# Break

10 mintues

# A note on Chapter 4: Disclosure considerations for synthetic data

- Disclosure risk is the risk of inappropriate release of data or attribute information of a record (often individual).
- Although no record in a (fully) synthetic data file corresponds to a real person or household, there is concern that attribute and identification disclosure risk could still be present.
- **Recommendation:** NSOs should choose additional disclosure controls based on their own legislative and operational frameworks.
- The purpose of this chapter is to present disclosure control options available to NSOs and their synthesizers

Privacy Preserving Techniques:

- K-anonymity
- $\ell$-diversity
- $t$-closeness
- Differential privacy

Disclosure Risk Measures:

- Peer review
- Feature Mean Scaled Variance
- Rates related to database reconstruction

# Sli.do Poll: #034032

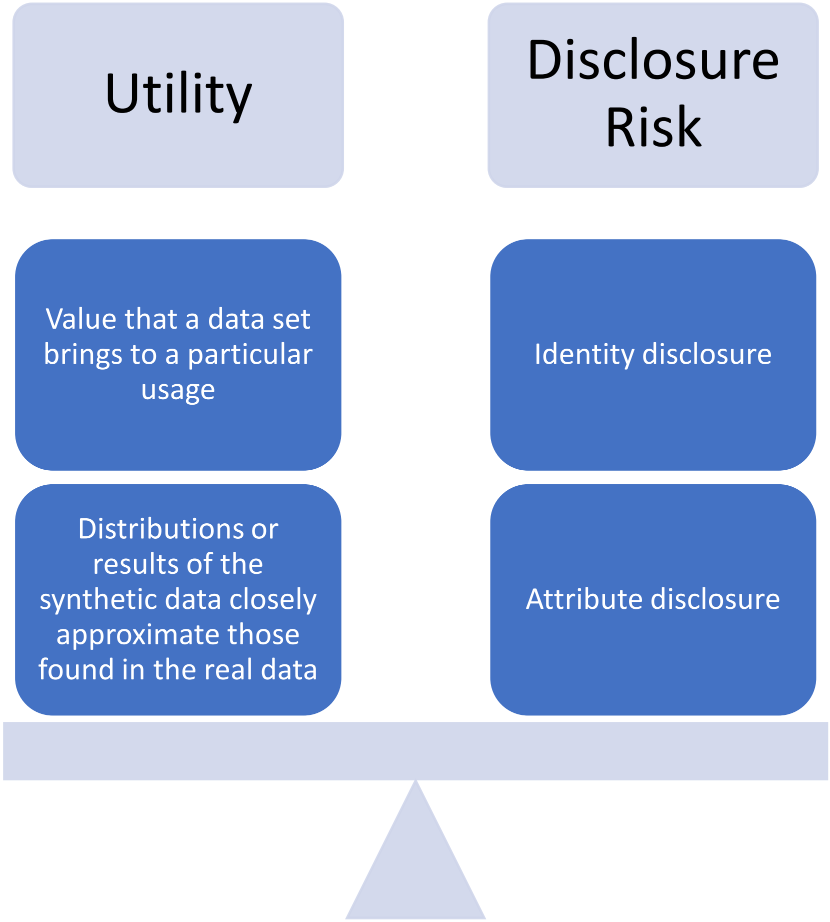
Tell us more about disclosure considerations in your NSOs for real data and synthetic data.

# Methods and Measures to assess Utility

Does your synthetic data meet user needs?

# What is utility in the synthetic data context?

- How useful a synthetic data set is to the purpose of the data
- A great challenge with synthetic data is balancing utility and disclosure risk



