# Hybrid Simulator for Projecting Synthetic Households in Unforeseen Events

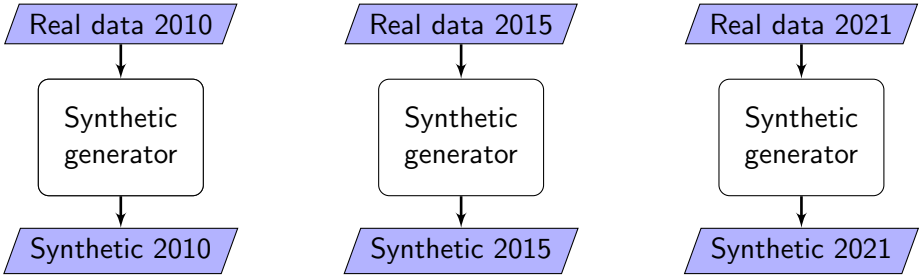
Marija Kukic  Michel Bierlaire

16 May, 2024

STRC | 24th Swiss Transport Research Conference
Monte Verità / Ascona, May 15-17, 2024

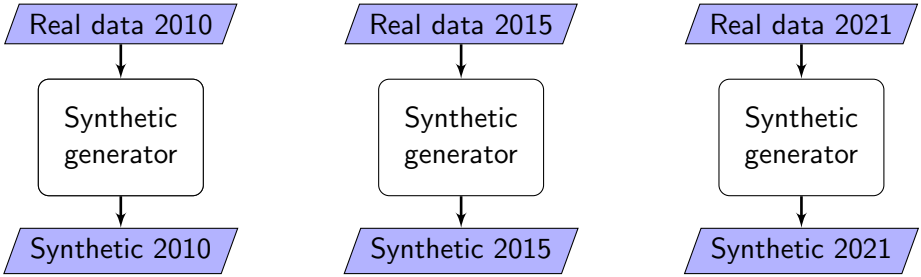
# Outline

# Regenerating Synthetic Population




**Synthetic Snapshot** = tabular data on individuals and households at a given point in time

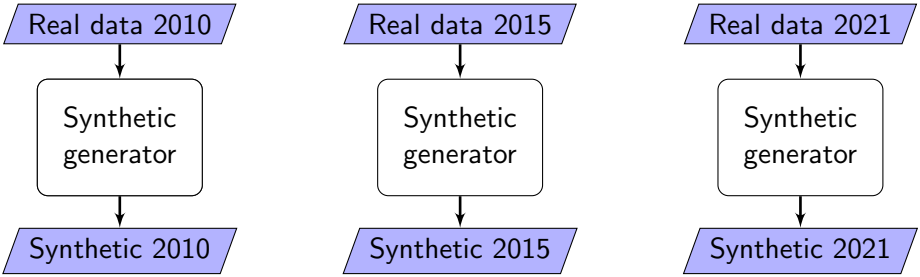
# Regenerating Synthetic Population: Problems




