SYLLS

SYNTHETIC DATA ESTIMATION FOR UK LONGITUDINAL STUDIES

Scottish Longitudinal Study 2011 Census Linkage Launch Event

SYnthetic data estimation for the UK LongitudinaL Studies

Beata Nowok, Gillian Raab & Chris Dibben

Administrative Data Research Centre - Scotland
University of Edinburgh

Queen's University Belfast
UCL
The University of Edinburgh
Northern Ireland Statistics & Research Agency
Office for National Statistics
National Records of Scotland
E·S·R·C Economic & Social Research Council
CALLS-HUB Census & Administrative data Longitudinal Studies Hub

What are synthetic data?

- Data that look (structurally) and behave (statistically) like original confidential data but contain artificial units only

Why synthetic data?

- Facilitate access to sensitive microdata sets while protecting confidentiality

The UK Longitudinal Studies (LSs)

- Sensitive microdata:

  Sample from the Census linked to administrative data (births, deaths, marriages, health and other)

- Restricted access:
  - Safe settings
    - ONS LS (England & Wales): London, Titchfield and Newport
    - SLS (Scotland): Edinburgh
    - NILS (Northern Ireland): Belfast
  - Remote access
    - Only variable names and labels are provided to the researcher in order to build syntax
    - A Support Officer run syntax on real data set

