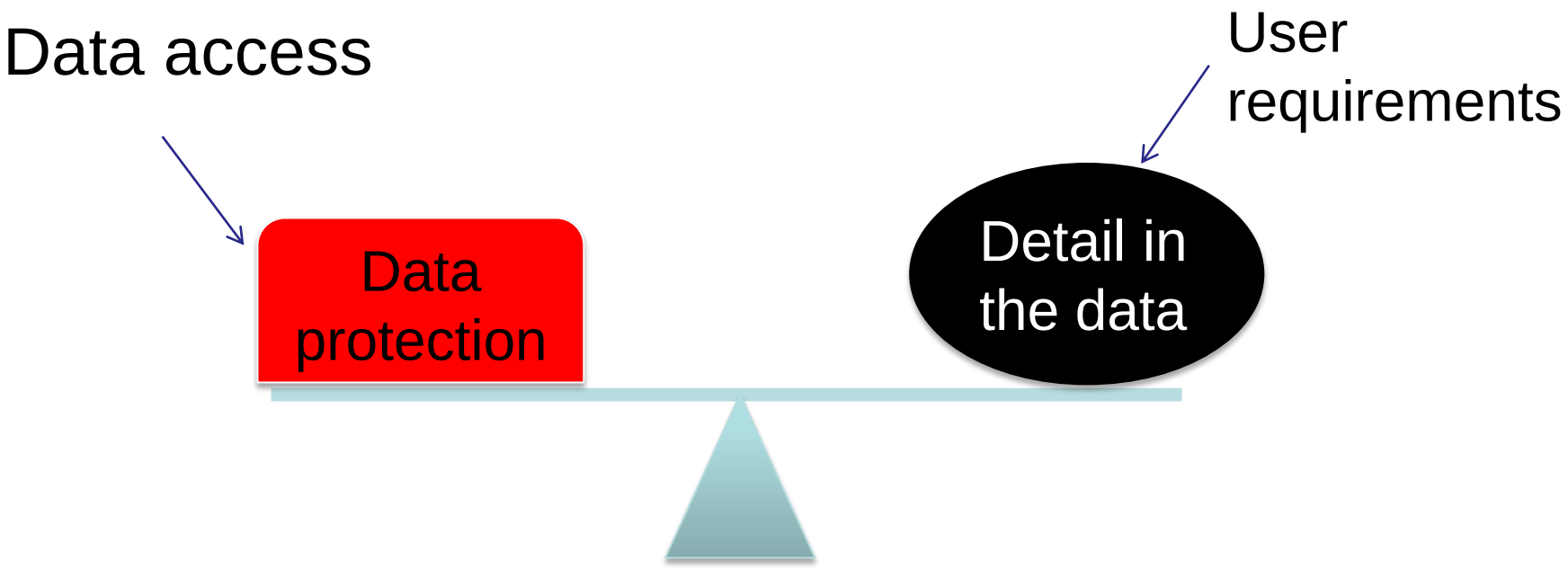
Creation of synthetic microdata in 2021 Census Transformation Programme (proof of concept)