# Where to Start?

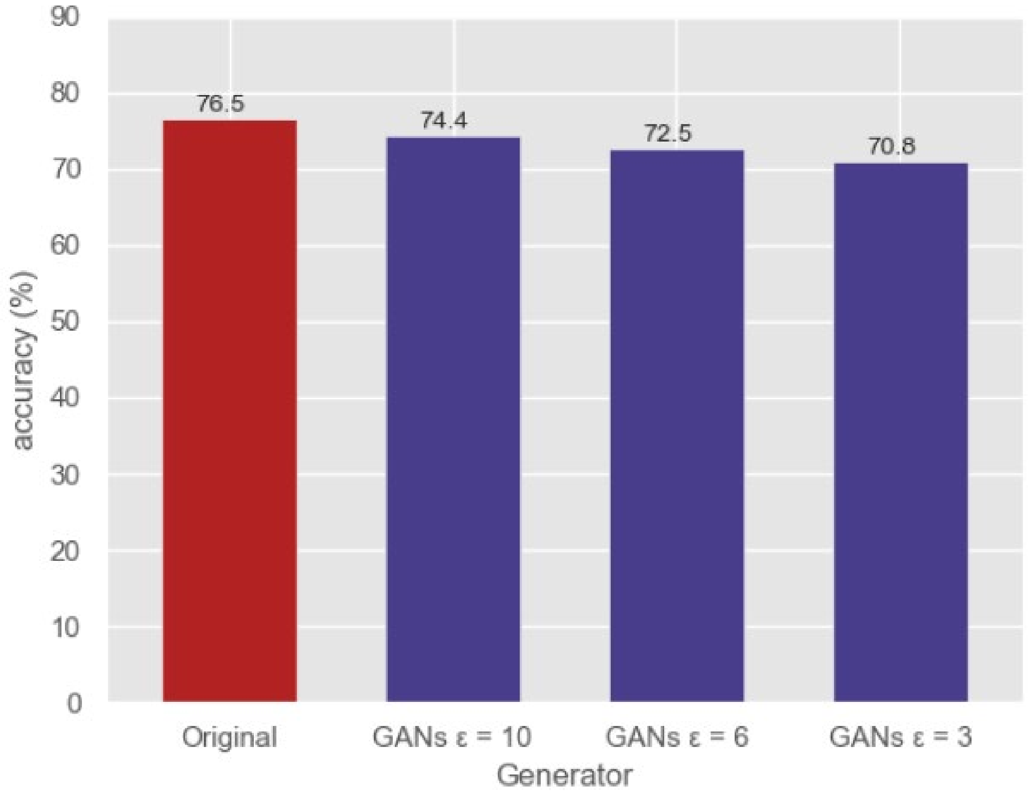
- Why do you want to measure utility?
  - To compare different syntheses of the same original
    - General measures
  - To improve the utility of your synthesis method (tuning)
    - Combinations of general and specific measures

# Specific Measures

- Specific utility measures compare the results of statistical models fitted to the synthetic and the original data.
- The results of any statistical analysis can be used to create a utility measures
  - Impact on policy decisions
  - Difference in means of variables
  - Differences in correlations
  - Tables and cross-tabulations
  - Task accuracy differences
  - Generalised Linear Models (GLMs)

**Example of task accuracy**

Classification accuracy trained on original US Adult Income data set and synthetic data sets generated with GANs, with different values of privacy loss ε (Kaloskampis et al. 2020).