**Outdated sample**

# Regenerating Synthetic Population: Problems

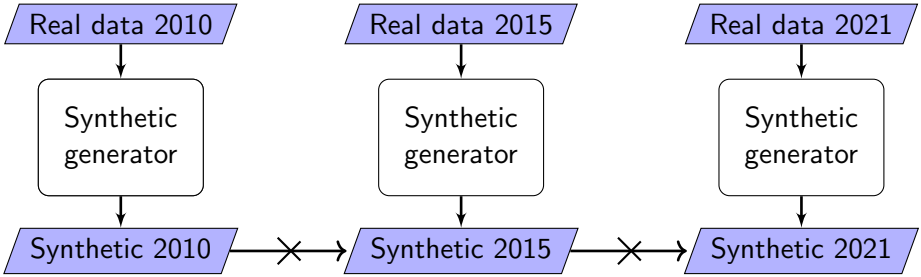
**Complicated and Repetitive**




**Outdated sample**

# Regenerating Synthetic Population: Problems

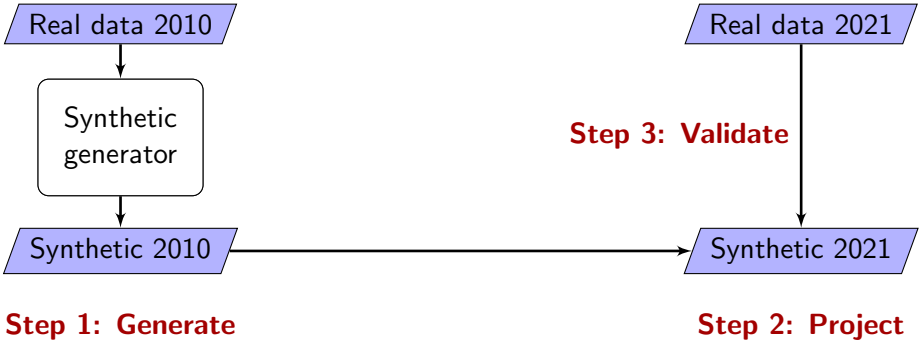
**Complicated and Repetitive**




**Outdated sample**

**Costly**

# Synthetic Projections



# Outline

# Existing Projection Methods

**Dynamic projection**[1, 2]

- Evolves population.
- Heterogeneous sample.

**Resampling**[3]

- Copying of data instead of evolving.
- Lack of heterogeneity over time.

# Existing Projection Methods

**Dynamic projection**[1, 2]

- Evolves population.
- Heterogeneous sample.

- Propagation of the generation bias.
- Increase of the error over time.

**Resampling**[3]

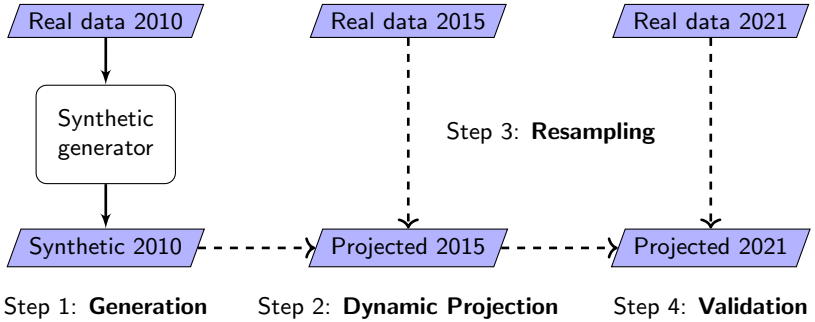
- Copying of data instead of evolving.
- Lack of heterogeneity over time.

- Can achieve a perfect fit of marginals.

# Outline

# Previous work

- Combine **dynamic projection** (i.e., model-driven) and **resampling** (i.e., data-driven) approach[4].
- Simulate **death, birth, migration** of synthetic individuals described by **age, gender, employment**.



# Drawbacks of Dynamic Projection

Dynamic projection is **dependent** on input demographic rates and **not robust** to the **unusual events**.

| | Report from 2010 [5] | | | Report from 2020 [6] | | |
|---|---|---|---|---|---|---|
| **Year** | **Births** | **Deaths** | **Balance** | **Births** | **Deaths** | **Balance** |
| 2020 | 82,7 | 66,4 | **16,3** | 89,4 | 67,5 | **21,9** |
| 2025 | 81,2 | 70,7 | **10,5** | 92,2 | 70,5 | **21,7** |
| 2030 | 78,3 | 76,2 | **2,1** | 93,8 | 74,9 | **18,9** |
| 2035 | 76,6 | 82,1 | **-5,5** | 94,9 | 80,0 | **14,9** |
| 2040 | 77,2 | 87,7 | **-10,5** | 96,3 | 85,5 | **10,8** |
| 2045 | 78,4 | 92,8 | **-14,4** | 98,4 | 90,6 | **7,8** |
| 2050 | 79,0 | 97,5 | **-18,5** | 100,3 | 95,5 | **4,8** |

Table: Demographic estimates in $10^3$ provided by Swiss Federal Statistical Office (BFS)

# Drawbacks of Dynamic Projection

Dynamic projection is **dependent** on input demographic rates and **not robust** to the **unusual events**.

