Office of Tax Analysis
Technical Paper 6
June 2015

Re-weighting to Produce State-Level Tax Microsimulation Estimates

Robin Fisher and Emily Y. Lin

RE-WEIGHTING TO PRODUCE STATE-LEVEL TAX MICROSIMULATION ESTIMATES

June 3, 2015

Robin Fisher and Emily Y. Lin

Revenue and distributional effects of a Federal tax law change can vary widely across states and the results are of policy relevance. A number of data files might be used for state-level tax simulation but each file has its deficiencies. For example, the Treasury's Individual Tax Model (ITM) is used to simulate the revenue and distributional impact but its sample is not representative at the state level. Hence, while the file includes a variable for state of residency, the sampling errors of state-level statistics based on this variable are likely to be unacceptably large. The Internal Revenue Service's Individual Returns Transaction File (IRTF) contains the population of individual income tax returns filed for a tax year so its state-level statistics are free of sampling errors. However, the file is unedited, contains random data errors, and more importantly, without substantial edits and modeling, lacks any capacity to project a tax law's effect over the budget window.

In this paper, we use a statistical method to impute the information about each state's tax return filers (e.g., number of filers and the distribution of their income, filing status, and number of dependents) from the population IRTF to the ITM. To the extent that the effect of a tax law change differs across states due to states' differences in the size of the filing population and other tax characteristics, imputing this information to the ITM provides key identification for state-level tax effects. Specifically, we calculate the distributions of filers across states—unconditional and conditional on a set of tax variables—from the IRTF, and impute 52 state weights to each ITM record to generate filer distributions that resemble those calculated from the population file.

By re-weighting ITM records, our method facilitates a straightforward way to estimate state-level revenue and distributional effects of Federal tax law changes because the same ITM calculator developed for analyzing the national effect can be used for state-level microsimulation. We also propose a method to evaluate the model fit of the state weight estimation and apply the state weights to an ITM run to produce state-by-state effects of repealing the alternative minimum tax (AMT). Lastly, we note that the re-weighting technique can be used to address other instances in which the sample in the ITM is not sufficiently representative of a specific population of interest.

Robin Fisher: Office of Tax Analysis, U.S. Department of the Treasury, Washington, DC (Robin.Fisher@treasury.gov)
Emily Y. Lin: Office of Tax Analysis, U.S. Department of the Treasury, Washington, DC (Emily.Lin@treasury.gov)

1. Introduction

Revenue and distributional effects of a proposed or enacted change in Federal tax law can vary widely across states. If every state had the same distribution of taxpayer characteristics that impact tax liabilities (such as income, filing status, and number of dependents) and all other factors for the Federal tax effect were identical across states, a Federal tax law change would result in a proportionally larger change in the Federal tax liability for larger states than for smaller states. To the extent that taxpayer characteristics and other factors differ across states, changes in Federal tax liabilities resulting from a Federal tax change will vary by state for reasons other than differences in state population size. This geographic distribution of Federal tax liability changes can play an important role in lawmakers' decision-making. Despite the policy relevance, due mainly to lack of appropriate data, analyses that show effects of Federal tax law changes by state are scarcely available. Using tax information derived from the population file of Federal individual income tax returns, we impute state weights to the Treasury's Individual Tax Model (ITM) to facilitate state-by-state tax microsimulation analysis. In particular, we re-weight the ITM records to create 52 (for 50 states, the District of Columbia, and other areas) weighted samples that resemble the filing population of each state along certain important tax characteristics.

A number of existing data can potentially be used to provide state-level tax analysis but each data has its own deficiencies. The Treasury Department uses the Individual Tax Model (ITM) to simulate the revenue and distributional impact of Federal tax law changes. The sample of tax return filers in the ITM is based on the Individual and Sole Proprietorship (INSOLE) returns drawn by the Statistics of Income (SOI) Division of the Internal Revenue Service (IRS), and the data are extrapolated and modeled by the Treasury Department's Office of Tax Analysis (OTA) to provide microsimulation estimates of the effects of tax changes over the 10-year period of the budget window. Although tax return data include a variable indicating a taxpayer's state of residency, the INSOLE sample is not randomly drawn across states and is not intended to be representative at the state level. Hence, while state-level tax statistics can be made available on the ITM based on the taxpayer's state of residency shown on the return, the sampling errors of some of these

statistics are likely to be unacceptably large, substantially reducing the precision of the point estimates.

Alternatively, state-level tax statistics can be directly obtained from the IRS' Individual Returns Transaction File (IRTF), which contains the population of Federal individual income tax returns processed by the IRS each year. The advantage of using the population file is elimination of sampling errors because it includes all returns, rather than a sample of returns, filed for a tax year. However, unlike the extensively edited INSOLE file, this transaction file is unedited and is known to include random data errors. More importantly, without any of the extrapolation, imputation, and modeling of the data, as in the ITM, the file provides historical statistics by state but has a limited capacity to project the state-by-state effect of a Federal tax law change that will occur during the budget window.

Another method in common use is to treat the states as the unit of measurement and use a model together with auxiliary information to find estimators with improved characteristics. Well-known examples include the Small Area Income and Poverty Estimates (SAIPE) and Small Area Health Insurance Estimates (SAHIE) by the Bureau of the Census.[1] Henry et al. (2007) applied the same method to the current situation. While these methods can produce state-level estimates with good properties, they are labor-intensive and can produce estimates only for a small set of variables, since each variable for which small-area estimates are desired must be explicitly included in the modeling. It is also not easy to see how to use these estimates for microsimulation problems like those encountered in the OTA; in general, the microsimulation would necessarily be run first, and then the small area model parameters would be estimated.

One way to form state-level microsimulation estimates is to use the ITM for its edited data, extrapolation, imputation and modeling advantage, coupled with the information obtained from the IRTF about the filing population for each state. From the population IRTF, we know the unconditional distribution of filers across states (e.g., X

[1] Information about SAIPE and SAHIE estimates can be found at http://www.census.gov/did/www/saipe/ and http://www.census.gov/did/www/sahie/.

percent of all filers live in state Y) and the distribution of filers conditional on observable tax characteristics, such as filing status, income classes, or number of dependents (e.g., Z percent of the nation's married-file-jointly couples who have income below $50,000 and two children live in state Y). To the extent that the effect of a proposal or law change on tax liability differs by the tax characteristics but is expected to be uniform among families with the same set of characteristics, we could use the ITM to estimate the projected total liability effect for each "cell" of family type partitioned by the selected characteristics, and then allocate the cell's aggregate effect across states based on the conditional distribution of filers we tabulate from the population file. By modeling on the ITM while utilizing geographic distributions of filers and their tax characteristics from the population file, this method addresses the issues that the ITM sub-samples are not representative of state populations and that the population file lacks certain modeling capacity.

Consider a tax cut for which the generosity increases with the tax unit's number of exemptions and decreases with the tax unit's adjusted gross income (AGI). We can model the tax cut on the ITM to calculate the amount of tax change for each family type, defined by the number of exemptions and AGI category. Note that the amount of tax cut varies by family type but is the same magnitude for tax units of the same type. From the ITM, we know the weighted count of tax units for each family type. From the IRTF, we know the number of tax units of each family type by state. With this information, we use the ITM to calculate the aggregate tax liability change for a family type and allocate the amount across states based on the conditional distribution of tax units across states we obtained from the IRTF tabulation. We then sum up the results across family types within a state to arrive at the state's total liability change.

In this paper, we use a statistical method to bring over the information about the distribution of tax units across states—unconditional as well as conditional on observed tax characteristics—from the IRTF to the ITM by splitting the weight of each ITM record into 52 state weights. The goal is to make the 52 re-weighted ITM samples to generate the conditional and unconditional distributions of tax units that resemble those calculated from the population file. The statistical method employed in this paper is closely related to that

described in Schirm et al. (1999) and Schirm and Zaslavsky (1998), which is an elaboration of the informal tabulation described in the previous paragraph. These two papers describe the utility of sample weight adjustments, especially in the microsimulation setting.

