Random Forests for Benefit Transfer

Robert J. Johnston

Department of Economics and George Perkins Marsh Institute, Clark University
R.Johnston@clarku.edu

Klaus Moeltner

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

October 25, 2024

Abstract

Benefit Transfer (BT) has evolved as the dominant non-market valuation method for large-scale environmental benefit-cost analyses, including those required of U.S. federal agencies. Yet, even best-practice approaches for BT based on Meta-Regression Models (MRMs) typically exhibit poor predictive fit and out-of-sample efficiency. This article introduces Random Forests (RFs) for nonparametric estimation of MRMs and construction of BT predictions. We compare the performance of a variety of RF models to current best practice approaches for BT, including a globally-linear MRM and Locally-Weighted MRM (LWR). We find that forest-based models substantially improve the within-sample accuracy of welfare predictions and tighten confidence intervals of predicted benefits for out-of-sample transfers. The best-performers reside within the family of Local Linear Forests (LLFs), essentially a hybrid approach that combines elements of RFs and LWR. We also examine the utility-theoretic properties of each specification. Results suggest that this new approach has the potential to substantially improve BT accuracy for environmental policymaking without sacrificing theoretic properties, while simultaneously reducing econometric and computational difficulties relative to leading alternatives.

keywords: benefit-cost analysis; environmental policy; water quality; machine learning

Suggested citation:
Paper presented at the Social Cost of Water Pollution Workshop, Washington, D.C., Oct. 2-4, 2024.

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

Introduction

Prospective or ex ante benefit cost analysis (BCA) is a cornerstone of U.S. environmental policymaking (Aldy et al., 2021). As mandated in Executive Orders since 1981 and outlined in Office of Management and Budget (OMB) Circulars A-4 and A-94, U.S. federal agencies are required to implement BCA for economically significant regulations and programs, including regulatory policies under the U.S. Clean Water Act (Office of the Federal Register, 1981, 1993; The White House, 2011; U.S. Environmental Protection Agency, 2024b).

However, quantifying environmental benefits to inform decision-making faces numerous practical challenges (Griffiths et al., 2012; Newbold et al., 2018b). Among these, a recurring concern has been an inability to reliably capture non-market environmental benefits. This is particularly true for large-scale water quality improvements wherein benefits are diffuse, spatially heterogeneous, and conditional on localized circumstances (Johnston et al., 2017; Keiser and Shapiro, 2019; Keiser et al., 2019). Yet, the capacity to accurately measure spatially explicit benefits can be a decisive factor in whether environmental policies and programs pass a benefit-cost test (Keiser, 2019).

For practical purposes, environmental benefit estimation within large-scale BCA almost universally requires benefit transfer (BT), characterized as the use of pre-existing empirical estimates of value from similar research settings (Griffiths et al., 2012; Johnston et al., 2021). As noted by Newbold et al. (2018b), p. 469, within the context of U.S. Environmental Protection Agency (EPA) policy analysis, "*it is impossible to conduct a prospective BCA without the use of at least some form of benefit (and cost) transfers.*" This reality is further acknowledged within updated guidance found in OMB Circulars A-4 and A-94, which gives increased emphasis and sanction to these methods.[1]

Yet there remains an unresolved and frequently acknowledged tension between the unavoidable use of BT for environmental benefits estimation within large-scale BCAs and the

[1] The earlier 2003 version of OMB Circular A-4 stated that BT should "*be treated as a last-resort option and not used without explicit justification.*" This language has now been deleted and replaced with a statement that "*benefit transfer methods are appropriate when more direct and specific valuations are unavailable or inferior, or when time, resources, or other constraints do not permit conducting studies specific to the regulatory context*" (Office of Management and Budget, 2023, p.37).

observation that even best-practice BTs still remain, on average, relatively inaccurate, as assessed by statistical fit to the underlying data (Newbold et al., 2018b; Johnston et al., 2021). This concern has been acknowledged since the establishment of BT as a formal subfield of environmental economics research (Brookshire and Neill, 1992), and remains true despite continued methodological advances and incremental improvements in accuracy and reliability (Newbold et al., 2018b; Johnston et al., 2021; Moeltner et al., 2023).

Existing best-practices for BT

In recent years, Meta-Regression Modeling (MRM) has emerged as a preferred route for BT by U.S. federal agencies (Griffiths et al., 2012; Newbold et al., 2018b). In essence, this method synthesizes data over many prior valuation studies in ways that collectively represent conditions at the "policy sites" (or contexts) for which value estimates are required (Johnston and Rosenberger, 2010; Kaul et al., 2013; Johnston et al., 2021). It lies at the core of benefit-estimation platforms such as U.S. EPA's prototype *BenSPLASH* model for water quality valuation (Corona et al., 2020).[2] From an econometric perspective, a valuation MRM can be interpreted as a secondary regression model, with comparable, study-specific welfare estimates on similar environmental goods and services as the outcome variable, and study-specific features and other observables (e.g., geospatial information, demographics, etc.) on the "right-hand-side." Advantages of MRMs for BT include an often wide geographic coverage of the supporting metadata, ability to capture study-level heterogeneity, and flexibility in accommodating ancillary spatially-explicit data, as described, *inter alia*, in Rolfe et al. (2015), Johnston et al. (2021), and Moeltner et al. (2023).

Parallel to the increased use of MRM for BT, considerable efforts have been undertaken to improve the accuracy (as gauged via cross-validation on the meta-sample itself) and efficiency (as measured in terms of predictive confidence bounds) of this method (Johnston et al., 2018). Examples include expansion of the metadata (Moeltner et al., 2023),

[2] Examples of recent EPA rulemaking based on BT-via-MRM include effluent guidelines for the construction and development category (U.S. Environmental Protection Agency, 2009), water quality standards for nutrients in lakes and rivers in Florida (U.S. Environmental Protection Agency, 2010), effluent limitations for the steam electric power generating sector (U.S. Environmental Protection Agency, 2015, 2020, 2024a), and effluent guidelines for the meat and poultry sector (U.S. Environmental Protection Agency, 2023).

inclusion of spatially explicit variables not captured in the original studies (Johnston et al., 2017, 2019), and incorporation of functional relationships to assure compliance of BT estimates with utility-theoretic considerations (Kling and Phaneuf, 2018; Newbold et al., 2018a; Moeltner, 2019). Yet despite these advances, the predictive properties of even best-practice MRMs remain arguably unsatisfactory for many applications, with large Cross-Validation (CV) errors against the actual metadata, and wide confidence intervals for BT predictions (Moeltner et al., 2023).

In the latest attempt to address these shortcomings, Moeltner et al. (2023) propose a locally-weighted version of the MRM, which they label Locally-Weighted MRM, or LWR. The key feature of the LWR is that the characteristics of the target policy context are now incorporated twofold in the prediction process: (i) via combination of policy settings with estimated coefficients, as in the generic, or Globally-Linear MRM (GL-MRM), and (ii) via assignment of weights to each meta-observation. These weights, in turn, capture how closely related to the policy point each actual observation is, in the Euclidean sense. Points more similar to the policy context receive a larger weight in the final regression step, while more distant points are down-weighted or completely eliminated from the data. Moeltner et al. (2023) show that this approach brings substantial gains in both predictive accuracy and efficiency compared to the GL-MRM.

However, these accuracy gains come with considerable implementation challenges. The LWR requires a time-consuming search for optimal weight settings and is susceptible to rank violations as observations are dropped from estimation as part of the weight construction process. Weights must be re-configured and a local regression re-estimated for each policy context. This can be cumbersome for large-scale applications, involving potentially tens of thousands of predictive contexts. This juxtaposition of potential accuracy gains with amplified practical difficulty exemplifies, at least arguably, the most important contemporary question for BT — *how can accuracy be advanced while maintaining methods that are feasible within applied BCA settings?*

A logical next step: Random Forests for BT

Addressing this challenge head-on, this article develops and evaluates a novel alternative to MRM BT that achieves both goals simultaneously — enhanced ease of application and (steeply) increased accuracy relative to extant methods. Specifically, we adapt a Machine Learning (ML) tool know as Random Forests (RFs), and a recently developed variant thereof labeled Local Linear Forests (LLFs, Friedberg et al. (2021)) to process valuation metadata in ways that enable (much) more efficient, accurate and straightforward BT value predictions, compared to current best-practice MRMs.

Random Forests were first introduced to the ML literature by Breiman (2001). As discussed in Hastie et al. (2017), Harding and Lamarche (2021), and Storm et al. (2020), RFs are among the most powerful and effective prediction techniques. They can detect highly nonlinear relationships, are robust to non-normality and outliers, provide algorithmic treatment of missing data, require little in terms of pre-processing or tuning, and are computationally less demanding than alternative ML approaches such as Neural Networks (Fernández-Delgado et al., 2014). Yet, they have only very recently entered the realm of environmental and resource economics, primarily as an alternative approach to estimate causal treatment effects in the policy evaluation literature (Miller, 2020; Harding and Lamarche, 2021; Stetter et al., 2022; Liu et al., 2023; Valente, 2023; Prest et al., 2023; Mink et al., 2024).

