Gibbs Sampler for Generating Longitudinal Synthetic Populations

Marija Kukic Michel Bierlaire

12-14 February, 2025

EPFL TRANSPORTATION SYMPOSIUM

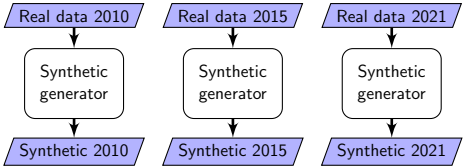
Outline

1. Motivation
2. Literature review
3. Methodology
4. Results
5. Conclusion

Traditional Synthetic Populations - Cross-sectional data

Snapshot of the population at the given point in time.

Static



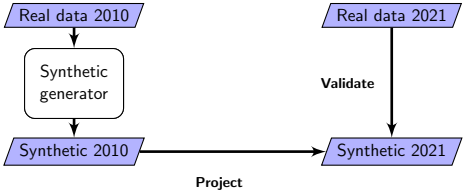
Challenges?

- Data quickly becomes outdated.
- Requires repetitive, costly regeneration.
- Dependent on real data availability.

Traditional Synthetic Populations - Projected data

Simulate changes based on the aggregated rates (e.g., births, deaths, migrations) over time.

Dynamic



- Maintains up-to-date synthetic data without regeneration.
- Trade-off between accuracy and efficiency.
- Does not depend on real data availability.

Outline

1 Motivation

2 Literature review

3 Methodology

4 Results

5 Conclusion

Traditional synthetic populations - Existing methods

Static

- Iterative Proportional Fitting (Beckman et al., 1996)
- Combinatorial Optimization (Abraham et al., 2012)
- Simulation-based (Farooq et al., 2013)
- Machine Learning (Xu and Veeramachaneni, 2018)

Dynamic

- Dynamic projection (Namazi-Rad et al., 2014)
- Static projection (Lomax et al., 2022)
- Resampling (Prédhumeau and Manley, 2023)
- Hybrid approaches (Kukic et al., 2023)

Contribution: Projected VS. Longitudinal data

Projected data

- Work only on the aggregated level.
- Lacks consistency.
- Simulate only common events.
- Lack of longitudinal data affects further development of other models. (Kukic et al., 2024)

Longitudinal data

- Tracks individuals over time.
- Ensures internal consistency.
- Enables scenario testing.
- Enables multi-period analysis for other models.

Idea: Generate **universal** variables once using Markov Chain Monte Carlo (MCMC) simulation and derive consistent **time-specific** synthetic populations.

Outline

1 Motivation

2 Literature review

3 Methodology

4 Results

5 Conclusion

Dynamic variables

Static variables

- Sex
- Age
- Driver licence
- Income
- Employment status
- Level of education
- Home location
- Work location

Dynamic variables

- Sex
- Age(t)
- Driver licence(t)
- Income(t)
- Employment status(t)
- Level of education(t)
- Home location(t)
- Work location(t)

Universal variables

Idea: Dynamic variables can be represented as:

- **Event** (e_i): The occurrence time of an event.
- **Duration** (d_i): The time until the next state change.

Examples:

- **Age:** e_1 (Birth year), d_1 (Lifespan).
- **Driving License:** e_2 (License acquisition year), d_2 (Time until revocation).
- **Education Level:** e_3 (Degree completion year), d_3 (Time until next degree).
- **Home Location:** e_4 (Relocation year), d_4 (Duration in residence).

Event-duration pairs define our **universal variables**.

Deterministic reconstruction of time-dependent variables

Given an **event** (e_i) and its corresponding **duration** (d_i), a time-dependent variable x_{it} is derived as:

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

The elapsed time since the event occurred is:

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

Example:

Age: Knowing birth date and lifespan, x_{1t} determines if a person is alive, and y_{1t} gives their age at time t.

Generating Universal Variables: Priors vs. Data Integration

Priors

Lifespan - Weibull distribution (Mahevahaja and Josoa Michel, 2023).

Getting driving licence - shifted lognormal distribution (Tefft et al., 2014).

Data integration

Repeated cross-sectional census data.

Distribution in 2010, 2015, 2020, etc. provide partial information on x and y.

However, not all of these data is relevant for each individual.

