Journal of Survey Statistics and Methodology (2024) ***12****, 435–457*

LEVERAGING PREDICTIVE MODELLING FROM MULTIPLE SOURCES OF BIG DATA TO IMPROVE SAMPLE EFFICIENCY AND REDUCE SURVEY NONRESPONSE ERROR

DAVID DUTWIN *
PATRICK COYLE
JOSHUA LERNER
IPEK BILGEN
NED ENGLISH

Big data has been fruitfully leveraged as a supplement for survey data—and sometimes as its replacement—and in the best of worlds, as a "force multiplier" to improve survey analytics and insight. We detail a use case, the big data classifier (BDC), as a replacement to the more traditional methods of targeting households in survey sampling for given specific household and personal attributes. Much like geographic targeting and the use of commercial vendor flags, we detail the ability of BDCs to predict the likelihood that any given household is, for example, one that contains a child or someone who is Hispanic. We specifically build 15 BDCs with the combined data from a large nationally representative probability-based panel and a range of big data from public and private sources, and then assess the effectiveness of these BDCs to successfully predict their range of predicted attributes across three large survey datasets. For each BDC and each data application, we compare the relative effectiveness of the BDCs against historical sample targeting techniques of geographic clustering and vendor flags. Overall, BDCs offer a modest improvement in their ability to target subpopulations. We find classes of

David Dutwin is a Senior Vice President of BVI and a Chief Scientist of AmeriSpeak at NORC, University of Chicago, Chicago, IL, USA. Patrick Coyle is a Statistician at NORC, University of Chicago, Chicago, IL, USA. Joshua Lerner is a Statistician at NORC, University of Chicago, Chicago, IL, USA. Ipek Bilgen is a Chief Methodologist of AmeriSpeak at NORC, University of Chicago, Chicago, IL, USA. Ned English is a Principal Research Methodologist at NORC, University of Chicago, Chicago, IL, USA.

*Address correspondence to David Dutwin, NORC, University of Chicago, 55 E. Monroe Street, Chicago, IL 60603, USA; E-mail: dutwin-david@norc.org.

https://doi.org/10.1093/jssam/smad016 Advance access publication 10 June 2023

predictions that are consistently more effective, and others where the BDCs are on par with vendor flagging, though always superior to geographic clustering. We present some of the relative strengths and weaknesses of BDCs as a new method to identify and subsequently sample low incidence and other populations.

KEYWORDS: Big data; Machine learning; Sampling.

Statement of Significance

Survey sampling has long utilized geographic clustering and vendor flags to increase the survey incidence of various target populations. Each of these methods has its limitations. A third approach, that of utilizing big data to more precisely target survey populations of interest, is in its infancy. There has been no systematic evaluation of this new methodology, particularly as to whether it offers significant improvement to survey sampling compared to more traditional and historical approaches. We provide an evaluation that spans the development and application of 15 different big data classifiers (BDCs), replicated over three test datasets. In most cases, BDCs are more effective than other methods at targeting populations when selecting probability samples, leading to greater sampling efficiency and likely lower costs.

1. INTRODUCTION

Nonresponse error is one of the most significant concerns today in survey research. Response rates have plummeted over the past 30 years (Czajka and Beyler 2016; Williams and Brick 2018; Leeper 2019; Kennedy and Hartig 2019) from about (for telephone surveys) the mid-30 percent range to under ten and even five percent. This has raised concerns not only about the very nature of probability samples (Gelman 2012) but has also driven up costs as many more sample elements need to be worked to attain even a single interview. While samples have not become appreciably worse in demographic representation, in general, they are consistently too educated, too old, and typically too White, both for telephone surveys (Dutwin and Buskirk 2021) and multimode (mail push-to-web and outbound telephone) address-based samples (Sherr et al. 2010; CHIS 2019).

Regarding nonresponse error, researchers have sought many different ways to attain more representative samples. Designed materials used in multimode address-based sampling, for example, have sought to encourage Americans with relative low response propensities to participate, and a fair amount of work has sought to increase participation of Spanish speakers (Trussell 2010; Ventura et al. 2018; Hughes 2019). Research into differential incentives has

also attempted to increase participation from persons with low response propensities (Singer et al. 1999; Singer and Ye 2013). Historically, a more efficacious approach has been using stratified sampling techniques to boost the probability of selecting groups anticipated to be underrepresented in final samples of completed interviews. Sample stratification is a decades-old approach (Kish 1965; Kalton and Anderson 1986; Dutwin et al. 2014; Dutwin and Lopez 2014; Lohr 2022) whereby certain sample elements are given a higher probability of selection than others. This has been executed over the years *a priori*, or via adaptive design (Thompson 1997). Either way, the challenge of mitigating nonresponse error, particularly regarding *a priori* sample stratification, has been the lack of data on households prior to interviewing. Moreover, while there are techniques (geographic clustering and vendor flagging, as described later) to make guesses as to the makeup of householders prior to sampling, such techniques are varyingly effective: while helpful, they are also characterized by problems with false-positive rates. What if there were a better way to predict who lives where, and thus, samples could be more effectively targeted based on accurately predicted demographics and even other attributes of households and householders?

Enter the promise of big data classifiers (BDCs). BDCs are a product of methods that utilize multiple sets of available data from public and commercial sources to predict any number of household and householder attributes, from demographics to attitudes and even behaviors. If such predictions can offer superior performance compared to historical sample stratification techniques, they could increase sampling efficiency and produce samples with lower survey nonresponse error. In the case of screening studies, where only a subset of the general population qualifies to be interviewed, BDCs may not only increase the effectiveness of targetability but significantly reduce costs and weighting variance.

In this article, we report on the building of BDCs to predict 15 different demographic or household attributes in the use of sample targeting and nonresponse error reduction in general population samples. We then assess their efficacy by applying them to three separate national datasets. We conclude by discussing the trade-offs of using BDCs by considering test set performance. Our research should offer researchers a clear picture of the consistencies and variabilities of BDCs across many predictions and different data applications.

2. A BRIEF HISTORY OF POPULATION TARGETING

Disproportionate stratified sampling techniques to target and oversample populations of interest (Groves et al. 2009) have been around for over 70 years (Cochran 1961). The two principal techniques in the United States are (1) the employment of Census data for populations that geographically cluster (Lohr 2022) and (2) using preconstructed sample vendor "flags" (Barron et al. 2015)

that indicate that a household likely holds a member falling into the class of the flag (e.g., Hispanic/not-Hispanic). These techniques are applicable to both RDD and address-based samples, though use within address-samples, as we investigate in this article, offer a more direct application since the data used in these techniques is principally based on address.

Geographic clustering can be an effective way to increase survey incidence for populations that cluster, as is the case for particular ethnic and racial populations like Hispanics in the United States (Dutwin and Lopez 2014; Howell et al. 2020). In short, while it is not known whether a specific household includes a Hispanic adult from such data, it is known what percent of households in a given Census block group, some 1,000 households on average, are Hispanic. Ergo, block groups with high incidence of Hispanic households can be oversampled relative to block groups with low Hispanic incidence. While effective, geographic clustering still produces plenty of instances of reaching non-Hispanic households in a block group high in Hispanic incidence, for example. Notably, the technique only works insofar as a population clusters. For instance, using geographic clustering to oversample young persons, who do not robustly cluster geographically (United States Census Planning Database 2021), is ineffective. And one cannot use geographic clustering to target populations for which the Census has no data, such as for religious groups, people with specific political party identification, or persons who consume marijuana.

