TPC TAX POLICY CENTER
URBAN INSTITUTE & BROOKINGS INSTITUTION

SAFELY EXPANDING RESEARCH ACCESS TO ADMINISTRATIVE TAX DATA: CREATING A SYNTHETIC PUBLIC USE FILE AND A VALIDATION SERVER

Leonard E. Burman, Alex Engler, Surachai Khitatrakun, James R. Nunns, Sarah Armstrong, John Iselin, Graham MacDonald, and Philip Stallworth

ACKNOWLEDGMENTS

We are grateful to John Abowd, Brian Balkovic, John Czajka, Dan Feenberg, Barry Johnson, Laura Kawano, Arthur Kennickell, Rob McClelland, David McClure, Alan Plumley, Rob Santos, Mike Strudler, Lars Vilhuber, Mike Weber, and members of the Statistics of Income Division Consultants Panel for helpful comments on an earlier draft; to James F. Raleigh for editing, 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.

A version of this paper was presented at the 8th Annual IRS/TPC Joint Research Conference on Tax Administration on Wednesday, June 20, 2018. This paper will be posted by the IRS in 2019.

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 their 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 Internal Revenue Service (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. 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 their output satisfies disclosure avoidance protocols. This paper discusses the current methodology used to produce public use datasets, surveys the literature on synthetic data and privacy protection, outlines our proposed plan to produce a synthetic file, and discusses special challenges.

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 et al. (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 Statistics of Income (SOI) Division of 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 location, the split of wages between spouses on joint returns, the ages of filers and family members, and many other variables that are available to the IRS, are either excluded altogether from the PUF, or are only available on certain returns in cursory form. In addition, the highest-income tax returns are aggregated, and information from other returns is deliberately "blurred" to protect against the risk of disclosing information that would allow users to identify the returns of specific taxpayers. Those kinds of restrictions have been in place for many years, but two developments have required the IRS to increase the restrictions, making the PUF less useful: (1) PUF users also have access to an increasing amount of microdata from various sources, which if compared (or combined) with the PUF data would potentially allow them to establish the identity of taxpayers in the PUF; and (2) modern computer processing capacities and software tools make it possible to match very large datasets and make inferences that were impossible to make just a few years ago. As these risks increase, the viability and usefulness of future PUFs is at risk. Thus, finding a replacement for the PUF is imperative.

Moreover, providing researchers greater access to administrative tax data could vastly expand our understanding of how tax policies affect behavior (and how those policies could be made more effective). The gold mine of administrative tax data is only available, however, inside select government agencies and via collaboration with analysts in those agencies or through highly restrictive arrangements with the IRS. Expanding access to administrative data would represent a major advance in the ability of the Tax Policy Center 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.

Synthetic data replace sensitive variables or observations with imputed values. Actual data are replaced in the public dataset by estimated values that are either generated by a statistical model such as regression or by nonparametric methods. Many public datasets are now partially synthetic, with some sensitive variables replaced by imputed values. Variables are deemed sensitive if an intruder could match them with unique data available in another database. By matching the externally available data to the administrative database, a particular record could be identified, disclosing all the other information contained in that observation.

Partially synthetic data always create a risk of disclosure, which grows as more data become available online. To address this, we propose creating a fully synthetic public use file. The goal is to approximate the statistical process that generates actual tax return data and to draw records randomly from this process.

In principle, a perfect synthetic dataset could exist. To see why, suppose that a random sample of tax filers disappeared, leaving only their tax returns behind. In this thought experiment, imagine that their disappearance means that they never actually existed. That sample of tax returns would represent the perfect synthetic tax dataset. It is as representative of the population of filers as any other random sample. And since, by definition, the filers in the dataset don’t exist, publishing their tax returns could not constitute a privacy violation.

The only question is whether this ideal sample, or something close enough to it, could be found. We propose a method to approximate the distribution of individual income tax returns and to test the utility of this synthetic dataset.

A number of synthesis techniques have been used in previous applications, including parametric and nonparametric models. 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 discrete variables.

We propose to use sequential regression-based methods to synthesize continuous variables similar to the sequential regression multiple imputation (SRMI) model that has been used to impute missing values. The method is computationally simple and relatively flexible.

We also propose to set up a secure method for researchers to submit statistical programs to run on a subset of the confidential administrative data. This model for research access to confidential data is referred to as a *validation server*. The structure of the administrative dataset would be the same as for the synthetic data, so programs that are developed using the PUF would work on the confidential data with minimal alteration. The IRS would review programs and output to protect against disclosure. A fee structure would be designed to defray costs and discourage data mining.

This paper is organized as follows. Section II discusses individual income tax data compiled by the SOI Division of 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. Section VI outlines our plan for producing a validation server at the IRS. Section VII presents concluding remarks.

We welcome 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 now used to create the PUF.

Federal individual income tax returns are filed annually with the IRS by most adults and some children. In 2017, 150.3 million individual income tax returns, covering 289.8 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 2016. Income tax returns may contain a large number of data entries. The basic income tax return, Form 1040, contained over 80 possible entries in 2016.[2] Tax returns may include one or more schedules; in 2016, there were 11 schedules to Form 1040, such as Schedule A for reporting itemized deductions, each with multiple possible entries. Other forms (more than 50 in 2016) provide 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.

Eighty-eight percent (131.6 million) of the individual returns filed in 2017 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 mathematical errors. The remaining 18.7 million returns in 2017 were filed on paper. The IRS records information from paper-filed returns electronically, then tests this information and sends notices to taxpayers if errors are detected.[3] The fraction of returns filed on paper does not vary much 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 Form 1040, so they 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, contains information on taxpayers' and dependents' dates of birth, genders, name changes, and, for deceased individuals, dates 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 (Cilke 2014).

