ASSESSING DISCLOSURE PROTECTION FOR A SOI PUBLIC USE FILE

Marianne Winglee, Richard Valliant, Jay Clark, and Yunhee Lim, Westat;
Michael Weber and Michael Strudler, Statistic of Income Division Internal Revenue Service
Marianne Winglee, Westat, 1650 Research Boulevard, Rockville, Maryland 20850

Key Words: Microdata, disclosure risk, subsampling, microaggregation, record linkage, information loss

1. Introduction

This paper describes an evaluation of the disclosure protection methods for the Individual Tax Model Public Use File (PUF) released by the Statistics of Income (SOI) program of the Internal Revenue Service. The purpose of this evaluation is to explore options to strengthen disclosure protection while limiting information loss for tax returns with high incomes. Section 1 presents an introduction and motivation of this study. Section 2 discusses the preparation of the PUF, options for subsampling high income returns (from samples in an internal nonPUF), and options for disclosure protection by microaggregation (grouping microdata in aggregates of three). Section 3 discusses the method and data used to measure disclosure risk and information loss. Section 4 discusses our results and recommendations for further research. Section 5 lists references used in this paper.

The first Individual Income Tax Return PUF was created in 1960. Needless to say, the issue of disclosure control was not the same hot topic then that it is today. Basic precautions were taken like the removal of obvious identifiers such as name, address, and social security number, but little more than that. During the mid-1980's SOI undertook a reevaluation of its disclosure control procedures (Strudler, Oh, and Scheuren, 1986). Subsequently, no record was given a weight of less than three, all amount fields were rounded to four significant digits, top coding was applied for selected codes, and some fields were eliminated for high-income records. In addition, certain fields were blurred or microaggregated in groups of three.

During the 1990's SOI, along with all of the other statistical agencies that release PUFs, reexamined its disclosure control procedures in light of technological changes (increased computer power, decreased storage costs, advances in record linkage techniques, and the proliferation of information networks such as the internet). SOI's current approach is to determine what items in the PUF can be obtained by an outside intruder. After the suspect fields have been identified, an extract from the IRS Individual Master File is made which contains these fields for all taxpayers. This extract and the, as of yet unreleased, PUF are then matched using record linkage software. If the results cause alarm, additional blurring or subsampling is performed.

This process provides SOI with what SOI believes is a limited but objective measure of disclosure risk. An obvious question that arises is what is the relative impact of the various disclosure procedures upon the risk of disclosure. For example, if the subsampling procedure limited records to a minimum weight of 5 instead of 3, how would the disclosure risk measurement change? If the records were microaggregated in larger groups and in a less rigid hierarchical order, how would the disclosure risk measurement change? Of course, the next obvious question that arises is what impact do disclosure control procedures have on data quality? In the end, the disclosure process is a constant effort to produce PUFs that retain as many qualities of the original data as possible while maintaining confidentiality. What follows are some of the results of our attempt to answer these questions.

2. Disclosure Protection of PUF

The creation of the PUF involves four steps: (1) preparation of an internal nonPUF and the application of SOI edits to the taxpayer-reported data, (2) subsampling of high-income returns that are included in the nonPUF with certainty (returns in the 100-percent sampling strata), (3) application of microaggregation procedures to sensitive data fields and other disclosure procedures (suppression, top coding, etc), and (4) the rounding of numeric values to four significant digits. Our evaluation examines options for subsampling high-income returns and for microaggregation.

2.1 Subsampling Options

SOI prepares two versions of the Individual Tax Model File each year, a nonPUF file for analyses by SOI, the Treasury Department and Congress' Joint Committee on Taxation and a PUF for public release. The nonPUF consists of an annual cross-sectional sample of individual tax returns. The chance that returns are sampled is determined by a composite income amount field (created by SOI for sampling), and the forms and schedules used for filing tax returns. For high-income returns with selection income (or loss) amount exceeding $5,000,000 and for returns with selection amounts of over $200,000 in nontaxable income, the nonPUF includes them with certainty and the PUF subsamples them at a rate of 1 in 3 for disclosure protection.

