TPC TAX POLICY CENTER
URBAN INSTITUTE & BROOKINGS INSTITUTION

A SYNTHETIC INCOME TAX RETURN DATA FILE: TENTATIVE WORK PLAN AND DISCUSSION DRAFT

Leonard E. Burman, Alex Engler, Surachai Khitatrakun, James R. Nunns, and Sarah Armstrong

June 30, 2017

ACKNOWLEDGEMENTS

We are grateful to John Abowd, Brian Balkovic, John Czajka, Dan Feenberg, Barry Johnson, Graham MacDonald, Rob McClelland, David McClure, Rob Santos, Mike Strudler, Lars Vilhuber, Mike Weber, and members of the SOI Consultants Panel for helpful comments and discussions, to John Iselin for expert research assistance, and to Lydia Austin and Ann Cleven for coordinating the project. This publication relies on the analytical capability that was made possible in part by a grant from the Laura and John Arnold Foundation.

The findings and conclusions contained within are those of the authors and do not necessarily reflect positions or policies of the Tax Policy Center or its funders.

ABSTRACT

Administrative tax data contain a wealth of information that is potentially valuable for research and analysis. However, the legal and ethical imperative to protect taxpayer privacy has restricted access to a small number of government analysts and select researchers. We propose to develop in consultation with the experts at the Statistics of Income division of the IRS a fully synthetic tax database – that is, a file that preserves many of the statistical characteristics of the restricted data without containing any identifiable tax return information. We will test our procedures using the existing public use file and adapt the procedures to run on the confidential tax data. Working with the IRS, we also hope to develop a procedure for researchers to submit their statistical programs, which have been tested on the synthetic data, to run on IRS computers subject to a review to guarantee that output satisfies disclosure avoidance protocols. A fee structure would be set to defray costs.

CONTENTS

I. INTRODUCTION

Administrative tax data (taken directly from individuals' and businesses' tax and information returns) are potentially enormously valuable for informing the public about a wide range of issues, some of which go well beyond tax policy. For example, Chetty and colleagues (2014) used tax data to illuminate the public debate about economic mobility across generations.

At present, however, researchers outside of government have very limited access to administrative tax data. After a lag of several years, the Internal Revenue Service (IRS) releases a public use file (PUF) based on a sample of individual income tax returns, but many potentially valuable variables, such as geographic location, the split of wages between spouses on joint returns, the ages of filer(s) and family members, and a wide range of other variables that are available to the IRS are either excluded altogether from the PUF, or only available on certain returns in cursory form. In addition, some high-income tax returns are aggregated and information from other returns is deliberately "blurred" to protect against the risk of disclosure. Over time, more and more restrictions have been added to successive PUFs, and the viability of future files is at risk because of the possibility of disclosing the identity of a taxpayer through matching PUF data with the growing volume of personal information available online. Thus, it is imperative that a replacement for the PUF be found soon.

Moreover, providing researchers greater access to administrative tax data could vastly expand our understanding about how tax policies affect behavior (and how those policies could be made more effective), but the goldmine of administrative tax data is only available inside select government agencies and via collaboration with analysts in these agencies or through highly restrictive arrangements with the IRS. Expanding access to administrative data would represent a major advance in the ability of TPC and the broader research community to develop economic knowledge that could be applied to public policy debates.

We propose a two-part approach to expanding researchers' access to administrative tax data: (1) creating one or more fully synthetic public use files that have been purged of personally identifiable information, such as unique tax return information, that could be matched with data from other sources, and (2) developing a secure process by which researchers could submit statistical programs that have been tested on the synthetic data to be executed on IRS computers with the statistical output emailed to the researchers after a disclosure review.

Fortunately, data science has made great strides in producing high quality synthetic data—that is, files that preserve important statistical characteristics of the administrative data while protecting privacy. We will experiment with several data-synthesis methods, including parametric and nonparametric models for replacing actual data with predicted values. A particularly promising nonparametric method, Classification and Regression Trees (CART), sorts observations into relatively homogeneous groups and draws from the empirical distribution of outcomes that occur for each group. There are computational and analytical challenges in implementing this method on a large scale, but we believe it could be a good option for certain variables.

We propose to create a fully synthetic dataset by applying an iterative technique, such as sequential regression multiple imputation (SRMI), that preserves the modeled relationships among variables.

We also propose to test a novel second round of refinement for the synthetic dataset. After synthesizing the roughly 200 million records in the population of tax records (including nonfilers who are represented on information returns like forms W-2 and Social Security Administration records), a subsample of, say, 500,000 records will be chosen to minimize the distance between sample and population statistics. We believe that this method will allow use of synthetic data for a broad class of problems related to the target statistics without requiring the correction of standard errors for valid statistical inference (as is required for data generated by standard multiple imputation methods). For example, if we target population means, variances, and covariances, the statistics based on those parameters (such as linear regression coefficients) in the sample should provide consistent estimates of population statistics.

This paper is organized as follows. Section II discusses individual income tax data compiled by the Statistics of Income division of the IRS, the nature of disclosure concerns, and how those are addressed in the PUF. Section III discusses the main methods used to produce synthetic data and surveys some of the applications to the release of administrative data by other agencies. Section IV discusses the special challenges for creating synthetic data, preventing disclosure, and addressing the computational demands in tax data. Section V outlines our proposed strategy for addressing those challenges and producing a synthetic tax database. We plan to develop our synthesis routines and computer programs using the public use file as a test database, as discussed in Section VI. Section VIII outlines our plan for producing a secure statistics server at the IRS. And Section IX presents concluding remarks.

This paper is very preliminary and intended to provide the basis for discussion and feedback. Please send any comments to Len Burman at lburman@urban.org.

II. ADMINISTRATIVE TAX DATA AND THE PUBLIC USE FILE

This section provides background on tax return filing, return population files, SOI sample files, disclosure risks, and the disclosure avoidance procedures that are currently used to create the PUF.

Federal individual income tax returns are filed annually with the IRS by most adults and some children. In 2016, 150.6 million individual income tax returns, covering 293.2 million taxpayers and dependents, were filed with the IRS.[1] Most of those returns covered the income, deductions, taxes, and credits of taxpayers in tax year 2015. Income tax returns may contain a large number of data entries: In addition to the basic income tax return, Form 1040 (which in 2015 contained over 80 possible entries),[2] tax returns may include one or more schedules (in 2015 there were 12 schedules to Form 1040, such as Schedule A for reporting itemized deductions), each with multiple possible entries, and other forms (of which there were over 60 in 2015) with multiple possible entries for providing additional information and computations that support the entries on form 1040. Further information may be supplied as marginal entries on form 1040, in attachments of information returns (such as for wage withholding reported to employees on form W-2), and in attachments prepared by taxpayers.

Nearly 88 percent (132.3 million) of the returns filed in 2016 were filed electronically, so the IRS has a complete electronic record of these returns and all return schedules, supporting forms, and attachments. Electronically filed returns are typically subject to consistency checks via tax preparation software and also receive some preliminary testing before they are accepted by the IRS, so they contain few typos or math errors. The remaining 18.3 million returns in 2016 were filed on paper. The IRS captures the information from paper-filed returns electronically, then tests this information and sends notices to taxpayers if errors are detected.[3] The fraction of returns that is filed on paper rather than electronically is relatively constant across income groups.

POPULATION FILES

The electronic records of all individual income tax returns filed each year are part of the Individual Master File (IMF).[4] The IRS also maintains electronic records of all information returns related to individual taxpayers (such as W-2s and 1099s) that are filed with IRS by employers, banks, and other entities each year for activity involving individuals in the preceding year. Most of these information returns are not filed with Forms 1040, so provide supplemental information to the IRS on taxpayers' income, deductions, taxes or credits.[5] In addition, the Data Master (DM-1) file, provided and regularly updated by the Social Security Administration (SSA), contains information on taxpayers' and dependents' date of birth, gender, name changes, and, for deceased individuals, date of death.

The IMF, information returns, DM-1 file, and other information[6] maintained at the IRS for each year represent the available administrative data on the entire individual income tax filing population as well as most nonfilers (based on the information from SSA).[7]

SOI SAMPLES

The annual sample of individual income tax returns drawn by the SOI Division of IRS is based only on the population of individual income tax returns processed by IRS during the year, and is drawn at an initial stage of processing, prior to the IMF being available.[8] SOI uses full sample, called the INSOLE file,[9] to prepare publications and other products, and by the Office of Tax Analysis (OTA) in the US Treasury and the staff of the Congressional Joint Committee on Taxation (JCT) in their microsimulation models and for other analyses. The INSOLE is also used to create the PUF. Neither the INSOLE nor the PUF contains information about nonfilers, but OTA and JCT create non-filer records from information returns and DM-1 information.

INSOLE

The INSOLE[10] sample is selected from all individual income tax returns processed during a year by IRS and posted to the IMF except tentative returns, amended returns, and returns that report no income.[11] Selection for the sample is based on the size of "total positive income" or, if larger in absolute value, "total negative income." These two amounts are the sum of nearly all positive and all negative items of income reported on a return. Based on the larger of these two amounts, all returns on the IMF that are eligible for selection are assigned to one of 24 strata: 9 negative income strata (strata 1 through 9) or one of 15 positive income strata (strata 10 through 24). Strata boundaries are dollar amounts that are expressed at their original values, established in 1991. However, in selecting the sample the dollar amounts of "total positive income" and "total negative income" on each return are deflated by the change in the chained gross domestic product (GDP) implicit price deflator between 1991 and the tax year of the sample.[12] Sample rates vary by strata, from a rate of about 0.1 percent (1 in 1,000) in strata 10 through 16, to 100 percent in strata 1, 2, 23 and 24. There are also two special strata for returns sampled at 100 percent, one for returns with gross receipts from one or more nonfarm or farm sole proprietorships reported on Schedules C and F of $50 million or more (stratum 201), and another for "high-income nontaxable returns" (HINTS), which are returns with income of $200,000 or more that report no income tax liability (stratum 101). In addition, periodically sample rates in certain strata are increased to insure an adequate sample of returns that claim an exclusion for foreign earned income on Form 2555.

