Synthetic Population Generation: Hierarchies and Activities

Marija Kukic
Michel Bierlaire
16.02.2023.
Internal seminar - Annecy

Synthetic Data: What? Why?

- Data collections: surveys, census, mobile phone tracking…

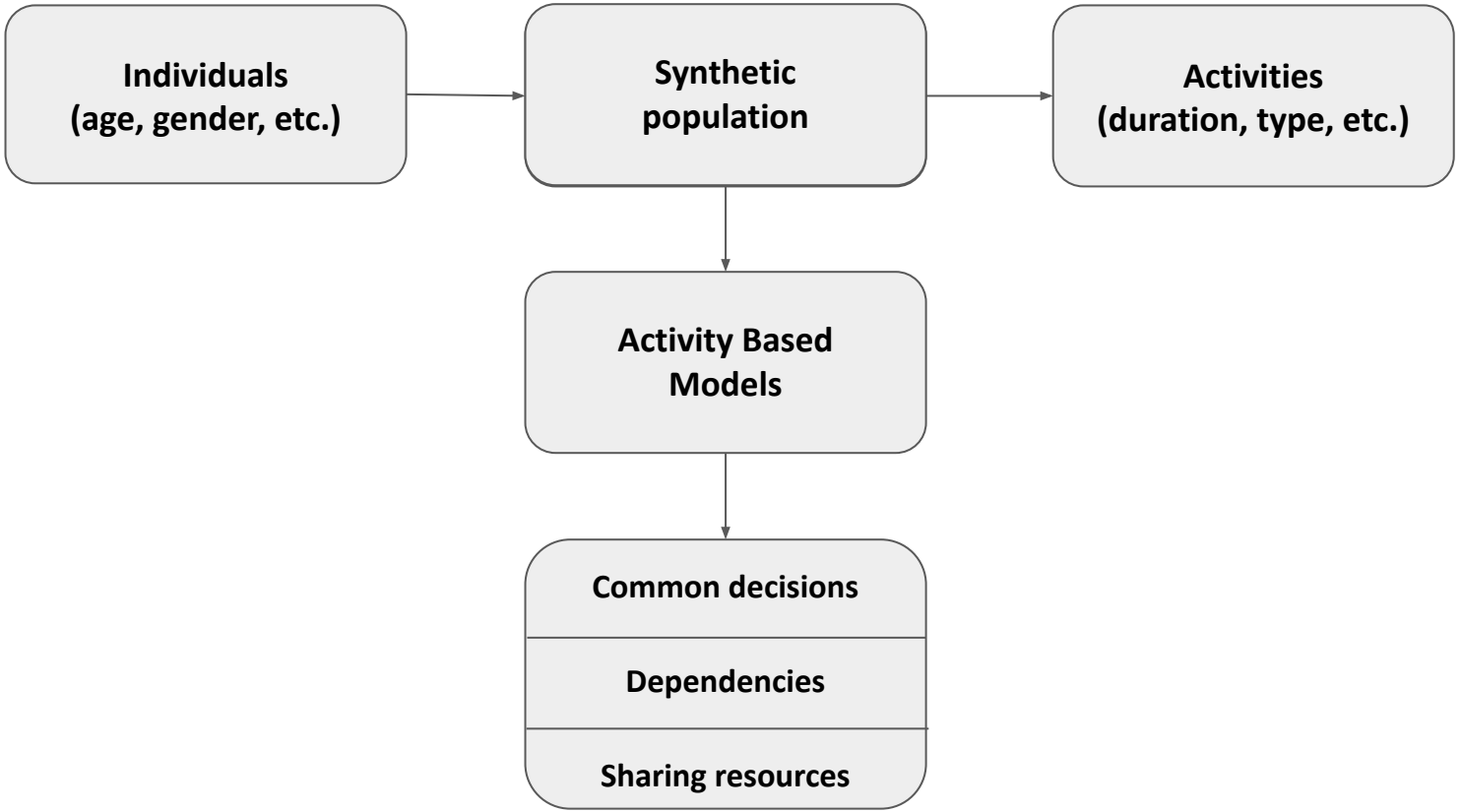
Limitations of real data

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

Let's use synthetic data!

- Open source
- Bias correction
- Privacy preservation

Synthetic population in transportation



Research Axes: Generating Tabular data

Representativity

Hierarchies

Sequence

Age	Gender	Driving licence
10	M	YES
20	F	NO

Size	Type	Cars
1	1	1
2	2	2

Type	Start	End
Home	00:00	**08:00**
Work	**08:00**	17:00

Literature Review: Gaps

Representativity

Beckman et al. 1996
IPF

Farooq et al. 2013
Simulation based
MCMC

GoodFellow et al. 2014
GAN
Xu et al. 2018
TGAN
Vega et al. 2020
CGAN
Lederrey et al. 2022
DATGAN

