Fully Conditional Specification (FCS)

Dataset: SATGPA

We used the `Synthpop` package with R to synthesize and evaluate the synthetic dataset.

Data Exploration

Running a simple linear regression on the `SATGPA` dataset shows the following results:

```
lm(formula = fy_gpa ~ sat_v + sat_m + hs_gpa, data = satgpa)

Coefficients:
(Intercept)        sat_v        sat_m       hs_gpa
-0.86693      0.01646      0.01240      0.58007
```

```
lm(formula = fy_gpa ~ sat_v + sat_m + hs_gpa, data = satgpa[satgpa$sex == 1, ])

Coefficients:
(Intercept)        sat_v        sat_m       hs_gpa
   -0.79878      0.01387      0.02057      0.43220
```

```
lm(formula = fy_gpa ~ sat_v + sat_m + hs_gpa, data = satgpa[satgpa$sex == 2, ])

Coefficients:
(Intercept)        sat_v        sat_m       hs_gpa
   -1.066058     0.017429     0.009975     0.684462
```

We see a clear difference in the factor weights depending on the `sex`. This means that there are significant differences in the relations between the columns depending on sex.

```
 sex            sat_v           sat_m          hs_gpa          fy_gpa
Min.   :1.000   Min.   :24.00   Min.   :29.0   Min.   :1.800   Min.   :0.000
1st Qu.:1.000   1st Qu.:43.00   1st Qu.:49.0   1st Qu.:2.800   1st Qu.:1.980
Median :1.000   Median :49.00   Median :55.0   Median :3.200   Median :2.465
Mean   :1.484   Mean   :48.93   Mean   :54.4   Mean   :3.198   Mean   :2.468
3rd Qu.:2.000   3rd Qu.:54.00   3rd Qu.:60.0   3rd Qu.:3.700   3rd Qu.:3.020
Max.   :2.000   Max.   :76.00   Max.   :77.0   Max.   :4.500   Max.   :4.000
```

Data Synthesis

We convert the `sex` column to a categorical variable and treat the rest as numeric variables.

In FCS data synthesis, the order of synthesis is up to discretion. We try 2 different orders and compare them.

Order 1: sex - sat_v - sat_m - hs_gpa - fy_gpa

```
 sex            sat_v           sat_m          hs_gpa          fy_gpa
Min.   :1.000   Min.   :24.00   Min.   :29.00   Min.   :1.800   Min.   :0.000
1st Qu.:1.000   1st Qu.:43.00   1st Qu.:49.00   1st Qu.:2.750   1st Qu.:1.940
Median :1.000   Median :49.00   Median :55.00   Median :3.200   Median :2.420
Mean   :1.483   Mean   :48.87   Mean   :54.28   Mean   :3.201   Mean   :2.439
3rd Qu.:2.000   3rd Qu.:54.00   3rd Qu.:59.00   3rd Qu.:3.700   3rd Qu.:3.000
Max.   :2.000   Max.   :76.00   Max.   :77.00   Max.   :4.000   Max.   :4.000
```

`utility.gen` produced propensity scores.

pMSE	S_pMSE
0.052702	1.662050

```
lm(formula = fy_gpa ~ sat_v + sat_m + hs_gpa, data = synthetic$syn)

Coefficients:
(Intercept)        sat_v        sat_m       hs_gpa
   -0.66371      0.01245      0.00757      0.65085
```

```
lm(formula = fy_gpa ~ sat_v + sat_m + hs_gpa, data = synthetic$syn[synthetic$syn$sex == 1, ])

Coefficients:
(Intercept)        sat_v        sat_m       hs_gpa
  -0.467549     0.007142     0.014192     0.532434
```

```
lm(formula = fy_gpa ~ sat_v + sat_m + hs_gpa, data = synthetic$syn[synthetic$syn$sex == 2, ])

Coefficients:
(Intercept)        sat_v        sat_m       hs_gpa
  -0.896868     0.015416     0.006148     0.720324
```

Order 2: sex - hs_gpa - sat_v - sat_m - fy_gpa

```
 sex            sat_v           sat_m          hs_gpa          fy_gpa
Min.   :1.000   Min.   :24.00   Min.   :29.00   Min.   :2.000   Min.   :0.000
1st Qu.:1.000   1st Qu.:43.00   1st Qu.:48.00   1st Qu.:2.800   1st Qu.:1.910
Median :1.000   Median :49.00   Median :55.00   Median :3.250   Median :2.420
Mean   :1.467   Mean   :48.94   Mean   :54.35   Mean   :3.207   Mean   :2.455
3rd Qu.:2.000   3rd Qu.:54.00   3rd Qu.:60.00   3rd Qu.:3.700   3rd Qu.:3.040
Max.   :2.000   Max.   :76.00   Max.   :77.00   Max.   :4.500   Max.   :4.000
```

`utility.gen` produced propensity scores.

pMSE	S_pMSE
0.048575	1.543723

```
lm(formula = fy_gpa ~ sat_v + sat_m + hs_gpa, data = synthetic$syn)

Coefficients:
(Intercept)        sat_v        sat_m       hs_gpa
   -1.02722      0.01805      0.01435      0.56734
```

```
lm(formula = fy_gpa ~ sat_v + sat_m + hs_gpa, data = synthetic$syn[synthetic$syn$sex == 1, ])

Coefficients:
(Intercept)        sat_v        sat_m       hs_gpa
   -0.78263      0.01421      0.01715      0.48421
```

```
lm(formula = fy_gpa ~ sat_v + sat_m + hs_gpa, data = synthetic$syn[synthetic$syn$sex == 2, ])

Coefficients:
(Intercept)        sat_v        sat_m       hs_gpa
    1.34464      0.02027      0.01641      0.61536
```

Conclusions

Dataset	Intercept	sat_v	sat_m	hs_gpa
SATGPA	-0.86693	0.01646	0.01240	0.58007

Dataset	Intercept	sat_v	sat_m	hs_gpa
SATGPA[sex=1]	-0.79878	0.01387	0.02057	0.43220
SATGPA[sex=2]	-1.066058	0.017429	0.009975	0.684462
Order1	-0.66371	0.01245	0.00757	0.65085
Order1[sex=1]	-0.467549	0.007142	0.014192	0.532434
Order1[sex=2]	-0.896868	0.015416	0.006148	0.720324
Order2	-1.02722	0.01805	0.01435	0.56734
Order2[sex=1]	-0.78263	0.01421	0.01715	0.48421
Order2[sex=2]	-1.34464	0.02027	0.01641	0.61536

One thing we noticed is that `hs_gpa` has much higher weight than either `sat_v` or `sat_m` when predicting `fy_gpa`. The result of this is that when we synthesize the data starting with `sat_v` and `sat_m`, we end up creating a dataset where `sat_v` and `sat_m` are underweighed and `hs_gpa` is overweighed. When we synthesized the data with `hs_gpa` before `sat_v` and `sat_m`, to reflect the relative order of importance, we got a dataset where the relative relationship are much closer to the original dataset.

Use Cases

1. Testing Analysis

For analysis purposes, FCS can potentially be problematic. Depending on the order of variable synthesis, the resulting dataset can have very different properties. This variance between datasets generated using different variable orders is large enough that drawing conclusions from analysis performed on synthesized data can be problematic.

2. Education

For the purpose of education, the FCS synthesis model preserves the most salient features of the dataset, such as the relative importance of variables, and the `sex` specific difference in factor weights.

3. Testing Technology

FCS synthesized data is perfectly fine for testing technology. There may be one small caveat which is that the synthesized data may fail to replicate the full range of data in the original dataset. For example, in the Order1 dataset, the max of `hs_gpa` is 4.0 whereas in the original dataset it was 4.5. This can potentially break certain components in a complex system that are sensitive to the range of its inputs.

4. Releasing to Public

FCS is fine for releasing data to the public as long as members of the public (for example, journalists) do not use the data for serious analytical purposes and shape public opinion based on it. Given that FCS synthesized data is not robust to changes in synthesis order, it could be a problem if members of the public propagate conclusions based on analysis of data synthesized with FCS.

