THE UNIVERSITY of EDINBURGH

Facilitating access to administrative records with synthetic data

Gillian Raab
Beata Nowok
Dawn Everington
Chris Dibben

Administrative Data Research Network

An ESRC Data Investment

Authors and affiliations

Beata Nowok

Chris Dibben

Gillian Raab

Dawn Everington

School of Geosciences, University of Edinburgh

Administrative Data Research Centre Scotland

SLS-DSU

Scottish Longitudinal Study Development & Support Unit

Administrative Data

- Collected by government departments and other organisations
  - registration, transaction and record keeping,
  - delivering a service or for day-to-day operations
  - not research-ready
- Important new resource for social scientists
  - coverage, methodology
  - better understanding of our society
  - better informed government policy

Administrative data research

- 4 Administrative Data Research Centres (ADRCs)

  [secure environments, research support, original research, local data negotiations]

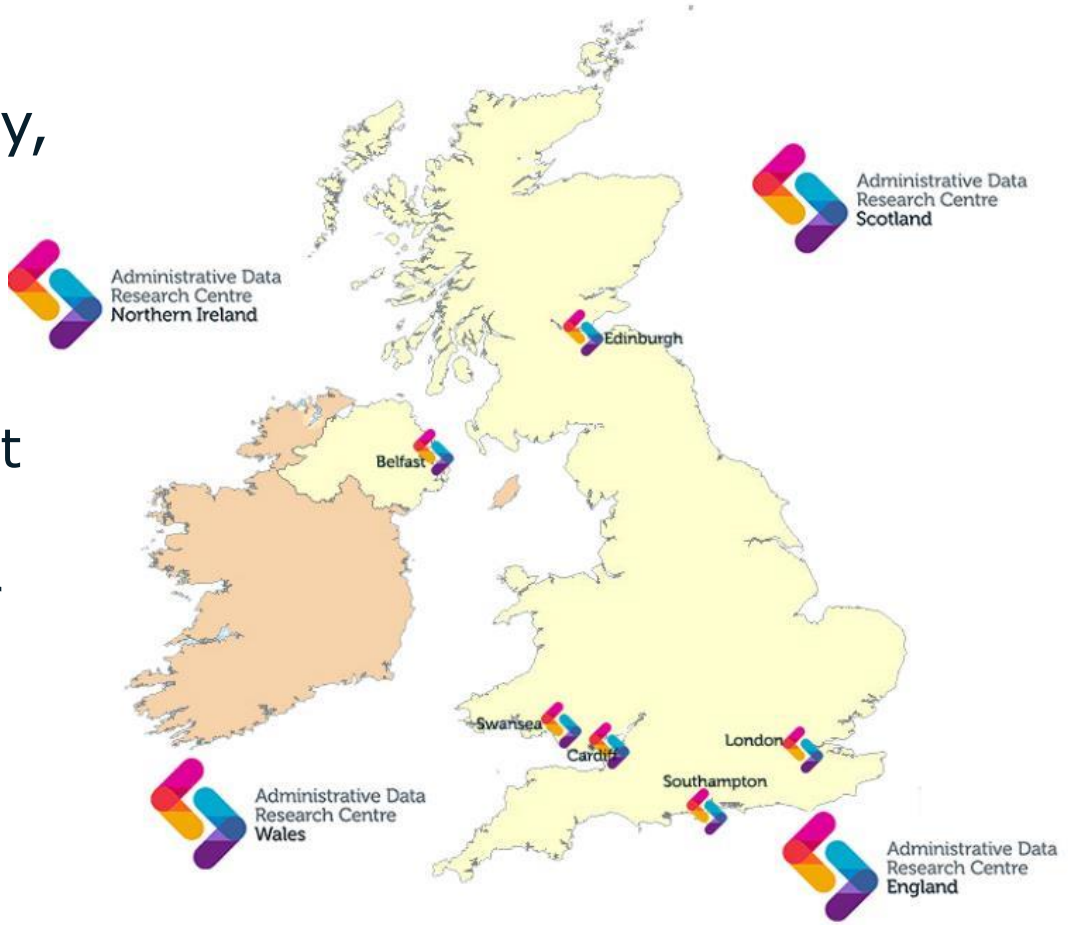
  - England – led by University of Southampton
  - Northern Ireland – led by Queens University Belfast
  - Scotland – led by University of Edinburgh
  - Wales – led by Swansea University

- Administrative Data Service (ADS) – led by UK Data Archive, University of Essex

