Introduction to Synthetic Data produced with the **synthpop** package.

Gillian M. Raab (gillian.raab@ed.ac.uk)

1 Introduction and summary

This is a very brief summary of our experience in producing synthetic data for users of potentially disclosive data for exploratory analyses. We [1] have developed the **synthpop** package [2] for the **R** program that is used by staff of the Scottish Longitudinal Study (SLS) to create synthetic extracts that can be released to users outside the safe setting at Ladywell House. This methodology has also been used to create data sets for teaching.

The first suggestion that synthetic data might be used for disclosure control was made 25 years ago [4]. The first papers showing how it can be done appeared around 10 years later (e.g [3]) and an extensive literature has developed since then. Some synthetic data products have been developed by various agencies, most notably by the US Bureau of the Census (e.g. [1]) The original proposal was that the synthetic data could be used in place of the original data in published reports, but it is now acknowledged that a final analysis for publication should always be carried out on the original confidential data.

The process of creating a synthetic data set from an original data set (we will refer to this as "real data" here) is not simple. Starting from scratch it requires a large amount of programming effort to produce and evaluate a single synthetic data set. This was the motivation for the creation of **synthpop** as a set of tools for creating and evaluating synthetic data, freely available to any R user. The synthetic data must have good utility (U) in that it must look like the real thing and analyses based on the synthetic data must lead to the same conclusions as those from the real data. It must also pose a low risk (R) of the disclosure of confidential information. We discuss here how the tools in **synthpop** control and evaluate R and U.

In Section 2 we introduce synthetic data by a highly simplified, completely artificial, example that illustrates how it can deliver high U and low R in ideal circumstances. In sections 3 and 4 we expand on each of U and R.

In later sections we do not include references to either the background literature or to our own published work, but a bibliography could be supplied to anyone who would like to follow up in more detail.

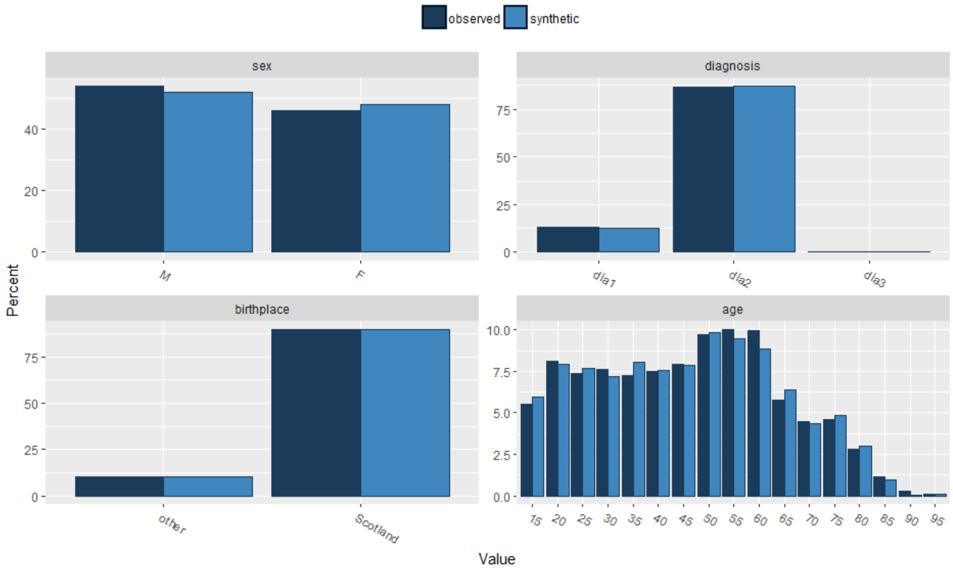

Figure 1: Comparison of real (black) and synthetic (blue or grey) data .

