REGRESSION METHODOLOGY BASED DISCLOSURE OF A STATISTICAL DATABASE

Michael A. Palley, Baruch College - CUNY, and Jeffrey S. Simonoff, New York University

ABSTRACT

A statistical database serves two major purposes: to provide the statistician with accurate aggregate statistical information, and to protect the confidentiality of individual database records. A technique is presented which utilizes regression methodology to compromise confidential information in a statistical database. In the case that a database management system precludes application of regression methodology, the research introduces the notion of a "synthetic database", created through legitimate means, which circumvents this control, and once again permits disclosure through regression methodology.

The approach is validated on various subsets of the 1980 U.S. Census microdata for the State of New York. Finally, the regression methodology approach is examined in its ability to cause disclosure even where various existing confidentiality protection measures are in effect.

I. INTRODUCTION

Statistical databases often maintain sensitive or confidential information. A statistical database is a database "from which aggregate information about large subsets of entities of an entity set is to be obtained, such as a database of census data..." [15]. Typical queries of the data are SUM, COUNT, and MEAN, of measured data, for subsets of the population.

Major statistical databases are maintained by federal and state governments, commercial organizations, non-profit organizations, etc. Burnham [4] states that approximately one-half of the workforce is employed by large corporations that collect detailed information, some of which is sensitive. This includes salary information, work evaluations, personnel references and the like. According to Burnham, two thirds of Americans have life insurance and nine tenths carry health insurance, both of which result in the collection of confidential information onto a database.

An overview of the problem of managing the inevitable conflict between society's need to know and process information, and the individual's right to privacy is found in [5]. The statistical societies have recognized this problem as well. In 1975, an Ad Hoc Committee on Privacy and Confidentiality was created by the American Statistical Association [1,2]. More recently, the Institute of Mathematical Statistics has encouraged its members to contribute to the discussion of this issue [8].

In this article, the use of regression methodology to compromise a statistical database is examined. This issue is related to, but not identical to, the widely recognized problem of preserving confidentiality in a released microdata file [13]. In the present situation, a user does not have access to an entire database; rather a user is permitted to ask queries of the database (controlled through a database management system (DBMS)) to acquire aggregate statistical information. As a fundamental control, the DBMS will not allow response to a query whose response set size is one. This fairly trivial control is easily undermined; an example is available in [9].

Section 2 describes an example of compromise using regression methodology. The approach is described in a situation where the DBMS precludes direct regression of the statistical database. Regression based disclosure is validated through application to subdatabases of the 1980 U.S. Census Microdata samples for the State of New York. Section 3 describes existing strategies that seek to deter the disclosure of confidential information contained in a statistical database as well as their impact on regression-based disclosure. Conclusions and recommendations are presented in Section 4.

II. INVADING PRIVACY USING REGRESSION METHODOLOGY

A statistical database contains a subset of the Cartesian product of domains of the attributes $(X_1, X_2, ..., X_m)$. The database is assumed to have n records (or tuples). Certain of the m attributes will be identifiers whose values are suppressed by the basic rules of the statistical database. That is, without loss of generality, $X_1, X_2, ... X_i$ will uniquely identify a record. These unique identifiers are not accessible by the statistical database user.

Application of regression to a statistical database is not a routine matter since we assume that the database management system precludes use of regression methodology. A database would refuse such strategies since a fitted regression based on the database could be used to predict confidential values in the database (as demonstrated later). As a result, some form of intermediate step is required. Our strategy shows that it is possible to circumvent database confidentiality controls by building a "synthetic database" which resembles the actual statistical database as much as possible. This synthetic database is created completely through legitimate queries (COUNT, MEAN, STANDARD DEVIATION) of the database, and through a minimum of contact with the actual statistical database. The synthetic database is then used to create a regression model (referred to as a "disclosure model") which is used to predict confidential variable values based on supplemental knowledge of non-confidential predictor variable values.

The creation of a synthetic database involves three steps: creation of histograms for candidate predictor variables in a disclosure model, querying of the database based on randomly generated key values, and transition to the synthetic database.

Creation of Histograms for Candidate Predictor Values in a Disclosure Model

Assume that a potential intruder seeks to learn the confidential salary of an individual in a database. The intruder would begin with the selection of several candidate variables that can be used to predict salary in a regression model. Selection is made on the basis of expectations, or supplemental knowledge (e.g. known correlations between variables).

The intruder asks queries of the form "COUNT WHERE X_a=value". This is repeated until all of the variable's values have been queried. Where there is a large domain of variables, value ranges can be used (e.g. AGE=30-35 etc.).

