Population synthesis at the level of households

One-step simulator for synthetic household generation

Marija Kukic
Supervisor: Michel Bierlaire
27.04.2022.
Workshop EPFL - ETHZ

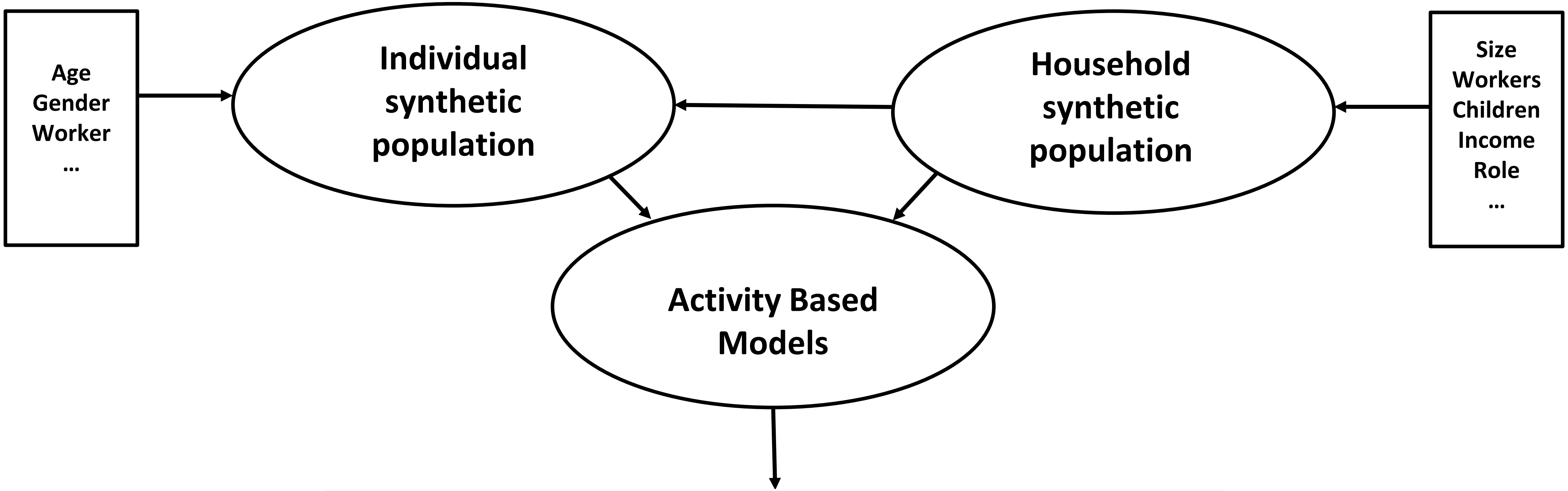
Outline

- Motivation
- Literature review
- Simulation approach for synthetic generation
- One-stage simulator for synthetic household generation
- Results and validation
- Divide and conquer Gibbs Sampler
- Conclusion

What are synthetic data and why do we need them?

- Data collections: surveys, census, mobile phone tracking…
- Why cannot we use those data?
  - High cost of data collection
    => reduce sample size
    => lack of representativity
  - Privacy preservation => data unavailability
- What is a solution? => **Let’s generate synthetic data!**

Why do we need synthetic data in transportation?



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 synthetic individuals to synthetic 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 Network **2018,** *Xu et al.* Tabular Generative Adversarial Networks **2019,** *Borysov et al.,* Variational Autoencoder **2020,** *Badu – Marfo et al.,* Composite Travel Generative Adversarial Networks **2022,** *Lederrey et al.,* DATGAN: Integrating expert knowledge into deep learning for synthetic tabular data		**2022** **...**

From synthetic individuals to synthetic households

Simulation methods

Model driven -> allows control within the generation process

Hierarchy generation -> accuracy of marginals and realistic rows

Curse of dimensionality-> the accuracy and efficiency drops with high dimensional datasets

