AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

Tree-Based Models for Political Science Data

Jacob M. Montgomery Washington University in St. Louis
Santiago Olivella University of North Carolina at Chapel Hill

Abstract: *Political scientists often find themselves analyzing data sets with a large number of observations, a large number of variables, or both. Yet, traditional statistical techniques fail to take full advantage of the opportunities inherent in "big data," as they are too rigid to recover nonlinearities and do not facilitate the easy exploration of interactions in high-dimensional data sets. In this article, we introduce a family of tree-based nonparametric techniques that may, in some circumstances, be more appropriate than traditional methods for confronting these data challenges. In particular, tree models are very effective for detecting nonlinearities and interactions, even in data sets with many (potentially irrelevant) covariates. We introduce the basic logic of tree-based models, provide an overview of the most prominent methods in the literature, and conduct three analyses that illustrate how the methods can be implemented while highlighting both their advantages and limitations.*

Replication Materials: The data, code, and any additional materials required to replicate all analyses in this article are available on the *American Journal of Political Science* Dataverse within the Harvard Dataverse Network at: https://doi.org/10.7910/DVN/8ZJBLI.

Social science scholars often work with data sets containing a large number of observations, many potential covariates, or (increasingly) both. Indeed, political scientists now regularly analyze data with levels of complexity unimaginable just two decades ago. Widely used surveys, for instance, interview tens of thousands of respondents about hundreds of topics. Scholars of institutions can quickly assemble data sets with thousands of observations using resources like the Comparative Agendas Project. Moreover, new measurement methods, such as text analysis, have combined with data sources, such as Twitter, to generate databases of almost unmanageable sizes. It is clear that political science, like all areas of the social sciences, will increasingly have access to a deluge of data so vast that it will dwarf everything that has come before.

What statistical methods are needed in this data-saturated world? Surely, there is no one correct answer. Yet, just as surely, traditional statistical models are not always equipped to take full advantage of new data sources. Traditional models—largely variants of linear regressions—are ideal for evaluating theories that imply specific functional forms relating outcomes to predictors. In particular, they excel in their ability to leverage assumptions about the data-generating process, or DGP (additivity, linearity in the parameters, homoskedasticity, etc.) to make valid inferences despite inherent data limitations. Although appropriate when testing theories that conform with these assumptions, standard models are often insufficiently flexible to capture nuances in the data—such as complex nonlinear functional forms and deep interactions—when no clear a priori expectations exist.

In this article, we introduce a family of tree-based nonparametric techniques from the machine learning literature. We argue that, under specific circumstances, regression and classification tree models are an appropriate standard choice for analyzing high-dimensional data sets. In particular, past research has shown tree-based methods to be very useful for making accurate predictions when the underlying DGP includes nonlinearities, discontinuities, and interactions among many covariates. Further, tree models require few assumptions. Rather than imposing a presumed structure on the DGP, tree-based methods allow the data to "speak for themselves." Thus, our goal in this article is to introduce political scientists to this promising family of methods, which are well suited for today's data analysis demands.

In the next sections, we discuss the promise and perils of high-dimensional, "large"-N data sets and introduce the basic logic of tree models. We then provide an overview of the most prominent methods in the literature.

Jacob M. Montgomery is Associate Professor, Department of Political Science, Washington University in St. Louis, Campus Box 1063, One Brookings Drive, St. Louis, MO 63130 (jacob.montgomery@wustl.edu). Santiago Olivella is Assistant Professor, Department of Political Science, University of North Carolina at Chapel Hill, Hamilton Hall 361, CB 3265, Chapel Hill, NC 27599 (olivella@unc.edu).

American Journal of Political Science, Vol. 62, No. 3, July 2018, Pp. 729–744

 DOI: 10.1111/ajps.12361

Next, we conduct three analyses that demonstrate both the advantages of tree models as well as their limitations. First, we conduct a simulation study to illustrate when tree-based methods are most appropriate and their performance relative to alternatives. We then apply them to a data set with many potential explanatory variables to generate estimates of the probability of campaigns "going negative," quantities that we subsequently use within a marginal structural modeling framework to estimate causal effects (Blackwell 2013). Finally, we replicate and extend Ghitza and Gelman (2013) and analyze a large collection of survey responses to estimate attitudes and behaviors of small demographic subgroups, which requires the efficient estimation of "deep" interactions between multiple covariates.

Before moving on, it is important to note that the precise role of tree-based models and other machine learning methods in the social science enterprise is an open question. Some scholars have argued that tree models are valuable tools for testing theories and estimating complex causal effects (e.g., Hill 2012; Imai and Strauss 2011). Yet, tree models were first and foremost designed for making accurate out-of-sample predictions rather than for testing theoretical claims. Moreover, the ability to "discover" subtleties in the data is not always a virtue. As we discuss below, the risk of confusing signal for noise in high-dimensional data is very real, and for some tasks, tree-based models are overly complex. If we are testing a theory adequately encapsulated by a parametric model, more traditional approaches are not only sufficient, but preferable. Lastly, large data sets and flexible models generally do not remove the burden from researchers for devising suitable theories and having a clear understanding about how predictors affect outcomes. Indeed, failing to take theoretical considerations seriously when building complex models often results in nonsensical findings (Lazer et al. 2014).

Thus, in this article, we advocate for the expanded use of tree models for characterizing complex DGPs where the goal is *not* direct theory testing but rather accurate prediction. That is, we believe tree models can serve as appropriate standard choices when researchers' primary goal is to correctly capture the nuances of a potentially complex but unknown data-generating process in a setting with many potential predictors related in nonlinear and interactive ways to the outcome. Superficially, our focus on prediction seems restrictive. As we show in our examples below, however, there are many instances in which tree models can contribute meaningfully to essential social science tasks, including estimating causal effects and improving measures of latent traits.

The Promise and Perils of Flexible Models of "Big Data"

When building predictive models using large data sets, quantitative scholars face two countervailing pressures. First, one wishes to leverage the richness of the data to correctly capture the data-generating process (DGP), thus avoiding model misspecification. A good model would allow for a large number of possible covariates and for complex interactions between them. So too it would allow for nonlinear functional forms and even discontinuous shifts in how a set of covariates is related to outcomes.

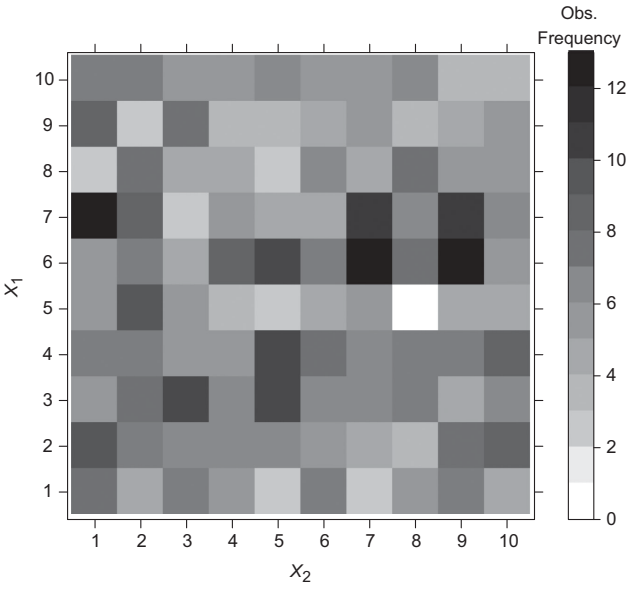
The second (and contrary) need is to avoid overfitting the data, a goal sometimes labeled *regularization*. Overfitting occurs when the model is so complex that it makes predictions based on idiosyncratic features of the data unrelated to the true DGP. In other words, we confuse the noise and the signal in our data, leading both to poor out-of-sample predictive performance and an incorrect understanding of the DGP. Overfitting, of course, is a potential problem for any statistical method, but it is particularly endemic for models that contain a large number of predictors, flexible functional forms, and deep interactions.

To understand how the goals of flexibility and regularization are at odds, consider a hypothetical example with $N = 600$ observations of a single outcome variable (y) with two possible predictor variables (x_1, x_2). Assume further that the predictors, which take on just 10 integer values, are distributed uniformly. How do we then determine what value of y should be associated with each unique combination of the covariates?

