Population synthesis at the level of households

Marija Kukic
Michel Bierlaire

14.09.2021.

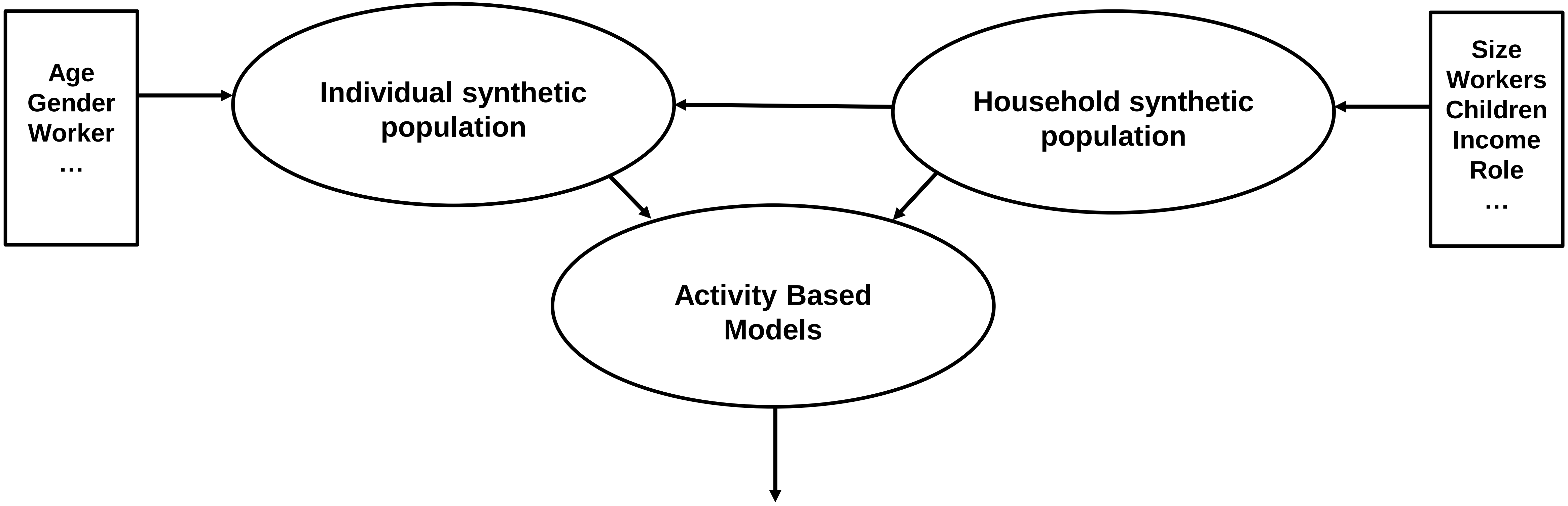
STRC | 21th Swiss Transport Research Conference
Monte Verità / Ascona, September 12 – 14, 2021

Outline

- Motivation
- Literature review
- Simulation approach for synthetic generation
- Synthetic households imputation
- Case study
- Future work

EPFL TRANSP-OR

Motivation: Activity based models and synthetic population



Data about individuals	Data about households
Discrete trips	Overall behavioral patterns
Decision of isolated individual	Decision at household level
Intersections are not analyzed	Dependencies, sharing resources and intersection constraints

Literature review: From individuals to households

	GENERATION OF INDIVIDUALS	GENERATION OF HOUSEHOLDS	ASSOCIATIONS BETWEEN INDIVIDUALS & HOUSHEOLDS
Iterative Proportional Fitting (IPF)	**1996** *Beckman et al*. Creating synthetic baseline populations	**2007** *Arentze et al*. Creating synthetic household populations	**2009** *Ye et al*. Iterative Proportional Updating
Simulation techniques (MCMC)	**2013** *Farooq et al.* Simulation based population synthesis		**2014**, *Anderson et al.*, Associations Generation **2015,** *Casati et al.*, Hierarchical MCMC
Machine Learning techniques	**2014,** *Goodfellow et al.* Generative Adversarial Networks **2018,** *Xu et al.* Tabular Generative Adversarial Networks **2019,** *Borysov et al.,* Variational Autoencoder **2020,** *Badu – Marfo et al.,* Composite Travel Generative Adversarial Neworks		**2021** **…**

Literature review: Synthetic population of households

	SAMPLE FREE	SAMPLE BASED
TWO – STAGE PROCESS	hMCMC	X
ONE – STAGE PROCESS	?	IPU

Literature review: Gaps and research questions

GENERATION OF INDIVIDUALS	GENERATION OF HOUSEHOLDS	ASSOCIATIONS BETWEEN INDIVIDUALS & HOUSHEOLDS

1. **How to design sample free methodology for creation of synthetic households in one – stage process?**
2. **How much control we can embed into generation process?**
3. **Do the existing state-of-the-art methodologies generate a consistent synthetic population?**

Simulation approach for synthetic population: existing approach

Simulation based population synthesis:

- Markov Chain Monte Carlo process

Sampling methods:

- Gibbs Sampling

Input preparation:

1. Conditional distributions constructed from:
   Data
   Models
   Assumptions

Assumptions:

- Given A, B is uniform across C,D

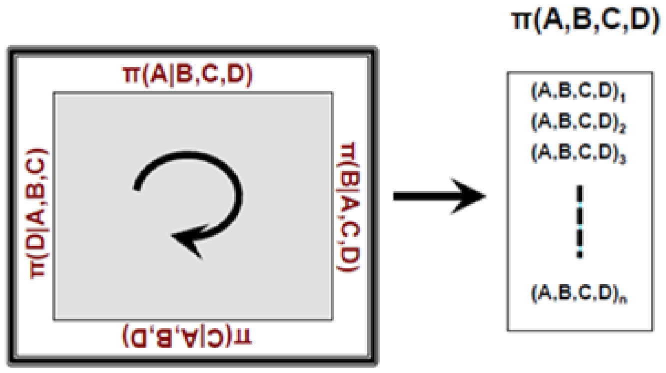
$\pi(A|B) = \pi(A|B,C,D)$

$\pi(A|B)$

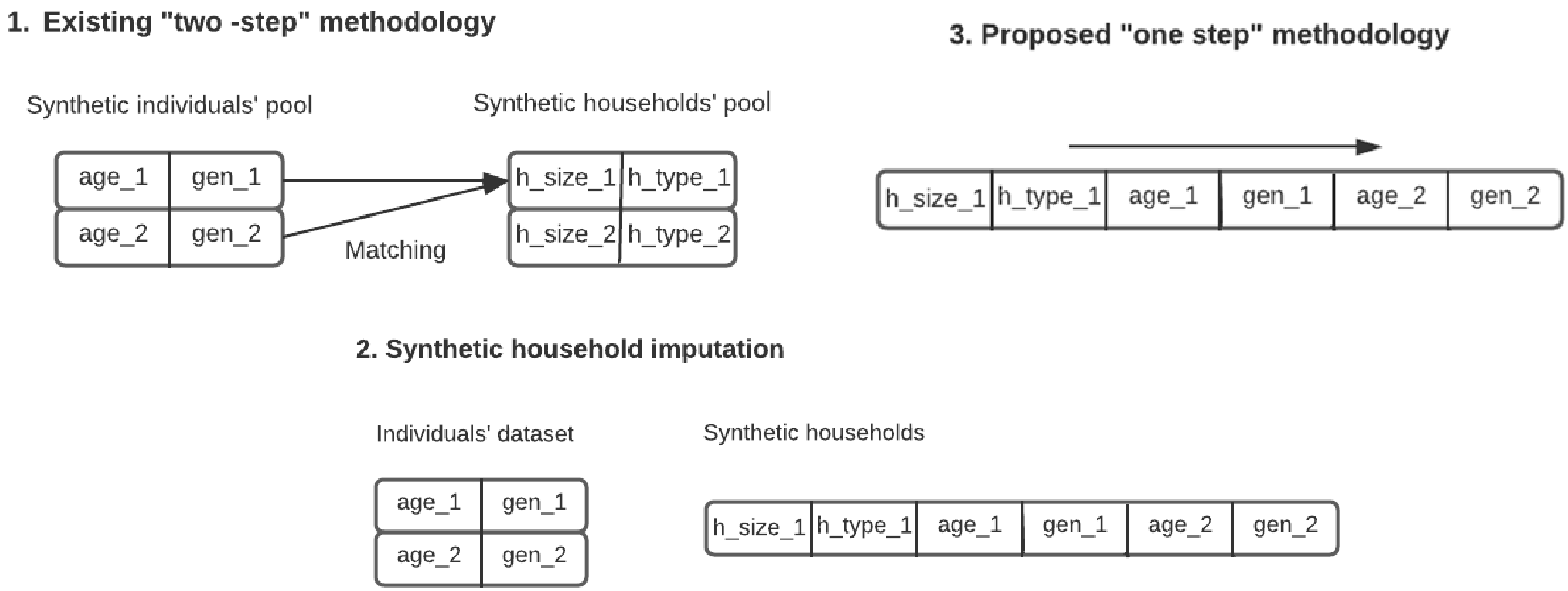
Age	Gender: Male	Gender: Female	Total	Target
0 to 16	11057	4069	15126	15012
17 to 25	21228	8335	29563	29567
26 to 55	6415	13762	20177	20234
56 and above	11209	23925	35134	35187
Total	49909	49932		
Target	50091	50155		
Total 0–25	32285	12404		
Target 0–25	32144	12435		

