synthpop
an R package for generating synthetic microdata

Beata Nowok
Administrative Data Research Centre – Scotland
What is **synthpop**?

- A software tool for producing synthetic versions of sensitive microdata
<table>
<thead>
<tr>
<th>Sex</th>
<th>Age</th>
<th>Education</th>
<th>Marital status</th>
<th>Income</th>
<th>Life satisfaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>FEMALE</td>
<td>57</td>
<td>VOCATIONAL/GRAMMAR</td>
<td>MARRIED</td>
<td>800</td>
<td>PLEASED</td>
</tr>
<tr>
<td>MALE</td>
<td>41</td>
<td>SECONDARY</td>
<td>UNMARRIED</td>
<td>1500</td>
<td>MIXED</td>
</tr>
<tr>
<td>FEMALE</td>
<td>18</td>
<td>VOCATIONAL/GRAMMAR</td>
<td>UNMARRIED</td>
<td>NA</td>
<td>PLEASED</td>
</tr>
<tr>
<td>FEMALE</td>
<td>78</td>
<td>PRIMARY/NO EDUCATION</td>
<td>WIDOWED</td>
<td>900</td>
<td>MIXED</td>
</tr>
<tr>
<td>FEMALE</td>
<td>54</td>
<td>VOCATIONAL/GRAMMAR</td>
<td>MARRIED</td>
<td>1500</td>
<td>MOSTLY SATISFIED</td>
</tr>
<tr>
<td>MALE</td>
<td>20</td>
<td>SECONDARY</td>
<td>UNMARRIED</td>
<td>-8</td>
<td>PLEASED</td>
</tr>
<tr>
<td>FEMALE</td>
<td>39</td>
<td>SECONDARY</td>
<td>MARRIED</td>
<td>2000</td>
<td>MOSTLY SATISFIED</td>
</tr>
<tr>
<td>MALE</td>
<td>39</td>
<td>SECONDARY</td>
<td>MARRIED</td>
<td>1197</td>
<td>MIXED</td>
</tr>
<tr>
<td>FEMALE</td>
<td>38</td>
<td>VOCATIONAL/GRAMMAR</td>
<td>MARRIED</td>
<td>NA</td>
<td>MOSTLY DISSATISFIED</td>
</tr>
<tr>
<td>FEMALE</td>
<td>73</td>
<td>VOCATIONAL/GRAMMAR</td>
<td>MARRIED</td>
<td>800</td>
<td>MOSTLY SATISFIED</td>
</tr>
<tr>
<td>FEMALE</td>
<td>54</td>
<td>PRIMARY/NO EDUCATION</td>
<td>MARRIED</td>
<td>NA</td>
<td>MOSTLY SATISFIED</td>
</tr>
<tr>
<td>MALE</td>
<td>30</td>
<td>VOCATIONAL/GRAMMAR</td>
<td>MARRIED</td>
<td>1500</td>
<td>MIXED</td>
</tr>
<tr>
<td>MALE</td>
<td>68</td>
<td>SECONDARY</td>
<td></td>
<td>-8</td>
<td>PLEASED</td>
</tr>
<tr>
<td>MALE</td>
<td>61</td>
<td>PRIMARY/NO EDUCATION</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Data that look (structurally) like original data but contain artificial units only
Generating synthetic data: method

Sequentially replacing original data values with synthetic values generated from conditional probability distributions.
Generating synthetic versions of sensitive microdata for statistical disclosure control

R package synthpop

http://cran.r-project.org/package=synthpop
synthpop: Generating Synthetic Versions of Sensitive Microdata for Statistical Disclosure Control

A tool for producing synthetic versions of microdata containing confidential information so that they are safe to be released to users for exploratory analysis. The key objective of generating synthetic data is to replace sensitive original values with synthetic ones causing minimal distortion of the statistical information contained in the data set. Variables, which can be categorical or continuous, are synthesised one-by-one using sequential modelling. Replacements are generated by drawing from conditional distributions fitted to the original data using parametric or classification and regression trees models. Data are synthesised via the function syn() which can be largely automated, if default settings are used, or with methods defined by the user. Optional parameters can be used to influence the disclosure risk and the analytical quality of the synthesised data.

