Journal of Official Statistics
Vol. 5. No. 1, 1989, pp. 57–67

Assessment of Regression Based Disclosure Risk in Statistical Databases

Michael A. Palley[1]

Abstract: Regression based disclosure can threaten the confidentiality of data stored in statistical databases. This holds true even when the database employs inference controls. This article describes database characteristics that correspond to high and low levels of risk of regression based disclosure. This leads to guidelines for the assessment of a statistical database's degree of risk. Identification of factors that contribute to disclosure risk is a first step in the development of new inference controls.

Key words: Confidentiality; statistical databases; disclosure; database management systems; security.

1. Introduction

The U.S. Census Bureau and other agencies collect data from individuals and later release the information in aggregate form. The respondent relies on the agency's promise of confidentiality when providing responses.

It is well known that promises of confidentiality are difficult to keep. Agencies that provide online statistical databases rely on various inference controls to maintain the confidentiality of collected data.

The literature discusses several inference controls for statistical databases. These include: limiting database responses to those with a minimum acceptable response set size; providing responses based on a random sample of the response set, (Denning (1980)); providing responses based on randomly perturbed (no bias) confidential data ("random data perturbation") (Beck (1980); Traub, Yemini, and Wozniakowski (1984)); and multidimensional transformation of the data matrix (Dalenius and Reiss (1982); Schlörer (1981)). An extensive discussion of the various existing inference controls is found in Chin and Ozsoyoglu (1982).

Inference controls offer only limited protection to the confidentiality of data in a statistical database. Note that restricting a statistical database's responses may conflict with the objective of a statistical database, i.e., to make aggregate statistical information available. Inference controls attempt to maintain the accuracy of database responses while protecting confidentiality.

When the confidentiality of data has been violated, "disclosure" has occurred. A method that can be employed to lead to the disclosure of confidential data is referred to as a "disclo-

[1] Associate Professor, Department of Statistics and Computer Information Systems, Baruch College – CUNY, 17 Lexington Avenue, Box 513, New York, N.Y. 10010, U.S.A.

sure technique." As will be discussed in this article, inference controls have only limited success in maintaining confidentiality when a regression based disclosure technique is applied to the database.

To understand the significance of our problem, it is necessary to further define statistical disclosure. Duncan and Lambert (1987) describe four ways to compromise the confidentiality of microdata that are considered in the literature.

Type I: Identification of a respondent from a released file (Duncan and Lambert call it an "identity disclosure"). (Spruill (1983); Paass (1985); Strudler, Oh, and Scheuren (1986).)

Type II: Obtaining information about a respondent by linking a record to the respondent ("attribute disclosure"). (Cox and Sande (1979).)

Type III: Gaining new information about a respondent from released data, even if no particular record is linked to the respondent, and even if the new information is inexact ("inferential disclosure"). (Dalenius (1974).)

Type IV: Disclosure of confidential information about a population or model, e.g., the relationship between employee characteristics and salary of a group (Palley and Simonoff (1986)). Duncan and Lambert call these "population disclosure" and "model disclosure" respectively.

Duncan and Lambert note that the tax compliance model of the U.S. Internal Revenue Service is a potential target for model disclosure. When a population has relatively few or insignificant outliers, individuals' confidential data can be estimated with these models, leading to inferential (Type III) disclosure as well. These issues are important for statistical agencies.

It has been shown (Palley and Simonoff (1987)) that a regression based technique can lead to inferential disclosure (Type III) as well as population or model disclosures (Type IV). The disclosure risk exists even when the online statistical database disallows the application of regression methodology. This disclosure technique has been shown effective even in the presence of existing inference controls. We identify characteristics that render a statistical database vulnerable to regression based disclosure.

We begin with a brief discussion of online statistical database systems. Section 3 summarizes the technique of regression based disclosure. Section 4 discusses factors that impede the use of the regression based disclosure technique. Section 5 presents assessment guidelines to determine a statistical database's risk of regression based disclosure. Finally, the impact on regression based disclosure control is presented in Section 6.

