Statistical Journal of the IAOS 33 (2017) 785–796
DOI 10.3233/SJI-150153
IOS Press

Providing bespoke synthetic data for the UK Longitudinal Studies and other sensitive data with the *synthpop* package for R[1]

Beata Nowok*, Gillian M. Raab and Chris Dibben
Administrative Data Research Centre – Scotland, School of GeoSciences, University of Edinburgh, Edinburgh, UK

Abstract. Synthetic data methods were designed to address the conflicting demands placed on data holders to unlock the research and policy potential of microdata while at the same time preserving the confidentiality of individuals. Recently, these methods have become more widely recognized in the UK and the provision of bespoke synthetic data has been approved to expand the use of one of the UK Longitudinal Studies. The process of producing useful synthetic data involves, however, a substantial investment of research time, as it always requires some customising for the characteristics of an individual data set. At the same time, a substantial part of it can be automated and this is essential when the process has to be conducted rapidly and on a regular basis. This paper describes the application of synthetic data to the UK Longitudinal Studies, details implementation process for the Scottish Longitudinal Study and presents methods used in an **R** package *synthpop* that has been developed to facilitate production of non-disclosive entirely synthetic data. A reproducible example using open data is given to illustrate the synthesising procedure and to provide insights into quality of synthetic data generated using different automated approaches.

Keywords: Synthetic data, confidentiality, statistical disclosure control, CART, UK Longitudinal Studies

1. Introduction

Ever since synthetic data was first proposed by Rubin [1] its main objective has been the reduction of disclosure risk in confidential microdata so that they can be released for wider use. The original sensitive data are used to estimate models for generating synthetic values to replace information judged as at risk of disclosure. When synthesis is complete, i.e. all values of all variables are generated from the models, the released synthetic data set includes only individual completely artificial units. At the same time conclusions that can be drawn from such data should not be compromised. Despite the attractiveness of this approach and recent developments in this area [2–9], the practical examples of synthetic data products are very limited [4,10,11] and their users are urged to request validation of results with original data. This is not surprising, because preserving all possible relationships between variables in a data set can be next to impossible. Nonetheless, data users can still benefit from having access to synthetic data for a preliminary analysis with approximate results if the associated risks are well understood.

With the aim of furthering developments in the area of synthetic data, this article presents a synthesising method and accompanying software that can be used to widen access to confidential microdata such as the UK Longitudinal Studies (LSs). The next section introduces the LSs, describes how synthetic data can be used in their context and how they are being implemented in one of the LSs. The following sections are more general in nature and they aim to explain and illustrate the process of generating synthetic data that can be applied also by other agencies. Section 3 dis-

[1] A paper presented at the NTTS 2015 conference.

*Corresponding author: Beata Nowok, Administrative Data Research Centre – Scotland, School of GeoSciences, University of Edinburgh, Drummond Street, Edinburgh EH8 9XP, UK. Tel.: +44 1316514014; Fax: +44 1316502524; E-mail: beata.nowok@ed.ac.uk.

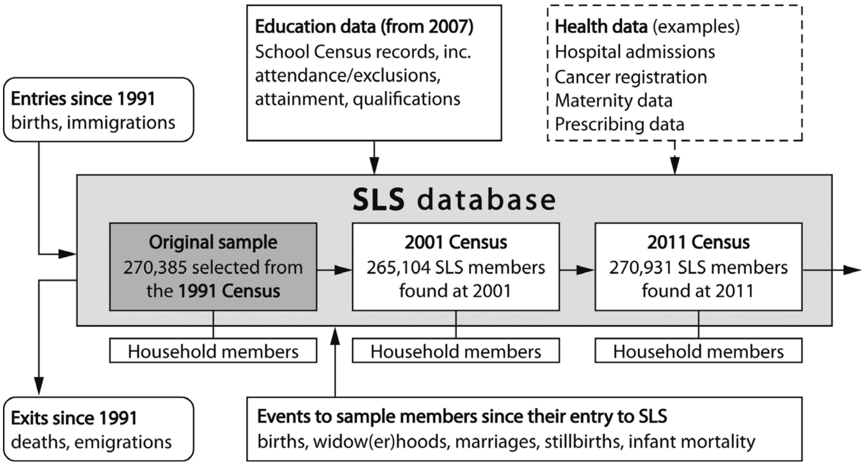

Education data (from 2007)
School Census records, inc. attendance/exclusions, attainment, qualifications
Health data (examples)
Hospital admissions
Cancer registration
Maternity data
Prescribing data
Entries since 1991
births, immigrations
SLS database
Original sample
270,385 selected from the 1991 Census
2001 Census
265,104 SLS members found at 2001
2011 Census
270,931 SLS members found at 2011
Household members
Household members
Household members
Exits since 1991
deaths, emigrations
Events to sample members since their entry to SLS
births, widow(er)hoods, marriages, stillbirths, infant mortality

Fig. 1. Structure of data sets linked into the SLS as of March 2016. (Dashed box indicates data that are linked in to the SLS on a project by project basis).

cusses the synthesising method. Section 4 presents the *synthpop* package for **R** and its main functionalities. The section that follows uses open data to provide a reproducible example in which the quality of synthetic data generated using different automated approaches is evaluated. The last section concludes the paper.

2. Synthetic data for the UK Longitudinal Studies

2.1. The UK Longitudinal Studies

The England and Wales Longitudinal Study (ONS LS) [12], the Scottish Longitudinal Study (SLS) [13] and the Northern Ireland Longitudinal Study (NILS) [14] are rich microdata sets linking samples from the national census in each country to administrative data for individuals and their immediate families across several decades. All of the LSs have a similar structure. At their core are data from the UK decennial census for the relevant country. Individuals are linked over time across censuses and to administrative data on births, deaths, marriages, records of immigration and emigration from the relevant country and other sources. Figure 1 illustrates the data that are currently linked or can be linked on a project by project basis to the SLS, including the censuses held in 1991, 2001 and 2011. Negotiations are ongoing for more data sources to be linked or linkable and individual projects may add extra data. The SLS also includes detailed geographic identifiers which allow linkage to environmental data and other sources aggregated for small areas. The other LSs have somewhat different structures and link to different ranges of administrative data. The ONS LS includes data on five censuses starting in 1971 while the NILS has data from the latest three censuses. More information about the three LSs and their support units can be found at http://calls.ac.uk/.

