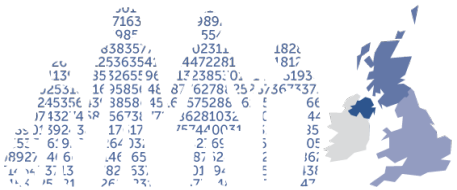


SYLLS

SYnthetic data estimation for uk LongitudinaL Studies

SYnthetic data estimation for the UK LongitudinaL Studies - SYLLS

Dr Adam Dennett

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

SYLLS

SYnthetic data estimation for uk LongitudinaL Studies

Outline

- What is SYLLS and why is it important?
- Our approach
  - The National Synthetic Data Spine
  - Individual Bespoke Synthetic Data

SYLLS

SYnthetic data estimation for uk LongitudinaL Studies

What is SYLLS?

- A project to generate synthetic versions of the national LSs which look and behave like the real thing, but are not subject to the same access restrictions

SYLLS
SYnthetic data estimation for uk LongitudinaL Studies

Why do we need synthetic data?

- ONS LS, Scottish LS and Northern Ireland LS are an unparalleled resource for social science research in the UK
- But compared to other Census data products, we have a very small user base

Census Data product	Unique users 2013
ONS LS	62 (open projects) 46 (active)
Flow data*	616
Aggregate data (Casweb)*	5781
Boundary data*	2873
*data from Q4 2012, Q3 2013 Census Support Service	

SYLLS

SYnthetic data estimation for uk LongitudinaL Studies

Why such a small user base?

Route to accessing flow data	Route to accessing ONS LS data
1. Formulate research question	1. Formulate research question
2. Turn on computer	2. Turn on computer
3. Go to cider.census.ac.uk	3. Go to ucl.ac.uk/celsius
4. Log on to WICID (now open access)	3. Download customer request form, data access agreement and approved researcher form
5. Choose your data	4. Fill out forms and submit for approval
6. Download to your own computer and analyse with preferred software	5. Wait for approval from LS research board
7. Repeat as necessary	6. Attend safe researcher certification course
	7. Ask research support officer to build your dataset from LS database
	8. Hop on train to London, Newport or Titchfield to attend VML
	9. Carry out analysis on VML terminal with old, slow software
	10. Ask for intermediate outputs to be cleared
	11. Seek final output clearance from LS research board
	12. Repeat as necessary

SYLLS

SYnthetic data estimation for uk LongitudinaL Studies

Why such a small user base?

- Complex data (compared to other cross-sectional census data products)
- Lack of exposure early in academic careers

SYLLS

SYnthetic data estimation for uk LongitudinaL Studies

Why do we need synthetic data?

- Access LS-like data on own computer
  - Iteratively refine research ideas, update analysis code etc.
- Use data in teaching and expose social science students to longitudinal data early in their research careers
- A UK longitudinal study dataset
- Methodological innovation for UK Census microdata – beyond 2011 agenda

SYLLS

SYnthetic data estimation for uk LongitudinaL Studies

Our approach

- Two project streams:
  - **National Synthetic LS Data Spine**
    - Adam Dennett, Belinda Wu, Nicola Shelton, Mike Batty and Rachel Stuchbury (UCL)
  - **Bespoke Synthetic Datasets**
    - Chris Dibben, Gillian Raab and Beata Nowok (Edinburgh)
- Ian Shuttleworth and Tony Gallagher also project partners (Queen's Belfast)

SYLLS

SYnthetic data estimation for uk LongitudinaL Studies

National Synthetic Data Spine

- Aims:
  - To create a core 'spine' dataset which:
    - Contains the same number of individuals (E&W 500,000 people + Scot 274,000 people + NI 500,000 people) as are in the LSs across 1991 > 2001 censuses
    - Has variable distributions which match those in the LS data for Age, Sex, Ethnicity, Limiting Long Term Illness, Marital Status, Births and Deaths
    - And has accurate spatial distributions of these individuals and their characteristics at the 1991 county district level

