

Distr.
GENERAL

CES/SEM.47/15 (Summary)
30 January 2002

Original: ENGLISH

**STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

**COMMISSION OF THE
EUROPEAN COMMUNITIES**

CONFERENCE OF EUROPEAN STATISTICIANS

EUROSTAT

**Joint UNECE/Eurostat Seminar on Integrated Statistical
Information Systems and Related Matters (ISIS 2002)**
(17-19 April 2002, Geneva, Switzerland)

Topic II: Secure communications and data confidentiality

CONFIDENTIALITY ISSUES FOR STATISTICAL DATABASE QUERY SYSTEMS

Invited paper

Submitted by the National Center for Health Statistics, United States ¹

Summary

1. National statistical offices (NSOs) are responsible to collect, verify and refine statistical data in order to make reliable information available to policy makers and the public. By law or regulation and ethical practice, the NSO must do so in a manner that preserves the *confidentiality* of data pertaining to individual entities such as persons, businesses, and health care providers. Notably exempt from confidentiality concerns are statistics pertaining to governmental units.

2. Prior to the 1960s, NSOs made statistical information available primarily in the form of computed or estimated *tabulations*, defined by cross-classification of only one, two or a small number of variables. The NSO decided which tabulations to release, first in printed form and later also in electronic form. Confidentiality protection, now named *statistical disclosure limitation*, was accomplished by suppressing or combining selected tabulations or entire sets of tabulations or, less frequently, by altering tabulations slightly through rounding or incorporation of random noise. In effect, the NSO first determined which tabulations were worth releasing and then released correspondingly less information in consideration of confidentiality, and data quality, concerns.

3. During the 1960s, first with the Continuous Work History Sample of the U.S. Social Security Administration, followed by Public Use Samples from the 1960 U.S. Decennial Censuses, NSOs began releasing *statistical microdata files* comprising records pertaining to individual entities (mostly, persons). The data user was now free to create all conceivable summaries from the unit record data and, equally important, to fit statistical, demographic or econometric models to the microdata. Statistical disclosure

¹ Prepared by Lawrence H. Cox (LCOX@CDC.GOV).

limitation became focused on altering or removing selected microdata records. Longitudinal data presented confidentiality problems that remain largely unsolved. Emerging research is directed towards fitting the data to complex statistical models and releasing instead model-derived *synthetic microdata* and/or the models themselves. Disclosure limitation for tabulations and microdata are provably complex, both theoretically and computationally.

4. NSOs are now considering allowing data users direct access to statistical data bases, on either a public or restricted access basis, via a *statistical data base query system*. This raises heightened confidentiality concerns and issues, and is likely to motivate statistical disclosure limitation research in coming decades. In this paper, we investigate through examples some of the confidentiality and data usefulness problems raised by the advent of statistical data base query systems. We focus on two query paradigms: tabulations from a data base organized as a large multi-dimensional contingency table and simple statistical models derived from the data base, namely, ordinary least squares regressions and best linear unbiased prediction (*kriging*) models for spatial data.