Creating Homogeneous Synthetic Individual Tax Files for Distributional Analysis

Emmanuel Saez, UC Berkeley and NBER

Gabriel Zucman, UC Berkeley and NBER

August 2018

Abstract:

This paper uses individual income tax data produced by the Statistics of Income Division of the IRS since 1962 to construct homogeneous synthetic annual files that can be used for distributional and tax analysis and can be disclosed online. First, we define a number of income, tax, and family status variables homogeneously across all years since 1962. Second, we apply the blurring method used in the most recent Public Use Files (PUF) retrospectively to the earlier years. Third, we create aggregate records grouping together returns with extreme values. Fourth, all other records are aggregated by groups of 5 returns chosen to be similar along some key family status and income variables. Fifth, we age the most recent PUF (2012) using tabulations of internal tax data by size of income for years 2013-2016 in order to create synthetic data for years 2013-2016. For all years 1962-2016, the records created are synthetic aggregations of individual tax returns and create no disclosure risk. As a result, they can be distributed online as tax statistics. The datasets can be used to estimate a number of distributional tax statistics. In particular, they can be used to generate Distributional National Accounts (DINA) files that are consistent with National Accounts as in Piketty, Saez, and Zucman (2018).

Emmanuel Saez, Professor of Economics, University of California, Berkeley, California. saez@econ.berkeley.edu. Gabriel Zucman, Assistant Professor of Economics, University of California, Berkeley, California, zucman@berkeley.edu. This paper was developed in the context of the external research contract TIRNO-15-P-00060 with the Statistics of Income (SOI) Division at the US Internal Revenue Service under the supervision of Mike Weber at SOI. I am grateful to Victoria Bryant, Barry Johnson, Mike Weber at SOI, and especially John Czajka for helpful comments and guidance in developing this project. All the statistics created in this project as well as any possible error are the sole responsibility of the author. We thank Sam Karlin and Carl McPherson for outstanding research assistance.

Introduction

The Statistics of Income (SOI) division of the IRS has a long tradition of producing a wide range of income tax statistics. Most of the statistics are published and posted online for wide use among researchers, the press, and the broader public.[1] SOI also currently produces Public Use Files (PUF) of individual tax return data. These data start from the internal individual income tax data files created by SOI (that we will refer to as "the SOI internal files" from now on) and apply a large number of subsampling and blurring methods to minimize disclosure risk.[2] These methods are described in complete detail in Bryant, Czajka, Ivsin, and Nunns (2014) as well as in the official PUF documentation (US Treasury, 2017). The PUF datasets are not homogeneous across years as variables change following changes in the tax code.

The goal of this project is to create a synthetized version of the PUF files that (1) does not create any disclosure risk and hence can be publicly posted online, (2) provides a set of variables consistent across years going back to 1962. Such files could then be easily used to compute a number of distributional and tax statistics. Therefore, these newly created files could greatly broaden the use of tax data for research and policy analysis and also familiarize more researchers with the use of tax micro-data.

The files we produce take the form of annual micro-files with a fixed set of homogeneous variables. They cover the years 1962, 1964, 1966-2012.[3] We also created files for 2013-2016 by aging the 2012 PUF file to meet tabulated statistics from internal data (these tabulated statistics are also posted online to allow replication for users). The variables include basic demographics (marital status, number of dependents, age 65+ status), the main income components (such as wages and salaries, pensions, dividends, interest, etc.), the main deductions (such as charitable giving, mortgage interest, state and local taxes), and basic tax variables (such as federal income tax, self-employment taxes, refundable tax credits).

Methodology

Our methodology proceeds in 5 steps.

1) Homogeneous variables across all years:

[1] The SOI tax statistics (both the most recent series as well as the universe of earlier series produced dating back over a century) are posted online at https://www.irs.gov/statistics.

[2] Complete methodological details are presented in the documentation accompanying each file (see US Treasury, 2017).

[3] These years are those for which micro-level data with a rich set of variables have been created and maintained by SOI.