Given their singular focus on prediction and the high stakes of "getting it right," BT problems represent an ideal laboratory for ML-type solutions, such as RFs. It is perhaps surprising that this integration has not previously been formalized. We are aware of only one published article in the environmental economics literature that exploits the predictive strength of RFs, within the context of detecting industrial water pollution violations (Hino et al., 2018). Ours is thus one of the very few contributions that use RFs, or, for that matter, any ML-based approach in the environmental and resource economics literature, and the first to introduce ML / RFs to the MRM-BT realm.[3] To our best knowledge, it is

[3] In agricultural economics, Sun et al. (2024) use a generic RF for a meta-analysis of meat preferences. In contrast to our study, they do not provide any econometric underpinnings that illustrate common ground

also one of the first studies adapting LLFs for any applied economics context.

The proposed methods are illustrated using an updated variant of the long-established metadata in Moeltner et al. (2023), on per household WTP for water quality improvements in U.S. waterbodies.[4] This application enables direct comparison of BT performance for essentially identical metadata (including out-of-sample CV using identical policy contexts). Using these metadata, we show that RFs, and especially the LLF variants, reduce predictive error by a factor of four to five compared to the LWR, and tighten confidence intervals for BT predictions by a factor of 10 to 20 relative to the LWR. We also examine utility-theoretic properties such as the Adding-Up (AU) restriction (Newbold et al., 2018a; Moeltner, 2019), and find that all of our forests empirically satisfy AU for small, but realistic incremental quality steps. As such, the proposed methods substantially outperform even the most accurate prior method for MRM BT (LWRs), for the same illustrative metadata and BT applications. They are also much easier to apply than LWRs, obviating the need for time-consuming specification searches and ad-hoc adjustments in the presence of rank violations.

The presented approach is applicable to any regression-type context where out-of-sample predictions are of central importance, of which BT is an archetype. Most importantly, the illustrated methods represent an approach that could revolutionize the ways that BTs are implemented for large-scale BCA, in that they are both more accurate than extant approaches and can be applied in straightforward fashion via readily adaptable methods and code.[5]

and differences between regression models and forests. Furthermore, they do not employ any LLF-type forests, do not use forests to generate truly out-of-sample predictions (i.e. outside the entire metadata), and do not appear to construct forests in a manner to ascertain asymptotic properties, as is the case for all of our forest versions. The latter is especially important if confidence intervals are desired for predictive constructs, as described below in more detail. However, they do observe pronounced accuracy gains in predictive fit to the actual metadata for forests versus common regression approaches, thus mirroring our findings in that respect.

[4] This metadata has been updated and improved continuously since first published in Johnston et al. (2003) and Johnston et al. (2005). It has been repeatedly used as the foundation for U.S. EPA regulatory BCAs (Moeltner et al., 2023), along with complementary analyses such as Newbold et al. (2018a), Johnston et al. (2017), Johnston et al. (2019), Moeltner (2019), and Newbold and Johnston (2020). It is thus the most heavily applied, published, and evaluated metadata in the valuation literature.

[5] Our entire analysis is coded in R and builds on existing R packages (Tibshirani et al., 2024a,b). This code will be made accessible to the broader research community to facilitate adoption of this promising method for other BT applications.

Modeling framework

Following Moeltner et al. (2023) we depart from a baseline GL-MRM that has been found to have desirable utility-theoretic properties while affording ease of estimation. For a given observation (or source study "site") i, with $i = 1 \dots n$, it can be written as:

$$\begin{aligned} & log\left(\frac{y_i}{q_{1,i} - q_{0,i}}\right) = \mathbf{x}'_{c,i}\boldsymbol{\beta} + \mathbf{m}'_i\boldsymbol{\gamma} + \delta\left(\frac{q_{0,i} + q_{1,i}}{2}\right) + \epsilon_i, \quad \text{with} \\ & \epsilon_i \sim n\left(0, \sigma^2\right), \end{aligned} \tag{1}$$

where y_i is the source-study estimated willingness-to-pay (WTP) for a water quality change from status quo level $q_{0,i}$ to policy level $q_{1,i}$, vector $\mathbf{x}_{c,i}$ comprises explanatory variables that are related to site- or population characteristics (referred to as "context-specific" in Moeltner (2019), Moeltner et al. (2019), and Moeltner et al. (2023)), $\mathbf{m}_i$ is a vector of (typically study-specific) methodological indicators (e.g. type of elicitation method, type of payment vehicle, time horizon for payments, etc.), and ϵ_i is an i.i.d. error term that is normally distributed with mean zero and variance σ^2.

Building on this standardized approach, Moeltner et al. (2023) introduce the LWR, a local specification of the GL-MRM model that estimates a separate version of (1) at each sample location, defined as a point or set of points that share identical settings for some vector of weight variables $\mathbf{z}$, typically comprised of most or all elements of $\mathbf{x}_{c,i}$, plus (optionally) additional variables not included in the baseline regression. As described in Moeltner et al. (2023), at each location $g = 1 \dots G$ there will be one or more "home observations" with weight values $\mathbf{z}_g$. All other points in the MRM will deviate from $\mathbf{z}_g$ in one or more dimensions. These deviations (in the Euclidean sense) are then converted to weights in the $[0, 1]$ interval via choice of distance function, window size, and weight function. Details for this selection / specification process are given in Moeltner et al. (2023).

Once each point in the metadata has received a weight (with home observations carrying a weight of one), the locally-weighted version of (1) at location g can be specified as:

$$\begin{aligned} &\sqrt{w_{i,g}}\, log\left(\frac{y_i}{q_{1,i}-q_{0,i}}\right) = \sqrt{w_{i,g}}\, \left(\mathbf{x}'_{c,i}\boldsymbol{\beta}_g + \mathbf{m}'_i\boldsymbol{\gamma}_g + \delta_g\left(\frac{q_{0,i}+q_{1,i}}{2}\right)\right) + \epsilon_i, \quad \text{with} \\ &\epsilon_i \sim n\left(0, \sigma_g^2\right), \end{aligned} \tag{2}$$

where $w_{i,g}$ is the weight assigned to observation i with respect to location g. As is clear from (2), all model coefficients now carry location-specific subscripts. Error terms are typically not weighted, but receive location-specific variance σ_g^2.

In essence, the LWR can be interpreted as a semi-parametric method that lends flexibility to estimation without proliferating on parameters and corresponding degrees of freedom (Pagan and Ullah, 1999; Fotheringham et al., 2002; McMillen and Redfearn, 2010). Moeltner et al. (2023) show that the LWR can greatly sharpen predictions for actual sample points within a convergent validity setting, and reduce variance at the BT stage (when the true value is unknown) compared to the GL-MRM. Hence, the LWR is more accurate than the GL-MRM for BT, as one might anticipate. However, the identification of promising weight settings (= combination of weight variables in $\mathbf{z}$, choice of distance function, window size, and weight function) requires a computationally intensive cross-validation (CV) process. An additional challenge when working with LWRs are possible rank violations that can occur when the explanatory data matrix is "cut" to the chosen window size, and some regressors become collinear. This can easily happen in metadata with its typically small to moderate sample size, and many (often sparse) binary covariates. In addition to these practical implementation challenges, the LWR still requires the explicit specification of a base function, and a fully specified statistical distribution for the error term.

In light of these challenges, RFs may offer an attractive alternative to extract signals from metadata and form BT predictions without the need to rely on a specific regression function. As mentioned above, they have made an entry into applied economic work in recent years, where they have been found to be a useful tool to predict outcomes and, to a larger extent, estimate causal treatment effects in the policy evaluation literature.[6] Fun-

[6] These "causal forests" were developed by Wager and Athey (2018) and generalized in Athey et al. (2019). A case study-type application is given in Athey and Wager (2019).

damentally, an RF is a fully nonparametric framework that can take as input a potentially very large set of explanatory variables (generally referred to as "features" in machine learning) and combine them in a flexible fashion to explain outcomes or isolate causal effects, without making any assumptions on underlying functional relationships.[7] As we show below in more detail, they do not suffer from risk of rank violations, and do not require the ex-ante construction of location-specific weights. Instead, as discussed in Wager and Athey (2018), Athey et al. (2019), and Friedberg et al. (2021), RFs produce local weights via an adaptive neighborhood kernel that is entirely data-driven. Greenwell (2022) gives a detailed and accessible introduction to regression trees and forests with many empirical applications.

In our context, an RF model can be generically written as

$$\begin{aligned} &log\left(\frac{y_i}{q_{1,i}-q_{0,i}}\right) = g\left(\mathbf{x}_{c,i}, \mathbf{m}_i, q_{0,i}, q_{1,i}\right) + \epsilon_i, \quad \text{with} \\ &E\left(\epsilon_i | \mathbf{x}_{c,i}, \mathbf{m}_i, q_{0,i}, q_{1,i}\right) = 0, \end{aligned} \tag{3}$$

where $g\left(.\right)$ is an unspecified nonparametric function, and the only assumption required for the error term is a conditional expectation of zero, implying unconfoundedness with observed / included variables. Note that we have preserved the nonlinear transformation of the dependent variable on the left hand side of (3) for a more even-footed comparison with the other two modeling frameworks. As shown below, this transformation also supports consistency with utility-theoretic properties of estimated welfare effects.

