Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 24-1043

Using Gaussian Process Regression in Two-Dimensional Regression Discontinuity Designs

Lily An
Harvard University

Zach Branson
Harvard University

Luke Miratrix
Carnegie Mellon University

Sometimes a treatment, such as receiving a high school diploma, is assigned to students if their scores on two inputs (e.g., math and English test scores) are above established cutoffs. This forms a multidimensional regression discontinuity design (RDD) to analyze the effect of the educational treatment where there are two running variables instead of one. Present methods for estimating such designs either collapse the two running variables into a single running variable, estimate two separate one-dimensional RDDs, or jointly model the entire response surface. The first two approaches may lose valuable information, while the third approach can be very sensitive to model misspecification. We examine an alternative approach, developed in the context of geographic RDDs, which uses Gaussian processes to flexibly model the response surfaces and estimate the impact of treatment along the full range of students that were on the margin of receiving treatment. We demonstrate theoretically, in simulation, and in an applied example, that this approach has several advantages over current approaches, including over another nonparametric surface response method. In particular, using Gaussian process regression in two-dimensional RDDs shows strong coverage and standard error estimation, and allows for easy examination of treatment effect variation for students with different patterns of running variables and outcomes. As these nonparametric approaches are new in education-specific RDDs, we also provide an R package for users to estimate treatment effects using Gaussian process regression.

VERSION: September 2024

Suggested citation: An, Lily, Zach Branson, and Luke Miratrix. (2024). Using Gaussian Process Regression in Two-Dimensional Regression Discontinuity Designs. (EdWorkingPaper: 24-1043). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/0g9t-gt96

Using Gaussian Process Regression in Two-Dimensional Regression Discontinuity Designs

Lily An[1], Zach Branson[2], and Luke Miratrix[3]

[1, 3]*Harvard University*
[2]*Carnegie Mellon University*

Sometimes a treatment, such as receiving a high school diploma, is assigned to students if their scores on two inputs (e.g., math and English test scores) are above established cutoffs. This forms a multidimensional regression discontinuity design (RDD) to analyze the effect of the educational treatment where there are two running variables instead of one. Present methods for estimating such designs either collapse the two running variables into a single running variable, estimate two separate one-dimensional RDDs, or jointly model the entire response surface. The first two approaches may lose valuable information, while the third approach can be very sensitive to model misspecification. We examine an alternative approach, developed in the context of geographic RDDs, which uses Gaussian processes to flexibly model the response surfaces and estimate the impact of treatment along the full range of students that were on the margin of receiving treatment. We demonstrate theoretically, in simulation, and in an applied example, that this approach has several advantages over current approaches, including over another nonparametric surface response method. In particular, using Gaussian process regression in two-dimensional RDDs shows strong coverage and standard error estimation, and allows for easy examination of treatment effect variation for students with different patterns of running variables and outcomes. As these nonparametric approaches are new in education-specific RDDs, we also provide an R package for users to estimate treatment effects using Gaussian process regression.

1 Introduction

Regression discontinuity designs (RDDs) were first introduced by Thistlethwaite and Campbell (1960) and are widely used in economics, education, political science, and statistics for estimating causal effects. In a classical RDD, treatment assignment is determined by a single covariate, called a "running variable," being above or below a cutoff. For example, education grants may only be provided to families with incomes below a certain amount (Li et al., 2015; Ludwig & Miller, 2007), or students may attend subject-specific summer school if their test score on that subject was below a certain value (Matsudaira, 2008). The goal of an RDD is to estimate the causal effect of treatment on outcomes, while accounting for the discontinuous nature of treatment assignment.

RDDs are valuable because, even though treatment assignment is completely confounded by the running variable, the causal effect *at the cutoff* is identifiable under the mild assumption that the treatment and control outcomes are continuous around the cutoff (Hahn et al., 2001). The causal effect can then be estimated by estimating the mean outcome under treatment and under control near the cutoff. Many have utilized nonparametric methods—in particular, local linear regression methods that upweight subjects near the cutoff—to do so (Cattaneo & Titiunik, 2022; Imbens & Lemieux, 2008; Pei et al., 2022). Local linear regression methods fit regressions of the outcome on the running variable (one below the cutoff, one above), extrapolate these regressions to the cutoff, and estimate the causal treatment effect as the difference between these two extrapolations.

Some RDDs have more than one running variable. For example, in education applications, it is often the case that a treatment—such as access to a tutoring program or high school diploma—is assigned to students if both of their two test scores (e.g., in math and English) fall either below or above some cutoff (Ou, 2010; Papay et al., 2014; Reardon et al., 2010). In geographic RDDs, treatment is assigned to units based on their geographic location which is defined by the running variables of latitude and longitude (Keele et al., 2015; Keele & Titiunik, 2015; Rischard et al., 2021). We focus throughout on RDDs that use two running variables to estimate the effect of one treatment. The cutoffs on the two running variables in a two-dimensional RDD form a boundary within a two-dimensional space, where units are treated on one side of the boundary but not the

other. While the causal estimand in a one-dimensional RDD is the average treatment effect at the cutoff, the causal estimand in a two-dimensional RDD is the boundary average treatment effect, or $BATE$. The $BATE$ represents an overall estimate of the weighted average of treatment effects that have been aggregated along the boundary using researcher-specified weights.

Estimating this causal estimand is not as straightforward as estimating an average treatment effect at the cutoff point in the one-dimensional case using local linear regression. Several common methods attempt to estimate causal effects in a two-dimensional RDD by extending methods for the one-dimensional case, but these fail to target the $BATE$. For example, one method collapses the two running variables into a single running variable, and then follows standard local linear regression methods (Cohodes & Goodman, 2012; Martorell, 2005; Robinson, 2011), while another estimates two separate one-dimensional RDDs for each running variable (Kane, 2003; Papay et al., 2010; Reardon et al., 2010). These approaches, though straightforward extensions of existing methods, reduce the problem to one dimension which loses valuable information (see Rischard et al. (2021) for further discussion). Response surface modeling methods have also been used to estimate two-dimensional RDD causal effects, but they can be very sensitive to model misspecification when conducted parametrically (Dee, 2012; Papay et al., 2014; Papay et al., 2011).

Given these challenges using existing methods for estimating the $BATE$ in two-dimensional RDDs, we propose using a nonparametric surface response modeling approach in education research. This has several advantages over current approaches, namely that it allows us to properly target the $BATE$ and that it flexibly estimates the mean outcome on both sides of the boundary. This makes model misspecification less of a concern. Uniquely, this approach also allows for easy examination of treatment effect variation along the boundary, which is not possible with existing methods. We compare current two-dimensional RDD approaches to nonparametric surface response methods that use Gaussian process regression and locally weighted regression (loess).

Our contributions are twofold: first, we examine how nonparametric surface response methods work for a specific type of two-dimensional RDD. Second, we make these tools more accessible by providing an R package.

In Section 2, we define the notation, setup, and causal estimands for two-dimensional RDDs. In Section 3, we review current methods for analyzing two-dimensional RDDs, and in Section 4.1, we present the Gaussian process regression and loess approaches. We compare these various methods via a simulation study in Section 5. We find that using Gaussian process regression shows slightly better performance in terms of bias, standard errors, and RMSE, in addition to much stronger coverage of treatment effects and estimation of standard errors. We further demonstrate these methods in Section 6 by analyzing an educational policy test-score cutoff in Wisconsin. Here, we show how using Gaussian process regression can help illuminate patterns of treatment effect heterogeneity for differently-scoring students, as well as provide more precise inference over prior analyses that use one-dimensional approaches. Finally, we conclude in Section 7.

2 Notation, Setup, and Causal Estimands

Consider N units, indexed by i, each with treatment assignment W_i, outcome Y_i, and running variables X_{i1} and X_{i2}. We focus on sharp two-dimensional RDDs, where

$$W_i = \begin{cases} 1 & \text{if } X_{i1} \geq c_1 \text{ and } X_{i2} \geq c_2 \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

for fixed, scalar cutoffs c_1 and c_2. This design requires that the individual receive one treatment on the basis of passing two score cutoff thresholds, and receive no treatment otherwise. For example, X_1 and X_2 might be math and reading scores, and the cutoffs represent eligibility to receive a high school diploma, such as with high school exit examinations. The outcome might be eventual college enrollment. Other policy designs may allow for different treatments if only one of the score cutoffs were met, but we do not focus on those here.

Let $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes of unit i under treatment $W_i = 1$ and control $W_i = 0$, respectively. The observed outcomes can then be written as a function of the potential outcomes: $Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$. The outcome $Y_i(1)$ is observed only if unit i is exposed to

Figure 1: An example of the two-dimensional RDD we consider in this paper. The cutoffs are $c_1 = 0$ along Running Variable 1 and $c_2 = 0$ along Running Variable 2. The treated region is in the top right quadrant and the control region is the region to the left and below the boundary defined by the black lines. The causal estimand is the $BATE$, which is estimated along the boundary.

treatment ($W_i = 1$), and $Y_i(0)$ is only observed if unit i is not exposed to treatment ($W_i = 0$).

In sharp RDDs, the probability of receiving treatment for every unit is 0 or 1. There are also fuzzy RDDs, where the probability of assignment to treatment at the threshold still changes, but can change by a smaller jump (see Imbens and Lemieux (2008) for a conceptual review of sharp and fuzzy RDDs). We focus on sharp RDDs because they are particularly common in education applications, where students gain access to a program or treatment if and only if two test scores fall to one side of their given cutoffs.

Figure 1 plots an example of this type of RDD. The data for this figure come from our applied example using Wisconsin student test scores (the two running variables) from students who are

English Language Learners (ELLs) and their treatment of reclassification into non-ELL status. This reclassification treatment involves a change in students' educational experience in schools, where formerly ELL students join "mainstream" non-ELL classrooms and lose access to ELL-specific supports. The cutoffs in this RDD are the $c_1 = 0$ and $c_2 = 0$ black lines, the treated units in the top right quadrant are colored in dark blue, and the control units are colored in light blue.

Other common applications in education assign one treatment if the student performs poorly on either test, such as access to summer school, which would change the inequalities in (1) to an *or* statement between running variable values that were less than the cutoffs. This setup was the focus of a simulation study in Porter et al. (2017) that compared various methods for analyzing two-dimensional RDDs, where their control units were in the top right quadrant and treated units were in the remaining three quadrants. That study simulated 5,000 data points from a standard two-dimensional normal distribution with varying correlations between the two running variables, and we will revisit this study design in Section 5.

Unlike a one-dimensional RDD where the causal estimand is the average treatment effect at a single cutoff value, the causal estimand in this two-dimensional RDD forms from the boundary created by the two cutoffs (colored in black in Figure 1). While geographic RDDs use irregular boundaries, two-dimensional RDD applications in education typically use test score cutoffs or other running variable cutoffs that are single values. Our focus in this paper is therefore on a rectangular boundary. The causal estimand is the weighted average of treatment effects along this boundary, known as the boundary average treatment effect ($BATE$), which depends on the choice of weights along the boundary. To formalize this causal estimand, define the boundary as

$$\mathcal{B} = \{\mathbf{x} : x_1 = c_1 \text{ and } x_2 \geq c_2 \text{ , or } x_2 = c_2 \text{ and } x_1 \geq c_1\} \tag{2}$$

where $\mathbf{x} = (x_1, x_2)$ is some realization of the two running variables.

The treatment effect $\tau(\mathbf{x})$ is a conditional average treatment effect, or $CATE$, function along

the boundary, and is defined as

$$\tau(\mathbf{x}) = \mathbb{E}[Y_i(1) - Y_i(0) | X_{1i} = x_1, X_{2i} = x_2] \tag{3}$$

Throughout, we assume that these conditional expectations are continuous at $\mathbf{x} \in \mathcal{B}$, which is one assumption needed to identify $\tau(\mathbf{x})$ (Hahn et al., 2001; Keele et al., 2015). Then, weights $w(\mathbf{x})$ can be specified for each $\mathbf{x}$ along $\mathcal{B}$ to estimate a the single quantity *BATE* τ^w, which is the weighted average of the treatment effects across all $\mathbf{x} \in \mathcal{B}$:

$$\tau^w = \frac{\int_{\mathbf{x} \in \mathcal{B}} w(\mathbf{x}) \tau(\mathbf{x}) \partial \mathbf{x}}{\int_{\mathbf{x} \in \mathcal{B}} w(\mathbf{x}) \partial \mathbf{x}} \tag{4}$$

Researchers can choose the weights $w(\mathbf{x})$ for the treatment effects along B. The simplest form of weights are uniform weights, defined as $w(\mathbf{x}) = 1$, which place equal weight on every point along the boundary. However, this may not be the most sensible choice if there is little data along some parts of the boundary, such as at the center-top or far-right portions of the boundary in Figure 1. We may not want to place much weight on these areas during inference.

Following this logic, researchers often place more weight on sections of the boundary with denser populations of units, i.e., setting $w(\mathbf{x}) = \rho(\mathbf{x})$, where $\rho(\mathbf{x})$ is the density of units at $\mathbf{x}$. This defines the following causal estimand:

$$\tau^{pop} = \frac{\int_{\mathbf{x} \in \mathcal{B}} \rho(\mathbf{x}) \tau(\mathbf{x}) \partial \mathbf{x}}{\int_{\mathbf{x} \in \mathcal{B}} \rho(\mathbf{x}) \partial \mathbf{x}} \tag{5}$$

Such weights have previously been considered by Wong et al. (2013) and Porter et al. (2017). A benefit of this estimand is that it corresponds to the average treatment effect of the super-population of units residing on the boundary (Imbens & Zajonc, 2011; Keele & Titiunik, 2015; Rischard et al., 2021). In practice, $\rho(\mathbf{x})$ must be estimated, such as by using a kernel density estimator.

There may be situations in which we are also interested in the average treatment effect across particular portions of the boundary. For example, the boundary in Figure 1 consists of a vertical

boundary (where the running variable X_1 is fixed) and a horizontal boundary (where X_2 is fixed), and researchers could estimate the average treatment effect for each of these two different boundaries. Define $\mathcal{B}_1 = \{\mathbf{x} : x_1 = c_1, x_2 \geq c_2\}$ and $\mathcal{B}_2 = \{\mathbf{x} : x_1 \geq c_1, x_2 = c_2\}$ as these sections of the boundary. Then, we have the following causal estimands:

$$\tau_1^w = \frac{\int_{\mathbf{x}\in\mathcal{B}_1} w(\mathbf{x})\tau(\mathbf{x})\partial\mathbf{x}}{\int_{\mathbf{x}\in\mathcal{B}_1} w(\mathbf{x})\partial\mathbf{x}}, \qquad \tau_2^w = \frac{\int_{\mathbf{x}\in\mathcal{B}_2} w(\mathbf{x})\tau(\mathbf{x})\partial\mathbf{x}}{\int_{\mathbf{x}\in\mathcal{B}_2} w(\mathbf{x})\partial\mathbf{x}} \tag{6}$$

Note that τ_1^w and τ_2^w in (6) are special cases of τ^w in (4), where for τ_1^w we have set $w(\mathbf{x}) = 0$ for $\mathbf{x} \in \mathcal{B}_2$, and likewise for τ_2^w we have set $w(\mathbf{x}) = 0$ for $\mathbf{x} \in \mathcal{B}_1$. However, the causal estimands in the two-dimensional RDD literature are usually not written in the form of (6). More typically, they are written in the following form:

$$\tau_1 = \mathbb{E}[Y_i(1) - Y_i(0) | X_{1i} = c_1, X_{2i} \geq c_2], \quad \tau_2 = \mathbb{E}[Y_i(1) - Y_i(0) | X_{2i} = c_2, X_{1i} \geq c_1] \tag{7}$$

For examples of this form for the causal estimands, see Porter et al. (2017), Equations (1) and (2), as well as the 13 papers they discuss in their literature review of two-dimensional RDDs. We prefer the form of (6) over (7) for defining these causal estimands because it is more flexible and more explicit. It is more flexible because it defines the large class of weighted average treatment effects that may be of interest depending on the application, and it is more explicit because it makes clear what kind of "averaging" is being done across the boundary. Meanwhile, the expectation in (7) is usually with respect to some hypothetical infinite population, meaning that it is (implicitly) using population-density weights to define a $BATE$.

