Synthetic populations and activity-based models: a dynamic perspective

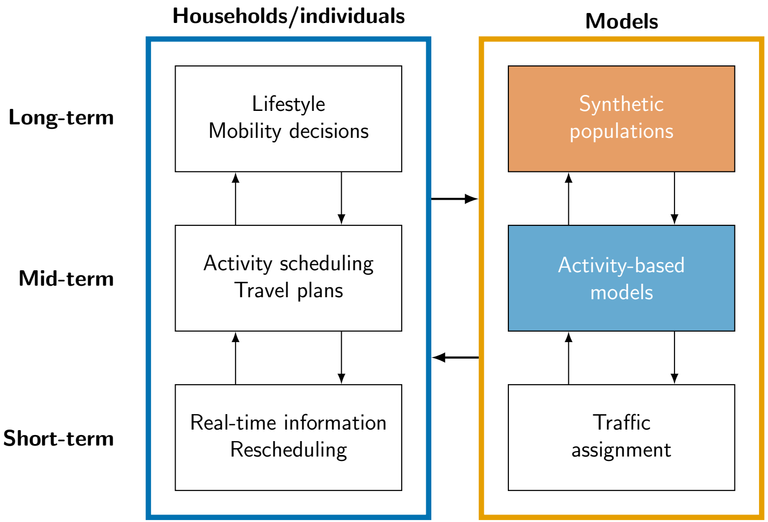
Marija Kukić   Michel Bierlaire

Transport and Mobility Laboratory
School of Architecture, Civil and Environmental Engineering
Ecole Polytechnique Fédérale de Lausanne

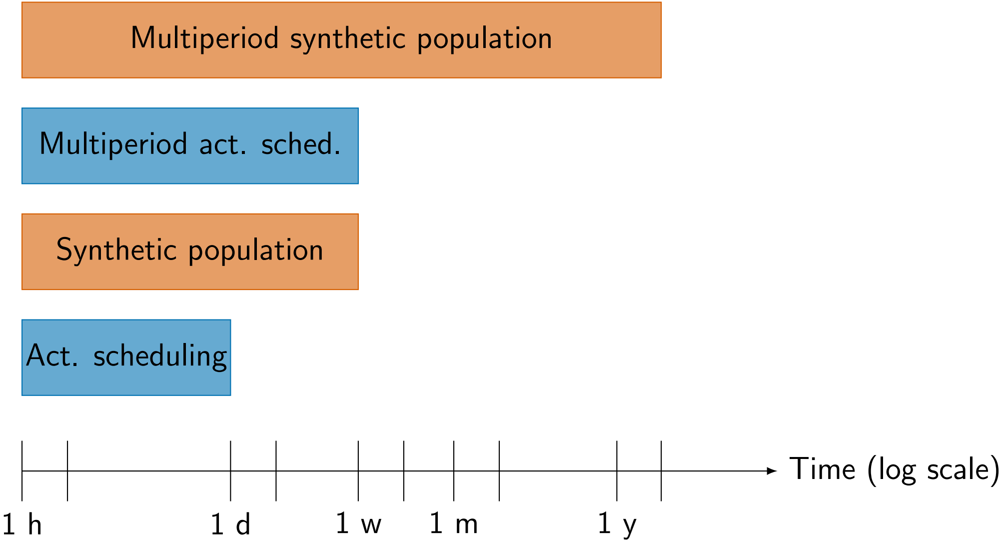
December 12, 2024

Outline

Travel demand modeling



Dynamic models



Outline

Synthetic populations

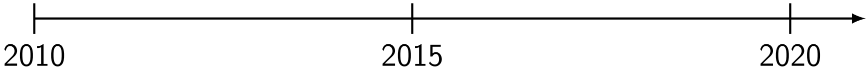
Cross-sectional

- Snapshot of the population at a given point in time.
- Based on an observed real population (census).
- Share the same statistical properties as the real population.
- Includes the status of long-term mobility decisions: home and work location, vehicle ownership, driver's license ownership, etc.
- Feed into activity scheduling models.

Multiperiod synthetic populations

Challenges

- Lack of panel data.
- Instead, repeated cross-sectional census data.
- Consistency (not necessarily the same individuals).



Traditional synthetic populations

Static

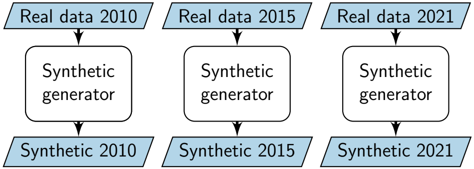
- Sex
- Age
- Income
- Employment status
- Level of education
- Home location
- Work location
- "Mobility tools" ownership
- Driver licence
- etc.