One naïve proposal might be to model the outcome based on the average value ($\bar{y}$) observed for each unique combination of the categories in x_1 and x_2, or "region." This fully interactive specification would be the ultimate in flexible models, allowing for almost any possible relationship between the covariates and the outcome. The problem, however, is that there is generally not sufficient data to execute this strategy. Figure 1 shows the number of observations that appear in each region in one simulated data set. In this example, the median region has just six observations, and the maximum number of observations in any region is 13. With so little data in each region, we increase the risk of overfitting.

Thus, even in a relatively simple world with only two covariates and a modestly large sample size, there is not sufficient data to make valid predictions about the expected value of y for each region. Obviously, this problem becomes exponentially worse as variables are added. With three similar covariates, there would be $10^3 = 1{,}000$

FIGURE 1 Number of Observations in Each "Region," or Unique Combination of x_1 and x_2 with N = 600 Randomly Generated Observations



possible regions, meaning that the majority of regions would be empty. Indeed, it is clear that in a data set with just 20 covariates, even "big data" on the grandest imaginable scale will not be big enough for this strategy to succeed.

Standard parametric models circumvent this problem by making assumptions about the DGP. Common regression models, for instance, assume that the value of y increases as a linear function of the (possibly transformed) covariates. The advantage is that we can accurately recover relationships between the covariates and the outcome despite the sparsity of data in each region. The disadvantage is that they eradicate aspects of data that do not conform with their underlying assumptions.

The trade-offs involved are illustrated in Figure 2. The upper-left panel shows the true DGP for the 600 observations shown in Figure 1. The upper-right panel shows the estimates generated by the naïve approach of estimating $\bar{y}$ for each combination of x_1 and x_2. As expected, this approach leads to significant overfitting: Estimates fluctuate wildly in response to random error rather than the true DGP. On the other hand, the bottom panels show estimates from a simple linear model and a model with polynomial terms and interactions. In both, the estimated relationships between the covariates and the outcome are clearly inadequate and would lead researchers to an incorrect understanding of the DGP.

Tree-based models are members of a growing class of methods from the machine learning literature designed to yield a balanced solution to this dilemma—allowing flexible functional forms while avoiding overfitting. Their goal is to specify regions of the covariate space such that the outcome is homogeneous and the number of observations in each region is sufficiently large, yet where the regions themselves are sufficiently numerous and unstructured to allow for complex relationships between covariates and the outcome. In this way, tree-based models are related to neural networks (Beck, King, and Zeng 2000), kernel regularized least squares (Hainmueller and Hazlett 2014), and other nonparametric techniques.

While each of these approaches has its own advantages, tree-based methods are particularly attractive in offering versatility and ease of use. Tree models are highly flexible, easily accommodating common problems such as missing data, interactions between many variables, and both continuous and discrete outcomes. Further, tree models are easy to interpret relative to other "blackbox" techniques, such as neural networks, although interpretation remains a challenge relative to, say, generalized additive models (Beck and Jackman 1998). Finally, although tree-based methods perform best when irrelevant variables are excluded, they are relatively adept at ignoring uninformative predictors. While imperfect—inevitably, some irrelevant predictors are chosen in splitting rules—tree-based methods tend to produce relatively parsimonious models even when offered many uninformative predictors.

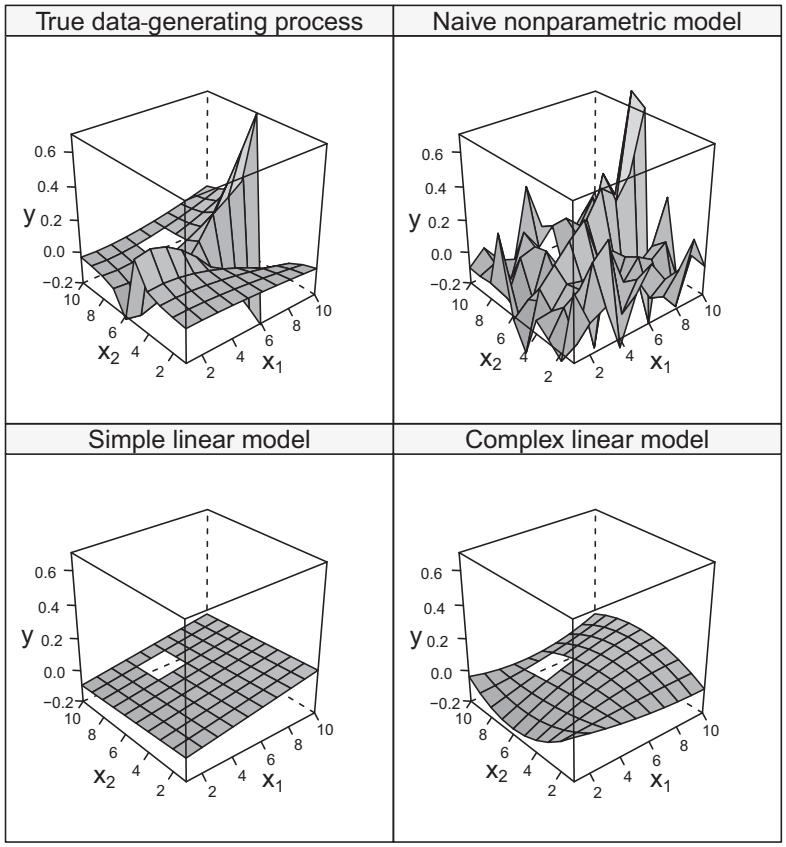
Although tree models are not unknown in the discipline (e.g., Green and Kern 2012; Imai and Strauss 2011; Kastellec 2010; Muchlinski et al. 2016), they have appeared rarely. At the same time, they are now "go-to" models in literatures focused on prediction and classification (Hastie, Tibshirani, and Friedman 2009). Our aim in the next sections, therefore, is to introduce the most prominent tree models and showcase their potential in analyzing political science data.

Single-Tree Models

At their core, tree-based models involve two basic steps.[1] First, they divide the covariate space into B nonoverlapping and exhaustive regions, $R_1, R_2, \ldots, R_B$, that are relatively homogeneous with respect to the outcome y. Second, they make a prediction, c_b, for all observations that fall within region R_b.

[1]For this section, we follow presentations in Faraway (2005) and Hastie, Tibshirani, and Friedman (2009).

FIGURE 2 True and Recovered Relationships in Simulated Data



Note: The true DGP is $y_i = \frac{1}{100} \times \sqrt{x_{i1}x_{i2}}\frac{(5.5-x_{i1})^2}{(5.5-x_{i1})(5.5-x_{i2})} + \epsilon_i$, where $\epsilon_i \sim N(0, 0.35)$. The simple linear model is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, whereas the complicated model is $y = \beta_0 + \text{poly}(x_1, 2) \times \text{poly}(x_1, 2)$ (where poly(x, d) is the sequential polynomial-generating function, d is the highest degree generated, and the $\times$ operator generates all main effects and interactions).

To understand this more clearly, consider the classification and regression tree (CART) model. Its first step consists of partitioning the covariate space into (hyper)rectangles. The left panel of Figure 3 displays a partition of a two-dimensional covariate space into 14 nonoverlapping and exhaustive regions using the same data as depicted in Figure 2. Each region corresponds to unique covariate value combinations, which can be succinctly represented in the form of a binary tree (shown in the central panel of Figure 3 for our example). At each internal node of the tree, the covariate space is split into two distinct regions depending on the splitting rule associated with the node (e.g., $X_1 \leq v_1$).