Subsampling for disclosure protection is a form of suppression. The lower the sampling rate, the less chance that a given rare return appears in the PUF. The consideration is how to select a suitable sample and maintain adequate sample size to ensure unbiased and accurate estimates of the population.

We compared two sampling options: (1) the current method of selecting a stratified systematic sample, and (2) the potential use of a balanced random sample at a lower sampling rate. The current method involves stratification by the type of tax forms filed with the return and the selection income amount. Within the certainty strata, individual tax returns are sorted by Adjusted Gross Income (AGI), very rare returns are removed, and the remaining returns are sampled systematically at a rate of 1 in 3 returns. This sampling method ensures that the sample units are evenly distributed and are representative of the population (Kish, 1965).

Valliant, Dorfman, and Royall (2000) refine the notion of a "representative sample" into the notion of a "balanced sample". One of the aims of balanced sampling is to provide better protection against bias in estimation (bias-robust estimation) under a class of superpopulation models. A sample is "balanced" for a given set of control variables if the sample moments equal the population moments. For first-order balance, the sample mean equals the population mean. Higher order balance can also be used. For example, samples can be restricted to ones where the first four sample moments (i.e., mean, variance, skewness, and kurtosis) are close to the population moments. The strategy for selecting a balanced sample involves the idea of randomization.

For the SOI individual tax model PUF, we drew a balanced sample using a stratified restricted random sampling plan with a sampling rate of 1 in 5 in each stratum. Stratification and the removal of very rare returns used the same current sample method. The proposed balanced sampling steps were (1) specify control fields and acceptance criteria for closeness to "balance", (2) select a stratified simple random sample without replacement, and (3) retain the sample if acceptance criteria are satisfied, otherwise replace the sample into the population and repeat step (2). We used as control fields the same fields selected for disclosure protection by microaggregation. The acceptance criteria we used were to retain samples for which the sample and the population moments differ by less than 5 percent for mean, and 10 percent for variance, skewness, and kurtosis per field. Among the collection of samples that met the acceptance criterion, we selected one of the balanced samples by further considering how well sample and population percentiles matched.

The proposed balanced sample for the certainty strata is 3/5 the size of the current PUF certainty strata sample. It should afford noticeably better disclosure protection. However, it does lose some analytic power because of the smaller sample size. We investigate this further in Section 3.

2.2 Microaggregation Options

Disclosure protection of individual tax returns in the PUF uses well-known statistical disclosure control (SDC) procedures including suppression, top coding, rounding, and microaggregation. Microaggregation is a perturbation disclosure technique introduced by Strudler, Oh, and Scheuren (1986) for the Individual Tax Model PUF. The idea is to apply the practice of the "rule of 3" to individual data. Any observed value with a frequency of less than three is deemed confidential.

Currently microaggregation is applied to sensitive data fields such as wages and salaries, real estate taxes, state and local taxes, and business net receipts for which external data may be available. The procedure involves forming aggregation classes defined by filing status (married filing jointly or other), number of exemptions, and income. Within each class, data fields for aggregation are individually ranked and aggregated in a fixed-group size of three (MicIR3). Relative to other perturbation techniques, the current SOI method of microaggregation ranks the best in limiting information loss but poorest in disclosure protection (Domingo-Ferrer and Torra, 2001).

Several recent researches have discussed the pros and cons of microaggregation (Defays and Anwar, 1998; Willenborg and de Waal, 2000) and proposed alternative methods of implementation (Sande, 2001, Domingo-Ferrer and Mateo-Sanz, 2002). Variants to the basic SOI approach include the use of (1) larger fixed-group size k, (2) variable-group size allowing k to vary according to data distribution (treating this as a clustering problem with a variable number of clusters and a minimum cluster size), and (3) multivariate microaggregation using distance or projection methods to form aggregate groups.