Method: Information Preserving Statistical Obfuscation (IPSO)

Method category: Sequential modelling

We produced a partially synthetic dataset based on the satgpa data using the IPSO method, treating the sex variable as the matrix of non-confidential variables X, and the remaining variables as the matrix Y of variables to be synthesized. We excluded the variable sat_sum when applying IPSO, as this variable is a sum of sat_v and sat_m in the original data. The synthetic version of sat_sum was generated by summing the synthetic versions of sat_v and sat_m.

We reasoned that sex on its own cannot be used to identify a specific person in the data set, and therefore it made sense to treat this as a non-confidential variable. Furthermore, many natural statistical analyses on the data would treat sex as an independent variable in a multiple regression model with grade variables as dependent variables. Therefore it seemed reasonable to use a similar regression model in the IPSO method. As is noted in the *HLG-MOS Synthetic Data Guide*, IPSO relies on the strong assumption of multivariate normality for all synthesized variables. We assessed the validity of this assumption by plotting histograms of the different variables for each sex. We observed that the distribution of the variables sat_v, sat_m and fy_gpa followed a roughly normal distribution, but distribution of hs_gpa did not. We therefore did not expect this method to emulate the distribution of hs_gpa well, and we evaluate the difficulty this may pose for specific use case scenarios below.

Tooling

We used the RegSDCipso function from the RegSDC R package to generate the synthetic dataset. To evaluate the utility of the synthetic dataset, we performed different analyses in base R as well as using the functions compare, utility.gen, utility.tables, and multi-compare from the synthpop R package. To evaluate the disclosure risk of the synthetic dataset, we used the function replicated.uniques from the synthpop package, as well as code written in base R for comparing matching records between datasets based on unique values for quasi-identifier variables (see details in 'Disclosure risk evaluation' section below).

Evaluation of data for different use cases

Various results relating to the utility and disclosure risk of the synthetic dataset are described in the results section below. In this section we evaluate the appropriateness of the synthetic dataset for the four different use cases, referring to the analyses described in the results section.

Use Case 1: Testing analysis

Utility: We did not find this method appropriate for producing synthetic data for users wishing to test their own specific analyses and models. By design, this method produces synthetic data that preserves means for variables, and the estimates of regression coefficients and the covariance matrix for a multiple regression model of the form $\boldsymbol{Y}' = \beta \boldsymbol{X} + \varepsilon$ are likewise preserved. Hence this data may perform well for some basic regression analyses. However, many basic statistical characteristics of the synthesized data differ quite obviously from the original data, such as the maximum and minimum values of variables and the distribution of the hs_gpa variable. Moreover, the pMSE score for the synthetic dataset indicates a strong lack of similarity between the synthetic and original data.

Privacy: In terms of disclosure risk, we did not identify any issues with the use of this synthetic dataset for this use case. There are no replicated uniques in the synthetic data, and for those unique values of the quasi-identifier pair consisting of sex and hs_gpa which appeared in the original set and also appeared in the synthetic dataset, there were no cases of agreement on the remaining variables. However, because of the low utility of this synthetic dataset, we would nevertheless not recommend it for this use case.

Use Case 2: Education

Utility: Similarly to the Testing Analysis use case, we do not find this synthetic dataset entirely appropriate for the purposes of teaching students and new learners the latest in data science methods. For this use case, it is very important for the data to yield realistic results, and as we observed, there are many very basic statistical characteristics of our synthetic data (such as maxima and minima of variables and the distribution of hs_gpa) that differed widely from the ground truth. A dataset where the synthesized variables follow a multivariate normal distribution by construction is not very useful when this is not consistent with reality, and the modern data science methods being practised will certainly extend far beyond basic multiple linear regression.

Privacy: Once again, our analysis did not identify any clear disclosure risk for this synthetic data. We believe this synthesized dataset meets the medium confidentiality requirements of education purposes, but we would nevertheless not recommend this dataset for this use case because of the utility shortfalls.

Use Case 3: Testing technology

Utility: This synthetic dataset may be appropriate for testing complex systems that have been built to transform the data holding into another product. Assuming these systems do not somehow rely on the particular distribution of the variables, or the non-existence of certain extreme values, the fact that the synthetic data does not closely emulate many statistical characteristics of the original data should not be an issue. However, for this use case it may be simpler and more efficient to use a less advanced method, such as simply creating a dummy file if the system that transform the data has no strong analytical requirements of the data.

Privacy: Once again, as we did not identify any disclosure risk from our synthetic data, we believe it should meet the medium confidentiality requirements for testing technology.

Use Case 4: Releasing microdata to the public

Utility: As with the Testing Analysis use case, we would not recommend this synthetic dataset for this use case because of failure of the synthetic data to emulate basic statistical properties of the original dataset.

Privacy: Although we found the disclosure risk of this synthetic dataset to be low, we would still not recommend the data for this use case because of the low utility.

Results

Utility results

Summary statistics for the synthetic dataset were very similar to those for the original dataset (Figure 1). In particular, the means matched almost precisely between the two datasets. Although the IPSO method preserves means precisely in the synthetic dataset, we see a very small difference in the sat_v mean as a

result of rounding that was applied to the synthetic dataset to match the formatting of the variables in the original dataset. However, the extreme values of variables in the synthetic dataset were noticeably more extreme than for the original dataset, with the minimum for the synthetic dataset being lower than for the original dataset and the maximum being higher for all variables. There was also one record in the synthetic dataset with a negative value for fy_gpa.

Similarly, summary statistics for the two sexes agree quite well between the synthetic and original datasets (Figures 2 and 3), with the exception of minima and maxima, which were more extreme for the synthetic data. In particular, differences in means of variables for the two sexes which were seen in the original data also appeared in the synthetic data. The means of sat_m and fy_gpa in particular showed large differences between the sexes.

Summary statistics for original and synthetic data for IPSO-synthesized variables

sat_v	sat_m	hs_gpa	fy_gpa
Min. :24.00	Min. :29.0	Min. :1.800	Min. :0.000
1st Qu.:43.00	1st Qu.:49.0	1st Qu.:2.800	1st Qu.:1.980
Median :49.00	Median :55.0	Median :3.200	Median :2.465
Mean :48.93	Mean :54.4	Mean :3.198	Mean :2.468
3rd Qu.:54.00	3rd Qu.:60.0	3rd Qu.:3.700	3rd Qu.:3.020
Max. :76.00	Max. :77.0	Max. :4.500	Max. :4.000
sat_v	sat_m	hs_gpa	fy_gpa
Min. :23.00	Min. :24.0	Min. :1.260	Min. :-0.470
1st Qu.:43.00	1st Qu.:49.0	1st Qu.:2.850	1st Qu.: 1.970
Median :49.00	Median :54.0	Median :3.220	Median : 2.470
Mean :48.94	Mean :54.4	Mean :3.198	Mean : 2.468
3rd Qu.:54.00	3rd Qu.:60.0	3rd Qu.:3.520	3rd Qu.: 2.970
Max. :80.00	Max. :81.0	Max. :5.300	Max. : 5.410

Figure 1. Summary statistics for IPSO-synthesized variables for original data (top) and synthesized data (bottom)

Summary statistics for original and synthetic data for IPSO-synthesized variables

Records with sex=1

sat_v	sat_m	hs_gpa	fy_gpa
Min. :24.00	Min. :29.00	Min. :1.800	Min. :0.000
1st Qu.:43.00	1st Qu.:51.00	1st Qu.:2.700	1st Qu.:1.935
Median :49.00	Median :57.00	Median :3.100	Median :2.380
Mean :49.25	Mean :56.46	Mean :3.124	Mean :2.396
3rd Qu.:55.00	3rd Qu.:62.00	3rd Qu.:3.500	3rd Qu.:2.910
Max. :71.00	Max. :77.00	Max. :4.500	Max. :4.000
sat_v	sat_m	hs_gpa	fy_gpa
Min. :23.00	Min. :26.00	Min. :1.260	Min. :-0.470
1st Qu.:43.00	1st Qu.:51.00	1st Qu.:2.770	1st Qu.: 1.938
Median :49.00	Median :57.00	Median :3.140	Median : 2.395
Mean :49.26	Mean :56.47	Mean :3.124	Mean : 2.396
3rd Qu.:55.00	3rd Qu.:62.00	3rd Qu.:3.470	3rd Qu.: 2.910
Max. :80.00	Max. :81.00	Max. :4.620	Max. : 4.340

Figure 2. Summary statistics for records with sex=1. Statistics for original data shown on top and synthesized data on the bottom.