Generating Universal Variables using Data Integration

$$Z = (x_{11}, x_{12}, \ldots, x_{1T})$$
$$X_t = (y_{1t}, x_{2t}, y_{2t}, \ldots, x_{Nt}, y_{Nt})$$

We aim to generate the vector of random variables:

$$(\underbrace{e_1, d_1, \ldots, e_N, d_N}_{\text{Universal variables}}, \underbrace{Z}_{\text{Life indicators}}, \underbrace{X_1, \ldots, X_T}_{\text{Time-dependent variables}})$$

We use the **Gibbs Sampler** to generate samples from this joint distribution by iteratively drawing from its conditional distributions.

Generating Universal Variables using Data Integration

1. **Deterministic Reconstruction**

$$Z, X_1, \ldots, X_T \mid e_1, d_1, \ldots, e_N, d_N$$

Given life events and durations, all time-dependent variables are deterministically derived.

2. **Duration Estimation**

$$d_i \mid e_1, d_1, \ldots, e_N, d_N, Z, X_1, \ldots, X_T \quad \Rightarrow \quad d_i \mid e_1, d_1, \ldots, e_N, d_N$$

Since cross-sectional data lacks duration information, priors define duration distributions.

3. **Sampling Life Events**

$$e_i \mid e_1, d_1, \ldots, e_N, d_N, Z, X_1, \ldots, X_T$$

Bayesian updating estimates life events based on likelihood and priors.

Outline

1 Motivation

2 Literature review

3 Methodology

4 Results

5 Conclusion

Data Generation: Priors VS. Data fusion

Data: MTMC from 2010 and 2020 Swiss Federal Office of Statistics (2023)

By integrating real data, we generate more realistic synthetic data.

No individuals born in 1961 are observed to live less than 50 years or more than 100, based on data from 2010 and 2020.

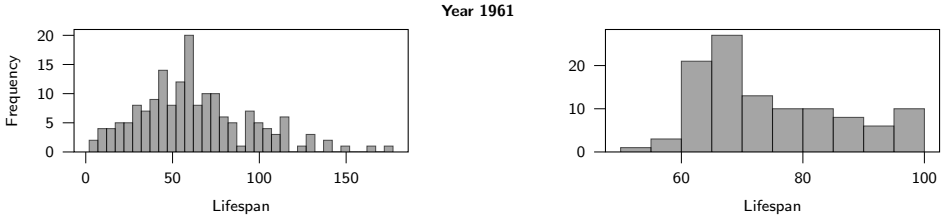
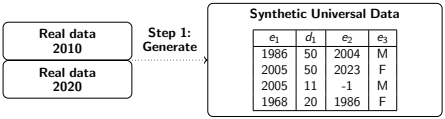


Figure: Conditional distributions of lifespan given birth year from synthetic universal datasets generated from priors (left) or real data (right)

Framework step by step - Hypothetical scenario testing

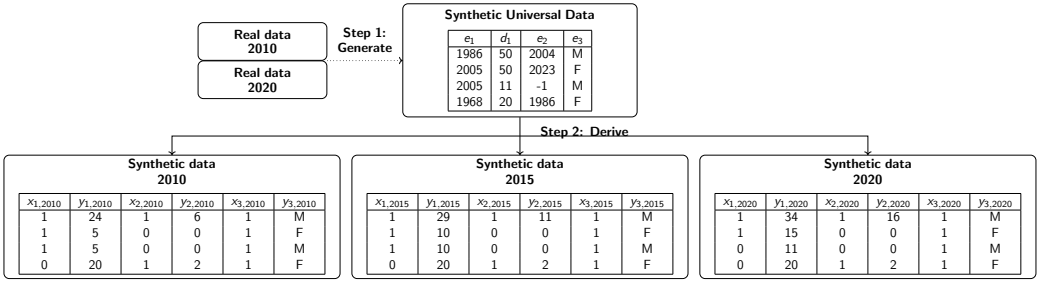
Demonstrate how unexpected events can be applied to the universal dataset and reflected in all derived datasets.

Generated universal variables: Year of birth, Lifespan, License acquisition year, Sex.



