Natalie Shlomo

TITLE: Releasing Microdata: Disclosure Risk Estimation, Data Masking and Assessing Utility

ABSTRACT: Statistical Agencies generally release sample microdata from social surveys according to different modes of access. The access methods range from “public use” files in the form of tables or highly perturbed datasets to “microdata under contract” for researchers and licensed institutions where data protection methods are less severe. Microdata Review Panels need to make informed decisions when releasing microdata and it is vital that they are able to assess correctly and objectively the disclosure risk.

The microdata contain individuals investigated in a sample where the population is unknown (or only partially known through some marginal distributions). The disclosure risk is a function of both the population and the sample, and in particular the cell counts for a contingency table defined by combinations of identifying discrete key variables, i.e. region, sex, age, occupation, etc. The disclosure risk is assessed using probabilistic models to estimate per-record disclosure risk measures based on the probability that a sample unique in a cell of the contingency table can be re-identified. Per-record risk measures can be used to target high-risk records for Statistical Disclosure Limitation (SDL) techniques and reduce information loss. Consistent global file-level disclosure risk measures are aggregated from the per-record risk measures and are particularly useful to Microdata Review Panels for determining tolerable risk thresholds and the amount of protection needed in the data according to the mode of access.

Based on the disclosure risk assessment, data custodians must choose appropriate SDL methods either by perturbing, modifying, or summarizing the data. The choice of the SDL method depends on the requirements of the users, the contents of the data and the impact on quality and information loss. Choosing an optimal SDL strategy is an iterative process where data custodians need to examine the trade-off between managing disclosure risk and obtaining high quality microdata. Examples of SDL methods for microdata include perturbative methods such as record swapping or a more general post-randomization method for categorical variables and additive noise for continuous variables, or non-perturbative methods such as recoding and coarsening variables or sub-sampling. Each of the methods impact differently with respect to information loss in the microdata and they should be combined and optimized to best preserve the properties of the data. Information loss measures quantify the effects of SDL methods on bias and variance and the impact on statistical analysis.