Random Forests for Contingent Valuation

Michela Faccioli

School of International Studies & Department of Economics and Management, University of Trento
michela.faccioli-1@unitn.it

Klaus Moeltner

Department of Agricultural and Applied Economics, Virginia Tech
*moeltner@vt.edu

December 15, 2024

Abstract

We introduce a novel, fully nonparametric estimation framework to process data from survey-based environmental valuation with a binary, referendum-style choice question, traditionally referred to as Contingent Valuation. Our approach combines the construction of choice probabilities via Random Forests (RFs) with welfare predictions via common distribution-free estimators. While popular as back-of-envelope alternatives to parametric estimation, these distribution-free methods are poorly suited for the incorporation of observation-specific heterogeneity. In contrast, our Random Forest Non-Parametric (RFNP) approach produces willingness-to-pay (WTP) estimates at the individual level, conditioned on a potentially large set of explanatory variables. Furthermore, our predicted choice probabilities as well as welfare estimates come with well-defined asymptotic properties. Using simulated data, we find that the RFNP estimator is robust to nonlinearities in the WTP function and can compete with correctly specified parametric models in terms of asymptotic efficiency. In our empirical application within the context of biodiversity enhancements on open land in the United Kingdom, we show that the RFNP is immune to negative WTP predictions by construction, and produces reasonable and efficient lower bound estimates for individual and sample-aggregated WTP. It can also generate welfare predictions that allow for long tails in individual WTP, without having to impose this feature on all observations. Our framework is well-suited for numerous extension, and readily implemented with existing software packages.

keywords: Machine learning, nonparametric methods, land use, biodiversity

Suggested citation:
Paper presented at the 2024 Experimental & Environmental Economics Workshop, Appalachian State University, Boone, NC, April 26-27, 2024. The online appendix to this paper is available here: online appendix

*Corresponding author: 208 Hutcheson Hall, Blacksburg, VA 24061; phone: (540) 231-8249

Introduction

The economic valuation of environmental amenities and services often requires survey-based, or Stated Preferences (SP) approaches to estimate societal benefits of planned policy interventions. Typically, respondents are asked to decide between current conditions, generally referred to as Status Quo (SQ), and one or more hypothetical policy scenarios that stipulate improved environmental quality or amenities, in exchange for a (hypothetically) binding payment ("bid") (Champ et al., 2017; Johnston et al., 2017). As originally recommended by a "blue ribbon" panel of prominent economists convened by the National Oceanic and Atmospheric Agency (NOAA) in the wake of the 1989 Exxon Valdez oil spill, a single, binary choice question involving the SQ and one policy alternative is widely considered the most robust elicitation approach in SP research, in terms of minimizing strategic response behavior and other undesirable survey design effects that could bias welfare estimates (National Oceanic and Atmospheric Administration, 1993; Freeman et al., 2014; Champ et al., 2017; Johnston et al., 2017; Phaneuf and Requate, 2017). This single choice, referendum-style format is generally referred to as "dichotomous choice contingent valuation." We will henceforth adopt the common abbreviation of "Contingent Valuation (CV)."

The vast majority of CV applications to date have been anchored in a Random Utility Modeling (RUM) framework, typically departing from a parameterized indirect utility function (IUF) paired with an additive stochastic error term (hence "random utility") with a specified statistical distribution, such as the Logit or variants thereof (Hanemann, 1984; Freeman et al., 2014; Champ et al., 2017; Phaneuf and Requate, 2017). Given the assumed error distribution, choice probabilities, say the probability of an observed YES response to the proposed policy, can then be expressed as an individual-specific evaluation of a cumulative distribution function (cdf). These cdf terms, in turn, then feed into a likelihood function that yields estimates of the preference parameters. In a final step, these parameter estimates are then combined with environmental quality attributes and (optionally) respondent characteristics to generate individual-specific welfare estimates, usually referred to as willingness-to-pay (WTP).

As noted in numerous articles that consider CV for the estimation of WTP, economic theory provides limited guidance on the form of the IUF and / or the distribution of the stochastic RUM component. In turn, mis-specification of either of both of these elements can lead to biased and inconsistent welfare estimates (e.g. Li, 1996; Creel and Loomis, 1997; Chen and Randall, 1997; Haab and McConnell, 2003; Crooker and Herriges, 2004; Watanabe and Asano, 2009; Watanabe, 2010; Lewis et al., 2024). To mitigate against mis-specification risks, several semi- and nonparametric estimators have been proposed in the existing literature.

The first set of these contributions focuses exclusively on relaxing distributional assumptions for the stochastic RUM component. For example, Li (1996) maintains an explicit form of the IUF, but applies Cosslett (1983)'s distribution-free Maximum Likelihood Estimation (MLE) estimator to recover utility parameters and construct WTP estimates. Similarly, Zapata and Carpio (2024)'s Semiparametric Iterated Linear Model (SPILM) assumes a parametric specification for the expected WTP function while using a nonparametric iterated procedure to estimate the error density. Watanabe and Asano (2009), in turn, completely abstract from any explicit IUF or WTP function, and center their approach to derive mean WTP around an assumed bid distribution.

A second set of papers relax both assumptions on IUF and error distribution, but focus squarely on the estimation of mean WTP for the underlying population without consideration of any explanatory variables. This includes the nonparametric Turnbull (1976) and Kriström (1980) estimators discussed in Haab and McConnell (1997), Haab and McConnell (2003) and Lewis et al. (2024), the aforementioned linear projection estimator of Watanabe and Asano (2009), and the related, but projection-free estimator of Watanabe (2010). However, as argued in Creel and Loomis (1997) and Watanabe (2010), it is often important to derive conditional WTP estimates, given an explicit set of quality changes and stakeholder attributes. This need arises, for example, in the context of Benefit Transfer (BT), where estimates from existing studies are used to predict benefits of a planned policy intervention elsewhere (e.g. Champ et al., 2017; Johnston et al., 2015; Moeltner et al., 2019, 2023; Johnston and Moeltner, 2024). Capturing individual heterogeneity is also important

when the distributional impact of interventions are of interest, as is increasingly the case in the environmental policy arena in light of growing environmental justice concerns (Banzhaf et al., 2019b,a; Andarge et al., 2024).

In this study, we relax both IUF and error specification assumptions while still allowing for the estimation of individual WTP for any desired combination of explanatory variables. There exist a handful of CV contributions that share the same objective, though none of them have become mainstream tools in the CV / SP arena. Creel and Loomis's (1997) Semi-Nonparametric Distribution Free (SNPDF) estimator utilizes a Fourier transform of the IUF in combination with an invertible error cdf to derive conditional WTP estimates. While their error distribution is still explicit (they use the Logistic for simplicity), it can be any invertible density. Chen and Randall (1997) take a similar approach, but depart directly from a WTP function, and approximate the error distribution with another series estimator. However, as also discussed in Crooker and Herriges (2004), the Fourier transform still requires "manual" specification of index vectors that are supposed to capture all possible elementary combinations of explanatory variables. This can become cumbersome for a high-dimensional covariate space, as is the case for our application. Furthermore, both Creel and Loomis's (1997) and Chen and Randall's (1997) method require MLE to estimate model parameters, and numerical integration to recover WTP.

Crooker and Herriges (2004) offer an alternative approach based on the Generalized Maximum Entropy (GME) estimator. While avoiding any type of series approximation, this method still requires the explicit specification of moment conditions that themselves include stochastic components, for which distributional assumptions have to be made. The model, if cast in its dual form, can be estimated via standard MLE (Crooker and Herriges, 2004). Watanabe (2010) builds on Watanabe and Asano (2009)'s Linear Projection estimator, but shows how covariates can be preserved in WTP construction. This approach still requires the linearity assumption for the projection step, and an explicit bid distribution. Moreover, this bid distribution must include the full support of WTP, which is unobserved by definition. Zapata and Carpio (2024) propose a Nonparametric Iterated Additive Model (NIAM). It builds on their SPILM model mentioned above, but specifies the conditional WTP via a

nonparametric additive model. Naturally, this approach requires choice of bandwidth and kernel functions, which have to be determined via cross-validation methods. This can be computationally burdensome with a large set of covariates, as is the default assumption in our methodological framework. Furthermore, their iterative optimization algorithm is rather complex, requiring eight individual estimation steps, while our proposed estimator can be implemented in two simple stages, largely building on existing `R` packages.

In this work we propose a novel, fully nonparametric strategy to estimate individual-specific WTP from CV data. Our method does not require any type of optimization or numerical approximation, can accommodate a large set of explanatory variables (potentially larger than the sample size), requires minimal tuning, and can be readily implemented in `R`. It builds on one of the most powerful and popular Machine Learning (ML) tools, Random Forests (RFs). In essence, we use an RF to predict acceptance probabilities for each respondent, and for each bid level offered in the survey to the sample at large (even though in actuality each survey taker only receives a single bid, as prescribed by the standard CV method). For each individual we then process the paired vectors of bid values and YES-probabilities using variants of the Turnbull (1976) and Kriström (1980) procedures. This, in turn, produces observation-specific welfare estimates, along with asymptotically guaranteed standard errors and confidence intervals.

We use both simulated data and an empirical application on biodiversity enhancements on open lands in the United Kingdom (UK) to showcase our framework. In the simulation exercise, we find that the RFNP estimator is robust to even severe departures from linearity in the WTP function, and generates asymptotic standard errors for WTP predictions that are of the same order of magnitude as those produced by a correctly specified Logit model. The RFNP out-competes the Logit in accuracy and coverage when we introduce nonlinearities in WTP. For our empirical application, we show that the RFNP, when combined with the Turnbull distribution-free estimator, generates reasonable and accurate lower bound estimates for individual and aggregate WTP. Implementing the RFNP with an adjusted Kriström component, as recently suggested by Lewis et al. (2024), allows for the modeling of longer tails for individual WTP, without having to impose this feature for the entire sam-

ple. As an added bonus, all of our RFNP estimators are immune to (nonsensical) negative WTP estimates, which we find to be a problem for their parametric counterpart.

To our best knowledge, this is the first study that applies RFs in the context of CV elicitation (or, for that matter, any SP approach), and the first that utilizes the Turnbull and Kriström (1980) estimators to generate welfare estimates at an individual level. Our framework is easy and fast to implement, and thus offers an attractive alternative to parametric modeling in any CV setting.

Modeling

In this study we compare the performance of a traditional parameterized Logit model and our RFNP estimator. We start by introducing the RUM framework for the Logit model. Consider the simple random utility model for individual i for status quo (subscript 0) and policy (subscript 1) conditions:

$$\begin{aligned} \tilde{U}_{0i}^{*} &= \mathbf{x}_{0i}'\boldsymbol{\beta}^{*} + \gamma m_i + \tilde{\epsilon}_{0i}^{*}, \\ \tilde{U}_{1i}^{*} &= \mathbf{x}_{1i}'\boldsymbol{\beta}^{*} + \gamma\left(m_i - P_i\right) + \tilde{\epsilon}_{1i}^{*} \\ \tilde{\epsilon}_{ji} &\sim EV\left(0,1\right), \quad j = 1,2 \end{aligned} \tag{1}$$

where k-dimensional regressor vector $\mathbf{x}$ comprises quality indicators, possibly interacted with household characteristics, m_i is income, and error term $\tilde{\epsilon}^{*}$ captures unobservables. If the intervention scenario is chosen, payment P_i is subtracted from income, as shown in the second equation. By standard convention, the error follows a standard Extreme Value (EV) distribution with mode zero and unit scale.