Dynamic

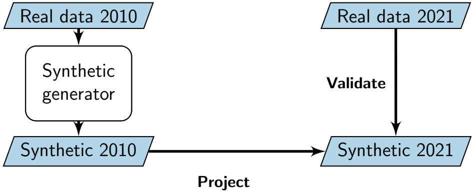
- Sex
- Age(t)
- Income(t)
- Employment status(t)
- Level of education(t)
- Home location(t)
- Work location(t)
- "Mobility tools" ownership(t)
- Driver licence(t)
- etc.

Traditional synthetic populations

Static



Dynamic



Traditional synthetic populations

Static

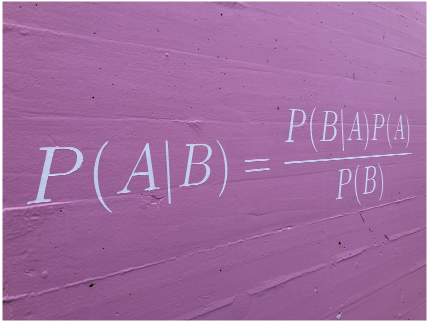
- Iterative Proportional Fitting. [Beckman et al., 1996]
- Combinatorial Optimization. [Abraham et al., 2012]
- Simulation-based. [Farooq et al., 2013]
- Machine Learning. [Xu and Veeramachaneni, 2018]

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]

Outline

Bayesian approach



Bayes theorem

- A: distribution of individuals, B: data.
- We need to draw from $A|B$.
- $\Pr(A|B) =$ likelihood $\cdot$ prior.

Priors: models

- Survival/duration models.
- Behavior models.
- Demographic models, etc.

Data fusion: MCMC

- Gibbs sampling.
- Metropolis-Hastings.

Proposed methodology

Variables

- Replace time dependent variables by time independent variables.
- Events and duration models.
- Examples:
  - age(t). Event: birth. Duration: lifespan.
  - home location(t). Event: last move. Duration: time until the next move.
  - driver's license(t). Event: acquisition of a driver's license. Duration: time until revocation.

Motivation

- Knowing birth date and lifespan, age(t) can be calculated for any t.
- Knowing the date of each move, home location(t) can be calculated for any t.

Mapping universal and time dependent variables

Universal variables

- Birth date b (continuous).
- Lifespan L (continuous).

Time dependent variables

- Being alive in 2010 $x_{2010}(b, L)$ (binary).
- Being alive in 2015 $x_{2015}(b, L)$ (binary).
- Being alive in 2020 $x_{2020}(b, L)$ (binary).
- Age in 2010 $a_{2010}(b, L)$ (continuous).
- Age in 2015 $a_{2015}(b, L)$ (continuous).
- Age in 2020 $a_{2020}(b, L)$ (continuous).

Prior: Event and duration models

Examples

- b: birth date. If $[t_b, t_e]$ is the time horizon of interest,

$$\Pr(b \leq t) = \frac{b - t_b}{t_e - t_b}.$$

- L: lifespan (in years) of an individual. [Gompertz, 1833]: For $\ell \geq 0$,

$$\Pr(L \leq \ell) = 1 - \exp\left(-b\frac{\exp(\eta\ell) - 1}{\eta}\right),$$

  - $b > 0$ is the scale parameter (e.g. $b = 0.0005$),
  - $\eta > 0$ is the shape parameter (e.g. $\eta = 0.1$).
- Age of driver's license: [Tefft et al., 2014]

Available data

- Repeated cross sectional census data.
- Distribution of $a_{2010}|x_{2010} = 1$.
- Distribution of $a_{2015}|x_{2015} = 1$.
- Distribution of $a_{2020}|x_{2020} = 1$.

Gibbs sampling

Objective

Generate draw from the random vector: (b, L)

Marginal distributions

- Draw from $b|L$.
- Draw from $L|b$.

Birth date

For illustration, assume that we have only one data point: 2010

$$\begin{aligned}\Pr(b = \alpha | L) = \Pr(a_{2010} = 2010 - \alpha | x_{2010} = 1, L) \Pr(x_{2010} = 1 | L) \\ + \Pr(a_{2010} = 2010 - \alpha | x_{2010} = 0, L) \Pr(x_{2010} = 0 | L)\end{aligned}$$

- $\Pr(x_{2010} = 1|L)$, $\Pr(x_{2010} = 0|L)$: deterministic:

$$\mathbb{1}[\alpha \leq 2010 < \alpha + L].$$

- $\Pr(a_{2010} = 2010 - \alpha | x_{2010} = 1, L)$: from the data.
- $\Pr(a_{2010} = 2010 - \alpha | x_{2010} = 0, L)$: use the prior. For instance, uniform distribution on

$$b \sim [t_b, 2010 - L[\cup]2010, t_e] \text{ or } a_{2010} \sim]L, 2010 - t_b] \cup [-t_e, 0[$$

Lifespan

$$\Pr(L = \beta | b)$$

- No information in the census data.
- It can be assumed that the lifespan does not depend on the birth date.
- Therefore,
$$\Pr(L = \beta | b) = \Pr(L = \beta).$$
- Prior models can be used.

