Comparison of weighting procedures in the presence of unit nonresponse: a simulation study based on data from the American Time Use Survey July 2018

Morgan Earp[1]
David Haziza[2]

[1]Bureau of Labor Statistics, PSB Suite 5930, 2 Massachusetts Avenue NE,
Washington, DC 20212, USA

[2]Universite de Montreal, 4253, Pavillon, André Aisenstadt,
Montreal, Quebec, Canada

1. Introduction. There are several weighting adjustment methods that can be used to adjust for potential bias resulting from survey nonresponse. Different types of weighting schemes can have different impacts on nonresponse bias and variance of the estimates. While some might lower the bias, they might also increase the variance, or where some have low variance, they might also result in higher average bias. When assessing and adjusting for nonresponse bias analysis, we can only examine the potential for bias so far as we have proxy variables related to key survey estimates. When we assess nonresponse bias, there are several reasons that we might not find evidence of nonresponse bias, four of which include: 1) there is no bias; 2) there was no bias for the key estimates for which we had proxy data; 3) we have controlled for all the variables that account for bias; or 4) we do not have very good proxy data. When we do find evidence of nonresponse bias, we can assess whether adding or removing variables from the current nonresponse adjustment method would reduce that bias without increasing the variance, using a simulation study.

When nonresponse adjustment is done in combination with calibration, or is followed up by calibration, it may be hard to see the impact of different weighting schemes, since calibration can smooth out differences across the different weighting scenarios. Not only can we change the variables we use to adjust for nonresponse, but we can also use a variety of different models to create nonresponse adjustment weights. In some cases, we might just use a simple raking method, in other cases we might create a propensity score model, and in other cases we might just use calibration to adjust for nonresponse. In the case of using raking versus a propensity score model, there are several variations to choose from. The impact of raking can depend on the number of variables and cells we include, and how related they are to the survey estimates, and the impact of using a propensity score model may vary depending on the variables, how related the variables are to nonresponse and key survey estimates, and the type of model used to generate the propensity scores. And since both raking and propensity score adjustment are often followed by calibration, it can sometimes be difficult to see any differences in bias or variance caused by using one method versus the other if we only compare the final estimate. However, for some surveys, calibration might not be an option if there are no or very limited targets on which to calibrate. This paper focuses on comparing the impact of different nonresponse adjustment scenarios prior to calibration.

2. Discussion of Survey Weighting and Nonresponse. Survey data are generally stored in rectangular data files, each row corresponding to a sample unit (a household, a person, or an establishment) and each column corresponding to a survey variable. In addition to the survey variables, the data file will also include two sets of weights; the base weights and the final weights. The process leading to the final weighting system generally consists of three main stages: 1) base weights, 2) nonresponse adjustment, and 3) calibration. In the first weighting stage, a base weight is assigned to every sample unit. Most often, the base weights are defined as the inverse of the inclusion probability in the sample. Virtually all the surveys face the problem of missing data. In particular, some sample units may not be reachable or may refuse to respond. This is referred to as unit nonresponse. Survey nonresponse always raises concern about the potential for nonresponse bias.

The goal of the second stage of the weighting process is to reduce the potential nonresponse bias. If certain households or persons within households are systematically less likely to respond to a survey, and they have common attributes related to what is measured by the survey, then there is potential for nonresponse bias. In order to adjust for nonresponse, we divide the base weights by the estimated response propensity.

After the weights undergo nonresponse adjustment, an additional modification is performed to ensure the consistency between survey estimates and known population totals. This process is referred to as calibration (Haziza & Beaumont, 2017). We provide an overview of the second stage below:

Let *U* denote a finite population of size *N* and let *s* be a sample, of size *n*, selected according a sampling design $p(s)$ with first-order inclusion probabilities $\pi_k,\ k=1,\ldots,N.$ The base weight attached to unit *k* is defined as $d_k = 1/\pi_k$ and the system $\{d_k; k \in s\}$ constitutes the basic weighting system. In this work, we are interested in estimating finite population totals. Let *y* be a generic survey variable. The total of the population *y*-values is $t_y = \sum_{k \in U} y_k.$ In the absence of nonresponse, the basic weighting system ensures that, when applied to the *y*-variable, the resulting estimator, $\hat{t}_y^F = \sum_{k \in s} d_k y_k.$ is design-unbiased for $t_y.$ This estimator is the well-known Horvitz-Thompson estimator.

In the presence of unit nonresponse, the information is collected on a subset s_r of *s*. The set s_r represents the set of respondents observed at the end of data collection. Conceptually, this set can be thought of as being generated according to a nonresponse mechanism $q(s_r \mid s)$, where the subscript *q* refers to the nonresponse mechanism. Let p_k be the probability of response associated with unit *k* and let $\hat{p}_k$ be an estimate of p_k. The weights adjusted for nonresponse are defined as

$$w_k^* = d_k / \hat{p}_k, \quad k \in s_r.$$

The weighting system adjusted for nonresponse is described as $\{w_k^*; k \in s_r\}$. Applying this weighting system to a *y*-variable leads to the so-called propensity score adjusted estimator

$$\hat{t}_y^{PSA} = \sum_{k \in s_r} w_k^* y_k.$$

To study the properties of $\hat{t}_y^{PSA}$, we consider its conditional nonresponse bias defined as

$$B_q(\hat{t}_y^{PSA}) = E_q(\hat{t}_y^{PSA} | s) - \hat{t}_y^F,$$

Where $E_q(.|\mathrm{s})$ denotes the expectation operator with respect to the nonresponse mechanism. The conditional nonresponse variance of

$$\hat{t}_y^{PSA} \text{ is } V_q(\hat{t}_y^{PSA} | s).$$

3. Analysis of Nonresponse Adjustment Simulation in the ATUS. There are a number of methods that exist for estimating the response probabilities $\hat{p}_k$ to ultimately construct the adjustment weights. We distinguish between parametric methods (that include logistic regression as a special case) from nonparametric methods (that include classification and regression trees as a special case).

