INSTITUTE FOR EMPLOYMENT RESEARCH
The Research Institute of the Federal Employment Agency

ITS COMPLICATED: KNOW YOUR DATA AND KNOW YOUR GENERATOR

Wiesbaden,
21. März, 2024

Jonathan Latner, PhD
Dr. Marcel Neuenhoeffer
Prof. Dr. Jörg Drechsler

SECTIONS

1. Introduction
2. Know your data (SD2011)
3. Know your generator
   - DataSynthesizer
   - CTGAN
   - Synthpop
4. Conclusion

Section 1: Introduction

COMPARE 3 SYNTHETIC DATA GENERATOR (SDG)

- DataSynthesizer (Bayes)
- CTGAN (GAN)
- Synthpop (CART)

Section 2: Know your data (SD2011)

REAL DATA

- Social Diagnosis 2011 (SD2011)
- Loads with Synthpop
  - `http://www.diagnoza.com/index-en.html`
  - Not entirely clear how original data is created or cleaned to create data in Synthpop
- Like real data, has ‘quirks’ or unusual values/variables
  - Includes missings
    - Informative (i.e. month married, but single)
    - Non-informative
  - Includes ‘errors’
    - `smoke` - Does not smoke is NO, but `nociga` - 20/22 cigarettes per day
  - Includes generated variables
    - `bmi,agegr`
    - Can be problematic for SDG

DATA (SD2011)

Number	Variable	Description	Type	Observations	Unique.Values	Missings	Negative.values	Generated	Quirks
1	sex	Sex	factor	5000	2	0	0		
2	age	Age of person, 2011	numeric	5000	79	0	0		
3	agegr	Age group, 2011	factor	5000	7	4	0	Yes	Yes
				...					
7	eduspec	Discipline of completed qualification	factor	5000	28	20	0		Yes
				...					
10	income	Personal monthly net income	numeric	5000	407	683	603		
11	marital	Marital status	factor	5000	7	9	0		
12	mmarr	Month of marriage	numeric	5000	13	1350	0		
13	ymarr	Year of marriage	numeric	5000	75	1320	0		
14	msepdiv	Month of separation/divorce	numeric	5000	13	4300	0		
15	ysepdiv	Year of separation/divorce	numeric	5000	51	4275	0		
				...					
22	nofriend	Number of friends	numeric	5000	44	0	41		
23	smoke	Smoking cigarettes	factor	5000	3	10	0		
24	nociga	Number of cigarettes smoked per day	numeric	5000	30	0	3737		Yes
				...					
27	workab	Working abroad in 2007-2011	factor	5000	3	438	0		
28	wkabdur	Total time spent on working abroad	numeric	5000	33	0	4875		Yes
				...					
33	height	Height of person	numeric	5000	65	35	0		
34	weight	Weight of person	numeric	5000	91	53	0		
35	bmi	Body mass index (weight/(height^2)*10000	numeric	5000	1396	59	0	Yes	Yes

Section 3a): Know your generator (DataSynthesizer)

TASKS

- Run default model - correlated attribute mode
- Hyperparameters
  - ϵ DP: 0 (default 0.1)
  - Bayesian network ($\mathcal{N}$) parents: 1, 2, or 3 (default is ‘greedy’)

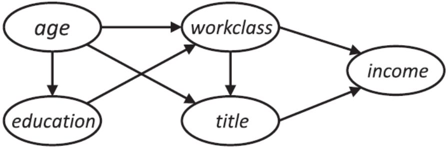
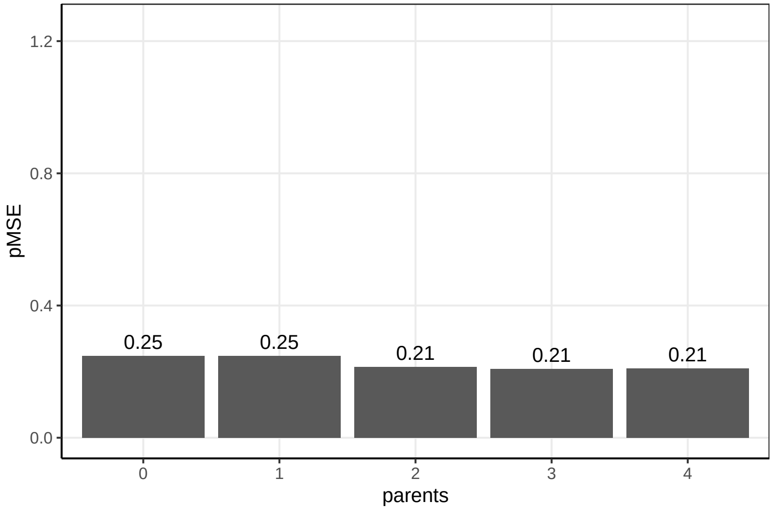