The new state weights generate 52 state samples in the ITM. Because we expect the liability effect of a tax change to vary by family type (defined by the selected tax characteristics), state-level microsimulation can be performed on the ITM with reduced sampling errors once the ITM's state samples have the population-like distributions of tax characteristics across states. Consequently, we use the same ITM tax calculator developed for analyzing the national effect and weight the sample with the newly imputed 52 state weights to simulate the effect of Federal tax changes state by state.

This paper is organized as follows. Section 2 describes the data files, the tax characteristics included in estimation, and the properties we expect the estimates to have. Section 3 describes the estimation model, with details provided in the Appendix. Section 4 presents the estimation results and verifies that the estimates generally have the desired properties. Section 5 shows state-by-state tax liability estimates (without behavioral change) resulting from repeal of the alternative minimum tax (AMT) as simulated by the re-weighted scheme on the ITM. Section 6 provides concluding remarks.

2. Data

The Individual Returns Transaction File (IRTF) contains records of all individual income tax returns filed for a given tax year. There is a variable for most (but not all) of the lines on the form 1040 and any schedules that may be attached. We use a set of X-variables (listed below) from the IRTF for tax year 2007 to estimate the probability a tax return was filed in a state *st* given the observed values of a set of those variables on that return. We denote this probability $p(ST = st|X)$. There are 52 "states" in our model: 50 states, the District of Columbia, and other areas. We rely on this population file to form the target distributions, including the distribution of returns across states and the distributions conditional on the selected tax variables. The variables we chose at this stage are

AGI: Adjusted Gross Income, categorized as in Table 1

EX: Number of exemptions, 1, 2, or 3 or more

AGE: Primary filer's age, grouped as in Table 2

FamType: Family type, recoded from filing status; see Table 3

SCHA: An indicator for filing Schedule A

SCHB: An indicator for filing Schedule B

STLCINTX: State and local income tax deduction, grouped as in Table 4

realEstTax: Real estate property tax deduction, grouped as in Table 5

Mort: Home mortgage interest deduction, grouped as in Table 6

The **AGE** variable is obtained from Social Security records, which include date of birth. We merge that variable onto the IRTF.

Each of the continuous variables in this list is converted to a discrete variable by grouping. We assume that the grouped variables are sufficient in the model. If this assumption is violated, information about the variables is lost and the estimates lose precision. The alternative is to treat some of the variables as continuous, but modeling the continuous distribution is a more complicated task.

For example, it may seem reasonable to use the continuous version of the X-variables as predictors in a multinomial logit model, for example, and estimate $log(\frac{P(ST=st|X)}{P(ST=Alabama|X)}) = f(X,\theta) + \varepsilon$, or some variation of that. In this case it is necessary to model the function *f* as well as the distribution of ε. We know nothing about *f*, however, and by representing the variables as continuous, we introduce more computational complexity. The presence of large outliers in the IRTF also makes statistical robustness a consideration. Since the variables are grouped into categories, even a large data error only changes an observation's category and thereby contributes to a relatively small model error.

Let *Y* denote other variables for which we wish to have state-level estimates but that are not used to calculate state conditional probabilities. We assume that the relationship between *Y* and the state probability is fully explained by variables *X* included in the model. Examples of *Y* include the child tax credit (CTC) or earned income tax credit (EITC) claimants or amounts. Violation of this assumption leads to a loss of precision.

In our application, we take the X-variables described above and tabulate them by the generating classes described in the Appendix. Once these tables are produced, we are finished with the IRTF for estimation purposes.

Table 1 Grouping Categories for AGI

Category label	Range
1	(−∞, 0]
2	(0,10000)
3	(10000,20000]
4	(20000,30000]
5	(30000,50000]
6	(50000,75000]
7	(75000,100000]
8	(100000,150000]
9	(150000,200000]
10	(200000,500000]
11	(500000,1000000]
12	(1000000,2000000]
13	(2000000,5000000]
14	(5000000,∞]

Table 2 Grouping Categories for Age

Category label	Range
1	(0,32)
2	(32,50]
3	(50,65]
4	(65,∞]

Table 3 Recode of Filing Status for Family Type

Category Label	Filing Status
1	Single
2	Head of Household
3	Married Filing Jointly
4	Married Filing Separately

Table 4 Recode of State and Local Income Taxes

Category Label	Range
1	0
2	(0,1522]
3	(1522,2672]
4	(2672,4402]
5	(4402,∞]

Table 5 Recode of Real Estate Property Taxes

Category Label	Range
1	0
2	(0,1232]
3	(1232,9407]
4	(9407,∞]

Table 6 Recode of Mortgage Deduction

Category Label	Range
1	0
2	(0,6382]
3	(6382,9150]
4	(9150,13316]
5	(13316,∞]

The Individual and Sole Proprietorship (INSOLE) file is a stratified Bernoulli sample of individual income tax returns drawn from the IRTF. The IRS produces the INSOLE file for each tax year. Ratio adjustments are applied to the file to enforce consistency of stratum totals with known population totals. The stratification and sampling rates in the INSOLE are designed so that the highest income (and other 'high-interest') returns are sampled at the highest rate; some high-interest returns are sampled with

certainty. Although the file contains a variable indicating a taxpayer's state of residence, the sample is not randomly drawn across states and its sub-samples are not intended to be representative of state populations.

Substantial effort has been made by the IRS to edit the INSOLE files so each record has some internal consistency and obviously erroneous fields are corrected. The editing eliminates occasionally large errors found in the IRTF that can affect statistical estimates. In our application we use the 2007 INSOLE but exclude the subset of records identified by the IRS as being filed only to receive economic stimulus payments.

The X-variables on the INSOLE are grouped into categories in the same manner as described above for the IRTF. Our goal is to estimate 52 state weights (for 50 states, DC and other areas) for each INSOLE record so that, when applying these weights to the INSOLE file, the resulting distributions of tax returns across states, unconditional or conditional on certain tax variables, resemble the target distributions we calculated from the IRTF. Because state-level statistics should add up to national totals in the INSOLE, our method, in practice, splits the INSOLE weight for each record into 52 state weights.

The Treasury Department's Individual Tax Model (ITM) is used to simulate the revenue and distributional impact of tax law changes. The ITM is based on the IRS' INSOLE file and is extrapolated by the Treasury's Office of Tax Analysis (OTA) to meet targeted population and income growth, among other trends, over the 10-year period in the budget window. The ITM was based on the 2007 INSOLE file when the current project started. The ITM sub-samples, just like the INSOLE, are not intended to be representative of state populations. Furthermore, the ITM contains non-filing tax units outside of the INSOLE sample. Some X-variables for non-filing tax units are based on information returns filed by third-parties to the IRS.

We merge taxpayers' state weights estimated on the INSOLE onto the ITM. As mentioned, the filing units in the ITM are based on the INSOLE sample so we bring over state weights estimated for each INSOLE record to the same tax unit in the ITM. We also impute state weights to non-filing tax units in the ITM (see the Model section). To estimate

state-by-state effects of tax law changes, we run the ITM as we do for national-level effects but use the imputed state weights (52 loops), instead of the file's original weights, to produce 52 state-level outcomes.

Once state weights are merged on to the ITM, we allow the weights to grow proportionally with the extrapolation of ITM weights for each tax unit over the budget window relative to the tax unit's weight in the base year 2007. The evaluations below involve comparisons between state-level estimates based on the re-weighting method on the 2007 INSOLE and the corresponding state statistics calculated from the 2007 IRTF by the IRS and published in Statistics of Income (SOI) Tax Stats (2008). Alternatively, one can compare the re-weighted state-level estimates from the ITM with the published IRTF statistics for tax year 2012, the most recent year for which the publication is available. To prevent confounding the effect of ITM edits and extrapolation on state-level estimates with model errors in the re-weighting estimation, we chose to evaluate state-level estimates produced by the re-weighted INSOLE, rather than by the re-weighted ITM, against published IRTF statistics.