Returns in the two special 100 percent strata (101 and 201) are selected for the full sample first. Sampling of remaining returns within each stratum is based on SSNs.[13] For returns in all strata, the last four digits of SSNs are examined (there are 9,999 such endings, since SSNs do not end in 0000). Any return with one of 10 specified endings is sampled (making the sample rate 10 in 9,999 or slightly more than 1 in 1,000, or 0.10001 percent). The 10 endings used for sampling are part of the Continuous Work History Sample (CWHS) designated by the Social Security Administration (SSA) for research purposes. For returns in strata 10 through 16, only CWHS returns are selected for the sample. For returns in the other strata (1 through 9 and 17 through 24), which have sample rates above the CWHS rate, non-CWHS SSNs are transformed (to correct for slight non-randomness in SSNs), and enough endings of the transform are selected to achieve the sample rate for the stratum (taking into account the CWHS portion of the sample).

The full SOI sample selected in 2015 (primarily of returns filed for 2014) included records for 343,748 returns (of 149.6 million filed), with 71,033 of those returns (nearly 21 percent of the total) selected with certainty (from strata 1, 2, 23, 24, 101, and 201).[14] Excluding HINTs (strata 101), 40,150 returns (12.8 percent of the total sample excluding HINTS) were selected with certainty. The minimum total positive (total negative) income required to be selected in strata 23 or 24 (1 or 2) was $7.7 million (-$7.7 million) in 2014.

SOI includes in the full sample file almost all of the information reported by taxpayers on their income tax return Forms 1040 (including marginal entries), all Form 1040 schedules, all attached forms (including W-2s), and supplements that return information with the date of birth, gender, and, for decedents, date of death of the taxpayer(s) and any dependents from the DM-1.[15]

PUF Sample

The PUF sample is a subsample of the full sample. Excluded from the PUF sample are all returns included in the INSOLE sample that were filed for taxable years more than three years prior to the current year and any oversampled Form 2555 returns. HINTS (strata 101) are placed in the strata (1 through 24 or 201) they would otherwise be assigned to and then subsampled for the PUF in the same manner as other returns in each stratum. All CWHS returns selected in the full sample are subsampled at a 70 percent rate for the PUF sample (i.e., only 7 of the 10 CWHS endings are included, making the sample rate from the population 0.07 percent). In strata 7 through 18 (total incomes between -$385,075 and $385,075 in 2014) only (the subsampled) CWHS returns are included in the PUF sample.[16] All returns in strata 5 and 6 (total negative income between -$1,540,300 and -$385, 076) and 19 and 20 (total positive income between $385,076 and $1,540,300), which for the full sample are sampled at rates of about 1 percent (strata 6 and 19) and 3 percent (strata 5 and 20), are included from the PUF (except the 30 percent of CWHS returns in these strata excluded in the earlier step). Returns in strata 3 and 4 (total negative income between -$7.7 million and -$1,540,300) and 21 and 22 (total positive income between $1,540,300 and $7.7 million), which for the full sample are sampled at rates of about 12 percent (strata 4 and 21) and 33 percent (strata 3 and 22), are selected for the PUF at a 10 percent rate (i.e., the PUF sample rates are about 1.2 percent and 3.3 percent). Some of the returns in the certainty strata (1 and 2, 23 and 24, and 201), those with “extremely large” values for one or more key variables, are aggregated into one of four records as a disclosure avoidance measure.[17] Generally, an amount of income (loss) from any one source that is among the 30 largest (smallest) amounts reported on all returns, and a deduction or credit that is among the 10 largest reported, is considered extremely large. The four aggregate records for the most recent (2011 tax year) PUF cover returns with any extremely large value for at least one key variable and (1) negative adjusted gross income (AGI); (2) AGI between $0 and $10 million; (3) AGI between $10 million and $100 million; and (4) AGI over $100 million.[18] The remaining returns in the certainty strata are subsampled at a 10 percent rate (i.e., the PUF sample rate for these returns is 10 percent).

The PUF sample for tax year 2011 included 163,790 return records (including the four aggregate records) representing the 145.2 million returns filed.[19]

In addition to including only a subsample of returns from the INSOLE, the PUF includes only a subset of the data items. The 2011 PUF, for example, includes entries from fewer than half the Form 1040 schedules and from only 8 of more than 60 attached forms that were included in the INSOLE file for 2011.[20] In addition, as discussed below, the data items from the INSOLE that are included in the PUF are further limited or altered to reduce disclosure risk.

DISCLOSURE RISKS IN TAX RETURN DATA

The Internal Revenue Code provides strict protections for the confidentiality of tax return information, and severe sanctions for disclosure of return information.[21] The definition of a disclosure is quite expansive; for example, the identification of a specific taxpayer's record on the PUF would be a disclosure under the Internal Revenue Code, even if all of the information on the PUF record was publicly available. All of the information on an individual income tax return and its associated schedules, other forms, and other attachments are all "tax return information" for disclosure purposes. Returns contain the name, address, and SSN of taxpayers, spouses, and dependents, information that directly identifies these individuals. But even with this information removed (as it is for the PUF), returns and their associated schedules, other forms, and other attachments can, as noted above, contain dozens, hundreds, and even thousands of entries on specific items of income, deductions, tax computations, and credits. Returns also contain information on the demographic characteristics of taxpayers and dependents (e.g., through filing status whether the taxpayer is married, through standard deductions whether the taxpayer, spouse, or both is age 65 or over, and through personal exemptions and certain credits the number and range of the ages of children); on the household's geographic location (e.g., state, even if the complete address is removed); and on the occupation of taxpayers. Other characteristics of a taxpayer (e.g., as a homeowner, employee, or retiree) can often be inferred merely from presence, or absence, of certain items reported on a tax return. Any of this tax return information, possibly in combination with other, publicly available information, potentially could be used to indirectly identify a specific taxpayer's record.

Some characteristics of many (perhaps most) taxpayers, such as their marital status, approximate age, number and approximate age of children, where they live, their occupation, and their status as a homeowner, employee, or retiree can fairly easily be observed (or inferred) directly or from readily available public information. These characteristics, by themselves or even in combination, are rarely so uncommon that they would identify a specific taxpayer precisely or even with high probability. But some combinations of taxpayer characteristics are quite rare, such as very large families with an elderly head. Little additional information about such taxpayers might be needed to identify them with a quite high level of confidence.

A great deal of additional information about individuals may be available from public sources. Wages of many individuals, such as employees of government agencies, are often publicly available or can be quite accurately estimated based on the individual's position and pay scales. The wages of officers and employees of nonprofit organizations and the officers of large companies are also frequently public information. In many other instances, it is not difficult to estimate an employee's wages from their position and employer (which typically are observable). Income from other sources may also be available from public records (e.g., business income reported on business registers and public licensing information), or estimated from other publicly available information. Further, employers, banks, other financial institutions, and other entities have direct access to specific amounts of income paid to an individual, and they often know other characteristics of the individual.

Amounts of deductions and credits might also be publicly available. The amount of charitable contributions made by specific individuals, particularly of large contributions (which are more likely unique), are sometimes made public by the recipient charitable organization. Property tax records are public, and state income tax records are public in Wisconsin. The installation of certain energy-efficient property that qualifies for a credit might also be readily observed.

A great deal of demographic and other information on individuals is publicly available. Many individuals voluntarily supply such information on social media and other websites. Such information may also be obtained or inferred from individuals' browsing, phone, texting, tweeting, and email activities.

Although some characteristics of individuals may not distinguish them from many other individuals, combinations of characteristics might. Over time, the scope of publicly available information on individuals has grown, especially through the internet, and the power of computers and software to link information has also grown. A particular risk is the growing threat of identity theft and theft of individual data which targets sensitive data, including the kind of information that might appear on tax returns. These trends significantly increase the likelihood that an individual represented on any microdata file like the PUF might be identifiable and therefore disclose all items on the record.

DISCLOSURE AVOIDANCE PROCEDURES FOR THE PUF

The highest sampling rate of the population of individual income tax returns included in the PUF is 10 percent, and nearly all of the return population is sampled at a rate of 0.07 percent.[22] Sampling at these relatively low rates in itself reduces disclosure risk because any individual return is unlikely to be in the PUF. The omission from the PUF of most of the schedules and forms that might be attached to a return also reduces the potential for identifying a filer.[23]

In addition to removing certain returns, subsampling remaining returns, and omitting many return entries that appear in the full sample, an additional set of disclosure avoidance procedures are applied to the PUF: some variables are deleted or modified, all amount variables are rounded, and returns are rebalanced.

Deleted Variables

State code is removed from all return records in the PUF because it could provide significant information for identifying taxpayers in small states (and in some circumstances taxpayers in large states). For records selected in strata 1–6, 19–24, or 201, alimony paid and received, the itemized deduction for state sales taxes,[24] and all age, gender, and earnings split variables[25] are removed. The ages of dependents are also removed from some records selected in strata 7–18.[26]

Modified Variables

Fiscal year returns are those filed for periods of less than 12 months or for noncalendar periods of 12 months. Because few individuals file fiscal year returns, for the PUF such returns are converted to calendar year returns for the most recent calendar year. Also, relatively rare surviving spouse returns are converted to joint returns, and head of household returns claiming no dependents[27] and selected in strata 1–6, 19–24 and 201 are converted to single returns.

As indicated above, a large number of dependents could help identify a taxpayer with little additional information. Whether the number of dependents on a return is large depends on filing status, because single and married filing separate filers are much less likely to have dependentsas head of household and joint filers. Those with dependents tend to have fewer as many of them. To address this potential disclosure risk, on the PUF the number of dependents is capped at three for head of household and joint returns, two on single returns, and one on married filing separate returns. These caps are carried through to other return items that are based on the number of dependents, such as personal exemption amounts, the earned income tax credit, and child tax credit.