SOI SAMPLES

The annual sample of individual income tax returns drawn by the SOI is based only on the population of individual income tax returns processed by the IRS during the year, and is drawn at an initial stage of processing, prior to the IMF being available.[7] The SOI uses the full sample, called the INSOLE file,[8] to prepare publications and other products, and both the Office of Tax Analysis in the US Department of the Treasury and the staff of the Congressional Joint Committee on Taxation use the INSOLE 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 the Office of Tax Analysis and the Joint Committee on Taxation create nonfiler records from information returns and DM-1 information.

INSOLE

The INSOLE sample is selected from all individual income tax returns processed during a year by the IRS and posted to the IMF.[9] 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 19 strata: 9 negative income strata (strata 1 through 9) or 10 positive income strata (strata 10 through 19).[10] Beginning with the tax year 2016 sample, strata boundaries are amounts expressed in 2016 dollars. In future years, the dollar amounts of "total positive income" and "total negative income" on each return will be deflated by the change in the chained gross domestic product implicit price deflator between 2016 and the tax year of the sample.[11] Sample rates vary by strata, from a rate of about 0.1 percent (1 in 1,000) in strata 10, 11 and 12, to 100 percent in strata 1, 2, 18 and 19. 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).[12] In addition, sample rates in certain strata are periodically increased to ensure 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 Social Security numbers (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 1 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 for research purposes. For returns in strata 10, 11 and 12, only CWHS returns are selected for the sample. For returns in the other strata (1 through 9 and 13 through 19), which have sample rates above the CWHS rate, non-CWHS SSNs are transformed (to correct for slight nonrandomness in the last four digits of the SSN), and enough endings of the transform are selected to achieve the target sample rate for the stratum (including the CWHS portion of the sample).

The full INSOLE SOI sample selected in 2017 (primarily of returns filed for 2016) included records for 351,049 returns (of 150.3 million filed), with 65,993 of those returns (nearly 19 percent of the total) selected with certainty (from strata 1, 2, 18, 19, 101, and 201).[14] Excluding HINTS (strata 101), 37,235 returns (12.6 percent of the total sample excluding HINTS) were selected with certainty. The minimum total positive (total negative) income required to be selected in strata 18 or 19 (1 or 2) was $8 million (–$8 million) in 2016.

SOI includes in the full sample file almost all of the information reported by taxpayers on their income tax return Form 1040s (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 INSOLE 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 (stratum 101) are placed in the strata (1 through 19 or 201) to which they would otherwise be assigned and then subsampled for the PUF in the same manner as other returns in each stratum. All CWHS returns selected in INSOLE 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 13 (total incomes between –$400,000 and $400,000 in 2016) only the subsampled CWHS returns are included in the PUF sample.[16] All returns in strata 5 and 6 (total negative income between –$1,500,000 and –$400,000) and 14 and 15 (total positive income between $400,000 and $1,500,000), which for the INSOLE sample are sampled at rates of about 1 percent (strata 6 and 14) and 3 percent (strata 5 and 15), are included in 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 –$8 million and –$1,500,000) and 16 and

17 (total positive income between $1,500,000 and $8 million), which for the INSOLE sample are sampled at rates of about 12 percent (strata 4 and 16) and 33 percent (strata 3 and 17), are subsampled for the PUF to achieve an effective sampling rate of 10 percent. Some of the returns in the certainty strata (1 and 2, 18 and 19, 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 in the PUF sample for tax year 2012 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 2012 included 172,411 return records (including the four aggregate records) representing the 144.6 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 2012 PUF, for example, includes entries from fewer than half the Form 1040 schedules and from only 8 of more than 50 attached forms that were included in the INSOLE file for 2012.[20] In addition, as discussed next, 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.[21] The definition of a disclosure is 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 were publicly available. All of the information on an individual income tax return and its associated schedules, forms, and attachments are "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, forms, and attachments may, as noted, 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 are 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 a taxpayer's occupation. Other taxpayer characteristics (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 identify a specific taxpayer's record indirectly.

Some taxpayer characteristics, 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, may 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 rare, such as very large families with an elderly head. Little additional information about such taxpayers might be needed to identify them with a 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 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 their other characteristics.

Amounts of deductions and credits also might 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.

Demographic and other information on individuals is also increasingly 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, social media, and email activities.

Although specific characteristics of individuals may not uniquely distinguish them, 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 rising threat of identity theft and data breaches that target and steal individuals' 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 put all the items in their record at risk of disclosure.

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 low rates 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, additional 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

The state of residence 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 through 6, 14 through 19, 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 through 13.[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 through 6, 14 through 19, and 201 are converted to single returns.

As the earlier discussion implies, 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 taxpayers filing separate are much less likely to have dependents than head of household and joint filers. Those who do have dependents tend to have few 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 the 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 returns selected in strata 1 through 6, 14 through 19, or 201, wages and salaries, state and local income taxes, and real estate taxes are blurred. Blurring for these records is "multivariate," meaning that all three variables, if present, are simultaneously blurred within a group.[29] For returns selected in strata 7 through 13, wages and salaries, state and local income taxes, and real estate taxes are also blurred, but so are alimony paid and received (which are 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 are 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 internally inconsistent. 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 generally will not give the same amount of AGI reported by the taxpayer. If reported AGI was retained in the return record, it might be possible to infer the true values for some components, negating the intended reduction in disclosure risk from the blurring procedure. For this reason, return records are "rebalanced" by recomputing 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 recomputing 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, and 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 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 to construct the PUF, 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 confidential while ensuring that inferences drawn from such data are, *ex ante*, statistically equivalent to inferences drawn from the actual data 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 high-quality data for a defined set of research questions while maintaining the required level of data confidentiality.

Most 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 are 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 while preserving 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, the generator need 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 a sequential process discussed next.

Different models may be appropriate for different variables. For example, different models may be more appropriate for discrete or categorical variables than for continuous variables. Even within the same data type, different underlying variable distributions may warrant different models.

