

ARTICLE

Not just a random forest: Using Bayesian additive regression trees to estimate willingness to pay to protect critical habitat

Lendie Follett¹ | Brian Vander Naald² | Sonja H. Kolstoe³

¹Department of Data Analytics, Drake University, Des Moines, Iowa, USA

²Department of Economics and Finance, Drake University, Des Moines, Iowa, USA

³USDA Forest Service, Pacific Northwest Research Station, Portland, Oregon, USA

Correspondence

Brian Vander Naald, Drake University, Des Moines IA, USA.

Email: brian.vandernaald@drake.edu

Abstract

We introduce a fully nonparametric estimator to estimate willingness to pay (WTP) using contingent valuation (CV) data. Traditional methods such as the probit model are used because of their roots in utility theory as well as their ease to implement. However, they are limited by their reliance on strict parametric assumptions, such as linearity and the requirement that the researcher pre-select important covariates. While these methods are widely used, they may fail to capture complex relationships, particularly in situations involving nonlinearity or interactions among variables. This paper introduces Bayesian Additive Regression Trees (BART) as a flexible, nonparametric way to estimate WTP from CV studies. Using a simulation study, we demonstrate that BART is competitive with traditional models under strict data-generating assumptions. When the underlying data-generating process is more complex, BART outperforms traditional models and even other machine-learning models. Moreover, BART automatically incorporates individual heterogeneity into estimation, making it more informative than traditional nonparametric estimators. We illustrate the breadth of the model's benefits with an empirical application estimating WTP to develop an Invasive Species Rapid Response Fund in the Hawaiian Islands. We find WTP distributional differences along the margins of gender, age, and income. We also construct willingness to pay curves out of Monte Carlo Markov Chain (MCMC) draws and calculate posterior means and credible intervals using empirical cumulative distribution functions (CDFs). We find these WTP to be lower than those from the probit. Future natural extensions of BART

make this method a compelling alternative for estimating WTP in CV experiments.

KEYWORDS

Bayesian additive regression trees (BART), contingent valuation (CV), invasive species, machine learning, nonmarket valuation, nonparametric estimators, willingness to pay (WTP)

JEL CLASSIFICATION

Q57, Q51

1 | INTRODUCTION

Nonmarket valuation is one of the primary methods used to estimate welfare values for environmental goods and services in the absence of market data. Stated preference studies represent a large proportion of nonmarket valuation work. Discrete choice experiments are useful when one is interested in valuing specific attributes of a policy or program. Contingent valuation (CV), on the other hand, is a binary choice method used when the object of valuation interest is the entire program or policy instead of its individual attributes. It has been shown to be less cognitively taxing for the respondent and therefore more likely to be incentive compatible (Carson et al., 2014). A CV task typically involves responding to a yes/no question such as “Would you pay \$A if it meant an environmental improvement of X would be made?” Each “yes” (“no”) survey response yields information about the lower (upper) bounds of a person’s willingness-to-pay (WTP) for that environmental improvement, which can be used to elicit estimates of WTP for a given population.

While the CV is supported by random utility theory, that theory is agnostic on the functional form the indirect utility should take. Possible estimators for CV data include both parametric and nonparametric estimators.¹ Estimation frequently begins with a linear probit or logit, sometimes with covariates of sociodemographic characteristics (Freeman III et al., 2014; Hanemann, 1984). While the computational burden of these models is low, the functional form and error term assumptions implicit in these models are frequently indefensible (Creel & Loomis, 1997; Crooker & Herriges, 2004). The literature has responded to these weaknesses in two different ways. The first way is to construct variations on a theme of parametric estimators. Czajkowski et al. (2024), for example, explore a wide range of parametric estimator assumptions to see which has the best model fit. Li (1996), on the other hand, relaxes the error distribution assumptions on the parametric estimator while Watanabe and Asano (2009) and Watanabe (2010) construct an estimator free of utility functional form assumptions. A second, parallel, branch of literature seized upon the weaknesses of parametric estimators to develop a suite of nonparametric estimators which do not require functional form or error distribution assumptions. The well-known Turnbull (Turnbull, 1976) and Kriström (Kriström, 1990, 1997) models are the most widely used of these. Although these models relax the assumptions of parametric models, one glaring weakness is that they are incapable of incorporating individual heterogeneity in welfare estimates. This is problematic because welfare effect heterogeneity has recently been gaining traction, both in the literature and policy (Abman et al., 2025; Faccioli et al., 2025).

Identifying heterogeneity in welfare estimates is typically accomplished in a few different ways. One method is to simply include sociodemographic and attitudinal covariates in parametric estimators such as logit or probit. Rulleau et al. (2012), for example, use binary probit to explore preference heterogeneity in WTP for beach quality. Nahuelhual et al. (2004), on the other hand, use a random

¹Parametric estimators impose error distribution and functional form assumptions whereas nonparametric estimators do not.

parameters logit to examine sociodemographic heterogeneity in CV data. Aldrich et al. (2007) segment their sample, using cluster analysis and latent class analysis to explore variability in WTP estimates for an endangered falcon and a sturgeon. These studies use parametric estimators; however, so they require the same distributional and functional form assumptions already discussed. The issue is more complicated when using nonparametric models since these do not accept covariates. Denstadli and Veisten (2020) simply split their sample between preference groups and run the same nonparametric estimator on the subsamples. Araña and León (2005) propose a mixture model composed of multiple normal distributions to allow the model greater flexibility in accommodating unobserved heterogeneity in CV data. However, these estimators do not explicitly include sociodemographic or attitudinal characteristics, limiting opportunities to model systematic heterogeneity.

A nascent branch of literature incorporating sociodemographics into nonparametric estimators has recently arisen. Zapata and Carpio (2024) develop a nonparametric model to estimate measures of WTP using double-bounded dichotomous choice data. They incorporate sociodemographics in their empirical application and conduct a simulation study that considers data with smooth functional forms. Meanwhile, Faccioli and Moeltner (2026) develop a random forest (RF) model whose ultimate goal is to estimate WTP from single-bounded dichotomous choice (SBDC) data. Their RF model is also able to incorporate sociodemographic variability and capture nonlinearities in estimation. We contribute to this budding branch of literature by introducing the application of Bayesian Additive Regression Trees (BART) to CV data. We evaluate our estimator in two ways. First, using a method similar to Crooker and Herriges (2004), we use simulation studies to demonstrate that BART is competitive with traditional parametric models under strict data-generating assumptions but outperforms them when the underlying relationships are complex. Second, we illustrate the strengths of our method with an empirical application focused on studying respondents' WTP to develop an Invasive Species Rapid Response Fund in Hawai'i.

The BART model has been used outside of economics to analyze binary outcome data (Hill et al., 2020; Linero & Yang, 2018; Tan & Roy, 2019) and is not new to the agricultural and natural resource economics field (Ramsey et al., 2026; Ramsey & Ghosh, 2025). It has been used extensively for forecasting in both the energy economics literature (Billio et al., 2024; Madadkhani & Ikonnikova, 2024; Wong et al., 2022) and macroeconomics (Clark et al., 2023; Marcellino & Pfarrhofer, 2025). This paper is the first to use the BART model on CV data, and one of only a handful of studies in the environmental and natural resource economics literature to use machine learning approaches to calculate welfare estimates. Ramsey and Bergtold (2021) use artificial neural networks (NN) to calculate marginal effects and WTP estimates using CV data, while Bravo-Moncayo et al. (2017) use NN to estimate WTP for transit options. As mentioned above, Faccioli and Moeltner (2026) use RF in the probability prediction stage of welfare estimation from SBDC data. As we show in the simulation exercise, our model outperforms both RF and NN for all types of simulated data.

Our BART model possesses many characteristics to recommend it. First, it is a flexible, machine learning approach that is able to incorporate sociodemographics, flexibly capture nonlinearities, and estimate inherent variability in WTP using SBDC data. It is not constrained by parametric assumptions, offering more flexibility to explore the relationships between response probabilities and sociodemographic factors than parametric models. Second, as our simulation study reveals, it consistently dominates the other machine learning methods in the literature on both metrics of root mean squared error (RMSE) and population bias. We attribute this improvement to the regularization induced by the prior structure, which stabilizes predictions and mitigates the tendency of other nonparametric approaches to fit idiosyncratic features of the data. Third, because BART is fully Bayesian, it naturally calculates the variability in WTP estimates using Monte Carlo Markov Chain (MCMC) samples. There is, therefore, no need to rely on asymptotic theory, which is especially valuable when sample sizes are small. Fourth, as our empirical application shows, BART is able to deal with fat tails, flat tails, and non-monotonicities in the data—three problems with which the CV literature has long wrestled. Finally, the model is ripe for possible extensions, such as using it with double-bounded dichotomous choice data and putting different priors on specific variables.

2 | BACKGROUND

To motivate our proposed methodology, and to set notation, we begin by reviewing a traditional probit framework for estimating WTP in CV studies (Haab & McConnell, 2002). We highlight its strong theoretical basis as well as its limitations. In the following sections, we introduce Bayesian and non-parametric extensions, culminating in our proposed use of BART.

2.1 | Traditional model

Suppose random variable Y_i takes on the value 1 if individual i responds “yes” to the question “would you choose to support this cause with \$A?” and 0 otherwise. x_i is defined as a vector of demographic traits for respondent i and A represents the bid amount presented to the survey respondent. Researchers often employ the model dictated by Equation (1) below.

$$\begin{aligned} Y_i &\sim \text{Bernoulli}(\pi_i) \\ \Phi^{-1}(\pi_i) &= x_i^T \beta + \alpha A_i \end{aligned} \quad (1)$$

Assume that the underlying WTP is generated via $\text{WTP} = x^T \gamma + \epsilon$ where the error terms are independent and normally distributed with mean 0 and variance σ^2 . The above model is strongly linked to utility theory in that:

$$\begin{aligned} P(Y = 1) &= \pi = P(\text{WTP} > A) = P(x^T \gamma + \epsilon > A) \\ &= P(\epsilon > A - x^T \gamma) \\ &= P\left(\frac{1}{\sigma} \epsilon \leq -\frac{1}{\sigma} A + x^T \left(\frac{1}{\sigma} \gamma\right)\right) \\ &= \Phi(-\alpha A + x^T \beta). \end{aligned} \quad (2)$$

Estimates of WTP can be recovered as $\text{WTP} = x^T \gamma$ where $\gamma = -\frac{\beta}{\alpha}$. Point estimates for WTP can be achieved by plugging in the maximum likelihood estimates. Standard errors are often calculated using the bootstrap method (Creel & Loomis, 1997; Zapata & Carpio, 2024).

2.2 | Bayesian extensions

Traditionally, the probit model is fit using maximum likelihood estimation. However, it can be argued that Bayesian methods offer several advantages over maximum likelihood estimation, including the ability to naturally estimate uncertainty, incorporate prior knowledge, and avoid reliance on asymptotic approximations for inference. The simplest Bayesian extension is to fit the probit model, as defined above, with the addition of priors being placed on the unknown parameters.

The data model shown in Equation (1), which provides the likelihood $p(Y|\alpha, \beta)$, remains unchanged. To use Bayesian estimation methods, we must also place distributions, reflecting our prior beliefs, on the unknown parameters. Following Gelman et al. (2008), we first standardize x by subtracting the means and dividing by two times the standard error of the numeric (non-binary) columns. With all x 's on the same scale, we can assign independent $p(\beta) = N(\beta|0, 2.5)$ priors for the coefficients corresponding to demographics and, independently, $p(\alpha) = N(\alpha|0, 2.5)$.²

²We also use a more diffuse prior ($N(0, 10)$) for the intercept term.