Predictions / Benefit transfer

Perhaps the clearest way to compare these three approaches - GL-MRM, LWR, and RF - is by examining the econometric underpinnings of how they generate predictions, for example in a BT context. Letting $\tilde{y}_i = log\left(\frac{y_i}{q_{1,i}-q_{0,i}}\right)$, collecting $\mathbf{x}_{c,i}, \mathbf{m}_i, q_{0,i}, q_{1,i}$ into vector $\mathbf{x}_i$, and all model coefficients in $\boldsymbol{\theta}$, the predicted outcome $\tilde{y}_p$ at some policy point $\mathbf{x}_p$ flowing from

[7] This is particularly useful for a BT context, wherein theory provides little guidance as to the functional relationships that enable calibration of welfare estimates for differences between valuation settings or sites.

the GL-MRM can be derived as

$$
\begin{aligned}
&E\left(\tilde{y_p}|\mathbf{x}_p, \mathbf{X}, \tilde{\mathbf{y}}\right) = \mathbf{x}_p'\hat{\boldsymbol{\theta}} = \mathbf{x}_p'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\tilde{\mathbf{y}} = \\
&\sum_{i=1}^{n} r_i\left(\mathbf{x}_p, \mathbf{X}\right)\tilde{y}_i,
\end{aligned}
\tag{4}
$$

where $\mathbf{X}$ is the full matrix of explanatory variables, $\tilde{\mathbf{y}}$ is the full n by 1 outcome vector, and $r_i\left(.\right)$ is the i^{th} element of $\mathbf{x}_p'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$. Thus, the predicted construct can be interpreted as a weighted sum of outcomes for all original sample points. Each weight is a function of the entire feature matrix and the policy setting $\mathbf{x}_p$.

Applying standard transformations, WTP predictions in dollars can be obtained as

$$
\hat{y}_p = \frac{1}{T}\sum_{t=1}^{T}\left(exp\left(\mathbf{x}_{p,t}'\hat{\boldsymbol{\theta}} + log\left(q_{1,p} - q_{0,p}\right) + 0.5 * s^2\right)\right),
\tag{5}
$$

where index t refers to a specific combination of methodological indicators in $\mathbf{m}$, and s^2 is the estimated variance of the error term. As is evident from (5), the final prediction is obtained by averaging over all possible combinations of methodological settings, as originally suggested in Moeltner et al. (2007), and applied in Moeltner (2019), Moeltner et al. (2019), and Moeltner et al. (2023). This neutralizes the effect of methodological indicators, which are de facto nuisance terms in the BT step. Naturally, for within-sample predictions, say of actual meta-outcome y_i, the original settings in $\mathbf{m}_i$ are used instead.

Predictions for the LWR can be obtained in similar fashion as weighted sum of sample observations. Specifically:

$$
\begin{aligned}
&E\left(\tilde{y_p}|\mathbf{x}_p, \mathbf{X}, \mathbf{W}\left(\mathbf{z}_p\right), \tilde{\mathbf{y}}\right) = \mathbf{x}_p'\hat{\boldsymbol{\theta}}_p = \mathbf{x}_p'\left(\mathbf{X}'\mathbf{W}\left(\mathbf{z}_p\right)\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}\left(\mathbf{z}_p\right)\tilde{\mathbf{y}} = \\
&\sum_{i=1}^{n} l_i\left(\mathbf{x}_p, \mathbf{X}, \mathbf{W}\left(\mathbf{z}_p\right)\right)\tilde{y}_i,
\end{aligned}
\tag{6}
$$

where indexing $\boldsymbol{\theta}$ by subscript p highlights that a local regression tailored to home observation $\mathbf{x}_p$ was employed, $\mathbf{W}\left(.\right)$ is an n by n diagonal matrix featuring weights $w_{i,g}$, introduced in equation (2) above, and $l_i\left(.\right)$ is the i^{th} element of $\mathbf{x}_p'\left(\mathbf{X}'\mathbf{W}\left(\mathbf{z}_p\right)\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}\left(\mathbf{z}_p\right)$. As is evident from the second line in (6), observation-specific prediction weights l_i are now also a function of the entire weight matrix $\mathbf{W}(\mathbf{z}_p)$. Specifically, points closer to the weight vector for the policy setting, $\mathbf{z}_p$, receive a larger weight in this averaging process, ceteris paribus. Actual welfare predictions, in dollars, can then be obtained via (5), using local $\boldsymbol{\theta}_p$ instead of global $\boldsymbol{\theta}$.

In contrast to these prior approaches, the building blocks of an RF are a large number of underlying "trees." Each tree $b = 1 \ldots B$ operates on a bootstrapped subset of the full data. A tree is "grown" by repeatedly and sequentially splitting the data into two segments. At each splitting occasion, the tree chooses a random subset of the explanatory variables, and within that set the feature and splitting point that best satisfy some optimization criterion. For example, in a standard regression RF the splitting objective is to maximize the reduction between the mean squared error (MSE) at the "parent" node, i.e. the combined data before the split, and the combined MSE in the two "child" nodes. This process is repeated at each new node until all observations are assigned to a terminal "leaf." This then allows to form leaf-specific predictions based on the values of the *outcome variable* for all observations that share the same leaf (e.g. simple average for standard regression forests). The RF at large then generates a final predictive value for a given combination of features, say $\mathbf{x}_p$, via a weighted average of tree-specific predictions, as will be shown next in more detail. A stylized example highlighting the mechanisms of a simple regression tree is given in the online appendix.

Formally, predictions from the RF composed of trees $b = 1 \ldots B$ can be derived as:

$$\begin{aligned}
&\tilde{y}_p | \mathbf{x}_p, \mathbf{X}, \tilde{\mathbf{y}} = \frac{1}{B} \sum_{b=1}^{B} \frac{1}{|L_b(\mathbf{x}_p)|} \sum_{i=1}^{n} \tilde{y}_i \; I(\mathbf{x}_i \in L_b(\mathbf{x}_p)) = \\
&\sum_{i=1}^{n} \tilde{y}_i \frac{1}{B} \sum_{b=1}^{B} \frac{I(\mathbf{x}_i \in L_b(\mathbf{x}_p))}{|L_b(\mathbf{x}_p)|} = \\
&\sum_{i=1}^{n} \alpha_i(\mathbf{x}_p, \mathbf{X}) \, \tilde{y}_i,
\end{aligned} \tag{7}$$

where $L_b(\mathbf{x}_p)$ is the terminal leaf of tree b that contains policy point $\mathbf{x}_p$, $|L_b(\mathbf{x}_p)|$ denotes the number of original training samples that were assigned to leaf $L_b(\mathbf{x}_p)$, and $I(.)$ is an indicator function taking a value of one if the condition it describes holds, and a value of zero otherwise. The first line in (7) simply indicates that the RF prediction is an average over trees of tree-specific predictions, which, in turn, are constructed as the average of outcomes for all observations sharing the same leaf as $\mathbf{x}_p$ within a given tree. The second line in (7) switches summation, and the third line makes explicit that RF predictions can again be interpreted as weighted sum of outcomes over all sample observations, akin to the two previous examples.

However, in this case individual weights $\alpha_i\left(.\right)$ represent the relative frequency with which $\tilde{y}_i$ shares the same leaf as the policy point. Thus, sample observations with features $\mathbf{x}_i$ that are close to $\mathbf{x}_p$ will be in the same leaf as the policy point relatively more frequently across trees, and thus carry a larger weight in the summation step. As is clear from this exposition, in contrast to the LWR these *adaptive neighborhood weights* do not require an extraneous weight matrix $\mathbf{W}$ to sharpen predictive focus towards the policy point. Furthermore, in contrast to both the GL-MRM and the LWR, RF predictions are not dependent on estimated model coefficients $\hat{\boldsymbol{\theta}}$, or, in other words, a prescribed combination of $\mathbf{x}_p$, $\mathbf{X}$, and $\mathbf{y}$ that represents the underlying regression model, as is the case for (4) and (6).[8]

Welfare estimates, in dollars, can then be obtained as in (5), by using the original forest to predict $\tilde{y}_p$ for each methodological setting, and omitting the correction term $0.5 * s^2$ in the exponent function.[9]

Local Linear Forests

Friedberg et al. (2021) propose a variant of RFs which they label Local Linear Forests (LLFs). As is argued in that article, RFs are well suited to detect nonlinear and high-dimensional signals in the data, but are less apt at modeling smooth, linear or close-to-linear

[8] An alternative approach to derive the result in (7) would be to consider a weighted regression on a constant term, as shown in the online appendix.

[9] This correction arises for the GL-MRM and LWR due to the assumed normality of the error term for the logged model. Converting to levels (dollars) implies moving from the normal to a log-normal distribution. The correct expression for the expectation of this log-normal density requires the addition of one half times the estimated error variance in the exponent, as given in (5) (e.g Greene, 2012, ch.4). Since the RF does not require normality, or, for that matter, any statistical distribution for the error term, this correction does not apply.

relationships. Furthermore, RF predictions may be unreliable at leaf boundaries and in sparse regions of the covariate space. Local Linear Forests combine the splitting mechanism of an RF to detect nuanced relationships with a linear regression at the prediction stage to more appropriately model smooth signals. This regression has two additional features: (i) it directly adjusts for covariate differences between data point $\mathbf{x}_i$ and policy point $\mathbf{x}_p$, and it includes a ridge penalty to prevent overfitting.

