Synthetic Population Generation:

From Hierarchical Snapshots to Panel Data

Marija Kukic
Seminar at Université Gustave Eiffel
Paris

3rd March, 2026

EPFL

Outline

Motivation: Synthetic Data Generation

The quality of the input data determines the quality of the model outputs

Data collections: surveys, census, mobile phone tracking...

Motivation: Synthetic Data Generation

The quality of the input data determines the quality of the model outputs

Data collections: surveys, census, mobile phone tracking...

Limitations

- High cost of data collection
- Lack of representativity
- Data privacy constraints

Motivation: Synthetic Data Generation

The quality of the input data determines the quality of the model outputs

Data collections: surveys, census, mobile phone tracking...

Limitations

- High cost of data collection
- Lack of representativity
- Data privacy constraints

Solution: Synthetic data!

- Open source
- Bias correction
- Privacy preservation

Motivation: Synthetic Data Generation

The quality of the input data determines the quality of the model outputs

Data collections: surveys, census, mobile phone tracking...

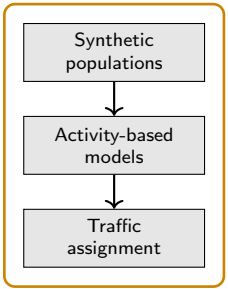
Limitations	**Solution: Synthetic data!**
High cost of data collection	Open source
Lack of representativity	Bias correction
Data privacy constraints	Privacy preservation

Synthetic population = tabular data describing socio-demographics of individuals and households

Motivation: Synthetic Population in Transportation

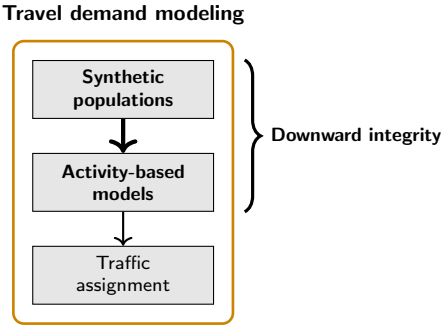
Literature: Synthetic populations (SynPop) provide the **input** needed for **Activity-Based Models** (ABMs) (Castiglione et al., 2014; La et al., 2025).

Travel demand modeling



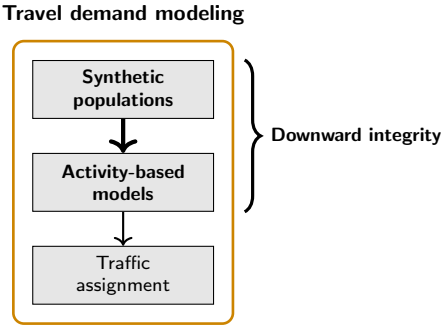
Motivation: Synthetic Population in Transportation

Literature: Synthetic populations (SynPop) provide the **input** needed for **Activity-Based Models** (ABMs) (Castiglione et al., 2014; La et al., 2025).



Motivation: Synthetic Population in Transportation

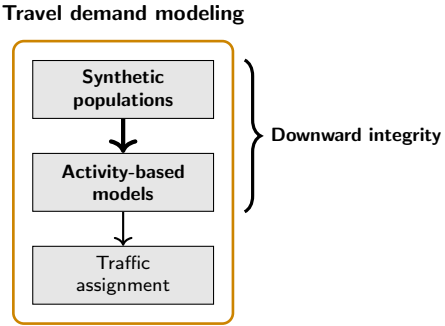
Literature: Synthetic populations (SynPop) provide the **input** needed for **Activity-Based Models** (ABMs) (Castiglione et al., 2014; La et al., 2025).



Practice: Most ABMs still **rely on real data** (Tajaddini et al., 2020) and **limitations** in SynPop methods **constrain** its **integration** into practical applications (Ramadan and Sisiopiku, 2019; La et al., 2025).

Motivation: Synthetic Population in Transportation

Literature: Synthetic populations (SynPop) provide the **input** needed for **Activity-Based Models** (ABMs) (Castiglione et al., 2014; La et al., 2025).



Practice: Most ABMs still **rely on real data** (Tajaddini et al., 2020) and **limitations** in SynPop methods **constrain** its **integration** into practical applications (Ramadan and Sisiopiku, 2019; La et al., 2025).

Theoretical connection → But better integration of SynPop needed?

Motivation: Literature review on SynPop & ABMs

Research questions

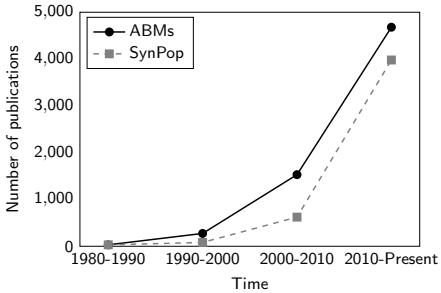
- What is the interconnected development in both fields?
- Can advances in SynPop enhance adoption of ABMs?

Kukic, Marija, Rezvany, Negar, Bierlaire, Michel. *A Review of Activity-Based Disaggregate Travel Demand Models*. **Findings**, December 2024. **https://doi.org/10.32866/001c.125431**.

Motivation: Literature review on SynPop & ABMs

Research questions

- What is the interconnected development in both fields?
- Can advances in SynPop enhance adoption of ABMs?

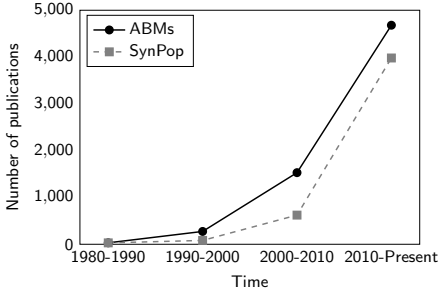


Kukic, Marija, Rezvany, Negar, Bierlaire, Michel. *A Review of Activity-Based Disaggregate Travel Demand Models*. **Findings**, December 2024. **https://doi.org/10.32866/001c.125431**.

Motivation: Literature review on SynPop & ABMs

Research questions

- What is the interconnected development in both fields?
- Can advances in SynPop enhance adoption of ABMs?



Key findings

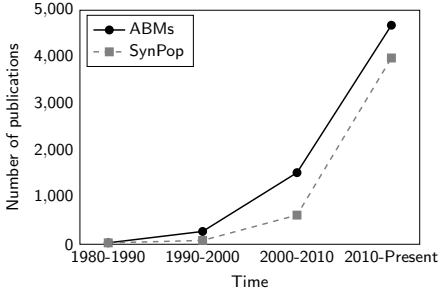
- ABMs and SynPop methods: co-evolved and coupled.
- Most existing work: **individual-level** and **single-period**.

Kukic, Marija, Rezvany, Negar, Bierlaire, Michel. *A Review of Activity-Based Disaggregate Travel Demand Models*. **Findings**, December 2024. **https://doi.org/10.32866/001c.125431**.

Motivation: Literature review on SynPop & ABMs

Research questions

- What is the interconnected development in both fields?
- Can advances in SynPop enhance adoption of ABMs?



Key findings

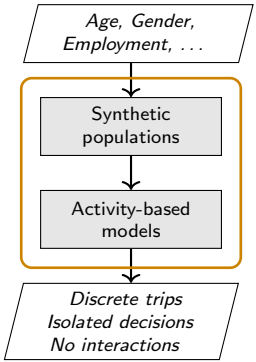
- ABMs and SynPop methods: co-evolved and coupled.
- Most existing work: **individual-level** and **single-period**.

What are the implications of ignoring **multi-person** *and* **multi-period** *modeling contexts in SynPop?*

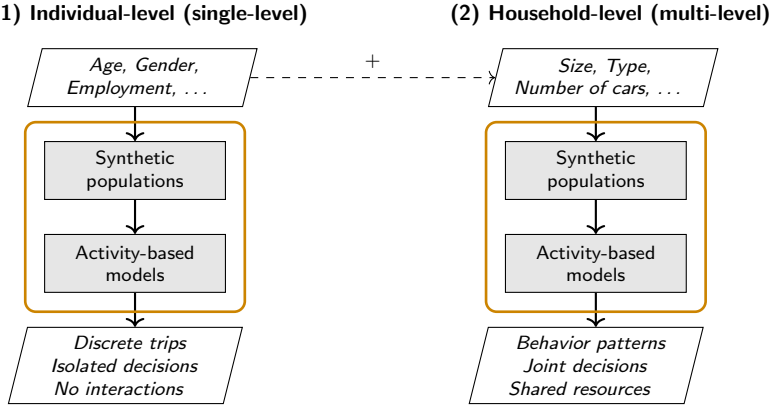
Kukic, Marija, Rezvany, Negar, Bierlaire, Michel. *A Review of Activity-Based Disaggregate Travel Demand Models*. **Findings**, December 2024. **https://doi.org/10.32866/001c.125431**.

Motivation: Multi-person (Hierarchical) modeling

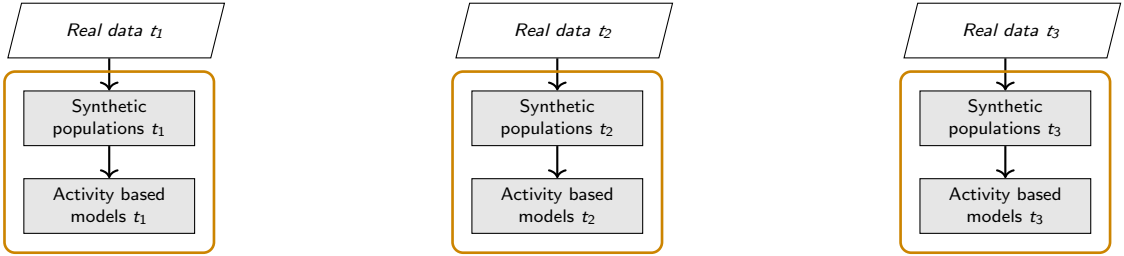
(1) Individual-level (single-level)