3. Model

Let ST represent the random variable *state*, X a set of variables present in both the IRTF and the INSOLE, and Y a set of random variables in the INSOLE. Lower case letters denote specific values of the random variables. We assume $ST \perp Y \mid X$ (which we read as "ST is independent of Y, conditioned on X" and which implies $p(st, y|x) = p(st|x)p(y|x)$); we further assume that $p(x) > 0$, so

$$p(st, x, y) = \frac{p(x,y)p(st,x)}{p(x)} \quad (1)$$

$$= p(st|x)p(x, y). \quad (2)$$

Since the conditional probability of returns across states, $p(st|x)$, can be written as

$$p(st|x) = \frac{p(x|st)p(st)}{p(x)} \quad (3)$$

and $p(x, y)$ is given by the INSOLE weights and not subject to manipulation, by assumption, our estimation task is equivalent to estimating $p(x|st)$, which indicates the distribution of x in a given state, and $p(st)$, which indicates the distribution of returns across states. The latter is easy; we calculate it directly from the IRTF. A description of the methodology for estimating $p(x|st)$ is given in the Appendix.

Combining (3) with INSOLE weights,

$$N\hat{p}(st, x, y) = \frac{p(x|st)p(st)w_i}{\sum_{st} p(x|st)p(st)}$$

$$= \hat{p}(st|x)w_i.$$

where N is the population count and w_i is the INSOLE weight for observation *i*. Thus the new weight for observation *i* in state *st* is

$$w_{i,st} = \hat{p}(st|x)w_i.$$

The new weight $w_{i,st}$ is just the INSOLE weight, split into smaller shares. Those shares are proportional to the estimated proportion of units that are in state st among all returns that are similar to the i^{th} return, in the sense that $X = x_i$ for those returns.

3.1 Non-filing Tax Units

We further assume that $p(st|x)$is the same for non-filing units as it is for filing units. Evaluating this assumption and the quality of the estimator for non-filers is the subject of ongoing research.

4. Evaluation on State-Level Estimates from the INSOLE

Using the state weights, we produce state-level estimates for several tax variables from the INSOLE and compare them to the corresponding set of estimates calculated by the IRS based on the IRTF and published in SOI Tax Stats (2008). Figures 1 through 9 present scatterplots of estimates based on the method in this paper (the *re-weighting*

method) versus the corresponding statistics from the IRTF published by the SOI. Recall that our estimation goal is to let the state weight estimators, when applied to the INSOLE records, produce similar cross-state distributions (unconditional or conditional) to those observed from the IRTF population file. Each dot on the graph represents a state, with X denoting the value calculated in the IRTF and Y denoting the value calculated from the INSOLE using our state weight estimators. Under a perfect fit, all dots would lie on the 45-degree line.

In addition to using scatterplots for visual evaluation, we also measure the model performance by calculating the correlation, as well as relative differences, between the two statistics in each scatterplot. Figure 10 contains a table of the correlation coefficients, estimated Coefficients of Variation (CV's), and Mean Absolute Relative Difference (MARD) for the same set of variables.

The correlation coefficients, *c*, is the usual one, defined on the linear scale, though the plots are on the *ablog* scale, where

$$ablog(c) = sgn(c)\log_{10}(|c| + 1).$$

This transformation has many of the properties of the $\log_{10}$ function, but it is continuous at zero and defined for (and preserves the sign of) negative numbers. This transformation is useful for visualizations of data from distributions with a long tail while supporting those that are not subsets of the positive real numbers. The CV's are estimated with

$$\widehat{cv} = \left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\tilde{y}_i - y_i}{y_i}\right)^2\right)^{\frac{1}{2}},$$

where n is the number of cells in the relevant table, y_i is the population proportion in that cell, and $\tilde{y}_i$ is its estimate formed from the re-weighting method. The MARD is

$$MARD = \frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{y}_i - y_i|}{y_i}.$$

The results presented in Figures 1 through 9 appear acceptable with the possible exception of the results for *other areas* and the Real Estate Tax variable. It is confirmed in Figure 10 that the correlation on the linear scale is low and the MARD is high for the Real Estate Tax, relative to the other variables. There are a number of possible reasons for these results. On the Real Estate Tax variable, it is possible that the categories are too wide when we group the variable and therefore the within-category variance is large, causing the excess variance in the estimates. For both *other areas* and Real Estate Tax, it is possible that the measurement differences between the INSOLE and the IRTF (particularly edits to the INSOLE) are large enough to matter.

5. State-Level Microsimulation from the ITM

In this section, we apply state weights to an ITM run to produce state-by-state tax liability effects of repealing the AMT. Note that neither the AMT status nor the amount is included in the X variables in the re-weighting estimation. However, because certain deductions and credits are preferences in the AMT, we expect that the AMT effect is explained in the model by taxpayer income and by itemized deductions for state and local taxes, the deductions of which are not allowed in the AMT.

The ITM predicts that, without behavioral changes, 4.35 million non-dependent taxpayers would be affected by a repeal of the AMT under 2016 law and income for a total decrease in the Federal individual income tax liability by $29.06 billion.[2] This liability effect takes place either through an elimination of AMT payments or increases in the tax credits lost under the AMT. The effect is unequally distributed across states. About 17 percent of the total liability effect resulting from the AMT repeal is attributable to California while another 11 percent and 7 percent are attributable to New York and New Jersey, respectively. Other states that are largely affected by the AMT include Illinois,

[2] This model run is performed on the version of the ITM based on the 2010 INSOLE file and FY2016 winter budget extrapolation.

Texas, Florida, Pennsylvania and Massachusetts, each of which contributes to at least 4 percent of the total liability effect. In contrast, the bottom 25 states (including the District of Columbia and Other Areas) combined make up just under 10 percent of the total AMT effect.

We divide each state's AMT liability effect by its total Federal individual income tax liability after refundable tax credits to account for the state's percentage change in tax liability from eliminating the AMT. According to this percent change, states are classified into four roughly equal sized groups. Figure 11 depicts the results. Within the top and bottom groups of states, 5 states are predicted to have their Federal individual income tax liability declined by more than 2.2 percent in 2016 with an AMT repeal, including New Jersey (with the greatest effect, 3.0 percent), California, Vermont, Maryland and Illinois, and 6 states below 1.2 percent, including Alaska (with the smallest effect, 0.7 percent), Wyoming, Nevada, Tennessee, South Dakota and Texas.

6. Conclusion

With the re-weighting method described in this paper, micro-simulations for tax proposals are conducted on the ITM in a straightforward way to produce state-by-state estimates. Based on this application, OTA estimated the number of workers by state who would benefit from the proposed EITC expansion for childless workers in the President's FY2015 Budget. The estimation results are released in a joint paper by the Executive Office of the President and U.S. Treasury (2014).

It is worth noting that, while we implement the re-weighting method to perform state-by-state analysis, the same method can be applied for other small domains whether geographic or not. For example, same-sex married couples were not allowed to file as married filing jointly in 2007. To project the tax consequence of joint filing for same-sex married couples, we re-weighted married-filing-jointly couples in the 2007 INSOLE-based ITM to make the key tax characteristics of these joint filers distributed similarly to those of same-sex married couples observed in the survey data in 2007. As a result, the tax liability of same-sex married couples had they filed jointly in 2007 can be inferred from the re-

weighted tax liability of married couples in the 2007 ITM.

The key assumption in this method is the conditional independence assumption. The results will still be valid if most, if not all, of the dependence between *Y* and *ST* is explained by *X*. One counter example, which gives rise to limitations in using the state weights, is when *Y* depends specifically on state policies, state-specific characteristics or state economic conditions whose effects are not captured by *X*. When using the state weights to produce state-level simulation estimates, one should exercise judgment on whether the Y variable(s) of interest, or the tax law's effect, can be reasonably explained by the X variables included in the estimation.

It is natural to consider extending the model by adding variables to the *X*-vector. There is trade-off, however. Specifically, *P(Y, X)* is estimated from the INSOLE sample, which has a limited sample size. As an illustration, consider the extreme case where *X* has *ST* as an element. Then *P(Y|X)* becomes *P(Y|ST).* As the INSOLE sample is not designed for state-level estimation, estimates of this density will be based on unsuitably small samples and it follows that estimates based on this density are likely to have unacceptably large variances. The problem persists for any *X* where the cells of the table defined by *X* are too small or, more generally, where the degrees of freedom in some sub-table are too small. Given the trade-off between model fit and variance, upon evaluating the re-weighting model, one should consider which *Y* variables are relevant and what level of error renders an estimate unacceptable.

References

Bishop, Yvonne M., Stephen E. Fienberg, and Paul W. Holland (1974). *Discrete Multivariate Analysis—Theory and Practice*, Cambridge, MA: MIT Press. Reprinted by Springer, 2007.

