THE UNIVERSITY of EDINBURGH

Generating synthetic data in R

Session 1 Introduction & background

Gillian Raab
& Beata Nowok

Administrative Data Research Centre – Scotland

Administrative Data Research Network

An ESRC Data Investment

Outline

- Origins of the SYLLS project and the *synthpop* package
- A very brief review of the literature and methods
- Experiences of other providers of synthetic data
- Our experience developing *synthpop*

How we got started

- Concern that the three Longitudinal Studies (LSs) were accessed less frequently than other data resources.
- What are the LSs?
  - **ONS-LS** (England and Wales) Started in 1974. All individuals born on 4 SECRET birthdays were included in the study. Data from the 1971, 1981, 1991, 2001 and 2011 Censuses (and members of their households) linked to records of births, deaths marriages, emigration and immigration and cancer registrations.
  - **SLS** (Scotland) Started in 1994. Has 20 SECRET birthdates 1991, 2001 and 2011 census data and in addition to the same data as the LS we can link our data to NHS data (e.g. hospital admissions and birth details) and most recently education data on school attendance and exam results.
  - **NILS** (N Ireland) Started in 2010. Initially 2001 census data only, but with 104 SECRET birthdays, but now extended to include 1991 and 2011 data.

What you have to do to access the SLS

- Decide what data you want to use by using resources on the SLS web site http://www.lscs.ac.uk/
- Discuss it with a member of the support team
- Complete training to be an approved researcher
- Complete an application form detailing exactly what files and variables you want
- Get approval from the SLS research board
- When all permissions obtained the SLS staff will provide you with an EXTRACT of data – usually involving linking data from various sources
- All of your team must sign further forms when data are available
- Come to Ladywell House to analyse your data or submit code to be run by support staff

How synthetic data can help

- It contains no real individuals, but is generated from the real data
- Users can be supplied with the synthetic data to analyse on their own computers
- Hence the SYLLS project to develop methods that LS staff can use to provide synthesised versions of extracts
- And hence the *synthpop* package for R

Synthetic data can overcome bits in green

Formal applications can delay your getting started

No access to data until all of these complete

Access to the data only available by visits to the safe haven in Ladywell House in Edinburgh or by submitting code to a support officer

Lots of forms to sign

No internet access in the safe setting

All output has to be cleared for potential disclosiveness by an SLS support officer and emailed to the user after transfer via an encrypted data stick

Synthetic data - background

- First proposed in 1993
- First papers suggesting how to do it from 2003 – mainly USA, but also Germany, New Zealand and Canada
- Many more theoretical papers up to now (see links to papers on course web site for references).
- Synthetic data products began to be available from around 2010

How does it work ?

- Staff with access to the real data fit a model to it
- The synthetic data are then generated from this model and synthetic data sent to users
- Initial theory was developed for examples like multivariate Normal data
- But no real data looks like this
- Very soon the idea of synthesising from a sequence of conditional models became the way synthesis was done in practice

A very simple example

- Suppose we have a data set with
  - **age**, **sex**, and **marital status**
- Sequence of models
  - First we take a bootstrap sample of **age** to make the first column of the synthetic data **age.syn**
  - Then we fit a logistic model to predict **sex** from **age**, using the real data and make the next column of the synthetic data by predicting **sex** from **age.syn** to get **sex.syn**
  - Then we fit a model of **marital status** in terms of **age** and **sex** with the real data and make the next column of the synthetic data by predicting from **age.syn** and **sex.syn** to get **maritalstatus.syn**

Types of model

- At each step we are fitting a conditional model, given the variables synthesised so far
- The example above used a parametric model at each step in the synthesis
- Sometimes such models may not fit the data well
- The use of more flexible models such as CART has been found to be a useful alternative to use for some or all of the conditional distributions

How should synthetic data be used?

- Initial papers suggested that it could be used INSTEAD OF the real data
- Methods of getting inference to the real data can require MULTIPLE synthetic data sets
- But this (at least for now) may be a step too far
- We can never be sure that our model of the data is the correct one

US synthetic data products

- **From the US Bureau of the Census**
  - Synthetic Longitudinal Business Database (SynLBD)
  - Survey of Income and Program Participation Synthetic Beta (SSB)
- You can apply to get them on the web
- But you are strongly discouraged from publishing anything based on only synthetic data
- You develop on synthetic data and Census Bureau staff run final analyses for you
- Only a single synthetic data set is available in each case – confidentiality reasons.

UK synthetic data products

- **UK**
  - UK Longitudinal Studies - sample from the Census linked to administrative data
- Users can request bespoke synthetic data sets for preliminary analysis
- Differs from the US case in that a new synthesis is needed for each user
- Hence the *synthpop* package we hope you will learn today

A software tool for producing synthetic versions of sensitive microdata

synthpop

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

Health warnings and disclaimers

- Synthetic data are only as good as the models used to create them and should always be checked
- To be able to synthesise any of the features of real data is a big challenge. For the US synthetic products a whole team of researchers worked to produce each single synthetic data product
- We don't claim that *synthpop* can synthesise every sort of data

synthpop is not perfect

- We are doing our best but some limitations remain.
  - Coping with very large and complex data sets
  - Structured data
  - Repeated event data
- We hope to learn more from users like you and we welcome your feedback
- We hope you will find *synthpop* helpful and not have too many problems today
- Good luck!