Querying of the Database Based on Randomly Generated Key Values

Assume that there are five variables selected as candidate predictor variables. Statistical database queries for salary are then based on a key composed of values for these five variables. A value for each of these variables is generated randomly from the above derived histograms, now serving as probability density functions. For each selected key value, queries of MEAN, COUNT, and STANDARD DEVIATION are applied to the statistical database. In our examples, the queries would be "MEAN SALARY WHERE X_1=value$_1$, X_2=value$_2$, ... X_k =value$_k$;" "STANDARD DEVIATION OF SALARY WHERE {key value} ;" "COUNT WHERE {key value} ." The key value, and the query responses are logged onto a table called the "interim tuple table". The process is repeated multiple times. A sample interim tuple table is shown at the top of Figure 1.

The issue of stopping point is discussed in detail in [9], and [10]. Other issues, principally the possible "combinatorial explosion" of key values are discussed there as well. In general we expect decreasing marginal utility of queries. Note that at the end of this stage, all contact between the intruder, and the actual statistical database ceases.

Transition to the Synthetic Database

Finally, the interim tuple table is used to create the synthetic database. In order to reflect variability of the confidential variable in the actual statistical database, a pooled estimate of the variance of the confidential attribute is calculated based on the standard deviation responses recorded in the interim tuple table. A random normal (0, s^2 pooled) is added to each mean confidential variable value f times, where f is the frequency of the key value in the actual database. This is repeated for all entries in the interim tuple table. This constitutes the completion of the synthetic database. The construction is demonstrated in Figure 1.

Once complete, the intruder is now free to apply regression to the synthetic database, unhindered by the restrictions imposed by the DBMS, and with little chance of being detected. A disclosure model is now built on the synthetic database, using stepwise regression, to model the behavior of the confidential variable. Ultimately the model is used to make predictions based on supplemental knowledge of the non-confidential predictor variables.

Validation of Approach

Clearly, the synthetic database model is feasible if it is possible to build a disclosure model on the synthetic database, which adequately describes the actual statistical database.

The approach was validated on subsamples of the United States Census Database, State of New York, C-sample, 1980. Due to hardware constraints, four subsamples of the microdata sample were used, corresponding to approximately 2,500 household records.

Family income for the year 1979 served as the confidential variable. Details of the database prototype and the "intruder" prototype system are described in [9]. Complete discussion of the derivation of the disclosure model, and its performance are described there, and in [10].

The research began by establishing the existence of a disclosure model derived directly on the statistical database. This was done to create a benchmark for

Figure 1 - Transition from the Interim Tuple Table to the Synthetic Database

X_1	X_2	.	.	X_k	F	$\bar{X}^c$	σ
$X_{1,i}$	$X_{2,i}$	.	.	$X_{k,i}$	F_i	$\bar{X}^c_i$	σ_i
.	.	.	.	.	.	.	.
$X_{1,j}$	$X_{2,j}$	.	.	$X_{k,j}$	F_j	$\bar{X}^c_j$	σ_j
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.

Interim Tuple Table

X_1	X_2			X_k	X^c
.	.	.	.	.	.
.	.	.	.	.	.
$X_{1,j}$	$X_{2,j}$	.	.	$X_{k,j}$	$\bar{X}^c_j + N_{j1}(0,\sigma^2_{\text{pooled}})$
$X_{1,j}$	$X_{2,j}$	.	.	$X_{k,j}$	$\bar{X}^c_j + N_{j2}(0,\sigma^2_{\text{pooled}})$
.	.	.	.	.	.
$X_{1,j}$	$X_{2,j}$	.	.	$X_{k,j}$	$\bar{X}^c_j + N_{j,F_j}(0,\sigma^2_{\text{pooled}})$
.	.	.	.	.	.
.	.	.	.	.	.
$X_{1,j+3}$	$X_{2,j+3}$			$X_{k,j+3}$	$\bar{X}^c_{j+3} + N_{j+3,1}(0,\sigma^2_{\text{pooled}})$
.	.	.	.	.	.

Synthetic Database

the comparison of a disclosure model to be built on the synthetic database. The best directly derived regression model was built on six non-confidential predictor variables. The model had an r-squared of .57, and highly significant f-statistic of 69.13 (df=7,366). This was after removal of nine outliers. These were either extremely low observations of family income (near zero or negative), or at or about the Census cutoff point of $75,000. A square root transformation was employed to remove heteroskedasticity.

At this point, the database management system precluded direct use of regression methodology. The synthetic database was then derived as described. Histograms were derived for six candidate predictor variables. Our research imposed a limit of 300 keys used for mean, standard deviation, and count queries. In order to limit the number of possible key combinations, one variable was dropped. Three hundred out of a possible 768 key values (built on five variables) were generated. After logging onto the interim tuple table, the synthetic database was created with random normals generated by IMSL routine GGNML.

