Deliverable 1 Methods for sunthetic data generation and utility and privacy measures and

S Y N T H E T I C DATA C H A L L E N G E

We have applied the following methods on the datasets:

Dataset SATGPA

- Fully Conditional Specification (CART with RSynthpop)
    - Two different ordering
- Fully Conditional Specification (Parametric RSynthpop)
- Differential Privacy (DP pgm – Minutemen)
- Deep Learning (SDGym) with GAN

Dataset ACS

- Fully Conditional Specification (CART with RSynthpop) – Dataset ACS
- With all variable on the default
- Fully Conditional Specification (CART with RSynthpop) – Dataset ACS
    - Stratified by state and by state and year
- Differential Privacy (DP pgm – Minutemen)
- Deep Learning (SDGym)

Data set SATGPA

We have applied different methods for the synthesis: FCS with the package synthpop, by applying the default method CART and with default for parametric methods, DPgm Minutemen and finally the SDgym Minutemen. In the application of the CART we have used different ordering in the sequence variables to be synthetised.

In the following we report the measures of utility and privacy.

As utility we focus on the propensity score based ones. For privacy we applied both the function replicate.uniques that measure the number of uniques from synthpop and the apparent match distribution privacy metrics.

Initial risk measures measured with sdcMicro

Key Variable	Number of categories	Mean size	Size of smallest (>0)
sex	2 (2)	500.000 (500.000)	470 (470)
hs_gpa	31 (31)	32.258 (32.258)	1 (1)

Number of observations violating

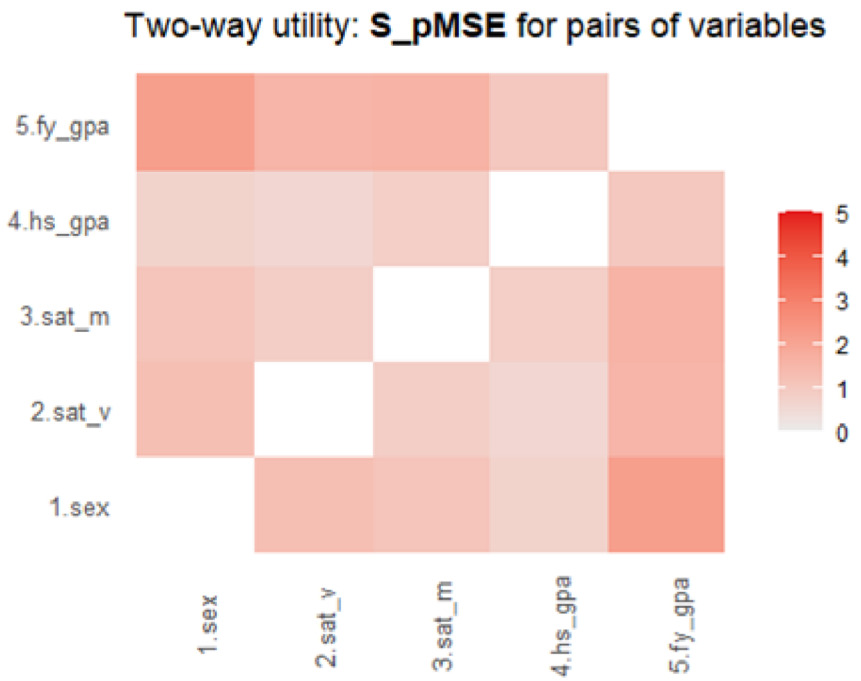
- 2-anonymity: 8 (0.800%)

- 3-anonymity: 14 (1.400%)

- 5-anonymity: 34 (3.400%)

Synthesis by CART

In this first synthesis, the ordering is as the original dataset, i.e.: sex, sat_v,Ssat_m, hs_gpa, fy_gpa. The method for synthesis is sample for the first variable and CART for the others.




```
Selected utility measures
##     pMSE   S_pMSE
## 0.052036 1.653087
```

```
## Medians and maxima of selected utility measures for all tables compared
##        Medians  Maxima
## pMSE    0.0012  0.0024
## S_pMSE  1.0523  2.1725
## df     24.0000 24.0000
```

```
Privacy measures:
As key variables:
sex,hs_gpa
sex,hs_gpa
 ****COUNTS****
33.333333    2
66.666667    1
**************
 Mean: 44.444444444444436 std: 19.245008972987524
max: 66.66666666666666
median: 33.33333333333333
Q3: 49.99999999999999
matched: 3
```

