

Motivation

Activity-based models require population data that are often unavailable due to confidentiality, leading to the use of synthetic populations. Existing methods mainly generate cross-sectional or pseudo-panel data, lacking a true temporal dimension and preventing consistent tracking of individuals over time. This work proposes a general framework to generate synthetic panel populations by construction. Building on [1], individuals are defined through time-independent variables encoding full life trajectories, ensuring temporal consistency. The framework is model-agnostic and produces realistic individuals at any scale. A Bayesian approach then incorporates cross-sectional data to update the distribution of trajectories, allowing inference on full life courses without longitudinal data.

Our goal is to generate realistic synthetic panel populations and develop a flexible framework that can be used data-free or recalibrated with observations.

Generating Panel individuals from a model

Time-independent space

The time-independent space defines a fundamental latent representation governing the life trajectories of individuals. Variables in this space are invariant with respect to the time of observation and are assumed to follow a stable distribution over the considered time horizon.

Time-dependent space

The time-dependent space is derived from the time-independent space through a deterministic mapping. At any time t , it corresponds to a projection of the time-independent variables, whose distribution is entirely induced by the underlying time-independent representation.

Age is defined relative to a time t , while the date of birth is time-invariant. Given the date of birth B and lifespan L , age is defined as

$$\text{Age}_t = t - B, \quad \text{for } B \leq t < B + L.$$

Individual life trajectories are described as sequences of **events** over time, such that the active state at any time can be uniquely identified. Events are organized into **dimensions**, grouping activities of similar nature. Within a dimension, events are mutually exclusive and occur sequentially, while events from different dimensions may overlap. Each realized event $e_{k,D}$ is represented by an interval

$$I_{\tau D,k} = [s^{(e_{k,D})}, s^{(e_{k,D})} + d^{(e_{k,D})}].$$

These intervals $I_{\tau D,k}$ satisfy three structural conditions:

- **Mutual exclusivity:** events within the same dimension cannot overlap:

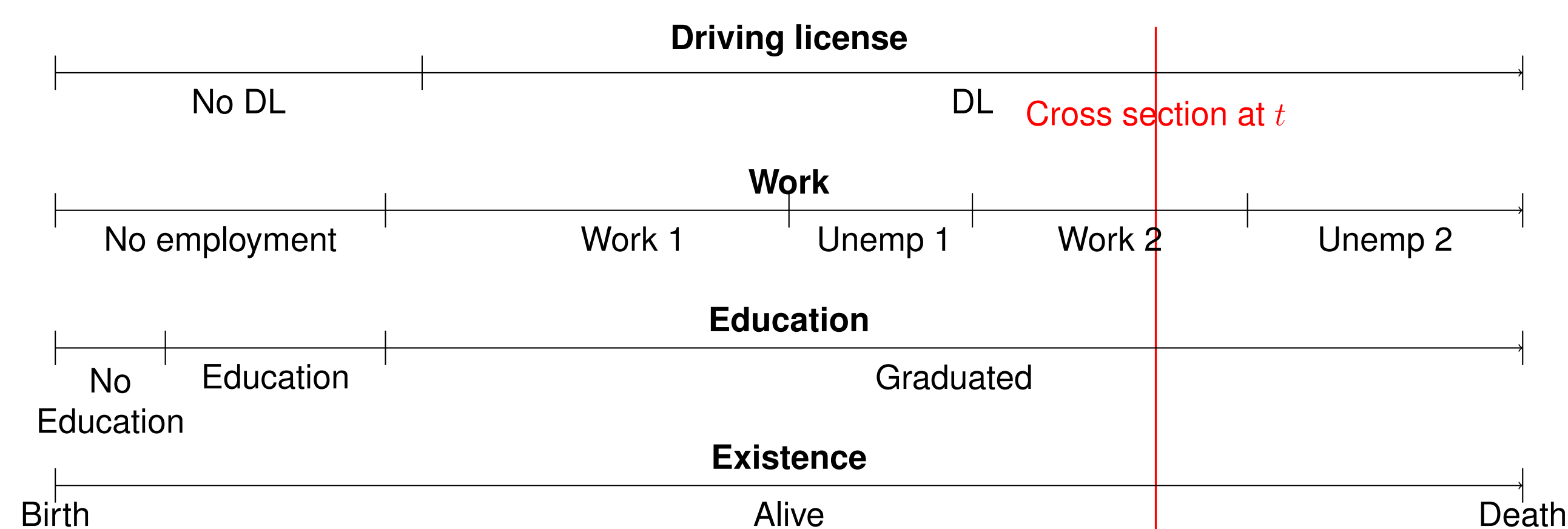
$$I_{\tau D,i} \cap I_{\tau D,j} = \emptyset \quad \text{for } i \neq j.$$

- **Full coverage:** the union of all event intervals coincides with the existence interval:

$$\bigcup_{k=1}^n I_{\tau D,k} = [b, b + L].$$

- **Positivity:** all main durations of occurring events are strictly positive:

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



Additionally, events may be subject to biological or legal constraints, such as no driving license before age 18, a maximum lifespan of 120 years, or car ownership only after obtaining a driving license. Together, these constraints define a convex admissible space.