Estimation can be done by sampling from the posterior distribution of the parameters. Letting $\xi = (\alpha, \beta)^T$ represent the combined parameter vector, and ϕ represent the standard normal cumulative distribution function (CDF), the joint posterior distribution can be written as:

$$\begin{aligned} p(\alpha, \beta | Y) &\propto p(\alpha, \beta) \times p(Y | \alpha, \beta) \\ &= \frac{1}{\sqrt{(2\pi)^2 \det(2.5I_{2 \times 2})}} \exp\left(-\frac{1}{2} \xi^T (2.5I)^{-1} \xi\right) \\ &\quad \times \prod_{i=1}^n \Phi(x_i^T \beta + \alpha A)^{y_i} \left(1 - \Phi(x_i^T \beta + \alpha A)^{1-y_i}\right). \end{aligned} \quad (3)$$

This joint posterior distribution is not available in closed form but we can obtain MCMC samples from $p(\alpha, \beta | Y)$ using a Gibbs sampler. We can use many different software packages (R, SAS Python, etc.) to obtain the set $\{(\beta^{(r)}, \alpha^{(r)})\}_{r=1}^R$, where R is the number of samples taken from the posterior.

An important advantage of the Bayesian framework is that we can obtain full posterior distributions of WTP for each individual (or, for a given demographic vector x). Mathematically, the collection of posterior samples allows us to examine $f(\text{WTP}(x) | \{Y_i\}_{i=1}^n) = f(-x^T \frac{\beta}{\alpha} | \{Y_i\}_{i=1}^n)$ for any vector x . These individual-level posterior samples can be aggregated across the whole sample to produce full WTP curve distributions because we effectively have R estimates of $\text{WTP}(x)$. This also allows us to quantify our uncertainty at every point along the curve, which can be a useful way to present the results because it maintains the ease of interpretation while adding transparency.

Probit regressions are used frequently in literature, though the Bayesian implementation is less common. They are both simple to implement (Bayesian or frequentist) and based on utility theory. Also, in the context of Bayesian estimation, neither the distributional assumptions nor the variance of the posterior distribution of WTP depends on asymptotic theory. However, the linear form and the need for the researcher to choose important inputs increase the chances of model misspecification. Because of these shortcomings, we turn our attention to nonparametric alternatives, starting with classification trees.

3 | A NON-PARAMETRIC, BAYESIAN APPROACH

3.1 | Classification trees

We will now briefly explain classification trees, the building blocks of our approach, along with their strengths and weaknesses. In the context of CVs, a classification tree partitions the demographic and pricing space through binary splits, aiming to create subsamples that are homogeneous in terms of survey responses (i.e., all “yes” or all “no” within subsamples). The process starts with the full survey data set, at which point we refer to it as the “root node”. This root node is then split (based on either a demographic factor or price) into two child nodes. For example, a split may be defined by numeric bid amount: observations with bid amount $\leq \$500$ would flow into child node 1 while those with bid amount $> \$500$ would be diverted to child node 2. This process continues, with child nodes being further split into more child nodes, until we reach the terminal nodes, at which point predictions are given. Splits are determined by optimizing the variable and the split point which maximizes the reduction in entropy. The algorithm will search over all variables and splitting points within each variable. More details on the building of classification trees can be found in Therneau et al. (1997).

After the general structure of the tree is built, a new observation can traverse down from the root node, through the internal nodes, until it arrives at one of the b terminal nodes, where it gets a prediction based on the training observations which live in that node. Unlike traditional linear models, like the ones described above (frequentist or Bayesian), classification trees automatically incorporate interactions between variables without requiring the researcher to pre-specify them. Figure 1 shows

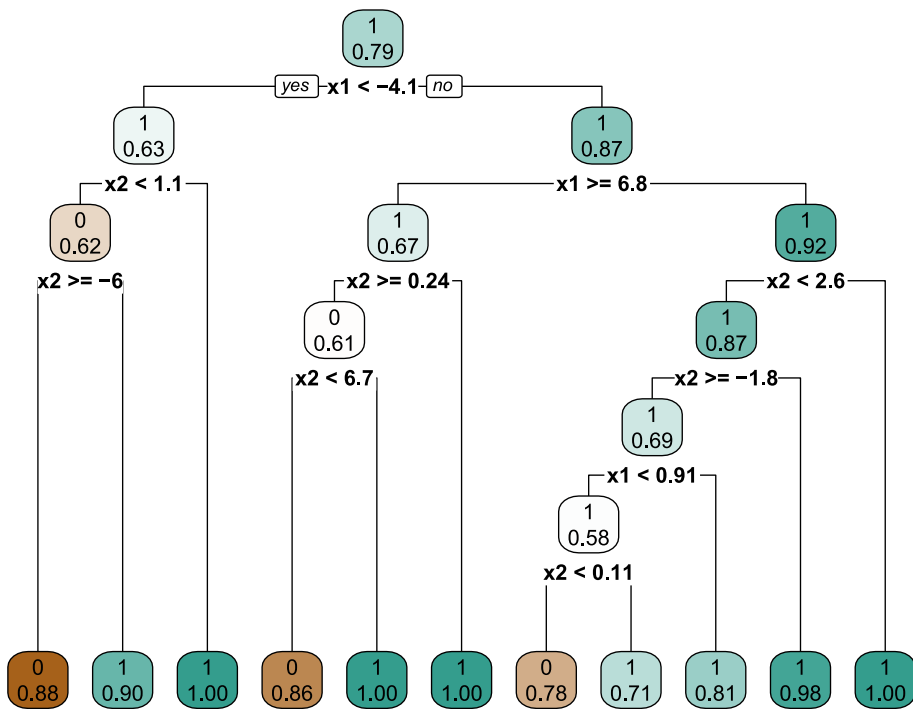


FIGURE 1 Illustration of a tree.

an example of a tree that predicts a binary outcome based on just two covariates, x_1, x_2 . Whenever more than one covariate is involved in a “branch” of a tree, this represents an interaction. For example, we see that $x_2 \geq -6$ is an important predictor of the binary response; however, this is only true when $x_1 < -4.1$ (the first splitting rule of the tree).

Figure 2 illustrates how the structural differences between trees and linear regression models result in different predictions, even when applied to the same data (and even when the researcher is able to correctly identify the presence of interactions). On the left, the probit predictions are a smooth probit function of x_1 and x_2 while the classification tree operates by “binning” (splitting, repeatedly) x_1 and x_2 . This allows for more flexible prediction functions because classification trees are able to capture multidimensional valleys and other non-monotonic structural relationships.

Mathematically, a single tree can be represented as $g(x_i | T, M)$, where T denotes the tree structure (splitting rules) and $M = \{\mu_1, \dots, \mu_B\}$ are the predictions at the B terminal nodes. The variables involved in T are chosen automatically based on minimizing entropy, which contrasts with traditional regression models such as Model (1).

While trees are flexible and able to fit training datasets to a high degree of precision, this trait makes them susceptible to overfitting. Ensemble methods, such as random forests (Breiman, 2001), can be used to combat these issues by combining multiple trees to improve predictions.

3.2 | Ensemble methods

A RF is a widely used ensemble learning method designed to mitigate the overfitting tendencies of individual classification trees. This method is well liked for its ease of implementation³ and also

³There are several R packages that can be used to fit forests (Chen & Guestrin, 2016; Liaw & Wiener, 2002; Wright & Ziegler, 2017), and so forth.

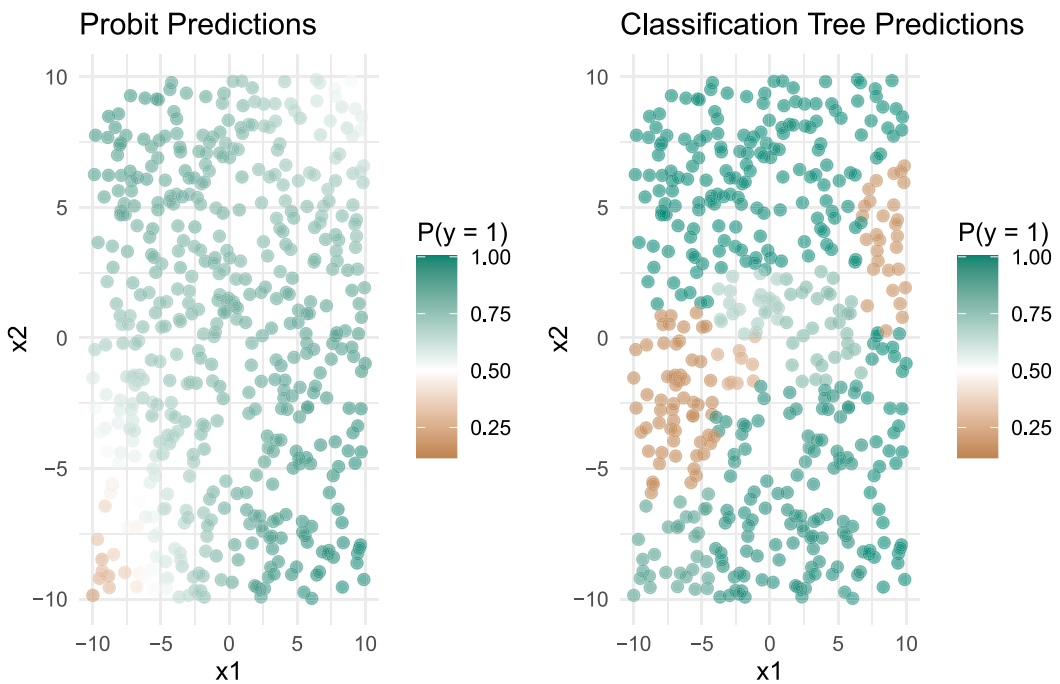


FIGURE 2 Predictive patterns—Probit versus classification trees. Illustration of prediction patterns for a logistic regression (left) and classification tree (right) when the data generating process involves interactions and non-linearities.

because it does not require extensive parameter tuning in order to have competitive predictive performance. As the name would suggest, an RF is a collection of many trees. This method's strength lies not in the number of trees in the forest but in the diversity of the trees in the forest. RFs introduce diversity among their trees in two ways. First, the trees are fit on different sets of training data (bagging—see Breiman (1996)). Specifically, for a training dataset of size n , each tree is grown on a random sample of n rows, sampled with replacement. Second, the trees are fit on different random sub-samples of the covariate space. This diversity de-correlates the trees, making for a more robust predictive model when the trees are considered in aggregate.

BART (Chipman et al., 2007, 2010), also a multi-tree-based method, further addresses the shortcomings of single trees by assuming priors on the parameters involved in the tree building and node predictions. These priors shrink the trees, keeping each one simple and preventing any single tree from explaining too much variability. This tree characterization is unlike RFs, which construct a large number of deep trees for optimal performance. A second difference is that RFs construct the trees independent of each other; BART, on the other hand, constructs the trees sequentially, allowing the model to share information across trees. A feature of BART—a Bayesian model—is that the model output is posterior distributions of probabilities while, for random forests, the output by default will be point estimates. Chipman et al. (2010) demonstrated that BART can outperform other methods, including RF and neural networks, in out-of-sample prediction problems. An important feature of BART is that, similar to random forests, it is able to achieve this level of competitive performance with very little tuning involved. In the next section, we detail BART and discuss how posterior distributions from the BART model can be translated to WTP. One practical differentiator between the models mentioned thus far is the computational expense associated with each of them. For most reasonably sized datasets, a probit model or classification tree will fit almost instantly. Other machine learning models such as RFs, neural networks, and BART have more complexity in their estimation process and can take longer. In the simulation study, we present average run-times for each of the models tested.