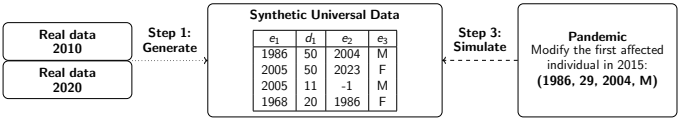
Framework step by step - Hypothetical scenario testing

Time dependent variables: Life indicator, Age, Driving licence status, Sex.



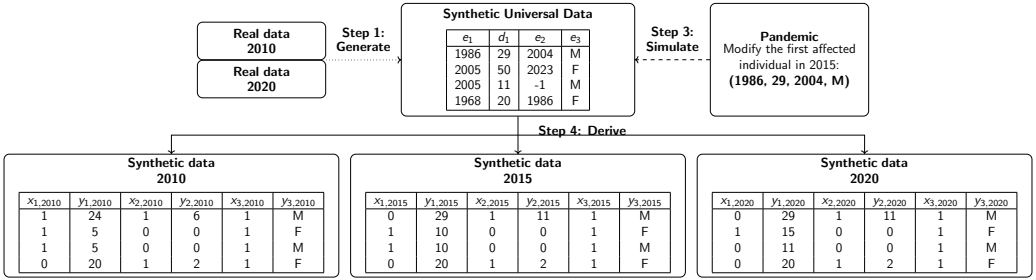
Framework step by step - Hypothetical scenario testing

Simulations: Simulate impacts of hypothetical scenarios on the universal dataset.



Framework step by step - Hypothetical scenario testing

Unexpected events applied to the universal dataset are reflected in all derived datasets.



e_1	d_1	e_2	e_3
1986	29	2004	M
2005	50	2023	F
2005	11	-1	M
1968	20	1986	F

Synthetic data 2010

$x_{1,2010}$	$y_{1,2010}$	$x_{2,2010}$	$y_{2,2010}$	$x_{3,2010}$	$y_{3,2010}$
1	24	1	6	1	M
1	5	0	0	1	F
1	5	0	0	1	M
0	20	1	2	1	F

Synthetic data 2015

$x_{1,2015}$	$y_{1,2015}$	$x_{2,2015}$	$y_{2,2015}$	$x_{3,2015}$	$y_{3,2015}$
0	29	1	11	1	M
1	10	0	0	1	F
1	10	0	0	1	M
0	20	1	2	1	F

Synthetic data 2020

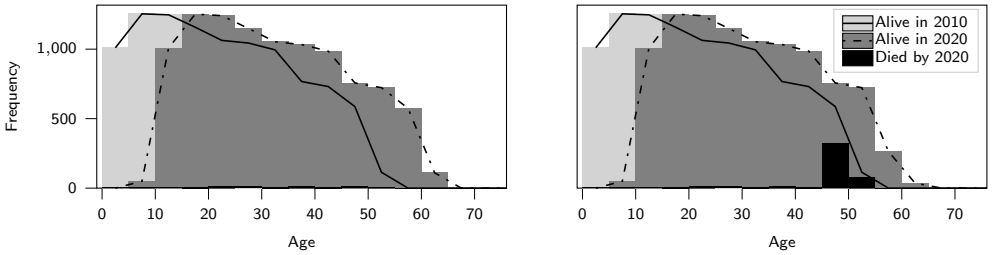
$x_{1,2020}$	$y_{1,2020}$	$x_{2,2020}$	$y_{2,2020}$	$x_{3,2020}$	$y_{3,2020}$
0	29	1	11	1	M
1	15	0	0	1	F
0	11	0	0	1	M
0	20	1	2	1	F

Simulation of hypothetical pandemic

Test the impact of hypothetical scenarios in both short and long-term simulations.

Normal: Derived datasets from 2010 and 2020 without any pandemic.

Pandemic: Simulate on universal dataset 70% mortality for individuals over 50 in 2015, then derive 2010 and 2020.