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

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

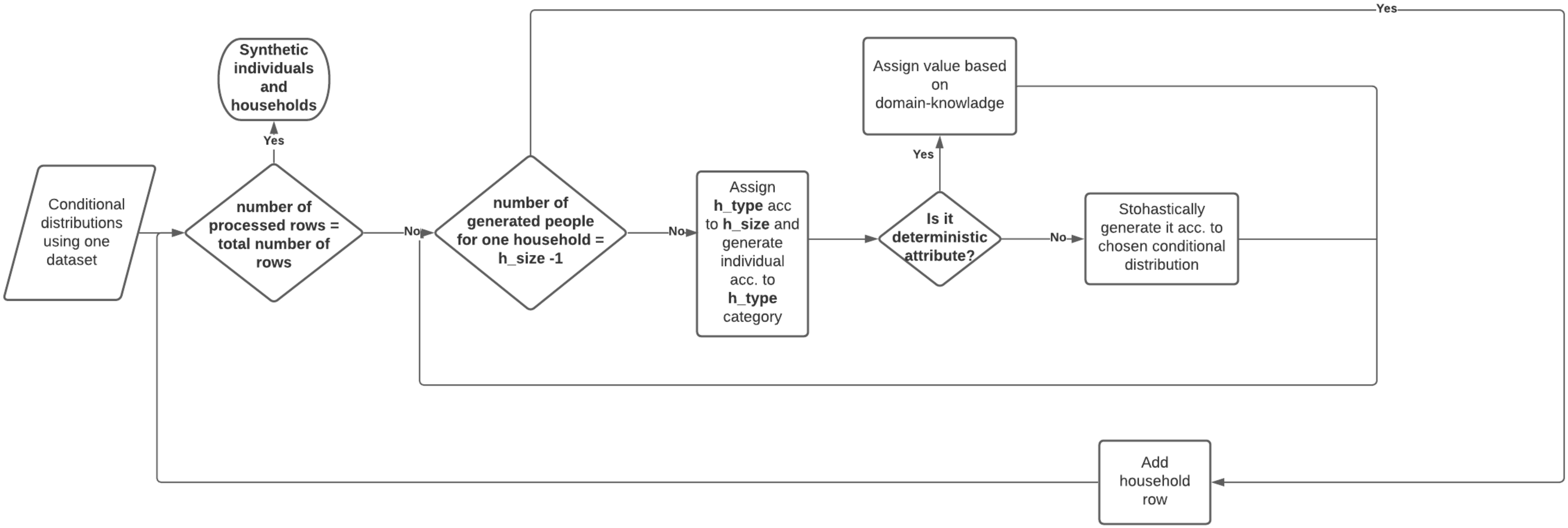


Simulation approach for synthetic population: contribution

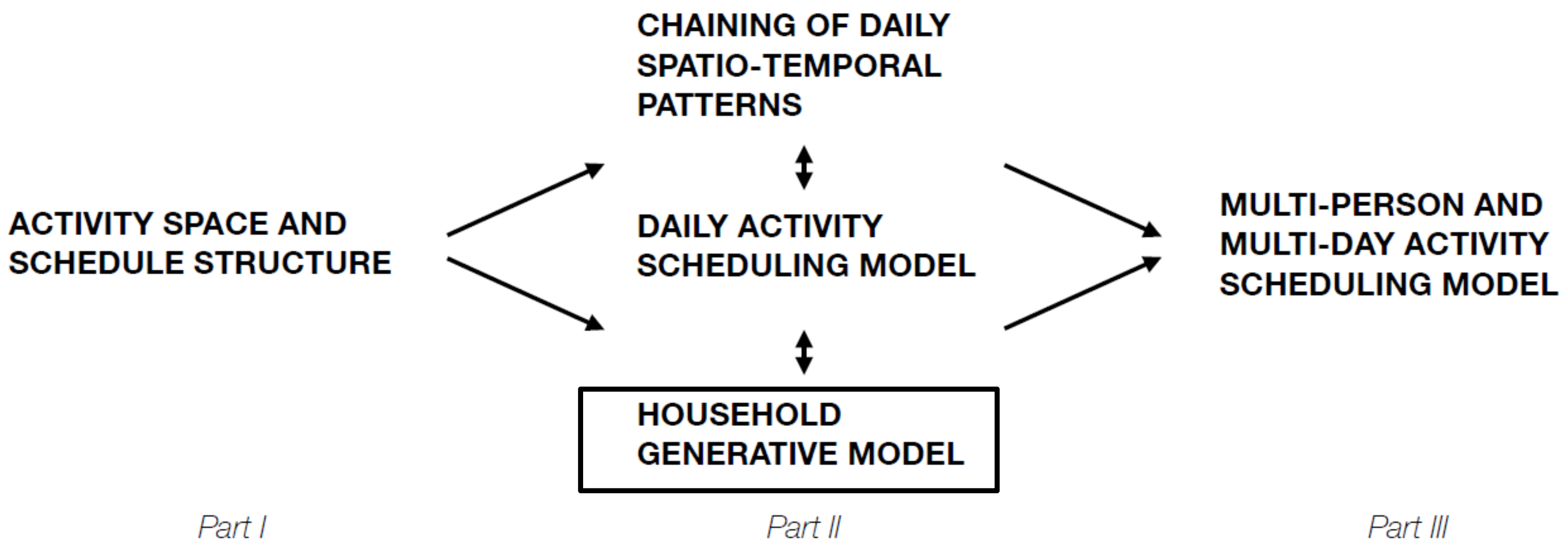


Synthetic households imputation: algorithm

- Household types: **single, couple, couple+children, single+children, non-family**
- Types of attributes: **deterministic** and **stochastically** assigned



