Is masked data suitable for purpose?

- Table 1. Synthetic Data
- Table 2. Univariate Blurring
- Table 3. Binary Rounding
- Table 4. Summary Quantiles

The utility of masked micro data is rarely addressed by the scholars advocating masking tecnhiques or agencies distributing masked data. Early evaluations of the quality of masked or synthesized files by scholars were through demonstrations of statistical similarity, such as comparisons of moments and cross moments (Reiter...). Woo et al propose testing the ability of propensity or logit models to distinguish between actual and synthetic records when the two files are stacked in a single analysis file. They also suggest measuring the distance between the two empirical distribution functions.

Advocates of masking methodologies sometimes provide more information. For example Reiter (2005) provides estimates of the reliability of several subgroup totals and linear regressions based on a synthetic dataset. Nevertheless, anyone using the masked data will have performed different calculations with the data and will wonder if subsets of the data and non-linear functions of the data are equally similar to the gold-standard file. None of these measures provide a path to generating a loss function units that has real, or even intuitive value.

For actual public use data even less information was typically provided. The SOI compares aggregates from the public use file if Individual Tax Returns (henceforth the PUF) with published aggregates, but that can only provide very weak evidence that masking has not affected a particular analysis. Alexander et al (2010) provide an example of a seriously misleading masked dataset.

The tables in this chapter seek to show the ability of synthetic files to reproduce the revenue scores of the public use version of the Individual Tax Model. We believe analysis of the ability to reproduce policy-relevant parameters is the most useful way to demonstrate the usability of masked data. Revenue scores are among the most common and important uses of tax micro-data, and lend themselves to this measure. They are also a good test of bias in masking, since the scores are non-linear functions of (typically) subsets of the data. We also offer statistics for various levels of micro-aggregation (univariate blurring) and binary rounding for comparison.

Tax return micro-data has multiple applications, apart from calculatings of NIPA statistics, the chief uses are estimating behavioral effects of taxation and revenue scoring. Only scoring is examined here, chiefly because it seems unlikely that a synthetic file could be useful for other applications.

Consider that classic research topic - the effect of marginal tax rates on charitable giving. Because the decision to itemize can depend on the amount of charitable giving a sample of all itemizers will induce sample selection bias in the estimates, It is therefore customary to include only those taxpayers whose non-charity deductions are sufficient to justify itemization. This avoids sampling on the dependent variable. This seems like a small point, but if ignored results will be substantially different.

Unless the synthetic dataset is explicitly created with a separate distribution for this particular subset of taxpayers, the ability to correct for this endogeniety will be lost. It is likely that most potential uses for synthetic data will run into one or more similar roadblocks as a CDF model general enough to anticipate the many possible demands placed on a dataset would be unmanageably complex.

Revenue scoring is somewhat simpler. As long as the file can maintain good correlation with marginal tax rates over the relevant range, the result will be accurate. There is a reasonable expectation that a synthetic file could do this.

Obviously we would like to have runs comparing synthetic files to the full (and confidential) Tax Model file, but for the time being we use the PUF (which is already subject to significant masking) for a proof of concept. This does mean that we can not test the effect of the current level of masking. The revenue scores we calculate are simple static scores assuming no taxpayer response. We don't intend to minimize the importance of behavioral response, however the ability to reproduce static scores is clearly the first step towards a believable result.

Synthetic data was provided by Max Ghenis (UBI Center), Anderson Frailey and Don Boyd (SUNY Albany). RF is a Random Forest seeded with

- AGI
- Exemptions
- Taxable Income
- AMT Taxable Income
- Income Tax before Credits
- Income Tax after Credits
- Earned income for EIC
- Total non-passive income

The file is reweighted to match several known totals.

To provide a comparison masking technique for the synthetic files we also apply the same analysis to a file with simple univariate blurring applied to to every dollar value, not just the sensitive ones already blurred in the PUF. Blurring groups include 6, 12, 24 or 48 taxpayers and no other homogeneity is imposed. That is, the file is not separated by marital status or other characteristic before the blurring groups are formed.

Another way to add noise to the data is rounding. In this case we round at the binary bit level, with 1, 2, 3 5 or 8 bits of precision. The original PUF file has about 10 bits of precision.

The last numeric column shows the coefficient of variation for each reform, evaluated with the Jackknife. Clearly, some of the synthetic data's worst scores are have a correspondingly low precision in the PUF itself.

The Tax Reforms

The selection of tax reforms is dictated by the availability of parameters in the existing tax calculator. Each tax rate or clawback rate is a potential reform and is varied by a single percentage point for this study. Each dollar threshold is raised or lowered by $1000. Each income, adjustment, or deduction is reduced by 10 percent of the taxpayer value. Credits are kept or abolished. Each reform is reported as a separate item. In the tables the column "Action" shows the nature of the tax reform. For example "0" means multiply the value by zero before calculating the tax. This gives us 100 potential reforms. While no individual reform is a likely political outcome, it is true that the majority of rates and thresholds modeled have changed in the recent decades, as has the taxability or deductibility of the income or deduction items studied, albeit not to the exact quantitative extent that we model. If the masking does not damage many of these scores, then a reader might take the reasonable conclusion that the the masked files are a good substitute for the actual data in the confidential files.

In the tables, each cell shows the percentage deviation in the revenue score of the masked file from the 2011 PUF applied to 2013 income. So, for example, in Table 1 we see that the revenue score calculated with the CARTFS synthetic file from abolishing the deduction for IRA contributions (Line 1) 2.95% smaller than the score based on the PUF itself. That is more accurate than most of the measured reforms, but most results are quite satisfactory given the likelihood of behavioral and other unknowns. The worst results are seen with income items that can be negative. Here the synthetic file does reasonably well on the net income, but if the positive and negative values are modified separately (see lines 78-79, for example) the results are much worse. By comparison, Table 2 shows that even with blurring groups of 48 returns only a few scores differ from the PUF estimate by more than 1%. Table 3 results show that even severe rounding leaves most of the scores quite accurate.