Cowell, Robert G., Philip Dawid, Steffen L. Lauritzen, and David J. Spiegelhalter (2003). *Probabilistic Networks and Expert Systems*. New York: Springer.

Executive Office of the President and U.S. Treasury Department (2014). *The President's Proposal to Expand the Earned Income Tax Credit*. *https://www.whitehouse.gov/sites/default/files/docs/eitc_report.pdf* .

Hojsgaard, Soren (2012). *gRim: Graphical Interaction Models, R Package Version 0.1-14*. http://CRAN.R-project.org/package=gRim.

Henry, Kimberly, Partha Lahiri, and Robin Fisher. "Using the Statistics of Income Division's Sample Data to Reduce Measurement and Processing Error in Small-Area Estimates Produced from Administrative Tax Records," *2007 Proceedings of the American Statistical Association: Section on Survey Research Methods*. 2007.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, http://www.R-project.org/.

Schirm, Allen L., Alan M. Zaslavsky, and John L. Czajka. "Large Numbers of Estimates for Small Areas," paper in 1999 Federal Committee on Statistical Methodology (FCSM) Research Conference. 1999. www.fcsm.gov/99papers/schirm.pdf, downloaded November, 2010.

Schirm, Allen L. and Alan Zaslavsky. "Model-Based Microsimulation Estimates for States When State Programs Vary," *1998 Proceedings of American Statistical Association*: *Section on Survey Research Methods*. 1998.

Statistics of Income, Internal Revenue Service. SOI Tax Stats- Historic Table 2. 2008. http://www.irs.gov/uac/SOI-Tax-Stats-Historic-Table-2, downloaded November 22, 2010.

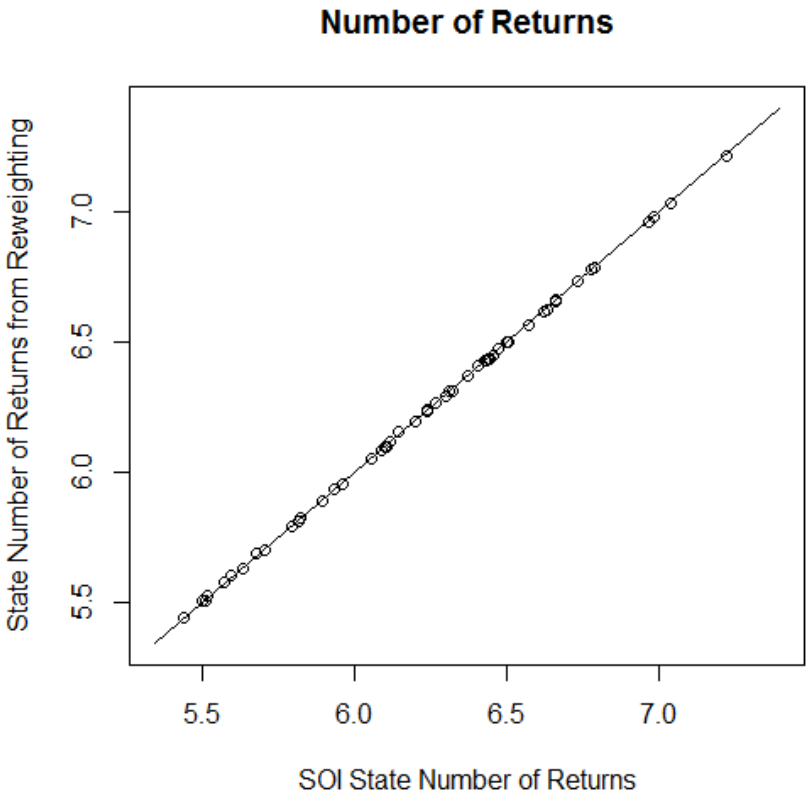

Figure 1: Total numbers of returns from reweighting method versus tabulations in the IRS publications. Log_{10} scale.

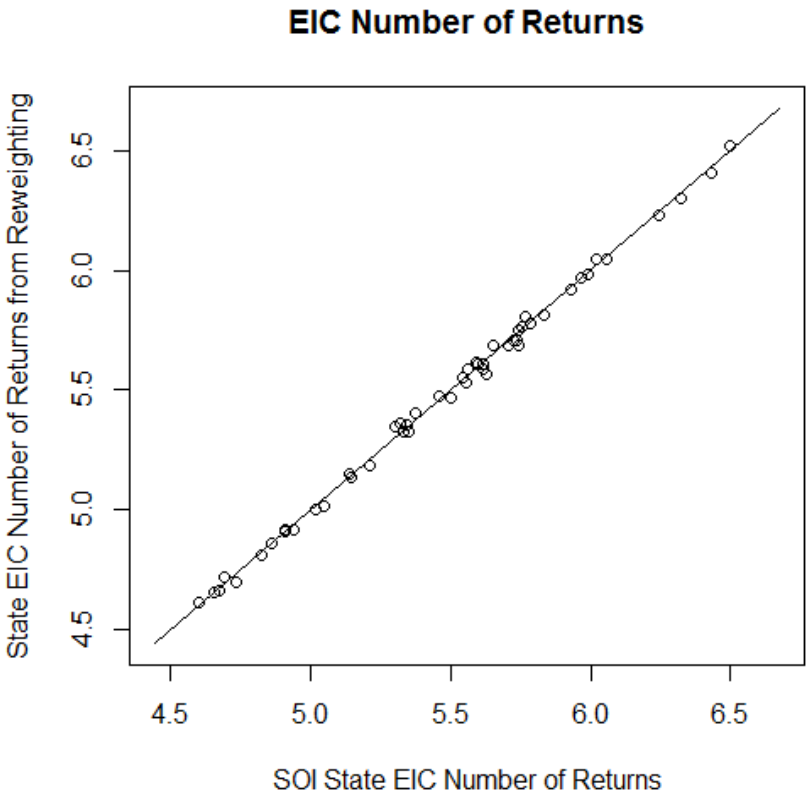

Figure 2: Numbers of Returns with EITC from reweighting method versus tabulations in the IRS publications. Log_{10} scale.

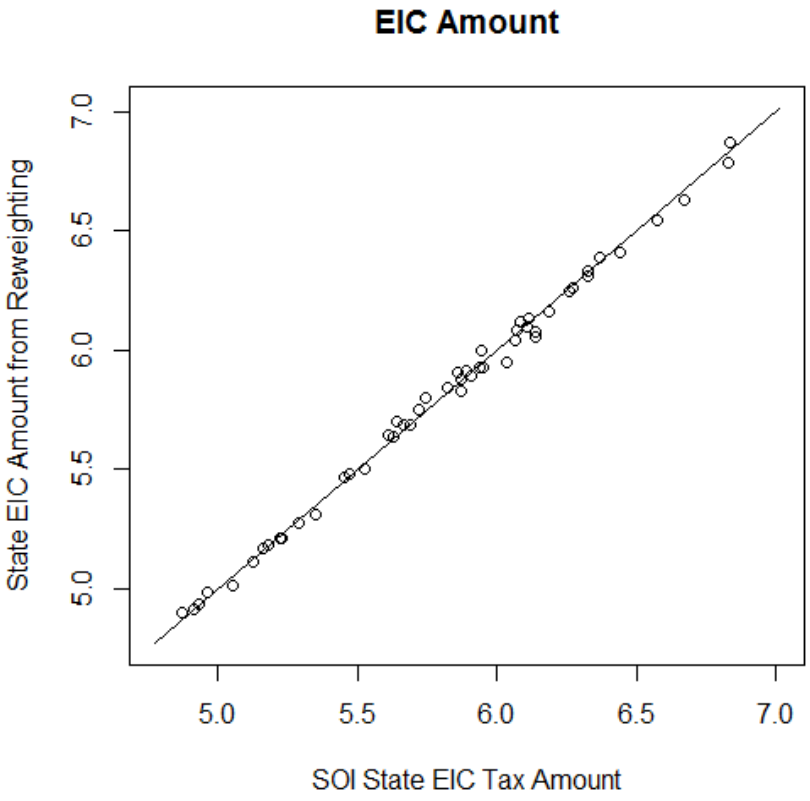

Figure 3: EITC payments from reweighting method versus tabulations in the IRS publications. Log_{10} scale.

