Working Paper No. 27 (Summary)
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint ECE/Eurostat work session on statistical data confidentiality
(Luxembourg, 7-9 April 2003)

Topic (i): New theories and emerging methods

NEW DEVELOPMENTS IN PERTURBING NUMERICAL MICRODATA

Contributed Paper

Submitted by the University of Kentucky and Oklahoma State University, United States[1]

[1] Prepared by Krish Muralidhar (krishm@uky.edu) and Rathindra Sarathy (sarathy@okstate.edu).

New Developments in Perturbing Numerical Microdata

Krish Muralidhar
Gatton Research Professor
Gatton College of Business & Economics
University of Kentucky
Lexington KY 40506
krishm@uky.edu

Rathindra Sarathy
Associate Professor
College of Business Administration
Oklahoma State University
Stillwater OK 74078
sarathy@okstate.edu

Abstract

Recently, we have developed several new techniques for perturbing numerical confidential data. These techniques include:

(1) General Additive Data Perturbation - A procedure that is capable of generating perturbed values such that all linear relationships between variables (both confidential and non-confidential) are maintained to be the same before and after perturbation. For variables that can be described by a multivariate normal distribution, this method provides very high data utility and very low disclosure risk. All previous methods of additive noise perturbation can be shown to be special cases of this method.

(2) Copula Based General Additive Data Perturbation - A procedure that is capable of generating perturbed values such that the marginal distribution and (linear and monotonic non-linear) relationships between variables is maintained to be the same before and after perturbation. This procedure also assures that the risk of disclosure is low.

(3) Data Shuffling – A procedure that combines the advantages of both perturbation and swapping in that it provides the same benefits of perturbation methods, but instead of modifying the values (like perturbation), it uses the original data (like swapping). However, unlike swapping, there is no direct exchange of values between records. The values are assigned between records so that the shuffled data preserves linear and monotonic non-linear relationships between variables and also provide protection against risk of disclosure. This approach also provides the missing theoretical link between perturbation and swapping.

In all of the above cases, it can be shown that the implementation of these techniques is straight-forward. Empirical evaluation indicates that these techniques fulfill their theoretical promise of high data utility and low disclosure risk.