www.synthpop.org.uk

Assessing, visualizing and improving the utility of synthetic data

Gillian M Raab, Beata Nowok

University of Edinburgh

gillian.raab@ed.ac.uk   beata.nowok@ed.ac.uk

SCOTTISH CENTRE FOR ADMINISTRATIVE DATA RESEARCH

Presentation to the UNECE/Eurostat expert meeting on Statistical Data Confidentiality, December 2021

History of synthpop

Year	version	What
2014	1.0	Creation of one or more synthetic data sets from **real** data by conditional models with many choices of methods and other options. But defaults make this easy. *syn_version <- syn(data)* Visualisation of univariate comparisons and of the results of fitting models to synthetic and original data (specific utility)
2015	1.1	Statistical Disclosure Control (SDC) added

2016	1.2 and 1.3	Functions to calculate general utility measures added
		utility.gen(syn_version, data)
		utility.tab(syn_version, data)

2018	1.5	New methods of synthesis for categorical data added. Not conditional models. ***method = "catall" or method = "ipf"***
2021	1.7	Added many more utility measures and methods for visualising them ***utility.tables(syn_version, data, tables = "twoway")*** ***syn_version*** is an R object that holds the synthetic data as well as information about the synthesis method All the utility functions also work with synthetic data created by other methods.

Why measure general utility?

- Because the synthesis may not have given a good result
  - Errors in your code/ program
  - The synthesis model being used is not correct

- To compare different synthesis methods
  - As scores in a challenge/hackathon
  - To decide which method is best
- To tune your synthesis
  - Modify the options to get better utility

New outputs from utility functions

`utility.gen()`

pMSE

SPECKS

PO50

U (Wilcoxon)

`utility.tab()`

VW

FT

G

JSD

MabsDD

WMabsDD

dBhatt

pMSE

SPECKS

PO50

U (Wilcoxon)

General utility functions in synthpop

Two approaches in the literature

- *Propensity scores* `utility.gen(syn, orig, method =, vars = )`

originally version gave propensity score mean-square error (pMSE)