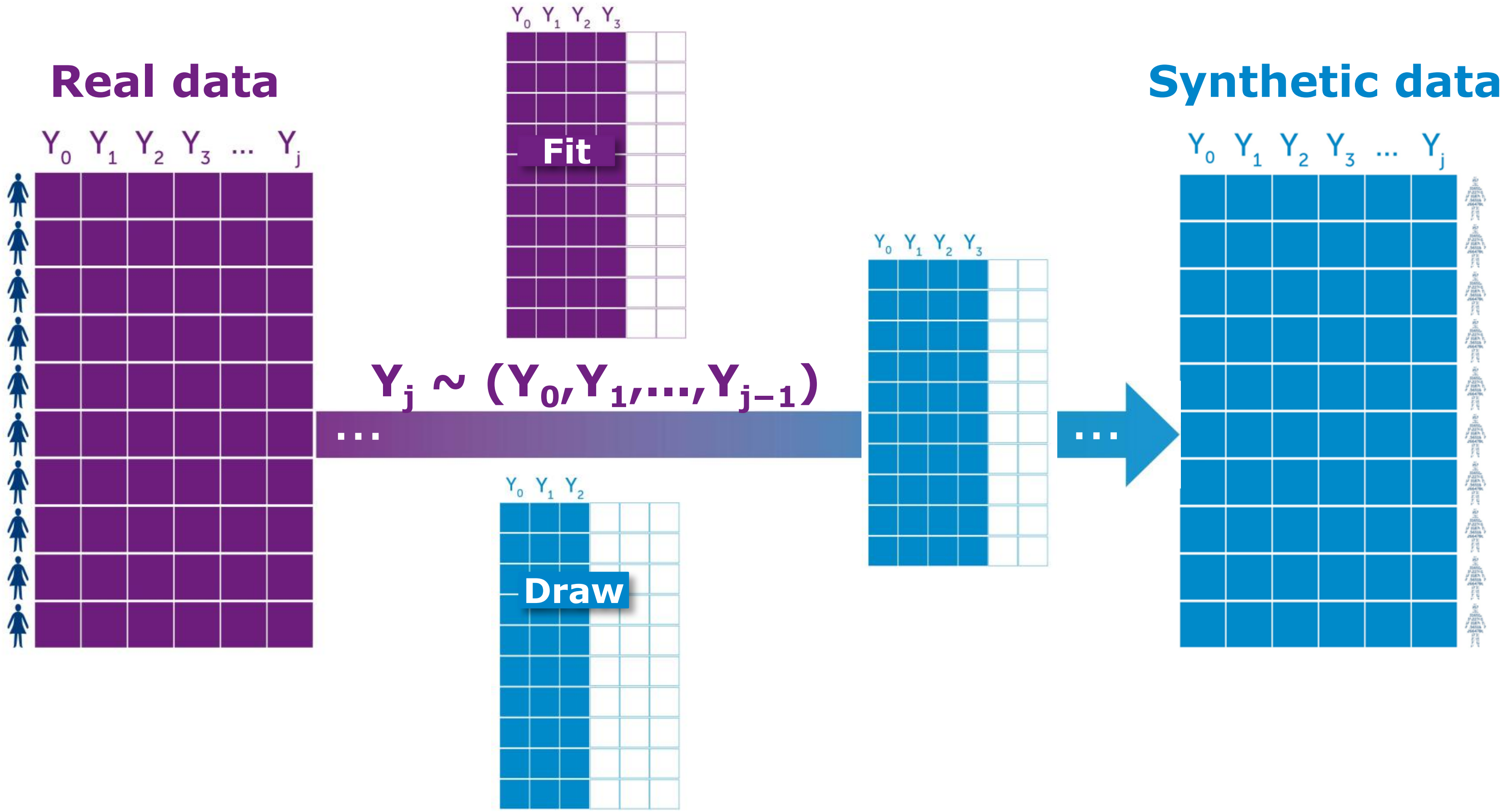
Synthetic data for the UK LSs

- Synthetic UK LS data spine (1991 & 2001)
  - Age, sex, marital status, ethnicity, limiting long term illness and geography
  - Open access via CALLS Hub and LS RSUs
- Bespoke synthetic data sets
  - Synthetic versions of data extracts to match individual user data requests
  - Provided to approved researchers for preliminary analysis, final analysis will be run on the real data in safe settings

Generating bespoke synthetic data

SYLLS
SYnthetic data estimation for UK LongitudinaL Studies

Sequentially replacing original data values with synthetic values generated from conditional probability distributions



Real data

Sex	Age	Education	Marital status	Income	Life satisfaction
WOMAN	57	VOCATIONAL/GRAMMAR	MARRIED	800	PLEASED
MAN	20	VOCATIONAL/GRAMMAR	UNMARRIED	350	MOSTLY SATISFIED
WOMAN	18	VOCATIONAL/GRAMMAR	UNMARRIED	NA	PLEASED
WOMAN	78	PRIMARY/NO EDUCATION	WIDOWED	900	MIXED
WOMAN	54	VOCATIONAL/GRAMMAR	MARRIED	1500	MOSTLY SATISFIED
MAN	20	SECONDARY	UNMARRIED	-8	PLEASED
WOMAN	39	SECONDARY	MARRIED	2000	MOSTLY SATISFIED
MAN	39	SECONDARY	MARRIED	1197	MIXED
WOMAN	38	VOCATIONAL/GRAMMAR	MARRIED	NA	MOSTLY DISSATISFIED
WOMAN	73	VOCATIONAL/GRAMMAR	WIDOWED	1700	PLEASED
WOMAN	54	SECONDARY	WIDOWED	2000	MOSTLY SATISFIED
MAN	30	VOCATIONAL/GRAMMAR	UNMARRIED	900	MOSTLY SATISFIED
MAN	68	SECONDARY	MARRIED	-8	DELIGHTED
MAN	61	PRIMARY/NO EDUCATION	MARRIED	-8	MIXED