3.3 | BART-based WTP estimates

BART model

As mentioned above, single classification trees are flexible but they are prone to overfitting. BART addresses this by summing over many shallow trees, each regularized by informative priors. This Bayesian ensemble approach can yield more stable predictions. In what follows, we describe BART and its implementation for estimating WTP. Our main contribution lies not only in using BART for prediction, but in transforming its probability-scaled outputs into meaningful measures of WTP.

Equation (4) describes the BART framework for classification problems. Line 1 describes the distributional assumptions for the binary response Y_i . Line 2 links the inputs— x_i (attributes seen by individual i) and $A_{j[i]}$ (where $j[i]$ is the index pertaining to the cost seen by individual i)—to the probability that a respondent says “yes.” In line 3, we take $g(x_i, A_{j[i]}|T_t, M_t)$ to represent a function (the i th tree) that inputs x, A and outputs a prediction μ_{bt} , which is the b th node parameter of the t th tree. Finally, we take $G()$ to be the function that inputs x, A and sums up the T trees.

$$\begin{aligned} Y_i &\sim \text{Bernoulli}(\pi_i) \\ \pi_i &= P(Y_i = 1|x_i, A_{j[i]}) = \Phi(G(x_i, A_{j[i]})) \\ G(x_i, A_{j[i]}) &= \sum_{t=1}^T g(x_i, A_{j[i]}|T_t, M_t) \end{aligned} \quad (4)$$

Based on this model, the expected (probit of the) probability that an individual responds “yes” is equal to the sum of all the terminal node parameters that $g()$ assigns based on values in x, A . In building the sum-of-trees model, the node parameters from the first tree ($\{\mu_{11}, \dots, \mu_{B^1-1}\}$) will explain some proportion of the variability in the responses; ($\{\mu_{12}, \dots, \mu_{B^2-2}\}$) will explain a proportion of the remaining variability, and so on.

In this model specification, there are no “coefficients” to place priors on as in the previous model. However, priors must still be placed on the unknown tree structures and node parameters, T_t and M_t , respectively. Trees are kept small by specifying a prior probability that a node at depth d is non-terminal, rather than allowing the tree to grow aiming to minimize entropy, as in RFs. For some $\alpha \in (0, 1)$ and $\beta \in [0, \infty)$, the prior probability distribution can be described as:

$$p(\text{node at depth } d \text{ is non-terminal}) = \alpha(1 + d)^{-\beta}. \quad (5)$$

By default, it is common to assume $\alpha = 0.95$ and $\beta = 2$. As d increases, the probability that it is non-terminal decreases rapidly. Likewise, the terminal node parameters also receive shrinkage priors. More details can be found in Chipman et al. (2010).

In the context of our model, the likelihood is given by

$$p(y|T, M) = \prod_i \Phi(G(x_i, A_{j[i]}))^{y_i} (1 - \Phi(G(x_i, A_{j[i]})))^{1-y_i} \quad (6)$$

and the corresponding joint prior is

$$p(T, M) = \prod_t p(T_t) p(M_t|T_t) = \prod_t \left[p(T_t) \prod_b p(\mu_{bt}|T_t) \right], \quad (7)$$

leading to a joint posterior distribution that is proportional to $p(y|T, M)p(T, M)$. An overall Gibbs and Metropolis-within-Gibbs sampler is used to jointly sample the tree and leaf parameters. We

direct the reader to Chipman et al. (2010) and McCulloch et al. (2018) for details on sampling. R package BART (Sparapani et al., 2021) performs sampling.

Estimating posterior distribution of WTP

While the BART model defined above has been used extensively in the literature as a predictive model, our contribution comes from the treatment of the output of the model; specifically, $G(x_i, A)$, the BART predictions of the probit transformation of the probability that a person with traits x_i would say “yes” to a bid of \$ A . WTP estimates are ultimately obtained through a two-step process. First, by fitting the predictive model (in our case, BART) on a training set, and second, the transformation of model predictions to a WTP scale.

Drawing on Parzen (1960)’s finding regarding CDFs of non-negative random variables, expected WTP can be thought of as the area under the curve defined by the probability of “yes” responses with respect to \$ A , the bid amount (illustrated in Figure 3). Let A_0 be the lowest bound considered, and A_K the maximum amount a person would reasonably pay. For an individual with demographics x , let $P(Y = 1|A, x)$ be the probability of a “yes” response to bid \$ A .

Once a BART model is trained on survey data, we can use it to estimate probabilities across an array of bid amounts ($A_0, \dots, A_K$), holding constant the demographics of a given individual, ultimately producing a model-based estimate of how that individual’s probability of saying “yes” would change as bid amount increases.⁴ Then, using model output $\{\Phi(G(x, A_j)) : j = 0, \dots, K\}$,⁵ we approximate

$$\text{WTP}(x) = \int_{A_0}^{A_K} P(Y = 1|A, x_i) dA$$

for each individual i . Much like we would see with other nonparametric methods that rely on the CDF result, this process yields a conservative estimate of WTP, that is, any remaining right-tail mass beyond this bound is excluded, implying a downward bias.

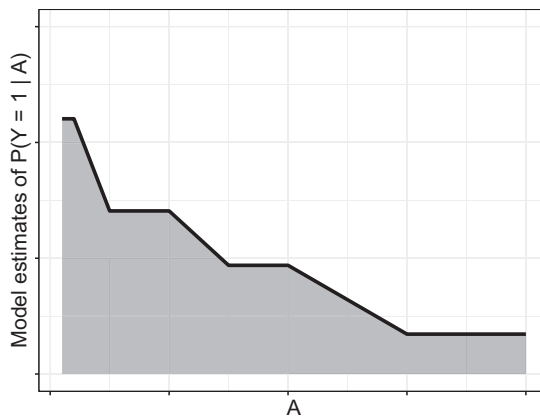


FIGURE 3 WTP as the area under a curve.

⁴Additional detail about training can be found in the appendix.

⁵Assume $\Phi(G(x, A_0)) = 1$ and $\Phi(G(x, A_K)) = 0$.

$WTP(x) = \int_{A_0}^{A_K} P(Y=1|A, x_i) dA$ can be approximated numerically over the grid of bid amounts; we use quadrature methods,⁶ following the approach of Ramsey and Bergtold (2021). In doing so, we assume a monotonic, decreasing relationship between the two inputs. In cases of non-monotonicity within a curve, we use smoothing algorithms to flatten.⁷

It is important to note that the original dataset is *labeled*, containing the observed yes/no responses, whereas the long-form dataset is *unlabeled* and used solely for prediction. In this long-form representation, each surveyed individual, along with their demographic characteristics, is paired with each unique bid amount ($A_0, \dots, A_K$) (i.e., `CV_donate_bid`). This construction allows the pre-trained model to generate predicted probabilities across the full range of bid amounts, facilitating the estimation of curves as illustrated in Figure 3, for all demographic profiles. The integration procedure, which uses the raw MCMC samples of the BART model, is used throughout both the forthcoming simulation study and empirical application to compute expected WTP.

Alternative treatments are available to convert the probabilistic output of the BART model to WTP. Specifically, Turnbull (Haab & McConnell, 2002), Krström (Krström, 1990), and the Krström adjustment (Lewis et al., 2024) are model-free WTP estimation techniques that can be applied, given a sequence of bids and corresponding probabilities. In these cases, similar logic regarding monotonicity holds. An exploration comparing the various second-stage approaches is beyond the scope of this research.

Estimating WTP variability

In Section 2, we illustrated how we can obtain full posterior distributions of an individual's WTP without resorting to asymptotic assumptions. This ability is maintained as we pivot to using BART; variability in the WTP estimates is introduced through the trees and root node parameters in the MCMC algorithm. MCMC estimation results in R posterior draws of tree sequences $\{T_t^{(r)} : t=1, \dots, T\}_{r=1}^R$ and corresponding terminal nodes $\{M_t^{(r)} : t=1, \dots, T\}_{r=1}^R$ and, therefore $\{P^{(r)}(Y=1|A, x) : j=1, \dots, K\}_{r=1}^R$ for a given x . Thus, the curves under which we integrate (using standard quadrature rules) vary from sample to sample across the MCMC chains as we approximate the integral $WTP^{(r)}(x) = \int_{A_0}^{A_K} P^{(r)}(Y=1|A, x) dA$ for $r=1, \dots, R$.

Noting that we typically take R to be a larger number such as 5000 or 10,000, we conceptually illustrate this in Figure 4 with just 5 MCMC draws and their resulting curves. That is, for any given demographic vector x , we can obtain posterior distributions of WTP by using the MCMC-implied variability on the estimated sequence of $P^{(r)}(Y=1|A, x)$. Expected WTP is computed by numerically integrating the predicted probability curve over the bid grid for each posterior draw. It is important to note that the resulting distributions are not based on asymptotic theory and, in general, do not make any parametric assumptions. In addition, the variability is estimated based on a single run of the model; this is in contrast to bootstrapping methods, which require repeated runs of the model on random samples of the training data.

We consider two ways of displaying the output of the proposed procedure. First, a general overview of the sample-weighted population's willingness to pay. For a given sample, we can obtain $\{WTP^{(r)}(x_i) : i=1, \dots, n\}_{r=1}^R$. That is, each fixed MCMC sample, r , yields an entire distribution of WTP for the sample population. These vectors, along with application of the empirical CDF, can be used to create R WTP curves (one for each MCMC sample) resulting in a distribution of WTP curves. The second way to study the output is to consider how WTP is affected by sociodemographic

⁶We use the `integrate()` and `approxfun()` functions in R.

⁷We use the `pava` (pool adjacent violators algorithm) function from the `Iso` package in R.

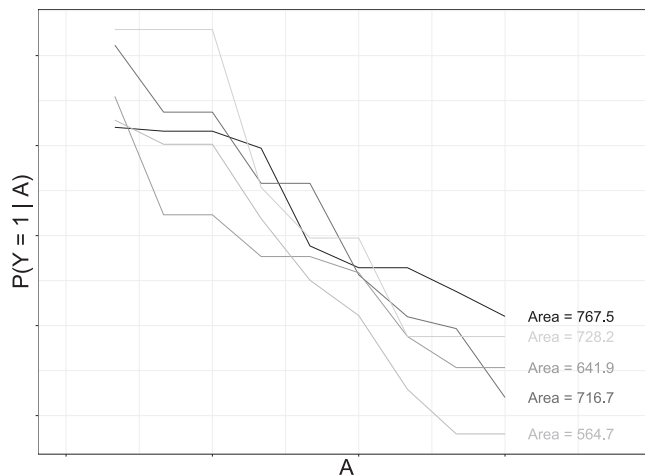


FIGURE 4 Predicted probabilities using BART. Illustrating the BART estimator with variability resulting from MCMC. For a given demographic vector x , five MCMC draws resulting in five areas under the curves.

factors. This requires calculating an estimated WTP as above and taking an exploratory approach to visualize the relationship between sociodemographic factors and WTP.

