Gibbs Sampler for Generating Longitudinal Synthetic Populations

Marija Kukic, Michel Bierlaire

Transport and Mobility Laboratory, EPFL, Lausanne, Switzerland
marija.kukic@epfl.ch, michel.bierlaire@epfl.ch

Extended abstract submitted for presentation at the 12th Triennial Symposium on Transportation Analysis conference (TRISTAN XII)
June 22-27, 2025, Okinawa, Japan

February 13, 2025

Keywords: Population synthesis and dynamics, Longitudinal data, Disaster scenario simulation

1 INTRODUCTION

Current state-of-the-art methods for generating synthetic populations typically create data at a single point in time, producing a synthetic snapshot. As demographic changes occur in the real population, synthetic snapshots quickly become outdated, requiring a complete regeneration to update which is both repetitive and computationally expensive. Additionally, generating snapshots independently leads to inconsistencies over time, which limits their usefulness for long-term forecasting. To address this issue, methods for evolving synthetic snapshots have been introduced (Lomax *et al.*, 2022, Prédhumeau & Manley, 2023). However, they often work at an aggregated level, focusing on changes in marginal distributions rather than capturing detailed individual-level dynamics. Also, they simulate only common demographic events such as births, deaths, and migrations, which may result in non-representative synthetic data during long-term forecasting that might involve unexpected events (e.g., COVID-19) (Kukic & Bierlaire, 2024). Limited data on the same individuals over time (i.e., longitudinal data) limit models that rely on individual-level insights (e.g., activity-based models), leading to overemphasizing past behaviors and focus on a single point in time (Zhang *et al.*, 2021).

To address these problems, we propose a novel method that utilizes the Gibbs sampler to generate longitudinal synthetic individuals, enabling us to follow the same synthetic individuals over time. Our method generates a universal set of time-independent synthetic variables $(X_0, X_1, \cdots, X_n)$ only once, from which we can then derive a set of time-dependent synthetic variables $(\hat{Y}_0^t, \hat{Y}_1^t, \cdots, \hat{Y}_n^t)$ at any point in time t. That way our model: (i) ensures internal consistency across time by using a single set of universal variables, avoiding discrepancies seen in independently generated snapshots, (ii) offers more efficient derivation of time-specific data compared to full data regeneration, (iii) provides disaggregated information on the same individuals over time, which offers richer insights compared to having only aggregated sociodemographic marginals, and (iv) enables flexibility, as changes to the universal dataset are reflected in all derived datasets, allowing for rapid testing of scenarios like natural disasters or pandemics. In the case study, we demonstrate the generation of an initial universal synthetic dataset, either using assumed priors or conditionals calibrated using Swiss Mobility and Transport Microcensus (MTMC) data (Swiss Federal Office of Statistics, 2012;2018;2023). Using a universal dataset, we simulate the effects of a pandemic that affects older individuals, ensuring the impact is reflected across all derived datasets with a single simulation.

2 METHODOLOGY

We adopt a model governed by a set of L time-independent universal variables X_ℓ, where $\ell = 0, \cdots, L-1$. This model enables the deterministic reconstruction of individuals' information described by a set of K variables $\hat{Y}_{\mathrm{k}}^{\mathrm{t}}$, where $\mathrm{k} = 0, \ldots, K-1$, at each time instance $\mathrm{t} \in N$, with time discretized into one-year intervals. We model the following universal variables that describe characteristics of individuals: X_0 denoting the year of birth; X_1 denoting the lifespan; X_2 denoting the sex that is assumed invariant over time; X_3 represents the age at which a driving license is acquired, which is assumed to be irrevocable once obtained. We generate them either using assumed priors or by integrating the real data as shown in Sections 2.1, 2.2 and 2.3.

While the method can accommodate a wider range of variables, we demonstrate its applicability using the following set of descriptors: $\hat{Y}_0^{\mathrm{t}}$, a binary variable that indicates if the individual is alive at time t; $\hat{Y}_1^{\mathrm{t}}$ represents the individual's age at time t if they are alive, or their age at death if they have passed away; $\hat{Y}_2^{\mathrm{t}}$, the individual's sex; and $\hat{Y}_3^{\mathrm{t}}$, a binary variable indicating whether the individual holds a driving license. These variables are derived from the previously generated universal variables $X_\ell, \ell = 0, \ldots, 3$ as follows: an individual is alive, $\hat{Y}_0^{\mathrm{t}} = 1$, if $X_0 \leq \mathrm{t} < X_0 + X_1$, otherwise $\hat{Y}_0^{\mathrm{t}} = 0$; the age $\hat{Y}_1^{\mathrm{t}}$ is calculated as $\hat{Y}_1^{\mathrm{t}} = \hat{Y}_0^{\mathrm{t}}(\mathrm{t} - X_0)$; sex $\hat{Y}_2^{\mathrm{t}}$ is given by $\hat{Y}_2^{\mathrm{t}} = X_2$; and driving license ownership $\hat{Y}_3^{\mathrm{t}}$ is set to $\hat{Y}_3^{\mathrm{t}} = 1$ if $\hat{Y}_0^{\mathrm{t}} = 1$ and $\mathrm{t} \geq X_0 + X_3$, otherwise $\hat{Y}_3^{\mathrm{t}} = 0$.