Motivation: Multi-person (Hierarchical) modeling

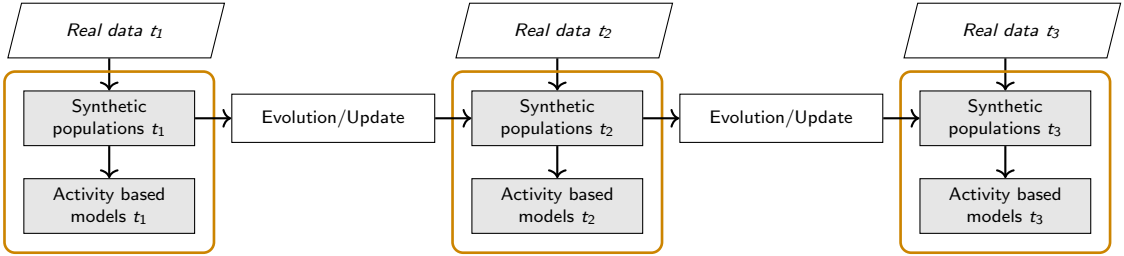


Motivation: Multi-period (Temporal) modeling



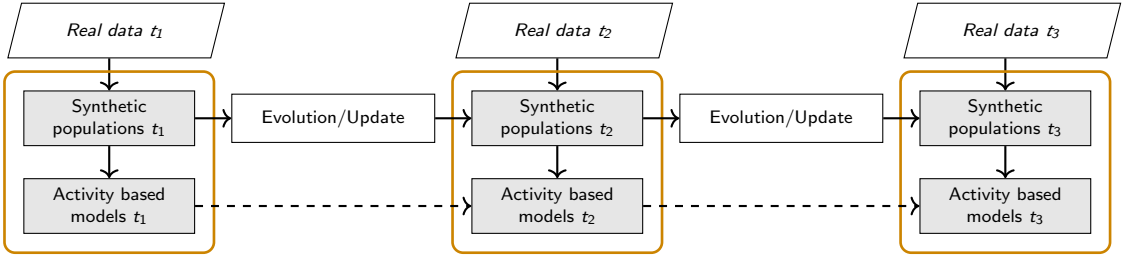
Synthetic static snapshots $\Rightarrow$ Regenerated from scratch!

Motivation: Multi-period (Temporal) modeling



Real populations **evolve gradually** $\Rightarrow$ Can we also evolve a synthetic population?

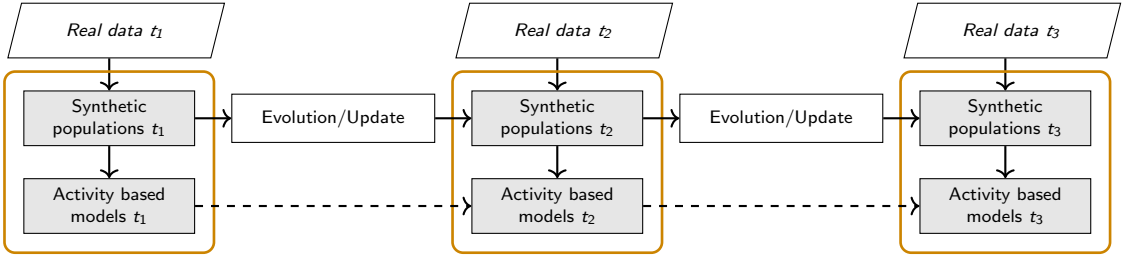
Motivation: Multi-period (Temporal) modeling



Most transport models recreate travel behavior from scratch each time the model runs (Ahmed and Moeckel, 2023).

Next-generation ABMs require **multi-year, evolving population inputs** (Haghighi and Miller, 2025).

Motivation: Multi-period (Temporal) modeling



Can we **generate hierarchical** synthetic populations
that **evolve** at both the **aggregated** and **disaggregated** levels,
providing a foundation for next-generation models that can also adjust incrementally?

Research Overview: Gaps and Contributions

Generation of Synthetic Households

Question

How to generate **diverse**, **consistent**, **extendable** households?

Contribution

One-step Gibbs sampler

Research Overview: Gaps and Contributions

Generation of Synthetic Households

Question

How to generate **diverse**, **consistent**, **extendable** households?

Contribution

One-step Gibbs sampler

Adaptive Synthetic Household Generation

Question

How to move from **regeneration** to **incremental** updates?

Contribution

Adaptive Gibbs-resampling

Research Overview: Gaps and Contributions

Generation of Synthetic Households

Question

How to generate **diverse**, **consistent**, **extendable** households?

Contribution

One-step Gibbs sampler

Adaptive Synthetic Household Generation

Question

How to move from **regeneration** to **incremental** updates?

Contribution

Adaptive Gibbs-resampling

Generating Synthetic Panel Data

Question

How to generate **synthetic panels** without real panels?

Contribution

Universal event–duration model

Research Overview: Gaps and Contributions

Generation of Synthetic Households	**Adaptive Synthetic Household Generation**	**Generating Synthetic Panel Data**
Kukic, Marija, Li, Xinling, Bierlaire, Michel. *One-step Gibbs Sampling for the Generation of Synthetic Households.* **Transportation Research Part C: Emerging Technologies**, 166:104770, September 2024. ISSN 0968-090X. **https://doi.org/10.1016/j.trc.2024.104770**.	Kukic, Marija, Bierlaire, Michel. *Adaptive Synthetic Generation using One-Step Gibbs Sampler.* **Transportation Research Interdisciplinary Perspectives**, 33:101597, September 2025. ISSN 2590-1982. **https://doi.org/10.1016/j.trip.2025.101597.**	Kukic, Marija, Bierlaire, Michel. *Simulation framework for generating synthetic panel data.* **Technical Report TRANSP-OR 251013**, Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 2025.

From Hierarchical Snapshots to Panel Data!

Research Overview: From Hierarchical Snapshots to Panel Data

Generating Snapshot of Individuals

t = 2021

$[a_1, g_1, e_1, \dots]$

$[a_2, g_2, e_2, \dots]$

$[a_3, g_3, e_3, \dots]$

$[a_4, g_4, e_4, \dots]$

Research Overview: From Hierarchical Snapshots to Panel Data

Generating Snapshot of Individuals

t = 2021

$[a_1, g_1, e_1, \dots]$

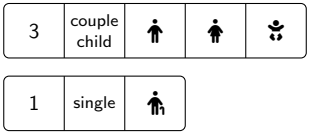
$[a_2, g_2, e_2, \dots]$

$[a_3, g_3, e_3, \dots]$

$[a_4, g_4, e_4, \dots]$

Generating Hierarchical Snapshot

t = 2021



Research Overview: From Hierarchical Snapshots to Panel Data

Generating Snapshot of Individuals

t = 2021

$[a_1, g_1, e_1, \dots]$

$[a_2, g_2, e_2, \dots]$

$[a_3, g_3, e_3, \dots]$

$[a_4, g_4, e_4, \dots]$

Generating Hierarchical Snapshot

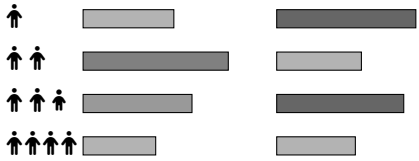
t = 2021

3	couple child			

1	single	

Evolution on Aggregated Level

t = 2021 t = 2025



Research Overview: From Hierarchical Snapshots to Panel Data

Generating Snapshot of Individuals

t = 2021

$[a_1, g_1, e_1, \dots]$

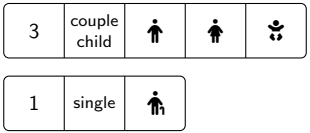
$[a_2, g_2, e_2, \dots]$

$[a_3, g_3, e_3, \dots]$

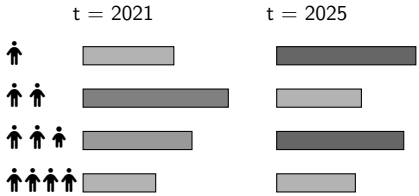
$[a_4, g_4, e_4, \dots]$

Generating Hierarchical Snapshot

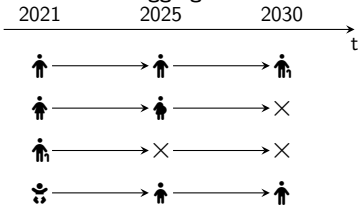
t = 2021



Evolution on Aggregated Level

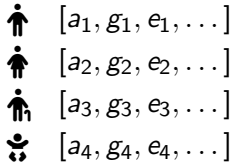


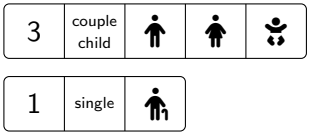
Evolution on Disaggregated Level



Outline

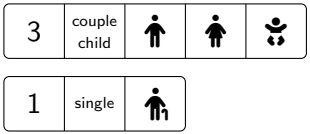
Motivation


$[a_1, g_1, e_1, \ldots]$
$[a_2, g_2, e_2, \ldots]$
$[a_3, g_3, e_3, \ldots]$
$[a_4, g_4, e_4, \ldots]$


3
couple
child
1
single

Motivation

$[a_1, g_1, e_1, \dots]$
$[a_2, g_2, e_2, \dots]$
$[a_3, g_3, e_3, \dots]$
$[a_4, g_4, e_4, \dots]$

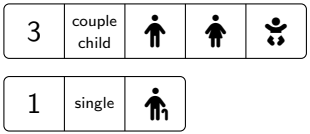


Literature (Yaméogo et al., 2021):

- Match marginals at both levels.
- Links within the same household.
- Relationships between households and individuals **most suitably?**

Motivation

$[a_1, g_1, e_1, \dots]$
$[a_2, g_2, e_2, \dots]$
$[a_3, g_3, e_3, \dots]$
$[a_4, g_4, e_4, \dots]$



Literature (Yaméogo et al., 2021):

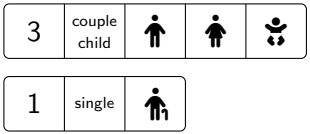
- Match marginals at both levels.
- Links within the same household.
- Relationships between households and individuals **most suitably?**

Simulation-based methods:

- No constraints on input data.
- No "zero-cell" problem.
- Work with small samples.
- Stochastic → heterogeneous population.

Motivation

$[a_1, g_1, e_1, \dots]$
$[a_2, g_2, e_2, \dots]$
$[a_3, g_3, e_3, \dots]$
$[a_4, g_4, e_4, \dots]$



Literature (Yaméogo et al., 2021):

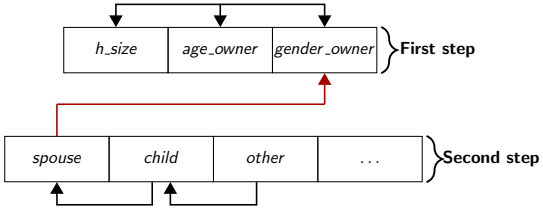
- Match marginals at both levels.
- Links within the same household.
- Relationships between households and individuals **most suitably?**

Simulation-based methods:

- No constraints on input data.
- No "zero-cell" problem.
- Work with small samples.
- Stochastic → heterogeneous population.