The data are extremely sensitive. The core census data are controlled by the Census Acts which means that public access is not available for 100 years. Inclusion in the studies is by a number of “secret” birthdays spread throughout the year that are known only to a very few core staff in each study. There are four such birthdays in the ONS LS, 20 in the SLS and 104 in the NILS, differing because of the different sampling fractions needed for each population size. No resident of the UK knows whether they are included in one of the LSs, which is justified ethically by the extremely secure conditions under which the data are held. Controls are in place to ensure no published analyses or even private tables taken out of the secure settings have the potential for disclosure of information about individuals or identifiable small subgroups.

Each of the LSs holds many thousands of variables in a series of files. No user has access to all of the data. Moreover, for data that cannot be permanently linked to LSs the linkage has to be done for each project separately by a third party. Following an application to use a study the user specifies the data required and an extract is prepared and linked to other data if requested. Whilst unique and valuable resources, the sensitive nature of the information they contain, and the legal restrictions that apply to census data, mean that access to the microdata is restricted to approved researchers and LSs support staff, who can only view and work with the data in safe settings controlled by the national statistical agencies (NSAs). The restrictive access regime has a detrimental impact on usage and limits potential impact of the three LSs.

2.2. *Application of synthetic data*

Completely synthesised data with all values generated from the models and with no real individuals, but which mimic the original observed data and preserve the relationships between variables and transitions of individuals over time has been recognized as a measure that may help to facilitate access to the LSs. Because they contain records of artificial individuals only, such data can be made available to trained and accredited researchers for a preliminary analysis and the development of analysis code on their own computers. This solution offers substantial costs and time savings related to visits to safe havens because data cleaning and preparation takes a significant part of the research time. Users can also opt for remote execution without the need to come to safe havens at all.

The synthetic data need to resemble the actual data as closely as possible, but will never be used in any final analyses. The users will carry out exploratory analyses and test models on the synthetic data, but they, or perhaps LSs support staff, will use the code developed on the synthetic data to run their final analyses on the original data set. This approach recognises the limitations of synthetic data to preserve all statistical properties of the observed data. A similar approach is currently being used for the synthetic products made available by the U.S. Census Bureau (see http://www.census.gov/programs-surveys/sipp/guidance/sipp-synthetic-beta-data-product.html and https://www.census.gov/ces/dataproducts/synlbd/) whereby no guarantee of validity is given unless the final analysis is run on the original data. These initiatives use the term "gold standard analysis" for the final analysis carried out on the original data and we follow this convention. When a final analysis will be run on the original data, we require inference about the gold standard parameters, rather than for the corresponding population parameters. The results from the synthetic data will then be used as a guide to what will be found from the original data, rather than as a means of obtaining inferences for population parameters.

Researchers will be provided with a synthetic version of an extract with just the variables they require for the population relevant to their research. Synthetic data will be created for each extract separately and will thus be project-specific. The need for bespoke syntheses results from the fact that every user of the LSs has a customised data set made available to them and the linked data can include variables that are not part of the LSs database. They can come from data sets held by other agencies or they can be provided by a user. The wide scope of data that can be linked to the LSs, and which are often updated frequently, makes a comprehensive one-off synthesis practically infeasible.

The need for a new bespoke synthetic data set for each user requires a largely automated means of data synthesis. Thus the **R** *synthpop* package [15] has been written as part of the UK Economic and Social Research Council funded SYLLS project (Synthetic Data Estimation for UK Longitudinal Studies; for more details see http://www.lscs.ac.uk/projects/synthetic-data-estimation-for-uk-longitudinal-studies/) to allow LS support staff to produce synthetic data tailored to the needs of each investigation. The package is still under development so that in future synthetic data of acceptable quality can be provided to all users of the LSs. At the current stage some complex data structures or problematic variables may preclude this.

2.3. *Implementation of synthetic data in the Scottish Longitudinal Study*

The three LSs are governed by different bodies that are developing their own set of principles and practices for the use of synthetic data. In England and Wales and in Northern Ireland the synthetic data methods and the potential for the provision of synthetic data sets to the ONS LS and the NILS users respectively are still under evaluation at the time of writing this paper (March 2016). In Scotland the proposal to supply synthetic data to the SLS users has been accepted and the implementation process is under way. Synthetic data have been produced and released for a first internal pilot project. A couple of other projects are going to be treated as test cases before an option to request a synthetic version of a data extract is available to all users.

Completely synthetic data such as those to be provided to the SLS users do not by definition include real units and would be expected to pose very little disclosure risk. This has been confirmed by studies evaluating their risks [11,16–19]. Elliot [20] came to a similar conclusion in his report on the disclosure risk associated with the synthetic data produced using the *synthpop* package. However, the actual and especially the perceived risks are not zero, thus some additional measures will be taken to minimise them. A single synthetic data set will be released for use outside the safe setting only to approved researchers who are granted access to the original sensitive data. It will be labelled appropriately to make it clear that the data are fake so no one mistakenly believe them to be real, with subsequent loss of reputation for the data collection agencies. A data file name will include "FALSE_DATA" and the first variable in the dataset will have value "FALSE_DATA" for all cases. In addition, any observations that are unique in the actual data and are replicated by chance as unique in the synthetic data will be removed from the latter.

3. Method

The key objective of producing synthetic versions of original data sets is to replace sensitive values with synthetic ones causing minimal distortion of the statistical information contained in the data set and to release the synthesised non-disclosive data for public use. A very high level of disclosure protection is achieved when all values are replaced and data comprise records of artificial individuals rather than actual ones. Such data will be produced for the extracts from the LSs. Alternatively, only those parts of the data considered to pose a disclosure risk could be synthesised leaving the rest of the data unchanged. Replacements are generated by drawing from synthesising models fitted to the original data and the correctness of the models determines the utility of the created data sets. The essential features of our synthesis procedure is presented below.