2.1 Prior distributions

For each universal variable X_ℓ, we can assume a prior distribution that enables our method to work, regardless of the availability of real sample data. We adopt that X_0 is uniformly distributed over the time horizon of interest, X_1 follows an exponential or Weibull distribution, as typical in survival analysis, X_2 is a Bernoulli random variable, and $X_3 - 18$ follows a lognormal distribution.

2.2 Data integration: generating year of birth X_0 and lifespan X_1

In this section, we discuss how to refine the priors and generate distributions of X_0 and X_1 using Gibbs sampling and information from the data. Assume that we have access to the distributions of life status Y_0 and age Y_1 at two different time points s and t. From this data, we can also derive conditionals $Y_1^{\mathrm{s}}|Y_0^{\mathrm{s}}$ and $Y_1^{\mathrm{t}}|Y_0^{\mathrm{t}}$. The general idea of Gibbs sampling is to draw from the joint distribution: $X_0, X_1, Y_0^{\mathrm{s}}, Y_1^{\mathrm{s}}, Y_0^{\mathrm{t}}, Y_1^{\mathrm{t}}$. Since the full joint distribution is not available, we decompose these draws into two sets of conditionals from which we draw: (i) $Y_0^{\mathrm{s}}, Y_0^{\mathrm{t}}, Y_1^{\mathrm{s}}, Y_1^{\mathrm{t}}|X_0, X_1$, which is deterministic, and (ii) $X_0, X_1|Y_0^{\mathrm{s}}, Y_1^{\mathrm{s}}, Y_0^{\mathrm{t}}, Y_1^{\mathrm{t}}$, which we sample using Bayes' theorem. The posterior distribution is updated iteratively, proportional to the likelihood times the prior distributions. Assuming that the two datasets are conditionally independent, we obtain:

$$\Pr(X_0, X_1|Y_0^{\mathrm{s}}, Y_1^{\mathrm{s}}, Y_0^{\mathrm{t}}, Y_1^{\mathrm{t}}) \propto \Pr(Y_0^{\mathrm{s}}, Y_1^{\mathrm{s}}|X_0, X_1) \Pr(Y_0^{\mathrm{t}}, Y_1^{\mathrm{t}}|X_0, X_1) \Pr(X_0) \Pr(X_1),$$

where $\Pr(X_0)$ and $\Pr(X_1)$ are the prior distributions provided in Section 2.1. We approximate:

$$\Pr(Y_0^{\mathrm{s}}, Y_1^{\mathrm{s}}|X_0, X_1) = \Pr(Y_1^{\mathrm{s}}|Y_0^{\mathrm{s}}, X_0, X_1) \Pr(Y_0^{\mathrm{s}}|X_0, X_1) \approx \Pr(Y_1^{\mathrm{s}}|Y_0^{\mathrm{s}}) \Pr(Y_0^{\mathrm{s}}|X_0, X_1),$$

where $\Pr(Y_1^{\mathrm{s}}|Y_0^{\mathrm{s}})$ is given by the data, and $\Pr(Y_0^{\mathrm{s}}|X_0, X_1) = 1$ if $X_0 \leq \mathrm{s} < X_0 + X_1$, otherwise is 0. Note that we simplify $\Pr(Y_1^{\mathrm{s}}|Y_0^{\mathrm{s}}, X_0, X_1)$ to $\Pr(Y_1^{\mathrm{s}}|Y_0^{\mathrm{s}})$ since knowing the fact that the individual is alive at year s is sufficient to draw the age, irrespectively of X_0 and X_1. The quantity $\Pr(Y_0^{\mathrm{t}}, Y_1^{\mathrm{t}}|X_0, X_1)$ is defined in a similar way. The draws from X_0 and X_1 are generated using Metropolis-Hastings algorithm.