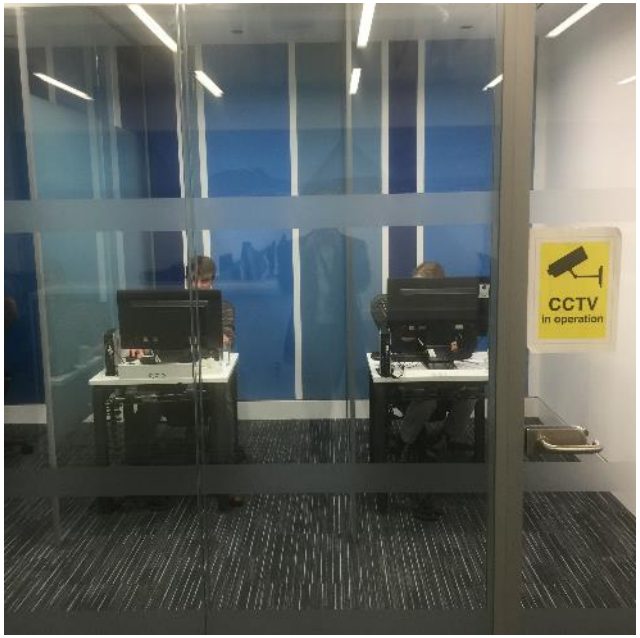
  [network coordination, first point of contact, UK data negotiations]

Safe settings

- On specific locations only, currently
  - England: London, Southampton, Titchfield
  - Northern Ireland: Belfast
  - Scotland: Edinburgh
  - Wales: Swansea, Cardiff
- For more details see adrn.ac.uk



Safe setting in the BioQuarter



Just you and the data!

Drawbacks of safe settings?

- Geography – need to travel
- Restricted work space
- No internet access
- Restrictions on what can be taken out
  - Any written material taken out must be inspected by safe-setting staff
  - Electronic output must pass disclosure rules
  - Safe-setting staff must review all such output
- The administrative data cannot be used for training courses

Synthetic data

What is it?

Data that resembles the original data
No records that correspond to real individuals or other units
But designed to make it give similar analytical results as would be found from the original data (good utility)

History

Originally proposed for disclosure control over 20 years ago
Many theoretical papers from the early 2000's
Real applications started to appear a few years later
US Bureau of the Census
Others in Canada, New Zealand, Germany

Disclosure risk

Not zero, but evaluations of applications suggest it is low.
Perceived risk may be as important as actual risk

Observed(input)

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 (output)

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

Data that look (structurally) like original data but contain artificial units only

Synthetic microdata questions

How to create it?

The distribution of the data is modelled and synthetic data generated from the models

Detailed specification of models required

Data needs to look plausible

And needs to reproduce the relationships of interest to researchers

How can it be used?

Originally proposal was to use it in place of real data.

But this is now thought to be a step too far

It can be made available to researchers

- To understand and explore the data structure
- To develop code to carry out their analyses
- Final analyses are carried out on the original data
- This closes the loop and helps to develop better methodology

To produce data for training courses

How freely can it be made available?

This is determined by the data holder

Data holders are concerned with perceived risk

Thus in most cases even synthetic data are restricted to specific researchers

SYLLS project – from 2013

- To develop tools that can be used by staff of the 3 UK longitudinal Studies with access to the original data to produce synthetic data extracts that can be made available more freely than the original data.
- Researchers can explore the synthetic data on their own computers and develop analysis code
- Teaching data sets are another use
- Originally we worked for the staff at the Scottish Longitudinal Study – and we still do
- But we now have a wider remit within ADRC-S to work with all staff making administrative data available