Variables are synthesised one by one using sequential regression modelling. It means that conditional distributions, from which synthetic values are drawn, are defined for each variable separately. As described in [21] synthetic values can be generated from distributions with parameters fitted to the observed data with or without sampling the parameters from their posterior distributions. In the former case we refer to "proper synthesis" and in the latter to "simple synthesis". Note that the fitted regression models are conditioned on the original variables that are earlier in the synthesis sequence. Consider as an example a default synthesis, i.e. synthesis with all values of all variables $(Y_1, Y_2, \ldots, Y_p)$ to be replaced. The first variable to be synthesised Y_1 cannot have any predictors and therefore its synthetic values are generated by random sampling with replacement from its observed values. Then the distribution of Y_2 conditional on Y_1 is estimated and the synthetic values of Y_2 are generated using the fitted model and the synthesised values of Y_1. Next the distribution of Y_3 conditional on Y_1 and Y_2 is estimated and used along with synthetic values of Y_1 and Y_2 to generate synthetic values of Y_3 and so on. The distribution of the last variable Y_p will be conditional on all other variables. Similar conditional specification approaches are used in most implementations of synthetic data generation. They are preferred to joint modelling not only because of the ease of implementation but also because of their flexibility to apply methods that take into account structural features of the data such as logical constraints or missing data patterns.

With practicality and flexibility in mind, classification and regression trees (CART) are used as the default conditional models for synthesis but various parametric alternatives are also available. CART methods [22] are an algorithmic modelling approach that can be applied to any type of data. The basic idea is to recursively split a data set into groups with increasingly homogeneous outcome. The splits are specified as yes-no questions referring to the predictor space. The values in each final group approximate the conditional distribution of the predicted variable for units with predictors meeting the criteria that define that group. The synthetic values are generated by sampling from an appropriate group. CART models were suggested for the generation of synthetic data by Reiter [9] and then evaluated as performing well by Drechsler and Reiter [23]. The key advantage of CART models is the ability to capture, in an automatic manner, non-linear relationships and interaction effects in the data that can be difficult to model using parametric approach.

4. The *synthpop* package for R

The *synthpop* package [24] is an add-on package to the statistical software **R** [25] and it is freely available from the Comprehensive **R** Archive Network (http://cran.r-project.org/package=synthpop). It utilises the structure and some functions of the *mice 2.18* multiple imputation package [26] but alters and extends it for the specific purpose of generating synthetic data. Although the *synthpop* package has been developed to enable staff of the LSs to generate bespoke synthetic data for users of these resources, the package and its functions can be used for a broad range of data sets. It also allows the user to extend the types of data that can be handled by writing their own routines that can be easily incorporated into the synthesising function.

Below we present a default synthesis and some tools available in the package to analyse synthesised data sets. Also, we describe briefly some of the additional features that allow synthesis to be tailored for a specific data set. A more exhaustive description of package functionalities can be found in the **R** documentation for the function `syn()` (command `?syn` at the **R** console) or in [15] that also provides a worked example.

4.1. Default synthesis and analytical tools

The *synthpop* package provides routines to generate a synthetic version of original data sets. Via the function `syn()` synthetic data are produced using a single command. To run a default synthesis only the data to be synthesised have to be provided as a function argument, e.g. `syn(mydata)`, where `mydata` is a name of a data frame that contains the original confidential microdata. As described in Section 3 variables are synthesised sequentially one by one and the default order of synthesis (`visit.sequence`) reflects the order of variables in the observed data set, i.e. column variables are synthesised from left to right. The default predictor selection matrix (`predictor.matrix`) is defined by the visit sequence and all variables that are earlier in the visit sequence are used as predictors. By default a single synthetic data set is generated (`m`). The default synthesising method `"cart"` uses an implementation of CART models from the `rpart` package [27], which in most details follows [22]. Setting the `syn()` function parameter `method` to `"parametric"` assigns default parametric methods to synthesised variables based on their types. The default parametric methods for numeric, binary, unordered factors and ordered factors are respectively normal linear regression preserving the marginal distribution, logistic regression, polytomous logistic regression and ordered polytomous logistic regression. To preserve marginal distributions of numeric variables, the data are transformed to normal scores calculated from their percentiles and regression is carried out on these values. The predicted values are then mapped back to the original data with the reverse transformation. The default parametric methods may be customised by setting the `default.method` parameter of function `syn()` to desired available methods. All synthesising methods available in *synthpop* can be consulted in the **R** help file for the function `syn()`. By default simple synthesis is used for all methods, but setting the parameter `proper` to `TRUE` will call the proper equivalent. Note that the first variable to be synthesised cannot have any predictors from the variables to be synthesised later and therefore a random sample with replacement is drawn from its observed values.

The *synthpop* package contains routines that can be used to summarise synthetic data and compare descriptive statistics of the original and synthetic data sets. The analyst with access to synthetic data only can use `summary()` function. The support officer, with access to the original data, can compare the distributions of the synthetic data to that of the actual data. Following a synthesis the `compare()` function can be used to check each variable using tabular and graphic output (it is a generic function for comparison of various aspects of synthesised and observed data, which invokes particular methods depending on the class of its argument `object`). Figure 2 in Section 5 is an example of a graphic output for variable `edu` (education), `marital` (marital status), `income` and `age`. Generalized linear models from synthetic data can be estimated using `glm.synds()` function. Model fits can be summarized using the `summary` method. The standard errors are estimated differently depending whether inference is made for the results that would be obtained from the observed data or for the parameters of the population that we assume the observed data are sampled from. The standard errors also differ according to whether synthetic data were produced using simple or proper synthesis (for details see [21]). If the original data are available, the gold standard analyses can be compared with those from the synthetic data using function `compare()`. The coefficients with confidence intervals for the observed data are plotted together with their estimates from synthetic data (see Fig. 3 in Section 5 for an example). When more than one synthetic data set has been generated appropriate combining rules are applied as described by Raab et al. [21]. The tabular output includes model estimates for both original and synthetic data and two measures assessing similarity of the results: a confidence interval overlap and a standardized difference in coefficient values.