Summary statistics for original and synthetic data for IPSO-synthesized variables

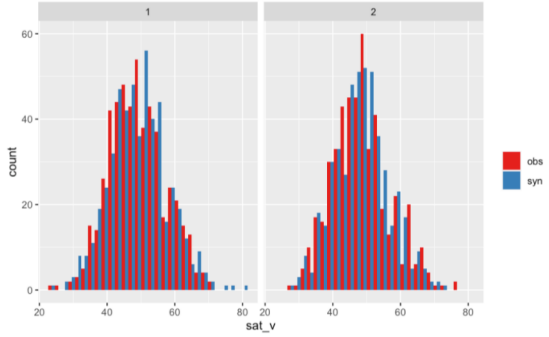
Records with sex=2

```
     sat_v           sat_m           hs_gpa          fy_gpa     
 Min.   :27.0   Min.   :31.00   Min.   :2.000   Min.   :0.000  
 1st Qu.:43.0   1st Qu.:47.00   1st Qu.:2.800   1st Qu.:2.020  
 Median :48.0   Median :52.00   Median :3.300   Median :2.560  
 Mean   :48.6   Mean   :52.19   Mean   :3.277   Mean   :2.545  
 3rd Qu.:54.0   3rd Qu.:58.00   3rd Qu.:3.750   3rd Qu.:3.120  
 Max.   :76.0   Max.   :77.00   Max.   :4.000   Max.   :4.000  
     sat_v           sat_m           hs_gpa          fy_gpa     
 Min.   :27.0   Min.   :24.00   Min.   :1.590   Min.   :0.450  
 1st Qu.:43.0   1st Qu.:47.00   1st Qu.:2.928   1st Qu.:2.027  
 Median :49.0   Median :52.00   Median :3.250   Median :2.565  
 Mean   :48.6   Mean   :52.19   Mean   :3.277   Mean   :2.545  
 3rd Qu.:54.0   3rd Qu.:57.00   3rd Qu.:3.592   3rd Qu.:3.053  
 Max.   :72.0   Max.   :76.00   Max.   :5.300   Max.   :5.410  
```

Figure 3. Summary statistics for records with sex=2. Statistics for original data shown on top and synthesized data on the bottom.

We also compared the distributions of variables by sex for the original and synthetic datasets using the multi.compare function from the synthpop package. The results are shown in Figure 4. The distributions of variables by sex are reasonably similar between the original and synthetic data, except in the case of the variable hs_gpa, where the IPSO method performed poorly due to non-normality of the original data.

Distributions of variables by sex in original and synthetic data sets



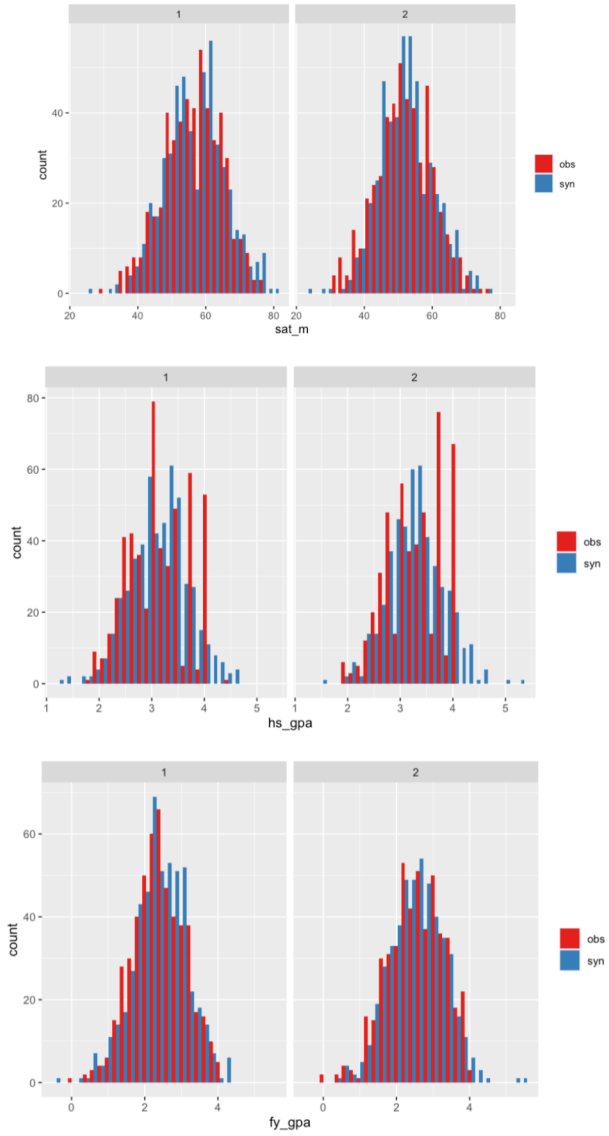


Figure 4. Distributions of variables by sex in original and synthetic datasets

The pMSE score for this synthetic dataset was 0.193 and the S_pMSE score was 4.94, as computed by the utility.gen function from synthpop using the CART method. These scores could likely be improved by using a method does not rely on the strong assumption of multivariate normality. We also produced

S_pMSE scores for pairs of variables using the utility.tables function from synthpop (Figure 5). We see that the S_pMSE score is worst for pairs of variables involving the variable hs_gpa, reflecting the fact that this variable had a particularly non-normal distribution. We used this information to **tune** our synthesis method to improve the synthesis of this variable. This is detailed in the description of the hybrid synthesis method below.

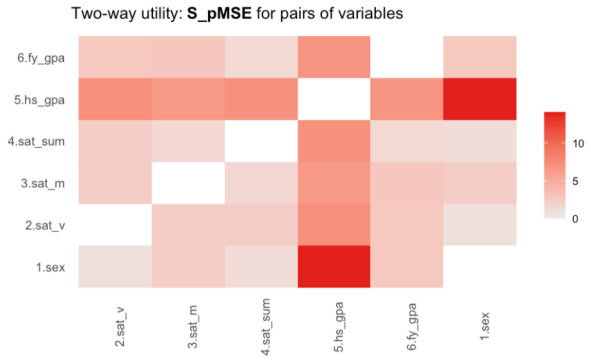


Figure 5. Heat map of two-way utility for the synthetic dataset

We also observed that for multiple regression models based on the original and synthetic data, the regression coefficients and the estimates of the covariance matrices were the same for the original and synthetic data. This is a built-in feature of the IPSO method.

Privacy results

In order to evaluate the disclosure risk of our synthetic dataset, we checked for any unique records in the original dataset which also appeared in the synthetic dataset. We did this using the replicated.uniques function from synthpop. There were no unique records in the original dataset which also appeared in the synthetic dataset. This is a basic measure which offers some assurance of low disclosure risk from our synthetic data.

As a further evaluation, we also considered pairs of values for sex and hs_gpa which were unique in the original dataset. There were eight such pairs. For each of the eight unique pairs, we checked if the pair also appeared in the synthetic dataset. In the case that a unique pair reappeared in the synthetic dataset, we checked for matches on the remaining variables. The eight unique pairs from the original dataset appeared fourteen times in the synthetic dataset. We observed that whenever a unique pair from the original dataset appeared in the synthetic dataset, there was total disagreement between the original and synthetic datasets on the remaining variables. This provided assurance that the values of other variables could not be inferred from knowing sex and hs_gpa alone.

Method: Hybrid between IPSO and Fully Conditional Specification

Method category: Sequential modelling

In order to improve on the IPSO synthesis method, we used Fully Conditional Specification with CART to synthesize the variable hs_gpa. Again, we produced a partially synthetic dataset, with all variables other

than sex being synthesized. As a first step, we synthesized the hs_gpa variable with FCS, with conditioning on the sex variable. We did not include conditioning on the other variables when applying FCS because these were synthesized subsequently using IPSO. We then used as the matrix X for the IPSO method the matrix containing the variables sex and the synthesized hs_gpa. The matrix Y was taken to be the matrix of all other variables excluding sat_sum as above.

Tooling

We used the same packages and functions as for the IPSO synthesis, but additionally we used the syn function from synthpop to perform the FCS with CART.