Taking the difference between utilities yields

$$\begin{aligned} &U_i^{*} = \tilde{U}_{1i}^{*} - \tilde{U}_{0i}^{*} = \left(\mathbf{x}_{1i} - \mathbf{x}_{0i}\right)'\boldsymbol{\beta}^{*} - \gamma P_i + \epsilon_i^{*} \quad \text{where} \\ &\epsilon_i^{*} = \tilde{\epsilon}_{si}^{*} - \tilde{\epsilon}_{0i}^{*} \sim LOG\left(0,1\right), \end{aligned} \tag{2}$$

where $LOG\left(0,1\right)$ denotes the stadard logistic distribution with a mean of zero and a

scale of one. To facilitate model comparison we convert the utility-differenced model into willingness-to-pay (WTP), or "surplus", space (Train, 2009; Cohen et al., 2016; Johnston et al., 2023). This requires dividing (2) by the price coefficient γ, yielding

$$\begin{aligned} U_i &= \frac{U_i^*}{\gamma} = \mathbf{x}_i'\boldsymbol{\beta} - P_i + \epsilon_i \quad \text{where} \\ \mathbf{x}_i &= (\mathbf{x}_{1i} - \mathbf{x}_{0i}), \; \boldsymbol{\beta} = \frac{\boldsymbol{\beta}^*}{\gamma} \quad \text{and} \\ \epsilon_i &= \frac{\epsilon_i^*}{\gamma} \sim LOG\left(0, \gamma^{-1}\right) \end{aligned} \tag{3}$$

Adjusted utility U_i now has the interpretation of "surplus," defined as the difference between the full WTP to obtain the policy scenario (given as $w_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i$) and the required payment P_i. The probability of an observed YES (= vote for the policy scenario) and NO response, respectively, can then be written as:

$$\begin{aligned} &p\left(YES\right) = p\left(y_i = 1\right) = p\left(w_i^* > P_i\right) = \\ &\left(1 + exp\left(\gamma * \left(P_i - \mathbf{x}_i'\boldsymbol{\beta}\right)\right)\right)^{-1} \\ &p\left(NO\right) = p\left(y_i = 0\right) = p\left(w_i^* < P_i\right) = \\ &\left(1 + exp\left(\gamma * \left(\mathbf{x}_i'\boldsymbol{\beta} - P_i\right)\right)\right)^{-1}, \end{aligned} \tag{4}$$

where y_i is the binary (1=YES, 0=NO) choice indicator collected in the survey. Given a sample of $i = 1 \ldots n$ observations, the model is estimated via standard Maximum Likelihood (MLE) procedures (e.g Cameron and Trivedi, 2007; Greene, 2012).

Random Forests

Random Forests were first introduced to the ML literature by Breiman (2001). They have proven themselves as a powerful and effective prediction tool in many applications (Hastie et al., 2017; Harding and Lamarche, 2021; Storm et al., 2020; Greenwell, 2022). Advantages over other methods include the ability to detect highly nonlinear relationships, robustness to non-normality and outliers, algorithmic treatment of missing data, and limited requirements in terms of pre-processing or tuning. They are also less computationally demanding than

alternative ML approaches such as Neural Networks (Fernández-Delgado et al., 2014). In the realm of environmental and resource economics at large, there are only a handful of empirical studies that have used RFs. Most of those focus on the estimation of causal treatment effects in the context of policy evaluation (Miller, 2020; Harding and Lamarche, 2021; Stetter et al., 2022; Liu et al., 2023; Valente, 2023; Prest et al., 2023; Mink et al., 2024). To our best knowledge, only Hino et al. (2018) exploit the predictive strength of RFs, within the context of detecting industrial water pollution violations. Recently, Johnston and Moeltner (2024) developed an RF-based framework for meta-regression modeling and BT. We are not aware of any published or unpublished study that uses RFs for CV modeling, or, for that matter, in any valuation context based on SP elicitation.

The conceptual starting point for the RF model is a direct relationship between observed choice y_i and all available explanatory variables. Combining covariates $\mathbf{x}_i$ and bid P_i into single vector $\mathbf{z}_i$ to simplify notation, it can be generically written as:

$$\begin{aligned} y_i &= g\left(\mathbf{z}_i\right) + \epsilon_i, \quad \text{with} \\ E\left(\epsilon_i | \mathbf{z}_i\right) &= 0, \end{aligned} \tag{5}$$

where $g\left(.\right)$ is an unspecified nonparametric function, and the only assumption required for the error term is a conditional expectation of zero, mirroring the equivalent assumption for the Logit. As is evident from (5) the model directly specifies observed choice y_i as outcome of interest. Covariates, or "features" $\mathbf{z}_i$ typically enter in differenced form for variables that change between SQ and policy, such as environmental quality, but can otherwise be used in their raw form, such as household characteristics. Furthermore, ordinal variables, such as Likert scores for attitudinal questions, can be fed into the model without pre-processing.

As described in Hastie et al. (2017) and Greenwell (2022), RFs build on a large number of underlying "trees." Each tree, in turn, is constructed using a bootstrapped sub-sample of the full data. The tree is then "grown" by repeatedly and sequentially splitting the data. The splitting point is a specific threshold value within one of the explanatory variables, e.g. "number of children > 2 vs. ≤ 2." The variable that produces the splitting point, in turn, is

chosen from a random subset of all available covariates. Using a specific optimization rule, such as maximum reduction in Mean Squared Error (MSE) (as applied in our case), the RF algorithm then searches over this subset and all possible splitting points contained therein to find the optimal way to separate the data. The splitting process ends when no further reduction in MSE can be achieved and / or minimum sample sizes are reached in what is referred to as "terminal leaves." Each of these leaves then contains a small sub-set of observations, typically three to ten. The placement into leaves of actual sample observations is thus purely determined by covariates $\mathbf{z}_i$.

Prediction of choice probabilities

Both the Logit and RF are suitable for predicting the probability of a YES response for an actual sample point or a new observation for which there is no observed outcome in the data. We will maintain subscript i for the first case, and subscript p (for "prediction") for the general situation that captures both within and out-of-sample scenarios.

For the Logit, the predicted probability of a YES vote for an observation with explanatory features $\mathbf{x}_p$ facing bid P_b, $b = 1 \dots B$, can be constructed via:

$$p\left(YES|\mathbf{x}_p, P_b\right) = \hat{y}_{p,L} = \left(1 + exp\left(\hat{\gamma} * \left(P_b - \mathbf{x}_p'\hat{\boldsymbol{\beta}}\right)\right)\right)^{-1} \tag{6}$$

where subscript L denotes the Logit model, and $\hat{\gamma}$ and $\hat{\boldsymbol{\beta}}$ are the MLE estimates of the corresponding parameters in (4).

For the RF, tree-specific predictions for a point with features $\mathbf{z}_p = \{\mathbf{x}_p, P_b\}$ are generated by averaging outcomes y_i for sample observations that *share the same leaf* as $\mathbf{z}_p$. The final prediction of a YES-probability, $\hat{y}_{p,F}$, with F denoting "Forest," is then derived by averaging these values over all trees in the forest. This can be formally written as (Athey and Wager,

2019; Friedberg et al., 2021; Tibshirani et al., 2024b; Johnston and Moeltner, 2024):

$$\hat{y}_{p,F}|\mathbf{z}_p, \mathbf{X}, \mathbf{y} = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{|L_b(\mathbf{z}_p)|}\sum_{i=1}^{n} y_i\, I\left(\mathbf{z}_i \in L_b(\mathbf{z}_p)\right) = \sum_{i=1}^{n} y_i \frac{1}{B}\sum_{b=1}^{B}\frac{I\left(\mathbf{z}_i \in L_b(\mathbf{z}_p)\right)}{|L_b(\mathbf{z}_p)|} = \sum_{i=1}^{n}\alpha_i\left(\mathbf{z}_p, \mathbf{X}\right) y_i, \tag{7}$$

where $\mathbf{X}$ and $\mathbf{y}$ denote, respectively, the covariate matrix and outcome vector for the full sample, $L_b(\mathbf{z}_p)$ is the terminal leaf of tree b that contains policy point $\mathbf{z}_p$, $|L_b(\mathbf{z}_p)|$ denotes the number of original observations that were assigned to leaf $L_b(\mathbf{z}_p)$, and $I(.)$ is an indicator function. As discussed in Johnston and Moeltner (2024) the first line in (7) performs the averaging over trees of tree-specific averages of outcomes within the target leaf. The second line switches summation, and the third expresses $\frac{1}{B}\sum_{b=1}^{B}\frac{I(\mathbf{z}_i \in L_b(\mathbf{z}_p))}{|L_b(\mathbf{z}_p)|}$ as observation-specific weight that determines the influence each actual outcome y_i has in the construction of final prediction $\hat{y}_p$. These weights are fully nonparametric and data-driven, and adjust for each new policy point $\mathbf{z}_p$. For this reason, Wager and Athey (2018) and Athey et al. (2019) refer to RF-based predictions as an "adaptive kernel method."

As shown in Wager and Athey (2018) and Athey et al. (2019) estimates flowing from RFs are asymptotically normal and consistent if forests are built following the "honesty" principle in tree construction. It requires using different portions of the data to, respectively, grow and populate a given tree (i.e. fill its leaves). We apply this honesty principle for all our RF models.

WTP predictions

For the Logit model a prediction of expected WTP (i.e. holding error noise at zero) for a specific combination of explanatory variables $\mathbf{x}_p$ can be obtained in straightforward fashion

via:[1]

$$\hat{w}_{p,L} = \mathbf{x}_p' \hat{\boldsymbol{\beta}}, \tag{8}$$

where $\hat{w_{p,L}}$ is predicted WTP, in dollars, and $\hat{\boldsymbol{\beta}}$ is the MLE estimate of the scale-adjusted coefficients $\boldsymbol{\beta}$ in (3) and (4).

For the RF, we proceed in two steps. First, we note that forests are well-designed to produce choice probabilities for an individual with any arbitrary mix of features $\mathbf{z}_p = \{\mathbf{x}_p, P_b\}$. This includes the case where quality and household characteristics $\mathbf{x}_p$ remain constant, but bid P_b takes on different levels. In essence, this mimics the ideal survey situation where each respondent answers sequentially and *independently* a set of choice questions for identical quality scenarios, but with bid varying over all bid levels represented in the survey, say P_b, $b = 1 \dots B$. In an RF predictive framework this is achieved by simply "replicating" each individual's (transposed) feature vector $\mathbf{x}_p$ B times, and adding a vector of sequentially increasing bids as a separate column. The resulting $n \times B$ by $k+1$ "augmented" feature matrix is then fed into the original forest to obtain B choice predictions per individual, one for each bid level.[2]

In the second step, we then employ the nonparametric methods discussed in Haab and McConnell (1997) and Haab and McConnell (2003), ch.3, to convert the B voting predictions into expected WTP. Specifically, we consider the lower-bound Turnbull estimator (Tb.low), the upper-bound Turnbull (Tb.up), and the linear interpolation estimator originally proposed by Kriström (1990) (K). Denoting the predicted probability of a YES response to bid P_b for an individual with features $\mathbf{x}_p$ as $\hat{y}_{p,b}$, and the corresponding probability of a NO response as $\hat{p}_{p,b} = 1 - \hat{y}_{p,b}$, the TB.low (henceforth captured with subscript "l") estimator for expected WTP flowing from the RF can be expressed as (Haab and McConnell, 2003):

$$\hat{w}_{p,l} = \sum_{b=0}^{B} P_b \left(\hat{p}_{p,b+1} - \hat{p}_{p,b} \right), \tag{9}$$

[1] We will henceforth use the terms "expected WTP" and "WTP" interchangeably. Both refer to the expectation over the stochastic component of a given estimator.

[2] Naturally, any arbitrary bid level not offered to anybody in the actual survey could be used to generate predictions. Here we limit ourselves to actual bids featured in the questionnaire to allow for a more even comparison with the Logit model.

where $P_b,\ b = 1 \dots B$ are the bids actually offered in the survey (to the sample at large), and lower-bound bid P_0 and corresponding NO-probability $\hat{p}_{p,0}$ are set to zero and one, respectively (following the standard assumption that nobody would reject the policy scenario if it came at no cost). Equation (9) also requires a value for the term $\hat{p}_{p,B+1}$, which is set to one by convention. This can be understood as the NO probability to a hypothetical "cut-off" bid P_{B+1}, i.e. the price point at which the individual under consideration would decline the policy scenario with 100% probability. As is evident from (9), and discussed in Haab and McConnell (1997) and Haab and McConnell (2003), an advantage of the Tb.low compared to other nonparametric WTP estimators (including Tb.up and K) is that this cut-off point can remain unspecified. As indicated by (9) estimated WTP $\hat{w}_{p,l}$ can be interpreted as the expectation of a discrete random variable with support points $P_b,\ b = 0 \dots B$ and corresponding point masses $\hat{p}_{p,b+1} - \hat{p}_{p,b}$.[3] This can also be seen as an approximation of a continuous distribution for WTP, where the entire probability mass for the interval $P_{b+1} - P_b$ is assigned to the lower of the two threshold bids. For this reason the Tb.low constitutes by construction a lower-bound estimate of expected WTP. Table A1 and the top panel of Figure A1 in the online appendix give a stylized example.