We note that identification of these causal estimands for multidimensional RDDs must meet RDD assumptions. These include the continuity of potential outcomes at the boundary, a discontinuity in the probability of treatment at the boundary, and inability of individuals to sort above or below the cutoff (e.g., by manipulating the value of their running variables). For discussion of these assumptions see Reardon and Robinson (2012) and Wong et al. (2013).

In Section 3, we review state-of-the-art methods for analyzing two-dimensional RDDs. These

methods typically use local linear regression to conduct inference for the estimands τ_1^w and τ_2^w instead of the overall treatment effect τ^w due to substantive interest in the different boundary's treatment effects as well as difficulties in targeting the whole boundary's treatment effect. However, τ^w is often the main causal estimand of interest in a two-dimensional RDD, because it reflects the aggregated treatment effects across the entire boundary, rather than sections of it. Furthermore, researchers may also be interested in $\tau(\mathbf{x})$ for each $\mathbf{x} \in \mathcal{B}$, as opposed to aggregating these effects across the boundary. A benefit of our approach is that—in addition to providing estimates for average effects such as τ_1^w, τ_2^w, or τ^w—it provides estimates for $\tau(\mathbf{x})$ for every $\mathbf{x} \in \mathcal{B}$. This allows researchers to assess if there is treatment effect heterogeneity along the boundary, instead of only focusing on average treatment effects (Rischard et al., 2021).

3 Review of Approaches for Two-Dimensional RDDs

Most current RDD methods leverage local linear regression approaches. First we will review local linear regression for one-dimensional RDDs, and then turn to two-dimensional RDD estimation.

3.1 Using Local Linear Regression for One-Dimensional RDDs

Consider the case where there is only one running variable X with cutoff c, and define treatment as $W_i = 1$ if $X_i \geq c$ and 0 otherwise. Because treatment assignment depends on this single variable, it is important to model $\mathbb{E}[Y_i(1)|X_i]$ and $\mathbb{E}[Y_i(0)|X_i]$ to estimate causal effects (Rubin, 1977). As there is no overlap in terms of X between treatment and control units, causal effect estimates will be particularly sensitive to model misspecification, especially the further we extrapolate. Thus, most RDD researchers focus on estimating the average treatment effect at the cutoff, defined as

$$\tau(c) = \mathbb{E}[Y_i(1)|X_i = c] - \mathbb{E}[Y_i(0)|X_i = c] \tag{8}$$

because it requires the least amount of extrapolation. As noted in Hahn et al. (2001), "any nonparametric estimator [can be used] to estimate" $\mathbb{E}[Y_i(1)|X_i = c]$ and $\mathbb{E}[Y_i(0)|X_i = c]$. Local linear regression is a popular method for analyzing one-dimensional RDDs because of its desirable boundary properties (Cattaneo & Titiunik, 2022; Hahn et al., 2001; Imbens & Lemieux, 2008).

Local linear regression solves two least-squares problems (one in the treatment group and one in the control group):

$$\min_{\alpha_T,\beta_T} \sum_{i:W_i=1} \left(Y_i - \alpha_T - \beta_T \cdot (X_i - c)\right)^2 K\left(\frac{X_i - c}{h}\right) \tag{9}$$

$$\min_{\alpha_C,\beta_C} \sum_{i:W_i=0} \left(Y_i - \alpha_C - \beta_C \cdot (X_i - c)\right)^2 K\left(\frac{X_i - c}{h}\right) \tag{10}$$

Here, $K(\cdot)$ is the kernel and h is the bandwidth, which dictate how much weight is placed on each unit during estimation. For example, one common choice for $K(\cdot)$ is the rectangular kernel, defined as $\mathbb{1}(c - h \leq X_i \leq c + h)$, which places an equal amount of weight on each unit within distance h of the cutoff and a weight of zero otherwise. Throughout, we focus on rectangular kernels when conducting local linear regression. A common choice for h is the MSE-optimal bandwidth of Imbens and Kalyanaraman (2012), but others—such as Ludwig and Miller (2007) and Porter et al. (2017)—have used cross-validation to select h, and work such as Calonico et al. (2020) has developed bandwidth choices tailored for inference in RDDs. After these least-squares problems are solved, the average treatment effect at the cutoff is estimated as the difference between the two estimated intercepts: $\hat{\alpha}_T - \hat{\alpha}_C$. Research such as Calonico et al. (2018), Calonico et al. (2014) have used bias corrections to improve local linear regression methods, and this approach has become popular in econometrics (Athey & Imbens, 2017).

There are also alternatives to local linear regression. The local randomization literature (Branson & Mealli, 2018; Cattaneo et al., 2015; Li et al., 2015; Sekhon & Titiunik, 2017) envisions the RDD as an as-if randomized experiment around the cutoff, due to randomness in the running variable. In this setup, they estimate the average treatment effect in a neighborhood of the cutoffs where assignment is assumed unconfounded given observed covariates. Cattaneo et al. (2017) compare

local randomization methods to local linear regression methods, and Díaz and Zubizarreta (2023) consider local randomization for complex RDDs with multiple treatment assignment rules.

Finally, there are Bayesian approaches to RDDs (Alcantara et al., 2024; Chib et al., 2023; Chib & Jacobi, 2016; Geneletti et al., 2015; Li et al., 2015). Branson et al. (2019) proposed using Gaussian process regression to analyze one-dimensional RDDs, and found that this approach exhibited promising coverage, interval length, and mean squared error over local linear regression methods. We propose the use of Gaussian process regression for two-dimensional education RDDs in Section 4.1, but first we will review current other approaches for analyzing two-dimensional RDDs.

3.2 Current Approaches for Analyzing Two-Dimensional RDDs

There are four leading methods in the literature for analyzing sharp two-dimensional RDDs like those presented in Figure 1. Wong et al. (2013) and Porter et al. (2017) reviewed them and conducted simulation studies comparing methods. The first three methods described below attempt to simplify the two-dimensional RDD into a one-dimensional RDD so that local linear regression methods can be used, while the fourth method attempts to model the RDD's two-dimensional nature. We discuss their implementation in Appendix 9.1.

1. **The binding score method**: This method for analyzing a two-dimensional RDD instead analyzes a single one-dimensional RDD. It uses both running variables to define a "binding score" under the rationale that one score (either the minimum or maximum of the running variables for each unit) will be responsible for a unit's assignment to treatment by falling either above or below the cutoff. The binding score then acts as the running variable in a one-dimensional RDD that uses local linear regression. In the case of Figure 1, where units receive treatment only if both running variables meet the threshold, the binding score BS_i is defined as $\min(X_{1i}, X_{2i})$ after X_{1i} and X_{2i} are centered around their cutoffs. This method implicitly makes the assumption that $\tau_1 = \tau_2$ (7) and will target some weighted average of τ_1 and τ_2 if they differ. This method has been discussed in works such as Martorell (2005), Robinson (2011), and Cohodes and Goodman (2012).

2. **The frontier method and pooled frontier**: The frontier method divides a two-dimensional RDD into two one-dimensional RDDs using standard approaches (e.g. local linear regression described in Section 3.1). First, units who are not assigned to treatment because of the values of both running variables are discarded, which may decrease precision due to smaller sample size. For example, the units in the lower-left quadrant of Figure 1 would be discarded because $X_{1i} < c_1$ *and* $X_{2i} < c_2$. Then, once the running variables are centered around their cut point values, the upper quadrants of Figure 1 can be viewed as a one-dimensional RDD across the boundary defined by $c_1 = 0$, where treatment is determined only by X_1 with X_2 being a nuisance variable. Similarly, only X_2 determines treatment across the $c_2 = 0$ boundary in the right-hand quadrants and X_1 is a nuisance variable. This method's two one-dimensional RDD analyses provide estimates for τ_1 and τ_2 in (7), respectively. This method has been discussed in works such as Kane (2003), Papay et al. (2010), and Reardon et al. (2010).

   The pooled frontier method, by contrast, produces a single overall $BATE$ per (4). This method estimates the $BATE$ as a weighted average of the treatment effects on both segments of the boundary, using the proportion of data close to each side of the boundary as the weights (see Wong et al. (2013), Equations (3) - (5)).

   Porter et al. (2017), who focused on τ_1^w and τ_2^w and did not aggregate them, recommend using the frontier method in most situations due to its strong performance and straightforward implementation in their simulation study. However, they did recommend comparing it to the binding score method if there is reason to believe that average treatment effects are similar along both segments of the boundary.

3. **The fuzzy IV method**: Like the frontier method, this method generates two separate one-dimensional RDD estimates, but it does so while using more sample data. By viewing Figure 1 as a one-dimensional RDD in terms of X_1 (and thus ignoring X_2), there is non-compliance to the right of c_1 because only some units receive the control condition while some are treated. This can therefore be viewed as a fuzzy one-dimensional RDD. Fuzzy RDDs can be

estimated by defining an instrumental variable (IV) for treatment assignment and targeting an average treatment effect among compliers at the boundary for X_1 and, separately, X_2 (Imbens & Lemieux, 2008). This method has been discussed in works such as Robinson (2011), Dee (2012), and Wong et al. (2013). Porter et al. (2017) found this method to more frequently exhibit bias and have higher variance than others, and we therefore do not investigate it.

4. **The parametric surface method**: This is the only method that views the two-dimensional RDD as, indeed, a two-dimensional RDD, instead of simplifying the problem into one-dimensional RDDs. Using a parametric approach, the method estimates separate response surfaces in the treatment and control groups by modeling the relationship between the two running variables and the outcome. This method is described in Reardon and Robinson (2012) and Wong et al. (2013). The treatment effect is then estimated as the average difference in outcomes for units on the treatment assignment boundary. A nonparametric approach would adaptively select a functional form as opposed to it being researcher-specified.

Although the surface method may seem to be the most appropriate method for analyzing a two-dimensional RDD, current implementations are not without their complications. The drawback most noted in the literature is that the surface method is very sensitive to model specification if conducted parametrically (Porter et al., 2017; Reardon & Robinson, 2012). Furthermore, it is unclear how to use a bandwidth to focus estimation on units near the cutoffs, as local linear regression does in one-dimensional RDDs. For example, one could select two bandwidths (h_1, h_2)—such as by cross-validation, as done in Papay et al. (2014), Papay et al. (2011)—and restrict analyses to the subset of data that falls within both of these bandwidths. However, this restricts analyses to data within a rectangular subset centered around the origin of Figure 1, which may discard many units that are very close to one of the borders ($\mathcal{B}_1$ or $\mathcal{B}_2$) but not both. More concerningly, (Wong et al., 2013) showed that the bandwidth is sensitive to arbitrary scaling decisions of the running variables. For these reasons, Porter et al. (2017) did not include the surface method in their simulation study.

Nonparametric surface response methods can avoid these concerns. We next explain the implementation of two such approaches: Gaussian process regression and loess regression.

4 Nonparametric Surface Response Approaches

Surface response methods, both parametric and nonparametric, support estimation of the two unknown functions in a two-dimensional RDD (the treatment response function and the control response function), yet nonparametric approaches are rare in applied educational RDD contexts. Nonparametric approaches offer advantages over parametric methods, such as flexibly model the treatment and control response surfaces, making model misspecification less of a concern as the researcher does not need to impose a structural form. In addition to this flexible fit, units near the boundary can be automatically upweighted, akin to using a bandwidth, and these weights can vary continuously across the two-dimensional domain recognizing the correlation between the running variables. Another unique benefit of nonparametric approaches is that they can be used to assess if there is treatment effect heterogeneity along the boundary.

4.1 Gaussian Process Regression for Two-Dimensional RDDs

To estimate the two unknown functions, we recommend the use of Gaussian process regression, because of its success in the machine learning and Bayesian modeling literature for estimating unknown functions (Rasmussen & Williams, 2006) as well as its success in one-dimensional RDDs (Branson et al., 2019) and geographic RDDs (Rischard et al., 2021).

Define $\mu_T(X_1, X_2) \equiv \mathbb{E}[Y_i(1)|X_1, X_2]$ and $\mu_C(X_1, X_2) \equiv \mathbb{E}[Y_i(0)|X_1, X_2]$ as the mean response functions for treatment and control, respectively. We will assume that the treatment and control response for each unit i are generated as

$$\begin{aligned} &Y_i(1) = \mu_T(X_{i1}, X_{i2}) + \varepsilon_{iT}, \text{ and } Y_i(0) = \mu_C(X_{i1}, X_{i2}) + \varepsilon_{iC}, \text{ where} \\ &\varepsilon_{iT} \overset{iid}{\sim} N(0, \sigma^2_{yT}), \text{ and } \varepsilon_{iC} \overset{iid}{\sim} N(0, \sigma^2_{yC}) \end{aligned} \tag{11}$$

Local linear regression methods make the same above assumption, but also specify a model for $\mu_T(X_1, X_2)$ and $\mu_T(X_1, X_2)$. Instead of specifying a functional form for $\mu_T(X_1, X_2)$ and $\mu_C(X_1, X_2)$,

we will place a Gaussian process prior on both of these functions:

$$\begin{aligned} \mu_T(X_1, X_2) &\sim \text{GP}(m_T(\mathbf{X}), K_T(\mathbf{X}, \mathbf{X}')) \\ \mu_C(X_1, X_2) &\sim \text{GP}(m_C(\mathbf{X}), K_C(\mathbf{X}, \mathbf{X}')) \end{aligned} \tag{12}$$

where we treat the two Gaussian process priors in (12) as independent.[1] The notation $\text{GP}(m(\mathbf{X}), K(\mathbf{X}, \mathbf{X}'))$ denotes a Gaussian process prior with mean function $m(\mathbf{X})$ and covariance function $K(\mathbf{X}, \mathbf{X}')$, where $\mathbf{X}$ is the $N \times 2$ matrix of running variables for all units. This states that the joint distribution of the unknown response function for the N units is a multivariate normal distribution with mean $(m(\mathbf{x}_1), \ldots, m(\mathbf{x}_N))$ and an $N \times N$ covariance matrix $K(\mathbf{X}, \mathbf{X}')$ whose (i, j) entries are $K(\mathbf{x}_i, \mathbf{x}'_j)$, where $\mathbf{x}_i \equiv (X_{i1}, X_{i2})^T$.

Following Branson et al. (2019) and Rischard et al. (2021), we will choose the linear mean functions $m_T(\mathbf{X}) = \alpha_T + \mathbf{X}\beta_T$ and $m_C(\mathbf{X}) = \alpha_C + \mathbf{X}\beta_C$ with standard multivariate Normal priors on (α_T, β_T) and (α_C, β_C). This does not force the estimated treatment and control responses to be linear in the running variables, but rather represents our prior information that test scores often have some association with the outcome of interest. Furthermore, we specify the covariance functions in treatment and control as the squared-exponential covariance function:

$$K_Z(\mathbf{X}, \mathbf{X}') = \sigma^2_{GP,Z} \exp\left(-\frac{(\mathbf{X} - \mathbf{X}')^T (\mathbf{X} - \mathbf{X}')}{2\ell_Z^2} \right) \text{for } Z = T, C. \tag{13}$$

This kernel is the most common covariance function in the Gaussian process literature, and it is often found to have good performance. See Rasmussen and Williams (2006, Chapter 4) for details about other covariance functions.