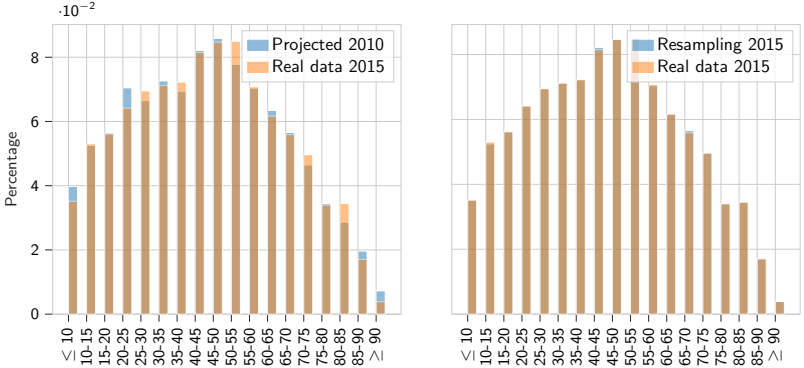
| | Report from 2010 [5] | | | Report from 2020 [6] | | |
|---|---|---|---|---|---|---|
| **Year** | **Births** | **Deaths** | **Balance** | **Births** | **Deaths** | **Balance** |
| 2020 | 82,7 | 66,4 | **16,3** | 89,4 | 67,5 | **21,9** |
| 2025 | 81,2 | 70,7 | **10,5** | 92,2 | 70,5 | **21,7** |
| 2030 | 78,3 | 76,2 | **2,1** | 93,8 | 74,9 | **18,9** |
| 2035 | 76,6 | 82,1 | **-5,5** | 94,9 | 80,0 | **14,9** |
| 2040 | 77,2 | 87,7 | **-10,5** | 96,3 | 85,5 | **10,8** |
| 2045 | 78,4 | 92,8 | **-14,4** | 98,4 | 90,6 | **7,8** |
| 2050 | 79,0 | 97,5 | **-18,5** | 100,3 | 95,5 | **4,8** |

Table: Demographic estimates in $10^3$ provided by Swiss Federal Statistical Office (BFS)

Although **rates** are frequently **updated**, **synthetic datasets** made using them **are not**.

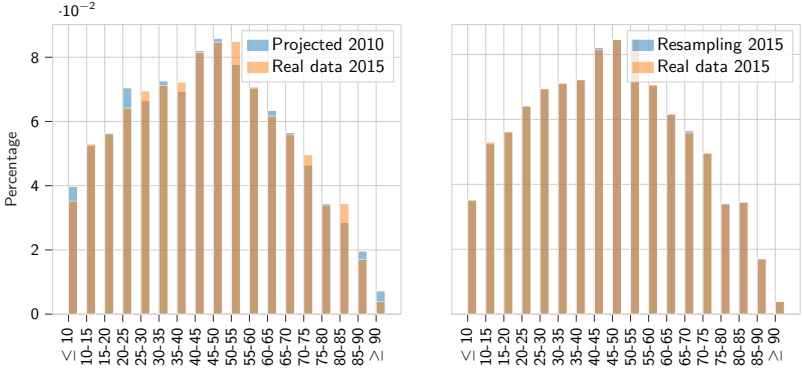
# Drawbacks of Resampling

**Resampling** procedure **randomly adds** and **deletes** people to achieve a perfect **marginal age fit**.



# Drawbacks of Resampling

**Resampling** procedure **randomly adds** and **deletes** people to achieve a perfect **marginal age fit**.




Marginal fit does not guarantee **heterogeneity** nor **representativeness** of the sample.