SYLLS

SYnthetic data estimation for uk LongitudinaL Studies

National Synthetic Data Spine

- Our method: Spatial Microsimulation

1. Take sample population from (publicly available) 1991 Individual SAR
2. Update values for SAR individuals according to LS distributions at county district level
3. Using transition probabilities from 1991 to 2001 (taken from LS data), age 1991 individuals on to 2001.
4. Finish with a full set of microdata records for all individuals in UK, with accurate transitions between 1991 and 2001 and accurate spatial distributions for 8 variables

SYLLS

SYnthetic data estimation for uk LongitudinaL Studies

Spine Data

1991

UID	AREAP	ETHGROUP	LTILL	MSTATUS	SEX	agegroup	age	death	birth
1	1	5	2	1	2	0	0	0	-9
2	1	5	2	1	2	0	0	0	-9
30	1	5	2	1	2	0	0	0	-9
31	1	5	2	1	2	0	0	0	-9
32	1	1	2	1					
33	1	1	2	1					
34	1	1	2	1					
1871	1	1	1	4					
1872	1	1	1	4					
1873	1	1	1	4					
1874	1	1	1	4					
1875	2	6	2	1					
1876	2	6	2	1					
1877	2	6	2	1					

2001

UID	AREAP	ETHGROUP	LTILL	MSTATUS	SEX	agegroup	age	death	birth
1	1	5	2	1	2	1	0	0	-9
2	1	5	2	1	2	1	0	0	-9
30	1	5	2	1	2	1	0	0	-9
31	1	5	2	1	2	1	0	0	-9
32	1	1	2	1	1	2	3	1	-9
33	1	1	2	1	1	2	3	1	-9
34	1	1	2	1	1	2	3	1	-9
1871	1	1	1	4	2	10	92	0	-9
1872	1	1	1	4	2	10	92	0	-9
1873	1	1	1	4	2	10	92	0	-9
1874	1	1	1	4	2	10	92	0	-9
1875	2	6	2	1	1	1	0	0	-9
1876	2	6	2	1	1	1	0	0	-9
1877	2	6	2	1	1	1	0	0	-9

SYLLS

SYnthetic data estimation for uk LongitudinaL Studies

National Synthetic Data Spine

- National Synthetic Spine almost complete:
  - Bespoke Spatial Microsimulation Software finished
  - E&W 1991-2001 data complete
  - Scotland 1991-2001 almost complete
  - NI in progress

SYLLS

SYnthetic data estimation for uk LongitudinaL Studies

Synthetic Spine Release Plans

- Currently in conversation with ONS, NRS and NISRA, but plans are for:
  - Open Access
  - Available through CALLS Hub and national research support units
- Completed software means potential for 2011 linkage in the future

SYLLS

SYnthetic data estimation for uk LongitudinaL Studies

Bespoke Synthetic Datasets

- Aims:
  - To develop a methodology and accompanying software which will allow the swift generation of statistically representative, but completely synthetic, versions of data requests submitted to the national LS Research Support Units
  - To make some bespoke synthetic datasets available for teaching, subject to disclosure control.

SYLLS

SYnthetic data estimation for uk LongitudinaL Studies

Bespoke Synthetic Datasets

- Our method: Conditional Simulation Models

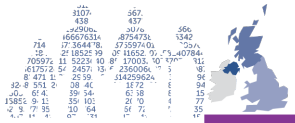
1. Take a data extract from one of the national LS datasets
2. Sequentially generate synthetic data from fitted conditional models
3. Final result is a completely synthetic representation of the joint distribution (if the models are true)

SYLLS

SYnthetic data estimation for uk LongitudinaL Studies

synthpop

- synthpop package developed in R
- Structure is based on the ‘mice’ multiple imputation package
- Range of parametric and non-parametric (classification and regression trees) options for data synthesis
- Allows for data rules, e.g. no married children
- Models missing data to produce missing data patterns like the real data



