Generating synthetic microdata to widen access to sensitive data sets

Beata Nowok, Gillian Raab & Chris Dibben

Administrative Data Research Centre – Scotland
Research context: ADMINISTRATIVE MICRODATA

- **SYLLS = SYnthetic data estimation for the UK Longitudinal Studies (LSs):**
  sample from the Census linked to administrative data (births, deaths, marriages, health)

- **ADRC-S = Administrative Data Research Centre - Scotland:**
  major Scottish administrative datasets (housing, transport, income, labour markets, health, crime and criminal justice, education, social services)
Research context: RESTRICTED ACCESS

• Safe setting
  • 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.

Small user base
Project aims

• Widening access to census-linked UK longitudinal studies while protecting confidentiality:
  • Devise a method of generating bespoke synthetic data extracts to match individual user data requests,
  • Bespoke data should look and behave (statistically) like real data so researchers can experiment and refine research without having to travel to safe settings.
• Make some bespoke synthetic data sets available for teaching
Synthetic data: background

• Similar initiatives in USA and Germany
• Previous work has focused on using multiple versions of synthesised data to make inferences to the population (proper synthesis)
• BUT most users will only wish to get results close to what would be found for the real data
• This needs a simpler approach with just a single synthetic sample
• It assumes users will run the final analysis on the real data
Bespoke synthetic data extract

**Original bespoke data extract**
- Requested variables

**Non-disclosive fully synthetic version**
- No one is real
- As many relationships as possible are preserved

Requested population

Final analysis in safe settings

Provided to user
Generating fully-synthetic data

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

Final result is a completely synthetic representation of the joint distribution (if the models are true)
Synthesising model choices

General choices
- Parametric
- Semi-parametric (preserving the marginal distribution)
- Non-parametric (CART)

Choice criteria
- Preserving as many relationships as possible while protecting data confidentiality
- Handling diverse data types
- Feasible for large datasets
- Easy to implement with little tuning required
CART models

• Build a tree
  $Y_j \sim (Y_0, \ldots, Y_{j-1})$

• Generate $Y_j$ by:
  • Running $Y_j, \ldots, Y_{j-1}$ down the tree
  • Sampling from the leaves
package for data synthesis

synthpop
synthpop: basic functionality

- A synthetic data set can be produced using a single command: `syn(data)`
- Can be run with default parameters according to the types of data encountered
- Or tailored for specific data sets, including options to match the structure of the real data
synthpop: basic functionality

- Optional parameters:
  - List of synthesising methods for each variable
  - Order in which variables should be synthesised
  - Detailed specification of predictors for each synthesised variable
  - Rules for dependencies between variables and structural zeros (e.g. rule age<16 sets marital status to "single")
  - Codes for missing values to be modelled (assuming MAR)
> test <- syn(data)
syn variables
1  sex age edu marital incomenm ls wkabint

> test
Call:
($call) syn(data = data)
Number of synthesised data sets:
($m) 1
First rows of synthesised data set:
($syn)
   sex age                  edu  marital incomenm                  ls wkabint
1 MAN  81 PRIMARY/NO EDUCATION  MARRIED     1500             PLEASED      NO
2 MAN  54 VOCATIONAL/GRAMMAR MARRIED     1700             PLEASED      NO
3 WOMAN 32 VOCATIONAL/GRAMMAR DIVORCED 870 MIXED NO
4 WOMAN 61 PRIMARY/NO EDUCATION MARRIED 800 MOSTLY DISSATISFIED NO
5 WOMAN 50 PRIMARY/NO EDUCATION MARRIED NA MOSTLY SATISFIED NO
6 WOMAN 37 VOCATIONAL/GRAMMAR MARRIED 158 PLEASED NO
Synthesising methods:
($method)
sex      age      edu  marital incomenm       ls  wkabint
"sample" "ctree" "ctree" "ctree" "ctree" "ctree" "ctree"
Order of synthesis:
($visitSequence)
sex      age      edu  marital incomenm       ls  wkabint
1        2        3        4        5        6        7
Matrix of predictors:
($predictorMatrix)

Administrative Data Research Centre - Scotland | Beata Nowok | 30 April 2014
R code to synthesise:  
\[
\text{test} \leftarrow \text{syn(data)}
\]
And compare to real data:  
\[
\text{compare.synds(test, data)}
\]
synthpop: example

R code to synthesise:
```r
test <- syn(data, m=10)
```

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

And compare to fit for real data:
```r
compare.fit.syn(fit.test, data, plot="Z")
```

Produces plot on RHS
Young men more likely to intend to work abroad – other factors do not matter. Same conclusion from synthetic data
synthpop in practice

• Effort required to produce realistic synthetic data can be substantial
  • Understanding the data
  • Derived variables
  • Rules for restricted values
  • Codes for missing values
  • …
Synthetic data: current status

• First version of the package now available – bugs being fixed
• More work needs to be done to overcome computing limitations and get formal permissions from LS to release such data
• Prof Mark Elliott will be carrying out and reporting a formal disclosure control evaluation of the package shortly
• LSs users should shortly be able to request bespoke synthetic data sets to accompany data requests