NISS

Data Quality and Data Confidentiality for Microdata: Implications and Strategies

Alan F. Karr and Ashish P. Sanil

Technical Report Number 149
December 2004

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

Data Quality and Data Confidentiality for Microdata: Implications and Strategies

Alan F. Karr and Ashish P. Sanil
National Institute of Statistical Sciences
PO Box 14006
Research Triangle Park, NC 27709-4006 USA
{karr, ashish}@niss.org

National statistical agencies (and other organizations) must fulfill two nearly contradictory missions. On the one hand, they must extract and disseminate—to other agencies, the research community and the public—useful information derived from sample surveys and censuses. But they must also protect the confidentiality of the data and the privacy of data subjects. Protecting confidentiality may be mandated by law, prescribed by agency practices or promised to respondents. Often, confidentiality must be preserved in order to ensure the quality of the data: respondents do not answer truthfully if they believe that their privacy is threatened.

In this paper we describe two formulations that balance data quality and disclosure risk. These formulations can inform the strategies used by agencies to construct microdata releases. The first, for data swapping, uses explicit quantitative measures of data quality—the utility of the released microdata— and disclosure risk to produce a risk-utility frontier of undominated candidate releases, to which the agency can restrict its attention. Given a "utility function" that trades off data quality for disclosure risk, an optimal release can be identified. The second, and rather different, setting is integration of distributed databases. There, arguably the quality is zero unless analyses that seem to but actually do not require integration of the data can be conducted safely.

Other risk–utility formulations have been devised for regressions [7], tabular data [3, 4] and other settings [5, 6, 12].

Data Swapping

Data swapping [2, 13] is a technique for statistical disclosure limitation (SDL) that protects confidentiality by modifying a fraction of the records in a database by exchanging a subset of attributes between selected pairs of records. Data swapping makes it impossible for an intruder to be certain of having identified an individual or entity in the database, because no record is certain to be unaltered.

Formulated as a decision problem [8], data swapping entails selection of one or more swap attributes and the swap rate, the fraction of records for which swapping occurs. More complex versions of the problem allow constraints on unswapped attributes. For example, an unswapped attribute may be forced to remain unchanged—preventing swapping across geographical boundaries, for example—or forced to change. Let $\mathcal{D}_{\text{pre}}$ and $\mathcal{D}_{\text{post}}(R)$ denote the pre-swap and post-swap contingency tables associated with the data.

Let $\mathcal{R}$ be the set of candidate data releases. To implement the risk–utility formulation, the agency must define both a *disclosure risk measure*—a function $\mathbf{DR} : \mathcal{R} \to \mathbb{R}$ with the interpretation that $\mathbf{DR}(R)$ is the disclosure risk associated with the release R, and a *data utility measure*—a function $\mathbf{DU} : \mathcal{R} \to \mathbb{R}$ with the interpretation that $\mathbf{DU}(R)$ is the utility of the release R.

In [8], which treats only categorical data, disclosure risk focuses on small count cells in the contingency table created by using all attributes in the data. One such measure is derived from the n-rule, which is widely used in SDL [13]. Disclosure risk is the proportion of unswapped records in small count cells in $\mathcal{D}_{\text{post}}(R)$:

$$(1) \qquad \mathbf{DR}(R) = \frac{\sum_{C_1,C_2} \text{Number of unswapped records in } \mathcal{D}_{\text{post}}(R)}{\text{Total number of unswapped records in } \mathcal{D}_{\text{post}}(R)},$$

where C_1 and C_2 are the cells $\mathcal{D}_{\text{post}}(R)$ with counts of 1 and 2 respectively. Other measures are based on the ease with which swapped records can be linked to an external database, as in [14].

In [8], data utility is the negative of *data distortion*, the latter given by

$$\mathbf{DD}(R) = d(\mathcal{D}_{\text{pre}}, \mathcal{D}_{\text{post}}(R)), \tag{2}$$

where d is a metric on an appropriate space of distributions. Specific measures include Hellinger distance, total variation distance and entropy change.

