The research program of the Center for Economic Studies produces a wide range of theoretical and empirical economic analyses which serve to improve the statistical programs of the U.S. Bureau of the Census. Many of these analyses take the form of research papers. The purpose of the Discussion Papers is to circulate intermediate and final results of this research among interested readers within and outside the Census Bureau. The opinions and conclusions expressed in the papers are those of the authors and do not necessarily represent those of the U.S. Bureau of the Census. All papers are screened to ensure that they do not disclose confidential information. Persons who wish to obtain copies of papers, submit comments about the papers, or obtain general information about the series should contact Peter Zadrozny, Editor, Discussion Papers, Center for Economic Studies, Room 3442, FOB 3, U.S. Bureau of the Census, Washington, DC 20233 (301-763-2490).

PUBLIC USE MICRODATA: DISCLOSURE AND USEFULNESS

by

Robert H. McGuckin and Sang V. Nguyen*

CES 88-3 September 1988

Abstract

Official statistical agencies such as the Census Bureau and the Bureau of Labor Statistics collect enormous quantities of microdata in statistical surveys. These data are valuable for economic research and market and policy analysis. However, the data cannot be released to the public because of confidentiality commitments to individual respondents. These commitments, coupled with the strong research demand for microdata, have led the agencies to consider various proposals for releasing public use microdata.

Most proposals for public use microdata call for the development of surrogate data that disguise the original data. Thus, they involve the addition of measurement errors to the data. In this paper, we examine disclosure issues and explore alternative masking methods for generating panels of useful economic microdata which can be released to researchers. While our analysis applies to all confidential microdata, applications using the Census Bureau's Longitudinal Research Data Base (LRD) are used for illustrative purposes throughout the discussion.

I. INTRODUCTION

Most official economic data publications are based on aggregations of microdata collected in statistical surveys of individual respondents. These data are used by policy makers, researchers, and market analysts as economic indicators, and as a source of information for developing economic policy and testing economic theory. As useful as these aggregate data are, the underlying microdata provide even more valuable information for the study of the economy. Many hypotheses concerning the nature of production, technical change, and the interaction of individual firms can only be tested using detailed microdata.[1] Moreover, the extent of aggregation bias can only be evaluated with the use of microdata. As a result, the demand for detailed microdata by public and private research communities has been increasing.

Faced with this, statistical agencies such as the Census Bureau have sought ways to make microdata available to outside researchers and policy makers without violating confidentiality commitments to individual respondents. Aside from legal issues, the confidentiality commitment to respondents is of great concern because statistical officials fear that low rates of response to statistical surveys will get lower if the released microdata reveal confidential information about individual respondents.

All masking techniques create surrogate data by adding either stochastic or systematic (or both) measurement errors to the data. In turn, undoing or correcting for such errors can only be

accomplished within the context of specific econometric models. Put differently, evaluation of the effects of measurement error on parameter estimates depends on the model describing the relationships among the variables associated with the unmasked data. Thus, determining the usefulness of a public use data file is essentially a problem in evaluating the effects of measurement error.

It would be convenient to have one public use file that could provide researchers with sufficient information to test hypotheses and estimate models, while maintaining confidentiality protection for respondents. Unfortunately, the masking techniques used to preserve confidentiality limit the economic studies that can be carried out with any particular public use data set. Thus, it is extremely unlikely that any single public use file will satisfy all users. Ideally, it would be best to release many different files to satisfy the needs of different researchers. But, this complicates disclosure analysis because the release of a particular public use file may make it possible to identify individual respondents in another file, which by itself would not reveal confidential information.

These issues are of clear importance to economists. Yet, the confidentiality issues are not widely understood, and there has been little research on the subject even within the statistical community. In this paper we discuss disclosure issues in the context of confidential economic microdata and we explore

alternative methods for generating useful public use microdata. We also provide a specific example of a deterministic transformation which generates masked microdata for estimating production function and other econometric models within the log-linear regression framework as well as for use in total factor productivity (TFP) analysis. Unfortunately, the degree of disclosure protection offered by this transformation is still an open question. Faced with unresolved disclosure issues we conclude that special "aggregate" tabulations in the form of variance-covariance matrixes offer researchers the best currently feasible method for obtaining public use data which allows them to obtain good estimates of microeconomic models.

II. PUBLIC USE DATA FILES AND CONFIDENTIALITY

The Census Bureau collects microdata under the authority of Title 13 United States Code which requires that all collected information must be kept confidential and used for statistical analyses only.[2] To protect confidentiality, Title 13 and the disclosure rules and regulations of the Census Bureau prohibit the release of information that could be used to identify or closely approximate data for individual establishments or enterprises. But, anytime data are released there is some, however slight, risk of confidentiality disclosure. Thus, Title 13 has two legitimate but conflicting objectives: promote wide use of the collected data in statistical analyses, while maintaining the confidentiality of

the data. 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.

