
Office of Tax Analysis
Technical Paper 11
January 2023

# Estimation of Race and Ethnicity by Re-Weighting Tax Data

Robin Fisher

*OTA Technical Papers* is an occasional series of reports on the research, models and datasets developed to inform and improve Treasury's tax policy analysis. The papers are works in progress and subject to revision. Views and opinions expressed are those of the authors and do not necessarily represent official Treasury positions or policy. OTA Technical Papers are distributed in order to document OTA analytic methods and data and invite discussion and suggestions for revision and improvement. Comments are welcome and should be directed to the authors. OTA Papers may be quoted without additional permission.

# Estimation of Race and Ethnicity by Re-Weighting Tax Data

Robin Fisher[1]

January 2023

U.S. tax forms do not collect information about race or ethnicity and thus the tax data available for use in tax policy analysis by the Treasury Department does not include such information. We impute information about race and Hispanic origin (RH) to a stratified random sample of taxpayers used in Treasury's Individual Tax Model to allow for tax policy analysis by race and Hispanic origin. Specifically, we use a set of explanatory variables, including total income, filing status, age, number of dependents, sex, first name, last name, and the ZIP Code Tabulation Area (ZCTA) of the residence, to make inferences about a taxpayer's race and Hispanic origin. We apply Bayesian inference to estimate the probabilities that each taxpayer in our sample is in each of the 6 groups— Hispanic, White, Black, American Indian or Alaska Native, Asian or Pacific Islander (API), and multiple-race—given the variables, which, in turn, form the 6 RH weights for each taxpayer.

Any taxpayer data used in this research was kept in a secured Treasury or IRS data repository, and all results have been reviewed to ensure that no confidential information is disclosed.

[1] Robin Fisher: Office of Tax Analysis, U.S. Department of the Treasury, Robin.Fisher@treasury.gov

## I. Introduction

Tax law can have different impacts on individuals in different racial and ethnic groups. Such disparate tax outcomes can occur to the extent that individual characteristics that are used in tax law—e.g., marital status, number of children in the family, the level and sources of family income, household expenses, etc.—vary across racial and ethnic groups. However, unlike some other Federal agencies, the Internal Revenue Service (IRS) does not collect information about the tax filer's race or ethnicity (Bearer-Friend (2019), Brown (2021)). As a result, the tax data available to the Office of Tax Analysis (OTA) for use in its analyses does not include information about race and ethnicity. To conduct tax policy analysis by race and ethnicity, we first must develop a method to impute information about race and Hispanic origin to a stratified random sample of taxpayers. These representative samples are used for tax analysis modeling and can be modified to facilitate a better understanding of tax outcomes by race and Hispanic origin.

The Individual Tax Model (ITM), established by the Office of Tax Analysis (OTA) of the Treasury Department, is based on a stratified random sample of tax returns from the Individual and Sole Proprietor sample (Statistics of Income (2016)) and a sample of non-filing tax units for a tax year. We use total income, filing status, age, number of dependents, sex, first name, last name, and the ZIP Code Tabulation Area (ZCTA) of the residence as the explanatory variables to make inferences about the race and Hispanic origin category of the primary taxpayer of a filing unit or family. We apply Bayesian inference to estimate the probabilities that each taxpayer in our tax sample is in each of the 6 racial and Hispanic origin groups—White, Black, American Indian or Alaska Native (Native), Asian or Pacific Islander (API), multiple-race, and Hispanic—given the explanatory variables. As used in this paper, these categories are mutually exclusive:

Hispanic people of any race are included in the Hispanic category and excluded from the other categories. Six racial and Hispanic origin weights are formed for each taxpayer by multiplying the estimated race and Hispanic origin probabilities by the taxpayer's sampling weight in the file.

## II. Imputation of Race and Hispanic Origan

Since information about race and Hispanic origin (RH) is not present on the tax data, we use external data sources to provide the information to make inferences. As mentioned earlier, OTA uses the Individual Tax Model (ITM) to calculate estimates of various tax quantities and simulate the revenue effects of tax law changes. The ITM sample is the union of a stratified random sample of Federal individual income tax units (the IRS Statistics of Income's Individual and Sole Proprietor sample, further described below) and a sample of nonfiling units. For any discrete variables X and Y, including variables calculated from the tax model or, for example, the name of the primary filer in the tax unit, the empirical distribution is:

$$\hat{E}(f(X,Y)) = \sum_k w_k f(x_k, y_k),$$

where $w_k$ is the normalized sample weight for record *k*. The estimates for the RH categories take the following form:

$$\hat{E}\big(f(RH,X,Y)\big) = \sum_k w_k P(RH = rh|X = x_k) f(rh, x_k, y_k).$$

An estimate for $E(f(RH,X,Y))$ can be calculated from the ITM by using a new weight, $w_k P(RH = rh|X = x)$, for each record. This reduces the problem to one of estimating $P(RH = rh|X = x_k)$.

Estimates of $P(RH = rh|X = x_k)$, using names and geography as the explanatory variables, have been in use for several years. In this paper, we build on the Bayesian Improved

First Name, Surname, Geocoding (BIFSG) by Adjaye-Gbewonyo et al. (2014), Haas et al. (2019), and Voicu (2018). There is evidence that the geocoding method alone is less effective in predicting the RH categories than an expanded method that incorporates names and other information. For example, an early version of the BIFSG method that uses the surname and address was shown to be 108 percent more effective than the address-only method (Elliot, et al. (2009)). This outcome may result because some RH groups, e.g., Asians, Hispanics, and Natives, are less geographically concentrated relative to Blacks. For the BIFSG model, the strategy is to use Bayesian updating to combine separate joint distributions of RH, first names, surnames, and geography. We have

$$P(RH|sname, fname, ZCTA) = C\frac{P(RH, fname)}{P(RH)}\frac{P(RH, sname)}{P(RH)}P(RH|ZCTA),$$

where *ZCTA* denotes the Zip Code Tabulation Area, defined by the US Census Bureau, *fname* denotes first name, and *sname* denotes surname. The constant $C$ is the normalizing constant, and the conditional independence of $sname, fname, ZCTA$ given $RH$ is an assumption in the BIFSG estimator. $P(RH, ZCTA)$ is tabulated from the 2010 Census, $P(RH, fname)$ is tabulated from mortgage records (Tzioumis, 2018), and $P(RH, sname)$ is tabulated from the 2010 Census. The BIFSG estimator is implemented in Python module *surgeo*. That package also contains files with tabulations of the relevant conditional distributions.