A second sample targeting technique, the use of vendor flags (Barron et al. 2015; English et al. 2019), is in some ways more optimal than geographic clustering, but has its own unique shortfalls as well. Vendor flags are indicators provided by sample and/or commercial data vendors. They have the advantage of making predictions often at the household or person level, compared to Census block group aggregations as in the case of targeting by geographic clustering. However, most flags are built with a mix of input data, and the techniques and variables that are used as inputs are not accessible to survey researchers. The process of generating flags is a proprietary black box utilizing data vendors' own algorithms to arrive at their predictions. And, in our experience, the predictions may differ across subsets.

Take, for example, most predictions of political party affiliation: in states where one can register for a political party, the data input into the flag is quite direct and exact; but for states where one does not register for a political party, it will typically be the case that a household is, at best, only predicted to be of a certain political party because the majority of households in that given precinct voted for one party versus another. Thus, while vendor flags can be attainable at the household and/or individual level, the data and method used to build the flags are opaque. Further, it is unclear to what degree vendor flags are up to date at any given time; as people move, it takes time for the flags to follow them. Finally, it is the case that vendor flags typically have both unit and item nonresponse issues. Valliant et al. (2014, table 2) estimated that 21.5

percent of the housing units were missing from commercial databases, whereas West et al. (2015) found a match rate ranging from 72.9 to 82.4 percent, depending on the vendor from which the flag was purchased. Furthermore, West et al. (2015) illustrated a wide range of *item* nonresponse rates from a number of vendors, from less than 10 percent for head of household to nearly 30 percent for the Hispanic indicator. Pasek et al. (2014) note that comparatively, different vendors flag different households on the same indicator. All of these findings present challenges for the use of vendor flags for sample targeting.

On the other hand, vendor flags offer a much greater range of possibilities for targeting compared to geographic clustering. First, they do not have to rely on aggregation at the block group or other levels. Second, there are many more indicators available. Common examples of household- or member-level vendor flag appends include matching telephone numbers associated with an address (Olson and Buskirk 2015), or indicators for the number of adults in a household (Roth et al. 2018), home tenure (Roth et al. 2018), the presence of children (English et al. 2014), or demographics such as age, race/ethnicity, or income (Roth et al. 2018; Pasek et al. 2014; DiSogra et al. 2010). It is clear that both the match-rate and accuracy of appended data vary depending on the variable of interest and specific geography of the households in question (Amaya et al. 2010; Pasek et al. 2014; Roth et al. 2018). Again, variability in efficacy depends upon how the data are modeled, compiled, and appended to an individual address (English et al. 2017).

BDCs offer a "third way," which may offer additional advantages and fewer disadvantages than geographic clustering and sample vendor flags. While relatively new to survey research, BDCs have been employed in election campaigns for some time. In 2008, Dan Wagner and the data science team within the Obama campaign used big data for household and person targeting or classification (Issenberg 2012). The principle is simple, although there are many complexities in practice: field a large survey to obtain a key self-reported attribute or classification of interest and append and use big data as the source of the independent variables in a model predicting that attribute. A new sample can then be drawn for a future study, big data appended, and the sample "scored" with predicted values from the prebuilt classifier model for whatever attribute was modeled. Wagner fielded surveys to gather two self-reported dependent variables: whether people intended to vote and which candidate they supported. Models were then built with as many as 1,000 variables from voter registration files and other databases—a task any machine learning model can easily handle. Once built, those models can subsequently be applied or "scored" to *any household or person* for which the same big data can be acquired. Political campaigns across the political spectrum have leveraged this technique to identify which households to approach with campaign interventions: those still potentially requiring persuasion to either get out to vote or to vote for a particular candidate (Nickerson and Rogers 2014).

Over a decade later, survey research largely has yet to incorporate a similar technique to target survey populations of interest, with a few preliminary efforts reported in the past. Dutwin (2020), for example, reported on the utility of using a BDC in predicting the religion of a household for Jewish Community studies. West et al. (2015) explored the utility of a limited amount of commercial data and voter registration data in predicting respondent eligibility in the National Survey of Family Growth. Each showed that commercial data could be an effective predictor of modeled attributes. Of course, the critical question concerning BDCs is: can they produce properties superior to geographic targeting and/or sample indicator vendor flags? On the face of it, BDCs should be an improvement for two main reasons. First, geographic population information and sample vendor flags are a form of big data that can be used as inputs to predictive modeling. Second, there are hundreds of big data variables available from a range of vendors *in addition* to geographic targets and appended vendor flags. As such, it behooves the survey research industry to seek to understand the ability to build and utilize such flags to attain more representative general population samples as well as to target hard-to-reach populations—or any population researchers wish to survey.

Our work reported here details a comparison of BDCs against the efficacy of geographic clustering and the use of vendor flags. We utilize a large national probability panel to build fifteen different BDCs and then compare the efficacy of the BDCs to clustering and flagging in three sets of survey data.

3. DATA

To uncover as many potential big data inputs as possible for our modeling, our exercise considered as wide a range of data as possible, including, broadly speaking, (1) publicly available federal data—including the American Community Survey, the Census Planning Database (Census PDB), and datasets from the U.S. Environmental Protection Agency, Federal Emergency Management Agency, Internal Revenue Service, and other agencies; (2) voter registration data from companies that provide such data; (3) social media data; and (4) general consumer data from data vendor sources.

In general, we found in our exploration of the potential to use these data that social media was limited in its ability to be linked to randomly selected addresses, and voter registration data held a limited number of variables specific to past voting histories and some demographics, many of which are modeled. Consumer data, however, has been gaining in popularity in recent years due to its abundance (GAO 2022). There are now dozens of companies that can provide a range of data, including but not limited to detailed housing purchase and value information, automotive purchasing behavior, social media use, demographics, detailed and extensive financial information including purchasing, travel, and investments, inferred behavior, and interest metrics (e.g.,

cat owner, interest in woodworking), charitable and other contribution behavior, detailed personal interests (classical music fan, football fan, etc.), magazine subscriptions, inferred lifestyle metrics (impulse buyer, etc.), media channel preferences, health metrics (conditions, healthcare data), and consumer purchase behavior (apparel types, furniture, jewelry, organic, etc.). In summary, consumer data sources offer considerable promise to those interested in targeting specific categories of households or individuals.

We utilized big data from two sources, the 2021 Census PDB (2021) and data from Merkle, a commercial data aggregator and modeler. The Census PDB is the dataset of choice for our "geographic clustering" technique, with variables indicating the percent of every block group in the United States on the number of households that are of a given race, ethnicity, age, educational status, and income, as well as the share of population and households that reside in public housing, that speak English, and other measures. Commercial data is widely available from many different providers, and our vendor provides nearly 600 variables on a range of topics, behaviors, and attitudes, similar to the list noted earlier. Due to space limitations, we cannot detail all the input variables here. However, we do provide a comprehensive list in an appendix in the supplementary data online.