At one level, analysis of disclosure is straightforward. Given the data to be released and an accurate list of publicly available information, one can either identify an individual establishment or not. Unfortunately, the process has some elements of uncertainty. Uncertainty arises because outsiders trying to uncover the identity of the individual entities use reverse transformations and estimations which introduce probabilistic elements to the matching process and confidentiality disclosure analysis.[3] Moreover, the extent of outside information is never perfectly known by the agency.

There is simply no easy way to know exactly what information is available to the public nor is there any easy way to evaluate its quality or how well it can duplicate the data in the confidential microdata file. Even if the agency had information on the extent and nature of outside microdata files, it would have to devote substantial resources to link outside data bases to the potential public use file to test for disclosures. In part, this is

because there exists an extremely large stock of "publicly" available information outside a particular federal agency. This becomes obvious when it is recognized that for purposes of confidentiality analysis at the Census Bureau, the IRS, BLS and other government agencies are outside users.

Moreover, it is impossible for an agency that wants to release a public use microdata file to keep track of new outside files and changes to existing ones. Finally, many publicly available data files have limited availability and hence the exact contents of the files would be unknown to the agency.[4]

CONFIDENTIALITY DISCLOSURE AND SUMMARY STATISTICS

The Census Bureau has well-defined procedures for evaluating disclosure in aggregate data tabulations. To the best of our knowledge, the confidentiality disclosure rule for aggregate data has been satisfactory for over 40 years. This policy is addressed for summary statistics or aggregate data with the Census Bureau's (n, k)-rule. For confidentiality reasons, the parameters n and k are not disclosed by the Bureau. In the rule n represents the minimum number of units or respondents represented in the cell and k is the maximum percentage of the value of the cell. The (n, k)-rule has been discussed at length elsewhere but a simple example of how it operates is useful.[5]

Table 1 provides a simplified but typical establishment data panel from the Center for Economic Studies' Longitudinal Research Data base (LRD).[6] If one wanted to release a table showing the

size distribution of plants by shipments in SIC 2011, then the (n, k)-rule guides the choice of size classes which can be used to display the data. Thus, for example the (n, k)-rule would allow for cells of no less than n plants accounting for k percent of total shipments in any publicly available display.

Table 1. Source Data for a Public Use File: An Example

	Data From the LRD Panel								Data from the Outside File				
	TS	IB	IE	K	L	E	...	C	R&D	PQ	PK	...	L'
MAJOR GROUP 20													
GROUP 201													
Industry 2011													
Plant 1													
1972	81	10	11	16	19	11	...	0	6	.815	.832	...	19
1973	85	11	15	19	21	13	...	0	5	.819	.841	...	21
.	.			.		.		.	.	.	.		.
1985	150	9	12	25	26	18	...	0	9	1.350	1.450		26
Plant 2													
1972	6	2	3	4	2	0	...	1	1	.801	.829	...	2
1973	9	3	2	5	2	0	...	2	1	.809	.838	...	2
.	.			.		.		.	.	.	.		.
1985	12	4	2	9	7	0	...	4	2	1.260	1.390	...	7
.	.			.		.		.	.	.	.		.
Plant n													
1972	82	13	15	51	13	0	...	21	2	.811	.817	...	3
1973	86	10	20	60	14	0	...	24	3	.819	.825	...	4
.	.			.		.		.	.	.	.		.
1985	151	4	26	100	17	0	...	30	6	1.270	1.310	...	7
Industry 2012													
Plant 1													
.	.			.		.		.	.	.	.		.
Industry 2026													
.	.			.		.		.	.	.	.		.
GROUP 203													
Industry 2032													
.	.			.		.		.	.	.	.		.
Industry 2037													
.	.			.		.		.	.	.	.		.
GROUP 209													
.	.			.		.		.	.	.	.		.
MAJOR GROUP 21													
.	.			.		.		.	.	.	.		.
MAJOR GROUP 39													
.	.			.		.		.	.	.	.		.

Definitions TS, IB, IE, K, L, E, and C denote total shipment, inventories at beginning and end of year, capital, labor, electricity, and coal in the LRD file, while R&D, PQ, PK, L' denote research and development (from the Census/NFS R&D file), and output price index, capital service price index (from the Bureau of Industrial Economics, BLS), and labor (from a trade

association.)

CONFIDENTIALITY DISCLOSURE AND MICRODATA

While the Bureau has well-defined rules for summary statistics, precise criteria for evaluating disclosure risk in economic microdata are not available. Without such criteria, the extent of confidentiality protection provided by any type of masked data is always uncertain.[7] Moreover, the problems involved in economic data are far more pronounced than those found for demographic data because of the nature of the data involved. The dummy data in Table 1 have been constructed to highlight a number of aspects of economic microdata which make disclosure of useful microdata difficult. We begin by focusing on two: uniqueness of particular information and the skewed size distribution of business units. Because of these characteristics use of expanded classifications and sampling procedures similar to those used to reduce disclosure risk for many demographic surveys are not very helpful in developing economic public use files.