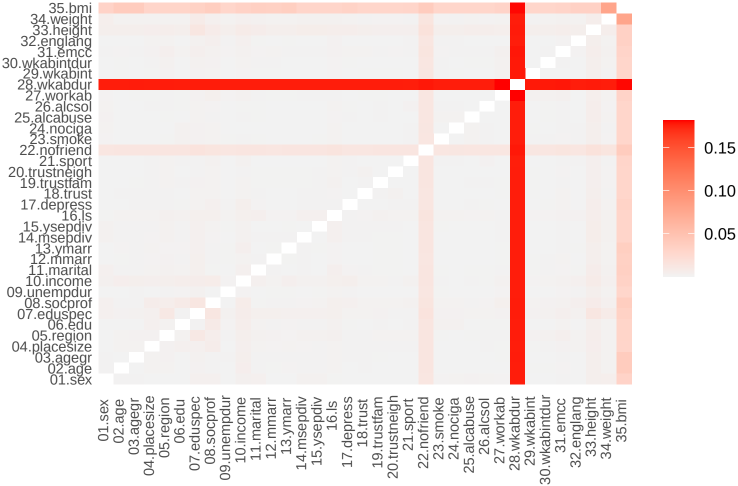
Fig. 1. A Bayesian network $\mathcal{N}_1$ over five attributes.

- In Fig. 1, $\mathcal{N} = 2$, but not known in reality

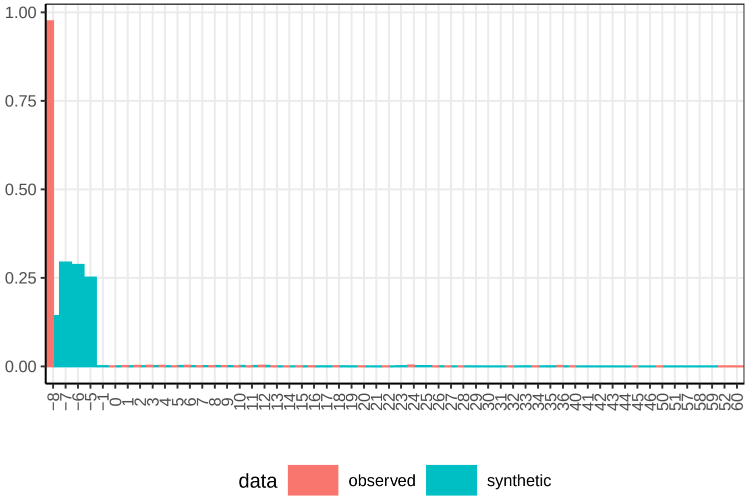
SD2011 - PMSE



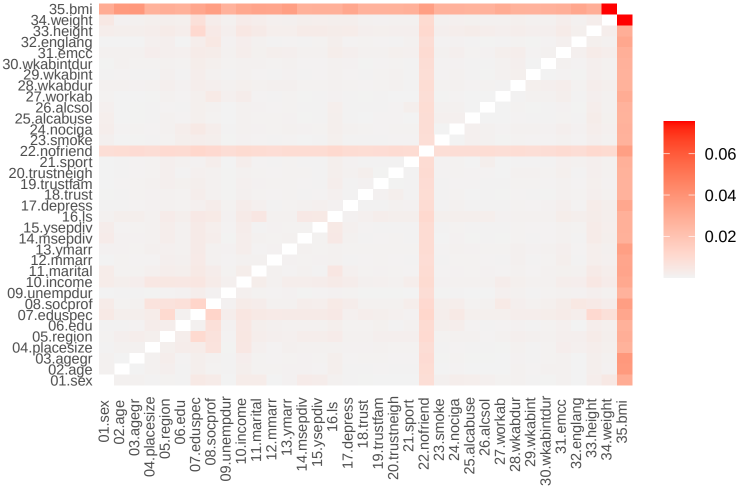
DATASYNTHESIZER - SD2011(A)