Hierarchies

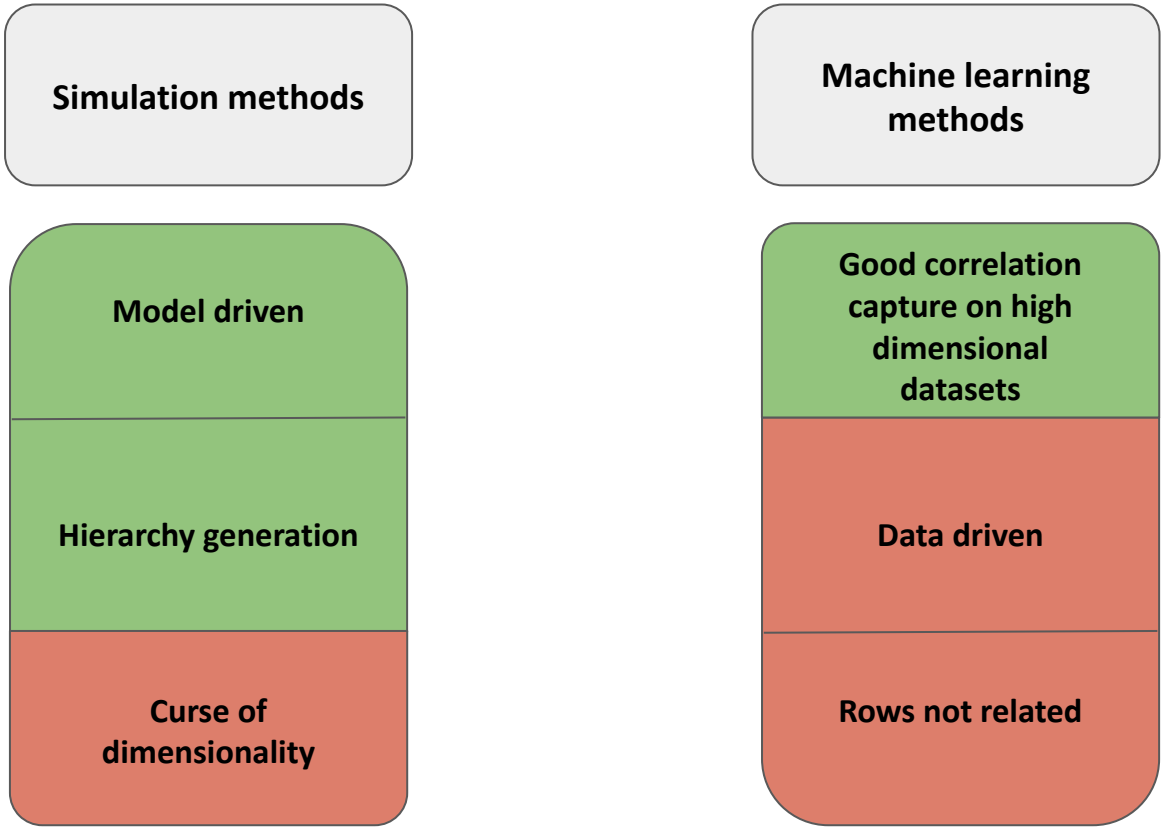
Divide and Conquer One-step Simulator for synthetic household generation

Sequence

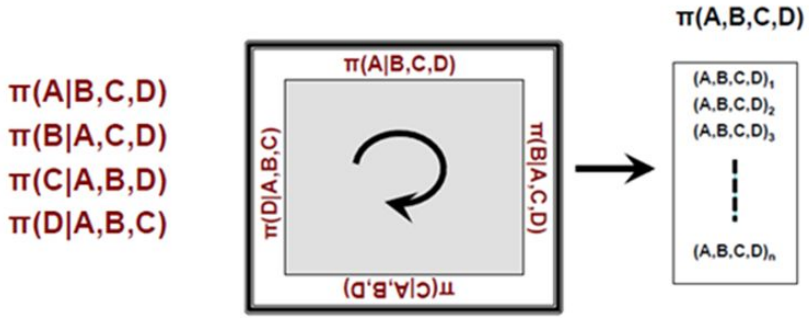
Machine learning techniques for activity sequence generation

Divide and Conquer One-Step Simulator for synthetic households generation

Why simulation?



Existing Methodology: Gibbs Sampler for Data Generation



Input: Conditionals

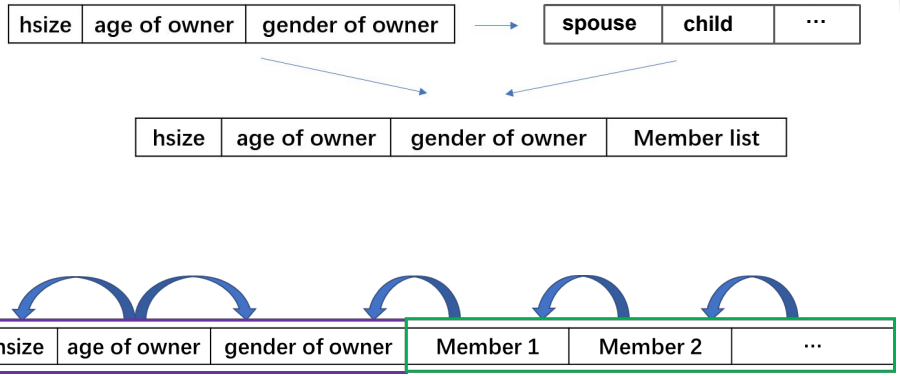
1. **Data**
2. Models
3. **Assumptions**

Assumptions:
Full conditionals can be simplified

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

Methodology: From two-steps to one-step

Two-steps

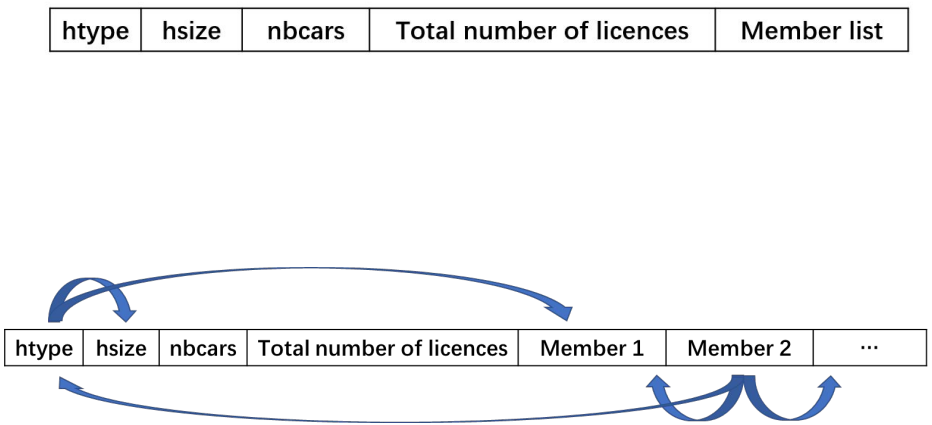


Low flexibility

Low efficiency

Not general enough

One-step



More flexible

Better efficiency

More general

Curse of dimensionality?

