Statistics of Income

2007 SOI Paper Series

Using the Statistics of Income Division's Sample Data To Reduce Measurement and Processing Error in Small-Area Estimates Produced from Administrative Tax Records

by Kimberly Henry, 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

Kimberly Henry, Internal Revenue Service, Partha Lahiri, University of Maryland, and Robin Fisher, Office of Tax Analysis

The approximately 133 million tax records on the Internal Revenue Service's (IRS) Individual Returns Transaction File have several uses to multiple government agencies. In particular, these data serve as the sampling frame for the Statistics of Income (SOI) Division of IRS, as well as a source of population data for other tabulations. For example, SOI publishes tabulated monetary amounts and the associated number of returns by State and Adjusted Gross Income (AGI) categories using these data (Table 2 in each spring issue of the *SOI Bulletin*). Also, the U.S. Census Bureau compiles the data to the county level for such uses as estimating county-to-county migration patterns (e.g., Gross, 2005) and auxiliary information in the Small Area Income and Poverty Estimation Program's (SAIPE) models to estimate the number of children in poverty within each U.S. county.

These population data, based on administrative tax records for the U.S. tax filing population, are not error-free. While estimates from these data are free from sampling error, the data contain various nonsampling errors, as discovered in prior SOI research comparing return records in the transaction file to records for the same returns in SOI's augmented and edited Form 1040 sample. Only those items necessary for computer processing of a tax return are retained on the transaction file, as opposed to items that might be needed for other purposes, such as auditing. Measurement errors exist between the IRS and SOI data values due to different data editing rules. For revenue processing purposes, IRS does not spend scarce resources correcting errors that do not affect tax liability in the approximately 130 million tax return records it processes each year. Since tax liability is correct, this approach does no harm to IRS's tax collection mission or to taxpayers, but it can adversely affect the usability of the data for statistical purposes. SOI's transcription and editing staff receive extensive training, and the sample of approximately 230,000 returns is augmented with additional items from the return, and more closely monitored and checked for data consistency. Errors occur particularly for variables that are indirectly related to tax liability, such as State and Local Income Taxes deducted on Schedule A. They were also discovered for variables such as Taxable Interest and Business Income/Loss from Sole Proprietors (as reported on Schedule C) in the Tax Year 2003 IRS data. To correct these errors, SOI had to delay its publication of Table 2 for several months. Other limitations in the IRS data include a smaller amount of information being available, compared to SOI's sample, and data are often provided to SOI in tabular form, with monetary amounts rounded to thousands, and certain high-income taxpayers are omitted.

In order to improve on design-based estimators, several indirect and model-based methods have been proposed in the literature. These improved estimation procedures essentially use *implicit* or *explicit* models which *borrow strength* from related resources such as administrative and census records and previous survey data. In order to estimate per-capita income for small areas (population less than 1,000), Fay and Herriot (1979) used an empirical Bayes method that combines the U.S. Current Population Survey data with various administrative and census records. In order to incorporate both the sampling and model errors, Fay and Herriot (1979) used a two-level model which can be either viewed as a Bayesian model or a mixed regression model. Their empirical Bayes estimator (also an empirical best linear unbiased predictor (EBLUP)) performed better than the direct survey estimator and a synthetic estimator used earlier by the U.S. Census Bureau.

In an EBLUP approach, the best linear unbiased predictor (BLUP) of the small-area mean is first pro-

duced, and the unknown variance component(s) is (are) estimated by a standard method (e.g., maximum likelihood (ML), residual maximum likelihood (REML), ANOVA, etc.). The resultant predictor, i.e., the BLUP with estimated variance component(s), is known as an EBLUP of the true small-area mean. A challenging problem in an EBLUP approach is to obtain a reliable measure of uncertainty of an EBLUP that captures all sources of variability. We do not attempt to cite all the papers that use the Fay-Herriot model or its extension; such references can be found in Rao (2003) and Jiang and Lahiri (2006).

▶ Data Description

The SOI Sample

SOI draws annual samples of tax returns to produce richer and cleaner data for population estimation and tax modeling purposes. Stratification for the finite population of tax returns for SOI's Tax Year 2004 Individual sample used the following categories:

1. Nontaxable returns with adjusted gross income or expanded income of $200,000 or more.

2. High combined business receipts of $50,000,000 or more.

3. Presence/absence of special forms or schedules (Form 2555, Form 1116, Form 1040 Schedule C, and Form 1040 Schedule F).

Stratum assignment priority was based on the order in which a return met one of these categories. For example, if a return met (1) and (2), it fell into strata based on (1). Within category (3), stratification also used size of total gross positive or negative income and an indicator of the return's "usefulness" for tax policy modeling purposes (Scali and Testa, 2006). The positive/negative income values in strata boundaries were indexed for inflation between 1991 and the current tax year (Hostetter et al. 1990). These criteria resulted in 216 strata.

Each tax return in the target population was assigned to a stratum based on these criteria, then subjected to sampling in a two-step procedure. Within each stratum, a .05-percent stratified simple random sample, called the Continuous Work History Sample (CWHS), was selected (Weber, 2004). For returns not selected for this sample, a Bernoulli sample was independently selected from each stratum, with sampling rates from 0.05 percent to 100 percent.

SOI's data capture and cleaning procedures resulted in a sample of 200,778 returns, including 65,948 CWHS returns, from an estimated population of 133,189,982 returns. We placed the 34,484 returns that SOI sampled with certainty into one certainty stratum, since they represented a census of tax returns. Thus, without loss of generality, we exclude this stratum from the population and develop our estimation method to estimate totals from all other strata. To estimate the entire population total, we simply add the total from the certainty strata to our estimate for the remaining population.

Small Areas and Variables of Interest

The reduced dataset for this analysis was created by first separating SOI's Tax Year 2004 (i.e., income reported in 2005 that was earned in 2004) sample into the certainty and noncertainty units. For both sets, the weighted sample data were tabulated to the State level, where "State" included the 50 U.S. States, Washington DC, and an "other" category that included returns filed by civilians and military individuals living abroad in U.S. possessions and territories, Puerto Rico, etc.

We selected six variables, which can be grouped into two categories: variables that are more or less susceptible to errors in the IRS data. They are listed, with their locations on Form 1040 and a brief description, in Table 1 below.

▶ Direct Estimators

Let y_k be the value of the characteristic of interest for the kth tax return, $k \in U$, the finite population of tax returns. We are interested in estimating the finite population total:

$$Y = \sum_{k \in U} y_k \ .$$

Table 1. Variable Names, Tax Form Location, and Description, by Variable of Interest

Susceptible to Error	*Variable*	*2004 Tax Form Location*	*Description* [a]
Less	Adjusted Gross Income	Line 36	Income reported from the calculation of total income (Line 22) (pp. 117-118).
	Taxable Interest Income	Line 8a	Taxable amount of interest from bonds, savings, etc. (p. 142).
	Earned Income Tax Credit	Line 65a	Taxpayer credit for lower-income working individuals (pp. 123-124).
More	Real Estate Taxes	Line 6, Schedule A	Amount of nonbusiness-related real estate taxes paid (p. 137).
	State and Local Income Taxes	Line 5a, Schedule A	An itemized deduction of the State/local income taxes withheld from taxpayer's 2004 salary (p. 144).
	State and Local General Sales Taxes	Line 5b, Schedule A	Deducted State and local general Sales tax (instead of state and local income tax deduction, p. 139).

a: page numbers from IRS 2006.

Let s denote the sample of tax returns drawn from the population of tax returns, $s_d \subset s$ denote the part of the sample that belongs to the domain d of interest, and w_k denote the sampling weight for the kth sampled tax return, $k \in s$. The sampling weight is simply the inverse of the inclusion probability and represents a certain number of population units in the finite population. In our case, we have *epsem* sampling within each stratum, i.e., the sampling weights are the same for all the sampled units belonging to the same stratum. The weights vary across strata, due to disproportional allocation of the sample into different strata. Our domain cuts across the design strata, so that weights of different sampled units inside a domain are generally different. Let

$$Y_d = \sum_{k \in U_d} y_k$$

denote the population total for the dth domain (excluding the tax returns belonging to the certainty stratum). Since N_d is known from the IRS records, our problem is equivalent to estimating the finite population mean for domain d:

$$\bar{Y}_d = Y_d / N_d .$$

We can consider the following design-based direct estimator of $\bar{Y}_d$:

$$\bar{y}_{dw} = \sum\nolimits_{k \in s_d} w_k y_k \Big/ \sum\nolimits_{k \in s_d} w_k . \qquad (1)$$

If the domain d is large, then a reliable design-based estimate of the sampling variance of $\bar{y}_{dw}$ can be obtained using the Taylor linearization technique using software like SUDDAN.

▶ EBLUP Estimators

In this section, we shall obtain an empirical best linear unbiased estimator of $\bar{Y}_d$ under the following area level model due to Fay and Herriot (1979). For $d = 1, \cdots, m,$ assume

$$\text{Level 1: } \bar{y}_{dw} \sim ind\ N(\bar{Y}_d, D_d);$$
$$\text{Level 2: } \bar{Y}_d \sim ind\ N(x_d^T \beta, \psi),$$

where D_d is the estimated sampling variance of $\bar{y}_{dw}$ and x_d is a $p \times 1$ vector of known auxiliary variables based on the IRS data.

