A Dynamic Perspective on Synthetic Populations: Simulation Framework for Longitudinal Synthetic Population Generation

Marija Kukic*[1] and Michel Bierlaire[2]

[1]PhD student, Transport and Mobility Laboratory, EPFL, Switzerland
[2]Professor, Transport and Mobility Laboratory, EPFL, Switzerland

Short summary

This paper introduces a novel framework for generating longitudinal synthetic populations that track individuals over time, addressing limitations of traditional snapshot-based synthetic population methods. We propose a Gibbs sampler-based approach that combines models and cross-sectional data to generate universal, time-independent variables, which enable the consistent derivation of time-specific synthetic populations at any point in time. A key advantage of this framework is that any changes to the universal dataset are automatically reflected in derived datasets, allowing for efficient scenario testing. The methodology is demonstrated using Swiss Mobility and Transport Microcensus data, by simulating the impacts of hypothetical events such as pandemics. This approach ensures temporal consistency, captures individual-level dynamics, and reduces the computational burden of regenerating populations, showcasing its potential for activity-based modeling and long-term policy analysis when real longitudinal data is unavailable.

Keywords: Activity-based models, Longitudinal data, Population dynamics, Pre- and post-disaster simulation, Simulation and modeling, Synthetic Population

1 Introduction

State-of-the-art methods for synthetic population generation typically produce cross-sectional data for a single point in time, creating a synthetic snapshot of socio-demographics and long-term lifestyle and mobility decisions of individuals. As demographic changes occur in the real population, synthetic snapshots quickly become outdated, requiring complete regeneration to update, which is both repetitive and computationally expensive. Moreover, generating snapshots independently leads to inconsistencies over time, limiting their usefulness for long-term forecasting. To address this issue, methods for evolving synthetic snapshots have been introduced (Bhat et al., 2004; 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 dynamics. Also, they simulate 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 focusing on a single point in time (Zhang et al., 2021).

In the literature, several studies have analyzed real longitudinal data to assess how life events impact travel behavior. For instance, Beige & Axhausen (2017) emphasize the interconnected nature of life choices, demonstrating through longitudinal data from Switzerland that decisions related to residence, employment, and commuting modes evolve together over time. Similarly, Ahmed & Moeckel (2023) show that while travel behavior is generally stable, life events such as employment changes or relocations lead to incremental changes rather than abrupt shifts. Both studies emphasize the importance of longitudinal datasets in distinguishing between attributes that remain stable over time (e.g., driving license ownership) and those that change in response to specific life events (e.g., commuting frequency). Their findings underscore the limitations of existing population models, which overlook the incremental and interconnected nature of such changes. To address these limitations, synthetic longitudinal data is needed to enable dynamic modeling that reflects both stability and behavior variability, thereby improving long-term forecasting accuracy.

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 only once, from which we can then derive a set of time-dependent synthetic variables 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

Rather than treating variables such as age, level of education, home location, and driving license as static attributes observed at a single point in time, we model them as dynamic variables that evolve over time. These variables can either be represented as functions of time, such as $\text{age}(t), \text{level of education}(t), \text{home location}(t), \text{driving license}(t)$, or described by associating specific events with their corresponding durations. For example:

- **Age(t)** is defined by the event of birth and the duration of a lifespan.
- **Level of education(t)** corresponds to the date of degree completion and the time until the next degree.
- **Home location(t)** is characterized by the date of relocation and the duration until the next move.
- **Driving license(t)** is associated with the date of acquisition and the time until revocation.

We denote e_i, $i = 1, \ldots, N$ as the list of events and d_i, $i = 1, \ldots, N$ as the corresponding durations. Together, these variables constitute what we refer to as universal variables. While the method can accommodate a wider range of variables, for simplicity, we demonstrate its applicability using $\{(e_i, d_i) \mid i = 1, 2, 3\}$ set of descriptors, where e_1 is the date of birth, d_1 is the lifespan, e_2 is acquisition date of driving license, d_2 is the time until its revocation. For simplicity, we assume that the driving license is irrevocable once obtained (i.e., $e_1 + d_1 = e_2 + d_2$) though this assumption can be relaxed if necessary. Also, we assume time is discretized into one-year intervals, meaning we consider the year of each event instead of exact dates. Although sex is a time-invariant variable, we can model it using the same framework for consistency. Thus, e_3 represents the assigned sex at birth and it is assumed to be invariant over time, i.e., $d_3 = d_1$.