VARIABLE: WKABDUR (WORK ABROAD DURATION)

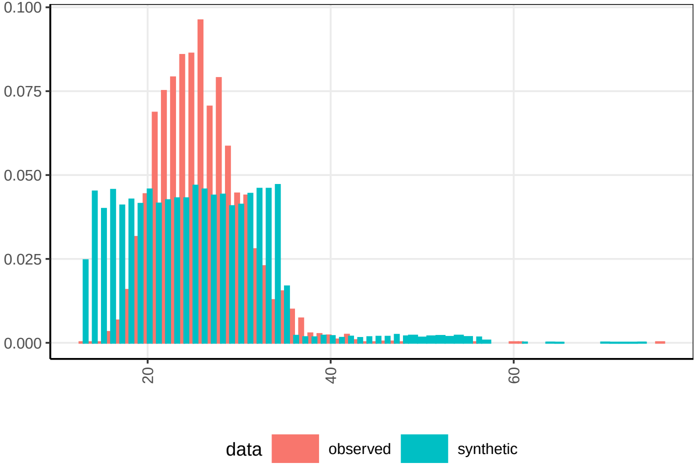


DATASYNTHESIZER - SD2011(B)

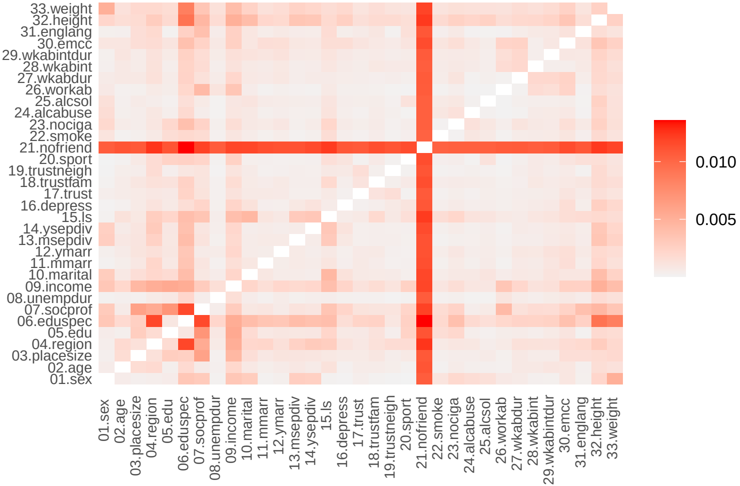


VARIABLE: BMI

Figure 1: BMI $<$ 20 is underweight/malnourished

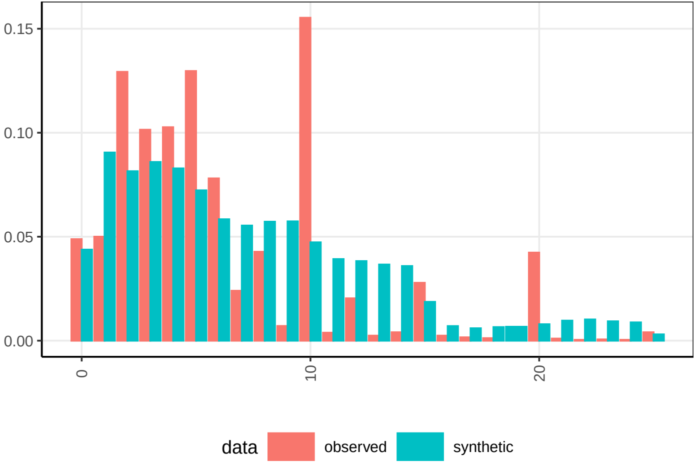


DATASYNTHESIZER - SD2011(C)

