Comparing the Ability of Regression Modeling and Bayesian Additive Regression Trees to Predict Costs in a Responsive Survey Design Context

James Wagner¹, Brady T. West¹, Michael R. Elliott¹, and Stephanie Coffey²

Responsive survey designs rely upon incoming data from the field data collection to optimize cost and quality tradeoffs. In order to make these decisions in real-time, survey managers rely upon monitoring tools that generate proxy indicators for cost and quality. There is a developing literature on proxy indicators for the risk of nonresponse bias. However, there is very little research on proxy indicators for costs and almost none aimed at predicting costs under alternative design strategies. Predictions of survey costs and proxy error indicators can be used to optimize survey designs in real time. Using data from the National Survey of Family Growth, we evaluate alternative modeling strategies aimed at predicting survey costs (specifically, interviewer hours). The models include multilevel regression (with random interviewer effects) and Bayesian Additive Regression Trees (BART).

Key words: Survey cost models; machine learning.

1. Introduction

Surveys are conducted in an environment of uncertainty. Many key design parameters are random variables. This makes it difficult to optimize a survey design before the field period begins. A new approach to survey design, called responsive survey design, is a direct reaction to this uncertainty. In a responsive design, the estimates of key design parameters are updated during the field period based on incoming data. This allows surveys to “recalibrate” designs toward optimality.

Responsive survey designs were first proposed by [Groves and Heeringa \(2006\)](#). They proposed that such designs should include the following elements: a) the pre-identification of key design features that have the largest potential impact on costs and errors in the survey, b) identification of a set of indicators for the cost and error properties of those design features to be monitored during the field period, c) changes to the design

¹ University of Michigan, 4053 ISR, 426 Thompson St., Ann Arbor, MI 48104, U.S.A. Emails: jameswag@umich.edu, bwest@umich.edu and mrelliot@umich.edu

² U.S. Census Bureau, 4600 Silver Hill Road, Suitland, MD 20746, U.S.A. Email: stephanie.coffey@census.gov
Acknowledgments: This work was supported by a grant from the National Institutes for Health (#1R01AG058599-01; PI: Wagner). The National Survey of Family Growth (NSFG) is conducted by the Centers for Disease Control and Prevention’s (CDC’s) National Center for Health Statistics (NCHS), under contract # 200-2010-33976 with University of Michigan’s Institute for Social Research with funding from several agencies of the U.S. Department of Health and Human Services, including CDC/NCHS, the National Institute of Child Health and Human Development (NICHD), the Office of Population Affairs (OPA), and others listed on the NSFG webpage (see <http://www.cdc.gov/nchs/nsfg/>). The views expressed here do not represent those of NCHS or the other funding agencies.

features over periods of time known as phases, based upon pre-specified decision rules, and d) combining data from the separate phases into a single estimate. The goal is to optimize cost-error tradeoffs over the phases. These tradeoffs are made within a total survey error framework (Biemer et al. 2017), where the total error is minimized for a fixed budget. For example, an initial phase might be highly successful at recruiting women but much less successful at recruiting men. A subsequent phase might be designed to complement this initial phase. The follow-on phase would aim to increase the participation of men.

Key to this process is monitoring the incoming data to know when a phase has reached its capacity and is no longer effectively reducing survey error. This might happen when a recruitment protocol is creating imbalances in who responds, or when the protocol becomes ineffective and very costly. The latter might occur, for example, in a face-to-face survey when continued contact attempts lose their effectiveness and become costly relative to the return.

The literature presents evidence of the successes that are possible when applying Responsive Survey Design (RSD) in practice. For example, the implementation of RSD in the National Survey of Family Growth led to a 25% reduction in the per-interview cost (Kirgis and Lepkowski 2013). Other projects have also implemented responsive survey designs with demonstrated success (Mohl and Laflamme 2007; Peytchev et al. 2009; Tabuchi et al. 2009; Kleven et al. 2010; Laflamme and Karaganis 2010; Lundquist and Särndal 2013; Barber et al. 2011; Finamore et al. 2013). However, these successes have been small in number and limited in scope (for a review, see Tourangeau et al. 2017).

One reason for the limited success may be that predictions about future outcomes (survey variables and costs) under alternative designs may be inaccurate. These sorts of predictions may be used to determine when a phase is over (Rao et al. 2008; Wagner and Raghunathan 2010; Lewis 2017; Paiva and Reiter 2017) and what the design of the next phase should be (Luiten and Schouten 2013; Rosen et al. 2014; Lynn 2016; Plewis and Shlomo 2017; Durrant et al. 2017). In either instance, the effectiveness of the responsive design is likely to be reduced if the predictions are inaccurate. Inaccurate predictions could lead, for example, to placing a phase boundary too early with the result that a more expensive design is implemented before the less expensive design has been fully exhausted. Inaccurate predictions about the types of respondents likely to be recruited over the different phases can lead to inefficiencies. It is also true that inaccurate predictions about the costs of a new phase could also lead to inefficient decisions. To date, no studies have attempted to evaluate methods for predicting costs in a responsive design framework.

This study attempts to address this gap by developing methods for predicting costs. These cost predictions would be a useful input for comparing alternatives and making design decisions. In this study, we assess our ability to accurately predict costs using different modeling approaches. We begin with a description in the Background section of the role of survey costs in the responsive design framework. We then discuss in the Data section the survey we will be examining and the data we use from that survey for our analyses. We then evaluate three different approaches to predicting costs in a face-to-face survey. These approaches are described in the Methods section. The approaches involve

using existing data to predict costs with three different modeling strategies. We then compare the results across replications from six different time points.

2. Background

Responsive survey designs optimize tradeoffs between survey errors and costs over a series of phases. Knowing about the likely errors and costs under each phase is an important aspect of an RSD. Groves and Heeringa (2006) specify that survey designers should choose a set of indicators for both costs and errors that will be monitored by the survey during data collection. These indicators are then used as inputs to the responsive design. Specifically, these indicators are used as inputs to decision rules about when to change design phases and what the design of the next phase should be.

Some studies have examined how to specify a rule for determining when the current design is not leading to changes in estimates (Rao et al. 2008; Wagner and Raghunathan 2010; Lewis 2017; Paiva and Reiter 2017). Other studies have examined inputs to decision rules, such as predicted response propensities (Luiten and Schouten 2013; Rosen et al. 2014; Lynn 2016; Plewis and Shlomo 2017; Durrant et al. 2017). To date, no studies have examined the most effective methods for the real-time prediction of costs. Groves and Heeringa (2006) report on the impact of responsive design on survey costs, but in a *post hoc* analysis of the costs. Other studies have focused on the error side of the cost-error tradeoff (see Groves 2006 for a review of studies on nonresponse bias; see also Biemer and Trewin 1997 for a review of studies on measurement error). In fact, the study of survey costs, in general, is limited. In particular, we are aware of no studies focusing on the prediction of survey costs.

Improving the predictions of costs and errors may be key to improving responsive survey designs. A recent review of responsive and adaptive designs found that the impact of these designs was often limited (Tourangeau et al. 2017). One issue that may blunt the impact of these designs is inaccurate predictions of future costs or errors. These inaccuracies could lead to less than optimal designs, thereby mitigating the impact of interventions. For example, an error in the prediction of the impact of a changed incentive may lead to a design change that occurs too late, thereby prolonging a relatively inefficient design and reducing the overall efficiency of the survey.

Techniques for improving predictions are a burgeoning area of research. In addition to standard regression techniques, machine learning techniques have been applied to prediction problems in many fields, but less so in survey research (for a review, see Kern et al. 2019). Exceptions include using machine learning to code open-ended responses (Schonlau and Couper 2016). In particular, a class of models known as Bayesian Additive Regression Trees (BART, Chipman et al. 2010) may be useful for the prediction of survey costs. BART models have been used in a variety of settings. For example, they have been used to detect spam (Abu-Nimeh et al. 2008), model treatment effects in an experiment with survey questionnaires (Green and Kern 2012), predict driving behavior (Tan et al. 2018), and inform survival analysis (Sparapani et al. 2016). These models can also be used for both continuous and binary outcomes. In this study, we evaluate the ability of BART models to improve predictions of survey costs (specifically, interviewer hours expended) relative to regression modeling in a responsive design framework.