Returning to the synthetic data, notice, for example, that the correlation of farm income with Taxable Income in the synthetic file (not shown but .83) is reasonably close to the correlation in the PUF, but the revenue score is wildly off - by a factor of 16 in the RF file and not much better in the other files. This is an unusual case, but indicative of the mischief that non-linear functions of data subsets can have on superficially similar datasets. The fields that are net of positive and negative values of similar magnitude are particularly vulnerable.

The column "CART subset" takes the CARTFS file and throws out any synthetic record which does not have at least 7 other observations "nearby" in the PUF. The idea is to eliminate records that might fool an intruder into thinking he had a real tax record by their similarity on multiple variables. The definition of nearby is "within twice the Euclidian distance of the closest record". All the variables are standardized to mean zero and standard deviation 1. That is, if the closest record is within x, then we want to see multiple records within 2x, so that the closest record is not a much better match than multiple other records, and therefore not a particularly high probability of being a disclosure. At this time the resulting file is not reweighted, and does poorly when measured by the 10 worst scores but is similar to the other files up to the 75th percentile.

A summary of the results is given in Table 4 which shows the quantiles of the absolute values of the percentage descrepencies. For example, the median descrepancy of the CARTFS score from the PUF calculated value is only 3.99% off the PUF score. At the 95th percentile the difference ranges from 163% to 200% of the PUF score across the various synthetic files, but it is still true that 75% of the results are within 25% of the PUF score in all of the synthetic files.

So, is the masked data suitable for purpose? One reasonable interpretation is that it is suitable for training and student exercises, but probably not for legislation. The scores calculated here offer a direct intuitive demonstration of the usability of the masked file.

Daniel Feenberg
feenberg@nber.org
26 August 2019

Reference
Inaccurate age and sex data in the Census PUMS files: Evidence and Implications [] Trent Alexander, Michael Davern, Betsey Stevenson NBER Working Paper No. 15703 Issued in January 2010

Table 1. Synthetic Data

% deviation of syn File Score from PUF Score 2013

Obs	name	RF	RF-REDUCED	CART	CARTFS	puf_cv	action	label
1	_03150	6.83	13.70	14.06	-2.95	2.96	*0	Deductible IRA Payments
2	_03210	5.08	30.99	30.72	28.40	1.69	*0	Student Loan Interest Deduction
3	_03220	12.36	9.54	8.35	-2.19	2.15	*0	Educator Expenses
4	_03230	3.97	26.00	16.81	14.62	4.11	*0	Tuition and Fees Deduction
5	_03240	68.35	24.07	20.58	50.36	2.83	*0	Domestic Production Activities Deduction
6	_03260	20.19	13.57	12.39	17.37	.	*0	Deduction for self-employment tax
7	_03270	19.63	12.36	14.78	14.30	2.06	*0	Self-employment Health Insurance
8	_03290	20.04	11.24	14.12	5.75	3.85	*0	Health Saving Account Deduction
9	_03300	23.16	21.27	15.50	18.24	2.54	*0	Payments to KEOGH
10	_03400	-1.68	-19.90	-20.79	-43.66	23.85	*0	Forfeited Interest Penalty
11	_03500	-31.67	4.16	15.51	-10.47	9.05	*0	Alimony Paid
12	_07180	-5.84	-5.29	-5.83	-6.37	1.72	*0	Child and dependent care
13	_07220	-9.39	0.95	10.13	24.76	4.73	*0	Child Tax Credit
14	_07230	-1.76	-6.24	-6.11	-4.03	1.34	*0	Education Credits
15	_07240	-5.06	-4.90	-4.02	-6.17	2.07	*0	Retirement Savings Credit
16	_07260	5.55	-2.82	4.07	-7.05	4.57	*0	Residential Energy Credit
17	_07300	-4.40	-2.44	-5.53	-5.25	3.52	*0	Foreign tax
18	_07600	37.23	24.00	23.08	39.72	7.10	*0	Credit for prior year minimum tax
19	_11070	2.19	-0.73	-0.38	1.46	0.97	*0	Additional Child Tax Credit
20	_adctcrt	-0.53	-6.02	-5.27	-2.19	1.85	-.01	Rate for additional ctc
21	_agcmax	-3.75	-8.26	-5.30	-2.07	-4.27	-1000	start of child care phase-down
22	_almsep	-17.26	-28.25	12.43	-16.01	15.79	-1000	Extra alminc for married sep
23	_almsp	-2.90	-1.83	-3.54	-8.01	5.78	-1000	AMT bracket
24	_amex	0.21	-0.04	-0.18	0.62	0.34	-1000	Personal Exemption
25	_amtex	-7.53	-12.14	-13.86	-11.97	4.88	-1000	AMT Exclusion
26	_amtx25	-5.31	-5.55	-7.23	-9.86	4.56	-1000	AMT exclusion phaseout rate
27	_amtys	-6.63	-10.98	-13.00	-11.76	5.23	-1000	AMT phaseout start
28	_brk1	0.05	0.14	0.01	0.15	0.29	-1000	10% tax rate thresholds
29	_brk2	0.32	5.80	5.85	0.81	0.60	-1000	15% tax rate thresholds
30	_brk3	5.31	33.65	33.92	9.70	1.28	-1000	25% rate thresholds