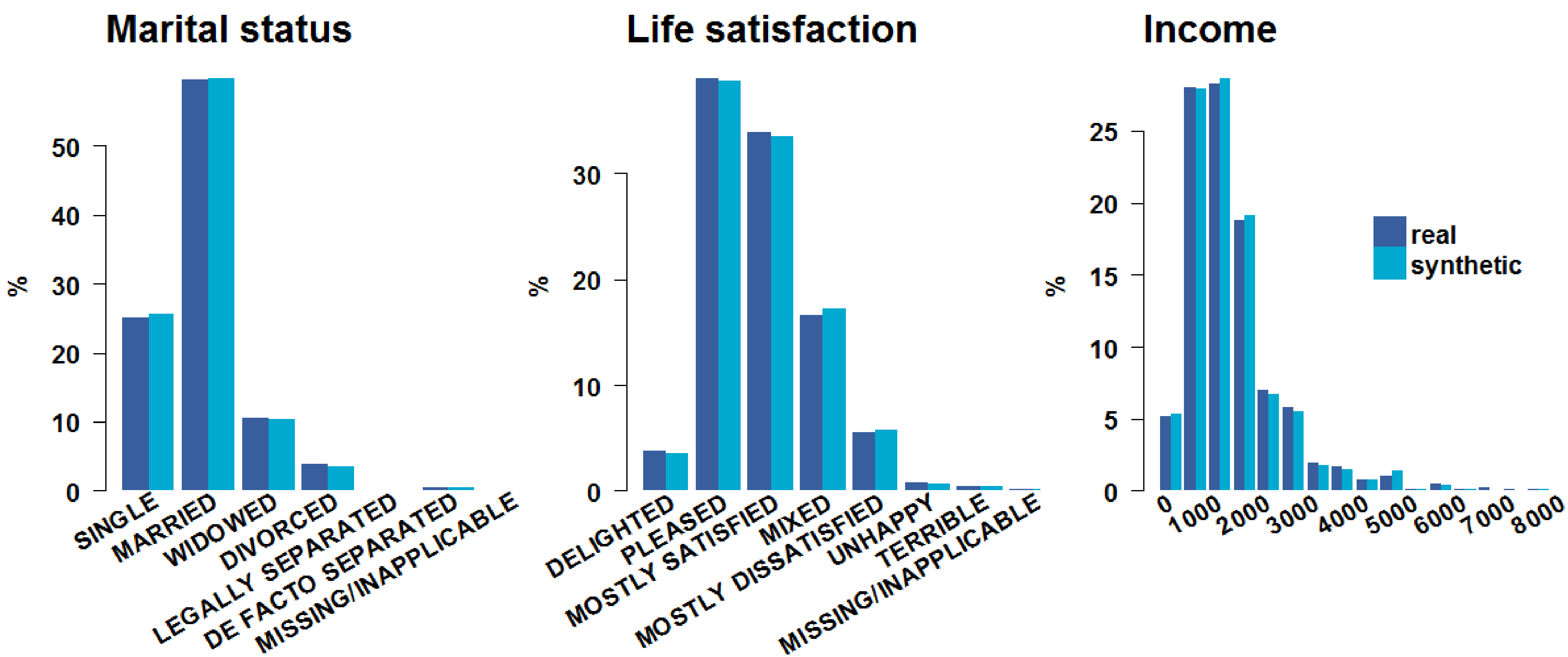
Real data

Sex	Age	Education	Marital status	Income	Life satisfaction
WOMAN	57	VOCATIONAL/GRAMMAR	MARRIED	800	PLEASED
MAN	20	VOCATIONAL/GRAMMAR	UNMARRIED	350	MOSTLY SATISFIED
WOMAN	18	VOCATIONAL/GRAMMAR	UNMARRIED	NA	PLEASED
WOMAN	78	PRIMARY/NO EDUCATION	WIDOWED	900	MIXED
WOMAN	54	VOCATIONAL/GRAMMAR	MARRIED	1500	MOSTLY SATISFIED
MAN	20	SECONDARY	UNMARRIED	-8	PLEASED
WOMAN	39	SECON			
MAN	39	SECON			
WOMAN	38	VOCATIONAL/GRAN			
WOMAN	73	VOCATIONAL/GRAN			
WOMAN	54	SECON			
MAN	30	VOCATIONAL/GRAN			
MAN	68	SECON			
MAN	61	PRIMARY/NO EDUC			