2. A Statistical Database

NAME	AGE	TITLE	YEARS	EDUC	DEPS	SALARY
JONES	32	VP	10	MBA	3	45 000
SMITH	40	AVP	10	BS	1	36 200
WATSON	27	MGR	2	MBA	0	37 500

Fig. 1. A section of a statistical database

A statistical database contains n records, each having m fields or "attributes" that contain values. Figure 1 presents a small section of a fictional statistical database. Some of these attributes (e.g., NAME) are unique record identifiers, called "keys" (Denning (1978)). Record values for keys are shielded by the database management system (DBMS) in order to prevent identification of described individuals. Other attributes (e.g., SALARY) contain confidential data. The attribute containing confidential data (referred to as the "confidential attribute") which the intruder seeks to disclose will be called "*Xc*."

The statistical database provides aggregate statistics to the user, for example MEAN SALARY = 35230, MEAN SALARY (WHERE AGE < 30) = 32000, etc. The parameters of the query, in our example AGE < 30, is referred to as the "characteristic" of the query (Denning (1978)). Note that a characteristic can involve several non-key attributes. The set of records that conform to a specific characteristic is referred to as the "response set." A statistical database is assumed to provide MEAN, COUNT (e.g., COUNT (where AGE = 30) = 35), and STANDARD DEVIATION (SD of SALARY (where AGE < 30) = 5010) query facilities. A simple technique to turn aggregate database responses into disclosure might be to identify records having unique characteristics. For example, the first record in Fig. 1 is the only record with AGE = 32 AND TITLE = VP AND YEARS = 10. Consequently, a typical inference control is to refuse to answer queries whose response set size is one, or relatively small.

3. Disclosure Using the Regression Based Technique

The technique of disclosing confidential data in a statistical database using regression methodology is developed, and described in detail in Palley (1986) and Palley and Simonoff (1987). Basically, a regression model is derived from the statistical database. The attribute storing confidential data serves as the dependent variable in the regression. Non-confidential, non-key attributes are candidates as predictor variables for the regression model. The model is built under the assumption that the DBMS precludes the direct application of regression methodology to the database. Hence the intruder must apply regression methodology using indirect means.

Based on the intruder's knowledge of predictor variable values (non-confidential "supplemental knowledge") relating to the individual whose confidential data is the "target," a statistical estimate of the target's confidential attribute value can be made. Disclosure of confidential data within a range of values constitutes inferential disclosure of the statistical database (notably Beck (1980); Traub et al. (1984); Denning (1978); Dalenius (1977); and Loynes (1979)). The following is a brief discussion of this technique.

3.1. The disclosure technique

Regression based disclosure involves three steps: selection of candidate predictor variables, generation of characteristic based queries, and derivation of what is called a "synthetic database." A synthetic database will exhibit regression relationships that mimic the statistical database. It is created using legitimate means, i.e., the technique does not violate any of the database's inference controls. Since the statistical database precludes application of regression methodology, the disclosure technique circumvents this control through derivation of a "synthetic" database.

3.1.1. Selection of candidate predictor variables

The technique begins with the selection of the confidential attribute of interest. Other useful

preliminary information is the list of database attributes and the number of records in the database. Henceforth, the statistical database will be referred to as the "actual database" (to be differentiated from the synthetic database). The intruder must then select a set of candidate predicator variables for the regression model from the non-key, non-confidential attributes. Selection is made on the basis of assumed attribute relationships or known correlations between attributes. For each candidate predictor variable, the intruder queries the actual database to create a histogram (i.e., a frequency distribution of values for each candidate predictor variable).

3.1.2. Generation of characteristic based queries

Once frequency tables are formed for each candidate predictor variable, the intruder randomly generates a value for each variable, based on the variable's frequency distribution. A combination of single values for each candidate predictor variable will constitute a characteristic. Each characteristic is used to generate three queries of the actual database, MEAN, COUNT, and STANDARD DEVIATION (SD). As an example, referring to Fig. 1, a possible characteristic might be (AGE = 35–40, TITLE = AVP, YEARS = 17, EDUC = BA, DEPS = 2). Let us call this characteristic P1. This characteristic is used to query the actual database: MEAN SALARY WHERE P1; SD SALARY WHERE P1; COUNT WHERE P1 (called F for frequency). The characteristic and responses to these queries are logged onto a table called the "interim tuple table (ITT)" (see Fig. 2). The strategy is repeated multiple times until an adequate percentage of database records are described. What constitutes an adequate percentage is a research issue and is discussed in Section 4.