Obs	name	RF	RF-REDUCED	CART	CARTFS	puf_cv	action	label
31	_brk4	6.80	59.04	61.15	14.14	1.55	-1000	28% rate thresholds
32	_brk5	0.91	23.50	25.22	2.06	1.07	-1000	33% rate thresholds
33	_brk6	0.04	24.05	24.95	0.57	1.01	-1000	35% rate thresholds
34	_cgrate1	-11.36	-4.27	-5.73	-8.81	2.28	+.01	Initial rate on long term gains
35	_chmax	1.47	-1.49	-1.46	1.41	0.65	-1000	Max Child Tax Credit per child
36	_cphase	0.98	4.01	3.23	2.16	2.17	-1000	Child Tax Credit Phase-Out
37	_crmax	0.29	-3.56	-3.46	-1.41	0.12	-1000	Maximum earned income credit
38	_dylim	515.98	3989.3	3292.1	2105.8	-18.04	-1000	Limits for EIC Disqualified Income
39	_ealim	-0.80	-4.66	-4.61	-3.75	1.55		Maximum earn ACTC
40	_edphhm	-7.03	4.87	6.53	8.36	4.33	*.9	end of education credit phaseout - married
41	_edphhs	-2.52	-6.32	-6.58	-7.13	4.10	*.9	End of education credit phaseout - singles
42	_exmpb	5.45	61.65	65.89	15.80	-2.29	*.9	Personal Exemption Amount start of phaseout
43	_joint	-3.29	-15.54	-14.26	-9.25	1.94	*.9	Extra to ymax for joint return
44	_med38	1.45	25.34	27.44	2.73	1.57	*.9	3.8% Medicare tax on property income
45	_med9	12.91	43.91	34.43	6.09	.	*.9	.9% Medicare tax on Excess Earning
46	_pcmax	-4.17	-9.13	-5.55	-2.28	4.34		Maximum Percentage for f2441
47	_rt1	0.06	-0.61	-0.70	0.20	0.27	+.01	10% rate
48	_rt2	0.25	2.12	2.08	0.37	0.40	+.01	15% rate
49	_rt3	1.93	13.57	13.61	2.66	0.76	+.01	25% rate
50	_rt4	5.94	45.02	46.17	11.19	1.31	+.01	28% rate
51	_rt5	2.08	35.72	37.99	5.30	0.99	+.01	33% rate
52	_rt6	0.74	19.81	21.60	1.68	1.02	+.01	35% rate
53	_rtbase	-0.13	-8.12	-7.14	-6.85	1.11	-.01	EIC base rate
54	_rtless	-2.29	-7.46	-7.50	-4.89	1.16	+.01	EIC phaseout rate
55	_ssb50	-0.48	-8.72	-9.21	-0.62	-1.43	+1000	SS 50% taxability threshold
56	_ssb50r	-1.36	-7.97	-8.57	-1.00	1.54	+.01	SS inclusion rate above _ssb50 threshold
57	_ssb85	-1.23	-4.12	-5.55	0.18	-2.04	+1000	SS 85% taxability threshold
58	_ssb85r	-1.55	8.29	7.28	-1.15	1.28	+.01	SS inclusion rate above _ssb85 threshold
59	_ssmax	29.44	20.05	18.71	13.13	.	+1000	SS Maximum taxable earnings
60	_stded	2.19	4.94	4.83	2.49	-0.37	+1000	Basic Standard Deduction
61	_tamtr2	-1.04	3.66	3.09	-3.24	3.71	+.01	AMT surtax rate
62	_tamtr26	-5.02	-4.78	-5.25	-8.80	3.02	+.01	AMT basic rate
63	_thresx	3.40	51.09	53.45	7.15	-3.65	-1000	Threshold for additional Medicare tax
64	_ymax	-0.62	-5.58	-5.39	-3.10	0.90	-1000	Start of EIC phaseout
65	c00100_total	-0.00	2.12	2.24	0.03	.		
66	c04800_total	-0.16	5.69	5.93	-0.14	.		

