Using Sample Data to Reduce Nonsampling Error in State-Level Estimates Produced from Tax Records

Kimberly Henry[1], *Partha Lahiri*[2], *and Jana Scali*[1]
[1]*Internal Revenue Service, P.O. Box 2608, Washington DC, 20013-2608*
[2]*University of Maryland, 1218 LeFrak Hall, College Park, MD 20742*

Abstract: For developing public policies and research purposes, income-related statistics are frequently needed for different small geographic regions. Previous research using the Statistics of Income (SOI) Division's Individual sample suggests that some IRS data, though free from the usual sampling error encountered in small area estimation, can be subject to nonsampling error. However, the SOI sample estimates, based on a large national sample of cleaned tax data, are subject to sampling variability for small domains. We use empirical and hierarchical Bayes methods to improve estimators of small-area totals and apply our estimators to data from SOI's 2004 and 2005 samples to evaluate the impact of an increased sample size.

Key words: Survey sampling, Administrative Records, Indirect Estimators

1. Introduction: Small Area Estimation with IRS Data and Associated Nonsampling Errors

The 135 million individual income tax records on the Internal Revenue Service's (IRS) annual individual returns transaction file have several uses to multiple government agencies. These data serve as the sampling frame for various IRS functions, including the Statistics of Income (SOI) Division of IRS. SOI uses the data to publish tabulated monetary amounts and the associated number of returns by state and Adjusted Gross Income categories (in 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 (2009) models to estimate the number of children in poverty in 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 error. Generally, only tax items necessary for computer processing of a tax return are retained on the IRS file, as opposed to items needed for statistical and tax policy research. Also, measurement errors can exist between IRS and SOI data values due to different data editing rules, as discovered when comparing records in the IRS file to the same returns in SOI's sample. For revenue processing purposes, IRS does not spend scarce resources correcting errors that do not affect tax liability in the more than 135 million individual income 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 can adversely affect the data's statistical usability for variables indirectly related to tax liability. Other IRS data limitations include a smaller amount of information available than SOI's sample, the IRS data are often provided to SOI in tabular form with monetary amounts rounded to thousands, and certain high income taxpayers are omitted.

The SOI Division of IRS draws large annual samples of tax returns to produce richer and cleaner data for population estimation and tax modeling purposes. SOI's transcription and editing staff receive more extensive training to transcribe, clean, and edit the data, the sample is augmented with additional items from the return, and the data is more closely monitored and checked for consistency. However, the state is not within the sample design, so sample-based state-level estimates have the usual sampling error problem.

To improve on design-based estimators, several indirect and model-based methods have been proposed in the literature. These estimation procedures essentially use *implicit* or *explicit* models that *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 (defined by populations less than 1,000), Fay and Herriot (1979) used an empirical Bayes (EB) method that combined the U.S. Current Population Survey data with various administrative and census records. 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 EB estimator (also an empirical best linear unbiased predictor, or 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 produced and the unknown variance component(s) is (are) estimated by a standard method [e.g., maximum likelihood, residual maximum likelihood, analysis-of-variance, 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. Rao (2003) and Jiang and Lahiri (2006) provide reviews of the Fay-Herriot method and its extensions.

In section 2, we describe the SOI sample data and analysis data descriptions. In section 3, we introduce the direct estimators that we used in our modeling and evaluation studies. In section 4, we introduce the area level model and the associated EBLUP methodology. To overcome the likelihood-based methods' problem of potential zero variance component estimates, we also introduce a simple hierarchical Bayesian approach in section 5. We describe our evaluation study and present results in section 6.

2. SOI Sample and Analysis Data Descriptions

The SOI Division selects large stratified Bernoulli samples of tax returns weekly, as they are processed by the IRS. Stratification for the sample uses various criteria, including size of total gross positive and negative income and an indicator for the returns' "degree of interest" for tax modeling purposes, to create 208 strata. The sample consists of two parts within each stratum. First, a 0.05 percent Bernoulli sample is selected, called the Continuous Work History Sample (CWHS, Weber 2004). A separate Bernoulli sample is also selected independently from each stratum, with rates ranging from 0.01 to 100 percent. The full sample, which itself is also a Bernoulli sample, consists of the CWHS plus all additional returns selected with unequal probabilities of selection across strata.

Each SOI study corresponds to a "Tax Year" (TY), which for individual tax returns involves income and financial information earned by U.S. taxpayers in the previous calendar year. For example, the TY 2004 sample, where 200,778 returns were selected from 133,189,982, reflected income earned in 2004 and reported to IRS by December 2005. For TY 2005, CWHS sampling rates were increased to 0.1 percent and 292,966 returns were selected from 134,494,440. More detail is given in Testa and Scali (2005).

The reduced dataset for this analysis was created by first separating the samples into the certainty (i.e., sample units with weights equal to one) and noncertainty (units with weights greater than one) units. We placed the 34,309 TY 2004 and 44,482 TY 2005 returns that SOI sampled with certainty each year into two certainty strata (one for each year), since they represented a census of tax returns. Thus, without loss of generality, we exclude these strata from the population and develop our estimation method to estimate totals from all other strata, then add the certainty strata total of SOI-transcribed values to our estimate from the remaining noncertainty strata for the entire population. For both the certainty and noncertainty datasets, the weighted sample data were tabulated to the state-level for the 50 U.S. states and Washington DC. We excluded the "other" state category of tax returns filed by civilians and military individuals living abroad, in U.S. possessions and territories, Puerto Rico, etc. This corresponded to 1,877 returns in 2004 and 5,186 in 2005, of which 683 and 3,543 were certainty units, respectively.

We selected six variables of interest from different parts of the 1040 tax return, which are more or less susceptible to errors in the IRS data: Adjusted Gross Income, Taxable Interest Income, Earned Income Tax Credit, Real Estate Taxes Deducted, State and Local Income Taxes Deducted, and State and Local General Sales Taxes Deducted. A description of each item is given in Table 1 (from IRS 2006).

As the domain is not in the SOI sample design, both the sample sizes across states and sample weights of units within the same state vary. Figure 1 shows the state sample sizes for both tax years. To see the impact of these varying sample sizes on sample-based estimates, Figure 2 shows the percent relative difference in the number population units from the TY 2004 and 2005 IRS frames and the corresponding population size estimated using the SOI sample weights. As the number of sample units decreases, the two state population sizes vary due to sampling error. The 2005 relative differences are lower for smaller states since the sample size is larger, but the pattern is the same. To overcome this, we use state-level poststratification adjustments to the SOI sample estimates of totals in our evaluation.

For evaluation purposes, we also collapsed the 51 states into groups based on different criteria of "similarity." The SOI-sample based state group-level estimates then have lower sampling error due to an increased group sample size, i.e., the direct estimates are more reasonable. Six states were considered large enough, with more than 5,000 noncertainty returns. The remaining states were grouped based on whether or

not the state had state income taxes, geographic region, and a relative size of income. This resulted in 21 groups, which are listed with the associated number of certainty and noncertainty sample units in Table 2.

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

Variable	*Description* [a]
Adjusted Gross Income	Income reported from the calculate of total income (Line 37, Form 1040) (pp. 119-120).
Taxable Interest Income	Taxable portion of interest received (Line 8a, Form 1040) (p. 146).
Earned Income Tax Credit	Taxpayer credit for working lower-income individuals (Line 66a, Form 1040) (pp. 125-126).
Real Estate Taxes Deducted	Taxes paid on real estate owned and not used for business (Line 6, Schedule A) (p. 138).
State and Local Income Taxes Deducted	Taxes withheld from salary, paid directly, or made to state disability funds (Line 5a, Schedule A) (p. 143).
State and Local General Sales Taxes Deducted	Sales Taxes incurred by individuals (Line 5b, Schedule A), (p. 143).

a: page numbers from IRS 2007.

