Locally-weighted meta-regression and benefit transfer

Klaus Moeltner

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

Roshan Puri

Ph.D. Student, Department of Agricultural and Applied Economics, Virginia Tech, puriroshan@vt.edu

Robert J. Johnston

Professor, George Perkins Marsh Institute, Clark University, rjohnston@clarku.edu

Elena Besedin

ICF Consulting, elena.besedin@icf.com

Jessica A. Balukas

ICF Consulting, jessica.balukas@icf.com

Alyssa Le

ICF Consulting, alyssa.le@icf.com

Selected Paper for presentation at the 2022 Agricultural & Applied Economics Association Annual Meeting, Anaheim, CA; July 31 - August 2

Locally-weighted meta-regression and benefit transfer

Abstract

We propose a locally-weighted regression approach to analyze meta-data and generate benefit transfer predictions in an environmental valuation context. We introduce the concept of locally-weighted meta-regression, provide econometric underpinnings, and discuss the construction of weight functions. We illustrate the use of cross-validation to decide between weight functions, and show how this framework can be applied in a benefit transfer setting. For our application on willingness-to-pay for water quality improvements we find that the proposed approach brings substantial gains in predictive accuracy in a leave-one-out setting, and vast improvements in predictive efficiency for benefit transfer. These gains are relatively robust to the choice of weight function, suggesting strong potential of this method to generalize to other meta-regression contexts.

Introduction

Borrowing from existing research findings to estimate societal values associated with prospective environmental policy has become an indispensable strategy in agency decision making in the U.S. and abroad (Kaul et al. 2013; Rolfe et al. 2015b; Rosenberger and Loomis 2017; Johnston et al. 2017a; Moeltner et al. 2019). In recent years, such Benefit Transfer (BT) approaches have been increasingly based on underlying Meta-Regression Models (MRMs), which treat welfare estimates and context-specific information from source studies as individual observations in a "re-cycled" econometric model. Estimates from the MRM, in turn, are used to generate BT predictions, much like out-of-sample predictions in standard regression based on primary data (e.g. Bergstrom and Taylor 2006; Moeltner et al. 2007; Johnston and Rosenberger 2010; Boyle et al. 2013).

MRMs have several advantages over simple value transfer or single-study functional transfer, such as the ability to explicitly capture study-specific heterogeneity, more extensive spatial coverage, and flexibility in econometric estimation (Rolfe et al. 2015a; Moeltner et al. 2007; Moeltner 2019; Moeltner et al. 2019). These strengths notwithstanding, the *predictive accuracy* of existing MRMs has generally left room for improvement. For example, typical Mean Average Percentage Errors (MAPEs) in leave-one-out (LOO) exercises have been in the 60-80% range when error variance is ignored, and even larger, up to 120-130% when it is incorporated in the prediction process (Stapler and Johnston 2009; Johnston et al. 2017a, 2019; Moeltner et al. 2019; Vedogbeton and Johnston 2020). In the same vein, in studies that report uncertainty or confidence bounds along with BT point estimates, the range of these bounds span between two and eight times the value of the respective mean estimate (Moeltner et al. 2007; Moeltner 2019; Moeltner et al. 2019). In many cases, this makes confidence bounds unsuitable to inform policy.

Poor predictive performance in a regression context is often driven by wrong or inflexible specification of the mean function, and lack of accommodation of observation- or group-specific heterogeneity in model coefficients. Both factors likely apply to MRMs, where in most cases a single model is fitted to all source points, and is asked to cover all possible BT scenarios in terms of out-of-sample predictions. Even MRMs that allow for some non-linearity in key regressors (Moeltner 2019), or those that apply model searches to identify the optimal mix of explanatory variables (Moeltner 2019; Moeltner et al. 2019) still rely, in the end, on a single underlying meta-function (e.g. linear in all or most coefficients). As evidenced by the wide range of resulting confidence bounds for BT predictions from even these advanced implementations, these adjustments are not sufficient to substantially improve predictive efficiency or - in case of LOO exercises - predictive accuracy.

In theory, flexibility could be improved by allowing for fixed effects and / or interaction terms corresponding to clusters of observations. However, given the typically modest sample sizes of existing MRMs (often under 100 observations) this will quickly exceed available degrees of freedom to identify all model parameters.

For the exact same reasons analysts in related fields, such as geography and urban / regional planning, have long resorted to Locally Weighted Regression (LWR) models as an attractive alternative to "global" models, such as the single-regression approach (e.g. Cleveland 1979; Cleveland and Devlin 1988; Loader 1999; McMillen and Redfearn 2010). The LWR approach can be interpreted as a form of weighted least squares, where a separate regression is fitted to each observation or cluster of observations. Observation-specific weights, in turn, are determined by a kernel function that itself is based on observable variables. This assures a fully flexible regression function without a proliferation of parameters and corresponding degrees-of-freedom. This strategy has been found particularly fruitful in hedonic analysis of housing markets, where LWRs (often implemented as "Geographically Weighted Regression" with a spatial weight kernel) have helped discover spatial variation in model coefficients that would have remained unnoticed otherwise, and improved predictive properties (Brunsdon et al. 1996; Fotheringham et al. 1997, 2002; Páez et al. 2002; Redfearn 2009; McMillen and Redfearn 2010; Páez et al. 2011)

To our knowledge, this study is the first to incorporate LWR techniques into MRM modeling and BT predictions.[1] Using meta-data on WTP for water quality improvements, we show how local weights can be constructed from observable variables that are already included in the meta-data, or publicly available variables that are extraneous to the meta-data itself. We further illustrate how Cross-Validation (CV) procedures can be employed to compare the performance of different settings for the weight function, and how the concept of LWR translates in intuitive fashion to the BT prediction stage. We find that LWR can vastly decrease predictive error compared to the single-equation approach in an LOO context, and substantially tighten the range of confidence bounds for BT estimates based on stylized but plausible policy settings.

Modeling framework

As starting point for our Locally-Weighted MRM (LW-MRM) we select a single or a set of "weight variables" that will determine how much influence each meta-observation is given in a specific local regression model. Taking the single-variable case for simplicity, this weight variable, say z, splits the data into $g = 1 \dots G$ groups that each share the same setting for z. This groups will often, but not always, overlap with un-

[1]Kaul et al. (2013) use a fully non-parametric framework to estimate a meta-regression of BT errors reported in a set of validation studies. Their work differs form ours in four dimensions: (1) Their MRM is not designed to explain WTP estimates themselves, but rather BT errors from studies that included a validation component, (2) they do not directly focus on BT errors related to MRMs, (3) their non-parametric model does not include any explanatory variables in the core regression function, and (4) they do not examine the impact of the choice of kernel function and weight variables on predictive performance.

derlying source studies. For example, our meta-data comprises $n = 188$ observations from $S = 58$ distinct source studies. However, for all weight variables we consider, G exceeds S, as many observations differ in the setting of z, even within the same source study.[2]

Let each of these G clusters of meta-data be labeled as a "location." Consequently, the LW-MRM estimates G separate regressions, one for each location. At each location, in turn, all "home" observations that share the same z will receive a weight of one, while all "guest observations" receive a weight smaller than one, with some observations potentially excluded depending on the chosen kernel function.

As a baseline MRM we choose the specification described in Moeltner (2019) and Moeltner et al. (2019) as "MRM2." This is a relatively simple linear model that has been found in those studies to comply well with important utility-theoretic properties, such as scope and adding-up. For a given observation (or "site") i flowing from study s it is formally given as:

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

where y_i is the willingness to pay estimated for observation i for a water quality change from $q_{0,i}$ to $q_{1,i}$, $\mathbf{x}_i$ a vector of context-specific regressors, i.e. variables that are usually available for any targeted policy site in a BT scenario, $\mathbf{m}_i$ is a vector of methodological indicators (usually fixed at the study level), and ϵ_i is an idiosyncratic error term that follows a normal distribution with mean zero and variance σ^2. While more complex error structures are certainly possible within our LWR framework, a simple independent error produced superior fit with the underlying meta-data in

[2]As pointed out in McMillen and Redfearn (2010), in theory LWR could proceed with local regressions for each individual observation. However, this substantially increases computational burden, often without commensurate gains in model performance.

