Selection of values to generate multiply imputed partially synthetic data for protecting confidentiality

Partially synthetic data is an increasingly popular approach used to protect the confidentiality of survey microdata. In this approach, the original units surveyed remain on file, but, prior to release, the data-holders replace certain values in the data with multiple imputations that are drawn from a plausible statistical model. As the released data sets now contain a mix of original and synthetic values, confidentiality has been protect to an extent. In addition, provided a plausible model has been used to generate the synthetic values the statistical properties present in the original data should be preserved in the released data. An under investigated area with this approach concerns the question of deciding which values in the microdata should be replaced with synthetic values. Typically entire key variable sets, which are variables that could lead to identification of units, are synthesised, and while this facilitates implementation of this approach, such an extreme implementation may be unnecessary to protect confidentiality. In this article, we consider an approach that identifies values amongst the key variables that distinguish a record from others in the sample, and hence are the values that need to be replaced. This leads to a framework that allows us to determine which values in the data set need to be synthesized. We illustrate the performance of this procedure to protect confidentiality in the Current Population Survey, and show that with a relatively small amount of synthesis there is a substantial reduction of confidentiality risk when adopting this selection strategy. Often the pattern of synthesis is non-monotone which complicates the imputation of standard parametric imputation procedures. We adapt the imputation model proposed in Lee and Mitra (2013) to generate synthetic values arising from a such pattern, the imputation model is further refined to handle point mass distributions present in the CPS.

Keywords: Bayesian statistics; Data confidentiality; Multiple imputation; Synthetic data