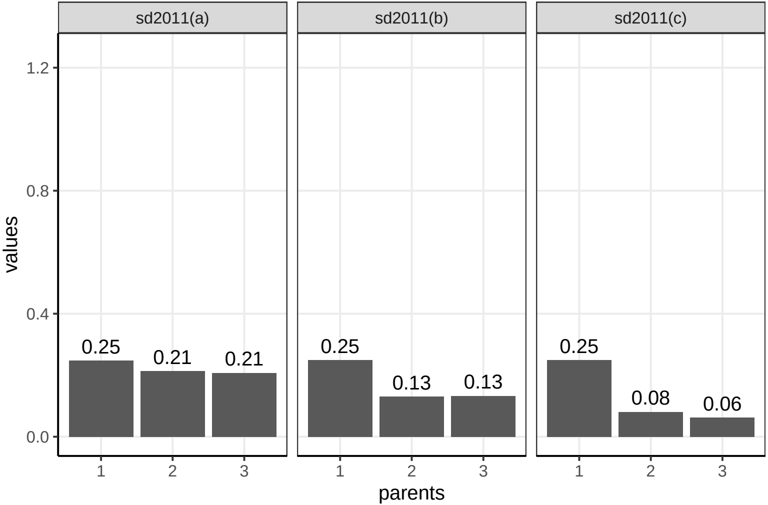


VARIABLE: NOFRIEND

Figure 2: Doesn't capture rounding/discontinuity



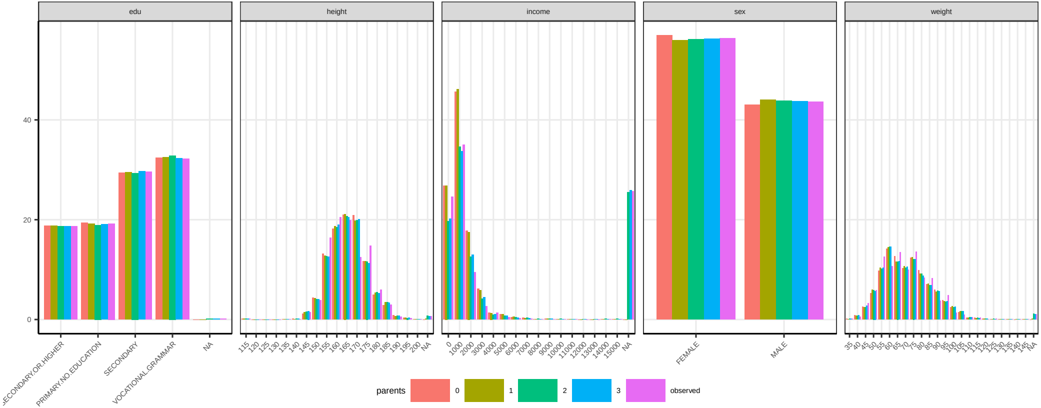
SD2011 - PMSE



sd2011(a)
sd2011(b)
sd2011(c)
values
1.2
0.8
0.4
0.0
0.25
0.21
0.21
0.25
0.13
0.13
0.25
0.08
0.06
1
2
3
parents

DATASYNTHESIZER: SELECTED VARIABLES

Figure 3: No missings if parents < 2



Section 3b): Know your generator (CTGAN)