In this work, we conduct an extensive simulation study based on the American Time Use Survey (ATUS) data to compare the bias and variance properties of both unadjusted and adjusted estimators. ATUS is an annual household survey sponsored by the Bureau of Labor Statistics and conducted by the U.S. Census Bureau using Computer Assisted Telephone Interviews (CATI). The ATUS is used to estimate how people spend their time. The ATUS sample is drawn from the population of households that responded to the Current Population Survey (CPS); and therefore CPS data is available for the full ATUS sample. Characteristics of ATUS respondents and nonrespondents can be modeled using the CPS frame data as a proxy for ATUS respondent and nonrespondent characteristics. The CPS frame data is strongly correlated with the ATUS sample respondent data on respondent age, household size, number of children, respondent sex, respondent employment status (see Table 1), and the ATUS key estimates are somewhat to moderately correlated with respondent CPS employment status (see Table 2). The CPS also collects data on a number of respondent and household characteristics (some that are collected on the ATUS, some that are not) that can be used to model characteristics of ATUS respondents versus nonrespondents (see Table 3).

Table 1: CPS & ATUS Variable Correlations

Variable Name	Pearson/Point Biserial Correlation
Respondent Age	.99
Household Size	.97
Number of Children < 18	.82
Respondent's Sex	.99
Respondent Employment Status	.84

Table 2: CPS Employment Status & ATUS Variable Correlations

Variable Name	Pearson/Point Biserial Correlation
Sleeping	-.14
Household Activities	-.11
Housework	-.07
Food Prep	-.12
Caring for and Helping Household Members	.04
Care of Household Children	.05
Caring for and Helping Household Children	.05
Socializing, Relaxing, and Leisure	-.34
Communicating	-.04
Watching TV	**-.26**
Sports, Exercise, and Recreation	.03

Table 3: ATUS Frame Variables

Variable Name
CPS Household Ownership Type
CPS Household Income
CPS Household Income Missing Indicator
CPS Education Level
CPS Respondent Sex
CPS Presence of Child
CPS Respondent Race
CPS Respondent Employment Status
CPS Respondent Age
CPS Household Type
CPS Household Size
CPS Race Indicator
CPS Respondent Hispanic (Y/N)
CPS Respondent Black (Y/N)
CPS Respondent Disability Flag

CPS Number of Children under 18
Geographic Division
Geographic Region
Metropolitan Status
ATUS Survey Incentive
ATUS Interview Reference Day (Weekend vs. Weekday)
ATUS Interview Reference Day (Sun-Sat)

3a. ATUS Imputation. Our data file included the entire ATUS sample for the first quarter of 2015. The ATUS sample was subject to unit nonresponse. That is, there were missing y-variables for a portion of the ATUS sample. We started our analysis by first completing the ATUS data file using imputation. We used the CPS auxiliary information which is available for all the ATUS sample units to create imputation values. The y-variables were imputed separately (this is often called marginal imputation) using random hot-deck imputation within classes to get a complete data file. The imputation classes were formed on the basis of CPS auxiliary information. The use of marginal imputation that does not preserve the relationship between variables is justified by the fact that we were only interested in univariate parameters (population totals) for each y-variable. After imputation, the data file consisted of 6,230 records representing the ‘‘full ATUS sample’’. Note that the original ATUS sample exhibited a response rate of 50 percent approximately.

3b. ATUS Nonresponse Simulation. For each method, we generated 1,000 replicates using two distinct nonresponse simulation models from our sample of 6,230 records; 1) logistic regression; and 2) a regression tree. We used both a logistic regression model and a regression model to simulate nonresponse, since we assumed that the weights created using a logistic regression model would perform better when nonresponse was simulated using logistic regression, and that weights created using a regression tree model would perform better when nonresponse was simulated using a regression tree model. Both nonresponse simulation models included all of the available CPS frame variables (see Table 3) and led to an approximate response rate of 50 percent in each replicate. We compared the correlation between actual ATUS 2015 response indicator and simulated response indicator to access how similar the nonresponse simulation models were to actual ATUS response. The simulation scenario using a logistic regression model appeared to be more strongly correlated with actual ATUS nonresponse than the simulation scenario using a regression tree model (Table 4).

Table 4: Simulated ATUS Nonresponse versus Actual ATUS Response Biserial Correlations

Nonresponse Simulation Model	Correlation
Regression Tree	0.21
Logistic Regression	0.28

3c. Nonresponse Adjustment. In order to assess the impact of adding additional variables to the ATUS nonresponse adjustment, we compared bias and variance across five weighting schemes: 1) current ATUS adjustment, 2) logistic regression, 3) weighting classes based on logistic regression , 4) regression tree using a CHAID growth method, and 5) regression tree using Gini growth method. In addition, we computed unadjusted estimators as follows:

Unadjusted: Constant estimated response probabilities: the probability of response for each unit was estimated by the overall response rate n_r/n, where n_r denotes the number of respondents to the survey. That is, $\hat{p}_k = n_r/n$ for all *k*.

