UCDAVIS
ENERGY ECONOMICS PROGRAM

DEEP WP 019

Regression Discontinuity in Time: Considerations for Empirical Applications

Catherine Hausman and David S. Rapson

April, 2018

Davis Energy Economics Program working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to review by any editorial board.

Regression Discontinuity in Time: Considerations for Empirical Applications

Catherine Hausman David S. Rapson*

April 2018

Abstract

Recent empirical work in several economic fields, particularly environmental and energy economics, has adapted the regression discontinuity (RD) framework to applications where time is the running variable and treatment begins at a particular threshold in time. In this guide for practitioners, we discuss several features of this "Regression Discontinuity in Time" (RDiT) framework that differ from the more standard cross-sectional RD framework. First, many applications (particularly in environmental economics) lack cross-sectional variation and are estimated using observations far from the temporal threshold. This common empirical practice is hard to square with the assumptions of a cross-sectional RD, which is conceptualized for an estimation bandwidth shrinking even as the sample size increases. Second, estimates may be biased if the time-series properties of the data are ignored (for instance in the presence of an autoregressive process), or more generally if short-run and long-run effects differ. Finally, tests for sorting or bunching near the threshold are often irrelevant, making the framework closer to an event study than a regression discontinuity design. Based on these features and motivated by hypothetical examples using air quality data, we offer suggestions for the empirical researcher wishing to use the RD in time framework.

Key Words: Regression discontinuity design; Autoregression; Treatment effects; Interrupted time series

JEL: C21; C22; C14; Q53

*(Hausman) Ford School of Public Policy, University of Michigan. Email: chausman@umich.edu. (Rapson) Department of Economics, University of California, Davis. Email: dsrapson@ucdavis.edu. We thank Michael Anderson, Max Auffhammer, Severin Borenstein, Matias Cattaneo, Lucas Davis, Stephen Holland, Ryan Kellogg, Doug Miller, and Jeff Smith for helpful comments. We also thank UC Berkeley's Energy Institute at Haas, where Rapson was a visitor while this research was conducted.

As noted by Lee and Lemieux (2010), the use of the regression discontinuity framework (RD) has expanded rapidly in economics for two reasons: (1) "RD designs require seemingly mild assumptions compared to those needed for other nonexperimental approaches"; and (2) "the belief that the RD design is not 'just another' evaluation strategy, and that causal inferences from RD designs are potentially more credible than those from typical 'natural experiment' strategies" (p. 282). An increasingly popular application of the RD design uses time as the running variable, with a treatment date as the threshold. Examples of this "regression discontinuity in time" (RDiT) are given in Table 1. Papers using RDiT span fields that include public economics, industrial organization, environmental economics, marketing, and international trade. Many are published in high-impact journals and, as of the writing of this paper, have received a cumulative citation count of around 1,100.[1] In this paper, we describe potential pitfalls affecting RDiT, illustrate these pitfalls via simulations, and offer practical advice to researchers considering RDiT.

Many—although not all—of the papers using RDiT are estimating treatment effects of environmental and energy policy.[2] As a hypothetical example, imagine that some regulation caused all coal-fired power plants in the United States to install an emissions control device at the same time, on the same day. The hypothetical RDiT researcher could then examine the impact on ambient air quality by estimating the change in pollution concentrations in the air at the time the policy occurred. The typical RDiT application in energy and environmental economics shares many features with this hypothetical example: (1) there is no available cross-sectional variation in policy implementation, so a difference-in-difference framework is not possible; (2) ambient air quality data are available at a daily or hourly frequency over a long time horizon (e.g., years); and (3) there are many potential time-varying confounders, which are assumed to change smoothly across the date of the policy change.

In contrast, a hypothetical cross-sectional RD might, for example, be based around a

[1] Citation counts were retrieved on Google Scholar on January 29, 2018. This represents a greater than 50 percent increase in citations since our last count in mid-2016.

[2] A possible explanation for the proliferation if RDiT in energy and environmental economics, relative to other applied microeconomic fields, is the availability of high-frequency time-series pollution data.