The first step selects and constructs a subset of key variables that are homogeneous across years across all original files created and maintained by SOI. For broadest applicability, we consider a basic set of demographic, income, deduction, and tax variables that are made homogeneous across years to the fullest extent possible. The files we create include the following variables (listed here with their names in the STATA files in brackets and bold):

Demographic variables:
[id] Unique record identifier in each annual file (id<=0 identifies aggregate records, see below)
[year] This is the year of the file, constant in each annual file. Incomes are earned during the year and the tax return is filed in year+1.[4]
[dweght] Population weight. Each record in the file represents dweght records in the population.
[married] Married joint filer dummy (married filing separately have the dummy set to zero). For married=1, incomes in the file are always the sum of incomes across the two members of the married couple.
[xded] Total number of dependents (capped at 3 and does not include the spouse for married filers). Incomes of dependents (if any) are generally not included in the income variables.
[xkids] Number of children at home dependents (capped at 3). Children dependents are minors or up to age 24 for fulltime students.
[oldexm] Age 65+ dummy for primary filer. The variable is imputed since 1996 using the size of the standard deduction for non-itemizers and the presence of social security benefits for itemizers.
[oldexf] Age 65+ dummy for secondary filer (in the case of married joint filers). The variable is imputed since 1996 using the size of the standard deduction for non-itemizers and the presence of social security benefits for itemizers.

Income variables:
[agi] Adjusted gross income (follows the tax definition, not consistent across years)
[waginc] Wages and salaries
[peninc] Taxable pension income (includes taxable IRA distributions but not social security benefits)
[penira] Taxable IRA distributions (start in 1987, included in peninc above)
[penincnt] Non-taxable pension income (excluded from fiscal income)
[divinc] Dividend income (both qualified and unqualified dividends, and gross of dividend exclusion that existed before 1987)
[intinc] Taxable interest income

[4] To be precise, the original PUF for year t is created from the universe of tax returns processed by the IRS in year t+1. This corresponds to tax returns for income earned during year t for 97-98% of cases and to tax returns for income earned in earlier years (typically year t-1) for 2-3% of cases (when the tax filer was very late).

[intexm] Tax exempt interest income (from state and local government bonds and excluded from fiscal income, only available since 1987)
[rentinc] Net rental income (from schedule E)
[mortrental] Mortgage interest paid for rental properties on Schedule E (this is a deduction factored into rentinc)
[rylinc] Net royalties income (from schedule E)
[estinc] Estate and trust net income (from schedule E)
[schcinc] Sole proprietorship and farm net income (from schedules C and F)
[partinc] Partnership net income (from schedule E)
[scorinc] S-corporation net income (from schedule E)
[kgagi] Realized capital gains in AGI (long-term capital gains are only partially included in AGI, 50% in 1960-78, 40% in 1979-86, and 100% in 1987+)
[kginc] Full realized capital gains (adjusts kgagi when only a fraction of capital gains was included in AGI before 1987)
[uiinc] Unemployment insurance (UI) benefits. The variable is available in 1979+ only (it is set to zero before 1979 when UI benefits were non taxable and not reported on tax returns). UI benefits are excluded from our fiscal income definition (see below).
[ssinc] Gross social security (SS) retirement and disability benefits. The variable is available in 1984+ only and complete only in 2006+ as taxpayers with zero taxable benefits often did not report their gross benefits in 1984-2005. The variable is set to zero before 1984 when SS benefits were non taxable and not reported on tax returns. SS benefits are excluded from our fiscal income definition (see below).
[sey] Self-employment income. This variable is available in 1984+ only and capped due to the social security tax earnings cap. sey is a part of schcinc and partinc. The variable is set to zero before 1984.
[seysec] Self-employment income of secondary earner. This variable is available in 1984+ only and capped due to the social security tax earnings cap. sey is a part of schcinc and partinc. The variable is set to zero before 1984. The variable is set to zero for non-married filers.
[agicrr] AGI correction defined for consistency across years. It is equal to -UI benefits in AGI – SS benefits in AGI + dividend exclusion (pre-1987).
[income] Total consistent fiscal income (this is the sum of fiscal income components excluding capital gains, UI and SS benefits). It is defined as income=agi+agicrr+agiadj-kgagi (see below for agiadj definition).
[othinc] Other fiscal income (not included in previously defined components above). It is obtained by subtraction: othinc =income- (waginc +peninc +divinc +intinc +rentinc +estinc +rylinc +schcinc +scorinc +partinc)