A software tool for producing synthetic versions of sensitive microdata

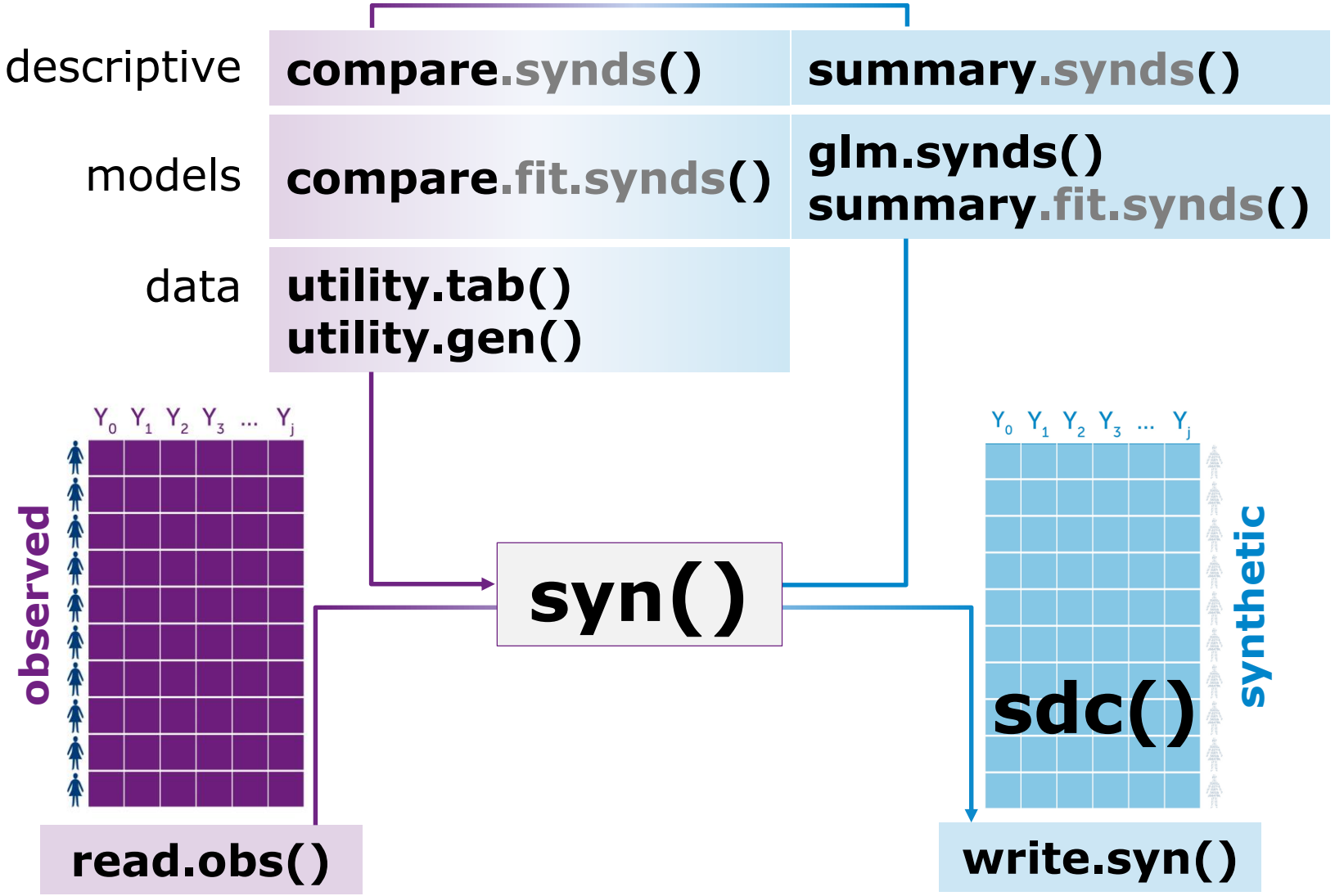
package

synthpop

version 1.4-0

http://cran.r-project.org/package=synthpop
https://github.com/bnowok/synthpop
https://www.jstatsoft.org/article/view/v074i11

Overview of **synthpop** functions



The Scottish Longitudinal Study

- One of three UK studies
  - ONS-LS (England and Wales)
  - SLS (Scotland)
  - NILS (Northern Ireland)

SLS-DSU
Scottish Longitudinal Study Development & Support Unit

- What data are held in the SLS?
  - Censuses data from 1991- 2001 2011 linked over time (5% of Scottish population)
  - Linked births, deaths, marriages. Migration records from GP registrations
  - Other administrative data sources e.g. education, health and others
- How can researchers obtain SLS data
  - Apply to do a project
  - Have it approved by the research board
  - Have safe-researcher accreditation
  - Each user gets a customised extract of linked data prepared for them to use with the variables they ask for
  - Visit the SLS safe setting to carry out analyses

