Synthetic Population Projections and Unforeseen Events: Hybrid Simulator for Capturing Dynamics

Marija Kukic Michel Bierlaire

05 February, 2024

Outline

1. Motivation
2. Literature review
3. Contribution
4. Methodology
5. Results: Case study of Switzerland
6. Conclusion and Future Work

Synthetic Population in Transportation: Why?

Real Data

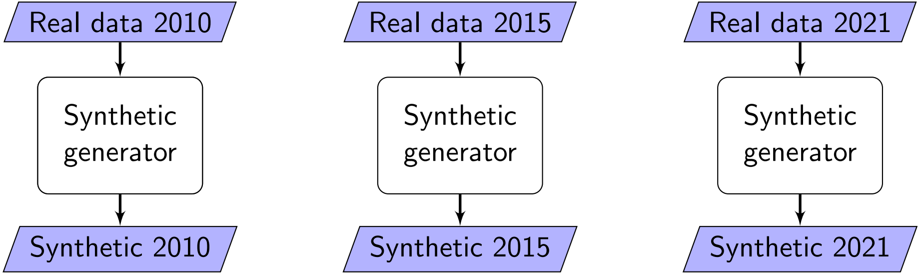
- High cost of data collection.
- Lack of representativity.
- Data privacy constraints.

Synthetic Data

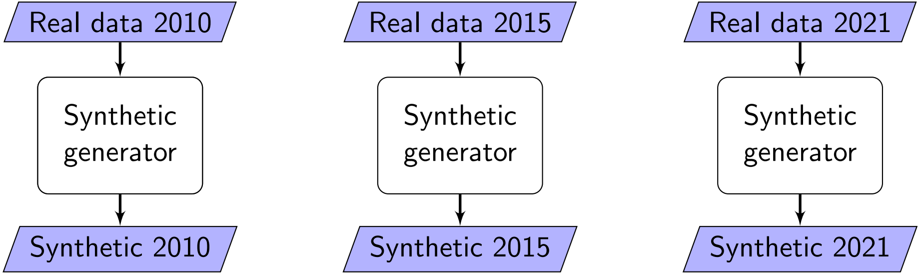
- Open source.
- Bias correction.
- Privacy preservation.

Synthetic Population = tabular data on individuals and households

Snapshot of Synthetic Population: Problems



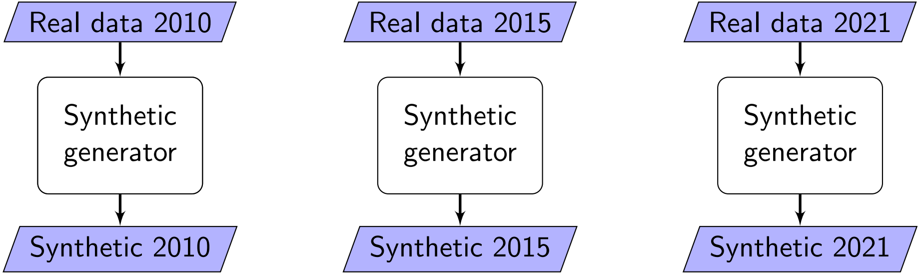
Snapshot of Synthetic Population: Problems