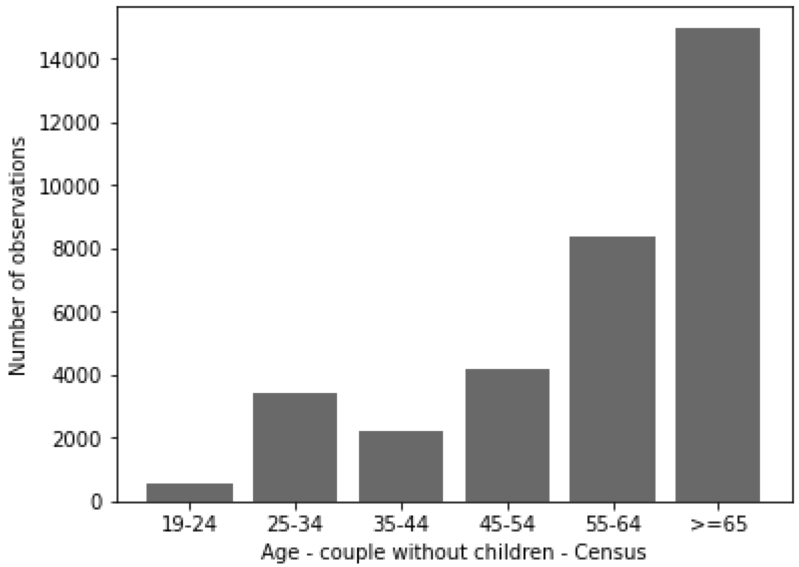
Case study: Multiday Activity Patterns and Schedules Owners

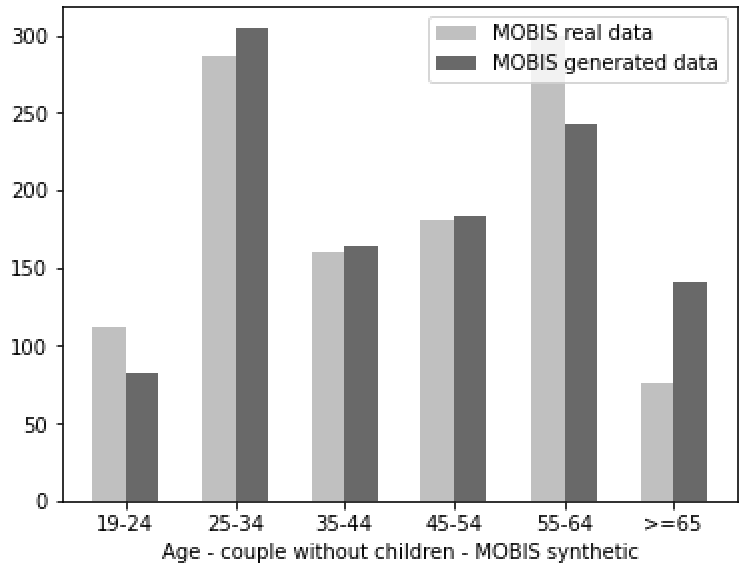


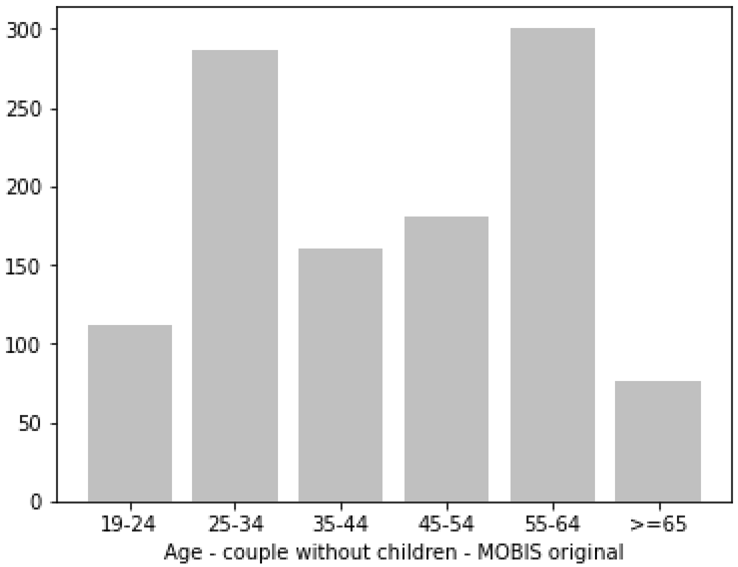
Case study: MOBIS and census datasets

	Synthetic dataset
Number of observations	10736 agents 3700 households
Area	Switzerland
Individual attributes	Age Gender Educational level Employment Income
Household attributes	Household size Owning car Household type Household role Number of children Language

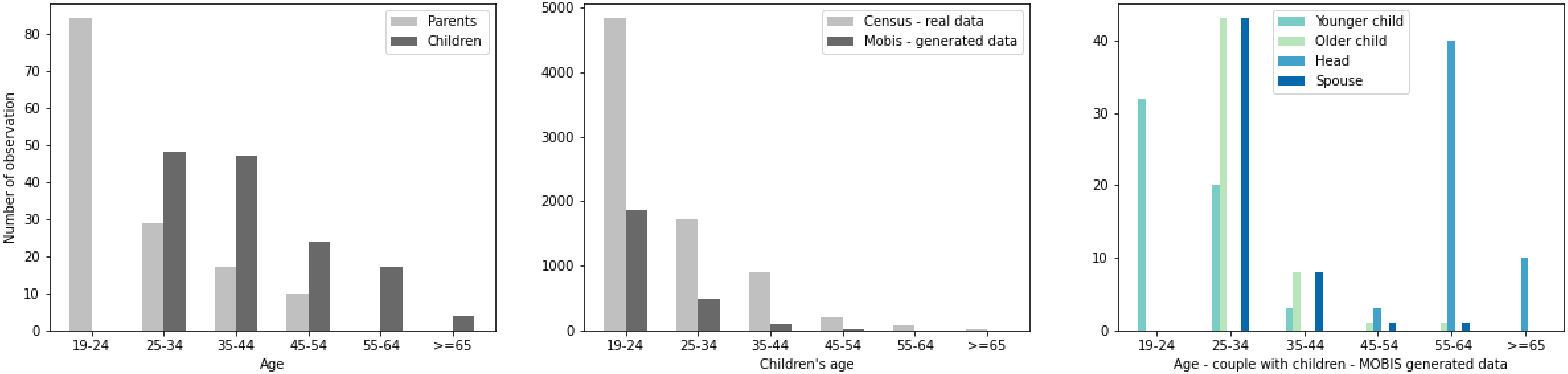
Results: Before and after imputation – MOBIS & census characteristics