Table 1: Examples of RDiT Papers

Paper	Journal	Setting
Anderson (2014)	American Econ Review	Traffic
Auffhammer and Kellogg (2011)	American Econ Review	Air quality
Bento et al. (2014)	American Econ Journal – Econ Policy	Traffic
Burger et al. (2014)	Trans Research Part A	Car accidents
Busse et al. (2006)	American Econ Review	Vehicle prices
Busse et al. (2010)	Marketing Science	Vehicle prices
Chen and Whalley (2012)	American Econ Journal – Econ Policy	Air quality
Chen et al. (2009)	J Marketing Research	Customer satisfaction
Davis (2008)	J Political Econ	Air quality
Davis and Kahn (2010)	American Econ Journal – Econ Policy	Vehicle sales
DePaola et al. (2013)	Empirical Econ	Car accidents
Gallego et al. (2013)	J Public Econ	Air quality
Grainger and Costello (2014)	J Environmental Econ & Management	Fisheries
Lang and Siler (2013)	Energy Policy	Energy use

policy that required the device only for power plants above a certain size threshold (say, 1000 megawatts). The RD researcher could then compare air quality near power plants just above and below the threshold (e.g., with size 990 to 1010). In this RD, the researcher observes many units near the threshold, and other confounders are assumed to change smoothly across the size threshold.[3]

In this paper, we argue that the most common deployment of RDiT faces an array of challenges—due primarily to its reliance on time-series variation for identification. The use of time-series variation renders the traditional RD toolkit (for example, as described in Lee and Lemieux (2010)) unavailable to the researcher. Researchers implementing RDiT are advised to be aware of the multiple ways in which the framework differs from an experimental ideal. We point to three potential pitfalls, conducting Monte Carlo simulations and providing recommendations for empirical researchers.

First, many RDiT designs in the existing literature have required observations far from the temporal threshold. Identification in cross-sectional RD designs rests upon a conditional expectation as one approaches the threshold—thus relying on a mass of cross-sectional units just above and below the threshold. However, because the RDiT framework is frequently

[3]As another example, not from energy and environmental economics, a cross-sectional RD might estimate the impact of a tutoring program on student educational outcomes. To qualify for the program, a student would need to have scored below some threshold (say, 500 points) on an exam. The researcher would then look at student outcomes after the program only for students with exam scores close to 500 points.

used in studies with no (or insufficient)[4] cross-sectional identifying variation, the sample is too small for estimation as the bandwidth narrows around the threshold. As a result, many RDiT researchers expand T to obtain sufficient power—thus relying on observations away from the threshold. In contrast, the typical cross-sectional RD can, in theory, increase the sample size by growing N even while approaching the threshold. The use in RDiT of observations remote (in time) from the threshold is a substantial conceptual departure from the identifying assumptions used in a cross-sectional RD, and we show it can lead to bias resulting from unobservable confounders and/or the time-series properties of the data generating process.

Second, RDiT requires that the researcher consider the time-series nature of the underlying data-generating process. The time-series literature has developed methods to account for processes that are autoregressive, for instance, but these methods have not generally been applied in the RDiT context. We demonstrate that such methods are relevant for proper estimation and interpretation of short-run versus long-run effects.

Finally, the McCrary (2008) density test, standard in cross-sectional RDs, is often not applicable when time is the running variable. This test allows researchers to assess whether observed units have sorted across the threshold, i.e. into or out of treatment, which would lead to bias in the empirical estimates. When the density of the running variable is uniform (e.g. time), the test becomes irrelevant. As such, the researcher can evaluate sorting into or out of treatment only indirectly.

Commenters have noted the similarity between RDiT and other methods (Shadish et al., 2002). For instance, the identifying variation in RDiT is similar to that in interrupted time series or a simple pre/post comparison. Davis (2008) argues that an advantage of RDiT over a pre/post comparison is the ability to include flexible controls, driven by the high-frequency nature of the data used in most RDiT papers. This high-frequency data also

[4]Some RDiT studies observe data series at multiple locations, but with policy implementation occurring at a single date for all locations. Thus any spatial correlation will undermine the cross-sectional variation. Other papers observe data series at multiple locations, but estimate a separate RDiT for each individual location, so that asymptotics are conceptualized in T.

invites comparisons to the widely-used "event study" framework. However, while the *context* in which RDiT is used is often similar to the context in which event studies are used, the specifications used in practice have tended to be quite different, as has the discussion of identification. For instance, the widely-cited event studies in environmental economics that have analyzed stock returns[5] generally (1) do not use long time horizons, instead focusing on short windows around the event; (2) do not use a single time series, but rather panel data; and (3) do not use high-order polynomial controls in time.

This paper clarifies how the RDiT framework as typically implemented differs from the cross-sectional RD. In doing so, we encourage researchers implementing RDiT to be aware of the ways the framework differs from an experimental ideal and to recognize that the failure of untestable assumptions can lead to biased results.

1 RDiT Framework & Comparison to Other Methods

1.1 RD Framework

In this section we briefly lay out and compare the cross-sectional RD framework and the RDiT framework. We begin by revisiting the cross-sectional RD framework, with the main purpose to reiterate the notation that has become common in the RD literature, following Lee and Lemieux (2010) and Imbens and Lemieux (2008). For context, suppose a researcher wishes to evaluate the impact on air quality of the policy described in the introduction: a policy that requires a pollution control device only for power plants above a certain size threshold, such as 1000 megawatts.