4.2. Additional features

The *synthpop* package offers a variety of options to customize the default synthesis presented in Section 4.1. They enable improvements of the utility of the synthetic data but also they provide tools for en-

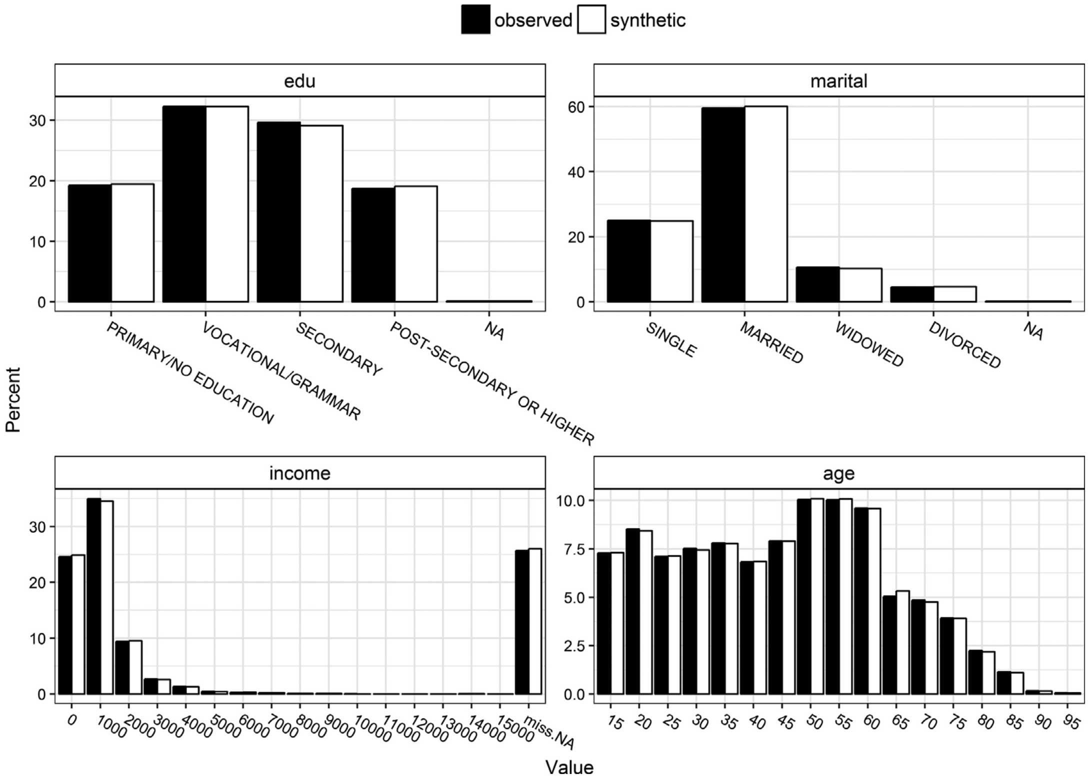
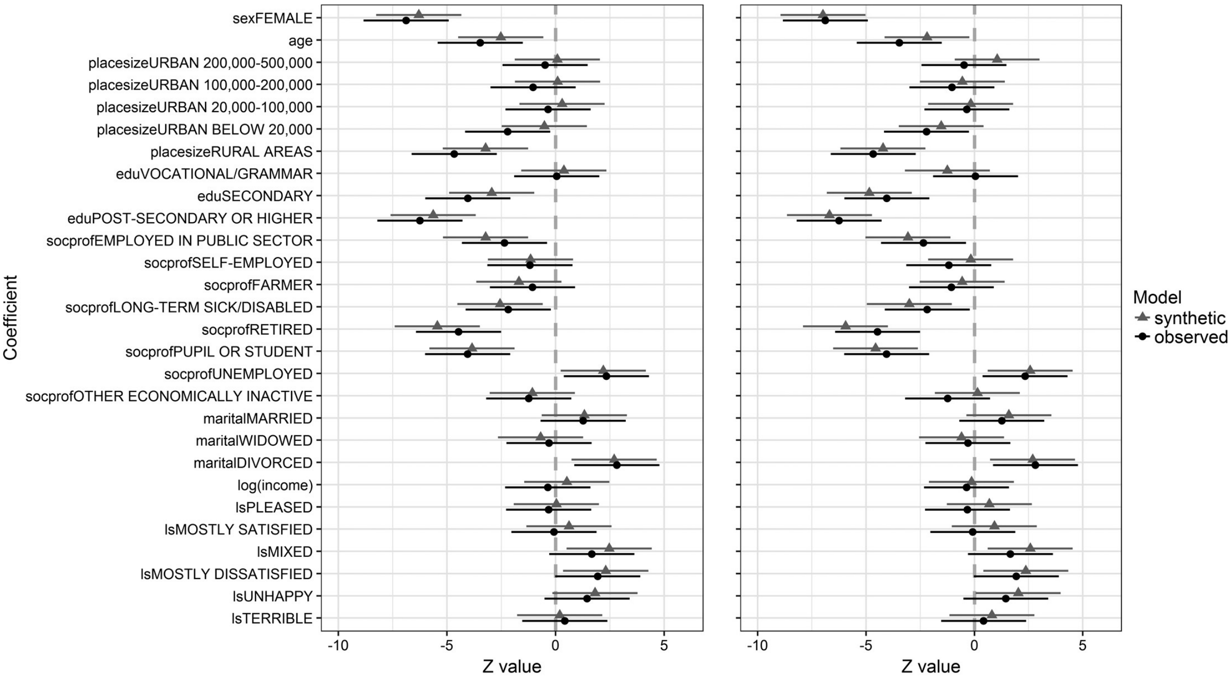


Fig. 2. Comparison of original and synthetic variables for CART models.

hanced protection of the information confidentiality, which may, however, impair the data quality. Users who want to pursue the development of more plausible synthesising models than the default ones can do it for each variable separately. A model type can be chosen out of the methods currently implemented or a new synthesising method can be introduced by writing a function named `syn.newmethod()` and then specifying elements of the `method` parameter of `syn()` function as `"newmethod"`. Next to model type specification, users can select the set of variables to include as predictors. The choice of explanatory variables is restricted by the order in which variables are synthesised. Variables that are not yet synthesised cannot be used in prediction models, but dependencies are maintained when those later in the sequence are themselves predicted. The synthesis order can be changed and it is also possible to include as predictors variables that do not belong to the data set to be synthesised. The user can also customise the predictor selection matrix to fit smaller models. An example when this is useful is when there are variables such as occupational categories with many possible values and thus potentially small cells. Those variables can be excluded from predictors and replaced by their values recoded into broad groups such as socio-economic status for occupation.