Deduction variables:
[agiadj] Total adjustments in AGI. It is the sum of above the line deductions for going from gross income to AGI. The list of allowed deductions varies across years depending on tax law.
[studentded] Student loan interest deduction (available in 1998+ and capped at $2500)

[item] Itemized deduction dummy (equal to one for itemizers and zero for non-itemizers)
[itemded] Total itemized deductions on schedule A. Set to zero for non-itemizers.
[mortded] Mortgage interest on schedule A. Set to zero for non-itemizers.
[intded] Total interest on schedule A. Set to zero for non-itemizers. intded is broader than mortded before 1987 when non-mortgage interest expenses (such as credit card interest) were deductible. It is equal to mortded starting in 1987.
[charit] Charitable giving on schedule A. Set to zero for non-itemizers.
[statetax] State and local income taxes on schedule A minus state tax refund from prior year in AGI. Set to zero for non-itemizers.
[realestatetax] Real estate taxes on schedule A. Set to zero for non-itemizers.

Tax variables:
[setax] Self-employment tax. This is the social security tax paid on self-employment earnings.
[fedtax] Federal income taxes paid net of all credits (but never negative)
[eictot] Total Earned Income Tax Credit (EITC) received
[eicrefn] Refundable EITC received
[ctctot] Total child tax credit received
[ctcrefn] Refundable child tax credit received

It is important to keep in mind that homogeneity across variables is not perfect as the legal definition and scope of income components can vary slightly across years. For example, wage income is always net of non-taxable fringe benefits such as health insurance benefits or retirement contributions. The cost of health insurance benefits (relative to total labor compensation) has greatly increased over time. Business income components such as sole proprietorship profits, partnership profits, or S-corporation profits follow the tax definitions and can vary slightly across years for example due to accelerated depreciation rules. Unemployment insurance benefits existed before 1979 but were non taxable and hence not available in our data. Similarly, Social security benefits were non taxable before 1984 and hence not available in our data.

2) Aggregate record with extreme values:
Starting in 2009, the PUF combines records with extreme values into aggregate records. We apply this procedure retrospectively for the pre-2009 years as follows. Extreme value records are defined as those records in the original PUF that contain one or more amount fields with deemed extremely large values. Values are considered extremely large if they are within the highest 30 (after population weighting) amounts reported for any income amount value or within the lowest 30 (after population weighting) amounts reported for any negative income. The rules for identifying extremely large values are applied to all $ variables that have a maximum of $500,000 or more or $ variables that have a minimum of -$50,000 or less (when the variable can be negative). Effectively, this implies that we exclude income variables that are capped (such as the EITC).

The numbers $500,000 and -$50,000 are applied for year 2012 and are adjusted for population wide average AGI for earlier years (to control for income growth and price inflation).

Extreme value records have been removed from the micro-data sample and are aggregated into one of four records identified with variable ID<=0. ID=-1 aggregates extreme value records with negative AGI. Then, extreme value records with positive AGI are ranked by AGI and divided into 3 equally sized groups (after weighting).[5] The three groups generate 3 aggregate records: those with lowest AGI have ID=-3, those with mid-level AGI have ID=-2, those with highest AGI have ID=0 for years 1962-1995 and ID=-4 for years 1996-2008. For years 1996-2008, another aggregate record with ID=0 had already been added based on the comparison of the PUF and the complete internal SOI file for all records with AGI above $10m.[6] Each of these 4 aggregate records represents a typically between 100 and 200 (after population weighting) and always strictly more than 40 records (after population weighting).

These aggregate records have all been set with marital status and age 65+ status and number of dependents and children dependents equal to their rounded values (based on the original extreme value records). Itemized status has been set to 1 for the aggregate records with positive AGI and 0 for the aggregate record with negative AGI.