Synthetic data

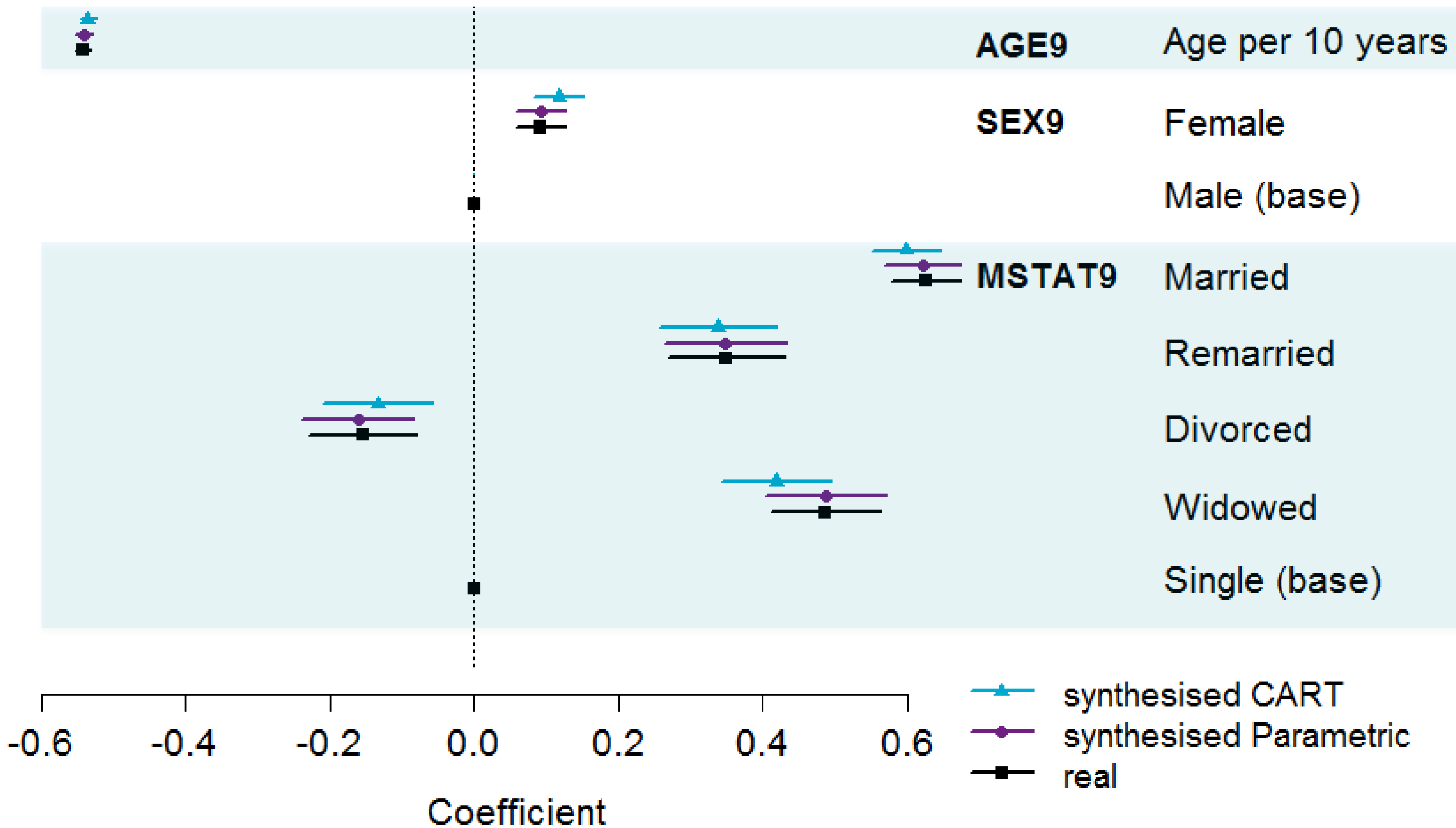
	Sex	Age	Education	Marital status	Income	Life satisfaction
false data	MAN	81	PRIMARY/NO EDUCATION	MARRIED	1500	PLEASED
false data	MAN	54	VOCATIONAL/GRAMMAR	MARRIED	1700	PLEASED
false data	WOMAN	32	VOCATIONAL/GRAMMAR	DIVORCED	870	MIXED
false data	WOMAN	61	PRIMARY/NO EDUCATION	MARRIED	800	MOSTLY DISSATISFIED
false data	WOMAN	50	PRIMARY/NO EDUCATION	MARRIED	NA	MOSTLY SATISFIED
false data	WOMAN	37	VOCATIONAL/GRAMMAR	MARRIED	158	PLEASED
false data	MAN	28	VOCATIONAL/GRAMMAR	NA	1500	MOSTLY SATISFIED
false data	WOMAN	62	PRIMARY/NO EDUCATION	MARRIED	830	MOSTLY SATISFIED
false data	MAN	78	PRIMARY/NO EDUCATION	MARRIED	NA	PLEASED
false data	WOMAN	29	SECONDARY	MARRIED	580	MOSTLY SATISFIED
false data	MAN	59	PRIMARY/NO EDUCATION	MARRIED	1300	MOSTLY SATISFIED
false data	MAN	41	SECONDARY	UNMARRIED	1500	MIXED
false data	MAN	58	SECONDARY	MARRIED	-8	PLEASED
false data	WOMAN	73	PRIMARY/NO EDUCATION	WIDOWED	1350	MOSTLY SATISFIED