Example

$$\Pr(b = 1.1.1950 | L = 66)$$

Deterministic life status

$$x_{2010} = 1, x_{2015} = 1, x_{2020} = 0.$$

$$\Pr(b = 1.1.1950 | L = 66) = \Pr(a_{2010} = 60 | x_{2010} = 1) \Pr(a_{2015} = 65 | x_{2015} = 1)$$

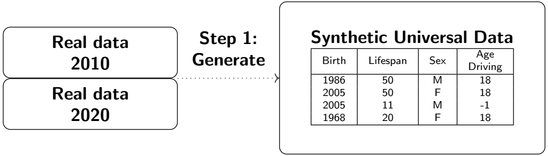
Outline

Illustration

Synthetic universal variables: birth date, lifespan, sex, driver's license acquisition age.

Process: for each variable define the conditionals and draw from them using real data.

Data: MTMC from 2010 and 2020 [Swiss Federal Office of Statistics, 2023]



Birth	Lifespan	Sex	Age Driving
1986	50	M	18
2005	50	F	18
2005	11	M	-1
1968	20	F	18

Illustration

Derived variables: life status, age, sex, and driver's license status.

Process: from universal variables we deterministically reconstruct derived variables.

Real data 2010

Real data 2020

Step 1: Generate

Synthetic Universal Data

Birth	Lifespan	Sex	Age Driving
1986	50	M	18
2005	50	F	18
2005	11	M	-1
1968	20	F	18

Step 2: Derive

Synthetic data 2015

Alive	Age	Sex	Driving licence
1	29	M	1
1	10	F	0
1	10	M	0
0	20	F	1

Synthetic data 2020

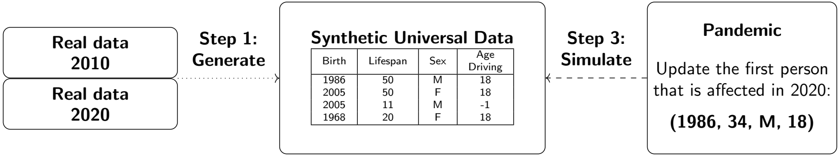
Alive	Age	Sex	Driving licence
1	34	M	1
1	15	F	0
0	11	M	0
0	20	F	1

Synthetic data 2025

Alive	Age	Sex	Driving licence
1	39	M	1
1	20	F	1
0	11	M	0
0	20	F	1

Illustration

Simulate impacts of hypothetical scenarios on the universal dataset.



Birth	Lifespan	Sex	Age Driving
1986	50	M	18
2005	50	F	18
2005	11	M	-1
1968	20	F	18

Illustration

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

Real data 2010

Real data 2020

Step 1: Generate

Synthetic Universal Data

Birth	Lifespan	Sex	Age Driving
1986	50	M	18
2005	50	F	18
2005	11	M	-1
1968	20	F	18

Step 3: Simulate

Pandemic

Update the first person that is affected in 2020:

(1986, 34, M, 18)

Step 4: Derive

Synthetic data 2015

Alive	Age	Sex	Driving licence
1	29	M	1
1	10	F	0
1	10	M	0
0	20	F	1

Synthetic data 2020

Alive	Age	Sex	Driving licence
0	**34**	**M**	**1**
1	15	F	0
0	11	M	0
0	20	F	1

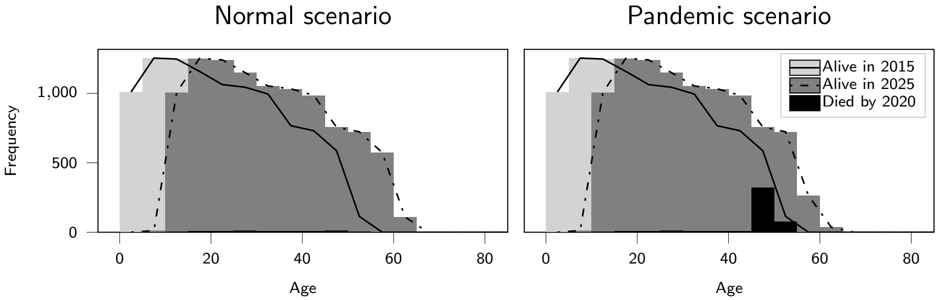
Synthetic data 2025

Alive	Age	Sex	Driving licence
0	**34**	**M**	**1**
1	20	F	1
0	11	M	0
0	20	F	1

Illustration

Normal: Derived datasets from 2015 and 2025 without any pandemic.