3. Data

3.1. Description of the Survey

The data for the present study come from the National Survey of Family Growth 2011–2019 (NSFG). This survey collects information on family formation, fertility, and other related topics. The survey population is persons living in the United States between the ages of 15 and 49 (prior to September 2015, the eligible population was persons in the United States between the ages of 15 and 44). A complete description of the survey, including questionnaires and field procedures, is available at <http://www.cdc.gov/nchs/nsfg/>.

The NSFG is a face-to-face survey that is conducted continuously. There are four quarters of data collection each year. In this analysis, we use data from 27 quarters (dating back to September 2011), but make predictions for six quarters (approximately January 2017–June 2018). The NSFG has a multi-stage area probability sample design. The primary stage units (PSUs) are counties and Metropolitan Statistical Areas. Each year, a new sample of PSUs is released. The second stage units (SSUs) are neighborhoods defined by Census Blocks. Each quarter, a new sample of SSUs is released within each of the sample PSUs. A sample of housing units is then released within each PSU. Interviewing is conducted in two stages. Interviewing staff first attempt to visit each housing unit and determine whether an eligible person lives there. This is known as “screening.” Once an eligible person has been identified and selected, an in-depth interview on fertility, family formation, and related topics is attempted. This second stage is known as the “main” interview.

The NSFG uses a responsive design approach. There are two phases. The first phase is defined by time and is completed in exactly ten weeks. The first phase design includes prenotification by standard mail, no token of appreciation for the screening interview, and a promised USD 40 token of appreciation for completion of the survey. The ten-week phase boundary was determined to be optimal using data from prior experience with NSFG data collection (Kirgis and Lepkowski 2013) and is fixed in advance.

The second phase lasts two weeks. In the second phase, a subsample of active cases is selected, and prenotification occurs using a Priority Mail package. This prenotification includes a token of appreciation for the screening interview (USD 5) for households that have yet to complete that stage and a prepaid token of appreciation (USD 40) for households where screening has been completed and an eligible person has been selected, with an additional token (USD 40, or USD 80 total) promised. This design has been shown to be effective in reducing the bias of NSFG estimates (Peytchev et al. 2010; Axinn et al. 2011).

In this design, the phase boundary is fixed (at ten weeks). In a responsive survey design, the phase boundary should be determined by observed changes in the field. The decision would be triggered by proxy indicators for costs and survey errors. In this study, our objective is to evaluate the ability of different methods to predict the costs associated with phase two of this design. The phase two costs include the cost of the mailed materials, the incentives, and the cost of interviewer time. If we can effectively predict the costs associated with this design phase, these predictions could be used to determine a more optimal time to switch to the second phase.

3.2. Description of Data

The data for this study are drawn from the following sources:

1. *NSFG sampling frame*. The sampling frame includes U.S. Census data on area characteristics and commercially supplied data on a large proportion of households.
2. *Paradata*. The NSFG paradata include interviewer observations about sampled neighborhoods, housing units, and persons, and level-of-effort data (e.g., number of call attempts with different types of outcomes, number of trips, etc.).
3. *Interviewer timesheet reports*. These data include the number of hours worked each day.

In this section, we describe briefly the variables drawn from each source and how they are summarized to the interviewer-week level in order to be used in the models. A full description of each of the variables is given in Appendix 1 (Subsection 7.1). The goal of the study is to compare the ability of different modeling approaches to predict survey costs. Therefore, we intend to use predictors that are available at the time that those predictions need to be made. In this case, that is two weeks prior to the week for which the predictions are to be made. In week ten of an NSFG quarter, predictions for weeks 11 and 12 (the weeks defining the second phase) would be needed in order to make a decision about whether to change the design and use the second phase protocol.

For this study, the main cost driver is interviewer hours expended. The number of hours depends, in part, upon the phase, as the second phase involves subsampling cases. There are other costs associated with the change in phase. These are the special mailing and the additional tokens of appreciation for the screening interview and the main interview. We do not predict these costs using the models. The sample size for the mailing is known prior to the phase. The costs of the increased incentive, which is prepaid, are also known. The cost of the post-paid incentive is a function of the number of interviews. We use response propensity models (described elsewhere, see [West et al. 2019](#)) as the basis of estimating these costs. Our primary focus in this study is on prediction of the hours that interviewers will expend once the second phase begins.

The sampling frame data include geographic characteristics, such as the Census Division (the United States is divided by the U.S. Census Bureau into nine Divisions), as well as area characteristics, such as estimated eligibility rates for the Census Tract (a geography defined by the U.S. Census Bureau containing roughly 2,500 to 8,000 persons) of the sampled unit from the American Community Survey (ACS) and estimated rates for the U.S. Census Block Group (a geography defined by the U.S. Census Bureau usually corresponding to between 600 and 3,000 persons) of ever being married from the ACS. The urbanicity of the sampled area is assigned at the county level using the U.S. Office of Management and Budget's (OMB) definition of Metropolitan Statistical Areas. Also available on the sampling frame are data from a commercial vendor, such as the ages of adults in the household. These data are linked to sample addresses in advance. There are a proportion of households for which these data are not available, and they may also be inaccurate (see [West et al. 2015](#) for an appraisal). The variables include the estimated age of one or two persons in the household, the estimated household income, and the quality of the match (i.e., likelihood of the match being correct).

These sampling frame data are attached to NSFG call record data. These predictors are summarized up to the interviewer-week level, either by taking the mean (for continuous variables, e.g., rates of ever being married at the U.S. Census Block Group level) or the mode (for categorical variables, e.g., Census Region – 50 States and the District of Columbia are grouped into four Census Regions, and nine Divisions nested with Region) of these characteristics for all the contact attempts each interviewer made in a given week. Since these contact attempt data are not available for “future” weeks, we lagged these variables by two weeks in order to make them available for prediction of the costs two weeks into the future. For example, we use the modal urbanicity of cases that were attempted two weeks prior as a predictor of interviewer hours in the current week. This variable could be observed in week ten in order to predict hours for weeks 11 and 12.

Paradata used in this study include interviewer observations of neighborhood characteristics, such as whether there are unimproved roads or seasonal hazards that make access to the neighborhood difficult and whether there is evidence of speakers of languages other than English. Interviewers also observe characteristics of housing units, such as whether it is a single-family home or a multi-unit structure, whether there is evidence of children in the household, and the likelihood that all persons in the household are over the age of 45. There are also level-of-effort paradata variables that are derived from records of call attempts and information about the number of active lines.

The paradata from any given week are highly predictive of the hours of effort in the same week ([Wagner 2019](#)). However, in this case, we are making predictions for the future and only have the values for these variables from previous time periods. Therefore, we included a series of variables that summarize counts of call attempts with different result types (e.g., main interviews, screening interviews, setting appointments) for the time period two weeks prior to the weeks for which predictions are being made. Since many of the predictors are lagged values, the first two weeks of each quarter were excluded from the analysis. We also have both the number of area segments visited and the number of active sample housing units in each interviewer’s workload two weeks prior to the week for which predictions are being made. These variables are available at the point in time when cost predictions are going to be made. We also include an indicator for whether the week is in the second phase of the NSFG responsive design. The phase variable could be used to predict costs under two different design options – the phase one design versus the phase two design.