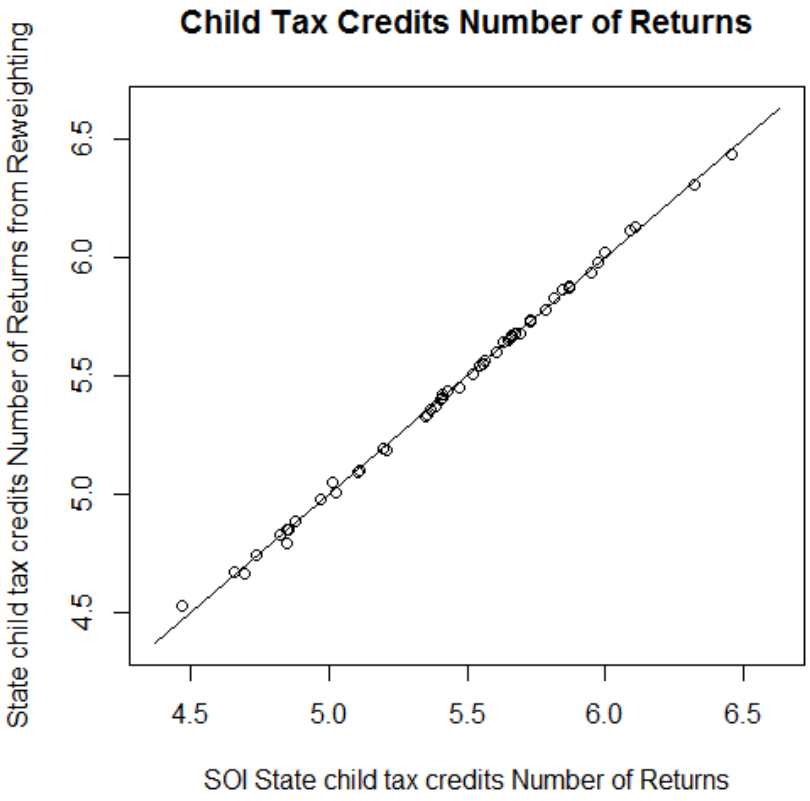

Figure 4: Numbers of returns with Child Tax Credits from reweighting method and from published tabulations from the IRTF. Log_{10} scale.

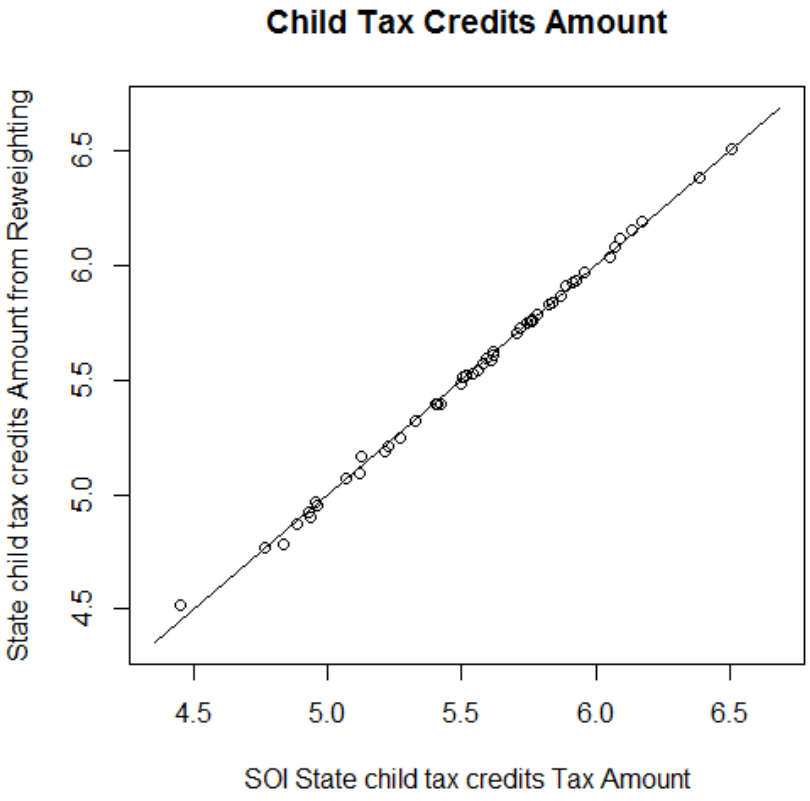

Figure 5: CTC from reweighting method versus tabulations in the IRS publications. Log_{10} scale.

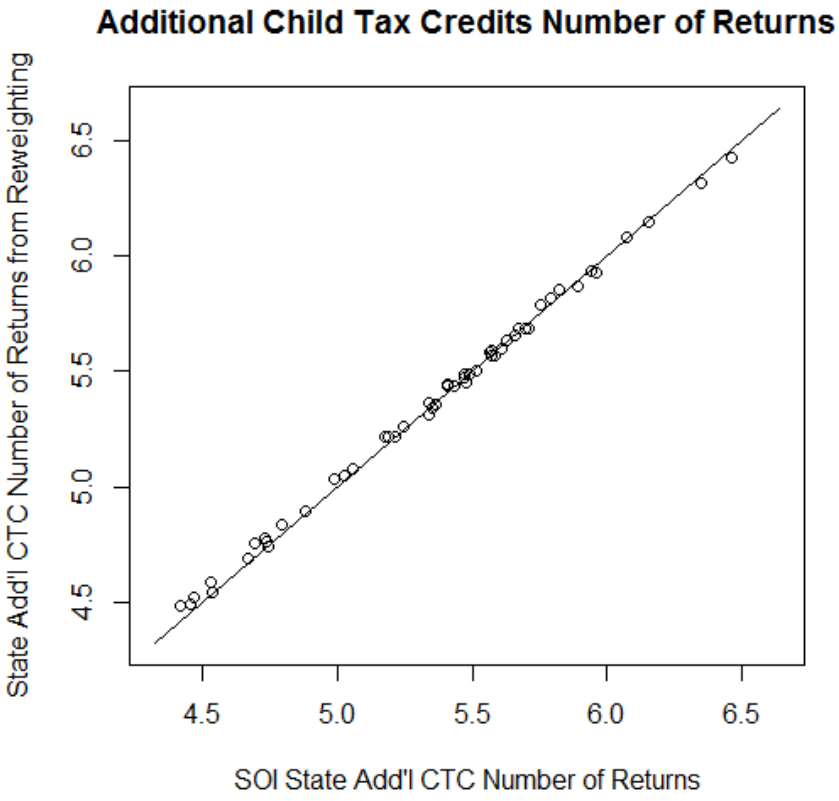

Figure 6: Numbers of returns with Additional Child Tax Credits from rewighting method and from published tabulations from the IRTF. Log_{10} scale.

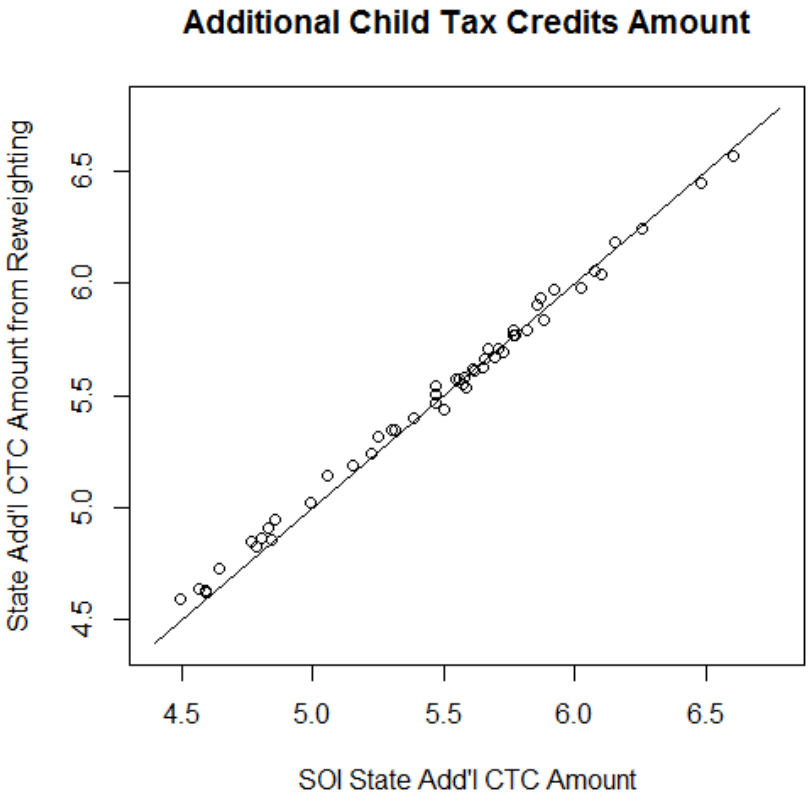

Figure 7: Additional CTC from reweighting method versus tabulations in the IRS publications. Log_{10} scale.

