INFERENTIAL DISCLOSURE-LIMITED MICRODATA DISSEMINATION

George T. Duncan[1]
Carnegie Mellon University
Pittsburgh, PA 15213

KEY WORDS: Confidentiality, data access

I. Introduction

Microdata files contain individual-level records captured by censuses, sample surveys, and administrative procedures. Increasingly, researchers who study crime, housing, health and other issues of public concern have found that microdata files provide the best factual base for policy analysis. The government agencies who collect and hold microdata files need effective programs for disseminating them to researchers. As data providers, these agencies are mindful of the need to protect the confidentiality of the data subjects. This agency concern is engendered by legal requirements of confidentiality, ethical issues involving commitments made to data respondents, and practical worries about response rates to statistical surveys. Thus an important part of a data-disseminating program is an adequate set of disclosure-limiting procedures. Disclosure limitation can be affected through various mixes of ethical, legal, administrative, and statistical controls. Statistical controls work directly with the microdata file, and specify the form in which it may be released. Using the disclosure-limiting framework of Duncan and Lambert [1986, 1989], this article investigates disclosure risk for microdata and *matrix masking* as a general class of statistical controls for microdata.

II. Motivation

At present some microdata files are released after disclosure limitation. The U.S. Census Bureau, for example, began providing public use microdata from the decennial census in 1963 when it released a one-in-one-thousand sample file for the 1960 Decennial Census. Also, Gates (1988) reports that the Census Bureau prepared a microdata file for the National Opinion Research Center (NORC) in which census tract characteristics are masked.

To provide a systematic basis for providing data to researchers while protecting the privacy of respondents, a decision-theoretic framework for disclosure-limited microdata dissemination can be built from the foundation of Duncan and Lambert [1986, 1989]. It begins with the definition of disclosure proposed by Dalenius [1977], recommended by the Subcommittee on Disclosure Avoidance Techniques [1978], and discussed in Jabine, Michael, and Mugge [1977]:

> If the release of a statistic S makes it possible to determine the (microdata) value more accurately than it is possible without access to S, a disclosure has taken place ...

This definition is also consistent with ones presented by Beck [1980] and Loynes [1979]. Disclosure in this form is called *inferential disclosure* by Duncan and Lambert [1989] and contrasted to *identity disclosure* (Spruill [1983], Paass [1988], Strudler, Oh, and Scheuren [1986]) and *attribute disclosure* (Cox and Sande [1979]). Generally, most confidentiality legislation is drafted using identity disclosure language. The Privacy Act of 1974, for example, says: "...and the record is to be transferred in a form that is not individually identifiable". Nonetheless, given evident public concern about privacy invasion, an agency concerned about its credibility with respondents may well wish to limit inferential disclosure as well. Avoiding jail sentences is not the sole motivation of a prudent data administrator. This paper deals exclusively with inferential disclosure.

In the microdata setting, the data set available to the agency is a file represented by an $n \times p$ matrix $\mathbf{X}$. Each of the n rows gives individual data on each of p attributes. Typically there are many attributes of respondents recorded in the file, including some which are either sensitive in themselves (as income, assets, or medical conditions of target individuals) or relate to sensitive attributes (as taxes paid by a business partner of a target individual).

The released statistic S is some transformation ψ of **X**. For disclosure limitation in the microdata dissemination case, the transformation ψ involves some masking of the data, through such methods as release only of a sample of the data (subtracting rows from **X**), inclusion of simulated data (adding rows to **X**), blurring (fuzzing individual values in **X** by random rounding, grouping, adding random error, etc.), exclusion of certain attributes (removing columns of **X**), and data swapping (exchanging blocks of rows in a certain subset of columns of **X**). Statistical controls specify a particular transformation ψ and ψ**X** is to be released as a complete file. We do not consider the problem of sequential access to a data base in this article.

The purpose of masking the data through ψ is to dissuade the data user from attempting to break the confidentiality of the database **X**. It is now generally accepted--perhaps reluctantly by researchers requiring access to certain data--that the simple transformation of removing columns of **X** that correspond to obvious identifiers or near identifiers (such as name, social security number, address, or telephone number) is insufficient to hamper a serious data spy (see Paass [1988]), just as locking car doors does not deter a serious thief. A careful consideration of the deterrence value of various forms of ψ is required if data custodians are to be convinced that microdata can be released under statistical controls.