Under the Fay-Herriot model, the best predictor (BP) of $\bar{Y}_d$ is given by:

$$\hat{\bar{Y}}_d^{BP} = (1 - B_d)\bar{y}_{dw} + B_d x_d^T \beta , \qquad (2)$$

where $B_d = \dfrac{D_d}{D_d + \psi}$. Note that the BP can be motivated without the normality assumption. If ψ is known, then β is estimated by the weighted least squares estimator:

$$\hat{\beta}(\psi) = \left(\sum_{d=1}^{m} \frac{1}{D_d + \psi} x_d x_d^T \right)^{-1} \left(\sum_{d=1}^{m} \frac{1}{D_d + \psi} x_d \bar{y}_{dw} \right).$$

Replacing β by $\hat{\beta}(\psi)$, we obtain the following empirical best predictor of $\bar{Y}_d$:

$$\hat{\bar{Y}}_d^{EBP} = (1 - B_d)\bar{y}_{dw} + B_d x_d^T \hat{\beta}(\psi). \quad (3)$$

Note that $\hat{\bar{Y}}_d^{EBP}$ is also the best linear unbiased predictor (BLUP) of $\bar{Y}_d$ under the following linear mixed model:

$$\bar{y}_{dw} = x_d^T \beta + v_d + e_d,$$

where the sampling errors $\{e_d\}$ and the random effects $\{v_d\}$ are uncorrelated, with $v_d \sim (0,\psi)$ and $e_d \sim (0, D_d)$.

When both β and ψ are unknown, we propose the following empirical best linear unbiased predictor (EBLUP) of $\bar{Y}_d$:

$$\hat{\bar{Y}}_d^{EBLUP} = (1 - \hat{B}_d)\bar{y}_{dw} + \hat{B}_d x_d^T \hat{\beta}(\hat{\psi}), \quad (4)$$

where $\hat{B}_d = \dfrac{D_d}{D_d + \hat{\psi}}$. In this paper, we consider the residual maximum likelihood (REML), Prasad-Rao simple method-of-moments (PR), and Fay-Herriot's method-of-moments (FH) estimators of ψ. We define the mean square prediction error of $\hat{\bar{Y}}_d^{EBLUP}$ as

$$MSPE(\hat{\bar{Y}}_d^{EBLUP}) = E\left(\hat{\bar{Y}}_d^{EBLUP} - \bar{Y}_d\right)^2,$$

where the expectation is taken over the joint distribution of $\bar{y}_{dw}$ and $\bar{Y}_d$ under the Fay-Herriot model. A naïve MSPE estimator is obtained by estimating the MSPE of the BLUP and is given by:

$$mspe_d^N = g_{1i}(\hat{\psi}) + g_{2i}(\hat{\psi}), \quad (5)$$

where $g_{1d}(\hat{\psi}) = \hat{B}_d \hat{\psi}$, $g_{2d}(\hat{\psi}) = \hat{B}_d^2 h_{dd}$, and

$$h_{dd} = x_d^T \left(\sum_{j=1}^{m} \frac{1}{D_j + \hat{\psi}} x_j x_j^T \right)^{-1} x_d.$$

Intuitively, this naïve MSPE estimator is likely to underestimate the true MSPE since it fails to incorporate the additional uncertainty due to the estimation of ψ. In fact, Prasad and Rao (1990) showed that the order of this underestimation is $O(m^{-1})$ under the following regularity conditions:

$$(r.1)\ 0 < D_L \le D_d \le D_U < \infty,\ d = 1, \cdots, m;$$
$$(r.2)\ \sup_{d \ge 1} h_{dd} = O(m^{-1}).$$

Interestingly, the naïve MSPE estimator even underestimates the true MSPE of the BLUP, the order of underestimation being $O(m^{-1})$. A second-order unbiased estimator of MSPE is given by:

$$mspe_d^{DL} = g_{1d}(\hat{\psi}) + g_{2d}(\hat{\psi}) + 2g_{3d}(\hat{\psi}) - g_{4d}(\hat{\psi}), \quad (6)$$

where $g_{3d}(\hat{\psi}) = \dfrac{\hat{B}_d^2}{D_d + \hat{\psi}} \mathrm{var}(\hat{\psi})$ and $g_{4d}(\hat{\psi}) = \hat{B}_d^2 bias(\hat{\psi})$. Here, $bias(\hat{\psi})$ and $\mathrm{var}(\hat{\psi})$ are the asymptotic bias and variance estimates of $\hat{\psi}$, respectively. For example,

$$\mathrm{var}(\hat{\psi}) = 2\left\{ \sum_{d=1}^{m} (D_d + \hat{\psi})^{-2} \right\}^{-1}$$ for the REML method,

$$\mathrm{var}(\hat{\psi}) = 2m^{-2} \sum_{d=1}^{m} (D_d + \hat{\psi})^2$$ for the PR method, and

$$\mathrm{var}(\hat{\psi}) = 2m \left\{ \sum_{d=1}^{m} (D_d + \hat{\psi})^{-1} \right\}^{-2}$$

for the FH method; $bias(\hat{\psi}) = 0$ for the REML and PR methods, and

$$bias(\hat{\psi}) = \frac{2\left[m\left\{ \sum_{d=1}^{m} (\mathrm{D_d} + \hat{\psi})^{-2} \right\} - \left\{ \sum_{d=1}^{m} (\mathrm{D_d} + \hat{\psi})^{-1} \right\}^2 \right]}{\left\{ \sum_{d=1}^{m} (\mathrm{D_d} + \hat{\psi})^{-1} \right\}^3}$$

for the Fay-Herriot method. See Datta and Lahiri (2000) and Datta, Rao, and Smith (2005) for details.

▶ Results: Descriptive Plots

Figure A.1 contains plots of $\bar{y}_{dw}$ versus the corresponding mean calculated from the IRS tabular data. The main sources of error in the IRS means are the measurement and processing errors. On the other hand, the SOI means are primarily subject to the sampling error. The magnitude of these errors varies depending on the variable, but the effect of both errors is that the points are further from the reference line drawn in each plot. However, a strong linear relationship is observed between these means for each variable, particularly for variables that are considered less likely affected by IRS errors. The relationship is weaker for State and Local Income Taxes and State and Local General Sales Taxes, where IRS data have more error. The State and Local Income Taxes variable also has an apparent outlier in one State's sample mean (TN).

Figure A.2 contains descriptive plots of the shrinkage factors for each variable and analysis method, sorted by the size of D_d. For each variable, as the estimates of D_d increase, the shrinkage factor increases, which implies that more weight is given to the IRS mean in $\hat{\bar{Y}}_d^{EBLUP}$. All three methods yield zero estimate of ψ, for Taxable Interest Income, which produces an estimate of 1 for the shrinkage factors for all areas. This is undesirable since, in such a situation, the EBLUP estimate is identical to the regression synthetic estimate, which does not directly use the SOI data.

Figure A.3 contains plots of the percentage-relative differences of the IRS totals, to various alternatives, for all six variables. For all variables, the States were sorted by the size of the estimated coefficient of variation (CV) of the direct estimate for the total. As the CV of the direct estimate increased, the direct estimate was further from the IRS-based total, shown by the points being further from zero on the right side of each plot.

Figure A.4 shows the percentage-relative gain of EBLUP estimates over that of the direct estimates for each variable. That is:

$$\% \text{ Rel Gain} = 100 \times \frac{CV(\bar{y}_{dw}) - CV(\hat{\bar{Y}}_d^{EBLUP})}{CV(\bar{y}_{dw})} .$$

For Adjusted Gross Income, the REML results had the largest percentage-relative gains, except for the two largest States (CA and NY), where the direct estimates were more precise. This was due to the fact that, in this case, EBLUP is identical to the regression synthetic estimate, since no weight was given to the direct estimate (as the shrinkage factor was equal to 1). All of the FH estimates were more precise than the direct, shown by positive gains for all States, while the PR were less precise for the five States with the lowest CVs. As expected, all EBLUP's gains in precision increased as the CV of the direct estimate increased.

For Taxable Interest Income, there were also some large gains in precision. Precision gains below -25 percent (or loss above 25 percent) were truncated, which occurred for nearly a third of the PR estimates and the REML and FH estimates for California. However, the REML and FH methods generally performed well for this variable.

For Earned Income Tax Credit, we obtained close positive percent relative gains for all States except "Other," where the PR performed best.

For Real Estate Taxes, all three EBLUPs performed well; PR and FH had higher gains in precision in States where the direct estimates had smaller CVs than REML, but all three performed equally well (and better than the direct estimates) for States where the direct estimates had higher CVs.

For State and Local Income Taxes, we see that only the REML performed well; the PR and FH methods produced negative percentage gains in precision for all States except the outlier point noted in Figure A.1, which had a much higher gain in precision. Thus, these methods appear to be sensitive to outliers: they adjusted our outlier, but at the expense of the other States.

For State and Local General Sales Taxes, we see lower (but positive) gains in precision that only slightly increased as the CV of the direct estimate increased.

Thus, when the relationship shown in Figure A.1 is much weaker, due to measurement and processing error in the IRS data, we see lower gains in precision in the EBLUPs.

▶ Conclusions, Limitations, and Future Considerations

We attempt to improve on population-based estimates that are subject to nonsampling error and sample-based estimates subject to sampling error. In general, our EBLUPs seem to produce preferable results, obtained by exploiting relationships between the sample and population variable means. This was demonstrated by high gains in precision and more stability in the estimates.

However, for four out of six of the tax return variables we examined, at least one of the REML, FH, and PR methods used to estimate ψ produced shrinkage factors equal to one for all States. This problem may be due to using unreliable design-based direct variance estimates for the D_d. The methods also appear to be sensitive when there are outliers, and performance is lower when the relationship is weaker. In order to overcome some of these problems and to make inferences more flexible, we plan to consider a hierarchical Bayes method in the future.

Starting in Tax Year 2005, SOI's individual tax return sample is expected to increase by approximately 65,000 noncertainty returns. This new sample will be useful to improve on the estimates. We can also use this new sample to develop a robust evaluation criterion to compare different model-based methods.

▶ References

Datta, G.S. and P. Lahiri (2000), "A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems," *Statistica Sinica*, 10, pp. 613-627.

Fay, R.E. and R.A. Herriot (1979), "Estimates of Income for Small Places: an Application of James-Stein Procedure to Census Data," *Journal of American Statistical Association* 74, pp. 269-277.