Machine Learning methods

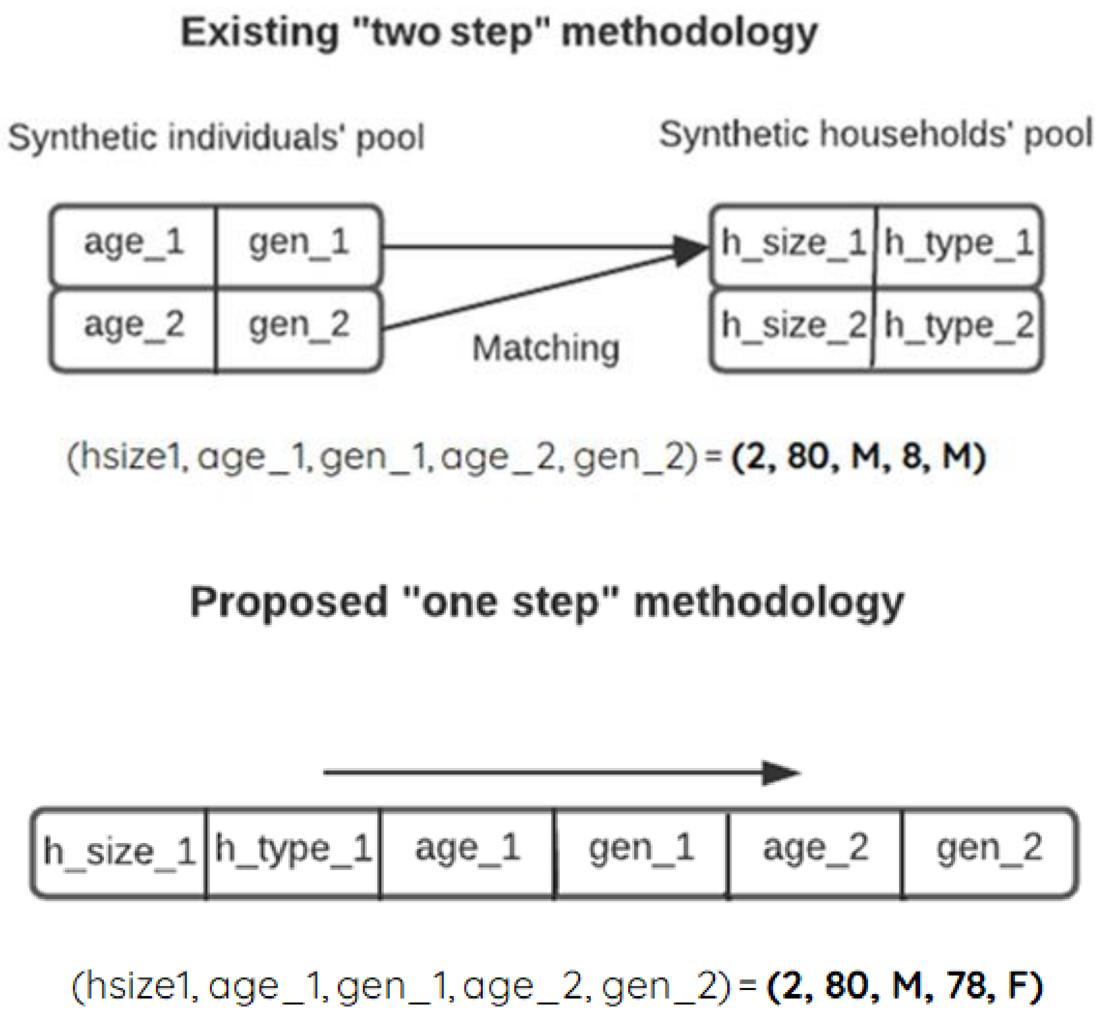
Good correlation capture on high dimensional datasets

Doesn't handle hierarchies -> marginals might seem accurate but unrealistic rows

Data driven-> black box solutions

Gaps in the literature – Why do we need one step simulator?

	METHODS
TWO – STAGE PROCESS	Hierarchical MCMC (hMCMC) Assuming independence between individuals
ONE – STAGE PROCESS	One-step simulator for synthetic household generation



Research questions

One-step simulator for synthetic household generation

How to design a methodology for creation of synthetic households in **one – stage** process?

How much **control** we can embed into generation process compared to other existing methodologies?

How to deal with the “**curse of dimensionality**”?

Existing approach - iMCMC

Simulation based population synthesis:

- Markov Chain Monte Carlo process

Sampling methods:

- Gibbs Sampling

Input preparation:

1. Conditional distributions constructed from:
   Data
   Models
   Assumptions

Assumptions:

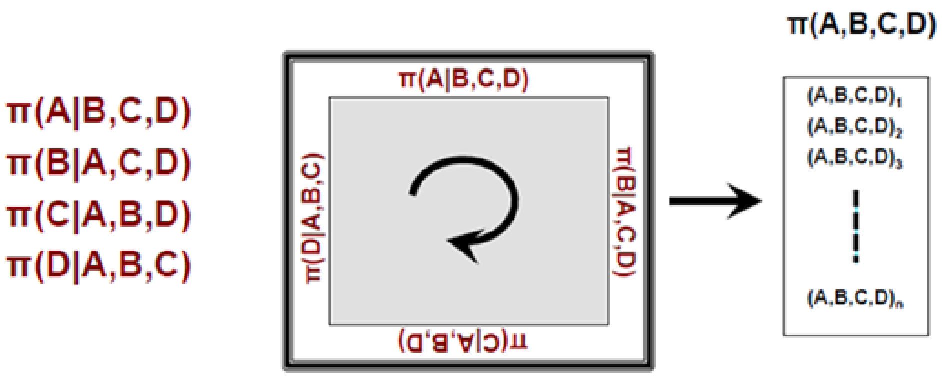
- Given A, B is uniform across C, D:

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

$\pi(A|B)$

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		

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



Contributions – Modeling part

Generalized approach:

H_Size	H_Type	H_Cars	H_Income	[Individuals]

Specific example:

2	Pair	1	6k - 8k CHF	29	M	28	F

P (**H_att_i** | H_att_i+1, H_att_i+2, H_att_i+3)

P (**age_owner** | H_type, H_cars, H_income)

P (**age_ind_2** | attributes_of_previous_ind)

Case study: MTMC 2015 dataset

	SYNTHETIC DATASET
Number of observations	163843 individuals 57090 households
Area	Switzerland
Individual attributes	Age Gender
Household attributes	Household size Household type Number of cars in household Household income

Case study: Validation methods

1. Visualization

- **Marginals** – verify aggregated values
- **Sub-distribution** – verify logic in the data

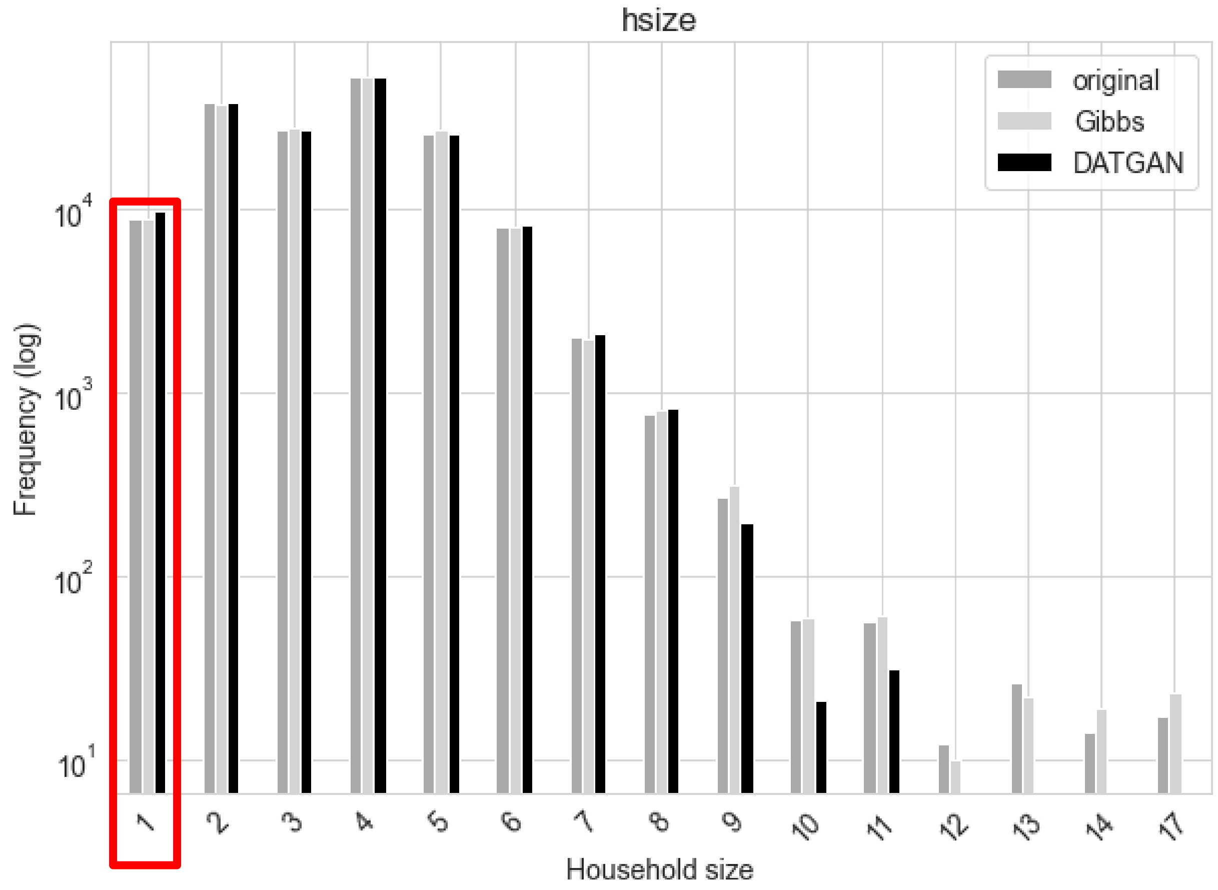
2. Statistics (Lederrey et al., 2022)