Moeltner et al. (2019), and also helps with computational speed, given the many models to be estimates as part of our weight search, described below in more detail.

As is clear from (1), this baseline MRM estimates a single set of model coefficients $\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\beta}' & \boldsymbol{\gamma}' & \delta \end{bmatrix}'$, assumed to hold for all study sites and observations. In turn, a single vector of estimated coefficients, say $\hat{\boldsymbol{\theta}}$, is then applied to any policy site for which transferred benefits are sought. Borrowing from the geography literature (e.g. Fotheringham et al. 2002; McMillen and Redfearn 2010), we will henceforth refer to this standard MRM as the "Globally Linear MRM," or GL-MRM."

In contrast, the LW-MRM relaxes this constraint by estimating a separate model at each sample location. Generically, our LW-MRM for any given location g can be expressed as:

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

where $w_{i,g}$ is the weight assigned to observation i with respect to location g. It should be noted that all model parameters, including the variance term, now carry a location-specific subscript. This stresses again that the LWR approach ultimately produces G different sets of parameter estimates. As we illustrate below, this not only greatly improves predictive accuracy, but also enhances flexibility and efficiency in the BT step.[3]

[3] In the terminology of McMillen and Redfearn (2010), our approach can also be described as "Conditionally Parametric Regression" (CPR), a special case of LWR that does not use all original explanatory variables in the weight function.

Weight construction

The weights in (2) are determined by a kernel function, generically given as

$$(3) \quad K(\mathbf{z}_i, \mathbf{z}_g, c)$$

where $\mathbf{z}$ denotes a set of one or more weight variables (that may or may not overlap with $\mathbf{x}$ in the main regression), and c is a "window size" parameter that dictates how many "guest observations" will be allowed to contribute to the regression at location g, i.e. will receive non-zero weight.[4] As discussed in Redfearn (2009), a smaller c will sharpen local predictions (reduce bias), but increase variance. The opposite effect occurs when c approaches n, the total sample size. This will blur the local differences in regression estimates and may thus lead to "over-smoothing." In an MRM context, an additional constraint on c is implicitly given by the modest sample sizes relative to a large set of regressors needed for BT, and resulting identification problems if too many observations are excluded from a given local regression. In our application below we thus experiment with window sizes in the 55-100% range.

An integral component of the kernel function is the "proximity" of point i to location g, usually expressed as Euclidean distance between $\mathbf{z}_i$ and $\mathbf{z}_g$ that is $d_{i,g} = \sqrt{(\mathbf{z}_g' \mathbf{z}_i)}$. In fact, the first step in constructing local weights is akin to nearest-neighbor matching (e.g. Ho et al. 2007; Abadie and Imbens 2011; Johnston and Moeltner 2019), in that $d_{i,g}$ is computed for all observations in the sample, observations are sorted in ascending order of distance, and only the c observations with the smallest distance measure are retained for estimation.

Once the local sample has been "cut" in that fashion, each retained observation

[4] Alternatively, the kernel function can be specified with a bandwidth parameter instead of window size. However, for the reasons discussed in Fotheringham et al. (2002) and McMillen and Redfearn (2010), i.e. risks of extreme over-or under-smoothing under a fixed bandwidth, we prefer to work with window size instead.

receives a continuous weight based on a chosen weight function. Three common choices are Gaussian, bi-square, and tri-cubic (Fotheringham et al. 2002; Farber and Páez 2007; Redfearn 2009). The first is given as:

$$(4) \quad w_{i,g} = exp\left(-\frac{1}{2}\left(d_{i,g}/d_{c,g}\right)^2\right),$$

where $w_{i,g}$ is the weight assigned to observation i for home location g, $d_{i,g}$ is the distance metric defined above, and $d_{c,g}$ is the maximum distance in the set of observations at location g, as (implicitly) imposed by window size c. As is evident from (4), the Gaussian function assigns a maximum weight of one to all home observations (since $d_{i,g} = 0$ for those cases), and a minimum weight of $exp\left(-\frac{1}{2}\right) = 0.6065$ for the observation(s) at the boundary of the chosen window size, with $d_{i,g} = d_{c,g}$.

A second standard weight function is the bi-square function, given as:

$$(5) \quad w_{i,g} = \left(1 - \left(d_{i,g}/d_{c,g}\right)^2\right)^2$$

In this case, the observation(s) with the largest distance relative to location g will be assigned a weight of zero, with all other observations receiving a weight between zero and one. A closely related third weight function we will consider is the tri-cubic:

$$(6) \quad w_{i,g} = \left(1 - \left(d_{i,g}/d_{c,g}\right)^3\right)^3,$$

which exhibits the same general properties as the bi-square.

Figure 1 depicts the weight trajectories over distance for four of our chosen weight variables in the empirical section below. The panels in the left column are based on a maximum window size of $c = 188$, our entire sample, while the panels in the right column pertain to a much smaller window size of $c = 110$. The three key insights to be gained from this figure are (i) the bi-square and tri-cubic both reduce weights to zero at the window size limit and otherwise exhibit similar trajectories, (ii) the Gaussian reduces weights at a much slower rate, with a minimum weight of approximately 0.6 reached at the window size boundary, and (iii) a tighter window size reduces the maximum distance at which a guest observation still receives non-zero (or non-minimum) weight, as is evident by comparing the range of the x-axis across panel rows.

Bayesian estimation

While the LW-MRM in (2) can be estimated in straightforward fashion via Weighted Least Squares (WLS), we prefer a Bayesian approach primarily for two reasons: (i) non-dependence on asymptotic theory for inferential purpose, especially given the modest size of our data, and (ii) ease of obtaining full posterior distributions for secondary estimation outputs, such as BT predictions for different policy settings.

Letting $y_i^* = \sqrt{w_{i,g}}\ log\left(\frac{y_i}{q_{1,i}-q_{0,i}}\right)$, $\mathbf{x}_i^* = \sqrt{w_{i,g}}\ \mathbf{x}_i$, $\mathbf{m}_i^* = \sqrt{w_{i,g}}\ \mathbf{m}_i$, and $q_i^* = \sqrt{w_{i,g}}\ \left(\frac{q_{0,i}+q_{1,i}}{2}\right)$, the likelihood function for home location g can be written as:

(7)
$$p\left(\mathbf{y}^*|\boldsymbol{\theta}_g, \sigma_g^2, \tilde{\mathbf{X}}^*\right) = (2\pi)^{-n_c/2}\left(\sigma_g^2\right)^{-n_c/2} exp\left(-\frac{1}{2\sigma_g^2}\left(\mathbf{y}^* - \tilde{\mathbf{X}}^*\boldsymbol{\theta}_g\right)'\left(\mathbf{y}^* - \tilde{\mathbf{X}}^*\boldsymbol{\theta}_g\right)\right),$$

where n_c is the set of observations with no-zero weight, as imposed by window size c, $\mathbf{y}^*$ is the vector of (weighted) outcomes, $\tilde{\mathbf{X}}^* = \begin{bmatrix}\mathbf{X}^* & \mathbf{M}^* & \mathbf{q}^*\end{bmatrix}$ is the full (weighted) matrix of explanatory data, and $\boldsymbol{\theta}_g = \begin{bmatrix}\boldsymbol{\beta}_g' & \boldsymbol{\gamma}_g' & \delta_g\end{bmatrix}'$ collects all corresponding model coefficients.