/ AGE	/ TITLE	/ YEARS	/ EDUC	/ DEPS	/ SALARY	/ SD	/ F	/
/ 35–40	/ AVP	/ 17	/ BA	/ 2	/ 32 171	/ 5011	/ 3	/
/ 40–45	/ AVP	/ 15	/ MBA	/ 1	/ 42 131	/ 2019	/ 8	/
/								/
/								/

Fig. 2. Interim tuple table

3.1.3. Synthetic database derivation

The next stage of the technique is the derivation of the synthetic database. Once created, the synthetic database is available for regression analysis, or any other type of analysis that the intruder desires. Creation of the synthetic database from the ITT proceeds as follows. The pooled variance (s^2_{pooled}) of the ITT is derived, based on our standard deviation findings for each query response. Each ITT record is copied F times into the synthetic database. For each of the F copies, we vary the value for its confidential attribute by adding random normals distributed over (0, s_{pooled}) to its mean value (from the ITT). The pooled variance is utilized since we seek to simulate the overall variability of the actual database in the synthetic database.

Interim Tuple Table

Characteristic AGE/TITLE/YEARS/ ...	$\bar{X}c$	SD	F
35–40 AVP 17 ...	32 171	5 011	3

Synthetic Database

	Characteristic AGE/TITLE/YEARS ...	$\hat{X}c$
⇒	35–40 AVP 17 ...	29 315
⇒	35–40 AVP 17 ...	37 208
⇒	35–40 AVP 17 ...	35 211

Fig 3. Transformation into the synthetic database

The synthetic database is complete when all of the records in the ITT have been transformed in this way. Regression based disclosure now applies stepwise regression analysis directly to the synthetic database. This regression model, referred to as a "disclosure model," is used to estimate values for the confidential attribute. The estimate is based on the intruder's supplemental knowledge of the non-confidential attribute values for the target individual.

It should be noted that despite the seeming complexity of the approach, the technique could be applied rather easily with the use of a microcomputer. Characteristic generation, and logging the responses onto the ITT occur independently of the actual statistical database. The only points of contact between the intruder and the actual database are the initial building of frequency distributions, and the characteristic based querying of the database. All other functions could be performed on a stand-alone microcomputer at the intruder's convenience.

3.2. The technique as a threat to confidentiality

Palley (1986) and Palley and Simonoff (1987) reported the results of validation of this technique. Regression based disclosure was attempted on several subsamples of the 1980 U.S. Census Microdatabase C-sample for the State of New York. Each of these subsamples was treated as a statistical database. The attribute FAMILY INCOME was considered to be confidential. Characteristic-based queries were applied to the actual databases using our technique. The specific findings are lengthy, and are recounted in detail in Palley and Simonoff (1987).

The analysis considered the following criteria. (a) The degree to which the synthetic database resembled the actual database. This was determined by cross-validating the synthetic database derived disclosure model against the actual database. (b) The quality of estimates of confidential data produced by our disclosure model. Regression based disclosure was found in our analysis to be an effective threat to statistical database confidentiality.

Palley and Simonoff (1987) found that the regression based disclosure technique performed well even in the presence of inference controls. It was shown theoretically that in situations where the "random sample queries" (Denning (1980)) and multidimensional transformation (Dalenius and Reiss (1982)) controls were employed, there was no effect on the performance of the regression based disclosure technique.

It was demonstrated empirically that the control of refusing to answer queries with small response sets had no significant effect on regression based disclosure. This held true until the minimum response set size was set to a threshold level, relative to the size of the database. It was also shown that at the same threshold level, accurate aggregate statistics would be withheld from legitimate users. Therefore, in reality the control offered no protection. Finally, random data perturbation was empirically shown to have little effect on the performance of the regression based disclosure technique. In fact, the research derived an adjustment factor to filter any bias that the perturbation may have added to the synthetic database (assuming the perturbation level of the actual database is known).