Gross, E. (2005), "Internal Revenue Service Area-To-Area Migration Data: Strengths, Limitations, and Current Trends," *Proceedings of the Section on Government Statistics*, American Statistical Association.

Hostetter, S.; J.L. Czajka; A.L. Schirm; and K. O'Conor (1990), "Choosing the Appropriate Income Classifier for Economic Tax Modeling," *Proceedings of the Section on Survey Research Methods Section*, American Statistical Association, pp. 419-424.

Internal Revenue Service (2006), "Explanation of Terms," *Statistics of Income–2004 Individual Income Tax Returns,* Internal Revenue Service, Publication 1304, pp. 117-147.

Jiang, J. and P. Lahiri, (2006). "Mixed model prediction and small-area estimation (with discussions)." Test, 15, 1, pp. 1-96.

Lahiri, P. (2001), *Model Selection,* IMS Lecture Notes/Monograph, Volume 38.

Lahiri, P. (2003b), "A review of empirical best linear unbiased prediction for the Fay-Herriot small-area model," *The Philippine Statistician*, 52, pp. 1-15.

Prasad, N.G.N. and J.N.K. Rao, (1990), "The Estimation of Mean Squared Error of Small Area Estimators," Journal of American Statistical Association 85, pp. 163-171.

Rao, J.N.K. (2003), *Small Area Estimation*, John Wiley and Sons, New York.

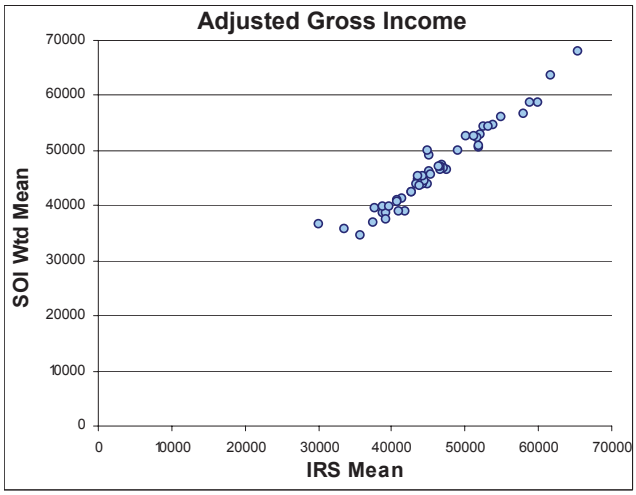
Sarndal, C.E. and M.A. Hidiroglou, (1989), "Small Domain Estimation: A Conditional Analysis," *Journal of the American Statistical Association*, 84, pp. 266-275.

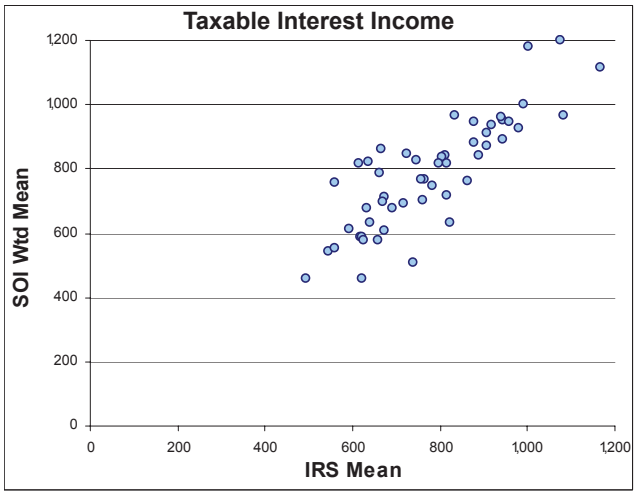
Scali, J. and V. Testa, (2006), *Statistics of Income–2004 Individual Income Tax Returns,* Internal Revenue Service, Publication 1304, pp. 23-27.

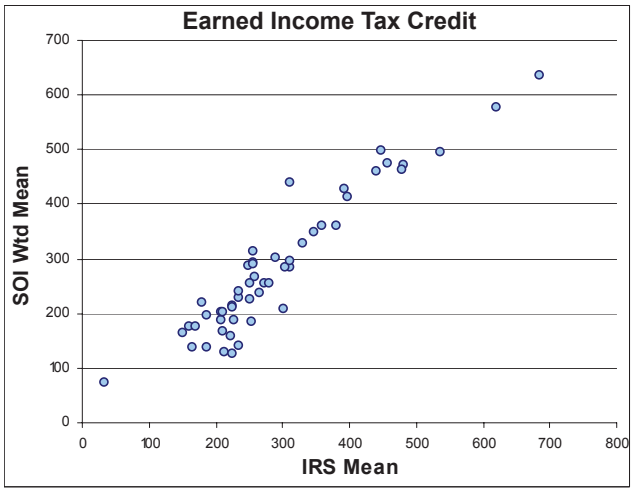
U.S. Census Bureau, Small Area Income and Poverty Estimation Division (SAIPE), http://www.census.gov/hhes/www/saipe/saipe.html.

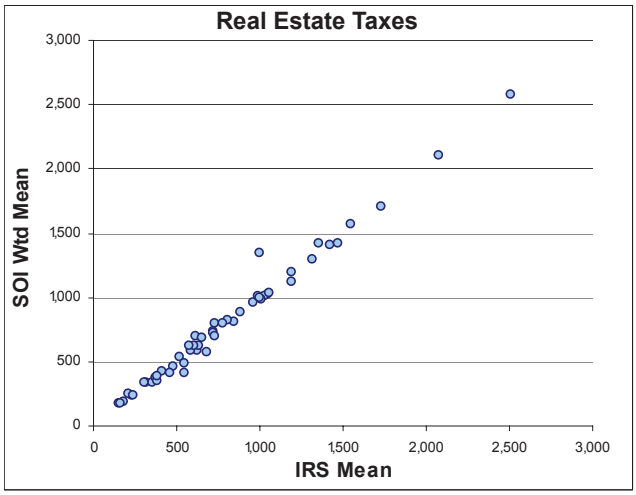
Weber, M., (2004), "The Statistics of Income 1979-2002 Continuous Work History Sample Individual Tax Return Panel," *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

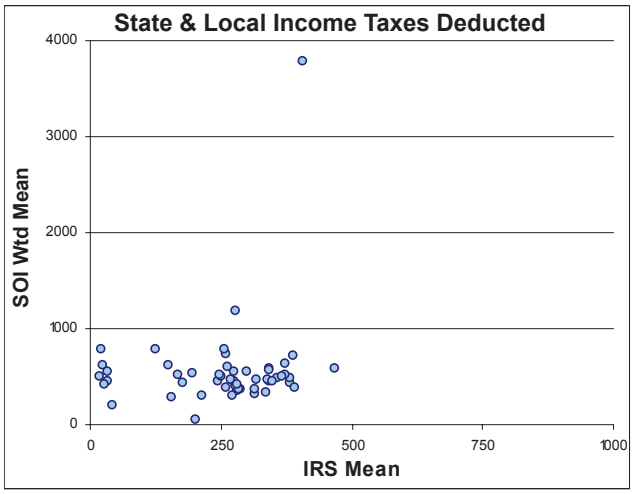
A.1. Mean Variable Plots, by Variable of Interest