3) Aggregating records by groups of five:

We aggregate all other records (except the aggregate extreme value records mentioned above) by groups of 5 returns chosen to be similar along some key family status and income variables. We proceed as follows. We first divide records into 12 cells based upon married dummy*itemizing dummy*(age 65+ vs. age<65 with dependent children vs. age<65 without dependent children). We do grouping within each cell so that we only aggregate records with the same marital status, same itemizing status, and same broad demographic category: primary filer aged 65+, primary filer aged less than 65 with children dependents, primary filer aged less than 65 without children dependents. Within each cell, we proceed as follows.

1) We start with the record with the highest AGI

[5] Starting in 2011, the PUF also creates 4 aggregate records from extreme value records: (1) records with negative AGI, (2) records with AGI in the $0,$10m range, (3) records with AGI in the $10m,$100m range, (4) records with AGI in the $100m+ range. In 2012, each of these three positive AGI groups has about the same number of (weighted) records. That is why we decided to split records with positive AGI into 3 equally sized groups for earlier years (using nominal or even indexed AGI thresholds is not as convenient due to large changes in the distribution of AGI since 1962).

[6] The construction of this aggregate record for the PUF 1996-2008 is described in earlier work Saez (2016) who produces the associated tables allowing users to add this aggregate record to the PUF 1996-2008. Adding this aggregate record is necessary because a small number (typically between 10 and 150) of extreme value records were excluded from the PUF sampling for these years (see US Treasury, PUF documentation). This created significant discrepancies between the PUF and the internal data for high income earners.

2) We then calculate the distance of each other record to this highest AGI record using 15 income variables (and normalizing the square distance by the standard deviation for each variable.[7]
3) We choose the closest 4 records to the highest AGI record and this creates a group of 5 records. We then remove these 5 records from the dataset and repeat the process 1)-2)-3) until the records are exhausted (and making sure the last group has at least 5 records).
4) Every 5th iteration, we instead sort from the lowest AGI (as long as there remain records with negative AGI).

Note that the matching will be highest quality for the highest AGI records and the negative AGI records (and lowest quality for small positive AGI records). This is the preferred strategy as (a) quality at the top of the distribution is most important (and negative AGI records often have large income and wealth components), (b) records with low positive AGI tend to be very simple.

After grouping, we take the (weighted) average value for all $ variables so that averages and aggregate $ values are left unaffected by the grouping. For categorical variables, we probabilistically assign round values. For example, if the secondary filer age 65+ dummy averages to 40% in the group, the grouped record is assigned age 65+ status with probability 40% (using seeded iid draws so that the probabilistic assignments can be replicated).

4) Applying the most recent PUF blurring method retrospectively:

We have applied the most recent blurring methods (used in the PUF 2012 as discussed in detail in US Treasury, 2017) retrospectively to all years (and adjusting them a little bit as discussed below whenever the grouping by 5 described above is a stronger blurring strategy. Let us go through the formal list of 10 blurring procedures described in the 2012 PUF documentation.

1) Record with tax year less than file year – 3 are removed. This does not apply as we do not keep the tax year variable in our file (hence the handful of records with tax year less than file year – 3 get anonymously grouped with others).

2) Extreme value returns: creation of aggregate record has been discussed above.

[7] The distance is computed based on the following 15 income variables: (1) AGI, (2) wages, salaries, and taxable pensions, (3) business profits (schedule C+F, partnerships, S-corporations), (4) capital income (interest, royalties, estate and trust, interest, rental income, and dividends), (5) full realized capital gains, (6) wage income, (7) dividend income, (8) interest income, (9) partnership income, (10) S-corporation income, (11) real estate taxes paid, (12) mortgage interest paid, (13) pre-tax national income, (14) post-tax national income, (15) imputed household wealth (double weight). The last 3 variables are not in the original PUF file but are constructed from the PUF and other publicly available sources in the project Piketty, Saez, Zucman (2018). We include them in order to have a better match of the distribution of these variables in the online files. There is a trade-off between choosing too few variables (grouped records will look very similar along these few variables but quite dissimilar along the other variables) vs. too many variables (grouped records will look only somewhat similar along the many chosen variables).