The cost data are derived from interviewer timesheets. There are three predictors derived from the timesheet data. The first is the number of hours worked in the week that was two weeks prior to the week for which predictions are being made. For example, the hours worked in week eight will be used to predict the number of hours worked in week ten. A second predictor is a categorical variable indicating whether the interviewer was involved with overnight travel two weeks ago. This measure is reported as the number of hours each week that are spent in overnight travel, but is categorized here as none, some, or all of the hours in the week two weeks prior to the current week. A third variable, days worked, is a count of the number of days in a week (two weeks prior to the current week) that the interviewer reported any time on their timesheet.

Our primary dependent variable is the hours worked by the interviewer each week, which is set by the interviewer in consultation with their manager. NSFG interviewers

agree to work a minimum of either 20 or 30 hours per week. However, interviewers set their own schedules and may deviate from the minimum specified for various reasons – either due to requests from sampled persons for specific appointment times, low contact rates on a particular day (Wagner and Olson 2018), or for personal reasons. Therefore, the hours worked for each interviewer is a random variable.

Each record in the dataset represents a week of an interviewer. In other words, an interviewer who works all weeks in a 12-week quarter will have ten records in the dataset (excluding weeks one and two due to the use of lagged values). Each interviewer is scheduled to work for at least one year (four quarters). This doesn't always happen (interviewers sometimes resign mid-quarter) and some interviewers worked in multiple years. In the data, the mean number of records (weeks) per interviewer was 47.6. The median was 27. The minimum and maximum were one and 253, respectively. Each record includes the number of hours worked that week (the outcome), a set of known characteristics (the week, the phase, the quarter, and the year), hours worked two weeks ago, whether the interviewer was travelling overnight two weeks prior, summary information about call attempts made two weeks prior, and characteristics of the interviewer's sample two weeks prior. Some records were excluded from the analysis. In some weeks, an interviewer might have reported hours in their timesheet, but did not make any call attempts. On rare occasions, call attempts were recorded in a week for which no hours were reported. In both of these situations, these records were excluded from the analysis.

4. Methods

4.1. Alternative Modeling Approaches

We will examine three different modeling approaches to predict the interviewer hours in a given week: a linear mixed (multilevel) model (MLM) with random intercepts for each interviewer, a Bayesian Additive Regression Trees (BART) model (Chipman et al. 2010) that includes an indicator variable for each interviewer, and a random intercept BART (RI BART) model (Tan et al. 2018). We also attempted to fit a linear regression model that did not include an indicator for each interviewer. This model produced poor predictions and, therefore, was not included.

From previous research into survey errors, we know that interviewers vary in their ability to recruit participants (see West and Blom 2017, for a review). Hence, we use multilevel models (MLM) in which each interviewer has a random intercept:

$$y_i = \sum_{p=1}^P x_{ip} \beta_p + \alpha_{g[i]} + \varepsilon_i$$

where g indexes the interviewer who may have repeated (i.e., clustered) measurements and p indexes the set of covariates, including a vector of 1's. The "random intercepts" associated with each interviewer are assumed to be draws from a normal distribution with a common variance. This allows for random variation among interviewers to be incorporated into the model. These models assume a linear relationship between the predictors and the outcome. In this case, that seems reasonable as the coefficients represent

estimated changes in hours when characteristics of the sample are changed. This method has been used previously to estimate costs in surveys (Wagner et al. 2017).

Our second method uses Bayesian Additive Regression Trees with an indicator variable for each interviewer. The BART approach uses the sum of a large number of regression trees, where each tree is constrained by specification of priors to include relatively few predictors. The model has a very general form:

$$y_i = \sum_{j=1}^m f(X_i; T_j, M_j) + \varepsilon_i$$

where $f(\cdot)$ defines the BART model, and T denotes a sequence of decision rules that split the sample into groups with each terminal node having a mean μ_i from the set M . The process can be repeated over m trees (indexed by j), with each tree reducing $\sum_{i=1}^n \varepsilon_i$ from the previous trees. Posterior distributions on parameters are developed using a Markov Chain Monte Carlo (MCMC) algorithmic approach. BART models have the ability to include or exclude a large number of interactions, including polynomials of continuous predictors, as the data will suggest. BART model predictions have been shown to perform well against other machine learning techniques (Chipman et al. 2010), and they allow for the calculation of a principled measure of uncertainty associated with the predictions. We would expect the BART model to produce more accurate predictions than MLM (Chipman et al. 2010). However, BART models do not produce easily interpretable model estimates in the same way that regression models produce estimated coefficients. Instead, BART models produce predicted outcomes.

We estimated BART models using one with interviewers as fixed effects and another incorporating interviewer effects as random intercepts (RI BART; Tan et al. 2018). The latter model can be specified as:

$$y_i = \sum_{j=1}^m f(X_i; T_j, M_j) + \alpha_{g[i]} + \varepsilon_i$$

where we use the same notation as earlier, with the addition of the random intercept, $\alpha_{g[i]}$, and g indexes the interviewer, who may work multiple weeks and therefore produces a clustered set of observations. Both the MLM and RI BART models allow for some consistency within each interviewer in their expected hours each week; the random effect approach formally treats the interviewers as being drawn from a population of potential interviewers, and can stabilize estimates of these “interviewer effects” compared with treating them as fixed effects.

We made predictions of hours worked during phase two for each of six quarters, where we trained the models using data from previous quarters and the first phase of the current quarter to predict the hours expended during the second phase of the current quarter. For example, we predicted the hours worked for phase two of Q22 using the hours from Q1 through Q22 phase one. We made predictions for phase two hours for Q22 through Q27. This allows us to assess the performance of our models using a form of temporal cross-validation. Given that we know the actual hours worked in these six quarters, we evaluate the predictions using two measures of prediction accuracy: mean squared error (MSE) and mean absolute error (MAE). We compare the observed and predicted values for our

dependent variable (interviewer hours expended) in the two weeks of phase two across six quarters of data collection.

All models included the predictors listed in Appendix 1. The multilevel models included a random intercept; the BART models included a dummy variable for each interviewer to accommodate interviewer effects if they were present; and the RI BART models included a random intercept for each interviewer.

All models were fit in R. The multilevel model was fit using the `lmer` function in the `lmer4` package. The BART model was fit using the `bart` function in the `dbarts` package (Dorie et al. 2019). In these BART models, the interviewer ID was included as a factor, so that an indicator for each interviewer could be included in the models. The RI BART model included a random intercept for each interviewer. These models were fit using the `rbart_vi` function in the `dbarts` package.

The BART models require the specification of priors on several parameters. In many cases, the default settings will work quite well (Chipman et al. 2010). In our case, in order to determine the values for priors, we performed cross-validation. The actual analyses were run on quarters 22 to 27. Therefore, in order to set priors, we tested a range of priors for quarters 16 to 21 and used training and test samples to determine the MSE and MAE for each combination of priors tested. Two important priors are the number of trees and k (the number of standard deviations $E(Y|x) = f(x)$ is away from ± 0.5). We tested a grid of possible values for trees of 45, 100, 150, 200, 250, 300, 350, and 400 crossed with the possible values for k of 1, 2, 3, and 4. We tested this grid of combinations of values across both the BART models. We calculated the MSE on the test sample for each of the six quarters (15 to 21) and then ranked the parameter pairs with lowest to highest MSE. In the case of the BART models, we found that 300 trees and $k=1$ performed the best with a mean rank of 4.0 across the six quarters. For the RI BART models, we found that 300 trees and $k=2$ performed the best with a mean rank of 4.0.

Once we had selected these two parameters for each of the two BART models, we also tested values for the priors for the variance parameters in a similar manner. The two parameters are the degrees of freedom for the error variance and the quantile of the error variance. We chose as the prior for the degrees of freedom for the variance as 3.1 and the quantile of the error variance prior as 0.96. These values are close to the suggested default values of 3.0 and 0.90.