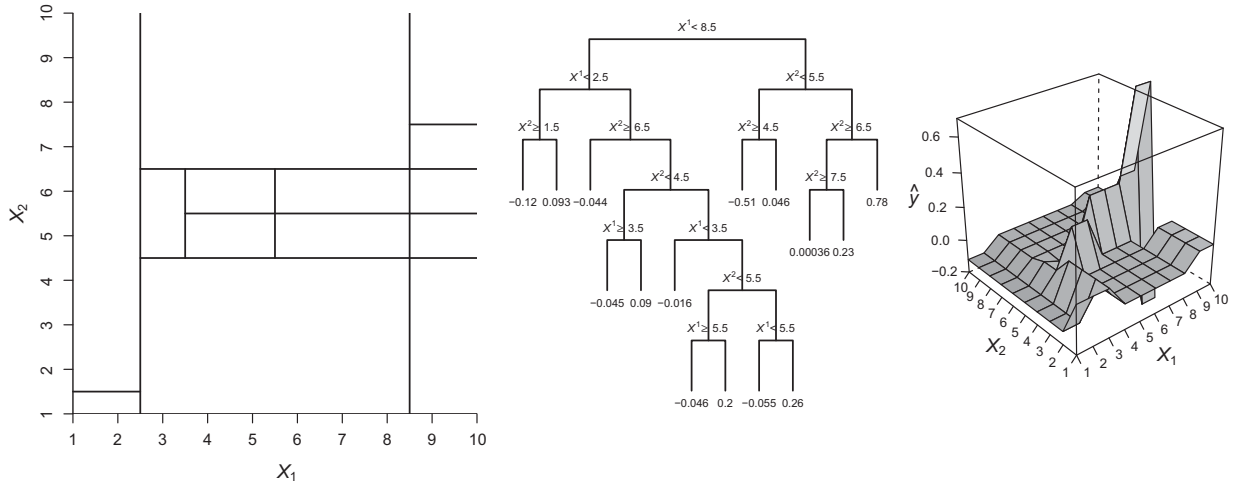
The terminal nodes, or "leaves," of the tree correspond to the regions, and constant predicted values (c_b) are assigned to each region/leaf. For a continuous outcome variable y_i, CART defines this constant as the mean outcome for all observations within region R_b ($\bar{y}_i \forall i \in R_b$).[2] Thus, the model produces a prediction surface for every possible combination of the explanatory variable values. For instance, the right panel of Figure 3 shows the predicted "response surface" corresponding to the regions defined in the left and center panels for our running example. More formally, a tree model for J covariates is a function

$$f(X_i) = T(X_i; \Theta) \equiv \sum_{b=1}^{B} c_b I(X_i \in R_b), \qquad (1)$$

where $I(\cdot)$ is the usual indicator function, and Θ is a set of parameters that contains the tree depth (or size), the region definitions (i.e., the splitting rules), and the predicted values (c_b).

[2]Although we focus on continuous outcomes, the same principles apply to categorical outcomes. The supporting information includes a discussion of trees for categorical outcomes.

FIGURE 3 Left: Example of a Partition of a Two-Covariate Space into 14 Rectangular Prediction Regions. Center: A Binary Tree Corresponding to the Partition Depicted on the Left. Right: 3-D Plot of the Prediction Surface Corresponding to Regions Defined in the Left and Center Panels



Choosing Θ optimally will yield a response surface that accurately captures the true relationship between the covariates and the outcome y while avoiding overfitting. Accurately retrieving the response surface is an optimization problem, namely,

$$\hat{\Theta} = \arg\min_{\Theta} \sum_{b=1}^{B} \sum_{X_i \in R_b} L(y_i, c_b),$$

where $L(\cdot)$ is a loss function that can be thought of as a measure of model fit. The usual loss function[3] in regression trees is the familiar sum of squared errors, $\sum_{i:X_i \in R_b} (y_i - c_b)^2$.

Since finding the *best* partition and predicted value combination for a given loss function is computationally prohibitive, CART adopts a heuristic known as recursive binary splitting to find an acceptable solution. This procedure is described in more detail in the supporting information.

For all its simplicity, CART is likely to fail in terms of preventing overfitting (Sutton 2005). After all, a tree with the same number of nodes as observations will produce a prediction surface that exactly matches observed outcomes, and that thus wildly overfits the data (as in the naïve example in the second section). A common strategy is to grow large trees and then "prune" them (Breiman et al. 1984). Complexity pruning involves finding a subtree T that minimizes the quantity $C_\alpha(T) = \sum_{b=1}^{B} L(y_{i:X_i \in R_b}, c_b) + \alpha B$, where $L(\cdot)$ is the loss function, B is the number of terminal nodes, and $\alpha \geq 0$ is a prespecified parameter that controls the trade-off between tree size and fit.

The advantage of CART models is that they can be built quickly and are relatively easy to understand and interpret. Depictions of binary trees are a very intuitive means of conveying modeling results (see, e.g., Kastellec 2010, 216). Moreover, single-tree models easily accommodate complicated interactive relationships, continuous and discrete predictors, and large numbers of irrelevant predictors. However, single-tree models perform very poorly when uncovering additive relationships (Fox 2000). Further, the algorithmic approach to building a tree leaves us with no means for assessing uncertainty in our estimates. Finally, the sequential nature of the binary recursive splitting algorithm means that the structure of the tree is often highly sensitive to small changes in the observations included.

Fortunately, the intuitive logic behind single-tree methods can be extended in order to successfully address these drawbacks by combining multiple trees that are aggregated to create superior ensemble models. In the next section, we review three methods for creating tree ensembles before turning to our empirical illustrations.

[3] In general, the choice of a loss function depends on the type of data being modeled, as well as on the inferential goal at hand. In certain instances, choice of a loss function carries implications regarding the conditional distribution of the outcome (e.g., negative log-likelihood loss functions are closely related to the distributional assumptions of classical Generalized Linear Models, or GLMs), so it is important for researchers to ensure their choice of a loss function is justified given the estimation problem at hand (see Hastie, Tibshirani, and Friedman 2009, 221–22).

Ensemble Approaches: A Family of Trees

Ensemble methods combine multiple trees of the type defined in Equation (1) in order to better approximate the outcome surface while reducing overfitting. Their general form is

$$f(X_i) = \sum_{m=1}^{M} T_m(X_i; \Theta_m), \quad (2)$$

where M is the number of trees, and Θ_m is the set of parameters that define each tree T_m. The approaches we review here differ only in the ways in which they construct the individual trees and weight them when forming the ensemble. Specifically, we discuss three of the most commonly used sum-of-trees models: random forests, gradient boosting machines, and Bayesian additive regression trees. Our aim is to provide an intuitive understanding of each model. We direct readers interested in more detailed presentations to James et al. (2013), Hastie, Tibshirani, and Friedman (2009), and Chipman, George, and McCulloch (2010).

Tree Bagging and Random Forests

Tree bagging—short for *bootstrap aggregating*—relies on the fact that single-tree methods can result in very different predictive surfaces depending on which observations are included. This is particularly true with "deep" trees with many terminal nodes. The intuition behind tree bagging is to conceptualize trees fit to different subsets of data as if they were independent draws of a random variable. By this logic, we can reduce the variance in single-tree estimates of the response surface by fitting many trees and combining them as defined in Equation (2). To the extent the independence assumptions hold, this ensemble model will provide a low-variance, low-bias estimate of the true response surface. More formally, tree bagging takes multiple simple random samples of the same data set (with replacement), of size equal to that of the original data set. It then fits a deep tree (with no pruning) to each bootstrapped sample. For M samples, the tree-bagging model is then $\hat{f}_{bag}(X_i) = \frac{1}{M}\sum_{m=1}^{M} T_m(X_i; \hat{\Theta}_m)$.

Although the power of tree bagging stems from the assumed independence of the trees, in practice, trees are highly correlated. Random forests (RF) address this pitfall by systematically lowering the level of correlation between trees (Breiman 2001). Specifically, at each splitting stage of the tree-growing algorithm, the RF selects an optimal splitting rule based on only a random subset of $a < j$ of the covariates. Moderately small values of a reduce the correlations among trees and improve the performance of the ensemble.

Tree Boosting and Multiple Additive Regression Trees

While superficially similar to bagging and random forests, tree boosting approaches the problem of creating multiple trees from a very different angle. First, the bagging procedure creates trees *independently*. Boosting, on the other hand, builds trees *sequentially*, such that each new tree improves the predictive power of the ensemble. Second, whereas bagging relies on fitting trees to *random samples* drawn from the data, boosting relies on fitting trees to *transformations* of the data. The result is a procedure that grows new trees specifically aimed at accommodating observations that the existing ensemble predicts poorly.

Boosting approximates a solution to the problem of fitting a sum of trees by adding new trees one at a time, while keeping all existing trees unchanged. At each stage m of this forward stagewise process, boosting solves

$$\hat{\Theta}_m = \arg\min_{\Theta_m} \sum_{i=1}^{N} L(y_i,\ f_{m-1}(X_i) + T_m(X_i; \Theta_m)), \quad (3)$$

where $f_{m-1}(X_i)$ is the value of the sum of trees that was estimated in the first $m-1$ stages. Intuitively, this stagewise approximation forces each new tree to focus on the errors of its predecessors—thereby progressively reducing lack of fit.