In examining the deterrence value of a particular transformation ψ, the beginning point of the disclosure-limiting (DL) approach of Duncan and Lambert [1986] is to model the decision problem of the statistical spy in inferring the value of a target Y from the released ψ**X**. The potential of the information in ψ**X** for inferring Y is a measure of disclosure risk. In the DL approach, disclosure risk is quantitatively assessed according to an uncertainty function U (see DeGroot [1962]).

The basic philosophy behind the DL approach is one of deterrence of the statistical spy. It is to raise the price of using the released information sufficiently high so that the spy will not use it to take actions that infer privacy-protected information. The point is not just to avoid having the spy make correct inferences. Since the act of making identifications in itself can be damaging to a data-disseminating agency and luring the spy to incorrect identifications can typically only be achieved by releasing misleading data which hurts legitimate researchers, the point is to insure that the spy does not make any identifications. The focus then is on the Bayes risk of the data spy. Based on statistical decision theory, the idea is for the agency to choose ψ so that the Bayes risk of inference is raised high enough so that the statistical spy prefers the option of no inference. This idea yields the threshold rule for the agency: release the data if the Bayes risk exceeds some threshold τ.

The data spy wants to use the information in the masked data ψ**X** to infer something about the sensitive target value **Y**. We pursue a decision-theoretic-based development of formal regression of **Y** on ψ**X** in which all probability distributions have the following interpretation: they are the subjective distributions of the statistical spy *as they are perceived by the data- disseminating agency*.

III. Certain Measures of Disclosure Risk

Consider the case of a multivariate target **Y**, multivariate data **X**, and arbitrary variance-covariance structure. We consider a class of measures of disclosure risk based on the eigenvalues of a conditional variance matrix. To obtain the generalization, let **Y** be a $t{\times}u$ matrix of target values sought by the data spy. Let **X** be the n-record attribute data file represented as an $n{\times}p$ matrix. The data spy is assumed to have a subjective joint multivariate normal distribution over **Y** and **X**. In this case, the statistical spy's posterior uncertainty about **Y** after release of **X** is a function only of the joint variance matrix of **Y** and **X**. Using the vec operation (see, e.g., Searle [1982; pp 332ff]) of stacking

(**Y**, **X**) vector-by-vector into a single vector of length tu+np, the joint variance matrix of **Y** and **X** can be expressed in partitioned form as

$$\boldsymbol{\Sigma} = \mathrm{Var}([\mathbf{Y}\ \mathbf{X}]) = \mathrm{Var}([\mathrm{vec}\ \mathbf{Y}\ \mathbf{X}])$$

$$= \begin{bmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{bmatrix}$$

where

Σ_{YY} = Var(vec **Y**), a tu×tu variance matrix;

Σ_{YX} = Cov(vec **Y**, vec **X**), a tu×np covariance matrix;

and

Σ_{XX} = Var(vec **X**), an np×np variance matrix.

The criterion for release is that the conditional variance matrix of **Y** given **X** be sufficiently large. Under joint multivariate normality of **Y** and **X**, the conditional variance matrix of **Y** can be expressed in terms of generalized inverses (Rao, 1973; pp. 522-523) as

$$\mathrm{Var}(\mathbf{Y} | \mathbf{X}) = \Sigma_{YY} - \Sigma_{YX}\, \Sigma^{-}_{XX}\, \Sigma_{XY}. \quad (3.1)$$

An important fact about this conditional variance matrix is that it does not depend on the specific realization of **X**.

As an aside on an informative special case: in typical applications, the statistical spy will view the n records as exchangeable (see, e.g., DeFinetti, 1975, pp 215ff). Hence the *np*×*np* matrix Σ_{XX} will have a block structure of n^2 *p*×*p* matrices, identical on the diagonal (being the variance matrix of the p attributes) and identical off the diagonal (being the covariance matrix of two p-dimensional records). Given this block structure, the use of the generalized inverse is only needed if the *p*×*p* variance matrix of the attributes is not of full rank.