# Current work

- Expand the method from the level of **individuals** to the **household** level.
- Adapting one-step **Gibbs Sampler** (GS) for **resampling** to enhance the **heterogeneity**.

# Current work

- Expand the method from the level of **individuals** to the **household** level.
- Adapting one-step **Gibbs Sampler** (GS) for **resampling** to enhance the **heterogeneity**.
- Evaluate **robustness** of hybrid simulator to **unforeseen events** (i.e., COVID-19) compared to state-of-the-art methods by testing **pre-pandemic** and **post-pandemic** scenario.

# Outline

# Hybrid Simulator

## Step 1: Generation

Markov Chain Monte Carlo Simulation. [7]

Synthetic households of size $N$, $X = (X_{\text{type}}, X_{\text{nb\_cars}}, [individuals_i]_{i \in [1...N]})$.

Synthetic individual described by $X_{\text{age}}, X_{\text{gender}}, X_{\text{empl}}, X_{\text{marital}}, X_{\text{dl}}$.

Bootstrap and convergence monitoring.

# Hybrid Simulator

## Step 1: Generation

Markov Chain Monte Carlo Simulation. [7]

Synthetic households of size $N$, $X = (X_{\text{type}}, X_{\text{nb\_cars}}, [individuals_i]_{i \in [1...N]})$.

Synthetic individual described by $X_{\text{age}}, X_{\text{gender}}, X_{\text{empl}}, X_{\text{marital}}, X_{\text{dl}}$.

Bootstrap and convergence monitoring.

## Step 2: Dynamic projection

When disaggregated data are not available.

Simulate events: **birth, death, migration, marriage, divorce, leaving the house**.

Use the rates provided by the Swiss Federal Statistical Office (BFS) [5, 6].

# Step 3: Resampling

When disaggregated data are available.

**Objective:** To align the projected and real marginal distributions generate and add a small portion of data using **one-step Gibbs sampler** (GS).

# Step 3: Resampling

When disaggregated data are available.

**Objective:** To align the projected and real marginal distributions generate and add a small portion of data using **one-step Gibbs sampler** (GS).

**Question:** How to determine the **quantities** of data that need **to be added** to each **household size** category so we obtain **the same probability distribution** as in real data?

# Step 3: Resampling

**Projected Data**

| | $c$ | $\lambda$ |
|---|---|---|
| 1 | 23287 | 0.16 |
| 2 | 43760 | 0.30 |
| 3 | 26397 | 0.18 |
| 4 | 37620 | 0.23 |
| 5 | 41995 | 0.10 |
| 6 | 52882 | 0.03 |

**Real Data**

| | $c'$ | $\lambda'$ |
|---|---|---|
| 1 | 49373 | 0.15 |
| 2 | 40286 | 0.32 |
| 3 | 29786 | 0.17 |
| 4 | 28228 | 0.22 |
| 5 | 40320 | 0.05 |
| 6 | 56493 | 0.05 |

**Objective:** Find $\mathbf{x} = (x_1, \ldots, x_n)$ where $x_i \geq 0$, representing additional observations needed by solving:

$$c_j + x_j = \lambda'_j \sum_{i=1}^{N} c_i + \lambda'_j \sum_{i=1}^{N} x_i \tag{1}$$

**Result:** New counts $\mathbf{c} + \mathbf{x}$ align with real data distribution $\lambda'$.

# Outline

# Generation of synthetic sample 2010 - Household level

**Validation:** Compare distributions with real data using statistics (e.g., SRMSE) and visualization.
Reference data: weighted **MTMC 2010, 2015, 2021** [BFS]