Percentage of replication

[1] 3.3

```
Standard risk measure from sdcMicro
```

Key Variable Number of categories Mean size Size of smallest (>0)

sex 2 (2) 500.000 (500.000) 470 (470)

hs_gpa 31 (31) 32.258 (32.258) 1 (1)

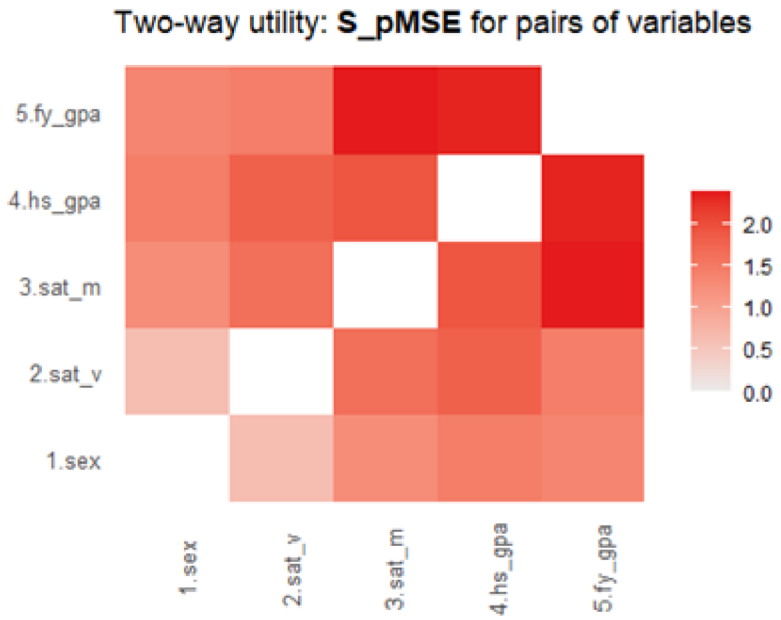
Infos on 2/3-Anonymity:

Number of observations violating

- 2-anonymity: 4 (0.400%)
- 3-anonymity: 12 (1.200%)
- 5-anonymity: 22 (2.200%)

Synthesis Parametric

This synthesis is applied with the same ordering as in the original dataset, I.e.: sex, sat_v,Ssat_m, hs_gpa, fy_gpa. The method for synthesis is sample for the first variable and default parametric for the others, that is: normrank.



Selected utility measures

pMSE S_pMSE

0.056873 1.661561

```
## Medians and maxima of selected utility measures for all tables compared
##        Medians  Maxima
## pMSE    0.0023  0.0036
## S_pMSE  1.5381  2.3673
## df     24.0000 24.0000
```

******COUNTS******

0.0 2

****

Mean: 0.0 std: 0.0

max: 0.0

median: 0.0

Q3: 0.0

matched: 2 25.0%

```
Form synthpop Percentage of replication by synthpop
## [1] 0
```

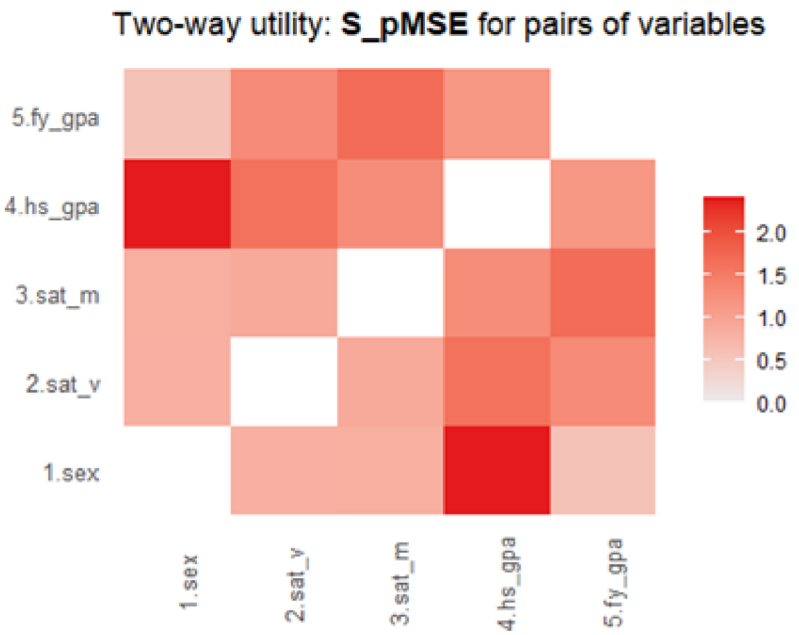
From sdcMicro Infos on 2/3-Anonymity:

Number of observations violating

- 2-anonymity: 6 (0.600%)
- 3-anonymity: 10 (1.000%)
- 5-anonymity: 25 (2.500%)

CART with different sequence for FCS

The last application of the FCS was done applying a different sequence for the synthesis: sex hs_gpa sat_v sat_m fy_gpa. The method for synthesis is sample for the variable sex and CART for the others.




```
## Medians and maxima of selected utility measures for all tables compared
##         Medians  Maxima
## pMSE     0.0015  0.0025
## S_pMSE   1.1974  2.3981
## df      24.0000 24.0000


Selected utility measures

    pMSE   S_pMSE

0.051693 1.634365
```

Privacy measures:

******COUNTS******

0.0 1

****

Mean: 0.0 std: nan

max: 0.0

median: 0.0

Q3: 0.0

matched: 1

From synthpop Percentage of replication:

[1] 1.5

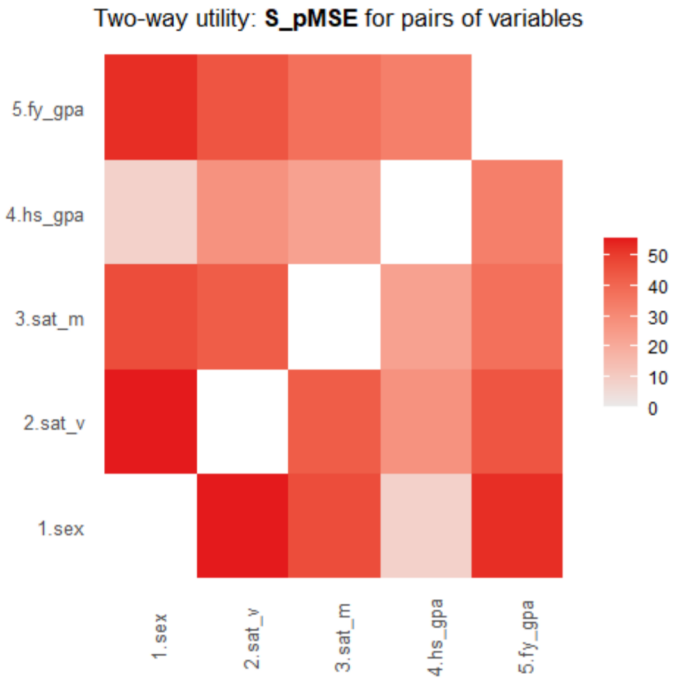
Risk evaluation:

- 2-anonymity: 2 (0.200%)
- 3-anonymity: 12 (1.200%)
- 5-anonymity: 33 (3.300%)

DPpgm (Minutemen)

For the application of the DPpgm, we used PUMA as geographical variable. We first applied the discretization made by default on numeric variables. To increase the performance on numeric variables in second run we changed the type into categorical and in the third run we increased the number of bins.

MINUTEMEN default discretization



Medians and maxima of selected utility measures for all tables compared

```
        Medians  Maxima
pMSE     0.0402  0.0707
S_pMSE 40.5115 55.6336
df     24.0000 24.0000
```