# Specific measures

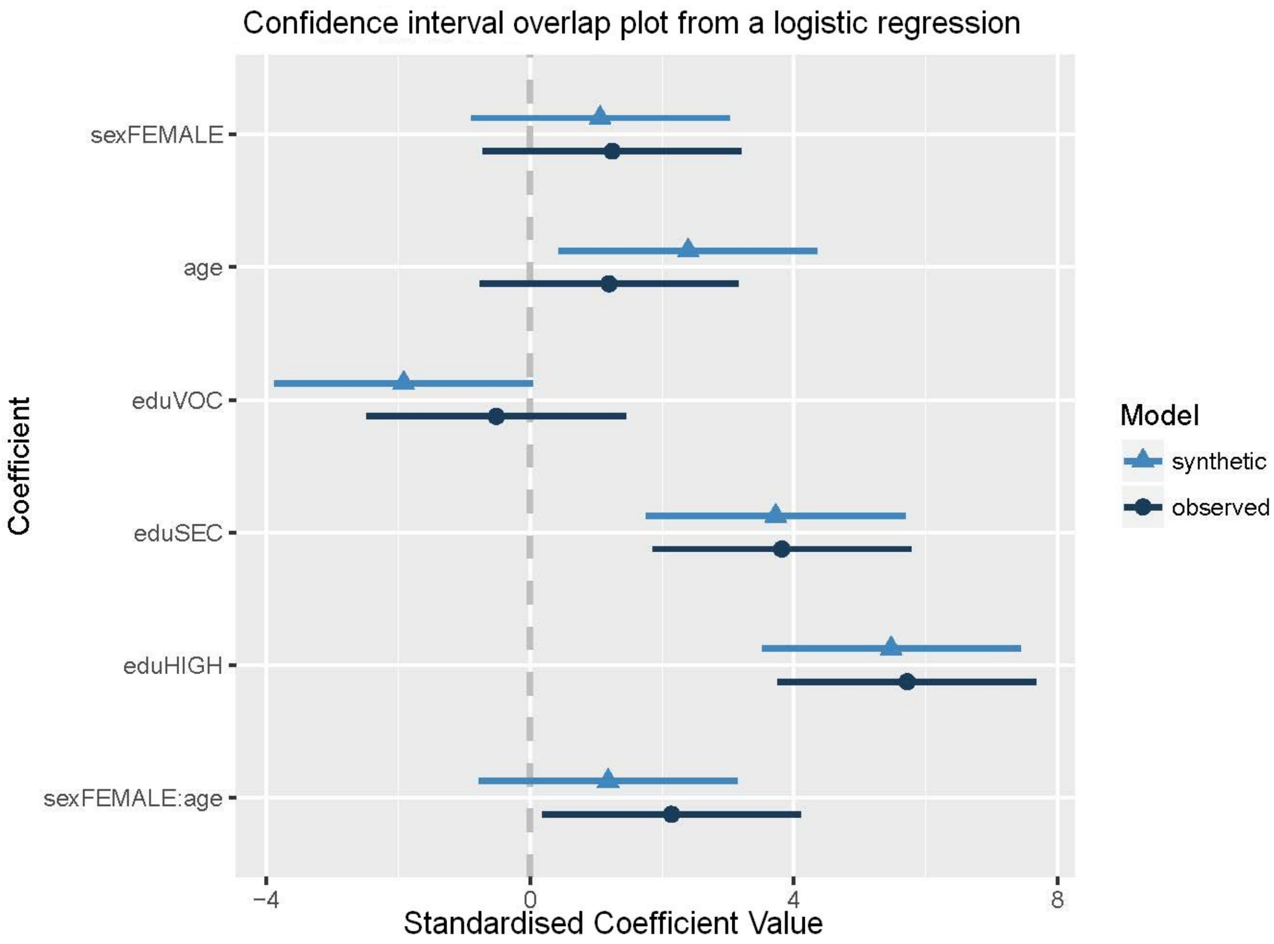
Any statistic with a standard error or interval estimate

- Summary measures
  - Standardised difference
  - Confidence interval overlap
- For statistical models
  - Average overlap for all coefficients
  - Mahalanobis distance ratio is a combined standardised difference

**Confidence interval plot**

Coefficients and interval estimates for a logistic model predicting the probability of not-smoking from age,sex and education..

Raab & Nowok, Inference from fitted models in *synthpop.*

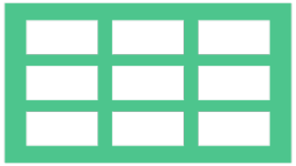


# What are General Utility Measures?

- Often the synthesizer does not know the specific use of the synthetic data
- A measure (just one number) that compares the whole distribution of the synthetic data to that of the original data

## Methods to calculate General Utility

Combine original and synthetic data and try to predict the synthetic from a **propensity score**



Make tables of original and synthetic data and calculate a measure based on their **differences**

| Propensity Score Measures |
| --- |
| Propensity score mean squared error (pMSE) |
| Kolmogorov- Smirnov Statistic comparing propensity scores for original and synthetic data (SPECKS) |
| Percentage over 50% of combined records correctly predicted by the propensity score (PO50) |
| Other comparisons of propensity scores for original and synthetic data , e.g Wilcoxon signed rank statistic (U) |

| Tabular Measures |
| --- |
| Voas-Williamson statistic |
| Freeman-Tukey |
| Likelihood ratio statistic from tables (G) and other members of the divergence family |
| Jensen-Shannon Divergence |
| Bhattacharyya metric |
| Mean absolute difference in densities |
| Weighted mean absolute difference in densities |

# How do you calculate the propensity score?

Any method to predict a binary variable will do.