Robert Rendell

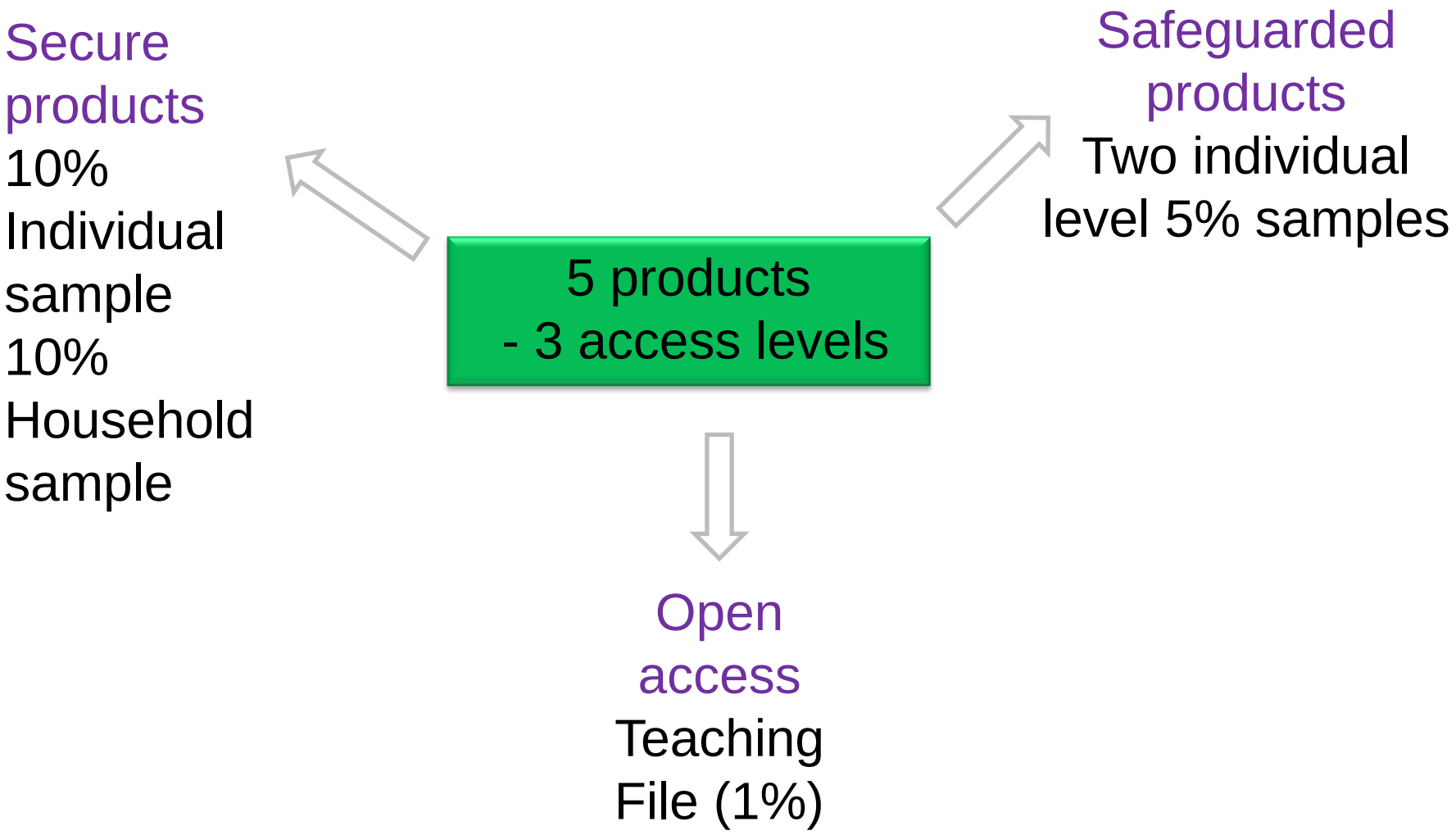
Outline

- Briefly introduce microdata and current microdata products
- Synthetic data generation methodology
- Utility analysis
- Next steps

Census Microdata Products



2011 Census Microdata Products for England and Wales



Synthetic Microdata

Proposed product: Safeguarded 5% partially synthetic 2011 Census household microdata sample for England and Wales

Enables wider access to meaningful census microdata

Access for international users

Access for commercial users

Synthetic Microdata Initial Research

- Initial test on one area (500,000 records)
- Select random 5% sample of test area to represent Microdata sample.
- Remove values from some variables
- Impute missing values using CANCEIS (Canadian Census Edit and Imputation System)