MINUTEMEN 2

With all categorical

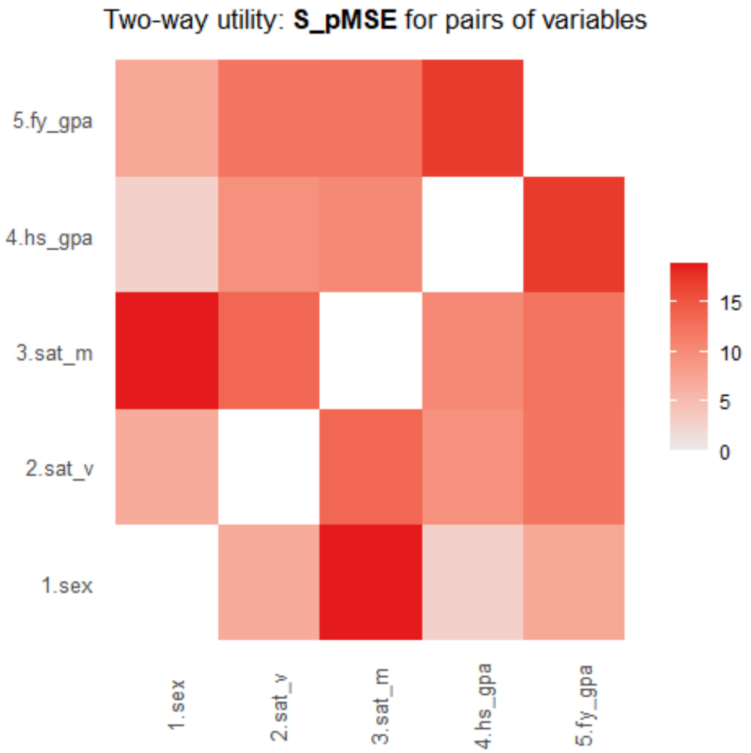
Medians and maxima of selected utility measures for all tables compared

Medians Maxima

pMSE 0.0152 0.0263

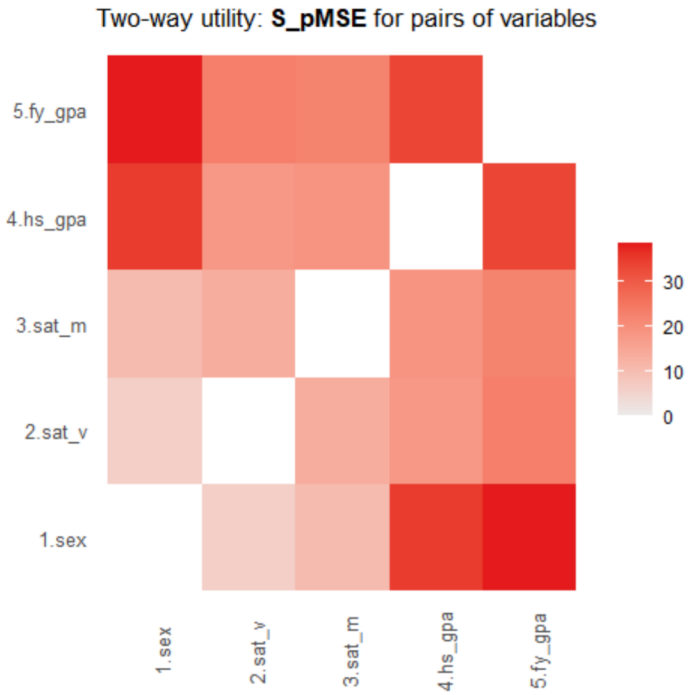
S_pMSE 11.3519 18.9010

df 24.0000 24.0000



MINUTEMEN with increased number of bins

```
Medians and maxima of selected utility measures for all tables compared

       Medians  Maxima
pMSE    0.0245  0.0505
S_pMSE 20.5952 38.5202
df     24.0000 24.0000
```



Private evaluation

sex,sat_v,hs_gpa

Mean: nan std: nan

max: nan

median: nan

Q3: nan

matched: 0

From synthtpop: Percentage of replication: 0

From sdcMicro

Infos on 2/3-Anonymity:

Number of observations violating

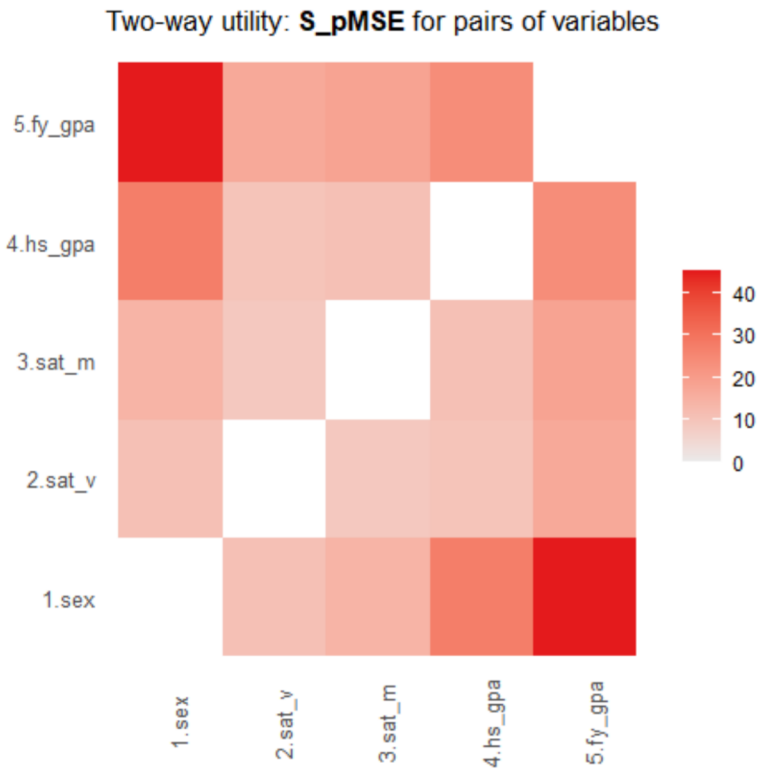
- 2-anonymity: 2 (0.200%)
- 3-anonymity: 4 (0.400%)
- 5-anonymity: 25 (2.503%)

DEEP LEARNING with GAN

We applied a tGAN (from the SDgym framework, ref: https://github.com/sdv-dev/SDGym) that uses a conditional GAN to address tabular table including discrete and continuous variables in the table.

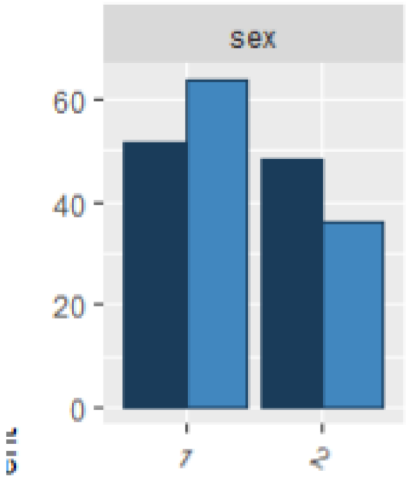
After initial experiments we set the training epochs equal to 500, we set the number of input tensors equal to 256^3.

We left the other parameters as default.3999999



```
Medians and maxima of selected utility measures for all tables compared

        Medians  Maxima
pMSE     0.0161  0.0357
S_pMSE 15.4588 44.9924
df     24.0000 24.0000
```

Note that in this latter case that the marginal variable SEX is not well reproduced.



****COUNTS****

0.0 1

****

Mean: 0.0 std: nan

max: 0.0

median: 0.0

Q3: 0.0

matched: 1

Percentage of replications=0

With tried also the version with the Gaussian copulas, but the results on the variables “sex” are still unreliable.

Data set IPUMS-ACS

For this dataset we have tested as synthesis techniques the CART in the R package synthpop, the DPpgm – Minutemen and finally the tGAN (from the SDgym framework, ref: https://github.com/sdv-dev/SDGym)

We decided to have a full synthesis of all variables.

First we created an age class variable as follows:

```
IL_OH_10Y_PUMS$AGEC<-ifelse(  IL_OH_10Y_PUMS$AGE<=25,1,

  ifelse(IL_OH_10Y_PUMS$AGE<=30,2,ifelse(

        IL_OH_10Y_PUMS$AGE<=35,3,ifelse(

        IL_OH_10Y_PUMS$AGE<=40,4,ifelse(the

        IL_OH_10Y_PUMS$AGE<=45,5,ifelse(

        IL_OH_10Y_PUMS$AGE<=50,6,ifelse(

       IL_OH_10Y_PUMS$AGE<=55,7,ifelse(

         IL_OH_10Y_PUMS$AGE<=60,8,ifelse(

           IL_OH_10Y_PUMS$AGE<=65,9,ifelse(

             IL_OH_10Y_PUMS$AGE<=70,10,11

       ))))))))))
```

Then we choose to create a cross classified class_age by sex variable to guarantee the joint distribution for age and sex:

```
IL_OH_10Y_PUMS$SEXAGEC<-as.factor(IL_OH_10Y_PUMS$SEX*100+IL_OH_10Y_PUMS$AGEC)
```

Synthesis by CART (synthpop)

The first synthetised data set, all variables, has been carried out without any stratification.