While the intuition behind boosting is straightforward, optimizing Equation (3) is not. However, the approximating procedure of choice—known alternatively as gradient boosting, multiple additive regression trees (MART), or gradient boosting machines (GBM)—is both extremely accurate and fast. In short, at the m^{th} stage of the process, GBM fits a new tree to the *negative gradient* of the loss function ($-\mathbf{g}_m$). That is, it approximates a solution to Equation (3) with

$$\hat{\Theta}_m = \arg\min_{\Theta_m} \sum_{i=1}^{N} (-g_{im} - T_m(X_i, \Theta_m))^2, \quad (4)$$

where g_{im} is the i_{th} component of $\mathbf{g}_m$ and serves as a more informative kind of residual.[4]

This strategy can create arbitrarily accurate models simply by increasing the number of trees. To prevent

[4]Let the gradient of the loss function be $\mathbf{g}_m = \partial L(y_i, f_{m-1}(X_i))/\partial f_{m-1}(X_i)$. When constructing the m^{th} tree, $\mathbf{g}_m$ is a vector pointing in the direction of steepest *increasing* loss. GBM's strategy is to add trees that move $f(X_i)$ in the opposite direction (Friedman 2001).

this, researchers can prespecify two parameters, in addition to selecting the number of trees. First, we can choose the number of terminal nodes in each tree, denoted B. Note that, given the additive form of GBM, the choice of B also determines the *maximum* order of interactions in the model—an upper limit that is usually justified substantively, and therefore held fixed at some (preferably low) level. Second, the regularizing role of B is complemented by tree shrinkage, which is achieved by scaling the contribution of each new tree by a factor, $0 < \nu < 1$, such that the running sum becomes $f(X_i) = f_{m-1}(X_i) + \nu T(X_i; \Theta_m)$. Setting ν close to zero limits each new tree's contribution to the model's prediction, which in turn increases the number of trees that need to be fit in order to approximate the outcome surface. Intuitively, setting ν to a low value allows the expansion to "learn" the outcome surface slowly (Hofner et al. 2014). On the other hand, setting ν too low can lead to slow rates of learning. In practice, it is common to preset both ν and B and find an optimal number of trees using cross-validation.

Bayesian Trees

Bayesian additive regression trees (BART) are similar to GBMs in that trees in the ensemble are grown to accommodate residuals of the current fit rather than outcomes themselves. Further, the contribution of each tree to the entire fit is regularized so that no one tree dominates the prediction of the response surface. Unlike GBMs, however, tree-growing and regularization goals are achieved by assuming that the parameters that govern tree construction can be estimated under a hierarchical Bayesian framework (Chipman, George, and McCulloch 2010). This provides both estimates of their expected predicted values and, uniquely, measures of uncertainty.[5]

The definition of BART begins by replacing Equation (2) with

$$y_i = \sum_{m=1}^{M} T_m(X_i; \Theta_m) + \epsilon_i, \qquad \text{with } \epsilon_i \sim N(0, \sigma^2), \quad (5)$$

for continuous outcome y_i. For a given number of trees M, we then place independent priors over all parameters in Θ_m: the depth of each tree, the variables used at each split, the value used for the corresponding split, the terminal-node values, and the error variance in Equation (5) (i.e., σ).[6] The definition of these prior probability distributions controls the influence of each tree in the prediction of the outcome surface such that no one tree can dominate the ensemble.

Relying on a Markov chain Monte Carlo algorithm, BART explores the space of all possible "forests" with M trees, producing a sample of M trees at every step and, with it, a sample of the outcome variable given predictors X_i. So, for instance, if we specify that the model should have 100 trees, BART creates a posterior sample of the 100-tree models that are likely given both the observed data and our priors. We can then summarize the posterior predicted outcomes in terms of their expected values and variability across draws. The result is a highly flexible, data-responsive ensemble method, which produces measures of uncertainty in the very process of finding a sum of trees that accurately reproduces a given outcome surface (Hill 2012).

In this section, our primary goal was to provide well-grounded intuition as to how tree-based models work. However, we provide additional discussions of some important practical considerations in our supporting information, including information about available `R` packages, approaches to choosing tuning parameters, and ideas of how to interpret the substantive effects.

Three Illustrations

We now provide three examples that showcase the advantages of tree models while also illustrating their relative strengths and weaknesses. First, we evaluate the performance of CART, RF, GBM, and BART models using synthetic data and compare them with several alternative approaches. Second, we use GBM and BART to predict when campaigns will engage in negative advertising. This example illustrates the advantages of tree models when the researcher's primary aim is to accurately recover the response surface for a specific outcome. We use these predictions to estimate the causal effect of negative campaigns on vote share in U.S. elections using the strategy outlined in Blackwell (2013). Finally, we replicate and extend the estimation of subgroup attitudes and behaviors in U.S. elections conducted by Ghitza and Gelman (2013) to illustrate the ability of tree methods to model deeply interactive relationships.

An Illustration Using Synthetic Data

We begin by creating 40 different potential covariates—including symmetric and asymmetric variables, continuous and categorical variables, and correlated and

[5] See Hofner, Kneib, and Hothorn (2014) on creating bootstrapped standard errors for non-Bayesian tree models.

[6] The structure of the regularization priors is discussed in the supporting information.

independent variables.[7] We then create outcomes under three different data-generating processes (DGPs): an additive and linear specification; a specification with both additive terms and interactions; and a more complicated specification that contains additive terms, interactions, nonlinearities, and discontinuities. In each case, we also ensure that at most 4 of the 40 features are actually related to the outcome of interest, and that each outcome contains some amount of Gaussian error. Finally, for each DGP, we create 100 training sets of 500 observations and a single test set with 3,000 observations. We use these test sets to evaluate the relative predictive strengths of the different methods.

We fit four different tree-based models to each of the training sets: CART, RF, GBM, and BART. In addition, we fit three non-tree-based models for comparison purposes: a (Gaussian) kernel-regularized least squares (KRLS) model, a single-hidden-layer neural network (NN), and a generalized additive model. Each of these models requires that we preselect various "tuning" or regularization parameters. For the BART model, we use the recommended default values for the model's prior hyperparameters discussed in Chipman, George, and McCulloch (2010).[8] KRLS uses an automated leave-one-out cross-validation procedure to choose its parameters. Further, for the Generalized Additive Model, or GAM, we used a thin plate to smooth over the *preidentified* relevant covariates in each DGP. This strategy allows us to show just how well tree-based models perform even compared to an unrealistically well-specified GAM.[9] Finally, for the CART, RF, GBM, and NN, we conducted a fivefold cross-validation to choose tuning parameters for each model for *each of the 100 training sets*.[10] For the tree models, we searched over the parameter values shown in Table SI-2 in the supporting information.[11]

To evaluate the relative performance of each, we calculate the out-of-sample root mean square error (RMSE) by first fitting the model using data from each of the 100 training sets and then evaluating their predictive accuracy using the test set. This results in 100 RMSE values for each model, for each DGP. To provide a meaningful scale to these RMSE values, we normalize them by the best observed RMSE, producing *relative* RMSE (RRMSE) measures. Thus, an RRMSE of 1.5 would indicate that a model has performed 50% *worse* than the best observed individual fit, and values closer to 1 indicate better performances.

Figure 4 presents box plots of the relative RMSE distributions across the 100 training sets, with a reference line indicating the relative RMSE achieved using a mean model (i.e., a model that predicts test observations using the mean of the test outcome). Distributions corresponding to tree-based models are shaded in gray.

The left panel shows the results when the underlying DGP is a simple additive relationship between the covariates and the outcome. Unsurprisingly, a smoothed linear model (viz., GAM) can outperform all others when its underlying assumptions are met. More interestingly, the tree-based models actually perform comparably well—even without the unfair advantage enjoyed by the GAM model. GBM, for instance, has a median RRMSE that is less than 5% larger than the best GAM, and BART performs similarly well on average. As discussed in the supporting information, the major shortcoming of single-tree CARTs is their inability to pick up the additive portions of a DGP—an issue illustrated by being one of the worst-performing models when confronted with a strictly additive DGP. RF tends to do much better than single-tree models in terms of predictive variance, but it still shares CART's weakness in performing relatively poorly for strictly additive DGPs.