We now describe the five weighting schemes in more detail:

(1) Current ATUS Nonresponse Adjustment: The current ATUS procedures that consists of creating 14 weighting cells based on the day of the week that a person is interviewed and whether or not they received an incentive for their interview. The ATUS nonresponse adjustment factor is calculated using the actual response rates observed within each of the 14 cells. That is $\hat{p}_k = n_{rg}/n_g$, where n_g and n_{rg} denote respectively the number of sample units and the number of respondents in each cell g. $g = 1, \dots, 14$. .

The current ATUS nonresponse adjustment uses a logistic regression model with three inputs (see Table 5), and relies on calibration to adjust for any remaining bias. Discussions with the BLS ATUS program office, indicated that nonresponse may also be related to the variables listed in Table 6, and therefore, we have added these variables to the logistic regression and regression tree models.

Table 5: Current ATUS Nonresponse Adjustment Inputs:

Variable Name
ATUS Survey Incentive
ATUS Interview Reference Day (Sun-Sat)
ATUS Survey Incentive * ATUS Interview Reference Day (Sun-Sat)

Table 6: Propensity Score Model Inputs:

Variable Name
CPS Household Ownership Type
CPS Household Income
CPS Household Income Missing Indicator
CPS Education Level
CPS Respondent Sex
CPS Presence of Child
CPS Respondent Race
CPS Respondent Employment Status
CPS Respondent Age
ATUS Survey Incentive
ATUS Interview Reference Day (Weekend vs. Weekday)

(2) Logistic Regression: The estimated response probabilities are given by

$$\hat{p}_k = \frac{\exp\left(\mathbf{x}_k'\hat{\boldsymbol{\beta}}\right)}{1+\exp\left(\mathbf{x}_k'\hat{\boldsymbol{\beta}}\right)},$$

where $\mathbf{x}_k$ is a vector of fully observed variables (shown in Table 6) attached to unit *k* believed to be associated with the probability of response.

(3) Logistic Regression Using Weighting Classes: Preliminary estimated response probabilities $\hat{p}_k$ are first obtained through a logistic regression model for $k \in s$based on the predictors listed in Table 6. These preliminary response probabilities are then ordered from the lowest to the largest. Weighting classes of equal size are then formed with respect to the $\hat{p}_k$'s. In our experiments, we used 10 weighting classes. In each class, the response probability of unit *k* in class *g* was estimated by the actual response rate observed in the same class. That is, $\hat{p}_k = n_{rg} / n_g,$ where n_g and n_{rg} denote respectively the number of sample units and the number of respondents in class g, $g = 1, \ldots, 10.$

(4) Regression Tree Model Using CHAID: The CHAID tree model uses a combination of a Chi-Square and an F test for to test for significance testing (depending on the level of the predictor variables included the model), and attempts to identify significant predictors of survey response. The regression tree sequentially evaluates each of the predictor variables with relation to survey response, using the p value to select the most significant predictor, as well as to identify the optimal breakpoint for classifying and distinguishing between respondents and nonrespondents. After the initial split is carried out, the model automatically identifies interaction, mediating, and moderating effects by continuing to further segment the data using recursive partitioning until no more significant splits can be identified, or until the growth limit in terms of number of leaves, branches, or the depth of the tree is reached. When the tree is complete, it will be comprised of potentially several leaves also referred to as end nodes, which are mutually exclusive groups that are each internally homogenous and externally heterogeneous with respect to propensity scores, and together comprise the full dataset. In each end node, the response probability of a unit was estimated by the actual response rate observed in the same end node.

(5) Regression Tree Model Using Gini Index: The Gini tree method is similar to the CHAID method, but instead of using statistical testing, it uses the Gini index as a measure of impurity (heterogeneity) to identify optimal tree splits. The lower the index is, the more pure (homogenous the groupings are). The Gini index is calculated as follows:

$$i(t) = \sum_{i \neq j} p(i|t) p(j|t),$$

where $p(j|t)$ is the predicted response propensity for class *j* with an assigned binary value of 1 for node t, and $p(i|t)$ is the inverse of that $[1 - (p(j|t))]$ with an assigned value of 0 for node t (Breiman, Friedman, Olshen, & Stone, 1984).

Note that weighting schemes two through five included the following additional variables not currently used in the ATUS weighting method shown in weighting scheme one: respondent age, household size, number of children in the household, respondent sex, respondent education level, reported household income, and whether household income was reported or imputed.

3d. ATUS Nonresponse Bias and Variance Estimation of Key ATUS Estimates. Using a simulation study, we were able to assess whether sample estimates were consistently over or under (bias) and how much the sample estimates varied (variance) using a given weighting scheme. The results of this paper will focus on how bias and variance compared across the five different nonresponse weighting adjustment methods. Nonresponse bias and variance were assessed for 13 key ATUS estimates. These 13 key ATUS estimates were identified by the ATUS program office and are shown in Table 7.