The squared exponential covariance function contains two parameters: the variance σ^2_{GP}, which affects the amplitude, and the lengthscale ℓ, which affects the smoothness. Following Rischard et al. (2021), we use maximum-likelihood to estimate these parameters. However, as shown in Branson et al. (2019), priors can also be placed on these parameters, thereby propagating their

uncertainty. As discussed in Branson et al. (2019), these covariance parameters play a role that is analogous to the bandwidth in local linear regression, in that they determine the weight of each unit in estimating the treatment effect. For example, if the lengthscale ℓ is large, then the prior says the response does not vary rapidly across values of the running variables; as a result, many units will play a role in estimating the response along the boundary. Likewise, if the lengthscale is small, then only units very close to the boundary will play a role in estimation. Meanwhile, the variance σ^2_{GP} determines how far departures from the mean function can be. Small values force the response surface to be relatively linear, and large values allow for departures from linearity. Following Branson et al. (2019) and Rischard et al. (2021), we will also make the assumption that $\sigma^2_{GP,T} = \sigma^2_{GP,C}$ and $\ell_T = \ell_C$; this is analogous to the common practice of letting the bandwidth in the treatment and control groups be the same when implementing local linear regression (Imbens & Lemieux, 2008). This assumption can be relaxed, though more data would be required for precise estimation of the larger number of parameters.

After the covariance parameters are estimated, the posterior distribution of $\mu_T(\mathbf{x})$ and $\mu_C(\mathbf{x})$ for any point $\mathbf{x} = (x_1, x_2)$ along the border $\mathcal{B}$ can be found via standard multivariate Normal theory. Thus, the posterior of the treatment effect $\tau(\mathbf{x}) = \mu_T(\mathbf{x}) - \mu_C(\mathbf{x})$ (defined in (3)) for any $\mathbf{x} \in \mathcal{B}$ can be readily obtained. In fact, a large benefit of this approach is that the joint distribution of any number of treatment effects along the boundary can be obtained as follows: consider R points $\mathbf{x}_1, \ldots, \mathbf{x}_R$ on the boundary where we would like to estimate the treatment effect, which we refer to as sentinel points. Then, under our Gaussian process model, the posterior joint distribution of $\tau(\mathbf{x}_{1:R})$, the vector of treatment effects to be estimated for $\mathbf{x}_1, \ldots, \mathbf{x}_R$, is

$$\begin{aligned} \tau(\mathbf{x}_{1:R}) | \mathbf{X}, \mathbf{y} &\sim \mathcal{N}_R\left(\mu_{\mathbf{x}_{1:R}|T} - \mu_{\mathbf{x}_{1:R}|C}, \Sigma_{\mathbf{x}_{1:R}|T} + \Sigma_{\mathbf{x}_{1:R}|C}\right) \\ &\equiv \mathcal{N}_R\left(\tau_{1:R|Y}, \Sigma_{\tau_{1:R}|Y}\right), \end{aligned} \tag{14}$$

where $\tau_{1:R|Y} \coloneqq \mu_{\mathbf{x}_{1:R}|T} - \mu_{\mathbf{x}_{1:R}|C}$, $\Sigma_{\tau_{1:R}|Y} \coloneqq \Sigma_{\mathbf{x}_{1:R}|T} + \Sigma_{\mathbf{x}_{1:R}|C}$, and, for $Z = T, C$,

$$\begin{aligned}
\mu_{\mathbf{x}_{1:R}|Z} &\equiv K_Z(\mathbf{x}_{1:R}, \mathbf{X}_Z)\left[K_Z(\mathbf{X}_Z, \mathbf{X}_Z) + \sigma^2_{yZ}\mathbf{I}\right]^{-1}\mathbf{y}_Z \\
\Sigma_{\mathbf{x}_{1:R}|Z} &\equiv K_Z(\mathbf{x}_{1:R}, \mathbf{x}_{1:R}) - K_Z(\mathbf{x}_{1:R}, \mathbf{X}_Z)\left[K_Z(\mathbf{X}_Z, \mathbf{X}_Z) + \sigma^2_{yZ}\mathbf{I}\right]^{-1}K_Z(\mathbf{X}_Z, \mathbf{x}_{1:R}).
\end{aligned} \tag{15}$$

To clarify notation of the above equations: $\mathbf{x}_{1:R}$ is an $R \times 2$ matrix (denoting the two dimensions for each of the R points), while $\mathbf{X}_T$ and $\mathbf{X}_C$ are $N_T \times 2$ and $N_C \times 2$ matrices, respectively, of the running variables of the treatment and control units. The vector $\tau_{1:R|Y}$ of estimates at the sentinel points is an R-vector. The $\Sigma_{\cdot|\cdot}$ covariance matrices are all $R \times R$.

There are three key properties of the posterior distribution shown in (14). First, the mean of this posterior distribution shows us that the average treatment effect at each $\mathbf{x}$ is estimated as simply the difference between the estimated treatment mean response and control mean response at that $\mathbf{x}$. Second, the covariance matrix in (14) shows us that the covariance of the treatment effects is the sum of the treatment and control covariances, because the two response surfaces are treated as *a priori* independent.. Third, the definitions of the treatment and control covariance matrices $\Sigma_{\mathbf{x}_{1:R}|T}$ and $\Sigma_{\mathbf{x}_{1:R}|C}$ (15) tell us that the covariance between any two $\tau(\mathbf{x})$ and $\tau(\mathbf{x}')$ depends on the proximity of $\mathbf{x}$ and $\mathbf{x}'$, where proximity is measured by $K_T(\mathbf{X}, \mathbf{X}')$ and $K_C(\mathbf{X}, \mathbf{X}')$ in (13).

The above formulation models the treatment effect as varying smoothly along the boundary. Estimates for the individual $\tau(\mathbf{x})$ can then be combined together with weights to conduct inference on the treatment effect for sections of the boundary, such as $\mathcal{B}_1$, $\mathcal{B}_2$, or even the entire boundary $\mathcal{B}$. As we show later in Section 6, it is useful to plot the point estimates of the average treatment effects for $(\tau(\mathbf{x}_1), \ldots, \tau(\mathbf{x}_R))$ and their corresponding confidence bands across the R sentinel points. Because we have modeled the *joint* distribution of $(\tau(\mathbf{x}_1), \ldots, \tau(\mathbf{x}_R))$, these confidence bands account for the covariance among these effects. This allows researchers to graphically diagnose the degree of treatment effect heterogeneity along the boundary, which—to our knowledge—cannot be done with other estimation approaches. One can also test if there is treatment effect heterogeneity across the boundary to a statistically significant degree; see Rischard et al. (2021), Appendix B.

4.1.1 Inference for Treatment Effects along the Boundary

Equation (14) provides a posterior for the joint distribution of $\tau(\mathbf{x}_{1:R})$ for any $\mathbf{x}_1, \ldots, \mathbf{x}_R \in \mathcal{B}$. To estimate τ^w for any set of weights w, we first take a weighted average of these $\tau(\mathbf{x}_{1:R})$. Technically, τ^w integrates the treatment effect across every infinitesimal point along the boundary, and we are approximating this integral with our sentinels. We can approximate more closely with more sentinels, if we choose. In particular, we estimate the treatment effect only for the finite number of sentinel points $\mathbf{x}_1, \ldots, \mathbf{x}_R$ by predicting outcomes from both the treatment and control surfaces at the sentinels. Assuming that these points are evenly spaced along $\mathcal{B}$, it is reasonable to approximate the causal estimand τ^w as

$$\tau^w = \frac{\int_{\mathbf{x}\in\mathcal{B}} w(\mathbf{x})\tau(\mathbf{x})\partial\mathbf{x}}{\int_{\mathbf{x}\in\mathcal{B}} w(\mathbf{x})\partial\mathbf{x}} \approx \frac{\sum_{r=1}^{R} w(\mathbf{x})\tau(\mathbf{x})}{\sum_{r=1}^{R} w(\mathbf{x})} \tag{16}$$

The above is a linear combination of the vector $\tau(\mathbf{x}_{1:R})$, whose posterior (14) is Normally distributed. Under our Gaussian process model, the posterior of τ^w is also Normally distributed:

$$\tau^w|\mathbf{X},\mathbf{y} \sim N\left(\frac{w(\mathbf{x}_{1:R})^T(\mu_{\mathbf{x}_{1:R}|T} - \mu_{\mathbf{x}_{1:R}|C})}{w(\mathbf{x}_{1:R})^T\mathbf{1}_R}, \frac{w(\mathbf{x}_{1:R})^T(\Sigma_{\mathbf{x}_{1:R}|T} + \Sigma_{\mathbf{x}_{1:R}|C})w(\mathbf{x}_{1:R})}{[w(\mathbf{x}_{1:R})^T\mathbf{1}_R]^2}\right) \tag{17}$$

where $w(\mathbf{x}_{1:R}) \in \mathbb{R}^R$ is a vector of weights, and the other parameters in (17) are already defined in (15). The posterior distribution of τ^w (17) can then be used to obtain a point estimate and credible interval for the overall $BATE$ defined by weighting w; we discuss this, along with weight selection, in Section 4.1.2. We compute $\hat{\tau}^w$ with a plug in estimator, plugging the posterior means $\hat{\mu}_{\mathbf{x}_{1:R}|T}$ and $\hat{\mu}_{\mathbf{x}_{1:R}|C}$ and estimated density weights $w(\mathbf{x}_{1:R})$ into (16).

4.1.2 Implementation

We use the *laGP* package to conduct Gaussian process regression for two-dimensional RDDs (Gramacy, 2016). Specifically, we use the functions *newGPsep*, *mleGPsep*, and *predGPsep*. *newGPsep* generates empirical Bayes priors and creates an initial Gaussian process, whose hyperparameters

are updated using *mleGPsep*. Then, *predGPsep* uses the Gaussian process model to make a prediction at each $\mathbf{x}_1, \ldots, \mathbf{x}_R$ sentinel.

We use 20 sentinels along each section of the treatment boundary depicted in Figure 1. The sentinels were evenly spaced between the cutoff along each running variable to the maximum running variable value. Using fixed sentinel points is an alternative choice for researchers.

We generate predictions in two ways: first, we use all the data and predict at the sentinels. Second, and this is the process we follow for results in Section 5, we set the mean of the priors in (12) as a linear model and fit a single Gaussian process before predicting at the sentinels. We implement this second approach by regressing the outcome Y on the two running variables and then treating the residuals of that regression as the outcome to predict at the sentinels. This two-step process of residualizing simplifies the analytic computation compared to estimating a full Bayes model with a built-in linear model; this is the method used by Rischard et al. (2021). Once the predictions are generated at each sentinel along the boundary on both the treated and control surfaces, the difference between a predicted treatment-side sentinel outcome, $\mu_T(\mathbf{x})$, and the predicted control-side sentinel outcome, $\mu_C(\mathbf{x})$, is the estimate of the treatment impact at that sentinel $\tau(\mathbf{x})$.

Choice of Weights Gaussian process estimates of the $BATE$ also rely on how the sentinels are weighted, as our outcome of interest is a weighted average of these sentinel-level estimates. We offer two ways of weighting the sentinels.

Our first approach is to weight the sentinels by the density of observations at that sentinel. This involves estimating the density from the distribution of the running variables. For our context, we fit a multivariate normal distribution to the running variables in our data and then calculate density weights given this distribution. Other means of estimating the density are possible; we take this approach because test score data often has a normal shape and fitting a parametric density model stabilizes the density estimation considerably.

As an alternative, we can generate a precision-weighted $BATE$, τ^{INV}, using the same sentinel-level effect estimates but weighted by the inverse of the sentinels' posterior variances, taking into

account their covariance. These (normalized) precision weights are

$$\vec{w} = (\mathbf{1}_R^T \Sigma^{-1}_{\tau_{1:R}|Y} \mathbf{1}_R)^{-1} \cdot \mathbf{1}_R^T \Sigma^{-1}_{\tau_{1:R}|Y}, \tag{18}$$

where $\Sigma^{-1}_{\tau_{1:R}|Y}$ is the covariance matrix of the posterior of the sentinels estimates (see (14)).

This precision weighting approach accounts for the correlation structure of sentinel estimates, such that the information gained from the sentinel estimates is maximized. In other words, this weighting approach targets an estimand, τ^{INV}, that can be most precisely estimated among all τ^w. However, τ^{INV} is arguably less interpretable than τ^{pop}, because it is a data-dependent estimand.

Because a greater density of samples often leads to more precision, precision weighting also tends to upweight areas of the boundary with high density, but while also accounting for the correlation structure of sentinel estimates. As a result, areas of the boundary with a higher density of sentinel points tend to not be "over counted," compared to the density weighting approach.

Using either approach, in our computation we chose to drop low-weight sentinels because using those sentinels would have extrapolated too far outside of the support of the distribution. Dropping these sentinels computationally adds a slight but negligible amount of bias to our $BATE$ estimate because of the small weight of data around these points.

Standard Errors We calculate the standard error of the $\widehat{BATE}$ using (17): we take the square root of the posterior variance as the standard error. In shorthand, we have, for weights $w(x_{1:R})$,

$$SE(\tau^w) = \left(\frac{1}{W^2} w(x_{1:R})^T \Sigma_{\tau_{1:R}|Y} w(x_{1:R}) \right)^{-1/2}, \tag{19}$$

with normalizing constant $W = w(x_{1:R})^T \mathbf{1}_R$.

For precision weights, our formula simplifies as the precision weights cancel with the covariance of our estimates. After some algebra, we produce standard errors for τ^{INV} of the form

$$SE(\tau^{INV}) = \left(\mathbf{1}_R^T \Sigma^{-1}_{\tau_{1:R}|Y} \mathbf{1}_R \right)^{-1/2}. \tag{20}$$

See Rischard et al. (2021), Section 2.3.3, for more details on the precision weighting approach.

4.2 Locally Weighted Regression

Gaussian process regression, at its root, is giving a flexible model to a two-dimensional function. Other choices are possible such as using local averaging, or loess (Cleveland & Devlin, 1986; Jacoby, 2000). This is a nonparametric technique that uses local weighted regression to fit a smooth curve through points. Here we present a straightforward way to use loess for two-dimensional RDDs, that, to our knowledge, is a novel application of this procedure.

The traditional one-dimensional loess can be extended to two dimensions: For each point $X = (x_1, x_2)$, we weight the data within a user-specified radius using $(1 - \sqrt[3]{(d)})^3$ where d represents the distance between the data point and (x_1, x_2). Using these weights, we fit a linear model with an interaction term to the treated and control units separately:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \varepsilon_i \tag{21}$$

to generate predictions at each (x_1, x_2) point. We then apply this smoothing function to each point on the boundary. We calculate the difference in predictions at the sentinels as the estimated treatment effect at each point and combine them as a weighted average to form the estimated $BATE$.

In our implementation, we use a radius of 0.5 around each sentinel, which is half the standard deviation of our simulated data, and require a minimum sample size of 8 data points within each radius to fit a model. This approach readily gives us point estimates, but generating standard errors and thus confidence intervals is more challenging and would require a bootstrap procedure.

5 Simulations

This paper uses simulations to compare the performance of the binding score, pooled frontier, and loess modeling methods discussed in Sections 3 and 4.2 to that of Gaussian process regression,

particularly in cases of heterogeneous treatment effects along the boundary.

We use the simulation data generating process of Porter et al. (2017), who compared binding score, non-pooled frontier, and fuzzy IV approaches. We were interested in seeing performance in a context we did not choose, to limit researcher degrees of freedom. As mentioned in Section 3, both Gaussian process regression and loess can be thought of as a "surface method" approach—an approach that Porter et al. (2017) discussed but did not include in their simulation study.

While we build upon the valuable insights from Porter et al. (2017)'s simulation study, our study differs in several ways. First, we do not use the fuzzy IV method due to its underperformance in Porter et al. (2017) and in Wong et al. (2013). Second, Porter et al. (2017) did not compare binding score's performance in scenarios where the treatment effect was unequal between the two segments of the boundary. We include binding score for comparison even in those cases because, in practice analyzing some treatment, we would not know that the treatment effect is unequal in different segments of the boundary. Third, we use pooled frontier estimates (Wong et al., 2013), as opposed to the individual frontier estimates described in Porter et al. (2017). Fourth, we include the loess curve modeling approach, which, as a flexible surface response model, serves as a more similar competitor to Gaussian process regression. Finally, though we use the same data generating models as found in Porter et al. (2017), the way we implemented each method is different. Porter et al. (2017) assumes the true model is known a priori and pass it to the regression discontinuity estimator. By contrast, we always fit the data using the same model for these methods which, except for the nonparametric surface response approaches, is a local linear regression on data within a given bandwidth. Thus, estimators' bias in our simulation study are higher than that in Porter et al. (2017), reflecting the deteriorated performance we would expect from estimators when we do not know the true data-generating process.

