Providing valid analytic properties in a synthetic version of confidential discrete data while nearly eliminating re-identification risk.

William E. Winkler, U.S. Census Bureau

Winkler (2003, 2008) provides theory for EM modeling of data to assure that all records satisfy edits and missing items are imputed, that aggregates from the data files preserve joint distributions, and, if necessary, that various aggregates can be scaled (using suitable convex constraints) to aggregates from external sources. By putting additional convex constraints on aggregates in the model (Winkler 2010), we can assure that any originally small cells are sampled with very small probability while much of their mass is dispersed across originally sampling zero cells in a manner that preserves analytic properties. While the methods are not differentially private, they are much easier to apply, much better preserve analytically properties, and may yield as low or even lower re-identification risk in number of situations. Features in the new generalized software are much larger data structures and tools for evaluating a large set of margins while maintaining extreme computational speed (500,000 cells with epsilon <10^-12 in 200 iterations in less than 60 seconds; 500 million cells in 1000 minutes). The methods are suitable for both cleaning up very large administrative files (100,000,000+ records) and creating synthetic copies that may be more easily shared among agencies.

Keywords: EM fitting, edit restraints, convex constraints, loglinear models, privacy