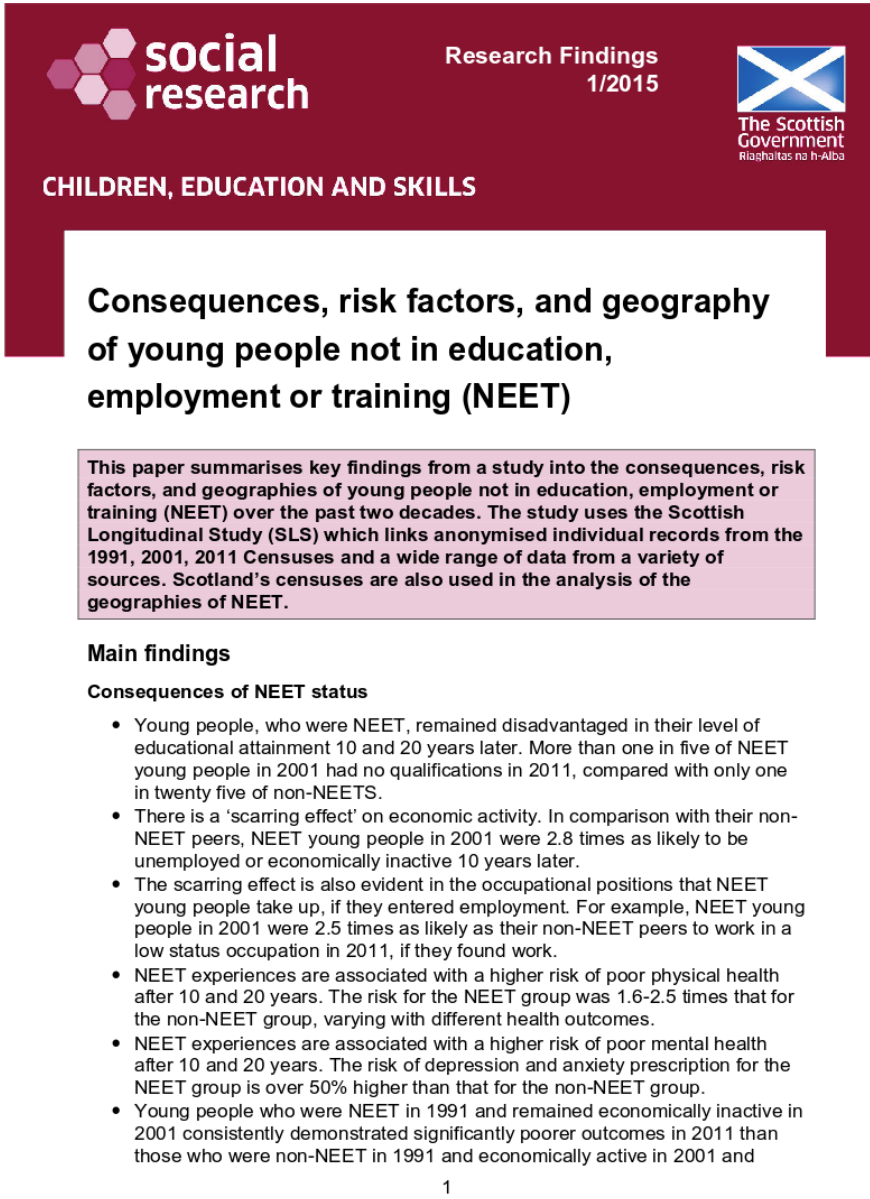
Scottish Longitudinal Study (SLS) safe setting

National Records of Scotland

SLS-DSU
Scottish Longitudinal Study Development & Support Unit

Example SLS project

- Collaborative project with Scottish Government Social Research
- Young people not in education or training (NEETS)
- ADRC-S staff are running a training course next week to teach researchers methods of handling administrative data
- Synthetic data sets have been prepared for the training course using some of the data from this project



social research

Research Findings 1/2015

The Scottish Government
Riaghaltas na h-Alba

CHILDREN, EDUCATION AND SKILLS

Consequences, risk factors, and geography of young people not in education, employment or training (NEET)

This paper summarises key findings from a study into the consequences, risk factors, and geographies of young people not in education, employment or training (NEET) over the past two decades. The study uses the Scottish Longitudinal Study (SLS) which links anonymised individual records from the 1991, 2001, 2011 Censuses and a wide range of data from a variety of sources. Scotland's censuses are also used in the analysis of the geographies of NEET.