3.4 | Simulation study

Setup

To evaluate the quality of BART-based WTP point estimates, we conduct a simulation study across a variety of data-generating mechanisms. These scenarios vary in complexity and alignment with traditional modeling assumptions. We include both traditional and Bayesian probit models as baselines. We also follow other studies in semi-nonparametric estimators of WTP (Ramsey & Bergtold, 2021) by incorporating a neural network. We use a grid search, along with cross-validation tools, to tune the neural network's nodes (varying between 2 and 8) and decay (varying between 0.01 and 0.5). This is done once within each repetition, so that each run of the neural network is tuned and optimized for prediction.⁸ Thus, averages for the neural network will be an aggregate of prediction performances from NNs of potentially different parameter values. Finally, to benchmark BART against another tree-based algorithm, we also include a RF. Much like BART, RFs are somewhat robust to tuning, and we use default settings.⁹

Each of the aforementioned competing models will be compared to a tuned BART model and an untuned BART model that is run using default hyperparameters. For the tuned BART model, we use a grid search, along with cross-validation tools, to tune the number of trees (trying 50, 100, 200) and k (trying 1, 2, 3). k is a parameter that controls the amount of shrinkage applied to the terminal node mean parameters; a higher k will shrink the terminal means harder toward an overall mean while a smaller k will allow individual nodes (and, as a result, trees) to have a larger impact on overall predictions. This is again done once within each repetition, so each run of the tuned BART is

⁸This level of tuning may not be practical for some applied settings.

⁹However, while the default output of RFs are numbers between 0 and 1, they are not technically probabilities associated with the positive event. That is, they are not always well-calibrated. Thus, we additionally attempt to calibrate the RF probabilities but find it does not improve performance. A detailed discussion of probability calibration is beyond the scope of this paper. More details can be found in Niculescu-Mizil and Caruana (2005), which notes that the quality of calibration of an RF can be context-dependent.

customized to the sample. As with the neural network, note that this allows for differing parameters across simulation runs, even within the same simulation settings.

For the Bayesian models (BART and Bayesian probit), we follow standard procedures by discarding the first half of each chain as burn-in; the latter half of each chain is used for prediction and inference (Gelman et al., 2013, pp. 282–283).

We simulate data according to four scenarios, as shown in equation set (8). While the first (Scenario 1) mimics the underlying assumptions of the probit WTP estimator, we introduce three additional simulation scenarios that feature more nuanced structures in the model's systematic component. The four simulation scenarios can be written as follows:

$$\begin{aligned}
 \text{Scenario 1: } WTP_i &= \beta_0 + \beta_1 * x_{1i} + \epsilon_i \\
 \text{Scenario 2: } WTP_i &= \beta_0 + \beta_1 * I(x_{1i} < 0.5) + \epsilon_i \\
 \text{Scenario 3: } WTP_i &= \beta_0 + 2 * \left(10 \sin(\pi x_{1i} x_{2i}) + 20(x_{3i} - 0.5)^2 + 10x_{4i} + 5x_{5i} \right) + \epsilon_i \\
 \text{Scenario 4: } WTP_i &= \beta_0 + \beta_1 * z_{1i} + \beta_2 * z_{2i} + \beta_3 * z_{3i} + \beta_4 * z_{4i} + \beta_5 * z_{5i} + \\
 &\quad \beta_6 * z_{1i} * z_{2i} + \beta_7 * z_{1i} * z_{3i} + \epsilon_i.
 \end{aligned} \tag{8}$$

Scenario 2 involves a simple step-function when the single covariate crosses a threshold. Scenario 3 is modified from the Friedman function (Friedman, 1991), commonly used to test predictive models in machine learning. The first two scenarios involve a single covariate, x_1 , while the third and fourth involve five covariates. The fourth scenario is different in that it involves discretized covariates, some of which are involved in interactions. That is, each z_{pi} is effectively simulated independently from a Bernoulli distribution by transforming the original x variable, for example, $z_{1i} = I(x_{1i} > 0.5)$.

In each case, the coefficients are calibrated so that, for average covariate values, the mean WTP is approximately \$100 and ranges between approximately \$25 and \$150. We let $\epsilon \sim N(0, \sigma_W^2)$, and allow σ_W to vary. We randomly generate 100 datasets of $n + 500$ rows from each scenario, and train our models on n rows, leaving 500 for test/evaluation. We are interested in evaluating performance on training sample sizes (n) varying from small to large. For the first three scenarios, each covariate, $x_{pi} : p = 1, \dots, 5$, is simulated from $\text{beta}(a, b)$ where we fix $a = 1$ and allow b to vary, introducing asymmetry into the covariates (thus, for scenario 4, the covariates are simulated from a Bernoulli(p) distribution with $p \neq 0.5$).

In real applications, the researcher would likely have multiple covariates ($x_1, \dots, x_p$), but would not know which of them are truly associated with WTP. Thus, we run the simulation study again, simulating the data in the exact same manner, but we simulate and serve $x_1, \dots, x_{10}$ as inputs to train the models. The purpose of this is to mimic real-world applications, rather than limiting the input to the relevant covariates involved in simulation. For example, for scenario 1, only one of the 10 input predictors (10%) actually influence WTP while in scenario 2, only five of the 10 input predictors (50%) influence WTP. Thus, we introduce the sparsity into the true coefficient vector for the second set of simulations. Table 1 lists all of the values tested for each parameter. Each unique parameter combination ($3 \times 2 \times 2 \times 2 = 24$), across four data-generating scenarios (96 total cases), is evaluated using 100 independent simulation replications. We provide pseudo code describing the data generating process in more detail in Appendix B.

With machine learning approaches, there are two stages to estimating WTP from CV data. First, the model is trained on survey data and produces probabilities of a “yes” response when applied to the test (hold-out) data. Second, those probabilities are translated to the WTP scale. We compare models at both stages; in each case, using the 500 observation hold-out test sample to calculate performance metrics. That is, the following is all based on data the models have not seen. To illustrate stage 1 model performance, we calculate the correlation between the true probability that the (simulated) respondent says “yes” and the modeled probability. For stage 2, we measure performance in

TABLE 1 Summary of parameters varied in simulation study.

Parameter	Values
n	250, 500, 1250
σ	7, 15
Sparsity	TRUE, FALSE
b	1, 3

Note: Parameters include sample size (n), error variance (σ_W^2), sparsity of the true model, and shape parameter(b) for covariate asymmetry.

two ways: (1) via the RMSE between the true WTP and the model-estimated WTP and (2) by calculating the average absolute bias between the true mean WTP of the test sample and the model-implied mean WTP. (1) reflects how accurately a model recovers individual-level WTP, which incorporates both bias and precision, while (2) measures the presence of systematic issues with the implied mean WTP, which is central to the practical implementation of CV studies. To illustrate practical significance, we divide both RMSE and bias by the true mean WTP of the distribution to provide a relative error term. In both cases, we use paired t -tests to summarize the statistical significance of the differences in model performance.¹⁰

Simulation results

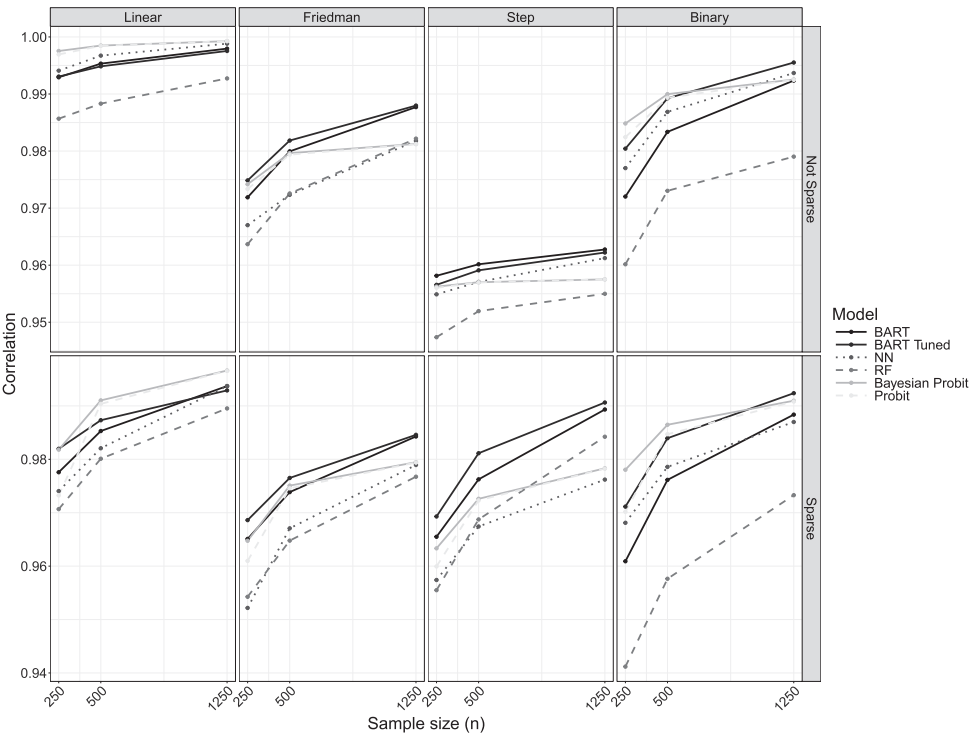
The simulation study provides valuable insights into the performance of our BART model compared with other nonparametric machine learning models and the traditional probit model in estimating WTP from SBDC data. We measure the predictive accuracy of our models in three ways: correlation between true and predicted probabilities, RMSE, and population bias. In each of the three Figures 5–7, the top set of facets illustrates results when the data are simulated with smaller error variance; the bottom when the error variance is larger. Within each set of facets, the top row shows results when the data are not sparse; the bottom when the data are sparse. Each column represents data generated according to the functions described in scenarios 1–4 above: Linear, Friedman, Step, and Binary. Each facet shows results as the sample size increases from left to right.

Figure 5 displays the average correlation between the true probabilities and the predicted probabilities from each of the models. In the linear scenario, probit is a clear winner with almost perfect correlation in some cases, though BART is not far behind and is always within 0.05. Because BART does not impose a fixed functional form on $P(Y = 1|A, x)$, linear relationships can be considered a special case of its functional capabilities. Further, for more complex data-generating-scenarios, BART often has a higher correlation.