Finally, for the BART models, we needed to set the MCMC parameters. We identified parameters that worked well for all quarters, and then ran the models using those parameters. We used trace plots of key parameters and a review of plots of autocorrelation functions to monitor the MCMC. In the case of the BART models, that meant 500 iterations for a burn-in, thinning to one in every 3,200 iterations, and running until 1,000 draws were obtained. For the RI BART models, we had 500 burn-in iterations, thinning to one in every 350 iterations, and running until we had 1,142 draws. We ran four chains for a total of 4,568 draws.

5. Results

First, we examine which predictors were important in each model. In the multilevel regression model, the interviewer IDs were important predictors. The proportion of the

total variance in costs that is due to the interviewers was between 0.21 and 0.25 across the six quarters we considered. These intra-class correlations are quite substantial relative to those for other survey outcomes (West and Blom 2017), suggesting that NSFG interviewers vary substantially in terms of these weekly hours during the second phase. Full model results for one quarter (Q27), as an example, are available in Appendix 2 (Subsection 7.2).

The BART models do not create scoring functions in the same way that a regression model does. Instead of presenting coefficients, we display the 20 predictors used most frequently in the BART modeling process. For the BART model, Figure 1 shows the 20 predictors that were used most frequently in splits in the Q27 model predicting the hours worked during an interviewer week. This is calculated as the number of splits based on that variable, divided by the total number of splits used. This quantity is averaged across all of the draws. Each predictor is described in more detail in Appendix 1 (Subsection 7.1).

The predictor most frequently included in the models is the number of hours worked in the week two weeks prior to the week of interest. The next most frequently used variable is an indicator variable for interviewer #78. In fact, five of the top ten variables are indicator variables for interviewers. Similar to the results from the multilevel model, this suggests that interviewers are relatively consistent in the hours that they charge each week. The number of active sample lines two weeks prior is the third most frequently used predictor for splits. An indicator variable for whether the interviewer was on full time travel two weeks prior to the week of interest is also important, as is an indicator for phase two. Figure 2 presents a similar figure for the RI BART model for Q27.

Fig. 1. The 20 predictors with the highest proportion of splits based on the variable in the ensemble of trees in the BART model for Q27 costs.

Fig. 2. The 20 predictors with the highest proportion of splits based on the variable in the ensemble of trees in the RI BART model for Q27 costs.

The number of hours worked two weeks prior to the current week is the most frequently included variable. Whether the interviewer was on full time travel or, alternatively, did not travel at all two weeks prior to the target week were also important predictors. An indicator for phase two was once again an important predictor. The number of screening interviews completed two weeks prior is also an important predictor. Other important predictors are indicators for several of the Census Divisions or Regions, sampling domain 1, several years (2012, 2015, and 2017), an indicator variable for quarter 12 (Q12), and the number of area segments visited and the number of appointments made two weeks prior to the current week.

Next, we examine the predictions of overall costs. Figure 3 presents the predicted total interviewing hours in phase two for each of the six “testing” quarters from each of the two modeling approaches. For each quarter, the predictions are based upon the data observed prior to each quarter’s phase two. The predictions of the total also have error bars. For the multilevel models, these are 95% bootstrap confidence intervals. These are 95% credible intervals for the BART models.

The models produce predictions of the total phase two hours with similar quality. All models do well in Q22, Q25, Q26, and Q27. Notably, the BART and RI BART models produce consistently narrower credible intervals. However, in Q23 and Q24, none of the models perform particularly well, and the intervals do not include the total hours that were actually observed. However, the intervals surrounding the predictions of the RI BART models come closer to including the observed hours than do those for the BART or MLM models. In all of the quarters, the RI BART model provides the predicted total closest to the observed total.

Fig. 3. Predicted total hours and 95% bootstrap or credible intervals from the MLM, BART, and RI BART models for each of six quarters.

Although the models predicted the total phase two costs fairly well, the differences between interviewers in pay rates mean that predictions at the interviewer level may be important in obtaining an accurate prediction of the overall cost. In [Figure 4](#), we examine the predictions for each interviewer-week in phase two of Q27. Each dot in [Figure 4](#)

Fig. 4. Q27 hours: Predicted values from three different models versus observed values.

Fig. 5. Mean squared error (MSE) and mean absolute error (MAE) for three models (multi-level model, BART, and RI BART) across six quarters (Q22–Q27).

represents one week worked by an interviewer. The observed hours are on the x-axis and the predicted hours are on the y-axis. The 45-degree line represents agreement between the predicted and observed values.

The RI BART model does appear to provide better predictions. To confirm this, we summarize the results at the interviewer-week level for all quarters using two error measures – the mean squared error and the mean absolute error. Figure 5 presents these measures for each of the six quarters.

The results in Figure 5 indicate that the RI BART model has the best performance for predicting interviewer-level costs. The RI BART model has the lowest MSE – and often substantially lower – than the other models in all six of the quarters. The BART model, on the other hand, produces predictions that have the second lowest MSE in three of the quarters (Q22, Q23, and Q26). The MLM models have the second lowest MSE in the other three quarters (Q24, Q25, and Q27).

6. Discussion

The accurate prediction of expected costs associated with a design change in a responsive survey design framework is necessary in order to make cost-error tradeoff decisions about potential design changes across phases of a survey. In our case, we are comparing two designs (phase one versus phase two). The predictions of costs – along with predictions about expected errors under different designs – could be used to optimize the design. In this case, it would be possible to use these predictions to determine when to switch from phase one to phase two. As a first step in this direction, we evaluate our ability to predict costs.

In some settings, the prediction of costs will be tightly tied to the completion rate. For example, in a web survey, the costs are largely driven by the payment of incentives and, to a lesser extent, the cost of sending email reminders. In that case, the predictions of costs are largely a function of expected completion rates under different designs. It would be

possible to accurately predict costs based on different incentive amounts if an accurate prediction of the completion rate at each incentive amount is available.

In our setting, prediction of costs is more complicated. Interviewers make frequent scheduling changes. Often, this is done in response to requests from sampled units. Interviewers seek to accommodate the schedules of sampled persons. For example, interviewers will set appointments with sampled persons at times that are convenient for the sampled person. This might mean making a change to their planned schedule for the week. Interviewers may also shorten their scheduled hours when they experience relatively low contact rates ([Wagner and Olson 2018](#)). These accommodations to the interviewers' schedules are necessary given the need to be flexible with the schedules of sampled persons. Further, the level of effort needed to obtain an interview can vary greatly from interviewer to interviewer and week to week within an interviewer. This can be a function of the choices the interviewer makes about when to work, the characteristics of the sample, and other factors that appear as noise in models of response propensity. These factors make it difficult to predict the number of hours that an interviewer will work in any given week. There are, on the other hand, factors that stabilize interviewers' effort. Mainly, project managers seek stable effort over time. This might include a certain number of hours as a goal to be worked each week. The hours that interviewers work each week may also vary when design changes are introduced over time. Our analysis focused on predicting hours as a function of stable design features.

We found that we could develop relatively accurate predictions with relatively simple models. The multilevel and BART models generated accurate predictions of the total hours worked in a given week. However, differences in wages paid across interviewers mean that these estimates of total hours might give very different cost estimates. The model that produced the most accurate results was the random intercept BART ("RI BART") model. The BART model that treated interviewers as a fixed effect also performed well. In initial analyses, we found that using a BART model (not with random intercepts) with many of the default settings performed about as well as the multilevel models. In our setting, the ability to tune the parameters proved valuable. We were able to do this since we had 27 iterations of the same survey. In other settings, the inability to test a variety of parameters might reduce the effectiveness of the BART approach, although [Chipman et al. \(2010\)](#) found that the default settings perform well in several different situations.

This article has focused exclusively on the ability of different modeling approaches to improve the prediction of survey costs. These predictions are intended to inform decisions about interventions in a responsive survey design (RSD) context. Although it is clear that measures of cost are always relevant for making design decisions, the exact way in which these predictions would be used to make decisions during data collection was outside the scope of this article. Here, we briefly outline an approach to including these predictions in a decision framework for RSD.