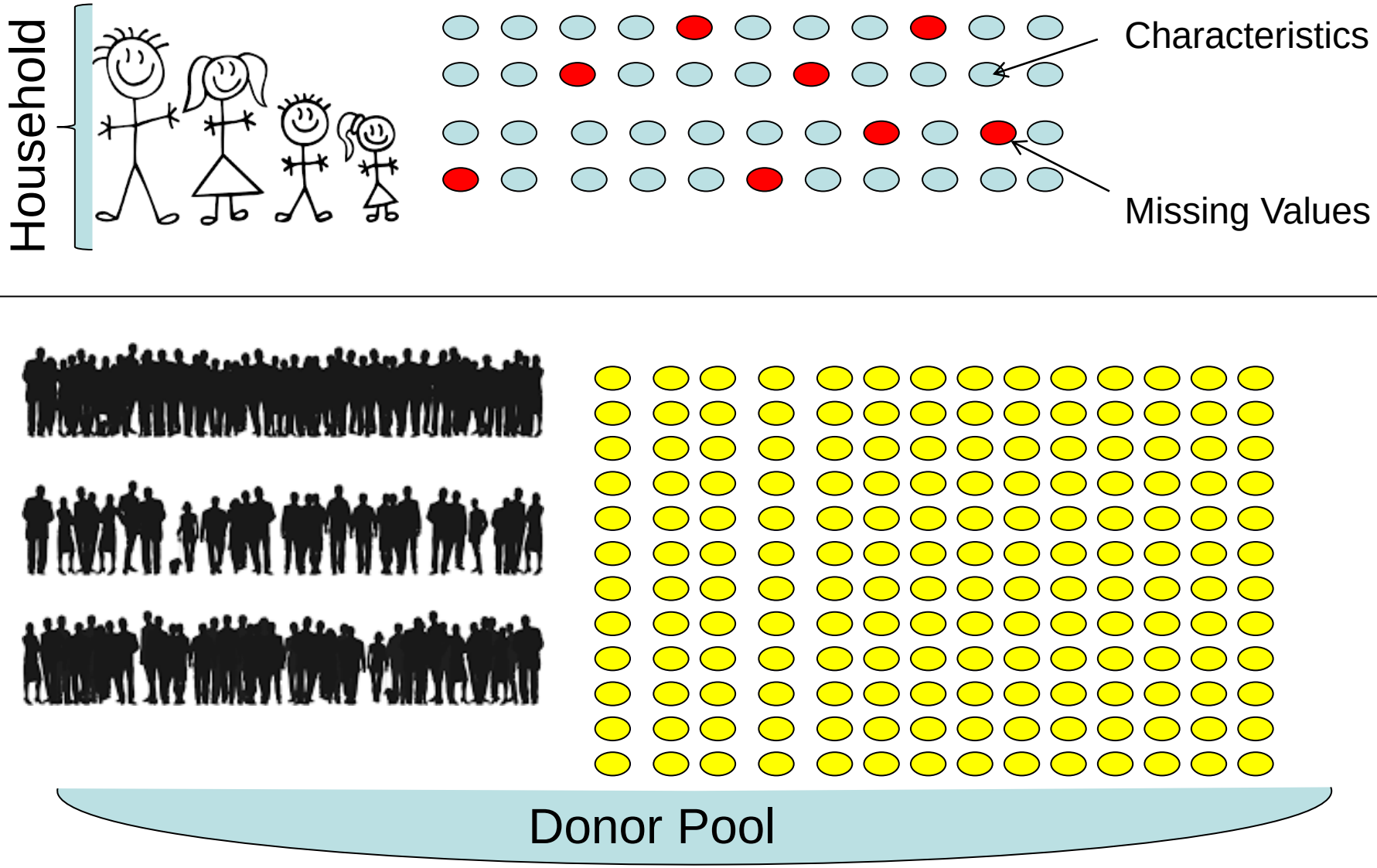
All records will undergo some level of imputation

Why CANCEIS?

The Imputation system used for the 2011 England and Wales Census.

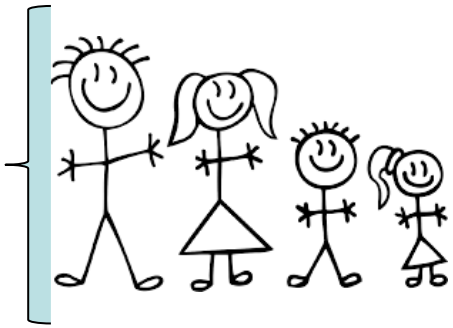
- Plausible by user defined edit rules
- Parameters can be manipulated as required
- Time saving as removes the need to develop a bespoke imputation model