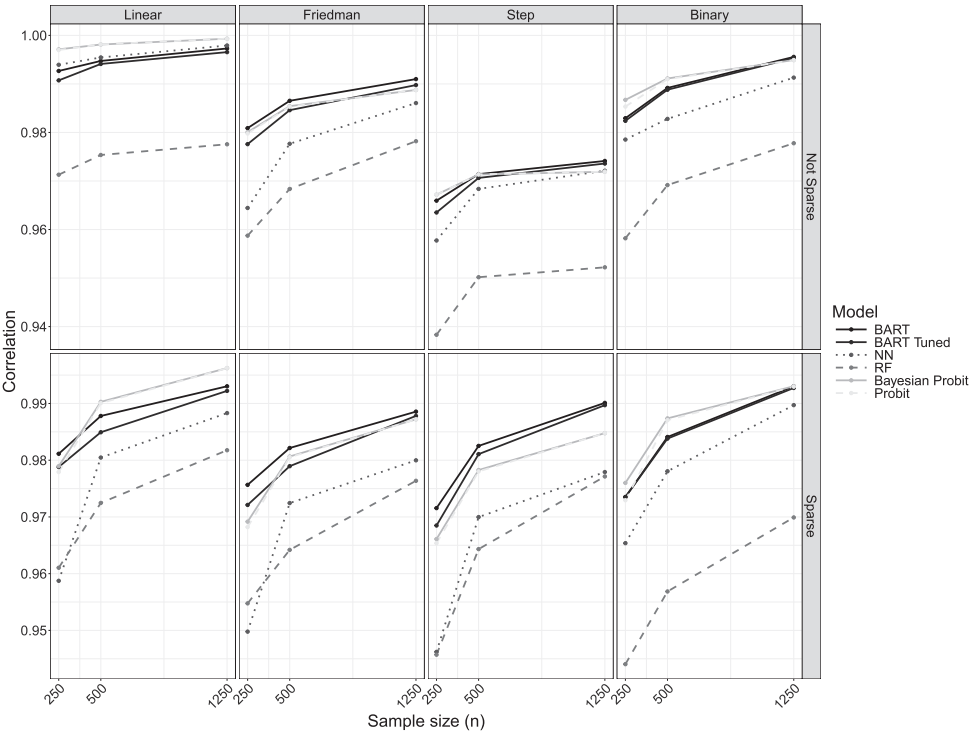
Figure 6 displays calculated RMSE values for each model.¹¹ This measure captures how well a model recovers individual-level WTP values; a lower value is better. Here we can see that when the (linear) assumptions of the probit model hold, the probit model performs well relative to more complex machine learning approaches while maintaining a substantially lower computational cost. The machine learning models typically achieve RMSE values within 0.5% of the probit model, but their additional complexity and computational demands make them less appealing when the data is well-behaved. This reinforces the idea that when the true relationship aligns with the probit model's assumptions, its efficiency and interpretability make it a strong default choice. However, when the functional form of $P(Y = 1|A, x)$ (probability of a yes) deviates from the restrictive form dictated by the probit model—such as when the underlying WTP functional form follows nonlinear, nonmonotonic, or discontinuous patterns, or when interactions are introduced—the machine learning models demonstrate clear advantages in terms of RMSE, particularly as training sample sizes increase. While the probit model sees some improvement when moving from small to moderate

¹⁰Code to generate simulation study can be found at <https://github.com/LendieFollett/Not-Just-A-Random-Forest>.

¹¹Results shown for symmetric covariates; asymmetry scenarios can be found in appendix.



(a) $\sigma = 7$



(b) $\sigma = 15$

FIGURE 5 Legend on next page.

sample sizes, its strict parametric form limits its adaptability in general. In contrast, BART and NN continue to improve with larger datasets, learning the underlying structure of the data more effectively. Across different data-generating processes, BART (tuned and default) and NN tend to emerge as the most accurate models in terms of RMSE, with BART performing particularly well in sparse settings where the relevant predictors are not “known” or pre-specified. BART provides strong predictive accuracy without the need for tuning, a critical step for the NN. This suggests that in more complex or poorly understood WTP estimation problems, BART is a robust choice.

Figure 7 illustrates the average absolute bias between the true mean WTP of the test sample and the model-implied mean WTP for each model. With the exception of RFs, all models exhibit bias that diminishes toward zero as the sample size increases. Both the probit and Bayesian probit models perform particularly well in terms of bias when their underlying assumptions are met, reflecting the advantages of correctly specified parametric structure. However, when covariates are asymmetric, those same models exhibit noticeably larger bias, as shown in the appendix, highlighting their sensitivity to distributional assumptions. In contrast, BART and the other semi-nonparametric approaches display consistently small negative bias across all four data generation scenarios. We expect this type of behavior given the quadrature-based second stage of WTP estimation, which yields conservative estimates by construction.

Across the simulation scenarios that were examined, we have seen little evidence that tuning the BART algorithm results in significant gains in predictive accuracy. This is consistent with the already-documented trait of the BART model: default BART settings tend to perform well. Still, the authors hesitate to recommend skipping tuning in real-world applications for which the true nature of the data-generating process is unknown.

To examine whether the differences between BART and probit shown in Figures 6 and 7 are statistically significant, we perform paired two-sided *t*-tests within each simulation scenario. The results of those tests are shown in Table 2. Panel (a) displays results of those tests for bias. While probit is favored in 61 scenarios, only 14 (22%) of those were statistically significant. Panel (b) shows the results of those same tests in terms of RMSE. BART is favored in 70 out of the 96 scenarios and 57 (59%) of those were by a statistically significant margin. The upshot is that while there is a tradeoff between bias and RMSE between the two models, the significance of increased bias attached to BART is less than the gains in predictive accuracy observed in RMSE. This suggests that BART strikes a favorable balance between accuracy and bias in more complex or less well-understood settings.

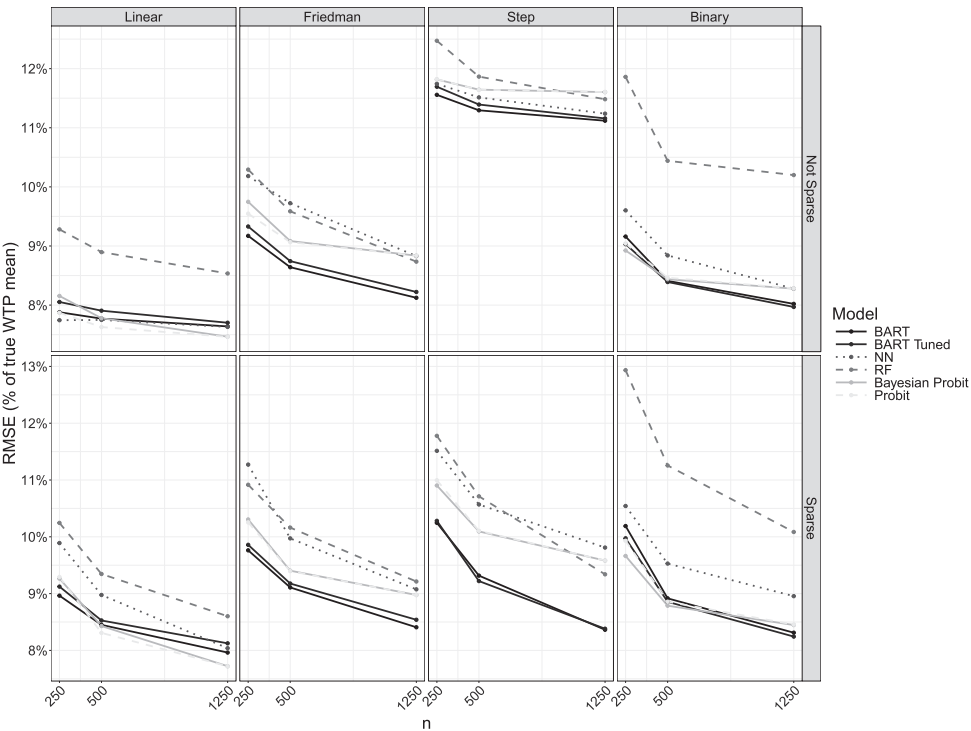
Lastly, we explore differences in computational time between models. While BART and other machine learning models are able to learn complex patterns found in data, they do so at the cost of computational time. This is especially true for Bayesian methods which rely on MCMC. How prohibitive the computational cost is will depend on the size of the data and complexity of the model. Table 3 shows the average run times of the various models for each training sample size considered. For NN models, this time includes the time it takes to tune over the number of nodes and decay. As expected, the Bayesian methods take considerably longer. However, the observed run times (20.85 s, at the longest, on average) are unlikely to be problematic for a typical researcher.

3.5 | Empirical application

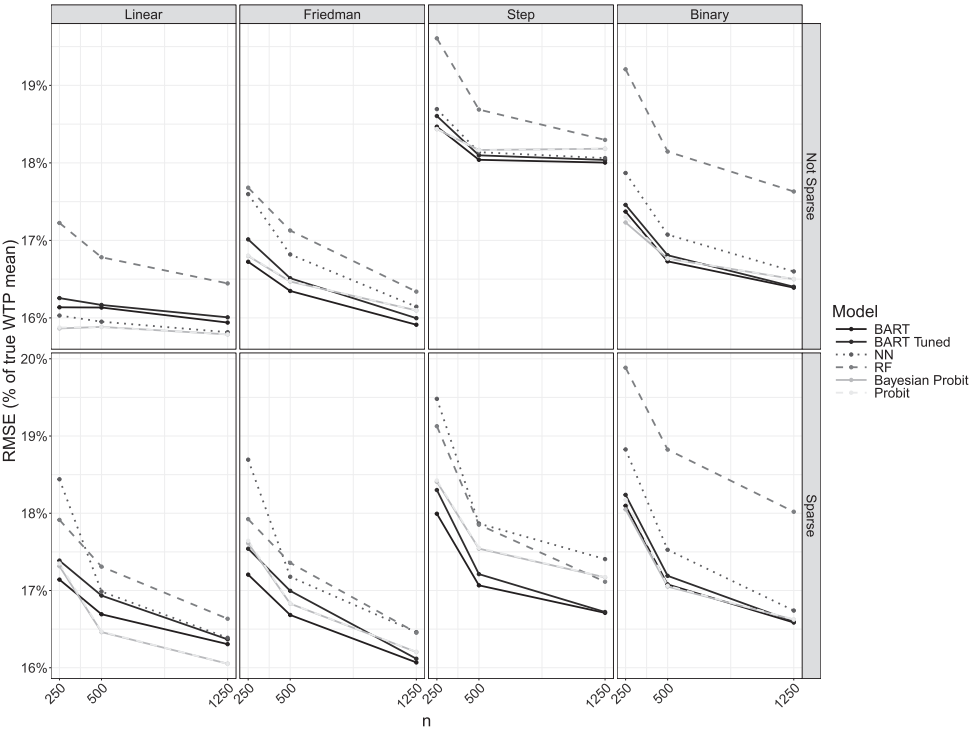
Data

The empirical application relies on CV data and sociodemographic information collected as part of a larger survey which also included a discrete choice experiment. Our study site is Waikamoi Preserve

FIGURE 5 Average correlation between true probability of saying “yes” and model-estimated probabilities. Shown for varying sample sizes (*x*-axis) and sparsity of coefficient vectors (vertical facets) and data generating scenarios (horizontal facets). Shown with error variance $\sigma = 7$ (top) and $\sigma = 15$ (bottom). This is shown only for symmetric covariates; the results for asymmetrically-distributed covariates are similar and shown in the appendix.



(a) $\sigma = 7$



(b) $\sigma = 15$

FIGURE 6 Legend on next page.

located on the Hawaiian island of Maui. The Nature Conservancy (TNC) manages the Waikamoi Preserve as the largest privately owned nature preserve in the state. Waikamoi Preserve's vast 'Ōhi'a-dominant forest is the last stronghold of 63 species of rare plants and 10 species of birds, five of them endangered. Since 2014, Rapid 'Ōhi'a Death (ROD) has threatened 'Ōhi'a trees, which is also a keystone species in other native Hawaiian wet forest ecosystems. ROD is mainly spread by humans through their gear. Once a tree shows symptoms of ROD, it usually dies within a few weeks.

ROD has not yet been detected in Waikamoi Preserve. However, because of the speed with which it kills and the growing risk at the time, TNC was interested in developing a system to identify and eradicate ROD-infected trees as quickly as possible.¹² Accomplishing these tasks is costly. For example, identification might involve the use of a drone to fly patterns over Waikamoi Preserve to identify potentially infected 'Ōhi'a trees. Once identified, infected trees are very expensive to remove because the terrain in Waikamoi Preserve is mountainous. Extirpation must be accomplished using a combination of helicopters and hiking.