The derived disclosure model was then cross-validated onto the actual statistical database in order to assess its ability to make predictions in the actual database. Recognizing that a square root transformation might be more appropriate, the synthetic database methodology was modified to incorporate transformations. Using the well-known Taylor Series approximations, the means and standard deviations of the transformed salary were approximated using the actual means and standard deviations. The result of cross validating this model onto the actual database is shown in the first column of Table 1. Predictive r-square of the disclosure model was .475. The result without the Taylor Series transformation was .454. These compare with the second column of Table 1, which shows the best regression had the DBMS allowed direct regression of the actual statistical database. The r-squared there is .54, indicating strong performance of the synthetic database creation strategy.

Table 1 - Cross Validation of Regression Models Derived from Synthetic Database onto Actual Statistical Database (SDB)

Item	Synthetic-Based Model Applied to Actual SDB	Directly-Derived Model
Sample Size:	374	374
Mean Salary:	21949	21891
Std. Deviation:	14882	14904
Residual SS:	4.339 E10	3.823 E10
Sum of Squares:	8.260 E10	8.308 E10
1-(RESID SS/SS):	.475	.540
(RESID SS/N):	1.160 E8	1.019 E8

Analysis of a larger database brought similar results. Again, using the Taylor Series transformation, a synthetic database derived disclosure model had predictive r-squared of .44 compared to a directly derived model with r-squared .50. Thus, it seems clear that our synthetic database compromise strategy is capable of compromise, even where the DBMS precludes direct use of regression methodology.

III. EXISTING CONFIDENTIALITY PRESERVATION STRATEGIES

Shoshani [12] divides existing statistical database confidentiality preservation strategies into five classes: limiting the response set size, limiting the intersection of response sets, random sample queries, partitioning of the database, and perturbation of data values. An additional strategy is multidimensional transformation, also known as "data swapping" [6].

Two of these strategies have not received much consideration in statistical database research. The first, limiting the intersection of response sets would require query logging (in order to identify intersecting queries). For most applications, this technique is operationally infeasible due to the necessity for a great deal of memory and CPU time for the maintenance and checking of the log. Partitioning of the database, another approach, is likely to over-restrict the legitimate user's access to statistical information [3].

Limiting the Response Set Size

Any statistical database will enforce a trivial control of the refusal to answer queries whose response set size is one. One of the first confidentiality strategies identified was the additional refusal of queries with small response set sizes (defined relative to the database size). Clearly, the larger the minimum response set size, the less a statistical database will achieve its need of providing useful aggregate statistical information for the legitimate user. Another problem with the approach is that this countermeasure has been shown to be subverted by a strategy called the "tracker" [11,7].

Our research found that this approach hindered regression methodology based disclosure only at a point where the minimum response set size was so large that it disrupted the ability of the database to provide useful statistics to the legitimate user. This threshold minimum response set size is a function of the database size, and number of variables composing a key. This is discussed in detail in [10].

Random Sample Queries

The random sample query strategy can be effective in certain database situations. Basically, rather than answering a database query (such as MEAN salary) by polling the entire response set (all cases meeting given criteria), the database will calculate the mean based on a random sample of the response set. The particular sample used to answer that query is logged for use upon later repetition of the identical query. In the absence of such logging, the query could be repeated multiple times in order to filter any difference between sample and full response set queries.

Denning [7] found that in order to maintain the database goal of providing useful aggregate statistics, samples must include a fairly large proportion of the complete response set. Superior results were found in the case of especially large databases. Statistics based on large response sets were found to be more accurate, and less likely to be compromised than statistics based on small response sets. As long as aggregate statistics remain representative of the full response set, this countermeasure will not impact regression methodology based compromise. When aggregate statistics are not representative, then both the ability to provide useful statistics to the legitimate user, as well as the compromise methodology are disrupted.

Random Data Perturbation

Data perturbation refers to the varying of database stored values by random error. Although individual observations of confidential data are perturbed, aggregate statistics should remain basically unchanged for "adequate" sized response sets, as long as the random error has no bias. One technique of varying data would be the additive perturbation of data, i.e., adding random errors (with mean 0) to the database values. This suffers from the problem of scale. A small perturbation to a large value (e.g., modifying salary from $100,000 to $105,000) would offer very little by way of confidentiality. Traub, et al. [14] propose the use of multiplicative perturbation, as a means of overcoming this problem of scale.

Palley and Simonoff [10] and Palley [9] offer an extensive discussion of the impact of this countermeasure on regression based disclosure. Perturbation is intended not to impact aggregate statistics. Therefore, it is found that the approach has little, if any, impact on our ability to compromise a database using regression methodology. It is shown that assuming that the intruder knows the level of perturbation, any possible bias of the aggregate statistics can be filtered. It is highly reasonable that the legitimate statistical user be informed of the database perturbation level. Palley [9] tested the effects of perturbation on regression based compromise on subsets of the Census database. It was found that the compromise strategy worked even where there was a multiplicative perturbation of the database with random normals (0,.25). In this case, even with fairly extreme perturbation, compromise was possible without adjusting for any bias.