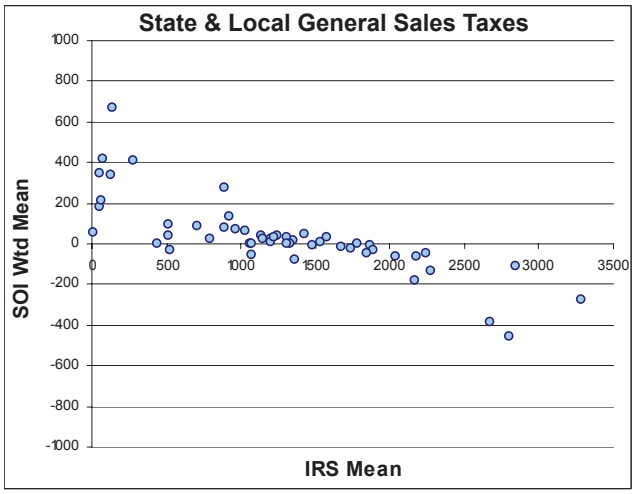






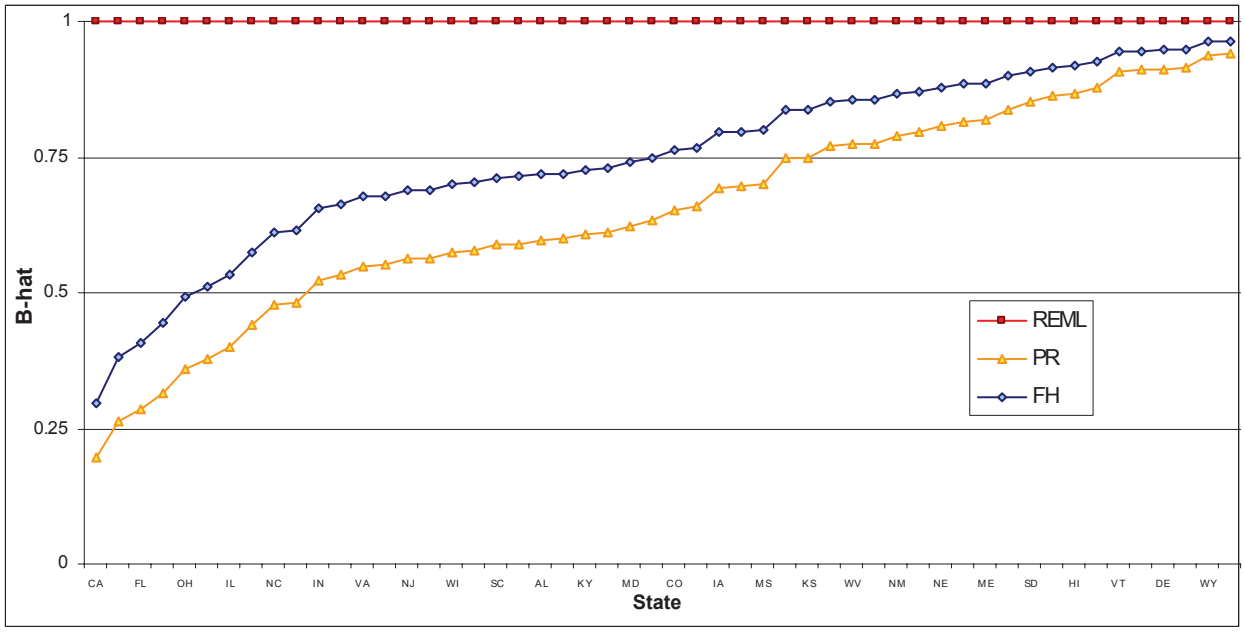




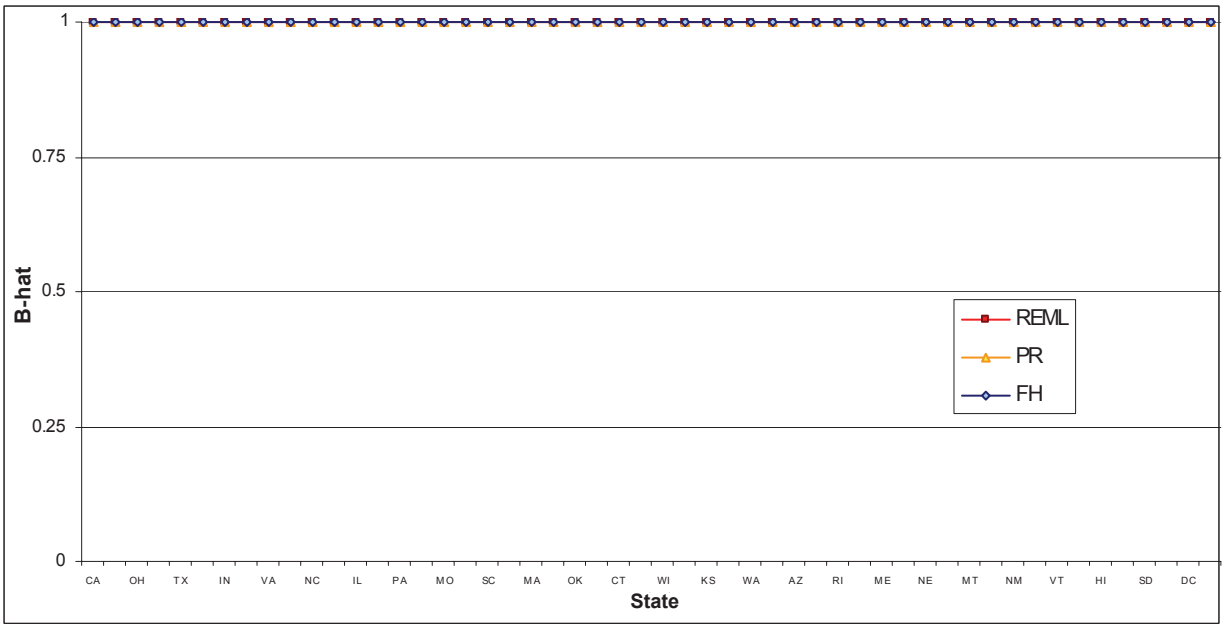


A.2. Shrinkage Factors, by Variable of Interest

Adjusted Gross Income

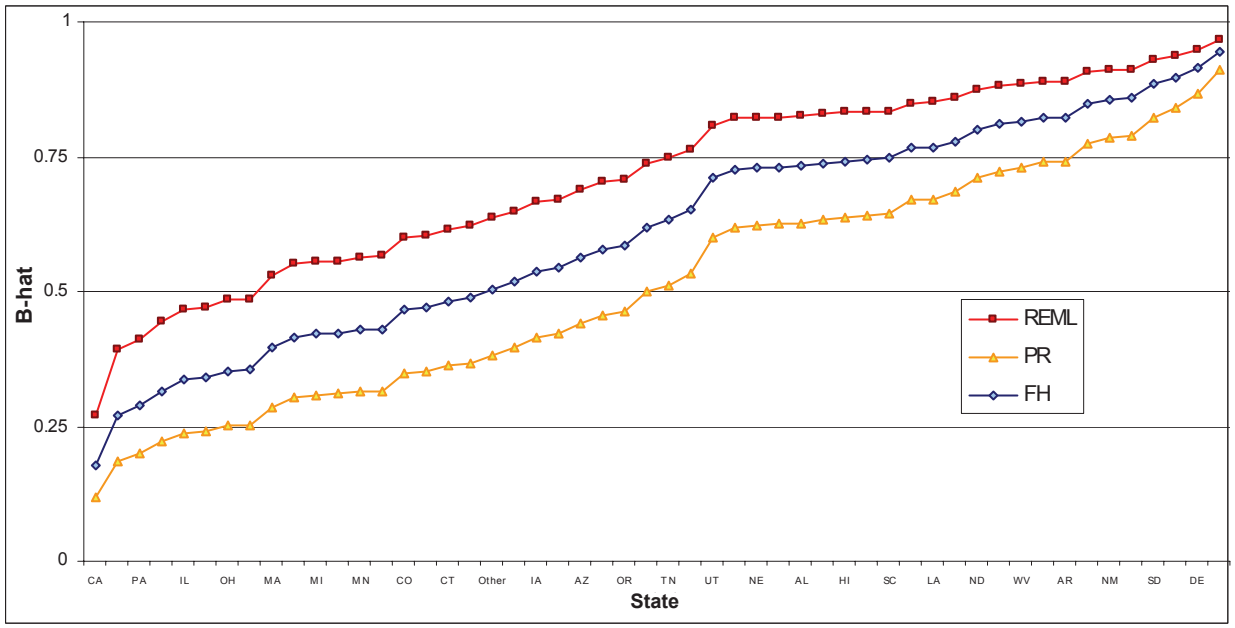


Taxable Interest Income

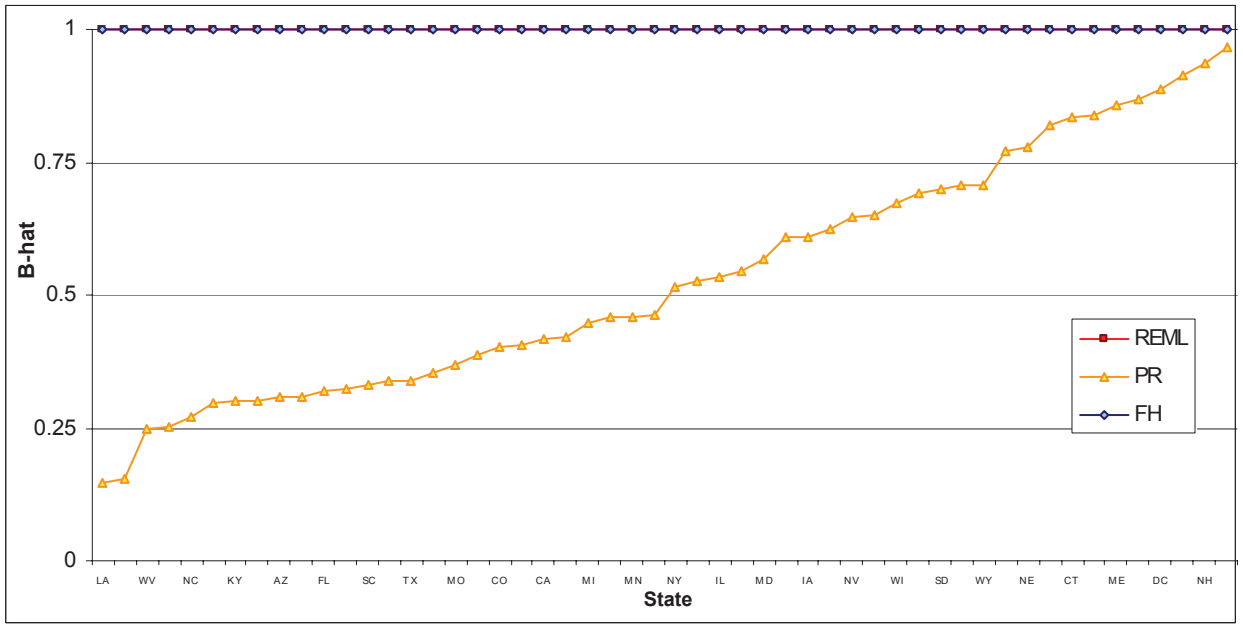


A.2. Shrinkage Factors, by Variable of Interest—Continued

Earned Income Tax Credit