The RH value for a tax unit in our model is the RH value for its primary filer. Here, the values for the RH variables are in the set {White, Black, Asian or Pacific Islander (API), Native American, Multiple Race, Hispanic}. This is the classification used in the BIFSG literature and the *surgeo* software, and it is also one of the official Census classifications. Nonetheless, it is one of several potential choices; other choices include the cross-classification of race with Hispanic

origin or something more detailed. However, the use of more detailed classifications will likely result in estimation difficulties from low counts of unweighted records in some cells.

ZCTAs are assigned to records in the ITM by first assigning Census blocks to the records, then using Census files to match those blocks to ZCTAs. As a result, about 78 percent of records have ZCTAs assigned. When the ZCTA could not be assigned, the 9-digit ZIP code was filled in, where available, leaving about 6 percent of the records in the ITM file without a ZCTA or ZIP code value.

Nearly all records have names assigned, but some names do not match to the BIFSG database. This is the case with about 13 percent of the first names and less than 1 percent of the surnames. When a value is missing, the relevant ratio in the equation above is set to one. This is consistent with an assumption that the variable is missing at random, which may not hold. Ongoing work centers on increasing the match rates for ZCTAs and names as well as investigating the association between the presence of missing values and the RH variable.

We use Markov Random Fields (MRFs) for our modeling framework.[2] The dependence relationship for these models is represented by a graph $\mathcal{G} = \{V, E\}$, where $V = \{v_i, \ldots, v_d\}$ is a set of vertices which correspond to a set of variables $\mathbf{X_V}$ in a $d$-dimensional multivariate distribution $P$, and $E = \{e_{i,j}, i \in \{1, \ldots, d\}, j \in \{i+1, \ldots, d\}\}$ is a set of undirected lines. If it is true that $e_{i,j} \notin E$, only if $v_i$ is independent of $v_j$ given $S$ for any set $S \subseteq V - \{v_i, v_j\}$, then $\mathcal{G}$ is *Markov* with respect to $P$.

A *clique* is a maximal complete subgraph. If $\mathcal{G}$ is *Markov* with respect to $P$, then $P(\mathbf{X_V}) = \prod_c \phi\left((X_c)\right)$, where $c$ indexes the cliques in $\mathcal{G}$. If $\mathcal{G}$ is *decomposable*,

[2] See Koller and Friedman (2009) for a detailed discussion.

$$P(\mathbf{X_V}) = \prod_c \frac{P(X_c)}{P(X_c \cap X_d)},$$

where $d$ and $c$ index cliques and $d < c$, for some ordering of the cliques. This corresponds to decomposability in log-linear models for contingency tables (Haberman (1974)). We call the arrays $\{P(X_c)\colon c\ is\ a\ clique\ in\ \mathcal{G}\}$ *clique tables*. The sufficient statistics are the collection of frequency tables for the cliques (the empirical clique tables), and the maximum likelihood estimator for complete data is just the factorization above, where the empirical clique tables stand in for the true clique tables under the model.

Decomposable models have some advantages. First, we can control the complexity of the model by controlling the maximal clique size. This is true in both an algorithmic and statistical sense. Statistically, controlling the clique size provides some protection against overfitting. Algorithmically, the computational cost of parameter estimation and the calculation of conditional probabilities increase very quickly as the size of the cliques increases. Second, we can estimate each clique table separately and combine them later with decomposable models. This is very helpful, since we don't have to re-estimate all the parameters in the model whenever we change the set $X$ of covariates. Note that, with $P(RH, X)$ estimated, conditioning is possible in any direction so $P(RH|X)$ is directly estimated, as is $P(X_1|RH)$ for $X_1 \in X$. For $Y \in T \backslash X$,

$$P(Y|RH) = \sum_x P(Y|X = x)P(X = x|RH),$$

or, similarly,

$$P(RH|Y) = \sum_x P(RH|X = x)P(X = x|Y).$$

We present the standard BIFSG model and one proposed expanded model in Figure 1. The chart on the left shows the Markov random field (MRF) for the BIFSG model; there are three cliques corresponding to the terms in the equation. The generating class for this model is

$\{[RH, fname], [RH, sname], [RH, ZCTA]\}$, and the probability distribution can be represented as

$$P(RH, X) = P(RH, sname, fname, geo)$$
$$= P(RH, geo)\frac{P(RH,sname)}{P(RH)}\frac{P(RH,fname)}{P(RH)}.$$

The MRF for the expanded model is presented on the right. The generating class for this model is $\{[RH, fname], [RH, sname], [RH, ZCTA], [RH, X_1, X_2]\}$, and the conditional distribution of $RH$ is

$$P(RH|fname, sname, ZCTA, X_1, X_2) =$$
$$C\frac{P(RH|X_1, X_2)}{P(RH)}\frac{P(RH|fname)}{P(RH)}\frac{P(RH|sname)}{P(RH)}P(RH|ZCTA),$$ where $C$ is a constant.