These controls failed since they preserve the overall statistical characteristics of the database. These controls must permit the database to answer queries without significantly distorting the responses. If responses are significantly distorted, the database will fail to provide accurate data to legitimate users. As long as accurate responses are provided by the database, regression based disclosure can occur, regardless of these inference controls.

4. Complicators of the Regression Based Approach

This research identifies factors that complicate disclosure of confidential data in a statistical database using the regression based technique. Knowledge of these factors will assist us in the evaluation of the disclosure risk of statistical databases. Future research based on this work may suggest new inference controls that can deter regression based disclosure.

Disclosure using the regression based technique requires the existence of a regression relationship in the actual database. This regression relationship must significantly describe the confidential attribute. The lack of such a statistical relationship will render the disclosure technique harmless. We proceed while assuming the existence of this relationship.

The major factors that complicate disclosure using the regression based technique are: (a) combinatorial explosion of possible characteristics, (b) uniform distribution of actual database records corresponding to the possible characteristics, and (c) minimum response set size that is large relative to the actual database. All of these factors eventually lead to an ITT that describes little of the actual database. Creation of a synthetic database from such an ITT will result in a synthetic database, and consequently a disclosure model, that bears little resemblance to the actual database. These factors are now discussed in detail.

4.1. Combinatorial explosion of possible characteristics

Regression based disclosure begins with the querying of the actual database to build histograms of candidate predictor variables. The next step is the random generation of characteristics used for querying the actual database. Combinatorial explosion, used here, is the presence of a large number of possible characteristics relative to the actual database size. For example, if a characteristic consists of attributes AGE (perhaps 50 possible values), TITLE (10 possible values), and YEARS-WITH-FIRM (30 possible values), then there would be 15 000 potential characteristics. The number of combinations worsens drastically with each additional attribute being used in a characteristic.

When there are a large number of possible characteristics relative to the actual database size, few database records (i.e., individuals described in the database) will conform to any given characteristic, hence small response set sizes. If an inference control that prevents responses to queries with small response sets is employed, many of these queries to the actual database will go unanswered. Even if the small response set queries are answered, the marginal value of asking these queries is relatively low. Considering the risk of detection related to asking many queries, the combinatorial explosion problem is potentially detrimental to the regression based disclosure technique.

There are two possible causes of combinatorial explosion. The first is the presence of many predictor variables in the regression model that exists in the actual database. An intruder's strategy would be to utilize best subset regression in order to limit the number of variables in the model. However, when a regression model has many predictor variables that each contribute relatively little to estimating the confidential attribute (measured by R^2), best subset regression would not be helpful. In addition to combinatorial explosion, a large set of predic-

tor variables requires that the intruder have a great deal of non-confidential data (supplemental knowledge) about the target in order to exploit the disclosure model. The more supplemental knowledge missing, the less useful the disclosure model will be for inferring an individual's confidential data.

Wide domains of predictor variable values would also contribute to combinatorial explosion. An intruder's strategy to remedy this would be to cluster values into subsets, e.g., AGE: 35–40 ... However, some continuous variables may have wide value ranges that will not form meaningful clusters.

4.2. *Interim tuple table insufficiency*

Combinatorial explosion will lead to an insufficient ITT. By insufficient, we mean that the ITT accounts for a small number of actual database records relative to the actual database size.

DATABASE SIZE	374					752				
PCTG OF RECORDS ACCOUNTED IN ITT	63%	50%	30%	20%		68%	61%	52%	47%	41%
PREDICTIVE R^2	.47	.42	.42	−.05		.44	.42	.42	.39	.25

Fig. 4. Relationship of ITT records to predictive R^2

Figure 4 emanates from the Palley and Simonoff (1987) study of two U.S. census subsamples. It is presented for the first time here. As the regression based disclosure technique was applied, the problem of ITT insufficiency was demonstrated. Various ITTs were created from each of two statistical databases. The number of possible characteristics (number of attributes in a characteristic; and number of possible values for each of the attributes in a characteristic) varied in the creation of each of these ITTs. The different number of possible characteristics led to ITTs that accounted for different percentages of records from their respective actual database (middle row of Fig. 4). These ITTs were employed to create disclosure models. The last row of Fig. 4 indicates the quality of the disclosure model derived from each synthetic database, as applied to the actual database, measured as predictive R^2.