The Tb.up (henceforth denoted by subscript "u") estimator for expected WTP, in turn, is derived as (Haab and McConnell, 2003):

$$\hat{w}_{p,u} = \sum_{b=0}^{B} P_{b+1} \left(\hat{p}_{p,b+1} - \hat{p}_{p,b} \right), \tag{10}$$

In this case, an arbitrary cut-off bid P_{B+1} needs to be chosen by the analyst. Compared to the Tb.low, all probability masses are then shifted to the right by one bid level for the construction of expected WTP $\hat{w}_{p,u}$. This situation is depicted in the bottom panel of Figure A1 in the online appendix, with a corresponding numerical example shown in Table A1. Thinking again in terms of a continuous underlying distribution of WTP, the Tb.up estimate is, by construction, an upper bound for the true expectation, conditional on knowing P_{B+1}.

[3] It can be easily verified mathematically, and empirically from online appendix Table A2, that these point probabilities add to one, as required.

The third nonparameteric WTP estimator we consider was originally suggested by Kriström (1990), and has since been employed in several empirical valuation studies (e.g. Ready and Hu, 1995; Creel and Loomis, 1997; Haab and McConnell, 1997; Richardson and Lewis, 2022; Lewis et al., 2024). Instead of assigning point-mass probabilities to each bid level, the K estimator approximates WTP density between bids via linear interpolation. This is shown in Figure A2 in the online appendix, with a corresponding numerical example given in Table A2. Mathematically, the K estimate for expected WTP (henceforth denoted by subscript "k") can be expressed as (Kriström, 1990; Haab and McConnell, 2003):

$$\hat{w}_{p,k} = \sum_{b=0}^{B} \left(P_{b+1} - P_b\right) \left(\hat{y}_{p,b+1} + \tfrac{1}{2}\left(\hat{y}_{p,b} - \hat{y}_{p,b+1}\right)\right), \tag{11}$$

where, as previously noted, $\hat{y}_{p,b}$ is the estimated probability of a YES response for an individual with feature vector $\mathbf{x}_p$, facing bid P_b.

As discussed in detail in Haab and McConnell (2003), all three nonparametric estimators assume ex ante that YES-probabilities are monotonically decreasing over bids (or, conversely, NO-probabilities are increasing with bid levels). If this monotonicity is violated, a *smoothing procedure* needs to be applied that drops bids with non-conforming associated NO or YES probabilities. The standard formula as given in (9) to (11) can then be applied to the remaining bid levels and corresponding probabilities. We apply a monotonicity check and, if required, a smoothing adjustment for all individuals and models in our analysis, as discussed below in more detail.

To date, the Turnbull and Kriström estimators have primarily been employed to derive a single WTP estimate for the sample at large, using sample proportion of NO or YES responses to a given bid in lieu of our estimated individual probabilities. In essence, we generalize this method by applying it at the individual level, obtaining a full set of n WTP predictions for the actual sample, and "personalized" WTP estimates for any desired composition of quality and household characteristics in $\mathbf{x}_p$. As mentioned above, and discussed in Haab and McConnell (2003), the traditional Tb and K estimators have only very limited ability to incorporate any form of observed heterogeneity, requiring splitting the data into

arbitrary sub-groups and applying the proportion-based estimator to each group. In stark contrast, our approach produces heterogeneous WTP estimates at the individual level by default.

Choice of cut-off bid

As is clear from the discussion above, both the Tb.up and K estimator require the ad-hoc specification of a cut-off bid P_{B+1}. In our case, this value has to be selected or empirically constructed for *each individual* for which a WTP prediction is sought. For our simulation exercise below, where we know the actual underlying distribution of WTP for the sample at large, we can mitigate against poor choices of P_{B+1} by setting up the ladder of "actual" bids such that only a very small proportion of individuals chooses the policy option at the highest administered level B. In other words, for most simulated "respondents" the highest bid offered in our hypothetical survey would be a reasonable (and observed) cut-off value (see also Glenk et al., 2024). For the remaining cases we select a common P_{B+1} near the maximum "observed" WTP.

The situation is less clear-cut for our empirical application, where a considerable share of the sample still votes for the policy option at the highest offered bid, and actual WTP is unknown by default. Here we apply two variants of the K estimator. The first version simply truncates a given individual's WTP distribution at the highest observed bid B. This approach is referred to as the "truncated Kriström" estimator (K.tr) in Lewis et al. (2024), and we will adopt this label for our application. As is the case for the Tb.low, the K.tr will produce an expected WTP estimate that is biased downward, but to a lesser degree than for the Tb.l, given its linear interpolation between all interior bids when constructing the estimate for expected WTP (Lewis et al., 2024). It can thus be interpreted as a less conservative lower bound estimate of expected WTP, if this linear interpolation assumption is correct.

For the second K-variant we follow Whitehead (2017), Richardson and Lewis (2022), and Lewis et al. (2024) and adopt their "adjusted Kriström estimator" (K.adj). This approach considers the slope of a linear regression of prob(YES) on observed bid values. The authors

then linearly extrapolate the WTP distribution from the last observed point $\{P_B, \hat{y}_{p,B}\}$, using this slope estimate to identify the cut-off bid P_{B+1}. The original approach taken in Lewis et al. (2024) only uses interior points to derive the slope estimate. In our case, this still leads to excessively high cut-off bids for some individuals due to relatively flat prob(YES)-regions across interior bids. This problem is mitigated when the hypothetical (but realistic) first point $\hat{y}_{p,0} = 1$ is used in the slope computation (see Figure A2 in the online appendix).[4] Since no such empirical guidance for the determination of P_{B+1} exists for the TB.up, we do not consider this estimator in our empirical application.

Standard errors and uncertainty bounds

Wager and Athey (2018) show that predictions flowing from random forests are asymptotically unbiased and normally distributed under a set of standard assumptions, most notably the honesty principle discussed above. Wager and Athey (2018) also illustrate how estimates of the asymptotic variance can be obtained, allowing for the construction of standard errors and confidence intervals. These asymptotic guarantees apply directly to our forest-generated choice probabilities in (7). In fact, the `R` package `grf`, which we use for the forest portion of our analysis, automatically generates a full set of individual-level standard errors along with point-predictions (Tibshirani et al., 2024a,b). This may be useful in situations where the prediction of voting probabilities for different segments of the target population is of central interest.

Here we are primarily interested in WTP estimates, for which we also seek asymptotically valid standard errors and confidence intervals. Clearly, all three of our nonparametric estimators in (9) through (11) are linear combinations of bid values and the estimated probabilities flowing from the forest. Since the latter are, in theory, independently and asymptotically normally distributed, we could derive estimated standard errors for WTP via simple algebraic manipulation (e.g. Greene, 2012, app. B). However, we prefer fol-

[4] In his original exposition, Kriström (1990) takes a similar approach, but only uses the last two points at the two highest observed bids to compute the slope estimate. As observed by Lewis et al. (2024) this can lead to excessive cut-off values and WTP estimates when acceptance probabilities do not change much across these final two bid values, a situation they call the "fat tails" problem.

lowing Creel and Loomis (1997) and Zapata and Carpio (2024) and use bootstrapping to obtain standard error for each WTP estimate. This guards against deviations of the independence assumption in finite samples, especially under smoothing adjustments (Haab and McConnell, 2003), and better captures the notion that each block of B bid-probability pairs fed into the Turnbull and Kriström estimators corresponds to a single individual. This is similar in spirit to the panel or block bootstrap discussed i.e. in Cameron and Trivedi (2007), section 11.6.2., where individuals represent the panel or block level, and the predicted YES probabilities form the panel-specific observations. To align post-estimation procedures across all models we also use bootstrapping to derive standard errors for the Logit. The exact steps of this bootstrap are given in the online appendix.

Simulation

To motivate our approach as guarding against misspecification we simulate data for three different models. Each model has a logistic error to conform to the standard Logit assumptions, but differs in the composition of its expectation function. Model M1 has a simple linear form, such that the generic Logit is the correct specification. It's corresponding latent WTP (i.e. surplus minus bid as in equation (3)) can be written as:

$$
\begin{aligned}
& w_i^* = 1 + x_{2,i} * \beta_2 + x_{3,i} * \beta_3 + \mathbf{x}'_{r,i}\boldsymbol{\beta}_r + \epsilon_i, \quad \text{with} \\
& \epsilon_i \sim LOG\left(0, \gamma^{-1}\right), \;\; \gamma = 0.5 \\
& x_2 \sim n\left(1.5, 0.4^2\right), \;\; x_3 \sim n\left(2.5, 0.4^2\right), \\
& x_{r,j} \sim n\left(0,1\right), \; j = 1 \ldots 7, \quad \beta_{r,j} = 0, j = 1 \ldots 7
\end{aligned}
\tag{12}
$$

where error scale γ^{-1} is the inverted marginal utility of income, to align with our "surplus" representation in (3). Our main focus rests on the specification of covariates x_2 and x_3, with the remaining seven regressors and corresponding null coefficients added to pose an adequate challenge in finding split points for our RF specifications.[5] The bid vector is chosen to correspond to the following percentiles of w_i^*: 5, 18, 31, 44, 57, 70, 83, and

[5] As noted in Tibshirani et al. (2024b), RFs perform best if trained on > 3 variables.

96. This assures a monotonic decline in YES probabilities over bids, which (largely) limits the need for smoothing adjustments in our nonparametric estimators. As discussed above, setting the highest bid to the 96^{th} percentile implies that very few individuals still choose the policy option at this price. This, in turn, guarantees that our simulation results will not be materially affected by the assumed cutoff bid (P_{B+1} in the previous section) for the K and Tb.up estimators. We choose this cutoff price as at the 99^{th} percentile of latent WTP, and use the same value for all individuals. "Observed" response indicator y_i is then constructed following the usual decision rule as captured in equation (4) above.

Model M2 is generated in the same fashion as M1, but introduces piecewise-linearity into the expectation function via:

$$\begin{aligned} &\beta_2 = 0 \ \text{ if } \ x_{2,i} \leq \bar{x}_2, \quad \beta_2 = 2 \ \text{ otherwise} \\ &\beta_3 = -1 \ \text{ if } \ x_{3,i} \leq \bar{x}_3, \quad \beta_3 = 4 \ \text{ otherwise} \end{aligned} \tag{13}$$

where the "bar" notation indicates the sample mean. This mimics a situation where preferences for certain quality attributes change around specific threshold points. For example, water clarity (or algal concentration) may not matter much until visibility increases above (or drops below) a certain threshold. Other attributes may switch from a disamenity to a benefit, as conveyed by the β_3 case. For example cold water temperatures in a lake or river may be perceived as undesirable for the purpose of water-based recreation, while warmer temperatures beyond some threshold may suddenly increase the value of such a site. Of course, the opposite could also hold, as exemplified by the problem of wildlife over-abundance in certain recreation areas or neighborhoods. Random Forests are generally well-suited to capture these types of nonlinearities, while the standard linear Logit will be misspecified by construction.

