Statistics of Income

2007 SOI Paper Series

Measuring Disclosure Risk and an Examination of the Possibilities of Using Synthetic Data in the Individual Income Tax Return Public Use File

by Sonya Vartivarian, John L. Czajka, and Michael Weber

Measuring Disclosure Risk and an Examination of the Possibilities of Using Synthetic Data in the Individual Income Tax Return Public Use File

Sonya Vartivarian and John L. Czajka, Mathematica Policy Research, Inc., and Michael Weber, Internal Revenue Service

Each year, the Statistics of Income (SOI) Division of the Internal Revenue Service (IRS) draws a sample of individual and sole proprietorship tax returns, abstracts and edits a large number of data items, and prepares a microdatabase that the Treasury Department and the Congress use for tax policy analysis. The SOI Division also produces a public use file (PUF) version of this sample so that tax policy researchers in academia and the private sector can have access to some of the same information for their own tax policy analysis. In addition to being stripped of all identifying information and limited to about 200 fields, the data included in the PUF are masked in a number of ways to further reduce the risk of disclosure.

Disclosure avoidance has been an ongoing topic of study by the SOI Division. The IRS is legally obliged to assure that no taxpayer information is ever disclosed. This legal obligation drives the producers of the PUF to address known risks of disclosure and also reinforces the need to push toward increased protection as technology and data that pose risks become more available and increasingly sophisticated.

▶ Disclosure Protection

As a source of data for research purposes, the tax return has a number of limitations that work in favor of efforts to limit the risk of disclosure. First, personal demographic information is very limited, consisting only of marital status (married versus not). Similarly, family demographic information includes only the number of claimed dependents, which is broken out into children who lived with the taxpayer, children who did not live with the taxpayer due to a divorce or separation, and other dependents who met both residency and support tests. Geographic information is captured in the mailing address, but this is not always residential (or even the taxpayer's). Income is reported by source (for example, wages and salaries, business income, capital gains), but, for most sources, only taxable income is collected, and the incomes of spouses are combined. Furthermore, income from wages and salaries, which may pose the greatest risk of disclosure for the most people, is aggregated over employers.

Further benefiting efforts to limit disclosure risk, most of the items reported on tax returns serve a unique administrative purpose and, for that reason, do not appear in other databases. Some items that do appear in other databases may appear in a disaggregated form, with only partial representation in a given database.

The SOI Division's strategy for producing a PUF reflects the Division's assessment of the relative sources of risk. The strategy recognizes the limited number of items that recur in other databases, whether identifiable or not, and the lower quality of other data sources, generally.

While a number of different methods of disclosure avoidance are applied to a wide range of variables, the most substantial masking is applied to those variables that are considered to present the highest risk. For many years, the SOI Division has relied on microaggregation, or blurring, as the method of choice. With micro-aggregation, values of similar observations are averaged and replaced by their mean values. Recently, the previous univariate masking was replaced with a multivariate masking method. Multivariate masking (Mateo-Sanz and Domingo-Ferrer, 1998) was first incorporated in the 2002 PUF and used again in creating the 2003 and 2004 PUFs. The groupings of observations selected to be blurred were based on a multivari-

ate measure that took into account several variables. Other disclosure avoidance methods applied to the PUF include subsampling returns with the highest selection probabilities, suppressing geography and selected other variables for the highest income returns, and excluding outliers from selection.

▶ Measuring the Risk of Disclosure

In measuring the risk of disclosure, the SOI Division is in a much better position than most statistical agencies because it has access to the full population from which the annual sample was drawn. Consequently, it is possible to use data from the population file in an attempt to reidentify records in the PUF. With enough variables, however, this is simply too easy—and unrealistic. To obtain useful information from this exercise, then, the SOI Division must restrict the record linkage attempts to a reduced number of fields representing a plausible threat scenario.

