Recent Confidentiality Research Related to Access to Enterprise Microdata

Arnold P. Reznek[1]
U.S. Census Bureau, Center for Economic Studies[2]

Prepared for the Comparative Analysis of Enterprise Microdata (CAED) Conference
2006 – Chicago, IL, USA.
September 18-19, 2006

[1] Arnold.Phillip.Reznek@census.gov, (301)763-1856.

[2] Disclaimer and Acknowledgment: This paper was written by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. Any views, findings or opinions expressed in this paper are those of the author and do not necessarily reflect those of the Census Bureau. I would like to thank Ron Jarmin for helpful comments.

Recent Confidentiality Research Related to Access to Enterprise Microdata
Arnold P. Reznek, U.S. Census Bureau

1. INTRODUCTION AND BACKGROUND

Statistical Disclosure Control (SDC) in general has made significant progress in the last few years and this progress has begun to affect how data users, including researchers, access and analyze enterprise[3] data. The 2003 and 2005 CAED conferences each devoted a session to this topic. The main goals here are to describe some of this progress and some of the context in which it has taken place. Too much has happened to discuss it all in a single conference session. This paper presents some developments that seem to me to be important and that provide context and background for the other two papers in this session. I also am writing from a U.S. perspective, and within that from a U.S. Census Bureau perspective. I apologize in advance for this uneven coverage.

The paper is intended primarily for readers who fit the profile of most attendees of this CAED conference: those who may be familiar with the SDC field, but not necessarily experts in it. I also assume the reader is familiar with Research Data Centers (RDCs);[4] this session includes a paper (Ritchie 2006) on SDC in RDCs. Several good publications now provide a way for interested persons to learn about the SDC field, including Willenbourg and de Waal (1996, 2001), Eurostat (1996), Doyle et al (2001), and FCSM (2005). An excellent series of relatively nontechnical articles in *Chance* magazine[5] introduces some of the recent developments described below.

Statistical agencies have always recognized the need to balance the conflicting requirements to protect the data against the requirement to provide information for decision makers. The environment has changed in recent years: on the one hand, the computer revolution has made it much more feasible for researchers to analyze large micro data sets using improved statistical techniques. On the other hand, that same computer revolution, which has also spawned an explosion of available data outside statistical agencies, has made it much easier for intruders to use try to identify confidential information (Sweeney 2001). A recent U.S. study (National Research Council 2005) summarizes changes that have taken place in the last ten years (particularly since the well-known study by Duncan et al. (1993). Until recently, few attempts had been made to make the risk-access tradeoff scientifically. Recent research has begun to explicitly take this tradeoff into account, and this has influenced my choice of developments to report.[6]

[3] For this paper, the terms "enterprise" and "firm" mean the same thing.
[4] For background on RDCs, see Dunne (2001). More information on the U.S. Census Bureau's RDCs is available at http://www.ces.census.gov.
[5] Vol. 16, no. 3, summer 2004. *Chance* is published jointly by the American Statistical Association and Springer-Verlag; its website is http://www.amstat.org/PUBLICATIONS/chance/.
[6] See Lane (2005) for a discussion of this tradeoff from an economist's point of view.

In keeping with the themes of this conference, the discussion is largely restricted to enterprise data, which means that little is said about a parallel set of significant developments related to public use microdata files of data on households and individuals. (The major exception relates to synthetic microdata.) The remainder of the paper provides an overview of developments in several research areas: SDC for tables, SDC for models, synthetic data, model servers, and "Virtual RDCs." (All of these terms are defined below.) References are provided for those who wish to explore further.

Though the progress has been significant, this paper and the others in this session will make clear that there is much to be done in this field. I hope to help stimulate more researchers' interest and participation in the SDC field, and encourage them to work with statistical agencies to improve the situation. Much of the important recent research has been done by academics who have learned that there are interesting research problems in this field and have worked with statistical agencies to solve them. We need more of this type of participation.

2. SDC FOR TABLES

Most enterprise data in most countries is released in the form of tables. One type of tables displays *magnitudes* such as total value of shipments (or sales), profits, employment, or capital investment, displayed by well-defined characteristics such as industry or geography. Another type displays *frequencies* (counts or percentages) of units that fall into ranges of variables (e.g., size classes of shipments, sales, profits, employment, or investment.) Often, both types of tables are hierarchical, with different levels of industries or geographies displayed (e.g., the bread industry within the food industry; or counties within states in the U.S.)