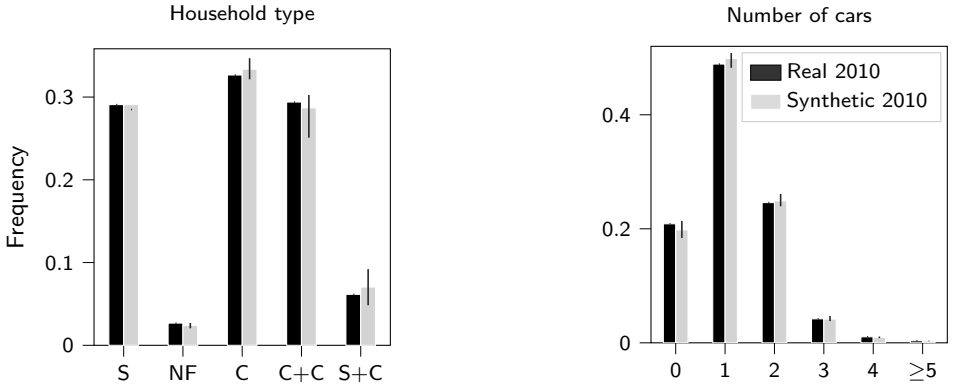
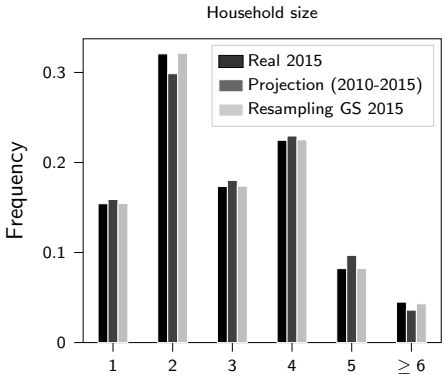



Figure: The comparison of household marginals between synthetic and real sample from 2010

# Dynamic Projection (2010-2015) and Re-sampling (2015)



# Dynamic Projection (2010-2015) and Re-sampling (2015)

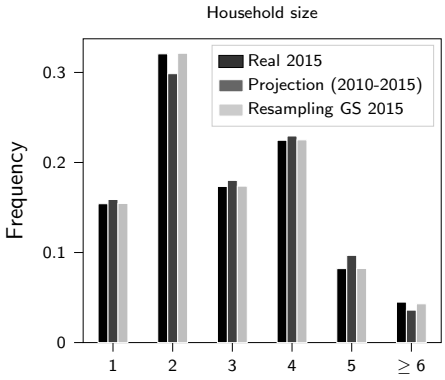



Table: Comparison of resampling methods compared to the real data from 2015

| | First order | Second order | Third order |
|---|---|---|---|
| GS | 0.05 | **0.12** | **0.25** |
| Random | 0.05 | 0.16 | 0.32 |

# Comparison of Dynamic Projection and Hybrid Approach - 2021

| Variable | Pre-pandemic scenario | | Post-pandemic scenario | |
|---|---|---|---|---|
| | Dynamic projection | Hybrid simulator | Dynamic projection | Hybrid simulator |
| Household size | 0.22 | 0.15 | 0.19 | 0.12 |
| Household type | 0.24 | 0.1 | 0.15 | 0.08 |
| Number of cars | 0.32 | 0.18 | 0.24 | 0.12 |
| Age | 0.24 | 0.07 | 0.04 | 0.02 |
| Gender | 0.01 | 0.01 | 0.01 | 0.01 |
| Driving licence | 0.1 | 0.1 | 0.1 | 0.1 |
| Marital status | 0.07 | 0.06 | 0.07 | 0.06 |
| Employment | 0.26 | 0.25 | 0.16 | 0.15 |
| Average SRMSE | 0.18 | 0.11 | 0.12 | 0.08 |

# Comparison of Dynamic Projection and Hybrid Approach - 2021

The hybrid simulator achieved a better score (i.e., lower) for each attribute in both scenarios.