Table 7: ATUS Y Variables:

Variable Name
Sleeping
Household Activities
Housework
Caring for and Helping Household Members
Care of Household Children
Caring for and Helping Household Children
Socializing & Communicating
Watching TV
Sports, Exercise, and Recreation
Participating in Sports, Exercise, and Recreation
Travel
Volunteering
Secondary Childcare

The performance of the adjustment weights varied across the 13 key estimates; meaning that some weights did a better job reducing bias and/or variance for some variables and a worse job for other variables. However, since it is standard practice to use a single weighting adjustment procedure for all the estimates, we were more interested in how the adjustment weights performed overall across all 13 key estimates than we were in how they performed for a single estimate. In order to compare the performance of the five weighting adjustment methods in both of the simulation scenarios described above, we plotted bias and variance values for each of the 13 key estimates and compared the overall performance across each of the five weighting schemes.

As a measure of weight variation, we calculated the variance of each of the five weighting methods.

$$VAR_{(W_k^{NR})} = \frac{\sum\left(w_k^{NR} - \widehat{w}_k^{NR}\right)^2}{n-1}$$

We also computed the relative variance of a given weight adjustment method, using the ATUS estimator as the reference:

$$RVAR_{(W_k^{NR})} == 100\, X \frac{VAR_{(w_k^{NR})}}{VAR_{(w_k^{ATUS})}}.$$

As a measure of relative bias of an estimate ($\hat{t}_y^{NR}$), we calculated an average of the Monte Carlo percent of relative bias across all 1,000 replicates:

$$RB_{mc}(\hat{t}_y^{NR}) = 100\ X\ \frac{E_{mc}\,(\hat{t}_y^{NR} - \hat{t}_y^{F})}{\hat{t}_y^{F}}\,,$$

where

$\hat{t}_y^{NR}$ denotes an estimator after nonresponse treatment;
$\hat{t}_y^{F}$ denotes the full sample estimate from ATUS;
E_{mc} (.) denotes the Monte Carlo average over the 1000 replicates.

Relative bias less than zero indicates negative bias, meaning we underestimate given nonresponse. Relative bias greater than zero indicates positive bias, meaning we overestimate given nonresponse. We also computed the absolute value of the relative bias of a given weight adjustment method, using the ATUS estimator as the reference:

$$RBC_{mc}(\hat{t}_y^{NR}) = ABS\ (100\ X\,\frac{RB_{mc}(\hat{t}_y^{NR})}{RB_{mc}(\hat{t}_y^{ATUS})}).$$

If $RBC_{mc}(\hat{t}_y^{NR})$= 100, the estimator based on a weight adjustment procedure exhibits the same bias as the ATUS estimator. If $RBC_{mc}(\hat{t}_y^{NR})$ < 100, the estimator based on a weight adjustment procedure is more efficient than the ATUS estimator.

We also computed the average Monte Carlo Mean Square Error:

$$MSE_{mc}(\hat{t}_y^{NR}) = E_{mc}(\hat{t}_y^{NR} - \hat{t}_y^{F})^2$$

Lastly, we computed the relative efficiency of a given weight adjustment method, using the ATUS estimator as the reference:

$$RE(\hat{t}_y^{NR}) = 100\ X\,\frac{MSE_{mc}(\hat{t}_y^{NR})}{MSE_{mc}(\hat{t}_y^{ATUS})}.$$

If $RE(\hat{t}_y^{NR})$ = 100, the estimator based on a weight adjustment procedure exhibits the same efficiency as that of the ATUS estimator. If $RE(\hat{t}_y^{NR})$ < 100, the estimator based on a weight adjustment procedure is more efficient than the ATUS estimator.

3e. Methodological Overview. An overview of the steps taken in this paper to assess and compare alternative weighting schemes to the current ATUS weighting scheme are laid out in Figure 1.

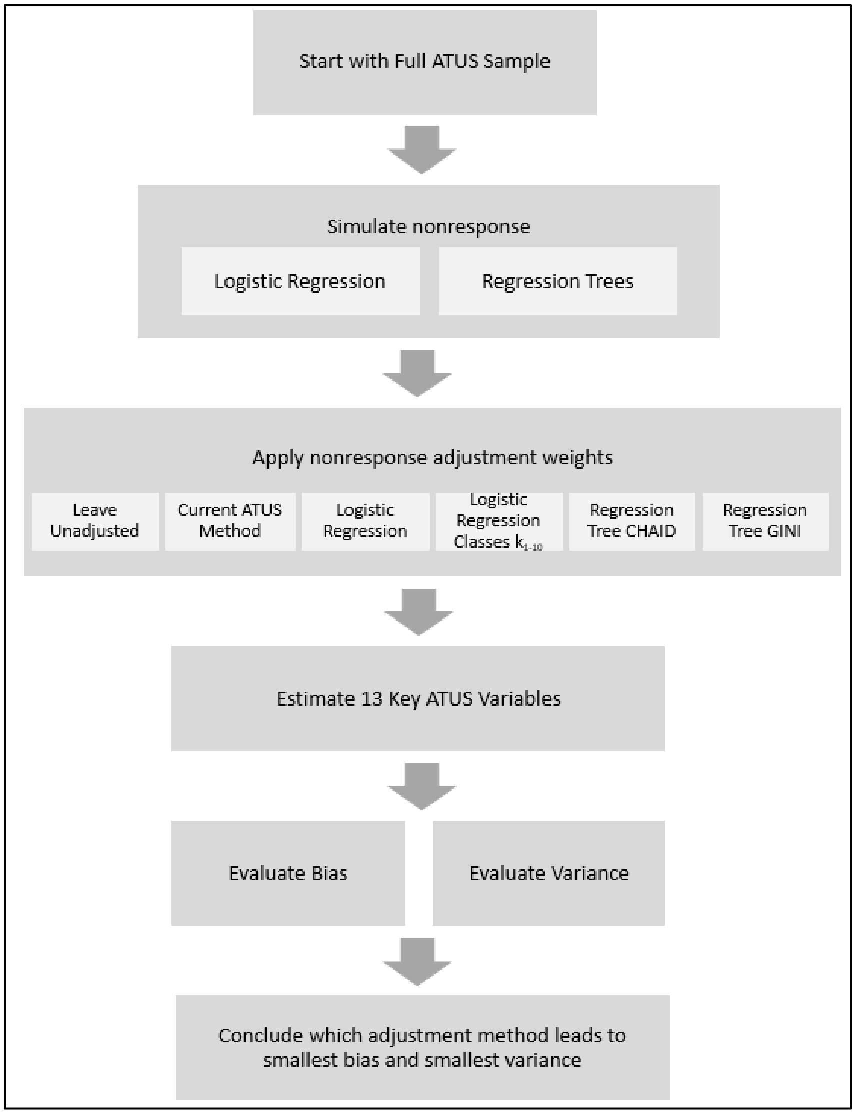
Start with Full ATUS Sample
Simulate nonresponse
Logistic Regression
Regression Trees
Apply nonresponse adjustment weights
Leave Unadjusted
Current ATUS Method
Logistic Regression
Logistic Regression Classes k_{1-10}
Regression Tree CHAID
Regression Tree GINI
Estimate 13 Key ATUS Variables
Evaluate Bias
Evaluate Variance
Conclude which adjustment method leads to smallest bias and smallest variance