With these fields, the SOI Division employs a distance-based algorithm to determine the PUF record that most closely matches a given record from the population and how this compares in rank order to the true match, if present in the PUF. The application of a distance-based algorithm in this manner is intended to test the PUF's exposure to reidentification from a would-be intruder who has access to accurate values for a limited number of variables. The distance-based algorithm yields results for a given record that are independent of the results for any other record. This contrasts with a probabilistic record-linkage approach requiring large numbers of records from both the external data and the PUF.

With the distance-based algorithm, protection against reidentification is measured in terms of the number of PUF records that lie at least as close to a record from the population as the true match. The minimum protection that is sought is having at least two records that are at least as close to a record from the population as the true match, if the true match is in the PUF. Because of the exceedingly small likelihood that a would-be intruder would have access to multiple variables as accurate as the population records, this methodology is more useful for evaluating the comparative risk of alternative disclosure avoidance schemes than the absolute risk associated with any one method.

As a disclosure avoidance technique, microaggregation becomes more effective as sample size decreases. The magnitudes of the adjustments (the masking) increase as the observations that are being pooled move farther apart. Multivariate microaggregation enhances that effect, but, beyond a certain point, the magnitudes of the adjustments may begin to have an adverse effect on data quality. In the SOI application, microaggregation is carried out within classes defined by important categorical variables, which reduces sample sizes even further. In the SOI experience, the adoption of a multivariate micro-aggregation method has provided an important improvement on the previous univariate methodology with respect to disclosure avoidance, but the enhanced protection does not come without a price in terms of data quality. While this is being addressed in the context of the multivariate design, the limits of the current masking method and the necessity for ever-increasing protection—including the likely need to expand the most extreme masking as additional variables become sensitive—provide cause for investigation of alternative data masking procedures.

In the remainder of this paper, we consider synthetic data as an alternative masking methodology.

▶ Synthetic Data

A dataset that has n observations sampled from a population of size N is to be prepared for release, but concerns of sensitive information being disclosed create the need for masking the data. Synthetic data involves the release of implicates that are generated through imputation and released instead of the source data as proposed in Rubin (1993) and Raghunathan, Reiter, and Rubin (2003). Synthetic imputation methodology follows the multiple imputation methodology where missing items are imputed through a statistical model, and missing data imputation can be combined with synthetic data imputation in producing synthetic data (see Reiter, 2004 for details). Generally, to create synthetic data first, $N - n$ observations, or alternatively,

all N observations, are generated through a synthetic data model to produce complete, imputed data populations. This process is repeated M times, where $M \geq 1$ is the number of synthetic datasets that will be created. Then, M implicates are produced, where the mth implicate dataset is created by randomly selecting k observations from the mth synthetic complete data populations, and the M implicate datasets are released for public use instead of the original data.

As described in Raghunathan, Reiter, and Rubin (2003), synthetic data are created by independently drawing from a posterior predictive distribution of the sensitive variables, thus conditional on any observed data and the model assumptions. Though the method does not require any nonsensitive variables, if such variables exist, they may be released unaltered. Methods of drawing from a posterior predictive distribution include a Bayesian Bootstrap methodology, the Federal Reserve Imputation Technique Zeta (FRITZ) used in the Survey of Consumer Finances (SCF) as described in Kennickell (1991, 1998), and the Sequential Regression Multivariate Imputation (SRMI) methodology of Raghunathan et al. (2001)

Fully synthetic or partially synthetic data can be produced, where the latter distorts only a subset of values that are considered at risk for disclosure (e.g., Rubin, 1993; Raghunathan, Reiter, and Rubin, 2003; Reiter 2003, 2004). Partially synthetic data may distort certain key variables, such as through the Selective Multiple Imputation of Keys (SMIKe) presented in Little and Liu (2002), where key variables are those an intruder may have access to (though not sensitive) and that would allow for identification of respondents through linkage to sensitive variables. Or, partially synthetic data may impute for only a subset of observations for sensitive variables (e.g., Kennickell, 1991, 1998; Reiter, 2004). Fully synthetic data assure high protection of sensitive information and may increase the release and thus use of such data since fewer data restrictions would be necessary. The assurance that the whole record is "made up" and thus does not risk disclosure is certainly appealing. However, the modeling may be complicated as it may involve a large set of variables and large number of records. The benefit of using a partial method such as SMIKe is that it requires less modeling and a fewer observations, thus reducing the sensitivity of inferences due to model misspecification (Little and Liu, 2002). Further, partial synthetic data might limit the loss of information due to imputation since fewer records and variables are imputed.