Assigning standard multivariate-normal and inverse-gamma priors to $\boldsymbol{\theta}_g$ and σ_g^2, respectively, produces a posterior kernel from which draws of $\boldsymbol{\theta}_g$ and σ_g^2 are obtained via a basic Gibbs Sampler (GS). Detailed econometric underpinnings for this sampling procedure are given in Moeltner et al. (2007), Moeltner (2019), and Moeltner et al. (2019). This process is then repeated for each of the $g = 1 \dots G$ locations, as defined by the current weight variable z.

Cross-validation

We repeat the entire estimation process for all G locations for different choices of weight variables, window sizes, and weight functions. Let each of these unique combinations of weight settings be labeled as $j = 1 \dots J$. This then allows for the use of Cross-Validation (CV) to examine the comparative performance of each setting. As described in Fotheringham et al. (2002), and Farber and Páez (2007), CV is similar to a standard LOO exercise, where each local regression is repeated without the home observation, and local regression output is then used to predict that omitted observation. If a given location comprises multiple home observations, the local process is repeated sequentially for each of these cases.

In other words, each location produces a set of $\mathcal{L}_g$ out-of-sample predictions, where $\mathcal{L}_g$ is the number of home observations at location g. The total number of predictions across all locations is equal to the sample size n, if all local regressions are identified in the main estimation stage, and all observation-deficient local regressions are identified in the CV stage, as discussed below in more detail.

The output from this CV exercise can then be used to compute summary statistics for predictive fit. We follow the bulk of the MRM / BT literature and use the Mean Absolute Percentage Error (MAPE) as accuracy criterion (Stapler and Johnston 2009; Johnston et al. 2017a, 2019; Moeltner et al. 2019; Vedogbeton and Johnston 2020). Letting the total number of predictions across all locations be n_{CV}, the MAPE associated with a given weight setting j can be constructed as:

(8)

$$
\begin{aligned}
&MAPE_j = \frac{100}{n_{CV}} \sum_{g=1}^{G} \sum_{i=1}^{\mathcal{L}_g} |\frac{\bar{\hat{y}}_{i,j,g} - y_i}{y_i}| \quad \text{with} \\
&\bar{\hat{y}}_{i,j,g} = \frac{1}{R} \sum_{r=1}^{R} \left(exp\left(\mathbf{x}_i' \boldsymbol{\beta}_{j,g,r} + \mathbf{m}_i' \boldsymbol{\gamma}_{j,g,r} + \delta_{j,g,r} \left(\frac{q_{0,i} + q_{1,i}}{2} \right) + \epsilon_r \right) + (q_{1,i} - q_{0,i}) \right), \\
&\epsilon_r \sim n\left(0, \sigma^2_{j,g,r}\right)
\end{aligned}
$$

where r denotes a specific iteration of the GS. In words, we compare the mean of the posterior predictive distribution of $y_{i,\hat{j},g}$ to its actual value. As is evident from (8), we follow Moeltner et al. (2019) and include draws of the regression error ϵ_i to derive our predictions for a more conservative estimate of predictive error.

Benefit transfer

The final step of our analysis is a stylized BT exercise with hypothetical policy settings $\mathbf{x}_p$ and a hypothetical water quality change from $q_{0,p}$ to $q_{1,p}$. For a given (implicit) weight setting j, location g, and the r^{th} draw of model parameters, this produces a draw from the posterior predictive distribution (PPD) of estimated WTP, in dollars, written as:

(9)

$$
\hat{y}_{p,j,g,r} = \frac{1}{T} \sum_{t=1}^{T} \left(exp\left(\mathbf{x}_p' \boldsymbol{\beta}_{j,g,r} + \mathbf{m}_t' \boldsymbol{\gamma}_{j,g,r} + \delta_{j,g,r} \left(\frac{q_{0,p} + q_{1,p}}{2} \right) + \epsilon_r \right) + (q_{1,p} - q_{0,p}) \right),
$$

where ϵ_r is, again, drawn from the normal density as for the MAPE construction above. We follow Moeltner et al. (2007), Moeltner and Rosenberger (2008), and Moeltner et al. (2019) and neutralize the effect of moderator variables, which are de facto "nuisance terms" in the BT step, by averaging predictions across all $t = 1 \dots T$ permissible combinations of moderator vector $\mathbf{m}$, with each combination receiving uniform weight $1/T$.

We repeat this process for each location. Thus, we originally treat the policy context as if it were a "home" observation at each of the G locations, and obtain a full posterior distribution of un-weighted predictions for each location. In a second step, we select one of our $j = 1 \dots J$ weight settings (= combination of weight variable(s), window size, and kernel function) and compute the relative weight of policy context p vis-a-vis each of the G locations, that is we derive

$$(10) \quad w_{p,j,g} = \mathcal{K}_j \left(d_{p,j,g} / d_{c_j,j,g} \right), \quad \forall g,$$

where $\mathcal{K}(.)$ is one of our three kernel functions. Naturally, computation of the Euclidean distance $d_{p,j,g}$ rests on the assumption that the value of the weight variable(s) under consideration for weight version j is (are) known for the policy context.

Next, still within a given weight setting j, we obtain a draw from the *location-weighted PPD of WTP* as the weighted average of all location-specific draws, formally given as

$$(11) \quad \hat{y}_{p,j,r} = \sum_{g=1}^{G} \left(\frac{w_{p,j,g}}{\sum_{g=1}^{G} w_{p,j,g}} \right) \hat{y}_{p,j,g,r}$$

Repeating this process for all $r = 1 \dots R$ iterations of the GS produces the full location-weighted PPD. As can be seen from (11), we apply *normalized weights* $\left(\frac{w_{p,j,g}}{\sum_{g=1}^{G} w_{p,j,g}} \right)$ in this step to capture the notion of a discrete probability mass function of BT predictions across locations.[5]

[5] We also perform this analysis using raw weights $\left(\frac{w_{p,j,g}}{G} \right)$. The latter reduce BT estimates slightly relative to the normalized approach (since $G > \sum_{g=1}^{G} w_{p,j,g}$ by construction), with relative differences in the 5-10% range.

To avoid having our BT predictions driven by a single weight setting, we repeat this entire step for a selected subset of J_p weight functions that showed promise in the MAPE exercise to derive draws of “weight-setting-averaged” BT predictions as:

$$(12) \quad \hat{y}_{p,r} = \frac{1}{J_p} \hat{y}_{p,j,r}$$

As before, performing this averaging over all R draws generated by all underlying Gibbs Samplers produces the full PPD for this construct. This distribution, in turn, forms the basis for further statistical examination, such as measures of central tendency and confidence bounds.

Empirical implementation

Data

We illustrate the LW-MRM framework using meta-data on WTP for water quality improvements in rivers and lakes. The core of this set has been used by U.S. EPA for rule making related to the steam electric power generation (U.S. Environmental Protection Agency 2015, 2020). It was also featured in recent academic publications to highlight MRM-related challenges and methods (e.g. Newbold et al. 2018; Johnston et al. 2019; Moeltner 2019). The original data comprised 140 observations from 51 source studies. For this application, we added 73 additional observations from 14 new contributions (recently published work, or older studies that were deemed suitable for inclusion upon further inspection). Conversely, we dropped 25 observations and seven studies from the original set, primarily due to overlap in underlying populations and study sites with more recent contributions. In total, the updated set thus includes

188 observations from 58 studies.[6]

We also changed the composition of explanatory variables slightly compared to Newbold et al. (2018) and Moeltner (2019) to optimize the within-fit of the augmented MRM. The resulting list of contributing variables along with summary statistics is given in Table 1. As is evident from the table, we separate our set of regressors into 15 variables for which exact settings will be known in a typical BT context, and seven moderator variables over which we average our predictions, as illustrated in equation (9). The combined set of 22 regressors is used in the GL-MRM, as well as all versions of the LW-MRM.