Despite beneficial properties → Rarely used?

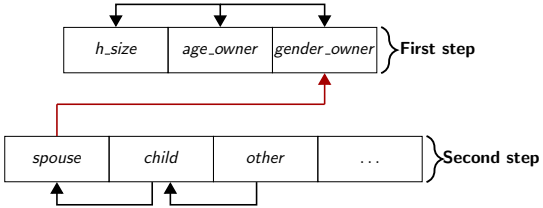
Existing methods: Gaps



Existing "two-step" method (Casati et al., 2015):

- **Role assumptions** on household structure
  $\Rightarrow$ Limited diversity, illogical households.

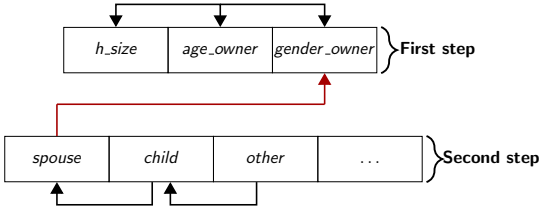
Existing methods: Gaps



Existing "two-step" method (Casati et al., 2015):

- **Role assumptions** on household structure
  $\Rightarrow$ Limited diversity, illogical households.
- Works with a **limited set** of attributes
  $\Rightarrow$ Curse of dimensionality when adding attributes.

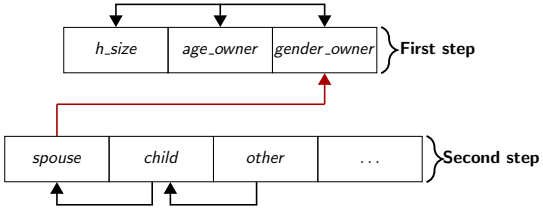
Existing methods: Gaps



Existing "two-step" method (Casati et al., 2015):

- **Role assumptions** on household structure
  ⇒ Limited diversity, illogical households.
- Works with a **limited set** of attributes
  ⇒ Curse of dimensionality when adding attributes.
- **Limited flexibility**
  ⇒ Unclear how to introduce new attributes.

Existing methods: Gaps



Existing "two-step" method (Casati et al., 2015):

- **Role assumptions** on household structure
  ⇒ Limited diversity, illogical households.
- Works with a **limited set** of attributes
  ⇒ Curse of dimensionality when adding attributes.
- **Limited flexibility**
  ⇒ Unclear how to introduce new attributes.

Is there an alternative way that addresses these limitations?

Methodology: One-step method

Population as a multidimensional random vector capturing all relevant attributes

N=1	Z_t	Z_c	$[X_{1a}, X_{1g}, X_{1m}, X_{1e}, X_{1l}]$	
N=2	Z_t	Z_c	$[X_{1a}, X_{1g}, X_{1m}, X_{1e}, X_{1l}]$	$[X_{2a}, X_{2g}, X_{2m}, X_{2e}, X_{2l}]$
⋮				
N=n	Z_k	$\cdots$	$[X_{1a}, X_{1g}, X_{1m}, X_{1e}, X_{1l}]$	$\cdots$ $[X_{na}, X_{ng}, X_{nm}, X_{ne}, X_{nl}]$

One-step method:

- Joint model for the household (sorted by age)
  $\Rightarrow$ flexible for extension.

Methodology: One-step method

Population as a multidimensional random vector capturing all relevant attributes

N=1	Z_t	Z_c	$[X_{1a}, X_{1g}, X_{1m}, X_{1e}, X_{1l}]$	
N=2	Z_t	Z_c	$[X_{1a}, X_{1g}, X_{1m}, X_{1e}, X_{1l}]$	$[X_{2a}, X_{2g}, X_{2m}, X_{2e}, X_{2l}]$
⋮				
N=n	Z_k	$\cdots$	$[X_{1a}, X_{1g}, X_{1m}, X_{1e}, X_{1l}]$	$\cdots$ $[X_{na}, X_{ng}, X_{nm}, X_{ne}, X_{nl}]$

One-step method:

- Joint model for the household (sorted by age)
  $\Rightarrow$ flexible for extension.
- Household type explicitly modeled
  $\Rightarrow$ diverse and realistic household structures.

Methodology: One-step method

Population as a multidimensional random vector capturing all relevant attributes

N=1	Z_t	Z_c	$[X_{1a}, X_{1g}, X_{1m}, X_{1e}, X_{1l}]$	
N=2	Z_t	Z_c	$[X_{1a}, X_{1g}, X_{1m}, X_{1e}, X_{1l}]$	$[X_{2a}, X_{2g}, X_{2m}, X_{2e}, X_{2l}]$
⋮				
N=n	Z_k	$\cdots$	$[X_{1a}, X_{1g}, X_{1m}, X_{1e}, X_{1l}]$	$\cdots$ $[X_{na}, X_{ng}, X_{nm}, X_{ne}, X_{nl}]$

One-step method:

- Joint model for the household (sorted by age)
  $\Rightarrow$ flexible for extension.
- Household type explicitly modeled
  $\Rightarrow$ diverse and realistic household structures.
- Extended to more attributes
  Too many attributes $\rightarrow$ curse of dimensionality
  Too few attributes $\rightarrow$ missing relationships
  Inconsistent choices $\rightarrow$ illogical observations

Methodology: One-step method

Population as a multidimensional random vector capturing all relevant attributes

N=1	Z_t	Z_c	$[X_{1a}, X_{1g}, X_{1m}, X_{1e}, X_{1l}]$	
N=2	Z_t	Z_c	$[X_{1a}, X_{1g}, X_{1m}, X_{1e}, X_{1l}]$	$[X_{2a}, X_{2g}, X_{2m}, X_{2e}, X_{2l}]$
⋮				
N=n	Z_k	$\cdots$	$[X_{1a}, X_{1g}, X_{1m}, X_{1e}, X_{1l}]$	$\cdots$ $[X_{na}, X_{ng}, X_{nm}, X_{ne}, X_{nl}]$

One-step method:

- Joint model for the household (sorted by age)
  $\Rightarrow$ flexible for extension.
- Household type explicitly modeled
  $\Rightarrow$ diverse and realistic household structures.
- Extended to more attributes
  Too many attributes $\rightarrow$ curse of dimensionality
  Too few attributes $\rightarrow$ missing relationships
  Inconsistent choices $\rightarrow$ illogical observations

How to handle these problems?
Simplification of full conditionals + Decomposition on household size!

Methodology: Modeling simplifications & Decomposition

$\pi(A, B, C, D)$

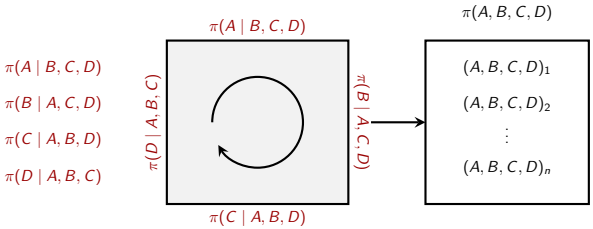
$(A, B, C, D)_1$

$(A, B, C, D)_2$

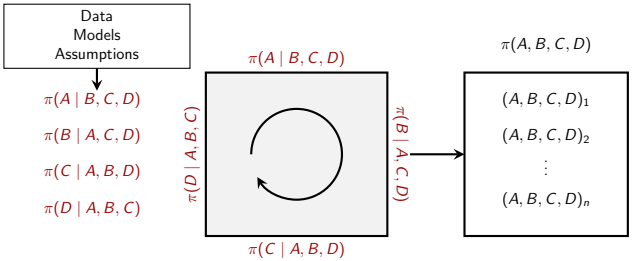
$\vdots$

$(A, B, C, D)_n$

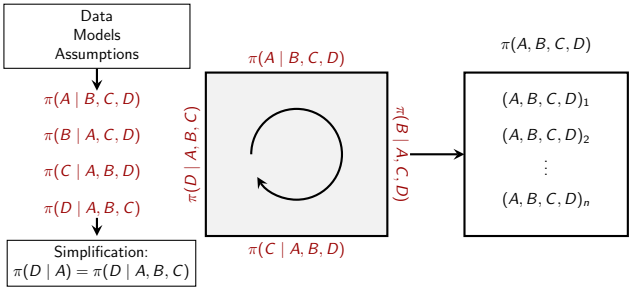
Methodology: Modeling simplifications & Decomposition



Methodology: Modeling simplifications & Decomposition



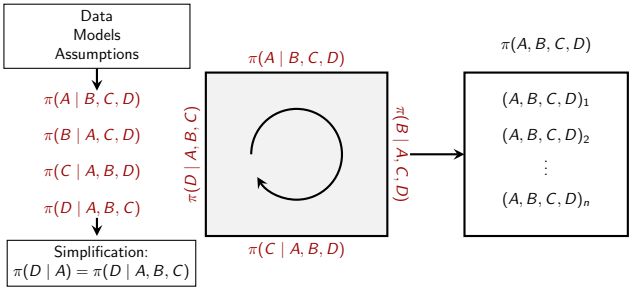
Methodology: Modeling simplifications & Decomposition



Modeling simplifications:

- Preserve relationships between households and members.
- Preserve relationships between individuals from the same household.
- Avoid illogical observations.

Methodology: Modeling simplifications & Decomposition



Modeling simplifications:

- Preserve relationships between households and members.
- Preserve relationships between individuals from the same household.
- Avoid illogical observations.

Decomposition:

N = 1

1	single	

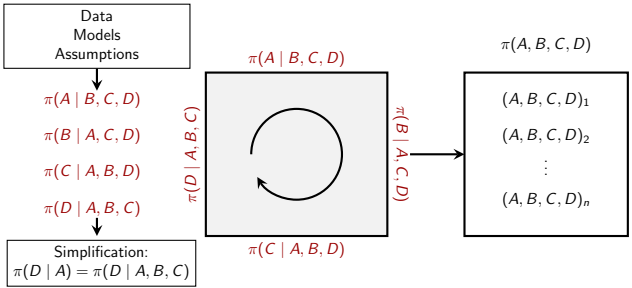
N = 2

2	single child		

N = 3

3	couple child			

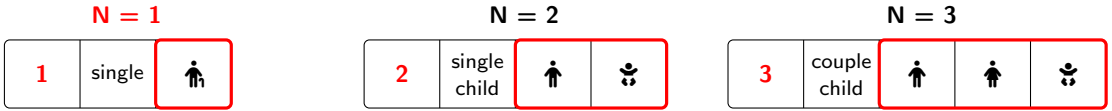
Methodology: Modeling simplifications & Decomposition



Modeling simplifications:

- Preserve relationships between households and members.
- Preserve relationships between individuals from the same household.
- Avoid illogical observations.