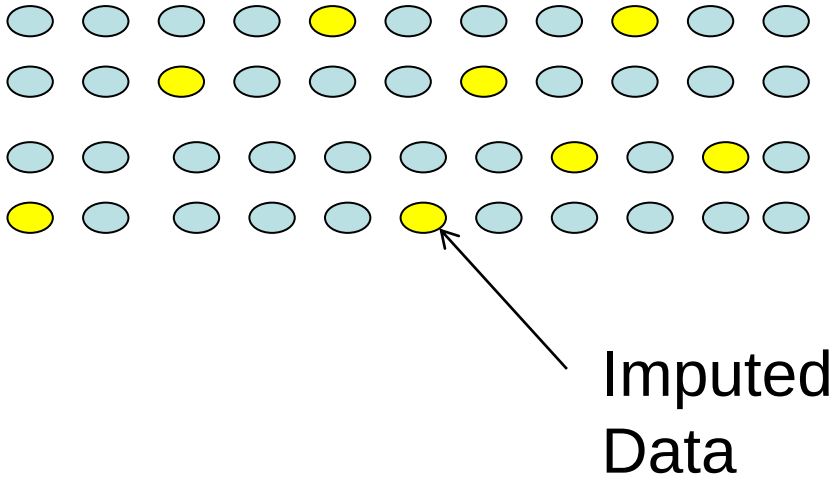
How it works



Output

Household





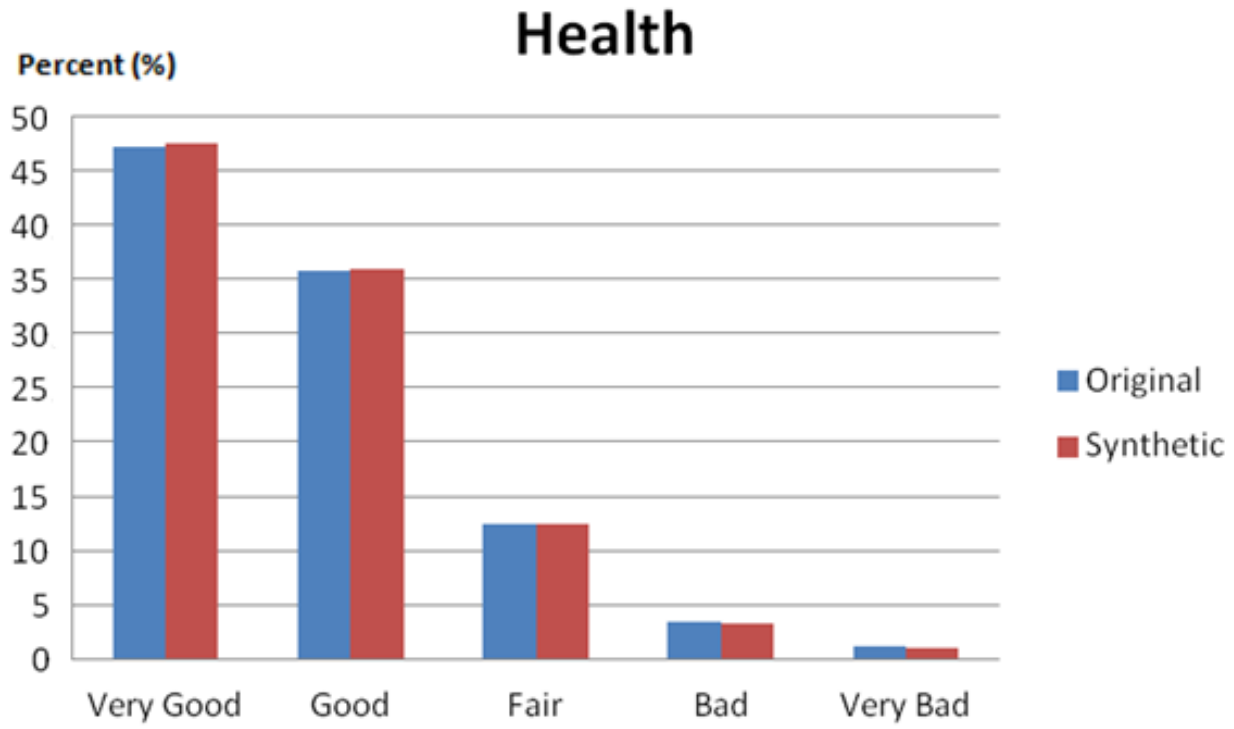
Utility Analysis

- Narrow Measures -