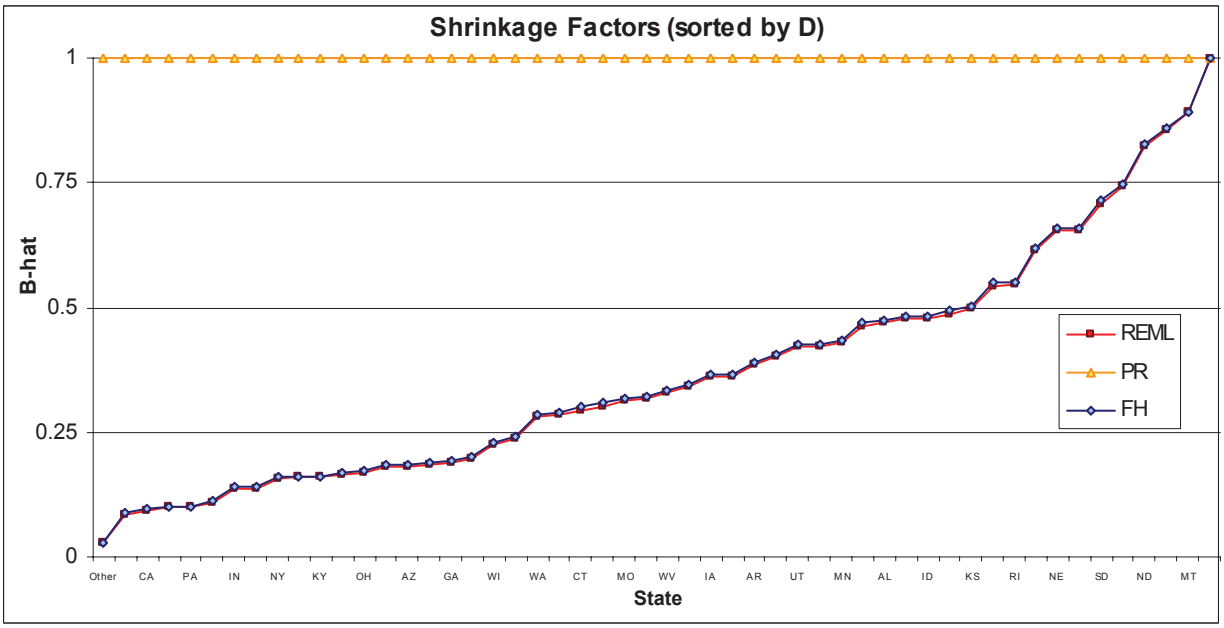


Real Estate Taxes

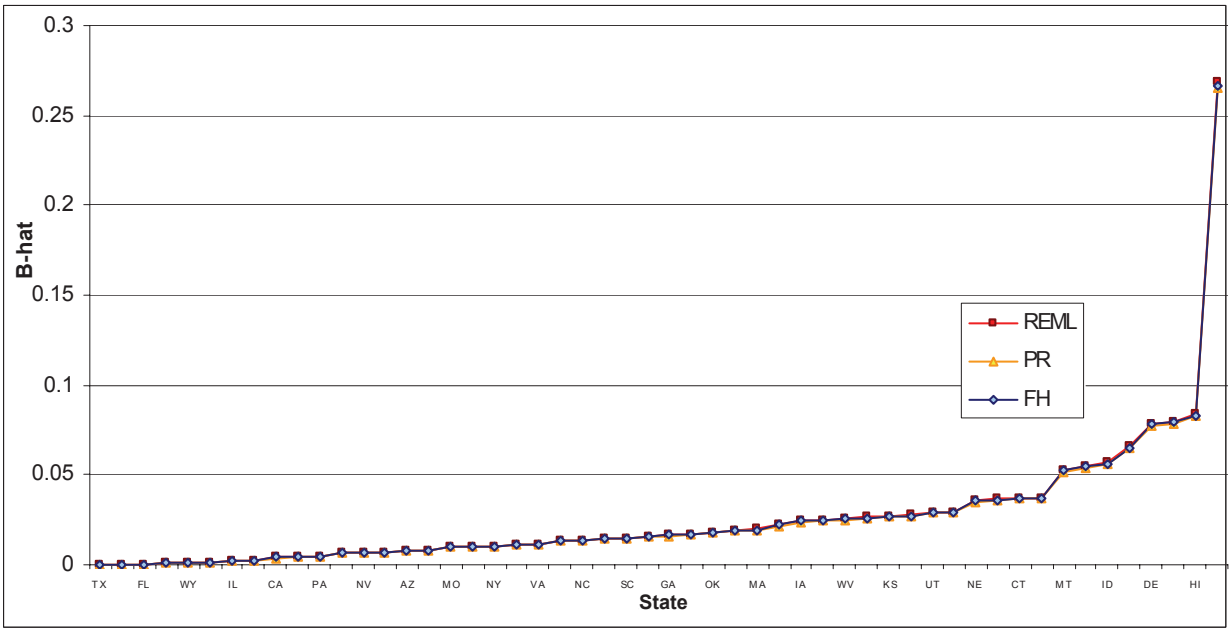


A.2. Shrinkage Factors, by Variable of Interest—Continued

State & Local Income Taxes

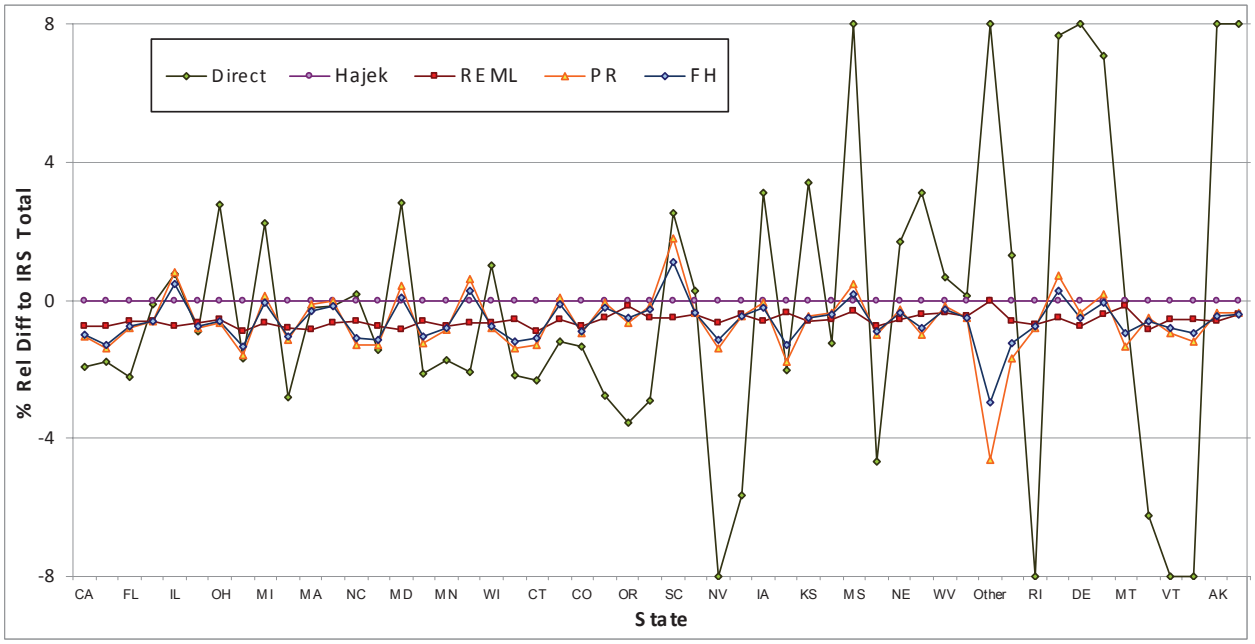


State & Local General Sales Taxes

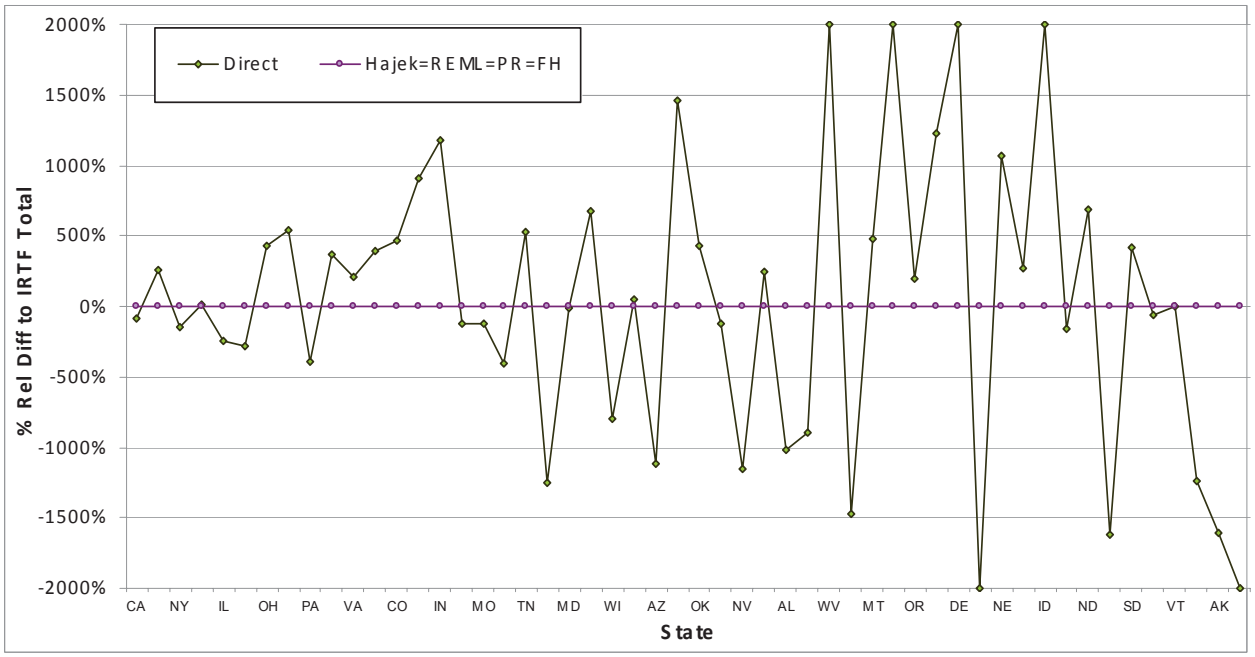


A.3. Percentage Relative Difference Between Various Estimates and IRTF Totals, by Variable of Interest