The simplest procedure is to add random noise to the independent variables in a dataset. If the induced errors are relatively small, much of the relationship among variables may be maintained, but preserving this information increases the risk of disclosure (Fuller 1993). Errors large enough to prevent disclosure create a 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 need 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.

More recent research has developed machine learning methods such as 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 using these methods.

Emerging research in this field applies a mixture of Poisson distributions 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).

When developing parameterized models, conditioning and grouping are often necessary. *Conditioning* is the selection of independent variables for the estimation of each variable to be synthesized (Benedetto, Stinson, and Abowd 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, filing status, number of dependents, and age. Grouping allows for more flexible estimation and significantly speeds computation by breaking up the very large full dataset into manageable subsets.

Groups must be of sufficient size to prevent overfitting of imputed values. Benedetto et al. (2013) recommend that the number of observations within subgroups (created by grouping) be at least 15 times the number of conditioning variables, or 1,000—whichever is greater.

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

Multiple Imputation

In the most general terms, *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 make the synthetic data suitable, they must be changed substantially to protect the anonymity of the original dataset 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 expanded 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 using multiple imputation to generate synthetic data, although their approaches differed. Rubin (1993) suggested merging a smaller dataset onto a larger dataset, such as a census file. If *X* is observed in both datasets and *Y* only in 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 nonsensitive 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 our example, consider the case where *Y*, a sensitive variable (subject to disclosure risk), and *X* both exist in a dataset. Use the process just described to create multiple implicates of *Y** in that dataset, and then replace the actual *Y* with the synthesized values *Y**. Assuming that the model fit 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 *classification and regression trees* 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 distributions and nonlinear relationships among explanatory variables that might be hard to identify and model explicitly (Reiter 2005). However, CART may be computationally 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 to get a single output for each observation. For classification trees this is generally 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 create more accurate out-of-sample predictions.

SRMI

Raghunathan et al. (2001) proposed a variant on multiple imputation that he called *sequential regression multivariate imputation* as a way to consistently impute values for missing data. 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. Raghunathan et al. sorted the variables by number of missing values and performed the imputations in order, starting with the variable with fewest missing values.

A joint multivariate probability distribution can be represented as the product of a sequence of marginal distributions:

$$f(Y_1, Y_2, \dots, Y_k \mid X, \theta_1, \theta_2, \dots, \theta_k) \; = \\ f_1(Y_1 \mid X, \theta_1) \; \cdot f_2(Y_2 \mid X, Y_1, \theta_2) \cdots \; f_k(Y_k \mid X, Y_1, Y_2, \dots, Y_{k-1}, \theta_k)$$

where Y_i are the endogenous variables, X is a matrix of independent variables, and θ_i are vectors of model parameters such as regression coefficients.

In the first iteration, the nonmissing values of Y_1 are modeled as a function of X; Y_2 is modeled as a function of Y_1 (with missing values imputed) and X; Y_3 is estimated as a function of Y_1, Y_2, and X; and so on.

In the second and subsequent iterations, each Y_i is modeled as a function of the other imputed $\{Y_k\}, k \neq i$, and iterations continue 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 to deal 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.

Sequential Regression

Our problem differs from Raghunathan et al. (2001) in two ways. First, missing data are not an important problem in the tax dataset.[34] Second, since we are proposing to create a fully synthetic dataset, there is no matrix of exogenous variables, *X*.

We thus propose several adaptations. We will initially order the variables to synthesize from most common to least common. The first variable, Y_1, will be synthesized by making *n* random draws from its empirical distribution. The second variable, Y_2, will be modeled as a function of Y_1. Model parameters estimated using the observed data will be used to synthesize Y_2 as a function of the synthetic Y_1 with errors drawn from the empirical error distribution. Y_3 will be estimated as a function of Y_1, Y_2, and so on until Y_k is estimated as a function of Y_1, Y_2, … , Y_{k-1}. Since *k* will be large, exclusion restrictions will need to be applied (i.e., leaving many of the Y_j variables out of the model) to prevent problems of singularity or overfitting.

We plan to use simple models to estimate the relationships such as linear regressions on a polynomial expansion of the right-hand-side variables. For censored variables (such as interest income, which is always nonnegative) we will use limited dependent variable models such as Tobit or the Heckman two-step estimator.[35]

EVALUATING SYNTHETIC DATA

Ideally, synthetic data should be high quality while protecting data confidentiality. These goals 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 next discuss how data quality and data confidentiality are measured.

Evaluating Data Quality

Quality of synthetic data can be measured by their usefulness for general purposes and for drawing inferences. In the privacy 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 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 an original dataset and its synthetic counterpart (Karr et al. 2006). An examination of the Statistics of Income PUF evaluated changes in the first four moments (mean, variance, skewness, and kurtosis) of variables of interest (Winglee 2002). That 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 synthesis methods will create datasets that are suitable for different analyses. Abowd and Lane (2004) showed that a single confidential file could be used to generate several different synthetic files for different statistical purposes.

Certain 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 dataset. Vilhuber and Abowd (2016) show that the same graph based on the synthetic version of this dataset exhibits no discontinuity. This smoothing occurs because the model that generates the synthetic data was not designed to account for such discontinuities. There are many similar kinks, notches, and other discontinuities in tax data that will not be reflected in the synthetic records unless the synthesis process is explicitly designed to capture them.

Evaluating Disclosure Risk of Synthetic Data

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

An intruder might attempt to use available information (potentially from public government wage information, public property tax valuations, or known number of dependents) to uniquely identify an individual row of public administrative data (Reiter 2005). Distance metrics can be used to reidentify observations in partially synthetic data (Domingo-Ferrer et al. 2005, 2006; Torra, Abowd, and Domingo-Ferrer 2006).

However, some taxpayers with relatively common items might have them in unusual combinations. Excluding such observations or aggregating several variables into one aggregate variable can mitigate this risk, but at the cost of data utility.

