Synthetic populations: why and how?

Michel Bierlaire

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

October 2, 2025

EPFL

Outline

Context and Motivation

Travel demand models

- Rapidly evolving mobility patterns.
- Travel needs under resource scarcity.
- Decision-makers face increasing complexity in mobility [Delhoum et al., 2020].

Activity-Based Models (ABMs)

Definition

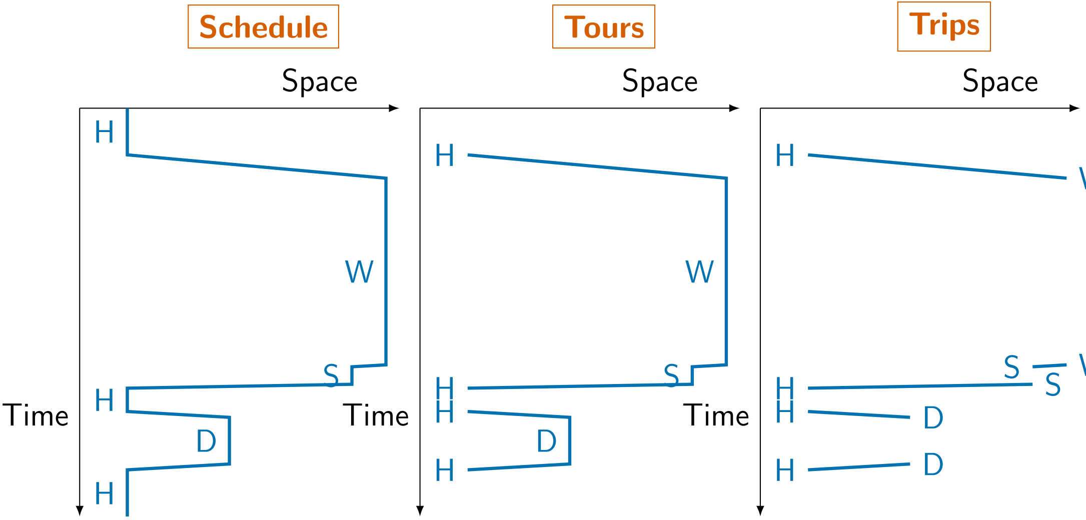
Disaggregate travel demand models that represent each individual/household and simulate their daily sequence of activities and trips, capturing heterogeneity and interactions between activities.

Motivation

- Represent travel demand as the result of **activities in space and time**.
- Contrast with trip-based models: trips are linked within daily schedules, not independent.
- Capture interdependencies between activities, time constraints, and household/social interactions.
- Provide a richer behavioral representation of travel demand.

[Castiglione et al., 2014], [Rezvany et al., 2024]

Travel demand models



H: Home, W: Work, S: Shop, D: Dining out [Source: M. Ben-Akiva]

Why Synthetic Populations?

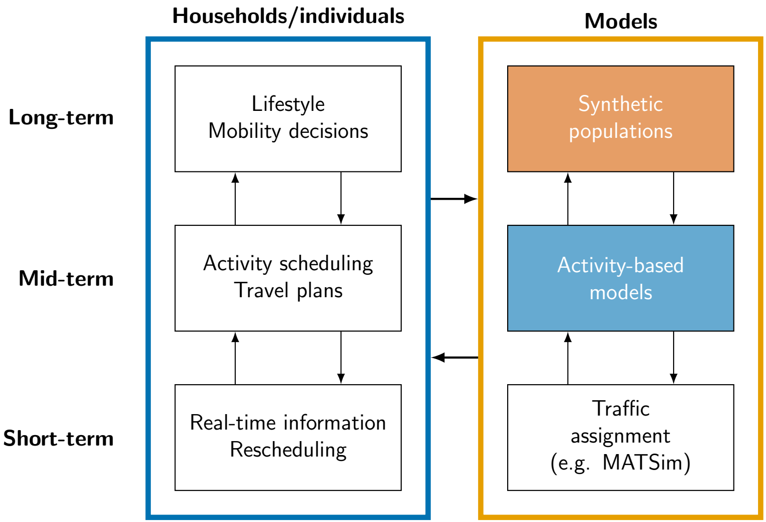
Role in ABMs

- Long-term structural choices (car ownership, residential location, workplace choice, etc.)
- Scarce longitudinal data tracking individuals and households over years.
- Need for individuals and households with consistent socio-demographic profiles and long-term attributes.

Advantages of Synthetic Data

- Realistic travel demand dynamics without causal models.
- Provide diverse and detailed datasets for ABMs.
- Overcome limitations of survey data: representation gaps, anonymization, and bias.
- Merge multiple sources to generate realistic, privacy-compliant, and unbiased synthetic datasets.