Our third stylized model continues along these lines by adding piecewise-*nonlinearity* to the original setup. Specifically:

$$\begin{aligned} &\beta_2 = 0 \ \text{ if } \ x_{2,i} \leq \bar{x}_2, \quad \beta_2 = 0.2 * x_{2,i} \ \text{ otherwise} \\ &\beta_3 = 3 \ \text{ if } \ x_{3,i} \leq \bar{x}_3, \quad \beta_3 = -0.5 * x_{3,i} \ \text{ otherwise} \end{aligned} \tag{14}$$

For each model we generate $n = 4,000$ observations, which we split evenly into a training and test sample. The training sample is used in actual estimation, while the test sample is used for predictions of choice probabilities and welfare measures, and corresponding model fit statistics. We estimate all three specifications with a standard linear Logit and a classical regression RF (with an MSE-based splitting rule), respectively.[6] As described in Tibshirani et al. (2024b) and discussed in Johnston and Moeltner (2024), RFs can be adjusted via several tuning parameters. We choose the `grf` default of 2000 underlying trees after observing that predictive fit stabilizes around 1,000 trees. We let the `grf` algorithm determine optimal settings of other key tuners, such as minimum leaf size (i.e. when a node becomes too small to be split further) and the number of randomly sampled covariates considered at each split occasion, via built-in cross-validation (Johnston and Moeltner, 2024; Tibshirani et al., 2024a,b). Before the bundle of predicted probabilities for a given individual generated by the RF gets processed by the Turnbull and Kriström estimators, we check for monotonic smoothness and eliminate divergent bid-probability pairs (Haab and McConnell, 2003). We keep track of these corrections and report summary statistics for these smoothing adjustments, as discussed below.

Predicted probabilities

We first examine the predicted YES probabilities produced by the different models. Figure 1 shows a scatter plot of M1 results for the Logit (top panel) and RF (bottom panel) vis-a-vis the corresponding true probabilities, with the superimposed 45-degree line indicating perfect concordance. As expected, the Logit, which is the correct specification in this case, predicts underlying YES probabilities very accurately, with all point-pairs tightly arranged along the parity line. However, the RF also performs reasonably well, especially in the tail areas. Not surprisingly, there is more variability in predictions near the 0.5-boundary, where unobservables have a relatively stronger influence on the mix of observed votes y_i within a given leaf (see (7)). This picture changes dramatically for scenario M2 with its piecewise-

[6] For comparison, and given the binary nature of our outcome variable, we also estimate classification forests, which use a slightly different splitting rule for its underlying trees (Hastie et al., 2017; Greenwell, 2022). The results are essentially identical to those produced by the regression forests.

linear expectation function. As can be seen from Figure 2, the linear Logit model is unable to track the correct probabilities, and produces an almost random pattern of predictions. In contrast, the RF predictions are still tightly clustered around the parity line, with very few exceptions. A similar picture emerges for scenario M3, as shown in Figure A3 of the online appendix. This provides a first indication of the robustness of RFs to deviations from the customary linear specification in indirect utility and expected WTP.

Predicted WTP for test set observations

Turning to welfare estimates, our first round of simulations produces WTP predictions for each individual in the test sample, and captures predictive fit for the sample at large via MSE and Mean Absolute Percentage Error (MAPE).[7] Results are given in Table 1. The table has three blocks of rows, one for each simulation scenario. The first row in each block shows results corresponding to the correct expected WTP value for each test observation. The first four numeric columns give the mean, standard deviation, minimum, and maximum WTP across all test observations. As is evident from the table, mean and standard deviations are of comparable magnitude across all models and simulations, and located in close proximity to the corresponding true values. As expected given the underlying formulas (equations (9) through (11)), the mean Kriström estimate is nested between the lower and upper Turnbull, respectively.

In contrast to means and standard deviations, the range between minimum and maximum predicted WTP (third-to-last-column) is considerably larger for the Logit model compared to the forest-based versions and the true WTP dispersion. This divergence is especially pronounced for the two scenarios (M2, M3) for which the Logit is misspecified by construction. Specifically, the Logit range is 75% wider than the true dispersion for M2, and over double the width of the actual spread of WTP for M3. In stark contrast, all ranges generated by the forest-based models, while under-predicting actual dispersion, are within 5-30% of the true range, with the TB.low performing especially well in that respect. One

[7] Formally, the MSE is given as $\frac{1}{n_T}\sum_{i=1}^{n}(\hat{w}_i - w_i^*)^2$, where n_T is the size of the test sample, $\hat{w}_i$ is the predicted WTP (in dollars) for sample observation i. The MAPE, in turn, is derived as $\frac{100}{n}\sum_{i=1}^{n_T}|\frac{\hat{w}_i - w_i^*}{w_i^*}|$.

reason for the excess dispersion of individual welfare estimates exhibited by the Logit for scenarios M2 and M3 is the vastly increased error variance under mis-specification. While the actual variance corresponding to a stipulated scale of $\gamma^{-1} = 2$ amounts to $\frac{\gamma^{-2} * \pi^2}{3} \approx 13$, the estimated variance for M2 and M3 equals (approximately) 36 (scale of 3.3), and 25 (scale of 2.8), respectively.

The last two columns of Table 1 show MSE and MAPE statistics for all scenarios and models. Not surprisingly, the Logit exhibits the best fit (= smallest MSE, MAPE) when it is correctly specified, with the Kriström estimator performing best among the nonparametric specifications. This picture is flipped again for scenarios M2 and M3, for which the Logit produces by far the largest MSE and largest to second-largest MAPE, respectively. For both nonlinear scenarios, the Kriström estimator stands out as the most accurate predictor amongst all models as measured by both goodness-of-fit statistics.

Additional intuition on model performance is given by Figures 3 and 4, which depict histograms for the entire distribution of true vs. estimated expected WTP values. Each figure contains four panels, one for each model. In each panel, the distribution of true WTP values is given in grey shading, with the estimated distributions super-imposed in a darker (blue) shading. For each panel, the x-axis gives WTP in dollars, while the y-axis captures normalized frequencies, such that the area under each histogram sums to one. Figure 3, which corresponds to the linear scenario M1, mirrors the results from Table 1: The Logit tracks the true empirical distribution of WTP very closely. As expected, the WTP distribution for the lower-bound Turnbull is shifted to the left, and that for the Tb.up is (slightly) shifted to the right compared to the true pattern. The distribution produced by the K estimator is largely centered on the correct version, but missed some mass in the right tail.

Figure 4 shows analogous panels for nonlinear scenario M2. By construction, the true distribution of individual (expected) WTP is starkly bimodal, given the central location of the kink points for β_2 and β_3, as stipulated in (13). As is clear from the upper left panel, the generic Logit produces an approximately symmetric distribution, which completely misses the correct pattern. In contrast, all three forest-based models reproduce the bi-modality

of the true distribution, but with the Tb.low shifting the entire histogram to the left, and the Tb.up moving the entire distribution to the right. The K estimator, in turn, matches both shape and range of the true WTP distribution quite well, which translates into the superior goodness-of-fit statistics compared to the other specifications observed in Table 1. Scenario M3 generates very similar patterns, which can be inspected in Figure A4 of the online appendix.

Predicted WTP for selected policy points

In our second simulation exercise we generate predicted (expected) WTP estimates for three selected out-of-sample points, set at the first, second (median), and third quartile of the empirical distribution of covariates x_2 through x_{10} in (13). This mimics a situation where point-specific predictions are sought, for example in a Benefit Transfer (BT) context (Johnston et al., 2015; Johnston and Moeltner, 2024). Results are given in Table 2, which features the same block-of-rows structure as Table 1. For each target point, the table captures the expected WTP estimate, along with its asymptotic standard error (derived via 500 bootstrapped samples, as discussed above), and the lower and upper bound of the corresponding 95% confidence interval (C.I.). For the point estimates, the value closest to the true E(wtp) is highlighted with a frame. We also added grey shading to confidence intervals that contain the true WTP value.

As can be seen from the table, the Logit generally produces the smallest asymptotic standard errors in most instances. This is as expected, given that the RF-models require two estimation steps to arrive at WTP, while for the Logit WTP estimates can be derived directly as a linear combination of estimated model coefficients (see equation (8)). However, in most cases the forest-generated standard errors are of the same order of magnitude as those produced by the Logit, even for the linear model. This is reassuring, as it indicates that the switch from a parametric, single-step estimator to a nonparametric, two-step approach does not come at the cost of prohibitively high efficiency losses. In terms of accuracy, the Logit estimate is, not surprisingly, closest to the true value for the linear model, but for all other cases one of the forest models, usually the K-estimate, exhibits higher proximity to

the true target. In terms of coverage, we note that the Logit and at least one of the forest estimators generate confidence bounds that contain the true WTP for the linear model. For the nonlinear scenarios, the K-generated C.I. contains the target estimate in three of six cases, with the upper Turnbull achieving the same for the third quartile point and M2.

Smoothing analysis

A remaining issue to examine is the RF's ability to generate YES-probabilities that are monotonically decreasing with increasing bid amount, at the individual level. Table 3 gives an overview of incidences where smoothing adjustments were necessary to impose monotonicity. Each pair of columns refers to one of our simulation scenarios (M1, M2, M3). The first column for each pair shows the sample count for each possible number of adjustments (given eight offered bids), as listed in the very first column of the table. As is evident from the table, for scenario M1 only 67 observations, or 3.35% of the test sample, required *any* smoothing intervention. More smoothing is needed for the nonlinear scenarios M2 and M3, but the number of adjustments per individual remain largely in the one to two range, with zero adjustments continuing to exhibit the highest frequency. These results indicate that smoothing remains the exception rather than the norm at least for our stylized data, such that we can be confident that our simulation results are not overly influenced by these monotonicity adjustments.

In sum, based on the totality of these simulation results we conclude that our RFNP estimator is capable of competing with a "workhorse" parametric approach in terms of accuracy and efficiency even in a situation where the parametric model is correctly specified. Furthermore, the RFNP remains robust under deviations from linearity in covariates in IUF and expected WTP, for both predicted choice probabilities and WTP estimates. It is able to detect multi-modality and gaps in empirical WTP distributions, and generate out-of-sample predictions with a reasonable to excellent degree of accuracy and efficiency (in terms of uncertainty bounds). Perhaps most reassuringly, the RFs appear to be sufficiently refined to (largely) assure that predicted YES probabilities decrease with increasing bids, a critical prerequisite for using the nonparametric estimators in the second step without the

need for excessive smoothing. In the next section we will put our RFNP framework to test in an actual empirical context.

Empirical Application

Data

For the empirical application of this study we rely on CV data and socio-demographic / attitudinal information that were collected as part of a larger SP project, which also included a Discrete Choice Experiment (DCE). The DCE component is analyzed in Faccioli et al. (2024). Here, we consider the CV portion of their survey, which has not yet been examined in any empirical context. To provide some policy background, the survey aims to elicit United Kingdom (UK) residents' preferences for biodiversity-enhanced open land areas to offset lost lands due to housing developments. Specifically, recently implemented UK laws stipulate that housing developers must ensure a minimum 10% "uplift" in biodiversity at or near a given construction site. In the survey, these "Net Gain" projects are interpreted as enhancing a nearby tract of open land with elements conducive to increased biodiversity, such as trees, shrubs, and, possibly, water features. Inspired by ongoing initiatives of several local planning authorities, who aim to achieve a greater than 10% biodiversity enhancement, the survey proposed to respondents a policy option, at extra cost, of different degrees of additional "uplift" (Faccioli et al., 2024). As described in the original paper, the survey data were collected online through a market research company that contacted a representative sample of UK residents during Spring 2022. This produced 3,600 completed questionnaires. The final sample after some cleaning steps comprises 3,203 observations (Faccioli et al., 2024).