To each event, one can attach attributes characterizing the event, such as income for a work spell or residence location for a residence spell.

Include information from observations

Using Bayes' rule, the posterior distribution of the time-independent variables X given the observed data $\tilde{Y}_t$ is

$$P(X|\tilde{Y}_t) \propto L(\tilde{Y}_t|X)P(X),$$

where $P(X)$ denotes the prior distribution over life trajectories, and $L(\tilde{Y}_t|X)$ is the likelihood of observing the data under X .

We model the observed data as a noisy measurement of the induced time-dependent process. Let $T(\cdot, t)$ be the deterministic mapping from time-independent variables to time t , and define

$$\tilde{Y}_t = T(X, t) + \varepsilon, \quad Y_t = T(X, t), \quad \varepsilon \sim p_\varepsilon.$$

That is, Y_t represents the true time-dependent variables, while $\tilde{Y}_t$ are noisy observations. Let $\mathcal{I}_t$ denote the set of individuals alive at time t , and $M_t(X) = |\mathcal{I}_t|$. Assuming independent observations, the likelihood is

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

Results

Time-independent variables from the prior model

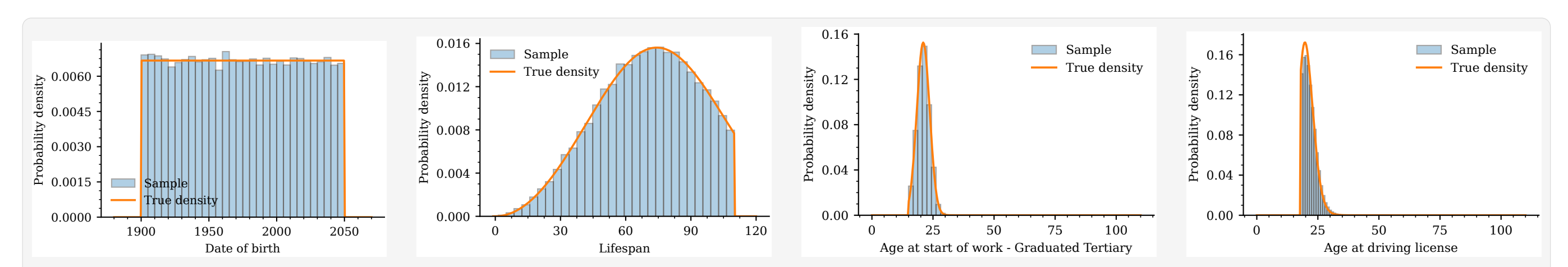


Figure 1: Empirical samples and theoretical densities for key time-independent variables.

Time-dependent variables induced by the prior model

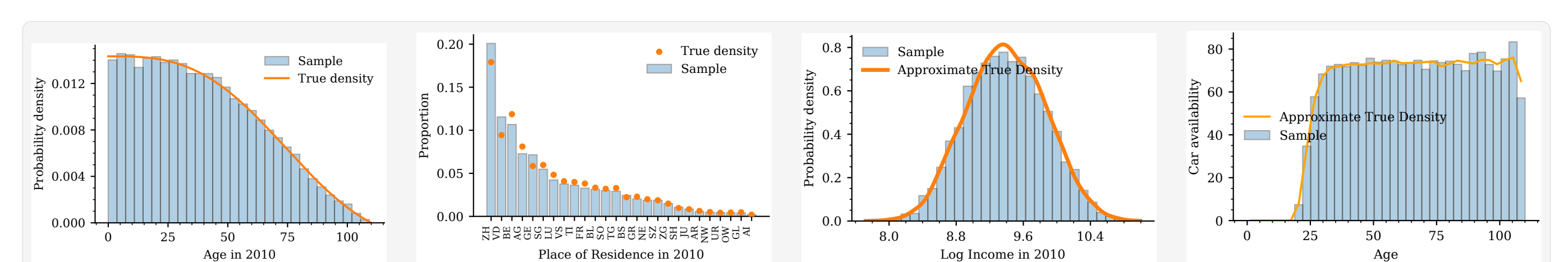


Figure 2: Age, canton of residence, log-income, and car availability distributions in 2010.