Knowing these variables allows for the deterministic reconstruction of time-dependent variables at any time. Thus, we define time-dependent variables x_{it} and y_{it}, for each i, and any time t, as:

$$x_{it} = \mathbb{1}(e_i \leq t < e_i + d_i)$$

and

$$y_{it} = t - e_i.$$

The indicator function x_{it} equals 1 if the event e_i has occurred and the duration has not elapsed at time t, and 0 otherwise. Similarly, y_{it} represents the elapsed time since the event. For each previously defined $\{(e_i, d_i) \mid i = 1, 2, 3\}$, we define corresponding $\{(x_{it}, y_{it}) \mid i = 1, 2, 3\}$. For example, if we know the date of birth e_1 and the lifespan d_1, x_{1t} is an indicator that equals 1 if the individual is alive at time t and 0 otherwise, and y_{1t} represents the age at time t. Similarly, if we know the acquisition date of a driving license e_2 and the time until its revocation d_2, x_{2t} indicates if the individual has a driving license at time t, and y_{2t} measures how long the person has had the license at time t. Knowing e_3 and d_3, x_{3t} indicates that sex is assigned, and y_{3t} encodes the sex itself.

The objective of our dynamic synthetic population method is to generate universal variables $e_1, d_1, \cdots, e_N, d_N$ to describe the lifetime of each individual. If the data is not available, each universal variable can be generated using assumed priors. We assume that e_1 follows a uniform distribution over the time horizon of interest, as it provides a neutral prior and avoids favoring any specific period when no empirical data is available (Gilbert & Troitzsch, 2005). For d_1, we use the Weibull distribution, as it is a common choice in survival analysis and aligns well with observed lifespan data, providing an accurate representation of the variability in human lifespans (Mahevahaja & Josoa Michel, 2023). e_2 follows a shifted lognormal distribution starting at age 18, as it captures the skewed nature of the age distribution for drivers, where licensure rates increase at 18 but a significant minority delays licensure due to socioeconomic and motivational factors (Tefft et al., 2014). Sex (e_3) is modeled as a Bernoulli random variable to reflect its binary nature and to provide a simple yet realistic representation consistent with available demographic data.

However, access to real cross-sectional data enables the integration of insights from observed distributions and facilitates sampling from posterior distributions rather than relying on assumed priors. Assume that snapshots of cross-sectional data are available at specific time points, providing partial information about the distributions of variables x and y. For instance, if the age distribution of individuals alive at time t is known, it represents the conditional distribution $y_{1t} \mid x_{1t}$. Our proposed method aims to generate $(e_1, d_1, \ldots, e_N, d_N)$ by incorporating the information on x and y extracted from the available data, leveraging the concept of state augmentation.

To formalize the methodology we first define the "life indicators". They play a crucial role in the model because they determine which cross-sectional data are relevant for each individual. While the variables $x_{11}, x_{12}, \ldots, x_{1\mathrm{T}}$ were introduced earlier, they are now grouped into the vector to highlight their specific function in the model as follows:

$$Z = (x_{11}, x_{12}, \ldots, x_{1\mathrm{T}}).$$

Also, we denote X_t the set of relevant variables at time t, without the life indicator:

$$X_t = (y_{1t}, x_{2t}, y_{2t}, \cdots, x_{Nt}, y_{Nt}).$$

We build on static synthetic population methods by assuming that, for each $t = 1, \cdots, T$, we can draw from the random vector

$$(X_t | x_{1t}) = y_{1t}, x_{2t}, y_{2t}, \cdots, x_{Nt}, y_{Nt} | x_{1t},$$

where x_{1t} is the indicator of being alive at time t. This can be done using Gibbs sampling methods.

The objective now becomes to draw from the vector of random variables:

$$e_1, d_1, \cdots, e_N, d_N, Z, X_1, \cdots, X_T.$$

To implement that, we have to draw from each group of variables conditional to others as follows:

1. $$Z, X_1, \cdots, X_T | e_1, d_1, \cdots, e_N, d_N$$

   is deterministic. Indeed, as previously mentioned, if we are given the time of each event, and its duration, we can deterministically reconstruct the x and y variables.

2. $$d_i | e_1, d_1, \cdots, e_N, d_N, Z, X_1, \cdots, X_T$$

   is simplified to

   $$d_i | e_1, d_1, \cdots, e_N, d_N,$$

   as the cross-sectional data do not provide any information about duration. Thus, we can derive the distribution of duration using the priors.

