THE UNIVERSITY *of* EDINBURGH

Generating synthetic data with the synthpop package for R

Introduction & background

Gillian Raab
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) **SLS** (Scotland) **NILS** (N Ireland)
  - Provide users with samples of Census data linked over time and to administrative data (births, deaths, marriages and other sources)
  - The data are extremely sensitive, Census data has legal protection, and a knowledge of who is in the LSs would be a major breach
  - Access requires users to visit safe settings with no internet access and other restriction
- Synthetic data has been used in other places, e.g. US Bureau of the Census
- Perhaps it could help to make the LSs more accessible

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 - 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

What is/are synthetic data?

- We will be discussing *completely* or *fully* synthetic data
- Every data item from every case is replaced by a synthesised value
- Some type of model is fitted to the real data and the synthesised values are replaced by data generated from the model
- No record in the synthesised data can be associated with a case in the real data

How does it work ?

- Some real data, even though anonymised, are too sensitive to be released to researchers
- Staff in an agency fit a model to the real data
- The synthetic data are then generated from this model and synthetic data produced that can be made more freely available
- 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 most promising approach

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 multinomial 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
- These can sometimes work well, but need to be selected carfully
- 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
- This generated many statistical papers proposing rather complicated methods of doing this some requiring multiple synthetic data sets to be released.
  - They have been very little used in practice
  - We can never be sure that our model of the data is the correct one
  - Agencies are unwilling to release more than one synthetic data set

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.

Our approach for the LSs

- So far only implemented for the SLS
- A trained and accredited user can request bespoke a synthetic data set for preliminary analysis
- They must sign agreements not to share them beyond named members of their study team
- The final analysis will be run on the real data by visiting the safe setting, or by users submitting code to be run by SLS staff
- US Census Bureau products (2 in all)
  - each produced by a whole team of analysts
- UK LSs
  - 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

R package

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.
- As *synthpop* is open source it is being used by others beyond the LSs
- Several groups we know of have used it to provide data sets to be used for teaching.

Recent developments now in *synthpop*

- Methods to assess the utility of synthetic data (session 1)
  - Comparing tables produced from real and synthetic data – chi-squared statistics
  - Calculating a general utility measure
  - Graphical tools
- Stratified synthesis (session 2)
- Synthesising groups of variables together (session 2)
  - As a complete cross-tabulation
  - To produce a data set where the margins are well-fitted

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!

Now over to Beata for How to use synthpop.

Copies of the slides and some sample code can be found at

https://www.geos.ed.ac.uk/homes/graab