The `syn()` function also includes procedures to mimic the characteristics of the original confidential data in many possible ways. With optional parameters it can take into account structural features of the data such as missing data patterns or restricted values. Both issues and methods to deal with them are described below.

Values representing missing data in categorical variables are treated as additional categories and reproducing them is straightforward. Continuous variables with missing data are modelled in two steps. In the first step, we synthesise an auxiliary binary variable specifying whether a value is missing or not. Depending on the method specified by a user for the original variable a logit or CART model is used for synthesis. If there are different types of missing values an auxiliary categorical variable is created to reflect this and an appropriate model is used for synthesis (a polytomous or CART model). In the second step, a synthesising model is fitted to the non-missing values in the original variable and then used to generate synthetic values for the non-missing category records in our auxiliary variable. The



(a) CART

(b) Parametric

Fig. 3. Coefficients from a logistic regression of smoking comparing estimates and 95% confidence intervals for Z statistics for original data and default synthesis using (a) CART and (b) parametric models.

auxiliary variable and a variable with non-missing values and zeros for remaining records are used instead of the original variable for prediction of other variables. The missing-data codes have to be specified by a user in `cont.na` parameter of the `syn()` function, unless only the **R** missing data code `NA` is used. Otherwise numeric missing-data codes for a continuous variable are treated as non-missing values. This may lead to erroneous synthetic values, especially when standard parametric models are used or when synthetic values are smoothed to decrease disclosure risk. The problem refers not only to the variable in question, but also to variables predicted from it.

Restricted values are those where the values for some cases are determined explicitly by those of other variables. For instance, the number of cigarettes smoked per day by non-smokers should be zero or marital status of young individuals (an exact age is country dependent) should be `SINGLE`. In such cases the rules and the corresponding values should be specified using `rules` and `rvalues` parameters. The variables used in `rules` have to be synthesised prior to the variable they refer to. In the synthesis process the restricted values are assigned first and then only the records with unrestricted values are synthesised.

Other original data problems that can be handled with optional parameters include semi-continuous variables, deterministic relations and linear constraints between the variables (e.g. the number of children in a household cannot be larger than the number of all household members).

The completely synthetic data produced using the *synthpop* package implementing the methods presented above do not by definition include real units and should provide sufficient disclosure protection. Nonetheless, there are a number of options that are designed to protect the data further and limit the perceived disclosure. For the CART method, the classification may produce a group which includes only a very small number of individuals and this may elevate the risk of replicating by chance a real person. To avoid this, a user can specify a minimum size of a final group that a CART model can produce. There is also an option that allows smoothing of continuous variables, to prevent rare original values being reproduced in the synthetic data. An additional precautionary option is built into the package in the `sdc()` function, which enables the user to remove from the synthetic data set any cases with unique variable sequences that are identical to unique individuals in the original dataset. This should reduce the chances of a person who is in the

Table 1
Z statistics for fits of smoking model to original data, CART and parametric syntheses

Coefficient	Z gold-standard	Z CART	Z parametric
sexMALE *(ref. category)*			
sexFEMALE	−6.88	−6.30	−6.99
age	−3.47	−2.52	−2.19
placesizeURBAN 500,000 AND OVER *(ref. category)*			
placesizeURBAN 200,000–500,000	−0.48	0.08	1.05
placesizeURBAN 100,000–200,000	−1.04	0.09	−0.56
placesizeURBAN 20,000–100,000	−0.35	0.30	−0.17
placesizeURBAN BELOW 20,000	−2.21	−0.52	−1.53
placesizeRURAL AREAS	−4.67	−3.23	−4.22
eduPRIMARY/NO EDUCATION *(ref. category)*			
eduVOCATIONAL/GRAMMAR	0.05	0.38	−1.25
eduSECONDARY	−4.04	−2.94	−4.85
eduPOST-SECONDARY OR HIGHER	−6.24	−5.64	−6.69
socprofEMPLOYED IN PRIVATE SECTOR *(ref. category)*			
socprofEMPLOYED IN PUBLIC SECTOR	−2.35	−3.22	−3.07
socprofSELF-EMPLOYED	−1.18	−1.15	−0.17
socprofFARMER	−1.06	−1.69	−0.57
socprofLONG-TERM SICK/DISABLED	−2.18	−2.56	−3.01
socprofRETIRED	−4.47	−5.44	−5.95
socprofPUPIL OR STUDENT	−4.05	−3.85	−4.56
socprofUNEMPLOYED	2.34	2.19	2.57
socprofOTHER ECONOMICALLY INACTIVE	−1.23	−1.07	0.14
maritalSINGLE *(ref. category)*			
maritalMARRIED	1.27	1.32	1.59
maritalWIDOWED	−0.30	−0.69	−0.59
maritalDIVORCED	2.82	2.70	2.69
log(income)	−0.36	0.52	−0.14
lsDELIGHTED *(ref. category)*			
lsPLEASED	−0.32	0.03	0.69
lsMOSTLY SATISFIED	−0.07	0.61	0.92
lsMIXED	1.66	2.47	2.58
lsMOSTLY DISSATISFIED	1.93	2.31	2.37
lsUNHAPPY	1.45	1.81	2.02
lsTERRIBLE	0.42	0.19	0.81

original data believing that their actual data are in the synthetic data. This function can also be used for labelling synthetic data as “false” to prevent intruders from believing that they are real.