Certain variables are also "blurred" on the PUF.[28] Blurring reduces disclosure risk by replacing the value of one or more variables on a group of returns with the average value(s) for the variable(s) for returns in the group. For records of returns selected in strata 1–6, 19–24, or 201, the variables blurred are wages and salaries, state and local income taxes, and real estate taxes. Blurring for these records is "multivariate," meaning that all three variables, if present, are simultaneously blurred within a group.[29] For records of returns selected in strata 7–18, wages and salaries, state and local income taxes, and real estate taxes are also blurred, but in addition so are alimony paid and received (variables deleted from other return records) and itemized deductions for medical and dental expenses. The blurring on these records is "univariate," with each variable blurred independently within a group.[30]

Rounding

Rounding reduces disclosure risk in a manner similar to blurring, and can be more effective than univariate blurring if variable amounts are clustered. All dollar amounts on every return is rounded as follows: amounts (in absolute value) over $100,000 are rounded to four significant digits; amounts between $10,000 and $100,000 are rounded to the nearest $100; amounts between $5 and $10,000 are rounded to the nearest $10; and amounts less than $5 are rounded to $2 (with the sign retained).

Rebalancing

Deleting, modifying, and rounding variables changes relationships among some of the variables on a tax return, making them out of balance. For example, AGI is computed as gross income less adjustments to income, so changes to items of income and adjustments will mean that summing income and subtracting adjustments will generally not give the same amount of AGI that was reported by the taxpayer; AGI will be "out of balance". If reported AGI was retained in the return record, it might be possible to infer the effects of, say, blurring, negating the intended reduction in disclosure risk from the blurring procedure. For this reason, return records are "rebalanced" by re-computing gross income, adjustments for education expenses, AGI, taxable income, regular tax, the alternative minimum tax, the child tax credit, the education credits, and tax after credits. However, the deleted (or blurred) income and adjustment items (alimony paid and received) could not be recovered from re-computing AGI in any event because they effectively become part of an implied residual PUF variable that includes "other" income and some adjustments to income. Similarly, itemized deductions are part of an implied residual PUF variable for the sum of total deductions (standard or itemized) and personal exemption amounts, so the deletion (or blurring) of certain itemized deductions is absorbed into this implied residual variable.

As a result of changes made to return entries during IRS and SOI processing, subsampling and omitting many variables from the full sample in constructing the PUF, the application of disclosure avoidance procedures, no entry on a PUF record will exactly match the amount actually reported by the taxpayer, and in some cases (e.g., wages on a high-income return) the discrepancy may be quite large. These differences reduce the likelihood that the record of any specific taxpayer could be identified on the PUF. But these differences also affect how well the PUF represents the return population it is designed to represent,[31] and the research and analysis that can be performed using it. Also, because of the complex set of rules used, it is probably not possible to correct statistical estimates to reflect the measurement error that is introduced.

III. CREATING SYNTHETIC DATA

The goal of data synthesis is to produce data that do not contain any information deemed to be confidential yet at the same time ensure that inferences drawn from such data are, ex ante statistically equivalent to inferences drawn from the actual sample used as a basis for producing the synthetic data. In practice, this task involves constructing a data generating process that contains sufficient details to produce data of high quality for a defined set of research questions using appropriate statistical techniques while maintaining the required level of data confidentiality.

Many synthetic datasets are partially synthetic: only variables or observations deemed especially sensitive are replaced with synthetic values. The current PUF is a kind of partially synthetic file. Aside from rounding, most of the data reported on most observations is reproduced intact. Partially synthetic data have obvious analytical advantages because they may preserve more information than fully synthetic files, but they create significant disclosure risks. In the case of tax data, the existence of actual data means that a number of other useful variables, such as state of residence, must be suppressed.

For those reasons, our goal is to produce fully synthetic data files.

OPTIONS FOR CREATING SYNTHETIC DATA

Much of the research on synthetic data derives from methods for imputing missing data in surveys. The problems are similar in the sense that the goal in each case is to produce values that are statistically unbiased and preserve the relationships among variables. In fact, synthesis is easier because the data generator knows the actual values for the data so, unlike in the missing data case, needs not assume that relationships between imputed and actual values are similar between the records with missing values and others. The challenge in the case of fully synthetic data is that relationships between all of the imputed (synthetic) values should represent the relationships in the actual data. We propose to solve that problem via an iterative process as discussed below.

Choosing the right model or models is key to the success of data synthesis. The validity of statistical inference based on synthetic data depends on the underlying methods of simulation. Bias will result if the models used for simulation do not preserve the critical relationships between variables of interest or if estimation does not account for measurement error introduced by the synthesis process.

Different models may be appropriate for different variables. For instance, different models may be necessary to account for varying data types (real, categorical, or mixed). Even within the same data type, different variable distributions may warrant different models.

The simplest procedure is simply adding random noise to the independent variables in a data set. If the induced errors are relatively small, much of the relationship among variables may be maintained, but this information is preserved at the cost of higher risk of disclosure (Fuller 1993). Errors large enough to prevent disclosure create significant measurement error, which biases statistical estimates. Nonetheless, simple masking might be useful for modeling minor variables that are not likely to be the focus of empirical studies, especially in large samples or where relationships among variables needs to be masked to prevent disclosure risk.

An obvious alternative is to use regression-based imputation (or some other statistical model) to replace actual data with predicted values. Reiter (2004) showed that multiple imputation—using a set of predicted values plus random draws from the empirical error distribution (called replicates)—is appropriate both for imputing missing data and for generating synthetic data. Variance estimates from the imputed data need to be corrected to account for the error introduced in the imputation process, but this is straightforward in most cases.

Bayesian methods offer a natural tool for imputation by drawing from the posterior distribution conditional on other variables in the dataset. Rubin (1978) developed a Bayesian bootstrap approach, but Allison (2000) showed that it could lead to bias in estimation because bootstrapping may not preserve the underlying statistical relationships. Fienberg (1994) proposed using random draws from a regression-based posterior distribution—predicted values plus an error drawn from the empirical distribution of regression errors.

More recent research has developed machine learning methods such as Classification and Regression Trees (CART) (Reiter 2005) and random forests (Caiola 2010). These methods can improve on parameterized approaches for synthetic data under certain circumstances. Other work has found that support vector machines and neural networks can outperform parameterized methods for missing data imputation (Richman, Trafalis, and Adrianto 2009). However, infusing uncertainty into machine learning estimation is not a well-explored task, which poses potential issues for use of these methods.

Emerging research in this field applies a mixture of Poisson distributions in order to impute continuous variables in which marginal sums need to be fixed or consistent. This approach, termed interval-protected multiple imputation, can ensure that totals created from aggregated synthetic data match those of the original confidential data (Wei 2016; Wei and Reiter 2016).

In developing parameterized models, considering grouping and conditioning is often necessary. Conditioning is the selection of independent variables for the estimation of each variable to be synthesized (Benedetto et al 2013). Grouping allows for separate model building and estimation on different subgroups of observations, which allows different relationships among variables within different subgroups. In tax data, it makes sense to group by income and filing status. This not only allows for more flexible estimation, but can significantly speed computation by breaking up the very large full data set into manageable subsets.

Groups must be of sufficient size to prevent over-fitting of imputed values. Benedetto Stinson, and Abowd (2013) recommends that the number of observations within subgroups (created by grouping) must be greater than 15 times the number of conditioning variables, or 1,000—whichever is greater.

We discuss some of the key methods in more detail below.

Multiple Imputation

The term "synthetic data" refers to data resulting from some statistical process that has been applied to an original dataset. This originating dataset is generally restricted, with the goal of the simulation being to create a new dataset that is suitable for public release. To be suitable for public release, the synthetic data must be substantially changed to protect the anonymity of the observations (often individuals). To be informative for research purposes, the synthetic data must also maintain the underlying statistical relationships of the originating data.

Most modern applications of synthetic data generation build upon multiple imputation methodology, a statistical technique originally designed to impute missing data. Multiple imputation was introduced by Rubin (1978), who went on to expand on the topic in a book (Rubin 1987) and several articles (Rubin 1996, 2004). In essence, multiple imputation for missing values recognizes that missing values are stochastic by nature and this stochasticity can be captured by providing multiple imputed values from the empirical distribution of predicted values. Incorporating uncertainty via multiple imputation, improves the reliability of statistical inference, assuming the model that produces the imputations is valid and that the derived statistics are corrected to account for the imputation process (Rubin 1996).[32]

Rubin (1993) and Little (1993) suggested that multiple imputation could be a useful technique to generate synthetic data, although they suggested different approaches. Rubin (1993) suggested merging a smaller dataset onto a larger dataset, such as a census file. If *X* is observed on both data sets and *Y* only on the smaller dataset, a regression of *Y* on *X* (or some other model) could be used to create multiple implicates of *Y**, the predicted value, on the larger dataset. These synthetic data could be used to expand the larger dataset with no concerns about confidentiality because the values would be based entirely on the non-sensitive variable *X*. This is a partially synthetic data file, but the synthesized variable does not exist in the larger dataset before imputation.

Little (1993) proposed using this methodology as a way to replace sensitive data, which is more relevant to our problem. To follow on the example above, consider the case where *Y*, a sensitive variable (subject to disclosure risk), and *X* both exist in a dataset. Use the process described above to create multiple implicates of *Y** in that dataset and then replace the actual *Y* with the synthesized values *Y**. Assuming that the model is not too good, these data protect confidentiality of *Y* while providing information that is useful for analytical purposes (assuming appropriate statistical techniques).

CART

Reiter (2005) suggested applying CART to generate partially synthetic data. The procedure basically involves setting criteria that repeatedly split a sample (analogous to branches on a tree) until each observation is assigned to a branch. There should be many observations (leaves) on each branch to protect against disclosure. The CART algorithm selects a leaf at random using the Bayesian bootstrap algorithm (described in chapter 4 of Rubin 1987).

Alternatively, to avoid releasing actual values, an empirical distribution function may be fitted to the leaves. One synthetic observation is randomly drawn from the empirical distribution for each replicate in a multiple imputation process.[33]

CART models are more flexible than regression-based or other parametric models. They can account for unusual variable distribution and nonlinear relationships among explanatory variables that might be hard to identify and model explicitly. (Reiter 2005) The process of defining branches (or data partitions) can in principle be automated. However, CART may be computationally quite intensive and parametric models may perform better when the relationships between variables can be accurately modeled.