- *Tables* `utility.tab(syn, orig, vars = )`

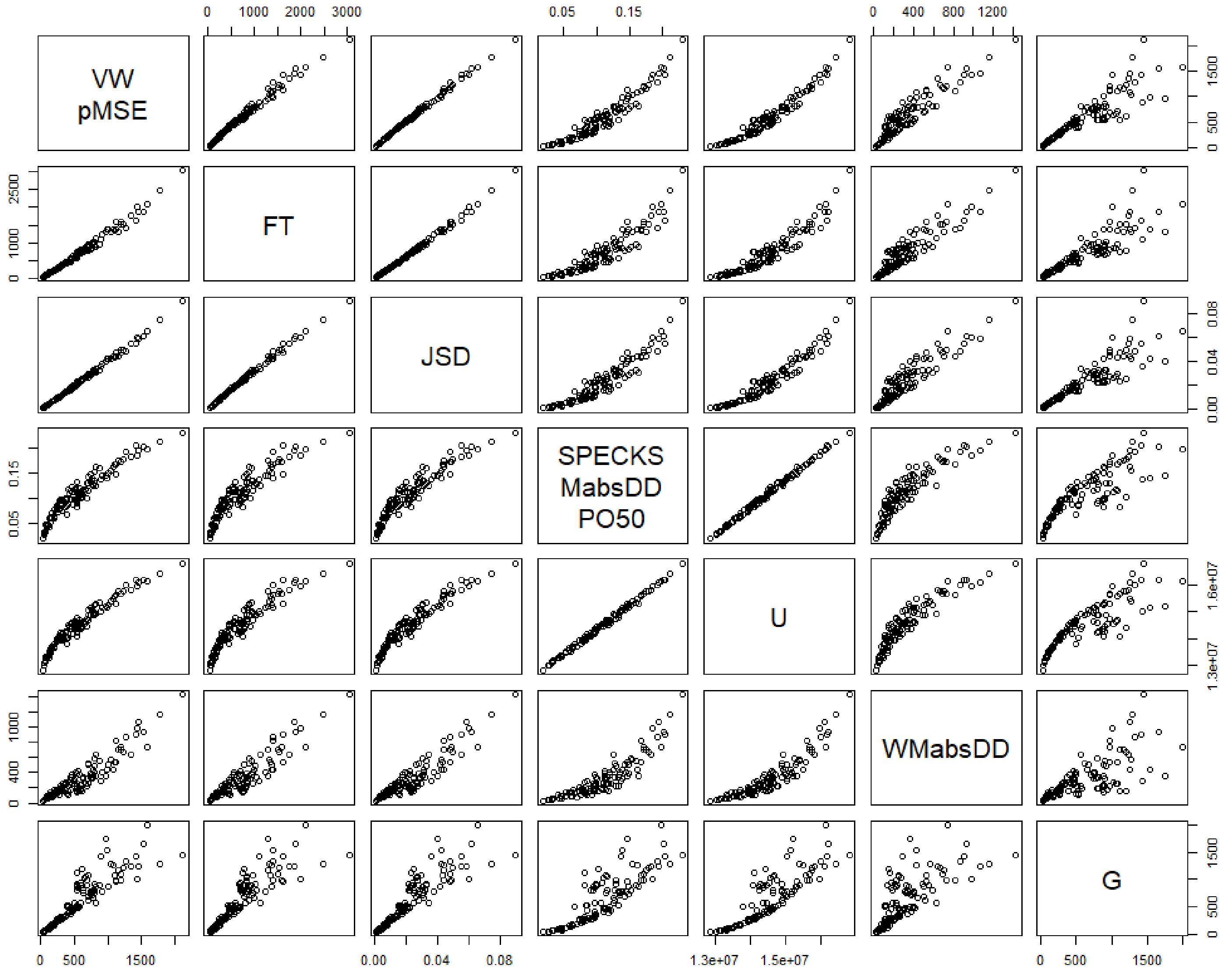
originally version gave Voas-Williamson Statistic (VW)

But tables can be framed as prediction models

- *Compare tables*
- *Proportion of synthetic counts gives the propensity score*
- *An n-way table is equivalent to fitting a logistic regression with all interactions up to order n*

This means that any measure calculated from propensity scores can also be calculated for tables

Comparing utility measures



Plots of 7 utility measures comparing the 120 3-way tables formed from all combinations of 10 variables.