Table 2 shows the 12 different weight variables or variable combinations we consider to group the data into locations and construct our regression weights. While initial weight selection is inherently arbitrary in our context, we ex ante focus on weight variables that are (i) continuous in nature to yield a large number of locations, and (ii) capture features of a source study's underlying population or spatial characteristic. Only two of these weight variables, `lnag` (the log of the affected resource area that is agricultural based on the USGS's National Land Cover Database), and `lnincome` (logged median household income in the sampled area based on census data) are internal to the MRM (and thus captured in both Tables 1 and 2).

The remaining individual weight variables are extraneous to the revised MRM. Two of them, `lnpop` (logged population residing in the affected resource area) and ln_sz_ratio (logged ratio of the size of the affected resource area over the population-averaged distance from the resource) had been incorporated in earlier installments

[6]Compared to the original meta-data, we added Bockstael et al. (1989) (one observation), Whitehead et al. (1995) (one observation), Schulze et al. (1995) (two observations), Johnston et al. (2002) (one observation), Collins et al. (2009) (one observation),Zhao et al. (2013) (three observations), Johnston and Ramachandran (2014) (three observations), van Houtven et al. (2014) (32 observations), Nelson et al. (2015) (two observations), Interis and Petrolia (2016) (ten observations), Holland and Johnston (2017) (six observations),Johnston et al. (2017b) (three observations),Moore et al. (2018) (two observations), and Choi and Ready (2021) (six observations).

of the MRM (e.g. Johnston et al. 2019), but were no longer found to significantly contribute to model fit in the current version. The remaining four were extracted from the EPA's database on the land features of catchment areas for lakes and streams, as described in Hill et al. (2016) and Hill et al. (2018). We merged this data with spatial information from our source studies to derive the percentage of catchment areas covered by a given study with developed land (`pctdev`), open and agricultural land (`pctopen`), forest land (`pctfor`), and wetlands (`pctwet`).

The remaining four weight scenarios are combinations of single variables. The first combination (`combo 1`) uses the first four variables listed in the table, that is all variables that are or have previously been incoporated in the water quality MRM. The second combo (`combo 2`) collects the four catchment area variables, and the remaining combinations (`combo 3` and `combo 4`) pair one MRM variable and one catchment variable, respectively, as described in the last column of the table. Overall, these combinations were chosen in part to simply illustrate the (straightforward) use of combined weights, and in part based on CV performance after initial estimation rounds.

As can be seen from the second column of Table 2, these weight choices produce a relatively large number of locations, as desired, ranging from 65 for the catchment variables to 79 for the first combination weight. Thus, on average, every 2-3 sample observations will share a common location.

Cross-validation results

We combine each of the 12 weight variables with the three kernel functions discussed above, and, within each variable-function pair, seven window sizes, ranging from the entire sample ($c = n = 188$) to $c = 100$, approximately 53% of total sample size. For each of the resulting 12 x 3 x 7 = 252 weight settings (denoted as $j = 1 \dots J$ above) we run the corresponding LW-MRM across all underlying locations, and perform the Cross-Validation (CV) exercise as described above.

In terms of Bayesian estimation, every single model (including the GL-MRM) is estimated with the same vague prior settings of a mean vector of zeros and a variance of ten for the coefficients, and a setting of $1/2$ for both the shape and scale of the inverse-gamma prior for the error variance. In each case, we discard the first 100,000 draws as "burn-ins" to countervail any lingering effect of starting values, and retain the following 100,000 draws for further analysis.

Table 3 shows CV results for one of our weight variables, `lnpop`. The first column gives the stipulated window size c. For some locations multiple observations share the same limiting distance. We include all of these "boundary" observations, which produces a slightly larger "effective window size" for at some locations, as shown in the second column of the table. The third column gives the number of identified local regressions. In absence of any identification problems, this will be the number of locations for the sample at large. However, at smaller window sizes, the data may be too truncated to fully identify all regressors at certain locations. This leads to a reduction in the number of local regressions that can be estimated. Similarly, the fourth column shows how many observations ultimately feed into the computation of MAPE as given in equation (8). This will be the total sample sizes minus observations lost due to unidentified local regressions in the first place, minus an additional attrition of observations that, in the leave-one-out stage of the MAPE procedure, produce singularities in the underlying regression. The remaining four columns then report the mean, median, minimum, and maximum Absolute Percentage Error (APE) over all underlying locations. In the following discussion we primarily focus on the mean, i.e. the MAPE.

The first row of the table gives results for the generic GL-MRM. We note that it generates a MAPE of approximately 112%. As is evident from the remaining entries of the "mean" column, MAPEs for most LW-MRMs remain far below this benchmark, with accuracy gains of up to 45% relative to the GL-MRM. Specifically, MAPEs tend to improve with decreasing window sizes, especially in the 110-100 range. However, this comes with the caveat that the effective sample size drops abruptly at $c = 100$, with only 119 observations contributing to the MAPE statistic due to identification problems. It is therefore problematic to interpret the corresponding MAPE as "representing the bulk of the data," and diminishes its comparability to the GL-MRM.

We also note that the bi-square and tri-cubed kernel function generate some of the lowest observed MAPEs. However, they are also prone to producing particularly poor out-of-sample predictions, especially in the mid-range of window sizes. This stresses the importance of ex ante considering a large range / set of window sizes to gauge the performance of the LW-MRM for a given weight setting.

Figure 2 provides a comprehensive overview for all 252 experimental scenarios. Each of the 12 panels shows CV results for a given weight variable or combination. Within each sub-panel, we plot the MAPE for the GL-MRM as solid line with a y-intercept of 112.378, as well as the 21 MAPEs corresponding to the LW-MRM versions, with each line representing a different kernel function, and each marker a specific window size. The main take-home points conveyed by this figure are: (i) Most LW-MAPEs fall below the GW-MRM benchmark, (ii) MAPEs can vary erratically across window sizes for some weight variables (e.g. `lnpop`, `pctwet`, `combo 1`), (iii) window sizes generally have a stronger effect on MAPE than kernel functions, and (iv) for a given weight variable, the smallest MAPE can occur at any window size.

Overall, we conclude from this exploratory analysis of out-of-sample predictive accuracy that LW-MRM has the potential to vastly outperform a standard MRM.

Reassuringly, every single weight variable or combination we examine outperforms the GL-MRM benchmark for at least one window size, as is evident from Figure 1. That said, these findings also stress the need to condider a large number of weight settings to find combinations of variable, kernel function, and window sizes that "click," and jointly enhance the fit of the meta-model.

Benefit transfer results

To examine if these efficiency gains hold up for a "fully out-of-data" prediction, we construct three stylized BT scenarios. Specifically, we set "study year" to 2021, "non-annual payments" to 0, "tax payment" to 1 (mimicking an annual tax increase), "non-users" to 1 (as has been standard practice in recent EPA applications), "central" to 1 and other regions to 0, "swimming" and "fishing" affected to 1, "user fees" to 0, and all continuous variables to their sample mean. In terms of water quality change, we consider a common baseline of $q_0 = 65$, and three endpoint scenarios for q_1 of 67.5, 70, and 75, respectively. These settings correspond to a moderately polluted water body and modest improvements, as they are typical in many policy contexts.[7] All weight variables feeding into the BT process are also set to the sample mean, as given in the fourth column of Table 2.

For each of these three policy scenarios we generate locally-weighted predictive distributions for all 12 weight variables / combinations. Within each combination, we select the kernel function that performed best for window sizes of 170 or 188, after preliminary estimation rounds revealed unstable predictive behavior at smaller window sizes for some weight variables. Maintaining larger window sizes also assures a (desirable) large number of identified local regressions, and conveys the notion that we are drawing from the full or near-full meta-data when constructing the transfer.

[7] In the terminology of Johnston and Besedin (2009), moving from an index of 65 to 70 or higher would render the resource from "fishable" to "swimmable."