5.1 Setup and Data Generating Processes

Using the data setup of Porter et al. (2017), we generate running variables from a bivariate normal distribution with a mean of 0 and marginal standard deviations of 1, and outcomes from four

different data generating processes.

Consider N simulated data points with running variables X_1 and X_2, treatment W, and outcome Y. Each running variable is generated as

$$\begin{pmatrix} X_{1i} \\ X_{2i} \end{pmatrix} \overset{iid}{\sim} \mathcal{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right), \tag{22}$$

and we vary the correlation parameter $\rho \in \{0.3, 0.5, 0.9, 0.95, 0.99\}$. Treatment is defined as

$$W_i = \begin{cases} 1 & \text{if } X_{1i} \geq c_1 \text{ and } X_{2i} \geq c_2 \\ 0 & \text{otherwise.} \end{cases} \tag{23}$$

for cutoffs c_1 and c_2. We use three combinations of locations for the cutoff parameters: the 50th percentile of the $\mathcal{N}(0,1)$ distribution for both running variables, the 30th percentile for both, and the 30th percentile for one running variable and the 70th percentile for the other.

In addition to matching some of Porter et al. (2017)'s simulation parameters, we also explore parameter setups that are a closer match to empirical applications in education research. This involves using high running variable correlation parameters $\rho \in \{0.95, 0.99\}$ and a smaller sample size of $n = 1{,}000$ in addition to Porter et al. (2017)'s sample size of $n = 5{,}000$.

We show results using $n = 1{,}000$ here and include results using $n = 5{,}000$, which were similar, in the Appendix Section 9.

Finally, the outcomes are generated as $Y_i = W_i + \mu(X_{i1}, X_{i2}) + \tau(X_{i1}, X_{i2}) + \varepsilon_i$, with $\mu(\cdot,\cdot)$ and $\tau(\cdot,\cdot)$ under four different settings as shown in Table 1, following Porter et al. (2017) with residuals $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0,1)$. In Models 1 and 2, the treatment effect is 0.4 everywhere along the boundary. In Models 3 and 4, the treatment effect is heterogeneous along the boundary, such that the choice of weights to estimate τ^w in Section 4.1 is consequential. Porter et al. (2017) focused on population-density weights for these models (see Appendix A of their supplementary material). For each data generating model and parameter setup, we generate 500 simulated data sets. We apply each method

Model Control Side Outcome $\mu_C(X_1, X_2)$	Treatment Effect $\tau(X_1, X_2)$
1) $Y = 0.4 + 0.5X_1 + X_2$	$\tau = 0.4$
2) $Y = 0.4 + 0.5X_1 + X_2 + 2X_1^2 + X_2^2$	$\tau = 0.4$
3) $Y = 0.4 + 0.5X_1 + X_2 + 2X_1^2 + X_2^2$	$\tau = 0.4 - 0.1X_1 - 0.2X_2$
4) $Y = 0.4 + 0.5X_1 + X_2 + 2X_1^2 + X_2^2$	$\tau = 0.4 - 0.1X_1 - 0.2X_2$ $-0.08X_1^2 - 0.08X_2^2$

Table 1: Simulation Outcome Generating Models. For all models, $Y_i(1) = Y_i(0) + \tau_i$.

to estimate the *BATE* and standard error for each simulated data set. A full list of parameters and *BATE* values for each data generating model is listed in Appendix Section 9.2. Our predictions, and correspondingly our treatment effect estimates, are scaled by the standard deviation of the control unit outcomes such that our results are in effect size units.

5.2 Implementation

We use the R package *rddapp* (Jin et al., 2021) to implement the binding score and pooled frontier methods (see Appendix Section 9.1). We specify sharp two-dimensional RDDs and include the running variables as covariates, which showed better precision in Porter et al. (2017). Gaussian process regression and loess are implemented as described in Sections 4.1.2 and 4.2.

5.3 Results

We compare the performance of each method on its average absolute bias, precision, RMSE, coverage, and standard error estimation across data generating models. We show results averaged over the three sets of running variable cutoff percentiles for the $n = 1{,}000$ case as results are similar across the three choices. We do not see substantial gains in using precision-weighted Gaussian process regression and therefore show results here relative to density-weighted Gaussian process regression. Results showing precision-weighted Gaussian process regression, precision-weighted residualized Gaussian process regression, and precision-weighted loess are in the Appendix 9.3.

In Figure 2, the columns correspond to the numbered models in Table 1. The models in the first two columns have a constant treatment effect, while those in the third and fourth columns

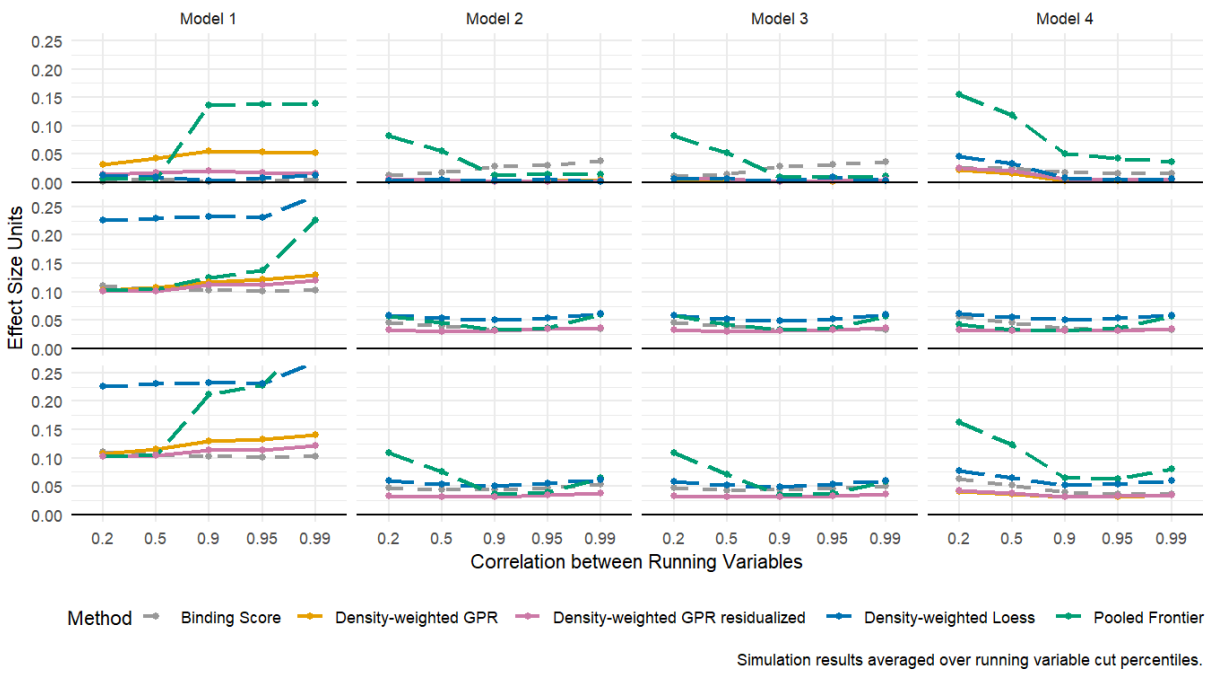

Figure 2: Simulation results averaged over running variable cutoff percentiles of 30/30, 50/50, and 30/70.

introduce heterogeneous treatment effects along the boundary. Additionally, models in columns two and four are more complex (they have curvature and interactions in the running variables) than the models in columns one and three. The rows of Figure 2 show our outcomes. In the first row, we compare the methods in terms of their absolute bias across simulations and simulation parameters. The second row shows their precision, and the third shows the RMSE.

From Figure 2, we see that Gaussian process regression does slightly better than other the methods in terms of absolute bias, particularly as the data generating models become more complex across the first row. Additionally, the difference between residualized and non-residualized Gaussian process regression's estimated bias in the first column shrinks in these three later columns. Looking across the middle row of the figures, density-weighted Gaussian process regression results are similarly precise between residualized and non-residualized regression. Density-weighted loess is always less precise than the other methods, but beyond this distinction, differences between methods are small, with Gaussian process regression results almost always matching or exhibiting smaller standard errors than the remaining methods. There is one exception in Model 1, where

binding score slightly outperforms Gaussian process regression under high running variable correlations. Finally, as shown in the third row of Figure 2, Gaussian process regression does well in terms of RMSE, with residualized and non-residualized Gaussian process regression performing similarly throughout. RMSE gives overall performance, taking any bias-variance trade-offs into account. Gaussian process regression ends up outperforming the other methods in the three more complex models in terms of RMSE as it balances these trade-offs most strongly. Like with the simulation's standard errors, Gaussian process regression's RMSE always matches or outperforms the other methods, except for the high correlation case in the least complex model, Model 1, where binding score slightly outperforms.

Figure 3 shows coverage results of the $BATE$ in the top panel using a 95% confidence interval averaged over the three running variable cutoff percentiles as results looked similar across them. The bottom panel of this figure shows how well the methods estimated standard errors. Note that we did not estimate standard errors for loess, which would require a bootstrap procedure. All methods except pooled frontier perform well for the simplest data setup, the first column. Gaussian process regression continues to do well, and outperforms the other methods, across the other three data generating models in terms of both coverage and estimation of standard errors. Results are fairly similar between residualized and non-residualized density-weighted Gaussian process regression.

In summary, residualized density-weighted Gaussian process regression performs slightly better than the non-surface response methods across our data generating models and simulated parameters in terms of estimated absolute bias, precision, and RMSE. Among nonparametric surface response approaches, Gaussian process regression is much more precise than density-weighted loess, particularly for data from the least complex model. Compared to binding score and pooled frontier, Gaussian process regression shows stronger coverage of the $BATE$ and more consistent recovery of the true standard errors. Taken together, Gaussian process regression exhibits the overall best performance of these methods in this simulation.

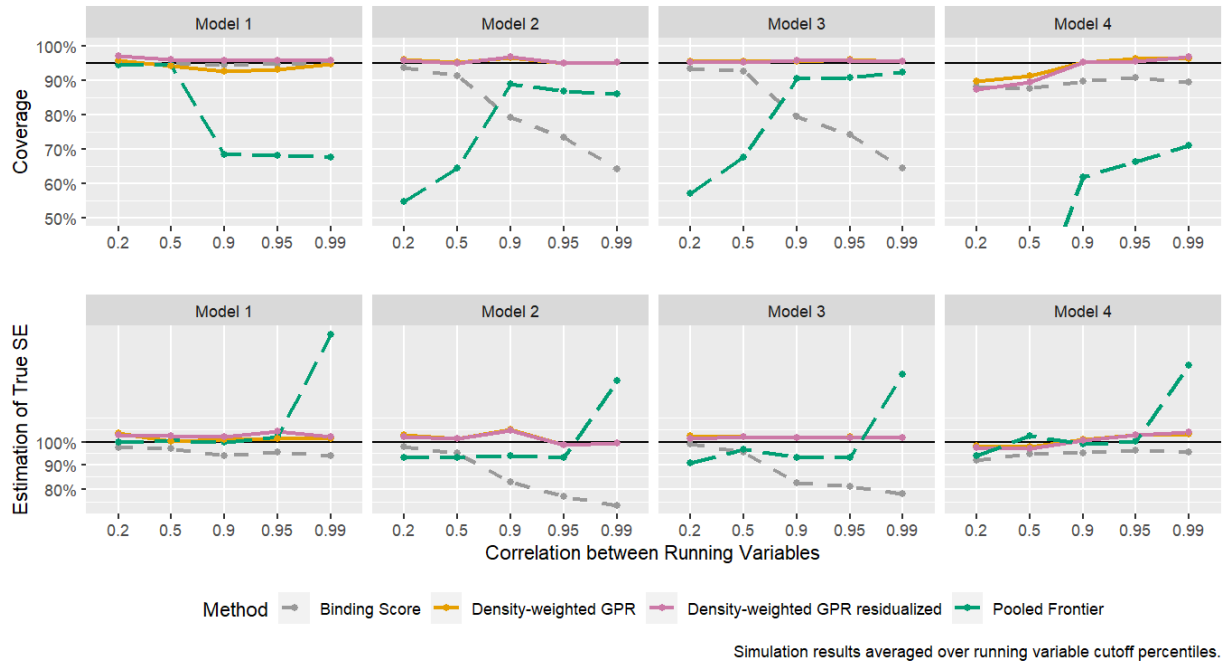

Figure 3: Simulation coverage and SE estimation across parameters.

6 Application: Wisconsin English Language Learner Reclassification

To understand the methods' comparative performance with real world data, we use Gaussian process regression, binding score, pooled frontier, and loess to analyze an empirical two-dimensional RDD. We describe the data and policy context, define the estimand of interest, and show the result of analyzing the two-dimensional RDD using the different methods. Since we do not know the true $BATE$, we compare the methods in terms of their precision. We end with a demonstration of the capacity of Gaussian process regression to investigate heterogeneous treatment effects.

6.1 Policy Context

The two-dimensional RDD we use centers on the assignment of English Language Learner (ELL) students to non-ELL status based on their test scores in public high schools in Wisconsin. This change is known as reclassification, which indicates that these students are now considered to be

fully English proficient. After reclassification, they typically join mainstream classes with native English speakers and no longer receive English language instruction or linguistic accommodations.

In Wisconsin, ELL students in grades K-12 are tested annually in the winter on the ACCESS examinations, which measure English proficiency. ELL students receive ACCESS scores in multiple English language domains: Speaking, Listening, Reading, and Writing. Literacy is calculated as $Literacy = 50\%Reading + 50\%Writing$. Students also receive an Overall ACCESS score, which is a weighted average of the four domains: $Overall = 35\%Reading + 35\%Writing + 15\%Listening + 15\%Speaking$ (Wisconsin Department of Public Instruction, 2023). ACCESS scores are provided in terms of scale scores as well as proficiency levels, which range from 1.0 (lowest) to 6.0 (highest). Each proficiency level score corresponds to the bottom of a range of scale scores. For example, a Literacy proficiency level of 5.0 requires a minimum scale score of 406 in the data we use.

From 2011-2016, Wisconsin state policy was to reclassify ELL students as non-ELL if they:

a) received a score of 6.0 on their Overall proficiency level score, or

b) received a score of 5.0 on their Overall proficiency level score AND a score of at least 5.0 on their Literacy proficiency level score.

We examine the two-dimensional case of b), where students were required to pass two thresholds to be eligible to receive the treatment of reclassification as a non-ELL. Even though there are technically two policy pathways to be labeled as a non-ELL, the rule of reclassification given at least a 5.0 on Literacy and a 5.0 on Overall fully determines who is reclassified and who is not, since anyone who scores a 6.0 on Overall also scores at least a 5.0 on Literacy. This means that we can conduct a two-dimensional RDD along the boundary defined by Literacy ≥ 5.0 and Overall ≥ 5.0 and employ the methods of this paper. Our research question is whether being eligible for mainstreaming into non-ELL classrooms helped students in terms of their academic outcomes, which in this case are their ACT test score results in the year after reclassification.

Similar data were used by Carlson and Knowles (2016) to examine the impacts of reclassification for ELL students in Wisconsin on their ACT achievement, likelihood of high school graduation, and post secondary enrollment following high school graduation, using the first rule (a) as

the discontinuity in a fuzzy one-dimensional RDD. In Wisconsin, students typically take the ACT in the spring of 11th grade, meaning that the ELL reclassification we and Carlson and Knowles study occurred one year prior to this outcome measure. Using both linear and quadratic functional form specifications above and below the singular 6.0 cutoff on the Overall proficiency level score, Carlson and Knowles found that reclassification in 10th grade had a positive effect on students' Composite ACT scores, mainly driven by increases in English and Reading ACT scores.