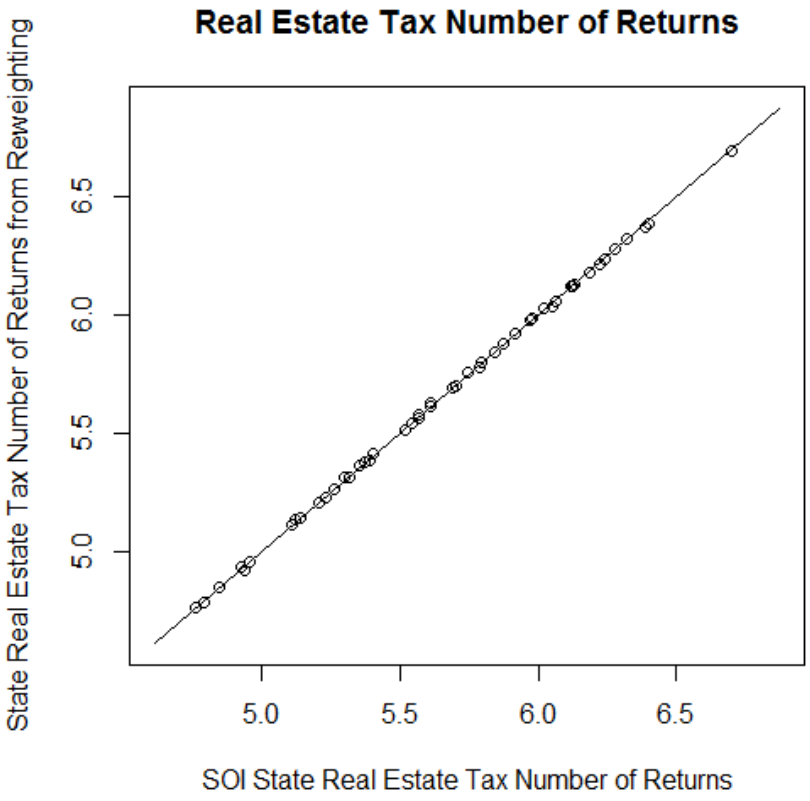

Figure 8: Numbers of returns with real estate tax property tax deductions from reweighting method versus tabulations in the IRS publications. Log_{10} scale.

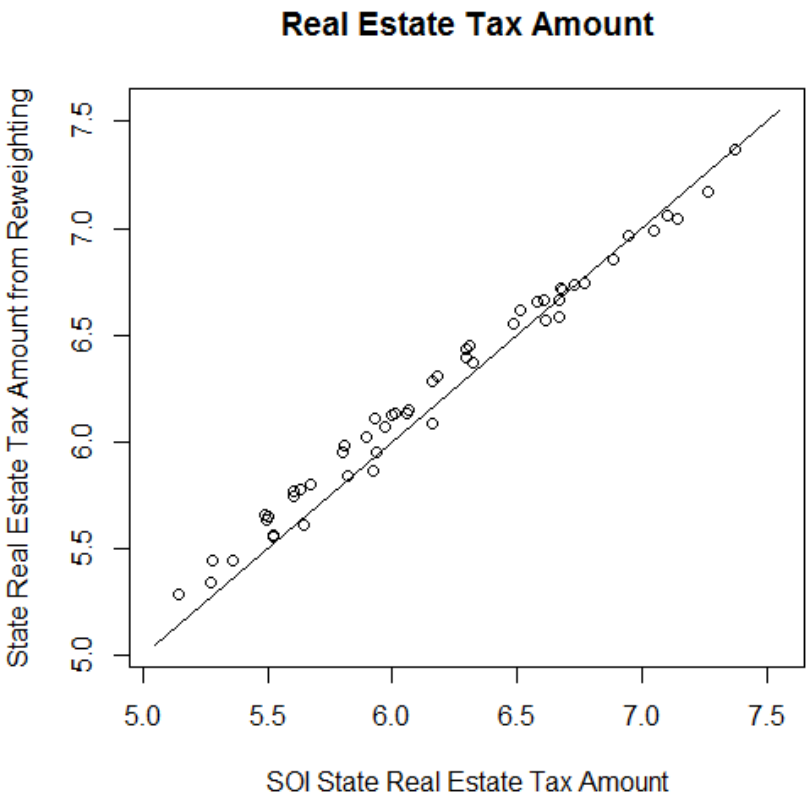

Figure 9: Amounts of real estate tax property tax deductions from reweighting method versus tabulations in the IRS publications. Log_{10} scale

Variable	Correlation, Linear Scale	Estimated CV	Mean Absolute Relative Difference
All Returns	0.999	0.026	0.020
EITC Number of Returns	0.998	0.051	0.039
EITC Amount	0.996	0.076	0.057
Mortgage Interest Deduction	0.999	0.093	0.076
Amount CTC Number of Returns	0.999	0.038	0.026
CTC Amount	0.998	0.042	0.028
Additional CTC Number of Returns	0.999	0.082	0.065
Additional CTC Amount	0.998	0.120	0.094
Number of Returns with Real Estate	0.997	0.047	0.038
Tax Amount of Real Estate Tax	0.986	0.26	0.22

Figure 10: Correlation and Mean Absolute Relative Differences between Re-weighting Estimates and Published Estimates without ”Other Areas.”

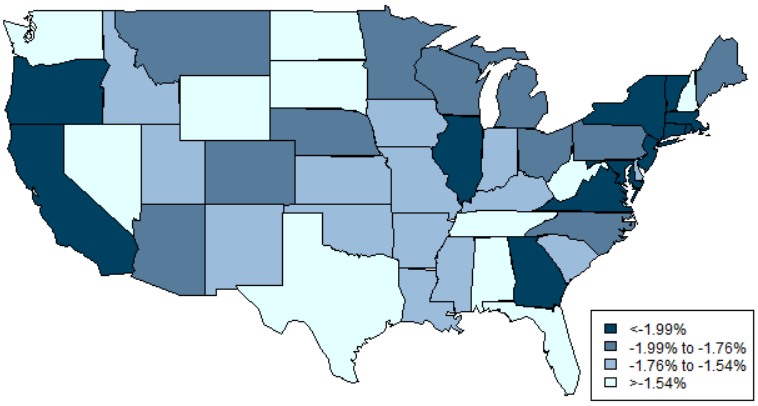

Figure 11: Liability Change from Repeal of the AMT

Appendix

Since the variables in the model are discrete, either naturally or by grouping, we represent the joint conditional distribution of the variables as a *table*. The table describing the joint distributions of the variables and *state* is potentially very large so our model potentially has a large number of parameters. To decrease this number, we search for parsimonious models for the relationships between the variables. In particular, we model independence relationships which take the form $p(X,Y|Z) = p(X|Z)p(Y|Z)$ for some variables $\{X,Y,Z\}$; we denote this relationship $X \perp Y \mid$ Z. There is an intuitive and rigorous method to map the independence assumptions in a model, representing the model as an undirected graph *G*=(*V*,*E*), where *V* is the set of *vertices* (or *nodes*) in *G*, and *E* is the set of *edges* connecting pairs of vertices. Each node in the graph represents a variable in the model and the edges are defined so that the absence of an edge between nodes, say A and B, represents the assumption that *A* is independent of *B* given all of the other variables, $V\backslash\{A,B\}$. That is, $p(A,B|V\backslash\{A,B\}) = p(A|V\backslash\{A,B\})p(B|V\backslash\{A,B\})$ or $A \perp$ B$|V\backslash\{A,B\}$. Such models are called *Markov Random Fields* or, alternately *Undirected Graphical Models*, and there is a substantial literature on their theoretical properties, and many efficient algorithms have been designed to take advantage of those properties (for example, see Cowell et al. (2003)). We are particularly interested in tools to search for simple models for the dependencies in these models and, generally, to reason about the joint distribution of the tax variables in our datasets. .