As a major application of partially synthetic data imputation, Kennickell (1997) notes that the Survey of Consumer Finances (SCF) has two serious disclosure risks: the survey obtains detailed family-level financial behavior and oversamples wealthy families. Multiple imputation for missing data in the SCF began in 1989 and included a modest amount of synthetic data imputation for sensitive variables, with synthetic imputations eventually applied to every dollar variable for sensitive cases in the 1998 SCF (Kennickell, 2000). Unlike the IRS data, the SCF does not have a population file available for use in assessing the performance of disclosure methods.

► Synthetic Data For SOI PUF

Some general concerns in creating and using synthetic data have been alleviated through advances in statistical methodology and technology. For example, the challenge of producing synthetic data has diminished since complicated analyses required to produce such data are no longer barred due to computing limitations (though other limitations may exist such as those described in Practical Implications of Producing Synthetic PUF). Further, common software can be implemented by data users to carry out analyses on synthetic data and can be combined with inferential formula that have been developed for fully and partially synthetic data (see Raghunathan, Reiter, and Rubin, 2003; and Reiter 2003, 2004). However, the SOI PUF released by the IRS can be said to have two additional concerns: the accounting relationship that must be maintained and the potential lack of variables to use in modeling the synthetic data. These topics are discussed in Accounting Relationships in SOI PUF Data and The Lack of Predictors, respectively.

▶ Accounting Relationships in SOI PUF Data

The structure of the IRS PUF data makes the task of producing synthetic data even more complicated. While only certain variables may require masking, virtually all of the variables in the file are part of various accounting relationships and nonlinear tax computations. For example, if total income is not sensitive, but some of the components that determine total income are sensitive (as are wages and salaries), then synthesizing the components may either distort the accounting relationship between the total and components or lead to synthesis of total income.

Another example of this problem is found in the itemized deductions reported on Schedule A. Given that all of the individual deductions must add up to the total amount of deductions, at least two variables must be masked to prevent an intruder from unmasking the masked variables. In addition, certain deductions are subject to an Adjusted Gross Income (AGI) limitation. Thus, if a variable subject to the limitation needs to be masked, then another variable also subject to the limitation must be masked as opposed to simply any other itemized deduction. Furthermore, the various itemized deductions are not functionally related to one another. For example, one may not be able to determine a useful statistical relationship among Medical and Dental Expenses, State Income Taxes, and Mortgage Interest. Yet if all three need to be masked, they must be jointly masked in such a way that their sum does not affect total itemized deductions allowing the total deduction amount to change may seem like a simple solution. Unfortunately, changing the amount of total itemized deductions affects taxable income and thus the computation of income tax, perhaps the most important variable in the file. It is very important that the PUF income tax variable produce an accurate estimate of aggregate income tax.

As a further complication, the Tax Code contains something called the Alternative Minimum Tax (AMT). The AMT is an independent calculation of tax that, unlike the regular income tax, does not allow income to be reduced by the amount of State income tax. Thus, if State income tax on Schedule A is masked, the amount of AMT will change. The AMT is perhaps the most vexing problem in tax policy today due to the exploding number of taxpayers who find themselves paying the AMT. For this reason, the PUF AMT variable must be accurate for a given return and bear the appropriate relationships to all of its determinants that are included in the PUF.

One example of accommodating the interrelationships among variables in the SRMI technique is through logical bounds imposed on imputations, where draws are obtained from a truncated predictive distribution, as described in Raghunathan et al. (2001) For example, the authors imposed bounds on the imputed value of the number of years smoked; they could be no more than the age of the respondent minus 18 unless school age smoking was reported. Possibly bounds similar to those described in the SRMI technique could be applied to help maintain the SOI PUF accounting relationships.