The specific CV question we use for this application proposed a policy scenario with a "Moderate Nature Enhancement" net gain project to offset lost open space due to a new housing development.[8] The sample was split according to three features of the housing

[8] While the survey asked four consecutive CV questions of each respondent, we consider only the first for this analysis to avoid potential ordering or sequencing effects (Champ et al., 2017; Johnston et al., 2017) and maintain the "gold standard" setting of a single binary choice questions offered to a given individual.

project: (1) distance from the respondent's location (2 miles / 50 miles), (2) scale of the new development (100 homes / 2000 homes), and (3) the wealth level (low, average, high) of the affected population, i.e. stakeholders that live near the open land that is lost due to development. Each split sample was also told that the offset parcel with enhanced biodiversity would be within the same distance from their residence as the new development, and benefit the same neighborhood, as defined by its wealth level, that would lose existing open space. Respondents were then asked if they preferred the SQ, i.e. the new development, as described, with an offset open land parcel that features only minimal legal enhancement requirements at no additional cost, or the policy scenario, i.e. the new development with an offset parcel that features a moderately enhanced ecosystem that goes beyond legal requirements, at a specific tax increase of P_b. The payment was specified as annual for a period of five years. Respondents were also assured that the enhanced project would be maintained for 30 years. Figure A5 in the online appendix shows an example CV question.[9]

In terms of utility-theoretic underpinnings, which are relevant for our Logit specification, development features and household characteristics are preserved in the analysis via interaction with the policy-specific constant term. The online appendix gives the explicit structure of the model. We would ex ante expect WTP to increase with proximity to a respondent's home, as this would raise opportunities related to potential use values (e.g. bird / wildlife watching on the biodiversity-enhanced offset parcel). By the same token, willingness-to-pay should also increase with the scale of the new development, as this would imply a larger offsetting and enhanced policy parcel. It may also increase with decreasing wealth of affected neighborhoods if WTP is driven by environmental justice concerns.[10]

Each survey taker was randomly assigned one of eight different bid values, ranging from £2 to £96. The distribution of YES and NO responses over bids for the sample at large is given in Table A5 of the online appendix. As is clear from the table, sample proportions of YES votes range from over 83% for the lowest bid to just over 32% for the highest tax

[9]The full survey instrument is given in the last section of the online appendix.

[10]In the Choice Experiment portion of the survey, Faccioli et al. (2024) find indeed that distance has a significantly negative effect on WTP, and that WTP increases if neighborhoods of lower wealth are affected. They did not analyze the effect of development / parcel scale, as the CE portion only referred to the 100-homes scenario.

level. Such a sizable share of YES responses to the highest bid, while challenging from an estimation perspective, is not uncommon in CV applications.[11] This relatively large share of YES responses to the highest observed bid raises the stakes for a reasonable choice of cut-off bids for our nonparametric estimators. As discussed above, we take a two-pronged approach to address this issue. Our K.tr estimator circumvents the problem by truncating the WTP survival function at the highest observed bid, accepting a downward bias and "lower bound" interpretation for expected WTP in exchange, assuming that linear interpolation between interior bids is not too far off the mark in tracking the unknown survival function. On the other hand, the K.adj approach takes an empirical route by deriving individual-specific cut-off bids via linear extrapolation, utilizing information on YES-responses to the observed bid amounts, and following recent literature (Richardson and Lewis, 2022; Lewis et al., 2024). We should note that this large share of acceptances at the highest bid and implied long tail for the expected WTP function also poses a challenge for the generic Logit model, which, for a given individual, is restricted to symmetry by definition. On the bright side, Table A5 also shows that empirical YES proportions are monotonically decreasing over the entire bid range, which should limit the need for smoothing interventions as discussed above and empirically verified below.

Estimation results

For purely illustrative purpose, we predict (expected) WTP for biodiversity-enhanced open space conditional on two "cornerstone" development scenarios, a (presumed) "high value" scenario S1, with features: 2 miles, 2000 homes, low wealth, and a (presumed) "low value" scenario S2, with characteristics: 50 miles, 100 homes, high wealth. In a first estimation round, we predict this welfare measure for each individual in the sample and report summary statistics. In a second round we predict the sample mean of expected WTP, along with asymptotic standard errors and confidence bounds, using bootstrapping as described above. For both rounds of predictions we adopt each household's characteristics and at-

[11] Parsons and Myers (2016) examine 86 CV studies conducted between 1995 and 2014, and find that 48% of them exhibit an acceptance probability of 30% or higher at the highest bid.

titudinal scores as collected in the survey, but set the development / affected population attributes to S1 or S2, respectively, for all individuals. Tables A3 and A4 in the online appendix give an overview of household-level variables. As before, we compare estimation results between the Logit and the different RFNP versions. Specifically, we derive expected WTP using the lower Turnbull (Tb.low), truncated Kriström (K.tr) and adjusted Kriström (K.adj.) estimators, as discussed in the previous section. As for the simulation exercise, we train our RFs with 2,000 trees and let `grf`'s built-in cross-validation feature determine the optimal setting for the remaining tuners. We use the entire sample of 3,203 observations to accomplish this task.[12]

Estimation results for individual-level WTP predictions are given in Table 4. The table shows the mean, standard deviation, minimum, maximum, and range of individual WTP for the sample at large, holding scenario-specific attributes at "S1" and "S2" levels, respectively. The upper block of rows gives results for scenario S1, with the lower block capturing estimates for S2. Within each block, the first four rows show results for each of our four estimators (Logit, TB.low, K.tr, K.adj), while the fifth row captures the empirically derived cut-off bid for the K.adj. We first note that the linear Logit model exhibits the well-known problem of negative WTP predictions, which are nonsensical in our context, especially of such large magnitude (down to -£58.6 for S1, and -£77.2 for S2). As discussed in Haab and McConnell (1997) this could, in theory, be circumvented with alternative parametric approaches, but at the cost of further increasing the sensitivity of welfare estimates to underlying distributional assumptions. As is evident from Table 4, this problem is avoided by construction for our nonparametric estimators, which exhibit reasonable minimum WTP values in the £25-£35 range.

Across scenarios, we observe the expected higher mean WTP for S1 compared to S2, though differences are negligible for the RF-based models. In contrast, the large gap between mean WTP estimates for S1 over S2 produced by the Logit appears to be inflated and is likely an artifact of negative WTP values, which are even more pronounced for the second

[12] To construct predictions for individual-specific YES probabilities, we consider only trees that were not "grown" with the help of the target observation to avoid over-fitting. This "out-of-bag" strategy is another built-in `grf` feature that is automatically activated when training and prediction samples are identical.

scenario. Across forest-based models, we note that the ranking of mean WTP estimates is Tb.low < K.tr < K.adj, as expected. While the Tb.low and K.tr produce similar means and distributional ranges, the K.adj exhibits a pronouncedly wider distributions, with maxima in the £700-1000 range across the two scenarios. As is evident from Figure 5 in the main text, and Figure A7 in the online appendix (both discussed below in more detail), these long right tails are driven by a relatively small number of outlier cases, with the bulk of WTP estimates confined to the £45-85 range. As a result, the mean WTP estimates for the K.adj are of the same order of magnitude as those generated by the other models for both policy scenarios. Specifically, they are approximately 45-50% higher than those produced by the K.tr, and 55-60% higher than the Tb.low counterparts.

As depicted in the last row for each scenario-specific block in Table 4, the mean empirical cut-off bid for the K.adj lies between £180 and £185 for both scenarios. For comparison, Figure A6 in the online appendix captures mean WTP results for S1 for the generic K model (as used in our simulation) for a series of *arbitrary* cut-off bids, applied to all individuals. At a common cut-off of £185, mean WTP generated by the K model for scenario S1 amounts to approximately £65, which is close to the value of just over £69 produced by the K.adj. Conversely, the common ad-hoc cut-off that would produce the same mean estimate as the K.adj amounts to (approximately) £205, which is not too far from the empirical mean cut-off of £183. We take this comparison as evidence that the mean WTP estimates generated by the K.adj model are not overly influenced by the outlier cases at the right-tail end of the sample distribution.

Figure 5 depicts histograms for the distribution of individual expected WTP estimates produced by our four models, for the first development scenario. For the three forest-based graphs, we super-impose the WTP estimates one would obtain from the traditional, "one-for-all" Tb.low and K-estimators, based on empirical sample proportions as vertical blue lines (Haab and McConnell, 1997, 2003; Lewis et al., 2024). The figure shows clearly that the RFNP models produce a right-skewed welfare distribution across individuals that is fully contained in the positive realm, while the Logit remains fairly symmetric and reaches into negative territory with its left-hand tail. The second key insight to be gained from the figure

is that the traditional nonparametric estimates (vertical (blue) lines), while located within proximity of the mean estimates of their counterpart RFNP specifications, fail to capture the considerable heterogeneity in welfare measures we observe for our sample. In other words, our RFNP models are much better suited to answer questions on distributional implications of land use policies, such as those considered in our scenarios. As mentioned at the onset, this is highly relevant given increasing awareness and concerns among policy-makers regarding equity and "environmental justice" issues related to envisioned environmental interventions.

Asymptotic results for scenario-specific mean WTP estimates are captured in Table 5. The table columns repeat the mean estimate from Table 4, followed by its asymptotic standard error (s.e.), lower and upper bounds of the corresponding 95% C.I., and the range of the C.I. As before, scenario-specific results are organized in blocks of rows, with each row corresponding to one of our four models. The main take-home message from the table is that none of the s.e.'s generated by the RFNP specifications are excessively inflated compared to the Logit, mirroring our insights from the simulation exercise above. In fact, s.e.'s for the Tb.low and K.tr are considerably smaller than those produced by the Logit, leading to C.I.'s that are two to three times tighter then the respective Logit counterpart. For the K.adj, in turn, standard errors are approximately three times larger than those generated by the other RFNP models. They are of similar magnitude to those flowing from the Logit for S1, and approximately twice as large as the Logit s.e.'s for S2. Arguably, uncertainty intervals for all our RFNP specifications are sufficiently tight to be informative from a policy perspective, with C.I. spans in the £7 (for Tb.low and K.tr) to £27 (for K.adj) range[13]

Smoothing analysis

Table 6 gives smoothing diagnostics akin to those shown for the simulation exercise (Table 3). As is evident from the table, over 85% of individual WTP predictions were generated without any need to enforce monotonicity via ad-hoc smoothing for both development

[13]Naturally, to what extent these ranges still allow for a clear decision rule in, say, a Benefit-Cost Analysis context will depend on the actual policy question.

scenarios. Of the remaining cases, the majority only requires a single adjustment (= removal of a single bid-prob(YES) pair), with very few observations necessitating a larger number of interventions. Nonetheless, the analyst needs to decide if cases with "excessive smoothing" should be dropped from estimation, and where this elimination threshold should be located. For our analysis, we excluded all cases with > 3 monotonicity violations, which implies an attrition of 47 observations for S1, and 36 observations for S2. We perform a sensitivity analysis to this rule, and compute individual and aggregate expected WTP for different elimination thresholds. Results are given in Table A6 of the online appendix for elimination thresholds of "drop none" to "drop if > 0 monotonicity violations." As is obvious from the table, our RFNP estimates are generally robust to these decisions, with the Tb.low and K.tr mean WTP estimates remaining essentially invariant, and changes in K.adj estimates remaining in the single-digit range.

For the latter, these subtle changes are largely driven by the (inadvertent) removal of extreme right-hand tail outliers that were "caught" in the smoothing rule. This is clearly depicted in Figure A7 of the online appendix, which gives box-and-whisker plots for all smoothing scenarios (and policy scenario S1). Interquartile ranges (IQRs, shown as a (very tight) box), and even the 1.5 times IQR bounds (shown as horizontal lines) remain essentially unchanged across the different interventions. The only visible change in the distribution of WTP is the reduction in extreme outlier points going from "drop none" to "drop if any violation." Overall, we conclude that monotonicity violations and corresponding smoothing requirements are a second-order concern for our empirical application.

Variable importance

While not of central importance for this study, we give a brief glimpse at RFs' ability to discern the influence of individual covariates on the outcome of interest, in our case the binary response to the choice question.[14] A list of such "post-hoc interpretability" measures is given in Greenwell (2022), ch. 6. Here we consider the perhaps simplest of these metrics,

[14] In theory, one could extrapolate from there to determine an individual covariate's effect on WTP, but we leave that to future research.

commonly labeled as Variable Importance (VI) score, as it is a standard feature of the `grf` package (Tibshirani et al., 2024a,b). It simply keeps track how often a given explanatory variable is chosen as the splitting variable in the constructions of underlying trees. This score is then normalized to add to one across all covariates.