Classification Criterion

Irrespective of the particular form in which data are released, one can gain some confidentiality protection by expanding the number of items in the class. This can be accomplished by reporting data, for example, at the
2-digit (or 3-digit) instead of the 4-digit SIC level of classification detail. Nonetheless, most economic models are specified based on certain assumptions about markets and the competitive relationships among the firms in the markets.

Therefore, classification schemes at the level of markets are generally better than the broader SICs typically used for the Census Bureau's and other official statistical agencies' publications.[8]

Sampling

In a similar vein, given the level of classification detail, confidentiality protection can always be increased by releasing a sample of the data file rather than a complete file. Use of sampling in this way will increase the variances of the estimates, but with a sufficient sample size, the precision of the estimates should be acceptable. Unfortunately, industries and markets are characterized by small sample sizes and extremely skewed size distributions. In fact, as already mentioned, in some instances the distributions may be so skewed that only one establishment uses a particular input or process.[9] For example in Table 1, only Plant 1 uses electricity (E), and all others in the industry use coal (C). This knowledge which could often be ascertained by public users would enable researchers to identify the electricity user's data.[10] Thus, in order to use sampling as a technique for increasing confidentiality protection, one would have to expand the plants included in the sampling frame beyond the 4-digit SIC level. This suggests that from the standpoint of data usefulness, sample public use files have limited applicability.

CONFIDENTIALITY AND THE COMPOSITION OF THE PUBLIC USE DATA FILE

We have referred at various points to a public use microdata file. Before proceeding, it is important to examine the issue of what data items are to be included in the public use data file. In some respects this represents the most difficult issue in creating a public use file. It is also an issue which, to our knowledge, has not been addressed explicitly by previous studies.

Most microdata files contain data for a large number of variables. For example, the LRD contains data for more than 80 reported variables. To conduct econometric analysis, researchers usually must generate certain constructed variables using two or more reported variables. As an example, consider output which is conventionally defined as

> Output = (total shipments) + (finished goods inventory at the beginning of year) - (finished goods inventory at the end of year) + (goods-in process inventory at beginning of year) - (goods-in-process at the end of year).

In this example output is the constructed variable, whereas inventories and total shipments are the reported variables. Does one mask the reported or constructed variables? If important relationships among the variables in the original data file are to be preserved, the constructed variable must be developed before masking. If one masks the data on total shipments and inventories before constructing the output series, the output variable in the public use file will not generally provide the same regression estimates as those obtained from the original data.

A similar difficult and important issue is what to do about

constructed variables that involve data from outside sources. As an example, when using LRD data to construct the service price of capital input, researchers need some outside data because the variables needed to construct it are not all available in the LRD. If the Census Bureau constructs with outside data a new variable not currently in the LRD, disguises it, and then releases it together with disguised data from the LRD file, confidentiality may be violated. Outside researchers can possibly use data from outside sources and the released transformed data to perform a reverse transformation which can help identify individual establishments in the original LRD file. If, however, the development of the constructed variable requires data from the original file and an outside file, disclosure risk is likely to be small if the constructed variable is based on multiple variables not included in the public use file. Also, if there is a series from an outside file that duplicates the data on a variable in the LRD file, then transforming and releasing this variable in a public use file would violate confidentiality. For example, in Table 1, the L series (in the LRD panel) and the L' series (in the outside file) are identical. If one transforms L and reports it in a public use file, then with L' available in an outside file, the researcher can successfully perform a reverse transformation and hence identify individual establishments in the public use file.

Thus, it is important to keep in mind that the data to be transformed must be carefully specified before any transformations

are undertaken. Moreover, inclusion of data from outside sources in the public use file can only be undertaken with extreme care. Finally, the choice of variables to be included in the public use file is part of the general problem of deciding on what aspects of the original file should be preserved in the surrogate public use file. As with the characteristics of the transformations discussed below, both the usefulness of the public use file and its disclosure risk must be considered.

III. TECHNIQUES FOR CREATING PUBLIC USE FILES

Proposals for creating public use microdata files by masking the original data can be broadly classified into two categories: tabulations or summary statistics and transformations of the original microdata which preserve the individual data unit. In assessing specific proposals within each category of masking technique, consideration of the usefulness of the transformed data must be weighed against the possibility of disclosure. In this section, we focus on utility issues. But, the reader should keep in mind that well defined disclosure criteria for microdata have not yet been developed.