The individual weights have been kept among the variables to be synthetised and in the models for the other variables.

The list of variables are:

"GQ","SEXAGEC","MARST","PERWT","CITIZEN","HISPAN","SPEAKENG","RACE","EDUC","WORKEDYR","LABFORCE","EMPSTAT","EMPSTATD","WRKLSTWK","ABSENT","LOOKING","AVAILBLE","WRKRECAL","INCTOT","INCWAGE","INCWELFR","INCINVST","INCEARN","POVERTY","DEPARTS","ARRIVES","PUMA","YEAR","HCOVANY","HCOVPRIV","HINSEMP","HINSCAID","HINSCARE"

The method for synthesis is sample for the first variable and CART for all the others. The default prediction matrix has been applied.

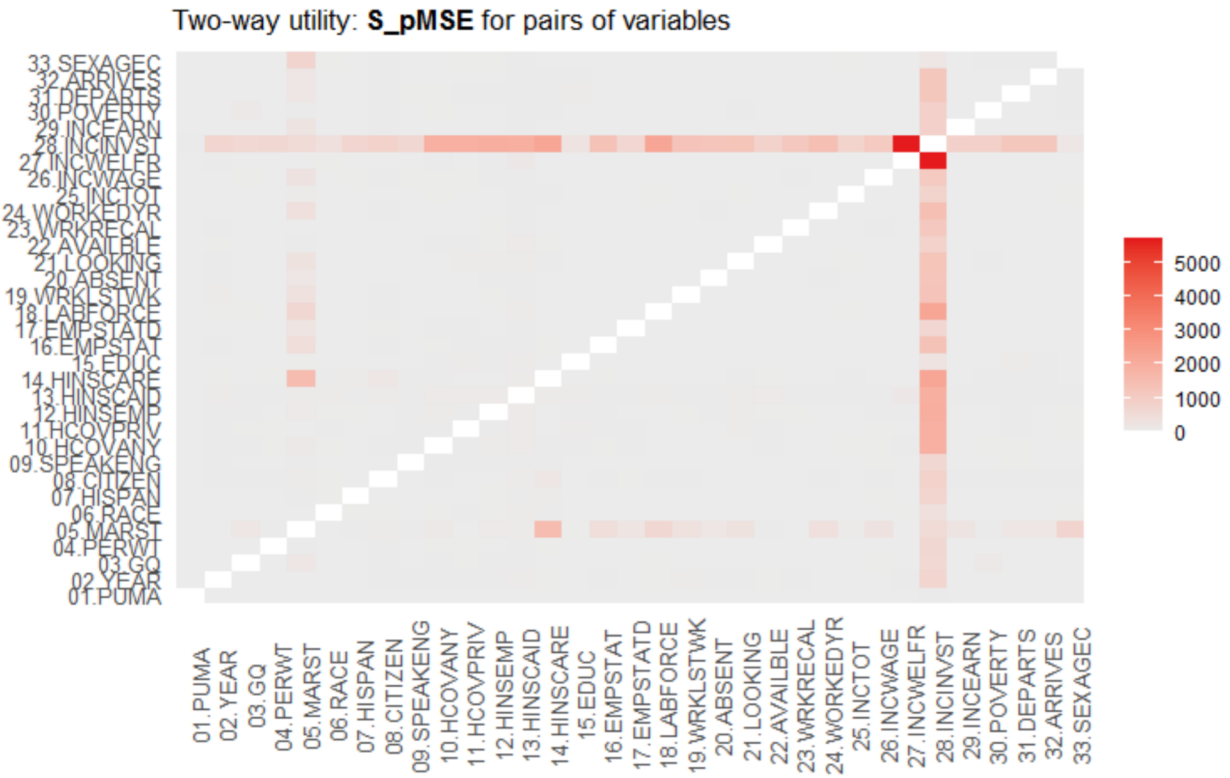
The resulting measures of utility are

```
Selected utility measures(utility.gen)
```

Utility score calculated by method: cart

pMSE	S_pMSE
0.002331	344.927791

Two-way utility: S_pMSE value plotted for 528 pairs of variables.