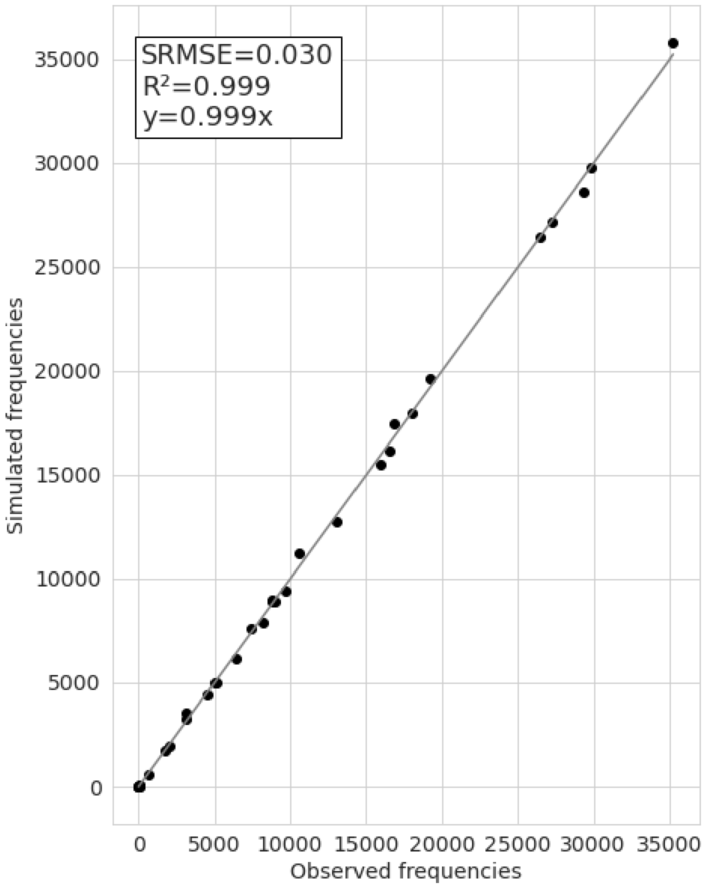
Results: Consistency



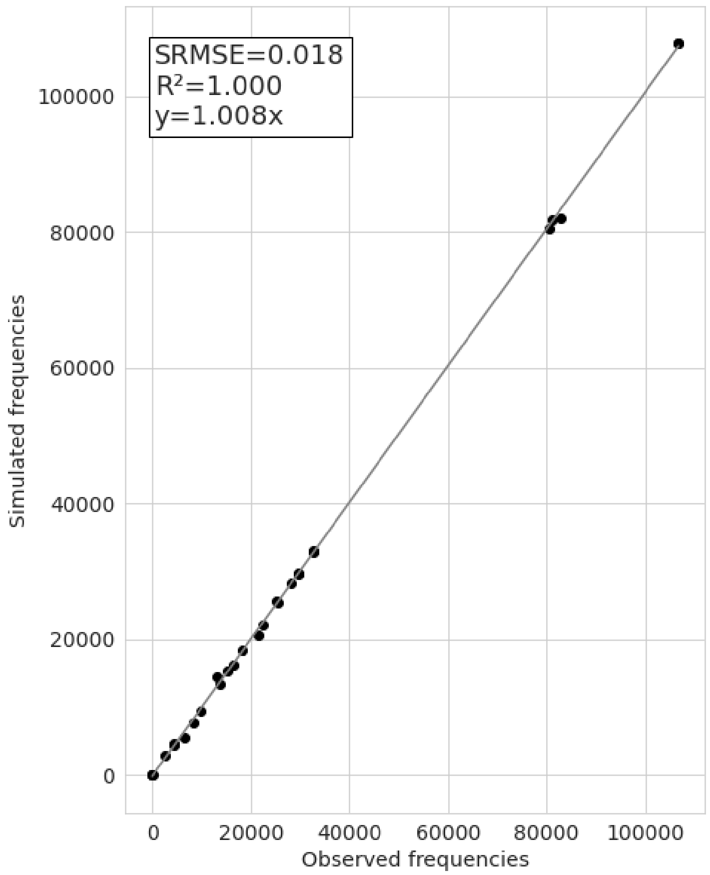
Case study: Goodness of fit – representativity

Standardized Root Mean Square Error

Results of generation – individuals and households

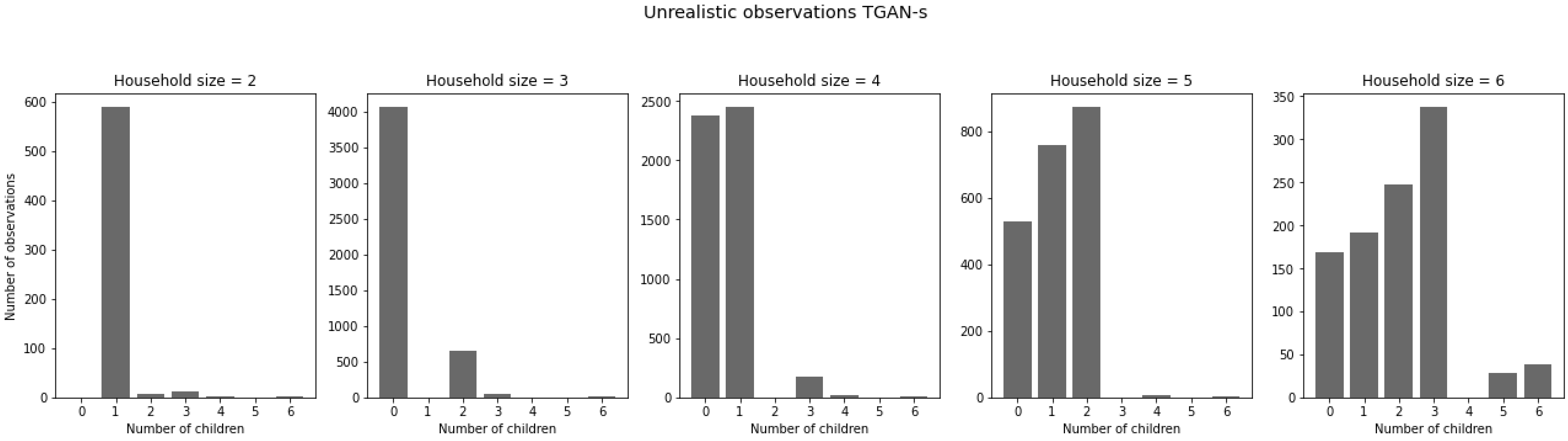


TGANs – individual dataset



TGANs – household dataset

Case study: Validation of consistency and realism



Case study: Is a consistency validated?

Standardized Root Mean Square Error – Does it validate multivariate distributions?

$$SRSME = \frac{\left[\sum_{i=1}^{m} \cdots \sum_{j=1}^{n} (R_{i..j} - T_{i..j})^2 / N\right]^{1/2}}{\sum_{i=1}^{m} \cdots \sum_{j=1}^{n} (T_{i..j}) / N}$$

Age : 0 – young, 1 – adult, 2 – old
Employment: 0 – school, 1 – employed, 2 - retired

AGE	EMPLOYMENT
0	0
1	1
2	2

Real dataset

AGE	EMPLOYMENT
0	2
1	0
2	1

Synthetic dataset

SRMSE = 0 => Synthetic columns values fit perfectly => Synthetic observations are unrealistic

Conclusion

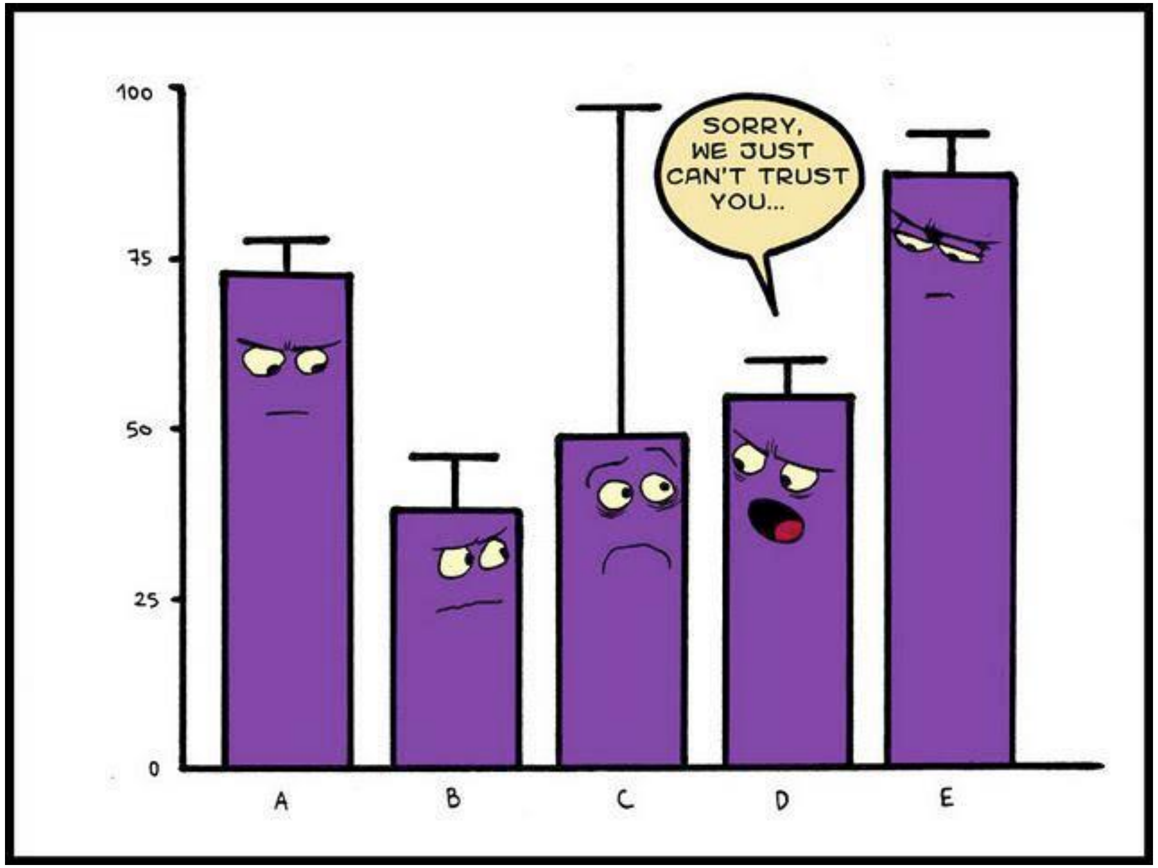
- Control can be embedded into generation process – consistency preserved
- Curse of dimensionality with complete generation