Use Case 1: Testing analysis

Utility: This method is likely not the most appropriate for producing synthetic data for testing analysis purposes, although it was a significant improvement over the pure IPSO method. We saw that the distribution of hs_gpa matched the ground truth much more closely, and the issue with extreme values for maxima and minima of variables in the synthetic data was also greatly improved. The pMSE score was less than half that of the pure IPSO synthetic data, indicating a much closer resemblance in statistical properties between the synthetic and original data with this method. However, since it is possible to emulate the joint distribution of the synthesized variables more closely by using an entirely Fully Conditional Specification method, that method should preferred for use cases where unknown statistical analyses will be performed.

Privacy: Once again, with this method we did not identify any issues with the use of this synthetic dataset for this use case. As with the pure IPSO method, there are no replicated uniques in the synthetic data, and for those unique values of the quasi-identifier pair consisting of sex and hs_gpa which appeared in the original set and also appeared in the synthetic dataset, there were no cases of agreement on the remaining variables. However, because of the possibility of achieving higher utility with other methods, we would nevertheless not recommend this hybrid method for this use case.

Use Case 2: Education

Utility: Similarly to the Testing Analysis use case, since we saw an improvement with the hybrid method over pure IPSO in terms of emulating statistical properties of the original data, we believe the hybrid method is preferable to pure IPSO for this use case. However, since the joint distribution of the synthesized variables in the original dataset can be more effectively reproduced with other methods (such as pure FCS), we would prefer those methods for this use case.

Privacy: Once again, our analysis did not identify any clear disclosure risk for this synthetic data. We believe this synthesized dataset meets the medium confidentiality requirements of education purposes, but we would nevertheless not recommend this dataset for this use case because of the utility shortfalls.

Use Case 3: Testing technology

Utility: This synthetic dataset may be appropriate for testing complex systems that have been built to transform the data holding into another product. However, if these systems to not have any strong analytical requirements of the data, it may be simpler and more efficient to use a less advanced method, such as simply creating a dummy file.

Privacy: Once again, as we did not identify any disclosure risk from our synthetic data, we believe it should meet the medium confidentiality requirements for testing technology.

Use Case 4: Releasing microdata to the public

Utility: As with the Testing Analysis use case, we would not recommend this synthetic dataset for this use case. Although it was an improvement over pure IPSO, the joint distribution of the original data can be better emulated by using other methods.

Privacy: Although we found the disclosure risk of this synthetic dataset to be low, we would still not recommend the data for this use case because of the low utility.

Results

Utility results

When looking at the summary statistics, again we see very good agreement between the original and synthetic datasets, but there is an improvement over the pure IPSO method for the maxima and minima values. These are not as extreme as we saw when using only IPSO (and there is no longer a negative value for fy_gpa). There is still good agreement in means, and in particular, the synthetic data mirrors the differences in means between sexes.

Summary statistics for original and synthetic data for synthesized variables

sat_v	sat_m	hs_gpa	fy_gpa
Min. :24.00	Min. :29.0	Min. :1.800	Min. :0.000
1st Qu.:43.00	1st Qu.:49.0	1st Qu.:2.800	1st Qu.:1.980
Median :49.00	Median :55.0	Median :3.200	Median :2.465
Mean :48.93	Mean :54.4	Mean :3.198	Mean :2.468
3rd Qu.:54.00	3rd Qu.:60.0	3rd Qu.:3.700	3rd Qu.:3.020
Max. :76.00	Max. :77.0	Max. :4.500	Max. :4.000
sat_v	sat_m	hs_gpa	fy_gpa
Min. :26.00	Min. :26.00	Min. :1.800	Min. :0.570
1st Qu.:43.00	1st Qu.:48.00	1st Qu.:2.800	1st Qu.:1.930
Median :49.00	Median :54.00	Median :3.200	Median :2.460
Mean :48.93	Mean :54.38	Mean :3.197	Mean :2.468
3rd Qu.:54.00	3rd Qu.:60.00	3rd Qu.:3.700	3rd Qu.:2.980
Max. :76.00	Max. :96.00	Max. :4.500	Max. :4.730

Figure 6. Summary statistics for synthesized variables for original data (top) and synthesized data (bottom)

Summary statistics for original and synthetic data for synthesized variables

Records with sex=1

sat_v	sat_m	hs_gpa	fy_gpa
Min. :24.00	Min. :29.00	Min. :1.800	Min. :0.000
1st Qu.:43.00	1st Qu.:51.00	1st Qu.:2.700	1st Qu.:1.935
Median :49.00	Median :57.00	Median :3.100	Median :2.380
Mean :49.25	Mean :56.46	Mean :3.124	Mean :2.396
3rd Qu.:55.00	3rd Qu.:62.00	3rd Qu.:3.500	3rd Qu.:2.910
Max. :71.00	Max. :77.00	Max. :4.500	Max. :4.000
sat_v	sat_m	hs_gpa	fy_gpa
Min. :26.00	Min. :33.00	Min. :1.800	Min. :0.650
1st Qu.:44.00	1st Qu.:52.00	1st Qu.:2.700	1st Qu.:1.870
Median :49.00	Median :57.00	Median :3.100	Median :2.370
Mean :49.23	Mean :56.44	Mean :3.119	Mean :2.396
3rd Qu.:55.00	3rd Qu.:62.00	3rd Qu.:3.500	3rd Qu.:2.890
Max. :76.00	Max. :81.00	Max. :4.500	Max. :4.500

Figure 7. Summary statistics for records with sex=1. Statistics for original data shown on top and synthesized data on the bottom.

Summary statistics for original and synthetic data for synthesized variables

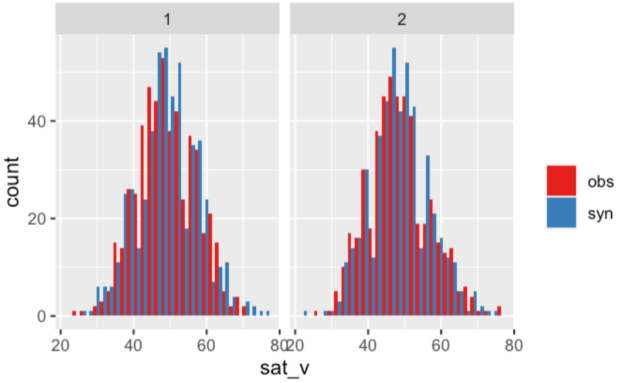
Records with sex=2

```
     sat_v           sat_m           hs_gpa          fy_gpa
Min.   :27.0   Min.   :31.00   Min.   :2.000   Min.   :0.000
1st Qu.:43.0   1st Qu.:47.00   1st Qu.:2.800   1st Qu.:2.020
Median :48.0   Median :52.00   Median :3.300   Median :2.560
Mean   :48.6   Mean   :52.19   Mean   :3.277   Mean   :2.545
3rd Qu.:54.0   3rd Qu.:58.00   3rd Qu.:3.750   3rd Qu.:3.120
Max.   :76.0   Max.   :77.00   Max.   :4.000   Max.   :4.000
     sat_v           sat_m           hs_gpa          fy_gpa
Min.   :30.0   Min.   :26.00   Min.   :2.000   Min.   :0.570
1st Qu.:42.0   1st Qu.:46.00   1st Qu.:2.900   1st Qu.:2.020
Median :49.0   Median :52.00   Median :3.315   Median :2.520
Mean   :48.6   Mean   :52.19   Mean   :3.281   Mean   :2.545
3rd Qu.:54.0   3rd Qu.:58.00   3rd Qu.:3.700   3rd Qu.:3.050
Max.   :76.0   Max.   :96.00   Max.   :4.000   Max.   :4.730
```

Figure 8. Summary statistics for records with sex=2. Statistics for original data shown on top and synthesized data on the bottom.

We see that there is significantly improved agreement in the distribution of the variables hs_gpa by sex between the original and synthetic dataset, as we would expect since the FCS with CART method does not rely on a normality assumption.