Variable combinations with worst 5 utility scores (S_pMSE):

27.INCWELFR:28.INCINVST	14.HINSCARE:28.INCINVST	18.LABFORCE:28.INCINVST	5717.303
2241.637	2239.843		

13.HINSCAID:28.INCINVST	12.HINSEMP:28.INCINVST

```
     1983.809          1957.466

## Medians and maxima of selected utility measures for all tables compared
##        Medians  Maxima
## pMSE    0.00    0.0059
## S_pMSE  8.6835  5717.303
## df     11.0000 3981.0000


Privacy measures:

Mean: 53.91862875758922 std: 18.652115505088286
max: 100.0
median: 54.166666666666664
Q3: 66.66666666666666
matched: 2251     21.54%
```

Then we performed another synthesis without individual weights: this new S1 will be the reference for the following comparisons.

As a second step we tried different synthesis stratifying according to state first (S2 data), and state by year secondly (S3 data).

On the finest stratification, the model was also applied on a different sequence of variables to address the issues encountered on the utility for the synthetic values of some of the variables (S4 data).

However, we could not improve as the following comparisons between the utility evaluations show. Indeed, the utility score is 1.6 times higher for the S1 synthesis (the basic not stratified one with weights) compared to the S3 and 1.3 times higher for the S2 synthesis compared to the S3 respectively.

```
Utility ratio S1 to S3: 1.6

Utility ratio S2 to S3: 1.3

Utility ratio S3b to S3: 2.5
```

Next step is to calculate utilities from all one-way marginal for the S2 and S3 model

```
S1:

Two-way utility: S_pMSE value plotted for 496 pairs of variables.

Variable combinations with worst 5 utility scores (S_pMSE):

31.INCWELFR:32.INCINVST 18.HINSCARE:32.INCINVST

              1496.9484                661.5452

22.LABFORCE:32.INCINVST    09.MARST:18.HINSCARE

               599.6349                597.2406
```

```
  11.HISPAN:18.HINSCARE
              574.3617


Medians and maxima of selected utility measures for all tables compared
       Medians    Maxima
pMSE     0.000    0.0031
S_pMSE  12.673 1496.9484
df      11.000 3695.0000


S2:
Two-way utility: S_pMSE value plotted for 465 pairs of variables.
Variable combinations with worst 5 utility scores (S_pMSE):
  09.MARST:18.HINSCARE 12.CITIZEN:18.HINSCARE
             1471.9184               840.8925
  09.MARST:22.LABFORCE  11.HISPAN:18.HINSCARE
              738.5248               702.8719
   09.MARST:39.SEXAGEC
              621.8690
Medians and maxima of selected utility measures for all tables compared
       Medians    Maxima
pMSE    0.0000    0.0049
S_pMSE 18.3994 1471.9184
df     11.0000 3695.0000
S3:
Two-way utility: S_pMSE value plotted for 465 pairs of variables.
Variable combinations with worst 5 utility scores (S_pMSE):
  09.MARST:18.HINSCARE 12.CITIZEN:18.HINSCARE   09.MARST:22.LABFORCE
             1845.2549              1105.1416               833.5808
```

```
11.HISPAN:18.HINSCARE    09.MARST:39.SEXAGEC

              821.4942               775.6502

Medians and maxima of selected utility measures for all tables compared

       Medians    Maxima
pMSE    0.0000    0.0061
S_pMSE 21.4829 1845.2549
df     11.0000 3695.0000
```

Privacy measure for the first stratification by state (S2)