In the specific case presented in this article, the decision is whether to switch the data collection protocol from phase one to phase two. In order to optimize that decision, we need to have estimates of the impact of phase two on nonresponse bias and the costs of the second phase. A further consideration would be sampling error. A simple form of the optimization problem would look at the design decision each week and compare outcomes for two design options under the same fixed budget: 1) continue with phase one, or 2) switch to phase two.

The outcome to be compared between the two designs would be the expected mean squared error of the survey estimates (incorporating both expected nonresponse bias and sampling error). Phase one could produce more interviews but risk higher nonresponse bias, while phase two could lower the risk of nonresponse bias but produce fewer interviews.

The focus of this article has been on accurate predictions of the costs involved in this decision. Other papers have focused on relative nonresponse errors associated with similar types of design changes (Peytchev et al. 2010; Axinn et al. 2011). With predictions of costs under the two approaches – which could be generated using the approach outlined in this article – and predictions of which cases are likely to be interviewed and their predicted responses under the two design options, the option that minimizes mean squared error for a fixed budget could be selected at a given point in time. The approach outlined here also allows for uncertainty in the cost estimation to be built into the decision-making process, for example, by making decisions based on a low or high percentile of an estimate, rather than just a mean or to make probability statements about predicted outcomes.

We also note that predictions have variance and accounting for this variance may be important. In our case, this variance was captured through repeated draws from the posterior distribution for the BART models. We used a bootstrap approach to assessing variance of the MLM predictions. We know that incorrect assumptions about underlying cost parameters can lead to inefficient designs (Burger et al. 2017). Capturing the variance may be a helpful input to design decisions. Designers, rather than focusing on point estimates, might use the variance to calculate the probability of achieving survey goals under alternative designs. Such an approach would lead to better decisions. The models developed in this paper and the predictions and prediction intervals they produce provide one “leg” of such a decision analysis.

Future research could look at combining predictions about costs, response propensities, and survey outcome variables under different designs. Then, design decisions can be informed by the predicted errors and costs under the alternative designs. This framework would allow us to move closer to a “total survey error” approach in practice. Of course, the underlying models need to be evaluated. This article is a step toward the development of models aimed at predicting costs. Many other models will need to be developed and tested for settings unlike the one used in this paper.

7. Appendix

7.1. Appendix 1. Available predictors used in all three types of models.

Source	Predictor	Description
TIME-SHEETS	NEWIWERID	Interviewer ID Number
	NHOURS_LAG2	Number of hours worked by the interviewer in the week two weeks prior to the current week
	TRAVEL_LAG2	How much did the interviewer participate in overnight travel in the week two weeks prior to the current week: NONE, SOME, or ALL of the week. Generally, interviewing staff is split into those who travel and

7.1. Appendix 1. Continued.

Source	Predictor	Description
	DAYS_ WORKED_ LAG2	those who do not. However, sometimes, under special circumstances such as a need to infuse more hours into production due to an unplanned staff shortage or if a PSU in need of hours happens to be geographically close to another PSU, non-travelling interviewers will travel. Therefore, this variable has three categories. The number of days worked (i.e., days with an entry in the timesheet) in the week two weeks prior to the current week
SAMPLING FRAME	QTR	The quarter of production (Q1–Q27)
	YEAR	The calendar year of production (2011–2018)
	CENSUS_DIV_ MODE_LAG2	The modal Census Division of the lines attempted by an interviewer in the week two weeks prior to the current week.
	CENS_REG_ MODE_ LAG2	The modal Census Region of the lines attempted by an interviewer in the week two weeks prior to the current week.
	EST_ELIG_ RATE_ MEAN_LAG2	This is the mean of the Census ZIP Code Tabulation Area (ZCTA) level data about the estimated eligibility rate. The data are at the ZCTA level, but the value here is the average over all contact attempts for the week that is two weeks prior to the current week.
	EST_ELIG_ 15_49_ACS_ MEAN_LAG2	The mean of the estimated eligibility rate for the Census Block Group reported in the American Community Survey. The data are at the Block Group level, but the value here is the average over all contact attempts for the week that is two weeks prior to the current week.
	ELIG_NEVER_ PCT_ MEAN_ LAG2	This is the percentage of eligible persons living in the Census Tract who have never been married. The data are at the Tract level, but the value here is the average over all contact attempts for the week that is two weeks prior to the current week.
	OCC_RATE_ MEAN_ LAG2	This is the Census Block level occupancy rate from the 2010 Decennial Census. The data are at the Block level, but the value here is the average over all contact attempts for the week that is two weeks prior to the current week.
	DOMAIN_ MODE_ LAG2	The domain is set at the Census Block Group level and assigned to housing units within each BG. All BGs are assigned to a domain based upon the following definitions: 1) <10% of Block Group African-American and <10% Hispanic, 2) >= 10% of Block Group African-American and <10% Hispanic, 3) <10% of Block Group African-American and >= 10% Hispanic, and 4) >= 10% of

7.1. Appendix 1. Continued.

Source	Predictor	Description
INTER- VIEWER OBSER- VATIONS	URBAN_MODE_LAG2	Block Group African-American and $\geq 10\%$ Hispanic. The mode is for the domain of the lines that are attempted in the week two weeks prior to the current week. The mode of the urbanicity (assigned at the case level) of the attempts made during the week that is two weeks prior to the current week, where 1 = Major Metropolitan Area, 2 = Minor Metropolitan Area, 3 = Non-Metropolitan Area, 4 = Remote Area.
	STRUCTURE_TYPE_MODE_LAG2	The mode of the structure type variable of the cases that were attempted in the week that is two weeks prior to the current week. 1 = Single family home, 2 = Structure with 2 to 9 units, 3 = Structure with 10 + units, 4 = Mobile home, 5 = Other.
	BLACCESS_GATED_MEAN_LAG2	The mean of an area segment-level observation about whether there is a gated community in the area segment. This is observed at the segment level, but the value here is average over all contact attempts for the week that is two weeks prior to the current week.
	BLACCESS_SEASONAL_HAZARD_MEAN_LAG2	The mean of an area segment-level observation about whether there is a potential seasonal hazard preventing access to the area segment (e.g., unplowed roads). This is observed at the segment level, but the value here is average over all contact attempts for the week that is two weeks prior to the current week.
	BLACCESS_UNIMPROVED_ROADS_MEAN_LAG2	The mean of an area segment-level observation about whether there are unimproved roads limiting access to the area segment. This is observed at the segment level, but the value here is average over all contact attempts for the week that is two weeks prior to the current week.
	BLACCESS_OTHER_MEAN	The mean of an area segment-level observation about whether there other (i.e., not gated, seasonal hazards, or unimproved roads) factors limiting access to the area segment. This is observed at the segment level, but the value here is average over all contact attempts for the week that is two weeks prior to the current week.
	LRESIDENTIAL_MEAN	The mean of an area segment-level observation about whether the area is completely residential or also includes some commercial structures. This is observed at the segment level, but the value here is average over all contact attempts for the week that is two weeks prior to the current week.