Correlations = 1.0000
VW / pMSE

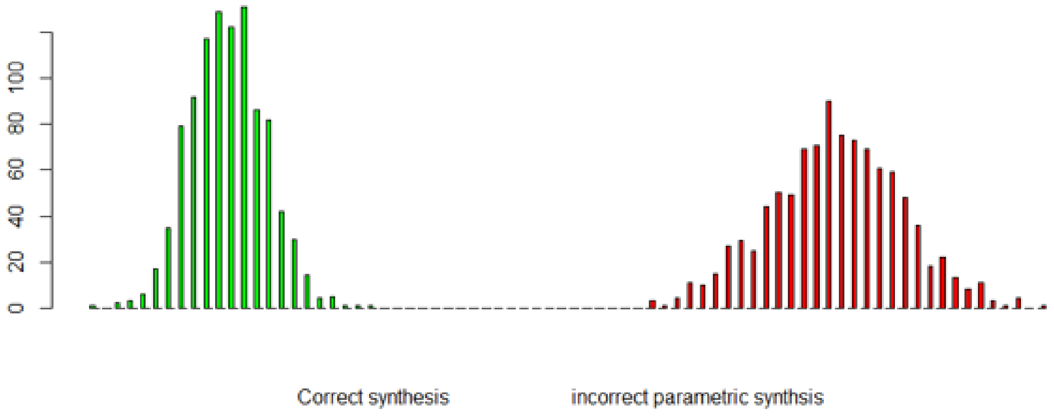
SPECKS/MabsDD/PO50

Correlations > 0.99
pMSE(VW) / FT(dBhatt) / JSD

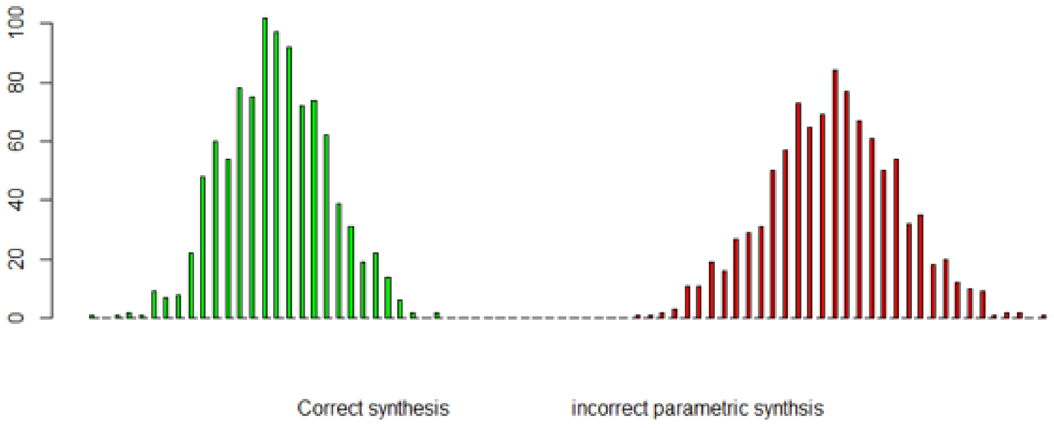
SPECKS(PO50 MabsDD) / U

Assessing discrimination power 4 variables

pMSE / VW measures



SPECKS / PO50 / MabsDD



- Make 1,000 syntheses of two synthetic data sets
- Original has just 4 categorical variables
- The correct synthesis makes a 4 way table and synthesises from that `(method = "catall")`
- The incorrect synthesis uses conditional synthesis with logistic or multinomial models

- Two utility measures
  pMSE/VW or SPECKS/PO50/MabsDD
- Power is difference in means standardised by the s.d. of the correct model

- Ratio for pMSE/VW 16.2
- Ratio for SPECKS/PO50/MabsDD 10.7

Which utility measure to choose?

- *Paper suggests pMSE/VW – but others e.g.*

  SPECKS/PO50/MabsDD could do
- *Can be computed from propensity score or from tables*
- *Has a known expected value so can be standardised with respect to expected stochastic variation for a correct model either from a formula or replication from a single synthesis (SPECKS needs multiple syntheses)*
- *Seems to have good power in some simulations*

What method should be used to get a single measure?

- *Practical considerations are important*
  - *Very large cross-tabulations run out of space*
  - *Parametric propensity score models often fail for many variables*
  - *CART propensity score models seem the best bet*

Does it have to be just one number?

Measures for tuning syntheses

A general utility measure does not have to be just one number

  - *Not helpful in diagnosing problems and tuning syntheses*
  - *Can be a collection of measures*
  - *Any set of summaries (means, correlations etc.) will do*
  - *Our preference is for sets of tables 1-way, 2-way 3-way with continuous variables grouped*

- *Tuning synthpop can involve*
  - *Changing methods for some variables*
  - *Changing order of conditional distributions*
  - *Restricting prediction matrix*
  - *Stratifying the synthesis*

Suggested procedure

- *Start with all 1 way tables*

  ```
  compare.syn(synthetic, original)
  ```

  now also gives a table of utilities

- *Fix if not OK*

- *Look at 2 way tables*

  ```
  utility.tables(synthetic, original, tables = "twoway")
  ```

  produces tables and averages of utilities (possible summary measure)