Adjusted Gross Income

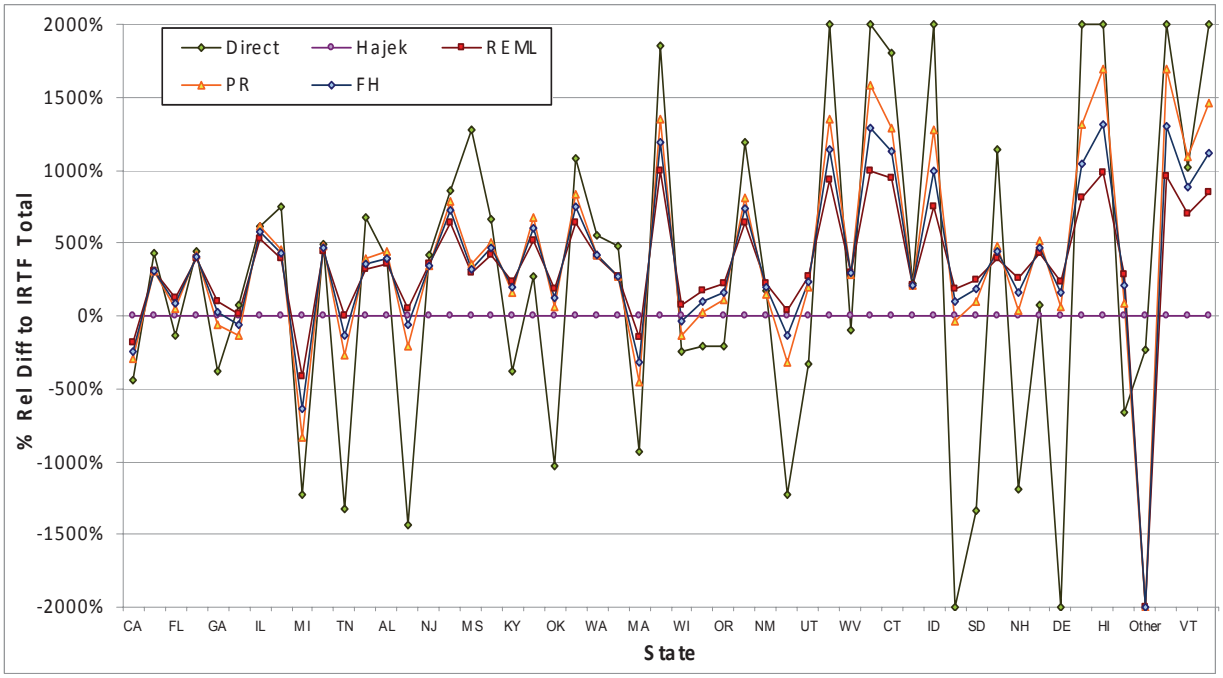


Taxable Interest Income

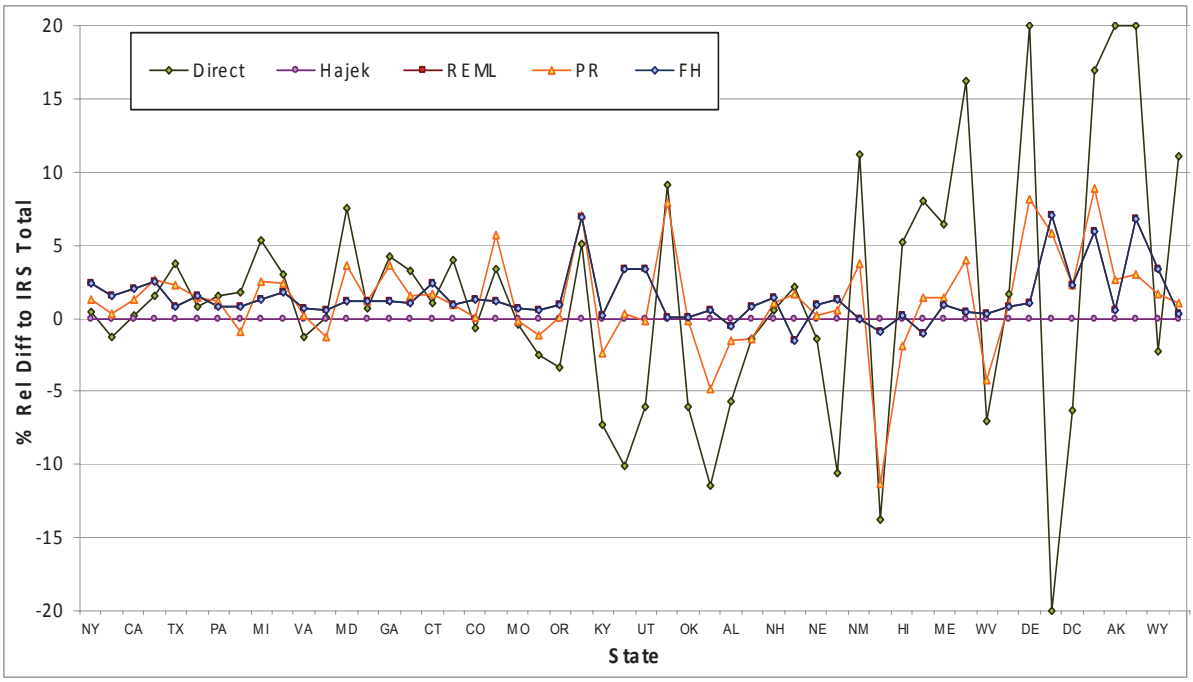


A.3. Percentage Relative Difference Between Various Estimates and IRTF Totals, by Variable of Interest—Continued

Earned Income Tax Credit

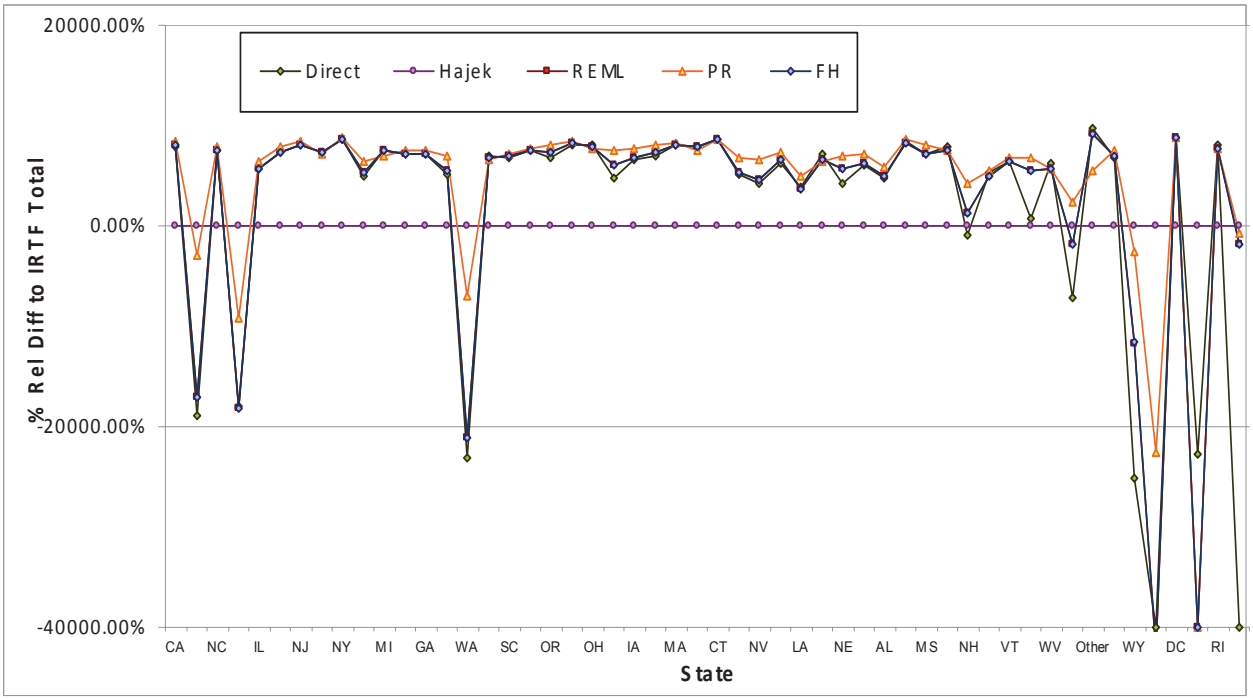


Real Estate Taxes

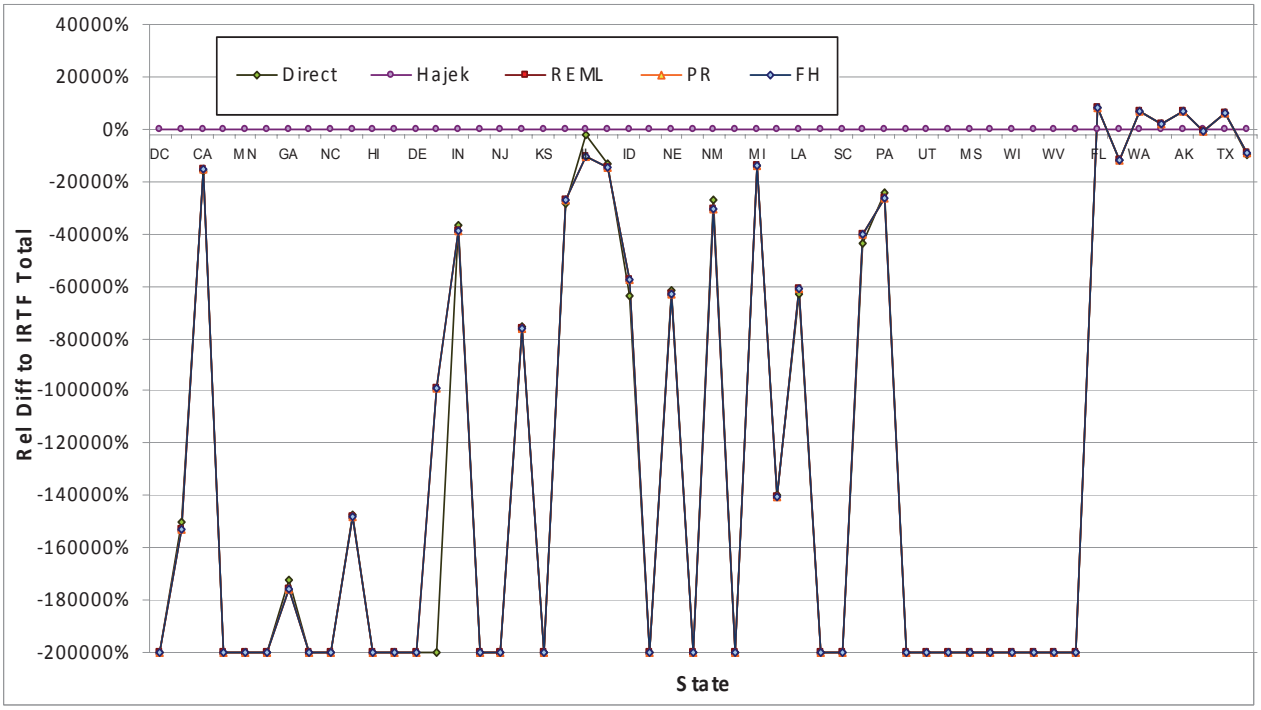


A.3. Percentage Relative Difference Between Various Estimates and IRTF Totals, by Variable of Interest—Continued

