THE UNIVERSITY of EDINBURGH

Synthetic data in practice: software, applications and challenges

Beata Nowok

Gillian Raab & Chris Dibben

Administrative Data Research Centre – Scotland

Administrative Data Research Centre Scotland

An ESRC Data Investment

Completely synthetic microdata

- Statistical disclosure control (SDC) method for individual level data
- All values of all variables are generated from statistical models - microdata set of artificial units only

Observed

Sex	Age	Education	Marital status	Income	Life satisfaction
FEMALE	57	VOCATIONAL/GRAMMAR	MARRIED	800	PLEASED
MALE	41	SECONDARY	UNMARRIED	1500	MIXED
FEMALE	18	VOCATIONAL/GRAMMAR	UNMARRIED	NA	PLEASED
FEMALE	78	PRIMARY/NO EDUCATION	WIDOWED	900	MIXED
FEMALE	54	VOCATIONAL/GRAMMAR	MARRIED	1500	MOSTLY SATISFIED
MALE	20	SECONDARY	UNMARRIED	-8	PLEASED
FEMALE	39	SECONDARY	MARRIED	2000	MOSTLY SATISFIED
MALE	39	SECONDARY	MARRIED	1197	MIXED
FEMALE	38	VOCATIONAL/GRAMMAR	MARRIED	NA	MOSTLY DISSATISFIED
FEMALE	73	VOCATIONAL/GRAMMAR	WIDOWED	1700	PLEASED
FEMALE	54	SECONDARY	WIDOWED	2000	MOSTLY SATISFIED
MALE	30	VOCATIONAL/GRAMMAR	UNMARRIED	900	MOSTLY SATISFIED
MALE	68	SECONDARY	MARRIED	-8	DELIGHTED
MALE	61	PRIMARY/NO EDUCATION	MARRIED	-8	MIXED

Synthetic

Sex	Age	Education	Marital status	Income	Life satisfaction
MALE	81	PRIMARY/NO EDUCATION	MARRIED	2100	PLEASED
MALE	54	VOCATIONAL/GRAMMAR	MARRIED	1700	PLEASED
FEMALE	32	VOCATIONAL/GRAMMAR	DIVORCED	870	MIXED
FEMALE	98	PRIMARY/NO EDUCATION	MARRIED	800	MOSTLY DISSATISFIED
FEMALE	50	PRIMARY/NO EDUCATION	MARRIED	NA	MOSTLY SATISFIED
FEMALE	37	VOCATIONAL/GRAMMAR	MARRIED	158	PLEASED
MALE	28	VOCATIONAL/GRAMMAR	NA	1500	MOSTLY SATISFIED
FEMALE	62	PRIMARY/NO EDUCATION	MARRIED	830	MOSTLY SATISFIED
MALE	78	PRIMARY/NO EDUCATION	MARRIED	NA	PLEASED
FEMALE	29	SECONDARY	MARRIED	580	MOSTLY SATISFIED
MALE	59	PRIMARY/NO EDUCATION	MARRIED	1300	MOSTLY SATISFIED
MALE	41	SECONDARY	UNMARRIED	1500	MIXED
MALE	18	SECONDARY	UNMARRIED	-8	PLEASED
FEMALE	73	PRIMARY/NO EDUCATION	WIDOWED	1350	MOSTLY SATISFIED