The following sections discuss metrics for disclosure risk in partially synthetic data and why fully synthetic data as we propose to construct would preserve privacy.

Distance-Based Record Linkage

Distance is a natural measure of identifiability. If a unique combination of attributes in the synthetic data file is too close to similar data possessed by an intruder, they may be able to infer a match in the records. Spruill (1982, 1983a, 1983b, 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 et al. 2006). Approaches to linking synthetic data have used Euclidean distance, standardized Euclidean distance, Mahalanobis distance, and in one case kernel distance (Torra et al. 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 (2012, 2016) 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. This is an extreme 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 disclosure risk, while outliers pose much greater risk.

Another approach is to take a best guess as to what set of variables a potential intruder might know and attempt to identify observations based on that information alone (Benedetto et al. 2013).

Disclosure is not only an issue for continuous variables. There might be attribute disclosure in a synthetic data file 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 the intruder could have high confidence that the record belonged to the target. However, if the file contains only 10 percent or less of observations in a particular group, then identification by attributes is much less of a risk because it is unlikely that any particular return is represented in the sample.

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 et al. (2006) found that for a small number of potentially identifying variables, distance-based and probabilistic record linking performed similarly. However, as the number of variables increased, probabilistic record linkages began to outperform the distance-based approach.

Disclosure Risk in Fully Synthetic Data

Since fully synthetic data do not include any actual tax return data, disclosure is in principle impossible. Even if some observations are close to actual information, this would occur purely by chance and would tell an intruder nothing about whether such observations exist in the administrative data. Our approach approximates the multivariate distribution that generates the actual population of tax returns and makes random draws from a smoothed version of that distribution. The only disclosure is about the characteristics of that distribution.

There could be a risk of disclosure if the empirical distribution has kink points that correspond to the underlying population, especially if the sampling rate is very high. Burman et al. (2018) show that smoothing the distribution between the kink points and sampling from that distribution effectively blurs the sample distribution in a way that makes reliable inference about kinks (and underlying data) impossible. Special issues arise for outliers, since sampling from the empirical distribution could provide information about extreme values. Burman et al. propose a sampling methodology for outliers that is consistent with the mean and variance of the underlying data in the tail, but does not reveal information about the particular values.

Our sequential regression synthesis methodology also protects against the most serious form of attribute disclosure (Burman et al. 2018). That is, while sampling from the estimated multivariate distribution may produce rare records with unusual combinations of attributes, they are extremely unlikely to be close to actual records. If by chance, synthetic records are close to actual records with rare attributes, the existence of the synthetic records provides virtually no information about whether such a combination actually exists.

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 to 1990, when the Decennial Census Summary Files used a version of synthetic data termed “blank and impute” (FCSM 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, 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 too computationally intensive to use on all the available variables, so they used quantile regression only for critical variables and hot-deck imputation for the rest.

In the creation of the 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, 2015; Hu, Reiter, and Wang 2014).

To either preserve the actual file structure, as in the SynLBD, or improve data utility, as in the SSB, a certain amount of actual data was retained in the synthetic datasets (Kinney, Reiter, and Miranda 2014; Benedetto et al. 2013; Kennickell 1997). In contrast, all variables in the customized-synthetic UKLS were synthesized to ensure a high level of data confidentiality (Raab, Nowok, and Dibben 2016).

To synthesize data, a data-generating process is estimated using a sequence of conditional regressions based on actual data. The customized-synthetic UKLS data-generating process directly incorporated the possibility of missing data. 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 sometimes parametric (ordinary least squares and logistic regressions for SSB, selected parametric specifications for customized-synthetic UKLS) and sometimes nonparametric (CART for SynLBD, Bayesian bootstrap for SSB, a choice of CART or random forest for customized-synthetic UKLS).

Extra measures were designed to improve data confidentiality. For example, to avoid releasing identifiable outcomes in 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 the establishment's last year cannot), and empirical distributions of outcomes were approximated. In addition, to derive synthetic values in SSB, regression coefficients were drawn at random from their respective estimated distributions.

There is no universally preferred method for producing synthetic data. For example, although nonparametric 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 special challenges to our design of data synthesis method. We discuss these synthesis and computational challenges in this section.

SYNTHESIS CHALLENGES

To design a suitable data synthesis method by precisely modeling the joint distribution of tax return data, we must account for the unique aspects of such data.

The first step is cleaning the data, especially the paper returns that are prone to mathematical errors, omitted entries, and transcription errors. While tax return data contain relatively few missing values, they do contain measurement error.[36] 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 many variables, most of which are dollar values. It may be infeasible to model the joint distribution of these variables in complete detail. In addition, information embedded in the joint distribution of the synthetic data might pose a disclosure risk. To reduce the risk, only some variables may be explicitly modeled, while the rest are populated but not strictly modeled.[37] For example, the variables that are not explicitly modeled could be aggregated into a few composite variables. This would help preserve the relationships between variables crucial for tax analysis. The aggregate variables could be decomposed using a simple formula to fill in values of components while masking any individual-level variation.[38]

Tax return variables are organized in a specific order because many variables are calculated based on information from preceding variables.[39] When modeling the joint distribution of variables as a sequence of conditional distributions, the order of variables in this sequence can be chosen arbitrarily. Prior work used the order that minimized computational burdens.[40] The explicit relationships between variables in the tax return 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 the sum of various income items minus 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 1. Alternatively, we might synthesize both the components of AGI and AGI itself and then scale the synthesized components so that they add up to synthesized AGI.

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—for the child tax credit, as an example, by 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 the decision to claim the credit since taxpayers may not claim all of the tax benefits to which they are entitled.[41]

Distributions of tax return variables vary with filing status. The likelihood of certain income items, tax deductions, and tax credits being reported may also depend heavily on income.[42] And the risk of disclosure is higher for taxpayers with very high incomes or big losses than for those with modest incomes and losses. All of these reasons argue for modeling the joint distribution of tax return variables by tax filing status and income group. This grouping will help speed up model runs, especially with parallel processing (as discussed in the Programming Languages section).