EVALUATING THE UTILITY OF PUBLIC USE DATA FILE

In the previous section, we focused on the properties of economic microdata as they related to disclosure. Here we try to be more specific about the general characteristics of the original

microdata file that the public use data file should preserve. Although it is unlikely that a public use data file will satisfy all users, there are at least three characteristics which a public use data file should possess.

First, because most empirical economic studies apply data to estimate the parameters of certain econometric models, we think it is most important that a public use data file should be capable of generating the same parameter estimates as those obtained from the original data. Consider the general production model

$$Y = F(X_1, X_2, \ldots, X_k ; ß_1, ß_2, \ldots, ß_k),$$

where Y denotes output and X_i denotes the ith input, and the ßs are the model parameters to be estimated. With masked data the model becomes

$$Y^* = F^*(X^*_1, X^*_2, \ldots, X^*_k; ß^*_1, ß^*_2, \ldots, ß^*_k),$$

where $Y^* = f^*(Y,u;1)$, $X^* = f^*(X,u;1)$, u is a random noise, 1 is a transformation parameter; and $ß^*$s are the model parameters associated with the masked variables X^*s.

In general, data masking will introduce stochastic and/or systematic measurement errors in variables that may lead to serious biases in model parameter estimates. Thus, ideally, the data should be masked in such a way that they can yield model parameter estimates that have the same properties as the original. While one can think of various possible relationships between $\hat{ß}^*$ and $\hat{ß}$, we emphasize $\hat{ß}^* = \hat{ß}$ throughout the remainder of the paper.

Second, it is important for outside researchers to be able to

link surrogate data with data available from outside sources for economic analyses because it is unlikely that any single file can provide all information needed for different studies. Third, an ideal public use data file should enable users to work with subsets of the public use panel. There are several aspects to this point. Researchers will often seek to edit out certain observations because they may, for example, represent outliers from the standpoint of the particular hypotheses under consideration. An important aspect of this issue is edit analysis. Typically, all microdata are subject to measurement errors. Moreover, the errors may often be large because the collection process is geared to producing aggregate statistics, and thus edits and imputations for the underlying observations are often neglected. Similarly, subsets of the data file are often examined to test for the stability of parameter estimates. Moreover, researchers will want to examine different dimensions of the data (e.g., panels, time-series, cross-sections).

Finally, it is useful to have a surrogate file which can easily be expanded to include data from new periods and industries. As we show below, the ability of a proposed public use file to accommodate users in these regards is extremely important.

TABULATIONS OR SUMMARY STATISTICS AS A PUBLIC USE FILE

Data Grouping

This approach involves tabulating average data values of

similar establishments grouped according to one or more criterion variables. In the most prominent example of this approach, Govoni and Waite (1985) suggest ranking establishments according to their values of shipments. Groups of establishments of size m would then be chosen from the ranked list of p x m establishments with p size classes, each contains m similar establishments needed to satisfy the (n, k)-rule for the release of summary data. The (n, k)-rule would be applied to each cell of the table where each cell contains the average value of the variable for those establishments in the group.[11] Thus the resulting public use data file will contain p (or less) data points (averages) for each variable in the original file.

This approach has two major advantages: it is easy to develop because it represents a simple sort and retabulation of the microdata that the Census Bureau uses regularly as the basis of published reports. More important, it takes the disclosure criteria into account directly in the calculation of the data.

While there are many possible variations to the specific proposal suggested by Govoni and Waite, each has the same fundamental problem: the linking of similar establishments is only valid with respect to criterion variables. For example, from Table 1 we see that Plant 1 and Plant n produce approximately the same amount of output (measured in total values of shipment, TS); however, they use different combinations of inputs. In particular, Plant 1 uses exclusively electricity (E) as its energy input, while

Plant n uses coal (c). With the data grouping approach and using total shipments as the criterion variable, one would put the two plants in the same group. Clearly, this procedure is only valid with respect to total values of shipments, but not valid with respect to electricity and coal because the resulting data will not reflect the exclusive use of energy inputs of the two plants. Thus for the variables other than the criterion variables, data grouping would introduce serious measurement errors in variables, leading to inconsistent model parameter estimates. This fundamental objection is valid for all tabulations and is one of the primary reasons researchers desire microdata files. In this regard the variance-covariance approach to which we now turn is a major improvement over the data grouping approach because it can provide the same regression estimates as those estimated using the original microdata.