For the SOI Individual Tax Model PUF, we explored a hybrid form of individual ranking microaggregation. Our approach is MicIR$g|k$, individual ranking with the partition group size g and aggregation group size k, for $g > k$. First, we formed aggregation classes similar to the current method. Within each class, data fields were again individually ranked and partitioned into contiguous groups of size $g = 30$. Within partition groups, returns were randomly reshuffled and aggregated by groups of three. This approach follows the same idea that no data value in the PUF has a frequency less than k, the minimum requirement for confidentiality. However, the units in an aggregate group are not necessarily of consecutive rank. This modification allows more variations within aggregation group. The maximum variation is controlled by the partition group boundaries.

3. Evaluation Method

SOI made available three data files with 1998 tax returns for this evaluation: (1) an abridged population source file (SF), (2) a nonPUF, and (3) a prerelease PUF. The SF includes 24,901 high-income individual tax returns and the original taxpayer-reported data on selected tax return fields (see fields used in record linkage analyses in Section 3.1). The nonPUF contains a sample of 1998 individual tax returns and data edited by SOI for data consistency. The rare tax returns in the SF are all included in the nonPUF sample. The PUF is prepared from the nonPUF by subsampling the high-income returns and applying disclosure protection procedures. Data from the three files allow us to systematically measure disclosure risk and information loss after the successive changes due to editing, subsampling, and microaggregation. A numeric return ID is included in each file to help us determine whether true matches can be made between the PUF (or nonPUF) and the SF.

3.1 Disclosure Risk

We used two methods to measure disclosure risk: (1) a record linkage approach to determine the potential risk of matching true data to perturbed data in the PUF and (2) an Euclidean distance measure to determine the potential risk that the perturbed data remain closest to the true data. Both methods depend on access to true data, and, even today, access to such data is not an easy task for most people. Therefore, our evaluation is considered conservative, measuring “potential” disclosure risks contingent upon data availability.

The record linkage approach used the commercial software AutoMatch (Matchware, 1996, Jaro, 1989). This package follows the Fellegi and Sunter (1969) framework of probability matching and is evolved partly from the match system used by the U.S. Bureau of the Census (Winkler, 1995). AutoMatch provides an iterative option for parameter estimation, calculates the log-odds match weights for record pairs assuming independence between matching fields, and uses a linear sum assignment algorithm to assign one-to-one matched pairs. This package includes a number of options that allows us to handle specific matching rules and allow for partial agreements in the matching fields (see Winglee, et al., 2000; Gomatam et al, 2002; Winglee and Valliant, 2002).

Record linkage in our evaluation used five match fields, the four key fields masked by microaggregation and a childcare earned income field masked by top coding. These fields were selected based on investigations of available data from outside sources. Linkage comparison allowed a tolerance for partial agreement (up to a 5% difference in the log scale) per match field. This procedure compared record pairs within blocks defined by marital status (married or single), number of children at home (none, one, two, and three or more children), and presence of foreign income (yes or no).

We also used a distance-to-self score to compare the Euclidean distance between pairs of returns. Specifically, the distance score d_{iI} between return *i* in the PUF (or nonPUF) and return *I* in the SF is computed as $d_{iI} = \sqrt{\sum_j (x_{ij} - X_{Ij})^2}, j = 1,2,3,4$ for the four fields perturbed by microaggregation, where x_{ij} is the masked data for field *j* and return *i* in the PUF and X_{Ij} is the reported data for the same field *j* and return *I* in the SF.

We defined linkage risk as the percent of returns in the high-income return population that are correctly matched with match weights exceeding a threshold level. We used a selection threshold weight where the chance of correct matches is close to 100 percent. Distance risk is defined as the percentage of returns where the distance-to-self score is the shortest or tied for shortest with fewer than three other record pairs.