Distributions of variables by sex in original and synthetic data sets



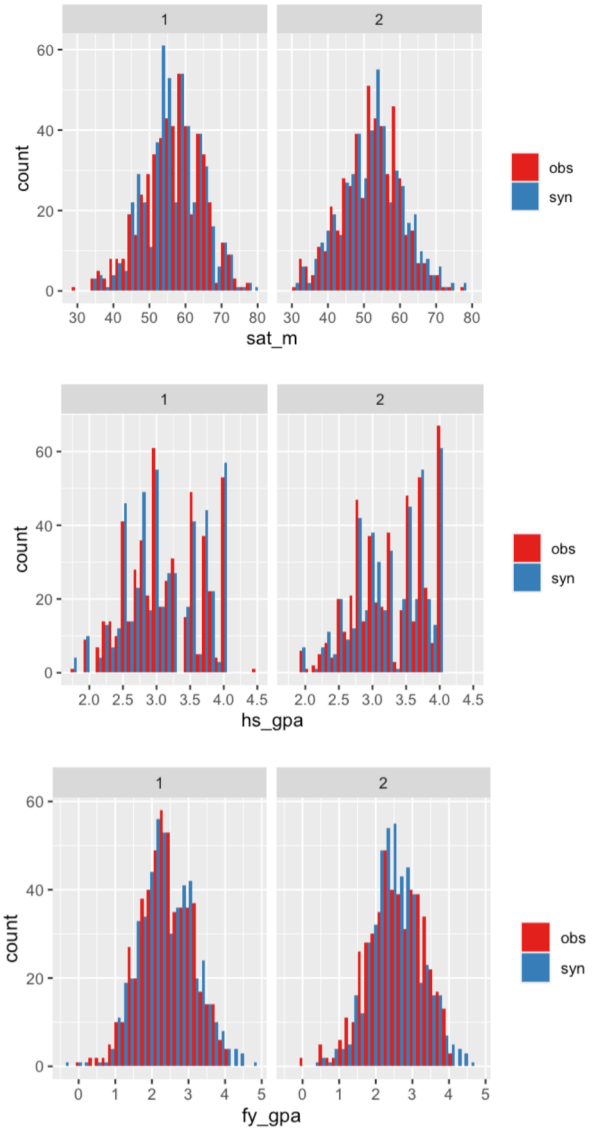


Figure 9. Distributions of variables by sex in original and synthetic datasets

The pMSE score for this synthetic dataset was 0.091 and the S_pMSE score was 2.48, again reflecting a significant improvement in utility over the pure IPSO method. Once again, we saw that for this dataset, pairs involving the variable hs_gpa had the worst S_pMSE scores. This may reflect the fact that in synthesizing this variable we only used conditioning on sex.

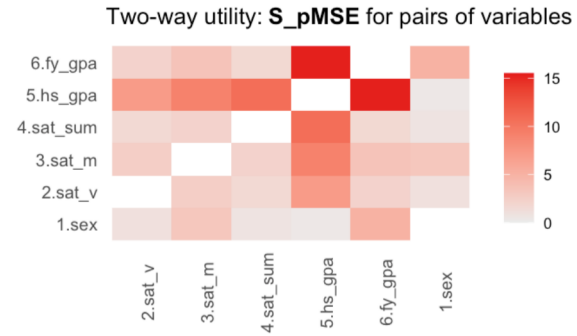


Figure 10. Heat map of two-way utility for the synthetic dataset

The hybrid method had one drawback over the pure IPSO method, which was that multiple regression models involving the hs_gpa variable no longer had the same regression coefficients and estimates of the covariance matrices for the original and synthetic data. This was to be expected since this variable was not synthesized using IPSO.

Privacy results

We used the same methods as for the pure IPSO synthetic dataset to evaluate the disclosure risk for this synthetic dataset. Using the replicated.uniques function from the synthpop package, we determined that there were no unique records in the original dataset that reappeared in the synthetic dataset. Again, this provided basic assurance of low disclosure risk from our synthetic data.

We also considered the eight unique pairs of sex and hs_gpa from the original dataset. These unique pairs appeared ten times in the IPSO/FCS hybrid synthetic dataset (fewer times than they appeared in the pure IPSO synthetic dataset). Again, we observed that whenever a unique pair from the original dataset appeared in the synthetic dataset, there was total disagreement between the original and synthetic datasets on the remaining variables. This provided assurance that the values of other variables could not be inferred from knowing sex and hs_gpa alone.

Simulation Method: Analytically Advanced Simulated Data

Alex Imbrogno, SIMD, Statistics Canada

For this synthesis method, two different synthetic versions of the satgpa data set were synthesized. In Both datasets, continous variables were synthesized using the simulation method for generating non-normal data as detailed by Fleishman (1978) and Vale and Maurelli (1983). In an attempt to increase analytical value, the second dataset was synthesized conditionally on sex, as an exploratory data analysis suggested an interaction between sex and the sat/gpa variables. This resulted in improved utility in the form of preserving this relationship.

The following R packages and functions were used during the synthesis and evaluation procedure.

Packages Used For Synthesis & Evaluation

- semTools (mvrnonnorm function)
- Synthpop (syn, utility.gen, utility.tab functions)

Methodology

The variables $satv, satm, fygpa$, and $hsgpa$, were synthesized using the *mvrnonnorm* function. Information from the sample in the form of vectors of means, kurtosis, skews, alongside a variance co-variance matrix were used during synthesis. The variable *satsum* was initially included but this caused the variance co-variance matrix to be non positive definite due to the deterministic relationship $satsum = satv + satm$. The *mvrnonnorm* function requires a positive definitie variance co-variance matrix so in response this variable was removed and synthesized as $satsum^{syn} = satv^{syn} + satm^{sym}$. The synthesized gpa scores were rounded to 2 significant digits to remain coherent with the original data. Since the Vale and Maurelli method is most appropriate for generating continous data, the variable *sex* was synthesized by drawing repeatedly from a $Bernoulli(p_{male})$ distribution with $p_{male} = Pr(sex = male)$ computed using the original data. In terms of preserving analytical value this approach is relatively naive since any relationship between *sex* and the rest of the variables was ignored. Preliminary data exploration suggested an interaction between *sex* and the gpa/ sat variables.

Synthetic data set 2 offers an improvement (tuning of the method) in an attempt to preserve certain relationships between *sex* and the rest of the data. The second data set was synthesized by partitioning the original data into males & females and then using sample information (for males & femles) and the *mvrnnorm* function to synthesize males and females independently. The number of synthetic males and females to generate was determined again by synthesizing sex from a $Bernoulli(p_{male})$ distribution. Using the Vale and Maurelli method, this approach can preserve the means, skews, kurtosis, and variance co-variance conditional on *sex*.

The utility of both synthetic data sets will now be evaluated with respect to each of the following use cases:

- Releasing synthetic microdata to the public & testing analysis
- Education
- Testing Technology

Utility Evaluation

The utility for the first synthetic data set will now be evaluated

Releasing Synthetic microdata to the public & testing analysis

Proportion of original & synthetic males/females

	male	female
original	0.516	0.484
synthetic	0.518	0.482

original and synthetic values for mean, kurtosis, and skew

	sat_v	sat_m	sat_sum	hs_gpa	fy_gpa
mean_orig	48.9340000	54.3950000	103.3290000	3.1981000	2.4679500
mean_syn	48.9521295	54.4906287	103.4427582	3.2125000	2.4787000
kurt_orig	-0.1202757	-0.1188956	-0.0779017	-0.8972017	-0.1947145
kurt_syn	-0.1067306	-0.2429580	-0.2486725	-0.9081970	-0.4308804
skew_orig	0.2089028	-0.1718894	0.0500327	-0.1718836	-0.2160610
kurt_syn	0.2077531	-0.1748772	-0.0009359	-0.1632309	-0.2048501