The complexity of accounting relationships and the implication for synthetic data imputation are evidenced in the National Center for Health Statistics National Health Interview Survey multiple imputation of family income and personal earnings (Schenker et al. 2006). The authors note that inconsistencies between family income and family earnings (income is less than earnings) are expected in general since family income is estimated by respondents instead of by summing up more detailed questions about personal earnings of family members. However, the completed data that include imputed values for missing items have a higher percentage of such inconsistencies than the respondent data.[1] Further, methods developed on restricting imputation of family income to be at least as large as imputed values of family earnings tend to distort the marginal distributions of family income and earnings. In this situation, marginal distributions were considered more important analytically. Thus it was decided to impute family income and family earnings without consistency restrictions, and the authors called for further research to resolve such inconsistencies. The NHIS experience reinforces concerns about the accounting relationships that must be accommodated in the SOI PUF.

[1] The imputation was produced through the SRMI technique

▶ The Lack of Predictors

Important variables in the modeling of synthetic data might include filing status and number of exemptions for children living at home, as well as occupation, geographic information such as Zip Code, education, age, and possibly AGI. Though the IRS file has many observations, it lacks a rich source of variables for use in modeling to produce synthetic observations such as education and often occupation. For example, human capital variables typically considered important in modeling financial variables are not available. Since synthetic data must be defined though meaningful models (analytically), the lack of predictors is an obstacle in considering synthesis as a data masking procedure.

▶ Analytic Usefulness: Synthetic Data Versus More Restrictive, Masked Data That Are Not Synthetic

Analytic validity of results based on synthetic data for anticipated analyses (e.g., through usual tax modeling and tax policy groups) and unanticipated analyses must be considered when assessing the usefulness of synthetic PUF data. Whether the synthetic model affects results or the potential for new analyses is a concern for both the usual tax model and policy groups as well as other, potential PUF users. This concern of sensitivity to model misspecification may be reduced through use of the SMIKe method and should be considered (Little and Liu, 2002). Yet the current method of microaggregation is already problematic in that there are too few observations to support even the variables currently blurred, and that distortions are quite large for certain types of records partly due to averaging dissimilar values with respect to a particular variable. If additional variables or observations are found to need disclosure protection, continued use of the current method of masking could lead to limiting the PUF release to certain observations or a reduced set of variables since satisfactory microaggregation may not be possible otherwise.

Such extreme measures may reduce the analytic utility of the PUF to such a degree that they are simply not viable as disclosure limitation strategies. While there are significant obstacles to be overcome in developing a new approach to disclosure limitation using synthetic data, an extrapolation of current trends leaves little question that a new approach is needed. Whether synthetic data will ultimately provide the answer remains to be seen, but the research attention being focused on synthetic data methodology at present makes this an option that must be given serious consideration.

▶ Practical Implications of Producing Synthetic PUF

Qualified staff resources must be available, and management must be able to commit those resources to a task that is expected to be large in scope. The budget for such an undertaking must also be well thought out and sufficient to produce a quality product. Time to release for a synthetic PUF is also an important practical factor. The first synthetic PUF would likely involve more extensive staff and labor hours, but, ideally, some of the time spent will be in developing the capability to produce future synthetic PUFs, thus reducing time and labor requirements for subsequent synthetic PUF releases.

Research into the duplicative or complementary nature of other such data releases should be assessed prior to undertaking the production of such synthetic data. For example, the Social Security Administration (SSA) has produced a partially synthetic Survey of Income and Program Participation (SIPP)/SSA/IRS PUF that should be examined. Methodology and lessons learned through the SSA PUF could be applied to the undertaking of an IRS PUF. In addition, the release of other synthetic data, such as the partially synthetic SSA PUF and synthetic monetary variables in the SCF, should increase awareness of synthetic data methods and allow public data users to consider synthetic data techniques as a viable disclosure avoidance methodology.