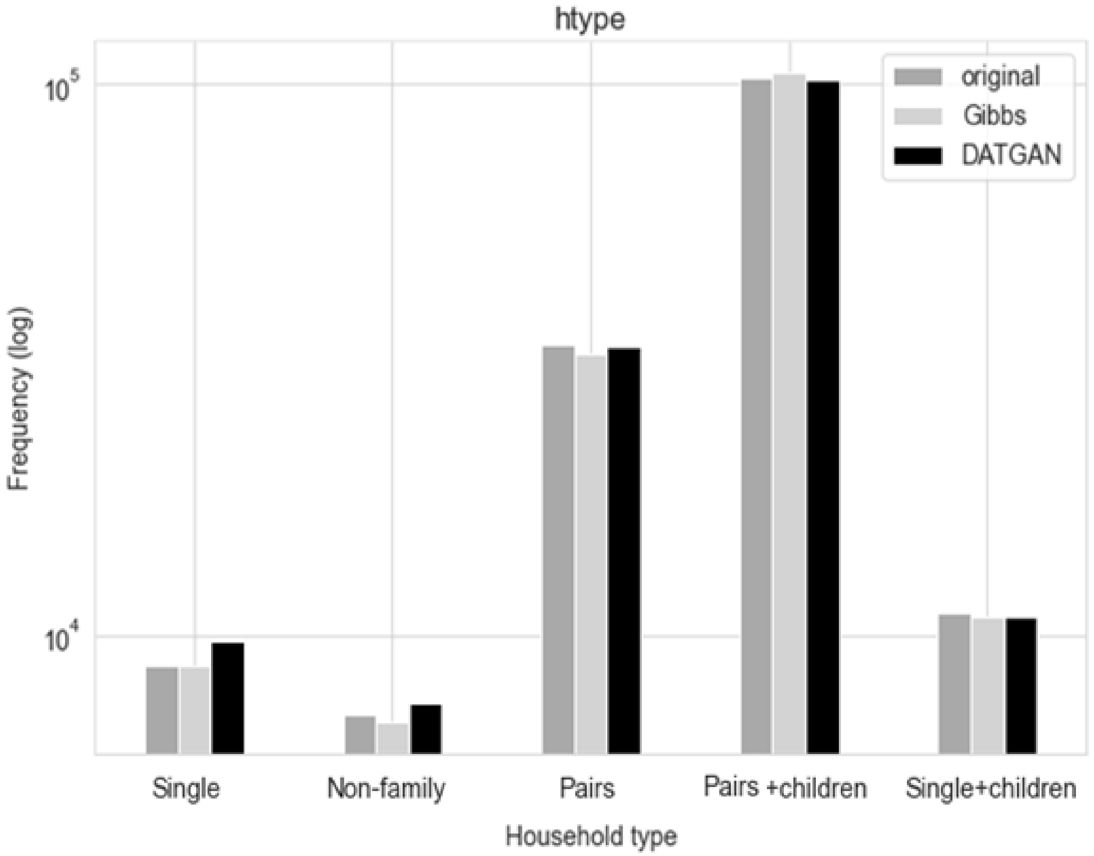
- **First level** – columns are compared one by one separately (verify aggregated values)
- **Second level** – columns are compared two by two (verify logic in the data)
- Calculating: MSE, RMSE, SRMSE, R^2, Piercon’s correlation

Comparison is done between:

- original dataset
- One-stage Gibbs simulator
- DATGAN (Lederrey et al.,2022)