Completely synthetic microdata

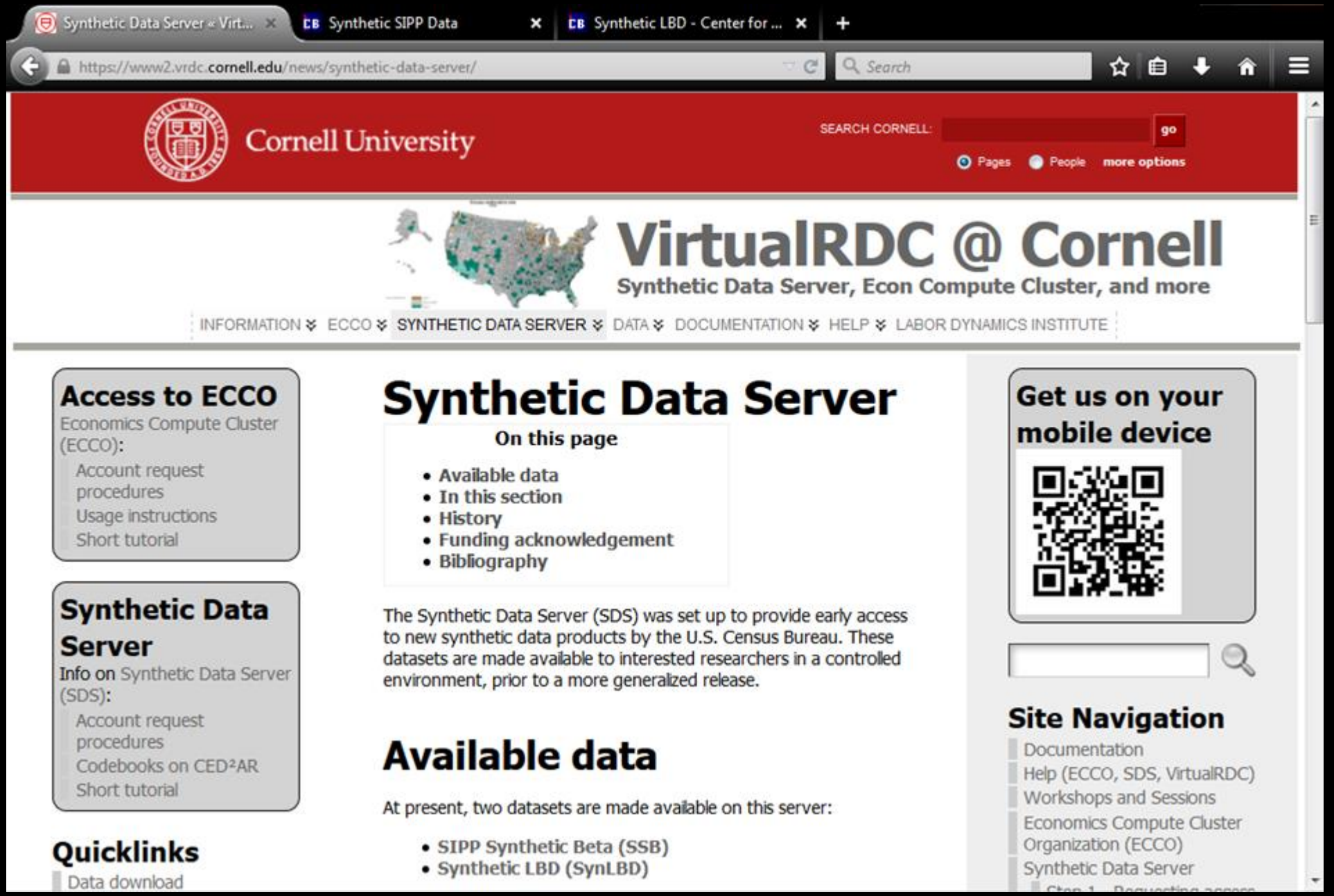
- Original data are used to inform the models (to reproduce their essential features)
- Synthetic data are only as good as the models that created them

The potential of synthetic data

- Research
- Teaching

Synthetic data in practise: research

- Initially meant to be used in the place of the real data but applications are now more cautious
- Validating results on the original, confidential version of the data
- For preparatory work and preliminary analysis (to develop code for data preparation and analysis)



https://www2.vrdc.cornell.edu/news/synthetic-data-server/

Scottish Longitudinal Study (SLS)

Sample from the Census linked to administrative data with restricted access (safe settings in Edinburgh)

SYLLS

SYNTHETIC DATA ESTIMATION FOR UK LONGITUDINAL STUDIES

SLS projects

- Researchers are provided with a single synthetic version of a project-specific extract
- Preliminary analysis and the development of analysis code
- Final analyses have to be run on the original data

SLS projects

- Only approved researchers who are granted access to the original sensitive data after signing a disclaimer
- Data are labelled appropriately to make it clear that they are synthetic
- Substantial costs and time savings related to visits to safe havens

The potential of synthetic data: teaching

- Instead of real data
- To develop practical data analysis skills

Synthetic data sets in practice: teaching

- Sample from the Scottish Morbidity Record (SMR01) - Medical Informatics course

- SLS extract - An introduction to using linked administrative data for social and health research (27 Nov – 1 Dec 2017)

How do we make synthetic data?

- Fitting statistical models to the original data and generating completely new records for public release
- Joint distribution is approximated by a set of conditional distributions

Observed

Sex	Age	Education	Marital status	Income	Life satisfaction
FEMALE	57	VOCATIONAL/GRAMMAR	MARRIED	800	PLEASED
MALE	41	SECONDARY	UNMARRIED	1500	MIXED
FEMALE	18	VOCATIONAL/GRAMMAR	UNMARRIED	NA	PLEASED
FEMALE	78	PRIMARY/NO EDUCATION	WIDOWED	900	MIXED
FEMALE	54	VOCATIONAL/GRAMMAR	MARRIED	1500	MOSTLY SATISFIED
MALE	20	SECONDARY	UNMARRIED	-8	PLEASED
FEMALE	39	SECONDARY	MARRIED	2000	MOSTLY SATISFIED
MALE	39	SECONDARY	MARRIED	1197	MIXED
FEMALE	38	VOCATIONAL/GRAMMAR	MARRIED	NA	MOSTLY DISSATISFIED
FEMALE	73	VOCATIONAL/GRAMMAR	WIDOWED	1700	PLEASED
FEMALE	54	SECONDARY	WIDOWED	2000	MOSTLY SATISFIED
MALE	30	VOCATIONAL/GRAMMAR	UNMARRIED	900	MOSTLY SATISFIED
MALE	68	SECONDARY	MARRIED	-8	DELIGHTED
MALE	61	PRIMARY/NO EDUCATION	MARRIED	-8	MIXED