In tabular output, a disclosure has taken place if users can estimate a responding company's value "too closely." *Linear sensitivity measures* (Cox 2001) have been developed for determining whether this has happened. Two examples of these rules are the "p% rule" - designed to ensure that a user cannot estimate a respondent's value to within p% of that value – and the "(n,k) rule" – designed to ensure that a small number of units does not "dominate" a cell. For tables based on (i.e., aggregated from) establishment-level data, data from all establishments within a cell are aggregated to the enterprise level before the rules are applied. That is, disclosure analysis is applied at the enterprise level.

Until relatively recently, statistical agencies have used *cell suppression* to avoid disclosure in tables. (See Cox 2001 and Giessing 2001.) Cells that are considered sensitive according to a linear sensitivity measure are not published. These cells are called *primary suppressions.* Because marginal totals are usually shown in the tables, other cells called *complementary suppressions* must be selected and suppressed, so that primary suppression values cannot be derived or estimated too closely via addition and subtraction of published values. Software based on variants of linear programming is used to find complementary suppressions that, according to some criterion, release the

most information while protecting the sensitive cells. For relatively small tables or tables with hierarchical structures, the suppression programs themselves can ensure that the sensitive cells are properly protected – i.e., such suppression programs for these tables can be "self-auditing." However, suppression programs for large and complex (e.g., non-hierarchical) tables are either not self-auditing or are cannot yet self-audit in reasonable time. In such cases, separate auditing programs must be run. These programs determine upper and lower bounds for each sensitive cell, based on the rest of the table. If these bounds are too narrow, then additional cells are suppressed to provide further protection.

In recent years, several important developments have taken place in cell suppression techniques. Perhaps the most notable are the methods of Fischetti and Salazar (2001, 2005), which are being incorporated into the ARGUS SDC software (Hundepool 2004). These methods have been developed as part of a major European research initiative, the Computational Aspects of Statistical Confidentiality (CASC) Project.[7]

However, the fundamental problem with cell suppression remains: from the user's point of view, it suppresses too much information, since many (if not most) of the secondarily suppressed cells are not actually sensitive. In large, complex tables, this problem causes users great difficulty. It is particularly serious for users in the field of regional science, who wish to have detailed geographic information on businesses. In a very recent, example, Isserman and Westervelt (2006) describe how they and their colleagues have developed techniques for estimating ("filling in") 1.5 million (!) suppressed cells in the U.S. Census Bureau's County Business Patterns data.

Because of the secondary disclosure problem, researchers and statistical agencies have intensified their search for techniques that avoid the cell suppression problem. Here, I mention two new techniques have been developed in recent years: noise addition and Controlled Tabular Adjustment. These and related techniques are sometimes called "perturbative" methods, because they prevent disclosure by perturbing the data in some way.

Adding Noise

In this technique,[8] random "noise" is added to the underlying microdata prior to tabulation (Evans, Zayatz, and Slanta, 1998). Each responding establishment's data are perturbed by a small amount, say 10% (the actual percent is confidential), in either direction. Noise is added in such a way that cell values that would normally be primary suppressions, thus needing protection, are changed by a large amount, while cell values that are not sensitive are changed by a small amount. To perturb an establishment's data by about 10%, the agency multiplies its data by a random number that is close to either 1.1 or 0.9. The agency could use any of several types of distributions from which to choose the multipliers, and the distributions remain confidential within the agency. The overall distribution of the multipliers is symmetric about 1.

[7] For more information, see the CASC website at http://neon.vb.cbs.nl/casc/.
[8] The discussion in the first two paragraphs relies heavily on Zayatz (2005).

The use of noise enables data to be shown in all cells in all tables, and it eliminates the need to coordinate cell suppression patterns between tables. It is a much less complicated and less time-consuming procedure than cell suppression. Because noise is added at the microdata level, additivity of the table is guaranteed. The noise procedure does not introduce any bias into the cell values for census or survey data. To protect the data at the company level, all establishments within a given company are perturbed in the same direction. The introduction of noise causes the variance of an estimate to increase by an amount equal to the square of the difference between the original cell value and the noise added value; and the agency can incorporate this into published coefficients of variation.