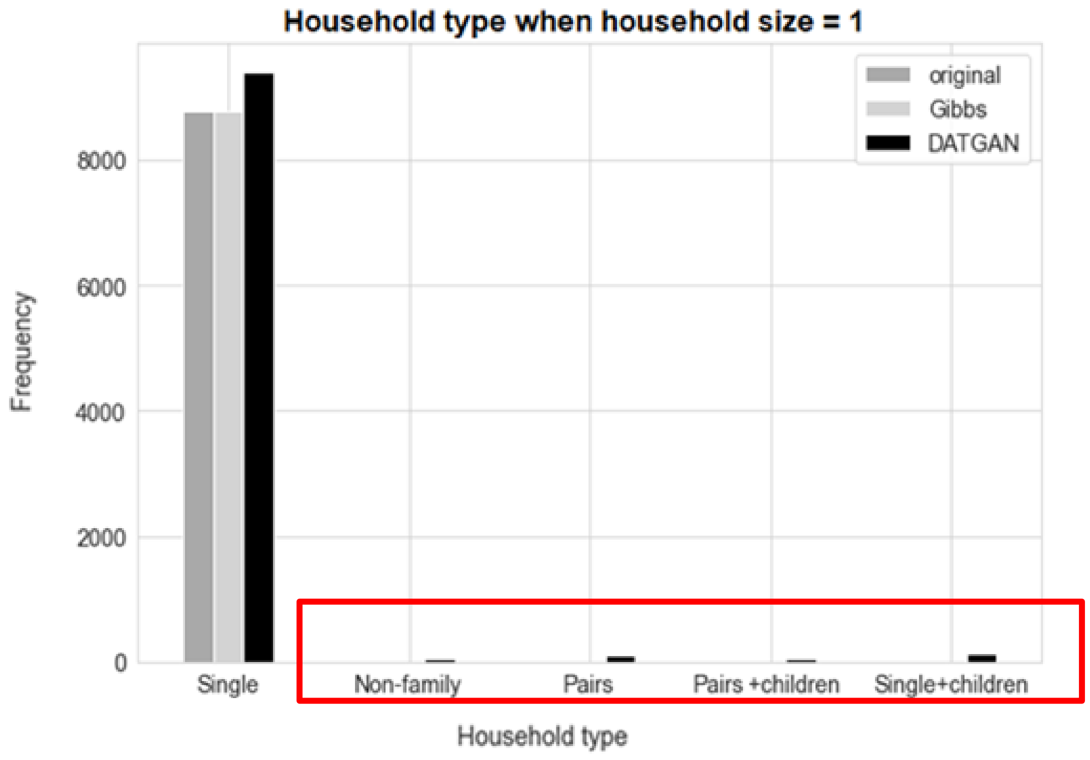
Results - Marginals



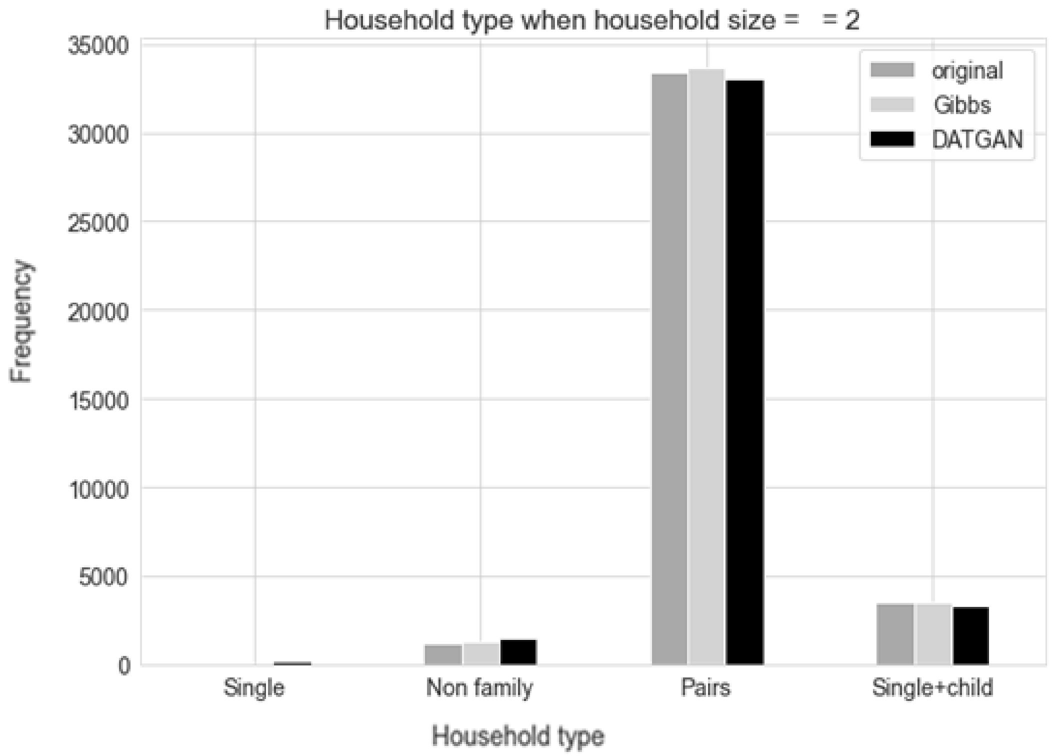


Based on **First order** statistics for **categorical** variables - **Gibbs** gave a better score than **DATGAN**