The results for these three BT scenarios are captured in tables 4 through 6, respectively. The tables give the weight variable in the first column, followed by window size ("ws"), the number of identified local regressions ("loc") and the kernel function type. The remaining four columns present the posterior mean of the predictive BT distribution, along with the lower and upper bound of the 95% Highest Posterior Density Interval (HPDI), which can be interpreted as the Bayesian analog of a confidence interval.[8] The last column shows the range spanned by the two bounds.

The first row of each table gives BT results for the GL-MRM. This is followed by results for the individual weight variables, and, in the last row, results for the aggregate BT distribution averaged over all weight settings, as captured in equation (12).

As is clearly evident from all three tables, the locally-weighted transfers vastly outperform those flowing from the GL-MRM, regardless of water quality scenario, and for essentially all weight settings considered. While posterior means are of comparable magnitude to those generated by the GL-MRM ($15-20 for scenario 1, $30-40 for scenario 2, and $50-70 for scenario 3), confidence bounds and ranges are by an order of magnitude tighter than those for the standard MRM. The aggregate model, which abstracts from relying on any single weight setting, performs especially well, with confidence ranges that are approximately 10 times tighter than the corresponding range for the GL-MRM for all three quality scenarios.

The same result is conveyed by Figure 3, which plots the full PPDs for all cases considered, with each scenario given in a separate panel. The PPD for the baseline MRM is again plotted as a solid line, PPDs for individual weight settings are dotted, and the PPD for the aggregate BT is given as a dashed line. While there is some

[8] As described inter alia in Koop (2003) the α-% HPDI is the smallest interval over the range of a given distribution that includes α % of the density mass. It is a common Bayesian statistic used to describe the spread of a given distribution, and/or confidence bounds for a given parameter of predictive construct of interest.

variability in predictive span and location across weight settings, most of them center around the same posterior mean, and all of them have posterior ranges that are a fraction of that covered by the GL-MRM. In fact, compared to the distributions for the LW-cases, the benchmark prediction appears as an almost flat line with a substantially larger range.

In sum, we take these combined results from the BT exercise as overwhelming evidence that there the LW-MRM approach has great potential to generate substantial efficiency gains in the BT process, for a wide range of weight function settings. In addition, and reassuringly, these efficiency gains are equally pronounced for the weight-aggregate transfer, which relaxes any constraint of basing BT estimates on a single weight setting.

Conclusion

This study introduces the method of locally-weighted regression to the meta-analysis and benefit transfer literature. We provide econometric underpinnings, discuss the construction of weight functions, and illustrate how CV can be employed to compare weight variables, kernel functions, and window sizes based on out-of-sample predictive accuracy. Reassuringly, we find pronounced reductions in predictive error for most of our weight settings. We then perform a simple BT exercise for selected weight functions, and find that gains in predictive efficiency can be staggering compared to a single-MRM approach.

We consider our analysis as encouraging starting point to further investigate locally-weighted MRM approaches, perhaps using different meta-data, different weight variables, alternative kernel functions, and an even wider range of window sizes. Another fruitful extension would be a validation exercise for situations where WTP estimates or "targets" from external primary studies are available. We believe that these initial findings provide ample ammunition to forge ahead with follow-up research on the use of local weighting to extract more information from environmental valuation meta-data.

References

Abadie, A., Imbens, G., 2011. Bias-corrected matching estimators for average treatment effects. Journal of Business & Economic Statistics 29, 1–11.

Bergstrom, J., Taylor, L., 2006. Using meta-analysis for benefit transfer: Theory and practice. Ecological Economics 60, 351–360.

Bockstael, N., McConnel, K., Strand, I., 1989. Measuring the benefits of improvements in water quality: The Chesapeake Bay. Marine Resource Economics 6, 1–18.

Boyle, K., Parmeter, C., Boehlert, B., R.W., P., 2013. Due diligence in meta-analysis to support benefit transfer. Environmental and Resource Economics 55, 357–386.

Brunsdon, C., Fotheringham, A., Charlton, M., 1996. Geographically weighted regression: A method for exploring spatial nonstationarity. Geographical Analysis 28, 281–298.

Choi, D., Ready, R., 2021. Measuring benefits from spatially-explicit surface water quality improvements: The roles of distance, scope, scale, and size. Resource and Energy Economics 63, 1101108.

Cleveland, W., 1979. Robust locally-weighted regression and smoothing scatterplots. Journal of the American Statistical Association 74, 829–836.

Cleveland, W., Devlin, S., 1988. Locally-weighted regression: An approach to regression analysis by local fitting. Journal of the American Statistical Association 83, 596–610.

Collins, A., Rosenberger, R., Fletcher, J., 2009. Valuing the restoration of acidic streams in the Appalachian Region, in: Thurston, H., Heberling, M., Schrecongost, A. (Eds.), Environmental Economics for Watershed Restoration, pp. 29–52.

Farber, S., Páez, A., 2007. A systematic investigation of cross-validation in gwr model estimation: Empirical analysis and Monte Carlo simulations. Journal of Geographical Systems 9, 371–396.

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

Fotheringham, A., Charlton, M., Brunsdon, C., 1997. Two techniques for exploring nonstationarity in geographical data. Geographical Systems 4, 59–82.

Hill, R., Weber, M., Debbout, R., Leibowitz, S., Olsen, S., 2018. The lake-catchment (Lake-Cat dataset: Characterizing landscape features for lake basins within the conterminous USA. Freshwater Science 37, 208–221.

Hill, R., Weber, M., Leibowitz, S., Olsen, S., Thronbrugh, D., 2016. The stream-catchment (StreamCat dataset: A database of watershed metrics for the conterminous United States. Journal of the American Water Resources Association 52, 120–128.

Ho, D., Kosuke, I., King, G., Stuart, E., 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Policitcal Analysis 15, 199–236.

Holland, B., Johnston, R., 2017. Optimized quantity-within-distance models of spatial welfare heterogeneity. Journal of Environmental Economics and Management 85, 110–129.

Interis, M., Petrolia, D., 2016. Loaction, location, habitat: How the value of ecosystem services varies across location and by habitat. Land Economics 92, 292–307.

Johnston, R., Besedin, E., 2009. Estimating willingness to pay for aquatic resource improvements using benefit transfer, in: Thurston, H., Heberling, M., Schrecongost, A. (Eds.), Environmental Economics for Watershed Restoration, pp. 29–52.

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., 2017a. 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., Moeltner, K., 2019. Special flood hazard effects on coastal and interior home values: One size does not fit all. Environmental and Resource Economics 74, 181–210.

Johnston, R., Ramachandran, M., 2014. Modeling spatial patchiness and hot spots in stated preference willingness to pay. Environmental and Resource Economics 59, 363–387.

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

Johnston, R., Schultz, E., Segerson, K., Besedin, E., Ramachandran, M., 2017b. Biophysical causality and environmental preference elicitation: Evaluating the validity of welfare analysis over intermediate outcomes. American Journal of Agricultural Economics 99, 163–185.

Johnston, R., Swallow, S., Allen, C., Smith, L., 2002. Designing multidimensional environmental programs: Assessing tradeoffs and substitution in watershed management plans. Water Resources Research 38, 479–510.

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.

Koop, G., 2003. Bayesian Econometrics. Wiley.

Loader, C., 1999. Local regression and likelihood. New York: Springer.

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

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., Rosenberger, R.S., 2008. Predicting resource policy outcomes via meta-regression: Data space, model space, and the quest for 'optimal scope'. The B.E. Journal of Economic Analysis and Policy: Contributions 8, Article 31.

Moore, C., Guignet, D., Maguire, K., Dockins, C., Simon, N., 2018. Valuing ecological improvements in the Chesapeake Bay and the importance of ancillary benefits. Journal of Benefit Cost Analysis 9, 1–26.