Random Forests

A random forest is a machine learning method that employs stochasticity and many classification or regression trees. A random forest works by running hundreds of decision trees, each predicting the same outcome variable but using different subsets of the rows and columns as observations and predictors, respectively. Each of the many trees run are thus slightly different, and their predictions must be aggregated back together to get a single output for each observation. Generally, for classification trees this is done by voting (the most common predicted outcome is chosen) and for regression trees the many predictions are averaged together. Breiman (2001) argues that random forests significantly reduce the risk of overfitting and has been shown to create more accurate out-of-sample predictions.

Sequential Regression Multiple Imputation

Raghunathan (2001) proposed a variant on multiple imputation that he called sequential regression multivariate imputation (SRMI). SRMI is simply a sequence of regressions designed to produce a set of imputed variables that preserves the conditional means and covariances in the synthetic data. In the first step, a sequence of imputation variables is specified. In Raghunathan's application, the variables were sorted by number of missing values and the one with the fewest missing values was imputed first. We do not have missing data, so we start with the quantitatively or most analytically important variables. The first variable, say Y_1, is regressed against actual values of all the other variables (or a subset if the dataset is quite large as discussed below), producing a predicted value, $\hat{Y}_1^1$, where the superscript refers to the iteration number. Y_2 is regressed against $\hat{Y}_1^1$, Y_3, Y_4, …, producing an imputed value, $\hat{Y}_2^1$, and so on.

In the second iteration, Y_1 is regressed against the imputed values from the first stage for Y_2, Y_3, …, to produce the fitted value $\hat{Y}_1^2$. Then Y_2 is regressed against the updated $\hat{Y}_1^2$ and the fitted values from iteration 1 for all other variables to produce $\hat{Y}_2^2$. Y_3 is regressed against $\hat{Y}_1^2$, $\hat{Y}_2^2$, $\hat{Y}_4^1$, etc. This procedure is repeated until the changes in fitted values from one iteration to the next become very small. van Buuren and Groothuis-Oudshoorn (2011) found that convergence usually occurs within 10 to 20 iterations.

Equations may be fit by various methods—including parametric or nonparametric methods for different variables—following a similar procedure. Thus the method provides a flexible way of dealing with different data types.

As in the standard multiple imputation method, replicates may be created by adding random errors from the empirical distribution for each model equation for a set of replicates. The standard errors are corrected via the same formula as in standard multiple imputation.

EVALUATING SYNTHETIC DATA

Ideally, synthetic data should be of high quality while protecting data confidentiality. These goals lead to conflict. The best quality data are those in the confidential dataset, but publicizing those almost guarantees that there will be a disclosure, especially for observations with unique values (not shared with other observations) where an independent source of information exists. Synthetic data deliberately introduce noise to protect against disclosure. The trick is in drawing the right balance (i.e., finding the best synthesis subject to disclosure constraints).

We discuss how data quality and data confidentiality are measured below.

Evaluating Data Quality

Quality of synthetic data can be measured by their usefulness for general purposes and for drawing inferences. In the literature, the former is commonly referred to as “general data utility,” which focuses on similarity between the joint distributions of synthetic and originating data; the latter is referred to as “specific data utility,” which focuses on similarity between specific analyses’ statistical inferences from synthetic and originating data (Snoke et al 2016). Differences between synthetic and originating data are referred to as information loss.

Some research has replicated prior studies that were originally completed on nonsynthetic data (Dreschler, Bender, and Rässler 2007). By reproducing estimates on the synthetic data, it is possible to measure the potential effect on a real research endeavor. In other studies, researchers chose to examine a plausible analysis of interest, such as running a Tobit regression estimating annual food expenditures in the Consumer Expenditure Survey (Raghunathan, Reiter, and Rubin 2003).

Other studies have looked at descriptive statistics across many perturbed variables, such as comparing confidence intervals between original dataset and synthetic dataset (Karr et al. 2006). An examination of the SOI PUF evaluated changes in the first four moments (mean, variance, skewness, and kurtosis) of variables of interest (Winglee 2002). This paper suggested a composite moments score, in which the first two moments were weighted as twice as important as the third and fourth moments. This same evaluation also used pairwise correlation coefficients and pairwise rank (spearman) correlation coefficients.

It is also likely that different methodologies of synthesis will result in datasets that vary in their suitability for various analyses. Abowd and Lane (2004) suggested that a single confidential file could be used to generate several different synthetic files with different statistical purposes.

And it is likely that certain kinds of models can only be estimated appropriately on the confidential data. For example, Bertrand, Kamenica, and Pan (2015) present clear graphical evidence of a discontinuity in the distribution of married couples at the point where the wife's income exceeds the husband's based on the restricted Survey of Income and Program Participation (SIPP) dataset. Vilhuber and Abowd (2016) show that the same graph based on the synthetic SIPP exhibits no discontinuity. This smoothing occurs because the model that generates the synthetic data was not designed to account for such discontinuities. There are numerous such kinks, notches, and other discontinuities in tax data and they will not be reflected in the synthetic records unless explicitly designed to capture them.

Evaluating Disclosure Risk of Synthetic Data

Despite the fact that synthetic datasets are entirely or largely generated, reidentification may be possible. Both distance-based and probabilistic record linkage techniques have been shown to reidentify synthetic data under certain circumstances.

The problem is typically formulated as one in which an intruder (Reiter 2005) is in possession of some data that are also on the confidential dataset. Can the intruder use the data to find a match in the synthetic data file with high probability? Two key issues are how much data an intruder could possess and what probability threshold constitutes disclosure. The more data assumed to be possessed by the intruder, the higher the likelihood of a match. And, of course, the lower the probability threshold for disclosure, the more challenging is data protection.

This section discusses two related metrics for disclosure risk: distance between synthetic data and data assumed to be possessed by an intruder and the probability of a match given the intruder's data.

Distance-Based Record Linkage

Distance is a natural measure of identifiability. If a unique combination of attributes in the synthetic data file are too close to similar data possessed by an intruder, he or she may be able to infer that the records are likely to match. Spruill (1982, 1983, 1984) developed the concept of distance-based record linkage, and Tendick (1992) and Fuller(1993) expanded on it. In the context of synthetic data, distance metrics are calculated by attributes shared across the released synthetic data source and a dataset presumed to be in the hands of an intruder. Under certain circumstances, distance metrics may be used to probabilistically identify observations in partially synthetic data (Domingo-Ferrer et al. 2005,. 2006; Torra, Abowd, and Domingo-Ferrer 2006). Approaches to linking synthetic data have used Euclidean distance, standardized Euclidean distance, Mahalanobis distance, and in one case Kernel distance (Torra, Abowd, and Domingo-Ferrer 2006). As indicated by the nature of these metrics, distance-based record linkages have been limited to continuous, rather than categorical, data in the existing research.

McClure and Reiter evaluate the worst-case scenario in which intruders have access to every observation in the confidential dataset except for one they are trying to identify (McClure and Reiter 2012, 2016). This is an incredibly high (and generally unrealistic) standard of disclosure protection, and thus the authors also assess disclosure risk as intruder knowledge decreases. Notably, they found that observations near the middle of the distribution are at low risk, while outliers pose much greater risk of disclosure.

Another approach is to take a best guess as to what set of variables a potential intruder might know, then set out to identify observations based on those columns (Benedetto, Stinson, and Abowd 2013).

Disclosure is not only an issue for continuous variables. There might be attribute disclosure if an intruder knows that someone is likely to file an unusual combination of forms and schedules. If the intruder also knows the person is in the dataset and finds one record with that combination of attributes, then he or she could have high confidence that the record belonged to the target. However, if the synthetic data file contains only 10 percent or less of observations in a particular group, then identification by attributes is much less of a risk.

Probabilistic Record Linkage

Another natural metric is the probability of a match. Probabilistic record linking can be used to quantify the degree of disclosure risk. Probabilistic record linkage methods were established to match two datasets in which some individuals were represented as observations in both datasets (Fellegi and Sunter 1969;Jaro 1989). Like distance-based record linkages, some columns must exist in both datasets. For each overlapping column, weights are calculated to measure their contribution to accurately matching observations.

Probabilistic record linking, unlike linking with distance metrics, is perfectly capable of handling categorical data. However, probabilistic methods are generally more computationally demanding. Torra (2006) found that for a small number of potentially identifying variables, distance-based and probabilistic-based record linking performed similarly. However, as the number of variables increased, probabilistic record linkages began to outperform the distance-based approach.

A key question is what probability threshold is deemed to constitute a disclosure.

APPLICATIONS TO ADMINISTRATIVE DATA

Synthetic datasets are becoming a common means of allowing public access to otherwise confidential administrative and survey information. The practice can be traced back as far as 1990 with the Decennial Census Summary Files using a version of synthetic data under the terminology "blank and impute" (Federal Committee on Statistical Methodology 2005). More recent and complex applications include Survey of Income and Program Participation Synthetic Beta (SSB), Synthetic Longitudinal Business Database (SynLBD), Survey of Consumer Finance (SCF) and customized-synthetic UK Longitudinal Studies (UKLS).

In an application of synthetic data to Iowa tax returns, Huckett and Larsen (2007) used quantile regression to model the complex conditional and marginal distributions. The authors found this approach to be too computationally intensive to use on all the available variables, so quantile regression was used only for critical variables and hot-deck imputation was applied to impute the rest of the dataset.

In the creation of the Synthetic Longitudinal Business Database, SynLBD, Kinney et al (2011) used dirichlet-multinomial models for categorical variables and linear regression for continuous variables. Dirichlet-multinomial regression has been shown to be effective in modeling highly dispersed categorical data generally (Guimareas and Lindrooth 2005) and specifically within the application of synthetic data (Hu, Reiter, and Wang 2014; Hu, 2015).

To either preserve the actual file structure in case of SynLBD or improve data utility in case of SSB and SCF, a certain amount of actual data was retained in the synthetic datasets (Kinney, Reiter and Miranda 2014; Benedetto, Stinson and Abowd 2013; Kennickell 1997). In contrast, given that customized-synthetic UKLS data sets' main use is for researchers to test out their analytical programs before applying such programs to the confidential UKLS, all variables in the customized-synthetic UKLS were synthesized to ensure a high level of data confidentiality (Raab, Nowok and Dibben 2016).