| | Pre-pandemic scenario | | Post-pandemic scenario | |
|---|---|---|---|---|
| **Variable** | **Dynamic projection** | **Hybrid simulator** | **Dynamic projection** | **Hybrid simulator** |
| Household size | 0.22 | **0.15** | 0.19 | **0.12** |
| Household type | 0.24 | **0.1** | 0.15 | **0.08** |
| Number of cars | 0.32 | **0.18** | 0.24 | **0.12** |
| Age | 0.24 | **0.07** | 0.04 | **0.02** |
| Gender | 0.01 | 0.01 | 0.01 | 0.01 |
| Driving licence | 0.1 | 0.1 | 0.1 | 0.1 |
| Marital status | 0.07 | **0.06** | 0.07 | **0.06** |
| Employment | 0.26 | **0.25** | 0.16 | **0.15** |
| Average SRMSE | 0.18 | **0.11** | 0.12 | **0.08** |

# Comparison of Dynamic Projection and Hybrid Approach - 2021

Some attributes are not affected by unforeseen events.

| Variable | Pre-pandemic scenario | | Post-pandemic scenario | |
|---|---|---|---|---|
| | Dynamic projection | Hybrid simulator | Dynamic projection | Hybrid simulator |
| Household size | 0.22 | 0.15 | 0.19 | 0.12 |
| Household type | 0.24 | 0.1 | 0.15 | 0.08 |
| Number of cars | 0.32 | 0.18 | 0.24 | 0.12 |
| Age | 0.24 | 0.07 | 0.04 | 0.02 |
| **Gender** | **0.01** | **0.01** | **0.01** | **0.01** |
| **Driving licence** | **0.1** | **0.1** | **0.1** | **0.1** |
| Marital status | 0.07 | 0.06 | 0.07 | 0.06 |
| Employment | 0.26 | 0.25 | 0.16 | 0.15 |
| Average SRMSE | 0.18 | 0.11 | 0.12 | 0.08 |

# Comparison of Dynamic Projection and Hybrid Approach - 2021

Using updated rates leads to better results for both methods.

| Variable | Pre-pandemic scenario | | Post-pandemic scenario | |
|---|---|---|---|---|
| | Dynamic projection | Hybrid simulator | Dynamic projection | Hybrid simulator |
| Household size | 0.22 | 0.15 | 0.19 | 0.12 |
| Household type | 0.24 | 0.1 | 0.15 | 0.08 |
| Number of cars | 0.32 | 0.18 | 0.24 | 0.12 |
| Age | 0.24 | 0.07 | 0.04 | 0.02 |
| Gender | 0.01 | 0.01 | 0.01 | 0.01 |
| Driving licence | 0.1 | 0.1 | 0.1 | 0.1 |
| Marital status | 0.07 | 0.06 | 0.07 | 0.06 |
| Employment | 0.26 | 0.25 | 0.16 | 0.15 |
| **Average SRMSE** | 0.18 | 0.11 | **0.12** | **0.08** |

# Comparison of Dynamic Projection and Hybrid Approach - 2021

The difference between pre and post-pandemic scenarios is smaller for the hybrid simulator.

| | Pre-pandemic scenario | | Post-pandemic scenario | |
|---|---|---|---|---|
| **Variable** | **Dynamic projection** | **Hybrid simulator** | **Dynamic projection** | **Hybrid simulator** |
| Household size | 0.22 | 0.15 | 0.19 | 0.12 |
| Household type | 0.24 | 0.1 | 0.15 | 0.08 |
| Number of cars | 0.32 | 0.18 | 0.24 | 0.12 |
| Age | 0.24 | 0.07 | 0.04 | 0.02 |
| Gender | 0.01 | 0.01 | 0.01 | 0.01 |
| Driving licence | 0.1 | 0.1 | 0.1 | 0.1 |
| Marital status | 0.07 | 0.06 | 0.07 | 0.06 |
| Employment | 0.26 | 0.25 | 0.16 | 0.15 |
| **Average SRMSE** | 0.18 | **0.11** | 0.12 | **0.08** |

# Outline

# Conclusion and Future work

**What do we do?**

Expand a hybrid simulator from the individual to **household level** by redefining the **dynamic projection** and **resampling** procedure.