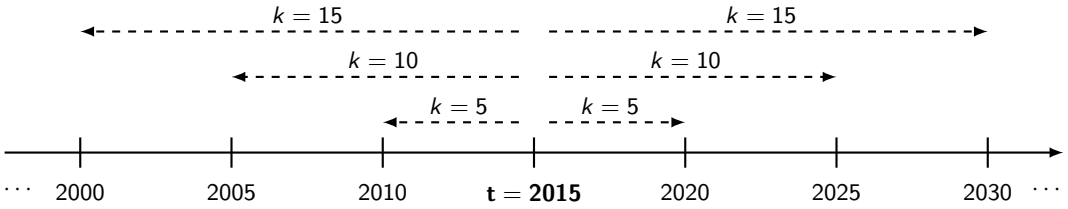
By comparing the two snapshots, we can identify the moment of the pandemic.

Simulation of hypothetical pandemic

How far apart should two datasets be to enable the detection of a pandemic?

Year of pandemic: **t**.

Time step: **k**.



Simulation of hypothetical pandemic

Compare **death rates** (DR) in normal and pandemic scenarios to evaluate the pandemic's impact at $t = 2015$.

$$DR = \frac{\text{Death \% After} - \text{Death \% Before}}{k}$$

DR_n: For **normal** scenario.

DR_p: For **pandemic** scenario.

k	DR_n	DR_p	DR_p/DR_n
5	0.17	0.94	**5.5**
10	0.87	1.18	1.4
15	1.16	1.32	1.1
20	1.33	1.43	1.1
25	1.48	1.54	1.0

Insights:

Pandemic is noticeable for small steps (e.g., $k = 5$, death rate is 5.5 times larger).

Larger steps hide the pandemic (e.g., $k \geq 25$, rates are nearly identical).

Outline

1 Motivation

2 Literature review

3 Methodology

4 Results

5 Conclusion

Conclusion and Future work

Conclusion

- First approach to track individuals over time in synthetic population models.
- MCMC approach allows to combine models and data.
- Enables efficient, consistent, and flexible scenario simulation.

Future work

- Accommodate a broader range of variables (e.g., level of education, home location, income).
- Synthetic populations of households.

Thank you! Questions?

Contact: **marija.kukic@epfl.ch**

References I

Abraham, J. E., Stefan, K. J., and Hunt, J. D. (2012). Population synthesis using combinatorial optimization at multiple levels. In *91st Annu. Meet. Transp. Res. Board*, Washington, DC, USA.

Beckman, R. J., Baggerly, K. A., and McKay, M. D. (1996). Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, 30(6):415–429.

Farooq, B., Bierlaire, M., Hurtubia, R., and Flötteröd, G. (2013). Simulation based population synthesis. *Transportation Research Part B: Methodological*, 58.

Kukic, M., Benchelabi, S., and Bierlaire, M. (2023). Hybrid simulator for capturing dynamics of synthetic populations. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pages 2646–2651.

Kukic, M., Rezvany, N., and Bierlaire, M. (2024). A review of activity-based disaggregate travel demand models. *Findings*.

References II

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

Mahevahaja, J. and Josoa Michel, T. (2023). Computation of human lifespan with a weibull distribution. *International Journal of Science and Research (IJSR)*, 12:1927–1932.

Namazi-Rad, M.-R., Mokhtarian, P., and Perez, P. (2014). Generating a dynamic synthetic population – using an age-structured two-sex model for household dynamics. *PLOS ONE*, 9(4):1–16.

Prédhumeau, M. and Manley, E. (2023). 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*. Bundesamt für Statistik (BFS), Neuchâtel.

References III

Tefft, B. C., Williams, A. F., and 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.

Xu, L. and Veeramachaneni, K. (2018). Synthesizing tabular data using generative adversarial networks. *arXiv:1811.11264 [cs, stat]*.

Backup slide

We use a Bayesian approach. The posterior distribution (6) is proportional to the likelihood times the prior distribution. The likelihood

$$X_1, \ldots, X_T | e_1, d_1, \ldots, e_N, d_N, Z, \tag{7}$$

is approximated by

$$X_1, \ldots, X_s | Z. \tag{8}$$

Assuming conditional independence, we use the static population synthesis (5) for each time t:

$$Pr(X_1, \ldots, X_s | Z) = \prod_{t=1}^{s} Pr(X_t | x_{1t}). \tag{9}$$

The prior

$$e_i | e_1, d_1, \ldots, e_N, d_N \tag{1}$$

is assumed to be given.