Main findings

Consequences of NEET status

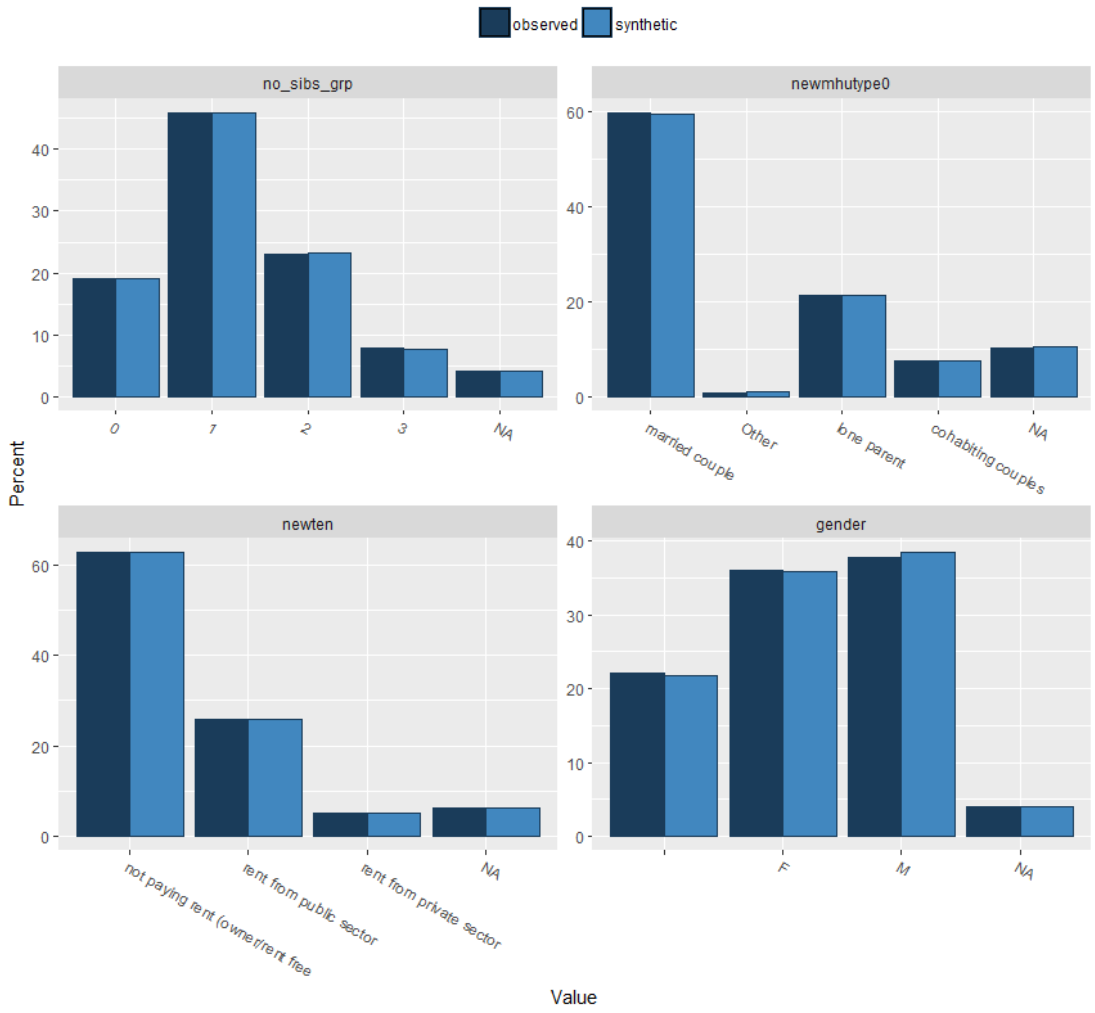
- Young people, who were NEET, remained disadvantaged in their level of educational attainment 10 and 20 years later. More than one in five of NEET young people in 2001 had no qualifications in 2011, compared with only one in twenty five of non-NEETS.
- There is a 'scarring effect' on economic activity. In comparison with their non-NEET peers, NEET young people in 2001 were 2.8 times as likely to be unemployed or economically inactive 10 years later.
- The scarring effect is also evident in the occupational positions that NEET young people take up, if they entered employment. For example, NEET young people in 2001 were 2.5 times as likely as their non-NEET peers to work in a low status occupation in 2011, if they found work.
- NEET experiences are associated with a higher risk of poor physical health after 10 and 20 years. The risk for the NEET group was 1.6-2.5 times that for the non-NEET group, varying with different health outcomes.
- NEET experiences are associated with a higher risk of poor mental health after 10 and 20 years. The risk of depression and anxiety prescription for the NEET group is over 50% higher than that for the non-NEET group.
- Young people who were NEET in 1991 and remained economically inactive in 2001 consistently demonstrated significantly poorer outcomes in 2011 than those who were non-NEET in 1991 and economically active in 2001 and