Figure 1. State Sample Sizes

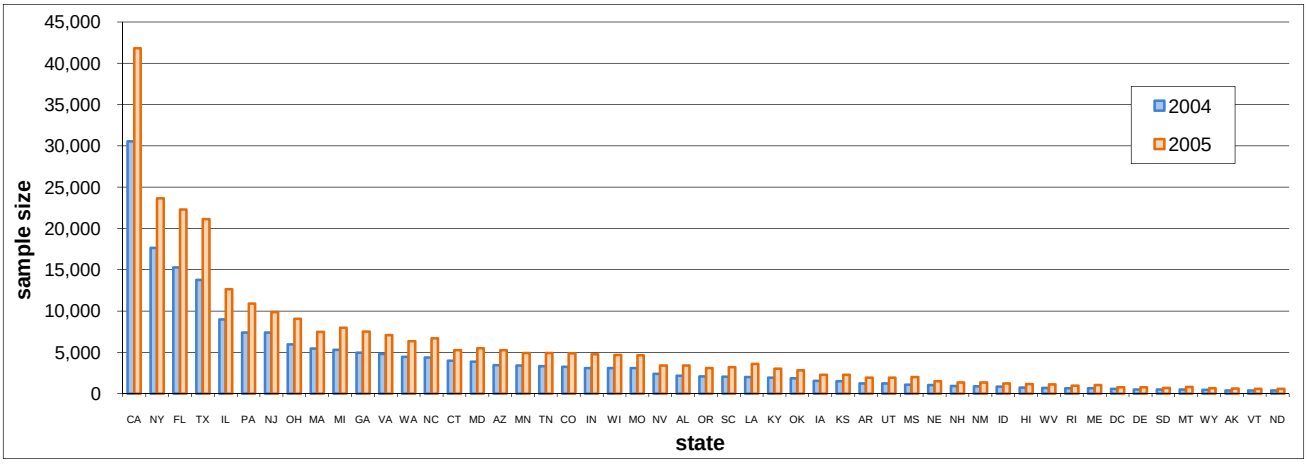


Figure 2. Percent Relative Differences Between IRS Frame and SOI Sample-Estimated Population Sizes

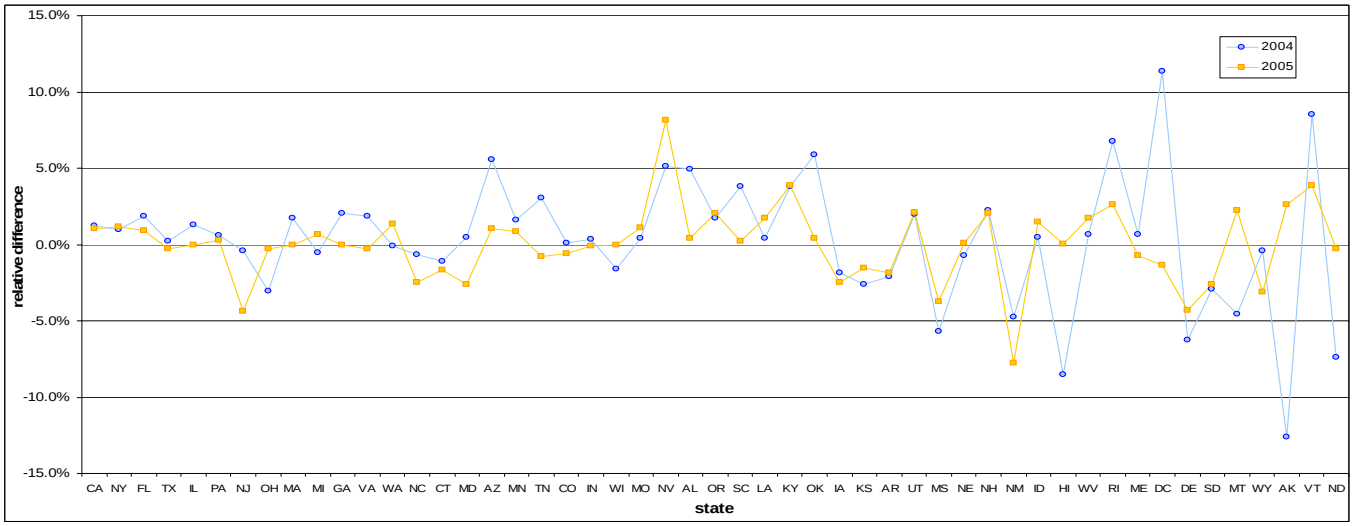


Table 2. Number (#) of Certainty and Noncertainty Sample Units, by State Groups

	Tax 2004			*Tax Year 2005*		
States Within Group	# certainty	# noncertainty	Total	# certainty	# noncertainty	Total
California	6,541	23,990	30,531	8,419	33,415	41,834
Florida, Tennessee	4,053	14,566	18,619	5,442	21,757	27,199
New York	4,528	13,101	17,629	5,283	18,347	23,630
Texas	2,319	11,427	13,746	3,439	17,700	21,139
Michigan, Wisconsin, Minnesota	1,448	10,379	11,827	1,814	15,797	17,611
Georgia, North Carolina, South Carolina	1,265	10,108	11,373	1,663	15,806	17,469
Indiana, Ohio, Kentucky	1,135	9,908	11,043	1,366	15,465	16,831
Connecticut, Rhode Island, Massachusetts	2,214	7,952	10,166	2,699	11,013	13,712
Iowa, Nebraska, Kansas, Missouri, Oklahoma	997	8,061	9,058	1,276	12,302	13,578
Illinois	1,539	7,451	8,990	1,816	10,832	12,648
Arizona, New Mexico, Utah, Colorado	1,432	7,415	8,847	1,967	11,482	13,449
Pennsylvania	932	6,480	7,412	1,210	9,695	10,905
New Jersey	1,273	6,138	7,411	1,520	8,367	9,887
Arkansas, Alabama, Mississippi, Louisiana	620	5,927	6,547	1,504	9,449	10,953
Virginia, West Virginia	731	4,798	5,529	927	7,278	8,205
Washington DC, Maryland, Delaware	777	4,180	4,957	1,024	6,028	7,052
Alaska, Washington	812	4,024	4,836	1,039	5,916	6,955
Montana, North Dakota, Idaho, Oregon	435	3,364	3,799	503	5,239	5,742
Nevada, Wyoming, South Dakota	935	2,450	3,385	1,144	3,600	4,744
Maine, Vermont, New Hampshire	215	1,770	1,985	299	2,656	2,955
Hawaii	108	621	729	128	1,025	1,153
Total	34,309	164,110	198,419	44,482	243,169	287,651

The state of Hawaii (HI) was previously grouped with the "other" category of states. Rather than group HI with dissimilar states, we retain it in a group by itself to illustrate the alternative estimators' performance under smaller sample sizes.

3. Direct Estimators

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

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

Let *s* denote the sample of tax returns drawn from the population of tax returns, $s_d \subset s$ the part of the sample in domain *d* of interest, and w_k the sampling weight for the *k*-th sampled tax return, $k \in s$. The sampling weight w_k is the inverse of the inclusion probability, adjusted for achieved population and sample sizes. As described in section 1, all formulas concern estimating noncertainty strata of the Tax Year 2004 and 2005 populations.

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. However, weights vary across strata and within a given domain. Let

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

denote the population total for the *d*-th domain (excluding the units belonging to the certainty stratum). We estimate the population total with the following design-unbiased *uncalibrated direct estimator*:

$$\hat{Y}_d = \sum_{k \in s_d} w_k y_k \qquad (1)$$

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 weighted sample mean as the design-based direct estimator of $\bar{Y}_d$:

$$\bar{y}_{dw} = \sum_{k \in s_d} w_k y_k \Big/ \sum_{k \in s_d} w_k . \quad (2)$$

Since the state is not in the sample design, we apply a simple state-level post-stratification adjustment to (1) and obtain the *direct calibrated* (CAL) *estimator*:

$$\hat{Y}_d^{CAL} = N_d \bar{y}_{dw} \quad (3)$$

Estimator (3), which is approximately design-unbiased in large samples, is used to compare alternative model-based state group-level estimates. Estimator (1) is used to evaluate estimates of national-level totals.