[1] The **synthpop** team consists of Beata Nowok, Gillian Raab and Chris Dibben

2 What are synthetic data?

The production of synthetic data can be considered as first fitting a statistical model (the signal) to the joint distribution of the real data as well as modelling the deviations from this model (the noise). The synthetic data consists of the signal plus a new random sample from the noise. Our artificial example consists of 2055 records with 4 items: sex, birthplace, diagnosis (1-3) and age. A synthetic data set of the same size as the real data[2] was produced using **synthpop** [3]. Figure 1 compares the real and synthetic data for each variable, giving our first test of U. We can see that the distributions are similar but not identical. We first focus on two aspects that might pose disclosure risks (R). The first is the small numbers of people with diagnosis 3 (top right panel in Figure 1) and the second is the small numbers of people over 95.

The real data had just 6 people with diagnosis 3 and this particular synthetic data set has 8. What information does this provide to the recipient of the synthetic data who does not know the number in the original? We must start with an assumption about the recipient's prior knowledge of how many diagnoses 3 are in the real data. One way of representing their ignorance is to say that they would consider any number between 1 and 20 equally likely. This is represented in Figure 2. Now, if the recipient knows how the synthetic data have been generated [4] then they can update their knowledge about the real data from what they have learned from the synthetic data. The resulting increase in knowledge is shown in Figure 3. We can see that, as expected, the actual value in the synthetic data is the most likely value, but other values in the range 4 to 14 are also quite likely, including, of course, the

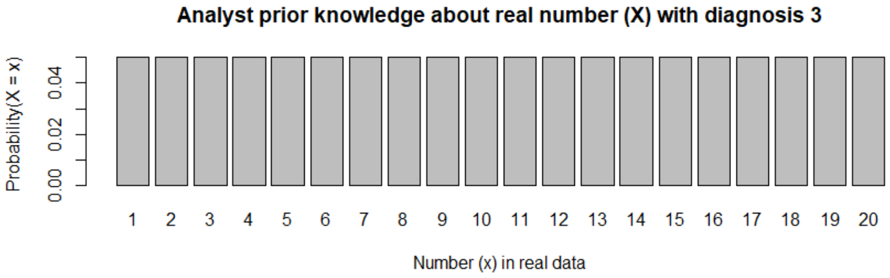

Figure 2: Prior distribution.

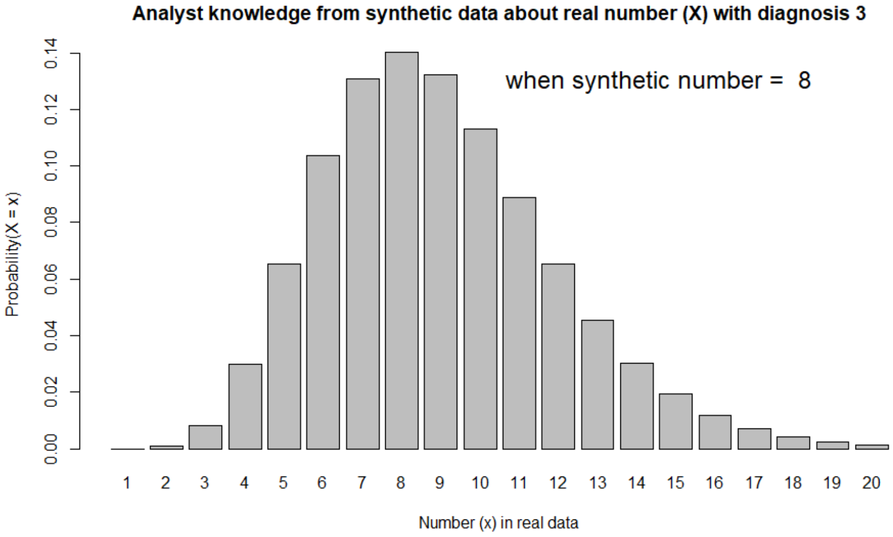

Figure 3: Posterior distribution , 95% interval 4-14 inclusive.