Predictive R^2 is a measure of the fit of a regression model created on one set of data as applied to another. The formulation for predictive R^2 is:

$$1 - \frac{\text{Residual sum of squares}}{\text{Total sum of squares}}$$

A "perfect" disclosure model would have no residual sum of squares, and therefore have a predictive R^2 of 1. Since this disclosure model is being applied to a different set of data than it was created on, predictive R^2 can potentially be negative (as seen in Fig. 4). This would be true if the disclosure model is a worse estimator of confidential values than the sample mean.

Figure 4 indicates a relationship between the percentage of database records described in the ITT and the quality of a derived disclosure model. We observe a decline in quality of the disclosure model (measured by the predictive

R^2), as the percentage of actual database records described in the ITT declines.

Part of the reason that ITT insufficiency leads to a poor disclosure model is that an insufficient ITT leads to a synthetic database that has low variability of the confidential attribute. Those ITTs that described relatively few actual database records led to disclosure models that described their synthetic database extremely well ($R^2 \geq 0.8$). However these disclosure models were very poor descriptors of the actual database, hence they had no utility to the intruder.

Two other factors may result in an insufficient interim tuple table, namely uniform distribution of actual database records corresponding to the possible characteristics, and a large minimum response set size control.

4.3. Uniformly distributed characteristic distributions

Disclosure through this technique presupposes that individuals described in the database tend to cluster among a relatively limited subset of the possible characteristics. The regression based technique acts to capture those characteristics that describe a large proportion of the records in the database. For example, let us assume that there are ten thousand records in a statistical database. We will also assume that there are one thousand potential characteristics. Regression based disclosure works relatively well if the database records cluster non-uniformly among a subset of those characteristics. However, if the records cluster uniformly among most of the characteristics, it will take a prohibitive number of database queries in order to build a sufficient ITT. Furthermore, if there is a uniform distribution, the typical response set size will be relatively small. This could cause problems if there is a minimum response set size.

The intruder's strategy would be analogous to the strategy for combinatorial explosion: to reduce the number of possible characteristics (i.e., by reducing the number of attributes comprising a characteristic, or by grouping characteristic attribute values), hoping that distributions among these fewer combinations of characteristics are less uniform. However, if database records remain relatively uniformly distributed among the new possible characteristics, the intruder will not be able to solve this problem.

4.4. High minimum response set size

If a statistical database employs an inference control that refuses queries with a minimum response set size that is large relative to the database size, many queries will go unanswered. This will result in ITT insufficiency. It is noted that the strategy of raising minimum response set size past a point is a "two-edged sword." The strategy will protect against regression based disclosure, but only at the expense of failing to provide the legitimate user with useful aggregate statistics.

5. Risk Assessment Guidelines

At this stage, we seek to assess the risk of regression based disclosure for a statistical database. The assessment guidelines generally parallel our discussion of the complicators of disclosure. A diagram of factors that contribute to the risk of regression based disclosure is presented in Fig. 5.

Here risk is described qualitatively. There currently exists no means of quantifying the relative risk level. It is proposed that agencies that provide online statistical database facilities can assess the level of regression based disclosure risk by answering the following.