4. EBLUP Estimators

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

$$\begin{aligned} &\text{Level 1: } \bar{y}_{dw} \overset{\text{ind}}{\sim} N(\bar{Y}_d, D_d); \\ &\text{Level 2: } \bar{Y}_d \overset{\text{ind}}{\sim} N(x_d^T \beta, A), \end{aligned} \quad (4)$$

where D_d is the estimated sampling variance of $\bar{y}_{dw}$, $\bar{Y}_d$ is the true population mean, and $x_d^T = [1 \quad \bar{x}_d]$, where $\bar{x}_d$ is the mean of the same variable based on IRS tabular data.

The main sources of error in the IRS means are the nonsampling error described in section 1, while the SOI means are subject to sampling error, which is reduced in the TY 2005 estimates due to the increased sample size. Figure A.1 contains plots of $\bar{y}_{dw}$ versus $\bar{x}_d$ for each variable in 2004 and 2005. Although the estimates $\bar{y}_{dw}$ are subject to sampling variability, a strong linear relationship is still observed between these estimates and $\bar{x}_d$ for each variable, particularly for variables less affected by IRS errors. We take advantage of this relationship in Level 2 of model (4).

Under model (4), the *best predictor* (BP) of $\bar{Y}_d$ is given by:

$$\hat{\bar{Y}}_d^{BP} = (1 - B_d)\bar{y}_{dw} + B_d x_d^T \beta, \quad (5)$$

where $B_d = \dfrac{D_d}{D_d + A}$. If A is known, then β is estimated by the weighted least squares estimator:

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

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

$$\hat{\bar{Y}}_d^{EBP} = (1 - B_d)\bar{y}_{dw} + B_d x_d^T \hat{\beta}(A). \quad (6)$$

Note that $\hat{\bar{Y}}_d^{EBP} \equiv \hat{\bar{Y}}_d^{BLUP}$, 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 $v_d \sim (0, A)$ and $e_d \sim (0, D_d)$. When both β and A are unknown, we propose the following *empirical best linear unbiased predictor* (EBLUP) of $\bar{Y}_d$:

$$\hat{\bar{Y}}_d^{EBLUP} = (1 - \hat{B}_d)\bar{y}_{dw} + \hat{B}_d x_d^T \hat{\beta}(\hat{A}), \quad (7)$$

where $\hat{B}_d = \dfrac{D_d}{D_d + \hat{A}}$ and $\hat{A}$ is any standard consistent estimator of A. In this paper, we consider the residual maximum likelihood (REML) estimator of A.

The EBLUP approach has several advantages for producing point estimates of the state-level means.. However, the standard likelihood and analysis-of-variance-based methods can numerically yield zero variance component estimates

We define the *mean square prediction error* (MSPE) 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:

$$mspe_d^N = g_{1i}(\hat{A}) + g_{2i}(\hat{A}), \tag{8}$$

where $g_{1d}(\hat{A}) = \hat{B}_d \hat{A}$, $g_{2d}(\hat{A}) = \hat{B}_d^2 h_{dd}$, and $h_{dd} = x_d^T \left(\sum_{j=1}^{m} \frac{1}{D_j + \hat{A}} x_j x_j^T \right)^{-1} x_d$. This is referred to as the "naïve MSPE estimator," since it does not incorporate the additional uncertainty due to the estimation of A. Prasad and Rao (1990) showed that the order of this underestimation is $O(m^{-1})$ under certain regularity conditions.

Figure A.2 shows the resulting estimated shrinkage factors ($\hat{B}_d$, the weight given to the regression estimate in (7)) for the 50 states and the District of Columbia in 2004 (patterns for 2005 were similar and thus omitted). For each variable, the states are sorted by D_d. The effect $\hat{A} = 0$ is that all of the weight is given to the regression estimate $x_d^T \hat{\beta}$, i.e., in estimating the state-level means, $\hat{B}_d = 1$ in (7), so that $\hat{\bar{Y}}_d^{EBLUP} = x_d^T \hat{\beta}(A)$. This applies to all states, regardless of the state's sampling variance, and occurred for six out of our twelve combinations of variables and tax years. This is unreasonable since we would like to use as much of the SOI sample information as possible, particularly for the larger states with lower sampling variance. Specifically, for 2004, this occurred for Adjusted Gross Income, Taxable Interest Income, Real Estate Taxes Deducted, and State and Local Income Taxes Deducted. For 2005, this occurred for Earned Income Tax Credit and Real Estate Taxes Deducted.

5. Hierarchical Bayes Models

To ensure that we always incorporate the SOI sample data in our state-level estimates, we now introduce a simple hierarchical Bayesian model to overcome the problem associated with the classical method of estimating A. However, this requires stronger model assumptions to evaluate. For $d = 1, \cdots, m$, assume

$$\begin{aligned} &\text{Level 1: } \bar{y}_{dw} \overset{\text{ind}}{\sim} N(\bar{Y}_d, D_d); \\ &\text{Level 2: } \bar{Y}_d \overset{\text{ind}}{\sim} N(x_d^T \beta, A); \\ &\text{Level 3: } \beta \sim Unif\, \Re^2,\ A \sim \pi(A). \end{aligned} \tag{9}$$

The first two levels of this model are identical to model (4) used to produce the EBLUP estimates described in section 4. Our theoretical motivation for model (9) is that, when A is known, the Uniform prior on the two hyperparameters in β produces the BLUP estimate (7) for $\bar{Y}_d$, i.e., $\hat{\bar{Y}}_d^{HB} = \hat{\bar{Y}}_d^{BLUP}$. We consider two alternative prior distributions for the hyperparameter A: a Uniform (*Unif*) prior, denoted by $\pi_1(A) = Unif\,(0, U)$ and an Inverse Gamma (*IG*) prior, denoted by $\pi_2(A) = IG(0.001, 0.001)$. The hierarchical Bayes estimators of $\bar{Y}_d$ corresponding to the priors $\pi_1(A)$ and $\pi_2(A)$ are denoted by $\hat{\bar{Y}}_d^{HB(1)}$ and $\hat{\bar{Y}}_d^{HB(2)}$, respectively.

These particular prior distributions are generally noninformative (or "flat") priors that have been used in similar variance component models (see, e.g., Gelman 2006). For both tax years and all variables, we chose $U = 10{,}000{,}000$. This upper bound creates a uniform prior that is very flat, while the $IG(0.001, 0.001)$ prior is a commonly used prior in the Bayesian literature. One advantage of the HB approach over the classical method is that it guarantees a strictly positive estimate of A. Another advantage of the HB approach is that we can use the posterior distributions for the parameters of interest for all related

inferences. We estimate the parameters by their posterior means and measure the uncertainty of the point estimators by the corresponding posterior variances. For interval estimation, the Bayesian approach uses credible intervals, which are much easier to compute and often easier for non-statisticians to interpret than the corresponding MSPE-based estimates described in section 4.

For both priors, the posterior distributions of the parameters of interest do not have closed-form solutions. We used 100,000 iterations (after a burn-in of 5,000 iterations) of the Markov Chain Monte Carlo (MCMC) algorithm to approximate the posterior distributions. The MCMC error in estimating A was found to be very low.