1

Synthetic school exclusions

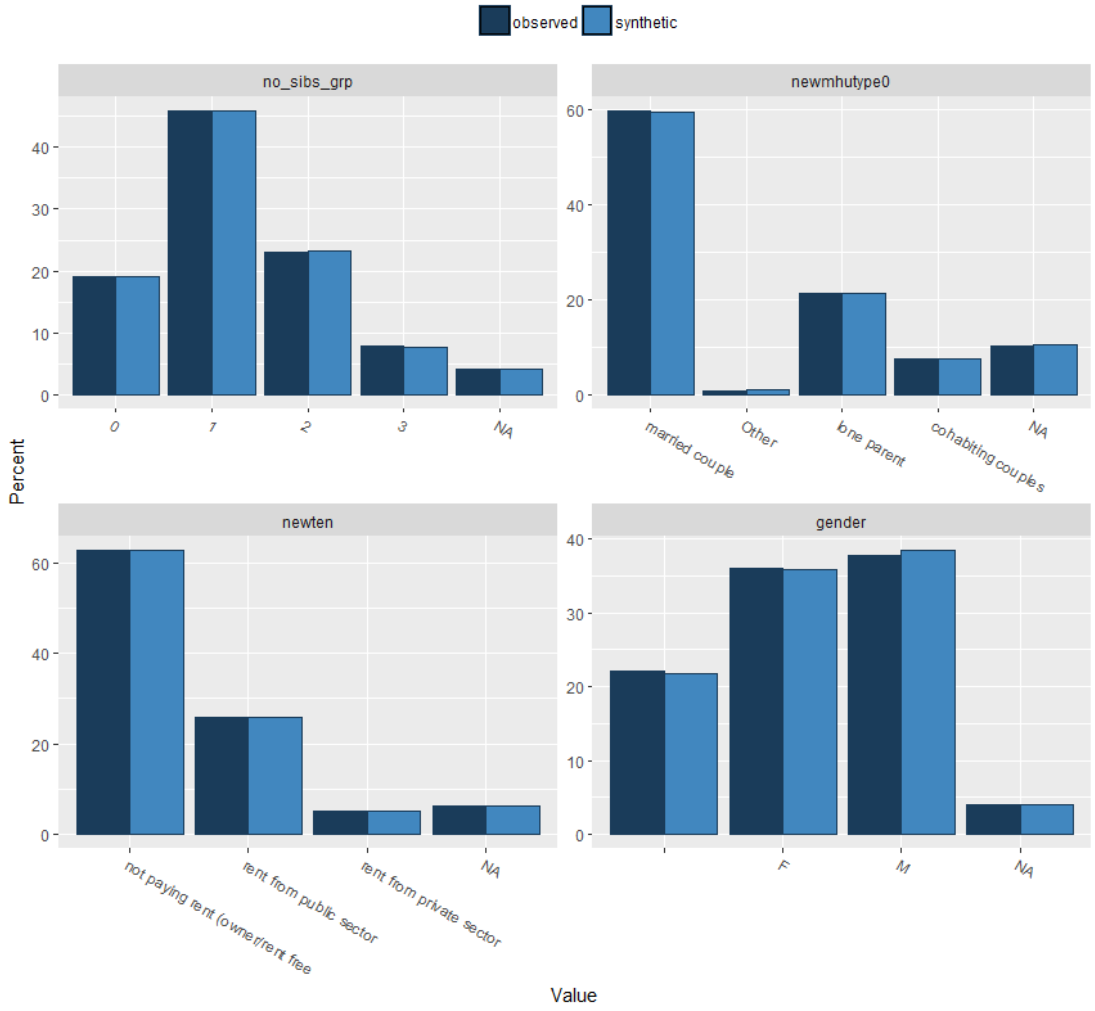
	flag	synid	startdate	finishdate	incidenttype	intaltprov	noprovdays
1	FALSE_DATA	4	10/03/2008	10/03/2008	Fighting	No provision	1
2	FALSE_DATA	7	19/01/2009	26/01/2009	Physical assault with no weapon against pupil	No provision	16
3	FALSE_DATA	7	04/04/2010	NA	Verbal abuse of staff	No provision	2
4	FALSE_DATA	7	04/04/2010	NA	Threat of physical violence, no weapon, against staff	No provision	2
5	FALSE_DATA	7	04/04/2010	NA	General or persistent disobedience	No provision	2
6	FALSE_DATA	17	20/05/2009	21/05/2009	Substance misuse â€“ alcohol	No provision	2
7	FALSE_DATA	20	07/03/2009	07/03/2009	Verbal abuse of staff	No provision	2
8	FALSE_DATA	20	16/03/2009	20/03/2009	Physical assault with no weapon against pupil	No provision	2
9	FALSE_DATA	20	27/04/2009	04/05/2009	Physical assault using improvised weapon against pupil	No provision	13
10	FALSE_DATA	20	03/09/2009	08/09/2009	Refusal to attend class	No provision	5
11	FALSE_DATA	20	03/09/2009	08/09/2009	General or persistent disobedience	No provision	5
12	FALSE_DATA	20	03/09/2009	08/09/2009	Physical assault using weapon against pupil	No provision	5
13	FALSE_DATA	22	08/05/2008	12/05/2008	Physical assault with no weapon against pupil	No provision	6
14	FALSE_DATA	36	01/11/2007	05/11/2007	General or persistent disobedience	Other	0
15	FALSE_DATA	55	28/10/2008	30/10/2008	General or persistent disobedience	No provision	4
16	FALSE_DATA	55	09/11/2008	10/11/2008	Verbal abuse of staff	No provision	2
17	FALSE_DATA	58	11/02/2010	12/02/2010	Fighting	No provision	2
18	FALSE_DATA	65	06/09/2007	11/09/2007	Threat of physical violence using weapon or improvised weapon, against pupil	No provision	6
19	FALSE_DATA	65	19/10/2008	26/10/2008	General or persistent disobedience	No provision	10
20	FALSE_DATA	65	29/01/2009	30/01/2009	Threat of physical violence, no weapon, against pupil	No provision	0

