Recognising real people in synthetic microdata: risk mitigation and impact on utility

Beata Nowok, Chris Dibben and Gillian M Raab

Administrative Data Research Centre - Scotland, University of Edinburgh, School of GeoSciences, Drummond Street, Edinburgh EH8 9XP, United Kingdom, {beata.nowok, chris.dibben, gillian.raab}@ed.ac.uk

Abstract. Completely synthetic microdata are generated from probability distributions and as such contain artificial units only. It is possible, however, that synthesising process will produce by chance synthetic individuals who are very similar or even identical to actual people. This has caused concern among data custodians about disclosure risk. Information on unique or rare individuals that are replicated in the synthetic data might be inherently disclosive. Besides, people may 'recognise' themselves in public use synthetic data files and mistakenly believe that data are real not synthetic, with subsequent loss of reputation for the data collection agency. As a result, data custodians in the UK usually request removal of such problematic units from synthetic data. In this paper we explore the scale of the deletion process depending on the number and type of variables used for comparing real and synthetic data and the damage that this disclosure control measure does to the utility of the synthetic data.

1 Introduction

Today, there is little doubt about advantages of sharing individual-level data for research, but a challenge of finding the best way to provide access while preserving confidentiality still exists. In 1993, Rubin argued that we can replace actual data with synthetic ones generated from the models, and data subjects can be assured that "[their] data will only be used to create synthetic data for public-use, and none of [their] data values will ever be released" (Rubin 1993). It is possible, however, that synthesising process will produce records that are identical to the records in the real data, either in the sample that was used for synthesis or in the whole population. This raises concerns among data custodians that such records might be disclosive or that they may create a false belief that a dataset is real and not synthetic. This has led to a number of rules being applied to synthetic data in the UK, one of them being the removal of unique real individuals that are replicated as unique in the synthetic data. The number of unique individuals and the chance of creating their identical copies depends on dataset characteristics and the synthesising method. This paper

uses an empirical example to explore the scale of the sample size reduction of the synthetic data due to such deletion and its impact on the synthetic data quality.

In the next section, we present the data and describe the synthesising method. In Section 3, we investigate the relationship between the number of unique individuals in the real dataset that are also present and unique in the synthetic data. The uniqueness is judged by all or a subset of variables. In the following Section 4 we look at the data quality consequences of suppression of such individuals from synthetic data. Section 5 concludes the paper with some final remarks. All calculations are done in **R** and the analysis code can be obtained on request from the authors.

2 Data synthesis

2.1 Data

We use a subset of freely available survey data collected within the Social Diagnosis project in 2011 (Council for Social Monitoring 2011) which aims to investigate objective and subjective quality of life in Poland. This subset is included in the **R** package *synthpop* (Nowok et al. 2016) as a data frame called `SD2011`. It contains information on a sample of 5,000 individuals aged 16 and over. For our example we have selected ten variables of various types and with different numbers of unique values. Details on these variables are presented in Table 1.

Variable name	Description	Data type	Unique values
sex	Sex	factor	2
smoke	Smoking cigarettes	factor	3
edu	Highest educational qualification	factor	5
marital	Marital status	factor	5
placesize	Category of the place of residence	factor	6
ls	Perception of life as a whole	factor	8
socprof	Socio-economic status	factor	10
age	Age	numeric	79
income	Personal monthly net income	numeric	406
bmi	Body mass index	numeric	1395

Table 1: Variables included in the dataset to be synthesised.

2.2 Synthesis

The data were synthesised with the **R** package *synthpop* using the classification and regression trees (CART) method (Breiman et al. 1984) in a series of conditional models (Reiter 2005). All values of all variables in our dataset were replaced. Missing values were not imputed prior to the synthesis and missing data were produced in the

synthetic datasets. The synthesising order was as follows: *sex*, *age*, *edu*, *placesize*, *socprof*, *marital*, *income*, *ls*, *smoke* and *bmi*.

Three sets of 100 synthetic versions of the original data were produced. The difference between these sets refers to the smoothing of the three numeric variables present in the data (*age*, *income*, *bmi*). In the first set (syn_1) no smoothing was applied. In the second set (syn_2) numeric variables were smoothed during synthesis using kernel density method and their smoothed values were used to derive models for generating synthetic values of variables that were later in the synthesising order. In the last set (syn_3) variables were smoothed after the synthesising process.