[2] This does not have to be the case in **synthpop**

[3] The first 3 variables were synthesised from a model that included all interactions between them and age was synthesised from a CART model depending on he first three

[4] This is consider good practice in the disclosure literature.

true value of 6 which the recipient does not know. Thus we have not revealed the true number to the recipient that might be used to find individuals, but only a range of possible numbers. Other synthetic data sets might have different numbers of diagnosis 3 but the pattern would be similar. Turning now to our second potential R. Each of the actual and synthetic data sets had exactly two people aged over 95. In the initial synthesis two individuals with these ages (97 years and 5 months and 96 years and 10 months) appeared in both the real and synthetic data[5] and so might potentially be disclosive of someone's identity. When the **synthpop** option *smoothing* is selected for age, these values become 96 years two months and 98 years 4 months. The statistical disclosure control (SDC) measures described in section 4 below can also reduce the disclosure potential of such records.

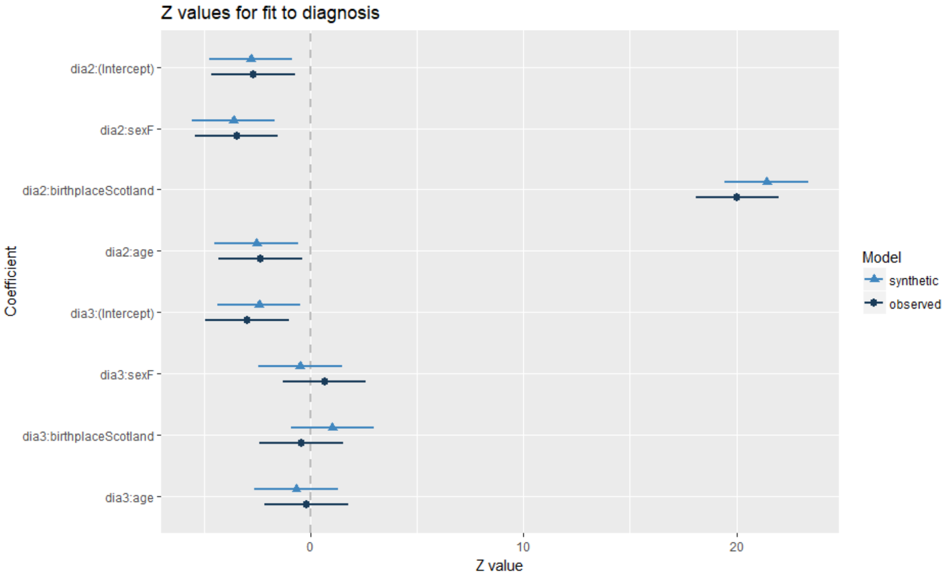

Figure 4: Coefficients with 95% intervals for a multinomial model used to investigate factors affecting diagnosis. Real data results (black lines) are compared to those from the synthetic data (blue/grey)

Finally we return to Utility, looking beyond the univariate summaries in Figure 1 to relatonships between variables. An analysis of factors influencing diagnosis was carried out with the synthetic data. Someone who also has access to the real data can compare the results from the two sources. A plot comparing the coefficients from the two analyses is shown in Figure 4. Both analyses tell the same story. Diagnosis 2, compared to the baseline diagnosis 1, is much commoner in those born in Scotland and less common among women and younger people. The numbers of diagnosis 3 are too small for the influence of any factors to be identified.

3 How useful (U) are synthetic data?

To make data useful the person creating them must be familiar with the data source. This allows the synthesis to be tailored to a particular data set using settings in **synthpop**. Logical constraints must be retained[6] and choices of method must be appropriate for each variable[7]. We have derived general and specific utility measures, as well asmeasures for tables, that can be used to evaluate the quality of synthetic data. They allow people producing synthetic data to diagnose which relationships in the data are causing a problem so that alternative synthesis methods can be tried. As well as the method used for each variable, **synthpop** allows the ordering of variables to be changed and synthesis to be carried out within strata defined by variables likely to be important to the recipient.

The creation of good synthetic data is as much an art as a science, and one that can improve with practice. But even the most experiences person, with all the tools available, cannot give a 100 % guarantee to have reproduced every relationship between variables present large complex data sets. Thus we recommend very strongly that no analysis from synthetic is published as if it were real.