For all variables and both years, the $IG(0.001, 0.001)$ prior created posterior distributions for A that were very skewed towards zero. Again, this affects the estimated shrinkage factor for all states in A.2. The effect here is similar to $\hat{A}^{REML} = 0$: given a very small $\hat{A}$ with the IG prior, most of the weight is given to the regression estimate $x_d^T \beta$ to produce $\hat{\bar{Y}}_d^{HB(2)}$. This also occurred regardless of the state's sampling variance and is unreasonable since we ignore most of the SOI sample information in $\hat{\bar{Y}}_d^{HB(2)}$ for larger states with lower sampling variance. However, the uniform prior assigned more weight (i.e., $\hat{B}_d < 0.5$) to the SOI mean for larger states, as illustrated in A.2.

6. Results

Here we consider three evaluations of our direct, EB and HB model-based totals: how well estimated state-level totals add up to the state-group totals, the national-level totals, and the precision of the state-level estimates. First, to evaluate the EB and HB estimates and the resulting totals, we calculated the alternative means $\hat{\bar{Y}}_d$, estimated the total of the noncertainty units with $N_d \hat{\bar{Y}}_d$, and added it to the variables' total from the certainty units for each state. We then collapsed the states into the twenty-one groups shown in Table 2 and used the difference relative to the calibrated total (1) to evaluate the alternatives:

$$\% \text{ Rel Diff} = 100 \times \frac{\sum_{d \in g} \hat{Y}_d^{CAL} - \sum_{d \in g} N_d \hat{\bar{Y}}_d}{\sum_{d \in g} \hat{Y}_d^{CAL}}, g = 1, \ldots, 21.$$

The calibrated SOI sample estimate in (3) is used to gauge how well the alternative model-based estimates estimate the state-group totals, since the large state-group sample sizes shown in Table 2 reduce the sampling error in the SOI estimates significantly. That is, preferable model-based state-level estimates, when added up within groups, are those closest to the SOI sample estimates. Figure A.3 shows the plots of the percent relative differences for the alternative state-group totals, for each variable of interest and alternative totals. For all variables, the states were sorted by descending state group sample size; as the group sample size decreases, the percent relative differences increase. For all groups, the model-based estimates are closer to the calibrated SOI state group total than the uncalibrated and IRS-based totals. The exception was HI, where the uncalibrated SOI sample total for this state was closest to the calibrated total (with exceptions, where the uncalibrated SOI sample total for HI was furthest from the calibrated total).

Table A.4 shows the absolute relative percent relative differences in A.3 for 2004, averaged across 20 of the state groups (HI was excluded for this summary measure). That is,

$$\text{Ave} \left| \% \text{ Rel Diff} \right| = \frac{1}{20} \sum_{g=1}^{20} \left| 100 \times \frac{\sum_{d \in g} \hat{Y}_d^{CAL} - \sum_{d \in g} N_d \hat{\bar{Y}}_d}{\sum_{d \in g} \hat{Y}_d^{CAL}} \right|.$$

The absolute value was used to avoid large positive and negative differences canceling each other out. For all variables and years, the HB uniform prior model had the lowest average percent relative difference across the state groups. The omitted 2005 relative differences were smaller, with the same patterns.

Second, we evaluate the alternative estimates at a national-level. When aggregating the state-level totals to the national-level, preferable model-based estimates should be close to the uncalibrated SOI sample-based estimates in (1), the estimator based upon the sample design strata. This total is used for evaluation of national-level totals since the sample is large enough to estimate them with low sampling error. Table A.5 shows the uncalibrated national-level total of each variable and the percent relative difference between it and the IRS data and alternative model-based estimates:

$$\% \text{ Rel Diff} = 100 \times \frac{\sum_{k \in s} w_k y_k - \sum_{d=1}^{51} \sum_{s_d} N_d \hat{\bar{Y}}_d}{\sum_{k \in s} w_k y_k} . .$$

For both years and all variables, the calibrated total in (3) was closest to the uncalibrated total, having the lowest percent relative difference. The HB model with the uniform prior was second closest. The IRS data-based totals are the furthest from the uncalibrated SOI totals due to the nonsampling error described in section 1. While all percent relative differences appear small, they correspond to very large differences in the totals measured in terms of dollar amounts (in the millions or billions).

Last, for the precision of the alternative state-level estimates, Figure A.6 shows the coefficients of variation (CV) for each variable in 2004. That is,

$$CV\left(\hat{Y}_d\right) = 100 \times N_d Var\left(\hat{Y}_d\right) \Big/ \hat{Y}_d \ .$$

For the EBLUP estimates, the MSPE estimates from (10) were used to estimate $CV(N_d \hat{\bar{Y}}_d^{EBLUP})$. For the HB estimates, the posterior variances were used to calculate $CV(N_d \hat{\bar{Y}}_d^{HB})$. For all variables and tax years, the SOI sample CVs increase as the state sample size decreases and become less stable for the smallest states, as the estimates of $Var\left(\hat{Y}_d\right)$ are also subject to sampling error. However, the model-based CV's are much more stable across states, due to the strong linear relationship noted between the IRS and SOI state-level means, used in Level 2 of the EB and HB models and shown in A.1. The HB IG prior model had the lowest CV's, but we need to consider the pattern: CV's for larger states are nearly identical to those for smaller states, which intuitively does not make sense. However, the HB uniform and EB REML-based CV's increase as the sample size decreases, as expected. Due to the larger 2005 sample sizes, the largest SOI sample CV's were 1-5% lower, but the 2005 results agreed with those for 2004, so they were omitted.

7. Conclusions and Limitations

We attempt to improve population-based estimates from administrative tax return data that are subject to nonsampling error and sample-based estimates subject to sampling error. Both EBLUP and HB approaches seem to produce results preferable over those produced using only the SOI sample or IRS frame data. They were obtained by exploiting relationships between the sample and population variable means and removing nonsampling errors in the certainty units' totals by using only the SOI data for these returns. This was demonstrated by gains in precision reflected by lower estimates of the coefficients of variation in the state-level totals and more stability in the estimates themselves when combined to the state-group level and compared to the calibrated SOI sample totals. In addition, when combined across all states, the model-basd estimates of state-level totals also produced national-level totals that were more consistent to those produced from the SOI sample than the IRS data.

Between our alternative estimators, while the prior specification in our small-area model is subjective, the resulting state-level mean estimates between the EB and HB Uniform prior results are very close when $\hat{A}$ is nonzero. This provides empirical support to the choice of model (7). Of the twelve tax return variable/tax year combinations we examined, six of the REML-based shrinkage factors were equal to one for all states, resulting in use of only the regression-based component to estimate the state-level mean. However, the Uniform and Inverse Gamma priors in the HB model both produced positive estimates of A in all cases. While the IG prior produced state-level estimates very close to the regression estimates and EB estimates when $\hat{A}^{REML} = 0$, the uniform prior seemed to work well for all six variables of interest and two tax years. The resulting model-based state-level estimates from this HB model use more SOI sample information for larger states, which is intuitively sensible from a design-based perspective. The HB approach of using the posterior distributions for all inference is also easier to both produce and interpret.

REFERENCES

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

Fay, R. E., and Herriot, R. A. (1979). "Estimates of Income for Small Places: an Application of James-

Stein Procedure to Census Data." *Journal of American Statistical Association,* **74**, 269-277.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2004). *Bayesian Data Analysis* (2nd edition eds.). Boca Raton: Chapman & Hall/CRC, pp. 182–184.

Gelman, A. (2006) "Prior distributions for variance parameters in hierarchical models." *Bayesian Analysis,* **1,** No. 3, pp. 515–533.

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

Internal Revenue Service (2007), "Explanation of Terms," *Statistics of Income – 2005 Individual Income Tax Returns, Internal Revenue Service, Publication 1304*, pp. 119-149.

Jiang, J., and Lahiri, P. (2006). "Mixed model prediction and small area estimation (with discussions)." *Test*, **15**, 1, 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 Rao, J.N.K. (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 & Sons: New York.