Assume a forest is evaluated and produces weights $\alpha_i\left(\mathbf{x}_p, \mathbf{X}\right),\ i = 1 \ldots n$. Continuing to label our outcome of interest as $\tilde{y}_p$, the regression problem in the predictive step focusing on policy point $\mathbf{x}_p$ can be formally written as (see Friedberg et al., 2021, equ.3):

$$\begin{aligned} &\begin{bmatrix}\hat{\tilde{y}}_p & \hat{\boldsymbol{\theta}}_p\end{bmatrix}' = \\ &\underset{\tilde{y}_p,\boldsymbol{\theta}_p}{\arg\min}\left\{\sum_{i=1}^{n} \alpha_i\left(.\right)\left(\tilde{y}_i-\tilde{y}_p-\left(\mathbf{x}_i-\mathbf{x}_p\right)'\boldsymbol{\theta}_p\right)^2+\lambda||\boldsymbol{\theta}_p||^2\right\}, \quad \text{where} \\ &||\boldsymbol{\theta}_p||^2=\sum_{j=1}^{k}\theta_{p,j}^2, \end{aligned} \tag{8}$$

λ is the ridge penalty, and k denotes the number of coefficients in $\boldsymbol{\theta}$, which is equal to the full set of features in the explanatory data. The solution to this penalized regression problem can be written as (see Friedberg et al., 2021, equ.5):

$$\begin{bmatrix}\hat{\tilde{y}}_p & \hat{\boldsymbol{\theta}}_p\end{bmatrix}' = \left(\mathbf{X}_p'\mathbf{A}\mathbf{X}_p + \lambda\mathbf{J}\right)^{-1}\mathbf{X}_p'\mathbf{A}\tilde{\mathbf{y}}, \tag{9}$$

where $\mathbf{X}_p$ is an n by $(k+1)$ matrix of centered features, with each row starting with a value of one, followed by $\left(\mathbf{x}_i - \mathbf{x}_p\right)',\ i = 1 \ldots n$, $\mathbf{A}$ is a diagonal matrix of dimension n featuring weights $\alpha_i\left(.\right)$, and $\mathbf{J}$ is a $(k+1)$ by $(k+1)$ diagonal matrix with zero in the first position, and ones along the remainder of the diagonal.

The sought prediction $\hat{\tilde{y}}_p$ is the first element of this $(k+1)$ by 1 solution vector. Consider $(k+1)$ by 1 vector $\mathbf{a}$ with one as its first element, and zeros elsewhere. We can then write

$$\begin{gathered}\hat{\tilde{y}}_p = \mathbf{a}' * \left(\mathbf{X}_p'\mathbf{A}\mathbf{X}_p + \lambda\mathbf{J}\right)^{-1}\mathbf{X}_p'\mathbf{A}\tilde{\mathbf{y}} = \\ \left(\mathbf{a}' * \left(\mathbf{X}_p'\mathbf{A}\mathbf{X}_p + \lambda\mathbf{J}\right)^{-1}\mathbf{X}_p'\right)\ \mathbf{A}\tilde{\mathbf{y}} = \\ \sum_{i=1}^{n}\boldsymbol{\gamma}_i\left(\mathbf{x}_p,\mathbf{X}\right)\alpha_i\left(\mathbf{x}_p,\mathbf{X}\right)\tilde{y}_i,\end{gathered} \tag{10}$$

where $\boldsymbol{\gamma}_i\left(.\right)$ is the i^{th} row of $\left(\mathbf{a}' * \left(\mathbf{X}_p'\mathbf{A}\mathbf{X}_p + \lambda\mathbf{J}\right)^{-1}\mathbf{X}_p'\right)$. This yields again a weighted-sum expression for the predicted policy outcome, as is evident from the last line in (10). However, in this case, two sets of weights are involved. Forest weight $\alpha_i\left(.\right)$, as before, and local adjustment weight $\boldsymbol{\gamma}_i\left(.\right)$. Intuitively, the former enhances the contribution of data point i in the formulation of the policy prediction based on joint membership in terminal tree leaves, while the latter fine-tunes this weight by considering the actual vector distance between data and policy point. In other words, if forest leaves containing $\mathbf{x}_p$ are sparse, unbalanced, or constrained by boundary restrictions, the second weight adds an additional tool to better calibrate the leaf-specific prediction to $\mathbf{x}_p$. With $\tilde{y_p}$ in hand, final transformations to obtain WTP in dollars can then be obtained as for the generic RF above.

Friedberg et al. (2021) show that the performance of LLFs can be further enhanced by adding a splitting rule that differs from the generic MSE-based rule in standard RFs. Specifically, they propose to use a ridge regression at each parent node to predict outcomes for all sample observations currently residing at that node. They then impose a standard, MSE-based split on the residuals flowing from this regression. As argued by the authors, this allows to model local / nonlinear effects in the construction of the forest, while capturing smooth, global effects at the prediction stage. Using simulations and empirical examples, they find that this LLF with ridge splitting produces substantially better predictive fit, especially in high-dimensional models with many smooth components. We will adopt ridge splitting for all our LLFs, and refer to the corresponding specification simply as "LLF."

The `R` package `grf`, which we use for all our forest models, offers an additional adjustment option for the LLF, which performs well in our application (Tibshirani et al., 2024a,b). Specifically, the analyst can replace the uniform penalty matrix $\mathbf{J}$ in (9) and (10) with the variance-covariance matrix of the centered regressors (i.e. $\mathbf{X}_p$ in (9) and (10)). This controls for potential differences in scale, i.e. undue influence of features that systematically take larger values than others. An analogous adjustment can be made at the splitting stage if the ridge regression approach is used. We henceforth label the LLF with covariance-adjusted ridge penalty in both splitting and prediction as "LLF.cov."[10]

Derivation of standard errors and confidence intervals for predictions

We estimate all regression models in a classical econometric framework, basing inference and uncertainty measures for BT predictions on standard asymptotic theory. For both the GL-MRM and LWR, standard errors and confidence intervals for dollar-valued BT predictions, as shown in (5), are derived via the simulation approach proposed by Krinsky and Robb (1986).

Wager and Athey (2018) and Athey et al. (2019) derive the asymptotic properties of estimates flowing from various types of RFs. Specifically, they show that these estimates are asymptotically normal and consistent. This opened the door for the use of RFs for statistical inference, including the computation of asymptotically valid standard errors and confidence intervals. Friedberg et al. (2021) shows that these asymptotic results also extend to LLFs. An important prerequisite for these asymptotic guarantees is the construction of forests via the "honesty" principle, i.e. by using different portions of the data to grow a given tree, and populate the leaves, respectively. We follow this honesty principle for all our forest-based models. We choose the delta method as the most straightforward approach to compute standard errors and confidence intervals for our dollar-valued welfare predictions (e.g. Greene, 2012, ch.4) .

Assessing model fit

We use MSE and Mean Absolute Percentage Error (MAPE) to assess the predictive performance of all models with respect to actual sample observations on (dollar-valued) WTP.

[10] This covariance adjustment to control for scale-imbalances can be seen as the analog to the covariance-adjusted distance metric suggested by Moeltner et al. (2023) for the LWR.

Deviations between predicted and observed WTP are commonly referred to as "transfer errors" in the MRM literature (Stapler and Johnston, 2009; Johnston et al., 2017, 2019; Moeltner et al., 2019; Vedogbeton and Johnston, 2020; Moeltner et al., 2023). For the GL-MRM we take a standard Leave-One-Out (LOO) approach to derive these statistics, as described in detail in Moeltner et al. (2019). A similar CV approach based on omitting the home observation from its respective local regression is used for the LWR, with details given in Moeltner et al. (2023). For forests, within-data predictions are generated by considering only trees that were not constructed with help of the target observation in the averaging formulas in (7) and (10), generally referred to as "out-of-bag" predictions (e.g. Tibshirani et al., 2024c).[11]

Utility-theoretic considerations

Kling and Phaneuf (2018), Newbold et al. (2018a), and Moeltner (2019) examine the utility-theoretic properties of different MRM specifications. The main questions in this context are if the chosen baseline MRM is guaranteed to generate WTP predictions that exhibit scope (larger WTP for a more pronounced change in quality), and Adding-Up (AU, the sum of benefits over incremental quality changes equals, approximately, total WTP for the entire change). Moeltner (2019) shows that the GL-MRM in the current context (deemed MRM2 in the original study) satisfies scope under mild and verifiable conditions, and approximately satisfies AU if the coefficient on the quality midpoint, i.e. δ in equation (1), is close to zero. Building on a series of policy simulations, Moeltner (2019) finds that δ is indeed sufficiently small to assure AU holds within negligible margins for their water quality metadata (essentially the predecessor to our data).

Formally, as shown in Moeltner (2019), the scope condition holds for the GL-MRM as long as $\delta\left(q_{p,1} - q_{p,0}\right) > -2$, for some quality levels $q_{0,p} < q_{1,p}$. Given that the quality

[11] Formally, the MSE is given as $\frac{1}{n}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2$, where $\hat{y}_i$ is the predicted WTP (in dollars) for sample observation i. The MAPE, in turn, is derived as $\frac{100}{n}\sum_{i=1}^{n}|\frac{\hat{y}_i - y_i}{y_i}|$. As mentioned in Moeltner et al. (2023), the MAPE has several statistical shortcomings compared to the MSE, but remains a popular metric to assess fit in the MRM/BT literature (Stapler and Johnston, 2009; Johnston et al., 2017, 2019; Moeltner et al., 2019; Vedogbeton and Johnston, 2020).

parameter δ is typically positive, this will be satisfied in most contexts.