Results – Sub-distributions



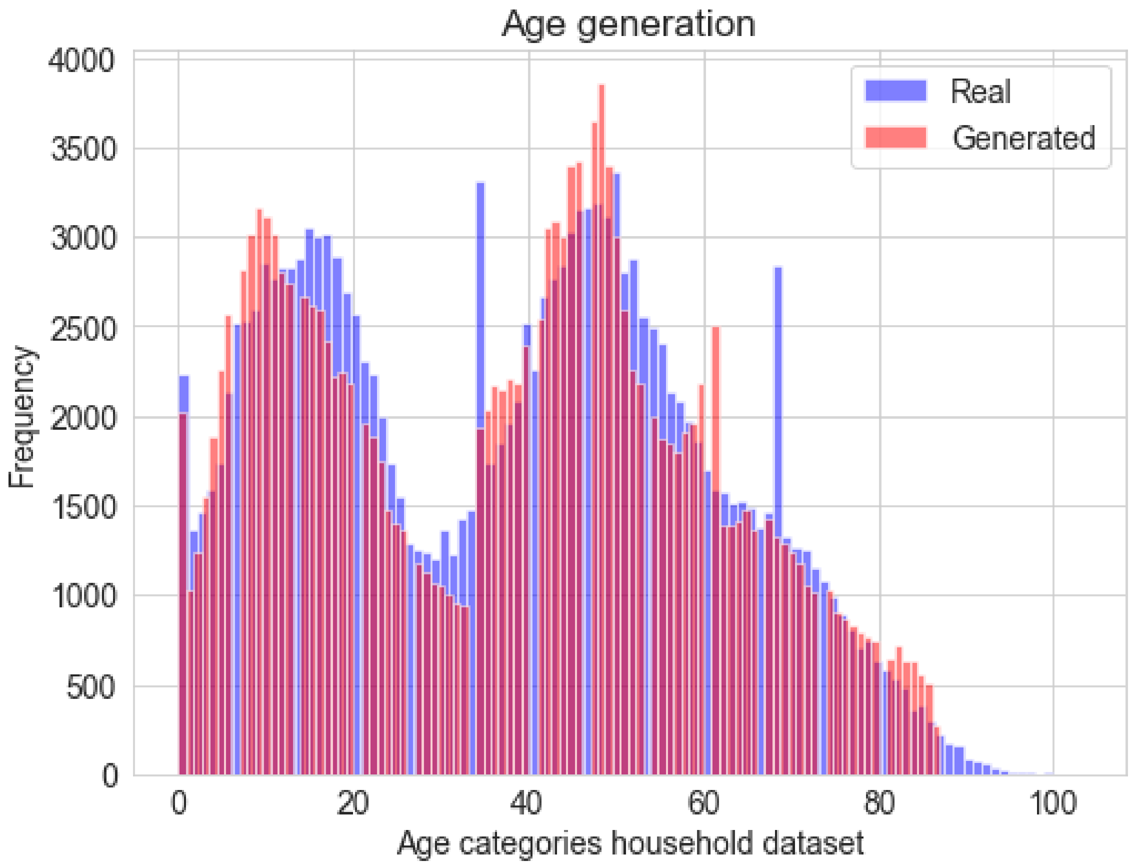
Deterministic part



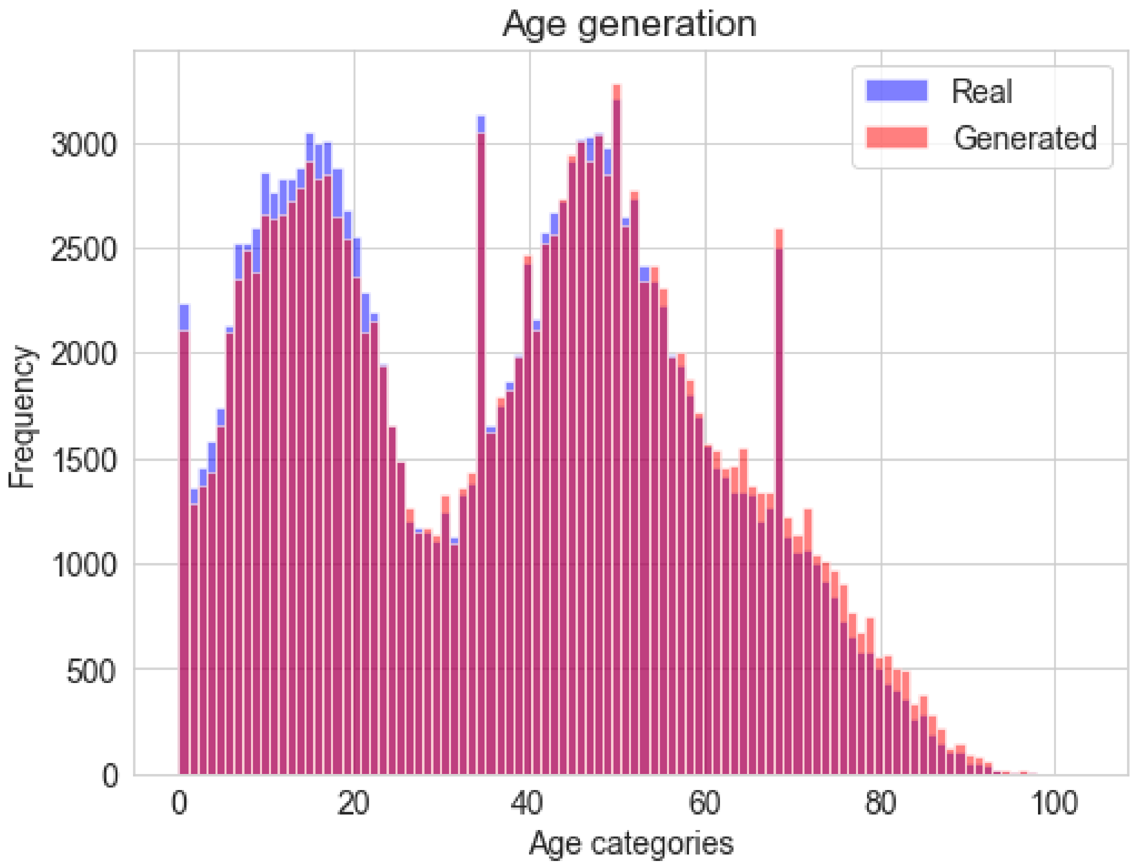
Stochastic part

Based on **Second** order statistics for **categorical** variables - **Gibbs** gave a better score than **DATGAN**

Results – Marginals individuals continuous



DATGAN



Gibbs sampler final model

Based on **First & Second** order statistics for **continuous** variables - **Gibbs** gave a better score than **DATGAN**