Decomposition:



Results: Comparison with two-step method

Swiss Mobility and Transport microcensus (MTMC) data from 2015.
Marginal fit is not enough.
Consistency between the individual and household levels and realism of synthetic data.

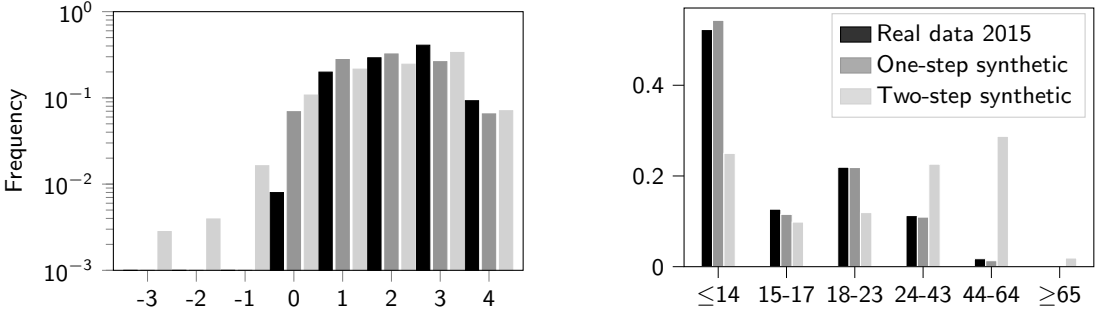


Figure: Distribution of age category gaps between a spouse and a child (left) and marginal distribution of age categories of children (right)

Summary

Idea

Adaptation of the Gibbs sampler for generating synthetic households by grouping all variables in the vector sorted in decreasing order of age.

Summary

Idea

Adaptation of the Gibbs sampler for generating synthetic households by grouping all variables in the vector sorted in decreasing order of age.

Contribution

- Add household type for realism.
- Sampler per household size for efficiency.
- Conditional distributions for control.

Summary

Idea

Adaptation of the Gibbs sampler for generating synthetic households by grouping all variables in the vector sorted in decreasing order of age.

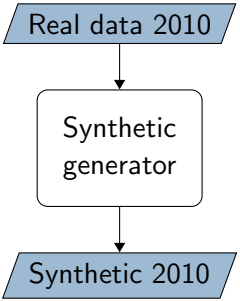
Contribution

- Add household type for realism.
- Sampler per household size for efficiency.
- Conditional distributions for control.

Can we use a one-step Gibbs sampler for updates?

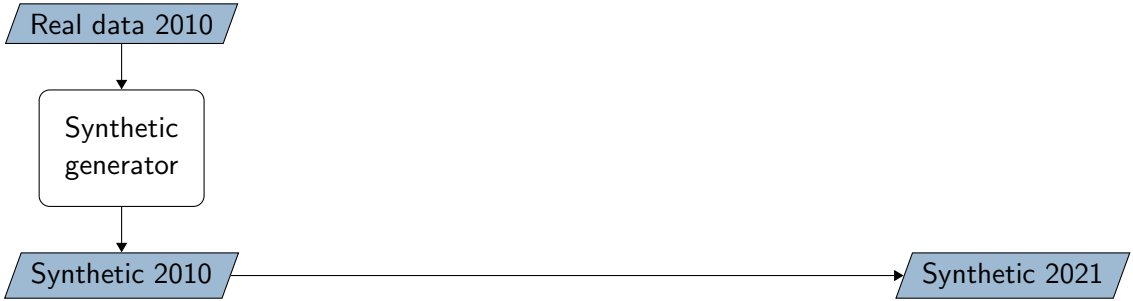
Outline

Existing methods



Step 1: Generation

Existing methods



Step 1: Generation **Step 2: Projection**

Existing methods

Existing methods: Gaps

Dynamic projection

Evolves population.

Heterogeneous sample.

Random resampling

Copying of data instead of evolving.

Lack of heterogeneity over time.

Existing methods: Gaps

Dynamic projection

Evolves population.

Heterogeneous sample.

Propagation of the generation bias.

Increase of the error over time.

Dependent on input rates.

Random resampling

Copying of data instead of evolving.

Lack of heterogeneity over time.

Can achieve a perfect fit of marginals.

Existing methods: Gaps

Dynamic projection

Evolves population.

Heterogeneous sample.

Propagation of the generation bias.

Increase of the error over time.

Dependent on input rates.

Random resampling

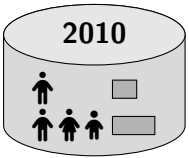
Copying of data instead of evolving.

Lack of heterogeneity over time.

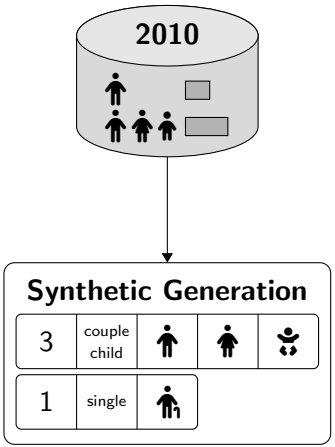
Can achieve a perfect fit of marginals.

Combine the strengths of both methods depending on the data availability!

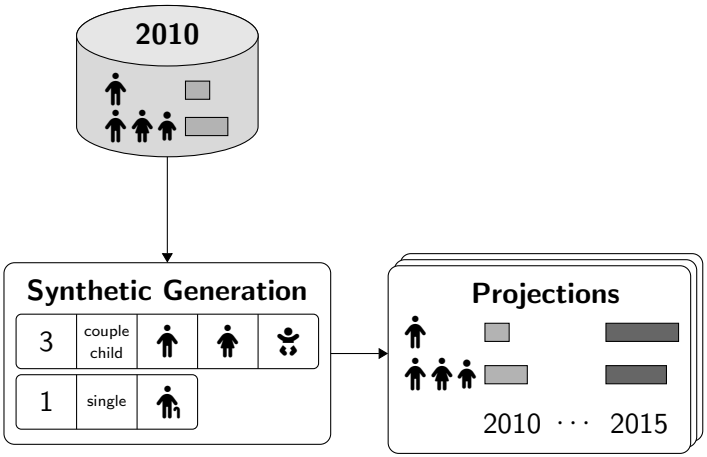
Methodology: **Adaptive** Synthetic Generation using **Gibbs Resampling**



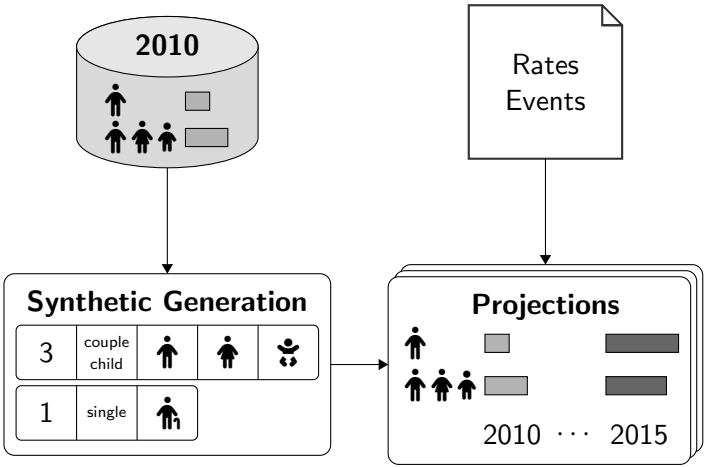
Methodology: **Adaptive** Synthetic Generation using **Gibbs Resampling**



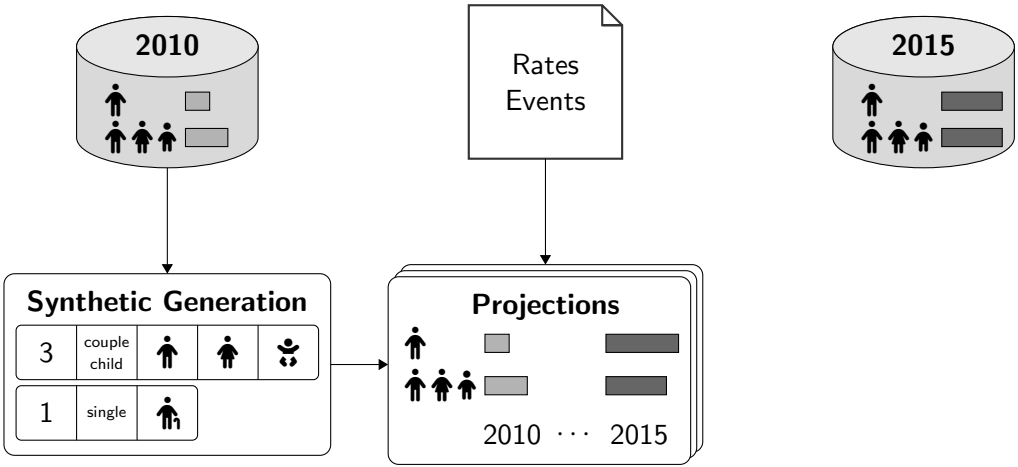
Methodology: **Adaptive** Synthetic Generation using **Gibbs Resampling**