State & Local Income Taxes

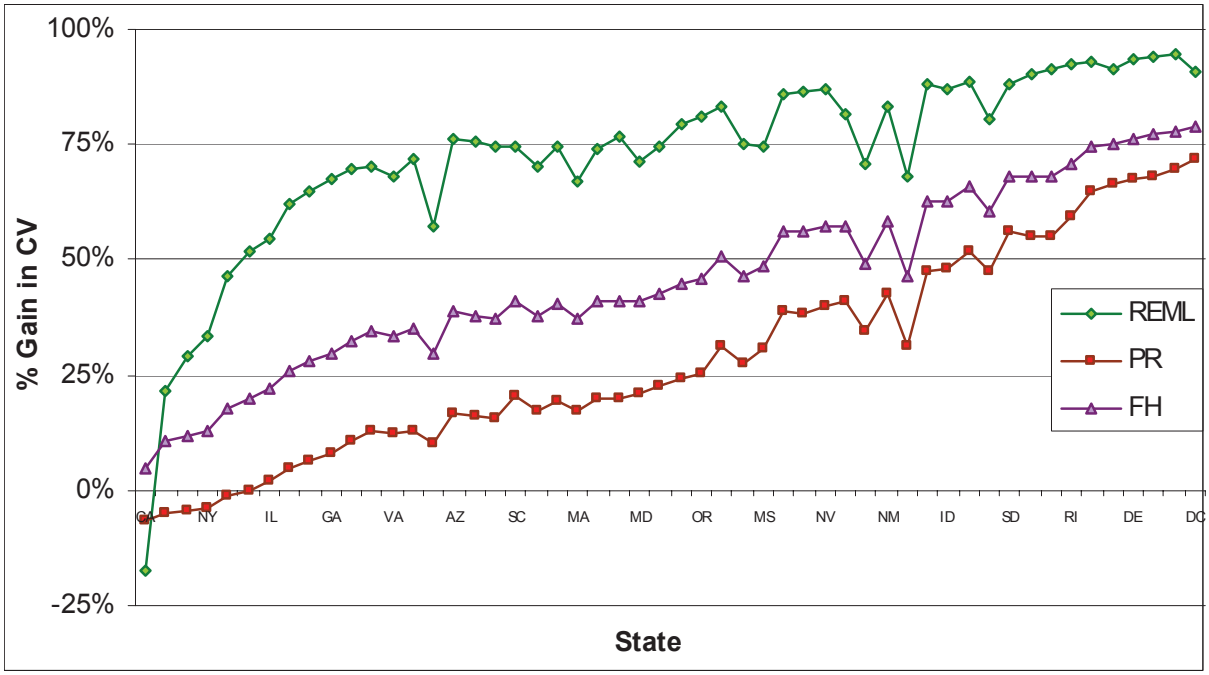


State & Local General Sales Taxes

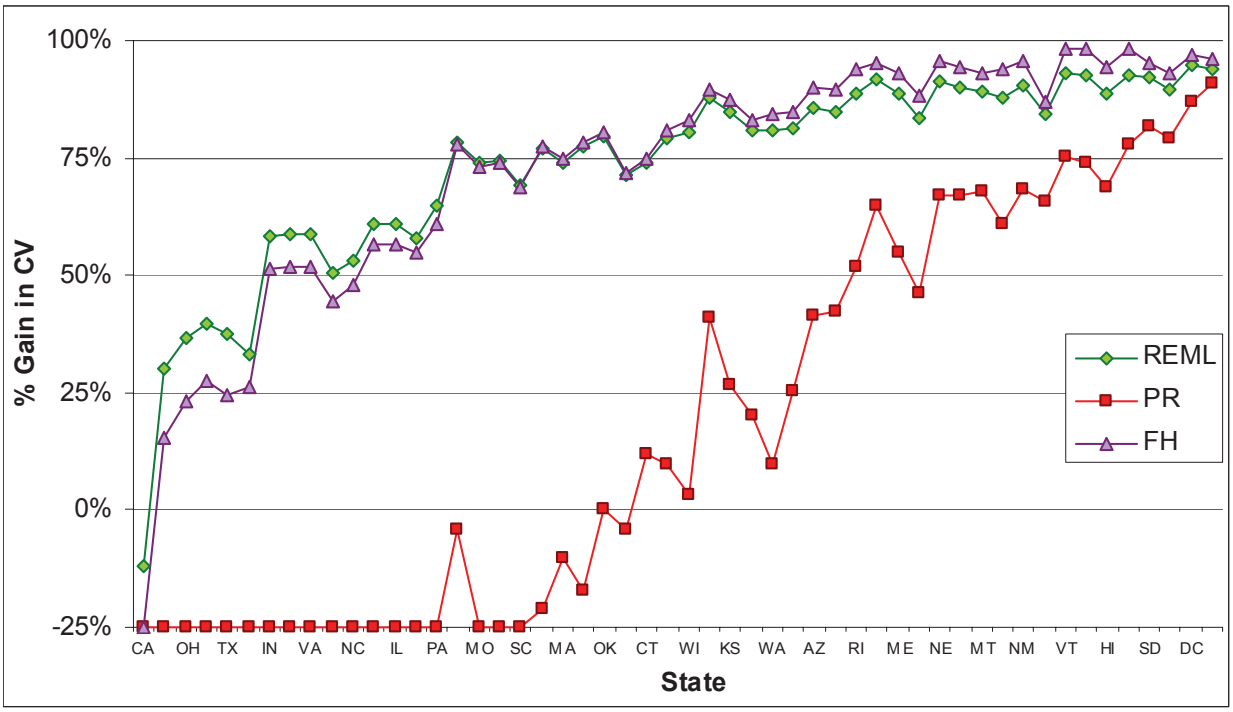


A.4. Percentage Relative Gain in EBLUP Estimates over the Coefficients of Variation of SOI Sample-Based Estimates, by Variable of Interest