The U.S. Census Bureau is currently using this technique in its Quarterly Workforce Indicators (QWI; see Abowd et al. 2006a, 2006b). The QWI are "developed by the Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) Program as a part of its Local Employment Dynamics partnership with U.S. state Labor Market Information offices. These data provide detailed quarterly statistics on employment, accessions, layoffs, hires, separations, full-quarter employment (and related flows), job creations, job destructions, and earnings (for flow and stock categories of workers). The data are released for NAICS industries (and 4-digit SICs) at the county, workforce investment board, and metropolitan area levels of geography. The confidential microdata - unemployment insurance wage records, ES-202 establishment employment, and Title 13 demographic and economic information– are protected using a permanent multiplicative noise distortion factor. This factor distorts all input sums, counts, differences and ratios. The released statistics are analytically valid – measures are unbiased and time series properties are preserved. The confidentiality protection is manifested in the release of some statistics that are flagged as "significantly distorted to preserve confidentiality." These statistics differ from the undistorted statistics by a significant proportion. Even for the significantly distorted statistics, the data remain analytically valid for time series properties... [additional] confidentiality protection is provided by the estimation process for the QWIs…" (Abowd et al 2006b, abstract)

In addition, some of the aggregate estimates turn out to be based on fewer than three persons or establishments. These estimates are suppressed and a flag set to indicate suppression. Suppression is only used when the combination of noise infusion and weighting may not sufficiently distort the publication data (Abowd et al 2006b, p. 46). Thus, in practice, some cell suppression may be used along with noise.

The U.S. Census Bureau is also considering using noise in its future publications of data on enterprises, including the County Business Patterns and the Economic Censuses. If this takes place, it would largely mitigate the need for users' efforts to "fill in" the vast majority of suppressed cells, like the work described in Isserman and Westervelt (2006). Again, a relatively small amount of cell suppression may still be used.

These developments are likely to affect the Census Bureau's RDCs. Within the Census Bureau, the question has arisen whether output (tables or model results) released from Census Bureau's Research Data Centers should be based on the data with or without noise added. Both are possible, since the "noise factors" are kept separate from the

original data. Another likely effect on the RDCs is to decrease the number of inquiries about prospective projects that really amount to efforts to fill in "holes" in the published tables. The Census Bureau's RDC system receives quite a few such inquiries, some from U.S. government agencies frustrated with cell suppression. We often must decline these requests. A decrease in data users' need to make these inquiries would be good for the Census Bureau, the RDCs, and the data users.

Controlled Tabular Adjustment

Controlled Tabular Adjustment (CTA)[9] is a relatively new approach. For magnitude data, a linear sensitivity rule (e.g., the P-percent rule) is used to determine which cells are sensitive. With CTA, each original sensitive value of a table is replaced with a safe value that is a "sufficient distance" away from the true value; and non-sensitive cell values are minimally adjusted to ensure that the published marginal totals are additive. A "sufficient distance" is the value that would have to be added to the "true" cell total to make the cell not sensitive (according to the sensitivity rule being applied). For frequency data, most linear sensitivity rules are equivalent to a threshold rule of 3 respondents and a "sufficient distance" from the true value would involve changing the value by either 1 or 2. That is, the value of a sensitive cell would be changed to either 0 or 3. This is identical to rounding to the base 3.

Cox, Orelien, and Shaw (2006) describe ongoing work to modify the CTA technique so that the adjusted table preserves as much as possible the distribution of the original (unadjusted) table. This work is aimed at preserving the analytic validity of the adjusted tables.