The dataset we used to build and initially assess the BDCs was a random half of AmeriSpeak (NORC 2021), NORC's probability-based research panel, inclusive of all panelists from 2019 through 2020. AmeriSpeak panelists are sampled from US households with a known, nonzero probability of selection from NORC's National Sample Frame, an area probability frame that provides coverage of over 97 percent of US households. Panelists are recruited via mail, phone, and in-person techniques, and any household member ages 18 and older are eligible. AmeriSpeak reports a recruitment rate of 29 percent (AAPOR RECR, see AAPOR 2016) and offers panelists a mode choice (web or phone) for completion of subsequent surveys during recruitment. Overall, at the time of analysis, there were 58,556 unique recruited household members. AmeriSpeak is an excellent candidate for this process given the large number of self-reported metrics available, from demographics to religious, political, health, and financial attitudes and behaviors.

To build a dataset for training models, we link the AmeriSpeak panel data (1) to the Census PDB using census geography and (2) to the commercial vendor data if there is a person-level match between the respondent and an individual reported by the vendor. Details on this matching process can be found in section A6.1 in the supplementary data online and in figure 1. After record linkage and filtering, we obtain a training dataset of 10,737 unique AmeriSpeak panelists with commercial vendor data appended. From this dataset, we built two sets of classification models: (1) a set of models utilizing a random half of AmeriSpeak data for training and a "holdout" random half for testing and (2) a set of models utilizing the entire dataset to be applied to test datasets.

Figure 1. Flowchart of the Big Data Classifier Routine.

Once the BDCs were built (see section 4), they were applied to three separate datasets for assessment. We also appended to these datasets geographic clustering variables from the Census PDB and vendor flags from the purchased commercial data. These datasets were (1) the other half of the AmeriSpeak 2019–2020 panelists, (2) the successful recruits from the AmeriSpeak 2021 recruitment effort, and (3) a combination of the General Social Survey 2016–2020 panel and 2021 cross-section. For brevity, we refer the reader to more detailed technical and methodological summaries (General Social Survey 2020; NORC at the University of Chicago 2022). These data are large randomized national probability samples with large sample sizes (AmeriSpeak Panel 2019–2020 test data $N = 5{,}226$; AmeriSpeak Mail Recruitment 2021 $N = 8{,}717$; GSS $N = 5{,}815$). The sample for the AmeriSpeak 2021 recruitment effort was stratified by using multiple BDCs, namely, predictions on (1) speaking Spanish, (2) African American, (3) Hispanic, (4) age 18–49, (5) high school diploma or less education, and (6) child in the household. As such, these indicators are assessed in this article. We also built additional BDCs for

predicting (7) teenager in the household, (8) Asian American, (9) income under $25,000 per year, (10) income over $100,000 per year, (11) age 65 and older, (12) Jewish, (13) Catholic, (14) Republican, and (15) whether the household utilized social media (Facebook or Instagram). These variables were chosen as outcome variables given their availability as self-reported variables in the AmeriSpeak datasets and because such variables provide a range of metrics from core demographics to religion, political party identification, and social media use. Each of these models has been developed by NORC for use either in a proposal to conduct a study or to target sample in an actual study.

4. METHODS

There are a range of potential approaches to creating BDCs. As is evident from its name, big data is "big." In this context, its size has two elements. First, it must be big in that it is available for the majority of the sample of interest. The second way big data is big is the large number of features (variables) available. Big data should have *depth* and *breadth*. We are using a dataset that contains hundreds of predictor variables, which makes conventional modeling options like standard logistic regression problematic due to overfitting and multicollinearity and instead calls for machine learning approaches that can handle considerably more parameters (Hastie et al. 2009). Another option would be to reduce the number of independent variables after some process of uncovering those that are most consequential. Variables can be reduced by using mechanisms to identify these consequential variables, or alternatively, reduction can occur through clustering and other data reduction techniques.

The many steps involved in building, evaluating, and applying BDCs for survey sampling are summarized in figure 1. There are several issues in the modeling process that are important to review. First is the machine learning algorithm to choose for modeling. There are many options, including regression-based methods, which have often been used to model survey samples, nonresponse, and other attributes of surveys (Rosenbaum and Rubin 1983; Rosenbaum 1987; Schonlau et al. 2004; Duffy et al. 2005). Other options are Bayesian models, kernel-based models, and tree-based approaches (Hastie et al. 2009). As reported elsewhere, we tested methods from these families (simple neural nets, SVM, random forest, and LASSO). Ultimately, our machine learning method of choice is a tree-based ensemble method—gradient-boosted decision tree ensembles.

XGBoost (Chen and Guestrin 2016)—which stands for "eXtreme Gradient Boosting"—is an ensemble method that combines decision/regression trees with gradient boosting (Friedman 2002; Breiman 2001; Hastie et al. 2009) and is our method of choice. Gradient boosting involves three elements: a loss function to be optimized, a weak learner to make predictions (simple decision trees), and an additive model to add weak learners to minimize the loss

function. For XGBoost, two added wrinkles are included: a dynamic regularization approach to make tree-pruning more accurate and efficient and an optimization approach that utilizes a Newton-Raphson based solver rather than the traditional gradient descent approach that previous gradient-boosted models used (Chen and Guestrin 2016; Sagi and Rokach 2021). First, a series of simple decision trees are applied to the training data to predict the outcome of interest. These predictions are then compared to the actual outcomes, and the model calculates the residuals. A new ensemble predictor is then created to target and correct these errors, which are the differences between the predicted and actual outcomes. The process of applying decision trees, calculating residuals, and building a new predictor is repeated multiple times until the model's overall accuracy is optimized. This iterative process enables XGBoost to improve its predictions with each iteration and make accurate predictions even with complex and high-dimensional data. The number of iterations is optimized via cross-validation—too many iterations will result in overfitting. XGBoost introduces many hyperparameters to be learned via cross-validation. This high level of model complexity, with relatively fast-fitting models, has often been cited as why XGBoost consistently attains superior results compared to other machine learning approaches (Chen et al. 2022; Yan et al. 2020; Jin 2022). In our test scenarios, XGBoost performed better than other candidate models although Random Forest was a close second. We prefer XGBoost due to its predictive performance and its high-quality implementation in R, which includes features like sparse matrix methods and parallel programming. These features make XGBoost run quickly, allowing us to implement cross-validation with many predictors. Researchers interested in building their own BDC should explore multiple machine learning approaches as it is possible that other methods would work better within other contexts and with different data.

Second, data must be made "tidy" to produce optimized models, which involves typical data cleaning procedures such as dealing with missing values and consideration of outlier responses. Because tree-based methods can overestimate the importance of predictors with many unique values, we binned all ordinal predictors so that there are at most 25 unique values before including them in the models.