Version: 1.1-1
Depends: lattice, MASS, methods, nnet, ggplot2, coefplot
Imports: graphics, stats, utils, rpart, party, foreign, plyr, proto
Published: 2015-07-08
Author: Beata Nowok, Gillian M Raab, Joshua Snoke and Chris Dibben
Maintainer: Beata Nowok <beata.nowok at gmail.com>
License: GPL-2 | GPL-3
NeedsCompilation: no
Materials: NEWS
In views: OfficialStatistics
CRAN checks: synthpop results

Downloads:

Reference manual: synthpop.pdf
Vignettes: Using synthpop
Package source: synthpop_1.1-1.tar.gz
Windows binaries: r-devel: synthpop_1.1-1.zip, r-release: synthpop_1.1-1.zip, r-oldrel: synthpop_1.1-1.zip
OS X Snow Leopard binaries: r-release: synthpop_1.1-1.tgz, r-oldrel: synthpop_1.1-0.tgz
OS X Mavericks binaries: r-release: synthpop_1.1-1.tgz
Old sources: synthpop archive
> install.packages("synthpop")
trying URL 'http://cran.r-project.org/bin/windows/contrib/3.3/synthpop_1.1-1.zip'
Content type 'application/zip' length 669748 bytes (654 KB)
downloaded 654 KB

package 'synthpop' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
   C:\Users\bnowok\AppData\Local\Temp\RtmpigR0o9\downloaded_packages

> library(synthpop)
Loading required package: lattice
Loading required package: MASS
Loading required package: nnet
Loading required package: ggplot2
Loading required package: coefplot
> |

Generating synthetic data: \textit{synthpop}
Generating synthetic data: synthpop

- Synthesis can be run with default parameters (CART – Classification and Regression Trees)

`syn(data)`
> sds <- syn(ods)
No seed has been provided and it has been set to 134.

syn variables
1  sex agegr placesize marital income smoke ls

> sds
Call:
($)call syn(data = ods)

Number of synthesised data sets:
($)n 1

First rows of synthesised data set:
($)syn

  sex agegr placesize marital income smoke ls
1 MALE 45-59 URBAN 20,000-100,000 MARRIED 1800 NO PLEASED
2 MALE 60-64 RURAL AREAS MARRIED 1300 YES MOSTLY DISSATISFIED
3 MALE 35-44 RURAL AREAS MARRIED 600 NO DELIGHTED
4 MALE 25-34 URBAN BELOW 20,000 SINGLE 1600 NO MOSTLY SATISFIED
5 FEMALE 25-34 RURAL AREAS MARRIED 1400 NO MOSTLY SATISFIED
6 MALE 60-64 RURAL AREAS SINGLE 2100 NO PLEASED
...

Synthesising methods:
($)method

  sex agegr placesize marital income smoke ls
"sample" "ctree" "ctree" "ctree" "ctree" "ctree" "ctree"

Order of synthesis:
($)visit.sequence

  sex agegr placesize marital income smoke ls
1 2 3 4 5 6 7

Matrix of predictors:
($)predictor.matrix

  sex agegr placesize marital income smoke ls
sex 0 0 0 0 0 0 0
agegr 1 0 0 0 0 0 0
placesize 1 1 0 0 0 0 0
marital 1 1 1 0 0 0 0
income 1 1 1 1 0 0 0
smoke 1 1 1 1 1 0 0
ls 1 1 1 1 1 1 0
Generating synthetic data sets

Description

Generates synthetic version(s) of a data set.

Usage

```
syn(data, method = vector("character", length = ncol(data)),
    visit.sequence = (1:ncol(data)), predictor.matrix = NULL,
    m = 1, k = nrow(data), proper = FALSE, minnumlevels = 5,
    maxfaclevels = 60, rules = NULL, rvalues = NULL,
    cont.na = NULL, semicont = NULL, smoothing = NULL,
    event = NULL, denom = NULL, drop.not.used = TRUE, drop.pred.only = FALSE,
    default.method = c("normrank", "logreg", "polyreg", "polr"),
    diagnostics = FALSE, print.flag = TRUE, seed = "sample", ...)
```

```
## S3 method for class 'synds'
print(x, ...)
```

Arguments

- `data`
  
a data frame or a matrix \((n \times p)\) containing the original data. Observations are in rows and variables are in columns.
syn() & common data problems