Frequency Tables

The above discussion has largely concerned tables of magnitude data, though some of the techniques also can be applied to tables of frequencies. Agencies release many frequency tables, but to protect confidentiality they often suppress "interior" cells in favor of marginal totals or conditional relative frequencies. Duncan et al. (2001) summarize research aimed at obtaining maximum analytic utility from frequency tables while maintaining confidentiality. They describe the "Risk-Utility (R-U) Confidentiality Map," which provides a quantitative way of measuring how different disclosure methods trade off disclosure risk against utility, for given measures of risk and utility. Skavkovic and Feinberg (2004) extend analysis of disclosure risk in frequency tables, using methods that characterize the set of all possible tables that could be generated from a given set of marginal or conditional tables. These methods use techniques from the mathematical field of algebraic geometry. The discussion here is obviously very sketchy and this area is developing rapidly; see the cited sources for more details. For relatively nontechnical sources, see Feinberg and Slavkovic (2004) for more on frequency tables, and Duncan and Stokes (2004) for more on the R-U confidentiality map.

[9] The discussion in this paragraph relies on FCSM (2005).

3. SYNTHETIC DATA

Statistical agencies have found it difficult to release useful public use microdata on enterprises. Although a relatively few such files have been released, agencies for the most part continue to believe that the skewness of the variable distributions and the public knowledge about the largest enterprises combine to present unacceptable disclosure risks. Synthetic data presents a possible way out. Given a data set, it is possible to develop models to generate synthetic data that have many of the same statistical properties as the original data (Abowd and Woodcock, 2001, 2004).[10] Generating the synthetic data is often done by sequential regression imputation, one variable in one record at a time. Using all of the original data, the agency develops a regression model for a given variable. Then, for each record, the agency blanks the value of that variable and uses the model to impute for it. Then, the process is repeated for the next variable.

Synthesizing data can be done in different ways and for different types of data products. One can synthesize all variables for all records (full synthesis) or a subset of variables for a subset of records (partial synthesis). A partial synthesis targets records and variables that are considered risky. The agency can generate one implicate from a file, or it can generate several implicates that can be analyzed using standard statistical software and combined using simple techniques.

Statistical agencies can synthesize data to release synthetic microdata or a product (such as a map) generated from the synthetic microdata. The U.S. Census Bureau and the U.S. Social Security Administration are involved in an effort (led by John Abowd) to develop a public use synthesized microdata file containing linked U.S. Social Security Administration earnings data and the U.S. Census Bureau's Survey of Income and Program Participation (SIPP) data with the goal of releasing multiple synthetic implicates. Both agencies are involved in judging the quality of the product and in disclosure-proofing the files.

In 2005, the Census Bureau approved the release of the Census Bureau's first data product based on partially synthetic data, called "On the Map," which is a set of maps that show where people work and where workers live. These are accompanied by reports on their age, earnings, industry distributions, quarterly workforce indicators, and more.[11]

In this CAED session, Kinney and Reiter (2006) report about ongoing work to develop synthetic public use data sets for longitudinal establishment data from the U. S. Census Bureau's Longitudinal Employer-Household Dynamics. For details, including more information about how data synthesis proceeds, see their paper.

[10] The first three paragraphs of this section rely on Zayatz (2005).
[11] This application is available at http://lehd.dsd.census.gov/led/datatools/onthemap.html.

4. SDC FOR REGRESSION MODELS AND MODEL SERVERS

The question of whether disclosure risks exist in regression-type models has become more important over the past decade as statistical agencies expand access to their micro data.[12] It is particularly important in Research Data Centers, where much of the research output consists of various types of regression results. Disclosure risks may arise from regression models, particularly in the standard linear regression model estimated using Ordinary Least Squares methods as well as in logit and probit models (which use binary (0,1) dependent variables) and other Generalized Linear Models (Reznek 2003, Reznek and Riggs, 2004, 2005). The risks in regression models that contain continuous variables on the right-hand side are small if the overall sample is large enough to pass tabular disclosure analysis. However, risks may exist in models that contain dummy variables as independent variables. Coefficients of models that contain only fully interacted (saturated) sets of dummy variables on the right-hand sides can be used to obtain entries in cross-tabulations of the dependent variable, where the cross-tabulation categories are defined by the dummy variables. The same types of cross-tabulations can also arise from correlation and covariance matrices of the variables, and from variance-covariance matrices of model coefficients, if these matrices include dummy variables. These research outputs present disclosure risks if the cross-tabulations present disclosure risks.