Sex distribution

Generate **Sex**

Sex
MALE
MALE
FEMALE
FEMALE
FEMALE
FEMALE
MALE
FEMALE
MALE
FEMALE
MALE
MALE
MALE
FEMALE

Observed

Sex	Age	Education	Marital status	Income	Life satisfaction
FEMALE	57	VOCATIONAL/GRAMMAR	MARRIED	800	PLEASED
MALE	41	SECONDARY	UNMARRIED	1500	MIXED
FEMALE	18	VOCATIONAL/GRAMMAR	UNMARRIED	NA	PLEASED
FEMALE	78	PRIMARY/NO EDUCATION	WIDOWED	900	MIXED
FEMALE	54	VOCATIONAL/GRAMMAR	MARRIED	1500	MOSTLY SATISFIED
MALE	20	SECONDARY	UNMARRIED	-8	PLEASED
FEMALE	39	SECONDARY	MARRIED	2000	MOSTLY SATISFIED
MALE	39	SECONDARY	MARRIED	1197	MIXED
FEMALE	38	VOCATIONAL/GRAMMAR	MARRIED	NA	MOSTLY DISSATISFIED
FEMALE	73	VOCATIONAL/GRAMMAR	WIDOWED	1700	PLEASED
FEMALE	54	SECONDARY	WIDOWED	2000	MOSTLY SATISFIED
MALE	30	VOCATIONAL/GRAMMAR	UNMARRIED	900	MOSTLY SATISFIED
MALE	68	SECONDARY	MARRIED	-8	DELIGHTED
MALE	61	PRIMARY/NO EDUCATION	MARRIED	-8	MIXED

Age predicted from **Sex**

Sex
MALE
MALE
FEMALE
FEMALE
FEMALE
FEMALE
MALE
FEMALE
MALE
FEMALE
MALE
MALE
MALE
FEMALE

Generate **Age**

Age
81
54
32
98
50
37
28
62
78
29
59
41
18
73

Observed

Sex	Age	Education	Marital status	Income	Life satisfaction
FEMALE	57	VOCATIONAL/GRAMMAR	MARRIED	800	PLEASED
MALE	41	SECONDARY	UNMARRIED	1500	MIXED
FEMALE	18	VOCATIONAL/GRAMMAR	UNMARRIED	NA	PLEASED
FEMALE	78	PRIMARY/NO EDUCATION	WIDOWED	900	MIXED
FEMALE	54	VOCATIONAL/GRAMMAR	MARRIED	1500	MOSTLY SATISFIED
MALE	20	SECONDARY	UNMARRIED	-8	PLEASED
FEMALE	39	SECONDARY	MARRIED	2000	MOSTLY SATISFIED
MALE	39	SECONDARY	MARRIED	1197	MIXED
FEMALE	38	VOCATIONAL/GRAMMAR	MARRIED	NA	MOSTLY DISSATISFIED
FEMALE	73	VOCATIONAL/GRAMMAR	WIDOWED	1700	PLEASED
FEMALE	54	SECONDARY	WIDOWED	2000	MOSTLY SATISFIED
MALE	30	VOCATIONAL/GRAMMAR	UNMARRIED	900	MOSTLY SATISFIED
MALE	68	SECONDARY	MARRIED	-8	DELIGHTED
MALE	61	PRIMARY/NO EDUCATION	MARRIED	-8	MIXED

Education predicted from **Sex** and **Age**

Sex	Age
MALE	81
MALE	54
FEMALE	32
FEMALE	98
FEMALE	50
FEMALE	37
MALE	28
FEMALE	62
MALE	78
FEMALE	29
MALE	59
MALE	41
MALE	18
FEMALE	73

Generate **Education**

Education
PRIMARY/NO EDUCATION
VOCATIONAL/GRAMMAR
VOCATIONAL/GRAMMAR
PRIMARY/NO EDUCATION
PRIMARY/NO EDUCATION
VOCATIONAL/GRAMMAR
VOCATIONAL/GRAMMAR
PRIMARY/NO EDUCATION
PRIMARY/NO EDUCATION
SECONDARY
PRIMARY/NO EDUCATION
SECONDARY
SECONDARY
PRIMARY/NO EDUCATION