For fixed Σ_{YY}, which is the case from the providing agency's viewpoint, the conditional variance matrix as given in Equation (3.1) suggests the following criterion: Release the data **X** provided

$$\mathbf{G} = \Sigma_{YX}\Sigma_{XX}^{-}\Sigma_{XY}$$

is small. In general, the matrix **G** is of dimension tu×tu. Making the criterion small requires defining some functional of **G**--most obviously some function of its eigenvalues, say its trace or determinant--and making it less than some value τ. In the case of a scalar target (tu = 1), setting the criterion less than τ defines an ellipsoidal region, and demonstrates how amiss can be setting standards for release marginally by individual attribute.

We can take functionals of **G** as measures of disclosure risk.

IV. <u>Matrix masking: Definition and Examples</u>

We explore a certain class of masking transformations ψ that are both important and analytically workable. We focus our attention on masking the n×p microdata file **X** through one of a class called matrix masks. Under matrix masking, the data user is provided the masked r×c microdata file **M** = **AXB** + **C**, and is not given the original data **X**. In this context, a *mask* is a triple of matrices: The linear part of the transformation is given by the r×n matrix **A** and the *p*×*c* matrix **B**; the affine part of the transformation is given by the *r*×*c* matrix **C**. The matrix **A**, as a matrix of row operators, directly transforms the data records in **X**; so we call **A** a *record transforming mask*. The matrix **B**, as a matrix of column operators, directly transforms the data attributes in **X**; so we call **B** an *attribute transforming mask*. The *r*×*c* matrix **C** displaces **AXB** by adding stochastic or systematic noise to the data; so we call **C** a *displacing mask*. In general, the mask (**A**, **B**, **C**) may depend on the particular values in **X**. That is, the mask components **A**, **B**, and **C** are not necessarily just fixed matrices with constant elements or random with elements that are independent of the values in **X**.

Generally, for reasons of data utility--the data must be analyzed--the data provider must also give the user either the complete specification or certain characteristics of the mask (**A**, **B**, **C**). It is an open question of disclosure-limitation methodology as to how much information should be given the data user about the mask in a particular context.

Connections to Commonly Proposed Disclosure-Limitation Methods

Matrix masking encompasses many commonly proposed disclosure-limitation methods.

Record Transforming Masks

By specializing the form of the record transforming mask **A**--with **B** an identity matrix and **C** a zero matrix--we can represent some currently proposed disclosure-limitation techniques, such as:

- *Aggregation across records.* For example, averaging all attributes over three similar records.
- *Supression of certain records.* For example, supression of records having extreme values on some attributes or suppression of records from small identifiable geographic units. Here the transforming mask is a function of the data file **X**.

We can also consider a *random* record transforming mask in which the matrix **A** has stochastic elements. Special cases of this that are of interest include the following:

- *Sampling.* In sampling r rows of **X**, the *r*×*n* matrix **A** has 0-1 random entries a_{ij} with a single 1 in each row. If just records 2 and 3, say, appear in the sample, then **A** has dimensions 2xn, $a_{12} = 1 = a_{23}$, and all other entries are 0.
- *Multiplication of records by random noise.* With the matrix **A** diagonal, each record is multiplied by a random variable.
- *Random aggregation across records.*

Attribute Transforming Masks

By changing the form of the attribute transforming mask **B**, we can represent the following disclosure-limiting procedures:

- *Aggregation across certain attributes.* For example, the release of total income, rather than salary income, business income, interest income, etc.
- *Supression of certain attributes.* For example, some attributes--such as personal identifiers or medical conditions like mental health or HIV infection indicators--may be supressed.
- *Multiplication of attributes by random noise.*

Displacing Masks

In the case of displacing masks (the matrices **A** and **B** are identities), adding **C** yields the following disclosure-limitation techniques:

- *Addition of random noise.* Adding a random variable to each entry.
- *Addition of deterministic noise.* Adding a specified quantity to each entry.

Since addition of deterministic noise has disclosure-limitation value only when **C** is not fully revealed to the data user, both techniques present measurement error or errors-in-variables problems for the user in analyzing the masked data.