Although reclassification under these policies in Wisconsin was intended to be automatic, there were students who passed these ACCESS test score thresholds but who were not classified as non-ELL students. This is partly due to the state's manual reclassification policies, where automatically reclassified students could be manually classified back as ELL students if their districts deemed them not to be fully proficient in English. Manual reclassification was also possible in the other direction for students who scored a 5.0 on their Overall ACCESS score but less than a 5.0 on their Literacy ACCESS score, if the district determined that the students were in fact proficient in English. These counts of students were "relatively few" in either direction (Carlson & Knowles, 2016, p. 562). A more likely explanation for discrepancies between passing ACCESS scores and reclassification status is that the automatic reclassification policy was a Wisconsin Department of Public Instruction recommendation without a strong forcing mechanism at the time. As such, there may have been delays between state-level automatic reclassification and district-level enacted reclassification of these students into non-ELL classrooms.

These discrepancies motivate our examination of intent-to-treat (ITT) effects of becoming eligible for reclassification after passing both score thresholds, as opposed to effects of reclassification itself, for this analysis. We believe that analyzing this sharp RDD is, in fact, preferable to a fuzzy RDD, as it involves more precise inference. Furthermore, analyzing the two-dimensional nature of the policy allows us to assess potential treatment effect heterogeneity along the boundary. This gives us a more detailed perspective about which particular students are affected by ELL reclassification. Utilizing sharp two-dimensional RDD methods therefore provides a more appropriate analysis of this data.

Average scores	All observations	Below threshold	Above threshold
ACT Composite	15.0	13.9	16.5
ACT Math	16.2	15.4	17.4
ACT Reading	14.6	13.5	16.2
ACT English	12.7	11.5	14.5
ACT Science	15.9	14.9	17.4
N	3,489	2,038	1,451

Table 2: ACT score descriptives for ELL students

6.2 Data

Our analytic sample consists of all 10th grade Wisconsin ELL students who were administered the ACCESS test between academic years 2010-11 through 2015-16 and who also took the ACT over that time, similarly to Carlson and Knowles (2016). This corresponds to 3,489 students.

In our data, 44% of students scored a 5.0 or above on Literacy, 52% scored a 5.0 or above on Overall, and 39% scored a 5.0 or above on both, meaning they passed the reclassification threshold. There were 2,038 students who did not meet the reclassification threshold in total. Of those students who did not pass, 3.9% scored a 5.0 or above on Literacy but not Overall (these were Overall scores in the 4.0-4.9 range), and 17.1% scored a 5.0 or above on Overall, but not on Literacy.

Table 2 presents statistics about these students' ACT scores, which can range from 0-36. On average, ACT scores were higher for students whose ACCESS scores fell above the reclassification threshold than for students whose scores fell below.

The test score data are displayed in Figure 1 in Section 2, with student ACCESS Literacy scores along the x-axis and ACCESS Overall scores along the y-axis. The estimated correlation between the two running variables is high, 0.96. Although the reclassification policy's two passing rules are defined using ACCESS proficiency levels from 1.0 to 6.0, we use the proficiency levels' corresponding scale scores to run our analysis following Carlson and Knowles (2016) and centered the running variables around their cutoffs. There is one scale score point that divides students who were eligible for treatment (the darker blue points in the top right of Figure 1) from those who were not eligible for treatment (the lighter blue points).

6.3 Estimand and Methods

We conduct inference on the τ^w defined in Section 4.1 to estimate the $BATE$. This corresponds to estimating the intent to treat effect of just barely passing the reclassification threshold boundary in 10th grade on ACT achievement. Our estimand is still the population estimand, and we examine results using both population density weights as well as precision weights. We consider all ACT skill area scores (English, Reading, Math, and Science) as well as students' Composite ACT scores.

We compare the results of using Gaussian process regression, loess, binding score, and pooled frontier methods. We implemented these methods the same way we had in the simulation study.

6.4 Results and Comparison

The results from all four methods are presented in Table 3. Each column represents one type of ACT test outcome estimate with standard errors in parentheses.[2]

The only statistically significant results are for ACT English scores using residualized density-weighted Gaussian process regression, as well as for ACT Math scores using binding score. These results are not consistently significant within outcome, across methods, though they are both negative. Directionally, Gaussian process regression, binding score, and precision-weighted loess results are almost all negative across ACT test results. Pooled frontier and density-weighted loess results are generally positive, though these are never significant.

When comparing the standard errors of the estimates across methods, precision-weighted Gaussian process regression estimates slightly outperform density-weighted Gaussian process regression, though they are similar. Meanwhile, pooled frontier and loess's standard errors are larger than Gaussian process regression results, while binding score results are slightly more precise. We conjecture that this performance is related to the 0.96 correlation between the running variables. They are so highly correlated because a student's Literacy score makes up part of their Overall score. These results in terms of precision are in line with the lower standard errors for binding score results and higher standard errors for pooled frontier and loess that we saw in the simulation study under high running variable correlations, see Figure 2.

ACT Score	Composite	English	Reading	Math	Science
Residualized GPR - density weights	-0.276	-0.471*	0.039	-0.228	-0.281
(standard error)	(0.213)	(0.265)	(0.264)	(0.249)	(0.272)
Residualized GPR - precision weights	-0.292	-0.393	0.096	-0.371	-0.277
(standard error)	(0.201)	(0.246)	(0.249)	(0.232)	(0.259)
Binding Score	-0.304	-0.319	0.102	-0.398*	-0.373
(standard error)	(0.187)	(0.231)	(0.245)	(0.213)	(0.250)
Pooled Frontier	0.094	0.140	0.205	-0.175	0.193
(standard error)	(0.238)	(0.315)	(0.311)	(0.259)	(0.386)
Loess - density weights	0.272	-0.138	0.179	0.061	0.785
(standard error)	(0.406)	(0.734)	(0.739)	(0.448)	(0.654)
Loess - precision weights	-0.102	-0.315	-0.024	-0.142	0.098
(standard error)	(0.475)	(0.644)	(0.420)	(0.342)	(0.578)

Table 3: ITT estimates using two-dimensional RDD methods. *$p < 0.1$

In Carlson and Knowles (2016)'s analysis of the fuzzy one-dimensional ELL reclassification RDD, they found estimated positive effects of reclassification that were 1 ACT Composite score point and 1.2-1.7 ACT English and Reading score points in magnitude. We, in contrast, found a significant but negative intent-to-treat effect of assignment to reclassification for ACT English scores when analyzing the two-dimensional case using Gaussian process regression, though note that this effect is one third the magnitude of Carlson and Knowles (2016). The difference in these two papers' results, as well as the differences among methods in this application, likely relate to differences in the methods' approaches to estimating treatment effects and hints at the existence of treatment effect heterogeneity. Namely, the binding score method assumes that the treatment effects of each segment of the boundary are equal, while the pooled frontier method combines the weighted average of each segment's treatment effects additively. Gaussian process regression, by contrast, would not double count units close to the intersection of the segments, nor would it require both segments to exhibit the same treatment effect. Across the two reclassification analyses, we may be seeing more non-statistically significant effects than Carlson and Knowles (2016) because of the existence of treatment effect heterogeneity, which we elaborate on further below.

Importantly, using Gaussian process regression, we are able to visually investigate evidence of any heterogeneous treatment effects. Conducting this analysis allows us to understand if treatment

functions differently for students with different values of running variables.

We show the ACT English test score outcome as an example here as it was found to be significant using Gaussian process regression. Figures for other ACT test outcomes are in the Appendix in Section 9.6. In some cases, $\widehat{BATEs}$ that are null overall can exhibit statistically significant effects at specific sentinels along the boundary, such as when sentinel-level results are both negative and positive and cancel out. This is more likely to happen when we estimate the $BATE$ along a whole boundary in two dimensions as opposed to a local average treatment effect at one point.

In Figure 4, the left panel shows the boundary defined by the sentinels that are numbered black dots from 1 to 39. Sentinel 20 is at the intersection of the two running variable cutoffs. The right panel shows the Gaussian process regression estimates (the light blue bars) using the ACT English score outcome. The accompanying orange whiskers represent +/- 1.65 standard errors on each estimate. As expected given the distribution of data in the left panel, the standard error bars widen for sentinels further away from sentinel 20. Note that the figure's right hand panel only shows estimates for sentinels 12-25, meaning that sentinels 1-11 and 26-39 had too little data to estimate treatment effects and were therefore dropped as explained in Section 4.1.2 .

We provide the following interpretation of Figure 4 as an example of how one could use a figure like this, with sentinel-level estimates from Gaussian process regression, to investigate heterogeneous treatment effects along the boundary. Looking at the right hand panel, we see that the magnitude and direction of the estimates vary over the sentinels, beginning with positive estimates for sentinels 12-15 and then showing negative estimates for the remaining sentinels. The standard error bars show that the estimated treatment effects were significantly negative only for sentinels 18 through 21, and not significant otherwise. This provides evidence that the ELL reclassification policy functioned differently in terms of students' English ACT scores for students who scored close to the ACCESS Literacy cutoff (those who just barely passed the Literacy cutoff, with an Overall score well above the cutoff along sentinels 12-15) or to the Overall cutoff with relatively high Literacy scores (along sentinels 22-25), compared to students who scored close to both cutoffs (along sentinels 18-21). Those students who scored closer to the intersection of the two running

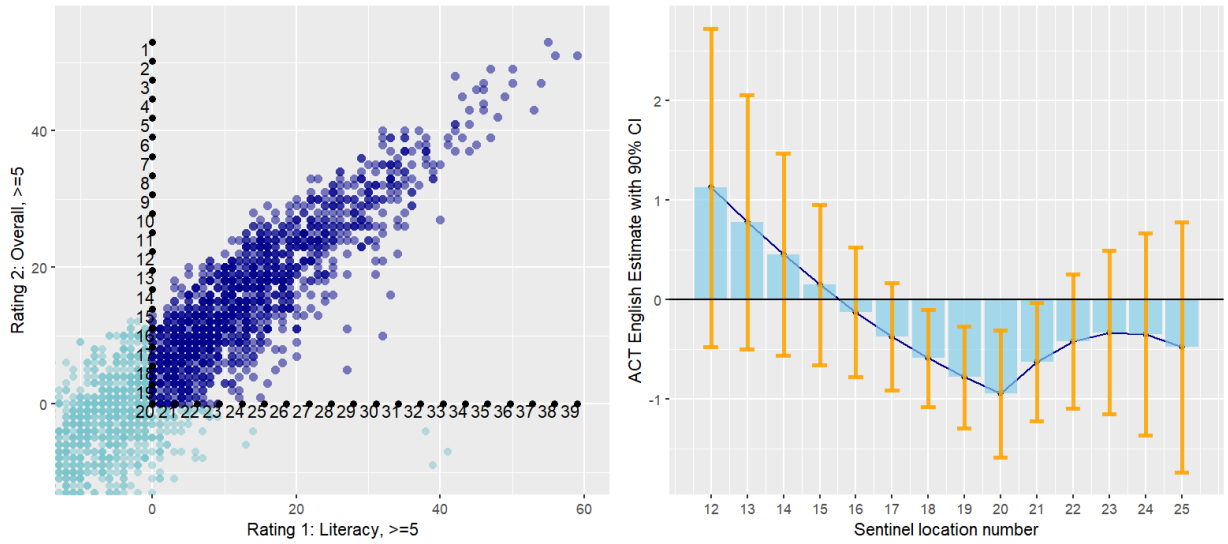

Figure 4: Treatment effect heterogeneity along the boundary.

variable cutoffs experienced negative, statistically significant treatment effects of assignment to reclassification on their ACT English scores. This implies that the reclassification policy may be less helpful for students with this pattern of running variable scores that are both close to their cutoffs. In contrast, students who scored away from this intersection did not experience significant effects of assignment to reclassification.

7 Discussion

Two-dimensional RDDs, such as requiring students to pass both an English and a math test score to receive a high school diploma, are fairly common in education research. Most existing methods to analyze these contexts collapse the two-dimensional RDD into a one-dimensional RDD, which may be inappropriate if there is treatment effect heterogeneity across the two-dimensional boundary. This article compares the performance of state-of-the-art two-dimensional RDD methods with that of previously unused nonparametric surface response approaches in education contexts, and provides code to implement them to estimate the boundary average treatment effect or $BATE$. These nonparametric approaches are Gaussian process regression, a Bayesian surface response method, and locally weighted regression (loess). The $BATE$ estimates a weighted average of the

treatment effect estimates at points along the boundary.

A key difference between surface response methods and the other methods is that the binding score and pooled frontier methods project the discontinuity to a one-dimensional space, which we argue is unduly limiting. By reducing the scope of the problem to one dimension, they lose information in order to use local linear regression. Meanwhile, our implementation of Gaussian process regression and loess analyzes treatment effects along the original two-dimensional boundary. Not only does this more accurately incorporate the two-dimensional nature of the data, but also it allows one to assess treatment effect heterogeneity along the two-dimensional boundary. This is particularly important when researchers suspect that the same treatment may impact subjects differently, depending on how they qualify for treatment.

Another state-of-the-art method that was not used by Porter et al. (2017), parametric surface response, may be a better comparison to Gaussian process regression, as it allows the researcher to jointly model the treatment and control response surfaces and estimate the $BATE$ as the difference. However, this method is difficult to use in practice given its sensitivity to the researcher's choice of functional form (Porter et al., 2017). Using nonparametric surface response methods reduces such concerns for model misspecification.

We target the $BATE$ in comparing the statistical properties of Gaussian process regression and loess to existing methods in simulated as well as empirical data. Our simulation approach extends the work of Wong et al. (2013) and Porter et al. (2017), with some key differences: We include additional high correlation contexts that mimic our empirical example, and we explore how these methods perform when the data generating models are unknown and potentially misspecified. We believe that this generates fairer comparisons between methods as well as to non-simulated data contexts where the researcher cannot predict the existence of heterogeneous treatment effects or nonlinear data in advance.

In simulations using similar parameter and data generating setups as Porter et al. (2017), we find that density-weighted Gaussian process regression performed slightly better in terms of absolute bias, RMSE, and precision compared to binding score, pooled frontier, and loess methods.

Though binding score exhibited slightly more precision in the least complex data scenario for high running variable correlations, Gaussian process regression showed the best consistent performance across all other running variable correlations, data generating processes, and running variable cutoffs. Gaussian process regression also demonstrated much better coverage and estimation of standard errors than the other methods.

Our simulation results partly agree with Porter et al. (2017) who found that results using binding score tended to be more precise than those using frontier as the correlation between running variables increased, though we used pooled frontier instead. Pooled frontier produces one singular estimate that combines the weighted average of the treatment effects along each individual frontier (Wong et al., 2013), as opposed to analyzing them separately as done Porter et al. (2017). However, Porter et al. (2017) recommended using the frontier method even if binding score's assumption that treatment effects are homogeneous across frontiers holds theoretical justification, whereas we saw either larger bias or less precision from pooled frontier compared to binding score across our simulation's parameters. In both our simulation as well as in Wong et al. (2013)'s simulation study, pooled frontier showed poor performance without using the true model when the authors applied a constant treatment effects model using pooled frontier to data with heterogeneous treatment effects. Therefore, we believe that frontier and pooled frontier methods are unsatisfactory approaches to estimating two-dimensional RDDs. Except for the least complex data setups, we recommend using Gaussian process regression which showed consistently strong properties. Future research could provide an expanded set of recommendations for when to use Gaussian process regression over other methods, as well as extending the analytic design to applications that involve more than two dimensions to assign students to treatment.

We next applied these methods to an educational data set for an English Language Learner reclassification policy. Gaussian process regression showed similarly precise results to binding score, which showed the smallest standard errors, across all ACT outcomes.

Importantly, we are able to use Gaussian process regression to visually investigate evidence of treatment effect heterogeneity for students with different patterns of running variables. In our

empirical example, our overall $BATE$ estimate for the ACT English test score outcome showed a statistically significant negative effect of assignment to reclassification. When we examined the effects at the sentinel level, we saw underlying heterogeneity: for example, students with a high Overall score but a Literacy score close to the cutoff boundary showed positive and insignificant treatment effect estimates for this outcome, while students with relatively lower Overall scores and Literacy scores that were both close to the boundary experienced negative and significant treatment effects. Their reclassification assignment lowered their ACT English test scores.