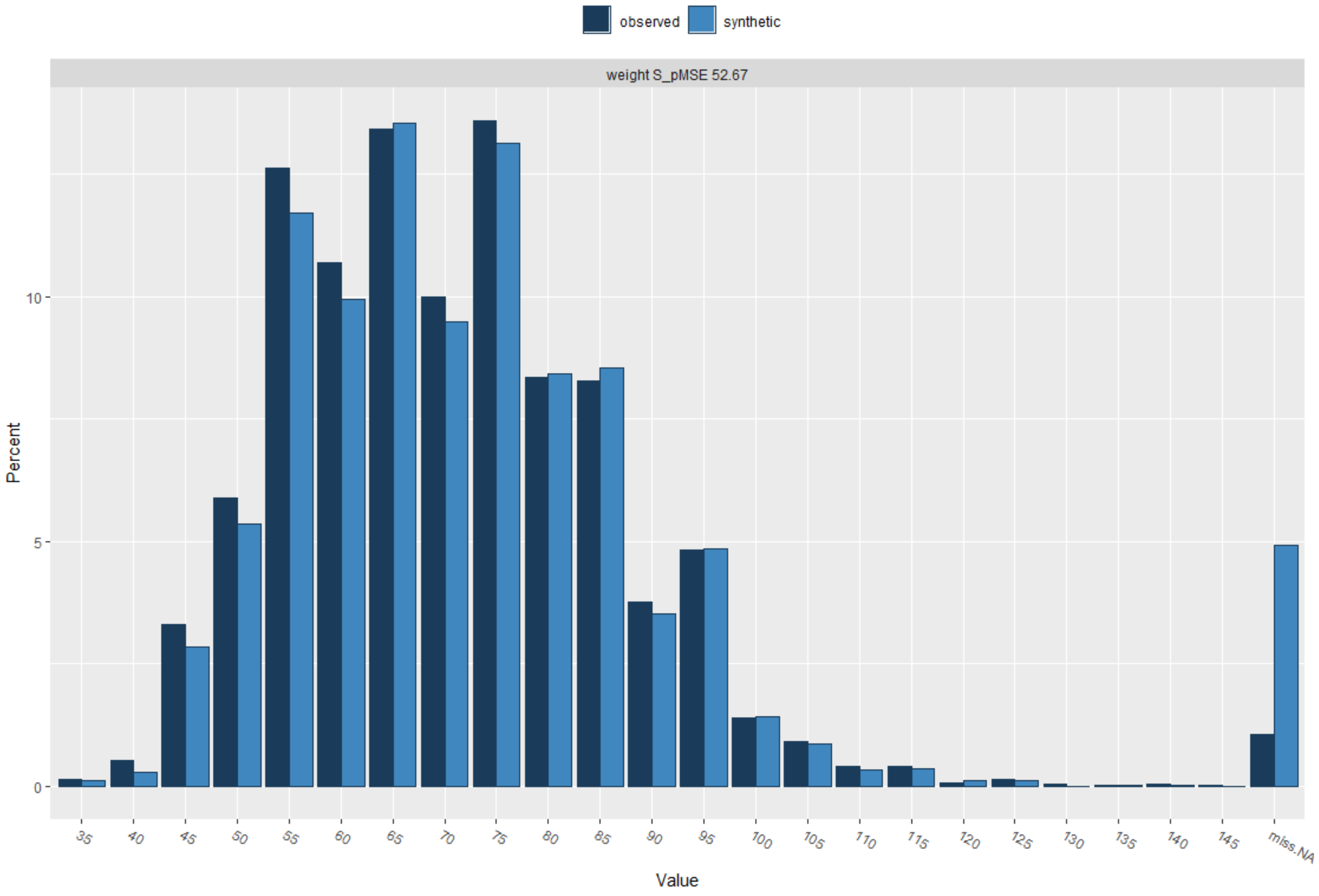
- *Fix if not OK*

- *Perhaps move on to 3 way tables*

  ```
  utility.tables(synthetic, original, tables = "threeway")
  ```

