A New Method of Generating Synthetic Data for Public Use Files

Inês Ribeiro Viana

FEP, Faculdade de Economia do Porto, inessrviana@hotmail.com, Porto, Portugal

*Pedro Campos**

LIAAD/INESC TEC e FEP, Universidade do Porto, Porto, Portugal, pcampos@fep.up.pt

Keywords: Public use Files (PUF), Synthetic Data, Statistical Disclosure Control

Abstract

The availability of statistical information in PUF (Public Use Files), which can be accessed by every citizen, endangers the protection of individual respondents. To ensure compliance to confidentiality, there are a set of statistical disclosure control methods that can be applied to the data sets before they are published, aimed at reducing the risk of individual identification. In this paper, we present a new method (MEAC - Minimum Absolute Error of Variable Combinations) to generate synthetic data in microdata files.

1. Introduction

The search of statistical information databases for different purposes (such as research, teaching, marketing, etc.), has led to a growing concern about the respect for privacy of respondents. In many cases, confidentiality issues constitute a barrier for the access to microdata, because it is possible to identity sensitive individual data that may relate to individuals, families or companies. Public Use Files (PUF) are part of the various solutions that have been found to circumvent access to microdata. These are files that can be formed by microdata (records contaiing individual information about respondents), but that do not put at risk the privacy of respondents.

2. Synthetic Data

Recently, several methods have appeared in the literature, for the production of public-use files, such as the synthetic data production methods (Nowok et al, 2014;. Dreschler and Reiter, 2010; Reiter 2005 Dreschler, 2011), with applications to official statistics (eg, Templ and Alfons, 2010). Synthetic data mimic the actual data and preserve the statistical relationship between the variables, but in some cases these files may contain a small percentage of records that can identify individual respondents.

In this paper we propose a method to generate synthetic data based on the calculation of conditional probabilities. The presented method aims to generate a fully synthetic micro file (X *) from an original micro file X, K observations and N-containing variables (each of which the variables have attributes kj). The variables can be quantitative or qualitative.

3. Method of Conditional Probabilities with Error Minimization (MEAC)

The generation of data takes into account a particular order in the choice of variables. First variables are generated directly and each of the subsequent variables is generated taking into account the previously generated variables and all possible combinations that may be established between them. The innovation of this method is that the combination is chosen which gives the lowest absolute error between the observations of the original data and the comments generated, ie:

$$x^{*(p)}_{k_j^{(1)}+k_j^{(2)}+\cdots+k_j^p} = P\left(x^{(p)}_{k_j^{(1)}+k_j^{(2)}+\cdots+k_j^p} \,/\, \min_h\left(\left|x_h^{(p)} - x_h^{*(p)}\right|\right)\right) * \; x^{*(p-1)}_{k_j^{(1)}+k_j^{(2)}+\cdots+k_j^{(p-1)}}$$

being:

. $x^{*(p)}_{k_j^{(1)}+k_j^{(2)}+\cdots+k_j^p}$ –number of observations generated for attribute j of the p[th] variable being generated, combining all the attributes previously generated.

. $x^{*(p-1)}_{k_j^{(1)}+k_j^{(2)}+\cdots+k_j^p}$ –number of observations generated in (p-1), verifying the combination of the several attributes previously generated. (3.2)

. $x^{(p)}_{k_j^{(1)}+k_j^{(2)}+\cdots+k_j^p}$ – number of observations in the original data file that verify that combination of attributes of the variables previously generated, according to attribute j of the p[th] variable being generated. (3.3)

. h - number of combinations of variables (3.4)

. $\min_h\left(\left|x_h^{(p)} - x_h^{*(p)}\right|\right)$ – minimum of the absolute erros between the number of observations of the priginal data set and the number of observations of the synthetic (generated) data set, in the case where the number h of combination of variables has been chosen.

This method proves to be suitable and easy to use. A comparison of the performance of this method and other methods of generation of synthetic data is also available in Table 1.

	Original data set	MEAC	Parametric	Sample	Passive	CTREE
Combinations	2432	2431	2583	3101	2551	2551
Critical combinations	1337	1297	1454	1551	1432	1432
New Combinations	-	0	878	1840	839	839
Deleted Combinations	-	1	727	1171	720	720
Critical Combinations unchanged	-	1297	482	217	486	486
Critical combinations deleted (1)	-	1	681	885	680	680
Critical combinations deleted (2)	-	39	174	235	171	171
New critical combinations (1)	-	0	145	135	164	164
New critical combinations (2)	-	0	827	1199	782	782

Table 1 – Comparisons of methods, in original and generated variable combinations

References

Drechsler J (2011). Synthetic Data Sets for Statistical Disclosure Control. Springer, New York.

Dreschler J, Reiter JP (2010). Sampling with synthesis: a new approach for releasing public use census midrodata" *Journal of the American Statistical Association*, 105(492), 1347– 1357.

Nowok, B., Gillian, M., Dibben, C., (2014), synthpop : Bespoke creation of synthetic data in R, accessed in 5/9/2014 and available at:
http://www.qub.ac.uk/research-centres/NILSResearchSupportUnit/FileStore/Filetoupload,431239,en.pdf

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

Templ, M., Alfons, A., (2010), Disclosure Risk of Synthetic Population Data with Application in the Case of EU-SILC, *Privacy in Statistical Databases*, *Lecture Notes in Computer Science*, Volume 6344, 2010, pp 174-186