Outdated sample

Snapshot of Synthetic Population: Problems

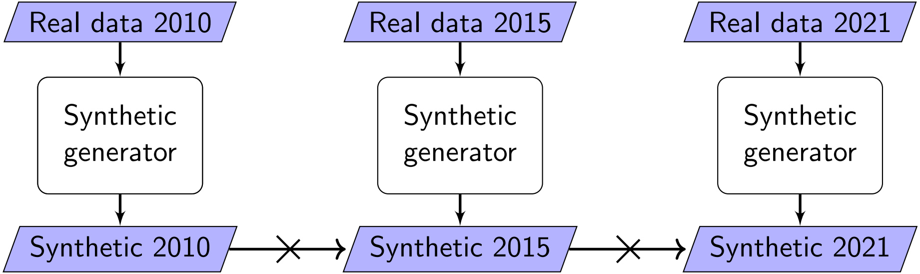
Complicated and Repetitive



Outdated sample

Snapshot of Synthetic Population: Problems

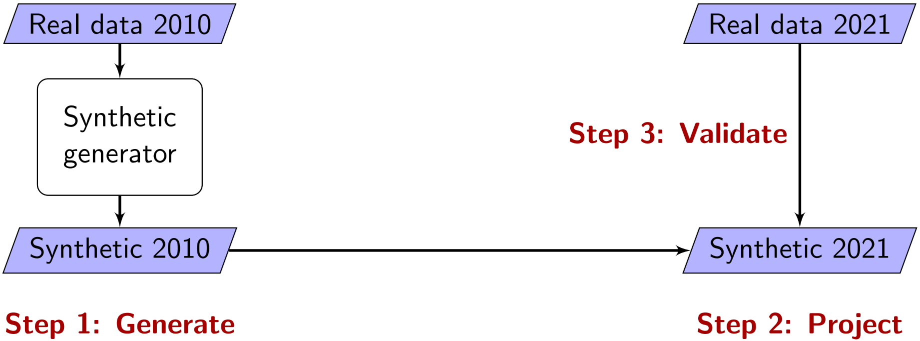
Complicated and Repetitive



Outdated sample

Costly

Synthetic Population Projections



Outline

1 Motivation

2 Literature review

3 Contribution

4 Methodology

5 Results: Case study of Switzerland

6 Conclusion and Future Work

Literature review - Generation and Projection

	Dynamic projection	**Static projection**	**Resampling**
Synthetic reconstruction	Fatmi et al.[1] **2017**	Lomax et al.[2] **2022**	Prédhumeau et al.[3] **2023**
Combinatorial optimisation	Namazi-Rad et el.[4] **2014**	**X**	**X**
Statistical learning	**Hybrid Simulator for Capturing Dynamics Model-driven**	**X**	**Hybrid Simulator for Capturing Dynamics Data-driven**

Literature Gaps

Dynamic projection

- Evolves population.
- Heterogeneous sample.

Re-sampling

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

Literature Gaps

Dynamic projection

- Evolves population.
- Heterogeneous sample.

- Propagation of the generation bias.
- Increase of the error over time.
- **Not robust to the unusual events.**
- **Dependent on input rates.**

Re-sampling

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

- Can achieve a perfect fit.

Literature Gaps

Dynamic projection requires demographic rates to simulate events!

Demographers provide reports on **demographic rates** every five years.

Assumption: Population trends remain **stable** over time.

Problems

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

Unforeseen events can result in projections that **do not represent** the real population.

Affects the **outcomes of transportation models** employing these samples.

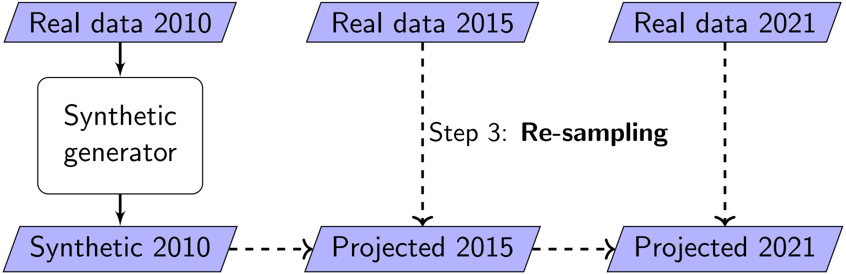
Problematic for **long-term** forecast.

Outline

1. Motivation
2. Literature review
3. Contribution
4. Methodology
5. Results: Case study of Switzerland
6. Conclusion and Future Work

Contribution - Previous work

- Combine **dynamic projection** and **resampling** at the level of **individuals** [5].



Step 1: **Generation** Step 2: **Dynamic Projection** Step 4: **Validation**

Model-based and Data-driven approach

Contribution - Previous work

- Simulate effect of **death, birth, migration** to synthetic individuals described by **age, gender, employment**.

What we showed?

- Maintenance of synthetic samples without regenerating.
- Access to up-to-date data and making use of the past.
- Trade-off between accuracy and efficiency.

Contribution - Current work

- Expand the method from the level of **individuals** to the **household** level.
- Evaluate **robustness** of hybrid simulator to **unforeseen events** (i.e., COVID-19) compared to state-of-the-art methods.

Outline

1 Motivation

2 Literature review

3 Contribution

4 Methodology

5 Results: Case study of Switzerland

6 Conclusion and Future Work

Hybrid Simulator for Capturing Dynamics

Step 1: Generation

Markov Chain Monte Carlo Simulation. [6]

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

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 Office (BFS) [7].

Hybrid Simulator for Capturing Dynamics

Step 3: Re-sampling

When disaggregated data are available.

Compare projected household-type marginals with real data.

Add or delete households to achieve desired fit.

Step 4: Validation

Compare marginal and sub-distributions with real data.

Statistics (e.g., SRMSE) and Visualization.

Evaluate projections to unforeseen events

Test two scenarios:

Pre-pandemic: Using rates from the report from **2010 without knowing** about the pandemic.

Post-pandemic: Using rates from the report from **2021 knowing** about the pandemic.

Goal:

Compare **dynamic projection** and **hybrid simulator** for these two scenarios by projecting samples from **2010** to **2021**.