# Conclusion and Future work

## What do we do?

Expand a hybrid simulator from the individual to **household level** by redefining the **dynamic projection** and **resampling** procedure.

## What do we show?

- Resampling using GS helps enhance the heterogeneity of the projected synthetic sample.
- The hybrid simulator is more robust to unforeseen events than the dynamic projection.
- The significance of validating and updating synthetic projected samples.

# Conclusion and Future work

## What do we do?

Expand a hybrid simulator from the individual to **household level** by redefining the **dynamic projection** and **resampling** procedure.

## What do we show?

- Resampling using GS helps enhance the heterogeneity of the projected synthetic sample.
- The hybrid simulator is more robust to unforeseen events than the dynamic projection.
- The significance of validating and updating synthetic projected samples.

## What do we do next?

Evaluating Incremental Generation of Hybrid Simulator vs. Complete Regeneration

# Thank you! Questions?

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

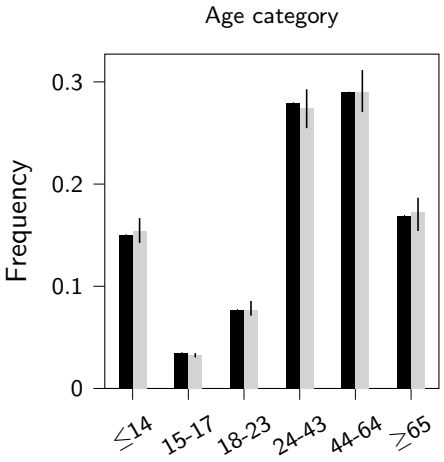
**References**

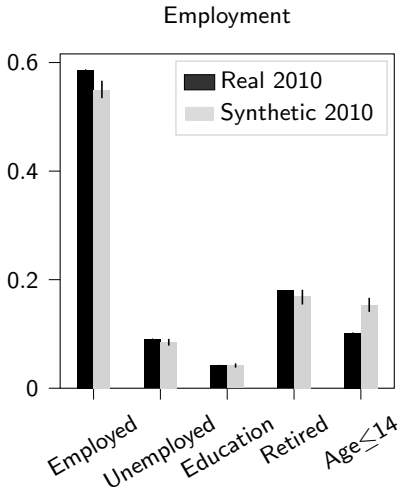
# Backup slides

## Literature review - Generation and Projection

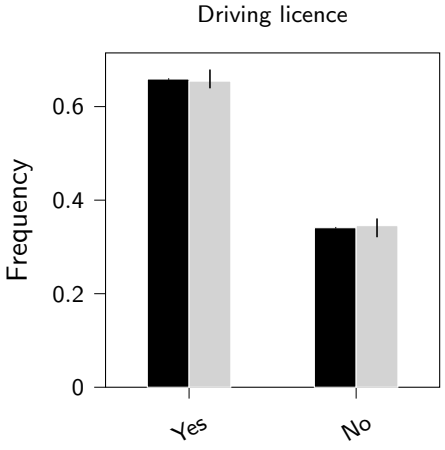
| | **Dynamic projection** | **Static projection** | **Resampling** |
|---|---|---|---|
| **Synthetic reconstruction** | Fatmi et al.[1] **2017** | Lomax et al.[8] **2022** | Prédhumeau et al.[3] **2023** |
| **Combinatorial optimisation** | Namazi-Rad et el.[2] **2014** | **X** | **X** |
| **Statistical learning** | **Hybrid Simulator for Capturing Dynamics Model-driven** | **X** | **Hybrid Simulator for Capturing Dynamics Data-driven** |