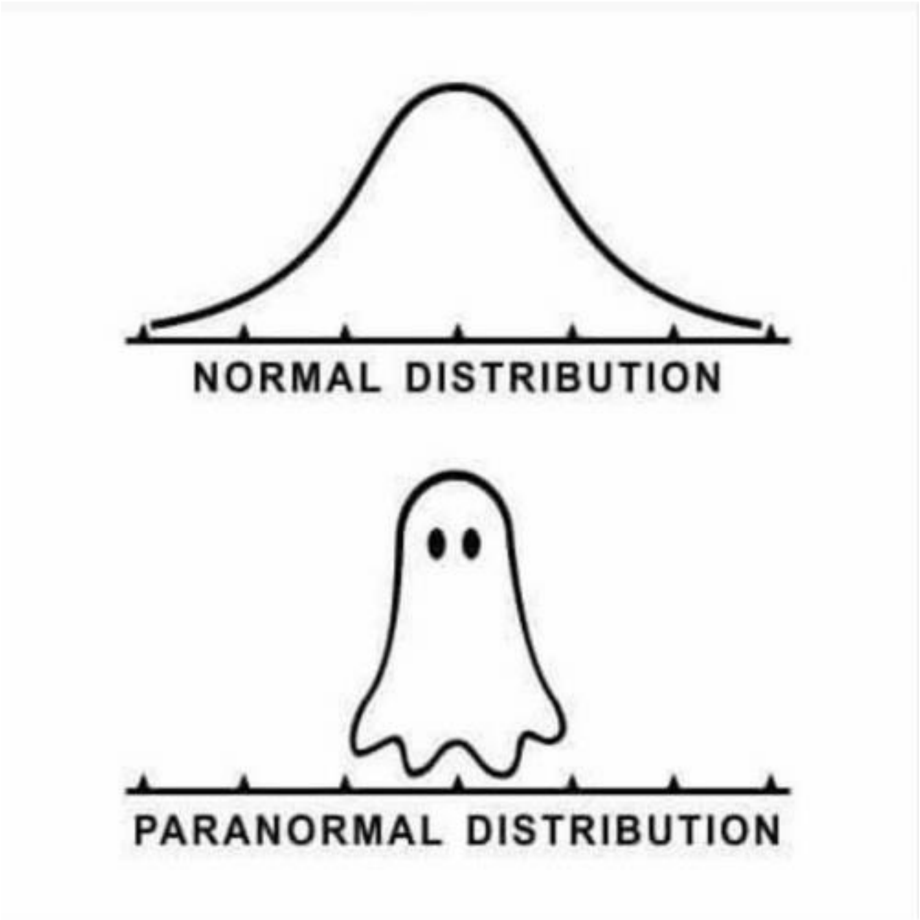
Future work

- From synthetic imputation to synthetic generator of households in one step – simulation or ML?
- Validation techniques for estimation of multivariate distributions

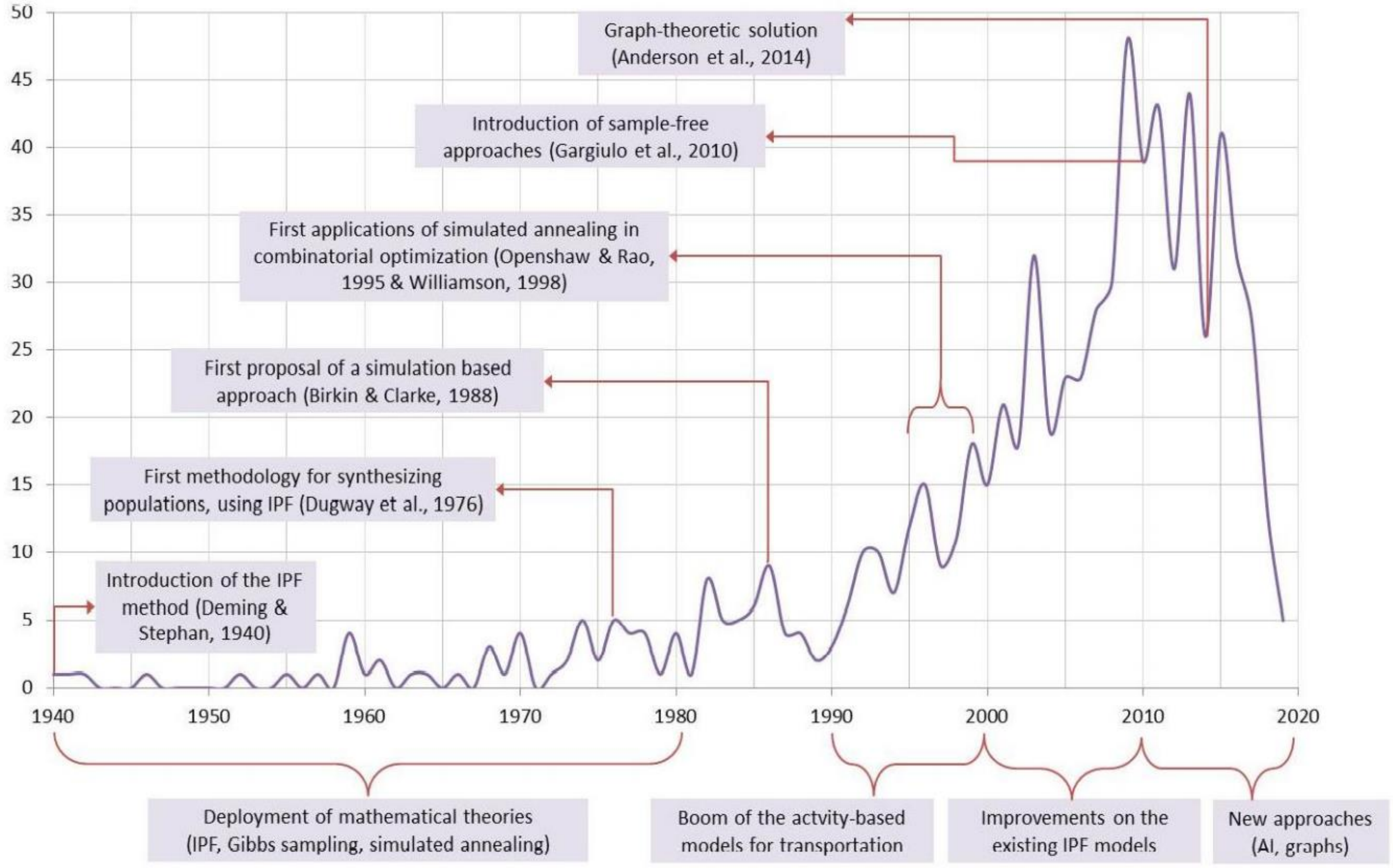
Q&A?