  - Marginal distributions
  - Level of agreement

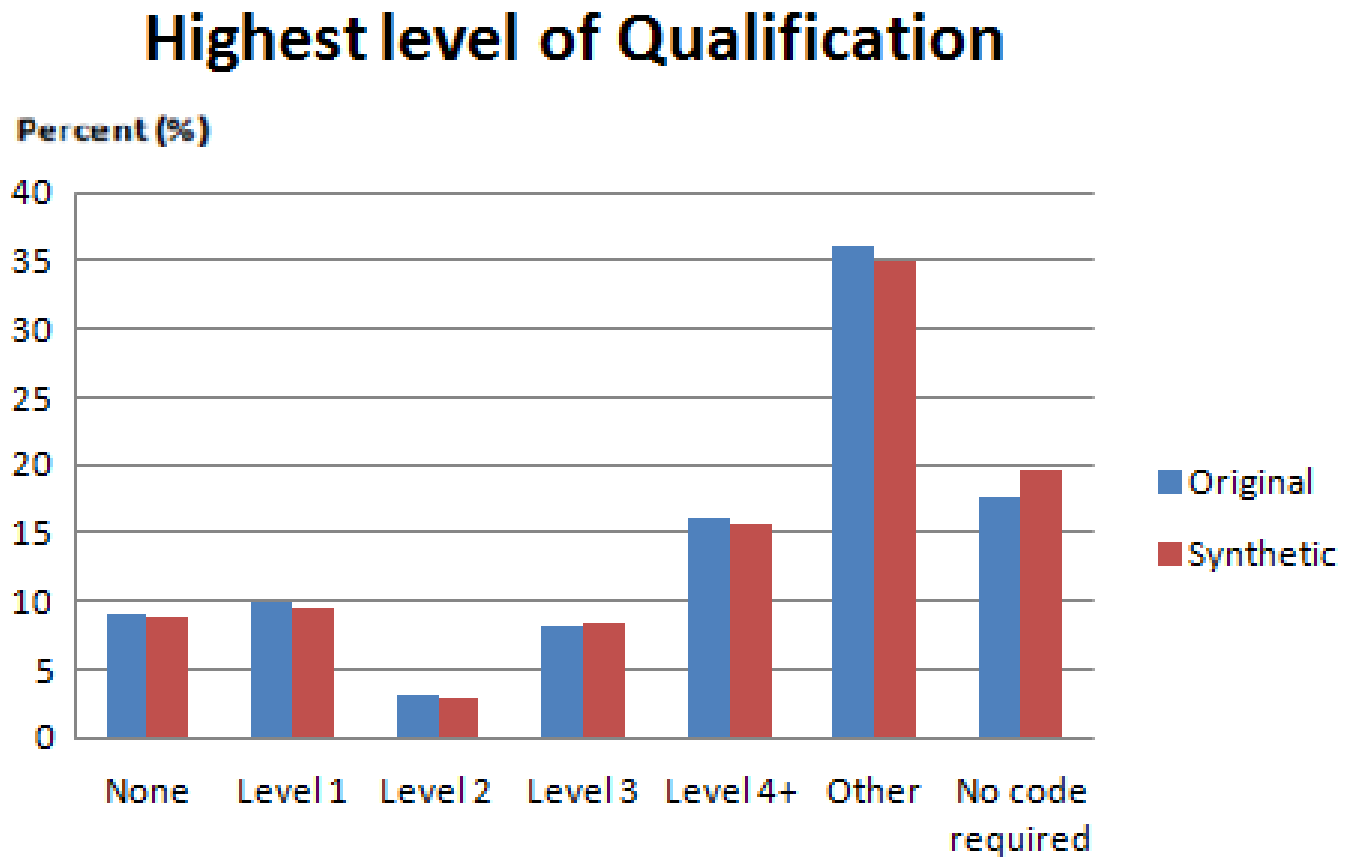
- Broad Measures -
  - Propensity Score measure