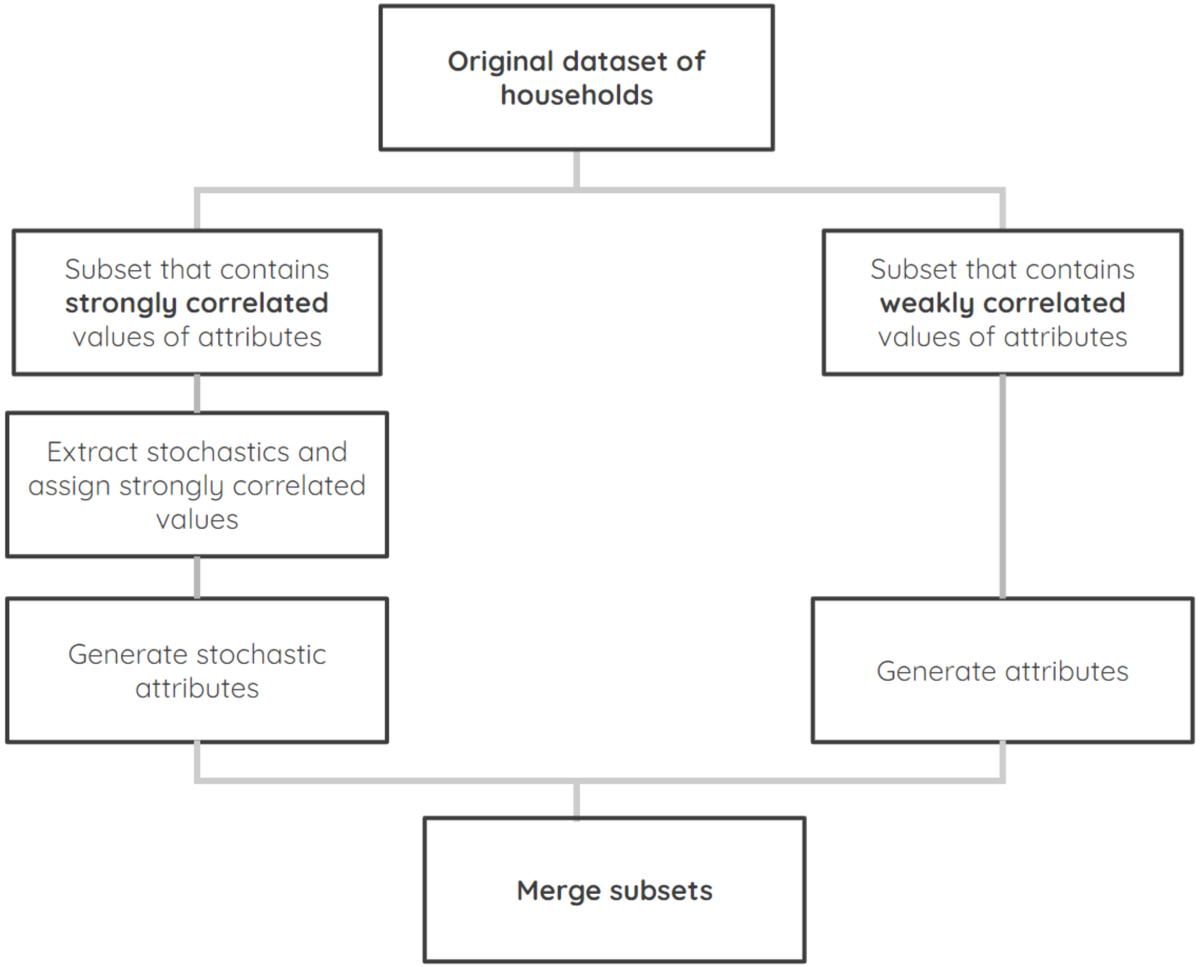
Contributions – Algorithmic part

- **Curse of dimensionality** breaks the algorithm by adding more dimensions

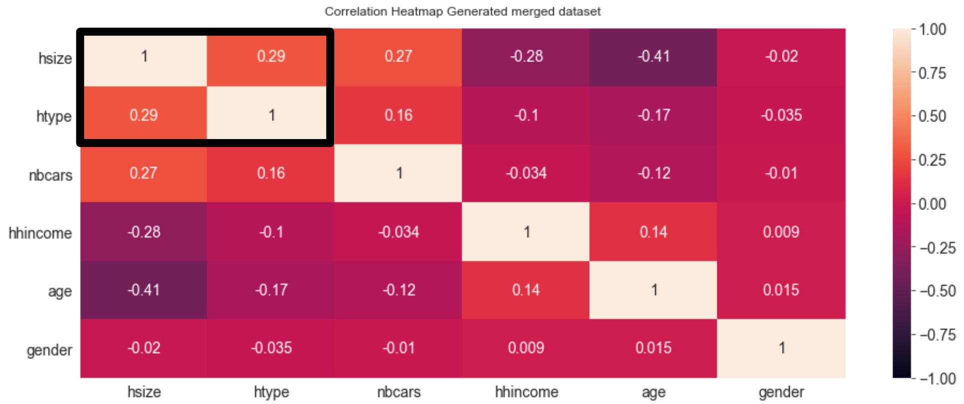
- Gibbs sampler gets stuck in highly correlated areas
  - long execution time
  - less accuracy by forcing "highly correlated" values and ignoring "weakly correlated" values

- Gibbs sampler completely fails if there is 1-1 correlation -> don't generate it, assume it, save time and be more accurate

Contributions – Divide and conquer simulator for synthetic household generation



On going work…



Conclusion & Future work

- Enforce rules -> control of generation process -> assume the correlations and let the model & data to do the rest
- Divide and conquer ->
    - Identify which values are causing strong correlation
    - isolate those areas
    - generate “strongly” and “weakly” correlated subsets in parallel
    - merge subsets
- Investigate convergence and influence on efficiency
- Revise all conditionals in order to simplify where needed

Thank you for your attention!