Thanks for your attention!





Appendix: Population Synthesis in transportation



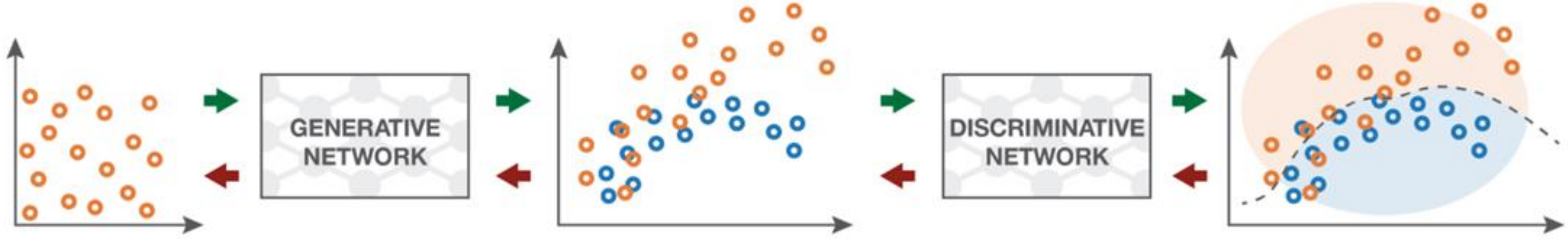
Case study: Comparison with TGANS

Generative adversarial network (GANs):

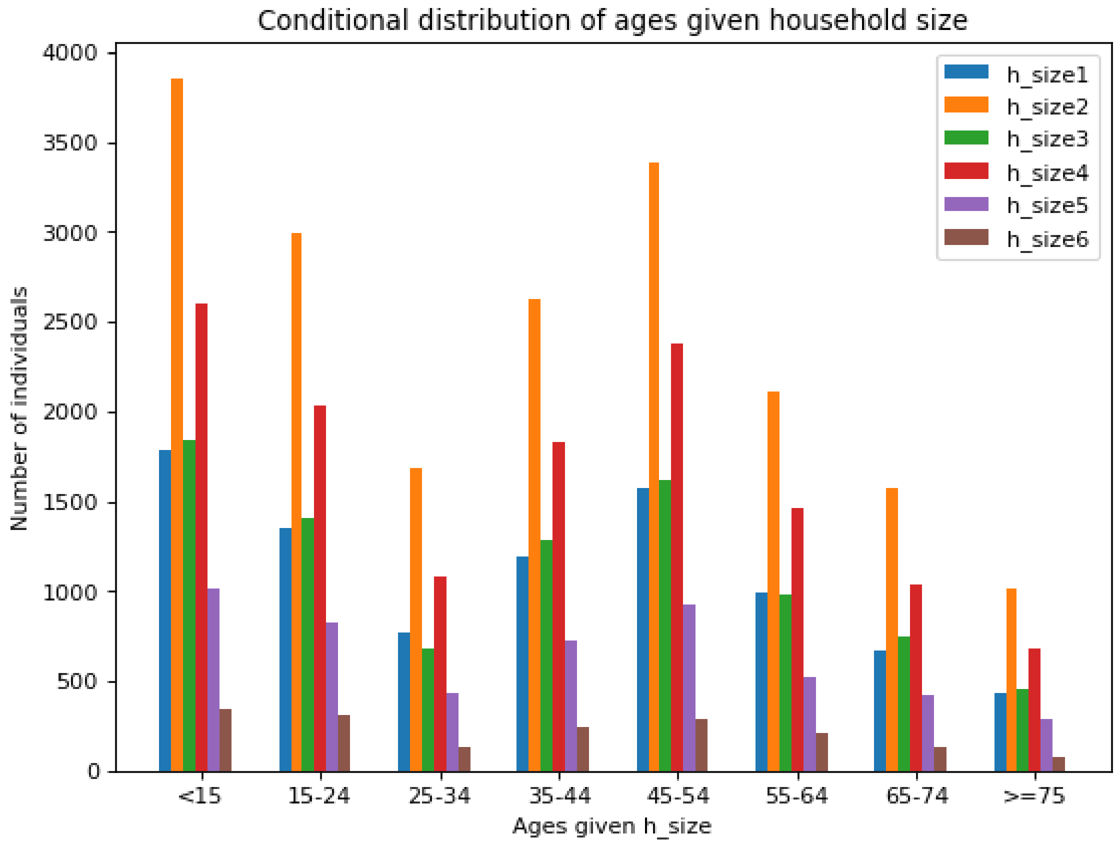
- Learn the probability distribution and draw samples from the distribution

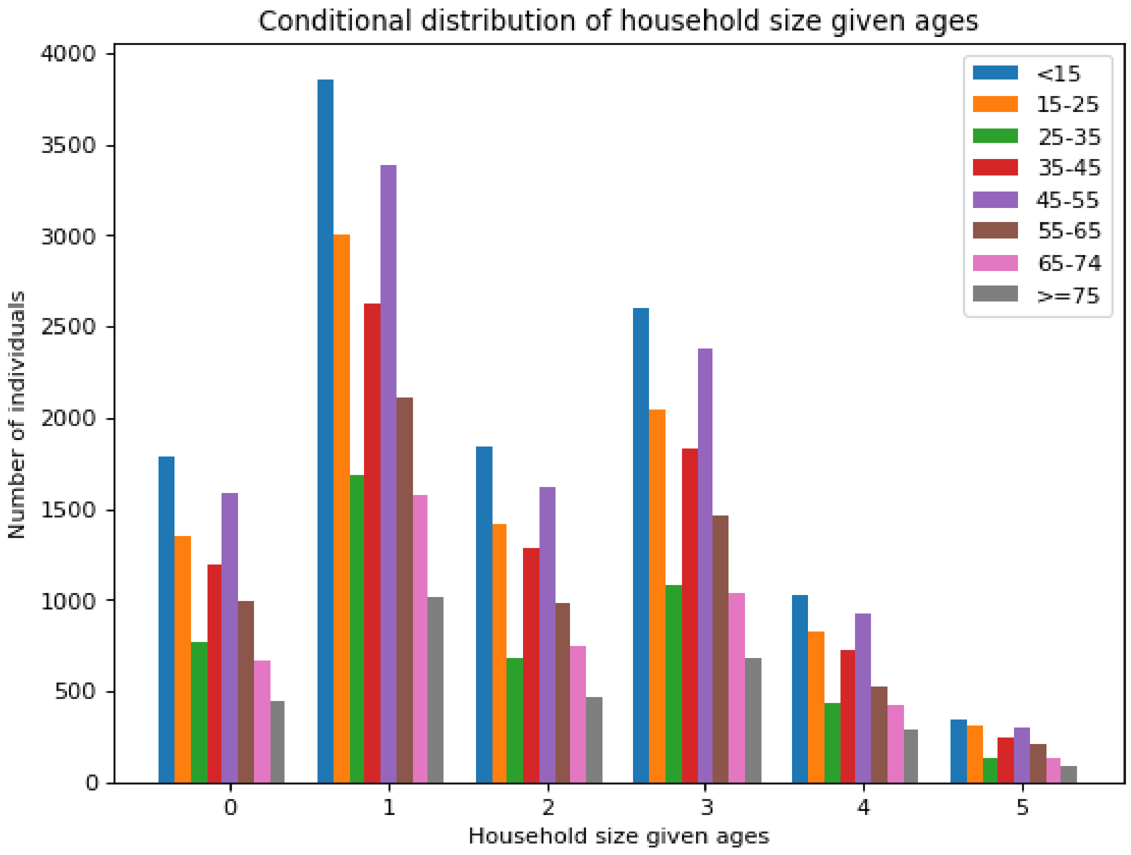
Tabular generative adversarial network (TGANs)