3. To draw from

   $$e_i | e_1, d_1, \cdots, e_N, d_N, Z, X_1, \cdots, X_T$$

   we can use a Bayesian approach. This posterior distribution is proportional to the likelihood times the prior distribution. As Z is given, we know what pieces of data are relevant. Assume

that it is $X_1, \cdots, X_s, s \leq t$, meaning that the individual is alive at time s, but not at time $s+1$. Given that, the likelihood

$$X_1, \cdots, X_T | e_1, d_1, \cdots, e_N, d_N, Z,$$

is approximated by

$$X_1, \cdots, X_s | Z.$$

Assuming conditional independence, we use the static population synthesis for each time t:

$$\Pr(X_1, \ldots, X_s \mid Z) = \bigcup_{t=1}^{s} \Pr(X_t \mid x_{1t}).$$

The prior $e_i | e_1, d_1, \cdots, e_N, d_N$ is assumed to be given.

3 Results and discussion

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 conducted case study. Since variables such as d_2 and d_3 are equivalent to lifespan d_1, they are excluded from the illustration.

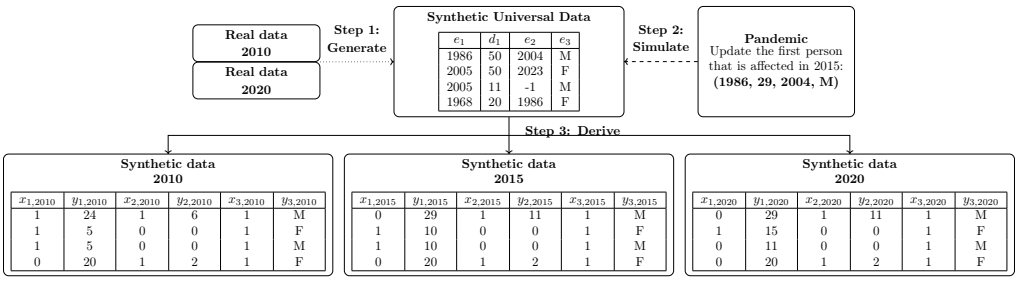

Figure 1 – *The framework for generating synthetic longitudinal data*

First, the synthetic universal dataset is generated using two approaches: priors alone and priors refined with insights from the MTMC data from 2010 and 2020, following the procedure outlined in Section 2. In Figure 2, we compare the resulting lifespan distributions from these approaches. Incorporating real data into the priors produces distributions that more closely align with observed patterns. For example, data from 2010 and 2020 indicate that individuals born in 1961 typically live between 50 and 100 years, with no observations outside this range. Real data define the bounds for lifespan, whereas using priors results in having more variability, with values beyond realistic lifespans.

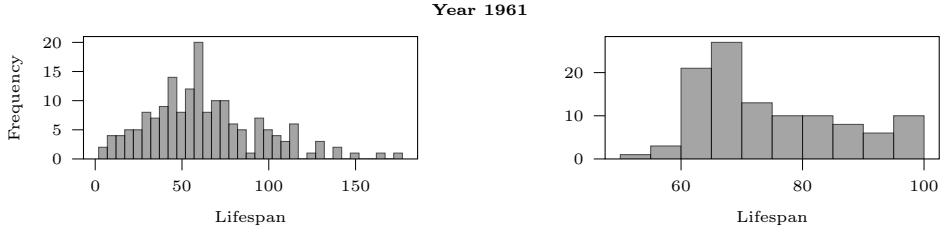

Figure 2 – *Conditional distributions of lifespan given birth year from synthetic universal datasets generated from priors (left) or data (right)*

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.

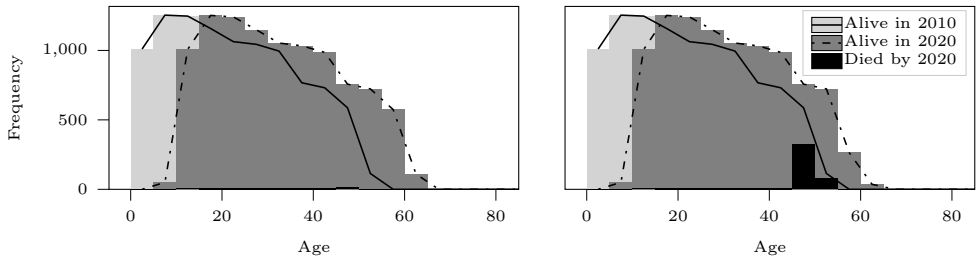

Figure 3 – *Simulation of the normal (left) and hypothetical disaster (right) scenarios*

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 some births in 2020 and a small percentage passing away by the year-end. Then, we simulate the 2015 pandemic on the universal dataset by imposing a 70% mortality rate on individuals aged over 50. We randomly select those considered elderly by 2015 and adjust their lifespans to reflect the event. 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 4 illustrates the setup for testing how the choice of time step s affects the visibility of pandemic effects relative to the year t in which the pandemic occurred. By comparing death rates at $t - s$ and $t + s$ (see Table 1), we analyze the extent to which the disaster's impact can be identified over varying temporal distances.