To synthesize the data discussed above, a data generating process was estimated using a sequence of conditional regressions based on actual data, generally after missing values were imputed with an exception of customized-synthetic UKLS where the data generating process directly incorporates the possibility of certain data being missing. Modeling the joint distribution of all variables synthesized as a sequence of conditional regressions ensures tractability and reduces computational burden since variables can be synthesized one at a time instead of simultaneously. These regressions were parametric in some case (ordinary least squares and logistic regressions for SSB, selected parametric specifications for customized-synthetic UKLS), or nonparametric in other cases (CART for SynLBD, Bayesian bootstrap for SSB, a choice of CART or Random Forest for customized-synthetic UKLS).

Extra measures were employed to improve data confidentiality. For example, to avoid releasing identifiable outcomes, for SSB and SynLBD, regressions were performed only when the numbers of underlying observations were larger than a relevant threshold and when there was sufficient variation of the underlying observations, the CART tree was pruned in certain directions (e.g. for SynLBD, the establishment's first year can have different branches for different locations while establishment's last year cannot), and empirical distributions of outcomes were approximated. In addition, to derive synthetic values in SSB, regression coefficients drawn from their respective estimated distributions were used instead of the point estimates of these regression coefficients.

There is no universally preferred method for producing synthetic data. For example, although non-parametric methods are better at capturing complexity in the data, increased computational burden when synthesizing a large number of variables as in SSB may render the approach infeasible. As a result, using a combination of parametric and nonparametric methods may be more practical. In some cases, parametric regressions perform as well as nonparametric methods, making regression the better choice (Nowok 2015).

IV. SPECIAL ISSUES RELATED TO TAX DATA

Unique aspects of tax return data pose unique challenges to our design of data synthesis method and data confidentiality evaluation. We discuss these synthesis and disclosure challenges in this section.

SYNTHESIS CHALLENGES

To design a suitable data synthesis method in general and to precisely model the joint distribution of variables in the tax return data in particular, we must account for unique aspects of tax return data.

A first step is cleaning the data, especially the paper returns that are especially prone to math errors, omitted entries, and transcription errors. While tax return data contain relatively few missing values, measurement error exists.[34] SOI edits the INSOLE file to try to correct these errors in a process that reflects decades of experience. We plan to build a data-check routine to identify potential errors in the IMF and correct them to the extent possible. Working with SOI staff, we will explore the possibility of using machine learning methods to learn from the corrections made on the INSOLE file and apply those corrections to the entire IMF. We will validate these measures by comparing the corrected IMF with the edited INSOLE for the records that overlap.

Tax return data files contain a large number of variables, most of which are dollar values. There are several hundred variables in the SOI tax return data files. It is simply infeasible to model the joint distribution of these variables in complete detail. In fact, even if it were possible, information embedded in the joint distribution of the synthetic data could be considered of a high risk in terms of data confidentiality. To reduce the confidentiality risk, we propose to produce fully synthetic data where only a selected number of variables are rigorously modeled while the rest are populated but not strictly modeled.[35] In particular, we will identify the variables that need to be modeled explicitly while aggregating values of the remaining variables into a few aggregate variables, and model the joint distribution of these selected and aggregated variables. This will help preserve the relationships between variables crucial for tax policy research. Finally, we will decompose aggregate variables to fill in values of their components in a way that guarantees data confidentiality.[36]

Tax return variables are organized in a specific order because many variables are calculated based on information from preceding variables.[37] When modeling the joint distribution of independent variables as a sequence of conditional distributions, the order of variables in this sequence can be chosen arbitrarily. Prior work determined this order so as to minimize computational burdens.[38] The explicit relationships between variables in the tax returns data, especially when certain variables are calculated from other variables, provides an additional restriction on how variables should be sequenced.

In addition, it is straightforward to synthesize tax return variables calculated directly from their components since we simply need to apply the relevant formulas. For example, AGI is sum of various income items minus the various above-the-line deductions. If those components are synthesized, then AGI may be calculated based on the synthesized values. However, in some cases, we may want to target calculated variables for imputation because of their importance. In this case, we would impute shares for the components subject to the constraint that they add up to one.

We will sometimes have to impose additional constraints. For example, if a tax credit is only available when AGI is less than a given amount, the model should not allow any taxpayers with AGI larger than that amount to have this tax credit. The solution is to model the underlying process—e.g., for the child tax credit, calculating the number of qualifying children and the estimated credit amount based on phase-in and phase-out rules. We will also have to model he decision to claim the credit taxpayers do not claim all of the tax benefits to which they are entitled.[39]

Distributions of tax return variables vary by tax filing status. Also, the likelihood of certain income items, tax deductions and tax credits being reported may depend heavily on taxpayers' income.[40] And the risk of disclosure is higher for taxpayers with very high incomes and big losses than for those with modest incomes or losses. All of these reasons argue for modeling the joint distribution of tax return variables by tax filing status and income group. A byproduct is that this grouping will help speed up model runs, especially with parallel processing (as discussed below).

We may also want to select by other criteria such as age, number of children, and region, although this creates the risk that some groups would be too small.

A separate issue is the income measure used to stratify tax returns. AGI is a natural candidate, but AGI must be calculated from synthesized variables. So, the data synthesis method must ensure that synthesized values of the AGI components are consistent with the AGI group that the observations belong to.

Under the progressive income tax, a relatively small number of high-income taxpayers pays a large share of total US federal income tax.[41] This requires that the joint distribution modeled must capture extreme values. Otherwise, the distribution of synthesized federal income tax liability may differ greatly from the distribution observed in the actual tax return data.[42]

A crucial challenge is how to capture extreme values while retaining a sufficient level of data confidentiality. A solution may lie in the fact that, for most taxpayers, tax liability can be approximated reasonably well with information of just a small number of variables. Thus, we can potentially model tax returns with extreme values separately from other tax returns. Then, similar to the discussion above, we will further identify a selected number of variables that we can afford to model explicitly while aggregating values of the remaining variables together into a few aggregate variables, and model the joint distribution of the selected and aggregated variables. Finally, we will need to decompose aggregate variables to fill in values of their components.

One advantage of starting with such a large database (the IMF) is that it is feasible to produce multiple versions of synthetic data from non-overlapping samples for different uses. The basic synthetic file will focus on details of US taxpayers' tax returns overall. A second file could be designed to be representative at the state level. To ensure data confidentiality, we may need to suppress many details about income and deductions in that file.

In the long run, we would like to produce a longitudinal file. The methods used to produce the SynLBD file and SIPP earnings histories might be useful here. This file could probably contain even less taxpayer detail while still preserving confidentiality.

Finally, we will have to assess whether we can treat tax return structure as given in creating the synthetic data. That is, the set of forms and schedules that are included on a tax return would not be synthesized even though the amounts reported on each form would be. As discussed above, this could create the risk of attribute disclosure. We would need to develop a procedure to identify unique or extremely rare combinations of returns and schedules and decide how to handle those situations.

With all the considerations above, it may turn out that we will be able to model directly a selected set of variables with the set containing fewer variables for higher income groups in order to preserve data confidentiality. However, the standard synthetic data could still be an improvement over the current PUF in certain respects, because of all the steps that distort data on the PUF to prevent disclosure.

DISCLOSURE CHALLENGES

An intruder might attempt to learn about a person by using available information (potentially from public government wage information, public property tax valuations, or known number of dependents) to match onto an individual row of public administrative data. There is significant evidence that distance metrics can be used to re-identify observations in partially synthetic data (Domingo-Ferrer et al. 2005, 2006; Torra, Abowd, and Domingo-Ferrer 2006). Since the intruder will have information limited to what they could assemble publicly, it would be nearly impossible for them to identify a person with relatively common tax characteristics. Work by McClure and Reiter (2012, 2016) found that observations near the middle of the distribution are at low risk, while outliers pose much greater risk of.

However, some taxpayers with relatively common items might have them in unusual combinations. We will design an algorithm that identifies rare combinations and evaluate the extent to which excluding such observations or imposing a certain type of data aggregation (such as aggregating several variables into one aggregate variable) impairs data quality.

Since fully synthetic data do not include any actual tax return data, disclosure is in principle impossible. However, it is possible that certain synthetic observations match closely to actual data by chance. We must make sure that this situation is a rare occurrence and prescribe a remedy (such as replacing the observations with an additional synthetic record or, if necessary, revise the data generating process). For example, we can use distance metrics to evaluate how far each row of synthetic data is from any real row of data. Having a sufficient distance between real rows of data and synthetic rows makes it far less likely that an intruder could draw valid inferences about individual taxpayers from the data. In addition, the process of sampling itself provides some protection. The sampling rate would never be higher than 10 percent, making it extremely unlikely that any particular taxpayer's information is in the sample.

Since disclosure is only a threat to relatively uncommon rows of data, we can jointly evaluate distinctness of an individual row of real data and its distance to any row of synthetic data. However, the matrix of distances between every feasible pair of observations could be enormous making it very expensive and time consuming to compute every element. We will thus have to implement a preliminary step where observations with common features, and hence have no disclosure risk, are identified and excluded from this test. The distance-matching test will only be performed on the remaining uncommon observations.

In practice, variables should be standardized before running the distance metrics, which should at least include Euclidean distance and cosine distance, the latter being valuable because of the high dimensionality of the problem.

Probabilistic linkage matching can also be used to evaluate disclosure risk. Similarly to distance metrics, if a synthesized row of data can be matched back onto an administrative row of data using a potentially public subset of variables, then it is possible an intruder could use public data to draw inferences about an individual. With probabilistic matching, a linkage score is calculated instead of a distance metric. These linkage scores include subjective information to determine if variables are close enough to indicate a match. For categorical variables, this might require an exact match, whereas a range might be more appropriate for continuous variables.

Having access to the administrative data allows us to evaluate the potential danger of this approach. In order to check whether this is a possible source of disclosure risk, we can develop a set of reasonable rules and then test (using the real administrative data) to see if they would predict actual matches. Probabilistic matching allows the intruder to incorporate significant subjective knowledge with statistical learning and thus should be evaluated carefully.

COMPUTATIONAL CHALLENGES

Given the scale and sensitivity of the synthetic data task at hand, several decisions have computational implications that affect both difficulty of coding and computation time, including

- the number of rows and variables to be synthesized,
- the statistical methods used in the synthesis and disclosure risk processes, and
- the programming language chosen.