- **Missing-data codes:** cont.na
  - categorical variables: additional factor level(s)
  - continuous variables: specified by cont.na and modelled separately

- **Semi-continuous variables:** semicont

- **Restricted values (interrelationships between variables):** rules & rvalues

- **Linear constraints:** denom

- **Non-negativity / non-normality:** method set to ‘lognorm’, ‘sqrtnorm’ or ‘cubertnorm’

- **Deterministic relations:** method set to “~I(...)”
> sds <- syn(odds,
+ m = 5,
+ method = "cart",
+ cart.minbucket = 10,
+ cont.na = list(income = c(NA, -8)),
+ smoothing = list(income = "density"),
+ seed = NA,
+ print.flag = FALSE)
>
Overview of synthpop functions

- **descriptive models**
  - `compare.synds()`
  - `summary.synds()`
  - `compare.fit.synds()`
  - `glm.synds()`
  - `summary.fit.synds()`

- **data structure**
  - `utility.synds()`

- **observed variable**
  - `read.obs()`

- **synthetic variable**
  - `syn()`
  - `sdc()`
  - `write.syn()`
```r
> compare(sds, ods, vars = "income")

Comparing percentage

income
  0   10
observed  24.58  34.9
synthetic  21.78  35.1
  9000 1000
observed  0.100  0.04
synthetic  0.108  0.06
miss.NA
observed  13.660
synthetic  13.084
```
```r
> compare(sds, ods, vars = "edu", msel = 1:3)

Comparing percentages observed

$edu

<table>
<thead>
<tr>
<th></th>
<th>PRIMARY/NO EDUCATION</th>
<th>POST-SECONDARY</th>
</tr>
</thead>
<tbody>
<tr>
<td>observed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>syn=1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>syn=2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>syn=3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

> |
```
```r
> ffsd <- glm.synds(I(smoke=="YES") ~ sex + agegr + placesize +
+ edu + socprof + marital + income,
+ data = sds, family = "binomial")

Call used to fit the model:
glm.synds(formula = I(smoke == "YES") ~ sex + agegr + placesize +
edu + socprof + marital + income, data = sds, family = "binomial")

Estimates for

(Intercept)
sexFEMALE
agegr25-34
agegr35-44
agegr45-59
agegr60-64
agegr65+
placesizeURBAN
placesizeURBAN
placesizeURBAN
placesizeURBAN
placesizeURBAN
agegr65+
agegr60-64
agegr45-59
agegr35-44
agegr25-34
sexFEMALE
```
utility.synds()

Method:
($method)
cart

Mean and SD of propensity score utility value for synthetized data:
($Utility.Summary)

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>cart</td>
<td>0.1266</td>
<td>0.0053</td>
</tr>
</tbody>
</table>

Individual propensity score utility values for each synthetic set:
($Utility.Raw)

cart

| syn=1  | 0.1183 |
| syn=2  | 0.1259 |
| syn=3  | 0.1282 |
| syn=4  | 0.1276 |
| syn=5  | 0.1331 |
sdc() & statistical disclosure control

- Data labelling: label
- Removing replicated uniques: rm.replicated.uniques
- Bottom- and top-coding: recode.vars, bottom.top.coding, recode.exclude
- At synthesis stage: smoothing, minbucket
```r
sdssdc <- sdc(sds,
+     ods,
+     label = "false_data",
+     rm.replicated.uniques = TRUE,
+     recode.vars = "income",
+     bottom.top.coding = c(200, 5000),
+     recode.exclude = -8)
```

m = 1
income: no. of bottom-coded values - 26, no. of top-coded values - 97

m = 2
income: no. of bottom-coded values - 24, no. of top-coded values - 72

m = 3
income: no. of bottom-coded values - 18, no. of top-coded values - 80

m = 4
income: no. of bottom-coded values - 21, no. of top-coded values - 90

m = 5
income: no. of bottom-coded values - 18, no. of top-coded values - 83

No. of unique units in the synthesised data that replicates unique real individuals:
98 94 103 112 92
They have been removed from the synthetic data set(s).

> |
Conclusions

The **synthpop** package for **R**: facilitating generation, evaluation and analysis of synthetic data