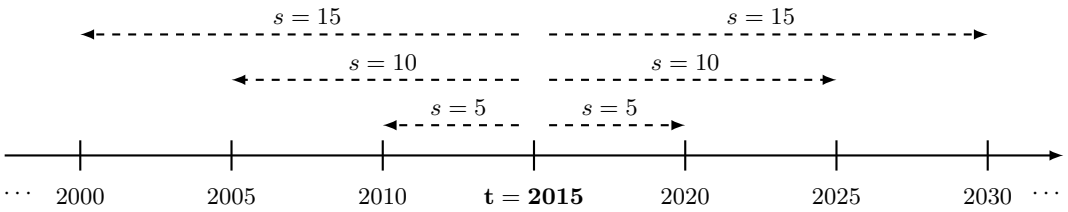

Figure 4 – *The effect of time step s on identifying pandemic impacts around year t*

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 becomes evident through a significant spike in the death rate for smaller time steps (e.g., $s = 5$), with the death rate in the disaster scenario being 5.5 times higher than in the normal scenario. For larger steps (e.g., $s \geq 25$), the natural rise in deaths hides short-term effects, making the disaster harder to detect.

Time Step s	**Death % at** $t - s$	**Death % at** $t + s$ **Normal**	**Death % at** $t + s$ **Disaster**	**Death Rate Normal** (DR_n)	**Death Rate Disaster** (DR_p)	$\frac{\mathbf{DR_p}}{\mathbf{DR_n}}$
5	0.17	1.02	4.86	0.17	0.94	≈ 5.5
10	0.12	8.83	11.91	0.87	1.18	≈ 1.4
15	0.10	17.50	19.92	1.16	1.32	≈ 1.1
20	0.07	26.66	28.63	1.33	1.43	≈ 1.1
25	0.07	37.07	38.47	1.48	1.54	≈ 1

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

4 Conclusions

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 demonstrate how a Bayesian approach can be adapted to integrate models and data, enabling the generation of synthetic longitudinal data that leverages insights from available real cross-sectional data. We also show its ability to simulate 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 (e.g., level of education, home location, income) and potentially be expanded from the individual to the household level.

References

Ahmed, U., & Moeckel, R. (2023, September). Impact of life events on incremental travel behavior change. *Transportation Research Record*, *2677*(9), 594–605. (Publisher Copyright: © National Academy of Sciences: Transportation Research Board 2023.) doi: 10.1177/03611981231159863

Beige, S., & Axhausen, K. W. (2017). The dynamics of commuting over the life course: Swiss experiences. *Transportation Research Part A: Policy and Practice*, *104*, 179-194. doi: https://doi.org/10.1016/j.tra.2017.01.015

Bhat, C. R., Guo, J. Y., Srinivasan, S., Sivakumar, A., Pinjari, A., & Eluru, N. (2004, November). *Activity-based travel-demand modeling for metropolitan areas in texas: Representation and analysis frameworks for population updating and land-use forecasting* (Research Report No. 4080-6). Austin, TX: Center for Transportation Research, The University of Texas at Austin. (Conducted in cooperation with the U.S. Department of Transportation and the Texas Department of Transportation)

Gilbert, N., & Troitzsch, K. G. (2005). *Simulation for the social scientist.* USA: Open University Press.

Kukic, M., & Bierlaire, M. (2024). Hybrid simulator for projecting synthetic households in unforeseen events. In *Conference in emerging technologies in transportation systems (trc-30).* Crete, Greece.

Lomax, N., Smith, A., Archer, L., Ford, A., & Virgo, J. (2022, 02). An open-source model for projecting small area demographic and land-use change. *Geographical Analysis*, *54*. doi: 10.1111/gean.12320

Mahevahaja, J., & Josoa Michel, T. (2023, 06). Computation of human lifespan with a weibull distribution. *International Journal of Science and Research (IJSR)*, *12*, 1927-1932. doi: 10.21275/SR23614150407

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

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

Tefft, B. C., Williams, A. F., & Grabowski, J. G. (2014). Driver licensing and reasons for delaying licensure among young adults ages 18–20, united states, 2012. *Injury Epidemiology*, *1*(4), 1927–1932. doi: 10.1186/2197-1714-1-4

Zhang, W., Ji, C., Yu, H., Zhao, Y., & Chai, Y. (2021). Interpersonal and intrapersonal variabilities in daily activity-travel patterns: A networked spatiotemporal analysis. *ISPRS International Journal of Geo-Information*, *10*(3). doi: 10.3390/ijgi10030148