Obs	name	RF	RF-REDUCED	CART	CARTFS	puf_cv	action	label
67	c09600_total	2.75	1.81	1.35	2.88	.		
68	e00200	-0.26	0.31	0.47	-0.41	-0.47	*.9	Salaries and wages
69	e00200_total	-0.01	0.06	0.14	0.01	.		
70	e00300	-2.13	17.75	21.46	-0.99	-1.79	*.9	Interest received
71	e00400	14.69	-22.77	9.44	-19.99	-8.66	*.9	Tax-exempt interest income
72	e00650	-1.64	4.42	4.88	-1.79	-1.84	*.9	Qualified Dividends
73	e00700	-2.04	5.66	4.59	-2.34	-1.24	*.9	State income tax refunds
74	e00800	-14.41	56.02	35.42	-31.97	-12.41	*.9	Alimony received
75	e00900	-1.28	11.02	10.16	-0.60	-1.88	*.9	Business or profession (Schedule C) net profit/loss
76	e01100	18.93	43.73	43.86	4.11	-9.78	*.9	Capital gain distributions reported on Form 1040
77	e01200	-300.2	3282.0	2900.6	-645.9	-56.08	*.9	Other gains (or loss) (+/-)
78	e01200_neg	162.29	71.61	103.87	265.02	5.10		
79	e01200_pos	125.72	429.00	417.65	166.05	-5.21		
80	e01400	-1.87	17.93	18.39	-1.35	-1.87	*.9	Taxable IRA distribution
81	e01700	-1.24	3.95	4.14	-0.65	-1.18	*.9	Pensions and annuities included in AGI
82	e02000	-0.83	18.08	18.58	-0.40	-1.29	*.9	Schedule E net income or loss (+/-)
83	e02100	1613.9	875.01	762.92	1191.6	58.70	*.9	Schedule F net profit/loss
84	e02100_neg	175.01	107.26	111.37	152.85	4.04		
85	e02100_pos	426.13	222.88	197.41	329.67	-4.86		
86	e02300	-0.51	23.81	25.30	0.59	-1.69	*.9	Unemployment compensation in AGI
87	e02400	-1.58	6.72	5.91	-0.97	-1.17	*.9	Gross Social Security benefits
88	e07400	27.00	5.08	9.62	1.51	5.37	*.9	General business credit
89	e08000	3.42	-35.20	-18.44	-23.06	20.29	*.9	Other credits
90	e09700	130.95	-18.99	14.27	34.66	-45.49	*.9	Recapture taxes
91	e09800	24.35	25.14	3.82	-8.75	-18.01	*.9	Social security tax on tip income
92	e09900	-11.78	6.76	8.70	-11.53	-3.20	*.9	Penalty tax on IRA
93	e17500	-5.18	-18.50	-20.23	-6.63	2.47	*.9	Medical and dental expenses
94	e18400	-2.20	-2.50	-2.50	-2.33	0.72	*.9	State and local taxes
95	e18500	-1.59	0.16	-0.46	-1.54	0.75	*.9	Real estate tax deductions
96	e19200	-0.84	-2.38	-2.33	-1.31	0.76	*.9	Total interest paid
97	e19800	-3.41	-1.18	-2.74	-5.10	1.28	*.9	Cash contributions
98	e20100	-6.78	-5.86	-7.03	-8.94	3.63	*.9	Other than cash contributions
99	e20400	0.25	0.79	0.92	1.24	1.42	*.9	Miscellaneous deductions before AGI limitation
100	e22250	14.73	182.53	199.17	56.17	-6.72	*.9	Short-Term Gain/Loss Net of Carryover
101	e23250	-1.98	18.04	18.61	-1.01	-1.85	*.9	Long term gains less losses
102	e24518	-21.15	29.30	34.48	-15.09	-13.42	*.9	28% Rate Gain or Loss
103	e26270	-6.80	-8.55	-7.28	-18.00	1.40	*.9	Combined partnership and S corporation net income

Obs	name	RF	RF-REDUCED	CART	CARTFS	puf_cv	action	label
104	e32800	-5.76	-4.61	-5.43	-6.63	1.85	*.9	Child Care Qualifying Expenses
105	e58990	34.52	18.86	8.61	2.59	-8.80	*.9	Investment income (Form 4952 part 2 line 4g)
106	e62900	-9.50	-7.39	-13.94	-14.81	8.33	*.9	Alternative tax foreign tax credit
107	e87521	-2.82	-7.40	-7.85	-5.59	1.77	*.9	American Opportunity Credit
108	e87530	-1.46	-0.96	-0.06	1.74	3.44	*.9	Lifetime Learning Total Qualified Expenses

Table 2. Univariate Blurring

% deviation of blur File Score from PUF Score 2011

Obs	name	BLUR06	BLUR12	BLUR24	BLUR48	puf_cv	action	label
1	_03150	-0.04	-0.08	-0.10	-0.05	2.96	*0	Deductible IRA Payments
2	_03210	-0.01	-0.03	-0.09	-0.11	1.69	*0	Student Loan Interest Deduction
3	_03220	-0.02	-0.01	-0.04	-0.14	2.15	*0	Educator Expenses
4	_03230	0.00	-0.05	-0.03	-0.44	4.11	*0	Tuition and Fees Deduction
5	_03240	0.01	-0.21	-0.21	-0.25	2.83	*0	Domestic Production Activities Deduction
6	_03260	-0.03	-0.04	-0.09	-0.10	.	*0	Deduction for self-employment tax
7	_03270	-0.02	-0.01	-0.02	-0.01	2.06	*0	Self-employment Health Insurance
8	_03290	-0.03	-0.06	-0.13	-0.14	3.85	*0	Health Saving Account Deduction
9	_03300	-0.01	-0.01	-0.02	-0.05	2.54	*0	Payments to KEOGH
10	_03400	2.15	-1.21	10.94	12.78	23.85	*0	Forfeited Interest Penalty
11	_03500	0.18	0.10	-0.13	-0.13	9.05	*0	Alimony Paid
12	_07180	-0.01	-0.03	-0.04	-0.04	1.72	*0	Child and dependent care
13	_07220	-0.27	-0.36	-0.18	-1.11	4.73	*0	Child Tax Credit
14	_07230	-0.00	-0.02	-0.02	-0.04	1.34	*0	Education Credits
15	_07240	-0.10	-0.21	-0.22	-0.44	2.07	*0	Retirement Savings Credit
16	_07260	0.06	0.08	0.09	-0.17	4.57	*0	Residential Energy Credit
17	_07300	-0.05	-0.05	-0.07	-0.10	3.52	*0	Foreign tax
18	_07600	0.09	0.45	0.55	0.64	7.10	*0	Credit for prior year minimum tax
19	_11070	0.00	0.00	0.00	0.00	0.97	*0	Additional Child Tax Credit
20	_adctcrt	0.01	0.02	0.04	0.08	1.85	-.01	Rate for additional ctc
21	_agcmax	-0.12	0.14	0.37	0.67	-4.27	-1000	start of child care phase-down
22	_almsep	-0.24	0.18	0.43	-0.14	15.79	-1000	Extra alminc for married sep
23	_almsp	-0.02	-0.11	0.06	0.04	5.78	-1000	AMT bracket
24	_amex	-0.01	-0.02	-0.04	-0.05	0.34	-1000	Personal Exemption
25	_amtex	-0.02	-0.14	-0.14	-0.21	4.88	-1000	AMT Exclusion
26	_amtx25	-0.01	-0.06	-0.07	-0.09	4.56	-1000	AMT exclusion phaseout rate
27	_amtys	-0.07	-0.06	-0.11	-0.10	5.23	-1000	AMT phaseout start
28	_brk1	-0.01	-0.02	-0.04	-0.06	0.29	-1000	10% tax rate thresholds
29	_brk2	-0.00	-0.00	-0.02	-0.04	0.60	-1000	15% tax rate thresholds
30	_brk3	-0.06	0.09	0.17	0.11	1.28	-1000	25% rate thresholds