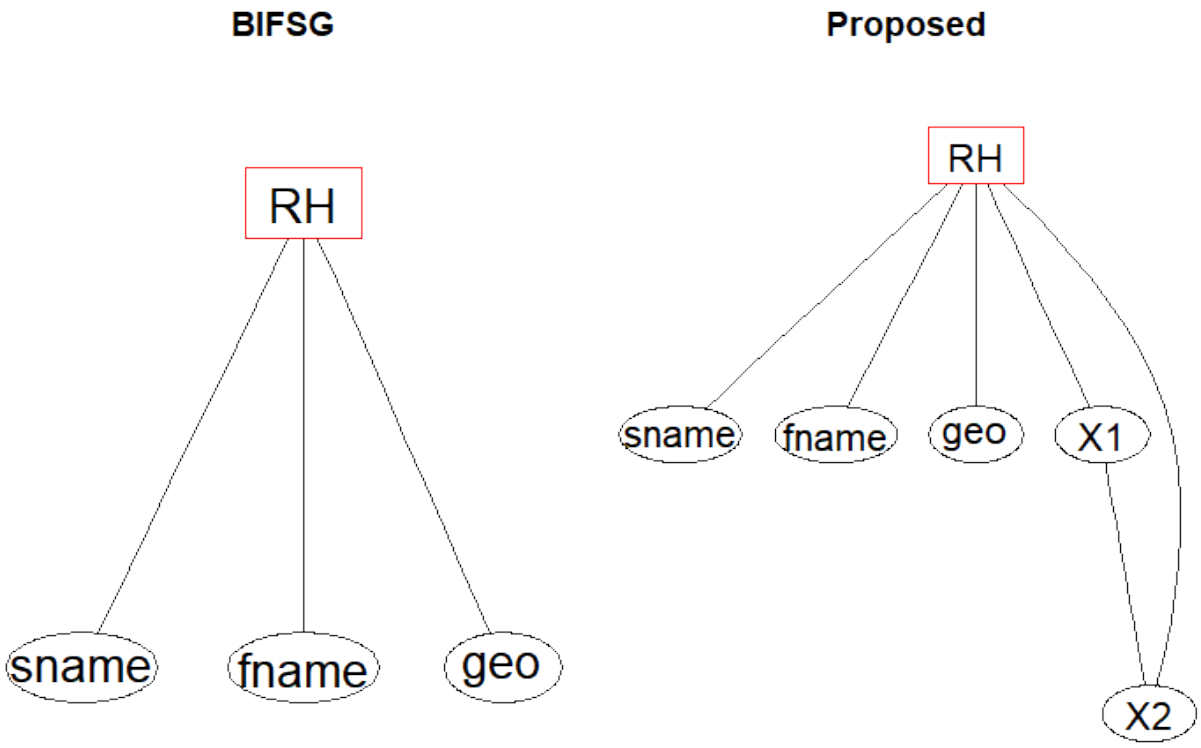



*Figure 1. MRFs for the BIFSG model (left) and a proposed expanded model (right)*

Thus, the BIFSG estimator is systematically different from the proposed extension by a factor $\frac{P(RH|x1,x2)}{P(RH)}$; we have just added a clique and, with it, a factor containing a clique table we can estimate separately. Further decomposable extensions of the model follow along similar lines. Note this is not the final model we apply in this paper, but a simpler version for illustrative purposes.

In the decomposable BIFSG model, given any set of imputation covariates $X$, the model has the (perhaps implicit) assumption that $RH \perp\!\!\!\perp Y|X$ for $Y \in T\backslash\{RH, X\}$. As a result, all association between $Y$ and $RH$ is mediated through $X$. In the BIFSG model illustrated in Figure 2, if variables of interest include *total income* and *marital status* (*totInc* and *MARS*, respectively), then those variables depend on $RH$ only through $(fname, sname, geo)$, their names and the location of their residence. If the analyst does not want to assume totInc and *MARS* are mediated through geo, then a model extension is needed.

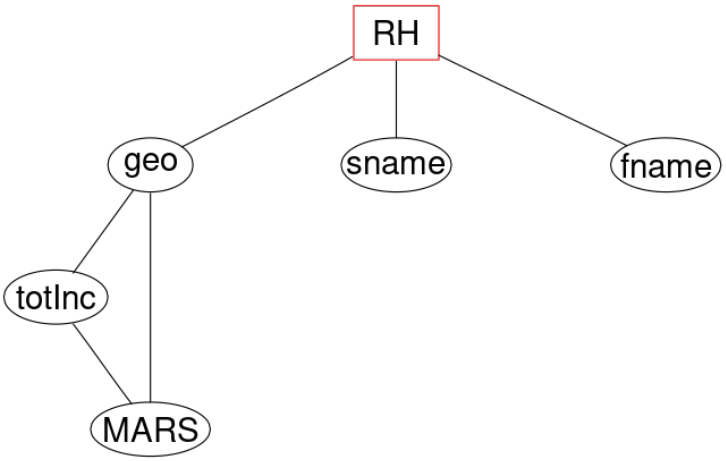



*Figure 2. Undirected Graph showing that (geo,sname,fname) separate (totInc,MARS) from RH in the BIFSG model. This implies that association between RH and (totInc,MARS) is mediated by (fname,sname,geo).*

In an extension, we can assume that, for example, $(totInc, MARS)$ are directly associated with $RH$ as in Figure 3.

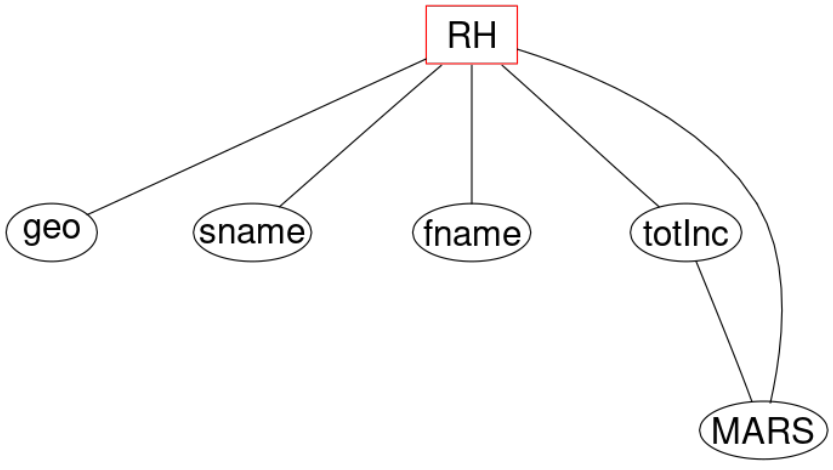



*Figure 3. Extended model where* $(totInc, MARS)$ *are directly associated with* $RH$.