We may also group by other criteria such as age, number of children, and region, although we will have to ensure that groups do not become 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. The data synthesis method therefore must ensure that synthesized values of the AGI components are consistent with the AGI group to which the observations belong. We will explore the possibility of using CART to select the income ranges subject to minimum size constraints.

Under the progressive income tax, a relatively small number of high-income taxpayers pay a large share of total US federal income tax.[43] 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.[44]

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 from 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 preceding discussion, we can further identify a selected number of variables that we can afford to model explicitly while aggregating values of the remaining variables into a few aggregate variables, then 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 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 Survey of Income and Program Participation earnings histories might be useful here. This file probably could contain even less taxpayer detail while still preserving confidentiality.

Finally, we will have to assess whether we can treat tax return structure as a given in creating the synthetic data. That is, the set of forms and schedules included on a tax return would not be synthesized even though the amounts reported on each form would be. As discussed, 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 these considerations, 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.

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 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 interacts 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 PUF has many more variables of interest to policy researchers than several prior synthesis tasks discussed 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.

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, SQL, and R and Python running on Apache Spark. We plan to use SQL for in database analytics, R (on a single large server), and R running on Apache Spark to work within the IRS's environment, with Stata as our fallback. We propose the following analytics pipeline in Table 1, then discuss its merits.

Table 1

Proposed Analytics Pipeline

<table>
<tr><th>1. In Database Analytics</th><th>1 [Optional]. Large single server processing</th></tr>
<tr><td>To the extent possible with the functionality provided, Urban staff will write SQL queries to merge, aggregate, subset, and sort the necessary data within the system in which the data are stored and indexed. Many database storage systems optimize and parallelize common data steps "under the hood." Reducing the size and complexity of the data before transferring over a network can improve processing time.</td><td>Sort, merge, transform, and other data-shuffling operations are key bottlenecks in Apache Spark. In addition, certain complex regression methods are unavailable. If certain operations prove impossible or too costly within the database or Spark, additional sort, merge, transform, modeling, and data-shuffling operations will be done in parallel on a large single-core machine using R (after as much data processing is done within the database as possible). The resulting data will be written out to the database of Apache Hadoop Distributed File System to be read by Spark for further processing.</td></tr>
<tr><th colspan="2">2. Apache Spark—Pre-processing and modeling</th></tr>
<tr><td colspan="2">After the data are prepared within the database and any additional sort, merge, transform, modeling, or shuffling operations are completed in the large single server, Apache Spark will be used for preparing any additional data and for running the remaining imputation modeling. Spark can optimize and parallelize over many more cores than a single server and is capable of running regression analyses as well as some of the more accurate machine learning models, such as random forests and gradient boosted trees.</td></tr>
</table>

Combining in-database SQL analytics with R has significant advantages. In-database analytics can significantly improve processing time; optimize some sort, merge, transform, and other shuffling operations that are difficult to parallelize; and reduce the amount of data transferred over the network to analysis servers.

R in general 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), when 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 or have the necessary packages available, either by direct internet download or from a data storage device.

Using SparkR or Sparklyr, the two options for running R on Apache Spark, dramatically improves the processing speed for some imputation methods, such as the generalized linear model, CART, random forests, and gradient boosted trees. However, the version of R that runs on Spark does not have as extensive a library of statistical methods as stand-alone 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. In addition, some of the most accessible R functions for Spark—in Sparklyr—are relatively newer than the original SparkR package, which might require the analysis team to adapt its programming to a style slightly foreign to base R users. Ultimately, we will program the more bespoke synthesis procedures in R on a single large machine and attempt to improve performance with parallel operations and other performance improvements, such as efficient kernel libraries. As much as possible, however, we prefer to implement the synthesis procedure in Spark using Sparklyr and SparkR.

Stata is an alternative option. Stata is widely used by researchers in and outside of government. 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 on powerful multicore machines. 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. The disadvantage of Stata is that it cannot scale to large datasets as well as Spark and cannot parallelize to more cores than the license allows.

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 undermines Python's viability in this scenario.

V. SUMMARY OF SYNTHESIS PROPOSAL

The synthesis methodology will be refined over time through testing.[45] Our current plan would synthesize discrete variables using CART and use a sequential regression method to synthesize continuous variables. Discrete variables would be top-coded to prevent release of information about extreme outliers.

As discussed, the sample would first be partitioned into relatively homogeneous groups subject to the constraint that none of the partitions be too small to protect against overfitting (Benedetto et al. 2013). For example, the sample might be partitioned by age, filing status, dependent return status, type of return filed, number of dependents, and income. We will test the feasibility of using CART to identify these groupings. We plan to produce a single synthetic dataset, though we may test the use of multiple implicates as part of the process.

For continuous variables, the methodology will order the variables from most to least common, though we will test the order of variables to maximize model accuracy. The first variable (Y_1) would be synthesized by randomly sampling from a smoothed version of the empirical distribution. To protect against disclosing the attributes of extreme values, the distribution for the tails would be modified to preserve the mean and variance while suppressing all other information. This makes meaningful inference about particular observations impossible and is consistent with the speculation of Machanavajjhala et al. (2008, 285): "We believe that judicious suppression and separate modeling of outliers may be the key since we would not have to add noise to parts of the domain where outliers are expected."

Subsequent variables (Y_2, Y_3, …, Y_k) will be synthesized by regressing each variable on a polynomial expansion of a subset of the prior variables. The synthesized variables are calculated by applying the regression parameter estimates to the prior synthesized values with the addition of an error term drawn from the smoothed empirical error distribution (with similar treatment of outliers as in the generation of Y_1). The models will be modified as appropriate to deal with censoring. For example, dividends only appear on a subset of income tax returns and must be positive.