XGBoost produces probability scores as outcomes, so a third step occurs after modeling is complete. To utilize a continuous prediction running from 0 to 1, one needs to generate a cut point so that sample scored with predictions can be stratified. One of the significant advantages of BDCs over vendor flags is the power to select the optimal cut point for a given study. For example, a general population study that desires to increase its share of African Americans may utilize a supplementary sample based on a BDC. Since the main sample is a full random sample of all addresses in the United States, the predictions could be "cut" at a point where survey incidence will be high, but coverage will be low. To develop "adaptive" cut points, we exploit Bayes' Rule, which

states that the ratio between coverage and incidence will be equal to the ratio between prevalence and the percent of cases we predict to be positive. For example, if a study's sample frame has a prevalence of 10 percent, and we want our incidence to be twice as high as coverage, then we label the highest 5 percent of the predictions as positive predictions. This Bayesian approach is safer than relying on a cut point selected via cross-validation, because XGBoost predictions are not necessarily well-calibrated, and they may be extremely polarized (very close to either 0 or 1) if the XGBoost model was trained for many iterations. The cutpoints used for this article's analysis are based on the prevalence from each model's training data; for example, if the training data had 10 percent prevalence for a particular label, then 10 percent of test data cases are predicted to be positive for that label.

A fourth issue is aggregation. Models are run at the individual level because consumer data obtained is at the individual level. To score models to a fresh sample, one would draw an address-based sample and send it to the vendor, which would then return all individual-level records associated with those addresses. An address-based sample of 10,000 would typically generate twice as many individual records. These data would be scored with the BDCs, but since most surveys sample at the household level, the data would then need to be aggregated to the household level. There are a range of potential approaches to accomplish aggregation. There is no need for aggregation for households with one record, but what about a household with three members, where only one of the members is predicted into a given classification? Concerning within-household aggregation, we considered several possibilities. For the analyses and applications in this article, we aggregated with an "any" rule—if any individual in a household is predicted to be Hispanic, the household is coded as Hispanic.

The analysis of the article unfolds in three main steps. First, we build and apply a model for each of our fifteen outcome variables noted earlier, using all available consumer data and Census data as independent variables, to the training dataset of 50 percent of the AmeriSpeak 2019–2020 panelists. We then report results from the three different testing datasets: the remaining 50 percent of AmeriSpeak 2019–2020 panelists, the 2021 AmeriSpeak recruitment dataset, and the GSS dataset. The 50–50 train-test split in our data is a feature of our own different constructed datasets, and different considerations for split severity would have to be made on a case-by-case basis. There is nothing in particular about the 50–50 split we use that is dispositive of using different splits in different applications. We wanted, for evaluation purposes, for our test split to be as large as we could manage.

We report the incidence, coverage, and other performance metrics for each model within each dataset. From the confusion matrices, we primarily want to understand two metrics: precision and recall. Precision is the percent of all cases predicted to be of a certain class (e.g. Hispanic) that are actually of that class. In the parlance of survey sampling, this would be what one would expect

as the *survey incidence*. Recall is the percent of all cases in a given class that the predicted values correctly identify with being in that class. In survey parlance, this is *coverage*, the degree to which a given sample (cases predicted to be in the class) covers all households/householders in that class. We also examine other measures of model performance: F1 score, PRAUC, and Cohen's kappa, which are standard machine learning metrics (Hastie et al. 2009). These metrics are defined and discussed in appendix 5 in the supplementary data online. In appendix 7 in the supplementary data online, we examine the stability of test set performance by plotting 95 percent bootstrap confidence intervals of the performance metrics over 200 bootstrap resamples.

5. RESULTS

Overall, we compare the results of the BDCs with outcomes from the vendor flags and geographic clustering since this article aims to assess BDCs against these historical oversampling methods. Figures 2–4 report the incidence, coverage, and F1 scores of the BDCs. Corresponding tables A3.1, A3.2, and A3.3 are provided in appendix 3 in the supplementary data online, offering tabular reports of the figures, as well as differences in the performance of the BDCs versus the vendor flags and predictions based only on Census data ("geographic clustering").

5.1 AmeriSpeak Training Dataset Results

Figure 2 shows that, for the AmeriSpeak 2019–2020 test data, the BDCs are 17 percent higher in incidence and 8 percent lower in coverage on average compared to vendor flags. The difference is similar to geographic clustering, where BDCs are 19 percent improved in incidence and 7 percent improved in coverage. For example, as illustrated in figure 2 and detailed in table A3.1 in the supplementary data online, the BDC for predicting whether there is a teenager in a household attained a 44 percent incidence, compared to 14 percent incidence for geographic clustering and 22 percent incidence for the vendor flag. In this case, the BDC attained 36 percent coverage, compared to 14 percent coverage for geographic clustering and 38 percent coverage for the vendor flag. PRAUC and Cohen's Kappa results also underscore the superiority of the "teenager in household" BDC to alternative methods. Use of this model would significantly increase the survey incidence attained in reaching households with a teenager and successfully find a larger percentage of all households with a teenager.

The BDC performs better than geographic clustering for all models for this test case. In most cases, we see that BDCs can offer an increase in incidence in exchange for a decrease in coverage, compared to the vendor flags. In other words, the BDCs successfully ordered the sample from "least likely" to "most

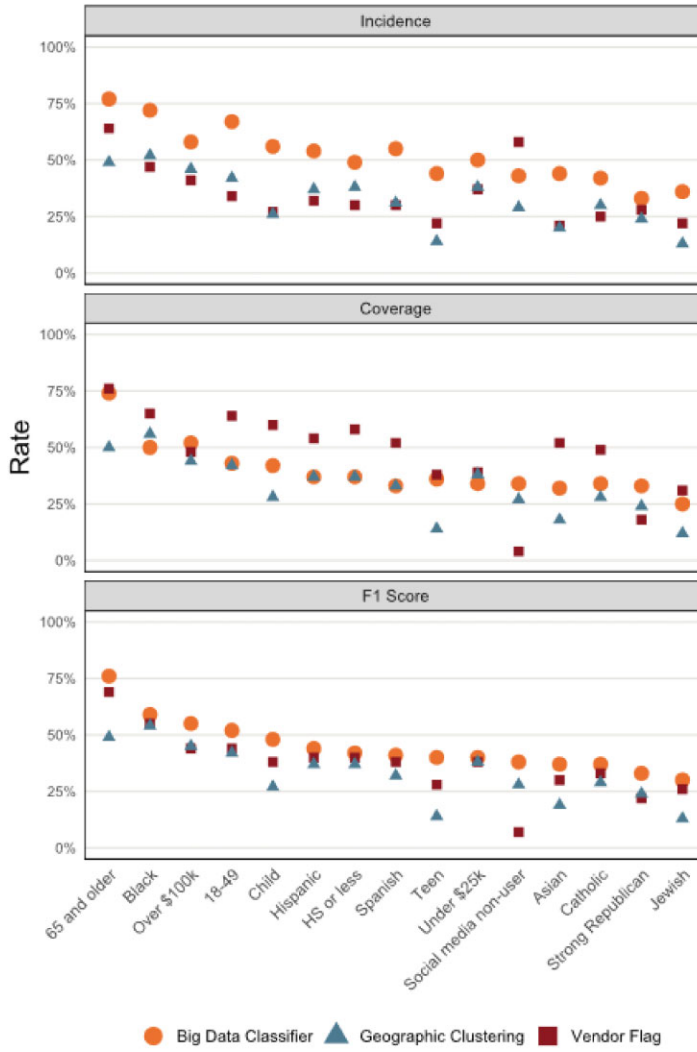

Figure 2. Model Performance from AmeriSpeak 2015–2020 Test Dataset.

likely", in a way that the binary vendor flag cannot. This allows us to select from the *m* most likely cases for sampling, where *m* may be less than the number predicted to be positive by the vendor flag.