It is worth emphasizing that the distributions $P(RH|sname)$, $P(RH|fname)$, and $P(RH|ZCTA)$ are taken as known, even though they are tabulations from data with possible error sources. These distributions are fundamental to the estimation procedure. If we sum over the variables $X_1$ and $X_2$ in the equation above, the result is the BIFSG estimator; the BIFSG estimator defines the expectations of $RH$ in this set of models. There is ongoing work to include uncertainty in the estimators for these distributions in the model.

The tax filing sample used by the ITM is a stratified Bernoulli sample of Federal individual income tax returns drawn by the IRS each year—the Individual and Sole Proprietor file (INSOLE), (Statistics of Income (2016)). The non-filing sample is generated by the OTA based on the income information submitted by third parties to the IRS. The stratum identifier, $CSAMP,$ is present in the file, and we include it in our model as a nuisance parameter. Specifically, we model the variation in the sample design as a collection of simple random samples of tax units from each stratum.

Consider the problem of estimating the clique-table $P(RH, X_1, X_2)$. For the ITM, the sampling strata need to be accommodated. In the case of complete data, it would typically be appropriate to use sampling weights. Here, with the latent class variable $RH$, it was not clear how to do that. Instead, we introduce a variable for the sampling stratum, CSAMP, and estimate the higher dimensional clique-table $P(RH, X_1, X_2, CSAMP)$ in order to estimate the marginal clique-table $P(RH, X_1, X_2)$.

The probability vector for $P(RH|X_1 = i, X_2 = j, CSAMP = k)$ is denoted $\theta_{i,j,k}$, and is a vector of probabilities, $\theta_{i,j,k} \in \mathbb{R}^6$. The unobserved sample counts of the RH categories for $(X_1 = i, X_2 = j, CSAMP = k)$ is denoted $\mathbf{n}_{i,j,k}$ with $\mathbf{n}_{i,j,k} \in \mathbb{N}^6$. The graph for this model is shown in Figure 4. This is the model structure we use for the parameter estimation for each

clique-table. The final model is assembled from the four clique-tables with ITM variables and the three 2-member clique-tables with name or geography information.

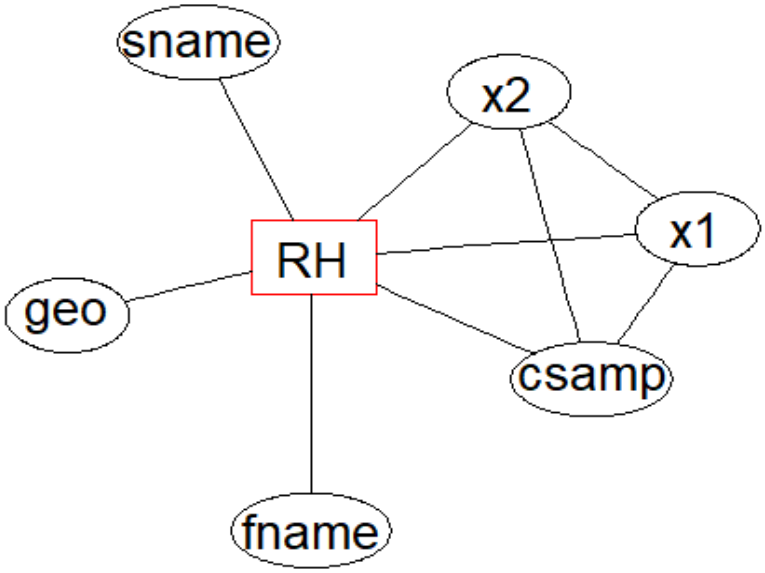



*Figure 4. Another proposed expanded model, including stratum variable CSAMP. This is also the model structure used to estimate the parameters for the each of the clique-tables* $P(RH, X_1, X_2)$.

## III. Current Extended Model

We use the following variables in the current model.

- total income, *totInc*,
- filing status, *MARS,*
- age,
- number of dependents, *ndep,*
- sex of the primary taxpayer, *gen1,*
- first name, *fname*

- last name, *sname,* and the
- ZIP Code Tabulation Area (ZCTA), ZCTA.

The graph for the current extended model is presented in Figure 5. It is decomposable with seven cliques total, three for the BIFSG variables on the right, and four for the ITM variables. The joint distribution function can be factorized as

$$P(RH, X, FSG)$$

$$= P(RH)\frac{P(RH, gen1, totinc)}{p(RH)}\frac{P(RH, gen1, ndep)}{P(RH, gen1)}\frac{P(RH, totinc, MARS)}{P(totInc, RH)}\frac{P(RH, age, MARS)}{P(RH, mars)} \cdot$$

$$\frac{P(RH, fname)}{P(RH)}\frac{P(RH, sname)}{P(RH)}\frac{P(RH, ZCTA)}{P(RH)}$$

As we noted earlier, this model includes the assumption that that $X \perp\!\!\!\perp FSG|RH$. It is plausible that this assumption is violated, especially with respect to geography. If this independence assumption is relaxed, our estimator for $P(RH|X)$ is biased. We are currently researching how to fit a model without these assumptions.

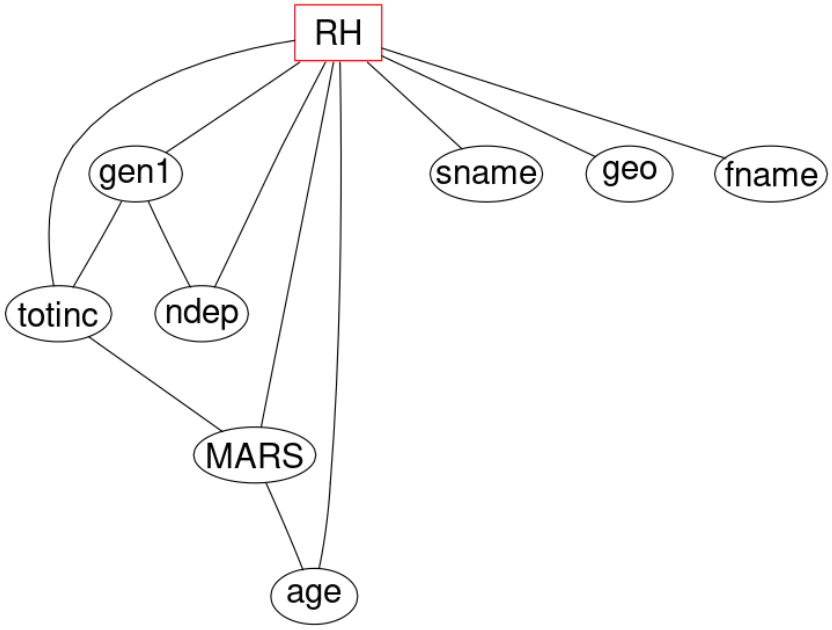



*Figure 5 Undirected graph for the current extended model.*

Here is the set of conditional distributions for the current model.