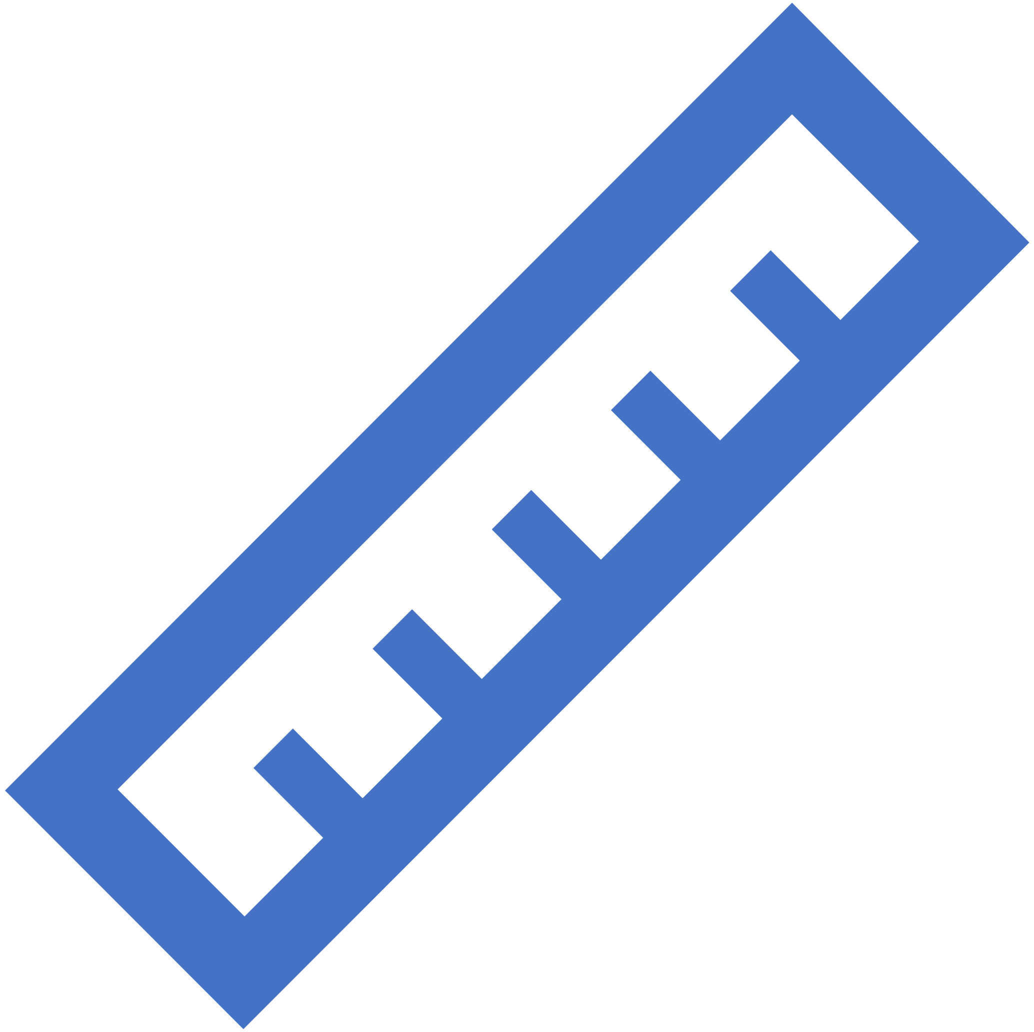
| Methods |
| --- |
| Logistic regression |
| Classification and regression trees (CART) |
| Any other classification method – neural nets, random forests,……. |
| Form tables and calculate proportions of synthetic to all records in each corresponding cell.<br><br>An n-way table is the same as a saturated logistic regression model for the n-variables |

# General Measures recap

- Many different measures have been proposed
- Some are the same as each other (e.g. pMSE from propensity scores and Voas Williamson from tables)
- All are highly correlated when calculated by the same method
- The method used to calculate the measures is more important than the choice of measure
- Raab Nowok and Dibben, 2021 *Assessing, visualizing and improving the utility of synthetic data*. https://arxiv.org/pdf/2109.12717.pdf
- It may be useful to calculate measures for subgroups of variables or subsets of the data