Shortcomings of Gibbs Sampler

“Curse of dimensionality”

- Highly correlated areas
- Total correlation
- Unavailability of full conditionals

Solution

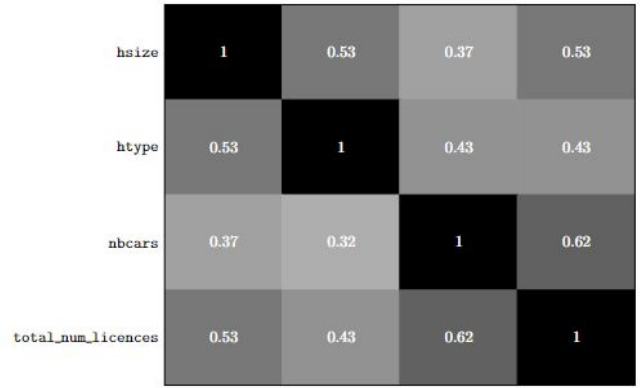
- Divide & Conquer approach
- Simplification of conditionals

Methodology: One-step Divide & Conquer Gibbs Sampler (D&C)

hsize <=

size >= 3

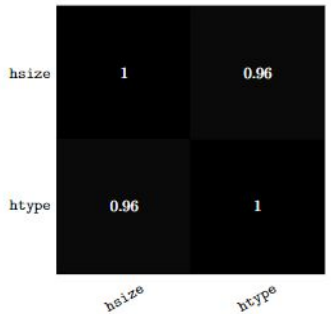
htype	hsize	nbcars	Total nu	Total number of licences	Member list

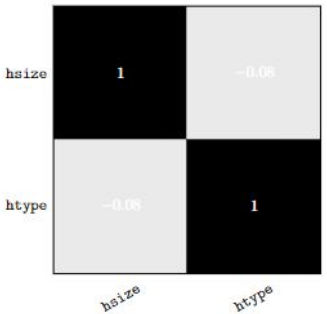


hsize <=2

hsize >=3

r list





Case study: MTMC 2015

163843 individuals, 57090 households

Household attributes:

- Household type
- **Household size**
- Number of cars
- Total number of driving licences

Individual attributes:

- **Age**
- **Gender**
- Marital status
- Employment status
- Driving licence

Comparisons:

- Two-steps VS. One-step
- One-step VS. One-step D&C
- One-step D&C VS. DATGAN

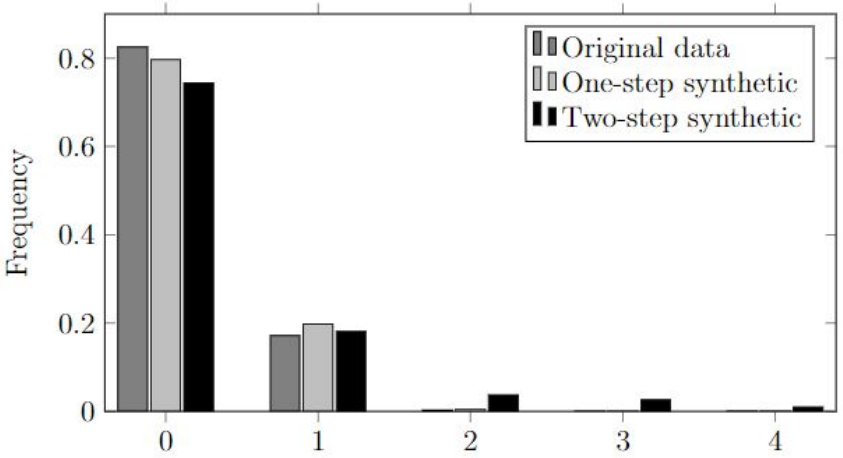
Case study: Validation Techniques

1. Visualization
    - Marginals - verify aggregated
    - Sub-distribution - verify logic in the data

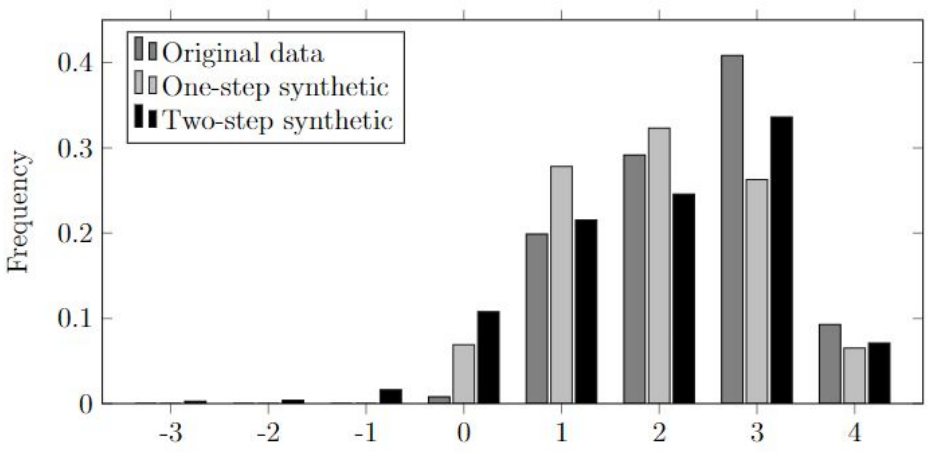
2. Statistics (Lederrey et al.,2022)
    - First level - one by one column
    - Second level - two by two columns
    - Calculating: SRMSE, MSE, RMSE, R^2, Pearson’s correlation