Table 1 shows disclosure risks under the current and proposed method of processing the PUF after editing, subsampling, and microaggregation. For linkage risk with the current method, 18.7 percent of rare returns in the SF are correctly matched to the nonPUF data after editing; 6.2 percent are matched after editing and subsampling; and 4.9 percent are matched to the PUF data after editing, subsampling, and microaggregation. The distance evaluation shows similar improvements by the successive processes. Relative to the current method, the disclosure risks under the proposed method are substantially lower, the potential linkage and distance risks after editing, subsampling, and micoraggregation are 0.4 percent and 1.1 percent.

Table 1. Potential disclosure risks

		Percent correct matches*	
Evaluation	Process	Current method	Proposed method
Record linkage	Editing	18.7	18.7
	Subsampling	6.2	3.9
	Microaggregation	4.9	0.4
Distance-to-self	Editing	47.0	47.0
	Subsampling	15.5	9.3
	Microaggregation	11.9	1.1

* Percent of tax returns in the population SF correctly matched to returns in the nonPUF and PUF.

Figure 1 shows histograms of record linkage match weights for true and false match pairs using the current and proposed methods to process the PUF. With the current method, nearly all pairs with match weights of 24 or greater are correct matches. In contrast, with the proposed method, relatively few record pairs had match weight above the threshold (0.4% of high-income returns in population), of which, 87 percent are true matches. Below the selection threshold of 24, 49 percent of pairs are false matches with the current method while 90 percent are false matches with the proposed method. Note that this threshold for risk assessment is less conservative than is sometimes used. For example, Yancey, Winkler, and

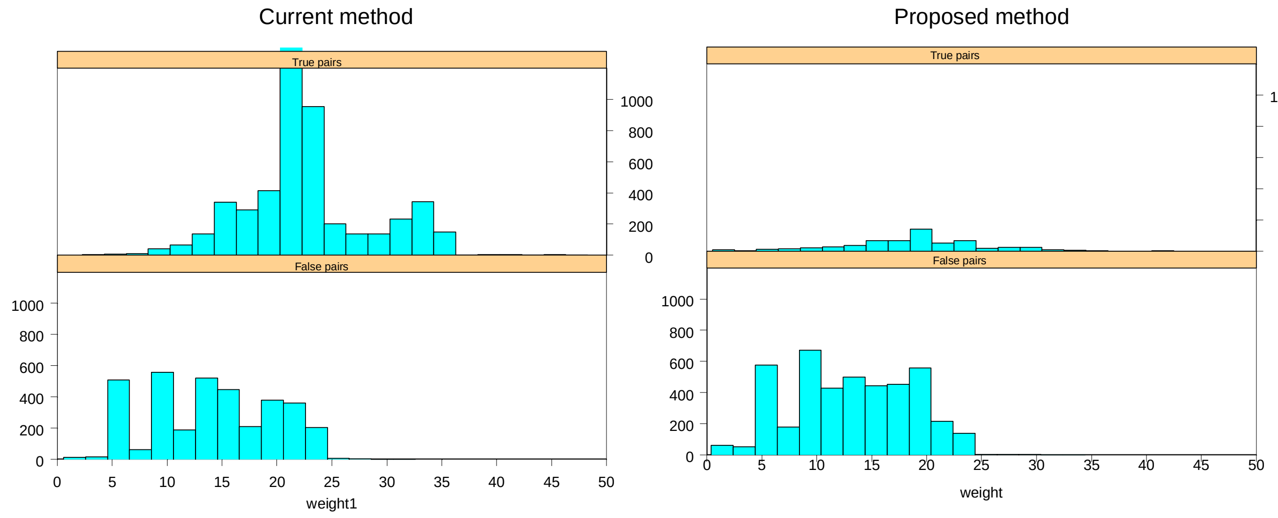

Figure 1. Histogram of match weights for true and false match pairs

Creecy (2002) identified cases as being at risk of disclosure if their probability of correct match was 20 percent or more.

3.2 Information Loss

To measure information loss, we also used two measurements. The first was a composite moments score to measure, for each field, the difference in population and sample moments (mean, variance, skewness, and kurtosis) resulting from disclosure procedures. The second was a measure of relationships between fields. We used a relative correlation score to measure differences in the population and sample pairwise product moment correlation and rank correlation for data fields that are often used in tax model analyses.