The Variance-Covariance Matrix

This approach involves tabulating the variances, covariances, and means of the original microdata. The advantage of this approach is that OLS estimates of both the intercept and slope-coefficients can be obtained directly from the public use data file. Also, the variance of the error term, and thus other test-statistics can be obtained from the information in the file.[12] This can be easily seen by noting that ordinary least squares (OLS) coefficients of a linear regression model

Y = " + ßX + , ,

can be estimated by computing

$$\hat{\beta} = \frac{\text{cov}(Y,X)}{\text{var}(X)},$$

and

$$\hat{\varepsilon} = Y - \hat{\beta}X.$$

The key point to recognize is that the variance-covariance approach provides OLS estimates, which are identical to those obtained from the original microdata, but involves a tabulation of summary statistics (i. e., variances, covariance, and means) with little risk of disclosure. In fact, the Census Bureau has already released an experimental variance-covariance file which was satisfactory for two microeconomics studies within a single-equation linear regression framework (see Griliches and Hall, 1982, and Griliches, 1986). But, the variance-covariance approach means that the researcher cannot obtain correct regression estimates for different subsets of the data file (for example, time-series data only) because the released statistics are computed using the data from the <u>whole</u> data set. Furthermore, if the researcher is concerned about problems such as missing data, outliers, industry and firm effects or the like, then multiple public use files will be needed for their research.[13] Thus the major disadvantage of this approach is that the agency eventually has to release an unreasonably large number of matrices because different users will require different matrices. Moreover, even the same user may require different matrices for hypothesis testing.[14] Also, as noted earlier, multiple public use files complicates the disclosure

analysis. The above shortcomings of both types of tabulation of summary statistics have led researchers to propose specific data transformations as techniques for creating public use files which mirror the original microdata.

MICRODATA SURROGATE FILES

Transformation techniques to create public use microdata files can be broadly classified into two categories which we distinguish according to whether the error introduced into each variable is stochastic or deterministic. Each of these methods has some merit; but, each is also subject to limitations with respect to the types of economic research it will support. Moreover, these methods are not perfect substitutes in terms of disclosure protection.

Stochastic Transformations

These techniques involve masking confidential data by transforming the original data into surrogate data by introducing random noise.[15] The simplest stochastic transformation is the addition of random noise to the original data. Such a scheme can be written as

$$X^*_i = X_i + u_i$$

where X^*_i and X_i denote the transformed and original variables and u_i is a random noise independently distributed with mean zero and variance F_i^2. Within the context of the simple regression model,

$$X_2 = " + ßX_1 + ,$$

estimated on the basis of surrogate file

$$X^*_2 = \alpha^* + \beta^* X^*_1 + \varepsilon^*.$$

the estimated slope coefficient

$$\hat{\beta}^* = \frac{\text{cov}\,(X_2, X_1)}{\text{var}\,(X_1) + \text{var}\,(u_1)},$$

is biased because of the var(u_1) in the denominator. However, if the var(u_1) is provided within the file, then $\hat{\beta}$ can be obtained from $\hat{\beta}^*$. Unfortunately, work by Paass (1985) based on re-identification experiments using discriminant analysis showed that adding random noise to the original data did not provide an acceptable level of disclosure risk in situations where the noise was small enough to preserve reasonably efficient estimates of relationships among the variables in the original file.

In the light of Paass' finding, a proposal to use a modified version of the additive random noise scheme was made in an attempt to provide more confidentiality protection.[16] Basically, this scheme focuses on imposing certain constraints on the variances and covariances of the random noise terms such that these statistics are equal or proportional to those found in the original variables. The advantage of the constrained random noise technique is that it can generate masked data that provide the same OLS estimates (including the intercept) as with the original data.[17] This result is in contrast to the biased estimates obtained from use of the simple additive random noise scheme.[18] However, the constrained random noise approach is subject to important limitations. These

limitations have been discussed at length in McGuckin and Nguyen (1988), but it would be helpful to mention here two particular shortcomings of this approach.

First, imposing restrictions on the variance-covariance matrix of the random noise will severely constrain the flexibility of the researcher to partition the public use file into useful subsets. For instance, if the public use file includes pooled time-series cross-sectional data as shown in Table 1, researchers cannot use either time-series or cross-sectional subsets of the panel alone. Similarly, researchers cannot suppress outliers due to erroneous responses or the like. Either of these actions will violate the variance-covariance restrictions built into the masked data, and hence will yield parameter estimates that do not reflect those obtained from the original data.[19]

Second, adding any random noise to the original data will distort the original variables and therefore the transformed data cannot be used to create new variables such as first differences, growth rates, and total factor productivity growth. Thus, as with the simple random noise scheme, the constrained random noise approach will rule out productivity analyses and other studies using first differences or rates of change.[20] These findings provide a good reason to examine microdata public use files based on deterministic transformations.