We recommend that researchers and policymakers use tools such as Gaussian process regression to understand these patterns of treatment effect heterogeneity when conducting program evaluations. In our application, the negative treatment effects on English test scores were only experienced by some students, while for other ACT outcomes and in other contexts, overall null $BATE$ estimates may mask a mix of positive and negative effects for students at different points along the treatment boundary. By recognizing the differences in both running variables and outcomes of students at different points along the boundary, future policymakers can direct programs differentially to students and therefore target students who may benefit the most.

Notes

1. Letting the treatment and control mean response functions be *a priori* independent is analogous to fitting two separate local linear regressions—one in the treatment group, one in the control group—which is by far the common practice in RDDs (Imbens & Lemieux, 2008). As an alternative to placing independent priors on the treatment and control mean response functions, work by Hahn et al. (2020) suggested placing priors on the control mean response function and the average treatment effect. This was recommended with the assumption that there is some covariate overlap between treatment and control, which is explicitly not the case in RDDs. Applying approaches such as Hahn et al. (2020) is outside the scope of this work, but is discussed in other recent work (Alcantara et al., 2024).

2. For loess, standard errors were estimated via bootstrap.

8 Funding

This work was supported by Grant R305B150010 to Harvard University from the Institute of Education Sciences, U.S. Department of Education.

References

Alcantara, R., Wang, M., Hahn, P. R., & Lopes, H. (2024). Modified BART for Learning Heterogeneous Effects in Regression Discontinuity Designs. *arXiv preprint arXiv:2407.14365*.

Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, *31*(2), 3–32.

Branson, Z., & Mealli, F. (2018). Local Randomization and Beyond for Regression Discontinuity Designs. *arXiv preprint arXiv:1810.02761*.

Branson, Z., Rischard, M., Bornn, L., & Miratrix, L. W. (2019). A nonparametric Bayesian methodology for regression discontinuity designs. *Journal of Statistical Planning and Inference*.

Calonico, S., Cattaneo, M. D., & Farrell, M. H. (2018). On the effect of bias estimation on coverage accuracy in nonparametric inferenc. *Journal of the American Statistical Association*, *113*(522), 767–779.

Calonico, S., Cattaneo, M. D., & Farrell, M. H. (2020). Optimal bandwidth choice for robust bias corrected inference in regression discontinuity designs. *The Econometrics Journal*, *23*(2), 192–210.

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, *82*(6), 2295–2326.

Carlson, D., & Knowles, J. (2016). The Effect of English Language Learner Reclassification on Student ACT Scores, High School Graduation, and Postsecondary Enrollment: Regression Discontinuity Evidence from Wisconsin. *Journal of Policy Analysis and Management*, *35*(3), 559–586.

Cattaneo, M. D., Frandsen, B. R., & Titiunik, R. (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the US Senate. *Journal of Causal Inference*, *3*(1), 1–24.

Cattaneo, M. D., & Titiunik, R. (2022). Regression Discontinuity Designs. *Annual Review of Economics*, *14*, 821–851.

Cattaneo, M. D., Titiunik, R., & Vazquez-Bare, G. (2017). Comparing inference approaches for RD designs: A reexamination of the effect of Head Start on child mortality. *Journal of Policy Analysis and Management*, *36*(3), 643–681.

Chib, S., Greenberg, E., & Simoni, A. (2023). Nonparametric bayes analysis of the sharp and fuzzy regression discontinuity designs. *Econometric Theory*, *39*(3), 481–533.

Chib, S., & Jacobi, L. (2016). Bayesian fuzzy regression discontinuity analysis and returns to compulsory schooling. *Journal of Applied Econometrics*, *31*(6), 1026–1047.

Cleveland, W. S., & Devlin, S. J. (1986). Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*, *83*, 596–610.

Cohodes, S. R., & Goodman, J. S. (2012). First degree earns: The impact of college quality on college completion rates. *HKS Faculty Research Working Paper Series*.

Dee, T. (2012). *School turnarounds: Evidence from the 2009 stimulus* (tech. rep.). National Bureau of Economic Research.

Díaz, J. D., & Zubizarreta, J. R. (2023). Complex discontinuity designs using covariates: Impact of school grade retention on later life outcomes in Chile. *The Annals of Applied Statistics*, *17*(1), 67–88.

Geneletti, S., O'Keeffe, A. G., Sharples, L. D., Richardson, S., & Baio, G. (2015). Bayesian regression discontinuity designs: incorporating clinical knowledge in the causal analysis of primary care data. *Statistics in Medicine*, *34*(15), 2334–2352.

Gramacy, R. B. (2016). laGP: Large-Scale Spatial Modeling via Local Approximate Gaussian Processes in R. *Journal of Statistical Software*, *72*(1), 1–46. https://doi.org/https://doi.org/10.18637/jss.v072.i01

Hahn, J., Todd, P., & Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, *69*(1), 201–209.

Hahn, P. R., Murray, J. S., & Carvalho, C. (2020). Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, *15*(3).

Imbens, G., & Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, *79*(3), 933–959.

Imbens, G., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*(2), 615–635.

Imbens, G., & Zajonc, T. (2011). Regression discontinuity design with multiple forcing variables. *Report, Harvard University.[972]*.

Jacoby, W. G. (2000). Loess: a nonparametric, graphical tool for depicting relationships between variables. *Electoral Studies*, *19*(4), 577–613.

Jin, Z., Liao, W., Papst, I., Zhang, W., Hochstedler, K., & Thoemmers, F. (2021). *rddapp: Regression Discontinuity Design Application* [R package version 1.3].

Kane, T. J. (2003). *A quasi-experimental estimate of the impact of financial aid on college-going* (tech. rep.). National Bureau of Economic Research.

Keele, L., Titiunik, R., & Zubizarreta, J. R. (2015). Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *178*(1), 223–239.

Keele, L. J., & Titiunik, R. (2015). Geographic boundaries as regression discontinuities. *Political Analysis*, *23*(1), 127–155.

Li, F., Mattei, A., & Mealli, F. (2015). Evaluating the causal effect of university grants on student dropout: evidence from a regression discontinuity design using principal stratification. *The Annals of Applied Statistics*, 1906–1931.

Ludwig, J., & Miller, D. L. (2007). Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *The Quarterly Journal of Economics*, *122*(1), 159–208.

Martorell, F. (2005). Do high school graduation exams matter? A regression discontinuity approach. *Unpublished manuscript. University of California Berkeley.*, *28*.

Matsudaira, J. D. (2008). Mandatory summer school and student achievement. *Journal of Econometrics*, *142*(2), 829–850.

Ou, D. (2010). To leave or not to leave? A regression discontinuity analysis of the impact of failing the high school exit exam. *Economics of Education Review*, *29*(2), 171–186.

Papay, J. P., Murnane, R. J., & Willett, J. B. (2010). The consequences of high school exit examinations for low-performing urban students: Evidence from Massachusetts. *Educational Evaluation and Policy Analysis*, *32*(1), 5–23.

Papay, J. P., Murnane, R. J., & Willett, J. B. (2014). High-school exit examinations and the schooling decisions of teenagers: Evidence from regression-discontinuity approaches. *Journal of Research on Educational Effectiveness*, *7*(1), 1–27.

Papay, J. P., Willett, J. B., & Murnane, R. J. (2011). Extending the regression-discontinuity approach to multiple assignment variables. *Journal of Econometrics*, *161*(2), 203–207.

Pei, Z., Lee, D. S., Card, D., & Weber, A. (2022). Local Polynomial Order in Regression Discontinuity Designs. *Journal of Business & Economic Statistics*, *40*(3), 1259–1267.

Porter, K. E., Reardon, S. F., Unlu, F., Bloom, H. S., & Cimpian, J. R. (2017). Estimating causal effects of education interventions using a two-rating regression discontinuity design: Lessons from a simulation study and an application. *Journal of Research on Educational Effectiveness*, *10*(1), 138–167.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Reardon, S. F., Arshan, N., Atteberry, A., & Kurlaender, M. (2010). Effects of failing a high school exit exam on course taking, achievement, persistence, and graduation. *Educational Evaluation and Policy Analysis*, *32*(4), 498–520.

Reardon, S. F., & Robinson, J. P. (2012). Regression discontinuity designs with multiple rating-score variables. *Journal of Research on Educational Effectiveness*, *5*(1), 83–104.

Rischard, M., Branson, Z., Miratrix, L., & Bornn, L. (2021). A Bayesian Nonparametric Approach to Geographic Regression Discontinuity Designs: Do School Districts Affect NYC House Prices? *Journal of the American Statistical Association*, *116*(534), 619–631.

Robinson, J. P. (2011). Evaluating criteria for English learner reclassification: A causal-effects approach using a binding-score regression discontinuity design with instrumental variables. *Educational Evaluation and Policy Analysis*, *33*(3), 267–292.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, *2*(1), 1–26.

Sekhon, J. S., & Titiunik, R. (2017). On interpreting the regression discontinuity design as a local experiment. In *Regression discontinuity designs: Theory and applications* (pp. 1–28). Emerald Publishing Limited.

Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, *51*(6), 309.

Wisconsin Department of Public Instruction. (2023). *WISEdash for Districts ACCESS Dashboards*. Retrieved December 1, 2023, from https://dpi.wi.gov/wisedash/districts/about-data/access

Wong, V. C., Steiner, P. M., & Cook, T. D. (2013). Analyzing regression-discontinuity designs with multiple assignment variables: A comparative study of four estimation methods. *Journal of Educational and Behavioral Statistics*, *38*(2), 107–141.

9 Appendix

9.1 Implementation of Comparison Methods

9.1.1 Binding Score

This method uses both running variables to define a new score, the "binding score,' which is the minimum or maximum of the running variables. The binding score then acts as the running variable in a one-dimensional RDD that uses local linear regression.

In addition to using the binding score BS_i as a running variable, one can also use the original running variables X_{1i} and X_{2i} as additional covariates to improve precision. We include them as covariates per Porter et al. (2017). Within a specified bandwidth, the following model is fit:

$$Y_i = \beta_0 + \beta_1 W_i + \beta_2 BS_i + \beta_3 W_i BS_i + \beta_4 X_{1i} + \beta_5 X_{2i} + \varepsilon_i \tag{24}$$

where $BS = \min(X_1 - c_1, X_2 - c_2)$. The estimated coefficient $\hat{\beta}_1$ is used as the causal effect estimator.

We use the R package *rddapp* (Jin et al., 2021) to estimate $\hat{\beta}_1$. *rddapp* calculates an optimal bandwidth using the Imbens and Kalyanaraman (2012) method. In contrast to the definition of treatment in (1), the package combines both running variables using an *or* statement. This means that instead of using the package to define the treated area as the top right quadrant in Figure 1, we use the package to define the treated area as the remaining three quadrants and then multiply the resulting $\hat{\beta}_1$ by -1. This allows us to identify the $BATE$ that applies to our design, where treated units are in fact in the top right quadrant. In particular, when using packages, it is important to establish that the treatment region defined by the package's functions aligns with the policy setup.

9.1.2 Pooled Frontier

The frontier method described in Section 3.2 reduces a two-dimensional RDD into two one-dimensional RDDs, one across each frontier, and uses local linear regression to estimate each frontier's ATE.

To produce a single overall $BATE$, we use the pooled frontier method described in Wong et al. (2013). To calculate the X_1 frontier's average treatment effect, we limit the data to units that fall above the cutoff, c_2, for the X_2 running variable. Similarly to binding score, we include the X_2 running variable as a covariate to increase precision. Then given the sample, we use units that are centered around their cut point values within a package-specified optimal bandwidth to fit the following model:

$$Y_i = \beta_0 + \beta_1 W_i + \beta_2 X_{1i} + \beta_3 W_i X_{1i} + \beta_4 X_{2i} + \varepsilon_i \tag{25}$$

We do this first for X_1, then again for X_2 (where we include X_1 as a covariate). These two resulting $\hat{\beta}_1$ are then combined, with weights proportional to the average density along each running variable, to calculate the $BATE$. Standard errors of pooled frontier estimates are obtained via bootstrapped samples. We note that Wong et al. (2013) cautions that calculating τ^w from the two boundary estimands τ_1^w and τ_2^w uses weights that are sensitive to the metric, or unit of measurement, and scale, or spread in standard deviation units, of the running variables. This calculation is complicated whenever $\tau_1^w \neq \tau_2^w$.

We also implement this method by using the R package *rddapp* (Jin et al., 2021). Therefore, it requires the same adjustment of the identified treated area and multiplication of the $\widehat{BATE}$ by -1 described above for the binding score method.

9.2 Simulation Parameters

Table 4 lists all parameter combinations used in the simulation as described in Section 9.2, along with $BATE$ values.

9.3 Precision-weighted Results

In Figure 5, we show the results of comparing Gaussian process regression approaches, both density-weighted and precision-weighted, as well as both residualized and non-residualized, in the simulation described in Section 5. These values are averaged over running variable cutoff per-

Model	Correlation	Cutoff 1 Location Percentile	Cutoff 2 Location Percentile	$BATE$
1	0.20, 0.50, 0.90, 0.95, 0.99	30	30	0.400
1	0.20, 0.50, 0.90, 0.95, 0.99	50	50	0.400
1	0.20, 0.50, 0.90, 0.95, 0.99	30	70	0.400
2	0.20, 0.50, 0.90, 0.95, 0.99	30	30	0.400
2	0.20, 0.50, 0.90, 0.95, 0.99	50	50	0.400
2	0.20, 0.50, 0.90, 0.95, 0.99	30	70	0.400
3	0.20	30	30	0.257
3	0.20	50	50	0.283
3	0.20	30	70	0.291
3	0.50	30	30	0.281
3	0.50	50	50	0.296
3	0.50	30	70	0.296
3	0.90	30	30	0.345
3	0.90	50	50	0.348
3	0.90	30	70	0.300
3	0.95	30	30	0.361
3	0.95	50	50	0.363
3	0.95	30	70	0.298
3	0.99	30	30	0.383
3	0.99	50	50	0.383
3	0.99	30	70	0.296
4	0.20	30	30	-0.832
4	0.20	50	50	-0.485
4	0.20	30	70	-0.794
4	0.50	30	30	-0.486
4	0.50	50	50	-0.304
4	0.50	30	70	-0.799
4	0.90	30	30	0.177
4	0.90	50	50	0.196
4	0.90	30	70	-0.646
4	0.95	30	30	0.278
4	0.95	50	50	0.285
4	0.95	30	70	-0.617
4	0.99	30	30	0.366
4	0.99	50	50	0.367
4	0.99	30	70	-0.591

Table 4: Simulation parameters.