To fund the system described above, TNC considered the creation of an Invasive Species Rapid Response Fund. Since this fund did not exist at the time of the survey, we developed a CV question to test how much the public would be willing to pay for it. Respondents were told that the fund would be used to detect and quickly mitigate incipient invasive species, including ROD. The CV question was a single-shot donation developed using the best practices presented in Johnston et al. (2017) and adapted from the work of Richardson and Lewis (2022). In it, respondents were asked if they would donate a fixed amount to help TNC develop an Invasive Species Rapid Response Fund. Bid amounts presented include: \$25, \$50, \$100, \$250, \$500, \$750, \$1000, \$1500, \$2000.¹³ We gathered sociodemographic information at the end of the survey. Table 4 contains the proportion and number of respondents who said yes to each bid amount.

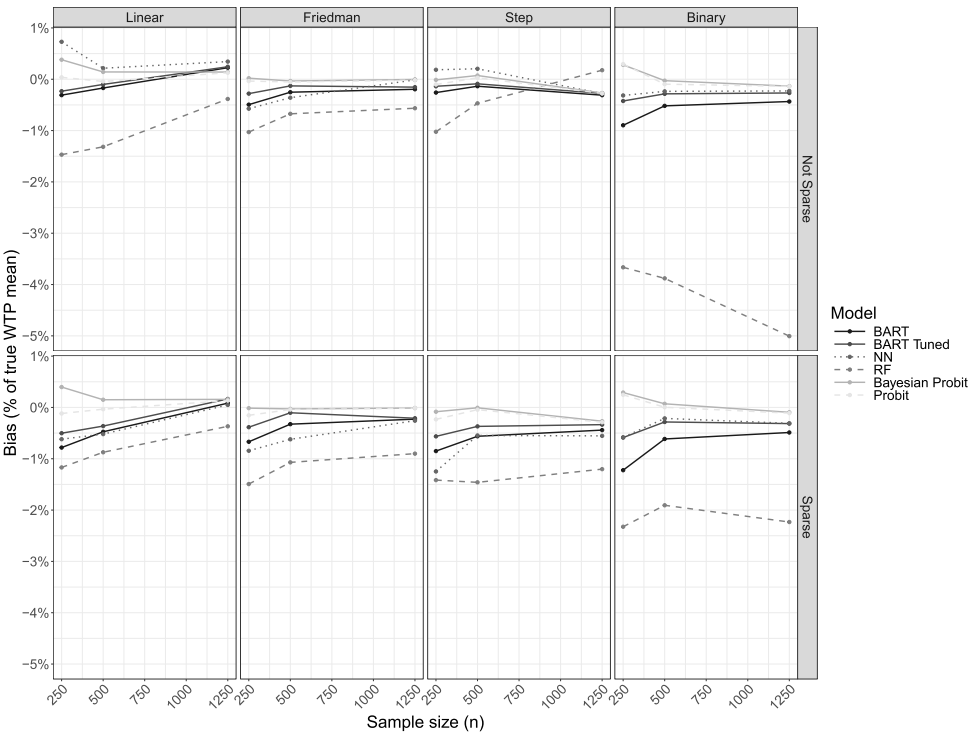
We took several steps to address incentive compatibility and reduce hypothetical bias. First, respondents were presented with an oath at the beginning of the survey in which they were required to pledge to give "thoughtful and honest" responses to all questions in the survey in order to continue (Jacquemet et al., 2017). Second, the CV task itself included a provision point mechanism (Poe et al., 2002), whose purpose is to attenuate inflated hypothetical "yes" bids. In it, respondents were told the fund would not be established if \$50,000 was not raised. Finally, following the CV task, we asked two follow-up questions. First, respondents were asked how certain they were in their responses (Blomquist et al., 2009; Penn & Hu, 2023). Second, respondents who rejected the bid amount were asked to select their reason for doing so. Based on these, we were able to classify respondents into those who said no out of protest and those who said no because they thought the program would not be consequential.

Some respondents may answer "no" to a CV question because they are "protesting" an aspect of the valuation exercise. Table 5 shows the proportion of each bid rejection by justification. Those who rejected the bid were asked why they rejected the bid. Those who indicated *I doubt that the Invasive Species Rapid Response Fund's work would succeed in protecting the species and biodiversity in Waikamoi from invasive species and diseases* or *I do not think there is a real risk of invasive species or*

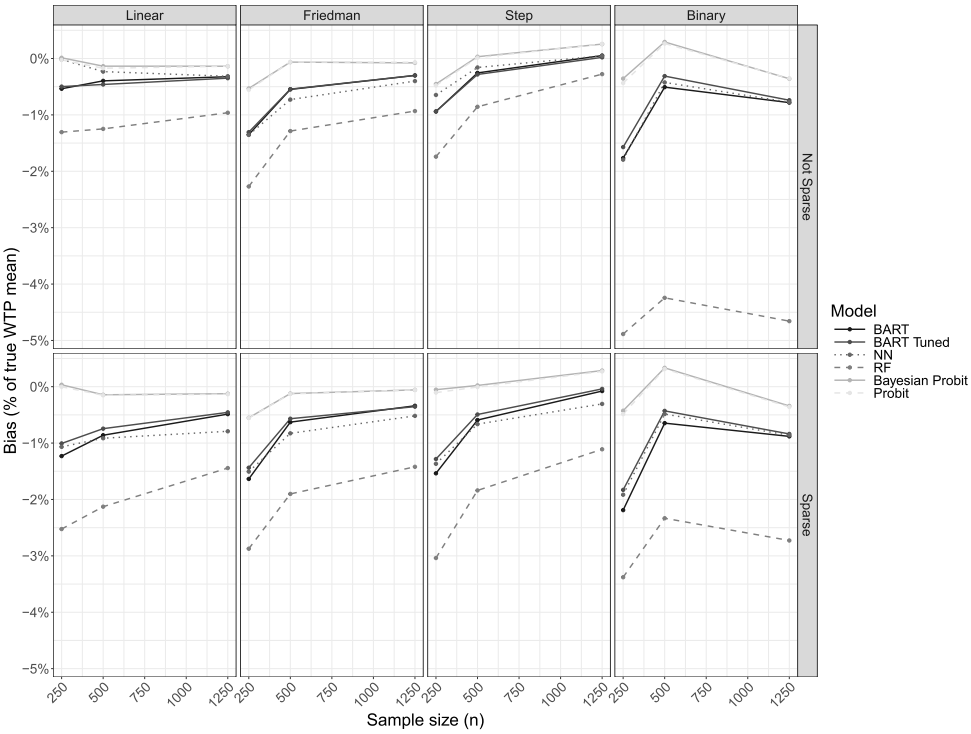
FIGURE 6 Average RMSE. Average RMSE for varying sample sizes (x-axis) and symmetry of covariate distributions (vertical facets) and data generating scenarios (horizontal facets). Shown with error variance $\sigma = 7$ (top) and $\sigma = 15$ (bottom). This is shown only for symmetric covariates; the results for asymmetrically-distributed covariates are similar and shown in the appendix.

¹²At the onset of the study, ROD had only been found on the islands of Hawai'i and Kauai'i. In July 2019, 5 trees were detected on Maui, subsequently destroyed and since then there have been no additional positive ROD detections on the island—cite <https://cms.ctahr.hawaii.edu/rod/>.

¹³The bid amounts of \$750, \$1500 and \$2000 were added after the pretest.



(a) $\sigma = 7$



(b) $\sigma = 15$

FIGURE 7 Legend on next page.

TABLE 2 Counts of simulation scenarios by *t*-test significance and direction of advantage.

(a) Bias			
	BART	Probit	Sum
Not significant	31.00	47.00	78.00
Significant	4.00	14.00	18.00
Sum	35.00	61.00	96.00
(b) RMSE			
	BART	Probit	Sum
Not significant	13.00	13.00	26.00
Significant	57.00	13.00	70.00
Sum	70.00	26.00	96.00

TABLE 3 Average run times of various models over sample sizes (in seconds).

<i>n</i>	Bart	NN	rf	Probit	Bprobit
250	5.92	0.46	0.17	0.01	0.98
500	20.91	0.80	0.36	0.01	3.44
1000	15.54	1.35	0.76	0.01	6.06
1500	20.85	2.14	1.13	0.02	5.44

TABLE 4 Frequency of survey responses to the contingent valuation question.

Bid	% yes	<i>n</i>
\$25	48%	115
\$50	41%	111
\$100	46%	119
\$250	39%	114
\$500	39%	122
\$750	31%	118
\$1000	34%	117
\$1500	28%	120
\$2000	32%	114

Note: Full sample.

diseases were classified as protest responses. Those who left similar comments in the “other” space were also classified as protesters and omitted from analysis. We also restricted the sample in another way. To assess whether they found the survey consequential, respondents were asked on a 3-point likert scale from “not at all” to “fully” *to what extent do you believe TNC will pay attention to people’s preferences, as expressed in this study, as they make their decisions about future hiking programs?* Responses from those who answered “not at all” were classified as inconsequential and omitted from the analysis.

FIGURE 7 Average absolute bias. Average absolute bias in terms of estimating mean WTP for varying sample sizes (*x*-axis) and symmetry of covariate distributions (vertical facets) and data generating scenarios (horizontal facets). Shown with error variance $\sigma = 7$ (top) and $\sigma = 15$ (bottom). This is shown only for symmetric covariates; the results for asymmetrically-distributed covariates are similar and shown in the appendix.

We conducted the survey during the summer of 2020 using Qualtrics’ panel of respondents. We finished the survey collection period with 1050 responses. We removed from the estimation sample protest votes and those who believed the survey to not be consequential. Of the 54 respondents removed, 33% of them were female and 40% were younger than 35 years of age. An equal proportion (30%) were aged 35–54 and greater than 55 years of age. 63% of the excluded sample is employed while 9% are retired. 52% have at least a bachelor’s degree while 26% achieved a graduate degree. 15% of the excluded sample earns less than \$25,0000 per year; 42% earns between \$25,000 and \$74,999 per year, and 39% earns over \$75,000 year. The restricted sample contains 996 observations. Slightly more than half of these respondents are female (52%). We have a roughly uniform distribution of ages: 31% are 34 or under; 34% are between 35 and 54; and 35% of the restricted sample is 55 and over. 58% of respondents are employed and 21% are retired. 57% hold a bachelor’s degree or higher, while 25% hold a graduate degree. 14% of this restricted sample earn less than \$25,000 per year; 39% earn between \$25,000 and \$74,999 per year, and 45% earn greater than \$75,000 per year. Figure 8 illustrates our data. They suffer from non-monotonicity (Haab & McConnell, 2002), fat tails (Lewis et al., 2024), and flat tails (Lewis et al., 2024). Together, these artifacts of the data lead to large WTP standard errors and imprecise WTP estimation when using traditional estimators.

TABLE 5 Reasons for a “No” response to the contingent valuation question.