5. Synthetic data quality

The characteristics and quality of synthetic data depend on the models used to generate them. Specifying appropriate models that capture all essential features of the original data is therefore crucial but it requires expertise in both the data to be synthesised and statistical methods. Moreover, it can be cumbersome and poses a major obstacle for data custodians with limited resources. With this in mind, we use the *synthpop* package and its `syn()` function to synthesise data with default settings for CART and parametric models and we evaluate and compare the quality of the results. In general, we aim to reproduce the logical structure of the data, univariate distributions and multivariate relations among the variables so that an analysis based on the synthetic data leads to the similar statistical inference as an analysis based on the observed data.

The *synthpop* package is designed for general application and is not specific to LSs data. Here, for illustrative purposes and ease of replication, we use a subset of freely available survey data collected within the Social Diagnosis project in 2011 [28] which aims to investigate objective and subjective quality of life in Poland. This subset is included in the *synthpop* package. It is called `SD2011` and it contains 35 selected variables of various types for a sample of 5,000 individuals aged 16 and over (see the **R** documentation for the data set `SD2011` for more details; the complete data set is available at http://www.diagnoza.com/index-en.html along with detailed documentation). For our example we have extracted data on sex, age, highest educational qualification, type of the place of residence, socio-economic status, marital status, personal monthly net

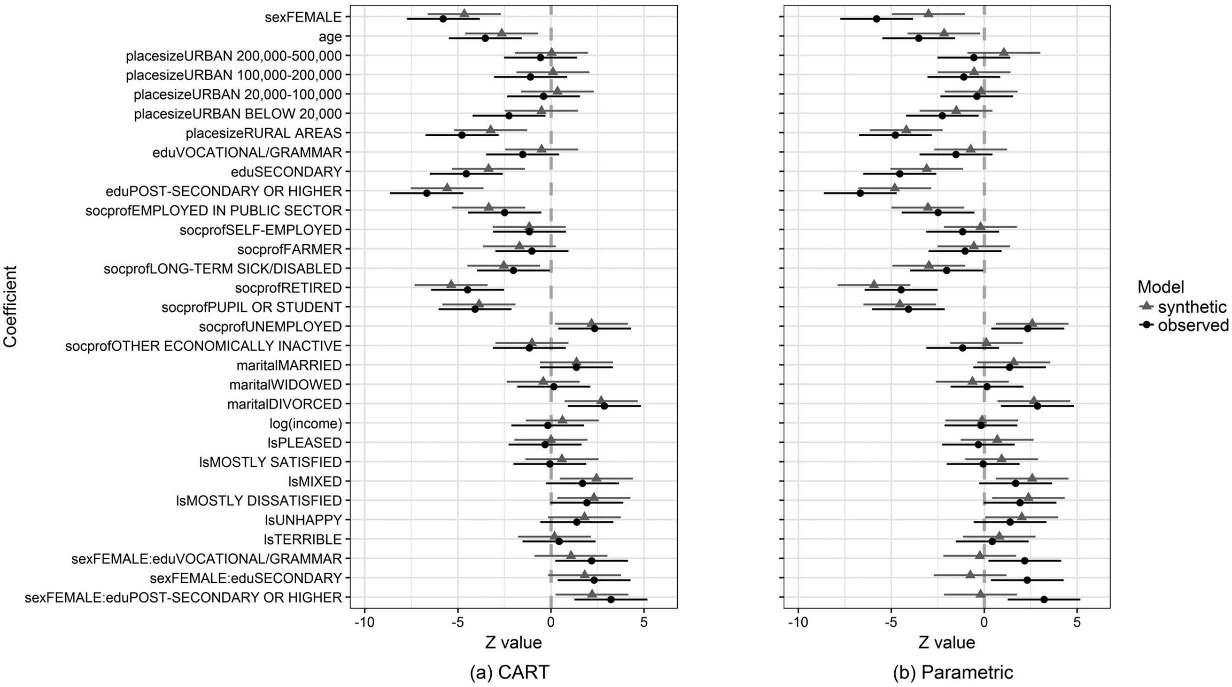


Fig. 4. Coefficients from a logistic regression of smoking with an interaction term comparing estimates and 95% confidence intervals for Z statistics for original data and default synthesis using (a) CART and (b) parametric models.

income, overall life satisfaction and cigarette smoking status. The names `sex`, `age`, `edu`, `placesize`, `socprof`, `marital`, `income`, `ls` and `smoke` are used to describe them. Our data set includes predominantly categorical variables, which is also the case for the LSs. The synthesis was conducted for all 5,000 individuals including cases with missing observations. Numeric missing data code (`−8`) for the income variable was, however, recoded to `NA` and due to small number of legally or de facto separated individuals (29) they were combined with the divorcees. Multiple synthetic data sets (`m=10`) were produced for both non-parametric (CART) and parametric synthesis. The **R** code to reproduce all results can be found in Appendix.

In terms of univariate distributions, both the CART and parametric methods gave satisfactory results. Comparison of the distributions of four variables synthesised using CART models to that of their observed values are shown in Fig. 2.

To investigate how well the parameters of models fitted to observed data are estimated by the synthetic data, we have modelled the factors that affect smoking by logistic regression with data synthesised by parametric and non-parametric methods. The outcome being modelled is the chance of being a smoker. In each case we use means of ten syntheses to get better estimates of the bias and variance of the gold standard parameters. Besides, we assume that the researcher is interested in estimating from the synthetic data the results which might be obtained from the original data, rather than in making inferences directly to population parameters. The function `glm.synds()` is used to calculate and store fits for each of the ten synthetic data sets, and `summary()` is used to combine them using the formulae from [21] (combined results can also be obtained using `compare()` function). Table 1 compares the Z statistics from the original data and its estimates from the two methods of synthesis, using simple synthesis in both cases. Figure 3, produced using the `compare()` function, plots the results of comparing the synthetic data with the actual data.