3 Replication analysis

Motivated by the requirement imposed by data custodians in the UK, we consider the case when unique individuals in a synthetic dataset that correspond exactly to unique individuals in the original data are removed from the former to enhance data protection. This is a very narrow definition of replications and in practice data holders may sometimes want to apply a broader one that would lead to more extensive removal of records and greater impact on data characteristics. It means, therefore, that we take a conservative approach to evaluation of data quality changes due to unit suppression.

There are ten variables in our dataset but in order to get a more general overview of the relationship between a number of unique individuals in the observed data and a number of their unique replications in the synthetic data we examine uniqueness for various subsets of variables (hereinafter, they are called key variables or key sets). We consider all possible combinations of the variables for all possible set sizes (between 1 and 10). There are 1,023 such non-empty combinations ($2^{10} - 1$). All calculations are made using *sdc()* function from the *synthpop* package and its *exclude* parameter is used to control variables for comparison.

The number of unique records in the original sample of 5,000 individuals ranges between zero (when only one of the categorical variable is taken into account or some two-variable combinations with binary variable *sex*) and 4,992 (when all ten variables are considered or 13 other combinations that include all three numeric variables). Figure 1 displays proportion of unique individuals in the original data and an average proportion of unique replications in 100 synthetic datasets for syn_1 for various sets of key variables. Results are ordered by the number of unique records in the original dataset (the same order is used for all Figures in this section). Due to the large number of possible combination of variables their labels are not shown but a full ordered list can be provided on request. As intuitively expected, the greater the number of possible combination of values, which is related to some extent to the number of variables, the greater the number of unique units in the data. What is more interesting to note is that the number of replications in the synthetic data

does not follow the increasing trend in the number of unique units in the observed data. The percentage of the replicates is determined mostly by the presence of numeric variables in the key variables. It reaches the highest levels when *bmi* or *bmi* and *age* are included (up to 17%). When all three numeric variables are used, the exact replication of their combinations becomes less likely and the proportion of replications can be lower than for categorical variables only (between 1.2-5%).

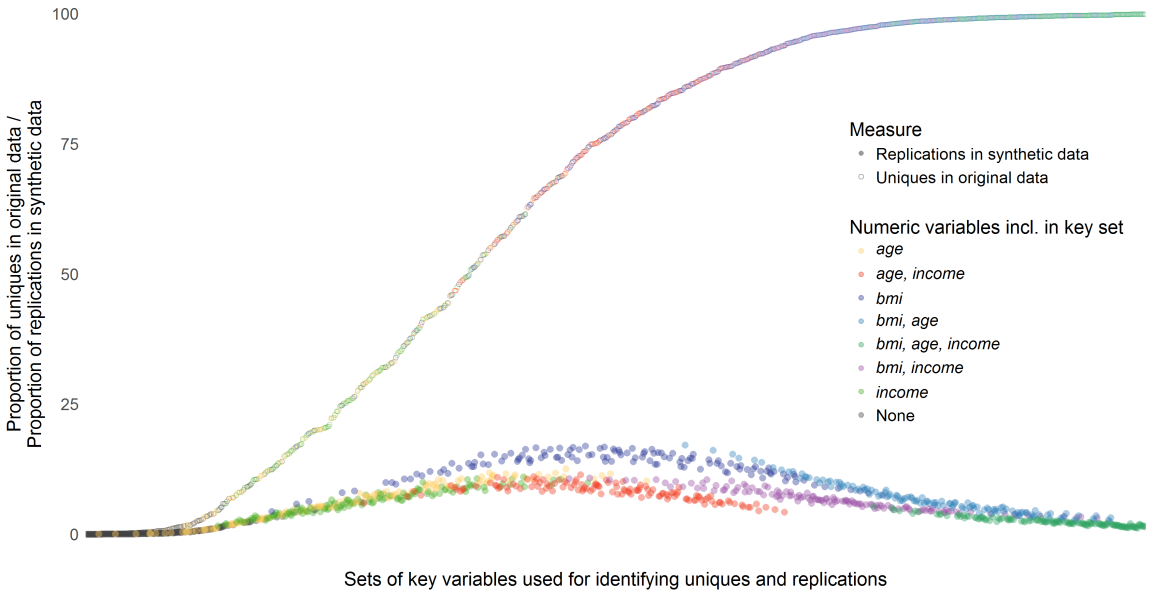

Figure 1: Proportion of unique individuals in the original dataset and average proportion of their unique replicates in the synthetic data for various sets of key variables.