3) High income tax returns with zero tax. We do not specifically sample down (beyond what is described below for returns in general) high income returns with zero tax.
4) No records should be sampled at a rate higher than 10%. We apply a 1/3 maximum sampling rate for years up to 2004 (the 10% maximum sampling is applied to the PUF starting in 2005). Combined with the grouping by 5, this is blurring as powerful as a sampling rate of 1/(3*5)=1/15 which is below the 10% maximum sampling.[8] This subsampling is applied after having constructed the aggregate extreme value records discussed in 2) and before the grouping by 5 discussed above.
5) Blurring of high sampling rate records. Alimony paid, alimony received, state sales tax variables are not in our data and hence are effectively removed. Personal exemptions are not part of the data (and hence do not need to be blurred). Multivariate blurring (by groups of 3 returns) for variables wages and salaries, state and local income taxes, and real estate taxes becomes redundant given the more powerful grouping by 5 that we do.
6) Blurring of normal sampling rate records. Univariate blurring (by groups of 3 returns) for variables alimony paid, alimony received, wages and salaries, medical and dental expenses, real estate taxes, and state and local income taxes becomes redundant given the more powerful grouping by 5 that we do.
7) Marital status surviving spouse converted to married filing joint is not relevant as we only have a married filing jointly dummy in our file.
8) Cap number of dependents at 3. We have capped the number of dependents and number of children at home variables at 3. We do not distinguish singles vs. head of households vs. married filing separately in our dataset so we do not need to apply the more stringent cap at 2 dependents for single returns (and at 1 for married filing separately).
9) Rounding: We have applied the PUF 2012 rounding method to all income variables (except the variables that are recomputed based on other variables in our file).
10) Rebalance for accuracy. We have not rebalanced our sample as all the procedures we have used are designed to keep the sample balanced.

5) Creating aged files for 2013-2016:
It is valuable to have access to datasets based on recent years. There is currently in a long lag in the creation of the PUF: by August 2018, 2012 is the most recent available PUF. Therefore, we have created aged PUF files based on the 2012 file created above and adjusted using tabulations by size of income created using internal data for years 2013-2016. We have proceeded as follows.

First, we use internal SOI data to create the following tabulations (presented in the associated excel worksheets) for each year 2013, 2014, 2015, and 2016. The statistics for 2013, 2014, 2015, 2016 are based on the INSOLE internal SOI files (from which future PUFs will be created).

[8] We choose this strategy because our analysis shows that sampling at a 10% maximum rate and then grouping by 5 weakens the quality of the data at the top of the distribution.

The tabulation made resembles the official SOI publication 1304 (IRS, 2017), Table 1.4 that displays by size of AGI, the number of returns, and total income values for a number of income components but it is disaggregated by 12 strata based on married dummy*itemizing dummy*(aged 65+ vs. age <65 with dependent children vs. age <65 without dependent children). The 12 strata are referred to by the indicator variable strata=1,..,12 in the tabulated file.

For each of the 12 strata, we divide returns based of 16 AGI groups: negative AGI, and among records with zero or positive AGI: 9 bottom deciles, P90-95 (percentile 90 to percentile 95), P95-99, P99-99.5, P99.5-99.9, P99.9-99.99, P99.99-100. The AGI group is denoted by the indicator variable agibin in the tabulated file. agibin=-1 indicates the group with negative AGI; agibin=0 the bottom decile (P0-10), agibin=.1 the second decile (P10-20),.., agibin=.9999 the top .01% (P99.99-100).

For each of the 12*16=192 cells, we compute the number of returns (population weighted) in each cell and the average amount in each cell for each variable in our file. We remove all cells with less than 10 returns (population weighted). We also remove all cells with less than 3 records (without weighting for population).[9] This last restriction is not binding as all cells contain at least 50 records. We denote cells by n=1,..,N=192. Within cell n, we denote by #n the (population weighted) number of returns in cell n and by yn the (population weighted) average of variable y in cell n. These tables are disclosed online as separate files so that users can replicate (or improve upon) our aging methodology described in the next steps just below.