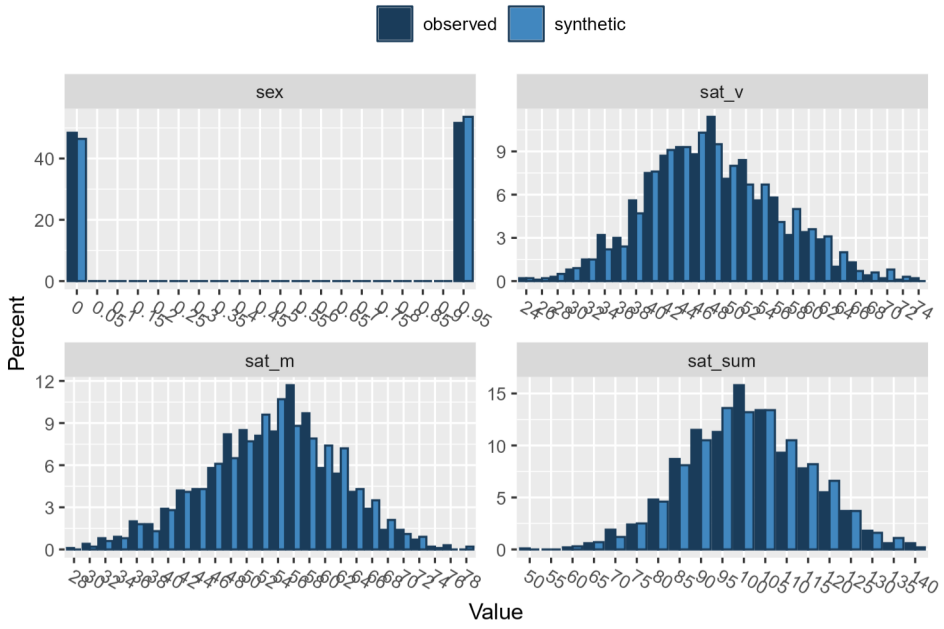
Original variance co-variance

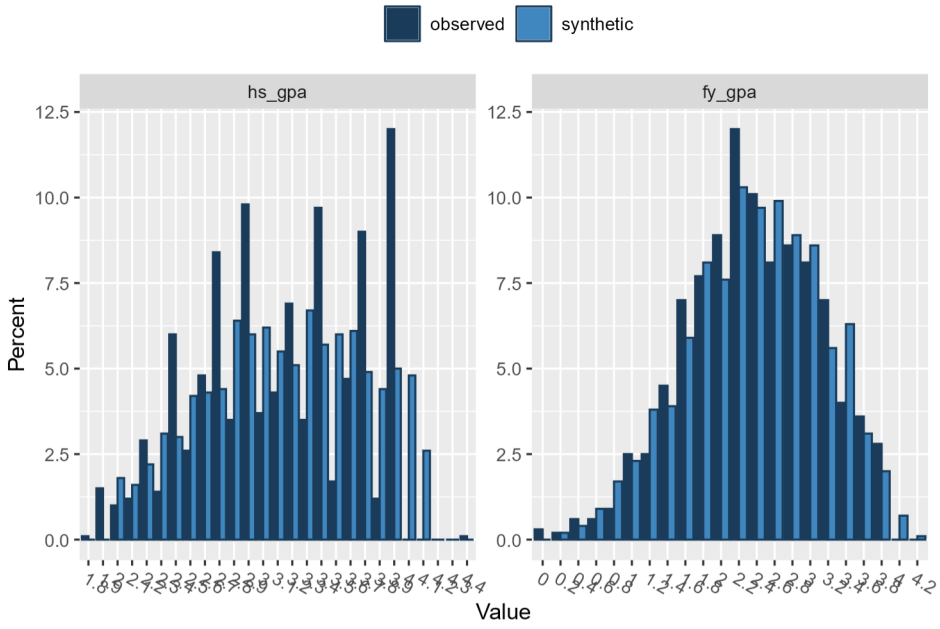
	sat_v	sat_m	sat_sum	hs_gpa	fy_gpa
sat_v	67.797441	32.463533	100.260975	1.6033279	2.4483731
sat_m	32.463533	71.404379	103.867913	1.7304209	2.4233131
sat_sum	100.260975	103.867913	204.128888	3.3337488	4.8716861
hs_gpa	1.603328	1.730421	3.333749	0.2933820	0.2180234
fy_gpa	2.448373	2.423313	4.871686	0.2180234	0.5487923

Synthetic variance-covariance

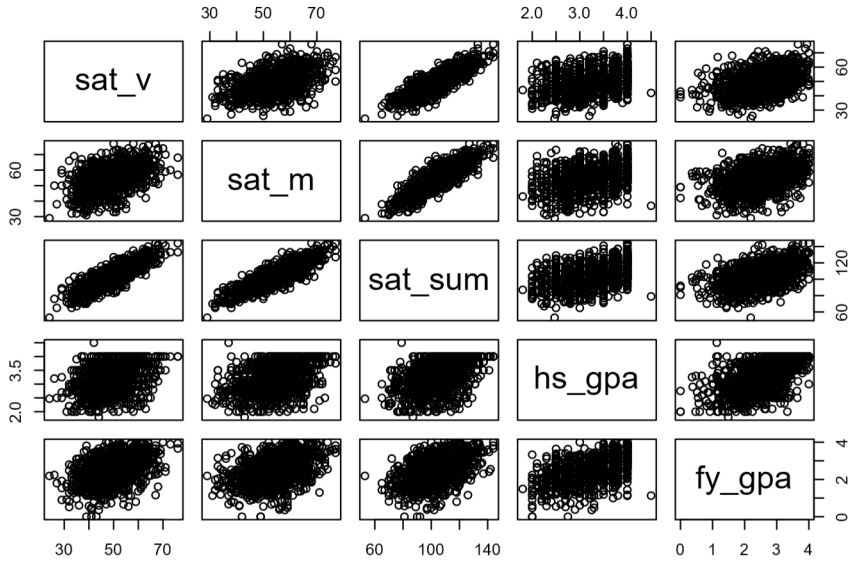
	sat_v_syn	sat_m_syn	sat_sum_syn	hs_gpa_syn	fy_gpa_syn
sat_v_syn	70.644200	31.543718	102.187917	1.6604175	2.5147468
sat_m_syn	31.543718	69.204155	100.747872	1.5125970	2.2224964
sat_sum_syn	102.187917	100.747872	202.935789	3.1730146	4.7372432
hs_gpa_syn	1.660418	1.512597	3.173015	0.2952832	0.2177186
fy_gpa_syn	2.514747	2.222496	4.737243	0.2177186	0.5716075

Original & synthetic histograms

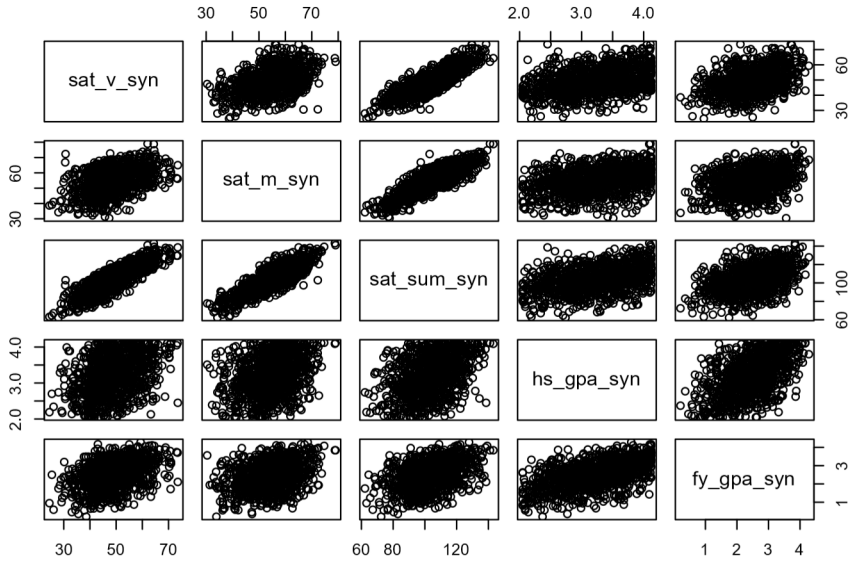

observed
synthetic
sex
sat_v
sat_m
sat_sum
Percent
Value



Original pairwise scatter plots



Synthetic pairwise scatter plots



utility.gen output

```
## Expectation of utility uses only coefficients involving synthesised variables: 6 from 21

##
## Utility score calculated by method:  logit
##
## Call:
## utility.gen(object = satgpa_syn_synthpop, data = satgpa)
##
## Utility score results
## Utility score: 143.85
## Expected value: 6
## Ratio to expected: 23.98
## p-value: < 0.0001
## pMSE (propensity score mean square error): 0.01
```

Overall the naive approach did a relatively good job at preserving the overall shape of the marginal distribution. This was illustrated through the overlapping histograms and comparing values of mean, skew, variance, and kurtosis. The pairwise plots seem to indicate that the overall linear relationship was preserved between the variables $satv, satm, hsgpa, fygpa$. Viewing the off diagonal (co-variances) elements of the variance covariance matrix adds to that. A more thorough approach would have been to compare all pairwise linear regression models between the original and synthetic data, but given the time constraint only linearity was verified.