Deterministic Transformations

There are numerous potential deterministic transformations

that could be used. But, the class of useful transformations is much smaller. Here we discuss a particular deterministic transformation possible for use with log-linear models, and some of its variants. The transformation disguises the original data using

$$X^* = MX^1 ,$$

where $M > 0$ and $1 > 0$.[21]

In general this scheme introduces systematic measurement errors in variables through the parameters M and 1. However, in the log-linear regression model, this transformation provides the same parameter estimates as with the original data.[22] Consider the widely used log-linear two-factor production function,

$$\text{Log}(Y) = " + ß_k \log(K) + ß_l \log(L) + ,,$$

where Y, K, and L denote output, capital, and labor, while $ß_k$ and $ß_l$ are the respective output elasticities to be estimated. Applying the masked data, the model becomes

$$\text{Log}(Y^*) = "^* + ß^*_k \log(K^*) + ß^*_l \log(L^*) + ,^*,$$

where $Y^* = MY^1$, $K^* = MK^1$, and $L^* = ML^1$. With some algebraic manipulations, it is easy to show that

$$ß\hat{}^*_i = \frac{1^2 \text{ cov } (Y,X)}{1^2 \text{ var } (X)} = ß\hat{}_i,$$

where i = k,l.

Choice of Transformation Parameters

The above deterministic transformation form includes two transformation parameters, M and 1, to reduce disclosure risk. However, it is important to point out that there is a trade-off

between disclosure protection and data utility in the choice of ways in which the transformation parameters are specified.

The parameter M is included to help disguise the level of the original data. The value of this parameter should vary across establishments for disclosure protection. For example, if large values of M are assigned to small establishments (and vice versa), then notably large (or small) establishments cannot be identified by their sizes.

The parameter 1 is included to disguise the first-order rates of change of the original data, but the way in which it is chosen defines the nature of the transformation scheme. Thus, depending on the choice of 1 the nonlinear scheme reduces to simple exponential, multiplicative or ratio transformation. One form of this transformation is of particular interest. If 1 is set equal to 1, then the transformation ratio reduces to

$$X^* = MX \ .$$

This is a ratio transformation scheme which was first proposed by Griliches (1985), and supported with suggested modifications by Monahan (1986).

The ratio approach has generated some user interest but there are several versions of it, depending on whether the parameter M is "predetermined" or randomly chosen, and whether it is fixed or varies across all establishments and over time. If M is fixed over time and varies across establishments, then the surrogate file will yield correct estimates for certain time-series regression models,

but generally will not yield correct estimates for cross-section models and hence pooled time-series cross-sectional models. An example of this ratio scheme is provided by Griliches (1985). As part of his proposal, Griliches suggested expressing all LRD data on a per-employee in 1977 basis (i.e., $M_i = 1/E_{i,1977}$). Monahan (1986) also suggested a similar scheme, but allowed outside researchers to pick any variable in the LRD file as the denominator of the ratio.

The advantage of Griliches' and Monahan's approach is that it provides disclosure protection by allowing M to vary across establishments while preserving the time series structure of the file in a form capable of handling TFP analysis. However, because the data for each establishment in the ratio file are divided by a constant, the rates of change of ratio data are identical to those of original data. Thus, unless there are many establishments of the same size within a given product class, outside researchers may be able to identify individual establishments if they can obtain data on one or more variables in a file from outside sources.[23]

A variant of Griliches' ratio approach is to assign "predetermined" values to M which vary across establishments and over time. For example, for each establishment, dividing each series of variables by the employment series (i.e., $M_{it} = 1/E_{it}$). We refer to E_{it} as "predetermined" values of M because once the series E is chosen, the M_{it} is determined by E_{it}. The advantage of the "predetermined" ratio form is that it allows M to vary over

time and across establishments for confidentiality protection while allowing outside researchers to use the panel data based on this scheme to conduct production and total productivity analysis. Unfortunately, because M varies through time, the growth rates of the transformed data differ from those of the original data. Thus, the growth rates of the transformed output and input data are not useful for TFP analysis unless one is willing to accept the constant returns to scale hypothesis. Similarly, the resulting transformed data are not valid for estimating non-constant returns to scale production functions.

Summary of Findings

Table 2 summarizes the various masking schemes based on their capacities to provide correct parameter estimates (i.e., the same as with the original data) for four common single-equation econometric models.[24] Turning initially to the random noise transformations, we see that if the variance of the noise is released along with transformed data both the simple random noise method and the constrained technique provide correct estimates for all the models.[25]

The variance-covariance matrix is a far better possibility for a public use file than either random noise approach. As shown in Table 2, estimates of economic models based on the underlying microdata can be obtained from the variance-covariance file. Moreover, as discussed earlier, the variance-covariance approach

provides more disclosure protection than the random noise techniques because variances, covariances and means are summary statistics. However, as with the constrained random noise public use file, the variance-covariance file does not provide researchers with flexibility in choosing subsets from the public use panel data file. This is not true of the simple random scheme which allows subsets, but does not appear to be disclosure free. These findings provide a good reason to examine public use microdata files based on deterministic transformations.