SYLL

SYnthetic data estima
uk LongitudinaL S

sex	age
WOMAN	57
MAN	20
WOMAN	18
WOMAN	78
WOMAN	54
MAN	20
WOMAN	39
MAN	39
WOMAN	43
WOMAN	63
WOMAN	38
WOMAN	73
WOMAN	54
MAN	30
MAN	68
MAN	61
MAN	84
WOMAN	87

	sex	age	edu	marital	incomenm	ls	wkabint	wkabintdur
false data	MAN	81	PRIMARY/NO EDUCATION	MARRIED	1500	PLEASED	NO	MISS/NA
false data	MAN	54	VOCATIONAL/GRAMMAR	MARRIED	1700	PLEASED	NO	MISS/NA
false data	WOMAN	32	VOCATIONAL/GRAMMAR	DIVORCED	870	MIXED	NO	MISS/NA
false data	WOMAN	61	PRIMARY/NO EDUCATION	MARRIED	800	MOSTLY DISSATISFIED	NO	MISS/NA
false data	WOMAN	50	PRIMARY/NO EDUCATION	MARRIED	NA	MOSTLY SATISFIED	NO	MISS/NA
false data	WOMAN	37	VOCATIONAL/GRAMMAR	MARRIED	158	PLEASED	NO	MISS/NA
false data	MAN	28	VOCATIONAL/GRAMMAR	MISS/NA	1500	MOSTLY SATISFIED	NO	MISS/NA
false data	WOMAN	62	PRIMARY/NO EDUCATION	MARRIED	830	MOSTLY SATISFIED	NO	MISS/NA
false data	MAN	78	PRIMARY/NO EDUCATION	MARRIED	1200	PLEASED	NO	MISS/NA
false data	WOMAN	29	SECONDARY	MARRIED	580	MOSTLY SATISFIED	NO	MISS/NA
false data	MAN	59	PRIMARY/NO EDUCATION	MARRIED	1300	MOSTLY SATISFIED	NO	MISS/NA
false data	MAN	41	SECONDARY	UNMARRIED	1500	MIXED	NO	MISS/NA
false data	MAN	58	SECONDARY	MARRIED	-8	PLEASED	NO	MISS/NA
false data	WOMAN	73	PRIMARY/NO EDUCATION	WIDOW(ER)	1350	MOSTLY SATISFIED	NO	MISS/NA
false data	WOMAN	70	SECONDARY	WIDOW(ER)	1313	MOSTLY SATISFIED	NO	MISS/NA
false data	MAN	54	VOCATIONAL/GRAMMAR	UNMARRIED	800	MOSTLY SATISFIED	NO	MISS/NA
false data	WOMAN	74	PRIMARY/NO EDUCATION	MARRIED	1300	MOSTLY SATISFIED	NO	MISS/NA
false data	WOMAN	48	POST-SECONDARY AND HIGHER	MARRIED	1500	MOSTLY SATISFIED	NO	MISS/NA
			PRIMARY/NO EDUCATION	MARRIED	-8	MIXED	NO	MISS/NA
			PRIMARY/NO EDUCATION	UNMARRIED	2000	PLEASED	NO	MISS/NA
			SECONDARY	DIVORCED	1400	MIXED	NO	MISS/NA

SYLLS

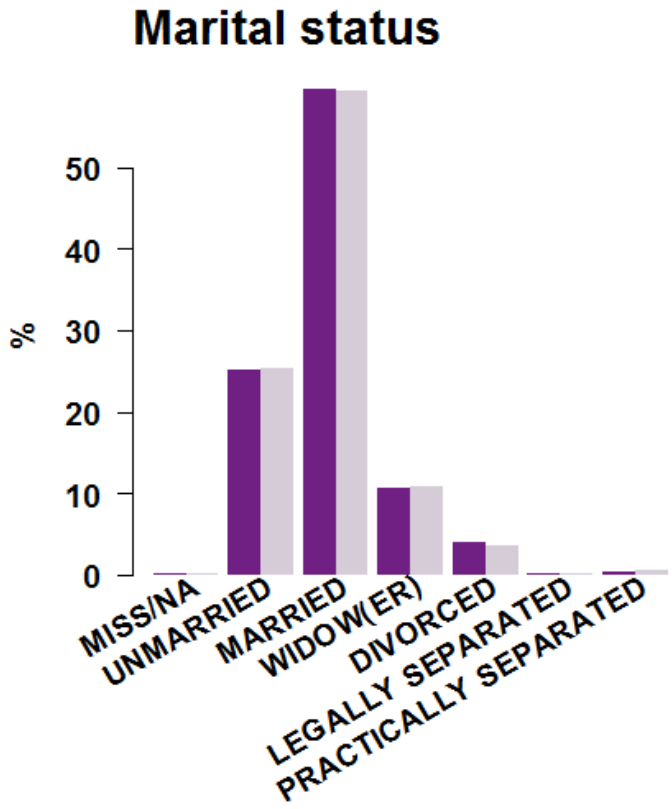
SYnthetic data estimation for uk LongitudinaL Studies