Initial results- Marginal Distributions

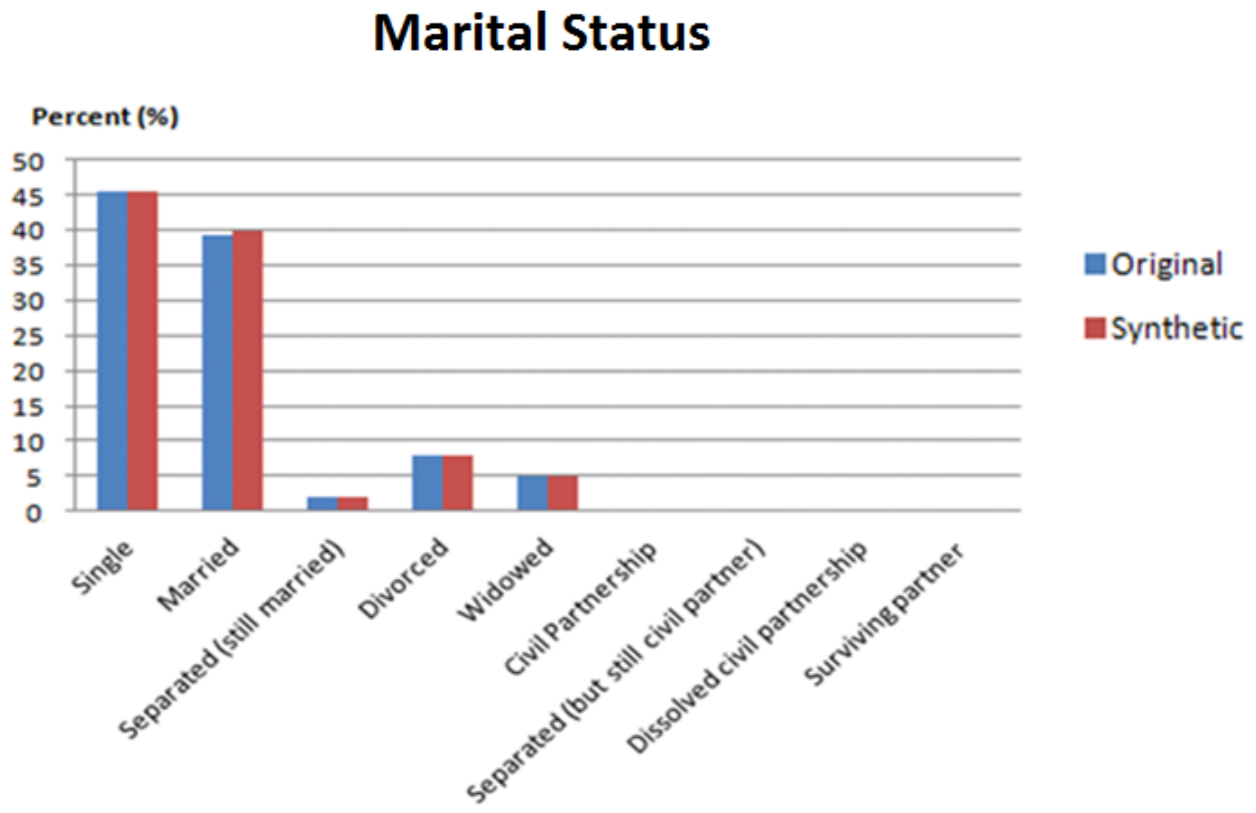
- Gives a macro level indication of how well the synthetic data matches the original data



Initial results- Marginal Distributions



Initial results- Marginal Distributions



Assessing Utility – Agreement

- Kappa Statistic – Measures agreement between original and synthetic variables, taking into account the probability of agreement by chance

- Calculated using a transition matrix

Transition Matrix

Marital Status		Before Imputation: Single	Married	Separated (still married)	Divorced	Widowed	Civil Partnership	Separated (still civil partnership)	Dissolved	Surviving
After Imputation	Single	43.5	1.2	0.1	0.5	0.1	0.0	0.0	0.0	0.0
	Married	0.8	37.8	0.1	0.4	0.0	0.0	0.0	0.0	0.0
	Separated (still married)	0.2	0.1	1.6	0.1	0.0	0.0	0.0	0.0	0.0
	Divorced	0.5	0.6	0.1	6.4	0.2	0.0	0.0	0.0	0.0
	Widowed	0.1	0.1	0.0	0.2	4.6	0.0	0.0	0.0	0.0
	Civil Partnership	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0
	Separated (Still civil partnership)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Dissolved	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Surviving	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

K = 0.83

K	Strength of Agreement
0.00	Poor
0.00 - 0.20	Slight
0.21 - 0.40	Fair
0.41- 0.60	Moderate
0.61- 0.80	Substantial
0.81-1.00	Almost Perfect

Kappa Scores

Variable	Kappa Score (K)
Health	0.77
Marital Status	0.83
Sex	0.69
Age	0.68
Carer	0.79
Student	0.86
Year last Worked	0.92

K	Strength of Agreement
0.00	Poor
0.00 - 0.20	Slight
0.21 - 0.40	Fair
0.41- 0.60	Moderate
0.61- 0.80	Substantial
0.81-1.00	Almost Perfect

Assessing Utility – Broad measure

Propensity score analysis

- Involves stacking the original and synthetic datasets and creating a dummy variable
- Dummy variable is used as the dependent variable in a logistic regression model
- Predicted probabilities close to 0.5

Propensity Score Analysis

- calculate the mean squared error to summarise overall level of utility

- When the original and synthetic data have similar distributions, U_p is near 0.

$U_p = 0.0020979$

Next Step - Disclosure Risk Analysis

- The purpose of synthesising data is to minimise the possibility of sensitive information being disclosed.
- Traditional disclosure control methods
- Intruder testing
- Make a decision

Conclusion

- Utility analysis suggests distributions between original and synthetic data are similar
- Suggesting a high level of analytical validity
- Disclosure risk analysis to assess whether the imputation provides sufficient protection

Robert.Rendell@ons.gov.uk