Across the remaining two panels, tree-based models—particularly tree ensembles—perform consistently well in comparison to other strategies. The other models are either less consistent or make poor predictions through insufficient regularization. The tree ensemble models show relatively good predictive performance under a DGP typical in political science research (viz., the additive and multiplicative DGP used in the central panel of Figure 4) and do even better when the DGP is simultaneously additive, multiplicative, nonlinear, and discontinuous.[12] Gradient-boosted tree ensembles perform well under all circumstances, followed closely by BART and random forests. Overall, then, tree-based models are shown to perform well under a variety of data-generating circumstances, offering very little room for researcher manipulation of model specifications and results. We next

[7] Additional details for this simulation are provided in the supporting information.

[8] Specifically, we use `sigdf = 3`, `sigquant = .9`, `ntree = 200`, and `k = 2`.

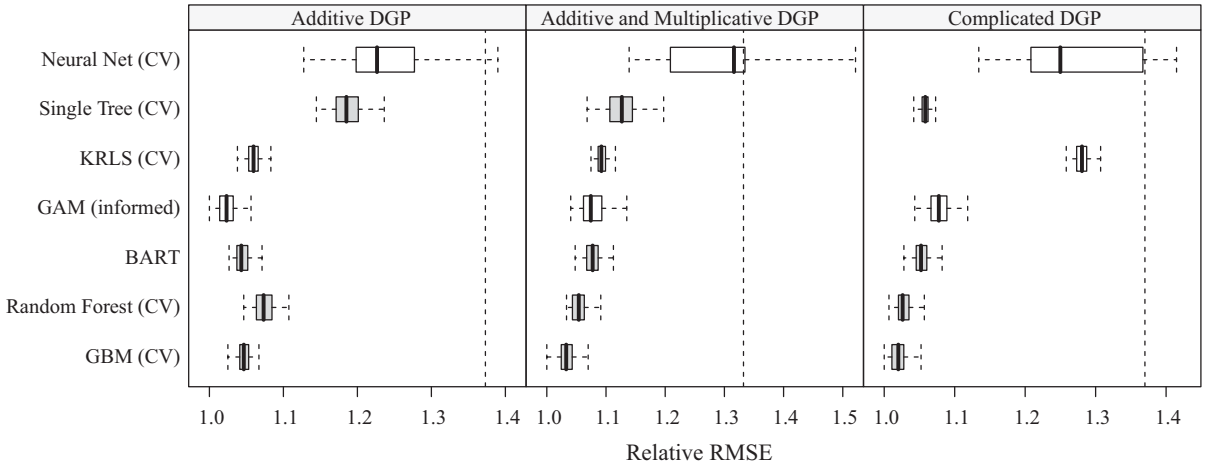
[9] There are not enough degrees of freedom to estimate a model that remains agnostic as to the variables over which to smooth.

[10] As noted in the supporting information, it is critical to consider many possible values for the number of trees in the GBM model. Due to the large number of data sets in this simulation, we consider values ranging from 1 to 2,500 by increments of 3. However, in applied settings an even finer grid is suggested.

[11] For the NN model, we held the decay constant at 0.1 and let the number of neurons vary from 1 to 15 by increments of 2.

[12] In the supporting information, we further illustrate that tree ensembles perform well when recovering the underlying prediction surfaces.

FIGURE 4 Relative RMSE across 100 Training Sets for Each Model and Each DGP Specification



Note: Lower values indicate better relative predictive accuracy with respect to test outcomes. The dashed vertical line indicates the RRMSE of a model that simply predicts the mean value of y in the test set (i.e., a mean model). GAM is estimated by smoothing over known predictors independently, and it should therefore be understood to have an unfair advantage over all other models.

turn to providing examples of how these models can be incorporated into political science research.

Accurate Predictions of Action Sequences in Marginal Structural Models

When studying dynamic processes, researchers wishing to make valid causal inferences are often faced with a difficult dilemma. On the one hand, failing to include important covariates leads to omitted variable bias. On the other hand, including many of the most important covariates in a dynamic setting may induce posttreatment bias.

To address this concern, Blackwell (2013) outlines a marginal structural modeling (MSM) approach for estimating the effects of time-varying covariates by relying on estimated inverse probability of treatment weights (IPTW). Blackwell (2013) applies this framework to estimate the effect of negative campaigning during the 5 weeks prior to the election on the two-party vote share for 144 Democratic candidates in the election cycles between 2000 and 2006. Specifically, the action sequence of interest is whether the candidate has "gone negative" in a given week. We direct interested readers to Blackwell (2013) for a fuller discussion and provide a brief summary of the data and methods here.

Assume that for observation i at time period t, we observe action $A_{it} \in \{0, 1\}$, which represents, for instance, whether campaign i engaged in negative campaigning in week t. To implement the MSM method, one can take the following steps: (1) Estimate the probability of the observed action based on a vector of confounders (X_{it}) and lagged values ($A_{i,t-1}$), denoted $\Pr(A_{it}|A_{i,t-1}, X_{i,t})$; (2) Estimate the probability of the observed action based only on a vector of lagged values $A_{i,t-1}$, denoted $\Pr(A_{it}|A_{i,t-1})$; (3) Calculate the "stabilized weight" for each observation as

$$SW_i = \prod_{t=1}^{T} \frac{\Pr(A_{it}|A_{i,t-1})}{\Pr(A_{it}|A_{i,t-1}, X_{i,t})}; \tag{6}$$

(4) Calculate the causal effect of the sequence of previous actions ($A_{i,t} \forall t \in [t^*, \ldots, T-1]$) on a given outcome at time T using a *weighted* least squares regression, $y_{i,T} = \alpha + \gamma \sum_{i=t^*}^{T-1} A_{i,t} + \beta X_{i,T} + \epsilon_i$, where the weights are calculated according to Equation (6).

The critical step is to build a "correct" model of the action sequence $\overrightarrow{A}_i = A_{i,1}, \ldots, A_{i,T-1}$. Accurately modeling $\overrightarrow{A}_i$ is especially important given that the MSM approach requires a sequential ignorability assumption, which states that the weights reflect the influence of *all* relevant time-varying covariates. Thus, MSM represents another instance of an increasingly common scenario in political science research where we wish to build models that provide accurate predictions for specific outcomes, but the set of covariates and the functional forms relating them to outcomes are not of direct substantive interest.[13]

[13] For an alternative approach to developing weights for MSM models, see Imai and Ratkovic (2015).

Table 1 Covariates Included in Predictive Models of Going Negative

Variable Description	Abbrev.	GAM (Incumbent Denom.)	GAM (Noninc. Denom.)	Numer.	Trees (Denom.)
Ave. polling support for Dem. in period $t-1$	Poll_{t-1}	X	X		X
Ave. polling support for Dem. in period $t-2$	Poll_{t-2}				X
Polling for undecided option in period $t-1$	Undec_{t-1}		X		X
Polling for undecided option in period $t-2$	Undec_{t-2}				X
Democrat ran negative ads in period $t-1$	DNeg_{t-1}	X	X	X	X
Democrat ran negative ads in period $t-2$	DNeg_{t-2}	X	X	X	X
Fract. of Rep. ads that were negative in $t-1$	RNeg_{t-1}	X	X		X
Fract. of Rep. ads that were negative in $t-2$	RNeg_{t-2}	X	X		X
Number of ads run by Democrat in period $t-1$	DAds_{t-1}				X
Number of ads run by Democrat in period $t-2$	DAds_{t-2}				X
Number of ads run by Republican in period $t-1$	RAds_{t-1}				X
Number of ads run by Republican in period $t-2$	RAds_{t-2}				X
Democratic fundraising in period $t-1$	DFund_{t-1}				X
Democratic fundraising in period $t-2$	DFund_{t-2}				X
Republican fundraising in period $t-1$	RFund_{t-1}				X
Republican fundraising in period $t-2$	RFund_{t-2}				X
Fract. of Dem. ads that were negative up to $t-3$	DNegFrac_{t-3}	X	X	X	X
Fract. of Rep. ads that were negative up to $t-2$	RNegFrac_{t-2}	X			X
Fract. of Rep. ads that were negative up to $t-3$	RNegFrac_{t-3}		X		X
Campaign length	Length	X	X	X	X
Baseline polling support for Democrat	Base poll	X	X	X	X
Baseline polling for undecided category	Base Undec.	X	X	X	X
Senate or gubernatorial race	Office	X	X	X	X
Year	y2000-2006	X	X	X	X
Incumbency status for Democrat	Deminc	NA	NA	NA	X
Weeks until Election Day	Weeks	X	X	X	X