Obs	name	BLUR06	BLUR12	BLUR24	BLUR48	puf_cv	action	label
31	_brk4	-0.05	0.22	-0.09	-0.19	1.55	-1000	28% rate thresholds
32	_brk5	0.05	-0.00	-0.14	-0.07	1.07	-1000	33% rate thresholds
33	_cgrate1	-0.04	-0.09	-0.11	-0.18	2.28	+.01	Initial rate on long term gains
34	_cgrate2	0.07	-0.04	-0.02	-0.16	.	+.01	Normal rate on long term gains
35	_chmax	-0.01	-0.01	-0.01	-0.02	0.65	-1000	Max Child Tax Credit per child
36	_cphase	-0.00	0.13	0.14	0.13	2.17	-1000	Child Tax Credit Phase-Out
37	_crmax	-0.00	-0.00	-0.00	-0.00	0.12	-1000	Maximum earned income credit
38	_dylim	0.03	-0.05	2.63	-0.35	-18.04	-1000	Limits for EIC Disqualified Income
39	_ealim	0.02	0.03	0.06	0.10	1.55		Maximum earn ACTC
40	_edphhm	-0.00	0.00	-0.03	0.09	4.33	*.9	end of education credit phaseout - married
41	_edphhs	-0.01	0.00	-0.00	-0.04	4.10	*.9	End of education credit phaseout - singles
42	_joint	-0.01	0.02	-0.02	-0.00	1.94	*.9	Extra to ymax for joint return
43	_pcmax	-0.11	-0.05	0.30	1.11	4.34		Maximum Percentage for f2441
44	_rt1	-0.01	-0.03	-0.04	-0.07	0.27	+.01	10% rate
45	_rt2	-0.01	-0.02	-0.03	-0.04	0.40	+.01	15% rate
46	_rt3	0.01	0.01	-0.01	-0.00	0.76	+.01	25% rate
47	_rt4	-0.02	0.03	-0.07	-0.05	1.31	+.01	28% rate
48	_rt5	0.01	-0.02	-0.35	-0.29	0.99	+.01	33% rate
49	_rt6	0.03	0.08	-0.25	-0.17	1.02	+.01	35% rate
50	_rtbase	0.01	0.03	0.05	0.11	1.11	-.01	EIC base rate
51	_rtless	-0.03	-0.04	-0.02	-0.13	1.16	+.01	EIC phaseout rate
52	_ssb50	-0.00	-0.07	-0.12	-0.13	-1.43	+1000	SS 50% taxability threshold
53	_ssb50r	-0.05	-0.16	-0.14	-0.30	1.54	+.01	SS inclusion rate above _ssb50 threshold
54	_ssb85	-0.01	-0.15	-0.18	-0.05	-2.04	+1000	SS 85% taxability threshold
55	_ssb85r	-0.02	0.00	-0.06	-0.08	1.28	+.01	SS inclusion rate above _ssb85 threshold
56	_ssmax	0.04	0.04	0.08	0.00	.	+1000	SS Maximum taxable earnings
57	_stddsi	0.00	0.01	0.02	0.03	1.84	-1000	Dependents standard deduction earnings addition
58	_stded	0.01	0.01	0.01	0.02	-0.37	+1000	Basic Standard Deduction
59	_tamtr2	-0.01	-0.01	0.04	-0.21	3.71	+.01	AMT surtax rate
60	_tamtr26	-0.03	-0.04	-0.01	-0.18	3.02	+.01	AMT basic rate
61	_ymax	-0.02	-0.03	-0.04	-0.11	0.90	-1000	Start of EIC phaseout
62	c00100_total	-0.00	-0.01	-0.01	-0.01	.		
63	c04800_total	-0.00	-0.00	-0.00	0.01	.		
64	c09600_total	-0.04	0.03	0.21	0.37	.		
65	e00200	-0.01	-0.01	-0.01	-0.00	-0.47	*.9	Salaries and wages