Real vs synthetic data



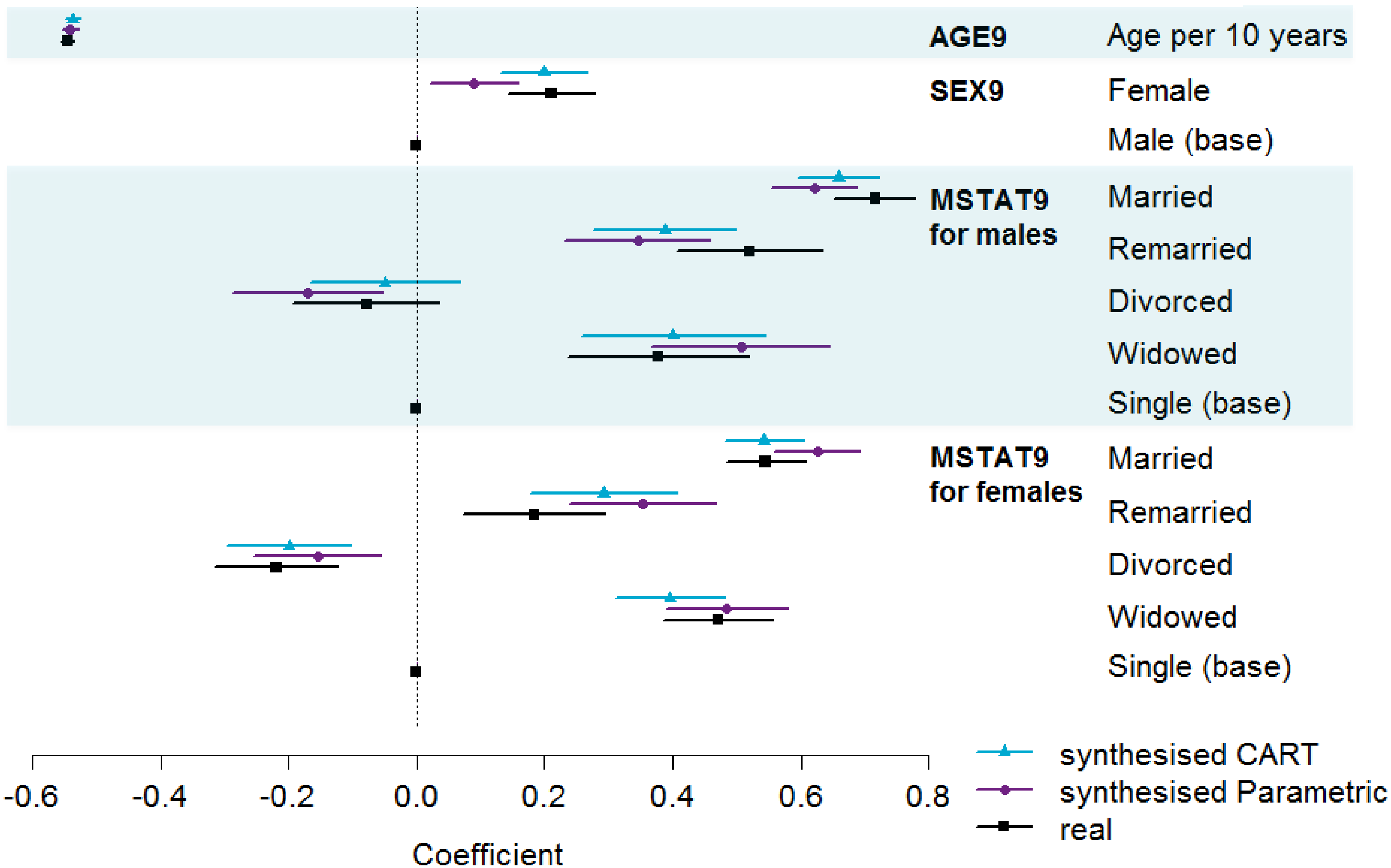
Real vs synthetic data

Logistic regression to absence of long-term illness in 1991 (ILL9), SLS



Real vs synthetic data

Logistic regression to absence of long-term illness in 1991 (ILL9), SLS



Generating synthetic versions of sensitive microdata for statistical disclosure control

R package

synthpop

http://cran.r-project.org/package=synthpop

CRAN - Package synthpop - Windows Internet Explorer

? http://cran.r-project.org/package=synthpop

CRAN - Package synthpop

synthpop: Generating synthetic versions of sensitive microdata for statistical disclosure control

A tool for producing synthetic versions of microdata containing confidential information so that they are safe to be released to users for exploratory analysis. The key objective of generating synthetic data is to replace sensitive original values with synthetic ones causing minimal distortion of the statistical information contained in the data set. Variables, which can be categorical or continuous, are synthesised one-by-one using sequential modelling. Replacements are generated by drawing from conditional distributions fitted to the original data using parametric or classification and regression trees models. Data are synthesised via the function