Hybrid Simulator for Capturing Dynamics of Synthetic Populations

Marija Kukic* Salim Benchelabi Michel Bierlaire

28 September, 2023

EPFL

Outline

Synthetic population in Transportation: Why?

Real Data

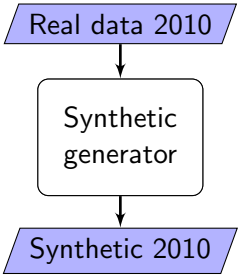
- High cost of data collection.
- Lack of representativity.
- Data privacy constraints.

Synthetic Data

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

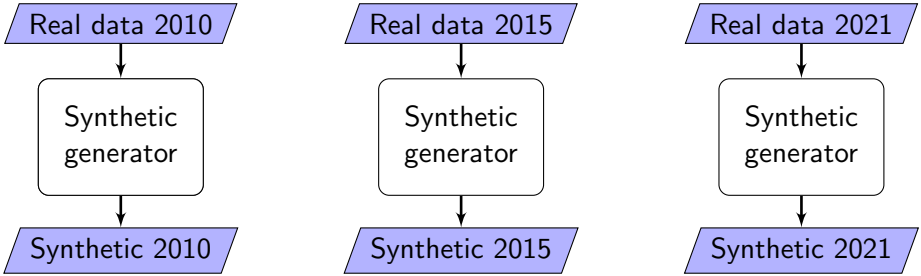
Synthetic Population = tabular data on individuals and households

Existing Generation Methods

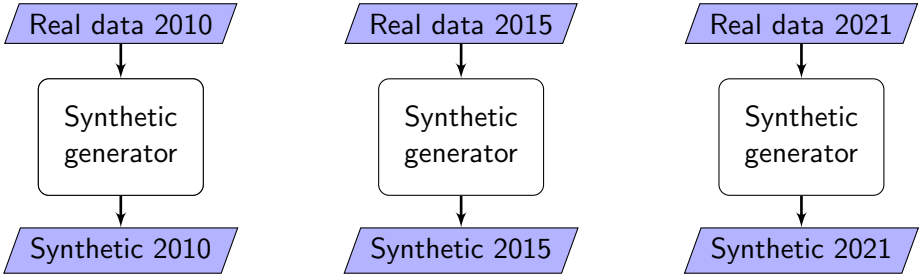


1. Statistical Reconstruction [1, 2]
2. Combinatorial Optimization [3]
3. Statistical learning
   Simulation methods [4, 5]
   Machine Learning methods [6, 7, 8]

Snapshot of Synthetic Population: Problems



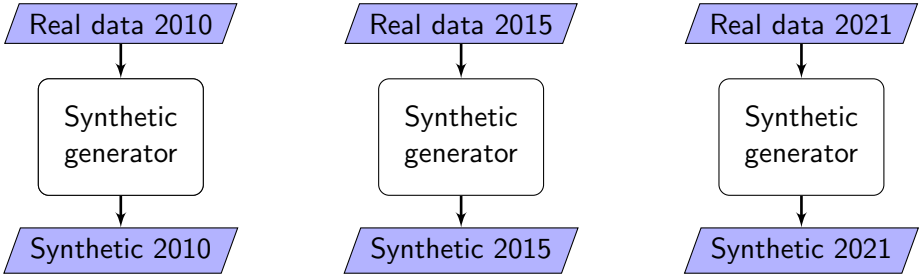
Snapshot of Synthetic Population: Problems



Outdated sample

Snapshot of Synthetic Population: Problems

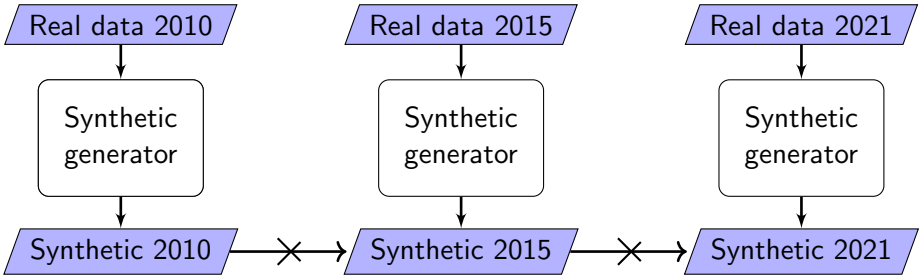
Complicated and Repetitive



Outdated sample

Snapshot of Synthetic Population: Problems

Complicated and Repetitive

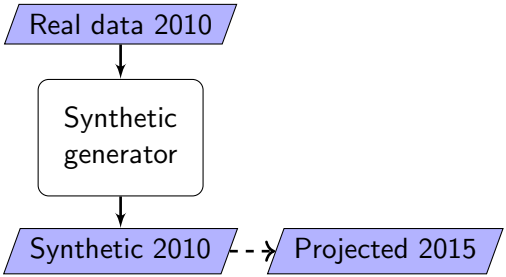


Outdated sample

Costly

Outline

Projection



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

1. **Dynamic Projection**[9, 10]
   Simulate life events
2. **Re-sampling**[11]
   Adjust marginals

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.