Figure A8 in the online appendix shows VI scores for the ten most "influential" variables in our RF. Is is evident from the figure, the bid amount dominates this list, accounting for over 25% of all splitting occurrences. The only other two variables that achieve a VI score of 0.1 or higher are `occ_plan` (1 = individual works in a planning-related occupation) and `env_memb` (1 = membership with an environmental organization). Without going into deeper discussion on variable-specific effects, we simply not that the heightened sensitivity of our forests to the bid amount is consistent with their ability to produce generally well-behaved bid-prob(YES) pairs feeding into the nonparametric portion of our estimators.

Conclusion

This study gives a first example how powerful ML tools such as RFs can be directly incorporated into the economic valuation of environmental assets and services based on SP methods. Focusing on the "gold standard" case of CV, we illustrate how RF estimation of choice probabilities can be combined with well-known nonparametric methods to generate individual-specific welfare measures. Most importantly, these can be obtained without the need to specify an explicit IUF or WTP function, and without having to choose a distribution for error terms. In a simulation exercise we show that our RFNP estimators are competitive with a standard Logit model even under correct specification of the WTP function, and clearly out-perform the parametric model when nonlinearities are introduced into the data generating mechanism. Using biodiversity enhancement in the UK as an empirical example, we show that the RFNP can generate reasonable estimates of individual and sample-averaged WTP, with sufficient asymptotic precision to be informative for policy applications. Compared to traditional nonparametric methods that typically only produce a single WTP estimate for the sample at large, our RFNP models map out the entire distri-

bution of individual WTP predictions, and thus allows for a closer look at distributional and equity implications of envisioned policy interventions. As an added bonus, our approach produces asymptotically guaranteed standard errors and confidence intervals, thanks to recent developments in deriving the asymptotic properties of RF estimates (Wager and Athey, 2018; Athey et al., 2019).

Naturally, there are trade-offs. Specifically, our fully nonparametric framework comes at the cost of having to impose (occasional) monotonicity corrections to assure acceptance probabilities of the policy scenario decrease with increasing bid levels, and, for our K-based versions, having to choose or derive a cut-off bid that "closes" the survival function for individual WTP (see Figure A2 in the online appendix). We find that smoothing interventions are a second-order concern for our empirical application, whereas handling the tail of the survival function has a more profound influence on WTP estimates. However, our Tb.low and K.tr estimates are not affected by these cut-off choices. They therefore offer a *guaranteed lower bound* (Tb.low) and a less conservative *plausible lower bound* (K.tr) of individual-level WTP. Our K.adj. estimator, in turn illustrates the ability of our framework to allow for long tails in individual WTP without imposing this assumption on the entire sample. It utilizes all known pairs of bids and YES-probabilities in the determination of the cut-off value, and thus follows recent "best-practice" recommendations in the literature (Richardson and Lewis, 2022; Lewis et al., 2024).

We consider our proposed framework a starting point that opens doors for extensions along multiple dimensions. For example, one could consider different RF variants in the first stage of our analysis, such as the Local Linear Forest (LLF) proposed in Friedberg et al. (2021) and applied in Johnston and Moeltner (2024), Boosted Regression Forests, or Classification Forests with different splitting rules (Tibshirani et al., 2024a,b). A logical next step would also be to compare RFNP estimators to more complex versions of binary parametric models, perhaps with built-in nonlinearities, different error distributions, or a built-in model search, as in Johnston et al. (2023). It may also be fruitful to explore combinations of RFs with some of the nonparametric estimators mentioned in the introduction section, other than the Turnbull or Kriström. Naturally, additional stress-tests of the RFNP

in other empirical contexts would be another logical extension.

On a final note, our RFNP framework is user-friendly and simple to implement given existing software packages. It is also computationally fast, at least for data sets of moderate size (say a few thousand observations), as they are typically encountered in SP research. This gives the analyst a lot of flexibility to experiment with "best judgment" tuning interventions as they may be needed in a given context, at reasonable computational costs. Overall, we believe the RFNP approach can be an attractive alternative or complement to parametric processing of CV data. Our first results, as reported in this study, are certainly encouraging.

References

Andarge, T., Ji, Y., Keeler, B., Keiser, D., McKenzie, C., 2024. Environmental justice and the clean water act: Implications for economic analyses of clean water regulations. Environmental and Energy Policy and the Economy 5, 70–126.

Athey, S., Tibshirani, J., Wager, S., 2019. Generalized random forests. The Annals of Statistics 47, 1148–1178.

Athey, S., Wager, S., 2019. Estimating treatment effects with causal forests: An application. Observational Studies 5, 36–51.

Banzhaf, S., Ma, L., Timmins, C., 2019a. Environmental justice: Establishing causal relationships. Annual Review of Resource Economics 11, 377–398.

Banzhaf, S., Ma, L., Timmins, C., 2019b. Environmental justice: The economics of race, place, and pollution. Journal of Economic Perspectives 33, 185–208.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Cameron, C., Trivedi, P., 2007. Microeconometrics. Cambridge University Press.

Champ, P., Boyle, K., Brown, T., 2017. A Primer on Nonmarket Valuation. Springer. 2 edition.

Chen, H., Randall, A., 1997. Semi-nonparametric estimation of binary response models with an application to natural resource valuation. Journal of Econometrics 76, 323–340.

Cohen, J., Moeltner, K., Reichl, J., Schmidthaler, M., 2016. Linking the value of energy reliability to the acceptance of energy infrastructure: Evidence from the EU. Resource and Energy Economics 45, 124–143.

Cosslett, S., 1983. Distribution-free maximum likelihood estimator of the binary choice model. Econometrica 51, 765–782.

Creel, M., Loomis, J., 1997. Semi-nonparametric distribution-free dichotomous choice contingent valuation. Journal of Environmental Economics and Management 32, 341–358.

Crooker, J., Herriges, J., 2004. Parametric and semi-nonparametric estimation of willingness-to-pay in the dichotomous choice contingent valuation framework. Environmental and Resource Economics 27, 451–480.

Faccioli, M., Tingley, D., Mancini, M., Bateman, I., 2024. Who should benefit from environmental policies? Social preferences and nonmarket values for the distribution of environmental improvements. American Journal of Agricultural Economics , 1–27.

Fernández-Delgado, M., Cernadas, E., Barro, S., 2014. Do we need hundreds of classifiers to solve real-world classification problems? Journal of Machine Learning Research 15, 3133–3181.

Freeman, A.I., Herriges, J., Kling, C., 2014. The Measurement of Environmental and Resource Values: Theory and Methods. Resources for the Future Press. 3rd edition edition.

Friedberg, R., Athey, S., Tibshirani, J., Wager, S., 2021. Local linear forests. Journal of Computational and Graphical Statistic 30, 503–517.

Glenk, K., Meyerhoff, J., Colombo, S., Faccioli, M., 2024. Enhancing the face validity of choice experiments: A simple diagnostic check. Ecological Economics 221, 108160.

Greene, W., 2012. Econometric Analysis. Pearson / Prentice Hall. 7th edition edition.

Greenwell, B., 2022. Tree-based methods for statistical learning in R. CRC Press / Taylor $ Francis group.

Haab, T., McConnell, K., 1997. Referendum models and negative willingness to pay: Alternative solutions. Journal of Environmental Economics and Management , 251–270.

Haab, T., McConnell, K., 2003. Valuing Environmental and Natural Resources: The Econometrics of Non-market Valuation. Edward Elgar Publishing.

Hanemann, W., 1984. Welfare evaluations in contingent valuation experiments with discrete responses. American JOurnal of Agricultural Economics , 332–341.

Harding, M., Lamarche, C., 2021. Small steps with big data: Using machine learning in energy and environmental economics. Annual Review of Resource Economics 13, 469–488.

Hastie, T., Tibshirani, R., Friedman, J., 2017. The elements of statistical learning. Springer.

Hino, M., Benami, E., Brooks, N., 2018. Machine learning for environmental monitoring. Nature Sustainability 1, 583–588.

Johnston, R., Boyle, K., Adamowicz, J., Bennett, J., Brouwer, R., Cameron, T., Hanemann, W., Hanley, N., Ryan, M., Scarpa, R., Tourangeau, R., Vossler, C., 2017. Contemporary guide for stated preference studies. Journal of the Association of Environmental and Resource Economists 4, 319–405.

Johnston, R., Moeltner, K., 2024. Random Forests for benefit transfer. Paper presented at the Social Cost of Water Pollution Workshop, Washington, D.C., Oct. 2-4, 2024.

Johnston, R., Moeltner, K., Peery, S., Ndebele, T., Yao, Z., Crema, S., Wollheim, E., Besedin, E., 2023. Spatial dimensions of water quality value in New England river networks. Proceedings of the National Academy of Sciences 120, e2120255119.

Johnston, R., Rolfe, J., Rosenberger, R., Brouwer, R., 2015. Benefit transfer of environmental and resource values. Springer.

Kriström, B., 1990. A non-parametric approach to the estimation of welfare measures in discrete response valuation studies. Land Economics 66, 135–139.

Lewis, L., Richardson, L., Whitehead, J., 2024. Fat tails, flat tails, and willingness to pay: Kriström 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.

Liu, B., Bryson, J., Sevinc, D., Cole, M., Elliott, R., Bartington, S., Bloss, W., Shi, Z., 2023. Assessing the impacts of Birmingham's clean air zone on air quality: Estimates from a machine learning and synthetic control approach. Environmental and Resource Economics 86, 203–231.

Miller, S., 2020. Causal forest estimation of heterogeneous and time-varying environmental policy effects. Journal of Environmental Economics and Management 103, 102337.

Mink, S., Loginova, D., Mann, S., 2024. Wolves' contribution to structural change in grazing systems among Swiss alpine summer farms: The evidence from causal random forest. Journal of Agricultural Economics 75, 201–217.

Moeltner, K., Balukas, J., Besedin, E., Holland, B., 2019. Waters of the United States: Upgrading wetland valuation via benefit transfer. Ecological Economics 164, 106336.

Moeltner, K., Puri, R., Johnston, R., Besedin, E., Balukas, J., Le, A., 2023. Locally-weighted meta-regression and benefit transfer. Journal of Environmental Economics and Management 121, 102871.

National Oceanic and Atmospheric Administration, 1993. Natural resource damage assessment under the oil pollution act of 1990. National Register 108, 4601–4614.

Parsons, G., Myers, K., 2016. Fat tails and truncated bids in contingent valuation: An application to an endangered shorebird species. Ecological Economics 129, 210–219.

Phaneuf, D., Requate, T., 2017. A course in Environmental Economics: Theory, Policy, and Practice. Cambridge University Press. 1st edition edition.

Prest, B., Whichman, C., Palmer, K., 2023. RCTs against the machine: Can machine learning prediction methods recover experimental treatment effects? Journal of the Association of Environmental and Resource Economists 10, 1231–1264.

Ready, R., Hu, D., 1995. Statistical approaches to the fat tail problem for dichotomous choice contingent valuation. Land Economics 71, 491–499.

Richardson, L., Lewis, L., 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.

Stetter, C., Mennig, P., Sauer, J., 2022. Using machine learning to identify heterogeneous impacts of agri-environmental schemes in the EU: A case study. European Review of Agricultural Economics 49, 723–759.

Storm, H., Baylis, K., Heckelei, T., 2020. Machine learning in agricultural and applied economics. European Review of Agricultural and Applied Economics 47, 849–892.

Tibshirani, J., Athey, S., Friedberg, R., Hadad, V., Hirshberg, D., Miner, L., Sverdrup, E., Wager, S., Wright, M., 2024a. Package 'grf': Generalized random forests. R package version 2.3.2.

Tibshirani, J., Athey, S., Sverdrup, E., Wager, S., 2024b. The GRF algorithm. Web site: https://grf-labs.github.io/grf/REFERENCE.html, last accessed 2024-10-11.