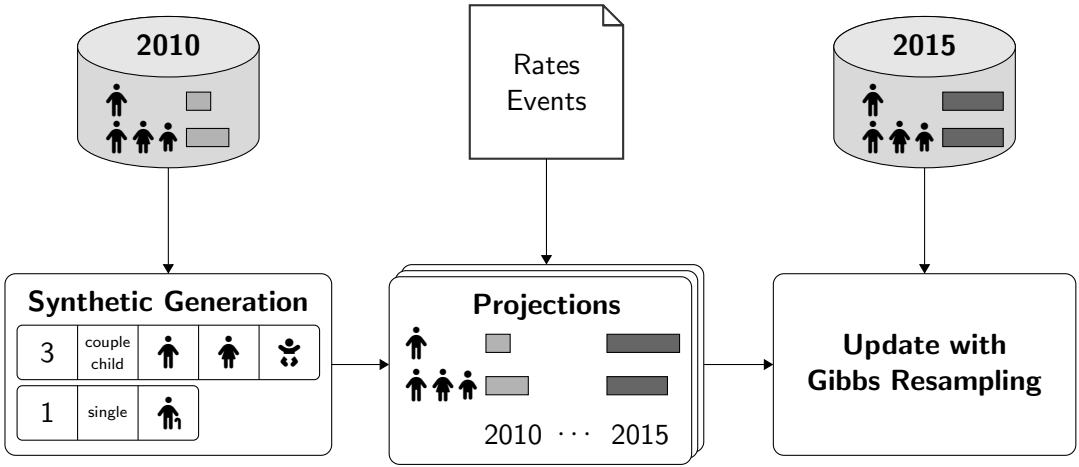
Methodology: **Adaptive** Synthetic Generation using **Gibbs Resampling**



Methodology: **Adaptive** Synthetic Generation using **Gibbs Resampling**



Methodology: **Adaptive** Synthetic Generation using **Gibbs Resampling**



Methodology: Gibbs Resampling

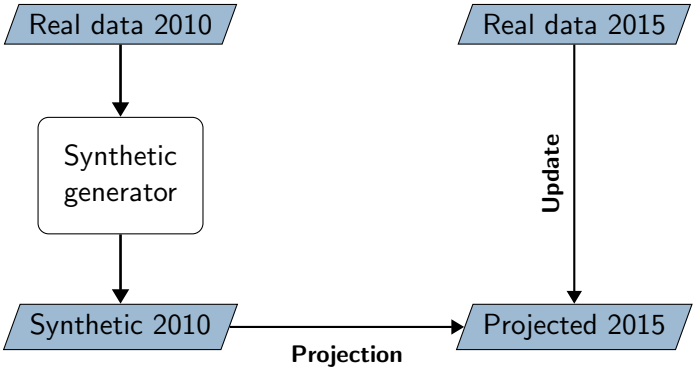
Objective
Align the projected synthetic population with the real marginal distributions observed at t.

Idea
Use the fact that the one-step Gibbs sampler takes the household size as input.

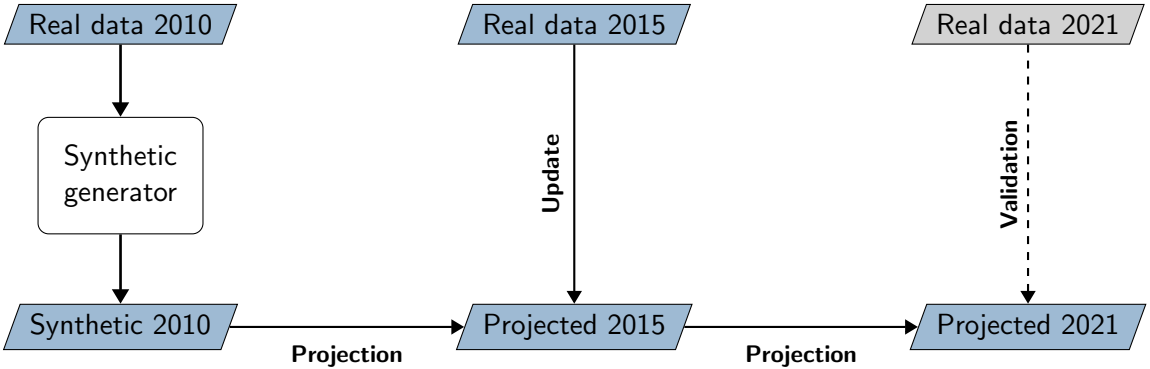
Method

1. **Determine** how much data is needed to match the real distribution.
2. **Generate** additional subsample calibrated on real data.
3. **Add** the new households to the projected synthetic population.
4. **Delete** households uniformly to restore the population size while keeping the adjusted distribution.

Methodology: **Adaptive** Synthetic Generation using **Gibbs Resampling**



Methodology: **Adaptive** Synthetic Generation using **Gibbs Resampling**



Results: Adaptive Synthetic Generation versus other methods

Data: MTMC 2010, 2015, 2021.
Experiment: Regenerate for 2010, 2015, 2021.
Result: more efficient and suitable for cases with limited or biased real data.

	Computational Efficiency
Independent regeneration	✗
Dynamic projection	∼
Random resampling	✓
Adaptive approach	✓

Table: Symbols: ✓ (High), ∼ (Moderate), ✗ (Low), - (not applicable)

Results: Adaptive Synthetic Generation versus other methods

Data: MTMC 2010, 2015, 2021.
Experiment: Projection from 2010 to 2021.
Result: incremental updates reduce accumulated errors over time.

	Computational Efficiency	Marginal Fit
Independent regeneration	✗	✓
Dynamic projection	∼	✗
Random resampling	✓	✓
Adaptive approach	✓	✓

Table: Symbols: ✓ (High), ∼ (Moderate), ✗ (Low), - (not applicable)

Results: Adaptive Synthetic Generation versus other methods

Data: MTMC 2010, 2015, 2021.
Experiment: Projection from 2010 to 2021.
Result: adding new subsample preserves heterogeneity.

	Computational Efficiency	Marginal Fit	Population Heterogeneity
Independent regeneration	✗	✓	✓
Dynamic projection	∼	✗	∼
Random resampling	✓	✓	✗
Adaptive approach	✓	✓	✓

Table: Symbols: ✓ (High), ∼ (Moderate), ✗ (Low), - (not applicable)

Results: Adaptive Synthetic Generation versus other methods

Data: MTMC 2010, 2015, 2021.
Experiment: Projection from 2010 to 2021.

	Computational Efficiency	Marginal Fit	Population Heterogeneity	Robustness to Unforeseen Events
Independent regeneration	✗	✓	✓	-
Dynamic projection	∼	✗	∼	✗
Random resampling	✓	✓	✗	∼
Adaptive approach	✓	✓	✓	✓

Table: Symbols: ✓ (High), ∼ (Moderate), ✗ (Low), - (not applicable)

Summary

Idea

Simulation framework that uses one-step Gibbs sampling for the **adaptive** generation and **updating** of synthetic households by incorporating new disaggregated data without regenerating.

Summary

Idea

Simulation framework that uses one-step Gibbs sampling for the **adaptive** generation and **updating** of synthetic households by incorporating new disaggregated data without regenerating.

Contribution

- Gibbs resampling: improves errors in long-term, more robust, enriching sample.
- Access to up-to-date synthetic samples.
- Making use of available real data.

Summary

Idea

Simulation framework that uses one-step Gibbs sampling for the **adaptive** generation and **updating** of synthetic households by incorporating new disaggregated data without regenerating.

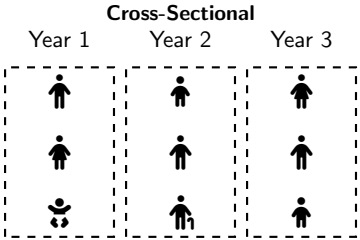
Contribution

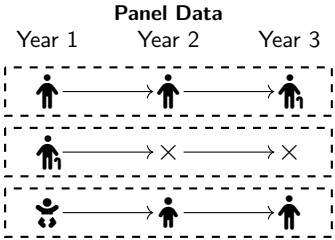
- Gibbs resampling: improves errors in long-term, more robust, enriching sample.
- Access to up-to-date synthetic samples.
- Making use of available real data.

Can we model the evolution at the disaggregated level?

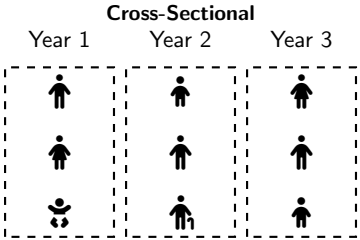
Outline

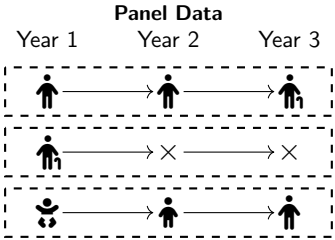
Motivation: Different data types





Motivation: Different data types





Why is there overemphasis on the cross-sectional data, although panel data are needed?

Motivation: Different data types

	Cross-sectional	**Panel**
Strengths	Low cost Easy to collect Accessible Large samples	Tracks behavior dynamics Captures hidden differences Enables causal insights
Limitations	No temporal aspect Misleading conclusions Assumption: *Variation = Dynamics*	Expensive to collect Attrition Panel fatigue Panel conditioning

Gaps, Research question, Contribution

Problem

Real panel inaccessible

$\Rightarrow$ Over-reliance on cross-sectional

$\Rightarrow$ Hinders other models!

Gaps, Research question, Contribution

Problem	Solution
Real panel inaccessible ⇒ Over-reliance on cross-sectional ⇒ Hinders other models!	Generate synthetic panel data ⇒ No practical problems ⇒ Accessible!

Gaps, Research question, Contribution

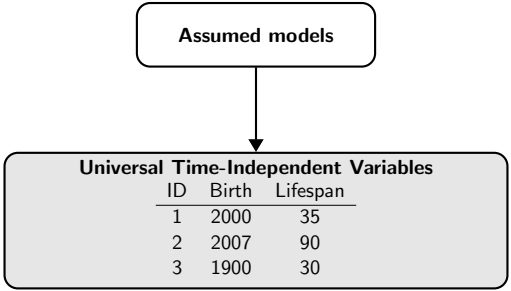
Problem	Solution	Question
Real panel inaccessible ⇒ Over-reliance on cross-sectional ⇒ Hinders other models!	Generate synthetic panel data ⇒ No practical problems ⇒ Accessible!	How to generate synthetic panel data when no real panel data is available?