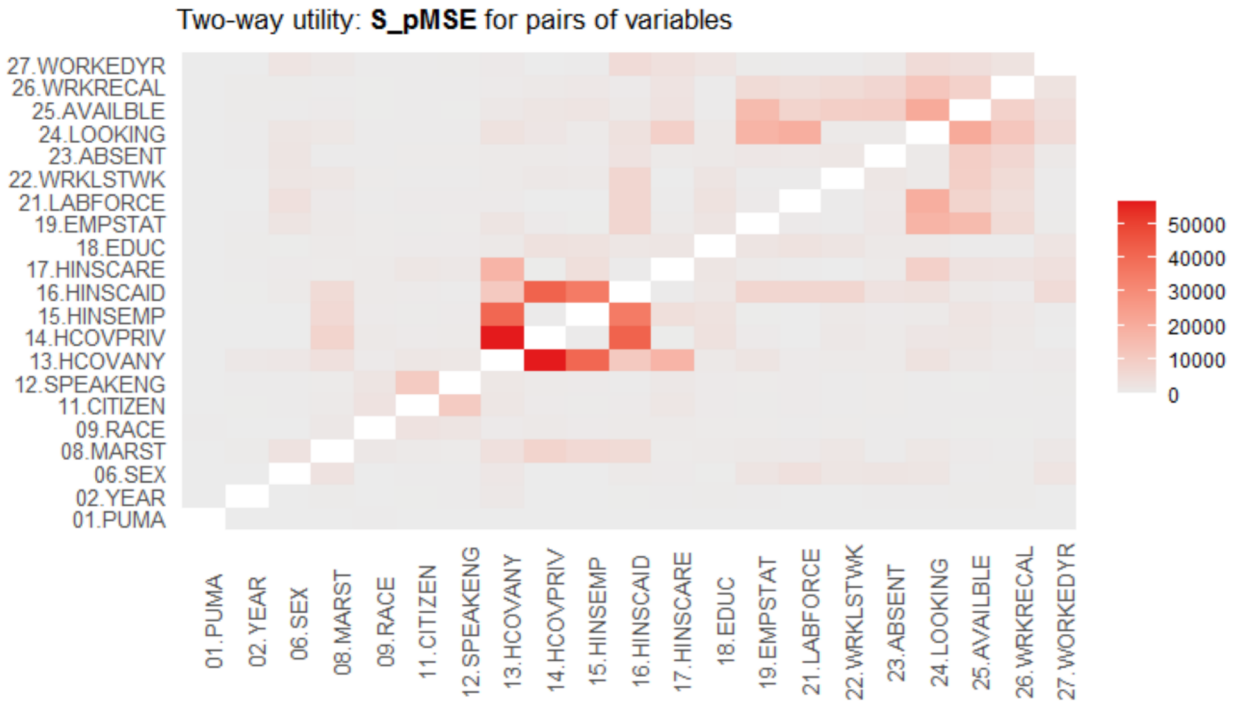
```
Mean: 49.879078450507016 std: 16.986848985137772
max: 88.46153846153845
median: 50.0
Q3: 65.38461538461539
matched: 1813
```

Privacy measure for the first stratification by state and year (S3)

```
Mean: 49.33496467616529 std: 16.962843673114435
max: 96.15384615384616
median: 50.0
Q3: 61.53846153846154
matched: 1949
```

DPpgm (MINUTEMEN)

```
utility.gen.data.frame(object = synPython, data = origg, vars = seque.Pyt)

Null utilities simulated from a permutation test with 50 replications.

************************************************************

Warning: null pMSE resamples failed to split 45 time(s) from 50

************************************************************

Selected utility measures

        pMSE       S_pMSE
1.900210e-01 1.953315e+05
```



Variable combinations with worst 5 utility scores (S_pMSE):

13.HCOVANY:14.HCOVPRIV	14.HCOVPRIV:16.HINSCAID	13.HCOVANY:15.HINSEMP	56280.39
42116.30	40841.40	15.HINSEMP:16.HINSCAID	
24.LOOKING:25.AVAILBLE			
35321.55	21002.93		

Medians and maxima of selected utility measures for all tables compared

	Medians	Maxima
pMSE	0.0006	0.0142
S_pMSE	599.4839	56280.3850
df	9.0000	1990.0000

Privacy measures

Mean: 40.33116883116883 std: 13.764062652408288

max: 66.66666666666666

median: 41.66666666666667

Q3: 54.166666666666664

matched: 1925

Conclusions

For both datasets the DPgym allowed better results in terms of measure of privacy (on the apparent matches) whereas the CART achieves better results in the utility measures. Suggesting that for educational purpose the dataset produced by DPgym is satisfactory but for analytical purposes still there is a need for improvement.

Further analysis intervening on the order of variables for cart models, or also changing methods, and as regards DPgym models regarding the discretization of numeric variables, are needed to improve the output and to obtain a better compromise between utility and risk.

The application of the tGAN had a very curious performance, producing a very unexpected distribution of the synthetic variable for "SEX", with a large unbalance distribution between female/males. We could not succeed in understanding and fixing this issues.

We thank the organizers for this very inspiring and useful challenge.