It is also possible [8] to define measures of data utility that account explicitly for quality of inferences drawn from the data using log-linear models [1]. Let $\mathbf{M}^* = \mathbf{M}^*(\mathcal{D}_{\text{pre}})$ be the "optimal" log-linear model of the pre-swap database $\mathcal{D}_{\text{pre}}$, according to some criterion, for example, the Akaike information criterion (AIC) or Bayes information criterion (BIC). Concretely, $\mathbf{M}^*$ can be thought of in terms of its minimal sufficient statistics—the set of marginal subtables of the contingency table associated with $\mathcal{D}_{\text{pre}}$ representing the highest-order interactions present. Let $\mathcal{L}_{\mathbf{M}^*}(\cdot)$ be the log-likelihood function associated with $\mathbf{M}^*$. Then as measure of data utility one can employ the log-likelihood ratio

$$\mathbf{DU}_{\text{llm}}(R) = \mathcal{L}_{\mathbf{M}^*}(\mathcal{D}_{\text{post}}(R)) - \mathcal{L}_{\mathbf{M}^*}(\mathcal{D}_{\text{pre}}); \tag{3}$$

the llm subscript abbreviates "log-linear model." The rationale is that higher values of $\mathbf{DU}_{\text{llm}}(R)$ indicate that $\mathbf{M}^*$ remains a good model for $\mathcal{D}_{\text{post}}(R)$.

Given disclosure risk and data utility measures, the decision problem can be solved in two distinct ways. The first is *utility maximization*: the optimal release R^* is chosen that maximizes data utility subject to an upper bound constraint on disclosure risk:

$$\begin{aligned} &R^* = \arg\max\nolimits_{R \in \mathcal{R}} \mathbf{DU}(R) \\ &\text{s.t. } \mathbf{DR}(R) \leq \alpha. \end{aligned} \tag{4}$$

A more flexible method is to define *risk–utility frontiers* using the partial order $\preceq_{\text{RU}}$ defined by

$$R_1 \preceq_{\text{RU}} R_2 \Leftrightarrow \mathbf{DR}(R_2) \leq \mathbf{DR}(R_1) \qquad \text{and} \qquad \mathbf{DU}(R_2) \geq \mathbf{DU}(R_1). \tag{5}$$

If $R_1 \preceq_{\text{RU}} R_2$, then clearly R_2 is preferred to R_1 because it has both lower disclosure risk and higher greater utility. Only release candidates on the risk–utility frontier of maximal elements of $\mathcal{R}$ with respect to the partial order (5) need be considered further. Calculation of the frontier can be done using existing algorithms for finding the maxima in a set of vectors.

Selection of a release on the risk–utility frontier can be done by assessing the risk–utility balance subjectively or quantitatively, by means of an objective function that relates risk and utility. Similar approaches have been used in economics to maximize consumer utility for the purchase of a combination of two commodities.

Data Integration

In a totally different setting, many scientific and policy investigations require statistical analyses that "integrate" data stored in multiple, distributed databases. But, the barriers to actually integrating the databases are numerous. One is confidentiality; others are regulation, proprietary data and scale.

Absent the ability to integrate the data, their (joint) quality diminishes. For many analyses it is not necessary to move the data to a single location. Instead, using techniques from computer science known generically as secure multi-party computation [15], agencies can share summaries of the data anonymously, but in a way that the analysis can be performed in a statistically principled manner.

We illustrate for linear regression on "horizontally partitioned data" [9, 10]. There are $K > 2$ agencies, each with the same numerical data on its own n_j data subjects—p predictors X^j and a response y^j, and that the agencies wish, *without sharing their data with each other or a trusted third party*, to fit the usual linear model

$$X\beta +$$

to the “global” data

$$X = \begin{bmatrix} X^1 \\ \vdots \\ X^K \end{bmatrix} \qquad \text{and} \qquad y = \begin{bmatrix} y^1 \\ \vdots \\ y^K \end{bmatrix}.$$

We embed the constant term of the regression in the first predictor: $X_1^j \equiv 1$ for all j.

Several assumptions about agency behavior are necessary. First, the agencies agree to cooperate to perform the regression, and none of them is specifically interested in breaking the confidentiality of the others’ data. Second, each reports accurately the results of computations on its own data, and follows the agreed-on computational protocols, such as secure summation, properly. And finally, there is no collusion among agencies.

Only one concept from secure multi-party computation is needed, that of secure summation—the agencies want to sum values v_j in a manner that lets each agency j learn only the minimum possible about the other agencies’ values, namely, the sum $v_{(-j)} = \sum_{\ell \neq j} v_\ell = v - v_j$. The secure summation protocol is almost more complicated to describe than to implement. Number the agencies $1, \ldots, K$. Agency 1 generates a very large random integer R, adds R to its value v_1, and sends the sum to agency 2. Since R is random, Agency 2 learns effectively nothing about v_1. Agency 2 adds its value v_2 to $R + v_1$, sends the result to agency 3, and so on. Finally, agency 1 receives $R + v_1 + \cdots + v_K = R + v$ from agency K, subtracts R, and shares the result v with the other agencies. Here cooperation matters: agency 1 is obliged to share v with the other agencies.