These results are sample size dependent. There is a finite number of all possible combinations of values (except when continuous variables are measured with high precision) and the number of unique individuals decreases with the increasing sample size. In other words, it is more difficult to be unique in a whole population than in a small sample survey. As regards replications, with larger sample size the chance of replicating a unique unit increases but so does the chance of replicating a synthetic unique. Figure 2 provides an illustration for population of 500,000 individuals. A hundred synthetic copies of our initial real data of 5,000 units were combined to form our new original data that were then resynthesised.

The relatively high level of replications for the case when numeric variables are taken into account results from the fact that synthesis with the CART method samples synthetic values from the real ones that are classified to a specific node of a tree. Figure 3 shows impact of smoothing of the numeric values, which would be usually

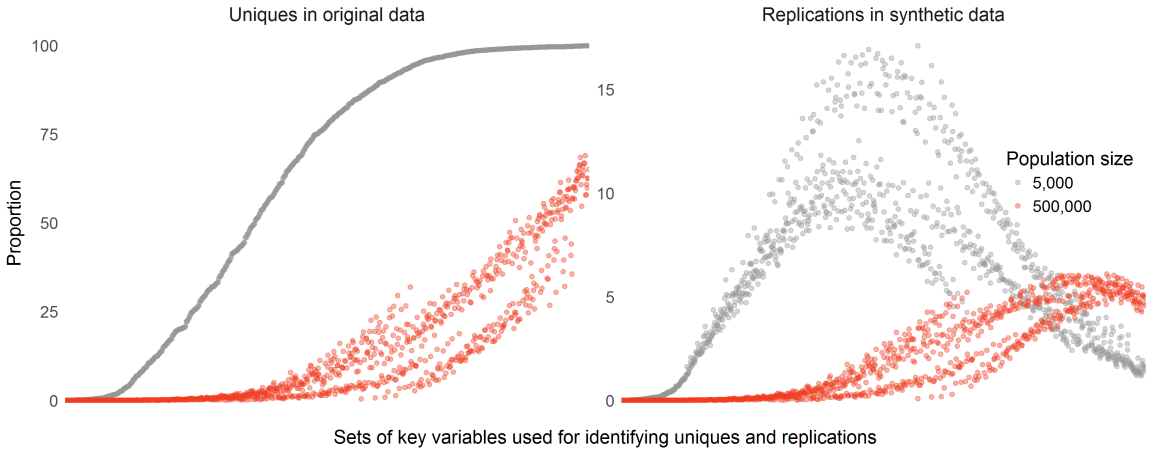

Figure 2: Proportion of unique individuals in the original data (left panel) and their replications in synthetic ones (right panel) for different population size.

applied in practice, on the number of replications. As illustrated and expected, the number of replications decrease substantially for both strategies of smoothing (see Section 2.2) with the exception for smoothing of *age* when its smoothed values are used to generate other synthetic variables (*bmi* is never used as a predictor because it is the last variable synthesised).

4 Impact on data quality

In evaluating impact on data quality we address the question whether suppression of selected replicated records from synthetic data (syn_1) distorts data distributions. The question is especially valid for cases with the highest proportion of identified and removed replications, but with practicality in mind we focus on cases when *income* and *bmi* are not included in the key variable sets (26 cases with proportion of replications exceeding 10%). Exact unsmoothed values of such variables would probably never be released. However, for comparative purposes, we analyse the extreme cases too (16 cases with proportion of replications greater than 16%). The complete lists of considered variable combinations can be found in the Appendix A.

The comparison of frequency tables for the original data, the complete synthetic data and the reduced synthetic data (without replications) does not reveal any serious distortions in the last case. Note that synthetic counts were averaged over one hundred versions of synthetic data and a greater impact may be observed for a single synthetic version. We can expect that removing unique individuals from a dataset would affect distributions to a greater degree but the condition that they have to be replications of unique individuals prevents the extensive removal of less