From the original data we can see that there is evidence that women smoke less as do those with higher education, those from rural areas, those who are single (as opposed to divorced) and older people. As regards socio-economic status, the chance of meeting a smoker is the highest for the unemployed and the lowest for students and retirees. The results from synthetic data generated by both the CART and parametric methods have similar patterns to the results from the original data. In most cases the estimates are very close to the real ones but as Fig. 3 shows there are some exceptions. The results from the model fitted to the synthetic data provide stronger evidence that low life satisfac-

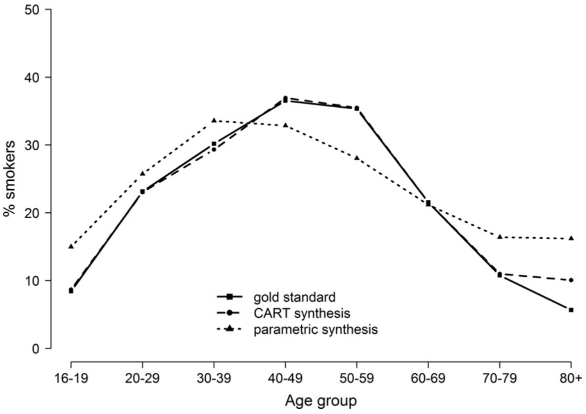


Fig. 5. Smoking rates by age group for original and synthesised data.

tion is conducive to smoking than the results from the observed data.

Despite some minor differences, based on the outcomes of the data analysis carried out so far we could conclude that both synthesising approaches produce satisfactory results. Further exploration undermines, however, the default parametric methods. An interaction term between sex and education added to the model presented above shows that the impact of increasing level of education on the prevalence of smoking is somewhat attenuated for women (see Fig. 4). The pattern of this effect is reflected in the estimates obtained from the data synthesised using CART methods but the parametric synthesis completely fails to capture it.

Next, note that the fitted models assume that smoking has a linear relationship with age but as we show in Fig. 5 this is very far from the truth. We can see that the parametric synthesis, using a linear model, misrepresent substantially smoking rates by age group. They are attenuated for people aged 40–59 and elevated for those aged 70 and over. The CART synthesis does a good job of capturing the age pattern, with only minor difference in smoking rate at the oldest ages.

This example confirms the advantage of CART synthesising models in their ability to capture in an automatic manner non-linear relationships and interaction effects in the data that would have to be specified explicitly in parametric models.

6. Final conclusions

Synthetic data offer a way to expand the use of confidential microdata such as the UK LSs. The *synthpop* package for **R** with its default CART method has been developed to facilitate generation of such data. It is not, however, a final product and feedback from package and synthetic data users is absolutely invaluable for further improvements. There is much still to be learned in the field of data synthesis, particularly its practical aspects. Next to facing the challenge of producing synthetic versions of complex data sets and further investigations of the strengths and weaknesses of different synthesising methods, more work on disclosure control is needed in order to address the concerns of data custodians responsible for protecting confidentiality. Finally, weaknesses and limitations of synthetic data have to be clearly communicated. We hope that the availability of the *synthpop* package may help to facilitate further research and development of best practices.

References

[1] D.B. Rubin, Discussion: Statistical disclosure limitation, *Journal of Official Statistics* **9** (1993), 461–468.

[2] G. Caiola and J.P. Reiter, Random forests for generating partially synthetic, categorical data, *Transactions on Data Privacy* **3** (2010), 27–42.

[3] J. Drechsler, Synthetic data sets for statistical disclosure control: Theory and implementation, New York: Springer Science+Business Media, 2011. doi: 10.1007/978-1-4614-0326-5.

[4] J. Drechsler, New data dissemination approaches in old Europe – synthetic datasets for a German establishment survey, *Journal of Applied Statistics* **39** (2012), 243–265.

[5] J. Drechsler and J.P. Reiter, Sampling with synthesis: A new approach for releasing public use census microdata, *Journal of the American Statistical Association* **105** (2010), 1347–1357. doi: 10.1198/jasa.2010.ap09480.

[6] S.K. Kinney and J.P. Reiter, Tests of multivariate hypotheses when using multiple imputation for missing data and disclosure limitation, *Journal of Official Statistics* **26** (2010), 301–315.

[7] J.P. Reiter, Satisfying disclosure restrictions with synthetic data sets, *Journal of Official Statistics* **18** (2002), 531–544.

[8] J.P. Reiter, Releasing multiply imputed, synthetic public use microdata: An illustration and empirical study, *Journal of the Royal Statistical Society*, Series A: Statistics in Society **168** (2005), 185–205. doi: 10.1111/j.1467-985x.2004.00343.x.

[9] J.P. Reiter, Using CART to generate partially synthetic public use microdata, *Journal of Official Statistics* **21** (2005), 441–462.

[10] J.M. Abowd, B.E. Stephens, L. Vilhuber, F. Andersson, K.L. McKinney, M. Roemer and S. Woodcock, The LEHD infrastructure files and the creation of the quarterly workforce indicators, in: *Producer Dynamics: New Evidence from Micro Data*, T. Dunne, J.B. Jensen and M.J. Roberts, eds, Chicago (IL): University of Chicago Press for the National Bureau of Economic Research, 2009, pp. 149–230.

[11] S.K. Kinney, J.P. Reiter, A.P. Reznek, J. Miranda, R.S. Jarmin and J.M. Abowd, Towards unrestricted public use business microdata: The synthetic longitudinal business database, *In-*

ternational Statistical Review **79** (2011), 362–384. doi: 10.1111/j.1751-5823.2011.00153.x.

[12] L. Hattersley and R. Cresser, The Longitudinal Study, 1971–1991: History, organisation and quality of data, LS Series no. 7, London: The Stationery Office, 1995.

[13] P. Boyle, P. Feijten, Z. Feng, L. Hattersley, Z. Huang, J. Nolan and G.M. Raab, Cohort profile: The Scottish Longitudinal Study (SLS), *International Journal of Epidemiology* **38** (2009), 385–392.

[14] D. O'Reilly, M. Rosato, G. Catney, F. Johnston and M. Brolly, Cohort description: The Northern Ireland Longitudinal Study (NILS), *International Journal of Epidemiology* **41** (2009), 634–641.