A. Does a regression relationship exist in the database, with a confidential attribute acting as dependent variable?

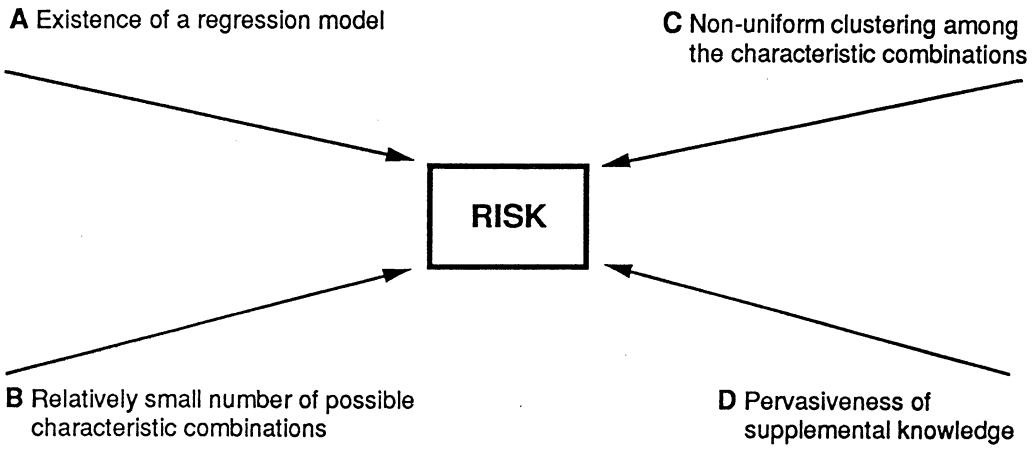

Fig. 5. Assessment model

Upon identification of those attributes which contain confidential data, the agency should perform correlational analysis to identify candidate non-confidential, predictor variables. Next, perform stepwise regression on the database. Existence of a sufficiently high R^2 regression model (sufficiency to be determined by the agency), and pervasiveness of supplemental knowledge (data for non-confidential attributes for individuals in the database) would be indicators of disclosure risk.

B. Is there a small number of characteristics relative to the database size?

B.1. Does the regression model require few predictor variables?

A minimum "safe" number of predictor variables is a function of the size of the database. As a general rule, the larger the database, the more predictor variables would be necessary to cause ITT insufficiency. In addition, a large number of predictor variables places an added burden on the intruder for extensive supplemental knowledge. The fewer predictor variables necessary for a disclosure model, the more at risk the statistical database. Note, as discussed, combinatorial explosion, based on too many predictor variables in a disclosure model, can be remedied by various intruder strategies. To assess risk under these strategies, an agency can assess the quality of regression models (in terms of R^2) that involve fewer predictor variables.This can be facilitated with best subset regression.

B.2. Do candidate predictor variables have few possible values?

The larger the domain of predictor variable values, the more difficult it is to create a disclosure model. A large number of predictor variable values will lead to a large set of possible characteristics, contributing to ITT insufficiency. Again, this is defined relative to database size. The ability to cluster wide variable value spreads into ranges is an effective way for the intruder to counter the combinatorial explosion problem. In order to assess the risk of this, an agency might try recording values into ranges, and test if the disclosure model remains relatively effective.

C. Are distributions of records among characteristics non-uniform?

The less uniform the distribution of database records among characteristics, the more regression based disclosure is facilitated. Uniform distributions result in query responses with small counts, making it difficult to adequately describe a large portion of the actual database with a reasonable number of queries (ITT insufficiency). An insufficient ITT leads to little variability of the confidential attribute data in the synthetic database, and likewise to a disclosure model that fails to describe the actual database. A database whose records cluster among relatively few characteristics has greater risk of regression based disclosure.

D. Is non-confidential data for the regression model's predictor variables generally available?

Inferential disclosure of a statistical database requires the intruder's knowledge of his target's values for predictor variables (supplemental knowledge). Alternative strategies are available to an intruder who has incomplete supplemental knowledge. The intruder may create a disclosure model involving only those predictor variables for which he has supplemental knowledge. Another strategy would be for the intruder to estimate missing values. Palley and Simonoff (1987) found that when supplemental knowledge was lacking for one or two (out of five) predictor variable values, an intruder could still perform inferential disclosure effectively. However, the higher a given predictor variable's t-value (measure of that variable's contribution to the regression model) in the disclosure model, the more impact its value's absence. Clearly, the more available the non-confidential data, the more disclosure risk.

These disclosure risk criteria are not necessarily exhaustive. It is possible that future research will yet determine other risk criteria.

6. Final Remarks

The notions of population disclosure and model disclosure run counter to the perceived goals of statistical agencies. Statistical agencies make information, and therefore, regression relationships publicly available. Legitimate users have a need for information. However, regression based disclosure can turn seemingly benign types of information into breaches of confidentiality. This is particularly a problem when a model disclosure is parleyed into inferential disclosure of an individual's information. This is a problem posed by the regression based disclosure technique.