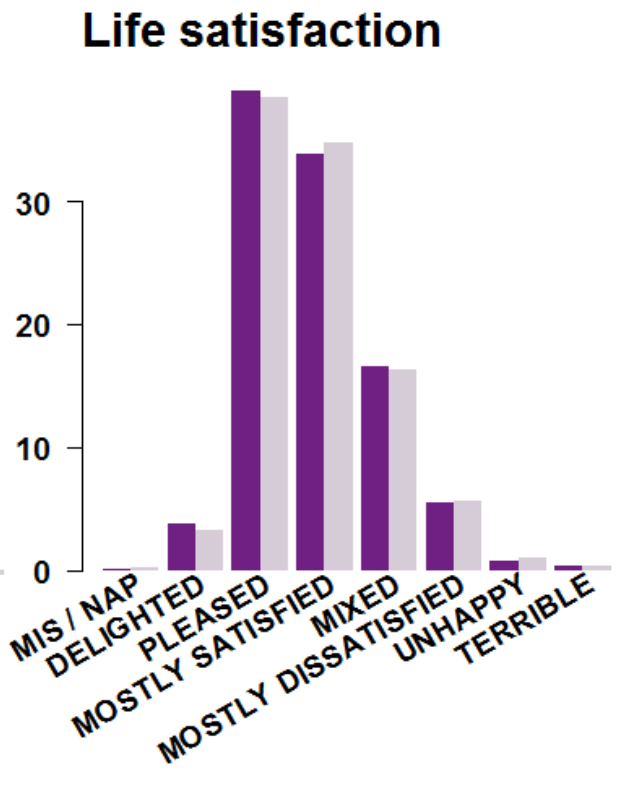
synthpop

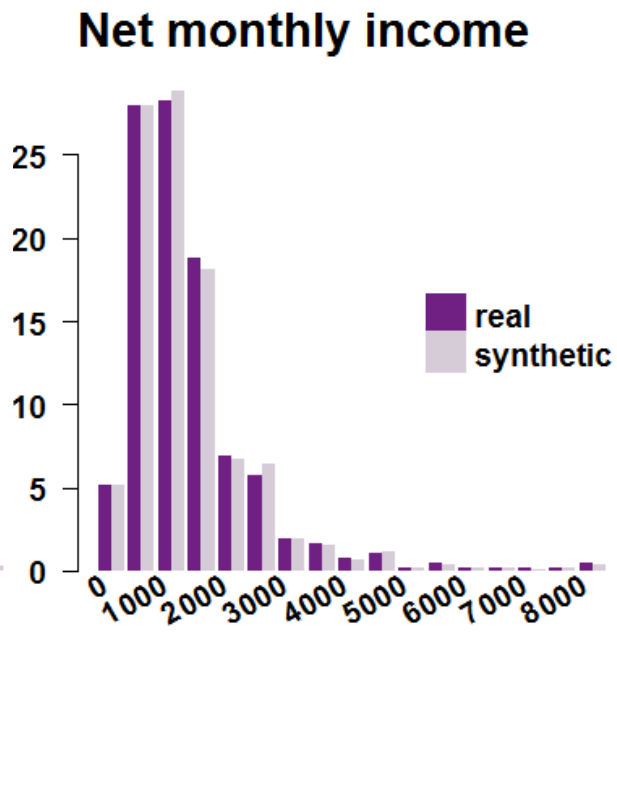
R code to synthesise: test <- syn(data)