Obs	name	BLUR06	BLUR12	BLUR24	BLUR48	puf_cv	action	label
66	e00200_total	-0.00	-0.00	-0.01	-0.01	.		
67	e00300	0.02	-0.18	-0.28	-0.12	-1.79	*.9	Interest received
68	e00400	-0.14	-0.03	0.10	0.10	-8.66	*.9	Tax-exempt interest income
69	e00650	-0.22	-0.04	0.00	-0.02	-1.84	*.9	Qualified Dividends
70	e00700	0.05	-0.10	-0.32	-0.26	-1.24	*.9	State income tax refunds
71	e00800	-0.32	-1.53	-4.05	-4.90	-12.41	*.9	Alimony received
72	e00900	-0.02	0.04	0.11	0.14	-1.88	*.9	Business or profession (Schedule C) net profit/loss
73	e01100	0.73	1.77	2.17	7.51	-9.78	*.9	Capital gain distributions reported on Form 1040
74	e01200	12.88	39.19	0.30	18.66	-56.08	*.9	Other gains (or loss) (+/-)
75	e01200_neg	-0.43	-0.68	-1.04	-1.95	5.10		
76	e01200_pos	0.04	0.20	-0.25	-0.23	-5.21		
77	e01400	-0.02	-0.02	-0.04	-0.05	-1.87	*.9	Taxable IRA distribution
78	e01700	-0.02	-0.03	-0.04	-0.02	-1.18	*.9	Pensions and annuities included in AGI
79	e02000	0.08	0.08	0.06	0.08	-1.29	*.9	Schedule E net income or loss (+/-)
80	e02100	-1.45	-2.07	-4.40	-18.17	58.70	*.9	Schedule F net profit/loss
81	e02100_neg	0.03	0.23	-0.00	-0.28	4.04		
82	e02100_pos	-0.10	-0.04	-0.03	-0.16	-4.86		
83	e02300	0.04	0.01	-0.08	-0.14	-1.69	*.9	Unemployment compensation in AGI
84	e02400	-0.01	-0.03	-0.05	-0.04	-1.17	*.9	Gross Social Security benefits
85	e07400	-0.23	-0.21	-0.71	-1.67	5.37	*.9	General business credit
86	e08000	-1.64	-6.14	-12.01	-19.66	20.29	*.9	Other credits
87	e09700	62.06	111.67	138.94	142.41	-45.49	*.9	Recapture taxes
88	e09800	-0.04	0.30	-0.05	-1.16	-18.01	*.9	Social security tax on tip income
89	e09900	-0.02	0.04	0.33	0.31	-3.20	*.9	Penalty tax on IRA
90	e17500	-0.02	0.04	-0.09	-0.21	2.47	*.9	Medical and dental expenses
91	e18400	0.10	0.03	-0.24	-0.35	0.72	*.9	State and local taxes
92	e18500	-0.00	-0.01	-0.09	-0.11	0.75	*.9	Real estate tax deductions
93	e19200	-0.01	-0.04	-0.06	-0.13	0.76	*.9	Total interest paid
94	e19800	-0.19	-0.39	-0.56	-0.62	1.28	*.9	Cash contributions
95	e20100	0.90	0.69	-0.15	-1.14	3.63	*.9	Other than cash contributions
96	e20400	0.08	0.05	-0.07	-0.01	1.42	*.9	Miscellaneous deductions before AGI limitation
97	e22250	-0.09	0.46	0.36	-1.50	-6.72	*.9	Short-Term Gain/Loss Net of Carryover
98	e23250	0.07	0.10	0.32	0.34	-1.85	*.9	Long term gains less losses
99	e24518	-2.10	-1.20	-2.33	-1.95	-13.42	*.9	28% Rate Gain or Loss
100	e32800	-0.01	-0.06	-0.05	-0.03	1.85	*.9	Child Care Qualifying Expenses
101	e58990	-3.58	-4.63	-6.41	-8.86	-8.80	*.9	Investment income (Form 4952 part 2 line 4g)

Obs	name	BLUR06	BLUR12	BLUR24	BLUR48	puf_cv	action	label
102	e62900	-1.32	-1.90	-3.44	-9.58	8.33	*.9	Alternative tax foreign tax credit
103	e87521	0.00	-0.01	-0.02	-0.03	1.77	*.9	American Opportunity Credit
104	e87530	0.01	0.01	-0.04	0.00	3.44	*.9	Lifetime Learning Total Qualified Expenses

Table 3. Binary Rounding

% deviation of round File Score from PUF Score 2011

Obs	name	R01	R02	R03	R05	R08	puf_cv	action	label
1	_03150	-0.24	-0.20	-0.49	-0.51	-0.47	2.96	*0	Deductible IRA Payments
2	_03210	0.30	0.06	-0.33	-0.39	-0.35	1.69	*0	Student Loan Interest Deduction
3	_03220	0.06	-0.26	-0.31	-0.41	-0.39	2.15	*0	Educator Expenses
4	_03230	0.40	-0.25	-0.47	-0.31	-0.34	4.11	*0	Tuition and Fees Deduction
5	_03240	0.15	-0.19	-0.25	-0.04	-0.07	2.83	*0	Domestic Production Activities Deduction
6	_03270	0.27	-0.34	-0.63	-0.75	-0.75	2.06	*0	Self-employment Health Insurance
7	_03290	0.01	-0.97	-1.00	-0.87	-0.89	3.85	*0	Health Saving Account Deduction
8	_03300	-0.50	-0.58	-0.63	-0.62	-0.62	2.54	*0	Payments to KEOGH
9	_03400	1.08	0.64	-0.06	0.10	0.10	23.85	*0	Forfeited Interest Penalty
10	_03500	0.26	-0.11	-0.05	-0.11	-0.14	9.05	*0	Alimony Paid
11	_07180	-20.63	2.20	2.27	-0.17	1.06	1.72	*0	Child and dependent care
12	_07220	-64.89	-65.94	-66.34	-66.15	-66.13	4.73	*0	Child Tax Credit
13	_07230	-8.87	-8.66	0.56	0.72	-0.24	1.34	*0	Education Credits
14	_07240	3.23	-0.83	-1.32	0.08	-0.03	2.07	*0	Retirement Savings Credit
15	_07260	-3.28	-0.53	-0.17	0.03	-0.00	4.57	*0	Residential Energy Credit
16	_07300	-4.32	-1.57	-0.23	0.01	-0.00	3.52	*0	Foreign tax
17	_07600	-6.35	-2.09	-0.83	0.12	-0.04	7.10	*0	Credit for prior year minimum tax
18	_11070	15.42	16.89	17.75	17.78	17.77	0.97	*0	Additional Child Tax Credit
19	_adctcrt	33.81	30.67	32.93	31.91	31.88	1.85	-.01	Rate for additional ctc
20	_agcmax	-3.40	3.31	0.21	1.80	1.03	-4.27	-1000	start of child care phase-down
21	_almsep	-6.04	-5.24	-1.21	-0.12	0.51	15.79	-1000	Extra alminc for married sep
22	_almsp	-2.01	0.02	-0.14	-0.52	-0.34	5.78	-1000	AMT bracket
23	_amex	1.25	0.53	0.12	0.13	0.16	0.34	-1000	Personal Exemption
24	_amtex	-2.94	-1.35	-0.56	-0.50	-0.43	4.88	-1000	AMT Exclusion
25	_amtx25	-2.01	-0.87	-0.77	-0.64	-0.58	4.56	-1000	AMT exclusion phaseout rate
26	_amtys	-3.08	-1.08	-0.60	-0.49	-0.42	5.23	-1000	AMT phaseout start
27	_brk1	0.66	0.26	-0.28	-0.30	-0.26	0.29	-1000	10% tax rate thresholds
28	_brk2	1.06	0.14	-0.40	-0.47	-0.46	0.60	-1000	15% tax rate thresholds
29	_brk3	4.32	0.55	-0.18	-0.96	-1.04	1.28	-1000	25% rate thresholds
30	_brk4	8.07	1.27	-0.62	-1.22	-1.53	1.55	-1000	28% rate thresholds
31	_brk5	4.45	0.53	-0.95	-1.05	-1.18	1.07	-1000	33% rate thresholds
32	_cgrate1	-2.54	0.00	-0.06	0.37	0.58	2.28	+.01	Initial rate on long term gains
33	_cgrate2	-3.07	-1.18	-0.36	-0.14	-0.10	.	+.01	Normal rate on long term gains
34	_chmax	7.43	7.56	7.72	7.73	7.73	0.65	-1000	Max Child Tax Credit per child
35	_cphase	0.02	0.09	-0.06	-0.11	-0.12	2.17	-1000	Child Tax Credit Phase-Out
36	_crmax	4.07	4.16	4.46	4.50	4.50	0.12	-1000	Maximum earned income credit