Gaps, Research question, Contribution

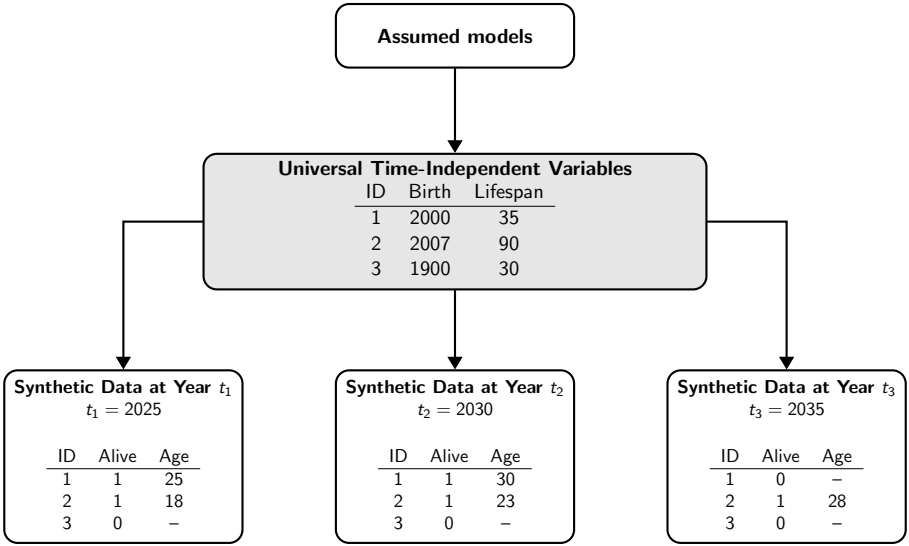
Problem	Solution	Question
Real panel inaccessible ⇒ Over-reliance on cross-sectional ⇒ Hinders other models!	Generate synthetic panel data ⇒ No practical problems ⇒ Accessible!	How to generate synthetic panel data when no real panel data is available?

Generate a universal **time-independent dataset** of individual **life sequences**, created either **without data** or by **integrating cross-sectional** data, enabling **synthetic panels** for any year.

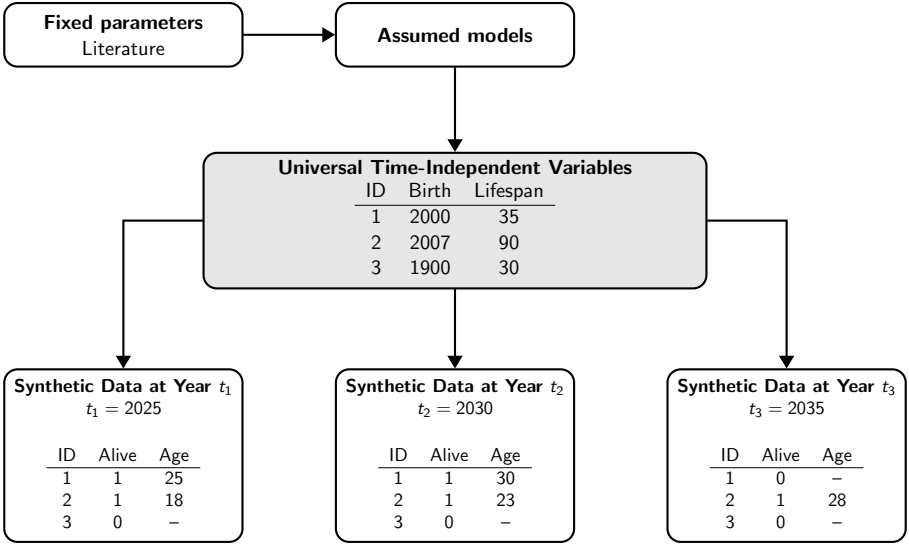
Methodology: Overview of the framework



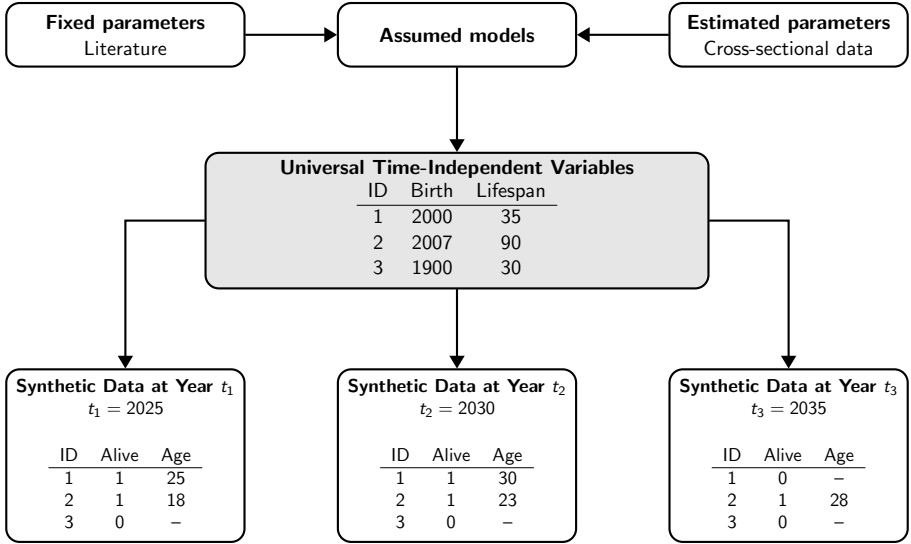
Methodology: Overview of the framework



Methodology: Overview of the framework



Methodology: Overview of the framework



Methodology: Design of universal variables

Time-independent variables

Birth year Lifespan	License year Validity	Education phases	Employment phases	Initial home phase	Home phases	Work phases

Methodology: Design of universal variables

Time-independent variables

Birth year Lifespan	License year Validity	Education phases	Employment phases	Initial home phase	Home phases	Work phases

Event-duration tuples

(e_1, d_1) (e_2, d_2) $\{(e_3, d_3), (e_4, d_4)\}$ $\{(e_m, d_m, \text{inc}_m)\}_m$ $(e_{M+5}, d_{M+5}, c_0^h)$ $\{(e_n, d_n, c_n^h)\}_n$ $\{(e_q, d_q, c_q^w)\}_q$

Methodology: Design of universal variables

Time-independent variables

Birth year Lifespan	License year Validity	Education phases	Employment phases	Initial home phase	Home phases	Work phases

Event-duration tuples

(e_1, d_1)	(e_2, d_2)	$\{(e_3, d_3), (e_4, d_4)\}$	$\{(e_m, d_m, \text{inc}_m)\}_m$	$(e_{M+5}, d_{M+5}, c_0^h)$	$\{(e_n, d_n, c_n^h)\}_n$	$\{(e_q, d_q, c_q^w)\}_q$

Assumed models

Uniform$(y_{\min}, y_{\max})$ Weibull (k, λ)	Mixture (π, μ, σ)	TruncNorm (μ_d^S, σ_d^S) Exp (μ_g^T) Weibull (k_T, λ_T)	Exp (μ_g) Exp $(\lambda(a_m))$ Exp (μ_b)	Population distribution	Income-boost model	Neighbour model (p_h, p_w)

Methodology: Design of universal variables

Time-independent variables

Birth year Lifespan	License year Validity	Education phases	Employment phases	Initial home phase	Home phases	Work phases

Event-duration tuples

(e_1, d_1)	(e_2, d_2)	$\{(e_3, d_3), (e_4, d_4)\}$	$\{(e_m, d_m, \text{inc}_m)\}_m$	$(e_{M+5}, d_{M+5}, c_0^h)$	$\{(e_n, d_n, c_n^h)\}_n$	$\{(e_q, d_q, c_q^w)\}_q$

Assumed models

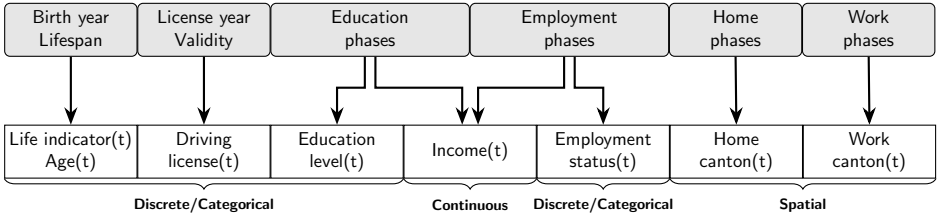
Uniform$(y_{\min}, y_{\max})$ Weibull (k, λ)	Mixture (π, μ, σ)	TruncNorm (μ_d^S, σ_d^S) Exp (μ_g^T) Weibull (k_T, λ_T)	Exp (μ_g) Exp $(\lambda(a_m))$ Exp (μ_b)	Population distribution	Income-boost model	Neighbour model (p_h, p_w)

Generated sequence

(1995, 30)	(2015, 10)	$\{(2010, 4), (2014, 4)\}$	(2018, 7, 3000)	(1995, 23, VD)	(2018, 7, ZH)	(2018, 7, ZH)

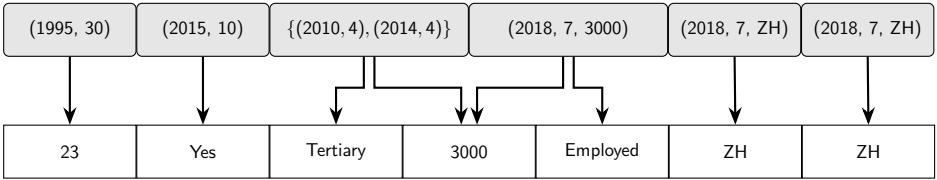
Within lifetime **Realistic** **Consistency**

Methodology: Specification of time-dependent variables



Deterministic mapping rules

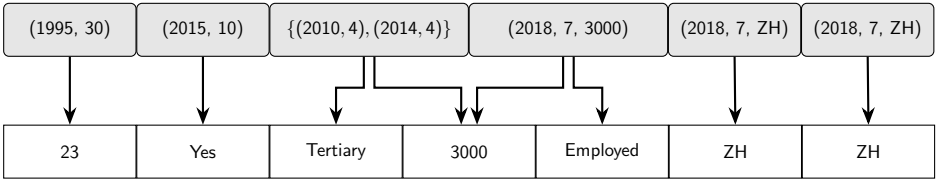
Methodology: Specification of time-dependent variables



$t = 2018$

- Derivation faster than regeneration.
- Control of consistency.
- Changes are automatically reflected.

Methodology: Specification of time-dependent variables

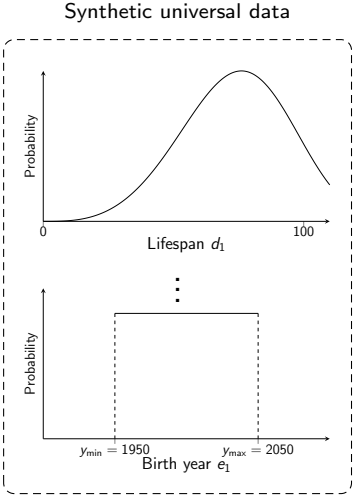


$t = 2018$

- Derivation faster than regeneration.
- Control of consistency.
- Changes are automatically reflected.