Reason	% responding
I cannot afford to donate anything	66%
I cannot afford to donate that much, but I would pay a different amount (please specify)	13%
I doubt that the Invasive Species Rapid Response Fund’s work would succeed in protecting the species and biodiversity in Waikamoi from invasive species and diseases	2%
I do not think there is a real risk of invasive species or diseases	1%
There are too many other good causes that need my donations more urgently	12%
Other	6%

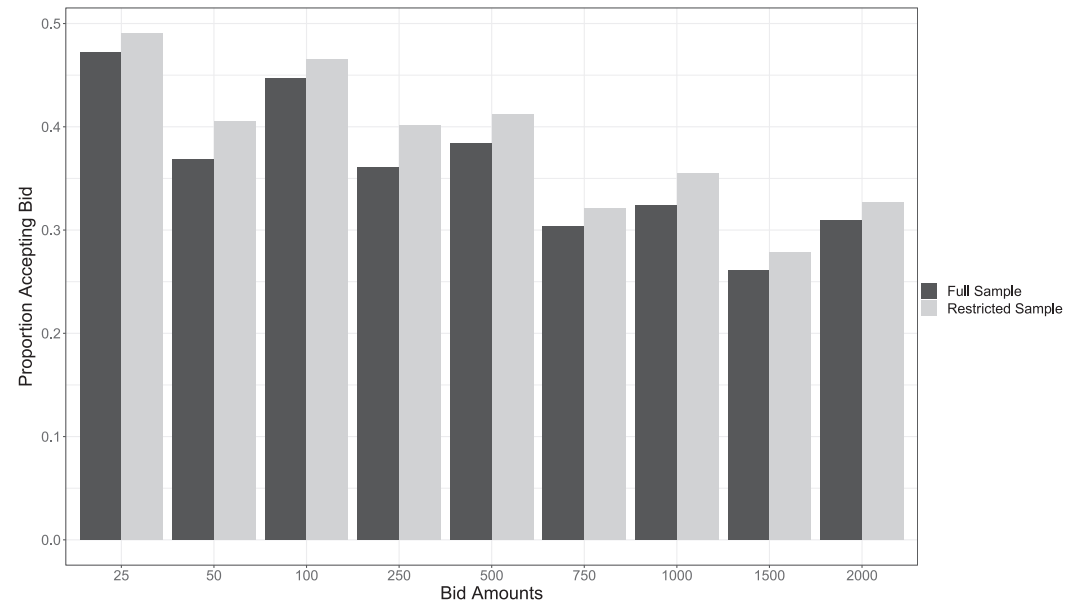


FIGURE 8 Percentage of yes-responses to the contingent valuation question. By bid amount, with and without protest and inconsequential responses.

Estimation results

For illustration, and because the results of our simulation study favor BART, we discuss WTP estimates from that model here, along with results from the Bayesian Probit for comparison. To estimate the BART model, we run 10,000 iterations, discarding the first 5000. We use default tuning parameters. In the second stage, we again use quadrature on the predicted probabilities to estimate WTP. To estimate the Probit model, we use the `rstanarm` package in R (Goodrich et al., 2024). Specifically, we run 4 chains of 5000 iterations of the MCMC algorithm, taking the final 1000 iterations from each, resulting in $R = 4,000$ iterations available for inference. In traditional CV applications, WTP is recovered via Equation (2), by taking the negative ratio of the appropriate coefficients. While this works well in situations where the data is well-behaved, it is not uncommon for irrational WTP values to result from this method. For our empirical data, we find that over half of individual-level WTP values are estimated to be negative using the traditional procedure. This motivates us to treat the probabilistic output from the probit model the same way we treat the BART output. We use quadrature to estimate the area under the curve, and, thus, WTP values. This constrains estimates to stay within rational bounds.

Figure 9 presents WTP estimates from the BART model and traditional probit model, illustrating the impact of demographics, but also revealing the effect of the modeling approach itself on the resulting estimates. Both BART and probit estimate a negative relationship between age and WTP and a positive relationship between household income and WTP, which is consistent with what we

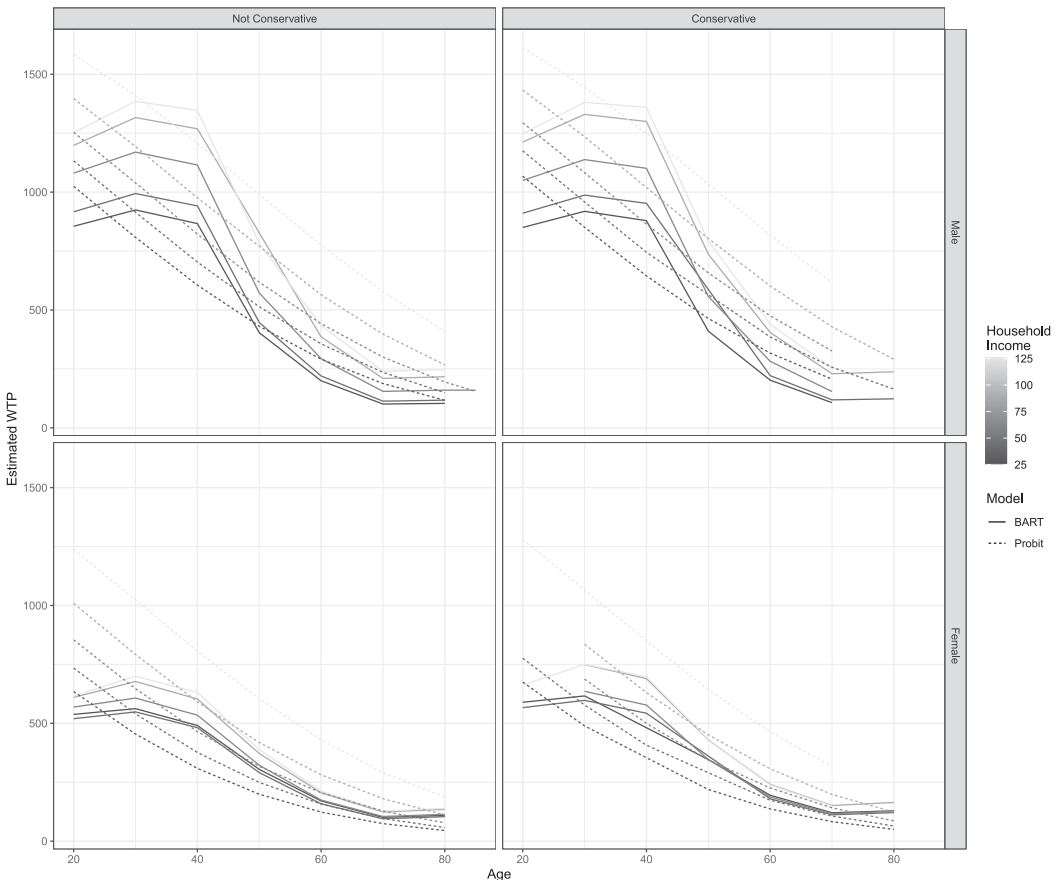


FIGURE 9 WTP curves. Estimated WTP based on probit (dashed) and BART (solid) across levels of age (x -axis) and household income (color). Shown separately for all combinations of political ideology and gender.

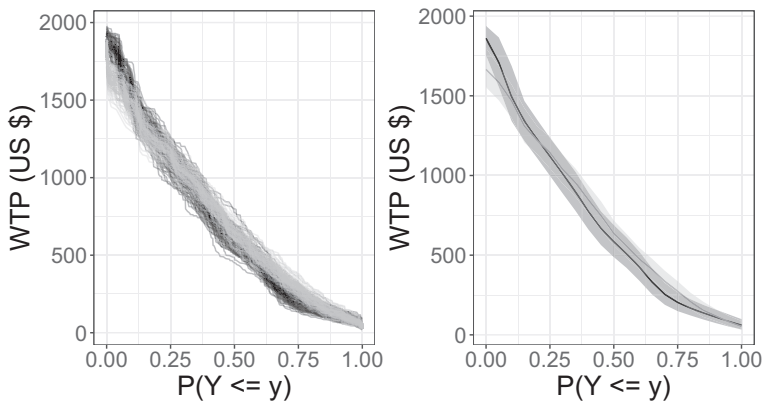


FIGURE 10 WTP envelopes. WTP curves for BART (black) and Bayesian probit (gray) with 90% credible intervals. Left panel shows willingness to pay curves resulting from individual MCMC iterations while the right panel summarizes at each probability threshold.

would expect. But BART picks up more nuance in these relationships. For age, BART suggests a slight increase in WTP between roughly ages 20 and 30 before the expected decline begins. Probit, constrained by linearity unless additional polynomial terms are specified, simply models a steady downward trajectory, effectively implying that individuals in their 20s have the highest WTP. We see the same kind of contrast with household income. Both models show higher WTP with higher income, but BART identifies a noticeably stronger income effect for males than for females (something probit cannot detect without an interaction being pre-specified by the researcher).

The most interesting disagreements between the two models appear in the tails of the covariates; in the youngest and oldest respondents and in the lowest- and highest-income households. In these regions, BART's flexibility allows it to capture nonlinearities and interactions that probit misses, highlighting how model choice can influence the resulting WTP estimates, especially where the underlying relationships are more complex.

Another powerful benefit of our BART model is its ability to quantify uncertainty. One way to illustrate this ability is shown in Figure 10, which shows what we call our WTP envelope. Using empirical CDFs, we calculate a WTP curve for each MCMC iteration $r = 1, \dots, 5,000$. These curves are illustrated in the left panel of Figure 10. Then, at each unique probability threshold, we calculate not only the posterior median, but also the 5% and 95% quantiles (90% credible interval); these quantiles are summarized in the right panel. Suppose we are interested in the price at which 50% of respondents would be willing to pay for this fund. Our best estimate of this price is \$587, but there is a 90% posterior probability that this number lies between \$490 and \$687. This can be compared to the Bayesian probit-based curve, for which the best WTP estimate is \$638 (\$562, \$720). BART suggests an average WTP that is \$51, or 8%, smaller. This particular curve could be of great interest to, for example, a nonprofit organization attempting to calculate how much revenue they might collect at a given donation amount. BART is able to make WTP projections without imposing distributional assumptions on the functional form. These WTP projections are lower than those implied by the probit model. As demonstrated in the simulation study, when probit assumptions are violated, these models can be sensitive to functional form misspecification.

4 | CONCLUSION

In this study, we introduce the BART model to estimate WTP using CV data. Using a simulation exercise, we show how BART compares with well-known parametric models such as probit when

the data-generating process is simple. When the data-generation process becomes more complicated, BART performs more consistently, compared not only to probit but to other non-parametric models such as RFs and NNs. Using a proposed invasive species mitigation fund in Hawai'i as an empirical example, we show that BART generates reasonable individual-level WTP estimates with individual-level precision unavailable from traditional parametric estimators. Moreover, BART automatically produces heterogeneity in estimates based on a full suite of sociodemographic and attitudinal covariates. The ability to capture such heterogeneity may be useful when conducting benefit-cost and distributional analyses.

Our proposed methodology comes with limitations. First, we rely on smoothing to ensure that the probability of acceptance decreases with increasing bid amount. This is an intervention that parametric models would not require. Second, BART does require slightly more computing time than a simpler tree-based method such as RFs. In addition, the second-stage WTP computations are slightly more involved when applied to MCMC samples rather than point estimates.

Our model paves the way for a number of different extensions. One such extension involves quantifying individual-level error variability. While our proposed method quantifies *parameter* and *demographic-based* variability, we do not fully quantify *random individual-level* error variability. Second, we quantify systematic variability in WTP attributable to demographic differences but we are unable to identify error variability on the WTP scale without imposing constraints. One could construct a semi-parametric BART model by enforcing linearity on the bid input. Obtaining good estimates of not only WTP, but all sources of the error variability of WTP, would allow researchers to more realistically represent the entire distribution of WTP of a sample.

Overall, the results of this study show promise for a new set of estimators that can estimate WTP with more precision. Unlike traditional approaches, nonparametric models follow a data-driven approach by not requiring the researcher to pre-specify key functional forms relating WTP to demographics. As computational resources become increasingly available, and Bayesian methodology becomes more accessible, methods like BART offer a practical alternative for estimating WTP.