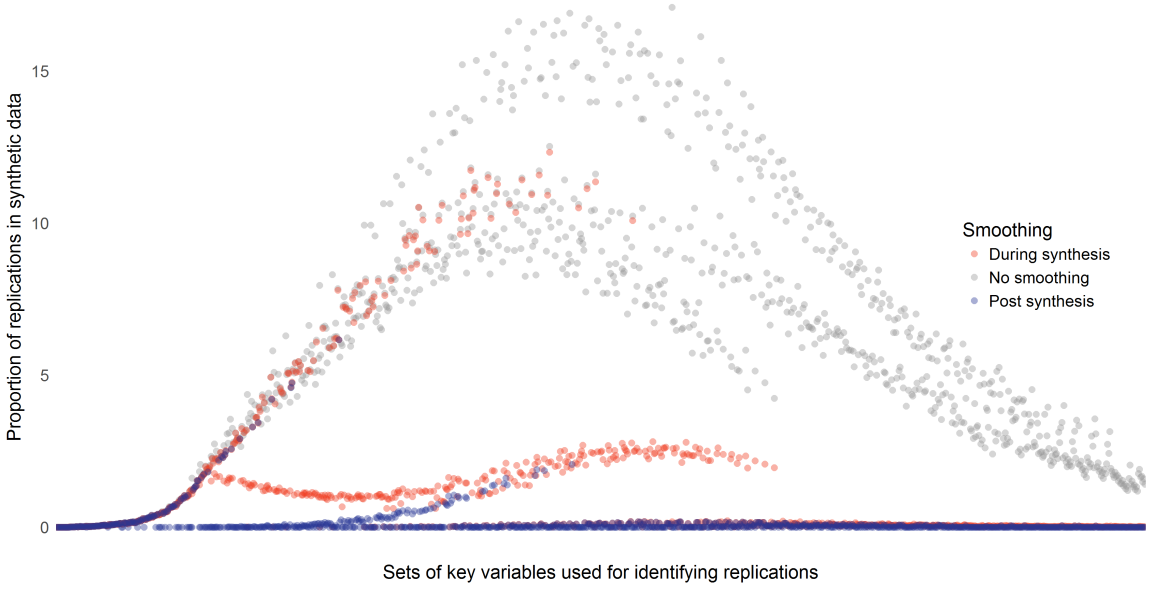

Figure 3: Average proportion of replicates of unique individuals in smoothed synthetic data for various sets of key variables.

frequent value combinations.

It is worth pointing out that in our example, the impact of smoothing during synthesis is greater than removal of replicated individuals (see Figure 4 for an example of frequency distributions of *age* variable), which may result, however from the method of smoothing that was applied.

To investigate further whether multivariate relationships are preserved in the reduced synthetic datasets we modelled the factors that affect smoking by logistic regression. The details of the model specification can be found in Nowok et al. (2016) and methods of inference from synthetic data are described in Raab et al. (2017). Differences are negligible and in most cases exactly the same conclusions regarding impact of various factors on chance of smoking can be drown based on complete and reduced synthetic datasets.

5 Concluding remarks

The analysis presented in this paper is only specific to this example and suppression of unique individuals replicated in the synthetic data may have a different impact on the data quality depending on the characteristics of the original dataset and the synthesising method used. Therefore, careful investigation of various options is recommended and tools available in the **R** *synthpop* package can facilitate the process.

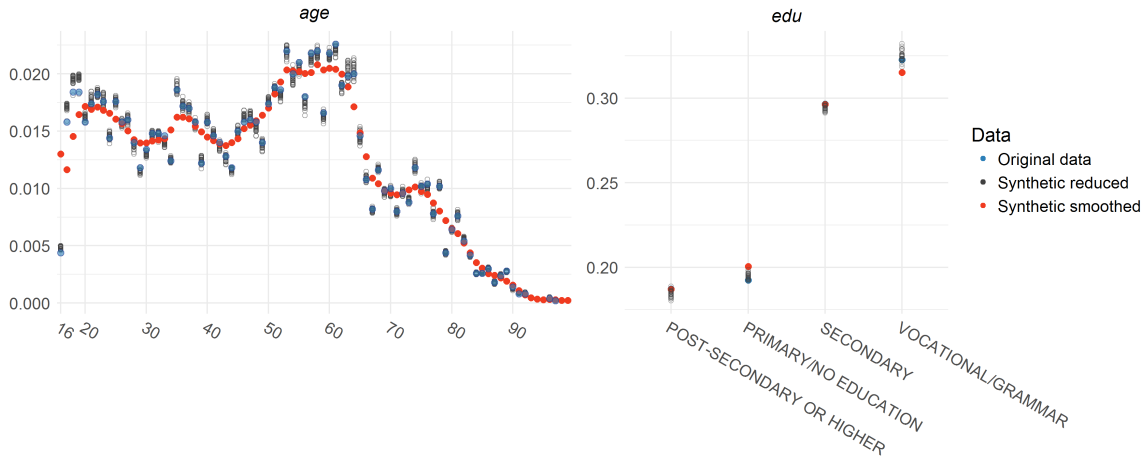

Figure 4: Frequency distribution of *age* and *edu*.