The faster that a synthetic data generation process can be coded, executed, and evaluated, the more times the research team can repeat the process and fine-tune the development of a high quality synthetic data product.

Sampling Methodology

The number of sampled rows from the gold standard file has a notable effect on computation time of the synthetic data process. As the sample size grows, the time taken for sequential regression methods will likely increase substantially. To reduce computational burden, the research team could use stratified sampling. For instance, in a synthetic file of state level estimates, a representative sample could be taken from each state independently (possibly grouping small states in the same region together to avoid disclosure risks). This would lead to a much more efficient synthesis process than drawing randomly from the entire IMF until each state had a sufficient sample.

The sample size also has interactions with the number of variables used for synthesis and the computational costs of the statistical methodologies. Each decision to change one of these key parameters may need to be weighed as a tradeoff against the others.

Number of Variables to Synthesize

Synthesizing more variables creates a longer process both in terms of programming and code execution. Each additional variable must be analyzed, and, for many statistical methods, a regression model must be carefully specified. It is worth noting that machine learning methods, though less understood in this context, may involve substantially less researcher time for model specification. Further, additional variables necessitate more computation time because each requires an additional regression step in each iteration of each implicate. This is particularly notable because the IRS SOI PUF has many more variables of interest to policy researchers than many prior synthesis tasks evaluated in the literature review.

Statistical Methodologies for Synthesis

There is a trade-off between computational expense and robustness of statistical methods. For instance, quantile regression is usually calculated using linear programming algorithms. This is substantially slower to calculate than ordinary least squares (simple matrix algebra) or maximum likelihood methods. Mixture models and most machine learning methods, depending on their implementation, often take longer than ordinary least squares for similar reasons.

Statistical Methodologies for Assessing Disclosure Risk

Calculating distance-based and probabilistic matching-based risk of disclosure involves intensive computation. The problem is that if there are, for example, 100,000 returns deemed unusual enough to be sensitive to disclosure, we might need to evaluate 10 billion record pairings (100,000 records on the original file and 100,000 synthesize records). Some initial screening might reduce the number of potentially problematic records by two-thirds, but that would still leave roughly 1 billion record pairings to evaluate.

However, the saving grace is that the disclosure metrics are easy to parallelize. For example, Euclidian distance is simply a sum of squares. For 1 billion record pairings, the problem could be partitioned into 200 chunks of 5 million records each, with each sum assigned to a separate thread on a large multicore processor or to a separate virtual machine on a cloud-based server. Even accounting for some overhead cost, this would cut down processing time by roughly two orders of magnitude. The process would still involve many calculations on each thread or virtual machine, but the processing time could be hours rather than days or weeks.

Programming Languages

The programming language used for this project will have a significant impact on the ease and speed of executing of many of the computational tasks. We have considered the potential advantages and disadvantages of using (either alone or in combination) R, Stata, Python, and Apache Spark. In short, R is the preferred choice if it is possible to work with this language in the IRS's environment, with Stata being the fallback. Further, Apache Spark could feasibly offer advantages for particularly demanding computational tasks, such as multiple imputations with complex machine learning models.

R has significant advantages. It has the widest breadth of statistical techniques that could prove valuable for sequential regression, including generalized linear methods, penalized regression methods, quantile regression mixture modeling, and robust libraries for Bayesian analysis and machine learning. In addition, several imputation frameworks have already been developed in R, including Multiple Imputation with Chained Equations and SynthPop. Although neither of these packages can completely solve the problems presented in this project, they can facilitate more effective and efficient programming. Further, a wide set of distance metrics and linking algorithms are available in R. R's strong parallelization libraries also offer an advantage over other languages for speeding up various computational tasks (especially running distance metrics, but potentially other aspects as well), if it may be run on a multicore machine.

Accessing R's libraries poses a potential problem for using the language. Because R is an open source language, most of the functionality exists in packages that must be downloaded individually from an online repository. This means that the environment used to create the synthetic data must be able to add R packages, either by direct download from the Internet or from a data storage device.

Stata may be a more feasible alternative to R. Stata is already available on IRS computers. Further, Stata has significant depth in the relevant statistical methods and some existing codebase for sequential regression using multiple imputation. One version of Stata, Stata MP,contains native parallel processing routines, which can significantly reduce the necessary computational time. Some programs may need to be adapted from R to run in Stata, but the program contains all the statistical and programming tools that would be needed to complete this project.

Python is another popular open source statistical language with many strong features. Although worthy of consideration, Python does not have the same expansive functionality in generalized linear models. The lack of existing frameworks for iterative imputation further undermine Python's viability in this scenario.

Apache Spark is not a programming language but rather a framework for distributed memory statistical analysis. Spark could be set up in place of the computing environment at the IRS, with R running on the platform. Setting up Spark requires that the data be accessible in a cloud environment, such as Amazon Web Services, that would fully take advantage of Spark's capabilities. If this was possible, Spark would dramatically increase the processing speed for some imputation methods, such as the generalized linear model, CART, and other machine learning methods. However, the version of R that runs on Spark does not have as extensive a library of statistical methods as standalone R. Thus, the trade-off for the speed gain would be the need to perform more bespoke programming or limit the range of synthesis algorithms tested.

V. TESTING THE METHODOLOGY USING THE PUBLIC USE FILE

Ultimately, we plan to work with SOI to produce a synthetic data file based on the IMF. To protect data security, that work will have to be done at the IRS. We are currently going through the clearance process to authorize that work. In the interim, however, we can refine and test our methodology using publicly available data. The obvious candidate for that work is the PUF.

The obvious advantage of using the PUF for testing is that it contains many of the same variables that are on the IMF (and INSOLE) and the relationships between those variables (and the associated forms and schedules) are the same. The disadvantage as discussed in Section II is that the PUF is only a fraction of the size of the IMF, many variables are suppressed, and some records are altered to prevent disclosure. The extreme outliers that pose the greatest challenge for disclosure prevention do not exist in the PUF.

Nonetheless, the PUF is a good basis for developing programs and debugging them. In consultation with our colleagues at the SOI, we will determine the software tools that we might have access to and determine which of those programs is most suitable for implementing the synthesis procedures.

Because many variables in the SOI data files are calculated from other variables, it is necessary to have a sufficiently accurate tax calculator to derive these calculated variables. Many calculations are straightforward (for example, AGI is the difference between the summation of taxable income components and the summation of above the line deductions), while some are rather complex (for example, taxpayers can choose between two alternative formulas to figure out their potential earned income credit). We plan to build this tax model by modifying the Tax Policy Center microsimulation model.

With a suitable test data file and tax model, we will be able to carry out series of tests on potential data-synthesis methods. We plan to carry out these tests as much as possible at Urban so we can utilize abundant computing resources and personnel. An ultimate product of these tests is a computation package—a set of programs and a troubleshooting guideline of a chosen data-synthesis method—that can be readily applied at the SOI facility.

VI. OUTLINE OF SYNTHESIS PROPOSAL

To summarize, here are the main steps we propose to follow in developing the synthetic tax data file.

1. Use the PUF to develop methodologies to apply to the confidential administrative tax data.
   a. This includes determining the synthesis method for each variable to be included based on the quality of synthetic data, speed of the synthesis process, and the importance of the variable. (Less important variables may be synthesized using quick and dirty methods to reduce cost.)
   b. Test the speed of different statistical programs and explore the possibility of parallel processing using hardware and software available at SOI.
   c. Determine how to handle calculated variables. Should they be calculated based on synthesized components, or should the calculated variables be synthesized first with an algorithm used to allocate the total to the components?
   d. Determine the partitions within which to synthesize the variables.
2. Adapt the programs created on the PUF so that they will run on the IRS databases using IRS computers. Adapt to include select variables that are not available on the PUF.
3. Create one or more synthetic data files based on the confidential administrative data (containing on the order of 200 million records) using procedures discussed above.
4. Assess risk of disclosure and, if necessary, exclude records or modify the synthesis procedure to reduce disclosure risks to acceptable levels.
5. Develop a procedure to extract the optimal sample of synthetic returns from the population of synthetic data records.
   a. Based on the population (IMF), calculate a set of statistical targets (for example, means, X'X, within income, filing status, age groups).
   b. Select the sample of size n_i within each partition *i* that minimizes the distance between sample statistics and population values.
   c. Evaluate data quality; compare to a random synthetic file of the same size and to the PUF.
   d. Test whether statistical inference in the refined sample matches inferences drawn from the population. We believe that this process would obviate the need to correct standard errors, as in multiple imputation.

VII. CREATING A SECURE MEANS OF RUNNING STATISTICAL PROGRAMS ON CONFIDENTIAL DATA

The final step is to create a secure way for authorized researchers to execute programs on the confidential administrative data as proposed by Reiter (2009). Because the synthetic data file would have the same structure as the restricted data, users could debug their programs with confidence on their own computers. Validated results must be checked for disclosure before being released to the researcher.

Computer scientists have developed a theory of data privacy that will be used to inform this project.[43] The basic idea is that each query of the confidential data extracts information that collectively could compromise privacy. For example, even with the addition of random errors to statistical estimates, it might be possible to rerun the same estimates multiple times to extract the true parameters. Researchers have developed a notion of a "privacy budget," that is drawn on each time the confidential data are accessed. Once exhausted, no further queries are permitted.

Because queries of the confidential dataset are a scarce resource, they entail a cost. One aspect of this project is to develop a pricing mechanism that reflects the shadow price of each statistical analysis of confidential data. (In a sense, this is analogous to imposing grazing fees to prevent a common resource from being depleted.) For estimates that only draw on a small sample from the underlying dataset, the shadow price would be very low. Estimates based on the entire population would typically have a higher price.

We still need to work out several technical issues for this project. For example, in theory, it would be possible to use the programs researchers submit to the IRS to improve the synthetic data file over time. Abowd and Schmutte (2015) describe such a methodology, but this could prove to be very computationally intensive.