***Figure* 1 Research Study Process Flow**

4. Results. We compared each of the four alternative weighting schemes to the current ATUS weighting method by looking at the variance of the weights, the bias, and the mean square error.

4a. Variance of the Weights. When we look at just the variance of the weights, we see that the logistic regression model using classes was the only weighting scheme to result in less variation in the adjustment weights than ATUS, whether we simulated nonresponse using a tree or a logistic regression model (see Figure 2). When we used a regression tree model to simulate nonresponse, the logistic regression weights had less variation than either set of weights produced using regression trees, but more variation than the ATUS weights or the class weights. When we used a logistic regression model to simulate nonresponse, the regression tree weights had less variation than the weights produced using a logistic regression model, but more variation than the ATUS weights or the class weights.

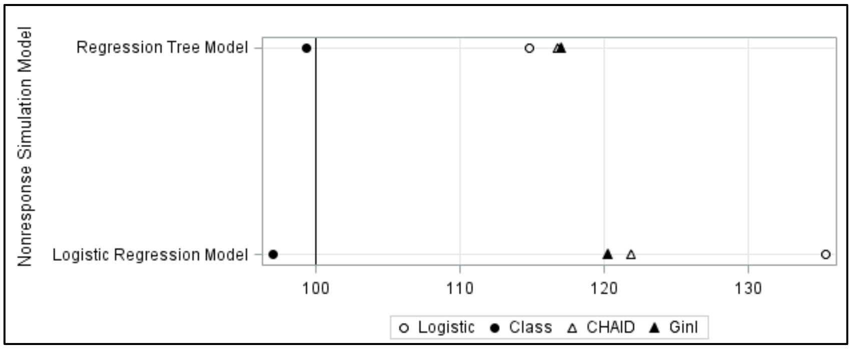

Figure 2: Relative Variance of the Weights by Nonresponse Simulation Model. Values less than 100 indicate less variation in the weights than ATUS, and values greater than 100 indicate more variation in the weights than ATUS. Note that CHAID and Gini have very similar levels of variation in the weights using a regression tree model to simulate nonresponse.

4b. Nonresponse Bias. When we compared the nonresponse bias using each of the alternative weighting schemes relative to the bias produced using the ATUS weights, trees tended to produce the least amount of bias across the thirteen key estimates (see Figure 3). When we simulated nonresponse using a regression tree model, both the logistic regression and the logistic regression class weights tended to perform about the same as ATUS; however, when we used logistic regression to simulate nonresponse, they varied in performance; the logistic regression weights tended to outperform the ATUS weights, where the class weights tended to perform relatively the same as the ATUS weights.