For an example, consider the graph in Figure A1. Here X has four components, and the graph corresponds to the model where

$$p(x_1, x_2, x_{3,} x_4) = \frac{p(x_1,x_2,x_3)p(x_1,x_2,x_4)}{p(x_1x_2)} . \qquad \text{(A1)}$$

Under this model, the original four dimensional table has been replaced with two smaller tables. Each table corresponds to a fully connected subgraph in the graph. Reductions like this can reduce the total number of parameters dramatically. Formally, the number of parameters increases asymptotically exponentially in the dimension of the largest table, so reducing the size of the largest table reduces the complexity, even if there are many such tables as a result. The *generating class* of this model is

[[X1,X2,X3],[X1,X2,X4]],

using notation like that in Bishop et al. (2007) (also see the reference for a detailed discussion of the interpretation of generating classes). The variables $S = (X_1, X_2)$ separate X_3 from X_4 in the graph and in the sense that $X_3 \perp X_4 \mid S$; S is therefore sometimes called a *separator*. The members of each set in the generating class are all connected to each other in the graph; in the graph theory literature, these sets are called *cliques*. They correspond to the largest tables we need to estimate. Equation (A1) can be generalized: for a set C of cliques with a set S of seperators,

$$p(\underline{x}) = \frac{\prod_{c \epsilon C} p(\underline{x}_c)}{\prod_{s \epsilon S} p(\underline{x}_s)}, \tag{A2}$$

where we abuse notation a little to overload $p(.)$ so it is the density of its arguments, whatever they are. The maximum likelihood (ML) estimate of $p(\underline{x})$ is just

$$\hat{p}(\underline{x}) = \frac{\prod_{c \epsilon C} \hat{p}(\underline{x}_c)}{\prod_{s \epsilon S} \hat{p}(\underline{x}_s)},$$

where the hat indicates the ML estimate, and, recall, the ML estimator for a saturated table is just the table of empirical proportions. This simplifies the task enormously, and eliminates the need for iterative methods like Iterative Proportional Fitting (see Cowell et al. (2003) for details).

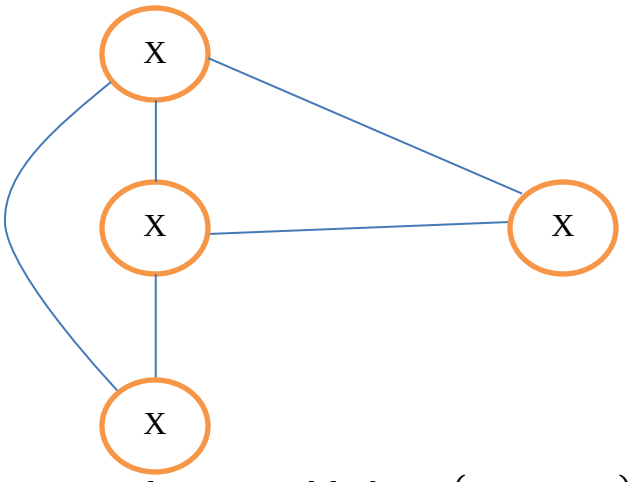

Figure A1. A graph corresponding to a model where $p(x_1, x_2, x_3, x_4) = \frac{p(x_1,x_2,x_3)p(x_1,x_2,x_4)}{p(x_1x_2)}$

In our application we partition the set of variables associated with each tax unit into three parts. The first part contains the state (or DC or Other Areas) of residence, which we denote *ST* , *ST* ∈{1, ..., 52}. The second and third parts contain set of variables *X* and *Y* for which we make the assumption that $ST \perp Y \mid X$, where *Y* are the remaining variables of interest in the INSOLE. That is, ST is independent of Y, conditioned on X. Denote the density (relative to the appropriate measure) of that distribution by $p(st, x, y) = p(ST = st, X = x, Y = y)$. To set the context, *ST* and *X* are variables available from the IRTF while *X* and *Y* are present on the INSOLE/ITM and we wish to construct weights on the INSOLE/ITM so that we can estimate state-level summaries of functions of (X, Y). For example, *X* may be *AGI*, *total exemptions*, and *state and local income tax*, and *Y* may be *taxes paid.* Then the assumption is that *taxes paid* is independent of ST given *AGI*, *total exemptions*, and *state and local income tax*. Put another way, *ST* adds no information about *taxes paid*, once *AGI*, *total exemptions*, and *state and local income tax* are observed. This illustrative example is, of course, unrealistic; the modeling exercise involves making good choices for *X* so the model holds well for a useful set of choices of Y.

These assumptions imply, assuming $p(x) > 0$,

$$p(st, x, y) = \frac{p(x,y)p(st,x)}{p(x)} \tag{A3}$$

$$= p(st|x)p(x,y). \quad \text{(A4)}$$

Since the conditional probability of returns across states, $p(st|x)$, can be written as

$$p(st|x) = \frac{p(x|st)p(st)}{p(x)},$$

and $p(x,y)$ is given by the INSOLE weights and is not subject to manipulation, our estimation task is equivalent to estimating $p(x|st)$, which indicates the distribution of *x* in a given state, and $p(st)$, which indicates the distribution of returns across states. The latter is easy; we calculate it directly from the IRTF data.

Note in equation (A3) that we group the denominator with the rightmost factor in the numerator, so information from the INSOLE informs the inference about *ST*, and the joint distribution preserves the marginal distribution $P(X,Y)$, which is the empirical distribution in the INSOLE. If we had grouped the denominator with the leftmost factor in the numerator, information would have propagated the other way, and the method would have been a post-stratification on states with raking, controlling to population estimates of $ST * X$ totals.

We assume the dependence structure of $X|ST$ can be represented by a graphical models as we have described. This is a weak assumption; it is implied by the generalized linear model (GLM). We also assume it is decomposable, which is a stronger assumption. See Bishop et al. (2007) and Cowell et al. (2003) for the implications. These models are all members of the class of log linear models. Finally, we also assume the Markov graph of $X|ST$ is the same for every value of *st* and that we can estimate that graph from a sample from the nation. This implies that, for the purpose of fitting the model structure, there are no aggregation issues.

We estimate the structure of the graph using a sequence of the forward and backward search algorithms implemented in R (R Core Team (2012)) in the package gRim (Hojsgaard (2012)). The transaction file has nearly 150 million records, so it will be

sensitive to even very small effects in the model. The resulting fitted model is likely to be over-fit for use with the INSOLE, which has a fraction of the sample. We form a sample by resampling without replacement from the INSOLE with probabilities proportional to the INSOLE sample weights, resulting in an approximately simple random sample without replacement. We do this because the theory is poorly developed for weighted data; we use the INSOLE in the first place to preserve the effect of the edits. This sample has 100,000 records in it. This sample size should be appropriate to detect dependencies useful for the INSOLE. After running the forward-backward procedure until the model stopped changing, which only took two iterations, the result was a log linear model with the following *generating classes*

[[*SCHB, STLCINTX, FamType, SCHA, AGE, AGI*],
[*EX, SCHB, FamT ype, STLCINTX, AGE*],
[*FamT ype, SCHA, SCHB, AGE, Mort, AGI*],
[*AGE, SCHA, realEstTax, SCHB, Mort, AGI*]].

Recall the only interactions in a log linear model are between variables within one member set of the generating class. In this model, for example, interactions are allowed between *SCHB* and *SCHA*, since they are common to the first set, but not *STLCINTX* and *Mort*, since they are not in any of the same sets in the generating class. The Markov Random Field associated with this model is shown in two forms in Figures A2-1 and A2-2. Figure A2-1 shows the whole graph, including *state*. Figure A2-1 shows the graph with *state* excluded; this is the graph for the model estimated within each state.