Variables that are a function of others, such as AGI or taxable income, will be calculated as a function of the synthesized variables.

We will test the methodology first by creating a synthetic nonfiler database, using information drawn from information returns about wages and salaries, Social Security income, and the like. This subsample by design has relatively low incomes and raises few disclosure risks. Assuming this synthesis is successful, we will next apply the methodology to the entire population of income tax returns.

The synthetic master file, which would not be shared with the public, would start with a very large number of observations, likely on the order of 10 million to start, though this number would be subject to testing. The synthetic PUF would be a stratified random sample of the synthetic master file with no records sampled at more than a 10 percent rate. Weights will be calculated as the inverse of the sampling rate. Sampling and weighting will ensure that the synthetic PUF is of a manageable size for researchers. Sampling is also an important protection against disclosing too much information about the underlying distribution (Burman et al. 2018).

The synthetic dataset will be evaluated for utility by comparing sample statistics with population values and those estimated using the INSOLE and the current PUF. In addition, we will run the Tax Policy Center Microsimulation Model on all three datasets, or some variant of the tax model depending on infrastructure constraints. We will estimate revenue, marginal tax rates, and tax burden distributions using the synthetic PUF and compare against the alternative benchmarks. We will use the following indicators to compare accuracy: confidence interval overlap, confidence interval length ratios, and difference in means (for the tax model).

The process of generating the synthetic data is designed to be disclosure proof, as discussed. However, we will apply various tests to guarantee that we haven't inadvertently created disclosure risks. Our tests will be especially sensitive to the risk of attribute disclosure. For example, if only one tax unit has a particular combination of tax forms and schedules, the inclusion of such a return in the synthetic database could constitute evidence that the unique unit had filed a return. We will identify such cases if they exist and address them. For example, we might reduce the number of forms and schedules in the synthetic database until attribute disclosure is impossible.

VI. 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 researchers.

Computer scientists have developed a theory of data privacy that will be used to inform this project.[46] 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 the common resource of a pasture 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 it is not clear this is computationally feasible at present.

VII. CONCLUSION

Administrative tax data are a potential trove of information that could inform research in public finance and other areas. However, taxpayers who are obliged to file have a legal right and expectation that their data 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 indistinguishable from the population data for many purposes while including no identifiable taxpayer information.

Still, like all synthetic data files, there are statistical problems for which our dataset 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 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." In *Privacy in Statistical Databases: PSD 2004, Lecture Notes in Computer Science*, vol. 3050, edited by Josep Domingo-Ferrer and Vicenç Torra, 282–89. Berlin: Springer.

Abowd, John M., and Ian M. Schmutte. 2015. "Economic Analysis and Statistical Disclosure Limitation." *Brookings Papers on Economic Activity* 46 (1): 221–93.

Allison, Paul D. 2000. "Multiple Imputation for Missing Data: A Cautionary Tale." *Sociological Methods and Research* 28 (3): 301–9.

Benedetto, Gary, Martha H. Stinson, and John M. Abowd. 2013. "The Creation and Use of the SIPP Synthetic Beta." Washington, DC: US 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. 2017. *General Description Booklet for the 2012 Public Use File*. Washington, DC: US Department of the Treasury, Internal Revenue Service, Statistics of Income Division.

Bryant, Victoria L., John L. Czajka, Georgia Ivsin, and Jim Nunns. 2014. "Design Changes to the SOI Public Use File (PUF)." Paper presented at the 107th Annual Conference of the National Tax Association, Santa Fe, New Mexico, November 13–15.

Burman, Leonard, Surachai Khitatrakun, Graham MacDonald, and Philip Stallworth. 2018. "Proposed Methodology for Creating a Fully Synthetic Dataset and Privacy Implications." Washington, DC: Urban Institute. Unpublished.

Caiola, Gregory, and Jerome P. Reiter. 2010. "Random Forests for Generating Partially Synthetic, Categorical Data." *Transactions on Data Privacy* 3 (1): 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.” Paper presented at the 107th Annual Conference of the National Tax Association, Santa Fe, New Mexico, November 13–15.

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): 355–69.

Domingo-Ferrer, Josep, Vicenç Torra, Josep M. Mateo-Sanz, and Francesc Sebe. 2005. “Empirical Disclosure Risk Assessment of the IPSO Synthetic Data Generators.” In *Monographs in Official Statistics: Proceedings of the UN/ECE-Eurostat Work Session on Statistical Data Confidentiality*, 227–38.

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* 1 (3): 105–30.

FCSM (Federal Committee on Statistical Methodology). 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* 64 (328): 1183–1210.

Fuller, Wayne A. 1993. “Masking Procedures for Microdata Disclosure Limitation.” *Journal of Official Statistics* 9 (2): 383–406.

Guimaraes, Paulo, and Richard Lindrooth. 2005. “Dirichlet-Multinomial Regression.” *Econometrics* 0509001. Economics Working Paper Archive at Washington University in St. Louis.

Hu, Jingchen. 2015. “Dirichlet Process Mixture Models for Nested Categorical Data.” PhD diss., Duke University. https://dukespace.lib.duke.edu/dspace/bitstream/handle/10161/9933/Hu_duke_0066D_12907.pdf;sequence=1

Hu, Jingchen, Jerome P. Reiter, and Quanli Wang. 2014. “Disclosure Risk Evaluation for Fully Synthetic Categorical Data.” In *Privacy in Statistical Databases*: *PSD 2014, Lecture Notes in Computer Science*, vol. 8744, edited by Josep Domingo-Ferrer, 185–99. Berlin: Springer.

Huckett, Jennifer, and Michael D. Larsen. 2007. “Microdata Simulation for Confidentiality of Tax Returns using Quantile Regression and Hot Deck.” In *2007 Proceedings of the Third International Conference on Establishment Surveys*, Montreal, Quebec, June 18–21.