# Scaling

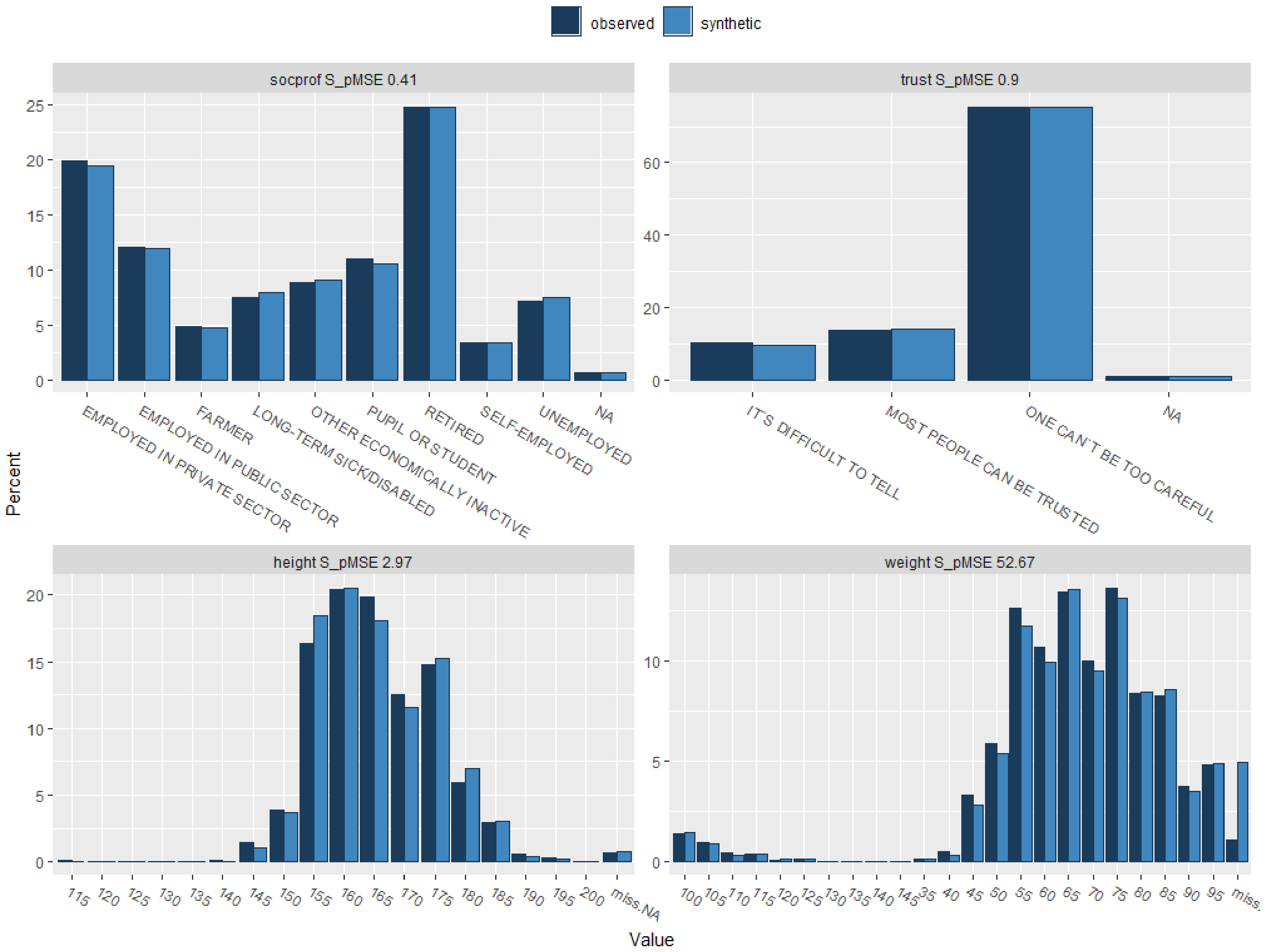
- It is helpful if the utility measures can be on a scale that makes them easy to interpret.
- For all the measures presented a large value indicates lack-of utility
- Another approach to scaling utility measures is to express them relative to the value that would be expected if the model used to synthesize the data was the “correct” model.
    - The expected value for the “correct” model can be termed the Null expectation.