3. Convergence investigation
    - Potential reduction scale
    - Effective number of draws
    - Computational time

Results: two steps VS. one step



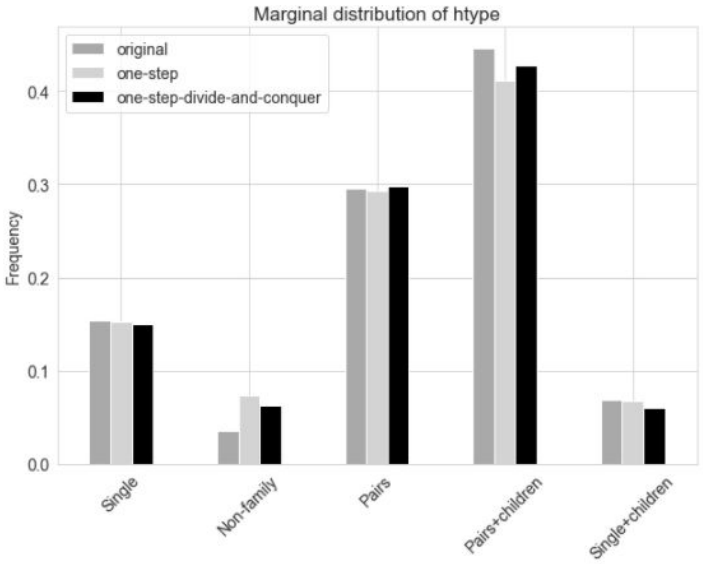
Age difference between partners in couple

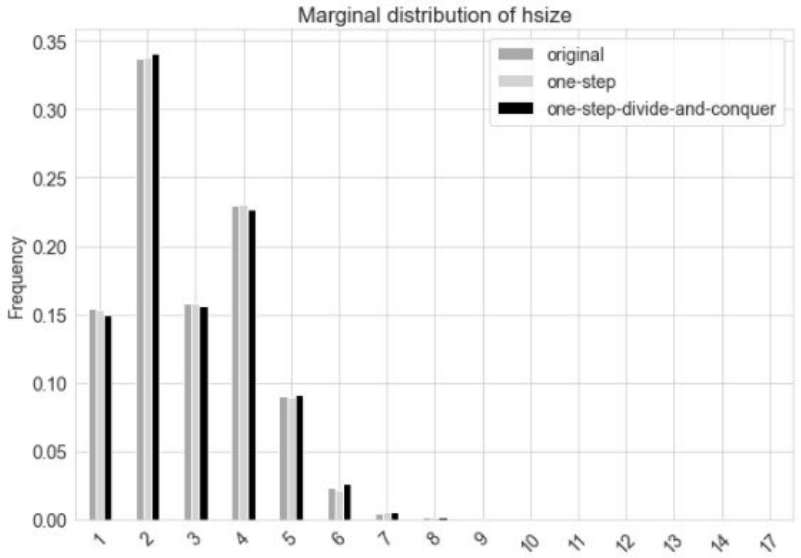


Age difference between spouse and child

First order	Second order	Computation time
One-step	One-step	One step 2x faster

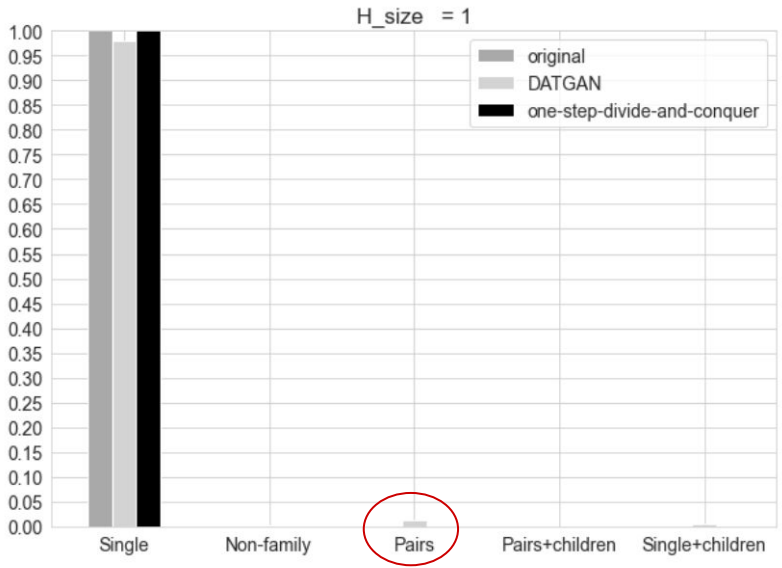
Results: one step VS. one step D&C

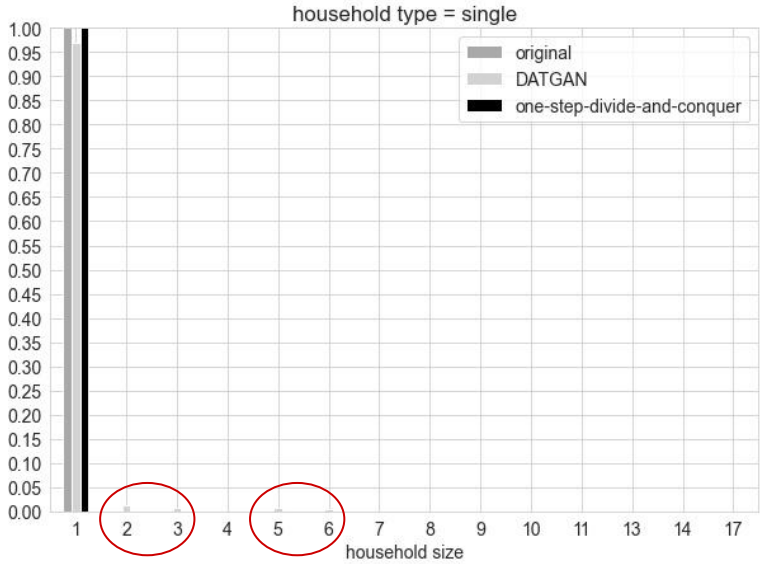




First order	Second order	Computation time
One-step D&C	One-step D&C	One step D&C 2x faster