Train, K., 2009. Discrete Choice Methods with Simulation. Cambridge University Press. 2nd edition.

Valente, M., 2023. Policy evaluation of waste pricing programs using heterogeneous causal effect estimation. Journal of Environmental Economics and Management 117, 102755.

Wager, S., Athey, S., 2018. Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association 113, 1228–1242.

Watanabe, M., 2010. Nonparametric estimator of mean willingness to pay from discrete response valuation data. American Journal of Agricultural Economics 92, 1114–1135.

Watanabe, M., Asano, K., 2009. Distribution free consistent estimation of mean WTP in dichotomous choice contingent valuation. Environmental and Resource Economics 44, 1–10.

Whitehead, J., 2017. Who knows what willingness to pay lurks in the hearts of men? A rejoinder to Egan, Corrigan, and Dwyer. Econ Journal Watch , 346–361.

Zapata, S., Carpio, C., 2024. Distribution-free methods to estimate willingness-to-pay models using discrete response valuation data. Journal of Agricultural and Resource Economics 49, 39–62.

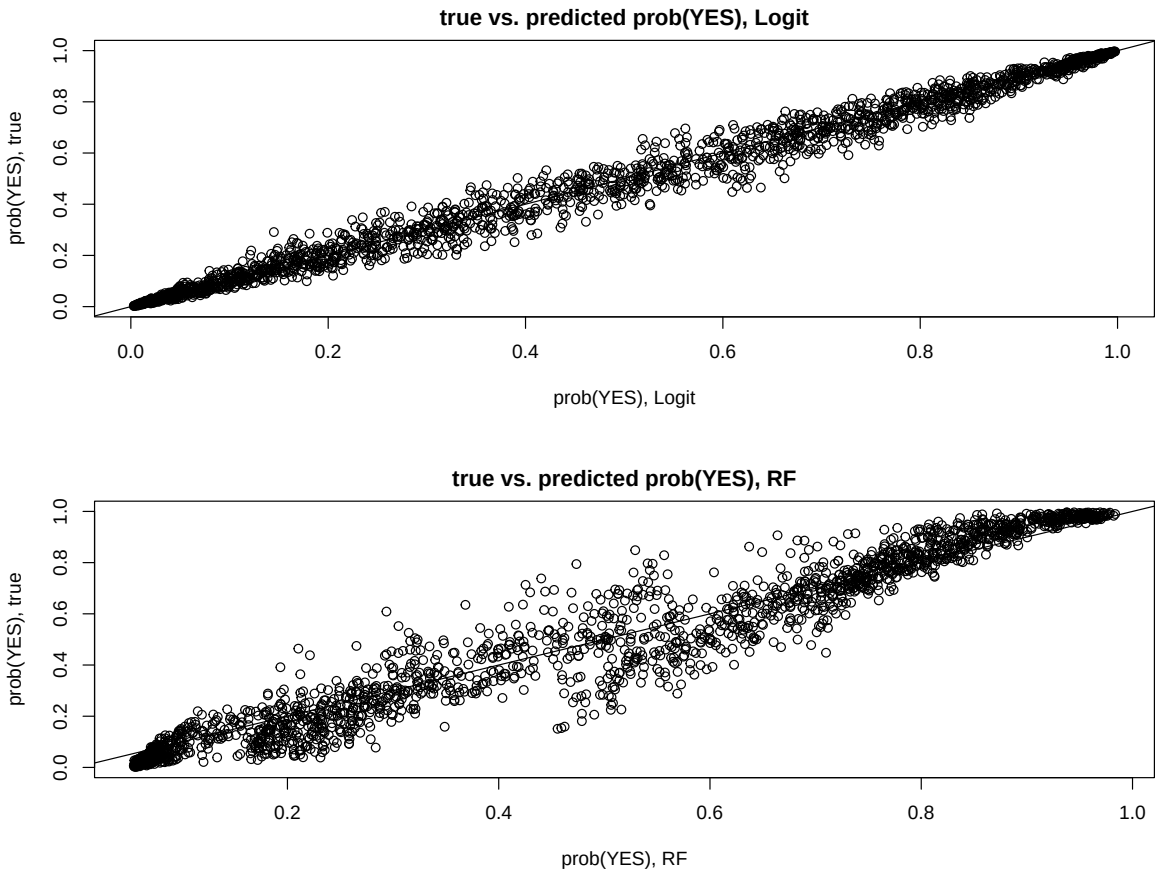

Figure 1: true vs. predicted prob(YES), linear model (M1)

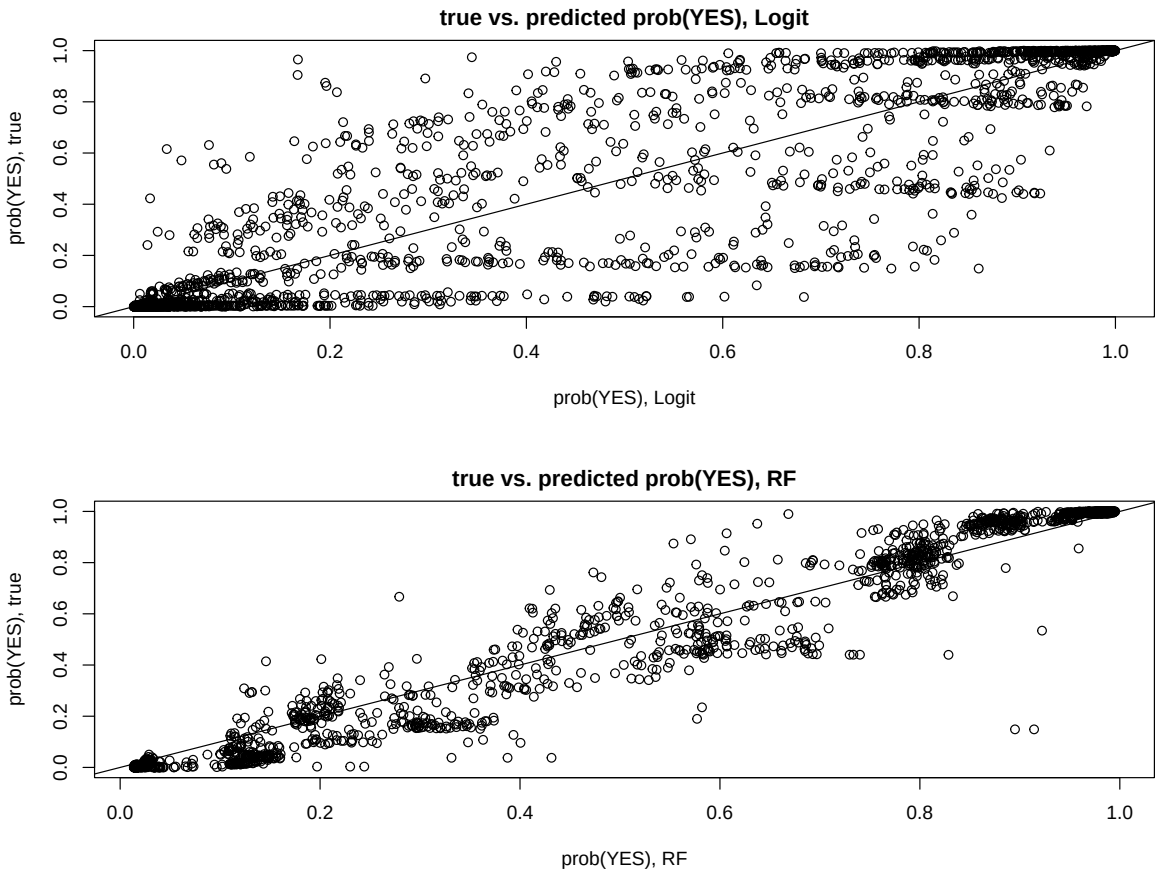

Figure 2: true vs. predicted prob(YES), piecewise-linear model (M2)

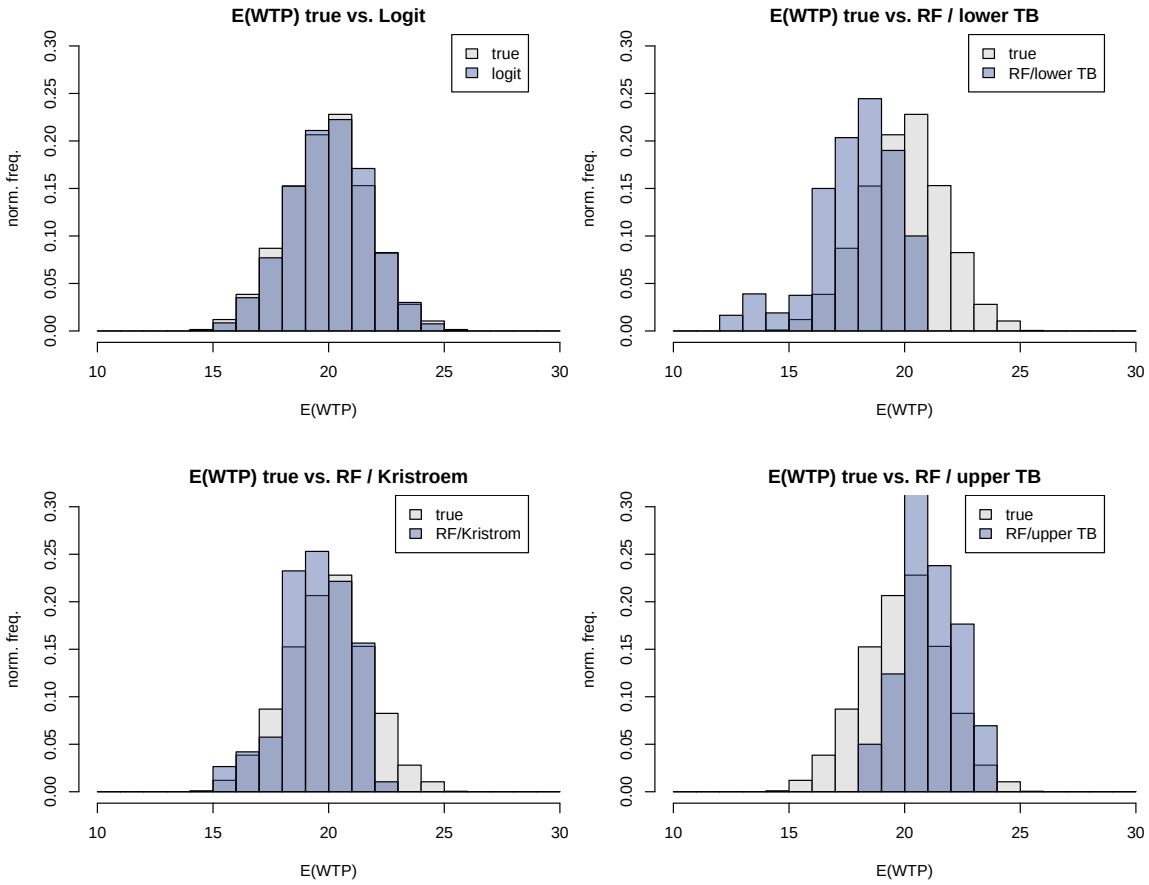

Figure 3: true vs. predicted WTP, linear model (M1)