syn() which can be largely automated, if default settings are used, or with methods defined by the user. Optional parameters can be used to influence the disclosure risk and the analytical quality of the synthesised data.

Version:	1.0-0
Depends:	lattice, MASS, methods, nnet
Imports:	rpart, party
Published:	2014-08-18
Author:	Beata Nowok, Gillian M Raab and Chris Dibben (first two authors in alphabetical order)
Maintainer:	Beata Nowok <beata.nowok at gmail.com>
License:	GPL-2 | GPL-3
NeedsCompilation:	no
CRAN checks:	synthpop results

Downloads:

Reference manual:	synthpop.pdf
Vignettes:	Using synthpop
Package source:	synthpop_1.0-0.tar.gz
Windows binaries:	r-devel: synthpop_1.0-0.zip, r-release: synthpop_1.0-0.zip, r-oldrel: synthpop_1.0-0.zip
OS X Snow Leopard binaries:	r-release: synthpop_1.0-0.tgz, r-oldrel: synthpop_1.0-0.tgz
OS X Mavericks binaries:	r-release: synthpop_1.0-0.tgz

R package synthpop 1.0-0

- Synthesis can be run with default parameters using command

  ```
  syn(mydata)
  ```

- Methods to summarise and to make inferences from synthetic data are included

Main message

- Access to LS-like data on own computer:
  - Following formal approval bespoke synthetic data should be available for SLS users in 2015
  - Spine datasets available soon via CALLS Hub and LS RSUs website