Next we will evaluate synthetic dataset 1's ability to preserve the relationship between *sex* and the rest of the data seeing as there appears to be a relationship (indicated in the original data). Many different analyses

were carried out to test this, some of which were: ANOVA (weak assumption of normality), comparing contional means, side by side histograms/ box plots, regressing *sex* onto the contious variables, however we will only present results from the logistic regression. All analyses were conclusive in determining that the relationship was not preserved in synthetic data set 1.

Original Logistic Regression: $sex = satv + satm + satsum + hsgpa + fygpa$

```
##
## Call:
## glm(formula = sex ~ ., family = binomial, data = satgpa)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3302  -1.0481   0.5471   1.0352   2.2395
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.600177   0.553805  -2.889 0.003859 **
## sat_v       -0.004676   0.009844  -0.475 0.634824
## sat_m        0.107299   0.010710  10.019  < 2e-16 ***
## sat_sum            NA         NA      NA       NA
## hs_gpa      -0.920906   0.161527  -5.701 1.19e-08 ***
## fy_gpa      -0.400117   0.117156  -3.415 0.000637 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1385.3  on 999  degrees of freedom
## Residual deviance: 1233.7  on 995  degrees of freedom
## AIC: 1243.7
##
## Number of Fisher Scoring iterations: 4
```

Synthetic Logistic Regression: $sex + satv + satm + satsum * hsgpa + fygpa$

```
##
## Call:
## glm(formula = sex_syn ~ ., family = binomial, data = satgpa_syn)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.417  -1.201   1.017   1.148   1.344
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.195281   0.518686   0.376   0.7066
## sat_v_syn    0.007625   0.008925   0.854   0.3929
## sat_m_syn   -0.011969   0.008844  -1.353   0.1759
## sat_sum_syn        NA         NA      NA       NA
## hs_gpa_syn   0.201767   0.142013   1.421   0.1554
## fy_gpa_syn  -0.198472   0.103936  -1.910   0.0562 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1385.0  on 999  degrees of freedom
## Residual deviance: 1378.6  on 995  degrees of freedom
## AIC: 1388.6
##
## Number of Fisher Scoring iterations: 3
```

Comparing the coefficients and p-values we can see that the relationship between sex and the rest of the data has not been preserved in the synthetic file. In an attempt to improve this, synthetic dataset 2 was synthesized conditional on sex as detailed at the start of the document. This resulted in the various tests mentioned above concluding that there was an interaction between sex and gpa/sat in the synthetic file. Below is the same original logistic model but compared to the model fit on synthetic data set 2.

Original Logistic Regression: $sex = satv + satm + satsum + hsgpa + fygpa$

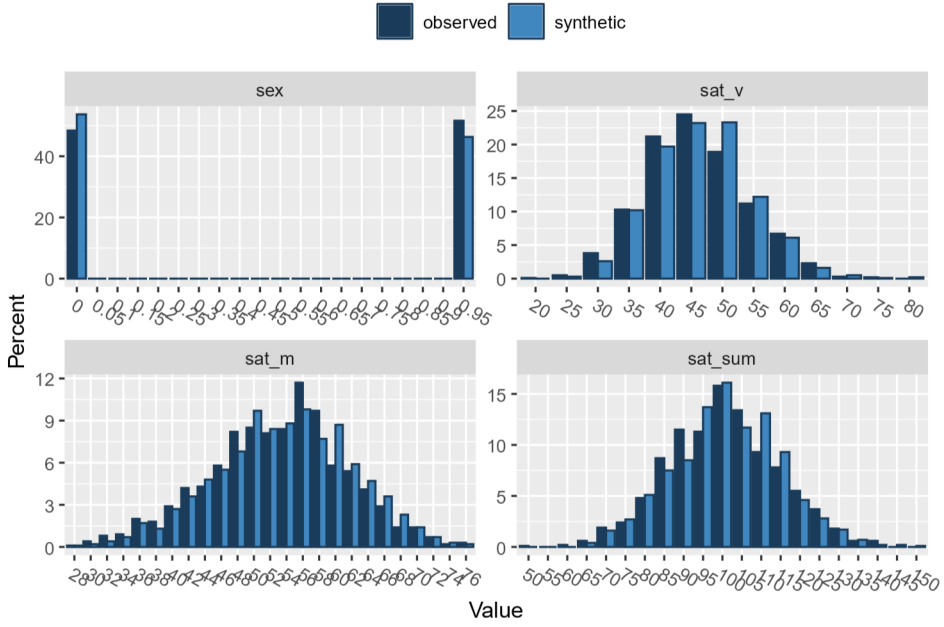
```
##
## Call:
## glm(formula = sex ~ ., family = binomial, data = satgpa)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3302  -1.0481   0.5471   1.0352   2.2395
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.600177   0.553805  -2.889 0.003859 **
## sat_v       -0.004676   0.009844  -0.475 0.634824
## sat_m        0.107299   0.010710  10.019  < 2e-16 ***
## sat_sum            NA         NA      NA       NA
## hs_gpa      -0.920906   0.161527  -5.701 1.19e-08 ***
## fy_gpa      -0.400117   0.117156  -3.415 0.000637 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1385.3  on 999  degrees of freedom
## Residual deviance: 1233.7  on 995  degrees of freedom
## AIC: 1243.7
##
## Number of Fisher Scoring iterations: 4
```

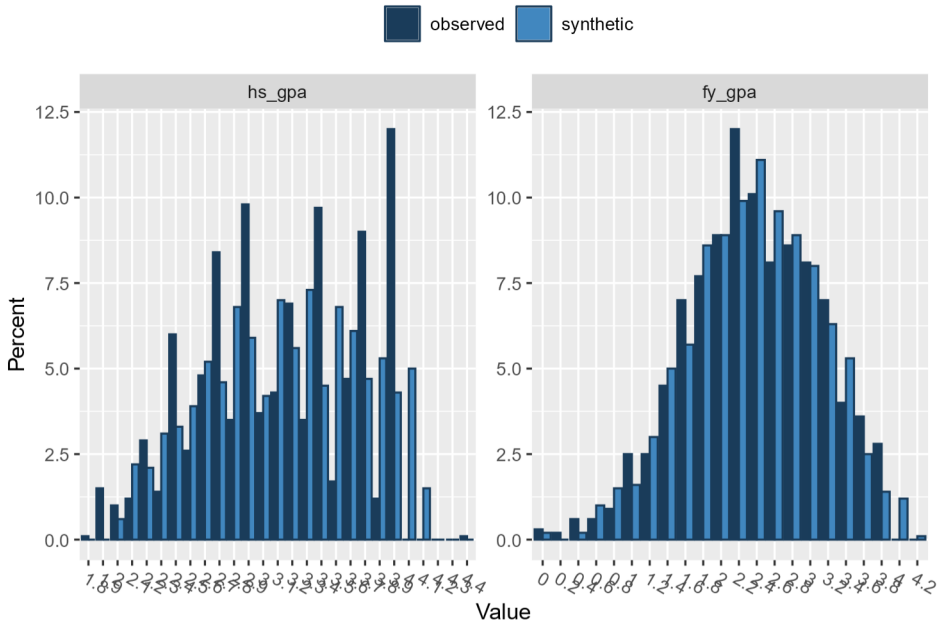
Synthetic 2 Logistic Regression: $sex + satv + satm + satsum * hsgpa + fygpa$

```
##
## Call:
## glm(formula = sex_syn ~ ., family = binomial, data = satgpa_syn)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1370  -1.0522  -0.4121   1.0526   2.1019
##
```

```
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.355774   0.571958  -2.370  0.01777 *
## sat_v_syn   -0.007894   0.010024  -0.788  0.43098
## sat_m_syn    0.106170   0.010729   9.895  < 2e-16 ***
## sat_sum_syn        NA         NA      NA       NA
## hs_gpa_syn  -0.992138   0.159290  -6.228 4.71e-10 ***
## fy_gpa_syn  -0.367063   0.116537  -3.150  0.00163 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1386.2  on 999  degrees of freedom
## Residual deviance: 1236.7  on 995  degrees of freedom
## AIC: 1246.7
##
## Number of Fisher Scoring iterations: 4
```