Obs	name	R01	R02	R03	R05	R08	puf_cv	action	label
37	_dylim	53.88	58.48	4.02	25.17	25.16	-18.04	-1000	Limits for EIC Disqualified Income
38	_ealim	26.57	21.95	23.76	23.14	23.13	1.55		Maximum earn ACTC
39	_edphhm	-2.63	-1.79	-0.23	-0.09	0.03	4.33	*.9	end of education credit phaseout - married
40	_edphhs	-3.66	-1.42	-0.07	0.12	0.10	4.10	*.9	End of education credit phaseout - singles
41	_joint	11.09	11.54	11.01	11.16	11.20	1.94	*.9	Extra to ymax for joint return
42	_pcmax	-3.64	3.78	0.41	1.57	1.02	4.34		Maximum Percentage for f2441
43	_rt1	0.69	0.24	-0.16	-0.14	-0.11	0.27	+.01	10% rate
44	_rt2	0.59	-0.03	-0.35	-0.38	-0.37	0.40	+.01	15% rate
45	_rt3	1.98	-0.02	-0.61	-0.69	-0.69	0.76	+.01	25% rate
46	_rt4	5.04	0.59	-0.91	-1.19	-1.27	1.31	+.01	28% rate
47	_rt5	6.14	1.75	-0.58	-0.87	-0.99	0.99	+.01	33% rate
48	_rt6	0.75	-0.57	-0.67	-0.69	-0.64	1.02	+.01	35% rate
49	_rtbase	14.09	18.88	10.72	14.34	14.40	1.11	-.01	EIC base rate
50	_rtless	4.55	4.67	4.63	4.68	4.67	1.16	+.01	EIC phaseout rate
51	_ssb50	1.24	0.97	0.52	0.48	0.43	-1.43	+1000	SS 50% taxability threshold
52	_ssb50r	0.33	0.29	-0.17	0.20	0.16	1.54	+.01	SS inclusion rate above _ssb50 threshold
53	_ssb85	0.29	0.01	-0.15	0.07	-0.00	-2.04	+1000	SS 85% taxability threshold
54	_ssb85r	-0.29	-0.24	-0.34	-0.31	-0.30	1.28	+.01	SS inclusion rate above _ssb85 threshold
55	_stddsi	0.23	0.20	0.24	0.27	0.36	1.84	-1000	Dependents standard deduction earnings addition
56	_stded	4.23	1.37	0.33	-0.04	-0.07	-0.37	+1000	Basic Standard Deduction
57	_tamtr2	-3.83	-1.86	-1.09	-0.71	-0.53	3.71	+.01	AMT surtax rate
58	_tamtr26	-3.29	-1.71	-0.90	-0.62	-0.51	3.02	+.01	AMT basic rate
59	_ymax	7.79	7.83	7.59	7.75	7.65	0.90	-1000	Start of EIC phaseout
60	c00100_total	-0.48	-0.37	-0.32	-0.31	-0.31	.		
61	c04800_total	0.00	-0.31	-0.39	-0.42	-0.42	.		
62	c09600_total	-3.17	-0.89	-0.20	-0.13	-0.11	.		
63	e00200	0.57	0.35	0.17	0.12	0.11	-0.47	*.9	Salaries and wages
64	e00200_total	.	0.00	0.00	0.00	0.00	.		
65	e00300	0.22	-0.02	-0.12	-0.17	-0.15	-1.79	*.9	Interest received
66	e00400	3.93	3.51	1.73	1.63	1.75	-8.66	*.9	Tax-exempt interest income
67	e00650	0.70	0.20	-0.08	-0.06	-0.14	-1.84	*.9	Qualified Dividends
68	e00700	1.21	0.53	-0.09	0.02	0.08	-1.24	*.9	State income tax refunds
69	e00800	0.80	0.59	0.70	0.55	0.48	-12.41	*.9	Alimony received
70	e00900	1.04	-0.16	-1.22	-1.95	-2.11	-1.88	*.9	Business or profession (Schedule C) net profit/loss
71	e01100	-0.73	-0.39	-0.45	-0.43	-0.40	-9.78	*.9	Capital gain distributions reported on Form 1040
72	e01200	8.52	6.73	1.31	0.11	0.24	-56.08	*.9	Other gains (or loss) (+/-)
73	e01200_neg	0.18	-0.28	-0.25	-0.28	-0.28	5.10		
74	e01200_pos	0.61	0.05	0.02	-0.18	-0.21	-5.21		
75	e01400	0.18	-0.01	-0.01	-0.07	-0.09	-1.87	*.9	Taxable IRA distribution