Date of birth resampled using data information

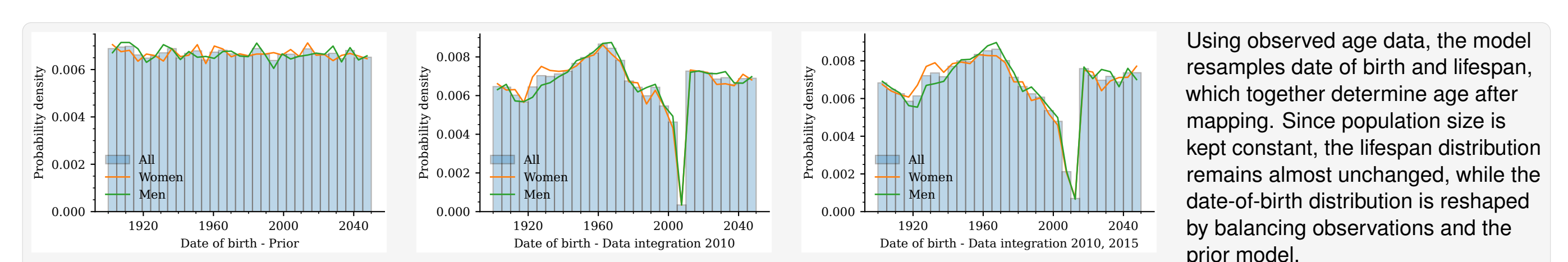


Figure 3: Date of birth distributions: prior and posterior updates, differentiated by gender.

Age resampled using data information

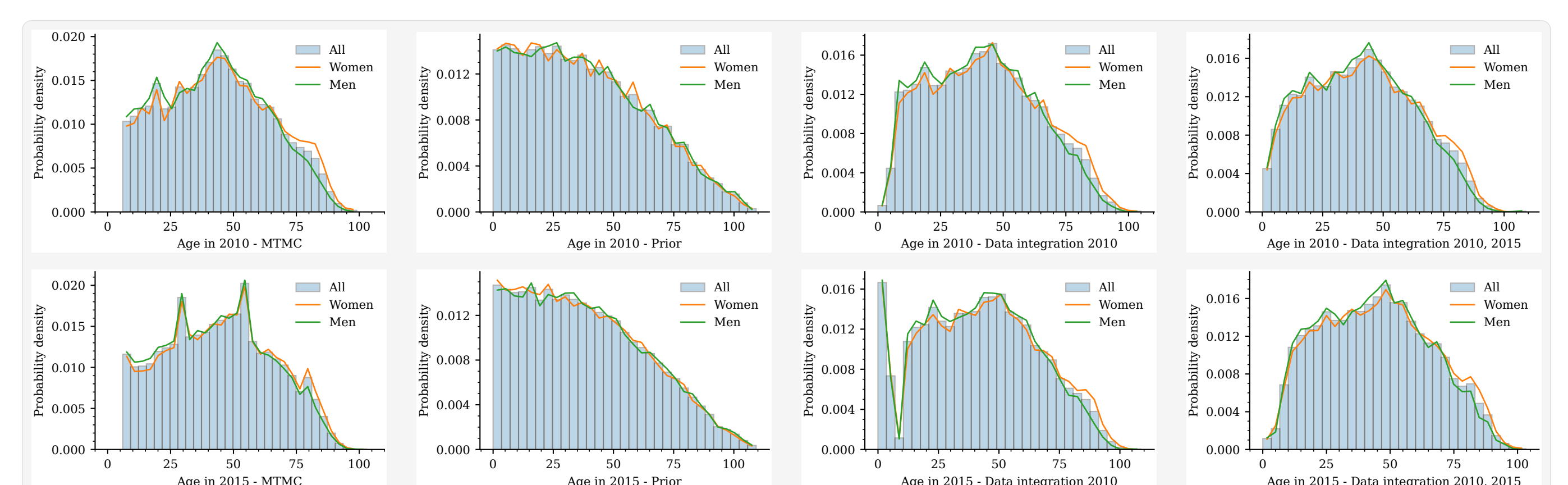
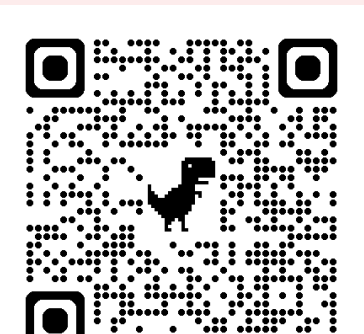


Figure 4: Age distributions. Columns correspond to data, prior, update with 2010 data, and update with 2010+2015 data. Rows correspond to years: 2010 on top and 2015 at the bottom.

This reshaping induces a change in the time-dependent variables. The posterior captures the correlation with gender: although men and women have the same prior age distribution, the model recovers the longer life expectancy of women.

Conclusion. The framework is general, since it can be adapted to different demographic models. It is realistic, since life-course constraints are enforced by construction. It is also flexible: synthetic populations can be generated from the model alone, or updated with observations to balance prior assumptions with observed reality.



Code and results