Outline

1. Motivation
2. Literature review
3. Contribution
4. Methodology
5. **Results: Case study of Switzerland**
6. Conclusion and Future Work

Generation of synthetic sample 2010 - Household level

Reference data: weighted **MTMC 2010, 2015, 2021** [BFS]

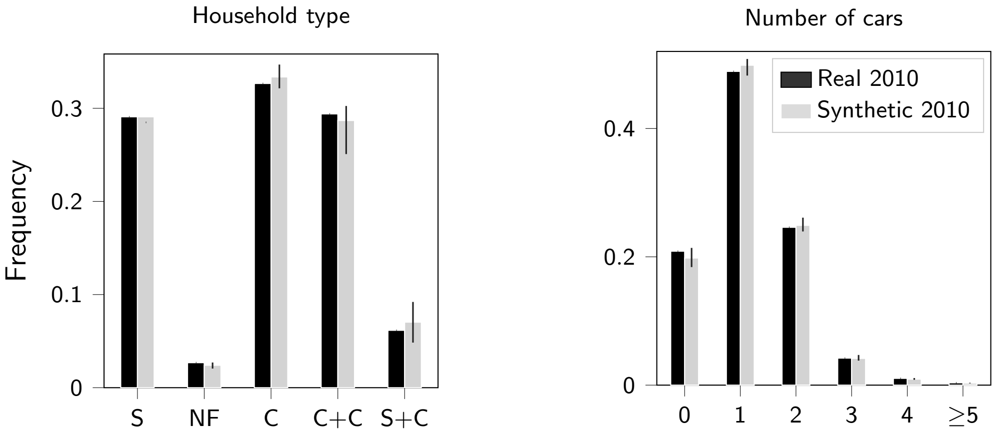
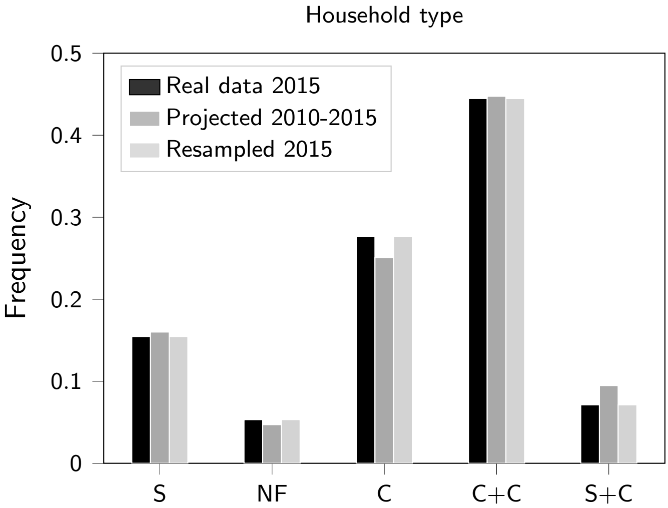


Figure: The comparison of household marginals between synthetic and real sample from 2010

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



Comparison of Dynamic Projection and Hybrid Approach - 2021

	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

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.

	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

Using updated rates leads to better results for both methods.

	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

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**

Comparison of Dynamic Projection and Hybrid approach - 2021

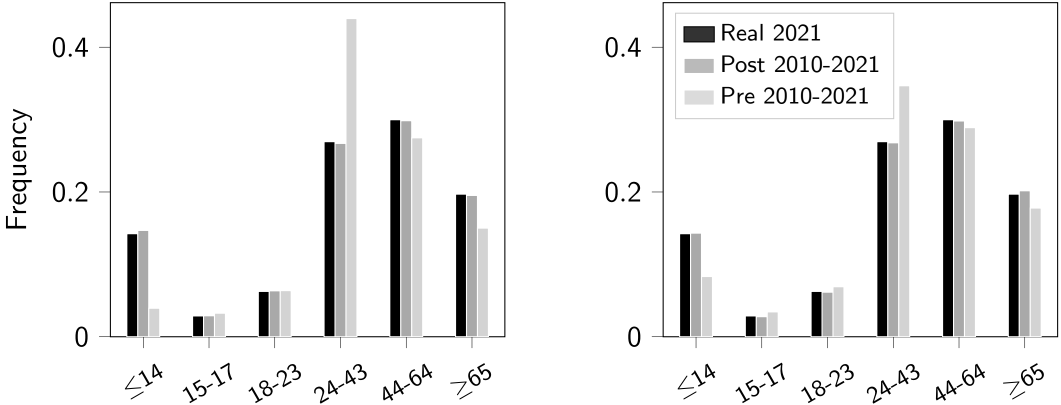


Figure: Marginal distribution of the age using pre and post-pandemic rates compared to the real data - (left) dynamic projection; (right) hybrid simulator

Outline

1. Motivation
2. Literature review
3. Contribution
4. Methodology
5. Results: Case study of Switzerland
6. Conclusion and Future Work

Conclusion and Future Work

We show:

- Resampling step helps reduce the accumulated projection error of dynamic projection.
- The hybrid simulator is more robust to unforeseen events than the dynamic projection.
- The significance of validating and updating synthetic projected samples.