5.2 2021 AmeriSpeak Recruitment Dataset and GSS Results

As a further test, we scored the models to the 2021 AmeriSpeak recruited sample (as presented in figure 3 and table A3.2 in the supplementary data online)

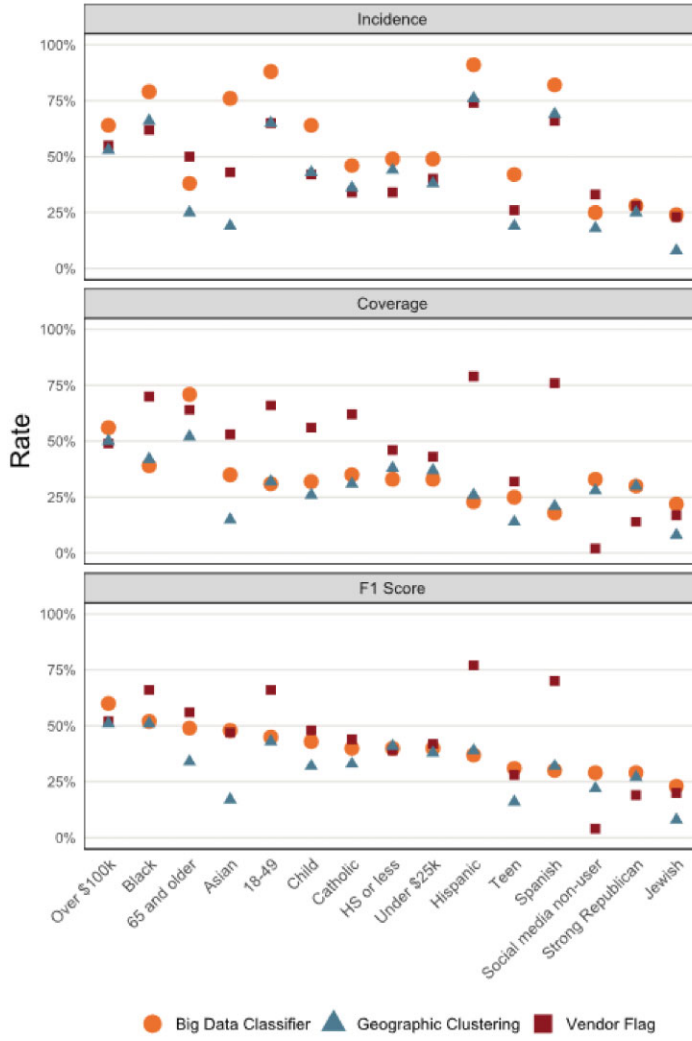

Figure 3. Model Performance from AmeriSpeak 2021 Recruitment.

and to the General Social Survey sample (figure 4 and table A3.3 in the supplementary data online). These samples were not used in either training or testing and therefore provide independent comparisons of the BDCs to vendor flags and geographic clustering (in machine learning terms this acts as a validation set). Overall, performance was slightly weaker for these test sets than for AmeriSpeak 2019–2020 test data, but with many similarities. Compared to

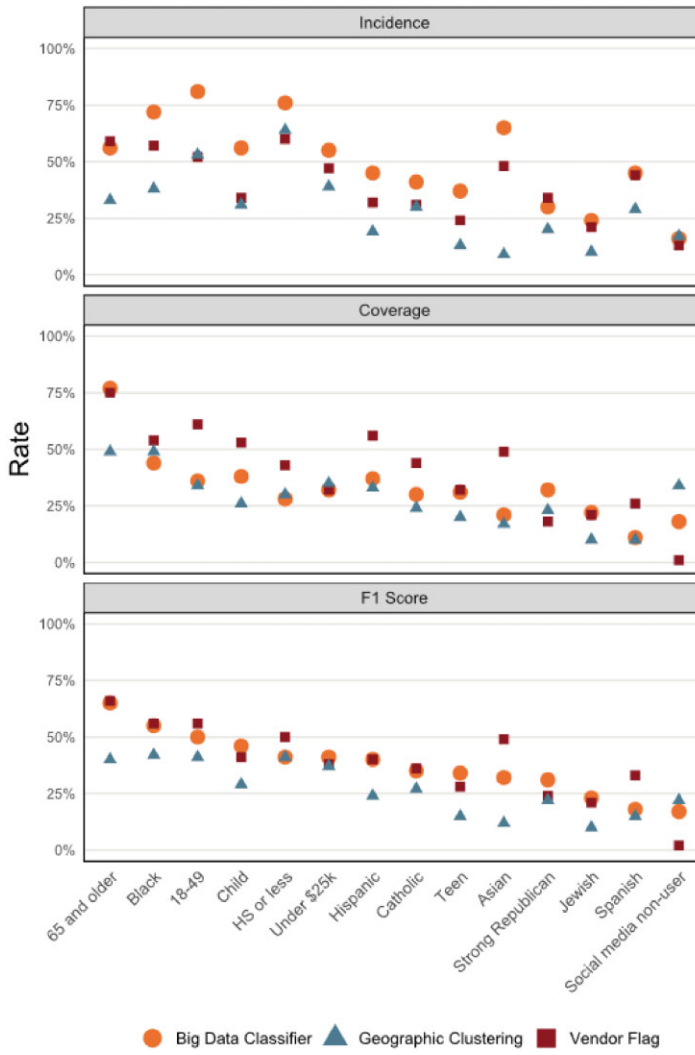

Figure 4. Model Performance for GSS.

vendor flags, BDCs offer an average of 11 percent incidence improvement with a 14 percent decrease in coverage for AmeriSpeak 2021 Recruitment data (figure 3/table A3.2 in the supplementary data online), and an average of 10 percent incidence improvement with an 8 percent decrease in coverage for GSS data (figure 4/table A3.3 in the supplementary data online). The improvement over geographic clustering was +16 percent/+5 percent (incidence and

coverage) for AmeriSpeak 2021 Recruitment data and +21 percent/+5 percent for GSS data.

BDC performance for the AmeriSpeak 2021 Recruitment data (figure 3/table A3.2 in the supplementary data online) showed higher incidence for finding household members under 18 compared to vendor flags, in exchange for lower coverage (+22 percent/−24 percent for children and +16 percent/−7 percent for teenagers). In this case, the BDCs were strong in identifying income over $100k (average of incidence and coverage = +8 percent) and Asians (+8 percent), but weaker for other racial/ethnic groups (average of incidence and coverage = −8 percent for Black, −20 percent for Hispanic, and −21 percent for Spanish speaking). Against geographic clustering, the BDCs continued to show consistent improvement for almost all models.

For the GSS data (figure 4/table A3.3 in the supplementary data online), the BDCs perform better than the vendor flag (when comparing average of incidence and coverage) for nine out of fourteen models. For finding household members under 18, the BDCs once again attained substantially higher incidence than the vendor flag by sacrificing coverage (+22 percent/−15 percent for children and +12 percent/−2 percent for teenagers).