Results: one step D&C VS. DATGAN





First order	Second order
DATGAN	DATGAN

Conclusion

- Trade-off between accuracy and efficiency
- Enforce the constraints -> realistic observations
- Dealing with curse of dimensionality

Future work

- Decorrelation of the variables
- Can we generate different data types (e.g. activity sequence)?

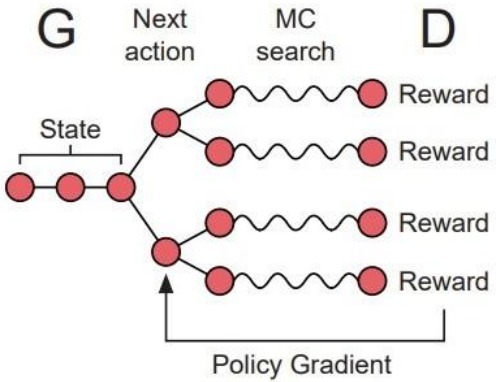
Target output	START (input)	END (input)
home	00:00	**08:00**
travel	**08:00**	08:20
work	08:20	17:00
travel	17:00	17:15
home	17:15	24:00

Machine learning techniques for activity sequence generation

Existing methods

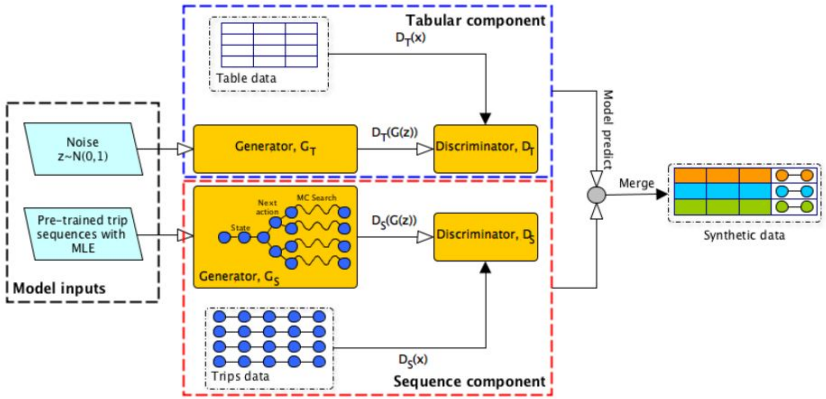
SeqGAN

Generator - RNN with Monte-Carlo tree search
Discriminator - CNN



CTGAN

Application of SeqGAN for trip sequence generation



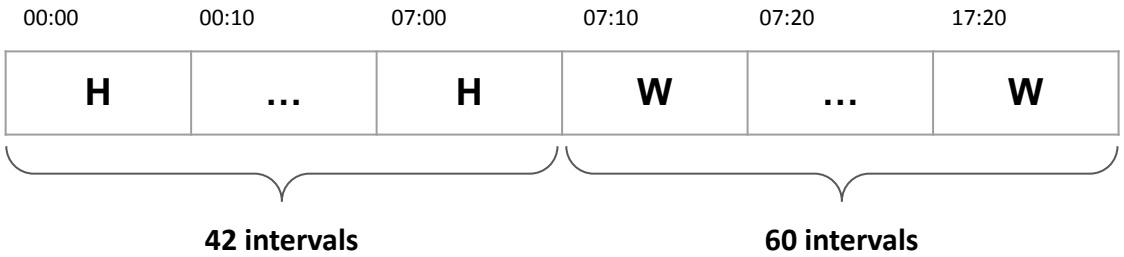
Existing methods - shortcomings

1. How to merge **socio-demographics** and **activities**?
2. **Fixed length** of activity sequence
3. **One** generated **value** for each point of sequence
4. Evaluation on the synthetic data