$$RH|(F = f_i) \sim MN(1, \phi_{f_i})$$

$$RH|(S = s_i) \sim MN(1, \phi_{s_i})$$

$$RH|(G = g_i) \sim MN(1, \phi_{g_i})$$

$$RH_i|(totInc_i = t, MARS = m, CS = c) \sim MN(1, \theta_{t,m,c}^{totInc,MARS,cs})$$

$$RH_i|(totInc_i = t, gen1 = g, CS = c) \sim MN(1, \theta_{t,g,c}^{totInc,gen1,cs})$$

$$RH_i|(MARS = m, age = a, CS = c) \sim MN(1, \theta_{m,a,c}^{MARS,age,cs})$$

$$RH_i|(Ndep = d, gen1 = g, CS = c) \sim MN(1, \theta_{d,g,c}^{Ndep,gen1,cs})$$

In the above equations, the parameters $(\phi_f, \phi_s, \phi_g)$ are taken as known. All that remains is to estimate the $\theta$ parameters. We use a Bayesian method to estimate these unknown

parameters. This approach requires starting with a prior distribution for $\theta_{i,j,k}$. Suppressing the superscripts, we specify that as a Dirichlet distribution,

$$\theta_{i,j,k} \sim Dir(\alpha_0 \mathbf{1}).$$

We let $\alpha_0 = 1.1$. The prior with $\alpha_0 = 1$ is uniform over the support of $\theta_{i,j,k}$, and is often considered noninformative. In the case with $\alpha_0 = 1.1$, we prohibit values of $\theta_{i,j,k}$ very close to the boundary, where one or more of the probabilities are close to zero. In particular, more work is needed since letting $\alpha_0 = 1.1$ may be too informative for the rare RH categories, *native American* and *Multiple Race NH*.

We calculate the posterior expectations in this model with a Gibbs Sample Markov Chain Monte Carlo method (Metropolis et al. (1953) and Geman and Geman (1984)). This is a way to simulate observations from any distribution from which we can generate random variables from the full conditional distributions, which are the conditional distributions of the target variables, given everything else in the model. For a parameter with a sufficient statistic, the 'everything else' reduces to the sufficient statistic. This sufficient statistic may include latent variables. That is the case for $\theta_{i,j,k}$, where the sufficient statistic is the $(RH, X_1, X_2, CSAMP)$ table. The variable $RH_m$, the Race and Hispanic origin assignment for record $m$, must be simulated for each record.

The full conditional distributions of $\theta$ and $RH$, given everything else (EE), are

$\theta_{i,j,k}^{X_1,X_2,CSAMP} | EE \sim Dir(\alpha_0 \mathbf{1} + \mathbf{n}_{i,j,k})$, where $\boldsymbol{n}_{i,j,k}$ is the vector of counts in the $RH$ categories for $(X_1 = i, X_2 = j, CSAMP = k)$ and

$RH_m | (X_1 = i, X_2 = j, CSAMP = k) \, | EE \sim$

$$MN\left(1, C\theta_{i,j,k}^{X_1,X_2,CSAMP} \frac{P(RH|geo)}{p(RH)} \frac{P(RH|fname)}{p(RH)} \frac{P(RH|sname)}{p(RH)}\right)$$

for record $m$.

An outline of the Gibbs Sampling algorithm in general and for this problem is presented in the Appendix.

The full current extended model uses the following variables: total income, filing status, age, number of dependents, and sex. All continuous variables are converted to discrete variables by grouping, so we can form a model on multidimensional contingency tables. Although the current extended model seems to do a better job than the standard BIFSG model in predicting the RH categories, determination of a set of variables which will perform well for general use of the tax model is the subject of ongoing research.

### III.1. Variance Estimation

Variance estimates are important in order to have a general understanding of the reliability of an estimated table. Even in the case where RH is not involved, some tables involve such small subpopulations that the sample may make it more difficult to evaluate the reliability. Sometimes estimated standard errors (SEs) may be large enough that we may not want to use the tables. It may also happen that the estimated effect of a policy change, or the differential effect between races, may be small compared to the standard error.

Test statistics in ordinary sampling situations, like i.i.d. samples, are relatively easily calculated from the log-likelihood in many situations. In the case of a designed sample like the ITM sample and the RH imputations that are not i.i.d., we need to do something else.

For the ITM sample and RH imputations there are a few sources of variance in the estimate. The first is the sampling error variance from the ITM sample. This is just the variance that arises from choosing a random sample from a fixed finite population. Second, there are the variances for the estimated parameters $\theta$ in the model.

An additional source of variance comes from the s variances associated with $P(RH|sname)$ and $P(RH|geo)$. These are tabulated using data from the 2010 decennial census and are likely to be different in current cross-sections. We can think of this as an extra component of variance or a bias term.[3] In addition, categories formed by names or geography may be very small. For example, there are many rare names, so $P(RH|fname)$ may be based on a small sample. These factors potentially increase the SEs.

Estimators for the ITM sampling variances have been available for a while. The IRS made coefficients of variation (CVs) available for several variables on the ITM. In addition OTA has implemented bootstrap-based replicate weights for the ITM, so variances related to quantities in the ITM can be calculated.[4] For at least a handful of variables, CVs calculated from the ITM replicates estimates match well with those produced by the IRS in the 2016 ITM. Note that these variances are on the ITM and may be used for any of the regular tables produced for the ITM.