A regression based technique has been found to defy existing inference controls. We have identified some critical factors that influence the risk posed by regression based disclosure. The reduction of statistical database disclosure risk has been investigated by, among others, Duncan and Lambert (1986 and 1987); Cox and Sande (1979); Dalenius and Reiss (1982); Traub et al. (1984). Nevertheless, the existing research does not specifically address disclosure risk posed by regression based techniques. Drastic controls, such as refusing to provide standard deviation responses, would also reduce the utility of the statistical database to legitimate users. Replacing standard deviation responses with minimum and maximum bounds, besides limiting the information available to legitimate users, would also pose new disclosure risks. Consequently, there are no simple solutions.

Nevertheless, identification of the critical factors in regression based disclosure is a step in the development of further controls. Future research will continue to address these problems. In the interim, this research highlights some limitations in our ability to protect the confidentiality of collected statistical data in online databases.

7. References

Beck, L. L. (1980): A Security Mechanism for Statistical Databases. ACM Transactions on Database Systems, 5, pp. 316–338.

Chin, F. Y. and Ozsoyoglu, G. (1982): Auditing and Inference Control in Statistical Databases. IEEE Transactions on Software Engineering, SE-8, 6, pp. 574–582.

Cox, L. H. and Sande, G. (1979): Techniques for Preserving Statistical Confidentiality. Proceedings of the 42nd Meeting of the International Statistical Institute, Manila, December, 1979.

Dalenius, T. (1974): The Invasion of Privacy Problem and Statistics Production – An Overview. Statistisk tidskrift, 12, pp. 213–225.

Dalenius, T. (1977): Towards a Methodology for Statistical Disclosure Control. Statistisk tidskrift, 15, pp. 429–444.

Dalenius, T. and Reiss, S. (1982): Data-Swapping: A Technique for Disclosure Control. Journal of Statistical Planning and Inference, 6, pp. 73–85.

Denning, D. (1978): A Review of the Research of Statistical Database Security. In Foundations of Secure Computation, edited by R. DeMillo et al., Academic Press, New York.

Denning, D. (1980): Secure Statistical Databases with Random Sample Queries. ACM Transactions on Database Systems, 5, pp. 291–315.

Duncan, G. and Lambert, D. (1986): Disclosure-Limited Data Dissemination. Journal of the American Statistical Association, 81, pp. 10–18.

Duncan, G. and Lambert, D. (1987): The Risk of Disclosure for Microdata. Statistics of Income and Related Administrative Record Research: 1986–1987. Internal Revenue Service Publication No. 1299, Government Printing Office, Washington, D.C., pp. 325–332.

Loynes, R. M. (1979): Discussion of the Papers by Professor Dalenius and Professor Durban. Journal of the Royal Statistical Society, Series A, 142, pp. 325–326.

Paass, G. (1985): Disclosure Risk and Disclosure Avoidance for Microdata. Paper presented at International Association for Social Service Information and Technology, May 1985.

Palley, M. A. (1986): Security of Statistical Databases: Compromise Through Attribute Correlational Modeling. Proceedings of the Second International Conference on Data Engineering, IEEE Computer Society, Washington, D.C., pp. 67–74.

Palley, M. A. and Simonoff, J. S. (1986): Regression Methodology Based Disclosure of a Statistical Database. American Statistical Association, Proceedings of the Section on Survey Research Methods, pp. 382–387.

Palley, M. A. and Simonoff, J. S. (1987): The Use of Regression Methodology for the Compromise of Confidential Information in a Statistical Database. ACM Transactions on Database Systems, 12, pp. 593–608.

Schlörer, J. (1981): Security of Statistical Databases: Multidimensional Transformation. ACM Transactions on Database Systems, 6, pp. 95–112.

Spruill, N. (1983): The Confidentiality and Usefulness of Masked Business Microdata. American Statistical Association, Proceedings of the Section on Survey Research Methods, pp. 602–607.

Strudler, M., Oh. H. L., and Scheuren, F. (1986): Protection of Taxpayer Confidentiality with Respect to the Tax Model. American Statistical Association, Proceedings of the Section on Survey Research Methods, pp. 375–381.

Traub, J. F., Yemini, Y., and Wozniakowski, H. (1984): The Statistical Security of a Statistical Database. ACM Transactions on Database Systems, 9, pp. 672–679.

Received September 1987
Revised March 1989