▶ Conclusion

There are several potential complications to producing a synthetic SOI PUF, such as handling the accounting relationships of these data, the nonlinearity of tax rates, and the uncertainty in relevant variables for use in the modeling. Yet we suspect the potential for damage to the quality of the data though the continued use of the current disclosure limitation methodology is one compelling reason to seriously explore a synthetic SOI PUF.

▶ References

Fries, G; L. Woodburn; and B. Johnson (1996), "Disclosure Review and Its Implications for the 1992 Survey of Consumer Finances," Proceedings of the Survey Research Methods Section, American Statistical Association.

Kennickell, A.B. (1991), "Imputation of the 1989 Survey of Consumer Finances: Stochastic Relaxation and Multiple Imputation," report prepared for the Joint Statistical Meetings.

Kennickell, A.B. (1997) "Multiple Imputation and Disclosure Protection: The Case of the 1995 Survey of Consumer Finances," *Record Linkage Techniques,* National Academy Press.

Kennickell, A.B. (1998), "Multiple Imputation in the Survey of Consumer Finances, "report prepared for the Joint Statistical Meetings.

Kennickell, A.B. (2000), "Wealth Measurement in the Survey of Consumer Finances: Methodology and Directions for Future Research, report prepared for the Annual Meetings of the American Association for Public Opinion Research.

Little, R.J.A. (1993), "Statistical Analysis of Masked Data," *Journal of Official Statistics,* 9, pp. 407–426.

Little, R.J.A. and D.B. Rubin (1997), "Should Imputation of Missing Data Condition on All Observed Variables?" Proceedings of the Survey Research Methods Section, American Statistical Association.

Little, R.J.A. and D.B. Rubin (2002), *Statistical Analysis with Missing Data,* Wiley, New York.

Liu, F. and R.J.A. Little. (2002), "Selective multiple imputation of keys for statistical disclosure control in microdata," Proceedings of the Survey Research Methods Section, American Statistical Association.

Mateo-Sanz, J.M. and J. Domingo-Ferrer (1998), "A Comparative Study of Microaggregation Methods," *Questiio*, 22, pp. 511-526.

Raghunathan, T.E.; J.M. Lepkowski; J. Van Hoewyk; and P. Solenberger (2001), "A Multivariate Technique for Multiply Imputing Missing Values Using a sequence of Regression Models," *Survey Methodology,* 27, pp. 85–95.

Raghunathan, T.E.; J.P. Reiter and D.B. Rubin (2003), "Multiple Imputation for Statistical Disclosure Limitation," *Journal of Official Statistics,* 19, pp. 1–16.

Reiter, J.P. (2002), "Satisfying Disclosure Restrictions with Synthetic Data Sets," *Journal of Official Statistics*, 18, pp. 531–543.

Reiter, J.P. (2003), "Inference for Partially Synthetic, Public Use Microdata Sets," *Survey Methodology*, 29, pp. 181–188.

Reiter, J.P. (2004), "Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation," *Survey Methodology*, 30, pp. 235–242.

Reiter, J.P. (2005), "Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study," *Journal of the Royal Statisitcal Society, A,* pp. 168, 185–205.

Rubin, D.B. (1993), "Discussion: Statistical Disclosure Limitation," *Journal of Official Statistics,* 9, pp. 461–468.

Schenker, N.; T.E. Raghunathan; P.L. Chiu; D.M. Makuc; G. Zhang; and A.J. Cohen (2006), "Multiple Imputation of Family Income and Personal Earnings in the National Health Interview Survey: Methods and Examples," http://www.cdc.gov/nchs/data/nhis/tecdoc.pdf.

U.S. Federal Committee on Statistical Methodology (2001), "Measuring and Reporting Sources of Error in Surveys," Statistical Policy Working Paper 31, U.S. Office of Management and Budget, Washington, DC.

Vartivarian, S. and R.J.A. Little (2002), "On the formation of weighting, adjustment cells for unit nonresponse," Proceedings of the Survey Research Methods Section, American Statistical Association.