Jörg Drechsler

TITLE: Data Utility vs. Disclosure Risk for Multiply Imputed Synthetic Datasets

ABSTRACT: Arguably fully synthetic datasets provide a very high level of disclosure protection since all released values are imputed values and if necessary the released data can be restricted to only those units that didn't participate in the survey. However, guaranteeing confidentiality is only one side of the medal and the best protection method is useless if the provided dataset can not guarantee a high level of data quality as well. Since all values in the dataset have to be imputed for fully synthetic datasets, the quality of the released data will strongly depend on the quality of the underlying model. To overcome this strong model dependency partially synthetic datasets replace only values at high risk with synthetic values for the price of a higher disclosure risk. Since true values remain in the dataset a careful disclosure risk evaluation is necessary in this context. Especially the number of imputations has to be chosen carefully since it has an impact on both, the data utility and the disclosure risk.
In this talk, we will compare the benefits and weaknesses of the two imputation approaches and discuss the impact of the number of imputations on the trade-off between data utility and disclosure risk using a genuine dataset: The IAB-Establishment Survey. We also propose a two stage imputation method to address this trade-off.