Often implemented procedures involve a combination of disclosure-limitation procedures. See, for example, Kim [1986] for a Census Bureau application to the Continuous Longitudinal Manpower Survey which is conducted for the Bureau of Labor Statistics to evaluate the effectiveness of the Comprehensive Employment and Training Act (CETA) of 1973. The public use files contain earnings data matched to Social Security Administration administrative records. The masking technique involved both addition of random noise and data transformation. In these cases, the transforming masks **A** and **B** are not identity matrices and the displacing mask **C** is not the zero matrix.

Given the richness of matrix masks, it is reasonable to ask, "What commonly used (or proposed) disclosure-limiting procedures are not matrix masks?" Here are some examples:

- *Attribute-specific aggregation over records.* Release of some attribute values unmasked, but aggregating other attribute values--say releasing only averages of interest income for similar records.
- *Data swapping.* Release of records with some, but not all, attribute fields interchanged.
- *Multiplication by random noise, in general.* Multiplying each element of **X** by np independent, say, random variables is not a matrix multiplication or addition.
- *Random rounding.* Rounding each entry to a certain base.
- *Grouping.* Condensing categories for some attributes.
- *Truncating.* Truncating distributions of certain attributes.

V. Matrix masks: Derivation and Evaluation

Generally, ad hoc arguments have been used to devise disclosure-limitation procedures and evaluate them in terms of disclosure risk and data utility. Particular implementations can result in significant differences between the information provided by the masked data and that available from the original file (see, e.g., Wolf [1988].

A threshold-rule release criterion can be expressed in terms of the covariance matrix $\mathbf{\Sigma}_{YM}$ (between the target matrix **Y** and the masked data matrix **M**) and the variance matrix $\mathbf{\Sigma}_{MM}$. Based on a functional Ω, it specifies release if

$$\Omega\{\Sigma_{YM}\Sigma_{MM}^{-}\Sigma_{MY}\} < \tau \qquad (5.1)$$

A matrix mask is τ-acceptable with respect to Ω and **Y** if it satisfies release criterion (5.1). Clearly, some masks immediately fail the release criterion and so are not τ-acceptable. For example, if r = n, **B** is the identity, **C** is the zero matrix, and **A** is of full rank and made known to the data user, then $\mathbf{A}^{-1}$ times the released **AX** recovers the original data **X**. Other configurations for **A** compromise some, but not all, records and some, but not all, attribute values.

The basic idea in disclosure limitation is to find a mask that leaves the maximum information about **X**, while at the same time preserves confidentiality. Choosing a mask (**A**, **B**, **C**) to, in some sense, minimize Var(**X**|**M**) while maximizing Var(**Y**|**M**), suggests minimizing Var(**X**|**M**) subject to the mask being τ-acceptable with respect to **Y**. This notion of constrained optimization can be considered consistent with what is reported to be Census Bureau policy: "In practice the Census Bureau has taken disclosure protection as a binding constraint and provided as much data to the public as is possible within this constraint." (McGuckin and Nguyen, 1988b).

For a specific illustration, consider the bivariate case with unit variances for X_1 and X_2 and correlation coefficient ρ. The correlations of Y with X_1 and X_2 are ρ_1 and ρ_2, respectively. We seek τ-acceptable A = [a, 1] to maximize the trace of Var(**X**|**M**). This is equivalent to the following optimization problem:

$$\max_a \frac{(1+\rho^2)a^2 + r\rho a + (1-\rho^2)}{a^2 + 2\rho a + 1}$$

subject to

$$\frac{\rho_1{}^2a^2 + 2\rho_1\rho_2 a + \rho_2{}^2}{a^2 + 2\rho a + 1} < \rho.$$

A boundary analysis of this optimization problem is informative. In the case where $\rho = 0$, there is no optimization problem, because the objective function is constant.

In the case when $\rho_1 = \rho_2 = 0$, the constraint function is zero so this is an unconstrained maximization problem. Solving the unconstrained problem yields the solutions: a = +1 when $\rho > 0$ and a = -1 when $\rho < 0$. Thus $X_1 + X_2$ is released when $\rho > 0$ and $X_1 - X_2$ is released when $\rho < 0$.