We control the **panel effect** → But how do we know that the **aggregated properties** are satisfied?

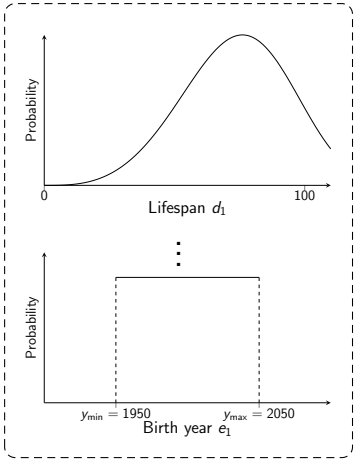
Methodology: Cross-sectional data integration



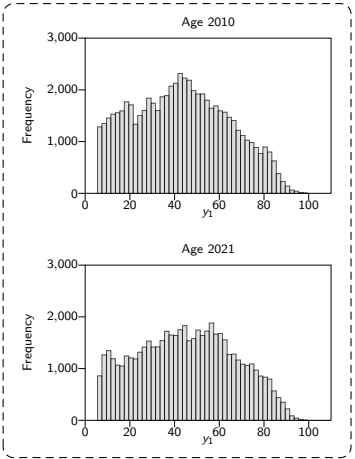
How do we know if the generated sequence produces **realistic data** at moment t?

Methodology: Cross-sectional data integration

Synthetic universal data

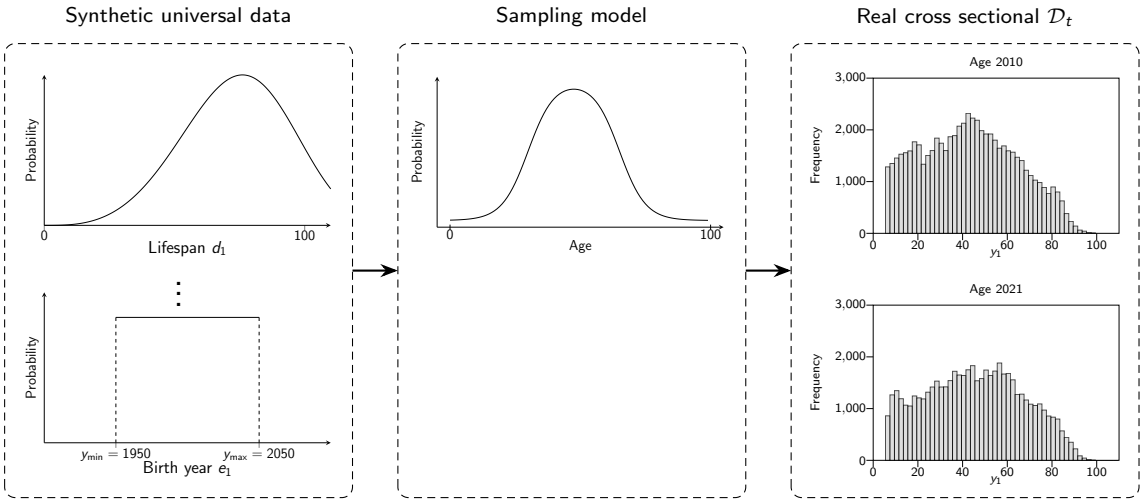


Real cross sectional $\mathcal{D}_t$



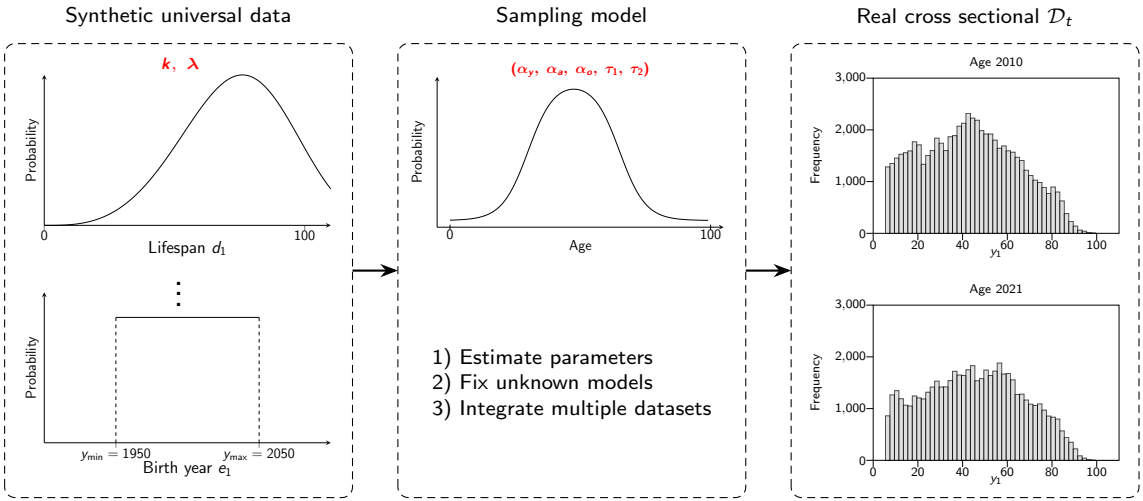
Can we use the information from **observed data** to improve **unobserved variables**?

Methodology: Cross-sectional data integration



Cross-sectional datasets **are outcomes of a sampling mechanism** applied to the universal population at moment t.

Methodology: Cross-sectional data integration



MLE to **estimate parameters** of universal-variable distributions using available cross-sectional datasets.

Cross-sectional Data Integration: MLE

$$\mathrm{L}(\theta) = \prod_{t \in \mathcal{T}} \prod_{i \in \mathcal{D}_t} \mathrm{P}(s_{i,t} = 1,\, z_{i,t} \mid \theta)$$

$s_{i,t} = 1$ sampled individuals

$z_{i,t}$ observed variables at time t (e.g., age, driving license status, ...)

θ parameters (e.g., $k, \lambda, \alpha_y, \alpha_a, \alpha_o, \tau_1, \tau_2$, ...)

Cross-sectional Data Integration: MLE

$$\mathrm{L}(\theta) = \prod_{t \in \mathcal{T}} \prod_{i \in \mathcal{D}_t} \mathrm{P}(s_{i,t} = 1,\, z_{i,t} \mid \theta)$$

$$\mathrm{L}(\theta) = \prod_{t \in \mathcal{T}} \prod_{i \in \mathcal{D}_t} \sum_{\mathbf{u}_\theta} \mathrm{P}(s_{i,t} = 1,\, z_{i,t},\, \mathbf{u}_\theta \mid \theta)$$

$s_{i,t} = 1$ sampled individuals

$z_{i,t}$ observed variables at time t (e.g., age, driving license status, ...)

θ parameters (e.g., $k, \lambda, \alpha_y, \alpha_a, \alpha_o, \tau_1, \tau_2$, ...)

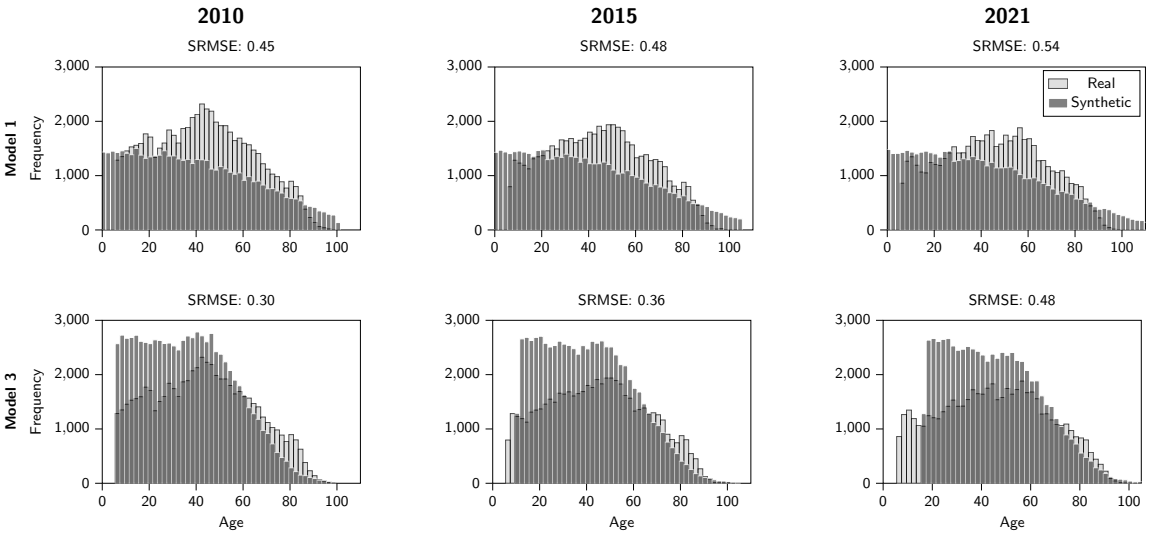
$\mathbf{u}_\theta$ universal variables (e.g., birth year, lifespan, age of getting driving license, ...)

Results: Data-free vs. Data-Integrated approach

Estimation of parameters - Model 1 (data-free) and Model 2 – 4 (data-integrated).
MTMC cross-sectional data 2010, 2015, 2021.

Parameter	Meaning	Model 1	Model 2	Model 3	Model 4
k	Shape (lifespan)	3.00	2.84	3.44	3.37
λ	Scale (lifespan)	85.00	69.28	77.36	74.41
α_y	Weight of young group	–	0.02	0.16	0.06
α_a	Weight of adult group	–	0.95	0.79	0.84
α_o	Weight of old group	–	0.04	0.05	0.10
τ_1	Young–adult threshold	–	33.98	38.06	37.73
τ_2	Adult–old threshold	–	84.89	82.50	87.76
π	Probability of never obtaining a license	0.15	0.17	0.19	0.21
μ	Log-location (license age)	3.02	2.89	2.87	2.90
σ	Log-scale (license age)	0.15	0.13	0.11	0.11

Results: Data-Free vs. Data-Integrated approach



Results: Data-Free vs. Data-Integrated approach

Model	Unseen data		
	2010	2015	2021
Model 1	**0.45**	**0.48**	**0.54**
Model 2	(0.38)	**0.43**	**0.50**
Model 3	(0.30)	(0.36)	**0.48**
Model 4	(0.29)	(0.36)	(0.47)

- Model 1: data-free $\rightarrow$ Worst fit

Results: Data-Free vs. Data-Integrated approach

Model	Unseen data		
	2010	2015	2021
Model 1	**0.45**↓	**0.48**↓	**0.54**↓
Model 2	(0.38)	**0.43**↓	**0.50**↓
Model 3	(0.30)	(0.36)	**0.48**↓
Model 4	(0.29)	(0.36)	(0.47)

- Model 1: data-free → Worst fit
- Model 3 & Model 4:
  more integrated data → Better fit.