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

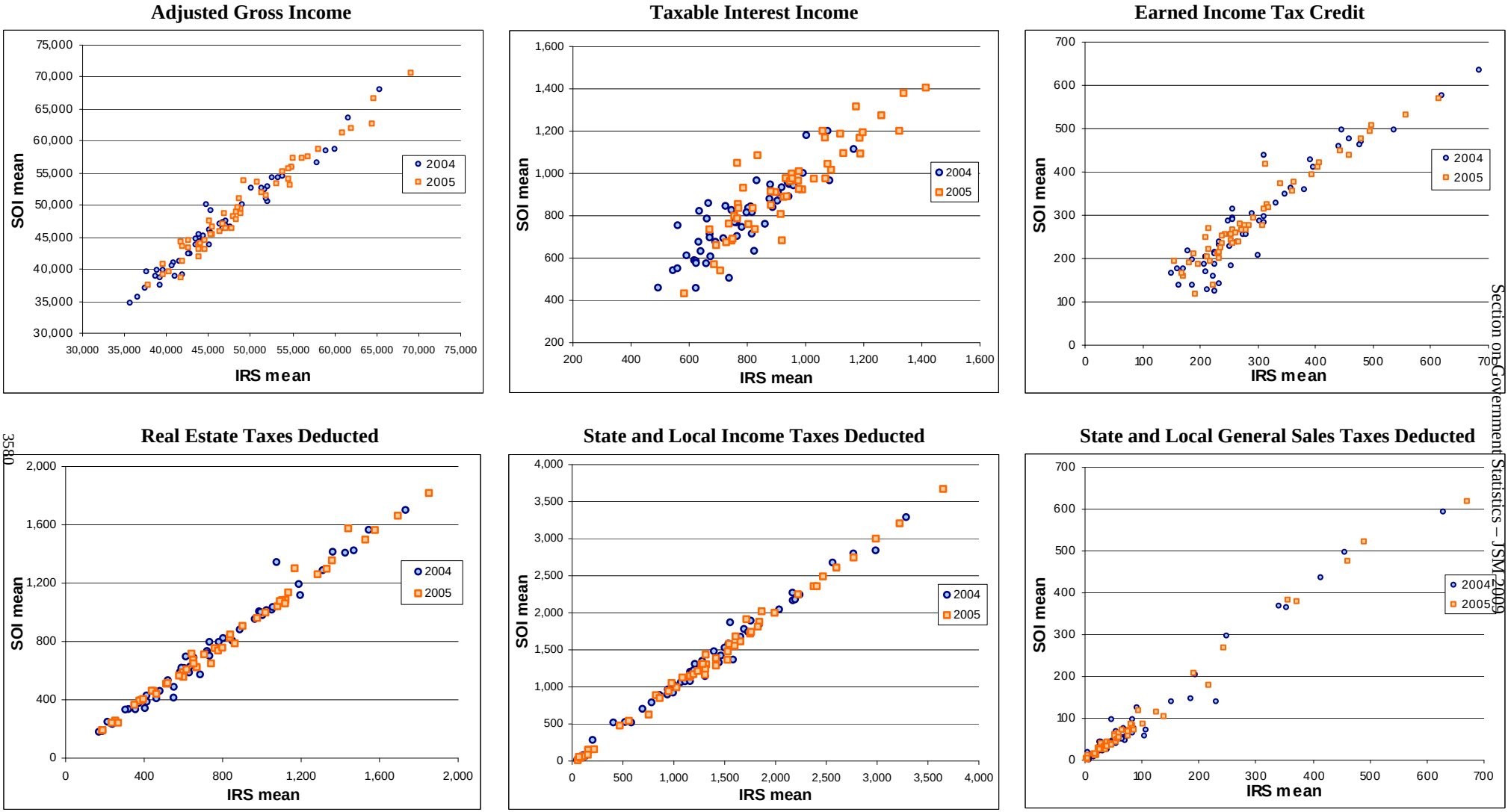
Scali, J. and Testa, V. (2006), *Statistics of Income – 2004 Individual Income Tax Returns, Internal Revenue Service, Publication 1304*, pp. 23-27.

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," http://www.irs.gov/pub/irs-soi/04webasa.pdf.

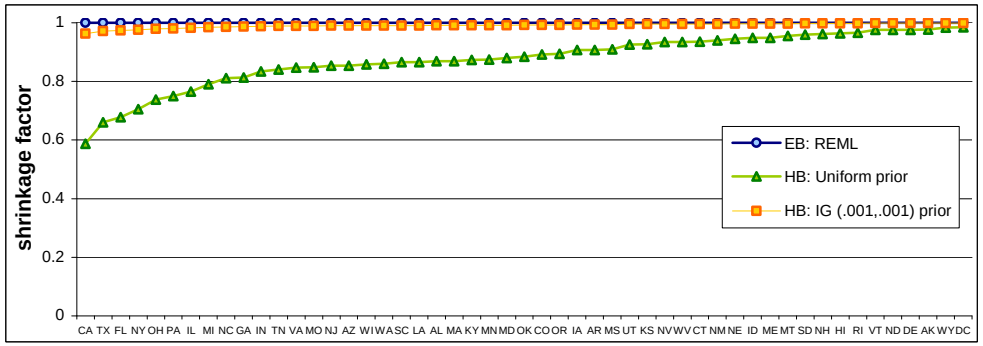
Wolter, K. M (1985). Introduction to Variance Estimation. New York: Springer-Verlag, pp. 201-217.

Figure A.1. IRS vs. SOI Mean Plots, noncertainty sample and frame units (note differences in scale)