In turn, the AU condition for the GL-MRM requires the following equality to be satisfied, for some quality levels $q_{0,p} < q_{1,p} < q_{2,p}$, context-specific policy settings $\mathbf{x}_{c,p}$ and holding methodological indicators at zero for ease of exposition (Moeltner, 2019):

$$\begin{aligned}
&WTP\left(q_{0,p} \rightarrow q_{2,p}\right) = \\
&exp\left(\mathbf{x}'_{c,p}\boldsymbol{\beta}\right) * exp\left(\delta\left(\frac{q_{0,p}+q_{2,p}}{2}\right)\right)\left(q_{2,i}-q_{0,i}\right) = \\
&WTP\left(q_{0,p} \rightarrow q_{1,p}\right) + WTP\left(q_{1,p} \rightarrow q_{2,p}\right) \approx \\
&exp\left(\mathbf{x}'_{c,p}\boldsymbol{\beta}\right) * \left(exp\left(\delta\left(\frac{q_{0,p}+q_{1,p}}{2}\right)\right)\left(q_{1,i}-q_{0,i}\right) + \right. \\
&\left. exp\left(\delta\left(\frac{q_{1,p}+q_{2,p}}{2}\right)\right)\left(q_{2,i}-q_{1,i}\right)\right),
\end{aligned} \tag{11}$$

where the approximation symbol in the third line indicates that (typically minor) income effects have been ignored in the summation of step-wide WTP measures. As can be easily verified, the AU equality only holds if $\delta = 0$, and likely holds within acceptable policy margins if δ is small.

For the LWR, scope will be satisfied as long as the local quality parameter δ_p exceeds $-\frac{2}{(q_{p,1}-q_{p,0})}$, analogous to the GL-MRM. Regarding AU, on the surface it appears that the same reasoning presented above for the GL-MRM should apply to the LWR in equation (2), as it essentially represents the GL-MRM specification with weighted data. However, this is not the case if starting and / or endpoint quality q are used in weight construction. In that instance, all weights will have to be re-configured with a change in q as the target observation's weight values now have changed as well. This could potentially lead to the selection of different sample points to contribute to the local regression. Even if the local sample remains the same, weight values for each observation will be different for different settings of q. This, in turn will affect the coefficient estimates for the local regression, i.e. $\boldsymbol{\beta}$ in (11) will be different for the full-step and the interim-step WTP models. Adding-up would then require close-to-equality for the different sets of coefficient vectors, in addition to $\delta \approx 0$ in all models involved.

In theory, the AU properties of the GL-MRM can be inherited by the LWR if quality is not used in weight construction. In that case the local regression coefficients remain unchanged, and (11) can be applied to the weighted data. However, it would be difficult to imagine a situation where a key variable such as quality would not be a relevant weight variable. In Moeltner et al. (2023), who use essentially the same metadata as we do in this study, quality was found to be an important component in all of their preferred weight settings.

For our forest models (RF, LLF, LLF.cov) a single forest is first "trained" using the actual metadata, with quality midpoint $\frac{q_{0,i}+q_{1,i}}{2}$ among the set of covariates. As mentioned above, we adopt the transformed outcome of the GL-MRM, i.e. $\tilde{y_i} = log\left(\frac{y_i}{q_{1,i}-q_{0,i}}\right)$. For any combination of $\mathbf{x}_{c,p}, q_{l,p}, q_{h,p},\ l < h$, the forest will predict $\tilde{y_p}\left(\mathbf{x}_{c,p}, q_{l,p}, q_{h,p}\right)$. Conversion to WTP, in dollars, then implies:

$$\hat{y_p}\left(\mathbf{x}_{c,p}, q_{l,p}, q_{h,p}\right) = exp\left(\tilde{y_p}\left(\mathbf{x}_{c,p}, q_{l,p}, q_{h,p}\right)\right) * \left(q_{h,p} - q_{l,p}\right) \tag{12}$$

If the different quality steps are small enough such that $\tilde{y_p}\left(.\right)$ is approximately equal across all quality change scenarios (recall $\mathbf{x}_{c,p}$ remains invariant), AU will (approximately) hold by construction, as can be easily deduced from (12). Scope, in turn, will be satisfied if the term in the exponent in (12) is non-decreasing in $q_{h,p}$ for some common baseline $q_{l,p}$. Since our forests use the quality mid-point as one of their features, this implies that scope will hold as long as otherwise identical policy points $\mathbf{x}_p$ generally end up in terminal leaves with sets of neighbors that exhibit outcome values that remain constant or increase with increasing midpoint. For AU, we need these neighbors to remain relatively invariant with changes in q.

In a nutshell, the satisfaction of core theoretic properties will largely be an empirical questions for the LWR (assuming quality is used in weight construction) and all of our forest versions. We examine this in more detail in a simulation exercise below.

Empirical application

Data

We use essentially the same metadata on WTP for water quality improvements in various water bodies across the U.S. as Moeltner et al. (2023).[12] As described there, these data (and earlier versions) have been employed in numerous publications and EPA rulemaking contexts (e.g. U.S. Environmental Protection Agency, 2015; Johnston et al., 2017; Newbold et al., 2018a; Johnston et al., 2019; Moeltner, 2019; U.S. Environmental Protection Agency, 2020). As these metadata and their progenitors are described in detail in these prior publications, we provide only a concise summary here. The metadata are drawn from primary stated preference studies that estimate per household (use and nonuse) WTP for water quality changes in U.S. water bodies. Studies were limited to those for which WTP estimates could be readily mapped to water quality changes measured on a standard 100-point Water Quality Index (WQI) from an identifiable baseline, following methods described in Johnston et al. (2017) and Johnston et al. (2019). Studies with primary focus on drinking water were not considered. The final data comprise 188 observations from 58 source studies. The dependent variable before transformation is WTP in 2019 dollars, to maintain comparability with the results reported in Moeltner et al. (2023).

Detailed variable descriptions are given in Moeltner et al. (2023). Table 1 gives an overview of these variables, along with descriptive statistics. As is evident from the third column, our metadata comprise a mix of continuous and binary features, for which tree-based methods are generally well-suited (Friedberg et al., 2021). At the same time, the presence of several continuous variables presents the possibility of globally smooth and quasi-linear effects of these features on WTP. This, in turn, supports consideration of LLF models of the type we use in our analysis. The last three columns, directly adopted from Moeltner et al. (2023), indicate which variables entered the three sets of weight combinations used for the LWR.

[12] The only (minor) adjustment we make to the data used in Moeltner et al. (2023) is a re-labeling of regional indicators "northeast," "central," and "south" to corresponding U.S. census regions, for compatibility with regional designations in agency rulemaking.

Forest tuning

As discussed in Tibshirani et al. (2024b), the main training parameters for forests in `grf` are the number of underlying trees, (`num.trees`) the fraction of the sample to be used for a given tree (`sample.fraction`), the minimum allowable node size (`min.node.size`), and the number of features to be considered at each splitting occasion (`mtry`). We choose the default setting of 2000 trees for `num.tree`, based on a preliminary examination of the effect of tree number on model fit, as summarized in the online appendix to this paper. We allow all other parameters to be optimally tuned via cross-validation (setting `tune.parameter=‘‘all’’` in grf). We also use cross-validation to tune parameters related to control "honesty" as described above, and to guide balance in splitting, as described in Tibshirani et al. (2024b).

As mentioned in Tibshirani et al. (2024a), parameter tuning is not (yet) available in `grf` for LLF's that use a splitting rule based on ridge regressions (our preferred approach). Thus, for all our LLF specifications we adopt default settings for all training parameters, as listed in Tibshirani et al. (2024a). Robustness checks for different parameter settings, given in the online appendix, illustrate that changes in key tuners do not have any material effects on predictive accuracy.

Model Fit

Table 2 shows results for predictive fit with respect to actual metadata points, as described above. The first row captures mean prediction errors for the GL-MRM, i.e. an MSE of 62.6 (in thousands), and a MAPE of approximately 124. The following ten rows give results for the best-performing LWR models that did not suffer from rank violations in the CV process and thus utilize all 188 observations.[13] The table also shows each LWR's weight setting components. For example, LWR-1 used the weight variables listed under C1 in Table 1, an un-adjusted distance function, a bi-square weight function, and a window size of 188 (= the entire sample). These weight settings parse the data into 143 distinct locations. As is evident from the table, and noted in Moeltner et al. (2023), the LWRs shave off

[13] We impose this full-sample restriction for better comparability to our other estimation frameworks, which, by default, operate on the entire data. As a result, the top ten models captured in the table, and their corresponding model fit metrics, differ slightly from those reported in Moeltner et al. (2023).

approximately two thirds of the GL-MRM's MSE, and reduce the MAPE by 35-45%.