5.3 Overall Results

Table 1 explores overall comparisons of the mean differences between BDCs, geographic clustering, and vendor flags, as well as the range of performance across the three datasets as measured by the standard deviation of the difference in outcomes (incidence, coverage, and average of incidence and coverage). In short, a low standard deviation indicates that the performance of a BDC versus geographic clustering or a vendor flag was consistent across all three test datasets. The table shows that, on average, there is a significant range in the efficacy of the BDCs versus the vendor flags and geographic clustering. The BDCs have significantly better incidence than vendor flags for targeting teenaged household members in exchange for reasonable coverage loss (average improvement of 16 percent incidence and loss of 4 percent coverage). However, the BDCs are modestly worse for identifying persons of low education (+17 percent/−16 percent), and Catholics (+13 percent/−18 percent).

Table 1 further supports (1) the strength of the BDCs for predicting income, given large mean improvement compared to the vendor flags, and (2) the weakness of the BDCs in improving identification of households with Hispanics and Spanish-speakers, and Asian persons, given low or negative mean improvement compared to the vendor flags. Compared to geographic clustering, the BDCs show large mean differences across all predicted variables, but most pronounced for age, Asian, and high income predictions.

Table 1. Cumulative Results Across All Test Data

Target	Incidence mean change (%)		Coverage mean change (%)		Total mean change (%)		Incidence deviations (%)		Coverage deviations (%)		Average deviations (%)	
	GC	Flag	GC	Flag	GC	Flag	GC	Flag	GC	Flag	GC	Flag
Child	25.1	24.1	10.8	−18.9	17.9	2.6	4	3.6	4.4	4.3	4.2	4
Black	22.1	18.5	−4.1	−18.5	9	0	11.1	5.3	1.7	11.2	6.4	8.2
Hispanic	19.6	17.3	0.4	−30.5	10	−6.6	5.7	4.7	3.5	21.9	4.6	13.3
Teen	25.4	16.4	14.7	−3.5	20	6.5	3.7	4.5	6.1	2.9	4.9	3.7
Asian	45.6	24.4	12.8	−22	29.2	1.2	18.6	8.4	8	5.3	13.3	6.9
HS or less	9.6	16.7	−2.2	−16.2	3.7	0.3	4	1.9	2.5	4.2	3.2	3.1
Under $25k	13.1	9.9	−3.4	−5.2	4.8	2.4	2.9	2.4	1.2	5.4	2	3.9
Over $100k	11.5	12.9	7	5.3	9.2	9.1	1.1	6	0.5	2.5	0.8	4.3
18–49	25.4	28.1	0.5	−27.1	13	0.5	2.6	5.1	1.9	7.5	2.2	6.3
65 and older	21.5	−0.3	23.7	2.3	22.6	1	8.1	12.9	4.7	4.1	6.4	8.5
Spanish	17.3	13.6	−0.5	−30.5	8.4	−8.4	5.7	12.5	2.2	23.3	3.9	17.9
Strong Republican	7.2	−0.1	5.8	14.8	6.5	7.4	3.5	5	5.1	1.4	4.3	3.2
Jewish	17.9	6.2	12.8	−0.2	15.3	3	4.7	7.6	1	6	2.8	6.8
Catholic	11	12.7	5.5	−18.3	8.2	−2.8	1	3.3	0.8	7.2	0.9	5.2
Social media nonuser	6.9	−6.8	−1.5	26	2.7	9.6	7.3	9	13.3	8.4	10.3	8.7
Average	18.8	12.9	5.4	−9.8	12.1	1.5	2.6	3.9	1.5	3.7	1.6	3.2

NOTE.—Table shows the average performance of BDCs relative to two alternative methods: geographic clustering (*GC*) and vendor flags (*Flag*).

Additional summary results for the BDCs are provided in appendices 1, 2, and 7 in the supplementary data online. Appendix 1 in the supplementary data online reports variable importance for each model. Appendix 2 in the supplementary data online contains graphs comparing precision and recall across all possible cut points. Appendix 7 in the supplementary data online examines the stability of model performance across test datasets using bootstrap resampling.

6. DISCUSSION

Big data, many initially thought, promised to transform the survey research industry. But over time, such thinking has become more nuanced, and the benefit of big data has moved from the overarching to the specific use-case. We offer details on one such use-case, the use of big data to create BDCs to effectively target households and people during survey recruitment on various attributes, from demographics to behavioral and even attitudinal variables. Our approach has been shown to be effective for many, but not all, metrics for this specific use-case.

Our study finds both variance and consistency. On one hand, we find variance in the degree to which BDCs for one metric are superior to the alternative methods of utilizing vendor flags or geographic clustering compared to other metrics. On the other hand, we find relative consistency in these differences across datasets. This suggests some test–retest reliability: if a classifier is superior to other methods in all three datasets, then there is higher confidence in its effectiveness in future applications. This we found true for predictions of age, income, social media use, and more modestly so for religion and political party. On the other hand, our data suggest that one will likely be able to use vendor flags just as effectively as the BDCs for race and ethnicity predictions.

Using the optimized cut-points, BDCs consistently improve incidence, but inconsistently improve coverage. Again, this is a function of the cut point created for the BDCs, which we regard as a relative strength of BDCs. We have also conducted research that contains a significantly large general population sample and then a modest oversample of a specific group. In such a use-case, researchers may choose to select a higher cut point, sacrificing some coverage in favor of incidence, given that the oversampled population already has some level of complete coverage in the general population sample. Of course, researchers should be mindful of the design effects created with different cut points and oversampling fractions.

A second reason we believe that vendor flags sometimes exhibit better coverage is that we have anecdotally found the item nonresponse in the vendor flags (within cases that can be matched) to have declined over time. Compared to West et al. (2015), we found consistently lower item nonresponse; for example, the Hispanic indicator, which West found to be 30 percent missing, is

missing for only 2 percent of cases in the consumer data we received (overall unit nonresponse, however, was similar).

It should be the case that BDCs offer improvement upon current techniques, such as the use of available sample vendor flags, given that there are additional costs to big data classification. While the only financial cost in utilizing sample vendor flags is licensing the flags themselves, there are two ways in which big data classification is potentially more costly. First is the specialized labor involved in building and scoring the models, though these costs diminish over time as the programming time decreases and the cost of building a model amortizes over several applications. Second is that models can only be scored to sample after sample generation: addresses must be sampled and sent to big data vendors to append their data, and only then can the models be applied so that their predictions can be appended to the sample. Thus, in a use-case where one desires to oversample a prediction by a factor of four, one would have to generate four times as much sample as is needed pay the append the big data to the sample, but then randomly discard some proportion of the sample not predicted to be in the class of the oversample to arrive at the appropriate sampling fraction.

But for these costs, the benefits are improved coverage and incidence properties, leading to fewer terminated interviews due to screening, which are typically very costly. The benefits of this efficiency improvement, in our experience thus far, outweighs (and often greatly so) the expense of acquiring and appending big data, which typically costs pennies per record. To further explore this, we offer a cost–benefit analysis in appendix 4 in the supplementary data online.