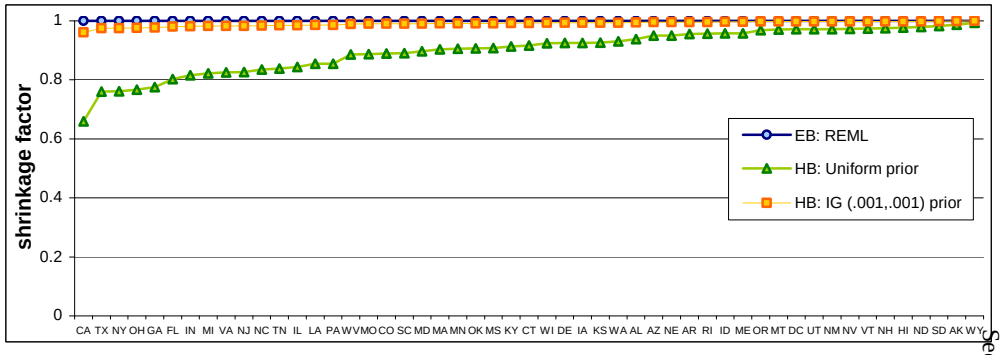


A.2. Estimated Shrinkage Factors, $\hat{B}_d$, Tax Year 2004

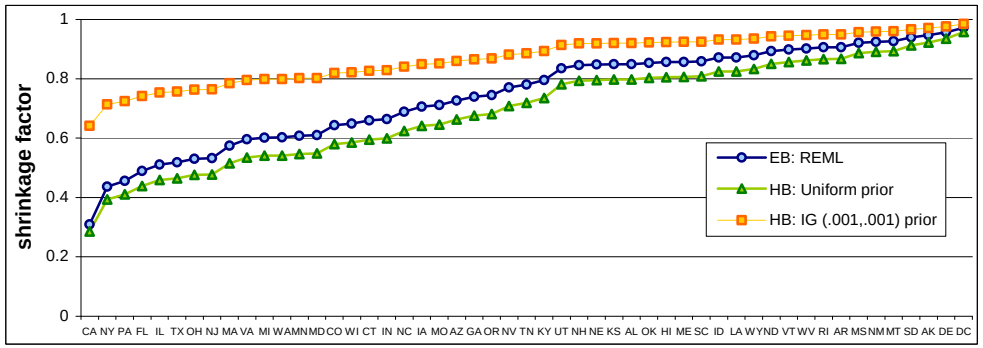
Adjusted Gross Income



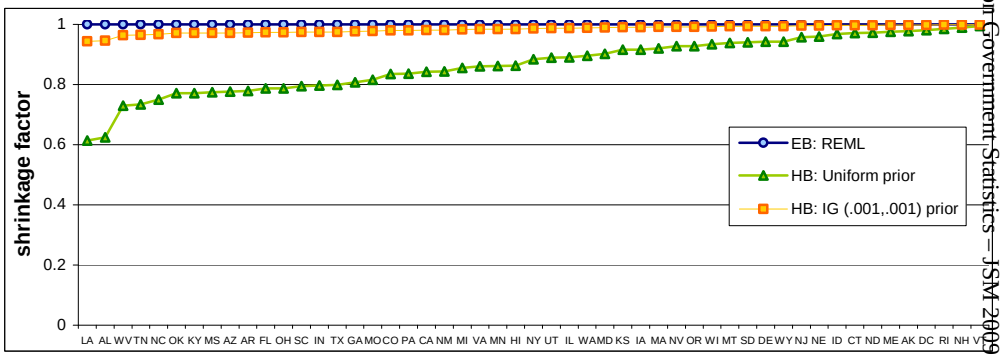
Taxable Interest Income



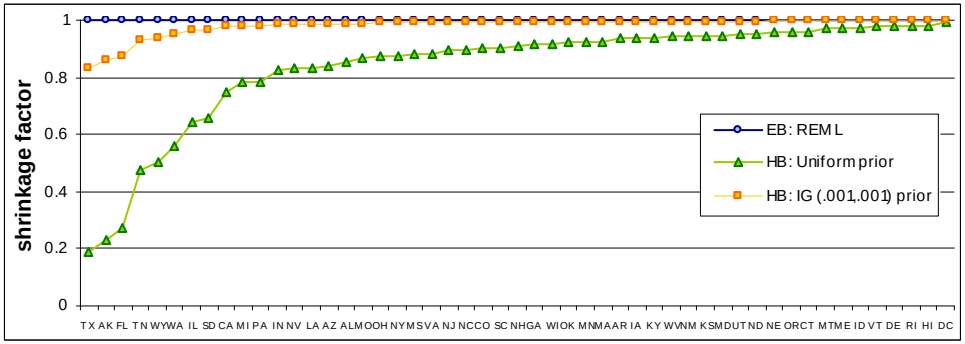
Earned Income Tax Credit



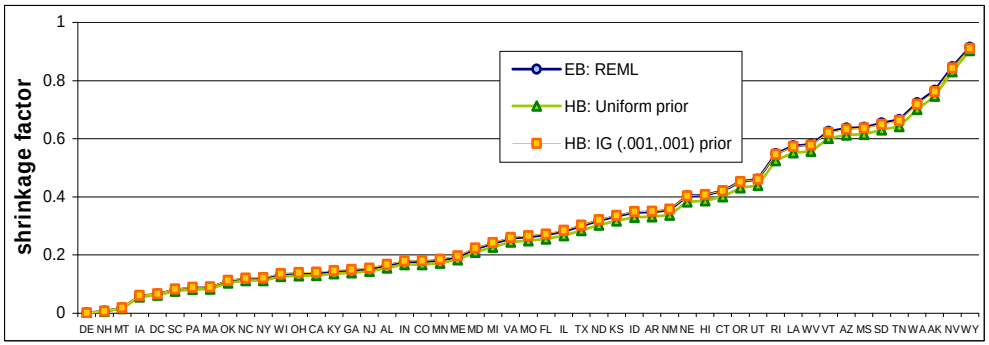
Real Estate Taxes



State and Local State Income Taxes

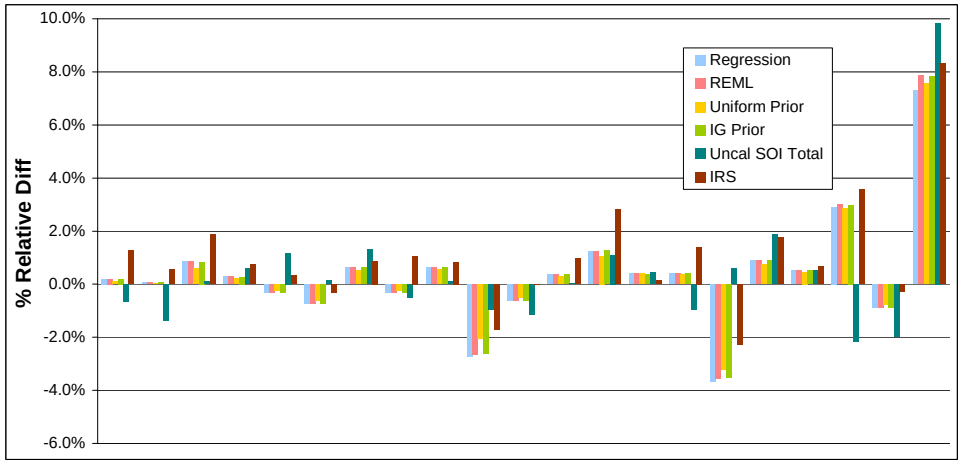


State and Local General Sales Taxes

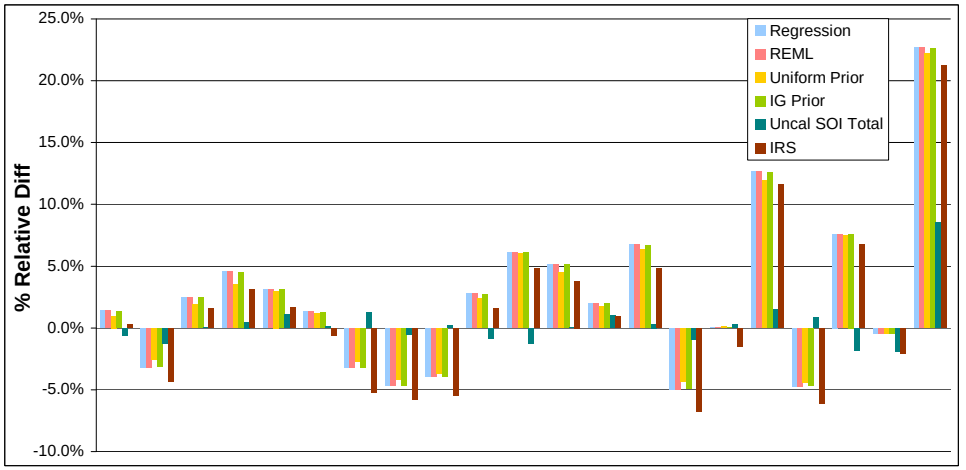


A.3. Percent Relative Differences Between Alternative Totals and SOI Sample Estimates to Calibrated State Group Totals, Tax Year 2004 (note differences in scale)