And compare to real data: compare.synds(test, data)

Produces the plots below







SYLLS
SYnthetic data estimation for uk LongitudinaL Studies

synthpop

R code to synthesise:

```
test <- syn(data, m=10)
```

Fit to synthetic data:

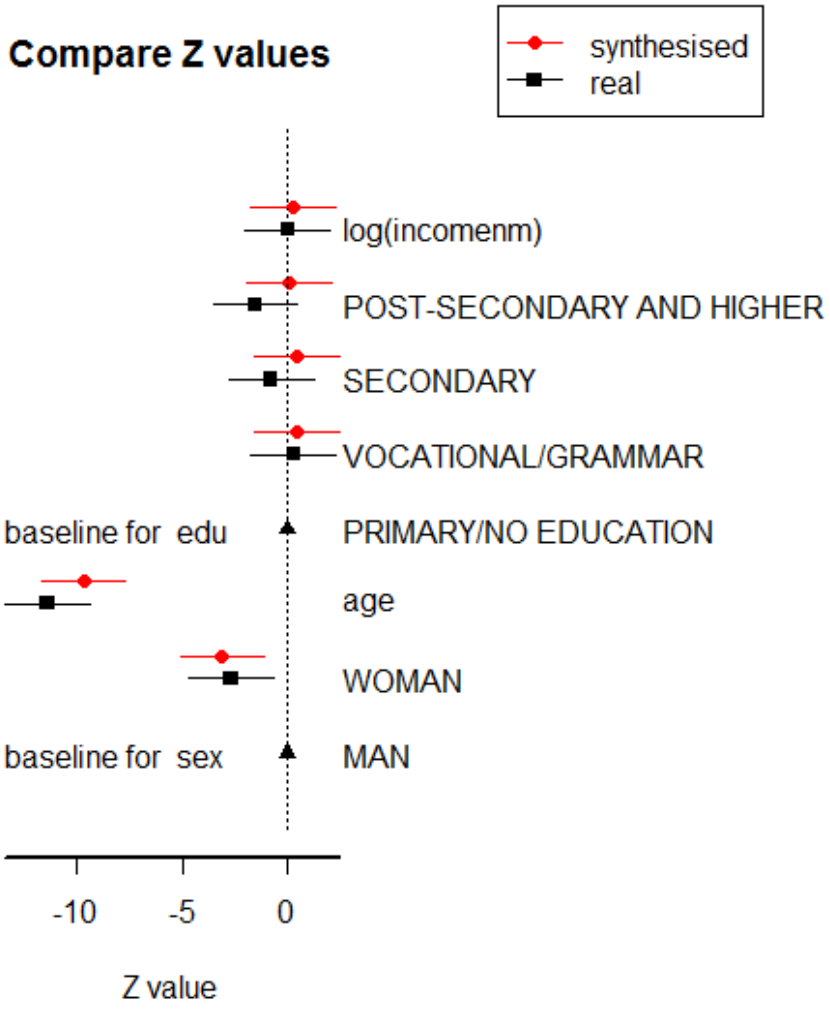
```
fit.test <- glm.synds(wkabint~ sex+age
+edu+log(incomenm),
object=test, family="binomial")
```

And compare to fit for real data:

```
compare.fit.syn(fit.test, data, plot="Z")
```

Produces plot on RHS

Young men more likely to intend to work abroad – other factors don't matter
Same conclusion from synthetic data



SYLLS

SYnthetic data estimation for uk LongitudinaL Studies

synthpop

- synthpop produces fully synthetic datasets which closely resemble the real longitudinal microdata
- Users who submit project proposals will be able to request synthetic datasets for personal research purposes

SYLLS

SYnthetic data estimation for uk LongitudinaL Studies

SYLLS

- When can we access SYLLS data?
  - Spine dataset(s) available soon via CALLS and RSUs
  - Users will shortly be able to request bespoke datasets from synthpop to accompany data requests **although a few software and disclosure control hurdles to jump first**

SYLLS

SYnthetic data estimation for uk LongitudinaL Studies

Thank you

SYLLS