WP.40
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 UNECE/Eurostat work session on statistical data confidentiality
(Manchester, United Kingdom, 17-19 December 2007)

Topic (v): Panel discussion on balancing data quality and confidentiality

PANEL DISCUSSION

BALANCING DATA QUALITY AND CONFIDENTIALITY

Chair: Lawrence H. Cox, National Center for Health Statistics, United States of America

Quality and Confidentiality Issues Associated with Tabular Data
Lawrence H. Cox

We separate issues surrounding the interplay between data quality for tabular data into two categories: local data quality and global data quality.

Local data quality is focused on characteristics of individual or small sets of cells. These include individual cell values and/or associated time trends (e.g., income or number of employees), or of quantities computed from two or more cell values (e.g., income per employee). Preserving local quality amounts to preserving local characteristics while satisfying confidentiality demands—or, balancing the two.

Global data quality is focused on characteristics of the data set or subsets. These include distributional parameters such as mean and variance, statistics such as quantiles, and distributional shape. Preserving global quality amounts to preserving global characteristics while satisfying confidentiality demands—or, balancing the two.

Local and global quality are not mutually exclusive. Two examples of the overlap are measurement error and correlation. Preserving all or most individual values to within measurement error arguably is perfect local quality, and arguably can be expected to preserve distributional parameters, statistics, and shape. Ensuring nearly perfect correlation between original and masked values is expected to preserve both local and global quality. Other examples are preserving time trends, rank order statistics, and usability/analyzability. These are examples of preserving quality in a *univariate* setting, viz., involving a single variable such as income or number of employees.

Local and global quality can also be *multivariate*, involving two or more variables such as income and number of employees. In addition to preserving univariate quality for each variable, in a multivariate setting the objective is to preserve relationships between variables such as covariance and regressions.

We observe the following data quality characteristics of familiar SDL methods.

Complementary cell suppression For univariate data, CCS preserves local quality perfectly for unsuppressed cells, poorly for suppressed cells; inhibits analyzability; pokes holes in the distribution. For multivariate data, its negative effects are magnified significantly.

Random rounding and perturbation Can preserve local and global quality well for both univariate and multivariate data.

Controlled tabular adjustment Can preserve local and global quality well for both univariate and multivariate data; QP-CTA and MDI-CTA have nearly opposite data quality strengths and limitations.

Perturbing underlying microdata Insufficient information to judge.

Data Masking Procedures for Numerical Microdata

Krish Muralidhar, University of Kentucky, Lexington KY 40506

Recently, there have been considerable developments in procedures that are used to mask numerical microdata. In this discussion, we attempt to assess the relative strengths and weaknesses of these procedures for the most general case in which there exists a data set consisting of both categorical and numerical variables. Since our focus is on masking numerical microdata, all categorical variables are treated as non-confidential or assumed to have been masked and some (or all) of the numerical variables are deemed confidential.

The primary objective of any masking procedure is to prevent disclosure of confidential information. We use the following definition of disclosure: "If the release of the data allows an intruder to estimate either the identity of an individual or the value of a confidential variable with a greater level of accuracy than before the release of the data, then disclosure has occurred." This definition is a combination of the disclosure risk definitions of Dalenius (1974) and Duncan and Lambert (1986). The risk of disclosure resulting from the masking procedure can be assessed as follows: "(Assuming that aggregate information regarding the entire data set and the confidential variables are already available to the user) Does the release of the masked microdata improve the user's predictive ability?" We can show that the answer to this question is "No" if, given the non-confidential variables, the original and masked confidential variables are (conditionally) independent. In this case, we can conclude that the masking technique minimizes disclosure risk since the release of the masked data does not improve the predictive ability of the intruder.

The secondary objective of any masking procedure is to provide the highest level of data utility (or lowest level of information loss) when the masked data is used in place of the original data. In order to achieve the lowest information loss, it is necessary that the response to any arbitrary query using the masked data should be identical to that using the original data. In practice, this measure is difficult to quantify and we usually employ empirical measures. These include the univariate characteristics of the confidential variable, linear and non-linear relationship measures between the variables, ability to maintain subset characteristics, and the ability to reach the same inferences using the masked data as that using the original data. Depending on the specific context, it is possible that alternative measures of data utility are used. Once we have assessed the masking technique on the primary objectives (disclosure risk and data utility), then we can use additional objectives such as ease of implementation, ease of use, etc.

It is often argued that there is an implicit trade-off between disclosure risk and data utility. This is not always true. Those techniques that minimize disclosure risk do not have this inherent trade-off. Only techniques that do not satisfy the minimum disclosure risk requirement have a trade-off between disclosure risk and data utility. Finally, the evaluation approach that we have presented allows us to identify future research opportunities.

1. Dalenius, T. 1977. Towards a methodology for statistical disclosure control. Statistisktidskrift 5 429–444.
2. Duncan, G. T., D. Lambert. 1986. Disclosure-limited data dissemination. J. Amer. Statist. Assoc. 81 10–18.

Quality and risk for different classes of synthetic data

Jerome P. Reiter

Synthetic data come in two flavors: fully synthetic or partially synthetic. To construct fully synthetic data, the agency (i) randomly and independently samples units from the sampling frame to comprise each synthetic data set, (ii) imputes the unknown data values for units in the synthetic samples using models fit with the original survey data, and (iii) releases multiple versions of these data sets to the public. Partially synthetic data comprise the units originally surveyed with only some collected values replaced with multiple imputations. For example, the agency might simulate sensitive variables or quasi-identifiers for units in the sample with rare combinations of quasi-identifiers; or, the agency might replace all data for selected sensitive variables or quasi-identifiers.

Selecting one of these two strategies involves a complex trade-off between disclosure risk and data usefulness. For fully synthetic data, identity and attribute disclosure risks are in general very small because the original units are not released. However, the validity of inferences with the synthetic data depends fully on the validity of the imputation models. The released data can be a simple random sample, eliminating the need for analysts to worry about the typically complex sampling design. But, it may be necessary to have large synthetic sample sizes or numbers of implicates to get inferences with good properties. For partially synthetic data, risks are higher since the collected units remain on the file. However, maintaining some genuine data weakens the reliance on imputation models for valid inferences. Analysts must worry about the original complex sampling design for inferences. A small number of datasets may be adequate when replacing only a fraction of the data.

Synthetic data involve another trade-off that appears most obviously in partially synthetic data: selecting the number of datasets to release. Increasing the number of datasets generally improves inferences (e.g., lowering variances and making normal approximations more plausible). But, it also increases disclosure risks, as intruders' can refine their guesses at the true values with more information. One approach is to release different numbers of replacements for different values, for example few replicates of highly identifying variables and more replicates of weakly identifying variables.

I believe that our current measures of data usefulness do not reflect all aspects of inference. We often focus on point estimation, when the real target should be inferential validity (e.g., 95% confidence intervals cover at least 95% of the time). With a few exceptions, we focus on the first two moments when many analyses rely on other features of distributions. We do not account for complex sampling designs and weights when evaluating (and proposing) disclosure limitation procedures, when these are the norm rather than simple random samples. Finally, we evaluate posited models when we also should examine the model-building process (e.g., do we get the same transformations, interactions, and variable selection). Developing usefulness metrics that incorporate these features is a challenging and important area of research.