The last three rows of the table present model fit statistics for our three forest specifications. Clearly, all three forests yield an additional, and substantial improvement in model fit, with MSE's reduced to the 6.4 to 7.7 range. This figures are essentially an order of magnitude smaller than those corresponding to the GL-MRM, and approximately one third to one fourth the magnitude of the MSE's produced by the LWR versions. The forests also exhibit smaller MAPE's than any of the other models, with the LLF versions performing especially well in that respect, more then halving the MAPE of the GL-MRM, and reducing the lowest LWR MAPE by another 18-20%.

In a nutshell, all three forest versions generate far superior model fits compared to their regression-based counterparts. Applying standard terminology from the BT literature as introduced above, this is akin to a substantial reduction in value transfer or generalization error. We examine next if this gain in predictive accuracy translates into commensurate improvements in predictive efficiency for out-of-sample points, such as a BT context.

Benefit Transfer comparison

To assess model performance in a BT context, where WTP for a given policy context is unknown by definition, we simulate a large number of BT scenarios with different combinations of values for several explanatory variables. The general strategy for this simulation is captured in the center column of Table 3. Specifically, we let `lnyear` be the log of (2024-1980) to convey the notion of a contemporary application. We hold weight variables `ln_size_ratio` and `lnpop` at their respective sample medians. Similarly, we also apply a close-to-median setting for the surface cover variables `pctdev` through `pctwet`, while still complying with the embedded adding-to-100 condition. (see variable definitions in Table 1). To ensure realism, settings for binary indicators and quality points are borrowed from the Des Moines watershed application given in Moeltner et al. (2023).

Our main focus rests on the four continuous context-specific variables `sub_proportion`, `ln_ar_agr`, `lnincome`, and `ln_ar_ratio`.[14] For each of these we vary values as shown in

[14] As discussed previously (Johnston et al., 2017), these variables capture key contextual dimensions when

Table 4. Specifically, we capture the minimum, median, and maximum value represented in the metadata, as given in the first, fourth, and last column of the table, respectively. We then divide both the lower and upper 50 percentile range, respectively, into two intervals of equal width, labeled as “med- -,” “med-,” “med+,” and “med++” in the table. The complete factorial of these seven settings for each variable yields 2401 unique combinations. We refer to this full factorial as our “full” scale simulation. We separately consider BT predictions generated by the factorial of the median and two adjacent values. We label this smaller set of 81 combinations our “center” simulation. Conversely, we separately bundle the four outside settings yielding 256 BT scenarios, and call it the “fringe” simulation. Together, the three simulations will illustrate how our models converge or differ in areas where data mass is highest (“center”), sparse (“fringe”), and over the full range of observations (“full”).

Results are given in Table 5. Each block of rows represents one of our three simulations (full, center, fringe). The first three columns give, respectively, the mean, minimum, and maximum prediction of WTP, in 2019 dollars, over all underlying BT scenarios. As expected, results are much more convergent across models for our “center” simulation, with means over scenarios ranging from $2.62 (LLF) to $7.65 (RF), and maxima in the $6-12 range. For the “full” simulation, we observe a two-threefold increase in means for the GL-MRM, LWR, and RF, and a dramatic increase in maxima for the linear regression models, exceeding $130 in both instances. These figures appear unrealistic given the small underlying quality change. In contrast, estimates for the LLF and LLF.cov remain of modest magnitude ($5-6), with considerably lower maxima, especially for the LLF.cov ($21). A similar trend relative to the “center” version is observed for the “fringe” simulation. It thus appears that the linear forest models, especially the LLF.cov, are less prone to extreme predictions in sparse / boundary ranges of the underlying data, compared to the generic RF and the linear regressions. Hence, an assessment of construct validity (here, whether the predicted value is consistent with theory and past findings (Bishop and Boyle, 2019))

seeking to predict WTP for water quality improvements. They characterize, respectively, the proportion of waterbodies of same hydrologic type in the state or region compared to the waterbody that is subject to the actual quality change, the (log of) the proportion of the improved area that falls under agricultural land cover, (log of) household income, and the (log of) the ratio of the extent of the market for WTP estimation (i.e. stakeholder population) and the area over which actual water quality improvements would occur.

strongly supports the LLF models over the other alternatives.

For each scenario, we also computed 95% C.I.'s for the corresponding BT estimate. Comparative statistics for the mean, minimum, and maximum of these intervals are shown in the second triplet of columns of Table 5. We first note that, as expected, the "center" simulation is associated with the tightest C.I. ranges for all models compared to the "full" and "fringe" version. The table also shows that the LWR produces a noticeable reduction in average interval range for all three simulations relative to the GL-MRM, as also observed in Moeltner et al. (2023). However, intervals generated by all three forest models are by an order of magnitude tighter than even the LWR constructs, remaining well in the single-digit dollar range. The covariance-corrected LLF is again the best performer, with the smallest mean C.I. range in all situations. This advantage is most pronounced at the fringe of the data space, as is evident from the last row of the table.

The last portion of Table 5 shows basic correlation coefficients for scenario-specific BT estimates across all models. As expected, the LWR is highly correlated with the GL-MRM. Both regression models, in turn, exhibit positive correlations of non-trivial magnitude with the LLF specifications. In contrast, the generic RF stands on its own, with much reduced, but still positive correlation with all remaining specifications. We take this as yet another indication that the uncorrected RF misses important linear / smooth signals in the metadata.

Adding-Up comparison

To examine the AU properties of the LWR and our preferred forest specifications (LLF and LLF.cov) relative to the GL-MRM, we adopt the Des Moines watershed settings from Moeltner et al. (2023). These values are given in the last column of Table 3. For the actual quality change, we consider three separate scenarios, each located at different segments of the 100-point water quality ladder. As indicated in the table, the first scenario stipulates a two-step improvement from $q_0 = 41$ to $q_2 = 44$ (as actually used in the Des Moines application), via intermediary point $q_1 = 42.5$. The second scenario operates with the same quality steps, but starts at $q_0 = 61$. Scenario three then adds another 20 quality points, starting at $q_0 = 81$. As discussed in Moeltner (2019) and Moeltner et al. (2023), these seemingly modest changes of single-digit quality points are quite realistic, and often reflect the upper practicable range of best management practices and / or water cleanup regulation.[15]

Results are presented in Table 6, which gives WTP (in 2019 dollars) for each step-wise and total change, respectively. While point predictions differ to some extent across models, the AU condition is fully or approximately satisfied by all specifications and for all scenarios. The linear forests perform especially well, with AU errors well below the 1-percent mark. Evidently, the stipulated quality changes are small enough for the AU-conducive conditions outlined in the previous section to hold, regardless of where along the 100-point ladder they occur. As an aside, we also observe from the table that the span between the lower ("low") and upper ("high") 95% C.I. interval for each prediction is by an order of magnitude tighter for the LLF specifications compared to the linear regression models, mirroring the findings from the BT simulation above. We examine uneven and larger quality steps in the online appendix. AU-compliance within 1-2% error continues to hold for the LWR and LLF. Errors for the LLF.cov range from six to 17%, confirming our conjecture that the AU error for forest-based models may increase with step size and / or step imbalances. The additional simulations based on uneven quality steps also allow for a cursory inspection of adherence to scope. Scope is satisfied for all models and quality simulations.

Conclusion

In this study we introduce different versions of RFs to the MRM-BT literature, and contrast their performance with respect to predictive accuracy and efficiency with existing best-practice approaches. We highlight key econometric differences and commonalities between forests and parametric regression models, and stress the interpretation of forests as nonparametric estimators based on policy-adaptive neighborhood kernels. We also discuss

[15] For the LWR we drop weight settings two, three, seven, and eight (as captured in table 2) due to egregiously large predictions. We employ the same MSE-based averaging across the remaining settings to obtain the estimates captured in table 6.

embedded utility-theoretic properties, and the structural conditions that need to hold for these properties to be satisfied. Ours is among the first applications in applied economics at large to put a novel variant of forests, LLFs, to the test in an empirical setting. Results overwhelmingly support these methods as an attractive and broadly applicable option for environmental BT. For our MRM application of WTP for water quality improvements, we find that forests, especially the LLF variants, bring vast gains in predictive fit and reductions in variance over even state-of-the-art MRMs. They also satisfy utility-theoretic properties on empirical grounds for small quality changes, as have been typical in recent rulemaking. Overall, we conclude that the new approach has the potential to substantially sharpen predictive inference for use in environmental regulation without sacrificing theoretical properties. Our analytical framework building on forests is applicable to any regression context where predictive performance is of central importance.

A logical next step would be to stress-test our forest-based approach with other meta-data and BT contexts, such as related to wetland valuation, recreation demand, or topics in health economics. We also note that while the primary focus of this study lies on prediction, forests are also suitable to provide nonparametric inspection of the relative importance of explanatory variables in the meta-data, e.g. via variable importance statistics, partial dependence plots, or Shapley values (e.g. Greenwell, 2022). We leave a closer examination of these RF features, and how they could possibly guide variable selection in a more structural MRM, to future work.

References

Aldy, J., Atkinson, G., Kotchen, M., 2021. Environmental benefit-cost analysis: A comparative analysis between the united states and the united kingdom. Annual Review of Resource Economics 13, 267–288.

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.

Bishop, R., Boyle, K., 2019. Reliability and validity in nonmarket valuation. Environmental and Resource Economics 72, 559–582.

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

Brookshire, D., Neill, H., 1992. Benefit transfers: conceptual and empirical issues. Water Resources Research 28, 651–655.