Future work

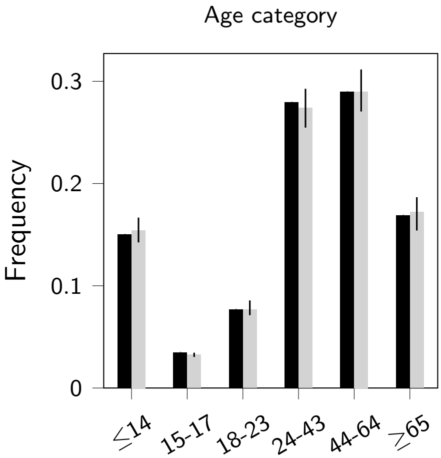
- How to model synthetic individuals over time using Gibbs Sampler?

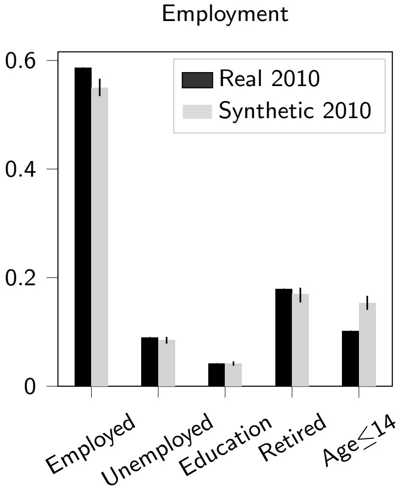
Thank you! Questions?

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

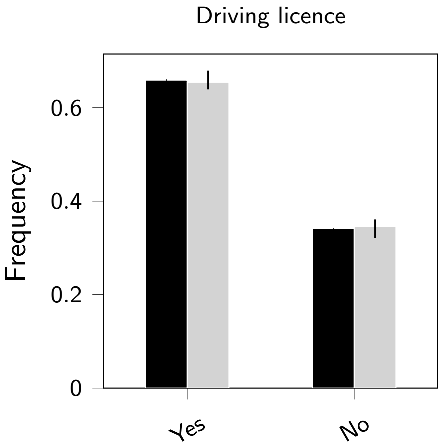
References

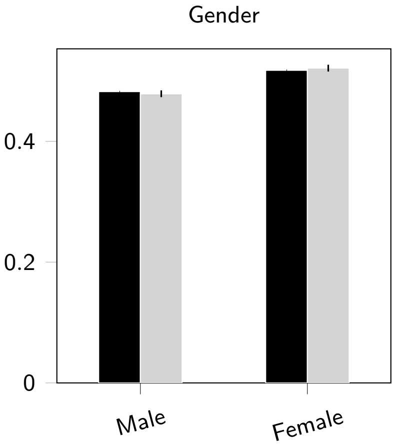
Backup slides



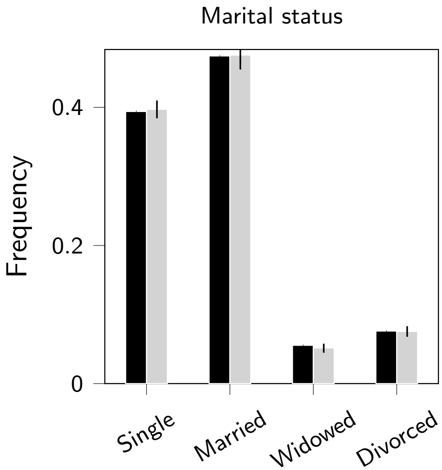


Backup slides





Backup slides



Backup slides - Dynamic projection - Births

Algorithm Births simulation

Require: P - synthetic population

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

Backup slides - Dynamic projection - Migrations

Algorithm Migration simulation

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

Backup slides - Dynamic projection - Marriages

Algorithm Marriages simulation

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

Algorithm Leaving the house 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_{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_{parent}
    Sample N candidates from C_{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

Backup slides - Resampling

Algorithm Resampling procedure

```
Input:
    counts_real - an array of frequency counts per household type in the reference sample
    counts_projected - an array of frequency counts per household type in the projected sample
    list_of_types - an array of existing household types
    num - total number of household types
Function Resample(counts_real, counts_projected, list_of_types, num, projected_sample)
result_sample ← projected_sample
for i ← 1 to num do
    nb_of_observation ← abs(counts_real[i] - counts_projected[i])
    if counts_real[i] - counts_projected[i] < 0 then
        Delete list_of_types[i], nb_of_observation from result_sample
    else
        Add list_of_types[i], nb_of_observation to result_sample
    end if
end for
end Function
```