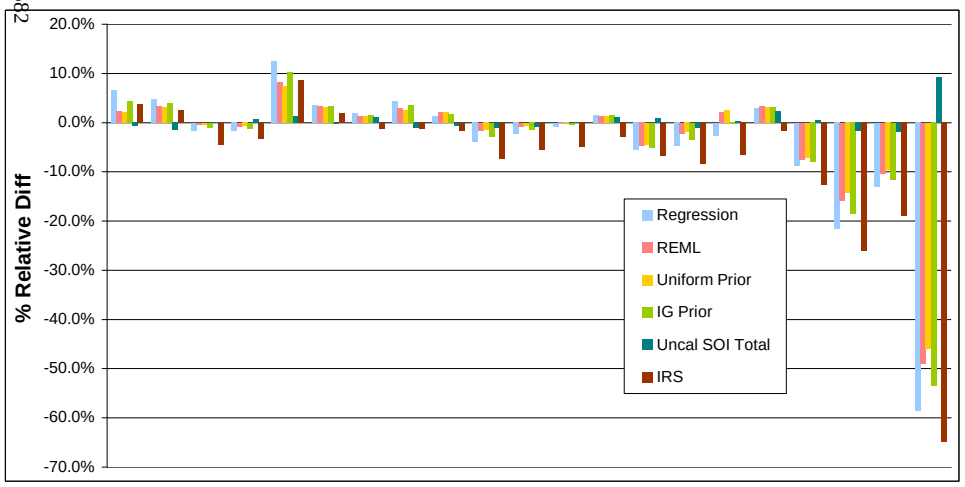
Adjusted Gross Income



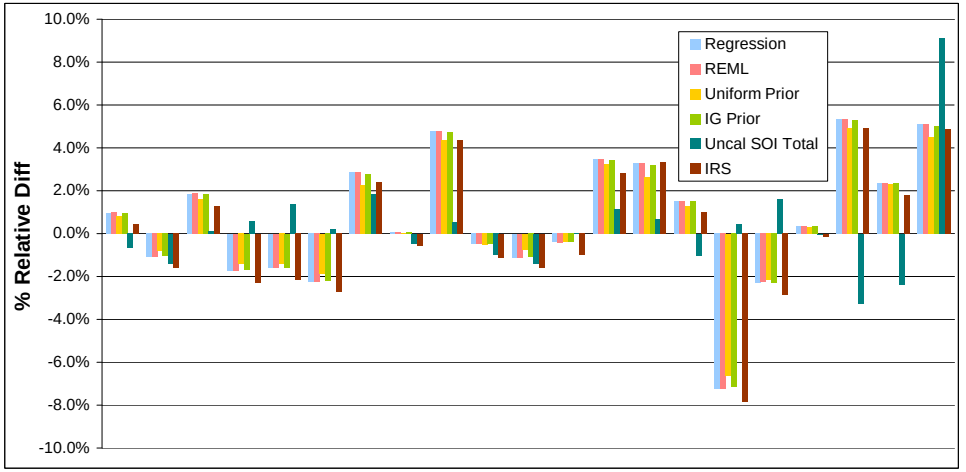
Taxable Interest Income



Earned Income Tax Credit

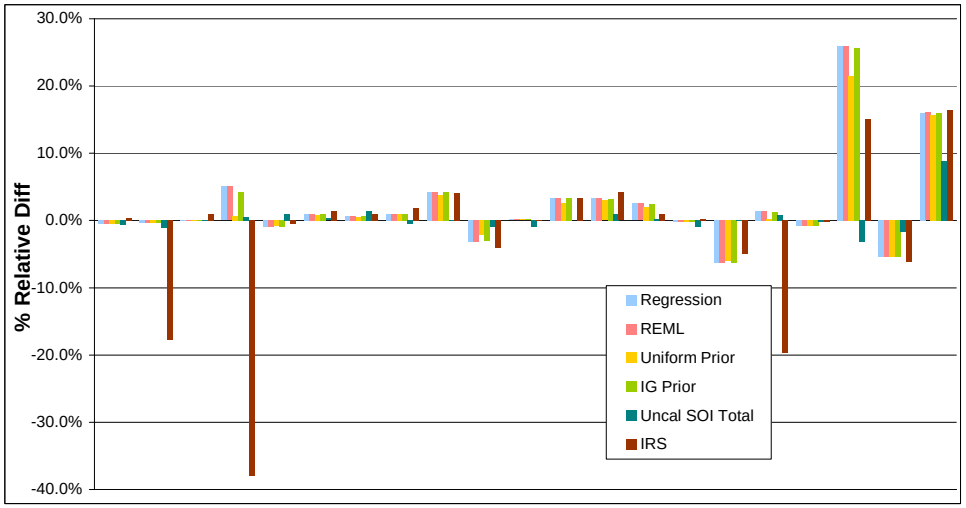


Real Estate Taxes Deducted



A.3. Percent Relative Differences Between Alternative Totals and SOI Sample Estimates to Calibrated State Group Totals, Tax Year 2004 (cont'd, note differences in scale)

State and Local State Income Taxes Deducted



State and Local General Sales Taxes Deducted

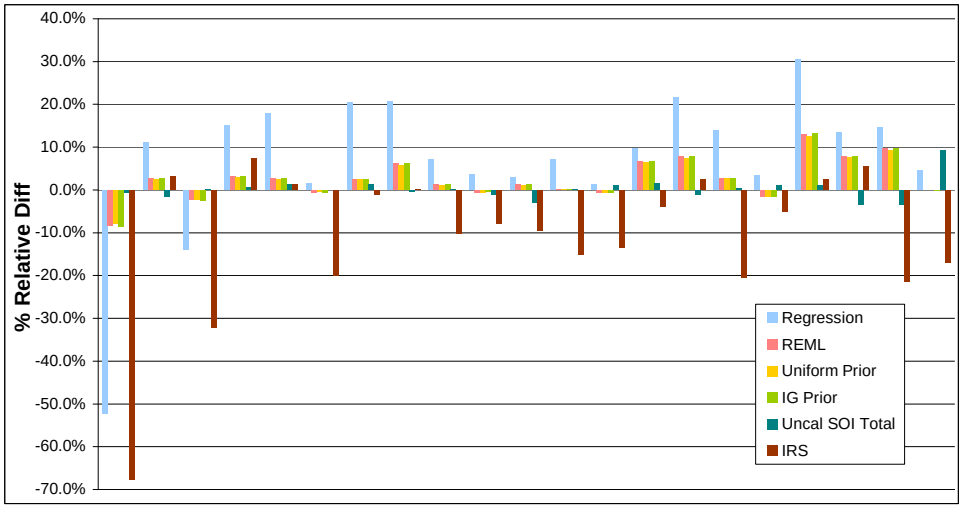


Table A.4. Average of Absolute Percent Relative Differences Across State-Groups (excl. HI) Between Alternative Estimates and Calibrated SOI Sample Total

Variable **2004 Ave (% Relative Difference)** **2005 Ave (% Relative Difference)**	*Uncalibrated SOI Total**	*IRS Total*	*Regression Estimate*	*REML*	*HB Unif* $(0,U)$	*HB IG* $(.001,.001)$
Adjusted Gross Income	0.89 1.05	1.16 1.08	0.93 0.82	0.93 0.67	0.79 0.59	0.92 0.79
Taxable Interest Income	0.82 0.91	3.93 2.76	4.06 2.61	4.06 2.36	3.67 2.02	4.02 2.52
Earned Income Tax Credit	0.93 1.13	6.51 3.35	5.30 2.98	3.70 2.98	3.44 2.37	4.37 2.89
Real Estate Taxes Deducted	1.01 1.18	2.31 2.07	2.25 1.49	2.26 1.48	1.99 1.32	2.22 1.45
State and Local Income Taxes Deducted	0.78 0.98	6.24 4.45	3.33 8.42	3.33 0.66	2.63 0.63	3.24 0.66
State and Local General Sales Taxes Deducted	1.14 1.22	12.51 14.05	14.12 12.64	4.05 6.21	3.88 5.82	4.06 6.46