The problem, of course, is that deciding on the appropriate set of covariates, interactions, and specific functional forms for this model is an uncertain process. For instance, Blackwell (2013, 513) writes, "In order to satisfy the assumption of sequential ignorability, we must gather as many covariates as possible that might influence the decision to go negative ... and are correlated with the election outcome." Building correctly specified models using traditional methods comes with a number of serious challenges and drawbacks. To begin with, there is a concern that researchers may search the space of potential model specifications until they arrive at one that generates weights that confirm their theory. Further, researchers may feel pressured to include a large number of potential covariates in the specification, which can lead to overly complex models. Tree-based methods offer a number of advantages to building models that accurately capture response surfaces in these settings, and they do so while requiring minimal researcher intervention in terms of choosing appropriate functional forms or relevant covariates.[14]

To illustrate this, we replicate the analysis in Blackwell (2013) using two of the best-performing tree-based models in our simulation study above: BART and GBM. As a first step, we replicate the models in Blackwell (2013). Originally, the numerator in Equation (6) was estimated using a logistic regression. However, the denominator quantities in Equation (6) were estimated in separate GAM models for incumbents and nonincumbents. The predictors included in all three of these models are shown in Table 1. The full model specification, which includes several interactions and nonlinear smoothing, is shown in the supporting information.

[14]While tree models tend to be sparser than traditional models, they are not specifically designed to identify parsimonious models. For this purpose, models like the LASSO offer a better alternative.

In all, the models included in Blackwell (2013) for calculating the weights are quite complex, requiring the construction of multiple models for subsets of observations as well as specifying multiple interaction and nonlinearities. As in all modeling exercises, these and many other choices must be made by researchers and then justified to readers. However, given the large number of possible model configurations and space constraints in standard articles, not all decisions can be adequately explained. Why, for instance, should we include lagged indicators for negative campaigning from the previous *two* periods but lagged polling data from only *one* previous period? Why not interact polling data with the number of weeks left before Election Day? We expect that these choices were made based on a deep familiarity with the data. Nonetheless, we believe that this is an example of a situation where relying on tree-based methods may provide an approach to model building that is easier to implement and to justify in terms of out-of-sample predictive power.

To construct our tree models, we first specified the set of potential predictors. The factors included for the numerator in Equation (6) in Blackwell (2013) were chosen for theoretical reasons, and we follow these recommendations. For the denominator, however, we expand the list of covariates to include the full set of 26 variables. Further, we estimate only a single model for incumbent and nonincumbent Democrats, relying on the models themselves to adequately capture any differences across groups. As noted above, the treatment of interest is whether a Democratic candidate has "gone negative" in a given week. Thus, the outcome is a binary indicator for negative advertising as measured for 2,598 candidate-weeks.

After specifying predictors, the next step consists of selecting appropriate tuning parameter values.[15] For GBM models, it is best to obtain these values using some form of cross-validation. Although BART can in principle be cross-validated, its default parameters have performed very well in a wide variety of settings. Accordingly, we estimate a GBM model using the best-fitting parameters resulting from a tenfold cross-validation (viz., number of trees $M = 638$, tree depth $B = 3$).[16] We estimate the BART model using the recommended default parameter settings (see note 8).

The next step is to asses the quality of the model fit. For tree-based models, it is crucial that model fit be assessed based on out-of-sample properties since it is possible to arbitrarily improve in-sample fit by allowing for increasingly complex models. In this case, we calculate fit statistics based on, first, a (separate) tenfold cross-validation of the same data used for choosing the tuning parameters. Second, we calculated fit statistics using 10% of the original data that were randomly selected to be held back during the process of choosing tuning parameters. Fit statistics for both analyses are shown in Table 2.

The first column of Table 2 shows that the predictions from each of the models are highly correlated. Yet, the remaining columns show that the GBM and BART models, despite requiring fewer decisions from the researchers, provide more accurate predictions. Specifically, the Brier scores[17] are lower for GBM and BART. Further, the areas under the receiver operator curve (AUROC), sensitivity curve, and specificity curve are all higher for the tree models. In general, therefore, these results indicate that the tree models are to be preferred in terms of out-of-sample predictive performance.

Although the predictions for specific candidate-weeks are highly correlated, the product in Equation (6) means that even small differences in predictions can cumulatively lead to different weights.[18] These differences are of substantive consequence. Table 3 shows the estimated effect of negative campaigning in the 5 weeks leading up to the election on two-party vote share for the 144 elections in the data set. Following Blackwell (2013), we estimate separate coefficients for incumbents and nonincumbent candidates. (Full model specifications are shown in the supporting information.)

The estimated effect of negative campaigning on vote share as well as bootstrapped confidence intervals for the unweighted regression and the GAM-weighted regression are shown in the top rows of Table 3. From these competing estimates,[19] Blackwell (2013, 514) concludes that the GAM-weighted MSM uncovers effects that are at odds with previous findings, suggesting that going negative has a discernibly positive and large effect on the vote share

[15]MSM models are consistent when the weighting model is consistent. Thus, in this instance, it is important that tuning parameters are chosen so that the model for the action sequence is regularized and not overly complex.

[16]We chose a shrinkage value of 0.005 and cross-validated only to identify the optimal tree depth and tree number. For tree depth, we considered values of 3 and 5, reflecting our uncertainty as to the maximum level of interactions between covariates. For the number of trees, we considered all integer values from 1 to 2,000.

[17]Let $y^*_{i,t}$ be the predicted probability of observing candidate i running a negative ad in week t; the Brier score is then $\sum_i \sum_t (y^*_{i,t} - y_{i,t})^2$.

[18]The final stabilized weights estimated by the GAM models are only loosely correlated with the GBM (Pearson's $r = .309$) and BART ($r = .589$) models, whereas there is a stronger correlation between the GBM and BART models ($r = .738$).

[19]Our estimates differ from those in Blackwell (2013) due to slight differences in missingness because of the expanded covariate list as well as Monte Carlo error.

Table 2 Predictive Fit Statistics for Competing Models of Going Negative

			Area under		
	Correlation with GAM	Brier Score	ROC Curve	Sensitivity Curve	Specificity Curve
Within-sample tenfold cross-validation fit statistics					
GAM	1.000	0.316	0.906	0.615	0.708
GBM	0.950	0.310	0.908	0.616	0.709
BART	0.964	0.309	0.909	0.616	0.710
N = 1,035					
Out-of-sample fit statistics					
GAM	1.000	0.355	0.866	0.610	0.713
GBM	0.931	0.333	0.906	0.622	0.740
BART	0.947	0.333	0.907	0.623	0.741
N = 115					

Note: The top panel evaluates predictions from the GAM, GBM, and BART models in a tenfold cross-validation within the training set. The bottom panel evaluates the same models using observations from the test set, which was excluded for the purposes of selecting tuning parameters. Lower values for the Brier score indicate superior fit, and higher values for the areas under the receiver operator curve (ROC), the sensitivity curve, and specificity curves indicate superior fit. In all cases, the tree-based methods provide the best fit.

Table 3 Effect of an Additional Week of Negative Advertising in the Last 5 Weeks of the Campaign on Democratic Percentage of the Two-Party Vote

	Democratic Nonincumbents	Democratic Incumbents
No weights	0.466	−0.953
95% Bootstrapped CI	[−0.180, 1.096]	[−1.807, −0.198]
90% Bootstrapped CI	[−0.066, 0.937]	[−1.679, −0.280]
GAM weights	0.667	−0.601
95% Bootstrapped CI	[0.109, 1.155]	[−1.505, 0.235]
90% Bootstrapped CI	[0.240, 1.070]	[−1.357, 0.120]
GBM weights	0.537	−0.688
95% Bootstrapped CI	[−0.071, 1.169]	[−1.468, −0.019]
90% Bootstrapped CI	[0.017, 1.034]	[−1.283, −0.119]
BART weights	0.509	−0.676
95% Bootstrapped CI	[−0.048, 1.073]	[−1.422, −0.010]
90% Bootstrapped CI	[0.020, 0.969]	[−1.283, −0.086]

Note: The outcome is the Democratic candidate's share of the two-party vote. Bootstrapped confidence intervals are in brackets. The models also controlled for incumbency, cycle fixed effects, the duration of the negative advertising campaign, campaign length, baseline polling, and the quality of the opposing candidate. Full model specifications are shown in the supporting information.

of nonincumbents, while having no reliable effects for incumbents.