Corona, J., Doley, T., Griffiths, C., Massey, M., Moore, C., Muela, S., Rashleigh, B., Wheeler, W., Whitlock, S., Hewitt, J., 2020. An integrated assessment model for valuing water quality changes in the United States. Land Economics 96, 478–492.

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.

Fotheringham, A., Brunsdon, C., Charlton, M., 2002. Geographically weighted regression: The analysis of spatially varying relationships. Wiley.

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

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.

Griffiths, C., Klemick, H., Massey, M., Moore, C., Newbold, S., Walsh, P., Wheeler, W., 2012. U.S. Environmental Protection Agency valuation of surface water quality improvements. Review of Environmental Economics and Policy 6, 130–146.

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., Besedin, E., Holland, B., 2019. Modeling distance decay with valuation meta-analysis. Environmental and Resource Economics 72, 657–690.

Johnston, R., Besedin, E., Stapler, R., 2017. Enhanced geospatial validity for meta-analysis and environmental benefit transfer: An application to water quality improvements. Environmental and Resource Economics 90, 773–795.

Johnston, R., Besedin, E.Y., Iovanna, R., Miller, C., Wardwell, R., Ranson, M., 2005. Systematic variation in willingness-to-pay for aquatic resource improvements and implications for benefit transfer: A meta-analysis. Canadian Journal of Agricultural Economics 53, 221–248.

Johnston, R., Besedin, E.Y., Wardwell, R., 2003. Modeling relationships between use and nonuse values for surface water quality: A meta-analysis. Water Resources Research 39.

Johnston, R., Boyle, K., Loureiro, M., Navrud, S., Rolfe, J., 2021. Guidance to enhance the validity and credibility of environmental benefit transfers. Environmental and Resource Economics 79, 575–624.

Johnston, R., Rolfe, J., Zawojska, E., 2018. Benefit transfer of environmental and resource values: Progress, prospects and challenges. International Review of Environmental and Resource Economics 12, 177–266.

Johnston, R., Rosenberger, R., 2010. Methods, trends and controversies in contemporary benefit transfer. Journal of Economic Surveys 24, 4–1 – 4–13.

Kaul, S., Boyle, K., Kuminoff, N., Parmeter, C., Pope, J., 2013. What can we learn from benefit transfer errors? Evidence from 20 years of research on convergence validity. Journal of Environmental Economics and Management 66, 90–104.

Keiser, D., 2019. The missing benefits of clean water and the role of mismeasured pollution. Journal of the Association of Environmental and Resource Economists 6, 669–707.

Keiser, D., Kling, C., Shapiro, J., 2019. The low but uncertain measured benefits of U.S. water quality policy. Proceedings of the National Academy of Sciences 116, 5262–5269.

Keiser, D., Shapiro, J., 2019. Consequences of the Clean Water Act and the demand for water quality. The Quarterly Journal of Economics 134, 349–396.

Kling, C., Phaneuf, D., 2018. How are scope and adding up relevant for benefit transfer? Environmental and Resource Economics 69, 483–502.

Krinsky, I., Robb, A., 1986. On approximating the statistical properties of elasticities. Review of Economics and Statistics 72, 189–190.

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.

McMillen, D., Redfearn, C., 2010. Estimation and hypothesis testing for nonparametric hedonic house price functions. Journal of Regional Science 50, 712–733.

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., 2019. Bayesian nonlinear meta regression for benefit transfer. Journal of Environmental Economics and Management 93, 44–62.

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., Boyle, K., Paterson, R., 2007. Meta-analysis and benefit-transfer for resource valuation: Addressing classical challenges with Bayesian modeling. Journal of Environmental Economics and Management 53, 250–269.

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.

Newbold, S., Johnston, R., 2020. Valuing non-market valuation studies using meta-analysis: A demonstration using estimates of willingness-to-pay for water quality improvements. Journal of Environmental Economics and Management 104, 102379.

Newbold, S., Massey, D., Walsh, P., Hewitt, J., 2018a. Using structural restrictions to achieve theoretical consistency in benefit transfer. Environmental and Resource Economics 69, 529–553.

Newbold, S., Simpson, D., Massey, D., Heberling, M., Wheeler, W., Corona, J., Hewitt, J., 2018b. Benefit transfer challenges: perspectives from us practitioners. Environmental and Resource Economics 69, 467–481.

Office of Management and Budget, 2023. Circular no. a-4: Regulatory analysis. Office of Management and Budget, The White House, Washington, D.C.

Office of the Federal Register, 1981. Executive order 12291 - Federal regulation. Web site: https://www.archives.gov/federal-register/codification/executive-order/12291.html, last accessed 2024-10-11.

Office of the Federal Register, 1993. Executive order 12866 of september 30, 1993. Federal Register 58, Oct. 4, 1993.

Pagan, A., Ullah, A., 1999. Nonparametric econometrics. Cambridge University Press.

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.

Rolfe, J., Brouwer, R., Johnston, R., 2015. Meta-analysis: Rationale, issues, and applications, in: Johnston, R., Rolfe, J., Rosenberger, R., Brouwer, R. (Eds.), Benefit transfer of environmental and resource values, pp. 357–382.

Stapler, R., Johnston, R., 2009. Meta-analysis, benefit transfer, and methodological covariates: Implications for transfer errors. Environmental and Resource Economics 42, 227–246.

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.

Sun, J., Caputo, V., Taylor, H., 2024. Using machine-learning methods in meta-analyses: An empirical application on consumer acceptance of meat alternatives. Applied Economic Perspectives and Policy , 1–27.

The White House, 2011. Executive order 13563 - Improving regulation and regulatory review. Web site: https://obamawhitehouse.archives.gov/the-press-office/2011/01/18/executive-order-13563-improving-regulation-and-regulatory-review, last accessed 2024-10-11.

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.

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

U.S. Environmental Protection Agency, 2009. Environmental Impact and Benefits Assessment for Final Effluent Guidelines and Standards for the Construction and Development Category. Technical Report EPA-821-R-09-012. United States Environmental Protection Agency, Office of Water.

U.S. Environmental Protection Agency, 2010. Economic Analysis of Final Water Quality Standards for Nutrients for Lakes and Flowing Waters in Florida. Technical Report November, 2010. United States Environmental Protection Agency, Office of Water.

U.S. Environmental Protection Agency, 2015. Benefit and cost analysis for the effluent limitations guidelines and standards for the steam electric power generating point source category. Technical Report EPA-821-R-15-005. United States Environmental Protection Agency, Office of Water.

U.S. Environmental Protection Agency, 2020. Benefit and cost analysis for revisions to the effluent limitations guidelines and standards for the steam electric power generating point source category. Technical Report EPA-821-R-20-003. United States Environmental Protection Agency, Office of Water.

U.S. Environmental Protection Agency, 2023. Benefit Cost Analysis for Revisions to the Effluent Limitations Guidelines and Standards for the Meat and Poultry Products Point Source Category. Technical Report EPA-821-R-23-013. United States Environmental Protection Agency, Office of Water.

U.S. Environmental Protection Agency, 2024a. Benefit and cost analysis for supplemental effluent limitations guidelines and standards for the steam electric power generating point source category. Technical Report EPA-821-R-24-006. United States Environmental Protection Agency, Office of Water.

U.S. Environmental Protection Agency, 2024b. Summary of the Clean Water Act. Web site: `https://www.epa.gov/laws-regulations/summary-clean-water-act`, last accessed 2024-10-11.

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

Vedogbeton, H., Johnston, R., 2020. Commodity-consistent meta-analysis of wetland values: An illustration for coastal marsh habitat. Environmental and Resource Economics 75, 835–865.

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

Table 1: Overview of variables

		variable				LWR weight variables		
variable	label	type	mean	min	max	C1	C2	C3
wtp	wtp for specified water quality change, 2019$	c	140.85	1.73	498.38	-	-	-
context-specific								
lnyear	log (years since earliest study in the sample (1980))	c	2.64	0.00	3.61	x	x	x
lump_sum	1 = payments are not annual (e.g. lump sum)	b	0.18	0.00	1.00		x	x
nonusers	1 = survey population only includes non-users	b	0.06	0.00	1.00		x	x
sub_proportion	proportion of waterbodies of same type in the state or region	c	0.35	0.00	1.00	x	x	x
census-S	1 = study conducted in the southern U.S. census region	b	0.35	0.00	1.00		x	x
census-MW	1 = study conducted in the midwestern U.S. census region	b	0.29	0.00	1.00		x	x
census-W	1 = study conducted in the western U.S. census region	b	0.09	0.00	1.00		x	x
swim_use	1 = changes in swimming use emphasized in survey	b	0.22	0.00	1.00		x	x
gamefish	1 = changes in game fishing emphasized in survey	b	0.19	0.00	1.00		x	x
ln_ar_agr	log of affected resource area that is agricultural	c	-1.64	-4.26	-0.08	x	x	x
lnincome	log (median houshold inc. in sampled area, census data, 2019$)	c	10.95	10.65	11.48	x	x	x
tax_only	1=payment mech = increased taxes	b	0.40	0.00	1.00		x	x
user_cost	1=payment mech = increased user cost	b	0.02	0.00	1.00		x	x
ln_ar_ratio	log of sampled area divided by affected resource area	c	-0.59	-8.48	6.65	x	x	x
q0	baseline water quality, 100-point WQI scale	c	46.90	10.00	85.00	x	x	x
q1	target water quality, 100-point WQI scale	c	59.99	12.50	95.00	x	x	x
additional weight variables								
ln_sz_ratio	log of (affected resource area / pop.-weighted avg. dist. from resource)	c	7.80	3.48	14.13	x		x
lnpop	log of affected resource area that is agricultural	c	-1.64	-4.26	-0.08	x		x
pctdev	log of population within the affected resource area	c	14.04	9.34	16.87	x		x
pctopen	percentage of catchment area that is developed land	c	9.35	0.17	77.06	x		x
pctfor	percentage of catchment area that is open and ag. land	c	29.22	0.26	91.40	x		x
pctwet	percentage of catchment area that is forest land	c	31.95	0.06	84.47	x		x
	percentage of catchment area that is wetland	c	11.10	0.01	64.30			
methodological variables								
thesis	1 = study was a PhD or Master's thesis	b	0.08	0.00	1.00			
volunt	1 = payment vehicle described as voluntary	b	0.05	0.00	1.00			
nonrev	1 = study was not published in a peer-reviewed journal	b	0.16	0.00	1.00			
oneshotval	1 = only 1 valuation question given	b	0.53	0.00	1.00			
rum	1=RUM model used	b	0.56	0.00	1.00			
ibi	1 = water quality derived from a biological index	b	0.08	0.00	1.00			

