Working Paper No. 24 (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

MICRODATA DISCLOSURE BY RESAMPLING – EMPIRICAL FINDINGS FOR BUSINESS SURVEY DATA

Contributed Paper

Submitted by the Centre for European Economic Research (ZEW), Germany[1]

[1] Prepared by Sandra Gottschalk (gottschalk@zew.de).

Microdata Disclosure by Resampling - Empirical Findings for Business Survey Data

by

Sandra Gottschalk

Centre for European Economic Research (ZEW)

February 25, 2003

Extended Abstract:

A problem which statistical offices and research institutes are faced with by releasing micro-data is the preservation of confidentiality. Official statistics are not allowed to pass on to external users outside the office, unless disclosure limitation is guaranteed. The same holds for survey-data, conducted by private or official research institutes, if confidentiality is promised to the respondents. Traditional methods to avoid disclosure often destroy the structure of data, i.d., information loss is more or less high. In this paper I discuss an alternative technique of creating scientific-use-files, which reproduce the characteristics of the original data quite well. It is based on an idea of Fienberg (1997 und 1994) [2], [3] to estimate and resample from the empirical multivariate cumulative distribution function of the data to get synthetic data. For estimation of the empirical cumulative distribution function non-parametric and semi-parametric methods are most useful, like kernel density estimators or a Bayesian approach (Fienberg, 1997 [2]). The procedure should create datasets which have the same charateristics as the original survey data. Means, variances, covariances, correlation and percentiles should not significantly differ. As the elements of the resample are drawn from the cumulative distribution function and do not necessarily correspond to any of those individuals in the original sample survey, an identification of true values is not possible. Nevertheless one cannot rule out the possibility of disclosure, as synthetic datasets could be very similar to real characteristics of observations. Especially, extreme values are at risk. To increase data protection the sample size of the resample could be raised until outliers repeatedly appear.

The typical user of scientific-use-files of official statistics or data from research institutes is interested in causal relationships between variables, perhaps for policy advices. Hence, statistical models are most useful. Econometric parameter estimates can be reproduced quite exactly with resamples, though econometricians have no disadvantages in using only "estimations" of the real datasets. However, the standard error of coefficient estimates in regression analysis could not be exactly estimated. Fienberg (1997) [2] proposes the use of replicated versions of the resample to get multiple parameter estimates.

To estimate the empirical cumulative distribution function is computationally difficult. The exact reproduction of multivariate relationships are not possible until now. For that reason I pick up a procedure, demonstrated by Devroye and Györfi (1995) [1] und Silverman (1986) [5], which avoid the necessity of estimation the distribution function, but directly generate a resample. The algorithm for the univariate case is shortly described in the following:

1. Draw observations X_Z of the data file X with replacement.
2. Compute u to have probability density function K, which is a kernel with window width h.
3. Generate $Z = X_Z + hu$.

The kernel function can be simulated from an epanechnikov kernel, for example[1]. The procedure resamples with replacement from the data and disturb the information in such a manner that the distribution of each variable is retained. The sample size of Z has to be large enough to approximate the distribution of the original data X. A higher dimensional version of the algorithm can be constructed by using directional information in the data, such as the covariance matrix of X. Therefore, the multivariate distribution can nearly be performed. A further modification yields first and second moment properties of the original sample.

In this paper I present some applications of this method with (a) simulated data and (b) innovation survey data, the Mannheim Innovation Panel, which is conducted by the Centre for European Economic Research (ZEW) within the German economy. The information refers to all German enterprises with at least five employees from the manufacturing and mining as well as the distributive and business oriented services sectors. The performance of resample is demonstrated. In a second step I compare resampling with a traditional method of disclosure control, disturbance with multiplicative error (see e.g. Hwang, 1986 [4]), concerning confidentiality on the one hand and the usage of the disturbed data for different kind of analysis on the other hand.

Keywords: dislosure control, resampling, econometric analysis, business survey data

[1] One can also think of the normal density.

References

[1] Devroye, L. and L. Györfi (1985), Nonparametric Density Estimation, New York.

[2] Fienberg, S.E. (1997), Confidentiality and Disclosure Limitation Methodology: Challenges for National Statistics and Statistical Research, Technical Report No. 161, Carnegie Mellon University, Pittsburgh.

[3] Fienberg, S.E. (1994), A Radical Proposal for the Provision of Micro-Data Samples and the Preservation of Confidentiality, Technical Report No. 611, Carnegie Mellon University, Pittsburgh.

[4] Hwang, J.T. (1986), Multiplicative Errors-in-Variables Models with Application to Recent Data Released by U.S. Department of Energy, Journal of the American Statistical Association, 81, 395, 680-688.

[5] Silverman, B.W. (1986), Density Estimation for Statistics and Data Analysis, Monographs on Statistics and Applied Probability 26, London.