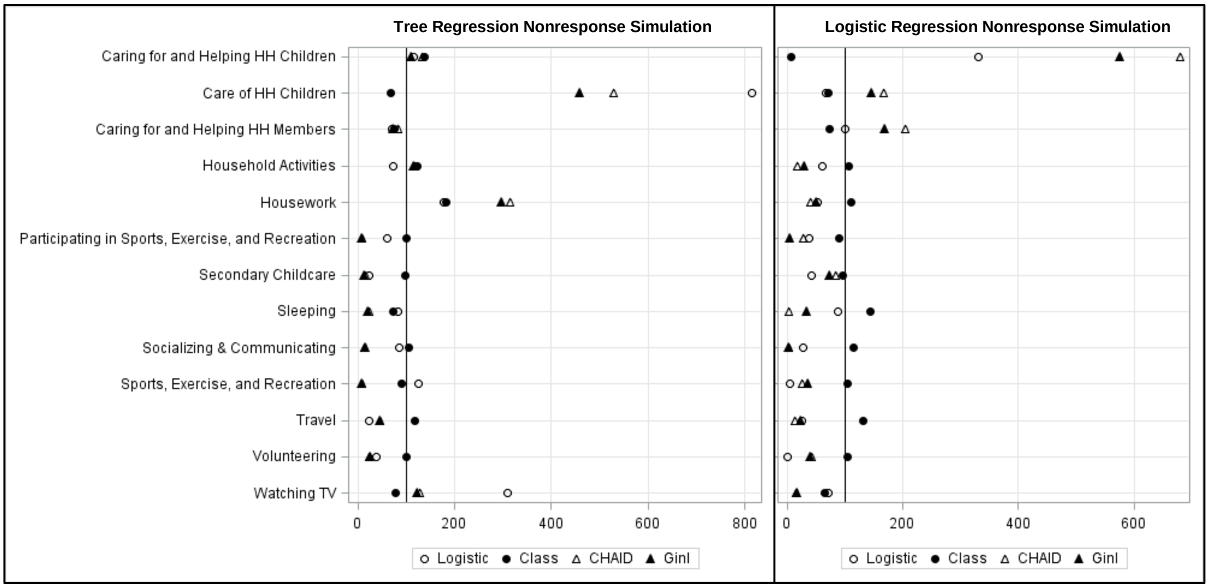

Figure 3: Monte Carlo Percent Relative Bias of the Adjusted Estimator using the Tree versus the Logistic Regression Nonresponse Simulation Scenarios with all CPS Frame Variables. Values less than 100 indicate less relative bias than using the ATUS weights, and values greater than 100 indicate more bias than using the ATUS weights.

4c. Mean Square Error. When we compared the mean square error using each of the alternative weighting schemes relative to the mean square error produced using the ATUS weights, the results varied depending on the model used to simulate nonresponse. When we used a tree model to simulate nonresponse, the tree weights tended to perform worse than the ATUS weights, and the logistic regression weights performed better for some estimates and worse for others; while the class weights tended to perform similarly to ATUS (see Figure 4). When we used a logistic regression model to simulate nonresponse, the tree and the logistic regression weights tended to outperform ATUS for about half of the estimates, and trees tended to slightly outperform logistic regression in those cases. In the cases where trees and logistic regression outperformed ATUS, the class weights tended to perform worse than ATUS, and in the cases where the tree and logistic regression weights performed worse than ATUS, the class weights tended to outperform ATUS (see Figure 4). Regardless of how nonresponse was simulated, trees tended to perform better than ATUS with regard to sports, exercise, and recreation type activities; secondary childcare; and travel; and they tended to perform worse than ATUS when adjusting for time spent caring for children and household members, as well as time spent socializing and communicating.

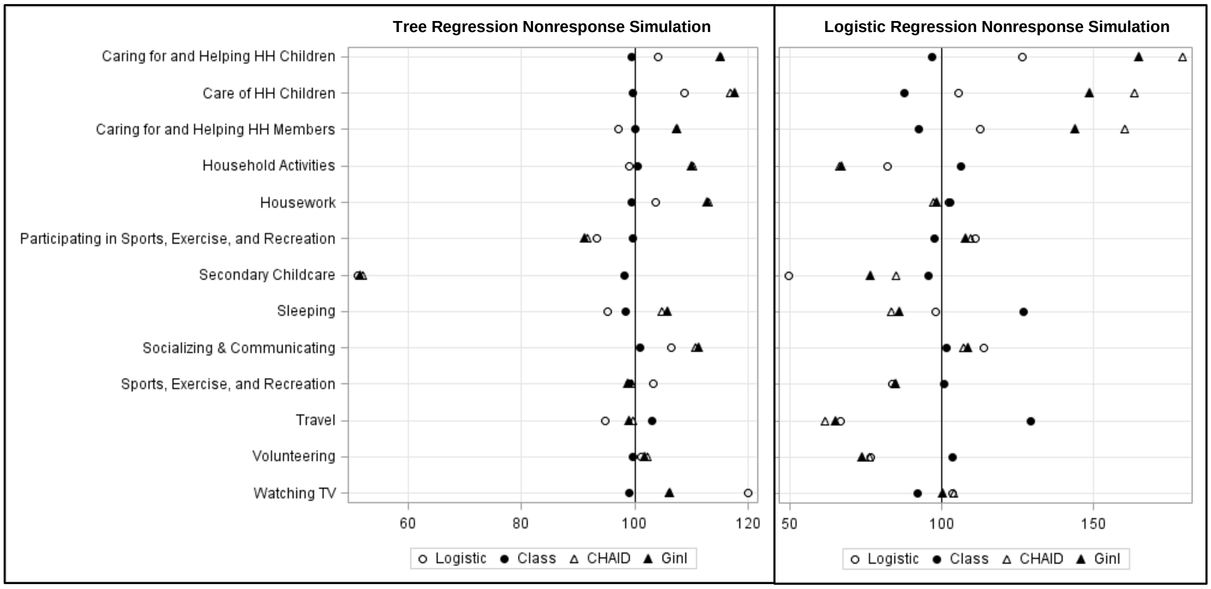

Figure 4: Mean Square Error of the Adjusted Estimator using the Tree versus the Logistic Regression Nonresponse Simulation Scenario with all CPS Frame Variables. Values less than 100 indicate less mean square error than using the ATUS weights, and values greater than 100 indicate higher mean square error than using the ATUS weights.