Table 2 shows the percentage difference in mean, variance, and composite moments score for selected fields. For example, the percentage difference in variance is computed by taking the weighted sample estimates of population variance minus the actual population variance divided by the actual population variance. The composite moments score m is a weighted average of the differences across all four moments computed as follows:

$$m=\frac{1}{6}\left(2*\frac{|m_1-M_1|}{M_1}+2*\frac{|m_2-M_2|}{M_2}+\frac{|m_3-M_3|}{M_3}+\frac{|m_4-M_4|}{M_4}\right),$$

where m_1 is the sample mean, M_1 is the population mean, m_2 is the sample estimate of the population variance, M_2 is the actual population variance, etc. This composite moments score is a weighted average of the relative difference in the four moments where differences in mean (m_1) and variance (m_2) are weighted twice as important as skewness (m_3) and kurtosis (m_4). This score is zero if the sample moments are exactly equal to the population moments.

Table 2. Percentage difference in mean and variance and a composite moments score for selected tax fields: current and proposed methods

	Percentage difference*				Composite	
	Mean		Variance		moments score	
Tax field	Current	Proposed	Current	Proposed	Current	Proposed
Wages and salaries**	4.2	2.5	17.2	(2.5)	0.09	0.08
Real estate tax**	1.8	(0.1)	(8.0)	(2.5)	0.18	0.11
State and local tax**	2.9	1.3	18.6	2.4	0.12	0.02
Business net receipts**	(16.4)	(8.6)	(55.5)	(38.2)	0.39	0.25
Adjusted Gross Income	1.9	(1.7)	14.2	(8.2)	0.13	0.10
Income tax before credits	2.3	(0.9)	13.4	(7.8)	0.08	0.07
Net capital gains	(0.0)	(3.2)	14.6	(15.2)	0.13	0.14

* Percentage difference between sample and actual population moments relative to the actual population moment. (Numbers in parentheses are negative).

**Fields perturbed by microaggregation and "balanced" under the proposed method.

For the control fields used in balanced sampling, the proposed method guarantees that the sample moments are "close" to the true population moments and the gains in sampling help to offset losses from the modified microaggregation procedure. As a result, the proposed method provides better estimates of the population moments. For fields not used as control fields for subsampling and not affected by microaggregation, the proposed method is not always better than the current method. For instance, the mean of sample net capital gains for the current method is

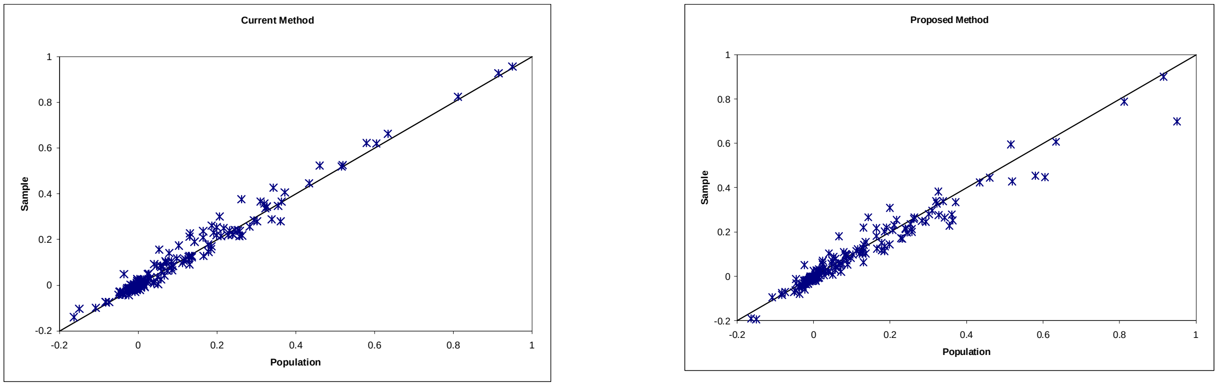

Figure 2. Pairwise product moment correlations for 20 variables after subsampling and microaggregation

equal to the population mean but is 3.2 percent less than the population mean for the proposed method.