Travel demand modeling



Synthetic Populations in Practice: MATSim

Microscopic Simulation Needs

- Tools such as **MATSim** [Axhausen et al., 2016] require a **synthetic population** as input.
- Demand is modeled at the level of **individual synthetic travelers**.
- Each traveler has a daily activity schedule and behavioral rules.

MATSim Users' Guide

"MATSim uses a microscopic description of demand by tracing the daily schedule and the synthetic travelers' decisions."

Outline

Travel demand models

Traditional methodology: IPF

Bayesian approach

Longitudinal synthetic population

Iterative Proportional Fitting (IPF)

Goal

- Adjust seed table to match **target marginals**
- Attributes: e.g. **Age** (rows), **Income** (columns)
- Preserve interaction structure

Algorithm

- Start with seed matrix $X^{(0)}$
- **Row scaling**: enforce row totals r_i
- **Column scaling**: enforce column totals c_j
- Alternate row/column updates until convergence

[Deming and Stephan, 1940]; [Beckman et al., 1996]

IPF Example: Age × Income

Setup

- Rows: $A_1 = 18-39$, $A_2 = 40+$
- Cols: $I_1 = \text{Low}$, $I_2 = \text{High}$
- Seed totals: row $= (100{,}100)$, col $= (100{,}100)$
- Targets: row $= (120{,}80)$, col $= (90{,}110)$

Seed $X^{(0)}$

	Low	High	Sum
18–39	60	40	100
40+	40	60	100
Sum	100	100	200

After Row Scaling

	Low	High	Sum
18–39	72	48	120
40+	32	48	80
Sum	104	96	200

$\Downarrow$ Continue alternating row/col scaling until targets are matched

Iterative Proportional Fitting (IPF)

Properties

- Converges under mild conditions
- Preserves zero cells
- Higher dimensions: iterate through dimensions

Limitations of IPF

Data Limitations

- **Sampling zeros** persist
- Sensitive to **measurement errors** in marginals

Modeling Limitations

- Many sampling zeros in high dimensions
- Only enforces **marginal distributions**
- Cannot capture higher-order interactions directly
- No correction for hidden biases in seed data

Practical Issues

- Output fractional $\rightarrow$ may require integerization
- Large sparse tables $\rightarrow$ convergence can be slow

Outline

Bayesian Approach: Population as a Random Vector

Concept

- Describe population by a high-dimensional **random vector**

$$X = (\text{age}, \text{income}, \text{household size}, \ldots)$$

- Distribution of X:
  - Complex
  - Unknown
- Individuals = instances of X.

Bayesian Approach: Methodology

Principle

- Approximate the unknown distribution of X
- Conditionals from: surveys, registers, fitted models (e.g. multinomial/logit)
- Use **simulation** to draw synthetic individuals / households

Simulation Algorithm

- **Gibbs sampling** (Markov Chain Monte Carlo)
- Iteratively sample each component of X conditional on the others
- Generates correlated samples from the joint distribution

[Farooq et al., 2013], [Kukic et al., 2024]

Gibbs Sampling with Conditionals (Age $\times$ Income)

Algorithm (keywords)

- Initialize $(\text{Age}^{(0)}, \text{Income}^{(0)})$
- For $k = 0, 1, \ldots$
  - Sample $\text{Age}^{(k+1)} \sim P(\text{Age} \mid \text{Income}^{(k)})$
  - Sample $\text{Income}^{(k+1)} \sim P(\text{Income} \mid \text{Age}^{(k+1)})$
- After burn-in: draws $\approx P(\text{Age}, \text{Income})$

Why This Captures Correlation

- Each update uses **informative conditionals** (from data/models)
- Complex patterns maintained: *age-specific income* and *income-specific age*
- Extends to high dimensions: sample each component given the rest

Gibbs sampling = sequential synthesis of individuals (Age × Income)

Gibbs sampler (individual-by-individual)

1. **Initialize** one attribute, e.g. $\text{Income}^{(0)} \sim P(\text{Income})$.
2. **For** $t = 1, \ldots, N$ **(each** t **creates one person):**
   - 2.1 Draw $\text{Age}^{(t)} \sim P(\text{Age} \mid \text{Income}^{(t-1)})$
   - 2.2 Draw $\text{Income}^{(t)} \sim P(\text{Income} \mid \text{Age}^{(t)})$
3. **Record** synthetic individual t: $(\text{Age}^{(t)}, \text{Income}^{(t)})$.

Outcome: a **disaggregate** synthetic population where each row is an individual. After burn-in, the sequence of pairs approximates the joint $P(\text{Age}, \text{Income})$.