If we focus specifically on the estimates used for press releases such as time spent on Caring for and Helping Household Members, Housework, Sleeping, Socializing & Communicating, and Watching TV, we see that the tree weights tend to result in less bias than the ATUS weights, except for time spent Caring for and Helping Household Members (according to both nonresponse simulation models) and Housework (according to the tree nonresponse simulation model). Using the regression tree nonresponse simulation model, we saw higher mean square error values in comparison to ATUS for all five of these estimates; however, when we used a logistic regression model to simulate nonresponse, trees resulted in lower mean square error for Housework and Sleeping, but not for time spent Caring for and Helping Household Members, Socializing & Communicating, and watching TV.

5. Discussion. Regression tree weights tended to result in less bias than logistic regression, class, or ATUS weights, however they also tended to have higher variance both in terms of the weights themselves, and in terms of the estimates with respect to mean square error values. Depending on how nonresponse is simulated, trees may perform worse overall with regard to mean square error or it may vary based on the estimate. The nonresponse simulation model based on regression trees, takes into account interaction effects, where the logistic regression simulation only models main effects. And while the logistic regression model does not account for interaction effects, the nonresponse simulated using logistic regression was more highly correlated with actual ATUS nonresponse than the nonresponse simulated using regression trees. If the nonresponse simulation model using logistic regression is a more accurate depiction of actual ATUS nonresponse (as the higher correlation might indicate), then using propensity score weighting with regression trees may be a good alternative both in terms of bias and mean square error; however, if actual ATUS nonresponse is closer to what was simulated using a regression tree model, then the current ATUS nonresponse adjustment model may work better when we take into account variation of the weights and mean square error.

Given that ATUS is used to produce trend estimates of how Americans spend their time, very careful consideration would have to be given to changing the weighting method used to adjust for nonresponse, since it would require reweighting previous datasets or making a break in the time series. One possible consideration might include exploring the use of regression trees to identify additional post strata adjustment targets as proposed by McConville and Toth (2017) for the Occupational Employment Statistics Survey. The ATUS program could consider using regression trees to explore and identify additional post adjustment strata beyond incentive and interview day to adjust for nonresponse.

6. Limitations. The focus of this paper was to compare the current ATUS weighting methods with alternative propensity score weighting methods. By taking a ratio of the relative bias and mean square error over the ATUS relative bias or mean square error, we are only assessing whether the propensity score method being compared is better or worse than the current ATUS weighting method. This is not an estimate of the actual bias or mean square error. A comparison of the actual estimate of the relative bias across all the weighting methods, as well as a comparison of the mean square error compared to the unadjusted weight can be found in the appendix.

7. References.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Boca Raton, FL; Taylor & Francis Group.

Haziza, D., & Beaumont, J. F. (2017). Construction of weights in surveys: A review. *Statistical Science*, *32*(2), 206-226. Can be accessed here: http://www.imstat.org/publications/sts/sts_32_2.pdf#page=11

Killion, R. A. (2006). Weighting Specifications for the American Time Use Survey (ATUS) for 2006. *US Bureau of the Census, Internal Memo (Doc.# ATUS-16)*.

McConville, K. S., & Toth, D. (2017). Automated Selection of Post-Strata using a Model-Assisted Regression Tree Estimator (submitted manuscript). Can be accessed here: https://arxiv.org/pdf/1712.05708.pdf .

8. Appendix.

Nonresponse Bias & Variance Assessment

As a measure of relative bias of an estimate ($\hat{t}_y^{NR}$), we calculated the average Monte Carlo percent of relative bias across all 1,000 replicates

$$RB_{mc}\left(\hat{t}_y^{NR}\right) = 100\ X\ \frac{E_{mc}\left(\hat{t}_y^{NR} - \hat{t}_y^{F}\right)}{\hat{t}_y^{F}},$$

where

$\hat{t}_y^{NR}$ denotes an estimator after nonresponse treatment;
$\hat{t}_y^{F}$ denotes the full sample estimate from ATUS;
E_{mc} (.) denotes the Monte Carlo average over the 1000 replicates.

Relative bias less than zero indicates negative bias, meaning we underestimate given nonresponse. Relative bias greater than zero indicates positive bias, meaning we overestimate given nonresponse. We expect relative bias to vary across different estimates, and that some nonresponse adjustment weights will adjust better for nonresponse than others, depending on the weight adjustment method.

Monte Carlo mean square error

We also computed the average Monte Carlo Mean Square Error:

$$MSE_{mc}\left(\hat{t}_y^{NR}\right) = E_{mc}\left(\hat{t}_y^{NR} - \hat{t}_y^{F}\right)^2$$

Lastly, we computed the relative efficiency of a given weight adjustment method, using the unadjusted estimator as the reference:

$$RE\left(\hat{t}_y^{NR}\right) = 100\ X\ \frac{MSE_{mc}\left(\hat{t}_y^{NR}\right)}{MSE_{mc}\left(\hat{t}_y^{Unadjusted}\right)}.$$

If $RE\left(\hat{t}_y^{NR}\right)$ = 100, the estimator based on a weight adjustment procedure exhibits the same efficiency as that of the unadjusted estimator. If $RE\left(\hat{t}_y^{NR}\right)$ < 100, the estimator based on a weight adjustment procedure is more efficient than the unadjusted estimator.