To monitor changes in relationships with multiple variables, we selected a set of 20 fields often used in tax model analyses (see for example Feenberg and Coutts, 1993) and computed product moment correlation and rank correlation for all possible pairs of fields (i.e., 190 correlations). We computed the correlations using population data and sample data with the current and the proposed method to process the PUF.

Figure 2 shows scatter plots of the correlations after subsampling and microaggregation under the current and proposed methods. Both methods preserve the correlations reasonably well, although the proposed sample method does yield sample correlations that are lower than those in the population in a number of cases. Recall, however, that the sample size in the proposed method is only 3/5 of that in the current method.

We used a relative correlation score to summarize the sample and population differences for all 20 fields as follows:

$$r = \frac{\sum_{j<j'}\sum \left| r_{jj'} - R_{jj'} \right|}{\sum_{j<j'}\sum R_{jj'}}, \; j, j' = 1, \ldots, 20 \,,$$

where $r_{jj'}$ is the correlation of the jj' pair of fields in the PUF and $R_{jj'}$ is the corresponding correlation of the same pair of fields in the nonPUF. A score of zero means that sample correlation is exactly equal to population correlation for the selected fields.

Table 3 shows the relative product moment and rank correlation scores using the current and the proposed methods of processing the PUF. After subsampling, the relative correlation score is 0.18 for the current systematic sample and 0.20 for the proposed balanced sample. The small difference may be a result of smaller sample size in the balanced sample. After microaggregation, the correlation scores for the two methods are 0.19 and 0.25, showing more perturbation from the proposed microaggregation scheme.

Table 3. Relative correlation scores: current and proposed methods

	Current method		Proposed method	
Relative correlation score	Sub-sampling	Sub-sampling and MicIR3	Sub-sampling	Sub-sampling and MicIR30/3
Product moment correlation	0.18	0.19	0.20	0.25
Rank correlation	0.06	0.06	0.06	0.06

4. Discussion

The need to strengthen disclosure protection is a pressing issue facing many federal agencies. For the SOI individual tax model PUF, the current method of disclosure protection is analyst-friendly (least information loss relative to alternative choices of data perturbation techniques). The concern is that if data became available for data linkage, some high-income returns might be correctly matched. Ad hoc changes to data fields are time consuming and are unreliable solutions.

This study explored two options to lower linkage risk for the SOI individual tax model file. First, we propose a smaller subsample of high-income tax returns in the PUF to lower the chance of exposure. Balanced sampling is a technique that allows us to control for "balance" in sample estimation and ensure that the sample of high-income returns is a good reflection of the population. We used a balanced random sample controlling the sample estimates for fields selected for perturbation. The list of control fields can be extended to include other tax modeling key items such as AGI, income tax before credits, and net capital gains. A smaller and better subsample of high-income returns could improve both disclosure protection and sample estimation.

Second, options to improve the perturbation of sensitive data fields are more complex because there is no easy solution to minimize information loss and maximize disclosure protection. The current method of individual ranking microaggregation has many desirable features suitable for the SOI tax model file. We considered a simple modification by forming larger rank-ordered contiguous partition groups and small random aggregate groups within partition groups. Aggregate group size is kept small to meet the minimum confidentiality requirement. Members of the aggregate group are more variable for better disclosure protection. The larger perturbation from modified microaggregation is offset to some extent by the improved subsampling method.

The combination of better subsampling and microaggregation can lower the potential disclosure risk for the Individual Tax Model file. The next steps are to consider further research to determine the impacts of different disclosure techniques on tax model analyses, ways to refine the balanced subsample of rare returns, and ways to determine suitable partition group size with individual ranking microaggregation or alternative data perturbation methods.