b = binary (0/1), c = continuous
LWR = Locally-Weighted MRM
C1, C2, C3: weight variable combinations used in the LWR

Table 2: Comparison of model fit

model	combo	dist	wf	ws	local regressions	mse	mape
GL-MRM	-	-	-	-	-	62.584	124.141
LWR-1	C1	U	B	188	143	21.752	78.686
LWR-2	C1	U	B	170	143	20.280	76.544
LWR-3	C1	U	T	170	143	19.066	73.100
LWR-4	C1	M	T	188	143	25.330	84.792
LWR-5	C2	IV	T	188	155	27.294	86.754
LWR-6	C2	M	B	188	155	19.776	76.463
LWR-7	C3	U	B	170	155	19.993	69.111
LWR-8	C3	U	T	170	155	20.686	72.087
LWR-9	C3	IV	T	188	155	29.690	75.619
LWR-10	C3	M	T	188	155	29.552	79.545
RF	-	-	-	-	-	6.503	69.018
LLF	-	-	-	-	-	7.707	56.642
LLF.cov	-	-	-	-	-	6.379	55.230

combo = set of weight variables (see text)
dist = distance function (U = unadjusted, IV = inverse variance, M = Mahalanobis)
wf = weight function (G = Gaussian, B = Bi-square, T = Tri-cubic)
ws = window size
local regressions = identified locations for LWR
mse = Mean Squared Error (in 1000's)
mape = Mean Absolute Percentage Error
GL-MRM = Globally-Linear MRM
LWR = Locally-Weighted MRM
RF = Random Forest
LLF = Local Linear Forest (with residual splitting)
LLF.cov = LLF with covariance-adjusted ridge penalty

Table 3: Benefit Transfer simulation settings

variable	BT simulation	AU simulation
context-specific		
lnyear	log(2024-1980)*	log(2024-1980)
lump_sum	0	0
nonusers	0	0
sub_proportion	0.00 to 1.00	0.166
census-S	0	0
census-MW	1	1
census-W	0	0
swim_use	0	0
gamefish	0	0
ln_ar_agr	-4.26 to -0.08	-0.239
lnincome	10.65 to 11.48	11.011
tax_only	1	1
user_cost	0	0
ln_ar_ratio	-8.48 to 6.65	2.08
q0	41.00	41, 61, 81
q1	44.00	42.5, 62.5, 82.5
q2	-	44, 64, 84
weight variables		
ln_sz_ratio	7.47	5.88
lnpop	13.99	17.65
pctdev	5.00	3.62
pctopen	45.00	82.66
pctfor	45.00	9.01
pctwet	5.00	2.13
methodological variables		
thesis	(averaged)	(averaged)
volunt	(averaged)	(averaged)
nonrev	(averaged)	(averaged)
oneshotval	(averaged)	(averaged)
rum	(averaged)	(averaged)
ibi	(averaged)	(averaged)

BT = Benefit Transfer
AU = Adding-Up

Table 4: Benefit Transfer simulation settings, details

	fringe		center			fringe	
variable	min.	med.–	med.-	median	med.+	med.++	max.
sub_proportion	0.00	0.04	0.08	0.12	0.41	0.71	1.00
ln_ar_agr	-4.26	-3.33	-2.41	-1.48	-1.01	-0.55	-0.08
lnincome	10.65	10.74	10.83	10.92	11.11	11.29	11.48
ln_ar_ratio	-8.48	-5.65	-2.83	0.00	2.22	4.43	6.65

min. (max.) = minimum (maximum) value in metadata
med–: min. + (1/3)*(median - min.)
med-: min. + (2/3)*(median - min.)
med+: median + (1/3)*(max. - median)
med++: median + (2/3)*(max. - median)

Table 5: Comparison of BT predictions

	BT estimate			BT C.I. range			correlations				
model	mean	min	max	mean	min	max	GL-MRM	LWR	RF	LLF	LLF.cov
					full (2401 scenarios)						
GL-MRM	10.044	0.483	139.401	39.221	2.102	625.397	1.000				
LWR	10.077	0.420	134.933	28.683	1.291	412.222	0.989	1.000			
RF	13.958	4.660	62.695	2.763	0.703	16.156	0.204	0.222	1.000		
LLF	5.312	0.252	49.616	2.704	0.094	32.225	0.862	0.898	0.403	1.000	
LLF.cov	6.044	0.832	20.999	1.531	0.221	5.408	0.629	0.659	0.580	0.817	1.000
					center (81 scenarios)						
GL-MRM	4.456	1.733	11.488	15.778	6.305	40.010	1.000				
LWR	4.236	1.564	10.839	10.880	3.970	28.243	0.994	1.000			
RF	7.646	6.022	10.430	1.279	0.796	2.157	0.436	0.482	1.000		
LLF	2.620	0.927	6.402	1.083	0.346	2.344	0.920	0.927	0.615	1.000	
LLF.cov	4.603	2.204	10.199	1.058	0.531	2.420	0.811	0.846	0.776	0.871	1.000
					fringe (256 scenarios)						
GL-MRM	16.473	0.483	139.401	67.386	2.004	614.187	1.000				
LWR	16.977	0.423	134.938	50.917	1.285	425.388	0.985	1.000			
RF	18.524	4.941	60.342	3.645	0.812	14.698	0.185	0.207	1.000		
LLF	7.900	0.256	49.616	4.293	0.103	29.570	0.838	0.896	0.352	1.000	
LLF.cov	7.085	0.832	20.567	1.866	0.221	5.408	0.648	0.695	0.537	0.841	1.000

mean, min., max. taken over BT point estimates for 5000 randomly drawn Census Block Groups (CBGs)
C.I. = 95% asymptotic confidence interval
range = upper minus lower 95% C.I. bound for a given CBG
GL-MRM = Globally-Linear MRM
LWR = Locally-Weighted MRM
RF = Random Forest
LLF = Local Linear Forest (with residual splitting)
LLF.cov = LLF with covariance-adjusted ridge penalty

Table 6: Adding-up examination

	GL-MRM			LWR			LLF			LLF.cov		
scenario (WQI)	low	mean	high	low	mean	high	low	mean	high	low	mean	high
41 to 42.5	0.286	1.657	6.048	0.375	1.353	3.496	0.404	0.538	0.671	1.342	1.531	1.720
42.5 to 44	0.284	1.626	5.985	0.372	1.320	3.433	0.381	0.519	0.656	1.306	1.491	1.677
41 to 44	0.594	3.283	11.858	0.752	2.683	7.164	0.783	1.055	1.327	2.632	3.013	3.395
adding-up error		0.00%			-0.37%			0.11%			0.29%	
61 to 62.5	0.223	1.286	4.619	0.254	0.932	2.412	0.271	0.390	0.509	0.550	0.656	0.762
62.5 to 64	0.221	1.262	4.554	0.260	0.909	2.385	0.271	0.385	0.499	0.547	0.654	0.761
61 to 64	0.453	2.549	9.048	0.523	1.877	4.877	0.541	0.774	1.007	1.094	1.308	1.522
adding-up error		0.00%			-1.60%			0.12%			0.20%	
81 to 82.5	0.172	0.999	3.592	0.184	0.670	1.758	0.336	0.420	0.505	0.510	0.573	0.636
82.5 to 84	0.167	0.980	3.653	0.180	0.656	1.715	0.346	0.430	0.514	0.510	0.573	0.636
81 to 84	0.342	1.978	7.167	0.373	1.334	3.551	0.682	0.850	1.018	1.020	1.146	1.272
adding-up error		0.00%			-2.53%			0.02%			0.00%	

GL-MRM = Globally-Linear MRM
LWR = Locally-Weighted MRM
RF = Random Forest
LLF = Local Linear Forest (with residual splitting)
LLF.cov = LLF with covariance-adjusted ridge penalty