Why might the BDCs not perform even better, relative to vendor flags specifically, than what our analysis found? While data vendors assert that their data is constantly updated, we cannot verify for each of the hundreds of variables provided that the data is genuinely up to date: these data come from multiple sources, Americans are a relatively mobile population, and their circumstances (employment status, number of kids, marital status, etc.) are dynamic. All these factors may lead to mismatches of true properties of persons and individuals to the data provided. We note the potential for error in that, at an individual level, we could only match about half of 2019–2020 AmeriSpeak panelists to their big data. While household rates are much higher at about 85 percent, error may be lower given a higher match rate at the individual level. Finally, while we have a sizable sample in the full modeling dataset of 10,737 panelists, some of the classifications we are predicting have small sample sizes. Generally, groups are substantial: there are 2,204 respondents reporting a child in the household, and 4,545 respondents aged 65 and over. There are three groups examined with particularly low prevalences in this dataset: Asian ($N = 194$), Jewish ($N = 140$), and Spanish speaking respondents ($N = 484$); all other groups examined have a sample size of at least 700 respondents. While we might note that the important sample size for predicting

the existence of any group is the overall sample size (e.g., both Asians and non-Asians), the smaller sample sizes of specific groups will limit the ability to find specific variables that reliably correlate to group membership. In the future, we hope to continue to model using ever larger datasets, minimizing any impact of small group sample sizes.

In addition to the ability to select one's own cut point, other advantage of big data classification exist. While vendor flags are limited to what is available, BDCs can be built for any attribute so long as one has an adequate survey to serve as training data where respondents self-report the attribute used as the outcome measure and that the survey dataset can be effectively merged with big data to be utilized as inputs. As such, we have built models on metrics not easily estimable by other means, such as smoking status and drug use. Additionally, one is not limited to survey self-reports; one could as well utilize interviewer observations as modelling targets (Sinibaldi et al. 2014). Of course, there are no guarantees that any model will be robust, but the possibility of building the model exists given adequate self-reported data. We feel the ability to create one's own custom model targeting a specific population not considered by sample vendors is a primary advantage of the big data approach.

While outside of the scope of this article, we have preliminarily explored the quality of samples attained via classification versus geographic clustering and vendor flag methods. So far, we have seen that cases successfully predicted of a specific attribute (e.g., Hispanics) by BDCs attain a more representative sample than in cases predicted by vendor flags or geographic clustering. We hope to explore this question in more detail in future research, but this makes theoretical sense in that BDCs utilize all the information from geography, vendor flags, and a host of other variables to make predictions. As such, there is some level of "averaging" going on in that bias in BDCs is not limited as a function of one method of targeting, given that geographic clustering data, vendor flags, and predictions from other variables are all leveraged in the BDCs.

With costs in surveys exploding due to declining response rates (Kennedy and Hartig 2019) and concerns for systematic nonresponse increasing, the potential of BDCs becomes ever more pronounced. Knowing who lives in a sampled household before fielding has always been tremendously beneficial to survey researchers. BDCs are the next step in the advancement of survey sample targeting and in the application of big data to survey science. We show that, on average, BDCs provide an advantage over other targeted sampling techniques. We hope this article provides a valuable resource to samplers in terms of to what degree they can expect a classifier to improve survey incidence and coverage and when to leverage BDCs over other disproportionate sampling techniques.

Supplementary Materials

Supplementary materials are available online at academic.oup.com/jssam.

REFERENCES

Amaya, A., Skalland, B., and Wooten, K. (2010), "What's in a Match?," *Survey Practice*, 3, 1–5.

American Association for Public Opinion Research (2016), *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys* (9th ed.), Chicago, IL: AAPOR.

NORC at the University of Chicago (2022), "Technical Overview of the AmeriSpeak® Panel, NORC's Probability-Based Household Panel." Available at https://amerispeak.norc.org/ content/dam/amerispeak/research/pdf/AmeriSpeak%20Technical%20Overview%202019%2002% 2018.pdf

Barron, M., Davern, M., Montgomery, R., Tao, X., Wolter, K., Zeng, W. K., Dorell, C., and Black, C. (2015), "Using Auxiliary Sample Frame Information for Optimum Sampling of Rare Populations," *Journal of Official Statistics*, 31, 545–557.

Breiman, L. (2001), "Random Forests," *Machine Learning*, 45, 5–32.

California Health Interview Survey (2019), *CHIS 2017-2018 Methodology Series: Report 1 – Sample Design*, Los Angeles, CA: UCLA Center for Health Policy Research.

Chen, T., and Guestrin, C. (2016), "XGBoost: A Scalable Tree Boosting System," in Proceedings on the 22nd SIGKDD Conference on Knowledge Discovery and Data Mining.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., and Yuan, J. (2022), "_xgboost: Extreme Gradient Boosting_." R package version 1.6.0.1. Available at https://CRAN.R-project.org/package=xgboost.

Cochran, W. G. (1961), "Comparison of Methods for Determining Stratum Boundaries," *Bulletin of the International Statistical Institute*, 38, 345–358.

Czajka, J. L., and Beyler, A. (2016), *Declining Response Rates in Federal Surveys: Trends and Implications (Background Paper)*, Washington, DC: Mathematica Policy Research.

DiSogra, C., Dennis, J. M., and Fahimi, M. (2010), "On the Quality Of Ancillary Data Available For Address-Based Sampling," in Proceedings Of The American Statistical Association, Section On Survey Research Methods (Vol. 417483).

Duffy, B., Smith, K., Terhanian, G., and Bremer, J. (2005), "Comparing Data from Online and Face-To-Face Surveys," *International Journal of Market Research*, 47, 615–639.

Dutwin, D. (2020), "Feedback Loop: Using Surveys to Build and Assess Registration-Based Sample Religious Flags for Survey Research," in *Big Data Meets Survey Science: A Collection of Innovative Methods*, eds. C. A. Hill, P. P. Biemer, T. D. Buskirk, L. Japec, A. Kirchner, S. Kolenikov, and L. E. Lyberg, Hoboken, NJ: Wiley, pp. 537–557.

Dutwin, D., Ben Porath, E., and Miller, R. (2014), "U.S. Jewish Population Studies: Opportunities and Challenges," in *The Social Scientific Study of Jewry: Sources, Approaches, Debates, Studies in Contemporary Jewry*, ed. U. Rebhun, New York: Oxford Academic, pp. 55–73.

Dutwin, D., and Buskirk, T. (2021), "Telephone Sample Surveys: Dearly Beloved or Nearly Departed? Trends in Survey Errors in the Age of Declining Response Rates," *Journal of Survey Statistics and Methodology*, 9, 353–380.

Dutwin, D., and Lopez, M. H. (2014), "Considerations of Survey Error in Surveys of Hispanics," *Public Opinion Quarterly*, 78, 392–415.

Enamorado, T., Fifield, B., and Imai, K. (2022), "Package 'fastLink.'" Available at https://cran.r-project.org/web/packages/fastLink/fastLink.pdf.

English, N., Allen, M., and O'Muircheartaigh, C. (2017), "Using Commercial Data to Enhance Survey Eligibility: The AmeriSpeak Experience," in 2017 Proceedings of the American Statistical Association, Survey Research Methods Section, Alexandria, VA: American Statistical Association.

English, N., Kennel, T., Buskirk, T., and Harter, R. (2019), "The Construction, Maintenance, and Enhancement of Address-Based Sampling Frames," *Journal of Survey Statistics and Methodology*, 7, 66–92.

English, N., Li, Y., Mayfield, A., and Frasier, A. (2014), "The Use of Targeted Lists to Enhance Sampling Efficiency in Address-Based Sample Designs: Age, Race, and Other Qualities," in Proceedings of the American Statistical Association, Survey Research Methods, Alexandria, VA: American Statistical Association.