Returning to the regression problem, to compute the least squares estimators $\hat{\beta} = (X^T X)^{-1} X^T y$, it is necessary to compute $X^T X$ and $X^T y$. Because of the horizontal partitioning of the data,

$$X^T X = \sum_{j=1}^{K} (X^j)^T X^j.$$

Therefore, agency j simply computes its own $(X^j)^T X^j$, which has dimensions $p \times p$, where p is the number of predictors, and these are combined entrywise using secure summation. The same is done for $X^T y$. Then, each agency can calculate $\hat{\beta}$ from the shared values of $X^T X$ and $X^T y$. Note that no agency learns any other agency’s $(X^j)^T X^j$ or $(X^j)^T y^j$, but only the sum of these over all the other agencies.

It is also possible to compute and share securely a variety of regression diagnostics [10], as well as deal with alternate problems, such as vertically partitioned data [11].

Acknowledgements

This research was supported by NSF grant EIA–0131884 to the National Institute of Statistical Sciences (NISS) and by the National Center for Education Statistics.

References

[1] Y. M. M. Bishop, S. E. Fienberg, and P. W. Holland. *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA, 1975.

[2] T. Dalenius and S. P. Reiss. Data-swapping: A technique for disclosure limitation. *J. Statist. Planning Inf.*, 6:73–85, 1982.

[3] A. Dobra, S. E. Fienberg, A. F. Karr, and A. P. Sanil. Software systems for tabular data releases. *Int. J. Uncertainty, Fuzziness and Knowledge Based Systems*, 10(5):529–544, 2002.

[4] A. Dobra, A. F. Karr, and A. P. Sanil. Preserving confidentiality of high-dimensional tabular data: Statistical and computational issues. *Statist. and Computing*, 13(4):363–370, 2003.

[5] J. Domingo-Ferrer, J. M. Mateo-Sanz, and V. Torra. Comparing SDC methods for microdata on the basis of information loss and disclosure risk. *Presented at UNECE Workshop on Statistical Data Editing*, May, 2001.

[6] G. T. Duncan, S. A. Keller-McNulty, and S. L. Stokes. Disclosure risk vs. data utility: The R-U confidentiality map. *Management Sci.*, 2004. Submitted for publication.

[7] S. Gomatam, A. F. Karr, J. P. Reiter, and A. P. Sanil. Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access analysis servers. *Statist. Sci.*, 2004. To appear. Available on-line at www.niss.org/dgii/technicalreports.html.

[8] S. Gomatam, A. F. Karr, and A. P. Sanil. Data swapping as a decision problem. *J. Official Statist.*, 2003. To appear. Available on-line at www.niss.org/dgii/technicalreports.html.

[9] A. F. Karr, X. Lin, J. P. Reiter, and A. P. Sanil. Analysis of integrated data without data integration. *Chance*, 17(3):26–29, 2004.

[10] A. F. Karr, X. Lin, J. P. Reiter, and A. P. Sanil. Secure regression on distributed databases. *J. Computational and Graphical Statist.*, 2004. To appear. Available on-line at www.niss.org/dgii/technicalreports.html.

[11] A. P. Sanil, A. F. Karr, X. Lin, and J. P. Reiter. Privacy preserving regression modelling via distributed computation. In *Proc. Tenth ACM SIGKDD Internat. Conf. on Knowledge Discovery and Data Mining*, pages 677–682, 2004. Available on-line at www.niss.org/dgii/technicalreports.html.

[12] M. Trottini. *Decision Models for Data Disclosure Limitation*. PhD thesis, Carnegie Mellon University, 2003. Available on-line at www.niss.org/dgii/TR/Thesis-Trottini-final.pdf.

[13] L. C. R. J. Willenborg and T. de Waal. *Elements of Statistical Disclosure Control*. Springer–Verlag, New York, 2001.

[14] W. E. Winkler. Re-identification methods for evaluating the confidentiality of analytically valid microdata. *Res. Official Statist.*, 1:87–104, 1998.

[15] A. C. Yao. Protocols for secure computations. In *Proceedings of the 23rd Annual IEEE Symposium on Foundations of Computer Science*, 1982.

RÉSUMÉ

Nous présentons deux methodes pour "trading off" la qualité et la confidentialité du microdata.