Results: Data-Free vs. Data-Integrated approach

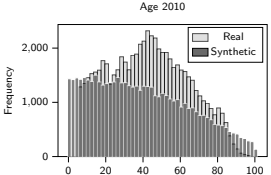
Model	Unseen data		
	2010	2015	2021
Model 1	**0.45**	**0.48**	**0.54**
Model 2	(0.38)	**0.43**	**0.50**
Model 3	(0.30)	(0.36)	**0.48**
Model 4	(0.29)	(0.36)	(0.47)

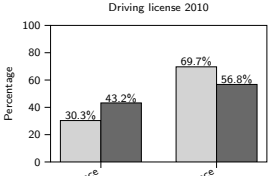
- Model 1: data-free → Worst fit
- Model 3 & Model 4:
  more integrated data → Better fit.
- Fit decreases over time
  (i.e., SRMSE increases).

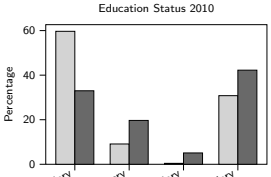
Results: Data-free vs. Data-Integrated approach

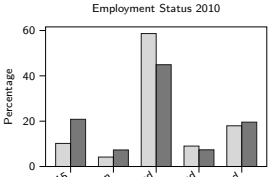
Aggregated validation: generate universal variables, derive year-specific datasets (e.g., 2010), and compare resulting marginals with observed data.

Data-free (Model 1)

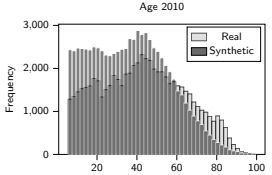


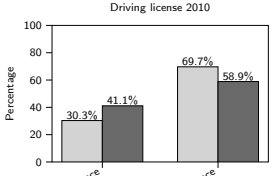


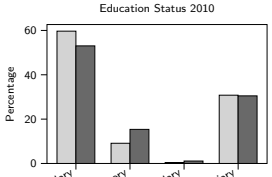


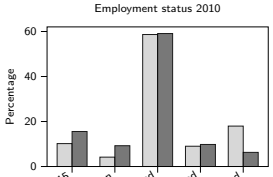


Data-integrated (Model 3)



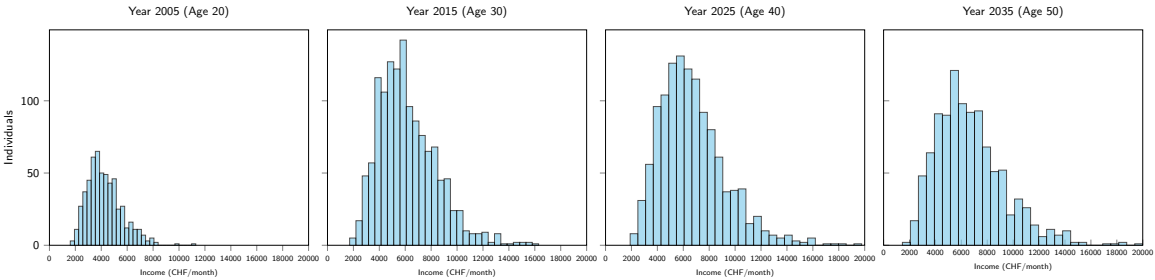






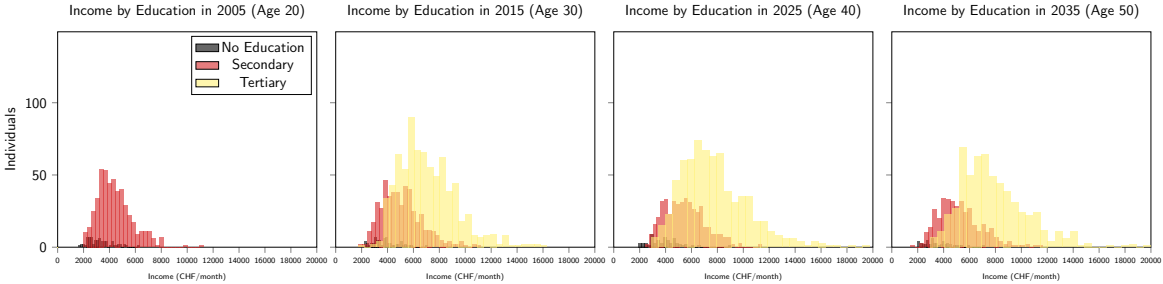
Results: Panel effect

Generate a universal dataset and select people born in 1985.
For these people, derive data for 2005, 2015, 2025, 2035.



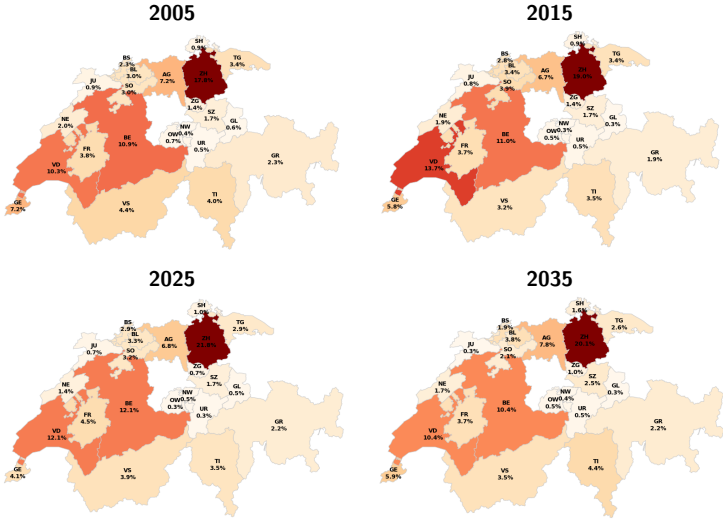
Income increases and the spread widens over time.
Longer tails mean more extreme incomes and greater income differences.

Results: Panel effect



Income levels strongly correlate with education level.

Results: Panel effect



Summary

Idea

Universal life-course event–duration model for generating synthetic panel data.

Summary

Idea

Universal life-course event–duration model for generating synthetic panel data.

Contribution

- Consistent individual tracking.
- Data-free and data-integrated generation.
- Automatic propagation of changes.

Summary

Idea

Universal life-course event–duration model for generating synthetic panel data.

Contribution

- Consistent individual tracking.
- Data-free and data-integrated generation.
- Automatic propagation of changes.

Limitations

- Simple models.
- Limited aggregated fit.

Summary

Idea

Universal life-course event–duration model for generating synthetic panel data.

Contribution

- Consistent individual tracking.
- Data-free and data-integrated generation.
- Automatic propagation of changes.

Limitations

- Simple models.
- Limited aggregated fit.

Future directions

Expand on tracking synthetic households.

Bayesian updating using defined models as priors.

User-defined panels for hypothetical testing.

Outline

Conclusion: Main Findings

Motivation

- Synthetic populations: **single-level** and **static**.
- No temporal **continuity** or **adaptability** to new data.

Conclusion: Main Findings

Motivation

- Synthetic populations: **single-level** and **static**.
- No temporal **continuity** or **adaptability** to new data.

Contributions

- Hierarchical generation.
- Adaptive updates.
- Synthetic panel data.

Conclusion: Main Findings

Motivation

- Synthetic populations: **single-level** and **static**.
- No temporal **continuity** or **adaptability** to new data.

Contributions

- Hierarchical generation.
- Adaptive updates.
- Synthetic panel data.

Vision

- Synthetic populations as **evolving** systems.
- **Adaptive**: **model-based** reasoning with **data-driven** calibration.

Conclusion: Future work

- Improving validation of synthetic data (e.g., rare households).
- Extension to other hierarchical structures.
- Formal more general life-trajectory framework with explicit structural constraints.
- Open-source software for generating synthetic panels under user-defined constraints.

Thank you! Questions?

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

References I

Ahmed, U. and Moeckel, R. (2023). Impact of Life Events on Incremental Travel Behavior Change. *Transportation Research Record*, 2677(9):594–605.

Casati, D., Müller, K., Fourie, P. J., Erath, A., and Axhausen, K. W. (2015). Synthetic population generation by combining a hierarchical, simulation-based approach with reweighting by generalized ranking. *Transportation Research Record*, 2493(1):107–116.

Castiglione, J., Bradley, M., and Gliebe, J. (2014). *Activity-Based Travel Demand Models: A Primer*. The National Academies Press, Washington, DC.

Haghighi, M. and Miller, E. J. (2025). Week-long activity-based modelling: a review of the existing models and datasets and a comprehensive conceptual framework. *Transport Reviews*, 45(1):119–148.

La, D. M., Vu, H., Kamruzzaman, L., and Miller, E. (2025). Population synthesis: a problem-based review. *Transport Reviews*, 45:1–24.

References II

Ramadan, O. E. and Sisiopiku, V. P. (2019). A Critical Review on Population Synthesis for Activity- and Agent-Based Transportation Models. In Luca, S. D., Pace, R. D., and Djordjevic, B., editors, *Transportation Systems Analysis and Assessment*. IntechOpen.

Tajaddini, A., Rose, G., Kockelman, K. M., and Vu, H. L. (2020). Recent Progress in Activity-Based Travel Demand Modeling: Rising Data and Applicability. In de Luca, S., Pace, R. D., and Fiori, C., editors, *Models and Technologies for Smart, Sustainable and Safe Transportation Systems*. IntechOpen.

Yaméogo, B. F., Gastineau, P., Hankach, P., and Vandanjon, P.-O. (2021). Comparing methods for generating a two-layered synthetic population. *Transportation Research Record*, 2675(1):136–147.