5. References

Defays D., and Anwar M.N. (1998). Masking microdata using micro-aggregation. *Journal of Official Statistics*. Vol. 14, No. 4, 449-461.

Domingo-Ferrer, J., and Mateo-Sanz, J.M. (2002). Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 14, No. 1, pp. 189-201.

Domingo-Ferrer, J., and Torra, V. (2001). *A quantitative comparison of disclosure control methods for microdata.* In Doyle P., et al. (ed.) Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, Elsevier Science, B.V., Netherlands, pp. 111-134.

Fellegi, I.P., and Sunter, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, pp. 1183-1210.

Feenberg, D. R., and Coutts E. An Introduction to the TAXSIM Model. *Journal of Policy Analysis and Management* 12, No. 1 (Winter 1993), pp. 189-194.

Gomatam, S., Carter, R., Ariet, M., and Mitchell, G. (2002). An empirical comparison of record linkage procedures. *Statistics in Medicine*, 21, pp. 1485-1496.

Jaro, M.A. (1989). Advances in record linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association,* 84, pp. 414-420.

Kish, L. (1965). *Survey Sampling.* New York: John Wiley and Sons, Inc.

Matchware Technologies Inc. (1996). *AutoMatch: Generalized Record Linkage System User's Manual.* Silver Spring, MD: Matchware Technologies, Inc.

Sande, G. (2001). Methods for data direction microaggregation in one or more dimensions. *Federal Committee on Statistical Methodology Research Conference*, Thursday A Sessions, November 15, pp. 1-10.

Strudler, M., Oh, H., and Scheuren, F. (1986). Protection of taxpayers' confidentiality with respect to the tax model. *Proceedings of the Sections on Survey Research Methods of the American Statistical Association*, pp. 375-381.

Valliant, R., Dorfman, A., and Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: John Wiley and Sons.

Willenborg, L. and de Waal, T. (2000). *Elements of Statistical Disclosure Control.* Lecture Notes in Statistics 155, Springer-Verlag, New York.

Winglee, M., Valliant, R., Brick, J.M., and Machlin, S. (2000). Probability matching of medical events. *Journal of Economic and Social Measurement*, 26, pp. 129-140.

Winglee, M., and Valliant, R. (2002*) Selection threshold and linkage errors.* Draft manuscript.

Winkler, W.E. (1995). *Matching and record linkage.* In Cox, B.G., *et al.* (ed.) Business Survey Methods, New York: J. Wiley, pp. 355-384.

Yancey, W.E., Winkler, W.E., and Creecy, R.H. (2002) Disclosure risk assessment in perturbative microdata protection. U.S. Bureau of the Census, Statistical Research Division, *Research Report Series*, (Statistics # 2002-01).

Note: Fields Used in Correlation Analyses: Salaries and Wages (Form 1040, line 7) (E00200), State and local income taxes (Form 1040, schedule A, line 5) (E18400), Real estate tax deductions (Form 1040, schedule A, line 6) (E18500), Business net receipts (Form 1040, schedule C, line 3) (E90040), Earned income for child care credit (Form 2441, line 4) (E32900),Taxable interest income (Form 1040, line 8a) (E00300), Investment dividends (Form 1040, line 9) (E00600), State tax refunds (Form 1040, line 10) (E00700), Net capital gain or loss (Form 1040, line 13) (E01000),Total pensions and annuities (Form 1040, line 16a) (E01500), Adjusted gross income (Form 1040, line 33) (E00100), Income tax before credits (Form 1040, line 40) (E05800),Foreign tax (Form 1040, line 46) (E07300), Self-employment tax (Form 1040, line 50) (E09400), Income tax withheld (Form 1040, line 57) (E10700), Balance due (overpayment) (Form 1040, lines 65 and 68) (E11900), Total interest paid deduction (Form 1040, schedule A, line 14) (E19200), Chari gifts deduction (Form 1040, schedule A, line 18) (E19700), Net casualty or theft loss (Form 1040, schedule A, line 19) (E20500), Business expenses (Form 1040, schedule C, line 28) (E90100).