Other risks may exist in regression models. Cox (2002) showed the following: Suppose we have two groups of (say) firms, group A and group B, where group A is included in group B. Then there is an implicit third group C, the difference between groups A and B. Suppose a researcher estimates the same regression model on groups A and B, and also calculates the mean of the dependent variable for group A. Then from the two sets of regression coefficients and the single mean for group A, it is possible to calculate the means of the dependent variable for groups B and C. Therefore, running a regression on one sample, and then running the same regression on a slightly larger or smaller sample can pose disclosure risks. Reiter (2004 p. 12-13) and Gomatam et al. (2005) show that still other disclosure risks can arise from a researcher's use of certain model specifications aimed at isolating a single firm's data (e.g., including an indicator for a single observation), and from applying certain data transformations that convert a single firm's data into an extreme outlier that "pulls" the regression line very close to itself.

All of the work above is applicable in the context of RDCs, where it is relatively easy to control the output that is released. However, much of the work has been carried out by researchers from NISS in the context of developing "model servers." These servers are computers which are set up to allow researchers to estimate regression parameter estimates and model diagnostics remotely, without having direct access to the microdata. In this setting, the agency potentially has much less control than at RDCs over the model output that is released. The NISS researchers have investigated how to release useful model results in the model server setting while not compromising confidential information (Gomatam et al, 2005, Reiter 2003, Reiter and Kohnen 2005). In particular, they propose the use of synthetic regression diagnostics, which are based on generating synthetic model residuals that mimic the properties of the actual residuals.

[12] This paragraph relies on FCSM (2005, pp. 84-85).

Statistical agencies in and outside the U.S. are at various stages of setting up versions of model servers. The 2005 United Nations Economic Commission for Europe had a session containing papers that describe efforts in the U.S. (Census Bureau and National Center for Health Statistics), Denmark, Sweden, and the Netherlands.[13] Rowland (2003) discusses developments and issues involved in remote access servers, which include model servers but also "table servers" that allow users to generate tables, but perhaps not models.

Further recent work by the NISS researchers has investigated how to estimate regressions using a combination of confidential data from several sources (e.g., statistical agencies) without compromising confidentiality (Karr et al, 2005). The combination of data could include different observations or different variables (or both) from each source.

5. VIRTUAL RDCs

Abowd and Lane (2004) describe an ongoing effort that combines many of these recent developments in SDC with the U.S. Census Bureau's RDC system and promises to benefit both the research community and statistical agencies. In principle, multiple public use synthetic data sets can be created from a single underlying confidential file and customized for different uses. The synthetic data can be maintained at a remote site that would be accessible from outside (e.g., from researchers' offices). This site is called a "Virtual RDC", in that the computer environment (including operating system, software, data file structures, variable names, and so on) are exactly the same as at the agency's restricted access RDCs. The RDCs contain the corresponding underlying confidential files, which are sometimes referred to as the "gold standard" files. Researchers can access the synthetic files at the Virtual RDC, to familiarize themselves with the structures of the "gold standard" files and carry out analyses to help them develop proposals to use these files at the RDCs. Researchers can carry out the same analysis on both the synthetic and "gold standard" files. Comparison of the results will stimulate the development of improved synthetic files that have better analytic validity. The virtual RDC is now operational. For more information, see http://vrdc.ciser.cornell.edu/news/.

6. CONCLUSIONS

This paper has described several recent developments in SDC. Taken together, this research promises to increase researchers' access to high-quality enterprise data – aggregate data and the underlying microdata -- that allows for valid statistical inferences while protecting the confidentiality of the microdata. It also promises to improve the process of releasing disclosure-free output from RDCs. However, a great deal needs to be done before this promise is realized.

[13] Available at http://www.unece.org/stats/documents/2005.11.confidentiality.htm.

7. REFERENCES

Abowd, J.M. and J.I. Lane (2004). "New Approaches to Confidentiality Protection: Synthetic Data, Remote Access and Research Data Centers." In Domingo-Ferrer, J. and V. Torra (eds.), *Privacy in Statistical Databases – PSD 2004.* LNCS 3050. Berlin: Springer-Verlag, pp. 282-289. Also Longitudinal Employer-Household Dynamics Technical Paper TP-2004-3, available at http://lehd.dsd.census.gov/led/library/techpapers/tp-2006-01.pdf.

