population) and <https://zenodo.org/records/21293242>. The repository is prepared at SRL4-5.

Some datasets cannot be publicly shared due to privacy restrictions. Therefore, synthetic example datasets are provided to reproduce the methodology.

## REFERENCES

- Bhat, C. R., & Koppelman, F. S. (1999). Activity-based modeling of travel demand. In R. W. Hall (Ed.), *Handbook of transportation science* (pp. 35–61). Boston, MA: Springer.
- Bierlaire, M., & Baud, C. (2026). A flexible and realistic synthetic panel population generation. In *Proceedings of the 2026 IEEE 29th international conference on intelligent transportation systems*. (In press)
- Borysov, S. S., & Rich, J. (2021). Introducing synthetic pseudo panels: Application to transport behaviour dynamics. *Transportation*, 48(5), 2493–2520.
- Deaton, A. (1985). Panel data from time series of cross-sections. *Journal of Econometrics*, 30(1–2), 109–126.
- Deschaintres, E., Morency, C., & Trépanier, M. (2022). Cross-analysis of the variability of travel behaviors using one-day trip diaries and longitudinal data. *Transportation Research Part A: Policy and Practice*, 163, 228–246.
- Farooq, B., Bierlaire, M., Hurtubia, L., & Flötteröd, G. (2013). Simulation-based population synthesis. *Transportation Research Part B*, 58, 243–263.
- Frees, E. W. (2004). *Longitudinal and panel data: Analysis and applications in the social sciences*. Cambridge: Cambridge University Press.
- Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410), 398–409.
- Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1), 97–109.
- Hsiao, C. (2014). *Analysis of panel data* (3rd ed.). Cambridge, U.K.: Cambridge University Press.
- Kukic, M., & Bierlaire, M. (2025). Adaptive synthetic generation using one-step gibbs sampler. *Transportation Research Interdisciplinary Perspectives*, 33, 101597.
- Kukic, M., Ilinov, P., & Bierlaire, M. (2025). *Simulation framework for generating synthetic panel data* (Tech. Rep. No. 251013). Transport and Mobility Laboratory (TRANSP-OR), EPFL.
- Kukic, M., Li, X., & Bierlaire, M. (2024). One-step gibbs sampling for the generation of synthetic households. *Transportation Research Part C: Emerging Technologies*, 166, 104770.
- Lovász, L., & Vempala, S. (2006). Hit-and-run from a corner. *SIAM Journal on Computing*, 35(4), 985–1005.
- Maier, C., Thatcher, J. B., Grover, V., & Dwivedi, Y. K. (2023). Cross-sectional research: A critical perspective, use cases, and recommendations for its research. *International Journal of Information Management*, 70, 102625.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6), 1087–1092.
- Müller, K., & Axhausen, K. W. (2010). Population synthesis for microsimulation: State of the art. In *Proceedings of the 10th swiss transport research conference (strc)*. Ascona, Switzerland.
- Rezvani, N., Kukic, M., & Bierlaire, M. (2024, December). A review of activity-based disaggregate travel demand models. *Findings*. doi: 10.32866/001c.125431
- Smith, R. L. (1984). Efficient monte carlo procedures for generating points uniformly distributed over bounded regions. *Operations Research*, 32(6), 1296–1308.
- Verbeek, M., & Nijman, T. (1998). Can cohort data be treated as genuine panel data? *Empirical Economics*, 23, 9–23.
- Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). Cambridge, MA: MIT Press.