Case study - MTMC 2015

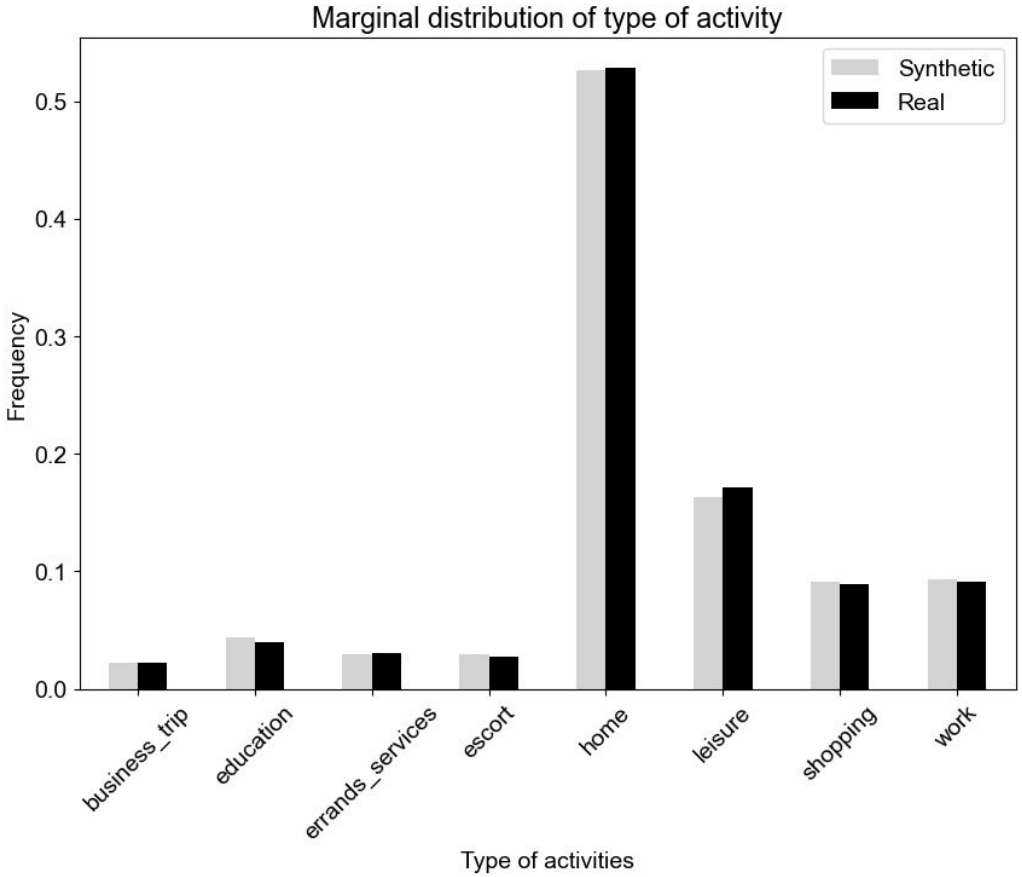
Preparation of input:

- Discretization 24 hours = 1440 min = **144** intervals * **10** minutes, **10 000** schedules

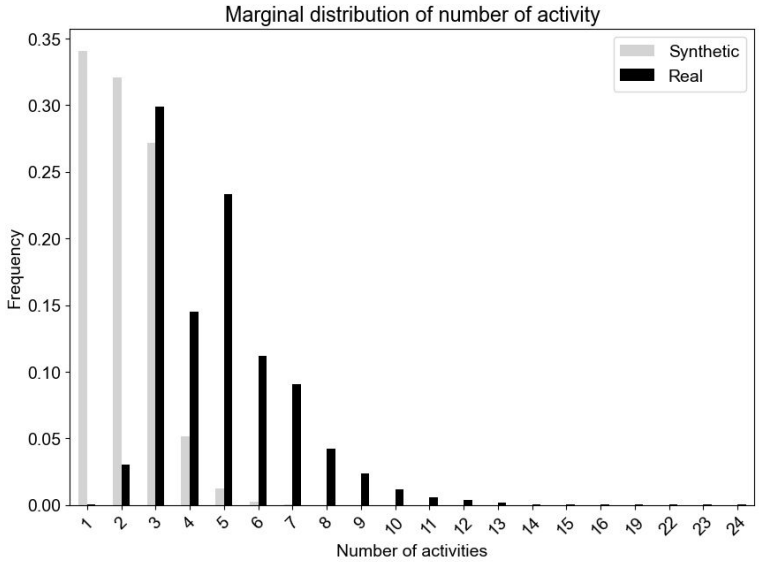


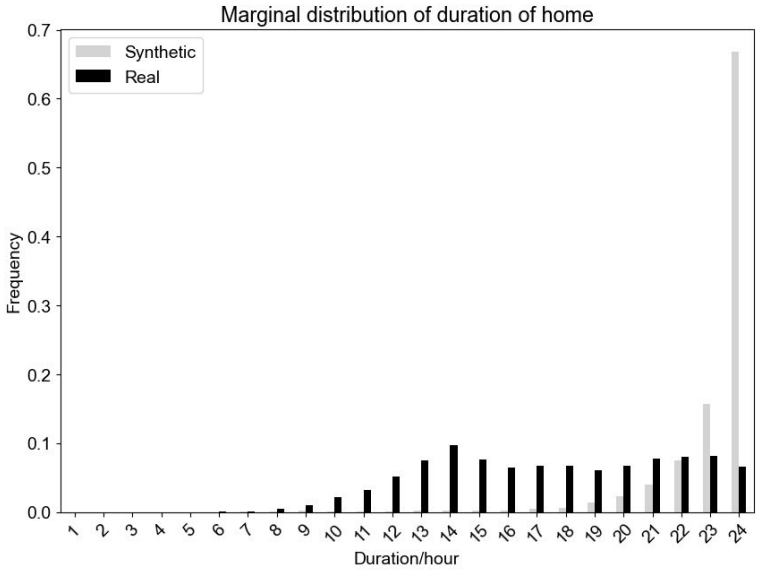
- **75%** observations travel time < **15 min**

Results


Marginal distribution of type of activity
Synthetic
Real
Frequency
0.0
0.1
0.2
0.3
0.4
0.5
business_trip
education
errands_services
escort
home
leisure
shopping
work
Type of activities