Abowd, J.M., B.E. Stephens, L. Vilhuber, F. Andersson, KL. McKinney, M. Roemer, and S. Woodcock (2006a). "The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators." LEHD Technical Paper TP-2006-01. Available at http://lehd.dsd.census.gov/led/library/techpapers/tp-2006-01.pdf.

Abowd, J.M., B.E. Stephens, and L. Vilhuber (2006b). "Confidentiality Protection in the Census Bureau's Quarterly Workforce Indicators." LEHD Technical Paper TP-2006-02. Available at http://lehd.dsd.census.gov/led/library/techpapers/tp-2006-02.pdf.

Abowd, J.M. and S.D. Woodcock (2004). "Multiply-Imputing Confidential Characteristics and File Links in Longitudinal Linked Data." In Domingo-Ferrer, J. and V. Torra (eds.), *Privacy in Statistical Databases – PSD 2004.* LNCS 3050. Berlin: Springer-Verlag, pp. 290-297.

Abowd, J.M. and S.D. Woodcock (2001). "Disclosure Limitation in Longitudinal Linked Data." Chapter 10 in Lane et al. (2001).

Cox, L. (2002). "Confidentiality Issues For Statistical Database Query Systems." Invited Paper for Joint UNECE/Eurostat Seminar on Integrated Statistical Information Systems and Related Matters (ISIS 2002). (17-19 April 2002, Geneva, Switzerland). Available at http://www.unece.org/stats/documents/ces/sem.47/15.s.e.pdf.

Cox, L. (2001). "Disclosure Risk for Tabular Economic Data." Chapter 8 in Doyle et al. (2001).

Cox, L., J.G. Orelien, and B.V. Shah (2006). "A Method For Preserving Statistical Distributions Subject To Controlled Tabular Adjustment." To be presented at *Privacy in Statistical Databases PSD – 2006*, Rome, Italy, December 2006. (Draft, July 2006)

Doyle, P., J.I. Lane, J.J.M. Theeuwes, and L.V. Zayatz (2001). *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*. Amsterdam: Elsevier Science B.V.

Duncan, G.T., S.E. Feinberg, R. Krishnan, R. Padman, and S.F. Roehrig (2001). "Disclosure Limitation Methods and Information Loss in Tabular Data." Chapter 7 in Doyle et al. (2001).

Duncan, G.T., T.A. Jabine, and V.A. De Wolf, Eds. (1993). *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics.* Washington, D.C.: National Academy Press.

Duncan, G.T. and S.L. Stokes (2004). "Disclosure Risk vs. Data Utility: The R-U Confidentiality Map as Applied to Topcoding." *Chance*, Vol. 3, No. 3, pp. 16-20.

Dunne, T. "Issues in the Establishment and Management of Secure Research Sites." (2001). Chapter 12 in Doyle et al. (2001).

Evans, B. T., L. Zayatz, and J. Slanta. (1998), "Using Noise for Disclosure Limitation for Establishment Tabular Data." *Journal of Official Statistics*, Vol. 14, No. 4, pp. 537-552.

Eurostat (1996). *Manual on Disclosure Control Methods.* Luxembourg: Office for Official Publications of the European Communities.

Federal Committee on Statistical Methodology (FCSM, 2005). *Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology*, Washington, DC, U.S. Office of Management and Budget. Revision of 1994 edition; available at http://www.fcsm.gov/working-papers/spwp22.html.

Fienberg, S.E. and A.B. Slavkovic (2004). "Making the Release of Confidential Data from Multi-Way Tables Count." *Chance* 17:3 (Summer 2004), pp. 5-10.

Fischetti, M. and J.J. Salazar (2001). "Solving the Cell Suppression Problem on Tabular Data with Linear Constraints." *Management Science* Vol. 47, No. 7, pp. 1008-1027.

Fischetti, M. and J.J. Salazar (2005). "A Unified Mathematical Programming Framework for Different Statistical Disclosure Limitation Methods." *Operations Research*, Vol. 53, No. 5, pp. 819-829.

Giessing, S. (2001). "Nonperturbative Disclosure Control Methods for Tabular Data." Chapter 9 in Doyle et al. (2001).