Adjusted Gross Income

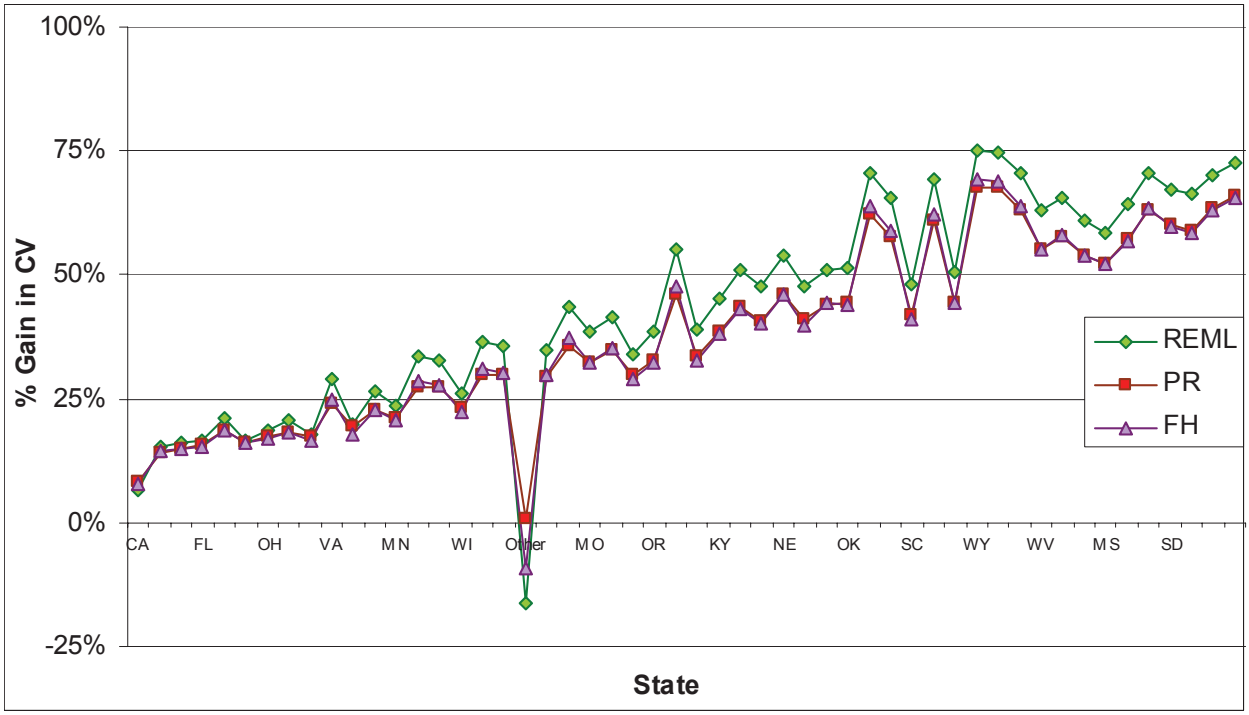


Taxable Interest Income

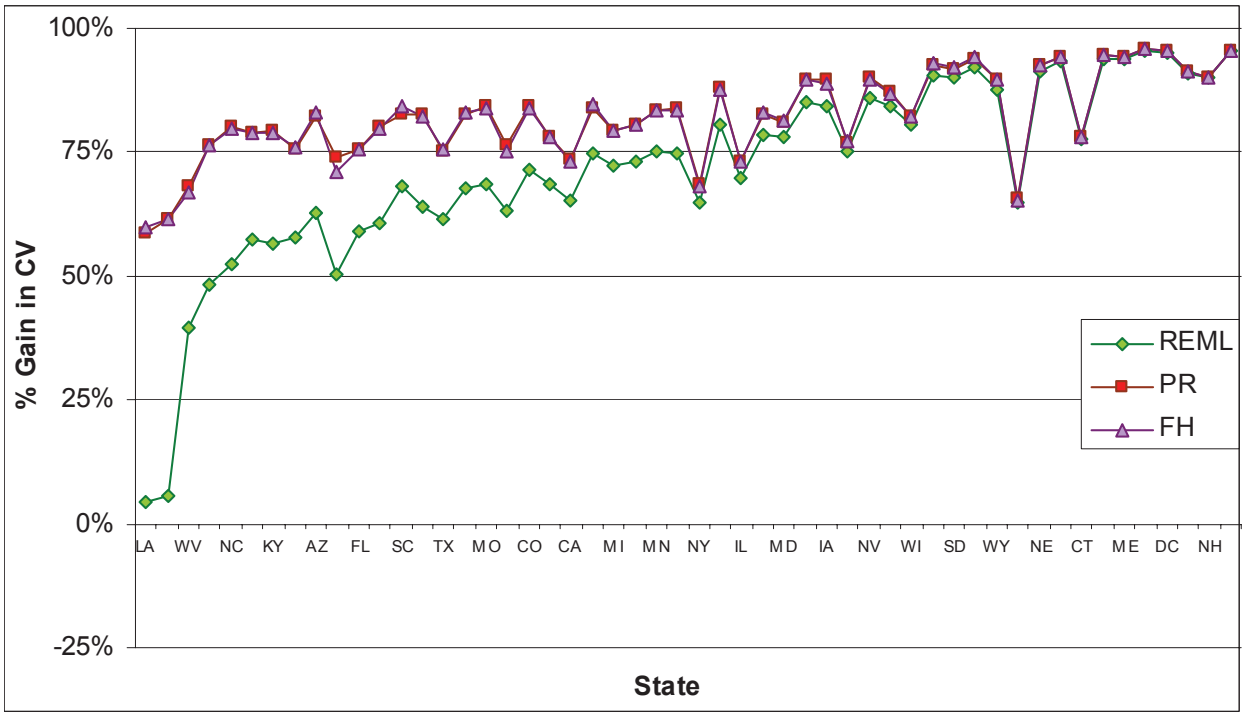


A.4. Percentage Relative Gain in EBLUP Estimates over the Coefficients of Variation of SOI Sample-Based Estimates, by Variable of Interest—Continued

Earned Income Tax Credit

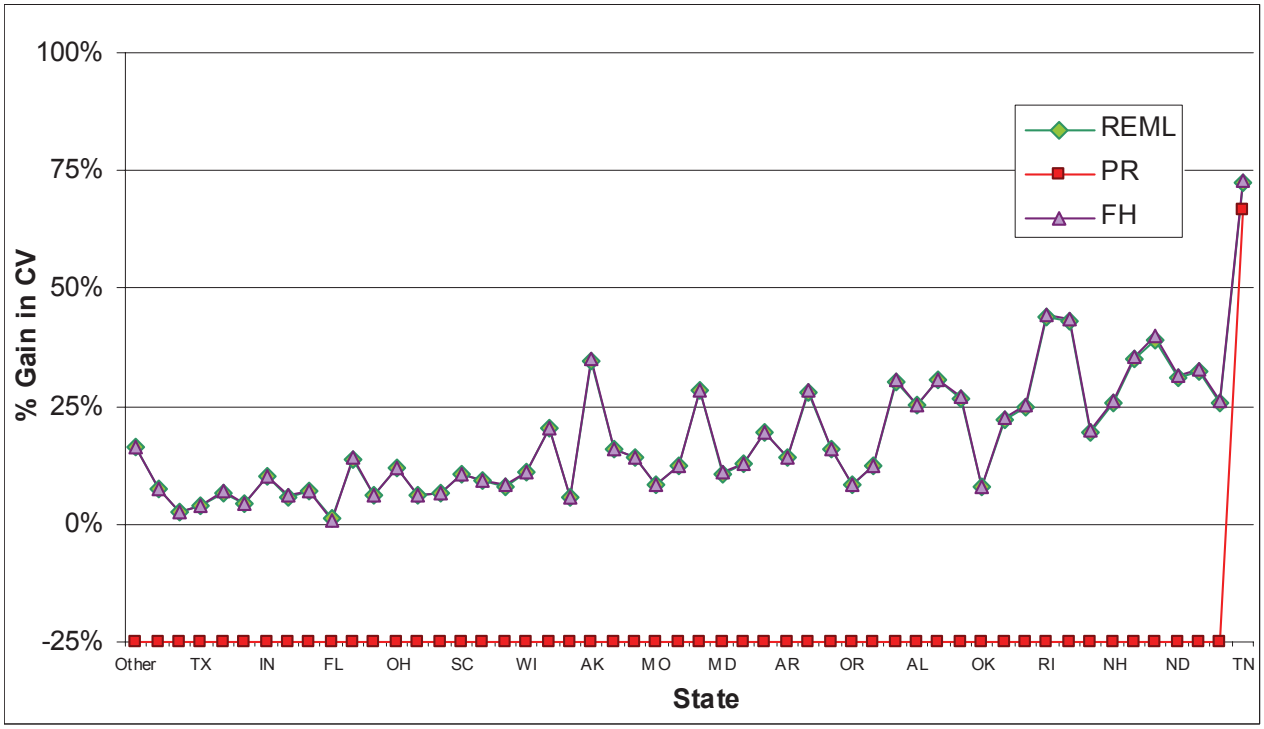
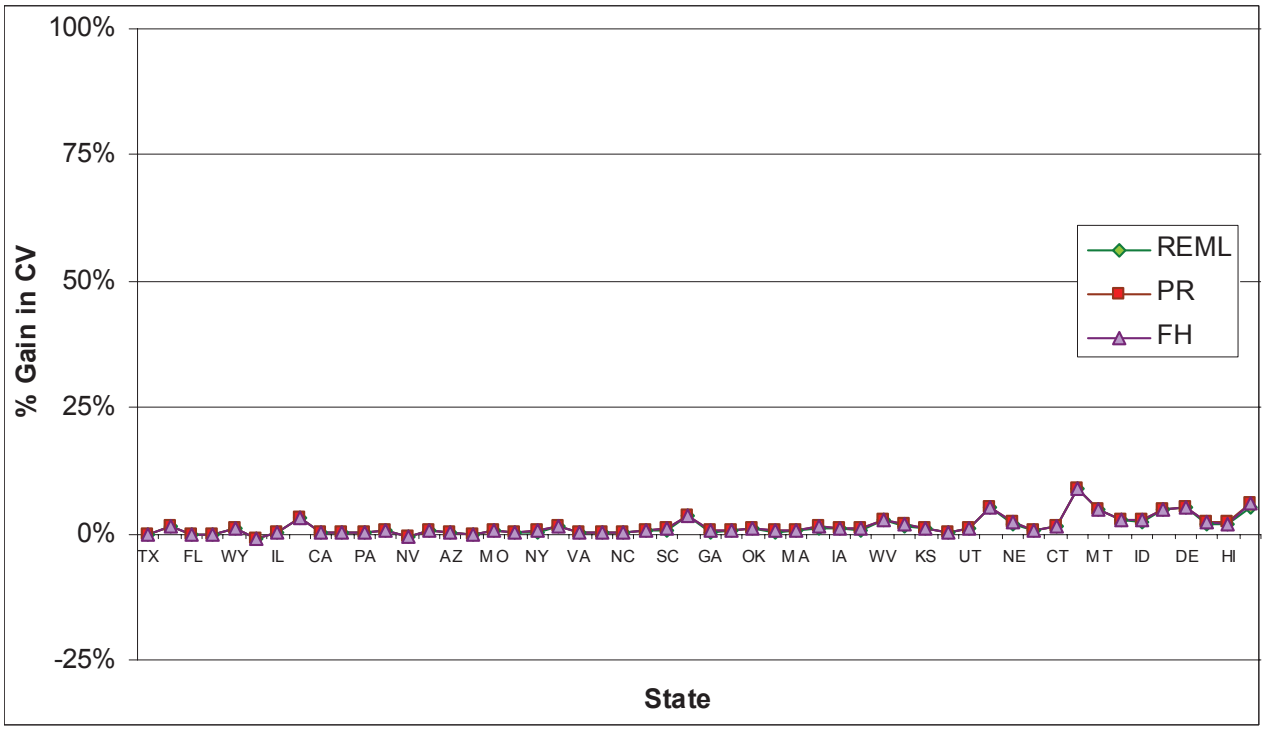


Real Estate Taxes



A.4. Percentage Relative Gain in EBLUP Estimates over the Coefficients of Variation of SOI Sample-Based Estimates, by Variable of Interest—Continued

State & Local Income Taxes



State & Local General Sales Taxes