Friedman, J. H. (2002), "Stochastic Gradient Boosting," *Computational Statistics & Data Analysis*, 38, 367–378.

Gelman, A. (2012), "Statistics in a World Where Nothing Is Random." Available at www.andrew-gelman.com.

General Social Survey (2020), Available at https://gss.norc.org/Get-The-Data.

Government Accounting Office (2022), "Consumer Data: Increasing Use Poses Risks to Privacy." Available at https://www.gao.gov/assets/gao-22-106096.pdf.

Groves, R., Fowler, F., Couper, M., Lepkowski, J., Singer, E., and Tourangeau, R. (2009), *Survey Methodology* (2nd ed.), Hoboken, NJ: Wiley.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Vol. 2). New York: Springer.

Howell, C. R., Su, W., Nassel, A. F., Agne, A. A., and Cherrington, A. L. (2020), "Area Based Stratified Random Sampling Using Geospatial Technology in a Community-Based Survey," *BMC Public Health*, 20, 16–78.

Hughes, T., Wells, B., Park, R.., and Sherr, S. (2019), "'Responda hoy': An Experiment in Recruiting Spanish Speakers for an ABS Web Survey," presented at the Annual Conference of the American Association for Public Opinion Research, Toronto, CA.

Issenberg, S. (2012), "How Obama's Team Used Big data to Rally Voters," *MIT Technology Review*. Available at https://www.technologyreview.com/2012/12/19/114510/how-obamas-team-used-big-data-to-rally-voters.

Jin, Y. (2022), "Tree Boosting With XGBoost – Why Does XGBoost Win 'Every' Machine Learning Competition?" *Synced*. Available at https://syncedreview.com/2017/10/22/tree-boosting-with-xgboost-why-does-xgboost-win-every-machine-learning-competition/.

Kalton, G., and Anderson, D. (1986), "Sampling Rare Populations," *Journal of the Royal Statistical Society, Statistics in Society Series A*, 149, 65–82.

Kennedy, C., and Hartig, H. (2019), "Response Rates in Telephone Surveys Have Resumed Their Decline." Available at https://www.pewresearch.org/short-reads/2019/02/27/response-rates-in-telephone-surveys-have-resumed-their-decline/.

Kish, L. (1965), *Survey Sampling*, New York: John Wiley and Sons.

Landau, W. M. (2021), "The Targets R Package: A Dynamic Make-like Function-Oriented Pipelinetoolkit for Reproducibility and High-Performance Computing," *Journal of Open Source Software*, 6, 2959.

Leeper, T. J. (2019), "Where Have the Respondents Gone? Perhaps We Ate Them All," *Public Opinion Quarterly*, 83, 280–288.

Lohr, S. (2022), *Sampling Design and Analysis*, London: Routledge.

Nickerson, D. W., and Rogers, T. (2014), "Political Campaigns and Big Data," *Journal of Economic Perspectives*, 28, 51–74.

Olson, K., and Buskirk, T. (2015), "Can I Get Your Phone Number? Examining The Relationship Between Household, Geographic and Census-Related Variables and Phone Append Propensity For ABS Samples," presented at the 70th Annual AAPOR Conference, Hollywood, FL.

Pasek, J., Jang, S. M., Cobb, C., Dennis, J. M., and DiSogra, C. (2014), "Can Marketing Data Aid Survey Research? Examining Accuracy and Completeness in Consumer-File Data," *Public Opinion Quarterly*, 78, 889–916.

Rosenbaum, P. R. (1987), "Model Based Direct Adjustment," *Journal of the American Statistical Association*, 82, 387–394.

Rosenbaum, P. R., and Rubin, D. B. (1983), "The Central Role of Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

Roth, S., Caporaso, A., and DeMatteis, J. (2018), "Variables Appended to ABS Frames: Has Data Quality Improved?" presented at the 73rd Annual AAPOR Conference, Denver, CO.

Sagi, O., and Rokach, L. (2021), "Approximating XGBoost with an Interpretable Decision Tree," *Information Sciences*, 572, 522–542.

Schonlau, M., Zapert, K., Simon, L. P., Sanstad, K. H., Marcus, S. M., Adams, J., Spranca, M., Kan, H., Turner, R., and Berry, S. H. (2004), "A Comparison between Responses from a Propensity-Weighted Web Survey and an Identical RDD Survey," *Social Science Computer Review*, 22, 128–138.

Sherr, S., Triplett, T., Rapoport, R., Long, S., and Dutwin, D. (2010), "A Comparative Study of ABS and RDD Sampling in Two Locations: The Commonwealth of Massachusetts and the District of Columbia," presented at the Annual Conference of the American Association for Public Opinion Research, Boston, MA.

Singer, E., Van Hoewyk, J., Gebler, N., Raghunathan, T., and McGonagle, K. (1999), "The Effect of Incentives on Response Rates in Interviewer-Mediated Surveys," *Journal of Official Statistics*, 15, 217–230.

Singer, E., and Ye, C. (2013), "The Use and Effects of Incentives in Surveys," *The Annals of the American Academy of Political and Social Science*, 645, 112–141.

Sinibaldi, J., Trappmann, M., and Kreuter, F. (2014), "Which is the Better Investment for Nonresponse Adjustment: Purchasing Commercial Auxiliary Data or Collecting Interviewer Observations?" *Public Opinion Quarterly*, 78, 440–473.

Thompson, S. K. (1997), "Adaptive Sampling in Behavioral Surveys," *NIDA Research Monograph*, 167, 296–319.

Trussell, N. (2010), "Spanish Respondents' Choice of Language: Bilingual or English?" *Survey Practice*, 3, 1–4.

United States Census Planning Database (2021), U.S. Census. Available at https://www.census.gov/topics/research/guidance/planning-databases.html.

Valliant, R., Hubbard, F., Lee, S., and Chang, C. (2014), "Efficient Use of Commercial Lists in US Household Sampling," *Journal of Survey Statistics and Methodology*, 2, 182–209.

Ventura, I., Bautista, R., and Henderwan, E. (2018), "An Experiment in Recruitment for Spanish Speaking Populations: The Amerispeak Case Study," Proceedings of the American Statistical Association, AAPOR Survey Research Methods, Alexandria, VA: American Statistical Association.

West, B., Wagner, J., Hubbard, F., and Gu, H. (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, 240–264.

Williams, D., and Brick, J. M. (2018), "Trends in U.S. Face-to-Face Household Survey Nonresponse and Level of Effort," *Journal of Survey Statistics and Methodology*, 6, 186–211.

Yan, L., Zhang, H.-T., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., Sun, C., Tang, X., Jing, L., Zhang, M., Huang, X., Xiao, Y., Cao, H., Chen, Y., Ren, T., Wang, F., Xiao, Y., Huang, S., Tan, X., Huang, N., Jiao, B., Cheng, C., Zhang, Y., Luo, A., Mombaerts, L., Jin, J., Cao, Z., Li, S., Xu, H., and Yuan, Y. (2020), "An Interpretable Mortality Prediction Model for COVID-19 Patients," *Nature Machine Intelligence*, 2, 283–288.