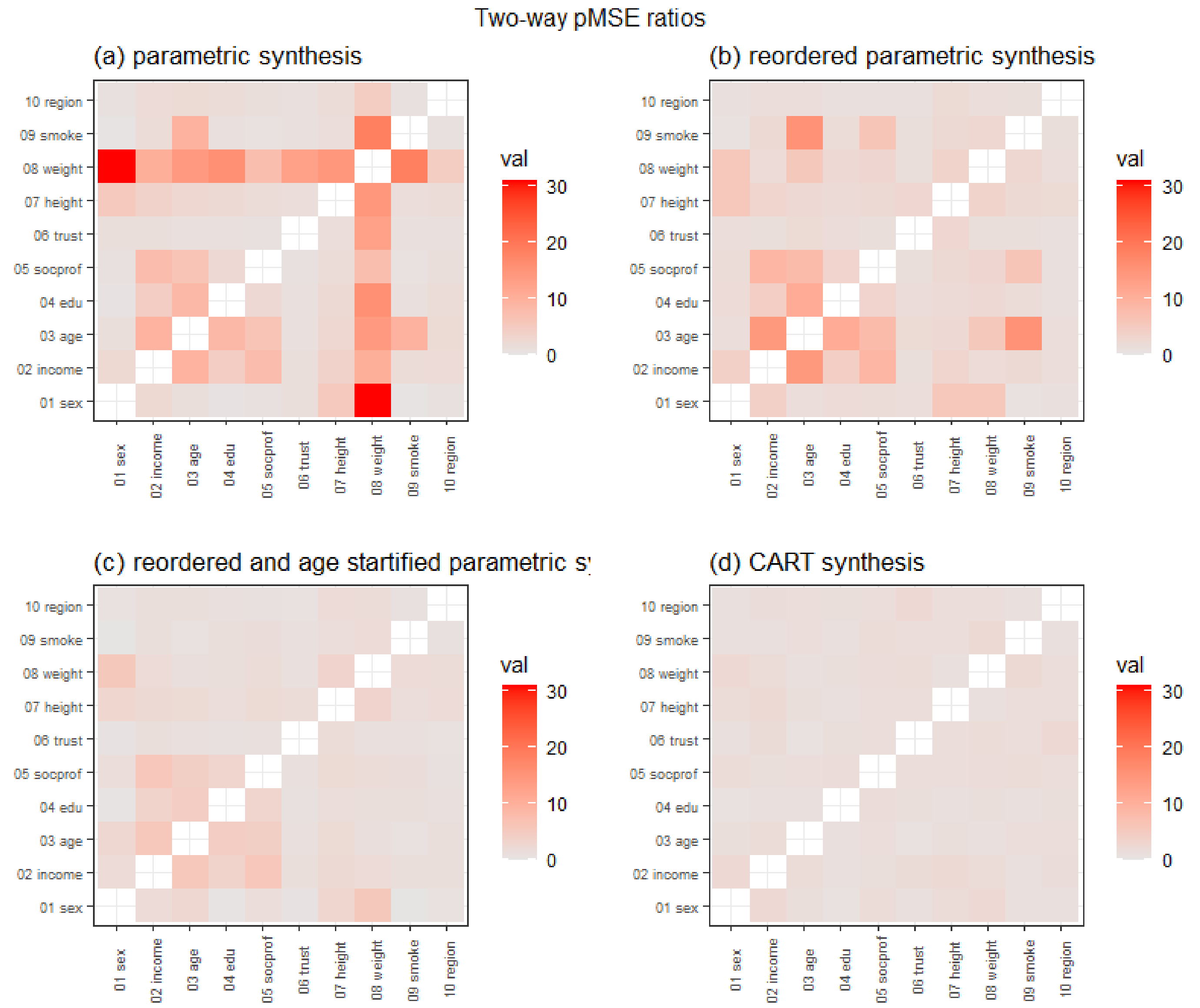
  produces tables and averages of utilities (possible summary measure)

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
```

Two way tables



ıetric synthesis shows
em with weight variable

omes it by reordering

ying by age helps a bit

ing a CART synthesis
have been better

Conclusions

- If you want a single number to compare syntheses then
  - either the pMSE/VW or SPECKS/MabsDD/PO50 from a CART model
  - Or get averages of tabular utilities from all possible 2 or 3 way tables
- For tuning
  - Visualise one-way, two-way or 3-way tables

More detail in our paper

Even more detail in the preprint

Assessing, visualizing and improving the utility of synthetic data

Gillian M Raab, Beata Nowok, Chris Dibben (2021)

Available from

https://arxiv.org/abs/2109.12717