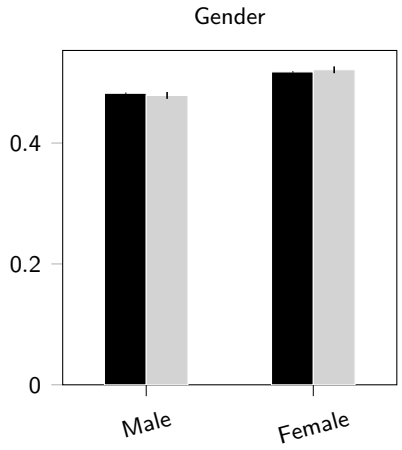
# Backup slides



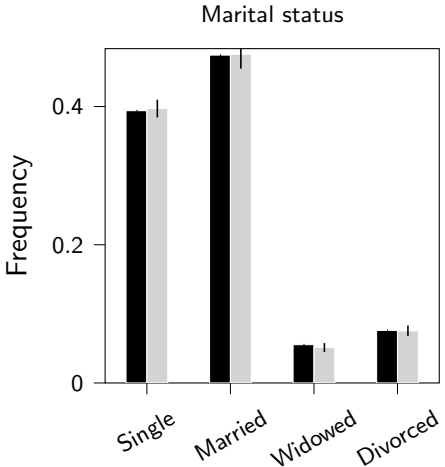



# Backup slides



# Backup slides



# Backup slides - Dynamic projection - Births

**Require:** $P$ - synthetic population

1: **for** $(a, m, o)$ in [age classes, marital statuses, birth orders] **do**
2: Extract mother candidates $M$ in $P$ with attributes $(a, m, o)$
3: Get the number $B$ of births with attributes $(a, m, o)$ {From BFS data}
4: Draw $B$ mothers from $M$
5: Add newborn in mothers' households
6: **end for**

**Algorithm 1:** Births simulation

# Backup slides - Dynamic projection - Migrations

$P$ - synthetic population
**for** $(a, g)$ in [ages, genders] **do**
  Get the net migration $N$ for attributes $(a, g)$ {From BFS data}
  **if** $N \geq 0$ **then**
    Draw $N$ individuals with attributes $(a, g)$ from $P$ {With replacement}
    Duplicate the $N$ individuals
    Build households from new individuals
  **else**
    Remove $N$ individuals with attributes $(a, g)$ from $P$
    Adapt modified households
  **end if**
**end for**

**Algorithm 2:** Migration simulation

# Backup slides - Dynamic projection - Marriages

1: $P$ - synthetic population
2: **for** $(h, w)$ in [husband ages, wife ages] **do**
3: Get marriage count $N$ for attributes $(h, w)$ {From BFS data}
4: Extract husband candidates $H$ from $P$
5: Extract wife candidates $W$ from $P$
6: Draw $N$ couples from product set $H \times W$
7: Create new households for each couple
8: Change couple marital status to "Married"
9: Adapt modified households
10: **end for**

**Algorithm 3:** Marriages simulation

$P$ - synthetic population
$r$ - official percentage of children in parental house
Extract individuals $C$ from $P$ with age in [15-28]
Extract individuals $C_{\text{parent}}$ from $C$ living in parental house
Compute the current percentage $r_{\text{cur}} = \frac{|C_{\text{parent}}|}{|C|}$
**if** $r_{\text{cur}} > r$ **then**
    $N \leftarrow \lfloor (r_{\text{cur}} - r) \cdot |P| \rfloor$
    Assign weights by age to $C_{\text{parent}}$
    Sample $N$ candidates from $C_{\text{parent}}$ with weights
    **for** each $c$ in candidates **do**
        **if** $c$ has children **then**
            Create a new house with type "Single-parent"
        **else**
            Create a new single household
        **end if**
    **end for**
    Adapt impacted household
**end if**

**Algorithm 4:** Leaving the house simulation