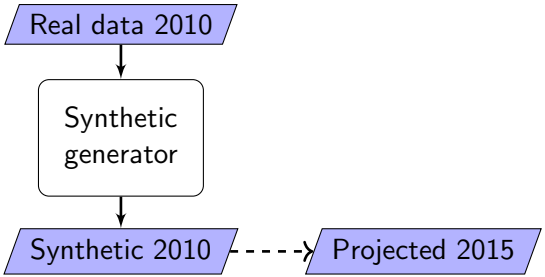
Re-sampling

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

- Can achieve a perfect fit.

Outline

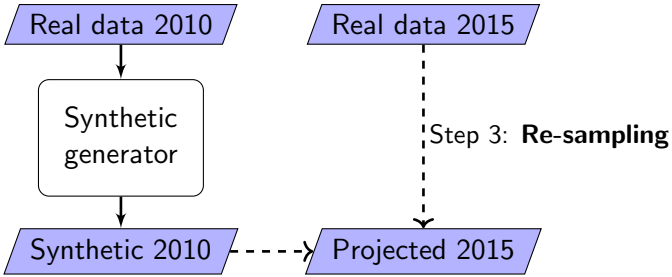
Hybrid Simulator for Capturing Dynamics



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

Model-based approach

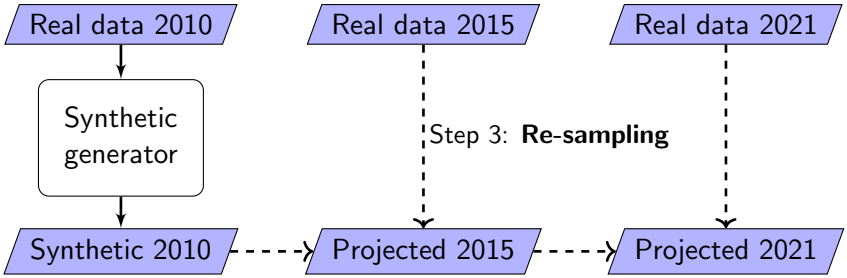
Hybrid Simulator for Capturing Dynamics



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

Model-based and Data-driven approach

Hybrid Simulator for Capturing Dynamics



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

Model-based and Data-driven approach

Outline

Hybrid Simulator for Capturing Dynamics

Step 1: Generation

Markov Chain Monte Carlo Simulation. [5]

Synthetic individuals $X = (X_{\text{age}}, X_{\text{emp}}, X_{\text{gender}})$.

Bootstrap and convergence monitoring.

Step 2: Dynamic projection

When disaggregated data are not available.

Simulate events: **birth, death and migration**.

Simulate impact on **age, gender and employment**.

Hybrid Simulator for Capturing Dynamics

Step 3: Re-sampling

When disaggregated data are available.

Compare age marginals with real most recent data.

Add or delete individuals to achieve desired fit.

Step 4: Validation

Compare marginal and sub-distributions with real data.

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

Outline

Generation and validation of synthetic sample - 2010

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

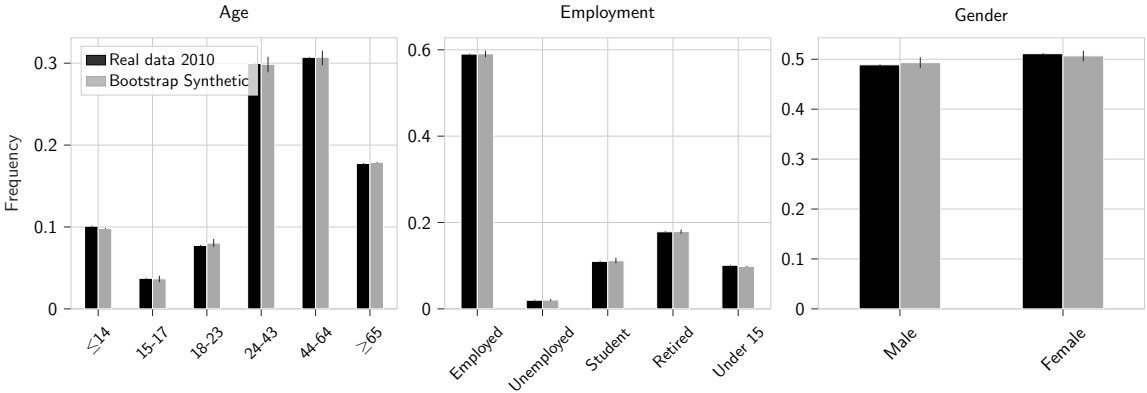


Figure: The comparison of the marginal distributions between synthetic and real sample from 2010