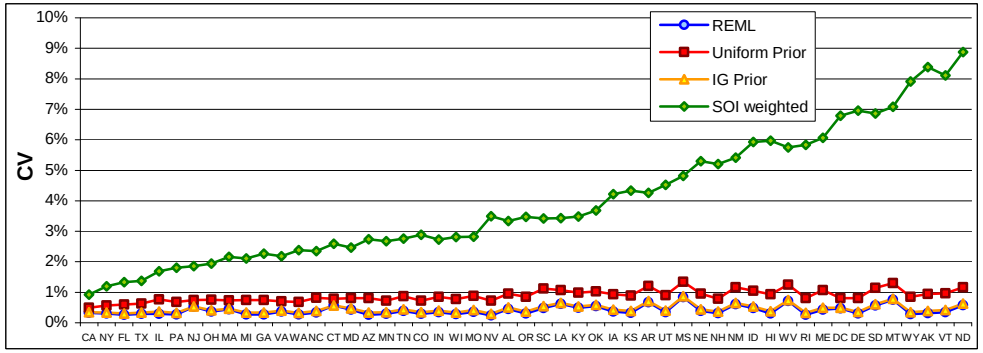
Table A.5. National-Level Totals Estimated Using Uncalibrated SOI Sample Total and Percent Relative Difference to Model-based Estimates

	National-Level Total*	**Percent Relative Difference to Uncalibrated SOI Total**					
Variable : 2004 result **2005 result**	*Uncalibrated SOI Total*	*Calibrated SOI Total*	*IRS Total*	*Regression Estimate*	*REML*	*HB Unif* $(0,U)$	*HB IG* $(.001,.001)$
Adjusted Gross Income	6,758,989,080 7,386,619,562	-0.003 -0.007	-0.747 -0.862	-0.036 -0.053	-0.036 -0.046	-0.034 -0.043	-0.038 -0.055
Taxable Interest Income	124,785,074 161,383,767	-0.055 -0.069	-0.183 -0.079	-1.510 -1.072	-1.508 -0.997	-1.372 -0.880	-1.503 -1.051
Earned Income Tax Credit	39,969,753 42,351,454	-0.038 0.002	1.728 0.559	-0.992 -0.304	-0.749 -0.304	-0.698 -0.268	-0.873 -0.298
Real Estate Taxes Deducted	132,120,007 144,546,368	0.029 0.113	0.221 1.565	-0.332 -0.036	-0.337 -0.039	-0.291 -0.018	-0.333 -0.039
State and Local Income Taxes Deducted	201,938,363 227,161,944	0.067 0.073	-0.373 0.633	-0.212 0.165	-0.210 0.032	-0.149 0.035	-0.204 0.027
State and Local General Sales Taxes Deducted	17,519,274 17,265,817	-0.371 -0.435	8.855 1.885	-2.691 -1.562	-1.261 -0.953	-1.224 -0.927	-1.274 -0.980

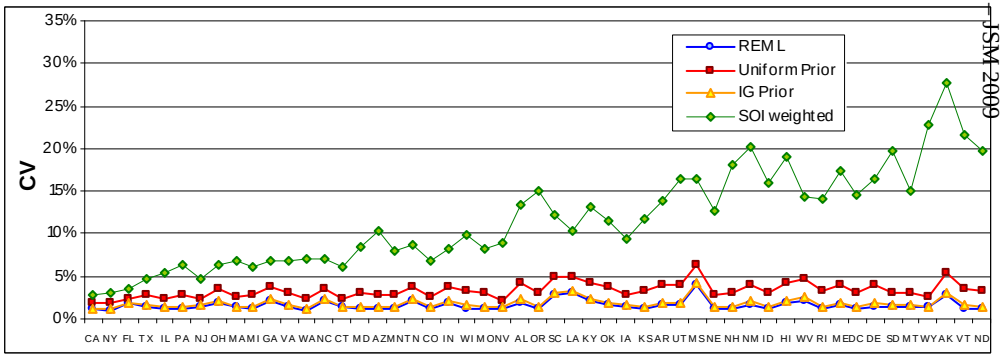
* SOI totals are rounded to the thousands of real dollars.

A.6. Coefficients of Variation for Alternative Estimates of State-Level Totals, Tax Year 2004 (note differences in scale)

Adjusted Gross Income

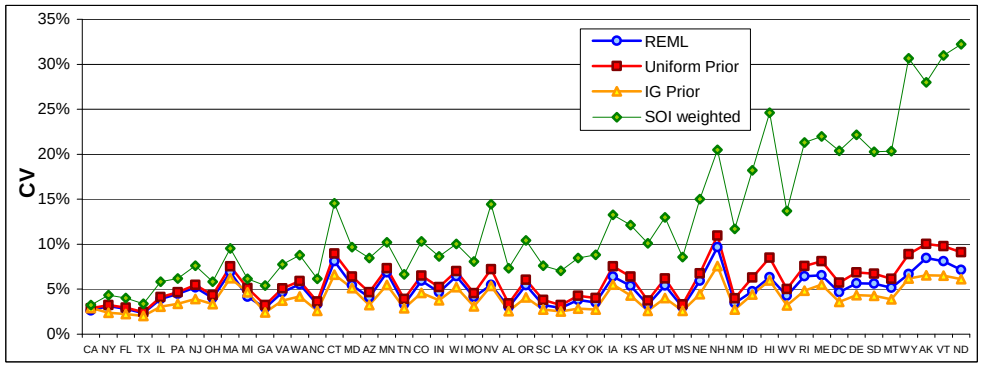


Taxable Interest Income

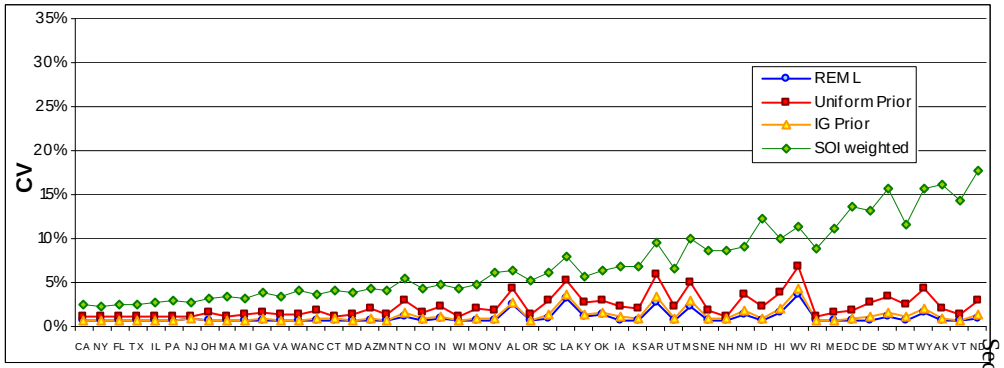


A.6. Coefficients of Variation for Alternative Estimates of State-Level Totals, Tax Year 2004 (cont'd, note differences in scale)

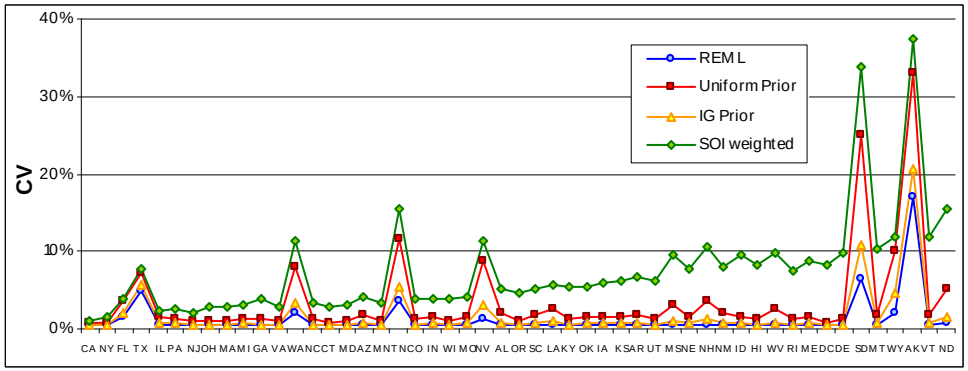
Earned Income Tax Credit



Real Estate Taxes Deducted



State and Local State Income Taxes Deducted



State and Local General Sales Taxes Deducted