2.3 Data integration: generating sex X_2 and driving licence age X_3

Assume now that we have access to the sex distribution Y_2 at two different points in time s and t, i.e., we have access to the distributions of $Y_2^{\mathrm{s}}|Y_0^{\mathrm{s}}$ and $Y_2^{\mathrm{t}}|Y_0^{\mathrm{t}}$. To generate X_2, we draw from:

$$X_2, Y_0^{\mathrm{s}}, Y_2^{\mathrm{s}}, Y_0^{\mathrm{t}}, Y_2^{\mathrm{t}},$$

that can be decomposed in the same way as described in Section 2.2. Similarly, by assuming that we have access to the driving license ownership distribution Y_3 in two moments in time s and t, we also have access to the distributions of $(Y_1^s|Y_0^s, Y_3^s)$, $(Y_1^t|Y_0^t, Y_3^t)$, $(Y_3^s|Y_0^s, Y_1^s)$, and $(Y_3^t|Y_0^t, Y_1^t)$, which allows us to generate age of obtaining a driving license X_3. As the driver's license ownership reveals something about the age, we want to draw from: $X_0, X_1, X_3, Y_0^s, Y_1^s, Y_3^s, Y_0^t, Y_1^t, Y_3^t$. To decompose those draws we use the same procedure as before.

3 Results

In this section, we aim to: (i) demonstrate the feasibility of generating a universal dataset with time-independent variables that enable the derivation of consistent time-specific synthetic populations, (ii) demonstrate how unexpected events can be applied to the universal dataset and reflected in all derived datasets, and (iii) test the impact of hypothetical scenarios in both short- and long-term simulations. In Figure 1, we illustrate the steps of the performed case study.

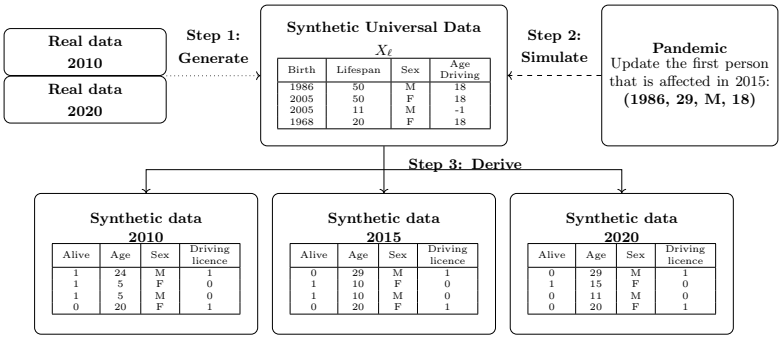

Figure 1 – *The framework for generating synthetic longitudinal data*

First, we generate the universal dataset using the conditionals defined in Sections 2.2 and 2.3. To estimate probabilities, we use MTMC real population data from 2010 and 2020. Note that these datasets cover the same population but do not track the same individuals. Figure 2 shows that refining the assumed priors with real data enables closer alignment with observed distributions. For instance, no individuals born in 1961 are observed to live less than 50 years or more than 100, based on data from 2010 and 2020. Real data define the bounds for lifespan, whereas using priors results in having more variability, with values beyond realistic lifespans.

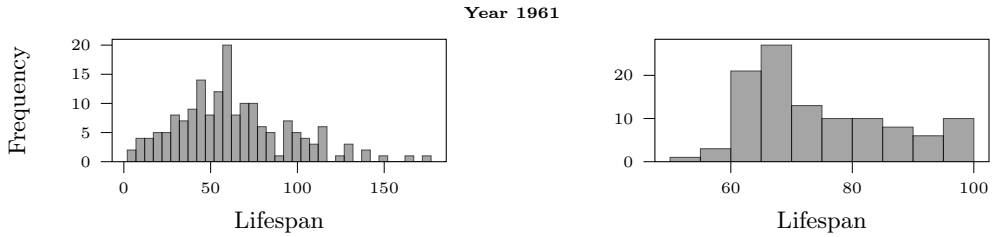

Figure 2 – *Conditional distributions of lifespan given birth year from synthetic universal datasets generated from priors (left) or data (right)*

After generating the universal dataset, we can derive synthetic datasets for any time t as shown in Section 2, enabling tracking of individuals over time. The key advantage is that the universal dataset is generated only once, and any change to it is reflected in all derived datasets. To illustrate this, we simulate a hypothetical pandemic scenario using the universal dataset as a baseline. Figure 3 shows the normal and disaster scenarios. In the normal case, we derive synthetic samples from the universal dataset for 2010 and 2020 that reveal the age distribution shift, with new individuals born between 2010 and 2020 and a small percentage passing away. Then, we simulate the pandemic in 2015 that impacted older people by randomly picking individuals from the universal dataset considered elderly by 2015 and shortening their

lifespan. Using this updated universal dataset, we derive new synthetic data. In the disaster scenario, the sample from 2010 remains the same, while more elderly people died by 2020.