TUNING CTGAN

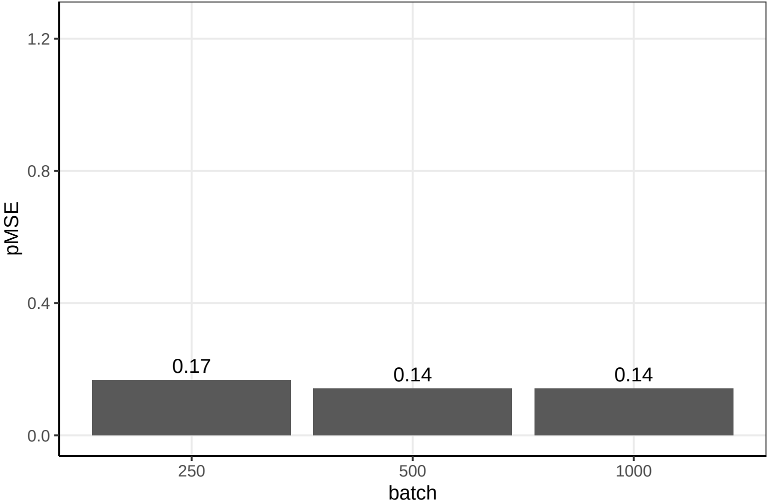
- Batch size (constant steps)
- Epochs (constant batch size)
- Dimensions (2 hyperparameters)
  - embedding_dim (int): Size of the random sample passed to the Generator. Defaults to 128.
  - dimensionality - 2 hyperparameters, but same value for each
    - discriminator_dim (tuple or list of ints): Size of the output samples for each one of the Discriminator Layers. A Linear Layer will be created for each one of the values provided. Defaults to (256, 256).
    - generator_dim (tuple or list of ints): Size of the output samples for each one of the Residuals. A Resiudal Layer will be created for each one of the values provided. Defaults to (256, 256).

BATCH SIZE, EPOCHS, AND STEPS

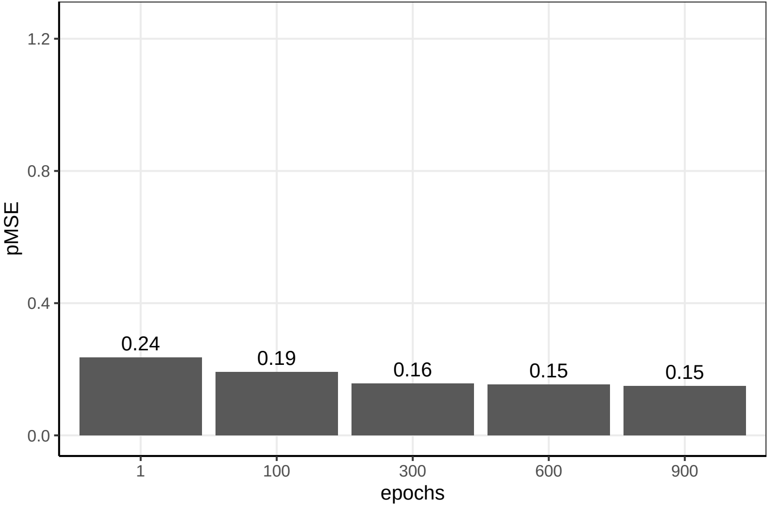
Table 1

N	Batch size	Steps per Epoch	Epochs	Actual Steps
5.000	100	50	60	3,000
5.000	250	20	150	3,000
5.000	500	10	300	3,000
5.000	1.000	5	600	3,000
5.000	500	10	100	1,000
5.000	500	10	300	3,000
5.000	500	10	600	6,000
5.000	500	10	900	9,000

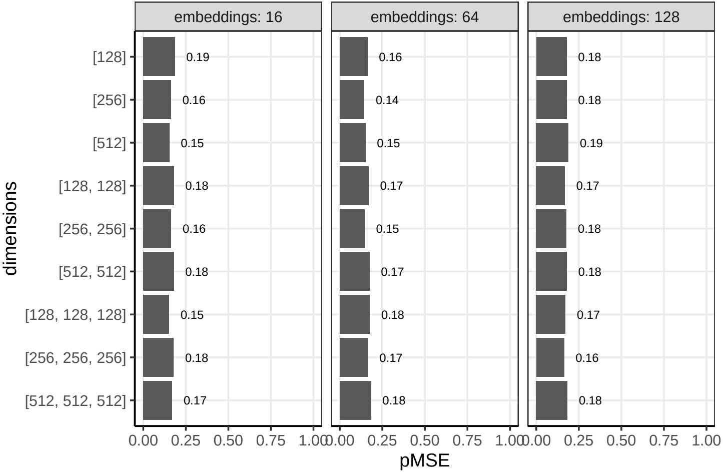
CTGAN: EFFECT OF BATCH SIZE (CONSTANT STEPS)



CTGAN: EFFECT OF EPOCHS (CONSTANT BATCH SIZE)



CTGAN: EFFECT OF DIMENSIONS



Section 3b): Know your generator (Synthpop)

Section 4: Conclusion