As with the random noise and variance-covariance schemes, the generalized and ratio deterministic transformation techniques both provide correct estimators for the log-linear production model if the M and 1 are constant.[26] Both deterministic approaches are able to provide useful measures of TFP from the transformed data with appropriate adjustment for 1. This contrasts with the random noise approaches and the variance-covariance approach. Unless the TFP model is estimated by the Census Bureau prior to release, these latter approaches cannot be used for TFP analysis.

As noted at the beginning of the section, in comparison with the constrained random noise transformation and the variance-covariance approach, a major advantage of the deterministic transformations is that they provide researchers with flexibility in using different subsets (both time-series and cross-sectional data) in their studies. Similarly, unlike the constrained random noise and variance-covariance approaches, deterministic surrogate

data files can be updated annually or expanded cross-sectionally without reconstructing the entire surrogate file.

IV. CONCLUDING REMARKS

In this paper, we addressed some important issues concerning confidentiality of microdata collected by official statistical agencies, and explored alternative data masking methods for constructing public use microdata files. While we outlined confidentiality issues, we did not attempt to carry out formal tests for risk of disclosure of confidential information. Instead, our focus was on the usefulness of the various data masking techniques in providing correct estimators for a particular class of single-equation econometric models. Our analysis indicates that it is extremely unlikely that any single public use file will satisfy all users.

Between the stochastic and deterministic transformations, there are good reasons for further research focusing on deterministic transformation techniques. A deterministic transformation such as the generalized transformation scheme proposed here may be able to provide a microdata file from which economic models can be estimated and which has the flexibility to allow researchers to use subsets of the data without the risk of disclosure associated with the simple random noise approach. Moreover, the deterministic approaches provide in a simple transformation, the possibility for estimating single equation production and TFP models.

A surprising conclusion of this analysis is that the variance-covariance approach is the most likely to offer immediate benefit since it provides correct regression estimates and can be easily obtained with little risk of disclosure. While the variance-covariance approach has many advantages, it limits the researcher to linear models. For example, if the model is linear in the original variables then the variances and covariances contained in the matrix must be calculated using data in level form. However, if the model is a log-linear model, then the data must be transformed into logarithms before computation of the variance and covariance matrix. Finally, a public use file based on this method cannot be used to create new variables such as growth rates of outputs or total factor productivity. This last property is very restrictive because it means that all imputations, data editing and related work must be done by the agency before release of the file.

The release of useful and disclosure free public use microdata files for empirical studies is important for both official statistical agencies and interested data users. Unfortunately, the issue is extremely complicated and requires much more research. Admittedly, in this exploratory paper we do not provide any solution that can satisfy all the needs of the many researchers; but we hope the paper has given sufficient motivations for further research. In this regard, it is extremely important to develop precise criteria for evaluating disclosure risk. Without such criteria, evaluating a microdata public use file in terms of disclosure is almost impossible. But, we emphasize that disclosure

free files are not enough. Such files must be useful and we think the best hope for developing a public use file lies in focusing research on surrogate data files which allow researchers to estimate common economic models.

Table 2.a
A Comparison of Competing Transformation Schemes
in Terms of Providing Correct Estimates for
Particular Econometric Models

Schemes	Economic Relations or Econometric Model in the Original Data			
	Linear $Y=" + ß_1X_1 + ß_2X_2$	Log-linear Production Model $lnY = " + ß_1lnX_1 + ß_2lnX_2$	Growth in TFP Models $TFP = ln(Y_i/Y_{it-1}) - Ew_iln(X_{it}/X_{it-1})$	Non-linear Quadratic Model $lnY = " + ß_1lnX + ß_2(lnX)^2$
Stochastic				
Simple[e,a]	Yes[a]	Yes[a,b]	No	Yes[a,b]
Constrained	Yes	Yes[b]	No	Yes[b]
Non-Stochastic				
Generalized[e]	No	Yes	Yes[d]	Yes[b]
Ratio[e]	Yes[c]	Yes[d]	Yes[d]	Yes[b]
Aggregate				
Var/cov	Yes	Yes[b]	No	Yes[b]
Data grouping	No	No	No	No

[a]No, if the variance of the error term is not released.

[b]No, if the appropriate transformations to the original variables cannot be done before the file is masked.

[c]Yes, for the linear if "a" is constant over time and establishments. Yes for the log-linear, if "a" is constant over time. Also, yes for the log-linear if "a" varies over time and across establishments and if constant returns to scale is applied (i.e., $ß_2 = 1-ß_1$).

[d] If "a" is constant across establishments and time it is equivalent to the generalized transformation. If "a" is fixed over time, then yes because the cross-sectional variation is not relevant in TFP construction. Yes, if constant returns to scale. If b = 1, then will get an unbiased estimate of TFP, and with $b \neq 1$, estimated TFP is biased but can be used in regression anslysis (See Appendix A.3).

[e] This technique allows use of subsamples of the surrogate file.