- Synthetic data generator based on GANs for tabular data

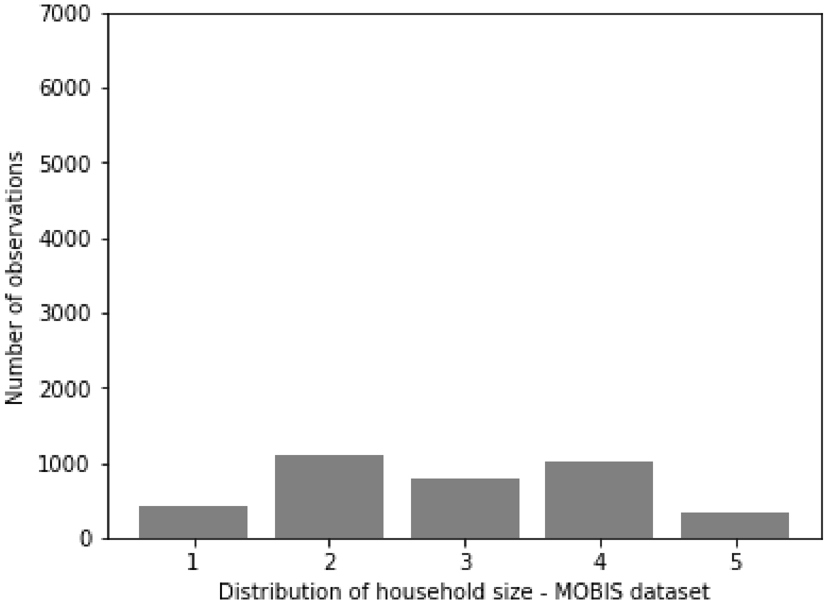


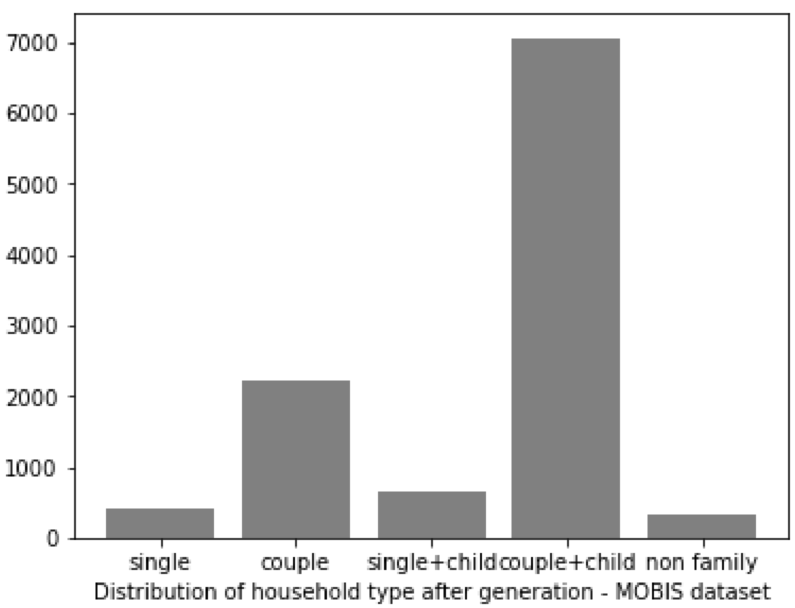
Case study: construction of conditional distributions