VIII. CONCLUSION

Administrative tax data are a potential gold mine of information that could inform research in public finance and other areas. However, taxpayers who are obliged to file have a legal and moral right that their data to be kept secure. This working paper has outlined a procedure for creating high-quality synthetic data files that appropriately balance data quality with privacy concerns. Because of the immense size of the IMF, we believe it is possible to produce synthetic data files that are statistically equivalent to the population data for many purposes while including no identifiable taxpayer information.

Still, like all synthetic data files, there are statistical problems for which this file would not provide consistent or efficient estimates. For that reason, we also propose that a secure way be established for users to submit statistical programs to run on the IRS computers for a fee. Output would be emailed to the researcher after being checked for violations of disclosure rules.

REFERENCES

Abowd, John M., and Julia Lane. 2004. "New approaches to confidentiality protection: Synthetic data, remote access and research data centers." *Lecture Notes in Computer Sciences* (3050): 282-289.

Abowd, John M., and Ian M. Schmutte. 2015. "Economic analysis and statistical disclosure limitation." *Brookings Papers on Economic Activity* 2015 (1): 221-293.

Allison, Paul D. 2000. "Multiple Imputation for Missing Data: A Cautionary Tale." *Sociological Methods and Research* (2000): 301-309.

Benedetto, Gary Martha H. Stinson, and John M. Abowd. 2013. "The Creation and Use of the SIPP Synthetic Beta." Washington, DC: U.S. Census Bureau.

Bertrand, Marianne, Emir Kamenica, and Jessica Pan. 2015. "Gender Identity and Relative Income within Households," *Quarterly Journal of Economics* 130 (2): 571-614.

Breiman, Leo. 2001 "Random forests." *Machine learning* 45 (1): 5-32.

Bryant, Victoria L., John L. Czajka, Georgia Ivsin, and Jim Nunns. 2014. "Design Changes to the SOI Public Use File (PUF)." In *107th Annual Conference of the National Tax Association, Santa Fe, New Mexico*.

Bryant, Victoria L. 2016. *General Description Booklet of the 2011 Public Use File*. Washington, DC: Statistics of Income Division, Internal Revenue Service.

van Buuren, Stef, and Karin Groothuis-Oudshoorn. "MICE: Multivariate imputation by chained equations in R." *Journal of statistical software* (45) 3.

Caiola, Gregory and Jerome P. Reiter. 2010. "Random Forests for Generating Partially Synthetic, Categorical Data." *Transactions on Data Privacy* (2010): 27-42.

Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. 2014. "Where is the land of opportunity? The geography of intergenerational mobility in the United States." *The Quarterly Journal of Economics* 129 (4): 1553-1623.

Cilke, James. 2014. "The Case of Missing Strangers: What We Know and Don't Know About Nonfilers." In *107th Annual Conference of the National Tax Association, Santa Fe, New Mexico*. 2014.

Dreschler, Jörg, Stefan Bender, and Susanne Rässler. 2007. "Comparing Fully and Partially Synthetic Data Sets for Statistical Disclosure Control in the German IAB Establishment Panel." *Transactions on Data Privacy* (2007): 105-130.

Federal Committee on Statistical Methodology (FCSM). 2005. “Report on Statistical Disclosure Limitation Methodology.” Washington, DC: FCSM.

Fellegi, Ivan P. and Allan B. Sunter. 1969. “A Theory for Record Linkage.” *Journal of the American Statistical Association* (1969): 1183-1210.

Domingo-Ferrer, Josep, Vincenc Torra, Josep M. Mateo-Sanz, and Francesc Sebe. 2005. “Empirical Disclosure Risk Assessment of the IPSO Synthetic Data Generators.” *Monographs in Official Statistics-Work Session on Statistical Data Confidentiality* (2005): 227-238.

Domingo-Ferrer, Josep, Anton Martínez-Ballesté, Josep M. Mateo-Sanz, and Francese Sebé. 2006. “Efficient Multivariate Data-oriented Microaggregation.” *The Very Large Databases Journal* (2006): 355-369.

Fienberg, Stephen E. 1994. “A Radical Proposal for the Provision of Microdata Samples and the Preservation of Confidentiality,” *Carnegie Mellon University Department of Statistics Technical Report* (1994).

Fuller, Wayne A. 1993. “Masking Procedures for Microdata Disclosure Limitation.” *Journal of Official Statistics* (1993): 383-406.

Guimaraes, Paulo, and Richard Lindrooth. "Dirichlet-multinomial regression." *Economics Working Paper Archive at WUSTL*, *Econometrics* 0509001 (2005).

Hu, Jingchen, Jerome P. Reiter, and Quanli Wang. " Disclosure Risk Evaluation for Fully Synthetic Categorical Data." In *International Conference on Privacy in Statistical Databases* (8744): 185-199.

Hu, Jingchen. "Dirichlet Process Mixture Models for Nested Categorical Data." PhD diss., Duke University, 2015.

Huckett, J. C., and Michael D. Larsen. "Microdata simulation for confidentiality of tax returns using quantile regression and hot deck." In *2007 Proceedings of the Third International Conference on Establishment Surveys*. 2007.

Jaro, Matthew A. 1989. “Advances in Record Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida.” *Journal of the American Statistical Association* (1989): 414-420.

Karr, Alan F., Christine N. Kohnen, Anna Oganian, and Ashish P. Sanil. 2006. “A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality.” *The American Statistician* (2006): 224-232.

Kennickell, Arthur B. "Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances." *Record Linkage Techniques* (1997): 248-267.

Kinney, Satkartar K., Jerome P. Reiter, Arnold P. Reznek, Javier Miranda, Ron S. Jarmin, and John M. Abowd. 2011. "Towards unrestricted public use business microdata: The synthetic longitudinal business database." *International Statistical Review* 79 (3): 362-384.

Kinney, Satkartar K., Jerome P. Reiter, and Javier Miranda. 2014. "Synlbd 2.0: improving the synthetic longitudinal business database." *Statistical Journal of the IAOS* 30 (2): 129-135.

Little, Roderick. 1993. "Statistical Analysis of Masked Data." *Journal of Official Statistics* (1993): 407-426.

McClure, David and Jerome P. Reiter. 2012. "Differential Privacy and Statistical Disclosure Risk Measures: An Investigation with Binary Synthetic Data." *Transactions on Data Privacy* (2012): 535-552.

_____. 2016. "Assessing Disclosure Risks for Synthetic Data with Arbitrary Intruder Knowledge," *Statistical Journal of IAOS*32 (1):109-126.

Nowok, Beata. 2015. "Utility of synthetic microdata generated using tree-based methods." Edinburgh: University of Edinburgh, School of Geosciences.

Parisi, Michael. 2017. "Individual Income Tax Returns, Preliminary Data, Tax Year 2015." *Statistics of Income Bulletin* (Spring 2017): 2-11.

Raab, Gillian M., Beata Nowok, and Chris Dibben. 2017. "Practical data synthesis for large samples." *Journal of Privacy and Confidentiality* 7 (3): 4.

Raghunathan, Trivellore E., James M. Lepkowski, and John Van Hoewyk, and Peter Solenberger. 2001. "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models." *Survey Methodology* (2001): 85-96.

Raghunathan, Trivellore E., Jerome T. Reiter, Donald B. Rubin. 2003. "Multiple imputation for statistical disclosure limitation." *Journal of Official Statistics* (2003): 1-16.

Reiter, Jerome P. 2004. "Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation." *Survey Methodology* (2004): 235-242.

_____. 2005. "Using CART to Generate Partially Synthetic Public Use Microdata," *Journal of Official Statistics* (2005): 441-462.

_____. 2009. "Multiple Imputation for Disclosure Limitation: Future Research Challenges." *Journal of Privacy and Confidentiality* 1(2): 223-233.

Richman, Michael B., Theodore B. Trafalis, and Indra Adrianto. 2009. "Multiple Imputation Through Machine Learning Algorithms." *Artificial Intelligence Methods in Environmental Sciences* (2009): 153-169.

Rubin, Donald B. 1978. “Multiple Imputation in Sample Surveys - a Phenomenological Bayesian Approach to Nonresponse.” *American Statistical Association Proceedings of the Section on Survey Research Methods* (1978): 20-40.

_____. 1987. *Multiple imputation for nonresponse in surveys*. Vol. 81. John Wiley & Sons, 2004.

_____. 1993. “Statistical Disclosure Limitation,” *Journal of Official Statistics* (1993): 462-468.

_____. 1996. "Multiple Imputation After 18+ Years." *Journal of the American Statistical Association* 434 (91): 473-489.

_____. 2004. “The Design of a General and Flexible System for Handling Nonresponse in Sample Surveys,” *The American Statistician* (2004): 298-302.

Snoke, Joshua, Gillian Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic. 2016. "General and specific utility measures for synthetic data." *arXiv preprint arXiv:1604.06651* (2016).

Spruill, Nancy L. 1982. “Measure of Confidentiality.” *Statistics of Income and Related Administrative Research:* 1982. Washington, DC: Department of the Treasury, Internal Revenue Service, Statistics of Income Division.

Spruill, Nancy L. 1983. “The Confidentiality and Analytical Usefulness of Masked Business Microdata,” *Proceedings of the Section on Survey Research Methods, American Statistical Association* (1983): 602-607

____. 1983. “Testing Confidentiality of Masked Business Microdata.” Alexandria, VA: Public Research Institute. 1983.

____. 1984. “Protecting confidentiality of business microdata by masking.” Public Research Institute, Alexandria, VA.

Statistics of Income Division. 2016a. *Individual Income Tax Returns 2014*. Washington, DC: Statistics of Income Division, Internal Revenue Service.

Statistics of Income Division. 2016b. *Program Documentation: Data Items by Forms and Schedules 2015*. Washington, DC: Statistics of Income Division, Internal Revenue Service.

Tendick, Patrick. 1992. “Assessing the effectiveness of the noise addition method of preserving confidentiality in the multivariate normal case.” *Journal of Statistical Planning and Inference* (1992): 273-282.

Torra, Vicenc, John M. Abowd, and Josep Domingo-Ferrer. 2006. “Using Mahalanobis Distance-Based Record Linkage for Disclosure Risk Assessment.” *Privacy in Statistical Databases* (2006): 233-242.