# Tuning

- To adapt the synthesis method being used in the light of utility findings about
  - Which variables, or combinations of variables, are contributing to the lack of utility
  - Any subsets of the data that may have poor utility
- Often one number is not enough: If the utility appears unsatisfactory they need to know which aspects of the distribution are causing the problem.
  - Univariate comparisons
  - Marginal comparisons
  - Comparing other statistics

# One way tables




Selected utility measures:

| | S_pMSE | df |
|---|---|---|
| sex | 0.55 | 1 |
| income | 1.00 | 6 |
| age | 1.05 | 4 |
| edu | 0.36 | 4 |
| socprof | 0.40 | 9 |
| trust | 0.90 | 3 |
| height | 2.97 | 5 |
| weight | 52.67 | 5 |
| smoke | 0.54 | 2 |
| region | 1.39 | 15 |

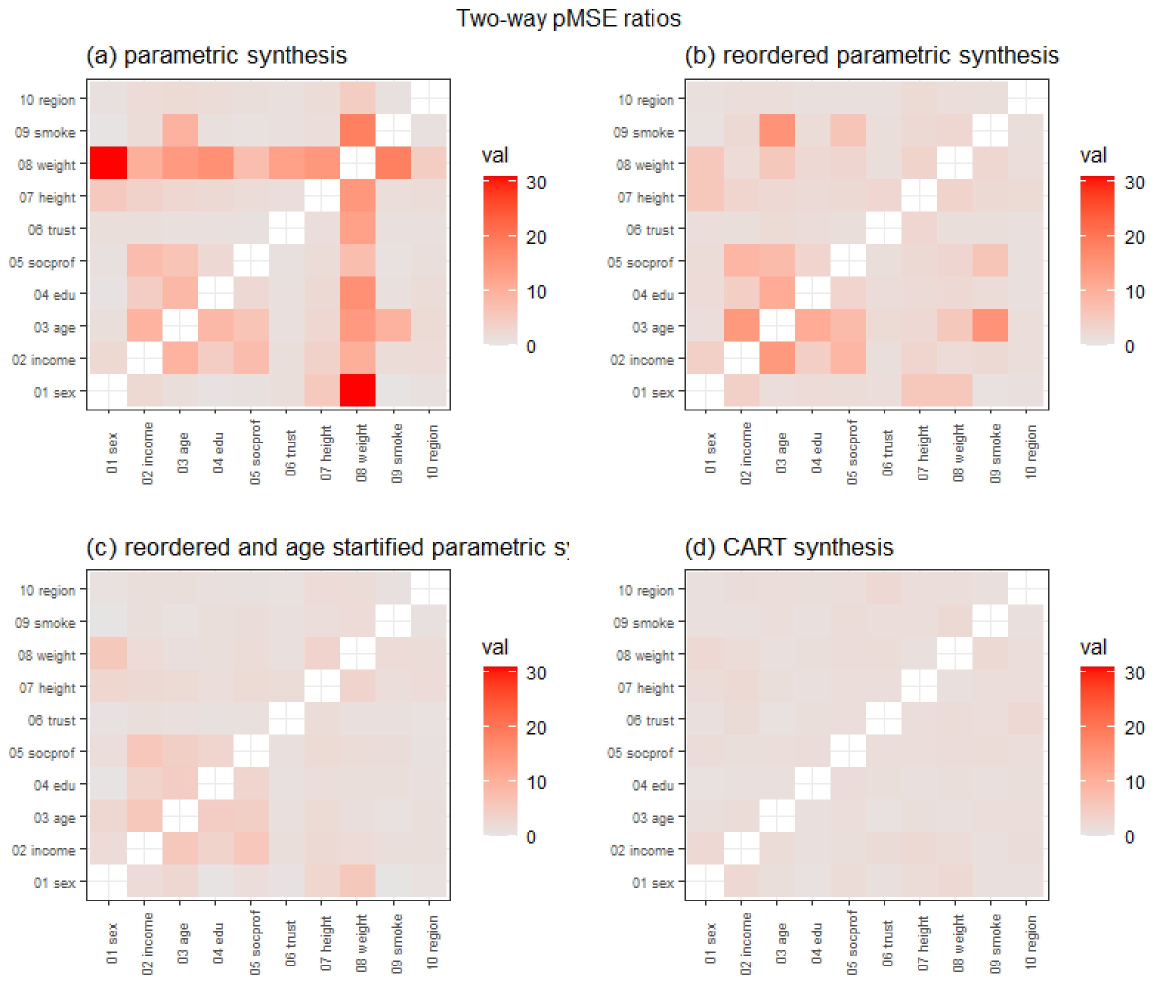
# Methods for exploring utility

- However, even after there is good performance on univariate distributions, there may be issues with correlations between features, or subgroups within features.