7.1. Appendix 1. Continued.

Source	Predictor	Description
	INON_ENGLISH_ SPEAKERS_ MEAN_LAG2	The mean of an area segment-level observation about whether the area has evidence of non-English speakers. This is observed at the segment level, but the value here is average over all contact attempts for the week that is two weeks prior to the current week.
	BLNON_ ENGLISH_ LANG_SPANIS_ MEAN_LAG2	The mean of an area segment-level observation about whether the area has evidence of Spanish speakers. This is observed at the segment level, but the value here is average over all contact attempts for the week that is two weeks prior to the current week.
	ISAFETY_ CONCERNS_ MEAN_LAG2	The mean of an area segment-level observation about whether the interviewer had concerns about their safety on the first visit. This is observed at the segment level, but the value here is average over all contact attempts for the week that is two weeks prior to the current week.
	MANYUNITS_ MEAN_LAG2	The mean of an observation at the housing unit level indicating whether the sampled housing unit has 1 = more than one unit, or 0 = 1 unit. This is observed at the housing unit level, but the value here is average over all contact attempts for the week that is two weeks prior to the current week.
	CHILDREN UNDER15_ MEAN_LAG2	The mean of an observation at the housing unit level indicating whether the interviewer believes that there are children under the age of 15 living in the housing unit (1 = Yes, 0 = No). This is observed at the housing unit level, but the value here is average over all contact attempts for the week that is two weeks prior to the current week.
	ALLAGE- OVER45_ MEAN_LAG2	The mean of an observation at the housing unit level indicating whether the interviewer believes that persons living in the housing unit are all over the age of 45 (1 = Yes, 0 = No). This is observed at the housing unit level, but the value here is average over all contact attempts for the week that is two weeks prior to the current week.
COMMER- CIAL DATA	MSG_ MATCHQUAL- ITY_ MEAN_LAG2	A variable indicating the estimated quality of the match of commercially-available data to the address (1–5). The data are at the case level, but the value here is the average over all contact attempts for the week that is two weeks prior to the current week.
	MSG_AGE_ MEAN_LAG2	The mean age of the first person from the commercially-available data where those data are available. The data are at the case level, but the value here is the average over all contact attempts for the week that is two weeks prior to the current week.

7.1. Appendix 1. Continued.

Source	Predictor	Description
	MSG_INCOME_ MEAN_LAG2	The mean of the estimated household income for cases with a match to commercially-available data. The data are at the case level, but the value here is the average over all contact attempts for the week that is two weeks prior to the current week.
LEVEL OF EFFORT PARA- DATA	PHASE	The phase of the NSFG design (first phase occurs in weeks 1–10, phase two during weeks 11–12).
	LAG2.ACTIVE_ LINES	The number of active lines from two weeks prior to the current week for each interviewer.
	TRIPS_LAG2	The total number of unique visits to an area segment (derived from call record data) from two weeks prior to the current week for each interviewer.
	FTFNOCON- TACT_LAG2	The total number of Face-to-face contact attempts that resulted in no contact from two weeks prior to the current week for each interviewer.
	FTFCONTACT_ LAG2	The total number of Face-to-face contact attempts that resulted in a contact with only agreement for a general callback from two weeks prior to the current week for each interviewer.
	FTFAPPT_ LAG2	The total number of Face-to-face contact attempts that resulted in setting an appointment from two weeks prior to the current week for each interviewer.
	MAINIW_LAG2	The total number of main interviews (all main interviews are completed face-to-face) from two weeks prior to the current week for each interviewer.
	FTFMAINCON- CERN_LAG2	The total number of Face-to-face contact attempts that resulted in the sampled person expressing concerns from two weeks prior to the current week for each interviewer.
	FTFMAINNI_ LAG2	The total number of Face-to-face contact attempts that resulted in a final noninterview from two weeks prior to the current week for each interviewer.
	FTFMAINNS_ LAG2	The total number of Face-to-face contact attempts that resulted in a final nonsample from two weeks prior to the current week for each interviewer.
	FTFSCRNIW_ LAG2	The total number of Face-to-face contact attempts that resulted in a screening interview from two weeks prior to the current week for each interviewer.
	FTFSCRNCON- CERN_LAG2	The total number of Face-to-face contact attempts that resulted in the sampled housing unit expressing concerns prior to completing a screening interview from two weeks prior to the current week for each interviewer.

7.1. Appendix 1. Continued.

Source	Predictor	Description
	FTFSCRNNI_LAG2	The total number of Face-to-face contact attempts that resulted in the sampled housing unit being finalized as a noninterview prior to completing a screening interview from two weeks prior to the current week for each interviewer.
	FTFSCRNNNS_LAG2	The total number of Face-to-face contact attempts that resulted in the sampled housing unit being finalized as nonsample prior to completing a screening interview from two weeks prior to the current week for each interviewer.
	FTF_MAINNS_INEL_LAG2	The total number of Face-to-face contact attempts that resulted in the sampled person being finalized as ineligible prior to completing a screening interview from two weeks prior to the current week for each interviewer.
	ACTIVE_LINES_LAG2	The number of active sampled units two weeks prior to the current week for each interviewer.
	TEL_ALL_LAG2	The total number of telephone attempts made by each interviewer two weeks prior to the current week.

7.2. Appendix 2. Q27 multilevel model predicting hours worked by an interviewer in a week: Estimated coefficients, confidence interval, and p-value.

Predictors	Estimates	CI	p
(Intercept)	32.73	24.32 – 41.13	< 0.001
Q10	-3.32	-5.02 – -1.62	<0.001
Q11	-1.60	-3.31 – 0.12	0.068
Q12	-0.23	-1.95 – 1.50	0.798
Q13	-2.71	-4.39 – -1.03	0.002
Q14	-1.83	-3.52 – -0.15	0.033
Q15	-1.26	-2.95 – 0.44	0.146
Q16	-1.40	-3.11 – 0.31	0.109
Q17	-1.29	-2.98 – 0.41	0.137
Q18	-1.05	-2.77 – 0.68	0.235
Q19	-2.34	-4.08 – -0.60	0.008
Q2	-3.37	-4.95 – -1.79	<0.001
Q20	-2.42	-4.16 – -0.67	0.007
Q21	-2.85	-4.68 – -1.02	0.002
Q22	-2.04	-3.82 – -0.25	0.025
Q23	-2.76	-4.56 – -0.95	0.003
Q24	-0.65	-2.47 – 1.17	0.483

7.2. Appendix 2. Continued.

Predictors	Estimates	CI	p
Q25	-2.30	-4.01 – -0.59	0.008
Q26	-2.74	-4.54 – -0.93	0.003
Q27	-3.35	-5.18 – -1.52	<0.001
Q3	-3.25	-4.86 – -1.64	<0.001
Q4	-0.95	-2.58 – 0.69	0.257
Q5	-2.46	-4.15 – -0.76	0.004
Q6	-2.45	-4.15 – -0.76	0.005
Q7	-3.33	-4.98 – -1.68	<0.001
Q8	-2.60	-4.27 – -0.92	0.002
Q9	-2.92	-4.59 – -1.25	0.001
PHASE_MODE2	-1.44	-2.05 – -0.83	<0.001
NHOURS_LAG2	0.13	0.09 – 0.17	< 0.001
DAYS_WORKED_LAG2	-0.10	-0.33 – 0.14	0.414
NONE	-0.84	-1.66 – -0.01	0.046
SOME	-1.31	-2.50 – -0.11	0.033
CENSUS_DIV_MODE_LAG22	0.74	-1.12 – 2.61	0.435
CENSUS_DIV_MODE_LAG23	0.06	-1.66 – 1.79	0.942
CENSUS_DIV_MODE_LAG24	-1.87	-3.92 – 0.18	0.074
CENSUS_DIV_MODE_LAG25	0.93	-1.17 – 3.03	0.384
CENSUS_DIV_MODE_LAG26	1.74	-0.66 – 4.15	0.155
CENSUS_DIV_MODE_LAG27	1.67	-0.78 – 4.12	0.181
CENSUS_DIV_MODE_LAG28	-0.52	-3.11 – 2.07	0.694
CENSUS_DIV_MODE_LAG29	0.95	-1.48 – 3.38	0.444
DOMAIN2_MODE_LAG22	0.65	-0.07 – 1.38	0.078
DOMAIN2_MODE_LAG23	-0.13	-0.92 – 0.65	0.738
DOMAIN2_MODE_LAG24	0.35	-0.47 – 1.17	0.406
URBAN_MODE_LAG22	-0.31	-1.11 – 0.48	0.436
URBAN_MODE_LAG23	0.23	-1.21 – 1.66	0.758
URBAN_MODE_LAG24	5.73	-1.50 – 12.96	0.121
STRUCTURE_TYPE_MODE_LAG22	-0.21	-1.17 – 0.75	0.663
STRUCTURE_TYPE_MODE_LAG23	-0.17	-1.13 – 0.78	0.723
STRUCTURE_TYPE_MODE_LAG24	0.98	-0.71 – 2.67	0.256
STRUCTURE_TYPE_MODE_LAG25	-5.87	-25.47 – 13.73	0.557
BLACCESS_GATED_MEAN_LAG2	0.09	-0.54 – 0.71	0.782
BLACCESS_SEAS_HAZARD_MEAN_LAG2	0.59	-0.61 – 1.80	0.335
BLACCESS_UNIMP_ROADS_MEAN_LAG2	-0.69	-1.60 – 0.21	0.133
BLACCESS_OTHER_MEAN_LAG2	-0.21	-1.33 – 0.90	0.709
LRESIDENTIAL_MEAN_LAG2	0.00	-0.65 – 0.66	0.988
INON_ENGLISH_SPEAKERS_MEAN_LAG2	-0.28	-0.89 – 0.33	0.372
ISAFETY_CONCERNS_MEAN_LAG2	0.73	0.11 – 1.35	0.020
MANYUNITS_MEAN_LAG2	-0.20	-1.63 – 1.23	0.785
CHILDRENUNDER15_MEAN_LAG2	0.42	-0.96 – 1.80	0.550