Table 2.b
A Comparison of Competing Transformation Schemes
in Terms of Satisfying Major Properties
of Useful Surrogate Economic Data Files

Properties of Useful Micro Surrogate Files

Transformation Schemes	Provide Correct Estimates for Log Linear Regression Models[a]	Provide Correct Estimates for Subsets of Data File	Obtain Correct Measures of Growth or TFP for Micro Economic Units[a,c]
Stochastic			
Simple	Yes[b]	Yes[b]	No
Constrained	Yes	No	No
Non-Stochastic			
Generalized	Yes	Yes	Yes
Ratio	Yes	Yes	Yes
Aggregate			
Var-cov	Yes	No	No
Data grouping	No	No	No

[a]See Table 1a for details. Throughout this table, we assume that the variables are in logarithms.

[b]If the variance of the noise is released.

[c]Each scheme will enable the researcher to obtain correct estimates for regressions involving linear transformation of surrogate variables except for the micro-aggregation approach.

REFERENCES

Abbott, Thomas A., Robert H. McGuckin, and Paul Herrick (1988), "Advanced Technology Products and the U.S. Trade Balance," (forthcoming).

Bethlehem, Jelke G. and Jeroen Pannekoek (1988) "Disclosure Control of Micro Data," Proceedings of the Annual Research Conference IV, U.S. Bureau of the Census (forthcoming).

Cox, Lawrence H. and Edwin L. Robinson (1982) "Confidentiality Issues Arising from the Longitudinal Establishment Data File," in Development and Use of Longitudinal Establishment Data, Economic Research Report ER-4, Bureau of the Census, U.S. Department of Commerce, Washington DC: 29-37.

Govoni, J. and P. J. Waite (1985), "Development of a Public Use File for Manufacturing," paper presented at the 1985 Joint Statistical Meetings, American Statistical Association, Las Vegas, Nevada, August 1985.

Griliches, Zvi (October 27, 1985), Internal Memorandum, U.S. Bureau of the Census.

Griliches, Zvi (1986) "Productivity, R&D, and Basic Research at the Firm Level in the 1970's," American Economic Review 76, 1: 141-154.

Griliches, Zvi and Bronwyn Hall (1982) "Census - NSF Data Match Project: A Progress Report," in Development and Use of Longitudinal Establishment Data, Economic Research Report ER-4, Bureau of the Census, U.S Department of Commerce, Washington DC: 51-61.

Kim, Jay (1986), "A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation," Proceedings of the Survey Method Research Section, American Statistical Association, pp. 370-74.

Kokkelenberg, Edward C. and Sang V. Nguyen (1988), "Modelling Technical Progress and Multifactor Productivity: A Plant Level Example," Journal of Productivity Analyses (1988, forthcoming).

Lichtenberg, Frank R. and Donald Siegel "Productivity and Changes in Ownership of Manufacturing Plants," Center for Economic Studies Working Paper, U.S. Bureau of the Census.

McGuckin, Robert H. and Stephen H. Andrews (1987), "The Performance of Lines of Business Purchased in Conglomerate Acquisitions," paper presented at the American Economic Association Meeting in Chicago, December 27-30, 1987.

McGuckin, Robert H. and George Pascoe (1988), "The Longitudinal Research Data Base (LRD): Status and Research Possibilities," paper presented at the Conference on Research in Income and Wealth, May 14, 1988.

McGuckin, Robert H. and Sang V. Nguyen (1988), "Use of Surrogate Files to Conduct Economic Studies with Longitudinal Microdata," Proceedings of the Annual Research Conference IV, U.S. Bureau of the Census (forthcoming).

Monahan, James L. (January 30, 1986), "Development of Microdata Public Use Data File from the Longitudinal Establishment Data File," Internal Memorandum, Center for Economic Studies, U.S. Bureau of the Census.

Nguyen, Sang V. and Edward C. Kokkelenberg (1987), "The Stock of Research and Development Knowledge and Multifactor Productivity Growth," paper presented

at the American Economic Association meeting in Chicago, December 27-30, 1987.

Paass, G. (1985), "Disclosure Risk and Disclosure Avoidance for Microdata," Working Paper, Institute for Applied Information Technique, Gesellschaft fur Matematik und Datenverarbeitung, Sankt Augustin bei Bonn, Federal Republic of Germany.

Phillips, Bruce D. and David A. Hirschberg (1982) "Longitudinal Data For Small Business Analysis," in Development and Use of Longitudinal Establishment Data, Economic Research Report ER-4, Bureau of the Census, U.S. Department of Commerce, Washington DC: 93-106.

Roberts, Mark J. (1988) "Disclosure Research: Discussion," Proceedings of the Annual Research Conference IV, U.S. Bureau of the Census (forthcoming).

Solow, John L. (1987), "The Capital-Energy Complementarity Debate Revisited," American Economic Review 77: 605-614.