In contrast, the bottom rows of Table 3 show that with more accurate predictions of the action sequence estimated using tree ensembles, the MSM estimates are less divergent from previous findings. Specifically, while the effect of negative advertising for nonincumbents is also positive and reliably discernible from zero at (approximately) the less stringent 90% level, the effect sizes are roughly 20% smaller than those reported in Blackwell (2013). In turn, when considering incumbents, the tree-weighted MSM models retrieve effects that are negative, large, and reliably distinguishable from zero.

Estimating Quantities for Demographic Subgroups in Large Surveys

Although survey data can provide important insight into how different sociodemographic traits covary with

political attitudes and behaviors, it is rarely the case that surveys are deployed at the level needed to produce estimates at the lowest levels of aggregation. For instance, how does one estimate the propensity to vote of white non-Hispanic men from Oregon? While such information may be of interest to campaigns or researchers, these quantities are difficult to estimate given standard techniques.

To create such estimates, Ghitza and Gelman (2013) use a model-based approach that produces estimates for small subpopulations using aggregate survey data. Treating subpopulations as cells in a cross-tabulation of sociodemographic and geographic traits, Ghitza and Gelman (GG) use multilevel models to estimate values as a function of these covariates (and their interactions), producing predicted cell values that are then reweighted (or poststratified) using census-based population counts to produce final estimates. Specifically, GG correctly argue that the multilevel and poststratification (MRP) approach improves upon previous strategies "by modeling deeper levels of interactions and allowing for the relationship between covariates to be non-linear and even non-monotonic" (2013, 773). While the benefits are clear, important questions remain unaddressed: Which demographic variables should be included when estimating attitudes with respect to different political issues? Are all cells different enough to warrant estimation of different values, or are some interactions not relevant? In general, questions of model specification in MRP models remain very important but largely unaddressed (Warshaw and Rodden 2012).

Given that tree ensembles are particularly well suited to model precisely these types of relationships (i.e., those that are highly interactive, nonlinear, and nonmonotonic), we argue that they should provide an *even better* alternative to multilevel models when it comes to estimating quantities of interest for small subpopulation groups. Indeed, as GG clearly articulate, current implementations of MRP struggle to estimate models with saturated, high-order interactions on large data sets. As we show below, however, tree-based methods can quickly and reliably estimate models with *even deeper* interactions than MRP, letting "the data define the appropriate level of nonlinearity and interaction between covariates" (Ghitza and Gelman 2013, 773).

To compare the performance of tree models to MRP, we implement an off-the-shelf (i.e., using the default parameter definitions) estimation of BART models on the same data as GG, including the same corrections for survey weights and self-report bias. The data consist of respondents to three waves of the National Annenberg Election Survey (NAES) for 2004 (N = 43,970) and 2008 (N = 19,170) and uses respondents' state of residence, ethnicity, income, and age to model turnout and support for Senator John McCain. While both outcomes are continuous, all predictors are categorical variables that are then contrast-coded to obtain a set of 64 binary predictors. More details of the model definition are given in Ghitza and Gelman (2013) and in the supporting information.

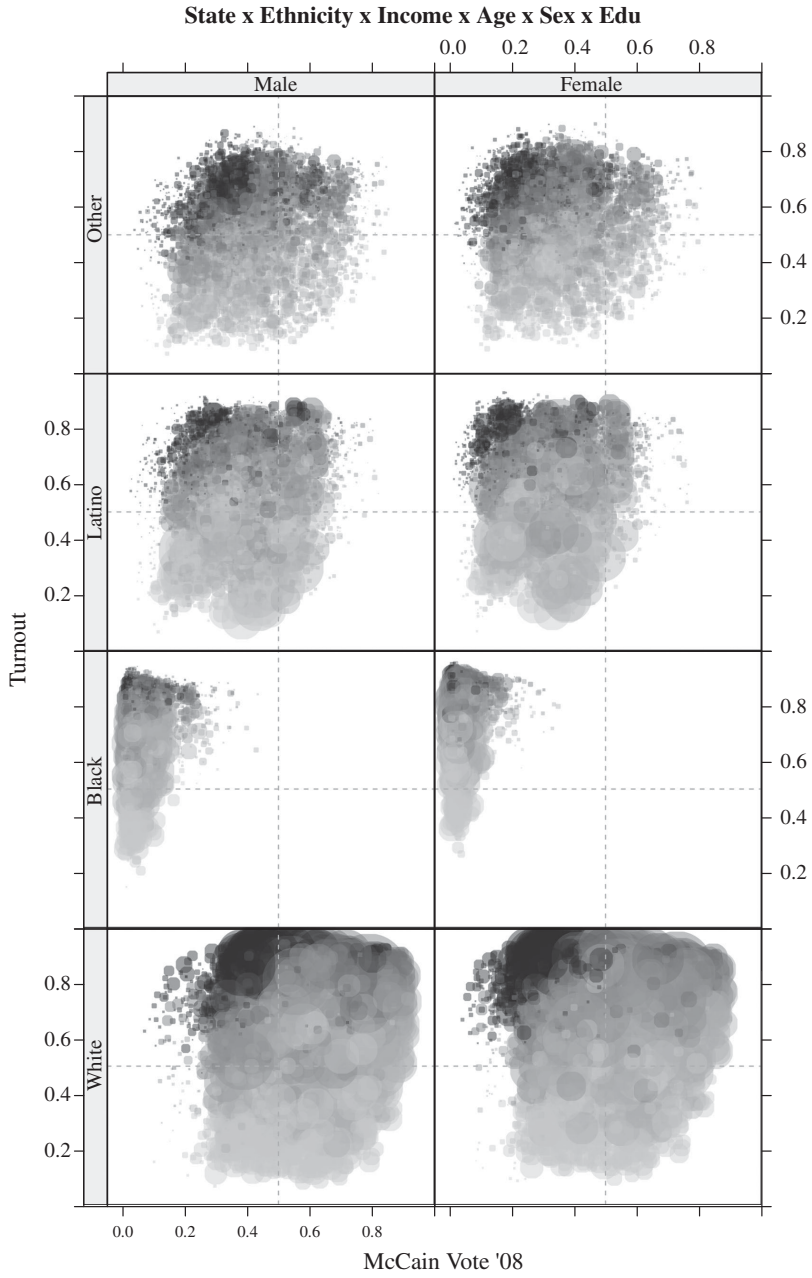
As a first step, we estimated a model using only the covariates included by GG. A simple comparison of the turnout and vote intention estimates for the subpopulations generated by the poststratified BART and MRP models using these data reveals that they are nearly identical. The correlations between the estimates generated using each method are .976 for the turnout and .975 for the vote choice.

However, the advantage of a poststratified BART is its ability to produce truly deep interaction models when the data call for them. To illustrate, consider a model that allows for interactions between state, ethnicity, income, age, sex, education, marriage status, and whether a person has children—the full array of demographic variables contained in the GG data set, which could not be included in their MRP implementation for computational reasons. BART is able to easily estimate such a model, producing interactions whenever the data support them in a way that requires minimal researcher intervention.

The results reveal *even more* nuance than GG's model originally displayed. For instance, GG focus on African American voters in North Carolina, who voted 95–5 for Obama in a state that went 50–49 for that same candidate. GG find that there was a significant difference between high-income African Americans, who voted 86–14 for Obama, and low-income African Americans, who voted 97–3 for Obama. However, further poststratification based on sex reveals that this 11-point gap in vote choice between rich and poor African American in North Carolina is primarily driven by men. The wealthiest African American *women* in the state are estimated to have gone 90% for Obama, whereas low-income African American *women* went 98% for Obama—an 8-point difference. Meanwhile, 84% of high-income African American *men* were estimated to vote for Obama, whereas 97% of low-income African American *men* did the same—a 13-point difference. Similarly, the gap between high-income and low-income African Americans in terms of turnout was estimated at 15% for women but 21% for men.