Data Swapping

The data swapping technique visualizes the statistical database as a very large matrix. If a query involves only one database attribute (e.g., COUNT WHERE AGE<30), it is referred to as a "1-order" statistic. A query that involves two variables (e.g. COUNT WHERE AUTOS=1 AND AGE<30) involves a "2-order" statistic. The data swapping technique [6] transforms the data matrix in such a way that k-order statistics are preserved. The technique suffers from a major problem in that no methodology has been devised to transform databases while preserving k-order statistics. The difficulty aggravates when we are dealing with matrices the size of a large statistical database e.g., hundreds or thousands of records.

Palley and Simonoff (1986) discuss that where the data swap preserves k-order statistics, regression based compromise is completely unaffected as long as the disclosure model requires, at most, k-1 predictor variables.

IV. CONCLUSIONS AND RECOMMENDATIONS

The shown regression-based methodology can compromise confidential information in a statistical database. The threat exists despite a database control precluding the direct application of regression methodology. The creation of a "synthetic database" facilitates disclosure. Regression based disclosure of a statistical database appears robust against our existing confidentiality preservation strategies, most notably the data perturbation approach.

It is not clear at this time, to what extent the threat is generalizable. In theory, and as validated against these subsamples of the Census database, the methodology tells the intruder information that would be deemed confidential by both database administrators, and by those described by the data.

Palley and Simonoff [10] identify various database characteristics that seem to increase and decrease risk of regression based disclosure. However, it is not clear at this time what measures can be taken to confound the performance of the regression based approach, on a "high risk" database. It is hoped that the awareness of the shortcomings of the existing confidentiality preservation strategies may lead to the development of more functional ones.

NOTE

Michael A. Palley is Assistant Professor of Computer Information Systems at

Baruch College - CUNY, 17 Lexington Avenue, Box 513, New York, N.Y. 10010.

Jeffrey S. Simonoff is Assistant Professor of Statistics at New York University, Graduate School of Business Administration, 100 Trinity Place, New York, N.Y. 10006.

REFERENCES

[1] American Statistical Association, "Report of the ASA Ad Hoc Committee on Privacy and Confidentiality", Amer. Statist., 31, 1977, pp. 59-78.

[2] _____, "Business directories: findings and recommendations of the ASA Committee on Privacy and Confidentiality", Amer. Statist., 34, 1980, pp. 8-10.

[3] Beck, L.L., "A security mechanism for statistical databases," ACM TODS, 5, 1980, pp. 316-338.

[4] Burnham, D., The Rise of the Computer State. New York: Random House, 1983, pp. 49-87.

[5] Dalenius, T., "The invasion of privacy problem and statistics production - an overview", Statistik Tidskrift, 12, 1974, pp. 213-225.

[6] Dalenius, T. and Reiss, S., "Data-swapping: a technique for disclosure control", Journal of Statistical Planning and Inference, 6, 1982, pp. 73-85.

[7] Denning, D., "Secure statistical databases with random sample queries", ACM TODS, 5, 3, 1980, pp. 291-315.

[8] Kempthorne, O., "President's Column", Institute of Mathematical Statistics Bulletin, 14, 1985, pp. 165-166.

[9] Palley, M.A., Security of Statistical Databases - Invasion of Privacy through Attribute Correlational Modeling. Ann Arbor, MI: University Microfilms International, 1986. (Ph.D. Dissertation, New York University Graduate School of Business Administration)

[10] Palley, M.A. and Simonoff, J.S., "The Use of Regression Methodology for the Compromise of Confidential Information in Statistical Databases", Working Paper #86-6-STAT, Baruch College - CUNY, 1986.

[11] Schlorer, J., "Disclosure from statistical databases: quantitative aspects of trackers", ACM TODS, 5, 1980, pp. 467-492.

[12] Shoshani, A., "Statistical databases: characteristics, problems, and some solutions", Proceedings of the Conference on Very Large Databases (VLDB), 1982, pp. 208-222.

[13] Spruill, N.L., "The confidentiality and analytic usefulness of masked business microdata", ASA Proceedings Sec. Survey Research Methods, 1983, pp. 602-607.

[14] Traub, J.F., Yemini, Y., and Wozniakowski, H., "The statistical security of a statistical database", ACM TODS, 9, 1984, pp. 672-679.

[15] Ullman, J.D., Principles of Database Systems. Rockville, MD: Computer Science Press, 1982.