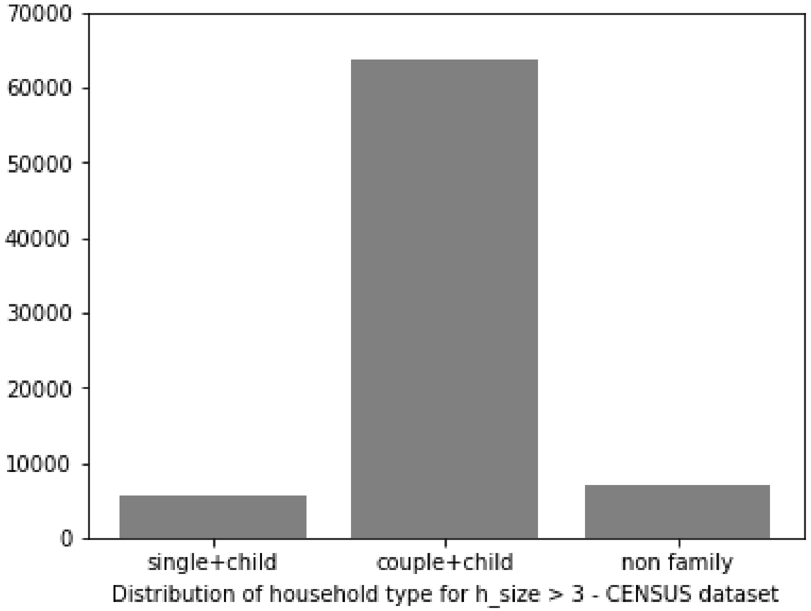




Results: Discrete and stochastic generation of attributes

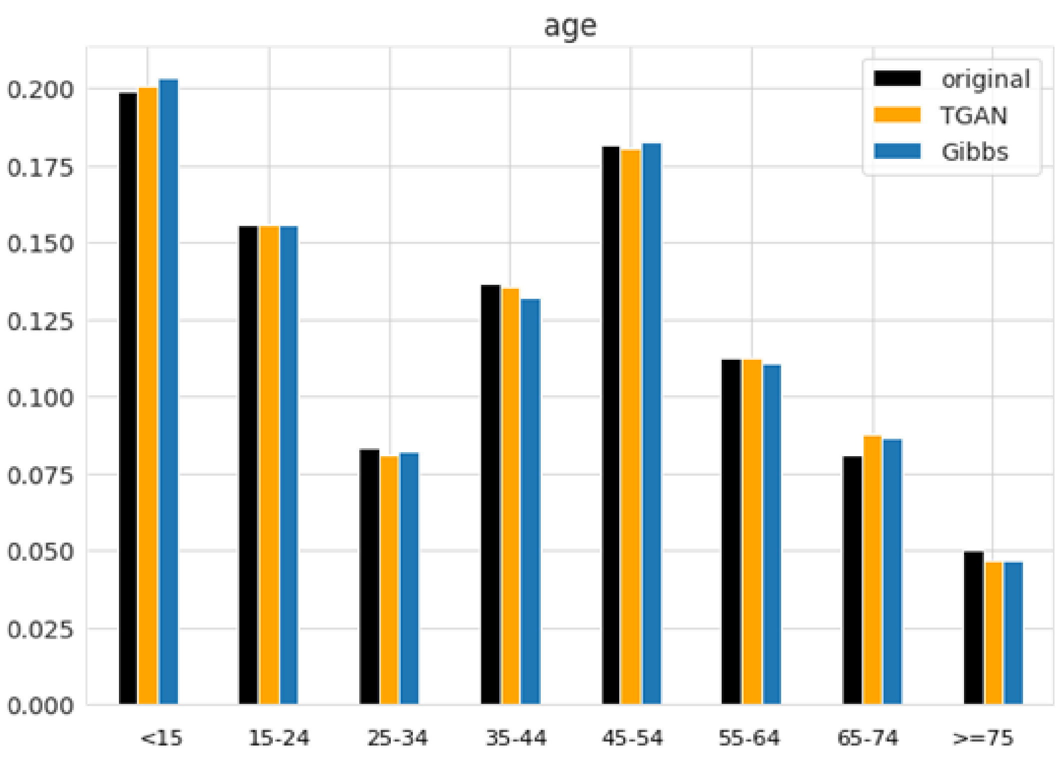


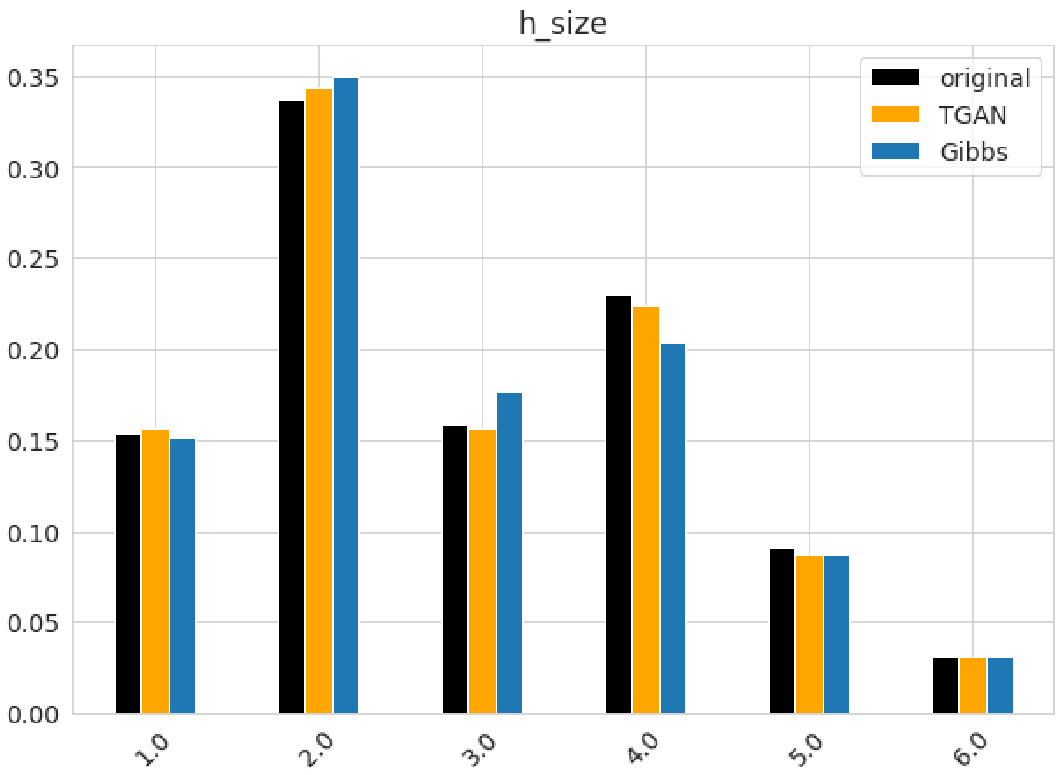




Case study: 2015 census data – Comparison with TGANS

Results of generation – individuals and households





Appendix

Algorithm 1: Household imputation

Data: X_{given}= (x^{age}_{given},x^{size}_{given},...,x^{n}_{given}) - the chosen row from the referenced dataset
n - number of the attributes of each individual
N - number of the individuals in referenced dataset
k - number of the processed rows
i - number of synthetic people in household
$\pi(X_i|X_j)$ - conditional distributions formed according to another dataset
Result: N * ($x^{size}_{given} - 1$) synthetic people grouped into N synthetic household

```
k ← 0
while k ≠ N do
    i ← 0
    while i < x_size do
        initialize synthetic individual X_i= (x_i^age, x_i^size, ..., x_i^n)
        if x_given^size = 1 then
            x_i^type ← single;
            x_i^role ← head;
            X_i = X_given
        else if x_given^size = 2 then
            generate_partner();
        else
            draw x_i^type following π( X_k^type | X_given^size > 2);
            if x_i^type = couple with children then
                generate_partner();
                generate_children();
            else if x_i^type = single parent with children then
                generate_children();
            else
                generate_person();
            end
            i ← i + 1;
            k ← k + 1;
        end
    end
end
```

Appendix

Algorithm 2: Generate partner

Data: X_{given}= (x^{age}_{given},x^{size}_{given},...,x^{n}_{given}) - the chosen row from the referenced dataset

n - number of the attributes of each individual

$\pi(X_i|X_j)$ - conditional distributions formed according to another dataset

Result: synthetic partner $X_k = (x^k_{age}, x^k_{size}, ..., x^k_n)$, k = 1

initialize X_k

if $x^{given}_{size} = 2$ **then**

  $x^{type}_k \leftarrow$ couple without children;

else

  $x^{type}_k \leftarrow$ couple with children;

end

$x^{language}_k = x^{language}_{given}$;

$x^{size}_k = x^{size}_{given}$;

$x^{car}_k = x^{car}_{given}$;

Generate x^{age}_k, x^{gender}_k,$x^{employment}_k$,$x^{education}_k$,x^{income}_k using Inverse Transform on chosen conditional distribution $\pi(X_i|X_j = x_{given})$;

if $x^{age}_k > x^{age}_{given}$ **then**

  $x^{role}_k \leftarrow$ head;

else

  $x^{role}_k \leftarrow$ spouse;

end

Appendix

Algorithm 3: Generate children

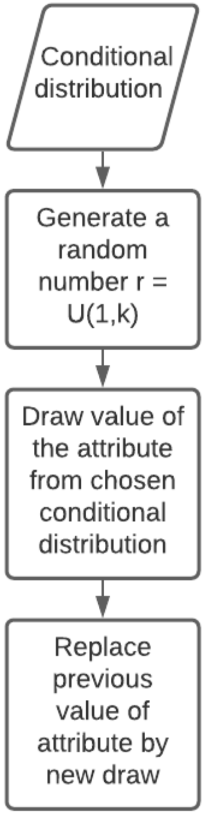
Data: X_{given}= $(x_{given}^{age}, x_{given}^{size}, ..., x_{given}^{n})$ - the chosen row from the referenced dataset
$\pi(X_i|X_j)$ - conditional distributions formed according to another dataset
Result: synthetic children $X_k = (x_{age}^k, x_{size}^k, ..., x_n^k)$
initialize X_k
$x_k^{type} \leftarrow$ couple with children; $x_k^{language} = x_{given}^{language}$;
$x_k^{size} = x_{given}^{size}$;
$x_k^{car} = x_{given}^{car}$;
$x_k^{role} \leftarrow$ child;
Generate x_k^{gender} draw from marginal distribution $\pi(X^{gender})$;
if $first\_child = True$ **then**
  Generate x_k^{age} using Inverse Transform on $\pi(X^{age\_child}|X^{age\_parent} = x_{age\_of\_younger\_parent})$;
else
  Generate x_k^{age} using Inverse Transform on $\pi(X^{age\_child}|X^{age\_parent} = x_{age\_of\_older\_sibiling})$;
end
Generate $x_k^{education}$ using Inverse Transform on $\pi(X^{education}|X^{age} = x_k^{age})$;
Generate $x_k^{employment}$ using Inverse Transform on $\pi(X^{employment}|X^{education} = x_k^{education})$;
Generate x_k^{income} using Inverse Transform on $\pi(X^{income}|X^{employment} = x_k^{employment})$;

Appendix

Data:
$\pi(X^i | X^j = x^j$, for $j = 1...k$ & $i \neq j), i = 1, ..., k$
iterations (integer): Size of the population pool
interval (integer): Acceptance interval
Result: Draws from $\pi(x)$
initialize X_{prev};
initialize X_pool;
initialize counter;
for $size\_pool \times interval$ **do**
    Generate a random number from $r = U(1, k)$;
    Generate x^r_{curr} using **Inverse Transform** on $\pi(X^r_{curr} | X^j = x^j_{prev}$, for $j = 1...n$ & $r \neq j)$;
    X_{curr} = X_{prev} with x^r_{prev} replaced by x^r_{curr};
    if *counter equals interval* **then**
        X_pool.Add(X_{curr});
    end
    X_{prev} = X_{curr};
end

Appendix

Gibbs Sampling Algorithm



Synthetic generator