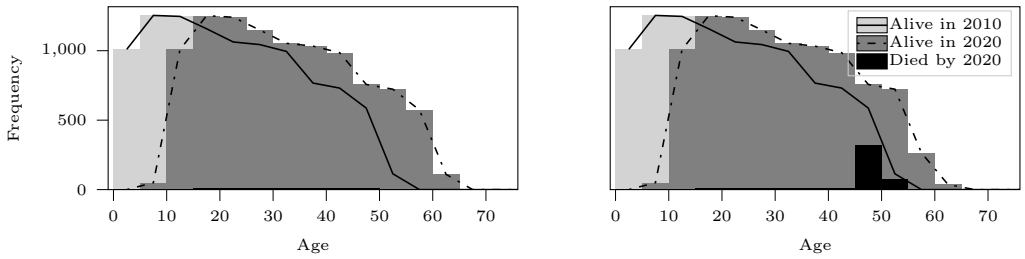

Figure 3 – *Simulation of the normal (left) and hypothetical disaster (right) scenarios*

In Table 1 we show the death percentage at moments $t - s$ and $t + s$, where t is the moment of the disaster and s is the time step. We calculate the death rate for both scenarios as the difference between the death percentage at $t + s$ and $t - s$, divided by the time step s. Since no pandemic has occurred before t, the death percentage at $t - s$ is the same for both scenarios. The disaster scenario shortens the lifespan distribution (X_1) for affected individuals, leading to an increase in deaths. The disaster effect is most pronounced for smaller s (e.g., $s = 5$), where the earlier end of lifespans causes a sharp rise in the death rate. For larger steps, the natural rise in deaths obscures short-term effects, making the disaster harder to detect.

Table 1 – *Comparison of cumulative death percentages and death rates for* $t = 2015$ *for different time steps in normal and disaster scenarios*

Time Step s	**Death % at** $t - s$	**Death % at** $t + s$ **Normal**	**Death % at** $t + s$ **Disaster**	**Death Rate Normal**	**Death Rate Disaster**	**Rate Difference**
5	0.17	1.02	4.86	0.17	0.94	0.77
10	0.12	8.83	11.91	0.87	1.18	0.31
15	0.10	17.50	19.92	1.16	1.32	0.16
20	0.07	26.66	28.63	1.33	1.43	0.10
25	0.07	37.07	38.47	1.48	1.54	0.06
30	0.05	46.66	47.57	1.55	1.58	0.03
35	0.05	56.67	56.67	1.62	1.62	0.00

4 Discussion

This paper introduces a model that generates synthetic universal variables, allowing the derivation of synthetic populations at any time point without recalibration while capturing individual-level changes. We show its capability to provide longitudinal insights and simulate both short- and long-term impacts of hypothetical scenarios, such as pandemics. The model is both efficient and flexible, as it ensures consistency over time and enables rapid scenario testing (e.g., war, hazards, etc.), making it valuable for analyzing trends when real longitudinal data is unavailable. In the future, the model should accommodate a broader range of variables and potentially be expanded from the individual to the household level.

References

Kukic, Marija, & Bierlaire, Michel. 2024. Hybrid Simulator for Projecting Synthetic Households in Unforeseen Events. *In: Conference in Emerging Technologies in Transportation Systems (TRC-30)*. Transport and Mobility Laboratory, EPFL, Crete, Greece.

Lomax, Nik, Smith, Andrew, Archer, Luke, Ford, Alistair, & Virgo, James. 2022. An Open-Source Model for Projecting Small Area Demographic and Land-Use Change. *Geographical Analysis*, **54**(02).

Prédhumeau, Manon, & Manley, Ed. 2023. A synthetic population for agent based modelling in Canada. *Scientific Data*, **10**(03).

Swiss Federal Office of Statistics. 2012;2018;2023. *Comportement de la population en matière de mobilité*. Neuchâtel: Bundesamt für Statistik (BFS).

Zhang, Wenjia, Ji, Chunhan, Yu, Hao, Zhao, Yi, & Chai, Yanwei. 2021. Interpersonal and Intrapersonal Variabilities in Daily Activity-Travel Patterns: A Networked Spatiotemporal Analysis. *ISPRS International Journal of Geo-Information*, **10**(3).