Nelson, N., Loomis, J., Jakus, P., Kealy, M., von Stackelburg, N., Ostermiller, J., 2015.

Linking ecological data and economics to estimate the total economic value of improving water quality by reducing nutrients. Ecologica Economics 118, 1–9.

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

Páez, A., Farber, S., Wheeler, D., 2011. A simulation-based study of geographically weighted regression as a method for investigating spatially varying relationships. Environment and Planning A 43, 2992–3010.

Páez, A., Uchida, T., Miyamoto, K., 2002. A general framework for estimation and inference of geographically weighted regression models: 1. location-specific kernel bandwidths and a test for locational heterogeneity. Envrionment and Planning A 34, 733–754.

Redfearn, C., 2009. How informative are average effects? Hedonic regression and amenity capitalization in complex urban housing markets. Regional Science and Urban Economics 39, 297–306.

Rolfe, J., Brouwer, R., Johnston, R., 2015a. 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.

Rolfe, J., Johnston, R., Rosenberger, R., Brouwer, R., 2015b. Introduction: Benefit transfer of environmental and resource values, in: Johnston, R., Rolfe, J., Rosenberger, R., Brouwer, R. (Eds.), Benefit transfer of environmental and resource values, pp. 357–382.

Rosenberger, R., Loomis, J., 2017. Benefit transfer, in: Champ, P., Boyle, K., Brown, T. (Eds.), A Primer on Nonmarket Valuation, pp. 431–462.

Schulze, W., Boyce, R., McClelland, G., Robert, D., Breffle, W., 1995. Contingent valuation of natural resource damages due to injuries to the Upper Clark Fork River Basin. Creative Media Partners, LLC.

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

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.

van Houtven, G., Mansfield, C., Phaneuf, D., von Haefen, R., Milstead, B., Kenney, M., Reckhow, K.H., 2014. Combining expert elicitation and stated preference methods to value ecosystem services from improved lake water quality. Ecological Economics 99, 40–52.

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.

Whitehead, J., Blomquist, G., Hoban, T., Clifford, W., 1995. Assessing the validity and reliability of contingent valuation: A comparison of on-site users, off-site users, and non-users. Journal of Environmental Economics and Management 29, 238–251.

Zhao, M., Johnston, R., Schultz, E., 2013. What to value? Ecological indicator choices in stated preference valuation. Environmental and Resource Economics 56, 3–25.