Nonetheless, no serious data damage caused by removal of replicated uniques was identified in the example considered in this paper. However, note that, as mentioned in Section 3, our definition of replication is very narrow and data custodians may request removal of not only unique replicates of unique real individuals but also replicated rare cases that are rare in the synthetic data too, which may lead to more pronounced changes in data quality.

References

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984) *Classification and Regression Trees*. Belmont, CA: Wadsworth.

Council for Social Monitoring (2011) Social Diagnosis 2000–2011: Integrated Database. Available at: `http://www.diagnoza.com/index-en.html`.

Nowok, B., Raab, G.M. and Dibben, C. (2016) synthpop: Bespoke creation of synthetic data in **R**. *Journal of Statistical Software*, **74**(11), 1–26. doi:10.18637/jss.v074.i11

Nowok, B., Raab, G.M. and Dibben, C. (2016) Providing bespoke synthetic data for the UK Longitudinal Studies and other sensitive data with the synthpop package for **R**. *Statistical Journal of the IAOS*, **Preprint**, 1–12. doi:10.3233/SJI-150153

R Core Team (2017) ***R****: A language and environment for statistical computing*,

R Foundation for Statistical Computing, Vienna, Austria, `https://www.R-project.org`.

Raab, G.M., Nowok, B. and Dibben, C. (2017) Practical data synthesis for large samples. *Journal of Privacy and Confidentiality*, 7(3):67–97. Available at: `http://repository.cmu.edu/jpc/vol7/iss3/4`

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

Rubin, D.B. (1993) Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, **9**(2), 461–468.

A Appendix - List of variable combinations with the highest level of replications

Percentage of replications exceeding 10% (*income* and *bmi* are not used to identify them).

$sex - age - edu - placesize - socprof - ls$
$sex - age - edu - placesize - socprof - smoke$
$age - edu - placesize - socprof - ls - smoke$
$sex - age - edu - placesize - socprof - ls - smoke$
$age - edu - placesize - socprof - ls$
$sex - age - edu - placesize - socprof - marital - smoke$
$sex - age - placesize - socprof - ls - smoke$
$sex - age - edu - placesize - ls - smoke$
$sex - age - edu - placesize - socprof - marital - ls$
$sex - age - edu - placesize - socprof - marital$
$sex - age - placesize - socprof - marital - ls - smoke$
$sex - age - edu - socprof - ls - smoke$
$sex - age - placesize - socprof - marital - ls$
$age - edu - placesize - socprof - marital - ls$
$sex - age - edu - socprof - marital - ls - smoke$
$sex - age - placesize - socprof - ls$
$age - edu - placesize - socprof - marital - ls - smoke$
$sex - age - edu - placesize - socprof$
$sex - age - edu - placesize - marital - ls - smoke$
$sex - age - edu - placesize - marital - ls$
$sex - age - edu - socprof - marital - ls$
$age - placesize - socprof - marital - ls - smoke$
$sex - age - edu - placesize - ls$
$age - placesize - socprof - ls - smoke$
$sex - age - edu - placesize - socprof - marital - ls - smoke$
$age - edu - placesize - socprof - marital - smoke$

Percentage of replications exceeding 16% (no restrictions on combinations content).

$age - bmi$
$sex - edu - socprof - bmi$
$edu - socprof - smoke - bmi$
$sex - socprof - marital - smoke - bmi$
$sex - socprof - marital - bmi$
$edu - socprof - bmi$
$sex - socprof - ls - bmi$
$sex - edu - socprof - smoke - bmi$
$sex - placesize - socprof - bmi$
$sex - socprof - smoke - bmi$
$socprof - marital - smoke - bmi$
$sex - edu - socprof - marital - bmi$
$placesize - socprof - bmi$
$edu - socprof - marital - bmi$
$socprof - ls - bmi$
$sex - age - bmi$