Observed

Sex	Age	Education	Marital status	Income	Life satisfaction
FEMALE	57	VOCATIONAL/GRAMMAR	MARRIED	800	PLEASED
MALE	41	SECONDARY	UNMARRIED	1500	MIXED
FEMALE	18	VOCATIONAL/GRAMMAR	UNMARRIED	NA	PLEASED
FEMALE	78	PRIMARY/NO EDUCATION	WIDOWED	900	MIXED
FEMALE	54	VOCATIONAL/GRAMMAR	MARRIED	1500	MOSTLY SATISFIED
MALE	20	SECONDARY	UNMARRIED	-8	PLEASED
FEMALE	39	SECONDARY	MARRIED	2000	MOSTLY SATISFIED
MALE	39	SECONDARY	MARRIED	1197	MIXED
FEMALE	38	VOCATIONAL/GRAMMAR	MARRIED	NA	MOSTLY DISSATISFIED
FEMALE	73	VOCATIONAL/GRAMMAR	WIDOWED	1700	PLEASED
FEMALE	54	SECONDARY	WIDOWED	2000	MOSTLY SATISFIED
MALE	30	VOCATIONAL/GRAMMAR	UNMARRIED	900	MOSTLY SATISFIED
MALE	68	SECONDARY	MARRIED	-8	DELIGHTED
MALE	61	PRIMARY/NO EDUCATION	MARRIED	-8	MIXED

Life satisfaction predicted from all other variables

Sex	Age	Education	Marital status	Income
MALE	81	PRIMARY/NO EDUCATION	MARRIED	2100
MALE	54	VOCATIONAL/GRAMMAR	MARRIED	1700
FEMALE	32	VOCATIONAL/GRAMMAR	DIVORCED	870
FEMALE	98	PRIMARY/NO EDUCATION	MARRIED	800
FEMALE	50	PRIMARY/NO EDUCATION	MARRIED	NA
FEMALE	37	VOCATIONAL/GRAMMAR	MARRIED	158
MALE	28	VOCATIONAL/GRAMMAR	NA	1500
FEMALE	62	PRIMARY/NO EDUCATION	MARRIED	830
MALE	78	PRIMARY/NO EDUCATION	MARRIED	NA
FEMALE	29	SECONDARY	MARRIED	580
MALE	59	PRIMARY/NO EDUCATION	MARRIED	1300
MALE	41	SECONDARY	UNMARRIED	1500
MALE	18	SECONDARY	UNMARRIED	-8
FEMALE	73	PRIMARY/NO EDUCATION	WIDOWED	1350

Generate **Life satisfaction**

Life satisfaction
PLEASED
PLEASED
MIXED
MOSTLY DISSATISFIED
MOSTLY SATISFIED
PLEASED
MOSTLY SATISFIED
MOSTLY SATISFIED
PLEASED
MOSTLY SATISFIED
MOSTLY SATISFIED
MIXED
PLEASED
MOSTLY SATISFIED

Observed

Sex	Age	Education	Marital status	Income	Life satisfaction
FEMALE	57	VOCATIONAL/GRAMMAR	MARRIED	800	PLEASED
MALE	41	SECONDARY	UNMARRIED	1500	MIXED
FEMALE	18	VOCATIONAL/GRAMMAR	UNMARRIED	NA	PLEASED
FEMALE	78	PRIMARY/NO EDUCATION	WIDOWED	900	MIXED
FEMALE	54	VOCATIONAL/GRAMMAR	MARRIED	1500	MOSTLY SATISFIED
MALE	20	SECONDARY	UNMARRIED	-8	PLEASED
FEMALE	39	SECONDARY	MARRIED	2000	MOSTLY SATISFIED
MALE	39	SECONDARY	MARRIED	1197	MIXED
FEMALE	38	VOCATIONAL/GRAMMAR	MARRIED	NA	MOSTLY DISSATISFIED
FEMALE	73	VOCATIONAL/GRAMMAR	WIDOWED	1700	PLEASED
FEMALE	54	SECONDARY	WIDOWED	2000	MOSTLY SATISFIED
MALE	30	VOCATIONAL/GRAMMAR	UNMARRIED	900	MOSTLY SATISFIED
MALE	68	SECONDARY	MARRIED	-8	DELIGHTED
MALE	61	PRIMARY/NO EDUCATION	MARRIED	-8	MIXED

Synthetic

Sex	Age	Education	Marital status	Income	Life satisfaction
MALE	81	PRIMARY/NO EDUCATION	MARRIED	2100	PLEASED
MALE	54	VOCATIONAL/GRAMMAR	MARRIED	1700	PLEASED
FEMALE	32	VOCATIONAL/GRAMMAR	DIVORCED	870	MIXED
FEMALE	98	PRIMARY/NO EDUCATION	MARRIED	800	MOSTLY DISSATISFIED
FEMALE	50	PRIMARY/NO EDUCATION	MARRIED	NA	MOSTLY SATISFIED
FEMALE	37	VOCATIONAL/GRAMMAR	MARRIED	158	PLEASED
MALE	28	VOCATIONAL/GRAMMAR	NA	1500	MOSTLY SATISFIED
FEMALE	62	PRIMARY/NO EDUCATION	MARRIED	830	MOSTLY SATISFIED
MALE	78	PRIMARY/NO EDUCATION	MARRIED	NA	PLEASED
FEMALE	29	SECONDARY	MARRIED	580	MOSTLY SATISFIED
MALE	59	PRIMARY/NO EDUCATION	MARRIED	1300	MOSTLY SATISFIED
MALE	41	SECONDARY	UNMARRIED	1500	MIXED
MALE	18	SECONDARY	UNMARRIED	-8	PLEASED
FEMALE	73	PRIMARY/NO EDUCATION	WIDOWED	1350	MOSTLY SATISFIED