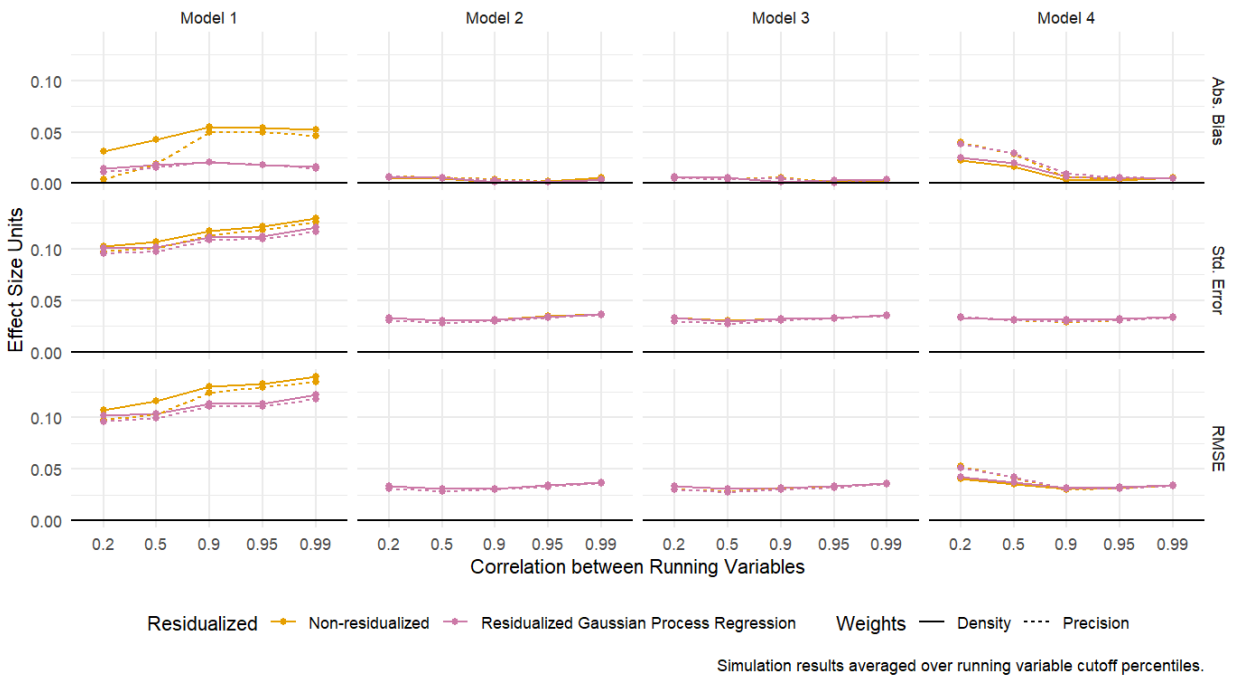

Figure 5: Comparison of residualized and non-residualized Gaussian process regression, both using density or precision weights for $n = 1,000$.

centile combinations as results were similar.

Residualized Gaussian process regression tends to outperform non-residualized in the first column, which represents the least complex data generating model, and shows comparable properties non-residualized Gaussian process regression otherwise. Then, looking within residualized Gaussian process regression, its absolute bias, standard errors, and RMSE are fairly similar between precision weights and density-based weighting. The only notable difference is in the top right corner, where density-weighted residualized Gaussian process regression shows lower absolute bias than non-residualized density weighted regression for low running variable correlations. As we do not see strong gains from using precision-weighted residualized Gaussian process regression, we only used density-weighted residualized Gaussian process regression results in our main results.

Next, we compare the properties of residualized Gaussian process regression and loess, varying their weighting method, for simulations using $n = 1,000$ in 6 and using $n = 5,000$ in 7. Results were similar over running variable cutoff percentile combinations and are therefore averaged in this figure.

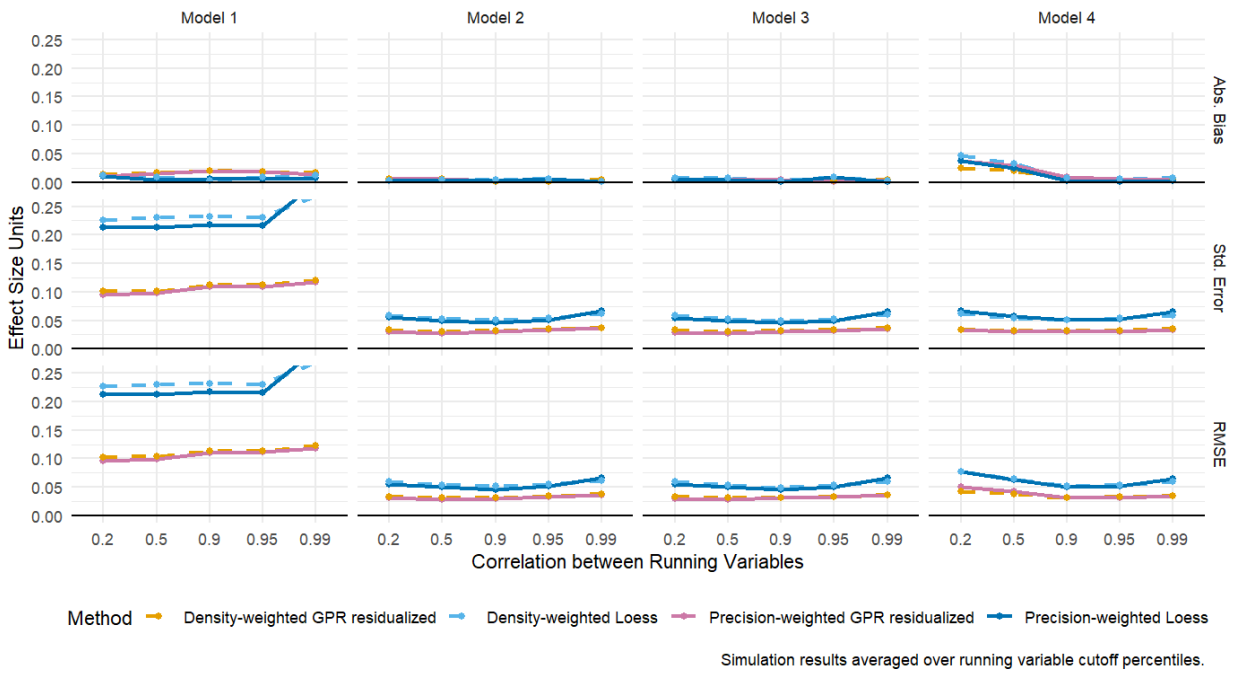

Figure 6: Loess and residualized Gaussian process regression results using both density and precision weights for $n = 1,000$.

Across both sample sizes, results are quite similar between the weighting approaches within methods. There is some evidence of underperformance of precision-weighted loess compared to density-weighted loess in model one in terms of standard errors, though this difference dissipates for the $n = 5,000$ case. Given these results and to make consistent comparisons with density-weighted Gaussian process regression, we used density-weighted loess in the main text of the simulation results.

9.4 Simulation Results $n = 5,000$

Porter et al. (2017) use a sample size of $n = 5,000$ in their simulation study. We chose to use $n = 1,000$ in our main text simulation results because this sample size may more closely approximate the size of typical empirical data sets. However, for completeness we include $n = 5,000$ results here in terms of bias, precision, RMSE, coverage, and estimation of standard errors. The results here are similar to those in the $n = 1,000$ case.

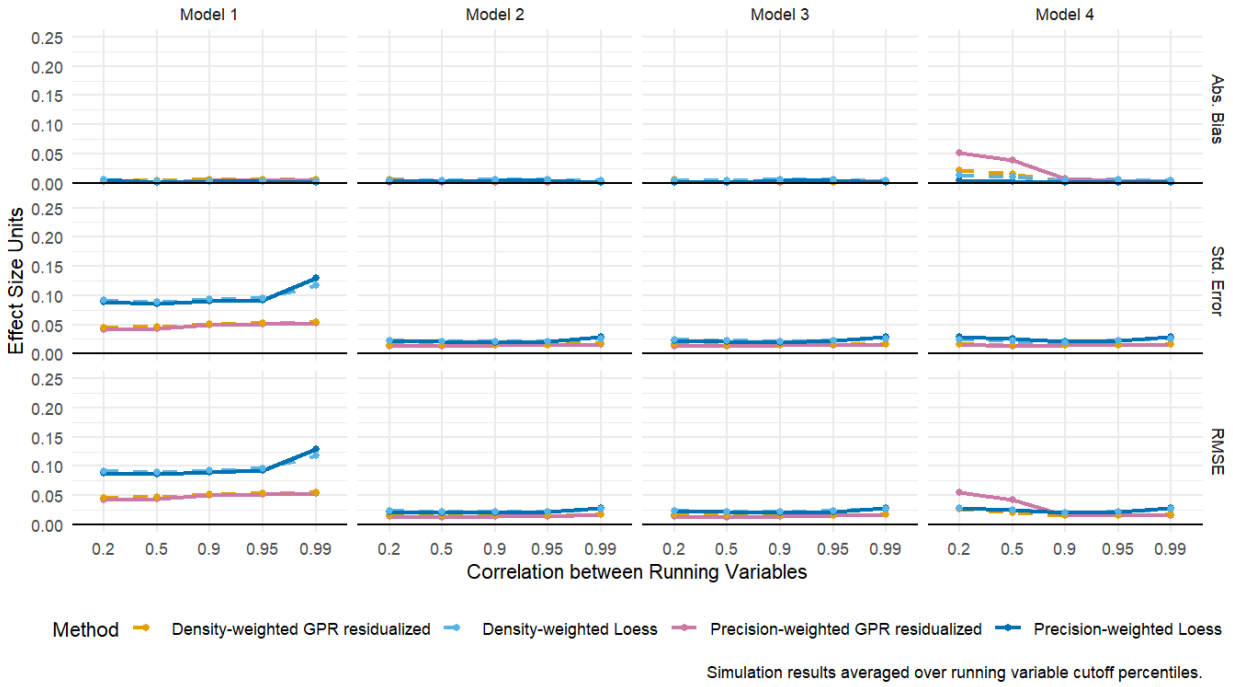

Figure 7: Loess and Gaussian process regression results using both density and precision weights for $n = 5,000$.

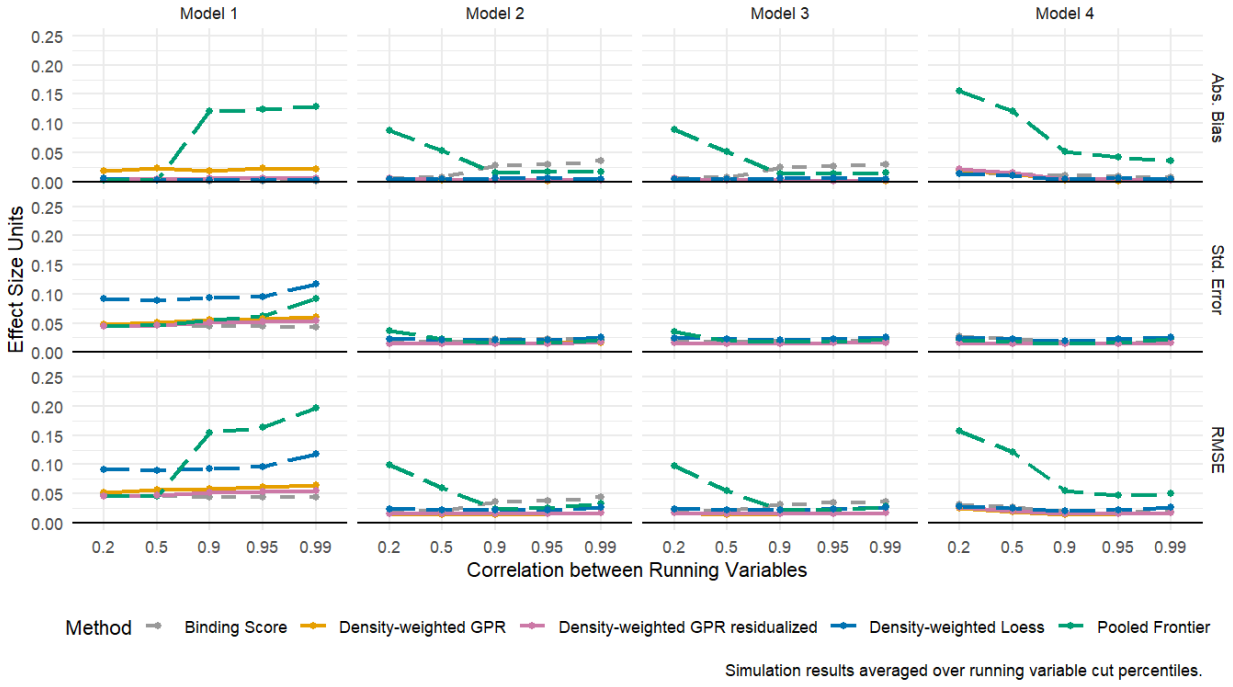

Figure 8: Simulation results across running variable cutoff percentiles for $n = 5,000$.

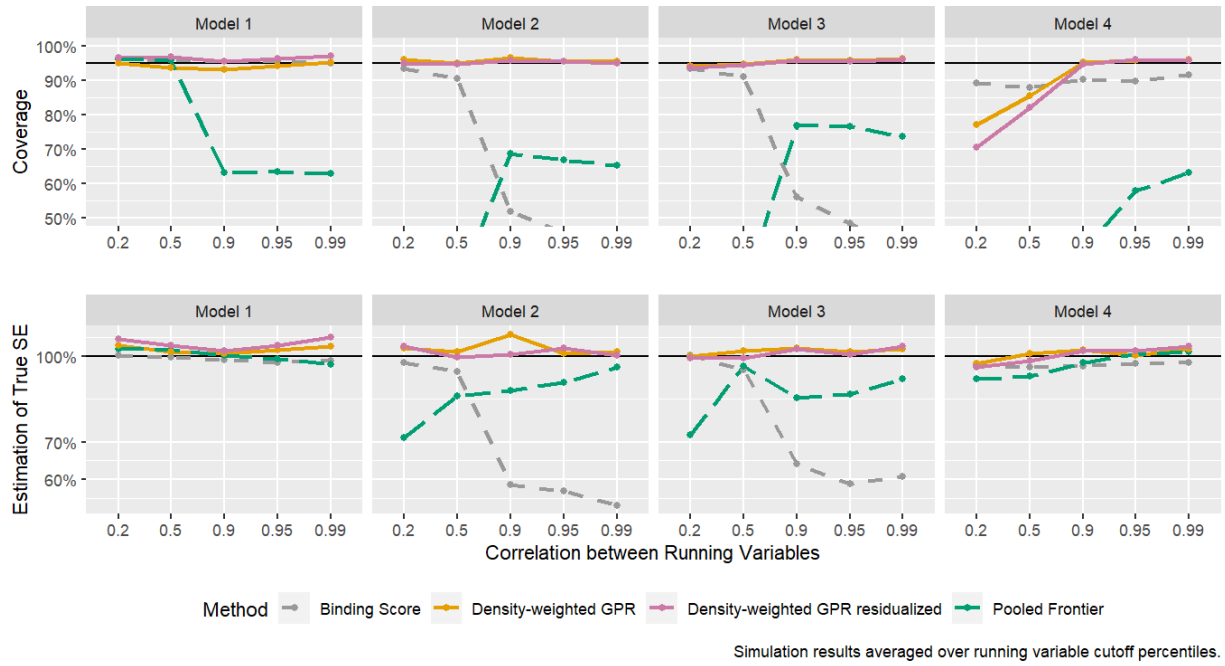

Figure 9: Simulation coverage across parameters.

9.5 Simulation Results $n = 1,000$ across Cutoff Percentiles

We, following Porter et al. (2017), averaged our main results over three different running variable percentile cutoffs: the 30th percentile on both running variables, the 50th percentile on both running variables, and the 30th percentile on one running variable with the 70th percentile on the other running variable. We include these results broken out by cutoffs here. Note that we are unable to estimate pooled frontier results under high correlations for running variable cutoff percentiles of 30/70 due to a lack of data around the 70th percentile cutoff for one of the running variables.