How are these unconstrained solutions affected by the constraint? The constraint function at a = +10 is

$$\frac{(\rho_1-\rho_2)^2}{2(1-\rho)},$$

which implies that there can be disclosure difficulty if ρ_1 and ρ_2 are both large, unless ρ is close to +1. Similarly, the constraint function at a = -1 is

$$\frac{(\rho_1-\rho_2)^2}{2(1-\rho)},$$

which implies that there can be disclosure difficulty if ρ_1 and ρ_2 are of opposite sign and of large magnitude, unless ρ is close to -1.

REFERENCES

Boruch, Robert. and Cecil, Joseph. [1979] "Report from the United States: Emerging Data Protection and the Social Sciences' Need for Access to Data." In E. Mochmann and P. Muller, eds. **Data Protection and Social Science Research** New York: Springer-Verlag, 104-128.

Cox, Lawrence H. and Sande, G. [1979] "Techniques for presenting statistical confidentiality," in **Proceedings of the 42nd Meetings of the International Statistical Institute** (Vol. 3). Manila: Philippine Organizing Committee, International Statistical Institute, pp. 499-512.

De Finetti, Bruno [1975] **Theory of Probability (Volume 2)**. New York: John Wiley.

DeGroot, Morris [1962] "Uncertainty, Information, and Sequential Experiments", **Annals of Mathematical Statististics**, 33, 404-419.

Duncan, George T. and Lambert, Diane. [1986] "Disclosure-Limited Data Dissemination", **Journal of the American Statistical Association** 81: 10-28. With discussion by L. Cox, O. Frank, J. Gastwirth, and H. Roberts. Applications Section Special Invited Paper at the ASA Annual Meeting, Las Vegas, August,1985.

Duncan, George T. and Lambert, Diane. [1989] "The risk of disclosure for microdata", **Journal of Business and Economic Statistics**, 7, 207-217.

Gates, Gerald W. [1988] "Census Bureau Microdata: Providing Useful Research Data While Protecting the Anonymity of Respondents" **Proceedings of the Section on Survey Research Methods**, American Statistical Association. Presented at the Annual Meeting of the American Statistical Association, New Orleans, August 22-25.

Govoni, J. P. and Waite, P. J. [1985] "Development of a Public Use File for Manufacturing", **Proceedings of the Section on Business and Economic Statistics**, American Statistical Association, 300-302.

Greenberg, Brian. [1988] "Disclosure Avoidance Research for Economic Data", Presented to the Joint Advisory Committee Meeting, October 13-14, Oxon Hill, MD.

Kim, J. [1986] "A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation". **Proceedings of the Survey Research Section**, American Statistical Association, 370-374.

McGuckin, R. and Nguyen, S. [1988a] "Use of 'Surrogate Files' to Conduct Economic Studies with Longitudinal Microdata", **Proceedings of the Third Annual Research Conference**, Bureau of the Census.

McGuckin, R. and Nguyen, S. [1988b] "Public Use Microdata: Disclosure and Usefulness", U.S. Census Bureau. Center for Economic Studies Discussion Paper CES 88-3, September.

Paass, G. [1988] "Disclosure Risk and Disclosure Avoidance for Microdata", **Journal of Business and Economic Statistics**, 6, 487-500.

Rao, C. R. [1973] **Linear Statistical Inference and Its Applications**, 2nd edition. New York: Wiley.

Searle, Shayle R. [1982] **Matrix Algebra Useful for Statistics.** New York: Wiley.

Spruill, Nancy L. [1983] "The Confidentiality and Analytic Usefulness of Masked Business Microdata," in **Proceedings of the Section on Survey Research Methods, American Statistical Association**, pp. 602-607.

Strudler, Michael, Oh, H. Lock, and Scheuren, Fritz [1986] "Protection of Taxpayer Confidentiality with Respect to the Tax Model," **Proceedings of the Section on Survey Research Methods, American Statistical Association**, pp. 375-381.

Wolf, Michael. K. [1988] "Microaggregation and Disclosure Avoidance for Economic Establishment Data", **Proceedings of the Section on Survey Research Methods**, American Statistical Association. Presented at the Annual Meeting of the American Statistical Association, New Orleans, August 22-25.

[1] Financial support of the National Science Foundation (SES8910513) is gratefully acknowledged. The comments of Diane Lambert are appreciated.