[15] B. Nowok, G.M. Raab and C. Dibben, Synthpop: Bespoke creation of synthetic data in **R**, *Journal of Statistical Software* **74** (2016), 1–26. doi: 10.18637/jss.v074.i11.

[16] J.M. Abowd, S. Hawala, B. Ricchetti and M. Stinson, Testing Disclosure Risk in the proposed SIPP-IRS-SSA Public Use Files, Document submitted to the Census Bureau's Disclosure Review Board on November 16, 2006. Available from: https://www2.vrdc.cornell.edu/news/wp-content/uploads/2007/11/drbmemonov2006.pdf.

[17] J. Drechsler, S. Bender and S. Rässler, Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB Establishment Panel, *Transactions on Data Privacy* **1** (2008), 105–130.

[18] J. Hu, J.P. Reiter and Q. Wang, Disclosure risk evaluation for fully synthetic data, in: *Privacy in Statistical Databases*, J. Domingo-Ferrer, ed., Lecture Notes in Computer Science 8744. Heidelberg: Springer, 2014, pp. 185–199.

[19] D. McClure and J.P. Reiter, Assessing disclosure risks for synthetic data with arbitrary intruder knowledge, *Statistical Journal of the International Association for Official Statistics* **32** (2016), 109–126.

[20] M. Elliot, Final report on the disclosure risk associated with the synthetic data produced by the SYLLS team, Report 2015-2, Cathie Marsh Institute for Social Research (CMIST), University of Manchester; 2014 Available from: http://www.cmist.manchester.ac.uk/research/publications/reports.

[21] G.M. Raab, B. Nowok and C. Dibben, Practical data synthesis for large samples, Submitted. Available from: http://arxiv.org/pdf/1409.0217v7.pdf.

[22] L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone, Classification and regression trees, Belmont (CA): Wadsworth, 1984.

[23] J. Drechsler and J.P. Reiter, An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets, *Computational Statistics and Data Analysis* **55** (2011), 3232–3243. doi: 10.1016/j.csda.2011.06.006.

[24] B. Nowok, G.M. Raab, J. Snoke and C. Dibben, Synthpop: Generating synthetic versions of sensitive microdata for statistical disclosure control, **R** package version 1.3-1; 2016. Available from: https://CRAN.R-project.org/package=synthpop.

[25] **R** Core Team. **R**: A language and environment for statistical computing. **R** Foundation for Statistical Computing, Vienna, Austria; 2016. Available from: https://www.R-project.org.

[26] S. van Buuren, K. Groothuis-oudshoorn, Mice: Multivariate imputation by chained equations in **R**, *Journal of Statistical Software* **45** (2011), 1–67. doi: 10.18637/jss.v045.i03.

[27] T. Therneau, B. Atkinson and B. Ripley, Rpart: Recursive partitioning and regression trees, **R** package version 4.1-10; 2015. Available from: https://CRAN.R-project.org/package=rpart.

[28] Council for Social Monitoring. Social Diagnosis 2000–2011: Integrated Database; 2011. Available from: http://www.diagnoza.com/index-en.html.

Appendix – Replication R code for results from the paper

```
library(synthpop)

# observed data set (ods)
vars <- c("sex", "age", "edu",
"placesize", "socprof",
"marital", "income", "ls", "smoke")
ods <- SD2011[, vars]
ods$smoke <-
factor(as.character(ods$smoke)=="YES")
levels(ods$marital)[levels(ods$marital)
%in%
c("LEGALLY SEPARATED",
"DE FACTO SEPARATED")] <- "DIVORCED"
ods$income[ods$income == -8] <- NA

# synthetic data set (sds)
sds.CART <- syn(ods, m = 10, method
= "cart", seed = 293)
sds.para <- syn(ods, m = 10, method
= "parametric", seed = 293)

# Fig. 2
compare(sds.CART, ods, vars =
c("edu", "marital", "income", "age"))

# Fig. 3 (a)
sfit.CART <- glm.synds(smoke ~
sex + age + placesize + edu + socprof +
marital + log(income) + ls,
family = "binomial", sds.CART)
(comp.CART <- compare(sfit.CART, ods))
# Fig. 3 (b)
sfit.para <- glm.synds(smoke ~ sex +
age + placesize + edu + socprof +
marital + log(income) + ls, family =
"binomial", sds.para)
(comp.para <- compare(sfit.para, ods))

# Table 1
tab1 <- cbind("Z gold-standard" =
comp.CART$coef.obs[,"Z"],
"Z CART" = comp.CART$coef.syn[,"Z.syn"],
"Z parametric" =
comp.para$coef.syn[,"Z.syn"])
round(tab1[-1,], 2)}
```

```
# Fig. 4 (a)
sfit2.CART <- glm.synds(smoke ~
sex + age + placesize + edu + socprof +
marital + log(income) + ls + sex:edu,
family = "binomial", sds.CART)
(comp2.CART <- compare(sfit2.CART,
ods, lwd = 1))
# Fig. 4 (b)
sfit2.para <- glm.synds(smoke ~
sex + age + placesize + edu + socprof +
marital + log(income) + ls + sex:edu,
family = "binomial", sds.para)
(comp2.para <- compare(sfit2.para,
ods, lwd = 1))

# Fig. 5
grper <- function(data, var = "smoke",
breaks = c(16, seq(20, 80, 10), Inf)){
data[,"age.group"] <- cut(data$age,
breaks = breaks, right = FALSE)
per <- (tapply(as.logical(data[, var]),
data[, "age.group"],
mean, na.rm = TRUE)) * 100
return(per)}
per.smoke.obs <- grper(ods)
per.smoke.CART <-
apply(sapply(sds.CART$syn, grper),
1, mean)
per.smoke.para <-
apply(sapply(sds.para$syn, grper),
1, mean)
per.smoke <- cbind(per.smoke.obs,
per.smoke.CART, per.smoke.para)
matplot(per.smoke, type = "b",
ylim = c(0, 50), pch = 15:17,
col = 1, las=1, xlab = "Age group",
ylab = "% smokers")
legend("bottom", c("gold standard",
"CART synthesis",
"parametric synthesis"),
lty = 1:3, pch = 15:17, bty = "n")
```