Second, we compute the same table in the 2012 PUF that gives for all cells, number of returns and averages for all variables.[10]

Third, we create the 2013 synthetically aged file by starting from the 2012 online synthetic file that we created above. Records from the 2012 online file naturally fall in one and only one of the 192 cells defined above by demographics and AGI group. Any record falling in cell n is re-weighted by the ratio #n(2013)/#n(2012). Variable y of the record is multiplied by the ratio yn(2013)/yn(2012).[11] This ensures that, for each cell n=1,..,N the total of number of weighted records in the 2013 aged file matches the actual 2013 file. This also ensures that, for each cell n=1,..,N, all variables have exactly the same mean in the aged 2013 file and the actual 2013 file. To avoid small cell issues, the ratios yn(2013)/yn(2012) are capped at 3 (maximum) and 0 (minimum). At the end, we rebalance each variable across all cells with a uniform multiplier to make sure aggregates match the 2013 statistics.

---

[9] We also zero out fractions based on categorical 0/1 variables such as oldexf (secondary filer being less than 65) whenever fewer than 3 records fall into the zero or one category. In practice, this constraint is never binding.

[10] This table is based on the entire 2012 PUF available through NBER rather than the 2012 synthetic online file that we created. This choice is made so that we obtain more precise estimates for our aging projection.

[11] This procedure is also applied to each of the aggregate records in the 2012 online file that were presented above. I.e., each of the aggregate records falls into a specific cell and is re-weighted accordingly.

Fourth, for categorical variables (such as number of children or age 65+ status of secondary filer), we do a probabilistic assignment to ensure that all categorical variables remain categorical. For example, suppose 25% of returns in a given cell have secondary filer with age 65+ in the 2012 file and 30% of returns in the same corresponding cell have secondary filer with age 65+ in the 2013 file. Then, we randomly assign age 65+ status to 5 percentage points of the 75% records without this status in the 2012 file in order to reach 30% as in the 2013 file. This random assignment is made iid with seeding.

The aged files for 2014, 2015, and 2016 are also produced using the same methodology and starting from the 2012 (most recent) PUF. We have tested that the aged files reproduce reasonably well key distributional statistics.

Associated files

There are two sets of files associated with this paper.

1) The synthesized PUF micro-files for 1962, 1964, 1966-2016. These files are in STATA format, one separate file for each year. They contain between 11,000 and 37,000 records and 47 variables described above. They are named pufonline1962.dta, etc. Each annual file is between 2.7MB and 11MB large.

2) The tabulations based on internal data for years 2013-2016. These are small STATA datasets of 192 cells and 49 variables with one dataset for each year. The 49 variables are AGI group indicator [agibin], the demographic strata indicator [strata] and the 47 variables used in the synthesized PUF micro-files. These tabulation files are named agetable2013.dta,.., agetable2016.dta

References

Bryant, Victoria L. John L. Czajka, Georgia Ivsin, and Jim Nunns. 2014. "Design Changes to the SOI Public Use File (PUF)", Prepared for the "New Resources for Microdata-Based Tax Analysis" Session, 2014 Annual Conference on Taxation, National Tax Association Santa Fe, New Mexico. Available online at https://www.irs.gov/pub/irs-soi/14rfpufredesignrecommen.pdf

Piketty, Thomas, Emmanuel Saez, and Gabriel Zucman. (2018) "Distributional National Accounts: Methods and Estimates for the United States", Quarterly Journal of Economics 133(2), 553-609. Data online at http://gabriel-zucman.eu/usdina/

Saez, Emmanuel. (2016) "Statistics of Income Tabulations: High Incomes, Gender, Age, Earnings Split, and Non-filers" SOI Working Paper, available online (with associated tables) at https://www.irs.gov/statistics/soi-tax-stats-soi-working-papers

U.S. Treasury Department, Internal Revenue Service, Statistics of Income Division. (2017) "2012 Statistics of Income Public Use Tax File" (Washington: DC).

U.S. Treasury Department, Internal Revenue Service, Statistics of Income. (2017b). Individual Income Tax Returns, Publication 1304. Annual publication, available online at https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-returns-publication-1304-complete-report