Results

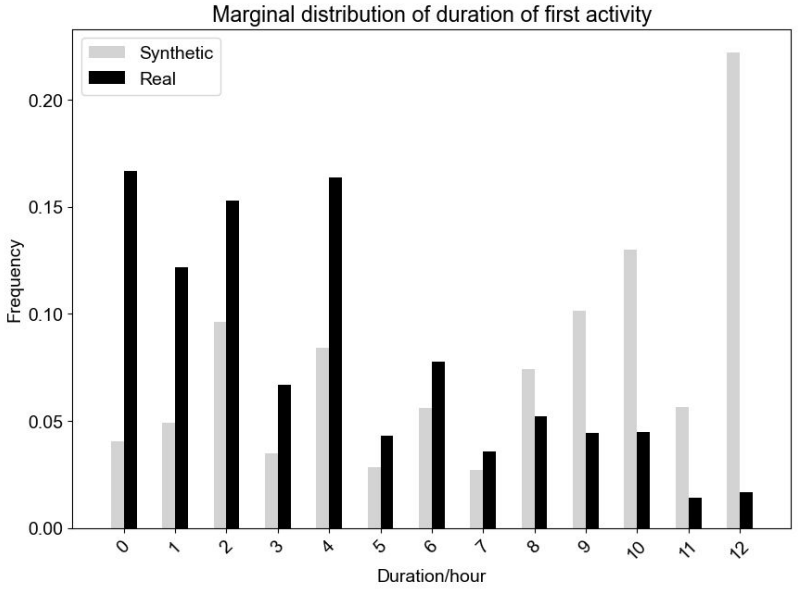




Home activity lasts the whole day - **unrealistic**

Results

Filter out home activity as much as possible - **08:00 - 20:00**



Activities that last the **whole day** still exist

Conclusion

- SeqGAN does not replicate sub-distributions of duration for different activity types
- Fixed length of activity sequence due to RNN structure

Future steps

- Adapt another seq2seq model (such as **Transformers**) for **activity generation**
- From text generation to discrete generation -> how to **tokenize** data?
- Is it possible to generate **multiple attributes** in sequence (e.g. activity type & duration?)

Thank you for your attention!

Appendix

Conditionals

1. Two step GS

- $P\ (hsize|age_{owner}, gender_{owner})$
- $P\ (age_{owner}|hsize, gender_{owner})$
- $P\ (gender_{owner}|hsize, age_{owner})$
- $P\ (age_{spouse}|hsize, age_{owner}, gender_{owner}, gender_{spouse})$
- $P\ (gender_{spouse}|hsize, age_{owner}, gender_{owner}, age_{spouse})$
- $P\ (age_{child}|hsize, age_{owner}, gender_{owner}, age_{spouse}, gender_{spouse}, gender_{child})$
- $P\ (gender_{child}|hsize, age_{owner}, gender_{owner}, age_{spouse}, gender_{spouse}, age_{child})$
- $P\ (age_{other}|hsize, age_{owner}, gender_{owner}, age_{spouse}, gender_{spouse}, age_{child}, gender_{child})$
- $P\ (gender_{other}|hsize, age_{owner}, gender_{owner}, age_{spouse}, gender_{spouse}, age_{child}, gender_{child})$

2. One step GS

- $P\ (htype|hsize, age_{owner}, age_{second})$
- $P(hsize)\ from\ marginal$
- $P\ (nbcars|htype, hsize, total\ number\ of\ licences)$
- $P\ (age_{owner}|htype, age_{second}, age_{third})$
- $P\ (gender_{owner}|htype, gender_{second})$
- $P\ (marital_{owner}|htype, age_{owner}, marital_{second})$
- $P\ (employment_{owner}|age_{owner})$
- $P\ (driving\ licence_{owner}|age_{owner}, nbcars)$

2. One step GS

- $P\ (age_{second}|htype, age_{owner}, age_{third})$
- $P\ (gender_{second}|htype, gender_{owner})$
- $P\ (marital_{second}|htype, age_{second}, marital_{owner})$
- $P\ (employment_{second}|age_{second})$
- $P\ (driving\ licence_{second}|age_{second}, nbcars)$
- $P\left(age_{member}\middle|htype, age_{previous\ individual}\right)$
- $P\ (gender_{member}|htype)$
- $P\ (marital_{second}|htype, age_{member})$
- $P\ (employment_{member}|age_{member})$
- $P\ (driving\ licence_{member}|age_{member}, nbcars)$

3. One step divide-and-conquer GS

- $P(hsize)\ from\ marginal\ for\ hsize \leq 2; otherwise\ P(hsize|htype)$
- Other conditionals the same as those in one step GS

- Previous approach -> **One attribute conditional to all others**

Generation of 4 household attributes:

- Household size **(HS)**
- Household type **(HT)**
- Number of cars **(NC)**
- Total number of driving licences **(TD)**

Full conditionals on full dataset:

- P(HS | HT, NC, TD)
- P(HT | HS, NC, TD)
- P(NC | HS, HT, TD)
- P(TD | HS, HT, TD)

	HS = 1	HS=2	HS >2
HT = Single	1	0	0
HT = Couple + children	0	P_1	P_4
HT = Non-family	0	P_2	P_5
HT = Single + children	0	P_3	P_6

Once it enters this vector it never goes out **-> degenerative state**

It fails to converge for the full dataset !

- **Problems while using full conditionals**
  - in case of total correlation it’s not possible to converge in reasonable time
  - unavailability of full conditionals - the more attributes we add, there are more unexisting categories while drawing
  - long computational time to access high dimensional tables
  - not all attributes have the same importance

“Curse of dimensionality”? => everyone is aware of the problem, but nobody is solving it

- **Solution 1: Simplify conditionals**
  - **Example :** P(HT | HS, NC, TD) ⇔ **P(HT | HS)**
    - **HT** is entirely defined by **HS**
    - other attributes only increase complexity (longer execution time, same accuracy)
    - fully stochastic process
  - **Expectations**: **accuracy will slightly drop but efficiency will increase**

- **Solution 2: Divide dataset based on the correlation between different categories of attributes**
  - deterministically assign what is possible, keep the rest of the process stochastic
  - isolate highly correlated areas
  - apply different conditionals for different subsets

Expectations: **accuracy and efficiency will increase**

3 scenarios applied on strong (hs = 1 or hs = 2) and week subsets (hs>2):