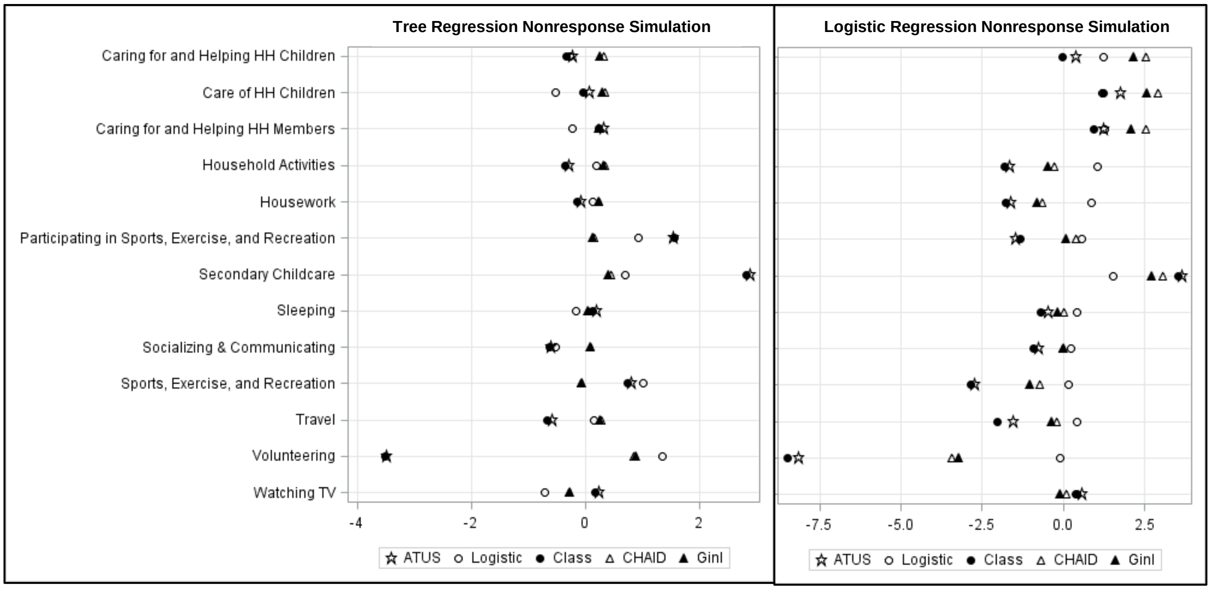

Figure A1: Monte Carlo Percent Relative Bias of the Adjusted Estimator using the Tree versus the Logistic Regression Nonresponse Simulation Scenarios with all CPS Frame Variables.

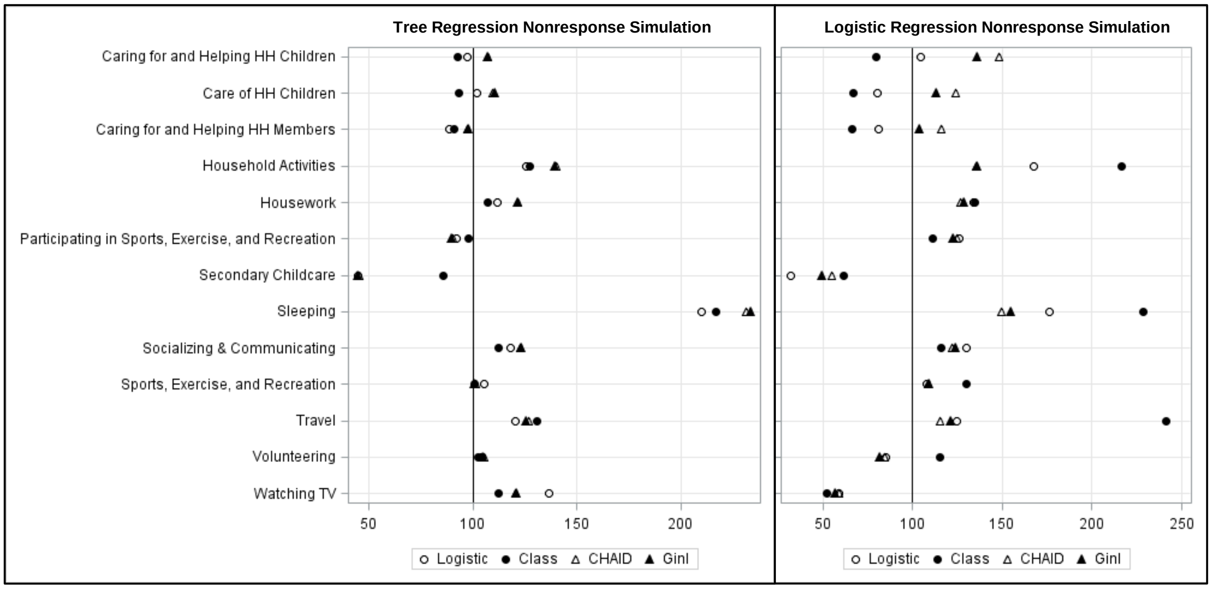

Figure A2: Mean Square Error of the Adjusted Estimator using the Tree versus the Logistic Regression Nonresponse Simulation Scenario with all CPS Frame Variables. Values less than 100 indicate less mean square error than using the unadjusted weights, and values greater than 100 indicate higher mean square error than using the unadjusted weights.