Obs	name	R01	R02	R03	R05	R08	puf_cv	action	label
76	e01700	0.34	0.11	0.02	-0.01	-0.01	-1.18	*.9	Pensions and annuities included in AGI
77	e02000	-0.01	-0.02	-0.06	-0.18	-0.18	-1.29	*.9	Schedule E net income or loss (+/-)
78	e02100	25.26	16.09	22.25	24.91	25.66	58.70	*.9	Schedule F net profit/loss
79	e02100_neg	0.24	-0.00	0.21	0.17	0.18	4.04		
80	e02100_pos	-1.28	-1.75	-1.70	-2.06	-2.15	-4.86		
81	e02300	1.72	1.31	1.18	0.94	0.99	-1.69	*.9	Unemployment compensation in AGI
82	e02400	-0.16	-0.11	-0.15	-0.23	-0.24	-1.17	*.9	Gross Social Security benefits
83	e07400	-6.54	-3.36	-1.11	-0.20	-0.06	5.37	*.9	General business credit
84	e08000	-10.72	0.37	-2.35	-1.27	-0.03	20.29	*.9	Other credits
85	e09700	-1.03	-4.54	-3.44	0.57	0.08	-45.49	*.9	Recapture taxes
86	e09800	-8.26	0.71	-0.40	0.76	0.02	-18.01	*.9	Social security tax on tip income
87	e09900	-4.89	-0.62	-0.20	0.06	0.05	-3.20	*.9	Penalty tax on IRA
88	e17500	-8.18	-2.59	-0.96	-0.33	-0.09	2.47	*.9	Medical and dental expenses
89	e18400	-5.19	-1.70	-0.68	-0.27	-0.12	0.72	*.9	State and local taxes
90	e18500	-4.46	-1.26	-0.55	-0.15	-0.12	0.75	*.9	Real estate tax deductions
91	e19200	-4.62	-1.54	-0.72	-0.27	-0.22	0.76	*.9	Total interest paid
92	e19800	-5.31	-1.72	-0.81	-0.30	-0.22	1.28	*.9	Cash contributions
93	e20100	-3.72	-1.25	0.16	-0.21	-0.07	3.63	*.9	Other than cash contributions
94	e20400	-4.79	-1.37	-0.89	-0.19	-0.09	1.42	*.9	Miscellaneous deductions before AGI limitation
95	e22250	-4.62	-2.89	-0.40	-0.15	-0.11	-6.72	*.9	Short-Term Gain/Loss Net of Carryover
96	e23250	-3.77	-0.98	-0.31	0.01	-0.02	-1.85	*.9	Long term gains less losses
97	e24518	-7.33	-2.79	-0.44	-0.52	0.01	-13.42	*.9	28% Rate Gain or Loss
98	e32800	-18.43	2.49	2.01	-0.10	0.97	1.85	*.9	Child Care Qualifying Expenses
99	e58990	-22.62	-11.04	-6.51	-2.22	-0.33	-8.80	*.9	Investment income (Form 4952 part 2 line 4g)
100	e62900	-39.61	-23.47	-9.48	-1.99	0.21	8.33	*.9	Alternative tax foreign tax credit
101	e87521	-6.75	-7.00	0.47	0.63	-0.14	1.77	*.9	American Opportunity Credit
102	e87530	5.90	-4.64	-0.04	-0.80	-0.01	3.44	*.9	Lifetime Learning Total Qualified Expenses

Table 4. Summary Quantiles

Summary of Masked Data Comparisons
Quantiles of absolute value of percentage deviation of revenue score

	10th Percentile	25th Percentile (Lower Quartile)	50th Percentile (Median)	75th Percentile (Upper Quartile)	90th Percentile	95th Percentile
	Mean	Mean	Mean	Mean	Mean	Mean
file						
Random Forest	0.24	0.89	2.75	12.19	34.71	162.84
CART Subset	2.07	4.87	9.76	24.12	44.20	184.41
CART	2.76	5.30	9.63	20.71	44.01	199.99
CARTFS	0.49	1.41	3.99	10.64	39.81	168.66
ROUND01	0.23	0.66	3.29	6.14	15.42	26.57
ROUND02	0.05	0.26	1.03	2.89	11.04	21.95
ROUND03	0.07	0.21	0.53	1.18	7.59	17.75
ROUND05	0.07	0.14	0.40	0.94	7.73	23.14
ROUND08	0.03	0.10	0.31	0.97	7.65	23.13
BLUR06	0.00	0.01	0.03	0.09	0.58	1.87
BLUR12	0.01	0.02	0.05	0.17	0.94	1.98
BLUR24	0.01	0.04	0.09	0.25	1.61	4.23
BLUR48	0.02	0.05	0.13	0.32	1.95	11.18