Dynamic Projection (2010 - 2014) and Re-sampling (2015)

Rates on **birth, death and migration** [OFS]

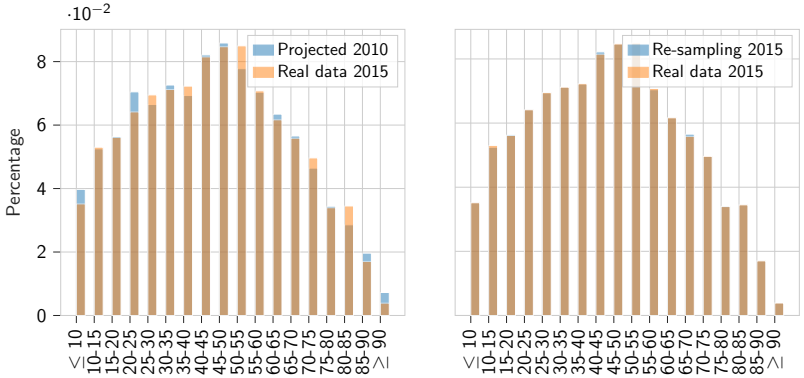


Figure: Comparison with real data 2015
(Left - **Dynamic Projection** results; Right - **Re-sampling** results)

Comparison of projection and hybrid approach - 2021

	Age $\cdot 10^{-2}$	Employment $\cdot 10^{-2}$	Gender $\cdot 10^{-2}$	Average $\cdot 10^{-2}$
Hybrid approach **2010 - 2021**	**7.35**	**5.26**	**0.61**	**4.41**
Dynamic Projection **2010 - 2021**	8.28	7.13	0.67	5.36

Table: SRMSE of projected samples against real sample 2021

Comparison of Dynamic Projection and Hybrid approach - 2021

	Age $\cdot 10^{-2}$	Employment $\cdot 10^{-2}$	Gender $\cdot 10^{-2}$	Average $\cdot 10^{-2}$
Dynamic Projection **2015 to 2021**	**5.76**	**3.71**	**0.48**	**3.31**
Hybrid approach **2010 - 2021**	7.35	5.26	0.61	4.41
Dynamic Projection **2010 - 2021**	8.28	7.13	0.67	5.36

Table: SRMSE of projected samples against real sample 2021

Outline

Conclusion and Future Work

Hybrid approach provides:

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

Future work

- Expand from individuals to households.
- Re-sample other attributes than age.
- Comparison with re-generation based on several criteria.

Thank you! Questions?

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

References

Backup slides

Problems of projection methods

- Arbitrarily chosen generators.
- Limited number of considered attributes.
- Lack of validation.

Backup slides

	Age $\cdot 10^{-2}$	Employment $\cdot 10^{-2}$	Gender $\cdot 10^{-2}$	Average $\cdot 10^{-2}$
Hybrid simulator 2010 - 2021	7.35	5.26	0.61	4.41
Dynamic projection 2010 - 2021	8.28	7.13	0.67	5.36
Re-sampling 2010 - 2021	**1.69**	**4.02**	**1.76**	**2.49**

Table: First order SRMSE between the real sample from 2021 and projected synthetic samples

Backup slides

	Age, Employment, Gender
Hybrid simulator 2010-2021	0.2
Resampling 2010-2021	0.33

Table: Third order SRMSE between the real sample from 2021 and projected synthetic samples

Backup slides

Algorithm 1 Dynamic projection

```
function DYNAMIC_PROJECTION(synthetic_sample, t_0, t_end)
    predictive_sample = synthetic_sample
    for i = t_0 to t_end do
        increment_age(predictive_sample);
        add_children(predictive_sample, i);            ▷ Birth rates
        remove_individuals(predictive_sample, i);      ▷ Death rates
        add_individuals(predictive_sample, i);         ▷ Migration rates
        remove_individuals(predictive_sample,i);
    end for
    draw_employment(predictive_sample);
end function                                           ▷ Return the updated sample
```

Backup slides

Algorithm 2 Resampling procedure

```
1:  function RESAMPLE(a,b,num,threshold)
2:      a - array of frequency counts per each age category in reference sample
3:      b - array of frequency counts per each age category in projected sample
4:      num - total number of age categories
5:      for i = 1 to num do
6:          if abs(a[i] − b[i]) > threshold then
7:              if (a[i] − b[i]) < 0 then
8:                  nb_of_observation = abs(a[i] − b[i])
9:                  for j = 1 to nb_of_observation do
10:                     randomly sample a person of the age i
11:                     remove a selected person from the projected sample
12:                 end for
13:             else
14:                 nb_of_observation = abs(a[i] − b[i])
15:                 for j = 1 to nb_of_observation do
16:                     randomly sample a person of the age i
17:                     add the selected person to the projected sample
18:                 end for
19:             end if
20:         end if
21:     end for
22: end function                                  ▷ Return the updated sample
```