7.2. Appendix 2. Continued.

Predictors	Estimates	CI	p
ALLAGEOVER45_MEAN_LAG2	-0.58	-1.75 – 0.60	0.335
EST_ELIG_RATE_MEAN_LAG2	0.22	-5.53 – 5.96	0.941
ELIG_NEVER_PCT_MEAN_LAG2	0.01	-0.02 – 0.03	0.585
OCC_RATE_MEAN_LAG2	1.34	-2.38 – 5.07	0.481
MSG_MATCHQUALITY_MEAN_LAG2	-0.85	-2.49 – 0.79	0.311
MSG_AGE_MEAN_LAG2	-0.10	-0.14 – -0.05	<0.001
MSG_INCOME_MEAN_LAG2	0.00	-0.00 – 0.00	0.919
EST_ELIG_15_49_ACS_MEAN_LAG2	-1.52	-5.16 – 2.12	0.413
TRIPS_LAG2	0.05	-0.02 – 0.12	0.191
FTFNOCONTACT_LAG2	-0.00	-0.01 – 0.01	0.368
FTFCONTACT_LAG2	-0.01	-0.06 – 0.05	0.820
FTFAPPT_LAG2	-0.17	-0.29 – -0.05	0.005
MAINIW_LAG2	-0.12	-0.24 – -0.00	0.050
FTFMAINCONCERN_LAG2	-0.01	-0.20 – 0.17	0.884
FTFMAINNI_LAG2	0.16	-0.38 – 0.70	0.561
FTFMAINNS_LAG2	2.11	-9.37 – 13.60	0.718
FTFSCRNIW_LAG2	-0.12	-0.16 – -0.07	<0.001
FTFSCRNCONCERN_LAG2	-0.02	-0.15 – 0.11	0.797
FTFSCRNNI_LAG2	0.33	-0.20 – 0.87	0.224
FTFSCRNNS_LAG2	0.02	-0.05 – 0.10	0.537
FTF_MAINNS_INEL_LAG2	-1.19	-3.04 – 0.66	0.208
ACTIVE_LINES_LAG2	0.03	0.03 – 0.04	<0.001
TEL_ALL_LAG2	-0.00	-0.02 – 0.01	0.498
Random Effects			
σ^2	98.58		
$\tau_{00newIwerID}^2$	32.31		
ICC	0.25		
N _{newIwerID}	187		
Observations	8843		
Marginal R ² / Conditional R ²	0.038 / 0.275		

In this table, the variable names are used. See Appendix 1 (Subsection 7.1) for a description of each variable.

8. References

Abu-Nimeh, S., D. Nappa, X. Wang, and S. Nair. 2008. . “Bayesian Additive Regression Trees-Based Spam Detection for Enhanced Email Privacy.” 2008 Third International Conference on Availability, Reliability and Security, Barcelona, Spain, 4–7 March 2008 IEEE. Available at: <https://ieeexplore.ieee.org/abstract/document/4529459> (accessed May 2020).

Axinn, W., C. Link, and R. Groves. 2011. “Responsive Survey Design, Demographic Data Collection, and Models of Demographic Behavior.” *Demography* 48(3): 1–23. DOI: <https://doi.org/10.1007/s13524-011-0044-1>.

- Barber, J.S., Y. Kusunoki, and H.H. Gatny. 2011. "Design and Implementation of an Online Weekly Survey to Study Unintended Pregnancies: Preliminary Results." *Vienna Yearbook of Population Research* 9: 327–334. DOI: <https://doi.org/10.1553/populationyearbook2011s327>.
- Biemer, P.P., de Leeuw, E.D., Eckman, S., Edwards, B., Kreuter, F., Lyberg, L., Tucker, C., and West, B.T. (Eds.). 2017. *Total Survey Error in Practice*. Hoboken, New Jersey: Wiley.
- Biemer, P.P., and D. Trewin. 1997. "A Review of Measurement Error Effects on the Analysis of Survey Data." In *Survey Measurement and Process Quality*, edited by L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin. (pp. 601–632). New York: Wiley.
- Burger, J., K. Perryck, and B. Schouten. 2017. "Robustness of Adaptive Survey Designs to Inaccuracy of Design Parameters." *Journal of Official Statistics* 33(3): 687–708. DOI: <https://doi.org/10.1515/jos-2017-0032>.
- Chipman, H.A., E.I. George, and R.E. McCulloch. 2010. "BART: Bayesian Additive Regression Trees." *The Annals of Applied Statistics* 4(1): 266–298. DOI: <https://doi.org/10.1214/09-AOAS285>.
- Dorie, V., H. Chipman, R. McCulloch, A. Dadgar, R.C. Team, G.U. Draheim, M. Bosmans, C. Tournayre, M. Petch, and R. de Lucena Valle. 2019. "dbarts: Discrete Bayesian Additive Regression Trees Sampler." Available at: [https://CRAN.R-project.org/package = dbarts](https://CRAN.R-project.org/package=dbarts) (accessed May 2020).
- Durrant, G.B., O. Maslovskaya, and W.F. Smith Peter. 2017. "Using Prior Wave Information and Paradata: Can They Help to Predict Response Outcomes and Call Sequence Length in a Longitudinal Study?" *Journal of Official Statistics* 33(3): 801–833. DOI: <https://doi.org/10.1515/jos-2017-0037>.
- Finamore, J., S. Coffey, and B. Reist. 2013. "National Survey of College Graduates: A Practice-Based Investigation of Adaptive Design." Annual AAPOR Conference, May 16–19, 2013. Boston, MA, U.S.A.
- Green, D.P., and H.L. Kern. 2012. "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees." *Public Opinion Quarterly* 76(3): 491–511. DOI: <https://doi.org/10.1093/poq/nfs036>.
- Groves, R.M. 2006. "Nonresponse Rates and Nonresponse Bias in Household Surveys." *Public Opinion Quarterly* 70(5): 646–675. DOI: <https://doi.org/10.1093/poq/nfl033>.
- Groves, R.M., and S.G. Heeringa. 2006. "Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 169(3): 439–457. DOI: <https://doi.org/10.1111/j.1467-985X.2006.00423.x>.
- Kern, C., T. Klausch, and F. Kreuter. 2019. "Tree-Based Machine Learning Methods for Survey Research." *Survey Research Methods* 13(1): 73–93. DOI: <https://doi.org/10.18148/srm/2019.v1i1.7395>.
- Kirgis, N., and J. Lepkowski. 2013. "Design and Management Strategies for Paradata-Driven Responsive Design: Illustrations from the 2006-2010 National Survey of Family Growth." In *Improving Surveys with Paradata: Analytic Uses of Process Information*, edited by F. Kreuter: 121–144. Hoboken, NJ: Wiley.