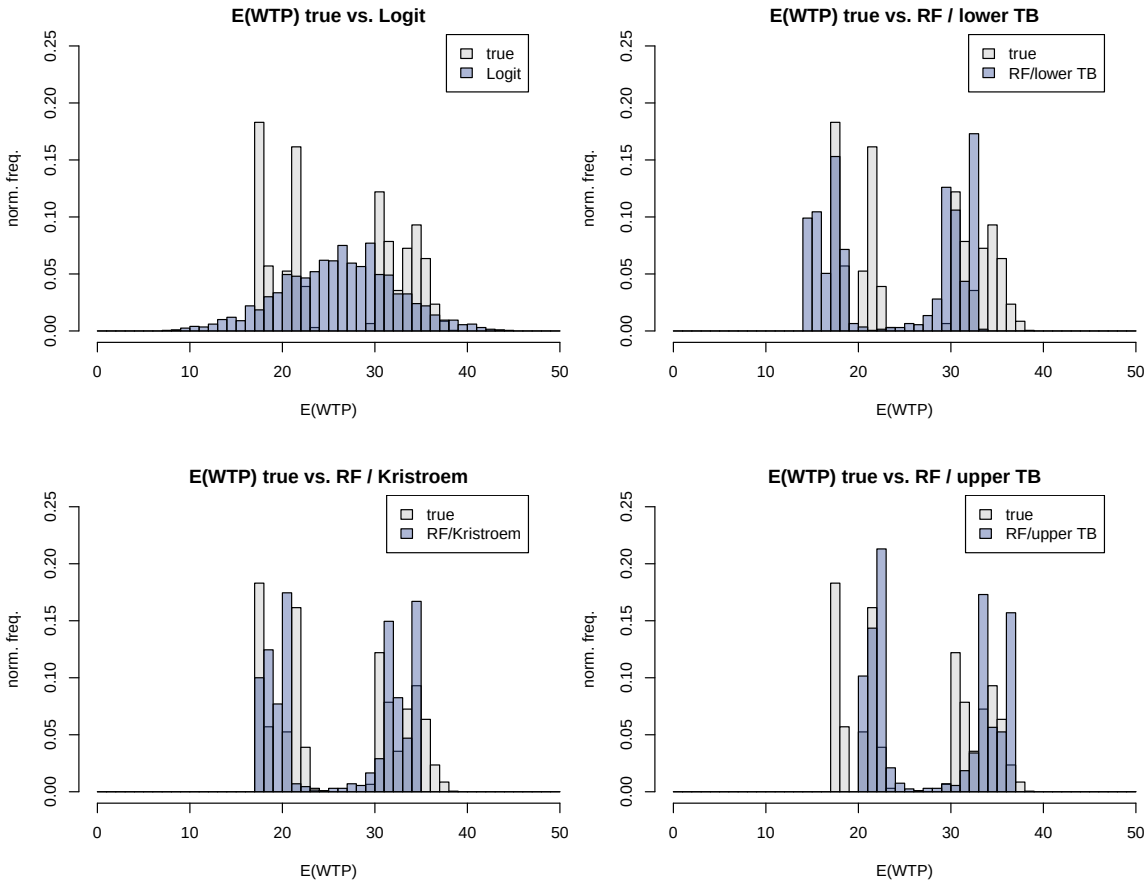

Figure 4: true vs. predicted WTP, piecewise-linear model (M2)

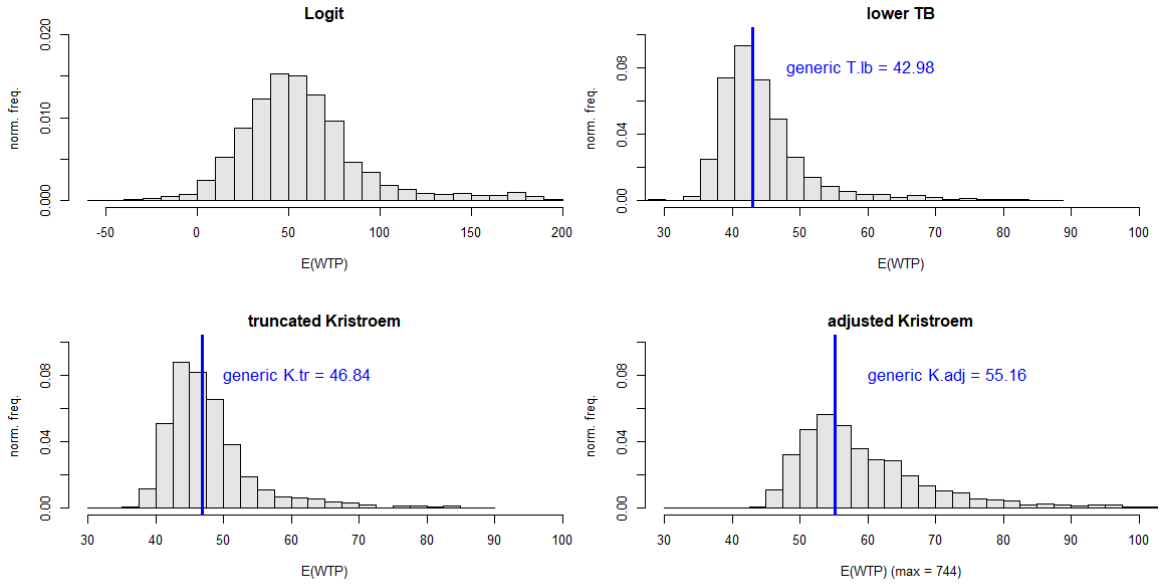

Figure 5: predicted indiv. WTP, biodiversity application, scen. S1

Blue lines and text capture the corresponding generic nonparametric estimators based on sample proportions.

Table 1: Simulation results for test sample

estimator	mean	std	min	max	range	mse	mape
			linear (M1)				
true E(wtp)	19.957	1.758	14.572	25.550	10.978	0.000	0.000
Logit	20.012	1.706	14.885	25.612	10.727	0.229	0.019
Turnbull.low	17.893	1.778	12.477	20.986	8.509	4.830	0.104
Kriström	19.487	1.467	15.488	22.237	6.749	0.656	0.030
Turnbull.up	21.082	1.193	18.499	23.528	5.029	1.844	0.063
			piecewise-linear (M2)				
true E(wtp)	26.397	6.912	17.509	38.550	21.041	0.000	0.000
Logit	26.205	6.082	7.425	44.327	36.903	16.037	0.142
Turnbull.low	23.772	7.216	14.158	33.032	18.874	8.669	0.113
Kriström	26.050	6.795	17.372	34.843	17.472	1.764	0.039
Turnbull.up	28.329	6.377	20.585	36.683	16.097	5.644	0.089
			piecewise-nonlinear (M3)				
true E(wtp)	26.579	5.294	16.997	33.465	16.467	0.000	0.000
Logit	26.624	4.651	8.120	42.156	34.036	11.526	0.108
Turnbull.low	23.673	5.942	14.573	30.447	15.874	9.822	0.119
Kriström	25.809	5.475	18.301	32.035	13.735	1.422	0.038
Turnbull.up	27.945	5.024	22.020	33.650	11.630	2.736	0.060

std = standard deviation
min / max = minimum / maximum
mse = mean squared error
mape = mean absolute percentage error

Table 2: Simulation results for quartile points

estimator	first quartile estimate	s.e.	lower	upper	median estimate	s.e.	lower	upper	third quartile estimate	s.e.	lower	upper
					linear							
true E(wtp)	18.373	0.000	0.000	0.000	19.989	0.000	0.000	0.000	21.624	0.000	0.000	0.000
Logit	18.649	0.270	18.247	19.319	20.040	0.122	19.909	20.402	21.466	0.269	21.025	22.044
Turnbull.low	16.955	0.823	14.098	17.378	18.476	0.500	17.272	19.208	19.961	0.469	18.307	20.134
Kristrom	18.529	0.581	16.607	18.862	19.709	0.382	18.923	20.434	21.107	0.380	19.953	21.367
Turnbull.up	20.103	0.360	19.073	20.450	20.941	0.295	20.601	21.715	22.254	0.326	21.504	22.739
					piecewise-linear							
true E(wtp)	17.773	0.000	0.000	0.000	32.989	0.000	0.000	0.000	34.624	0.000	0.000	0.000
Logit	21.028	0.569	20.000	22.196	26.323	0.238	25.903	26.840	31.676	0.544	30.708	32.884
Turnbull.low	14.915	1.099	12.062	16.334	24.571	1.601	21.177	27.552	32.510	0.517	30.907	32.906
Kristrom	17.954	0.743	16.024	18.956	26.690	1.511	23.707	29.486	34.378	0.510	32.734	34.732
Turnbull.up	20.993	0.453	19.896	21.706	28.809	1.446	25.982	31.590	36.247	0.520	34.569	36.571
					piecewise-nonlinear							
true E(wtp)	31.680	0.000	0.000	0.000	22.352	0.000	0.000	0.000	21.829	0.000	0.000	0.000
Logit	29.684	0.433	28.793	30.476	26.538	0.200	26.139	26.888	23.364	0.438	22.464	24.192
Turnbull.low	30.051	0.375	29.109	30.526	24.597	1.525	20.368	26.229	18.713	0.827	15.978	19.140
Kristrom	31.606	0.306	30.908	32.084	26.424	1.373	22.560	28.009	20.896	0.529	19.164	21.241
Turnbull.up	33.161	0.302	32.501	33.698	28.252	1.235	24.927	29.815	23.080	0.298	22.256	23.422

s.e. = standard error
lower / upper = lower / upper bound of 95% confidence interval
boxed mean = closest to true E(wtp)
shaded confidence bounds: contains true E(wtp)

Table 3: Smoothing diagnostics, simulation

	M1		M2		M3	
adjustments	count	%	count	%	count	%
0	1933	96.65%	865	43.25%	1514	75.70%
1	67	3.35%	744	37.20%	318	15.90%
2	0	0.00%	391	19.55%	84	4.20%
3	0	0.00%	0	0.00%	84	4.20%
4	0	0.00%	0	0.00%	0	0.00%
5	0	0.00%	0	0.00%	0	0.00%
6	0	0.00%	0	0.00%	0	0.00%
7	0	0.00%	0	0.00%	0	0.00%
total	2000	100.00%	2000	100.00%	2000	100.00%

Table 4: WTP predictions, individual level

estimator	mean	std	min	max	range
			Policy scenario S1		
Logit	57.071	34.478	-58.556	193.388	251.943
Turnbull.low	44.570	7.059	25.277	84.158	58.880
Kriström, trunc.	47.955	6.857	35.688	84.435	48.747
Kriström, adj.	69.407	50.324	37.450	744.082	706.632
K.adj. cutoff	£183	£101	£104	£1,620	£1,516
			Policy Scenario S2		
Logit	38.443	34.478	-77.183	174.760	251.943
Turnbull.low	43.144	7.241	25.137	87.794	62.658
Kriström, trunc.	46.599	7.153	33.056	87.967	54.910
Kriström, adj.	68.676	56.940	34.161	966.945	932.784
K.adj. cutoff	£184	£114	£97	£2,021	£1,924

std = standard deviation
min / max = minimum / maximum
range = (max - min)
Turnbull.low = lower bound Turnbull estimator
Kriström, trunc. = truncated Kriström estimator
Kriström, adj. = adjusted Kriström estimator
K.adj.cutoff = estimated cutoff bid for adjusted Kriström

Table 5: WTP predictions, sample mean

estimator	mean	s.e.	low	up	range
			Policy scenario S1		
Logit	57.071	5.960	46.442	68.457	22.015
Turnbull.low	44.570	2.190	38.062	45.982	7.920
Kriström, trunc.	47.955	1.997	42.024	49.422	7.398
Kriström, adj.	69.407	6.611	56.405	82.303	25.898
			Policy Scenario S2		
Logit	38.443	3.481	31.027	44.929	13.901
Turnbull.low	43.144	2.232	35.522	44.029	8.507
Kriström, trunc.	46.599	2.015	39.923	47.195	7.271
Kriström, adj.	68.676	6.658	53.037	79.777	26.739

s.e. = standard error
low / up = lower / upper bound of 95% confidence interval
range = (upper - lower)
Turnbull.low = lower bound Turnbull estimator
Kriström, trunc. = truncated Kriström estimator
Kriström, adj. = adjusted Kriström estimator

Table 6: Smoothing diagnostics, application

	scenario S1		scenario S2	
adjustments	count	%	count	%
0	2753	85.95%	2768	86.42%
1	336	10.49%	313	9.77%
2	42	1.31%	45	1.40%
3	25	0.78%	41	1.28%
4	27	0.84%	24	0.75%
5	20	0.62%	12	0.37%
6	0	0.00%	0	0.00%
7	0	0.00%	0	0.00%
total	3203	100.00%	3203	100.00%