Disaggregating further by education level also reveals interesting conditional relations between demographic characteristics, turnout, and vote choice. Figure 5 shows

FIGURE 5 Poststratified BART Estimates of 2008 Turnout and Vote for McCain



Note: Size represents population size, and shade represents education level (darker shades indicate more education).

turnout and vote choice estimates for subgroups defined by state, age, and income, as well as ethnicity and sex. The estimates are shaded to represent different education levels, with darker shades representing more educated subgroups (and bubble size indicating subgroup size). The association between turnout, McCain vote, and education level is strikingly clear. In general, more highly educated people tend to turn out more often. They also tend to support Obama more, although this tendency is strongest among women. This interaction itself (viz., the clustering of more educated groups in the upper-left portion of the plots for women) is strongest among Latinos and weakest among whites. Thus, using a ensemble of trees has enabled us to estimate attitudes and preferences of even smaller subgroups at very little additional computational costs.

Conclusion

In this article, we presented tree-based methods as a promising approach for modeling large data sets in political science. We argued that they are particularly valuable in settings where one wishes to make accurate predictions in the context of a generally unknown DGP with potential nonlinearities, interactions, and many (potentially irrelevant) covariates. In the spirit of other nonparametric strategies used in the discipline, these techniques make few assumptions about DGPs or functional forms relating outcomes to predictors, and the distributional assumptions they do make are often embedded in the chosen loss function. Since researchers are increasingly confronted with larger data sets containing many observations, many possible predictors, or both, we believe that regression and classification tree models are worth considering as a more standard tool in prediction tasks.

To that end, in this article, we have presented a necessarily brief tour of some of the most common tree-based methods, complemented by three illustrations and a supplementary information appendix aimed at providing applied analysts with both insight as to the strengths and weaknesses of the various models and guidance as to how these methods can be used in practice. It is, in that sense, an invitation to adopt these methods as part of the standard repertoire of statistical tools in the discipline.

Despite their advantages, it is worth emphasizing the limitations of tree-based models that we noted in our introduction. To begin with, we reiterate that for some tasks, tree-based models are overly complex and unnecessary. Indeed, tree models are inappropriate in the context of a well-understood DGP when our aim is to test for relationships with clearly hypothesized functional forms. If the theory can reasonably be represented and tested by a parametric model, tree models are not the right tools for the job.

Likewise, tree-based methods are no replacement for good research design and rigorous theory building. While the models we have discussed allow researchers to more easily model complexities in large data sets, they do not by themselves overcome common issues of endogeneity, posttreatment bias, and the like. Even a model with high levels of out-of-sample accuracy is of limited scientific value when the modeling strategy is poorly thought out. That is, tree models are no exception to the adage, "garbage in, garbage out." Even more, empirical regularities "discovered" by this nonparametric approach are not necessarily meaningful in a theoretical sense. That some set of variables is *predictive* of an outcome does not by itself indicate that they are *causal* or even theoretically of interest.

Despite these caveats, we feel that there are many potential uses for tree-based models in political science. In our illustrations above, we demonstrated how tree models can be incorporated into standard social science tasks such as accurate measurement and causal inference. Other applications include, for instance, imputing missing data (Stekhoven and Bühlmann 2012), identifying fraudulent vote returns (Montgomery et al. 2015), and using covariates to make individual-level predictions for effectiveness of interventions (Samii, Paler, and Daly 2016). More directly, tree models may prove to be particularly valuable in the context of improving prediction—an increasingly common task in political science research. Our hope is that our discussion and illustrations will entice quantitative students of politics facing increasing demands to make sense of large amounts of social data to explore the rich possibilities offered by tree-based methods.

References

Beck, Nathaniel, and Simon Jackman. 1998. "Beyond Linearity by Default: Generalized Additive Models." *American Journal of Political Science* 42(2): 596–627.

Beck, Nathaniel, Gary King, and Langche Zeng. 2000. "Improving Quantitative Studies of International Conflict: A Conjecture." *American Political Science Review* 94(1): 21–35.

Blackwell, Matthew. 2013. "A Framework for Dynamic Causal Inference in Political Science." *American Journal of Political Science* 57(2): 504–20.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45(1): 5–32.

Breiman, Leo, Jerome Friedman, Charles J. Stone, and Richard A. Olshen. 1984. *Classification and Regression Trees.* New York: CRC Press.

Chipman, Hugh A., Edward I. George, and Robert E. McCulloch. 2010. "BART: Bayesian Additive Regression Trees." *Annals of Applied Statistics* 4(1): 266–98.

Faraway, Julian J. 2005. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models.* New York: CRC Press.

Fox, John. 2000. *Multiple and Generalized Nonparametric Regression.* In *Quantitative Applications for the Social Sciences*, Number 131. Sage.

Friedman, Jerome H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics* 29(5): 1189–1232.

Ghitza, Yair, and Andrew Gelman. 2013. "Deep Interactions with MRP: Election Turnout and Voting Patterns among Small Electoral Subgroups." *American Journal of Political Science* 57(3): 762–76.

Green, Donald P., and Holger L. Kern. 2012. "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees." *Public Opinion Quarterly* 76(3): 491–511.

Hainmueller, Jens, and Chad Hazlett. 2014. "Kernel Regularized Least Squares: Reducing Misspecification Bias with a Flexible and Interpretable Machine Learning Approach." *Political Analysis* 22(2): 143–68.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer Verlag.

Hill, Jennifer. 2012. "Bayesian Nonparametric Modeling for Causal Inference." *Journal of Computational and Graphical Statistics* 10(2): 217–40.

Hofner, Benjamin, Thomas Kneib, and Torsten Hothorn. 2014. "A Unified Framework of Constrained Regression." *Statistics and Computing* 26(*1–2*): 1–14.

Hofner, Benjamin, Andreas Mayr, Nikolay Robinzonov, and Matthias Schmid. 2014. "Model-Based Boosting in R: A Hands-On Tutorial Using the R Package mboost." *Computational Statistics* 29(1–2): 3–35.

Imai, Kosuke, and Marc Ratkovic. 2015. "Robust Estimation of Inverse Probability Weights for Marginal Structural Models." *Journal of the American Statistical Association* 110(511): 1013–23.

Imai, Kosuke, and Aaron Strauss. 2011. "Estimation of Heterogeneous Treatment Effects from Randomized Experiments, with Application to the Optimal Planning of the Get-Out-the-Vote Campaign." *Political Analysis* 19(1): 1–19.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning.* New York: Springer Verlag.

Kastellec, Jonathan P. 2010. "The Statistical Analysis of Judicial Decisions and Legal Rules with Classification Trees." *Journal of Empirical Legal Studies* 7(2): 202–30.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014, March 14. "The Parable of Google Flu: Traps in Big Data Analysis." *Science* 343: 1203–5.

Montgomery, Jacob M., Santiago Olivella, Joshua D. Potter, and Brian F. Crisp. 2015. "An Informed Forensics Approach to Detecting Vote Irregularities." *Political Analysis* 23(4): 488–505.

Muchlinski, David, David Siroky, Jingrui He, and Matthew Kocher. 2016. "Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data." *Political Analysis* 24(1): 87–103.

Samii, Cyrus, Laura Paler, and Sarah Daly. 2016. "Retrospective Causal Inference with Machine Learning Ensembles: An Application to Anti-Recidivism Policies in Colombia." *Political Analysis* 24(*4*): 434–456.

Stekhoven, Daniel J., and Peter Bühlmann. 2012. "MissForest—Non-parametric Missing Value Imputation for Mixed-Type Data." *Bioinformatics* 28(1): 112–18.

Sutton, Clifton D. 2005. "Classification and Regression Trees, Bagging, and Boosting." In *Handbook of Statistics: Data Mining and Data Visualization*, Vol. 24, ed. C. R. Rao, Edward Wegman, and Jeffrey Solka. San Diego: Elsevier, 303–29.

Warshaw, Christopher, and Jonathan Rodden. 2012. "How Should We Measure District-Level Public Opinion on Individual Issues?" *Journal of Politics* 74(1): 203–19.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

SI-1. Recursive Binary Splitting
SI-2. Discrete Outcomes
SI-3. K-Fold Cross-Validation
SI-4. Practical Considerations in Fitting, Selecting, and Interpreting Tree Models
SI-5. Additional Details for Illustrations