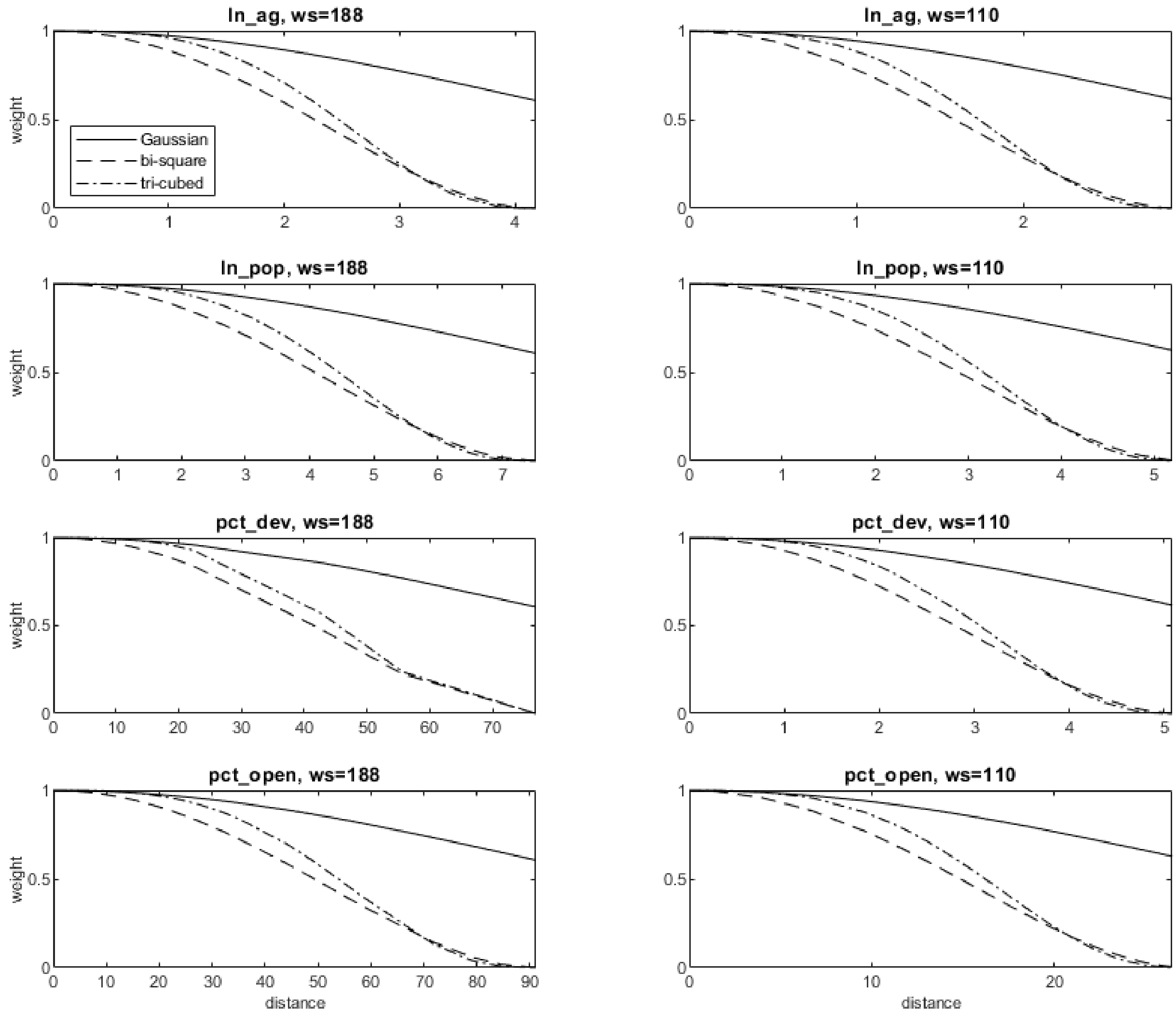

Figure 1: Distances and weights, selected examples

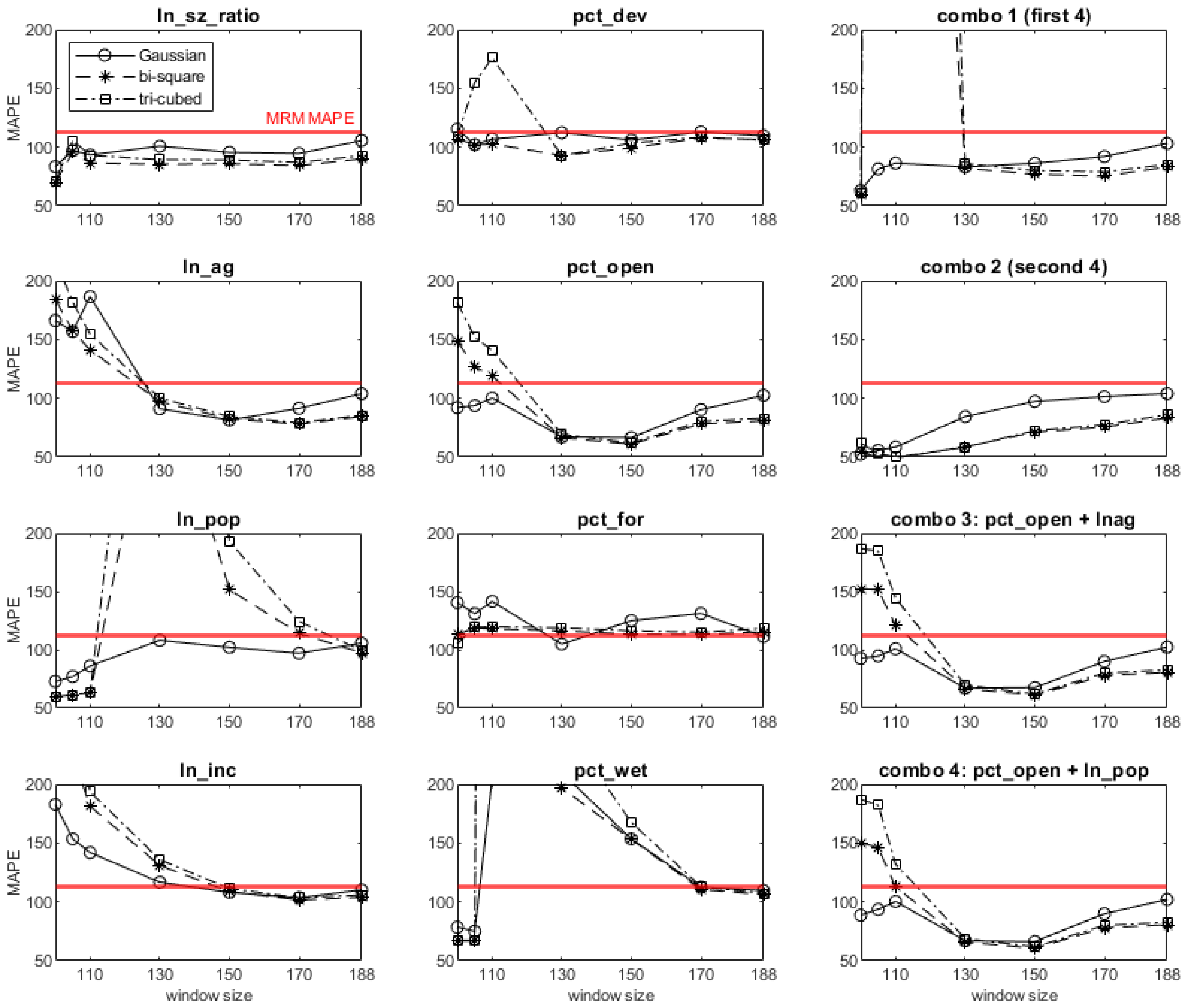

Figure 2: MAPE results by weight variable

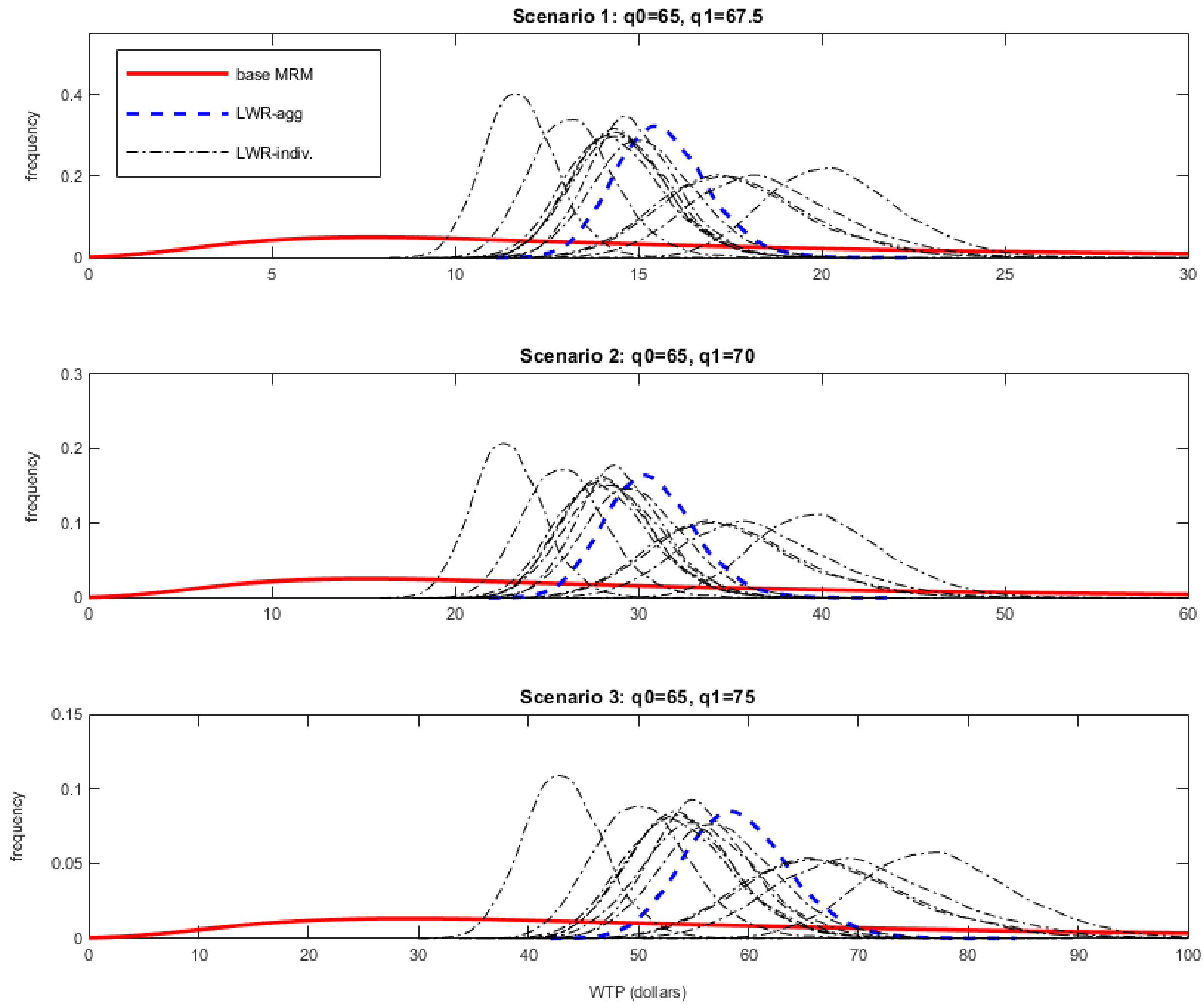

Figure 3: BT predictions by scenario and model

Table 1: MRM variables and statistics

label	mean	std	min	max	description
wtp					sample average or estimated wtp for specified water quality change, 2019$
known for policy context					
lnyear	2.64	0.98	0.00	3.61	log (years since earliest study in the sample (1980))
noann	0.18	0.39	0.00	1.00	1 = payments are not annual (e.g. lump sum)
nonusers	0.06	0.24	0.00	1.00	1 = survey population only includes non-users
northeast	0.13	0.34	0.00	1.00	1 = study conducted in the northeast region of the U.S.
central	0.30	0.46	0.00	1.00	1 = study conducted in the central (midwest or mountian plains) region of the U.S.
south	0.30	0.46	0.00	1.00	1 = study conducted in the southeast or southwest region of the U.S.
swim	0.22	0.41	0.00	1.00	1 = changes in swimming use emphasized in survey
fish	0.19	0.39	0.00	1.00	1 = changes in game fishing emphasized in survey
lnag	-1.64	0.90	-4.26	-0.08	log of affected resource area that is agricultural
lnincome	10.95	0.16	10.65	11.48	log (median U.S. income in year of data collection, 2019$)
paytax	0.40	0.49	0.00	1.00	1=payment mech = increased taxes
payuse	0.02	0.14	0.00	1.00	1=payment mech = increased user cost
ln_ar_ratio	-0.59	2.41	-8.48	6.65	log of sampled area ("market area") divided by the affected resource area
q0 (baseline)	46.90	15.16	10.00	85.00	baseline water quality, 100-point WQI scale
q1	59.99	14.07	12.50	95.00	target water quality, 100-point WQI scale
moderators					
thesis	0.08	0.27	0.00	1.00	1 = study was a PhD or Master's thesis
volunt	0.05	0.23	0.00	1.00	1 = payment vehicle described as voluntary
nonrev	0.16	0.37	0.00	1.00	1 = study was not published in a peer-reviewed journal
sametype	0.35	0.40	0.00	1.00	proportion of waterbodies of same type in the state or region
oneshotval	0.53	0.50	0.00	1.00	1 = only 1 valuation question given
rum	0.56	0.50	0.00	1.00	1=RUM model used
ibi	0.08	0.27	0.00	1.00	1 = water quality derived from a biological index