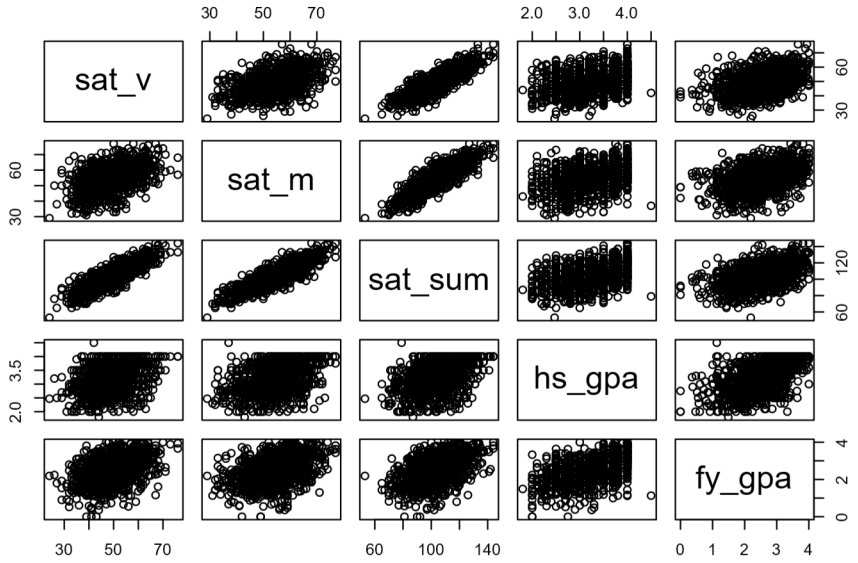
Again comparing the synthetic and original coefficients and p-values, the relationship appears to be better preserved in comparison to the first synthetic data set. As was done with the synthetic data set 1, the marginal and synthetic distributions of each variable was compared. The results were identical to the first synthetic data set. Below are the overlapping histograms between original and synthetic data set 2.



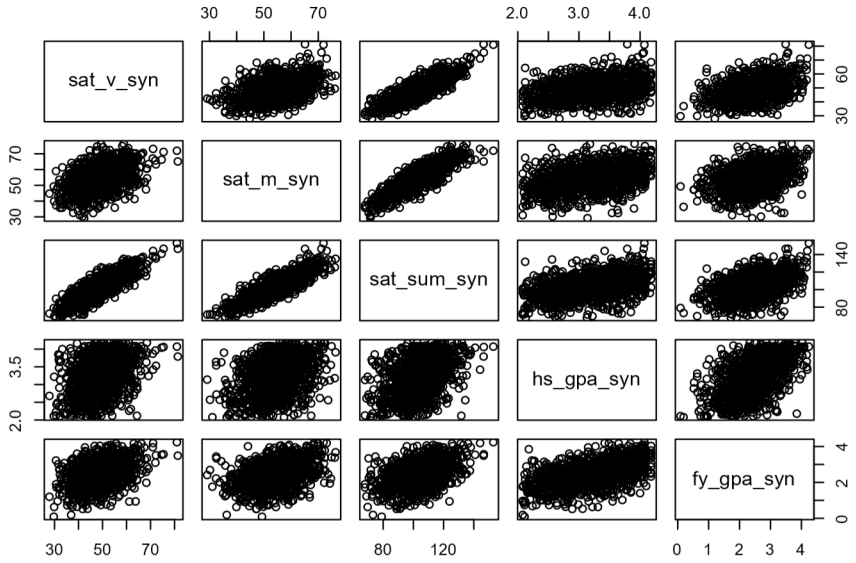


The linear relationships between the continous variables were also preserved similar to synthetic file 1.

Original pairwise scatter plots



Synthetic 2 pairwise scatter plots



Overall, the analytically advanced simulated data method preformed reasonably well at preserving marginal distributions and relationships between continuous variables but required some tuning in order to capture relationships between categorical and continuous variables. In terms of this use case I would only recommend this method when there is a priori knowledge about what relationships/ analyses need to be preserved (as was identified in synthesis 2). The method was relatively simple to understand and implementation was not very hard but there seems to be a limit on the amount of utility one can gain.

I thought about how I could extend the conditional synthesis conducted for data set 2 to more complex data sets. Just an idea but cells could be created by crossing variables which are believed to be informative with respect to the rest of the data and then carry out the Vale and Maurelli method within each class. This could risk small cells being created with poor privacy guarantees so one would need to possibly collapse small cells or address these concerns in a different way.

Education

It appears this method can be used for educational purposes if the synthetic data is synthesized in such a was as to preserve important relationships to be studied during the educational process. If synthetic data set 1 was used to teach how sex impacts sat and gpa, there would be no clear conclusion/ teaching points. Synthetic dataset 2 would have been a better choice for this type of educational exercise. If the educational purpose was to explore marginal distributions of certain continuous variables then synthetic data set 1 most likely would have been sufficient.

Testing Technology

Given that the method and implementation is rather straight forward, I would recommend this data set if technology needs to be tested. However, a dummy file would most likely be a more efficient approach

since high analogical value may not be needed to test technology. If the technology being tested requires some analogical value to be preserved then this method could be a reasonable solution between Sequential Modelling & Dummy Files.

Privacy Evaluation

To evaluate privacy on both synthetic data sets, the R function *replicated.uniques* from the *synthpop* package was used.

Synthetic Dataset 1

```
##
## Variable(s): sat_v_syn, sat_m_syn, sat_sum_syn, hs_gpa_syn, fy_gpa_syn not synthesised or used in pr
## CAUTION: The synthesised data will contain the variable(s) unchanged.
##
## Synthesis
## -----------
##  sex_syn sat_v_syn sat_m_syn sat_sum_syn hs_gpa_syn fy_gpa_syn

##
## Variable(s): sat_v, sat_m, sat_sum, hs_gpa, fy_gpa not synthesised or used in prediction.
## CAUTION: The synthesised data will contain the variable(s) unchanged.
##
## Synthesis
## -----------
##  sex sat_v sat_m sat_sum hs_gpa fy_gpa

## $replications
##    [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [49] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [61] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [73] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [85] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [97] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [109] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [121] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [133] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [145] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [157] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [169] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [181] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [193] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [205] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [217] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [229] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [241] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [253] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [265] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [277] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [289] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
##   [301] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [313] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [325] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [337] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [349] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [361] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [373] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [385] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [397] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [409] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [421] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [433] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [445] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [457] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [469] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [481] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [493] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [505] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [517] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [529] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [541] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [553] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [565] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [577] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [589] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [601] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [613] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [625] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [637] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [649] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [661] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [673] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [685] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [697] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [709] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [721] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [733] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [745] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [757] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [769] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [781] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [793] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [805] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [817] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [829] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [841] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [853] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [865] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [877] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [889] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [901] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [913] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [925] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [937] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
##  [949] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [961] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [973] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [985] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [997] FALSE FALSE FALSE FALSE
##
## $no.uniques
## [1] 1000
##
## $no.replications
## [1] 0
##
## $per.replications
## [1] 0
```

Synthetic Dataset 2

```
## $replications
##    [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [49] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [61] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [73] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [85] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [97] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [109] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [121] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [133] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [145] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [157] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [169] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [181] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [193] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [205] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [217] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [229] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [241] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [253] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [265] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [277] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [289] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [301] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [313] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [325] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [337] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [349] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [361] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [373] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [385] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [397] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [409] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [421] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
##  [433] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [445] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [457] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [469] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [481] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [493] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [505] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [517] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [529] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [541] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [553] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [565] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [577] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [589] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [601] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [613] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [625] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [637] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [649] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [661] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [673] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [685] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [697] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [709] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [721] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [733] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [745] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [757] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [769] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [781] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [793] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [805] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [817] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [829] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [841] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [853] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [865] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [877] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [889] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [901] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [913] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [925] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [937] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [949] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [961] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [973] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [985] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [997] FALSE FALSE FALSE FALSE
##
## $no.uniques
## [1] 1000
##
## $no.replications
## [1] 0
```

```
##
## $per.replications
## [1] 0
```

Both synthetic data sets appear to exhibit a high degree of privacy in terms of the unique match proportions. Both data sets have a unique match proportion of 0.

As mentioned above, the conditional method used for synthetic file 2 has the potential to have poorer privacy protection when the number of records in a response category is low. With *sex* being around a 50-50 split this issue was not observed in synthesis 2.