Illustration (first few individuals)

t	$\text{Income}^{(t-1)}$	Draw $\text{Age}^{(t)}$	Draw $\text{Income}^{(t)}$	Individual t
1	Low	18–39 (p=0.70)	High (p=0.35)	(18–39, High)
2	High	40+ (p=0.70)	High (p=0.65)	(40+, High)
3	High	18–39 (p=0.30)	Low (p=0.65)	(18–39, Low)
4	Low	…		
⋮				

Bayesian Approach: Advantages

Compared to IPF

- Uses **marginals** but also captures complex **correlation structures**
- Not limited to adjusting contingency tables

Probabilistic Nature

- Naturally incorporates **uncertainty**
- Can model **measurement errors** in data
- Produces distributions, not just point estimates

Outline

Travel demand models

Traditional methodology: IPF

Bayesian approach

Longitudinal synthetic population

Synthetic populations

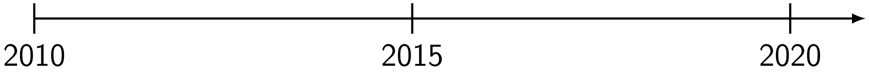
Cross-sectional

- 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

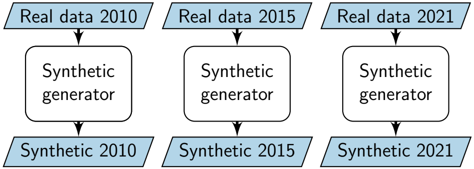
- Sex
- Age
- Income
- Employment status
- Level of education
- Home location
- Work location
- "Mobility tools" ownership
- Driver license
- etc.

Dynamic

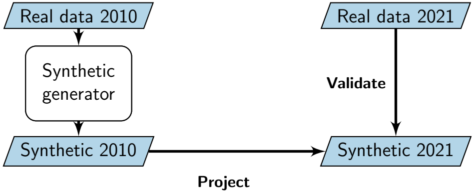
- Sex
- Age(t)
- Income(t)
- Employment status(t)
- Level of education(t)
- Home location(t)
- Work location(t)
- "Mobility tools" ownership(t)
- Driver license(t)
- etc.

Traditional synthetic populations

Static



Dynamic



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

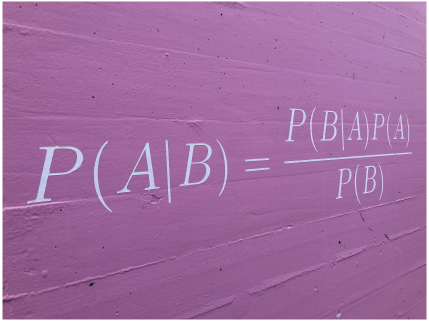
- Being alive in 2010 $x_{2010}(b, L)$ (binary).
- Being alive in 2015 $x_{2015}(b, L)$ (binary).
- Being alive in 2020 $x_{2020}(b, L)$ (binary).
- Age in 2010 $a_{2010}(b, L)$ (continuous).
- Age in 2015 $a_{2015}(b, L)$ (continuous).
- Age in 2020 $a_{2020}(b, L)$ (continuous).

Bayesian approach

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 license(t): date of acquisition [Nurul Habib, 2018].
- etc.

Bayesian approach



Cross-sectional data

- A: distribution of [time independent] individuals.
- B: data.
- We need to draw from $A|B$.
- $\Pr(A|B) =$ likelihood $\cdot$ prior.
- Prior: previous slide.
- Likelihood: mapping time independent variables with time dependent variables.

Data fusion: MCMC

- Gibbs sampling.
- Metropolis-Hastings.

Conclusion

Synthetic populations

- More and more important in travel demand analysis.
- Bayesian approach allows to combine models and data.
- From cross-sectional to longitudinal synthetic data.

Future research

- Synthetic populations of households.
- Integration with activity-scheduling models.

Bibliography I

Axhausen, K., Horni, A., and Nagel, K. (2016).
The multi-agent transport simulation MATSim.
Ubiquity Press.

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

Castiglione, J., Bradley, M., and Gliebe, J. (2014).
Activity-Based Travel Demand Models: A Primer.
Transportation Research Board, Washington, D.C.

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.

Delhoum, Y., Belaroussi, R., Dupin, F., and Zargayouna, M. (2020).
Activity-based demand modeling for a future urban district.
Sustainability, 12(14).

Bibliography III

Deming, W. E. and Stephan, F. F. (1940).
On a least squares adjustment of a sampled frequency table when the expected marginal totals are known.
The Annals of Mathematical Statistics, 11(4):427–444.

Farooq, B., Bierlaire, M., Hurtubia, R., and Flötteröd, G. (2013).
Simulation based population synthesis.
Transportation Research Part B: Methodological, 58:243–263.

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

Bibliography IV

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.

Kukic, M., Li, X., and Bierlaire, M. (2024).
One-step gibbs sampling for the generation of synthetic households.
Transportation Research Part C: Emerging Technologies, 166(104770).

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.

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.

Bibliography VI

Rezvany, N., Kukic, M., and Bierlaire, M. (2024).
A review of activity-based disaggregate travel demand models.
Findings.
Accepted on Nov 01, 2024.