Gomatam, S., Karr, A. F., Reiter, and J. P., Sanil, A. (2005). "Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access servers." *Statistical Science*, 20, pp. 163 - 177.

Hundepool, A. (2004). "The ARGUS Software in the CASC-Project." In Domingo-Ferrer, J. and V. Torra (eds.), *Privacy in Statistical Databases – PSD 2004.* LNCS 3050. Berlin: Springer-Verlag., pp. 323-335.

Isserman, A.M. and Westervelt, J. (2006). "1.5 Million Missing Numbers: Overcoming Employment Suppression in *County Busness Patterns* Data." *International Regional Science Review* 20, 2: pp. 311-335.

Karr, A.F., X. Lin, A. P. Sanil, and J.P. Reiter (2005). "Analysis of Integrated Data without Data Integration." *Chance,* Vol. 3, No. 3, pp. 26-29.

Kinney, S. and J.P. Reiter (2006). "Making Public Use, Synthetic Files of Longitudinal Establishment Data." Presented at CAED 2006, Chicago IL.

Lane, J. (2005). "Optimizing the Use of Micro-data: An Overview of the Issues." Paper presented at the 2006 Joint Statistical Meetings as part of a session organized to honor Pat Doyle. Available at http://client.norc.org/jole/SOLEweb/Accesstomicrodata%5B1%5D.pdf.

National Research Council. (2005). *Expanding Access to Research Data: Reconciling Risks and Opportunities.* Panel on Data Access for Research Purposes, Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Reznek, A.P. (2003). "Disclosure Risks in Cross-Section Regression Models." *American Statistical Association 2003 Proceedings of the Section on Government Statistics and Section on Social Statistics*, pp. 3444 -3451. (on proceedings CD).

Reznek, A.P., and T. L. Riggs (2005). "Disclosure Risks in Releasing Output Based on Regression Residuals." *American Statistical Association 2005 Proceedings of the Section on Government Statistics and Section on Social Statistics*, pp. 1397-1404. (on proceedings CD).

Reznek, A.P., and T. L. Riggs (2004). "Disclosure Risks in Regression Models: Some Further Results." *American Statistical Association 2004 Proceedings of the Section on Government Statistics and Section on Social Statistics,* pp. 1701-1708 . (on proceedings CD).

Reiter, J. (2003). "Model Diagnostics for Remote Access Regression Servers." *Statistics and Computing,* 13, pp. 371-380.

Reiter, J.P. (2004). "New approaches to data dissemination: A glimpse into the future (?)." *Chance,*17:3 (Summer 2004), pp.12 - 16.

Reiter, J. P. and Kohnen, C. N. (2005) Categorical data regression diagnostics for remote servers. *Journal of Statistical Computation and Simulation,* 75, pp. 889 - 903.

Ritchie, F. (2006). "Statistical Disclosure Control in a Research Environment." Presented at CAED 2006, Chicago IL.

Rowland, S. (2003), "An Examination of Monitored, Remote Microdata Access Systems" Paper presented at the U.S. National Academy of Science Workshop on Access

to Research Data: Assessing Risks and Opportunities, October 16-17, 2003. Available online at http://www7.nationalacademies.org/cnstat/Rowland_Paper.pdf.

Slavkovic, A.B. and S.E. Fienberg (2004). "Bounds for Cell Entries in Two-Way Tables Given Conditional Relative Frequencies." In Domingo-Ferrer, J. and V. Torra (eds.), *Privacy in Statistical Databases – PSD 2004.* LNCS 3050. Berlin: Springer-Verlag., pp. 323-335.

Sweeney, L. (2001). "Information Explosion." In Doyle et al. (2001), Chapter 3.

Willenborg, L. and T. de Waal (2001). *Elements of Statistical Disclosure Control*. Lecture Notes in Statistics, volume 155. New York: Springer-Verlag.

Willenborg, L. and T. de Waal (1996). *Statistical Disclosure Control in Practice*. Lecture Notes in Statistics, volume 111. New York: Springer-Verlag.

Zayatz, L. (2005). "Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update." Research Report RRS2005/06. Statistical Research Division, U.S. Census Bureau. Available at http://www.census.gov/srd/papers/pdf/rrs2005-05.pdf.