ACKNOWLEDGMENTS

This research was funded by The Nature Conservancy. The findings and conclusions are those of the authors, and should not be construed to represent any official USDA or U.S. Government determination or policy. All remaining errors are the responsibility of the authors.

REFERENCES

- Abman, R., E. C. Edwards, and D. Hernandez-Cortes. 2025. "Water, Dust, and Environmental Justice: The Case of Agricultural Water Diversions." *American Journal of Agricultural Economics* 107: 1041–58.
- Aldrich, G. A., K. M. Grimsrud, J. A. Thacher, and M. J. Kotchen. 2007. "Relating Environmental Attitudes and Contingent Values: How Robust Are Methods for Identifying Preference Heterogeneity?" *Environmental and Resource Economics* 37: 757–775.
- Araña, J. E., and C. J. León. 2005. "Flexible Mixture Distribution Modeling of Dichotomous Choice Contingent Valuation with Heterogeneity." *Journal of Environmental Economics and Management* 50(1): 170–188.
- Billio, M., R. Casarin, M. Costola, and V. Veggente. 2024. "Learning from Experts: Energy Efficiency in Residential Buildings." *Energy Economics* 136: 107650.
- Blomquist, G. C., K. Blumenschein, and M. Johannesson. 2009. "Eliciting Willingness to Pay without Bias Using Follow-up Certainty Statements: Comparisons between Probably/Definitely and a 10-Point Certainty Scale." *Environmental and Resource Economics* 43: 473–502.
- Bravo-Moncayo, L., J. L. Naranjo, I. P. García, and R. Mosquera. 2017. "Neural Based Contingent Valuation of Road Traffic Noise." *Transportation Research Part D: Transport and Environment* 50: 26–39.
- Breiman, L. 1996. "Bagging Predictors." *Machine Learning* 24: 123–140.
- Breiman, L. 2001. "Random forests." *Machine Learning* 45: 5–32.
- Carson, R. T., T. Groves, and J. A. List. 2014. "Consequentiality: A Theoretical and Experimental Exploration of a Single Binary Choice." *Journal of the Association of Environmental and Resource Economists* 1: 171–207.
- Chen, T., and C. Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16*. 785–794. New York, NY, USA: ACM.

- Chipman, H. A., E. I. George, and R. E. McCulloch. 2007. "Bayesian ensemble learning." In *Advances in Neural Information Processing Systems* 265–272. Cambridge, MA: MIT Press.
- Chipman, H. A., E. I. George, and R. E. McCulloch. 2010. "BART: Bayesian Additive Regression Trees." *Annals of Applied Statistics* 4: 266–298.
- Clark, T. E., F. Huber, G. Koop, M. Marcellino, and M. Pfarrhofer. 2023. "Tail Forecasting with Multivariate Bayesian Additive Regression Trees." *International Economic Review* 64: 979–1022.
- Creel, M., and J. Loomis. 1997. "Semi-Nonparametric Distribution-Free Dichotomous Choice Contingent Valuation." *Journal of Environmental Economics and Management* 32(3): 341–358.
- Crooker, J. R., and J. A. Herriges. 2004. "Parametric and Semi-Nonparametric Estimation of Willingness-to-Pay in the Dichotomous Choice Contingent Valuation Framework." *Environmental and Resource Economics* 27: 451–480.
- Czajkowski, M., E. Zawojka, N. Meade, R. S. da Motta, M. Welsh, and R. A. Ortiz. 2024. "On the Inference about a Willingness-to-Pay Distribution Using Contingent Valuation Data." *Ecological Economics* 222: 108207.
- Denstadli, J. M., and K. Veisten. 2020. "The Flight Is Valuable Regardless of the Carbon Tax Scheme: A Case Study of Norwegian Leisure Air Travelers." *Tourism Management* 81: 104150.
- Faccioli, M., and K. Moeltner. 2026. "Random Forests for Dichotomous Choice Contingent Valuation." *Journal of Environmental Economics and Management* 137: 103303.
- Faccioli, M., D. M. Tingley, M. C. Mancini, and I. J. Bateman. 2025. "Who Should Benefit from Environmental Policies? Social Preferences and Nonmarket Values for the Distribution of Environmental Improvements." *American Journal of Agricultural Economics* 107: 725–751.
- Freeman, A. M., III, J. A. Herriges, and C. L. Kling. 2014. *The Measurement of Environmental and Resource Values: Theory and Methods*. New York, NY: Routledge.
- Friedman, J. H. 1991. "Multivariate adaptive regression splines." *Annals of Statistics* 19: 1–141.
- Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2013. *Bayesian Data Analysis*, 3rd ed. Boca Raton, FL: Chapman & Hall CRC.
- Gelman, A., A. Jakulin, M. G. Pittau, and Y. S. Su. 2008. "A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models." *Annals of Applied Statistics* 2: 1360–83.
- Goodrich, B., J. Gabry, I. Ali, and S. Brilleman. 2024. "rstanarm: Bayesian applied regression modeling via Stan." R package version 2.32.1
- Haab, T. C., and K. E. McConnell. 2002. "Valuing Environmental and Natural Resources: The Econometrics of Non-market Valuation." In *Valuing Environmental and Natural Resources*. Northampton, MA: Edward Elgar Publishing.
- Hanemann, W. M. 1984. "Welfare Evaluations in Contingent Valuation Experiments with Discrete Responses." *American Journal of Agricultural Economics* 66: 332–341.
- Hill, J., A. Linero, and J. Murray. 2020. "Bayesian Additive Regression Trees: A Review and Look Forward." *Annual Review of Statistics and Its Application* 7: 251–278.
- Jacquemet, N., A. James, S. Luchini, and J. F. Shogren. 2017. "Referenda under oath." *Environmental and Resource Economics* 67: 479–504.
- Johnston, R. J., K. J. Boyle, W. Adamowicz, J. Bennett, R. Brouwer, T. A. Cameron, W. M. Hanemann, et al. 2017. "Contemporary Guidance for Stated Preference Studies." *Journal of the Association of Environmental and Resource Economists* 4: 319–405.
- Krström, B. 1990. "A Non-parametric Approach to the Estimation of Welfare Measures in Discrete Response Valuation Studies." *Land Economics* 66: 135–39.
- Krström, B. 1997. "Spike models in contingent valuation." *American Journal of Agricultural Economics* 79: 1013–23.
- Lewis, L. Y., L. Richardson, and J. C. Whitehead. 2024. "Fat Tails, Flat Tails, and Willingness to Pay: Krström Revisited." *Land Economics* 100: 639–651.
- Li, C. Z. 1996. "Semiparametric Estimation of the Binary Choice Model for Contingent Valuation." *Land Economics* 72: 462–473.
- Liaw, A., and M. Wiener. 2002. "Classification and Regression by randomForest." *R News* 2: 18–22.
- Linero, A. R., and Y. Yang. 2018. "Bayesian Regression Tree Ensembles that Adapt to Smoothness and Sparsity." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 80: 1087–1110.
- Madadkhani, S., and S. Ikonnikova. 2024. "Toward High-Resolution Projection of Electricity Prices: A Machine Learning Approach to Quantifying the Effects of High Fuel and CO2 Prices." *Energy Economics* 129: 107241.
- Marcellino, M., and M. Pfarrhofer. 2025. "Nonparametric Mixed Frequency Monitoring Macro-at-Risk." *Economics Letters* 255: 112498.
- McCulloch, R., R. Sparapani, R. Gramacy, C. Spanbauer, and M. Pratola. 2018. "BART: Bayesian additive regression trees." R package version 1
- Nahuelhual, L., M. L. Loureiro, and J. Loomis. 2004. "Using Random Parameters to Account for Heterogeneous Preferences in Contingent Valuation of Public Open Space." *Journal of Agricultural and Resource Economics* 29(3): 537–552.
- Niculescu-Mizil, A., and R. Caruana. 2005. "Predicting good probabilities with supervised learning." In *Proceedings of the 22nd International Conference on Machine Learning* 625–632. New York, NY: Association for Computing Machinery.
- Parzen, E. 1960. *Modern Probability Theory and its Applications*, Vol 10. New York, NY: Wiley.

- Penn, J., and W. Hu. 2023. "Adjusting and Calibrating Elicited Values Based on Follow-up Certainty Questions: A Meta-Analysis." *Environmental and Resource Economics* 84: 919–946.
- Poe, G. L., J. E. Clark, D. Rondeau, and W. D. Schulze. 2002. "Provision Point Mechanisms and Field Validity Tests of Contingent Valuation." *Environmental and Resource Economics* 23: 105–131.
- Ramsey, A. F., and S. K. Ghosh. 2025. "Bayesian Additive Regression Tree (BART) Models of Market Integration in the 19th-Century Trans-Atlantic Wheat Trade." Working paper, Agricultural and Applied Economics Association.
- Ramsey, A. F., J. Yu, and K. Moeltner. 2026. "Bayesian Econometrics in Agricultural and Resource Economics." *European Review of Agricultural Economics* 53(1): jba060.
- Ramsey, S. M., and J. S. Bergtold. 2021. "Examining Inferences from Neural Network Estimators of Binary Choice Processes: Marginal Effects, and Willingness-to-Pay." *Computational Economics* 58: 1137–65.
- Richardson, L., and L. Lewis. 2022. "Getting to Know you: Individual Animals, Wildlife Webcams, and Willingness to Pay for Brown Bear Preservation." *American Journal of Agricultural Economics* 104: 673–692.
- Rulleau, B., J. Dehez, and P. Point. 2012. "Recreational Value, User Heterogeneity and Site Characteristics in Contingent Valuation." *Tourism Management* 33: 195–204.
- Sparapani, R., C. Spanbauer, and R. McCulloch. 2021. "Nonparametric Machine Learning and Efficient Computation with Bayesian Additive Regression Trees: The BART R Package." *Journal of Statistical Software* 97: 1–66.
- Tan, Y. V., and J. Roy. 2019. "Bayesian Additive Regression Trees and the General BART Model." *Statistics in Medicine* 38: 5048–69.
- Therneau, T. M., E. J. Atkinson, et al. 1997. An introduction to recursive partitioning using the RPART routines Working paper, Technical report Mayo Foundation
- Turnbull, B. W. 1976. "The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data." *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 38: 290–95.
- Watanabe, M. 2010. "Nonparametric Estimation of Mean Willingness to Pay from Discrete Response Valuation Data." *American Journal of Agricultural Economics* 92: 1114–35.
- Watanabe, M., and K. Asano. 2009. "Distribution Free Consistent Estimation of Mean WTP in Dichotomous Choice Contingent Valuation." *Environmental and Resource Economics* 44: 1–10.
- Wong, J. C. Y., B. Blankenship, J. Urpelainen, K. Balani, K. Ganesan, and K. Bharadwaj. 2022. "Understanding Electricity Billing Preferences in Rural and Urban India: Evidence from a Conjoint Experiment." *Energy Economics* 106: 105735.
- Wright, M. N., and A. Ziegler. 2017. "Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R." *Journal of Statistical Software* 77: 1–17.
- Zapata, S. D., and C. E. Carpio. 2024. "Distribution-Free Methods to Estimate Willingness to Pay Models Using Discrete Response Valuation Data." *Journal of Agricultural and Resource Economics* 49: 39–62.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Follett, Lendie, Brian Vander Naald, and Sonja H. Kolstoe. 2026. "Not Just a Random Forest: Using Bayesian Additive Regression Trees to Estimate Willingness to Pay to Protect Critical Habitat." *American Journal of Agricultural Economics* 1–25. <https://doi.org/10.1002/ajae.70084>