[5] The synthetic values produced by CART are selected from those that appear in tha actual data. This is not true for all methods.

[6] Examples are that mothers must be female and older than their children by a specified amount

[7] One example is a categorical variable, like detailed diagnosis, with a large number of categories. We have shown how these can be synthesised by using broader groupings

4 What is the disclosure risk (R) from synthetic data?

Disclosure risk can be subdivided into identification risk (learning that a unit is in the data and, possibly, also identifying their record) and attribution risk (learning something new about an individual). These risks do not just depend on formal re-identification probabilities, as discussed in Section 2, but also the circumstances in which the data are released and the information about the data that is available to the recipient. For completely synthetic data there is no one-to-one correspondence between the real and synthetic data, however the synthetic data may produce records where values suggest an identification with a known individual or unit, leading to a false identification and potentially a wrong attribution. Thus it is important that anyone who sees synthetic data knows that it is not real. To ensure this, the output module from **synthpop** adds labels to the data to indicate that the records are false.

Attribution risk from real data can take place when an individual is identified from a unique combination of certain variables that reveal the value of other attributes. Such exact attribution is not possible for data known to be synthetic. However, the synthetic data can be used to obtain posterior distributions for the value of one variable, given the others, in the same manner as was done for the number of individuals with diagnosis 3 in the example in Section 2. Such probability calculations could provide partial information about particular individuals. This approach has been investigated for some simple cases and for a few real examples. The results have been fairly reassuring, showing that for most individual units the posteriors from the synthetic data will provide little additional information for most units. Furthermore, to obtain such information would require that someone with access to the synthetic data and partial information about a member of the real data had the skills and motivation to make such calculations: which seems an unlikely scenario.

A more realistic disclosure risk is that a certain individual, identified by their attributes from the synthetic data, is identified as present in the real data when this information, by itself, would be a disclosure risk. The oldest individuals in the unsmoothed synthesis described above would be an example. While smoothing can overcome this a risk may still exist if some information about an outlier in the real data is available. The SDC module in **synthpop** allows top and bottom coding to exclude outliers. Another example would be a very rare medical condition. Data with many very small categories is best dealt with by aggregating to larger groups. If the user wishes to see the range of such categories in the real data, the detailed categories can be provided from bootstrap samples and the user is informed of what has been done. Units which are unique in the real data which also appear as unique units in the synthetic data can also be excluded from by the SDC module.

The method of *differential privacy* (DP) adds random noise to the results of a query so as to ensure that the probability of a result will not be changed by more than a specified factor if any one unit is or is not in the data. The original DP method does not apply to the release of micro-data because the amount of noise added is a function of the query and data items may contribute to many queries. Several recent papers have attempted to generate synthetic data that obeys conditions similar to DP. These attempts have all shown that such synthetic data either provides very weak privacy protection (probability ratios only within very wide bounds) or have such poor utility that the analyses will make no sense. The disclosure risk reduction provided by DP and these methods makes no use of information that the data recipient holds about individuals in the data base. Thus it must protect against any such information, and so may be a stronger criterion than is realistic.

These formal definitions of disclosure risk do not measure the harm that may result from R. Such harm depends crucially on to whom and with what conditions data are released. Both for the SLS and for teaching data sets procedures are in place to ensure that those receiving microdata understand synthetic data and are aware of confidentiality constraints.

References

[1] Kinney, S. K., and Others. Towards unrestricted public use business microdata: The synthetic longitudinal business database. *International Statistical Review 79* (2011), 363 – 384.

[2] Nowok, B., Raab, G. M., and Dibben, C. synthpop : Bespoke creation of synthetic data in R. *Journal of statistical software 74* (2016), 1–26.

[3] Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics 19* (2003), 1–17.

[4] Rubin, D. B. Discussion: Statistical disclosure limitation. *Journal of Official Statistics 9* (1993), 462–468.