- Visualising two-way the utility of all two-way tables can be helpful

**Example of marginal comparisons**

Visualizations of the utility of all two-way relationships between variables



# Methods for exploring utility

- On larger problems (>10 variables), randomized three-way marginals can be algorithmically searched for poor-performing feature correlations, using Frequent Item-set analysis.

- Once basic evaluations are complete, subgroups in the population should be evaluated separately to ensure fair performance across the full population.

```
frequent_itemset:       support
0     1.0000             (ANC1P)
5     0.1250              (PUMA)
17    0.1250       (ANC1P, PUMA)
8     0.1125             (RAC2P)
20    0.1125      (RAC2P, ANC1P)
23    0.1000       (ANC1P, SCHL)
21    0.1000      (RAC3P, ANC1P)
9     0.1000             (RAC3P)
11    0.1000              (SCHL)
7     0.0875             (RAC1P)
```

Table 2: Stable Feature Race Analysis

| EVAL | MEAN | STD |
|---|---|---|
| Two or more races | 868.91 | 13.51 |
| Some other race alone | 908.55 | 5.34 |
| Asian alone | 908.46 | 7.78 |
| Black alone | 913.78 | 3.67 |
| White alone | 977.95 | 2.10 |

# Methods for exploring utility

- If different subgroups in the population have significantly conflicting patterns of correlations between their features, it can be difficult for the generative model to capture all groups adequately.
- When large discrepancies between the ground truth and synthetic data distributions are identified, further investigation including geographic-based evaluations and exploration of model dependencies can help diagnose the problem.

| | Variable | Description | Weight |
|---|---|---|---|
| Top Race Predictors–PUMA 3529 | POBP | Place of Birth | 0.283 |
| | HISP | Hispanic Origin | 0.217 |
| | ANC1P | Ancestry 1 | 0.188 |
| | OIP | Non-wage Income | 0.109 |
| | ANC2P | Ancestry 2 | 0.057 |
| | OCCP | Occupation | 0.021 |
| | LANP | Language | 0.018 |
| | JWAP | Commute Arrival | 0.015 |
| | PRIVCOV | Priv. Health Ins. | 0.013 |

| | Variable | Description | Weight |
|---|---|---|---|
| Top Race Predictors–Overall IL | ANC1P | Ancestry 1 | 0.489 |
| | HISP | Hispanic Origin | 0.139 |
| | POVPIP | Income/Poverty Ratio | 0.087 |
| | ANC2P | Ancestry 2 | 0.071 |
| | POBP | Place of Birth | 0.041 |
| | OCCP | Occupation | 0.028 |
| | JWAP | Commute Arrival | 0.020 |
| | PINCP | Total Income | 0.017 |
| | JWDP | Commute Departure | 0.013 |

# Methods for *improving* utility

When issues are identified, there are a variety of steps that can be taken to improve synthesis quality, depending on the synthesis approach.

- Synthesis can be partitioned so that distinct subgroups in the population are synthesized separately to better preserve their unique distributions.
- Variables can be redefined or post-processed to satisfy edit constraints. If a variable in the schema is computed deterministically from other variables, it should be recalculated rather than synthesized.
- Structural zeros/null values can be synthesized independently in a 2–step process: First synthesize a binary variable **IsNull_VarA**, then synthesize **Val_VarA**. This allows the model to better fit and reproduce these patterns.
- Adding supplementary variables (such as the median feature value in a given geographical area) can make important information more accessible and improve model performance.

# Slido exit poll

Event #**931050**

# We are looking for feedback on the guide

- Meeting on Statistical Data Confidentiality, December 2, 2021
- Data Challenge – test drive the guide!
  - Dates: January 24 to January 28, 2022
  - Problem: you are a NSO that is facing one of 4 disclosure problems. You must generate synthetic data and assess if it meets the disclosure and utility standards to release it.
  - You will be provided with an 'original' data file
  - Experts will be on hand to help.
- Registration now open: https://indico.un.org/event/1000359/
- Encourage your NSOs and networks to participate!

# Slido exit poll part 2

Event #**931050**