Vilhuber, Lars, and John M. Abowd. "Session 14: Usage and outcomes of the Synthetic Data Server." In *Understanding Social and Economic Data*. 2016.

Wei, Lan. "Methods for Imputing Missing Values and Synthesizing Confidential Values for Continuous and Magnitude Data." PhD diss., Duke University, 2016.

Wei, Lan and Jerome P. Reiter. 2016. “Releasing Synthetic Magnitude Microdata Constrained to Fixed Marginal Totals.” *Statistical Journal of IAOS* (2016): 93-108.

Winglee, Marianne, Richard Valiant, Jay Clark, Yunhee Lim, Michael Weber and Michael Strudler, 2002. “Assessing Disclosure Protection for a SOI Public Use File.” Paper presented at the American Statistical Association Meetings.

NOTES

[1] Parisi (2017); this count does not include tentative or amended returns. Based on more detailed data available for returns filed in 2015 (Statistics of Income Division 2016b), about 54 million returns filed in 2016 were joint returns (108 million taxpayers), and about 9 million were filed by children and other dependents.

[2] The shorter variants of Form 1040, Form 1040A and 1040EZ, have fewer lines and require fewer supporting schedules and forms.

[3] Notices are also sent to electronic filers if certain errors are detected in tests applied after the return is filed. Paper and electronically filed returns may also be selected, in various ways, for further testing, audit, and enforcement actions.

[4] The population file of year 2016 contains most but not all tax returns filed for tax year 2016 but also a good number of tax returns filed for prior years (some are for a tax year before 2013).

[5] Some information returns also pertain to dependents who do not file an income tax return.

[6] For example, the IRS has the returns and schedules filed by partnerships and Subchapter S corporations, which provide information in addition to what is reported by individual owners on their income tax returns.

[7] The information returns and DM-1 files also represent most of the non-filing population; see Cilke (2014).

[8] Historically there was a significant lag between "pipeline" processing of returns and recording in the Master File, but that lag has now essentially disappeared. Some return information (such as attached W-2s) is not captured in IRS processing (because IRS subsequently matches returns to W-2s supplied to it by employers), but is included in the SOI sample so is captured by SOI to speed completion of the sample file.

[9] This sample is referred to as the INSOLE, a combination of Individual and sole proprietor (a name that originated when sole proprietorships were the primary form of noncorporate business).

[10] The following description is taken, with minor modifications, from Bryant, et al. (2014).

[11] Tentative (preliminary, incomplete) returns are excluded from the sample because the final return will be subject to sampling; amended returns are omitted because the original return was subject to sampling.

[12] The chained gross domestic product implicit price deflator generally grows more slowly than consumer prices indexes, such as the CPI. The change in the deflator between 1991 and 2014 (the deflation factor for income on returns sampled in 2015) was 1.5403 (compared to a CPI-U ratio for the same period of 1.7097).

[13] Social Security numbers (SSNs) are issued by the Social Security Administration to every US citizen, every noncitizen who has permission to work in the United States, and certain other noncitizens who require an SSN to receive benefits from the federal government or a state or local government. Noncitizens who do not qualify for an SSN but who are required to file a tax or information return with the IRS are required to have an Individual Taxpayer Identification Number, which is a nine-digit number with a first digit of nine that is otherwise similar to an SSN. IRS also issues an Adoption Taxpayer Identification Number for a child legally placed for adoption with a taxpayer who does not know the child's SSN.

[14] Due to a processing error, 1,355 returns that should have been included in stratum 101 (HINTS) were not included in the sample (see Statistics of Income Division 2016a, 216). HINTS still accounted for 30,883 returns in the sample.

[15] The items included in the sample for tax year 2015 are detailed in Statistics of Income Division (2016b).

[16] The full sample rates for strata 7 and 8 and 17 and 18 are about .33 percent (design and achieved sample rates differ somewhat).

[17] Bryant, et al. (2014) describe how these returns are selected. A few returns in non-certainty strata are included in the aggregate record. For the (tax year) 2011 PUF, 1,155 returns, representing 1,300 returns in the population, were aggregated (Bryant 2016).

[18] Bryant (2016).

[19] The full SOI sample for tax year 2011 represented a population of 145.6 million returns, but 0.4 million of these were filed for taxable years prior to 2008 and therefore not in the population the PUF represents.

[20] Bryant (2016).

[21] See in particular Internal Revenue Code Section 6103 and the associated provisions imposing penalties for breach of confidentiality.

[22] For returns sampled in 2015, 97.6 percent of returns were in strata 7 through 18 (Table B, page 216, in Statistics of Income 2016a).

[23] Changes made during IRS and SOI processing to return entries made by taxpayers may also reduce disclosure risk, because these changes in themselves make some PUF records different from the taxpayer's return.

NOTES

[24] Taxpayers have the option of taking an itemized deduction for state income or state sales taxes. Because most states impose an income tax that is generally higher in amount for itemizers than their sales tax amount, use of the sales tax deduction (for which IRS supplies suggested amounts by income and family size) potentially could be used to infer state of residence.

[25] The PUF includes variables based on the DM-1 information on the full sample file for age of the primary taxpayer (in separate ranges for non-dependent and dependent primary filers), ages of dependents (in ranges), and the gender of the primary taxpayer. It also includes a variable computed from W-2 and Schedule SE information on the full sample file for ranges of the split of earnings between spouses on joint returns.

[26] Bryant (2017).

[27] A single filer who supports a parent in a separate household can file as a head of household.

[28] Blurring is also referred to as masking and microaggregation.

[29] Groups are defined by marital status and number of dependents. Within these groups. variables to be blurred are normalized, a distance metric among all returns is calculated, and then the returns are selected in subgroups of three, starting with the returns that are furthest apart, and the variables are multivariate blurred.

[30] For a summary description of the various blurring groups, see Bryant (2016).

[31] One measure of the statistical accuracy of the PUF is the difference between population estimates for variables based on the full sample and on the PUF. As shown in Bryant (2016) for the 2011 PUF, the differences can be quite large (e.g., for the itemized deduction for motor vehicle taxes the difference is over 60 percent of the amount estimated from the full sample), although for common items is quite small (e.g., the difference is only .13 percent for AGI).

[32] The basic problem with simply imputing missing values is that the imputed value has measurement error but is treated as a known constant, which creates bias. Multiple imputation reflects the uncertainty in imputation by providing a range of equally plausible estimates. However, standard errors must be adjusted upwards for proper statistical inference (see Rubin 1996, equation 2.2 for example).

[33] A kernel density function is fitted to the values of leaves on the terminal branch (Reiter 2005). The support of the density function is constrained to be between the minimum and maximum values unless those are so close that there would be a disclosure concern. In that case, the support may be extended (i.e., allowing for values outside those observed), but that process may produce unrealistic imputations. A better option may be to prune the tree (reduce M) so that each branch has a suitably diverse distribution of values.

[34] Taxpayers can file for an amendment in case that they need to make a correction. However, these corrections may not be recorded in the SOI data files, which are a snapshot of tax return data.

[35] This list of variables that will be modeled rigorously should include most of the variables currently present in PUF with at least a few more variables added.

[36] One possibility is to randomly select one of tax returns in this subgroup, calculate the ratio between each component's value and its associated aggregate value, apply any necessary blurring (e.g. not allowing any share to be more than 90 percent or less than 1 percent), then apply these shares to decompose the aggregate values into components. The goal of this method is to reasonably populate values of component variables, but not to preserve the relationships between components and aggregated variables to ensure data confidentiality.

[37] To complete an individual income tax return, a taxpayer must record her income sources, potential tax deductions and various expenses in detail in the main 1040 Form and its associated schedules, forms and supplemental calculation worksheets. These forms, schedules and worksheets embed in them tax-year-specific formulas to calculate taxable income, itemized deductions, alternative minimum tax, claimable tax credits, and other applicable taxes from the detailed information recorded. This taxpayer fills these calculated components in the main 1040 Form to figure out her tax liability and, after accounting for taxes that she has paid in advance, the available tax refund if she has paid more than her liability in advance or the additional tax that she needs to pay if she has not paid enough in advance.

[38] In their experiment, Raab, Nowok and Dibben (2016) noted that putting categorical variables with a large number of categories early in the sequence helped speed up the calculation of the synthesis model. Reiter (2005) proposes a variable order based on the number of values need to be synthesized.

[39] Some people do not claim credits or deductions to which they are entitled. In some cases, there is low participation because the benefits are quite small and eligibility rules complicated (e.g., the EITC for childless adults). Taxpayers may rationally forgo certain credits or deductions if their value is small relative to the time cost of completing the associated tax forms, schedules, or worksheets required to claim the tax benefit. In other cases, taxpayers may not know about certain tax benefits or that they are eligible for them.

NOTES

[40] According to a SOI tabulation, among the tax returns filed in 2014, the percentage of tax returns claiming itemized deduction was 19.8 percent among taxpayers with AGI at most $100,000 and 81.1 percent among taxpayers with AGI more than $100,000. Some tax credits (such as earned income credit and saver's credit) are only available to taxpayers whose AGI do not exceed a specified amount, while some taxes (0.9% Medicare surtax and 3.8 percent net investment tax) are only imposed on taxpayers with AGI above a certain threshold.

[41] According to a SOI tabulation, there were 148.6 million tax returns filed in 2014 with the total income tax after credit of $1.36 trillion. Out of these, 16,733 returns (0.01 percent of 148.6 million returns) had AGI of $10 million or more and their total income tax after credit was $123.6 billion (9.1 percent of $1.36 trillion).

[42] To see this, consider the following example. Suppose both A and B are single with no dependent, A had $5,000 and B had $95,000 of taxable income in 2016. In this case, A's and B's tax before credit would be $530 and $19,644, respectively, resulting in a total of $20,174.

Suppose that the synthesized data somehow swapped certain income values of A and B resulting in both A and B having $50,000 of taxable income in 2016. In this case, both A's and B's tax before credit would be $8,278, resulting in a total of $16,556, underestimating the actual tax liability by 18 percent!

[43] This discussion draws heavily on Abowd and Schmutte (2015).

The Tax Policy Center is a joint venture of the
Urban Institute and Brookings Institution.

For more information, visit taxpolicycenter.org
or email info@taxpolicycenter.org