Pandemic: Simulate on universal dataset 70% mortality for individuals over 50 in 2020, then derive 2015 and 2025.



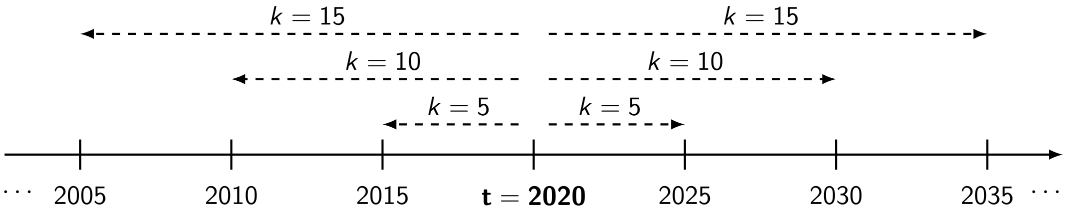
Looking at the two snapshots we can identify the moment of the pandemic.

Illustration

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

Year of pandemic: **t**.

Time step: **k**.



Illustration

Compare death rates (DR) in normal and pandemic scenarios to evaluate the pandemic's impact at $t = 2020$.

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

DR_n: For **normal** scenario.

DR_p: For **pandemic** scenario.

k	$\mathbf{DR_n}$	$\mathbf{DR_p}$	$\mathbf{DR_p/DR_n}$
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

Generalization

Time independent priors

- Age(t): birth date and lifespan.
- Income(t): income evolution models [Kaldasch, 2012].
- Employment status(t): choice of employment status [Kolvereid, 1996].
- Level of education(t): educational choice models [Manzo, 2013].
- Home location(t): last location, moving behavior [de Palma et al., 2015].
- Work location(t): firm relocation [Bodenmann and Axhausen, 2015].
- "Mobility tools" ownership(t): last vehicle, duration model [Gilbert, 1992].
- Driver licence(t): date of acquisition [Nurul Habib, 2018].
- etc.

Bringing it all together

Methodology

- Identification of the time-dependent variables and their event/duration counterparts.
- Identification of the prior models.
- Data fusion using MCMC algorithms.
- Result: synthetic population of individuals with time independent variables.
- Time dependent quantities can be directly derived from the time independent ones.

Conclusion

Current research

- Flexible methodology.
- Bayesian approach allows to combine models and data.
- Cross-sectional data can be integrated.

Future research

- Proof of concept and validation.
- Synthetic populations of households.
- Integration with activity-scheduling models.

Bibliography 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.

Bodenmann, B. R. and Axhausen, K. W. (2015).
Modeling life-cycle of firms and its effect on relocation choice.
In Bierlaire, M., de Palma, A., Hurtubia, R., and Waddell, P., editors, Integrated Transport and Land Use Modeling for Sustainable Cities, pages 201–218, Lausanne, Switzerland. EPFL Press.

Bibliography II

de Palma, A., de Lapparent, M., and Picard, N. (2015).
Modeling real estate investment decisions in households.
In Bierlaire, M., de Palma, A., Hurtubia, R., and Waddell, P., editors, Integrated Transport and Land Use Modeling for Sustainable Cities, pages 137–160, Lausanne, Switzerland. EPFL Press.

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

Gilbert, C. C. S. (1992).
A duration model of automobile ownership.
Transportation Research Part B: Methodological, 26(2):97–114.

Bibliography III

Gompertz, B. (1833).
On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies.
Proc. R. Soc. Lond., 2:252–253.

Kaldasch, J. (2012).
Evolutionary model of the personal income distribution.
Physica A: Statistical Mechanics and its Applications, 391(22):5628–5642.

Kolvereid, L. (1996).
Prediction of employment status choice intentions.
Entrepreneurship Theory and Practice, 21(1):47–58.

Bibliography IV

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.

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.

Bibliography V

- Manzo, G. (2013).
  Educational Choices and Social Interactions: A Formal Model and a Computational Test, volume 30 of Comparative Social Research, pages 47–100.
  Emerald Group Publishing Limited.

- 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.

Bibliography VI

Nurul Habib, K. (2018).
Modelling the choice and timing of acquiring a driver's license: Revelations from a hazard model applied to the university students in toronto.
Transportation Research Part A: Policy and Practice, 118:374–386.

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.

Bibliography VII

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(1):4.

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