Variances for the $\theta$ parameters of the RH imputations are calculated using the Gibbs Sampler. As noted above, the output of the Markov Chain Monte Carlo (MCMC) represents a sample from the distribution of $\theta|$data. We can use the values generated by the Gibbs Sampler by themselves to calculate variances of $\theta$, which is useful for model-checking, or we can combine them with the replicates for the ITM sampling error variances. We have 100 replicates from the sampling error variance estimator and 100 from the Gibbs Sampler. We combine the 100 replicates from the ITM bootstrap replicates with 100 replicates from the Gibbs Sampler to get 100 combined replicates. The ordering of the combinations does not matter. For the $b^{th}$ combined replicate,

[3] We expect new data from the Census in the next few months.
[4] We use the mrfbootstrap function in the R package survey (Lumley (2020)).

$$Y^{(b)} = \sum_i w_i^{(b)} P^{(b)}(RH|X = x_i) Y_i$$

Here, $Y_i$ may depend on any subset of $(T, RH)$.

$$\bar{Y} = \frac{1}{B} \sum_b Y^{(b)}$$

$$\widehat{var}(Y) = \frac{1}{B-1} \sum_b \left(Y^{(b)} - \bar{Y}\right)^2$$

### III.2. Comparison of Imputation Results to Census

To evaluate the current extended BIFSG imputation on the ITM, we compare the distribution of RH as measured by the U.S. Census Bureau's 2020 Decennial census for the U.S. residential adult population to the distribution of RH in Treasury's ITM for the primary filer on a return for 2023.[5] As shown in Table 1, the imputation is more likely to sort multiple race primary filers into a single RH designation while White primary filers appear to be overrepresented in the imputation results.

Some of the difference can be explained by conceptual differences between the two distributions. Census' adult population includes dependents over 18 and secondary filers on a joint return while the imputation does not include either of these populations. To the extent that RH of adult dependents and secondary filers is not distributed the same as other adults, the comparison will be imperfect.

[5] In Table 1, Native American includes Alaskan Native, Hawaiian Native and other Pacific Islander. Treasury does not have a category for "some other race alone."

**TABLE 1: COMPARING U.S. POPULATION BY RACE AND ETHNICITY**

| Race/Ethnicity | Census 2020 Adults | | Treasury 2023 Families/Primaries | |
|---|---|---|---|---|
| | Millions | % | Millions | % |
| Total | 259.5 | 100 | 186.5 | 100 |
| White | 157.4 | 61 | 124.1 | 67 |
| Hispanic | 43.4 | 17 | 28.2 | 15 |
| Black | 30.3 | 12 | 19.7 | 11 |
| Asian | 16.2 | 6 | 11.0 | 6 |
| Native American | 2.2 | 1 | 1.3 | 1 |
| Some other race alone | 1.2 | 0.5 | - | - |
| Multiple race | 8.8 | 3 | 2.2 | 1 |

Source: U.S. Census Bureau, Decennial Census Table P4, "Hispanic or Latino, and Not Hispanic or Latino by Race for the Population 18 years and over."

### III.3. Testing the methodology on U.S. Army applicant data

We have tested our extension of the BIFSG model to the original BIFSG model, with promising results, especially with regard to imputing the probabilities of being Black or Hispanic. With permission of the U.S. Army, we tested our imputation using a dataset of the universe of U.S. army applicants. These data include the applicant's first name, surname, address, marital status, and income as well as self -reported race. Our extended model results were very similar in predicting the joint distribution of marital status and income by race for White applicants, slightly improved for Hispanic applicants, much improved for Black applicants, but not improved for Asian applicants. This testing is encouraging but not definitive. Relative to the general population as represented in the tax data, Army applicants are much less likely to be Asian and are more likely to be low-income and single.

We estimate the Kullback-Liebler (KL) distance between the estimated joint distributions ($\hat{P}$) of income and marital status (Xs) under the BIFSG model and the extended model as compared to the distributions as tabulated using the actual RH values in the military applicant data. The KL distances for the BIFSG model and extended model are as follows (where $\xi$ represents parameters from the look-up tables for first name and last name).

$$KL(\hat{P}||P)_{BISFG} = \sum_x P(X|RH = rh) \log\left(\frac{P(X|RH = rh)}{\bar{P}(X|RH = rh, \xi)}\right)$$

$$KL(\hat{P}||P)_{extended} = \sum_x P(X|RH = rh) \log\left(\frac{P(X|RH = rh)}{\frac{\hat{P}(RH=rh|X,\theta)}{P(RH=rh)}\bar{P}(X|RH = rh, \xi)}\right)$$

The only difference is the extra term in the extended model: $(\hat{P}(RH=rh|X,\theta))/P(RH=rh)$. Thus, the difference between the KL measures for the extended model and the BIFSG measure can be estimated as follows:

$$\begin{aligned}
KL(\hat{P}||P)_{ext} - KL(\hat{P}||P)_{BISFG} &= -\sum_x P(X|RH = rh) \log\left(\frac{\hat{P}(RH = rh|X, \theta)}{P(RH = rh)}\right) \\
&= -\sum_x P(X|RH = rh) \log\left(\frac{\hat{P}(X|RH = rh, \theta)}{P(X)}\right) \\
&= -\sum_x P(X|RH = rh) \log\left(\frac{\hat{P}(X|RH = rh, \theta)}{P(X|RH = rh)}\frac{P(X|RH = rh)}{P(X)}\right) \\
&= \sum_x P(X|RH = rh)\left[\log\left(\frac{P(X|RH = rh)}{\hat{P}(X|RH = rh, \theta)}\right) - \log\left(\frac{P(X|RH = rh)}{P(X)}\right)\right. \\
&= \sum_x P(X|RH = rh)\left[\log\left(\frac{P(X|RH = rh)}{\hat{P}(X|RH = rh, \theta)}\right]\right. - \sum_x P(X|RH = rh) \log\left(\frac{P(X|RH = rh)}{P(X)}\right)
\end{aligned} \tag{5}$$

Let

$$D_{X|RH,th} = \sum_x P(X|RH = rh)\left[\log\left(\frac{P(X|RH = rh)}{\hat{P}(X|RH = rh, \theta)}\right)\right.$$

and

$$D_X = \sum_x P(X|RH = rh) \log\left(\frac{P(X|RH = rh)}{P(X)}\right).$$

Let

$$D_{X|RH,th} = \sum_{x} P\,(X|RH = rh)[\log(\frac{P(X|RH = rh)}{\hat{P}(X|RH = rh, \theta)}$$

And

$$D_X = \sum_{x} P\,(X|RH = rh)\log(\frac{P(X|RH = rh)}{P(X)}).$$

When the extended model performs better in terms of the KL measure, the difference between the KL distances should be negative, that is, $D_X > D_{X|RH,th}$.