- Kleven, Ø., J. Fosen, B. Lagerstrøm, and L.-C. Zhang. 2010. . "The Use of R-Indicators in Responsive Survey Design—Some Norwegian Experiences." Q2010 Conference, Helsinki, 3–6 May 2010. Available at: <http://hummedia.manchester.ac.uk/institutes/cmist/risq/kleven-2010b.pdf> (accessed May 2020)
- Laflamme, F., and M. Karaganis. 2010. "Implementation of Responsive Collection Design for CATI Surveys at Statistics Canada." Proceedings of the European Conference on Quality in Official Statistics, Helsinki, Finland, Helsinki, Finland, 3–6 May, 2010. Available at: https://q2010.stat.fi/media/presentations/1_Responsive_design_paper_london_event1_revised.doc.
- Lewis, T. 2017. "Univariate Tests for Phase Capacity: Tools for Identifying When to Modify a Survey's Data Collection Protocol." *Journal of Official Statistics* 33(3): 601–624. DOI: <https://doi.org/10.1515/jos-2017-0029>.
- Luiten, A., and B. Schouten. 2013. "Tailored Fieldwork Design to Increase Representative Household Survey Response: An Experiment in the Survey of Consumer Satisfaction." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 176(1): 169–189. DOI: <https://doi.org/10.1111/j.1467-985X.2012.01080.x>.
- Lundquist, P., and C.-E. Särndal. 2013. "Aspects of Responsive Design with Applications to the Swedish Living Conditions Survey." *Journal of Official Statistics* 29(4): 557–582. DOI: <https://doi.org/10.2478/jos-2013-0040>.
- Lynn, p. 2016. "Targeted Appeals for Participation in Letters to Panel Survey Members." *Public Opinion Quarterly* 80(3): 771–782. DOI: <https://doi.org/10.1093/poq/nfw024>.
- Mohl, C., and F. Laflamme. 2007. "Research and Responsive Design Options for Survey Data Collection at Statistics Canada." Joint Statistical Meetings, Salt Lake City, UT, 29 July–2 August, 2007. Available at: <http://www.asasrms.org/Proceedings/y2007/Files/JSM2007-000421.pdf> (accessed May 2020).
- Paiva, T., and J.P. Reiter. 2017. "Stop or Continue Data Collection: A Nonignorable Missing Data Approach for Continuous Variables." *Journal of Official Statistics* 33(3): 579–599. DOI: <https://doi.org/10.1515/jos-2017-0028>.
- Peytchev, A., R.K. Baxter, and L.R. Carley-Baxter. 2009. "Not All Survey Effort Is Equal: Reduction of Nonresponse Bias and Nonresponse Error." *Public Opinion Quarterly* 73(4): 785–806. DOI: <https://doi.org/10.1093/poq/nfp037>.
- Peytchev, A., E. Peytcheva, and R.M. Groves. 2010. "Measurement Error, Unit Nonresponse, and Self-Reports of Abortion Experiences." *Public Opinion Quarterly* 74(2): 319–327. DOI: <https://doi.org/10.1093/poq/nfq002>.
- Plewis, I., and N. Shlomo. 2017. "Using Response Propensity Models to Improve the Quality of Response Data in Longitudinal Studies." *Journal of Official Statistics* 33(3): 753–779. DOI: <https://doi.org/10.1515/jos-2017-0035>.
- Rao, R.S., M.E. Glickman, and R.J. Glynn. 2008. "Stopping Rules for Surveys with Multiple Waves of Nonrespondent Follow-Up." *Statistics in Medicine* 27(12): 2196–2213. DOI: <https://doi.org/10.1002/sim.3063>.
- Rosen, J.A., J. Murphy, A. Peytchev, T. Holder, J. Dever, D. Herget, and D. Pratt. 2014. "Prioritizing Low Propensity Sample Members in a Survey: Implications for Nonresponse Bias." *Survey Practice* 7(1). DOI: <https://doi.org/10.1.1.686.6795>.

- Schonlau, M., and M.P. Couper. 2016. "Semi-Automated Categorization of Open-Ended Questions." *Survey Research Methods* 10(2): 143–152. DOI: <https://doi.org/10.18148/srm/2016.v10i2.6213>.
- Sparapani, R.A., B.R. Logan, R.E. McCulloch, and P.W. Laud. 2016. "Nonparametric Survival Analysis Using Bayesian Additive Regression Trees (BART)." *Statistics in Medicine* 35(16): 2741–2753. <https://doi.org/> DOI: 10.1002/sim.6893.
- Tabuchi, T., F. Laflamme, O. Phillips, M. Karaganis, and A. Villeneuve. 2009. "Responsive Design for the Survey of Labour and Income Dynamics." Statistics Canada Symposium. October 27–30, 2009. Gatineau, Québec, Canada. Available at: <http://oaresource.library.carleton.ca/wcl/2016/20160811/CS11-522-2009-eng.pdf#page=149>.
- Tan, Y.V., C.A. Flannagan, and M.R. Elliott. 2018. "Predicting Human-Driving Behavior to Help Driverless Vehicles Drive: Random Intercept Bayesian Additive Regression Trees." *Statistics and Its Interface* 11(4): 557–572. DOI: <https://doi.org/10.4310-SII.2018.v11.n4.a1>.
- Tourangeau, R., J. Michael Brick, S. Lohr, and J. Li. 2017. "Adaptive and Responsive Survey Designs: A Review and Assessment." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180(1): 203–223. DOI: <https://doi.org/10.1111/rssa.12186>.
- Wagner, J. 2019. "Estimation of Survey Cost Parameters Using Paradata." *Survey Practice* 12(1): 1–10. DOI: <https://doi.org/10.29115/SP-2018-0036>
- Wagner, J., and K. Olson. 2018. "An Analysis of Interviewer Travel and Field Outcomes in Two Field Surveys." *Journal of Official Statistics* 34(1): 211–237. DOI: <https://doi.org/10.1515/jos-2018-0010>.
- Wagner, J., and T.E. Raghunathan. 2010. "A New Stopping Rule for Surveys." *Statistics in Medicine* 29(9): 1014–1024. DOI: <https://doi.org/10.1002/sim.3834>.
- Wagner, J., B.T. West, H. Guyer, P. Burton, J. Kelley, M.P. Couper, and W.D. Mosher. 2017. "The Effects of a Mid-Data Collection Change in Financial Incentives on Total Survey Error in the National Survey of Family Growth." In *Total Survey Error in Practice*, edited by P.P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L.E. Lyberg, N.C. Tucker, and B.T. West. New York. Wiley.
- West, B.T., and A.G. Blom. 2017. "Explaining Interviewer Effects: A Research Synthesis." *Journal of Survey Statistics and Methodology* 5(2): 175–211. DOI: <https://doi.org/10.1093/jssam/smw024>.
- West, B.T., J. Wagner, F. Hubbard, and H. Gu. 2015. "The Utility of Alternative Commercial Data Sources for Survey Operations and Estimation: Evidence from the National Survey of Family Growth." *Journal of Survey Statistics and Methodology* 3(2): 240–264. DOI: <https://doi.org/10.1093/jssam/smv004>.
- West, B.T., J. Wagner, S. Coffey, and M.R. Elliott. 2019. "The Elicitation of Prior Distributions for Bayesian Responsive Survey Design." *Historical Data Analysis versus Literature Review*. Available at: <https://arxiv.org/ftp/arxiv/papers/1907/1907.06560.pdf>.

Received August 2019

Revised February 2020

Accepted May 2020