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* 84 (406): 414–20.

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* 60 (3): 224–32.

Kennickell, Arthur B. 1997. "Multiple Imputation and Disclosure Protection: The Case of the 1995 Survey of Consumer Finances." In *Record Linkage Techniques, 1997*, edited by Wendy Alvey and Bettye Jamerson (1997): 248–67. Washington, DC: National Academies Press.

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): 129–35.

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): 362–84.

Little, Roderick. 1993. "Statistical Analysis of Masked Data." *Journal of Official Statistics* 9, 407–26.

Machanavajjhala, Ashwin, Daniel Kifer, John Abowd, Johannes Gehrke, and Lars Vilhuber. 2008. "Privacy: Theory Meets Practice on the Map." In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, 277–86. New York: IEEE Computer Society.

Maddala, G. S. 1983. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge, UK: Cambridge University Press.

McClure, David, and Jerome P. Reiter. 2012. "Differential Privacy and Statistical Disclosure Risk Measures: An Investigation with Binary Synthetic Data." *Transactions on Data Privacy* 5 (3): 535–52.

———. 2016. "Assessing Disclosure Risks for Synthetic Data with Arbitrary Intruder Knowledge." *Statistical Journal of the IAOS* 32 (1): 109–26.

Nowok, Beata. 2015. "Utility of Synthetic Microdata Generated using Tree-Based Methods." Edinburgh, UK: University of Edinburgh, School of Geosciences.

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

Raab, Gillian M., Beata Nowok, and Chris Dibben. 2017. “Practical Data Synthesis for Large Samples.” *Journal of Privacy and Confidentiality* 7 (3): 67–97.

Raghunathan, Trivellore E., James M. Lepkowski, John Van Hoewyk, and Peter Solenberger. 2001. “A Multivariate Technique for Multiply Imputing Missing Values using a Sequence of Regression Models.” *Survey Methodology* 27 (1): 85–95.

Raghunathan, Trivellore E., Jerome T. Reiter, and Donald B. Rubin. 2003. “Multiple Imputation for Statistical Disclosure Limitation.” *Journal of Official Statistics* 19 (1): 1–16.

Reiter, Jerome P. 2004. “Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation.” *Survey Methodology* 30 (2): 235–42.

———. 2005. “Using CART to Generate Partially Synthetic Public Use Microdata.” *Journal of Official Statistics* 21 (3): 441–62.

———. 2009. “Multiple Imputation for Disclosure Limitation: Future Research Challenges.” *Journal of Privacy and Confidentiality* 1 (2): 223–33.

Richman, Michael B., Theodore B. Trafalis, and Indra Adrianto. 2009. “Missing Imputation through Machine Learning Algorithms.” In *Artificial Intelligence Methods in the Environmental Sciences*, edited by Sue Ellen Haupt, Antonello Pasini, and Caren Marzban, 153–69. Berlin: Springer.

Rubin, Donald B. 1978. “Multiple Imputation in Sample Surveys—A Phenomenological Bayesian Approach to Nonresponse.” In *American Statistical Association Proceedings of the Section on Survey Research Methods*, 20–40.

———. 1987. *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and Mathematical Statistics, vol. 81. Hoboken, NJ: John Wiley & Sons.

———. 1993. “Statistical Disclosure Limitation.” *Journal of Official Statistics* 9 (2): 461–68.

———. 1996. “Multiple Imputation after 18+ Years.” *Journal of the American Statistical Association* 434 (91): 473–89.

———. 2004. “The Design of a General and Flexible System for Handling Nonresponse in Sample Surveys.” *The American Statistician* 58 (4): 298–302.

Snoke, Joshua, Gillian Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic. 2016. “General and Specific Utility Measures for Synthetic Data.” arXiv:1604.06651.

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

Spruill, Nancy L. 1983a. "The Confidentiality and Analytical Usefulness of Masked Business Microdata." In *American Statistical Association Proceedings of the Section on Survey Research Methods* (1983): 602–7.

———. 1983b. "Testing Confidentiality of Masked Business Microdata." Working Paper PRI 83-07.09. Alexandria, VA: Public Research Institute.

———. 1984. *Protecting Confidentiality of Business Microdata by Masking*. Alexandria, VA: Public Research Institute.

Statistics of Income Division. 2018a. *Individual Income Tax Returns 2016*. Washington, DC: US Department of the Treasury, Internal Revenue Service, Statistics of Income Division.

Statistics of Income Division. 2018b. *Individual Income Tax Returns: Line Item Counts 2016*. Washington, DC: US Department of the Treasury, Internal Revenue Service, Statistics of Income Division.

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* 31 (3): 273–82.

Torra, Vicenç, John M. Abowd, and Josep Domingo-Ferrer. 2006. "Using Mahalanobis Distance-Based Record Linkage for Disclosure Risk Assessment." In *Privacy in Statistical Databases: PSD 2006*, *Lecture Notes in Computer Science*, vol. 4302, 233–42. Berlin: Springer.

van Buuren, Stef, and Karin Groothuis-Oudshoorn. 2011. "mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software* (45) 3: 1–67.

Vilhuber, Lars, and John M. Abowd. 2016. "Session 14: Usage and Outcomes of the Synthetic Data Server." In *INFO7470 Understanding Social and Economic Data*, https://ecommons.cornell.edu/handle/1813/45065.

Wei, Lan. 2016. "Methods for Imputing Missing Values and Synthesizing Confidential Values for Continuous and Magnitude Data." PhD diss., Duke University, 2016. https://dukespace.lib.duke.edu/dspace/bitstream/handle/10161/12897/Wei_duke_0066D_13672.pdf?sequence=1.

Wei, Lan, and Jerome P. Reiter. 2016. “Releasing Synthetic Magnitude Microdata Constrained to Fixed Marginal Totals.” *Statistical Journal of the IAOS* 32 (1): 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 Meeting, Fairbanks, AK, July 10–12.