A software tool for producing synthetic versions of sensitive microdata

package

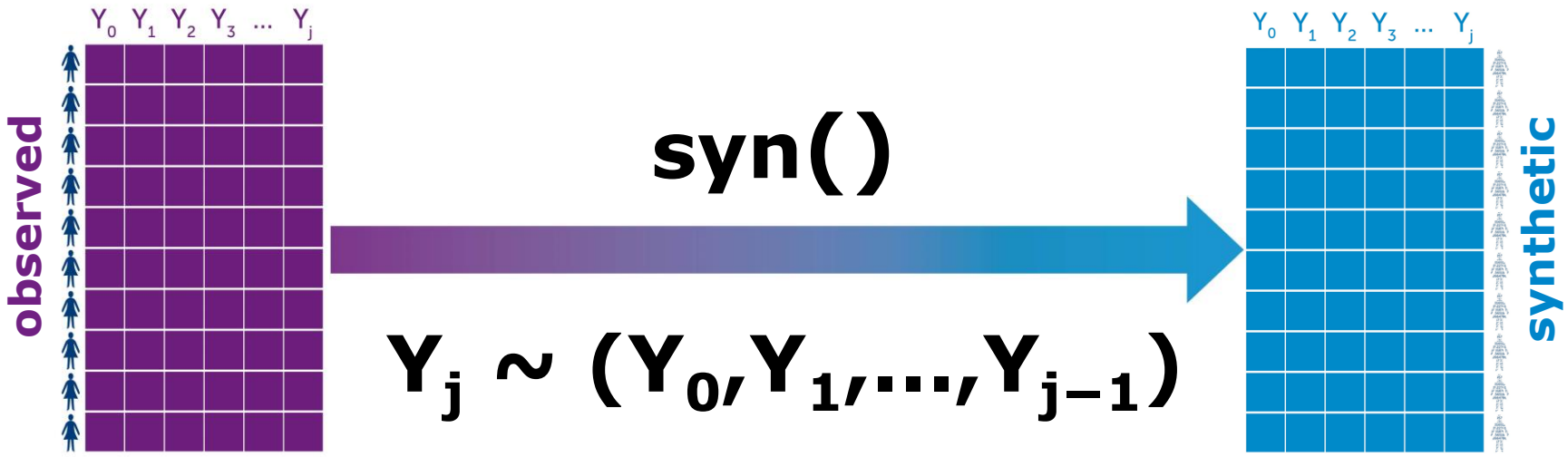
synthpop

version 1.3-2

http://cran.r-project.org/package=synthpop
https://github.com/bnowok/synthpop

https://www.jstatsoft.org/article/view/v074i11

Generating synthetic data: synthpop



syn(): Default parameters

- CART methods
- A single synthetic data set
- In the order of variables in the data set
- All previously synthesised variables as predictors
- No specification of rules
- No smoothing of continuous variables
- No coding of missing value indicators
- No stratification into subgroups