- **Full conditionals -** keeping the same conditionals as before for both subsets
- **Simplified conditionals -** remove attributes from full conditionals that are considered as less meaningful from modeling point of view for both subsets
- **Advanced conditionals -** use derived attributes from dataset that are more informative and revise conditionals based on the expert knowledge

	hs>2	hs=1 or hs=2	
Full	P(HS | HT, NC, TD) P(HT | HS, NC, TD) P(NC | HS, HT, TD) P(TD | HS, HT, NC)	HS marginal distribution HT deterministically wrt. HS P(NC | HS, HT, TD) P(TD | HS, HT, NC)	
Simplified	P(HS | HT) P(HT | HS) P(NC | TD) P(TD | NC)	HS marginal distribution HT deterministically wrt. HS P(NC | HS, TD) P(TD | HS, NC)	
Advanced	P(HS | HT, TD) P(HT | HS) P(NC | HS, TD) P(TD | HS,NC)	HS = 1 HT = Single P(NC | TD) P(TD | NC)	HS = 2 HT from marginals P(NC | TD) P(TD | NC)

Computational time:

1. generated_advanced_4d

2. generated_simple_4d
3. generated_full_4d

First order statistics:

1. generated_advanced_4d - 7.80e-03 ± 4.36e-03

2. generated_simple_4d - 1.11e-01 ± 1.15e-01
3. generated_full_4d - 1.42e-01 ± 1.05e-01s

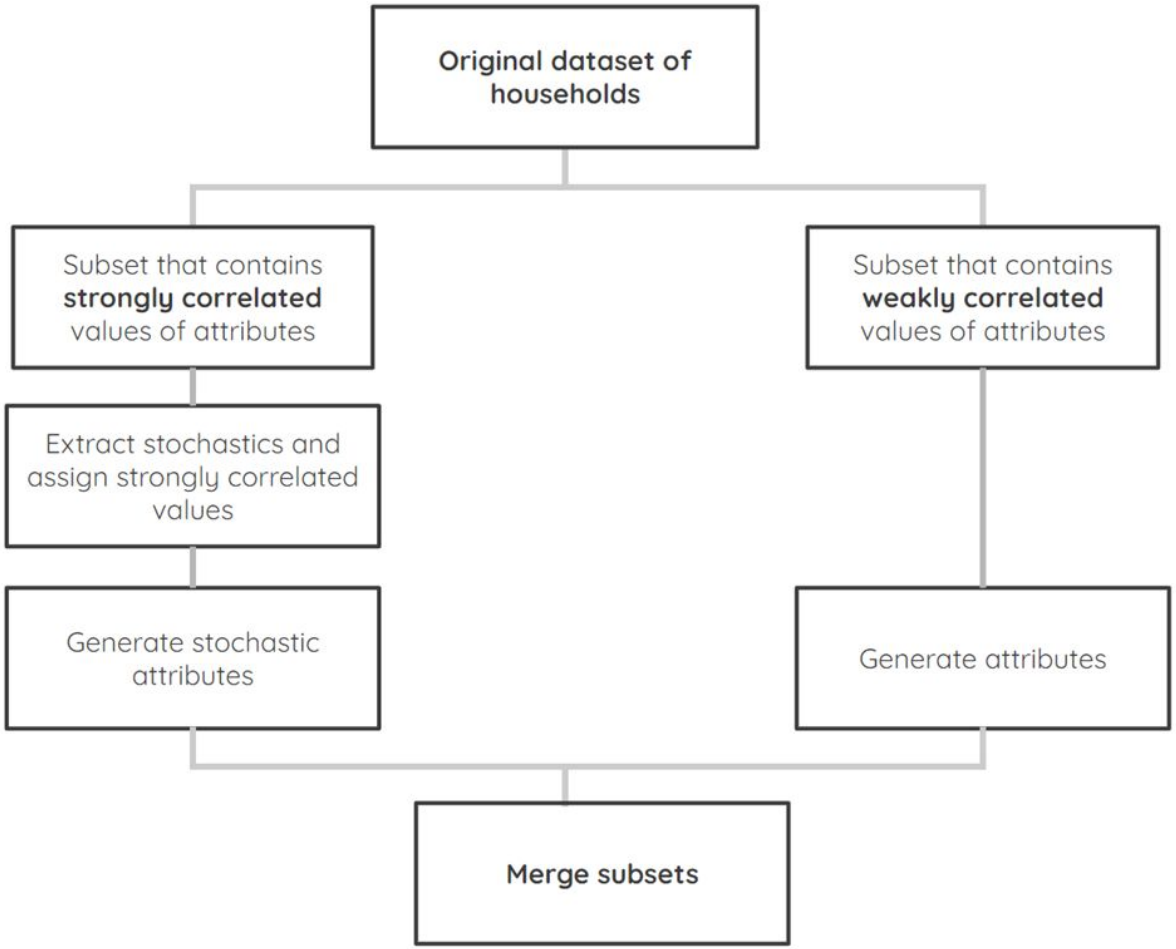
Second order statistics:

1. generated_advanced_4d - 7.24e-02 ± 4.05e-02

2. generated_simple_4d - 4.84e-01 ± 2.47e-01
3. generated_full_4d - 5.04e-01 ± 1.76e-01

Research questions

- How to design a methodology for creation of synthetic households in **one – step** process?
- How much **control** we can embed into generation process in order to generate **realistic households**?
- How to deal with the **curse of dimensionality** phenomena?


Original dataset of households
Subset that contains strongly correlated values of attributes
Subset that contains weakly correlated values of attributes
Extract stochastics and assign strongly correlated values
Generate stochastic attributes
Generate attributes
Merge subsets