The comparison of the KL measurements for the imputation and reported RH by Army applicants appears in Table 2. The KL distance under Treasury's extended model is virtually the same for White applicants, lower for Hispanic and Black applicants and higher for Asian applicants. Given that Asian applicants are only 1 percent of all military applicants but 6 percent of the U.S. resident population, we do not consider the results for Asian applicant RH from the military data to be definitive. It is also true that, since the sample has conditioned on being an Army applicant, the joint distribution of income, marital status, race, and geography are not the same as the ITM population. Therefore, although we are less confident in the Asian probabilities for the general model, this does not give us evidence to contradict our imputation model.

**TABLE 2: COMPARING KL DISTANCES**

| Race or Ethnicity | $D_X$ | $D_{X\mid RH,rh}$ |
|---|---|---|
| White | 0.0080 | 0.0079 |
| Hispanic | 0.0056 | 0.0024 |
| Black | 0.1190 | 0.0072 |
| Asian | 0.0610 | 0.0936 |
| Native | 0.0500 | 0.0840 |
| Mulitple | 0.0200 | 0.0130 |

## IV. Conclusion

In order to conduct equity analysis in taxation, we describe our method to address the omission of race and ethnicity from tax data used in our analyses. Our approach uses external information to facilitate inferences about an individual's race and ethnicity. Specifically, we impute six race and Hispanic origin weights for use with our sample of tax units that that enables the re-weighted ITM sample to represent the racial and ethnic composition of the U.S. population.

It is our goal to produce an imputation for Race/Hispanic Ethnicity that can be used on the ITM in support of the policy analysis OTA conducts for the administration. In pursuit of that goal Our imputation method is based on a set of explanatory variables that are both intuitive and which analysis shows are associated with both RH and other tax variables of interest. This results in valid tax analysis when most of the dependence between the tax estimate of interest and the RH fields is explained by the selected explanatory variables. However, the optimal choice of explanatory variables required for policy analysis may vary by the policy being examined so we continue to investigate additional sets of variables that will be best suited to our expanding modeling needs.

## V. References


Adjaye-Gbewonyo, Dzifa, Robert A. Bednarczyk, Robert L. Davis, and Saad B. Omer. 2014. "Using the Bayesian Improved Surname Geocoding Method (BISG) to Create a Working Classification of Race and Ethnicity in a Diverse Managed Care Population: A Validation Study." *Health Services Research* 49 (1): 268–83.

Bearer-Friend, Jeremy. 2019. "Should the IRS Know Your Race? The Challenge of Color-Blind Tax Data." *Tax Law Review* 73(1): 1-67.

Brown, Dorothy A. 1997. "The Marriage Bonus/Penalty in Black and White." *University of Cincinnati Law Review* 65(Spring): 787-798.

Brown, Dorothy A. 2021. *The Whiteness of Wealth*. Crown Publishing Group.

Elliott, Marc. N., Peter A. Morrison, Allen Fremont, Daniel F. McCaffrey, Philip Pantoja, and Nicole Lurie. 2009. "Using the Census Bureau's Surname List to Improve Estimates of Race/Ethnicity and Associated Disparities." *Health Services and Outcomes Research Methodology* 9(2): 69-83.

Geman, Stuart and Donald Geman. 1984. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6: 721–41.

Haberman, Shelby. 1974. *The Analysis of Frequency Data.* University of Chicago Press.

Haas, Ann, Marc N. Elliott, Jacob W. Dembosky, John L. Adams, Shondelle M. Wilson-Frederick, Joshua S. Mallett, Sarah Gaillot, Samuel C. Haffer, and Amelia M. Haviland. 2019. "Imputation of Race/Ethnicity to Enable Measurement of Hedis Performance by Race/Ethnicity." *Health Services Research* 54(1): 13–23.

Koller, Daphne, and Nir Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

LaLumia, Sara. 2008. "The Effects of Joint Taxation of Married Couples on Labor Supply and Non-Wage Income." *Journal of Public Economics* 92(7): 1698-1719.

Lumley, T. 2020. "Survey: analysis of complex survey samples" R package version 4.1-1

Metropolis, Nicholas, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. 1953. "Equation of State Calculations by Fast Computing Machines." *The Journal of Chemical Physics* 21(6): 1087–92.

R Core Team. 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Statistics of Income, Internal Revenue Service. SOI Tax Stats-Historic Tables 2. 2016 https://www.irs.gov/uac/SOI-Tax-Stats-Historic-Table-2, downloaded November 2020.

Tax Policy Center,.2001. *Biden's Expanded EITC Adds Significant Marriage Penalties*. https://www.taxpolicycenter.org/taxvox/bidens-expanded-eitc-adds-significant-marriage-penalties Retrieved on Oct. 20, 2021.

Tzioumis, Konstantinos. 2018. "Demographic Aspects of First Names." *Scientific Data* 5 (1): 1–9.

Voicu, Ioan. 2018. “Using First Name Information to Improve Race and Ethnicity Classification.” *Statistics and Public Policy* 5 (1): 1–13.