NOTES

[1] See Statistics of Income Division (2018a); this count does not include tentative or amended returns. About 54 million returns filed in 2017 were joint returns (108 million taxpayers), and about 9 million were filed for children and other dependents.

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

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

[4] The population file of returns processed by IRS in 2017 contains most but not all tax returns filed for tax year 2016 and also a good number of tax returns filed for prior years (some are for a tax year before 2014).

[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 individual partners and shareholders report on their income tax returns.

[7] 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 the IRS subsequently matches returns to W-2s supplied to it by employers), but is included in the SOI sample so captured by SOI to speed completion of the sample file.

[8] *INSOLE* is a combination of *individual* and *sole proprietor*. The name originated when sole proprietorships were the primary form of noncorporate business.

[9] 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 returns were subject to sampling.

[10] When the original strata boundaries were established in 1991, some of the positive income strata covered the same income ranges but covered returns with different levels of "interest" for tax analysis purposes. Within an income range, a higher sample rate was assigned to the stratum containing more interesting returns. As a result, there were 15 positive income strata. Over time the differential sampling rates were dropped, so SOI now uses only income ranges to define these strata.

[11] Prior to the tax year 2016 sample, strata boundaries were expressed in 1991 dollar levels, with amounts on returns adjusted by the change in the gross domestic product implicit price deflator since 1991. This deflator generally grows more slowly than consumer prices indexes, such as the consumer price index for all urban consumers CPI-U. The change in the deflator between 1991 and 2015 (the deflation factor for income on returns sampled in 2016) was 1.5874 (compared to a CPI-U ratio for the same period of 1.7402).

[12] The identification of high-income nontaxable returns, as defined in the Tax Reform Act of 1976, is based on the larger of Adjusted Gross Income or an expanded income definition measured in current dollars (that is, not deflated). When first established, this stratum captured 69 returns. It now includes about 30,000.

[13] Social Security numbers 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] HINTS accounted for 28,758 returns in the sample.

[15] The items included in the sample for tax year 2016 are detailed in Statistics of Income Division (2018b).

[16] The full sample rates for negative strata 8 and 9 and positive strata 13 are about 0.33 percent, while the rate for negative strata 7 is about twice that rate (design and achieved sample rates differ somewhat).

[17] Bryant et al. (2014) describe how these returns are selected. A few returns in noncertainty strata are included in the aggregate record. For the tax year 2012 PUF, on a weighted basis 1,260 returns were aggregated (Bryant 2017).

[18] Bryant (2017).

[19] The INSOLE sample for tax year 2012 represented a population of 145.0 million returns, but 0.4 million of these were filed for taxable years prior to 2009 and therefore not in the population the PUF represents.

[20] Bryant (2017).

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

[22] For returns sampled in 2017, 97.8 percent of returns were in strata 7 through 13 (Table C, p. 208, in Statistics of Income Division 2018a).

[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.

[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 INSOLE sample file for age of the primary taxpayer (in separate ranges for nondependent 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. 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 (2017).

[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 (2017) for the 2012 PUF, the differences can be large (e.g., for the itemized deduction for motor vehicle taxes, the difference is over 80 percent of the amount estimated from the full sample), although for common items it is small (e.g., the difference is only 0.14 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 upward for proper statistical inference (e.g., see Rubin 1996, equation 2.2).

[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] We may want to use some relatively less important variables that were included in electronically filed returns but were not captured from paper returns (unless they were selected for the INSOLE file). We plan to follow the synthesis procedure described hereafter and assume that the small fraction of returns for which those data are not available are otherwise similar to other returns in these respects.

[35] See, for example, Maddala (1983) for a survey of limited dependent variable models. More detail on our synthesis methodology and approach to privacy is in Burman, Khitatrakun, MacDonald, and Stallworth (2018).

[36] Taxpayers can file for an amendment if 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.

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

[38] One possibility is to randomly select one of the 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 to ensure data confidentiality by not preserving the relationships between components and aggregated variables.

[39] To complete an individual income tax return, a taxpayer must record her income sources, potential tax deductions, and various expenses in detail on the main 1040 Form and its associated schedules, forms, and supplemental calculation worksheets. These forms, schedules, and worksheets have embedded in them tax-year-specific formulas to calculate taxable income, itemized deductions, alternative minimum tax, available tax credits, and other applicable taxes from the detailed information recorded. The taxpayer enters these calculated components on the main Form 1040 to figure out her tax liability and, after accounting for taxes that she has paid in advance, obtains either 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.

[40] 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 that need to be synthesized.

[41] Some people do not claim credits or deductions to which they are entitled. In some cases, there is low participation because the benefits are small and eligibility rules complicated (e.g., the earned income tax credit 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 benefits. In other cases, taxpayers may not know about certain tax benefits or that they are eligible for them.

[42] According to an SOI tabulation, among the tax returns filed in 2016, the percentage of returns claiming itemized deduction was 19.9 percent among taxpayers with AGI through $100,000 and 74.6 percent among taxpayers with AGI more than $100,000. Some tax credits (such as the earned income credit and saver's credit) are only available to taxpayers whose AGI does not exceed a specified amount, while some taxes (0.9 percent Medicare surtax and 3.8 percent net investment tax) are only imposed on taxpayers with AGI above a certain threshold.

[43] According to an SOI tabulation, 150.3 million tax returns were filed in 2017, with a total income tax after credits of $1.43 trillion. Out of these returns, 16,087 (0.01 percent of all returns) had AGI of $10 million or more, and their total income tax after credit was $114.4 billion (8.0 percent of $1.43 trillion).

[44] To see this, consider the following example. Suppose both A and B are single with no dependents. 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!

[45] The methodology and an assessment of privacy issues are discussed in more detail in Burman et al. (2018).

[46] 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