std. = standard deviation
MRM = Meta-Regression Model

Table 2: Weight variables and statistics

version	label	identified locations	mean	std	min	max	description
1	ln_sz_ratio	78	7.80	2.52	3.48	14.13	log of (size of affected resource area / population-weighted aerage distance from resource)
2	lnag	74	-1.64	0.90	-4.26	-0.08	log of affected resource area that is agricultural
3	lnpop	75	14.04	1.64	9.34	16.87	log of population within the affected resource area
4	lnincome	74	10.95	0.16	10.65	11.48	log (median U.S. income in year of data collection, 2019$)
5	pctdev	65	9.35	15.52	0.17	77.06	percentage of catchment area that is developed land
6	pctopen	65	29.22	21.26	0.26	91.40	percentage of catchment area that is open and ag. land
7	pctfor	65	31.95	22.62	0.06	84.47	percentage of catchment area that is forest land
8	pctwet	65	11.10	12.51	0.01	64.30	percentage of catchment area that is wetland
9	combo 1	79	-	-	-	-	first four combined
10	combo 2	65	-	-	-	-	second four combined
11	combo 3	75	-	-	-	-	lnag + pctopen
12	combo 4	76	-	-	-	-	lnpop + pctopen

std. = standard deviation

Table 3: Within-predictive results using ln_pop as weight variable

weight function	eff. WS min	eff. WS max	identif. local. reg.	usable obs. for APE	APE mean	APE med	APE min	APE max
none (baseline MRM)	188	188	-	188	112.378	57.834	0.410	4337.237
Gaussian	188	188	75	188	105.722	56.326	0.390	3922.562
	170	173	75	188	97.057	54.331	0.081	2421.326
	150	157	75	188	102.094	53.031	0.109	3150.141
	130	138	75	188	108.161	46.936	0.489	4456.061
	110	114	75	187	86.182	50.957	0.686	1793.051
	105	112	75	187	77.090	50.866	0.534	1178.159
	100	111	53	119	72.871	45.512	0.388	1134.523
Bi-square	188	188	75	188	97.391	50.228	0.124	3945.745
	170	173	75	188	114.104	45.436	0.272	7439.872
	150	157	75	188	151.561	44.157	0.200	14910.479
	130	138	75	188	353.396	43.189	0.156	53317.155
	110	114	75	187	63.136	44.711	0.016	444.110
	105	112	75	187	61.251	43.798	0.033	451.350
	100	111	53	119	59.795	42.772	0.077	453.343
Tri-cubed	188	188	75	188	99.361	52.732	0.002	4019.933
	170	173	75	188	124.352	45.023	0.347	9097.335
	150	157	75	188	193.004	43.458	0.216	22377.800
	130	138	75	188	564.245	42.858	0.015	92757.878
	110	114	75	187	63.069	44.208	0.023	449.492
	105	112	75	187	61.491	43.949	0.164	453.948
	100	111	53	119	59.472	43.011	0.012	455.761

eff. WS = effective window size (allowing for ties)
identif. loc. reg. = number of identified local regressions
usable obs. = usable observations (=identified leave-one-out regressions) for APE
APE = absolute percentage error

Table 4: BT predictions for scenario 1: q0=65, q1=67.5

model	ws	loc	fun	low	mean	high	range
generic MRM				1.095	18.610	50.582	49.487
lnszratio	170	78	bi	16.899	20.366	24.174	7.274
lnag	170	74	bi	9.793	11.779	13.706	3.914
lnpop	188	75	bi	12.566	15.237	18.111	5.545
lninc	170	74	bi	11.919	14.614	17.303	5.384
pctdev	188	65	bi	13.826	17.768	21.951	8.126
pctopen	188	65	bi	11.700	14.388	17.096	5.397
pctfor	188	65	G	14.727	18.565	22.852	8.125
pctwet	188	65	tri	13.524	17.624	21.598	8.075
combo 1 (first 4)	170	79	bi	12.616	14.877	17.346	4.730
combo 2(second 4)	170	65	bi	10.915	13.250	15.536	4.621
combo 3 (lnag + pctopen)	170	75	bi	11.913	14.543	17.089	5.176
combo 4 (lnpop + pctopen)	170	76	bi	12.092	14.513	17.156	5.064
aggregate				13.233	15.627	18.135	4.902

ws = window size
loc = number of local regressions
fun = kernel function (G = Gaussian, bi = bi-square, tri = tri-cubed)
mean = posterior mean
low (high) = lower (upper) bound for 95% HPDI
range = (upper bound - lower bound)

Table 5: BT predictions for scenario 2: q0=65, q1=70

model	ws	loc	fun	low	mean	high	range
generic MRM				2.140	36.591	99.838	97.699
lnszratio	170	78	bi	33.044	40.065	47.354	14.309
lnag	170	74	bi	19.042	22.897	26.659	7.616
lnpop	188	75	bi	24.584	29.893	35.460	10.876
lninc	170	74	bi	23.738	28.784	34.360	10.621
pctdev	188	65	bi	27.308	34.945	43.276	15.967
pctopen	188	65	bi	22.912	28.143	33.440	10.528
pctfor	188	65	G	28.885	36.488	44.842	15.957
pctwet	188	65	tri	26.734	34.684	42.614	15.880
combo 1 (first 4)	170	79	bi	24.778	29.128	34.016	9.238
combo 2(second 4)	170	65	bi	21.460	26.077	30.569	9.109
combo 3 (lnag + pctopen)	170	75	bi	23.139	28.399	33.271	10.132
combo 4 (lnpop + pctopen)	170	76	bi	23.642	28.340	33.553	9.912
aggregate				25.953	30.654	35.566	9.613

ws = window size
loc = number of local regressions
fun = kernel function (G = Gaussian, bi = bi-square, tri = tri-cubed)
mean = posterior mean
low (high) = lower (upper) bound for 95% HPDI
range = (upper bound - lower bound)

Table 6: BT predictions for scenario 3: q0=65, q1=75

model	ws	loc	fun	low	mean	high	range
generic MRM				2.961	70.739	191.246	188.285
lnszratio	170	78	bi	64.213	77.538	91.904	27.691
lnag	170	74	bi	36.318	43.275	50.713	14.395
lnpop	188	75	bi	47.228	57.545	68.068	20.840
lninc	170	74	bi	45.683	55.841	66.356	20.673
pctdev	188	65	bi	52.777	67.600	83.605	30.828
pctopen	188	65	bi	43.939	53.842	64.110	20.171
pctfor	188	65	G	55.715	70.482	86.635	30.920
pctwet	188	65	tri	52.316	67.174	83.097	30.781
combo 1 (first 4)	170	79	bi	47.408	55.848	65.149	17.741
combo 2(second 4)	170	65	bi	41.338	50.522	59.078	17.740
combo 3 (lnag + pctopen)	170	75	bi	44.249	54.156	63.564	19.316
combo 4 (lnpop + pctopen)	170	76	bi	44.857	54.042	63.781	18.924
aggregate				49.887	58.989	68.412	18.525

ws = window size
loc = number of local regressions
fun = kernel function (G = Gaussian, bi = bi-square, tri = tri-cubed)
mean = posterior mean
low (high) = lower (upper) bound for 95% HPDI
range = (upper bound - lower bound)