Comparing real and synthetic data



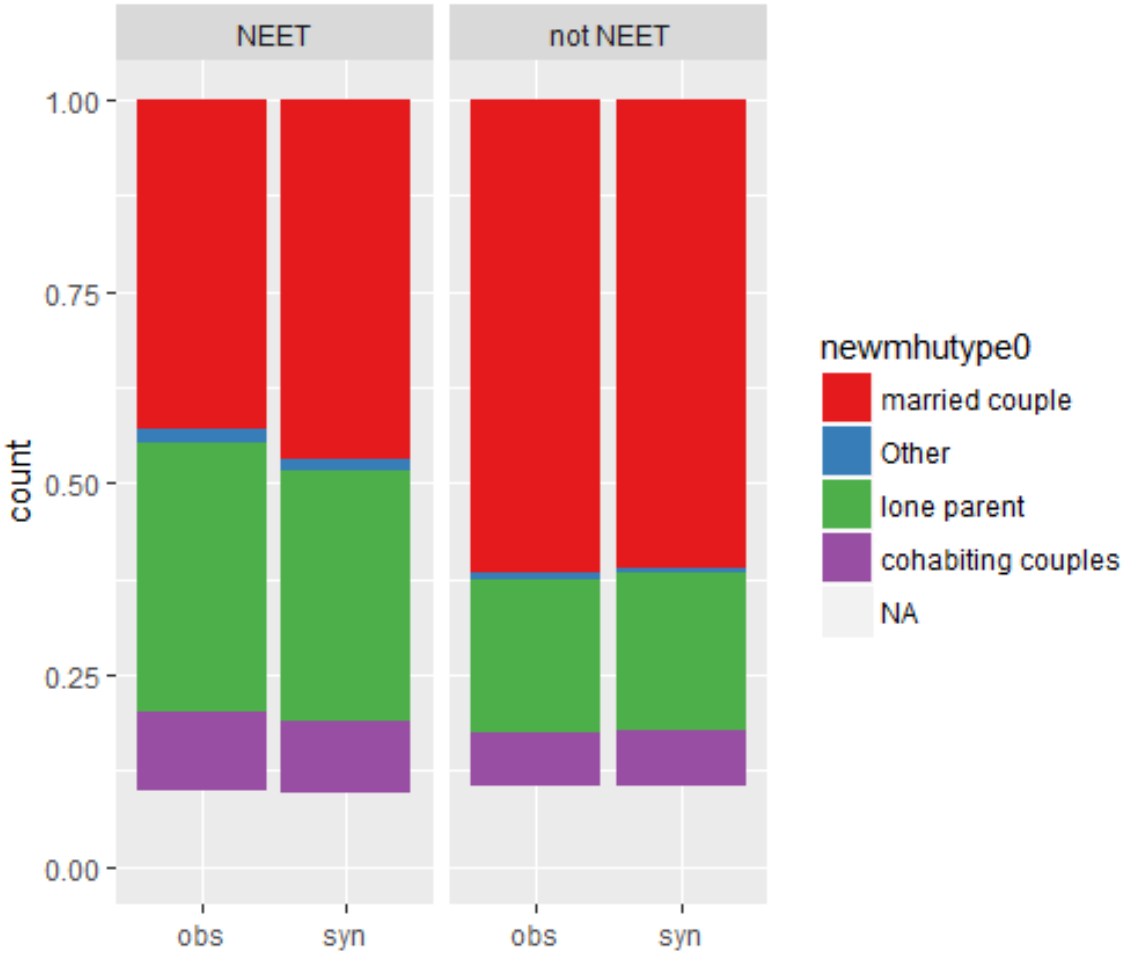
Source: Scottish Longitudinal Study

Comparing real and synthetic data



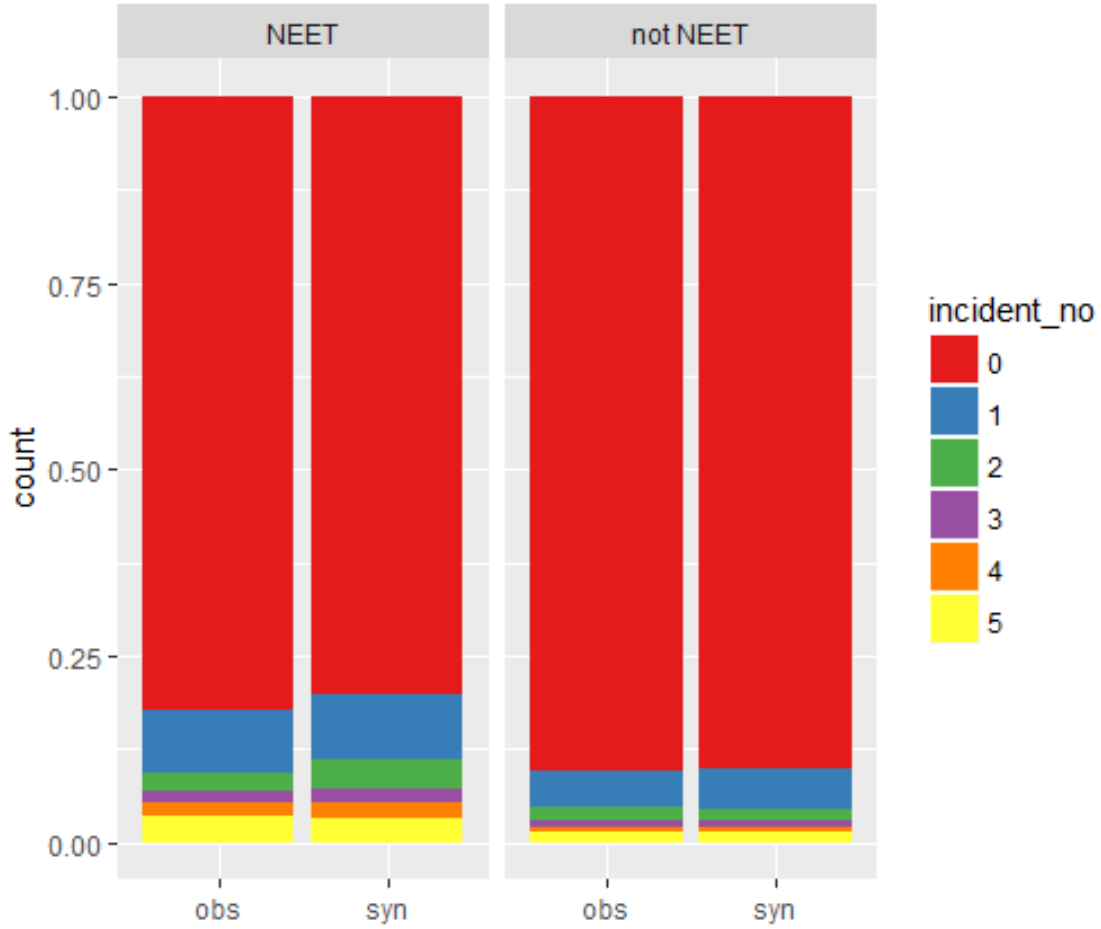
Source: Scottish Longitudinal Study

NEET 2011 by household type 2001



Source: Scottish Longitudinal Study

NEET 2011 by exclusions 2006-2010



Source: Scottish Longitudinal Study

Utility

- These examples worked OK
- Synthetic data are only as good as the models that created them.
- Creating synthetic data, even with *synthpop* is not easy
- Detailed specification is needed to provide useful and plausible data

Disclosure risk

- Perceived risk may be important After synthesis we carry disclosure control
- Including
  - Labelling the data as FALSE DATA
  - Removing any sample uniques that appear in the synthesised data
  - Top and bottom coding
- Making synthetic data available only to trained and approved researchers
- Ensuring that training data is only used for the course

Acknowledgements

The help provided by staff of the Longitudinal Studies Centre Scotland is acknowledged. The LSCS is supported by the ESRC/JISC, the Scottish Funding Council, the Scientists Office and the Scottish Government. The authors alone are responsible for the interpretation of the data. Census output is Crown copyright and is reproduced with the permission of the Controller of HMSO and the Queen's Printer for Scotland.