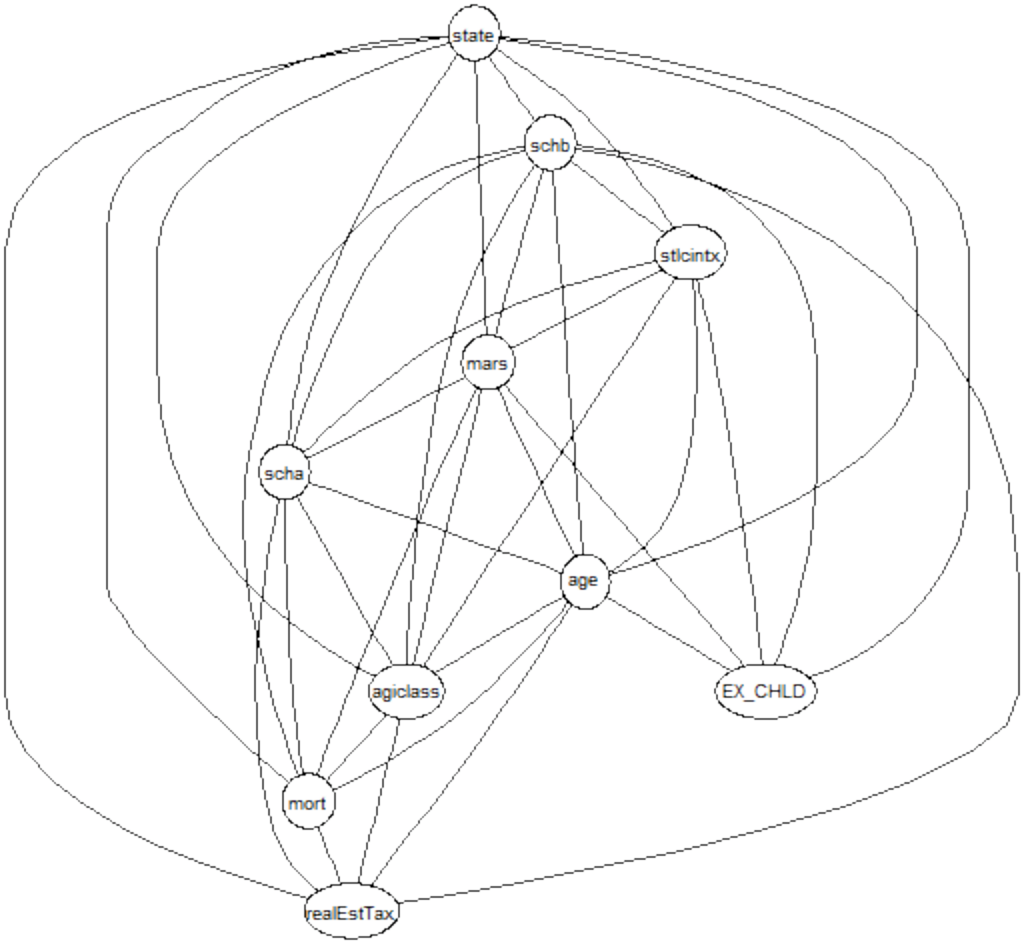

Figure A2-1. Markov Random Field for the model with *state.*

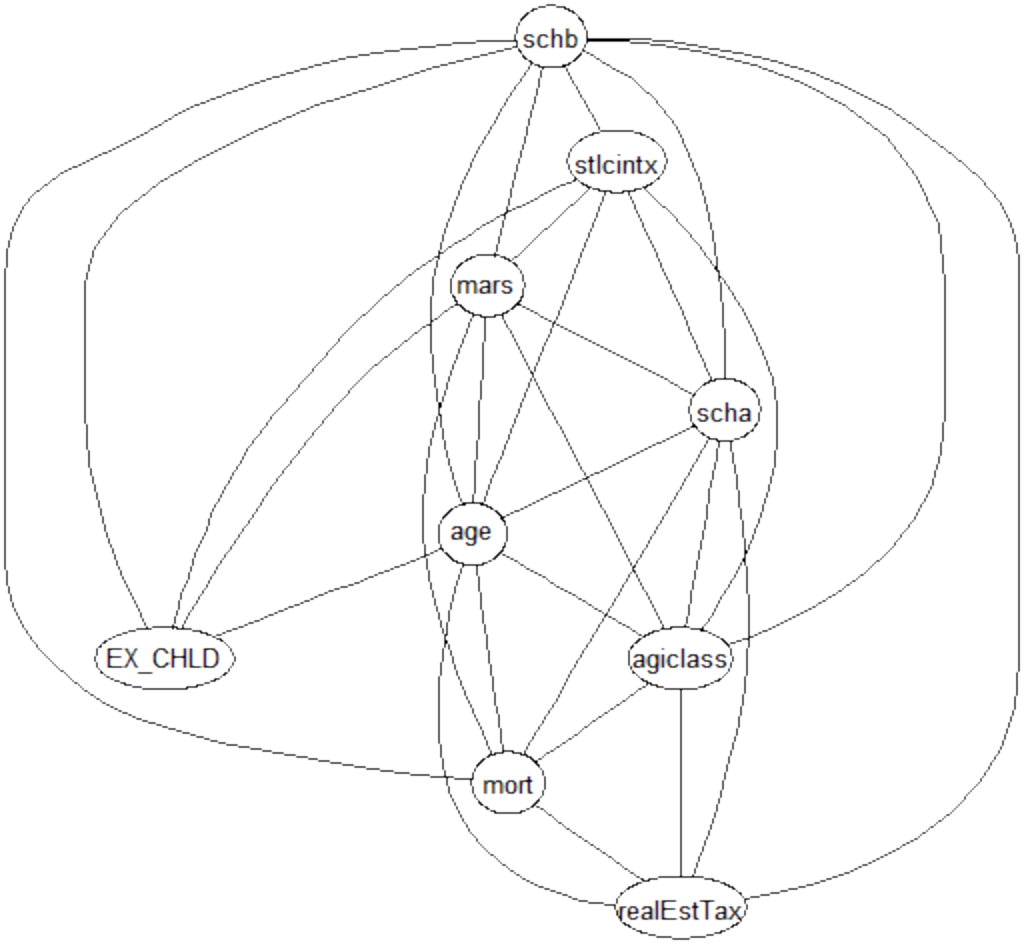

Figure A2-2. Markov Random Field for variables within *state*

It follows that we can summarize the data from the IRTF with the set of four tables with the variables listed in the generating class for each state.

We estimate the joint probabilities of the tables directly by forming the cross-tabulations from the IRTF separately for each state,

$$\hat{p}(X_c = x\_c | ST = st) = \frac{\sum_i I(X_{c,i}=x_c, ST_i=st)}{\sum_i I(ST_i=st)},$$

where *i* indexes the records in the IRTF. The distribution of x and y, normalized so the

total is the population size, evaluated at $(x_{i,}\, y_i)$, is $Np(x, y)$, and is represented by the weight w_i.

Putting it all together,

$$N\hat{p}(st, x, y) = \frac{p(x|st)p(st)w_i}{\sum_{st} p(x|st)p(st)}$$
$$= \hat{p}(ST|x)w_i.$$

Thus our new weight for observation *i* in state *st* is

$$w_{i,st} = \hat{p}(st|x_i)w_i.$$

The new weight $w_{i,st}$ is just the INSOLE weight, split into smaller shares. Those shares are proportional to the estimated proportion of units that are in state st among all returns that are similar to the i^{th} return, in the sense that $X = x_i$ for those returns. Note $\sum_{st}\sum_{i\epsilon st} w_{i,st} = N$. This fits the framework in Schirm et al. (2010), but we use more sources of data; in particular, we use the external tabulations and the model based on external data, rather than relying on data internal to the data set of interest to form the weight adjustment. We also use the equivalence of the decomposition of the Markov graph with the arrangement of the data into two datasets. In the decomposition of the graph into two graphs G_1 and G_2, G_1 corresponds to the first data set (here, the INSOLE) while G_2 corresponds to the second (here, the IRTF). The variables common to the two data sets (here, *X*) correspond to a separator S. This method should generalize to larger collections of data sets, each represented by a graph G_i, i indexing data sets, and a collection of models , M_i.