syn(): common data problems

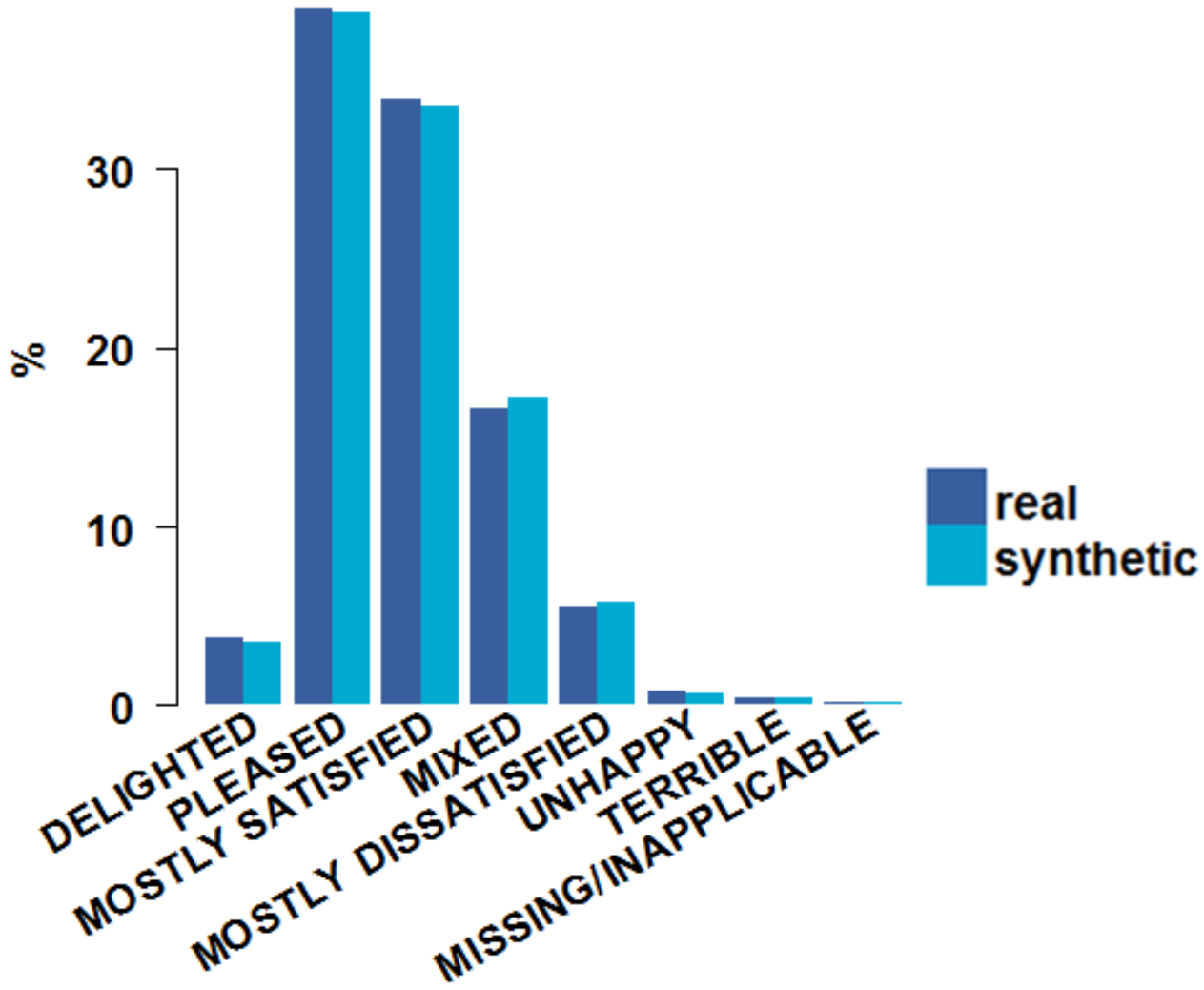
- Missing-data codes: `cont.na`
  - categorical variables: additional factor level(s)
  - continuous variables: specified by `cont.na` and modelled separately
- Semi-continuous variables: `semicont`
- Restricted values (interrelationships between variables): `rules` & `rvalues`
- Non-negativity / non-normality: `method` set to ‘`lognorm`’, ‘`sqrtnorm`’ or ‘`cubertnorm`’
- Deterministic relations: `method` set to “`~I(…)`”

Synthetic data utility

- Distribution of individual variables
- Multiway cross-tabulations
- Analysis-specific measures - model fits
- Propensity scores to discriminate between the original and synthetic data