**Appendix**

An outline of the Gibbs Sampling algorithm is provided below. Note the superscript $(b)$ indexes the iteration number; it is not an exponent. The Gibbs sampler for the general case is as follows:

- We have a set of random variables $\{X_k,\ k = 1, \dots, K\}$
    - For each $k$, we can generate a random variable from $P\left(X_k \middle| \{X_j, j \neq k\}\right)$
- Execute the following algorithm
    - Initialize the $X_k$'s to get $\left\{X_k^{(0)}, k = 1, \dots, K\right\}$
    - Repeat for $b = 1, \dots, B$ for some large B:
        - For each $k$
            - Generate $X_k^{(b)} \sim P\left(X_k \middle| \left\{X_j^{(b-1)}, j \neq k\right\}\right)$
        - Yields $X^{(b)}$
    - Yields a sample $\left\{X^{(b)}, \mathrm{b} = 1, \dots \mathrm{B}\right\}$
    - For large B, this sample approximates a sample from $P(X)$
        - $\frac{1}{B}\sum_b f(x^{(b)}) \rightarrow E(f(X))$ *almost surely*

To apply the Gibbs sampler to the current model, use the following algorithm:

- Initialize $RH_m$ for each ITM record $m$.
- For each record, indexed by $m$, generate $RH_m^{(0)}$ from $P(RH_m|EE) = MN\left(1, C\frac{P(RH|geo_m)}{p(RH)}\frac{P(RH|fname_m)}{p(RH)}\frac{P(RH|sname_m)}{p(RH)}\right)$, where $MN(n,p)$ represents the

multinomial distribution with size $n$ and probability vector $p$. $C$ is a constant such that the second argument sums to 1.

- Tabulate $\mathbf{n}_{i,j,k}^{(0)} = \sum_{X_m=(i,j,k)} RH_m^{(0)}$

- Main Loop
  - For $b = 1, \ldots, B$
    - Generate $\theta_{i,j,k}^{(b)} \sim Dir(\mathbf{1}, \alpha_0 + \mathbf{n}_{i,j,k}^{(b-1)})$ for each $i, j, k$
    - Generate $RH_m^{(b)}$ from $P(RH|EE) =$

      $MN(1, C\theta_{(i,j,k)_m}^{(b)} \frac{P(RH|geo_m)}{p(RH)} \frac{P(RH|fname_m)}{p(RH)} \frac{P(RH|sname_m)}{p(RH)})$
    - Tabulate $\mathbf{n}_{i,j,k}^{(b)} = \sum_{X_m=(i,j,k)} R\,H_m^{(b)}$

This generates a sequence of values $(\theta_{i,j,k}^{(b)}, b = 1, \ldots B)$. If the initial values, where $b = 0$, are far from the center of the posterior distribution, it may take several iterations for the sequence to move toward the mode of the posterior. As above, we can

Collapse $\theta^{(b)}$ over $Csamp$ for each value of $(i, j)$.

$$\theta_{i,j}^{(b)} = \frac{\sum_k \theta_{i,j,k}^{(b)} p(X_1=i, X_2=j, Csamp=k)}{p(X_1=i, X_2=j)},$$

Which is just $\hat{P}(RH, |X_1 = i, X_2 = j)$ in the $b^{th}$ iteration.

Then if $B$ is large,

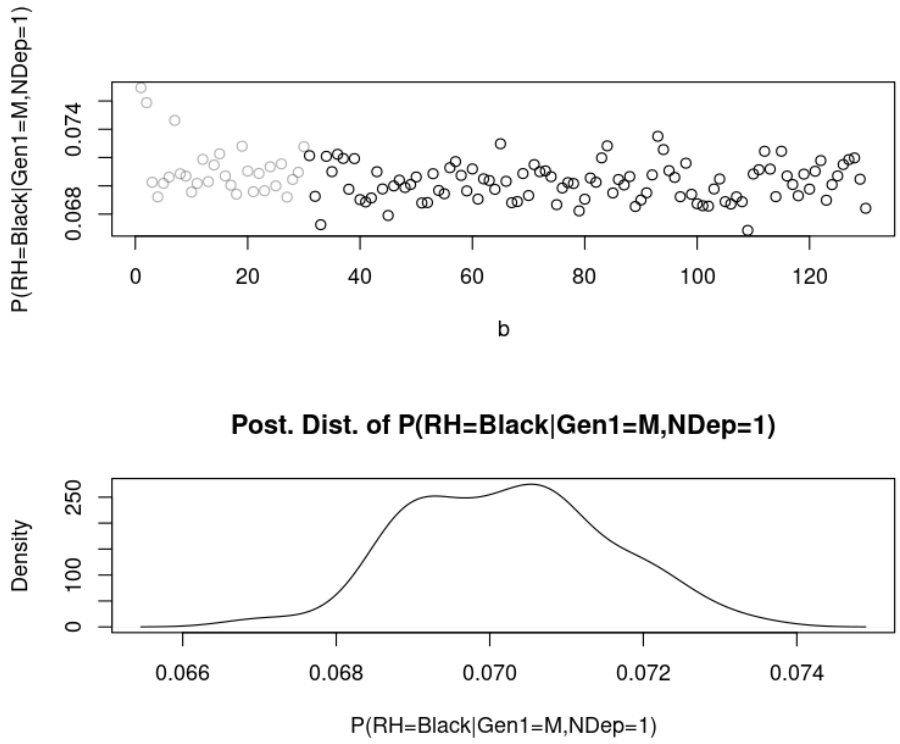
$$E(\theta_{i,j}|Data) = \frac{1}{B}\sum_1^B \theta_{i,j}^{(b)},$$

and

$$cov(\theta_{i,j}|Data) = \frac{1}{B}\sum_1^B \theta_{i,j}^{(b)}\, \theta_{i,j}^{(b)T}.$$

Note that if the initialized value is not typical, then there may be a few iterations at the beginning where the points from the process are not typical of the distribution of interest. It is common practice, and one we follow, to delete some of the output of the sampler at the beginning. In this case, we use B=130, and delete the first 30, leaving 100 points in the sequence for estimation.

For example, consider $P(RH = black|gen1 = 1, ndep = 1)$. The trace and estimated posterior are given in the figure below.




In the top panel, 130 points are represented; the first 30 are greyed out, which corresponds to the points that have been deleted. Examination of the points at the beginning

show that at least two seem to be atypical of the sequence, which is the effect of the initialization to the BIFSG estimates. The bottom panel shows a density estimate from the last 100 points.