

Working Paper No. 4 (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

BALANCING DATA QUALITY AND CONFIDENTIALITY FOR TABULAR DATA

Invited Paper

Submitted by the U.S. National Center for Health Statistics and OptTek Systems, Inc. (United States)¹

¹ Prepared by Lawrence H. Cox (lcox@cdc.gov) and James P. Kelly (Kelly@opttek.com).

BALANCING DATA QUALITY AND CONFIDENTIALITY FOR TABULAR DATA

Lawrence H. Cox
US National Center for Health Statistics
Hyattsville, MD USA
LCOX@CDC.GOV

James P. Kelly
OptTek Systems, Inc.
Boulder, CO USA
KELLY@OPTTEK.COM

Abstract

Controlled tabular adjustment (CTA) is a new disclosure limitation method for tabular data that can be used in lieu of cell suppression. Controlled tabular adjustment imputes *safe values* for sensitive tabular cells and then *adjusts* other cells to restore additivity of item detail to marginal totals. Ramesh Dandekar conceived CTA (first named synthetic tabular data) and provided the first software implementation. CTA can be formulated precisely as an integer linear program (Cox 2000). CTA clearly improves data usability compared to cell suppression. It is also relatively easy to implement, even for large and complex tabular structures. The question that remains is what are and how to control the effects of CTA on data quality and accuracy. Initial steps in this direction, from Dandekar and Cox (2002) and Cox and Dandekar (2003), impose within the (integer) linear program capacity constraints on adjustments to individual cell values (e.g., capacities equal to measurement error) and minimize a linear objective function representing overall change across the entire tabular system (e.g., sum of absolute values of individual cell value adjustments). Linear programming constraints contribute importantly to preserving data quality and can be used to preserve means, but fall short of preserving statistics such as variance, correlation or regression coefficients.

In this paper, we explore strategies and limitations on preserving these statistics within the CTA framework. We present two methods. The first is an extension of the linear programming paradigm. The second is based on search strategies, using Tabu search, developed by OptTek, Inc. (Kelly et al. 2003). The linear programming methods have the advantage of being easy to implement in a wide range of applications using standard software. The search methods require considerably more expertise, but offer the advantage of being capable of optimizing measures of data quality of arbitrary (nonlinear) functional form, such as Chi-square. In addition, we demonstrate that some of these objectives can be mutually incompatible, requiring objectives that balance combined effects. We illustrate practical results using simulated data.

Keywords: imputation; Tabu search; (integer) linear program; statistical disclosure limitation