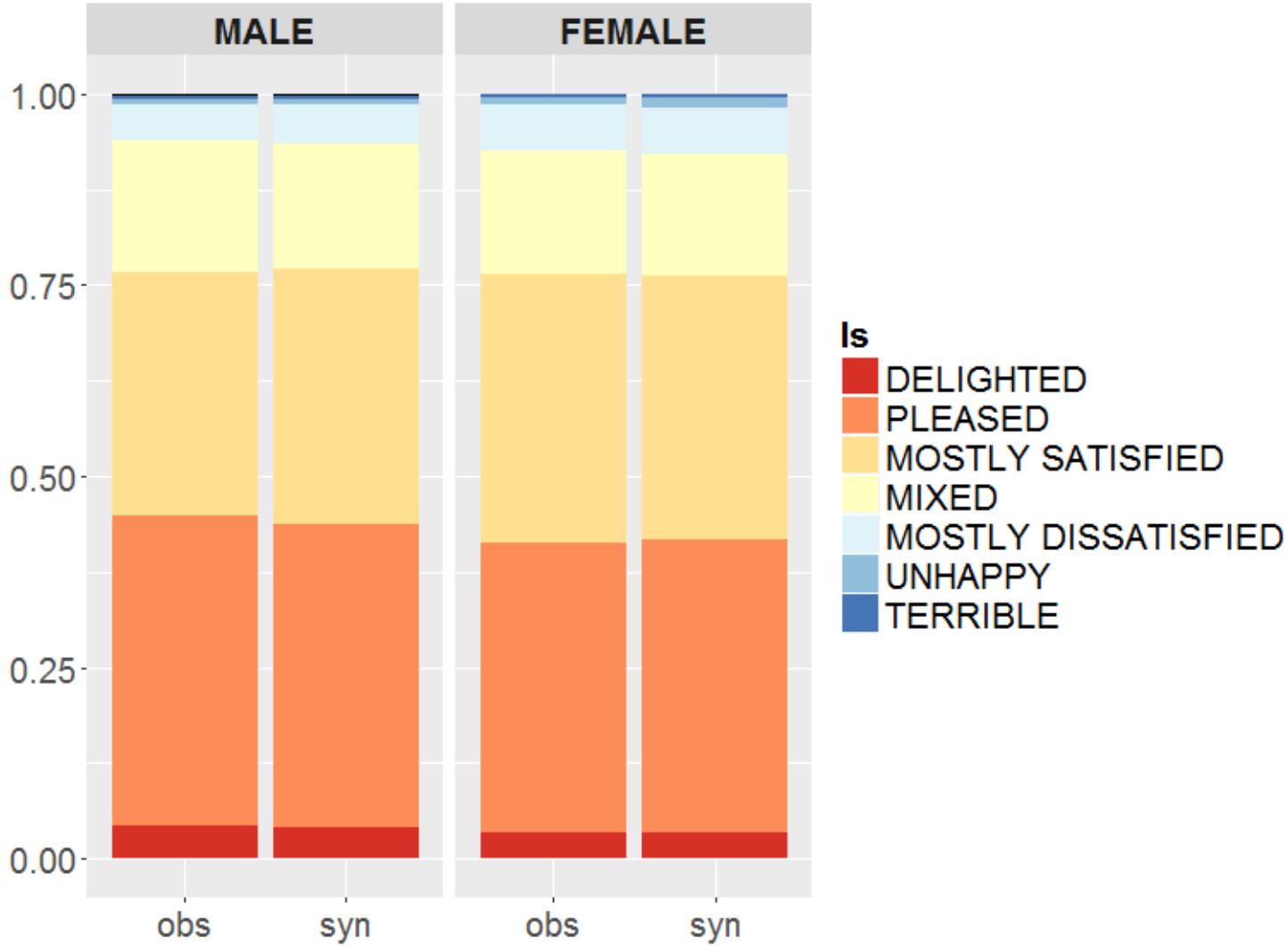
Observed vs. synthetic data

Life satisfaction



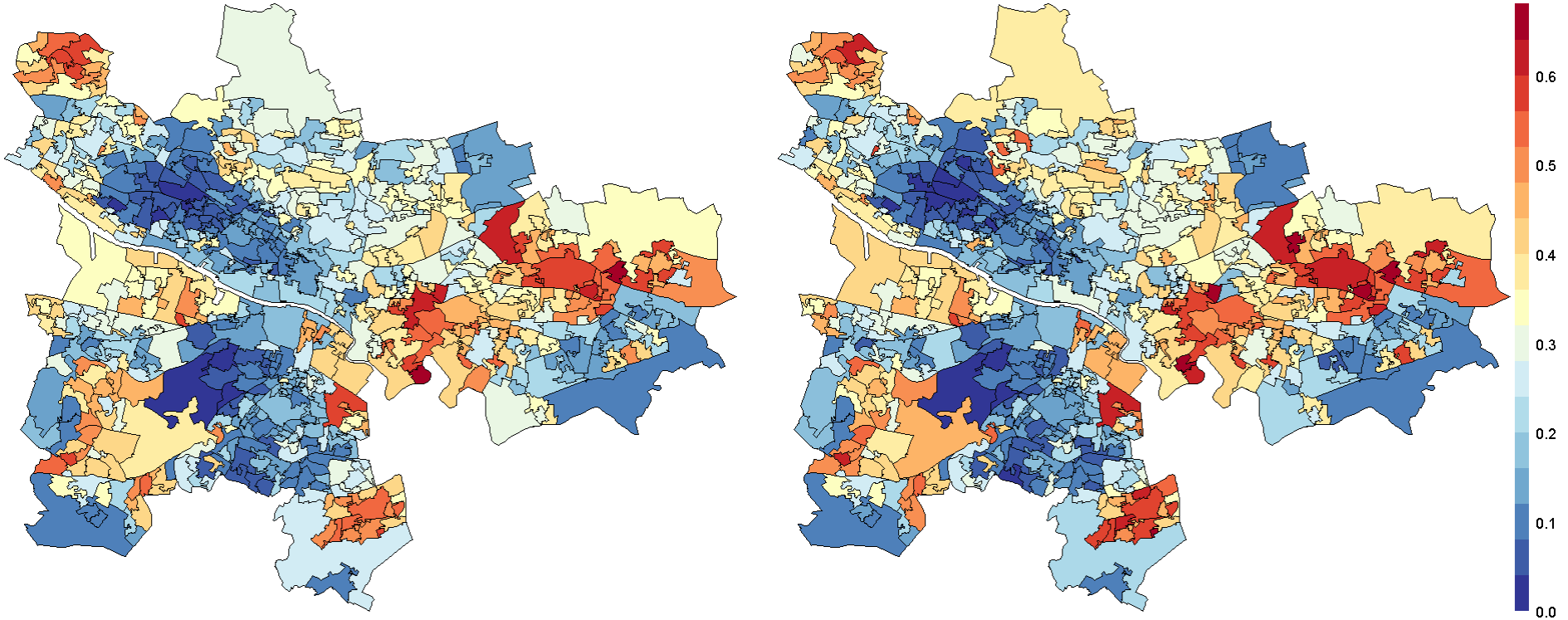
Observed vs. synthetic data

Life satisfaction by sex



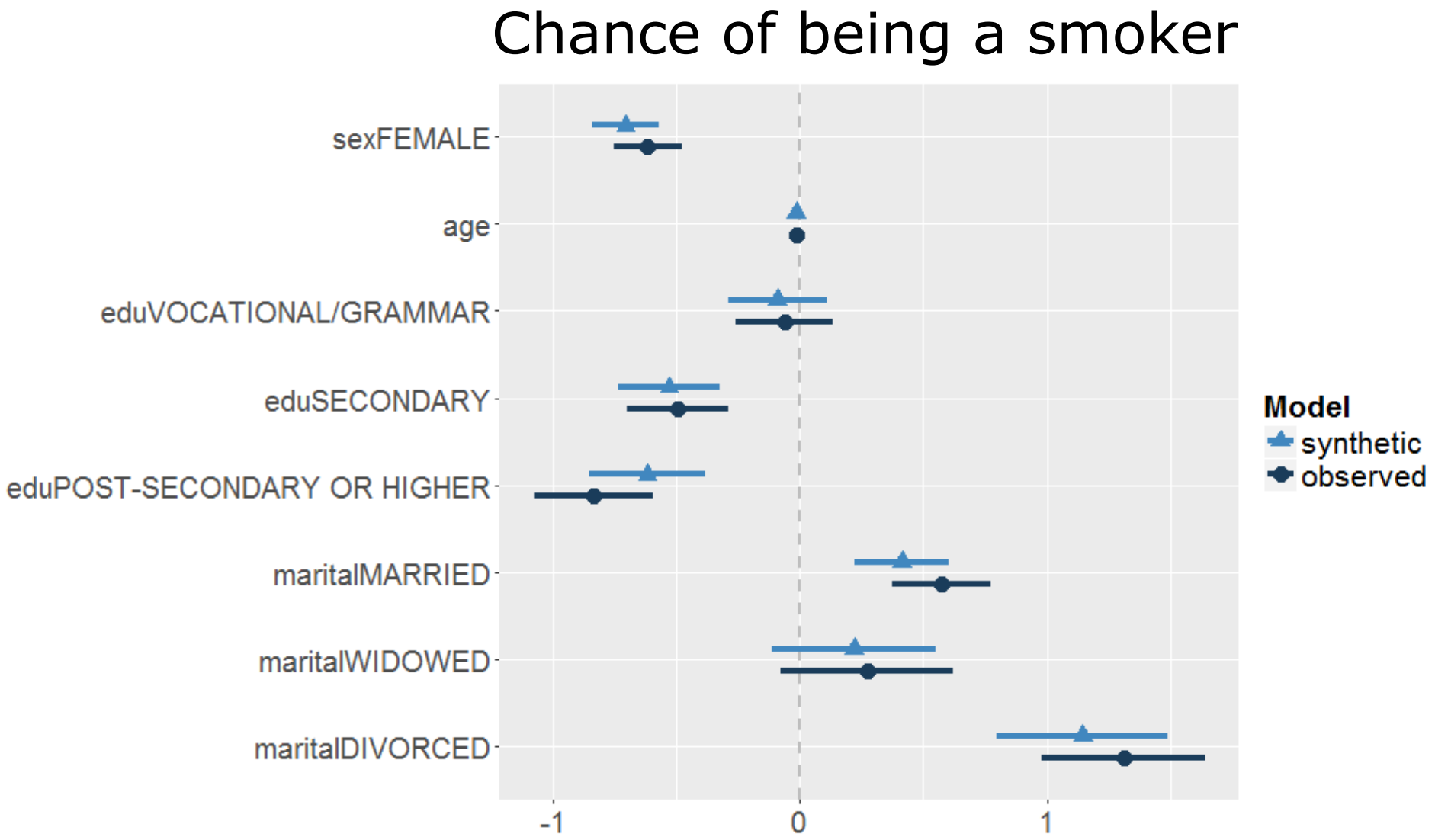
Observed vs. synthetic data

Smoking during pregnancy in Glasgow



real data **synthetic data**

Observed vs. synthetic data



sdc(): statistical disclosure control

- Data labelling: `label`
- Removing replicated uniques: `rm.replicated.uniques`
- Bottom- and top-coding: `recode.vars`, `bottom.top.coding`, `recode.exclude`
- At synthesis stage: `smoothing`, `minbucket`

Final thoughts - challenges

- Real data sets are complicated and large
- Persuading administrative data holders to allow the release of synthetic data
  - Hard to explain the process
  - Does not correspond to the usual methods (e.g. data swapping or top-coding) that are used by most data holders at present
  - Formal disclosure control measures are not available
  - Perceived risk considered as important as actual
- ADRC-S public panel was very positive