9.6 Treatment Effect Heterogeneity in ACT Test Outcomes

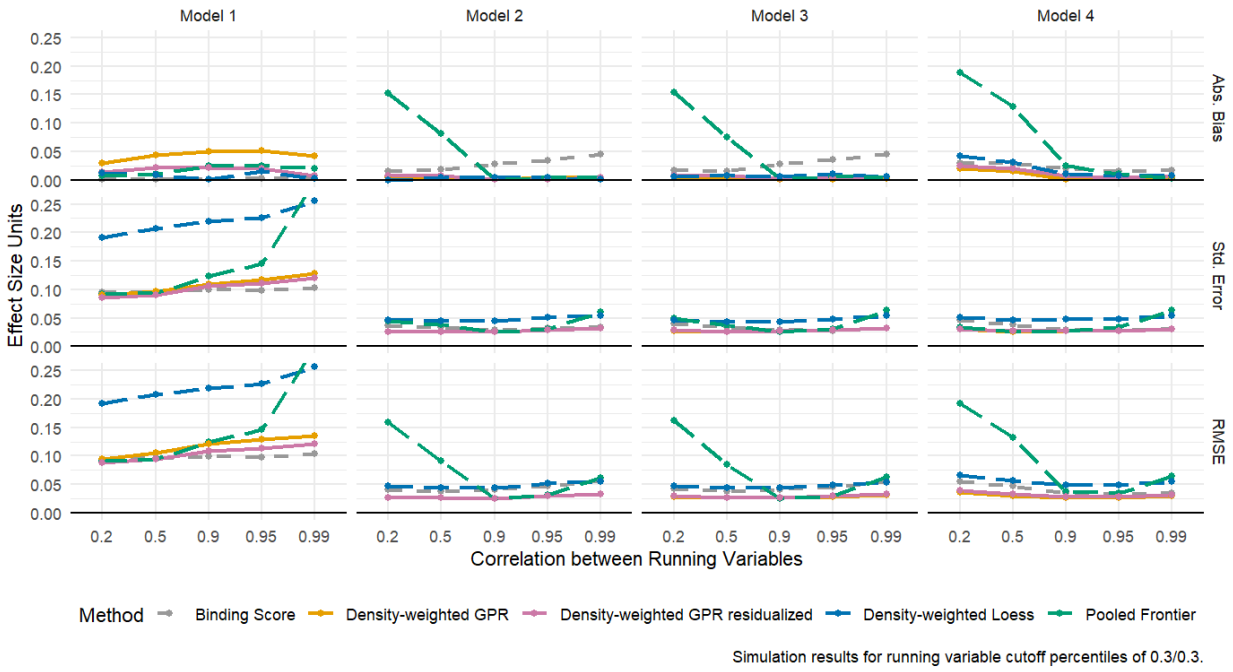

Figure 10: Simulation results for running variable cutoff percentiles of 30/30 for $n = 1,000$.

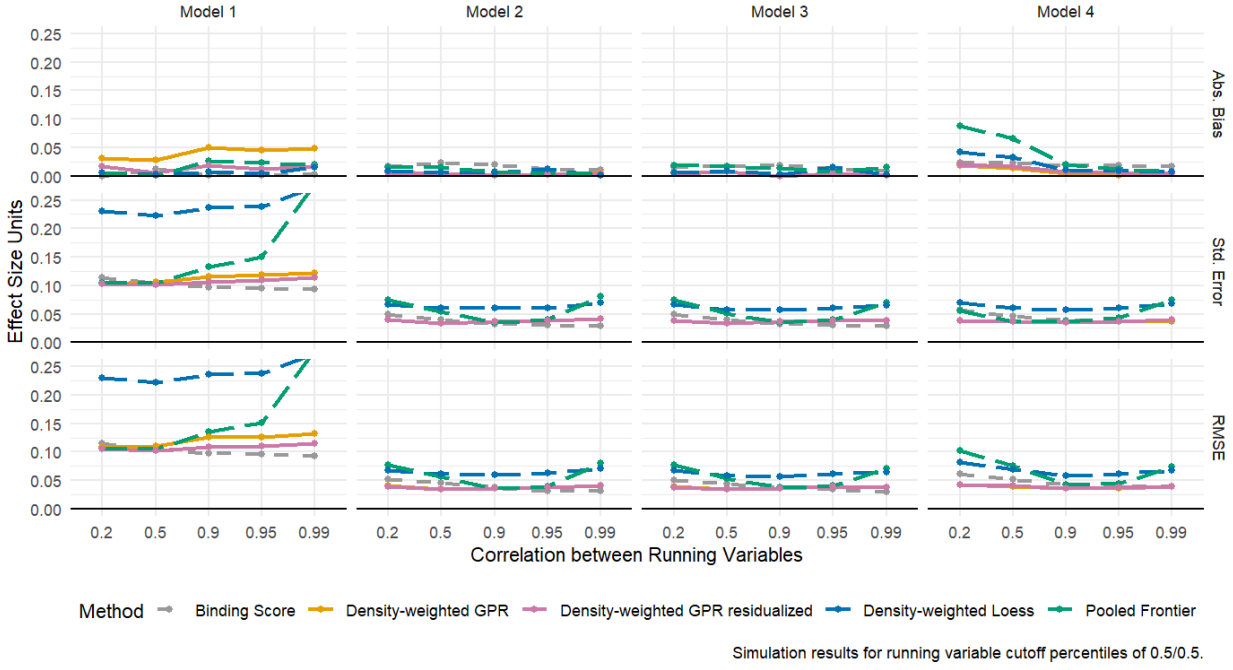

Figure 11: Simulation results for running variable cutoff percentiles of 50/50 for $n = 1,000$.

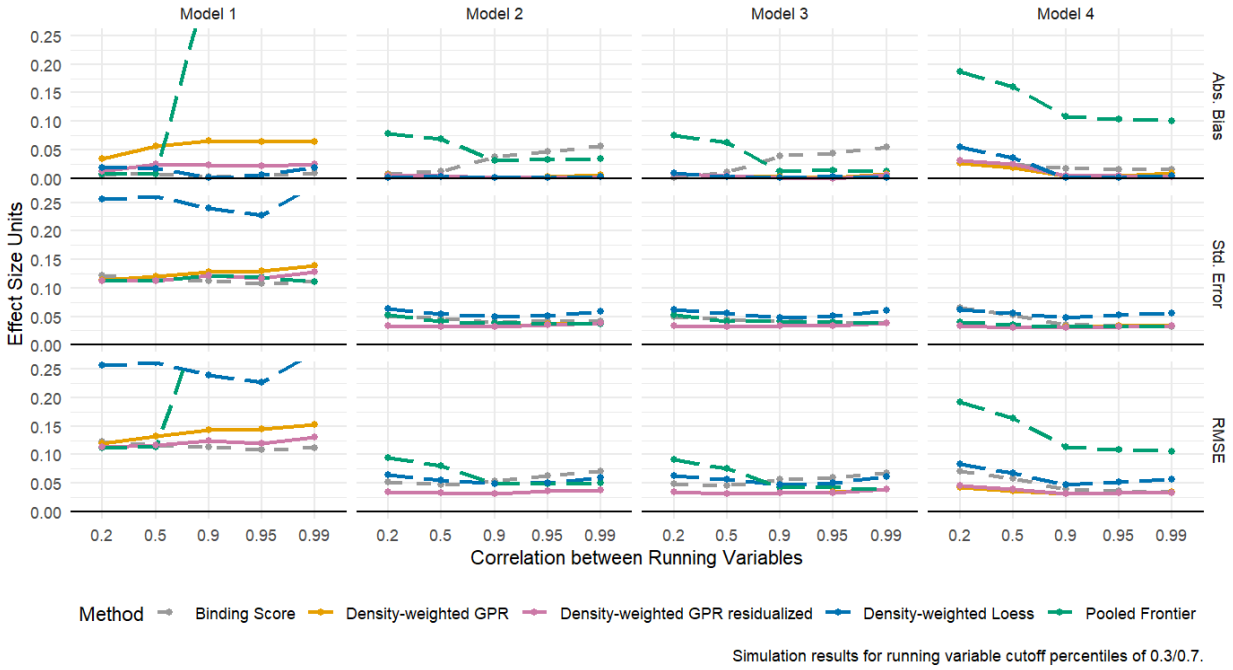

Figure 12: Simulation results for running variable cutoff percentiles of 30/70 for $n = 1,000$.

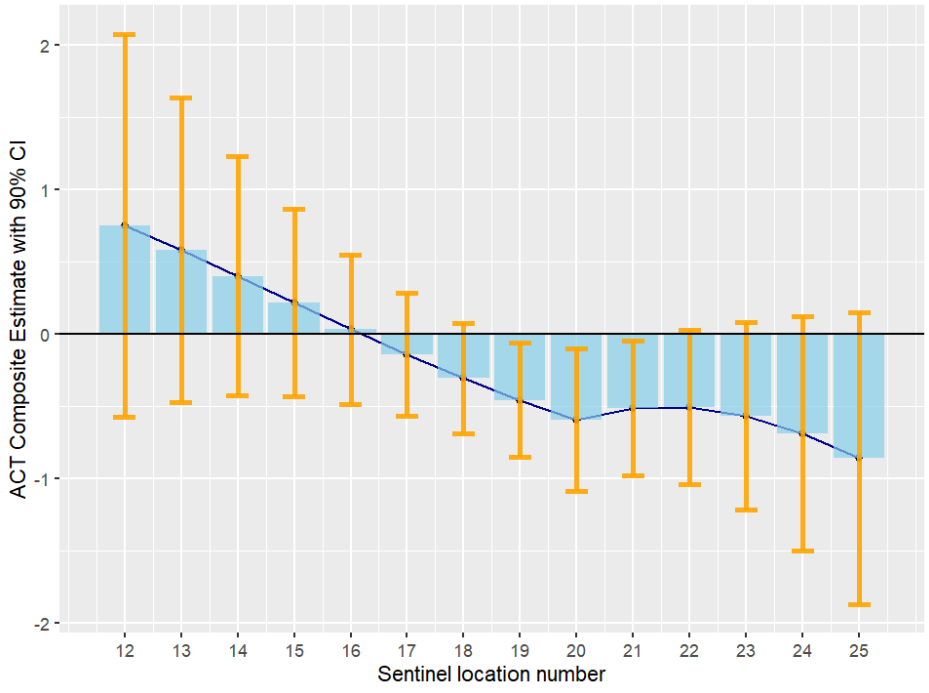

Figure 13: Gaussian process regression sentinel estimates for ACT Composite scores.

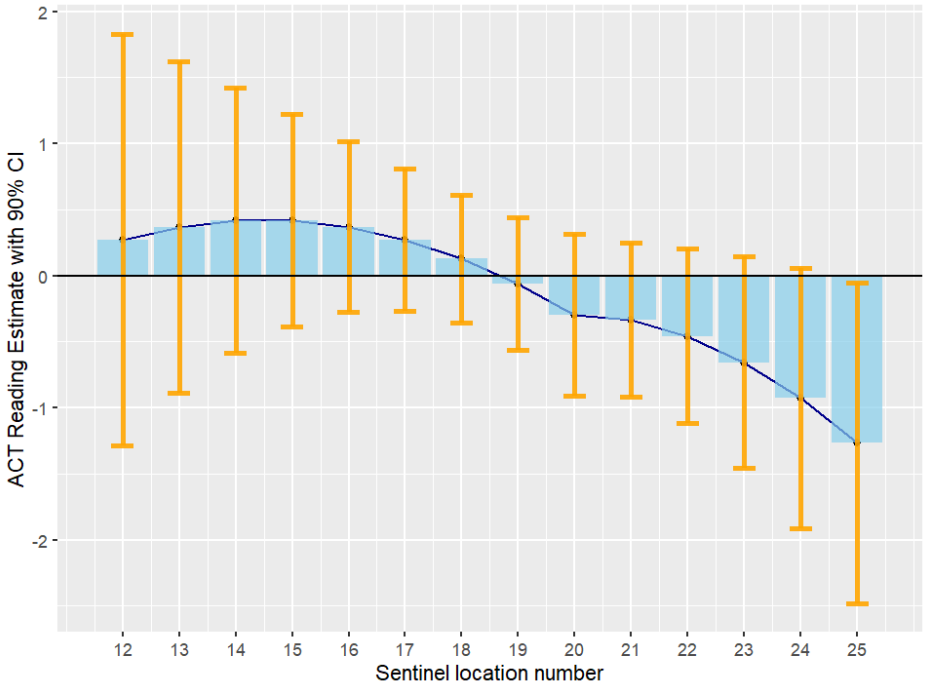

Figure 14: Gaussian process regression sentinel estimates for ACT Reading scores.

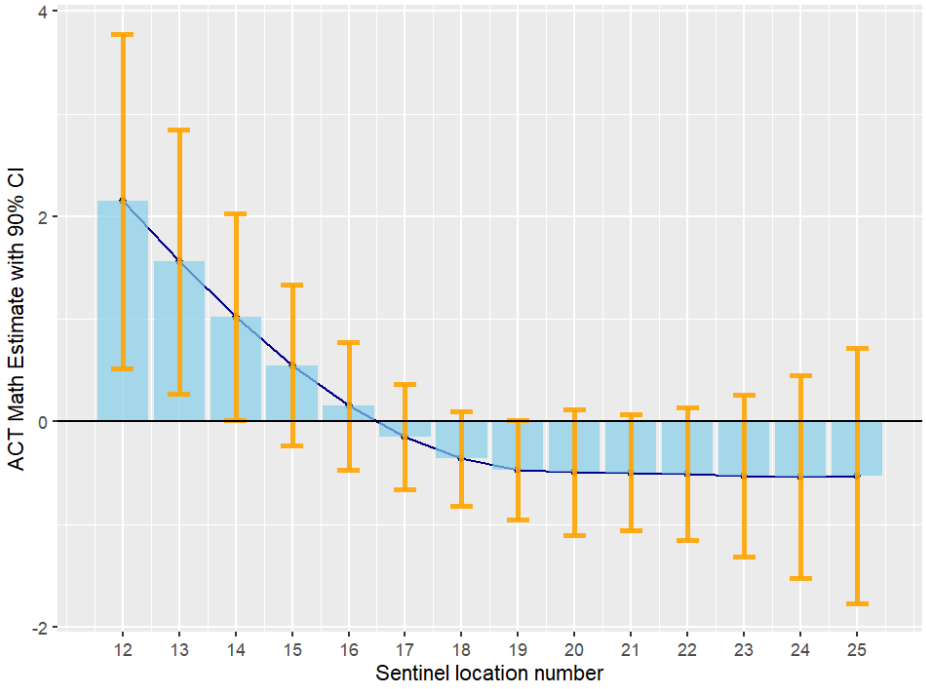

Figure 15: Gaussian process regression sentinel estimates for ACT Math scores.

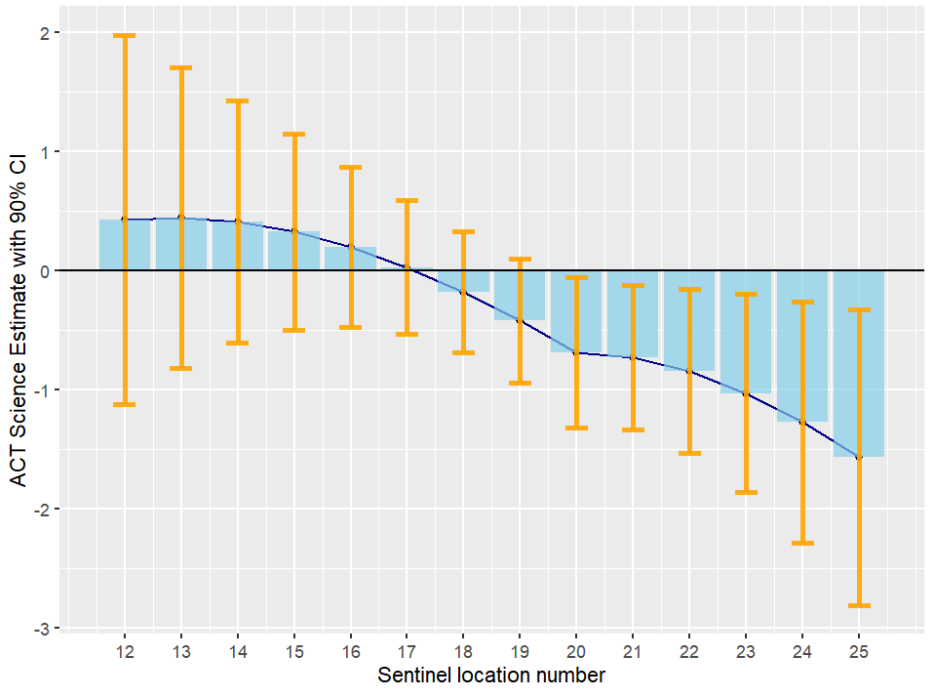

Figure 16: Gaussian process regression sentinel estimates for Science scores.