The appeal of the RD approach can be most clearly seen through the lens of the Neyman-Rubin potential outcomes framework (Rubin, 1974; Splawa-Neyman, 1923 [1990]) and the intuitive similarity to the randomized controlled trial (RCT). The empiricist cannot observe an individual power plant (i) in more than one state of the world at once, so she cannot

[5]Examples include Hamilton (1995); Konar and Cohen (1997); Dasgupta et al. (2001).

explicitly compare an outcome such as nearby air quality Y_i under both treatment status (T) and control status (C). The causal treatment effect is logically well-defined as the difference in the outcome variable in the two states of the world ($Y_i^T - Y_i^C$), but it is unobservable at the individual level. Instead, empiricists seek to estimate the average treatment effect $E[Y_i^T - Y_i^C]$ using some representative sample of the relevant population. A comparison of average outcomes of subjects who receive treatment to that of subjects who do not receive treatment (i.e., $E[Y_i^T] - E[Y_i^C]$) provides an estimate of the average treatment effect. If the expected outcome of treated subjects *had they not been treated* is equal to the expected outcome of non-treated subjects, then an unbiased estimate of the causal impact of the treatment can be obtained.[6]

A well-designed experiment randomly assigns subjects into treatment or control status. One way to randomize would be to assign each subject a draw, ν, from a distribution. If the distribution is uniform over the range, say, $[0, 1]$, then a simple assignment rule can allocate subjects into control and treatment. For a 50/50 split, subjects with a draw of $\nu > 0.5$ are given the treatment, and those with $\nu \leq 0.5$ are not.

The power and appeal of RD designs are derived from their close proximity to this experimental ideal. Lee and Lemieux (2010) go so far as to describe RD designs as a "local randomized experiment" (p. 289). In the RD context, treatment occurs on one side, but not the other, of some threshold c for a treatment assignment variable X.[7] That is, a treatment assignment variable X_i is observed for each individual i. Then, if $X_i > c$, the subject is treated, and if $X_i < c$, the subject is not. Typically many units of observation fall in the neighborhood of the threshold, allowing estimation via a cross-sectional comparison of subjects "just above" and "just below" the threshold.

In the previous example, rather than randomly assigning a power plant's treatment status

[6]In the power plant example, the mean potential air quality under treatment near treated power plants must equal the mean potential air quality under treatment near untreated power plants. Similarly, the mean potential air quality without treatment near treated power plants must equal the mean potential air quality without treatment near untreated power plants.

[7]We refer to this as the "running variable" throughout. It is also commonly referred to as the "forcing variable" in the RD literature.

by, say, flipping a coin, treatment in the RD framework is determined by whether the power plant is larger than 1000 megawatts. While very small power plants (e.g. 100 megawatts) are likely to differ from very large power plants (e.g. 3000 megawatts), the researcher may reasonably assume that power plants near the threshold are comparable (e.g., a 990 megawatt plant is comparable to a 1100 megawatt plant). A credible RD design thus is conceptually similar to the randomized experiment described above: in the neighborhood of the threshold, whether X is above or below the threshold is "as good as random." If true, RD designs can be analyzed, tested and trusted like randomized controlled trials.

Lee and Lemieux (2010) develop a checklist of recommended diagnostics that can provide evidence on the quality of causal inference using RD estimates. The purpose of the checklist is, basically, to determine that there are no other explanations than the treatment itself for differences in outcomes between treatment and control subjects. If the empirical setting passes the tests, the RD is generally viewed as producing an unbiased estimate of the causal average treatment effect.[8]

1.2 RD in Time Framework

In the typical RDiT framework, the researcher knows the date c of a policy change. She assumes that for all dates $t > c$, the unit is treated, and for all dates $t < c$, the unit is not. This RDiT set-up uses time-series data, for instance daily or hourly observations, for a given geographic region (e.g., a city as in Chen and Whalley (2012) or a building as in Lang and Siler (2013)).[9] A common application in the energy and environmental economics literature is the evaluation of policies impacting air quality (see Table 1).

Returning to the power plant example, suppose the policy requiring a pollution control device takes effect on a particular date and impacts *all* power plants. Suppose the researcher

[8]The estimate of the average treatment effect is *local* in the sense that it is representative only of the units of observation near the threshold; it may differ from the treatment effect for units far from the threshold.

[9]Some papers leverage cross-sectional variation by observing multiple geographic regions, although typically the timing of the policy change is common to the different regions. Other papers observe multiple geographic regions, but estimate a separate RDiT specification for each location. In either case, asymptotics are primarily or entirely in T, not N.

does not observe emissions from individual firms, but rather ambient air quality within a city. Rather than observing air quality near some power plants that are treated and others that are not (as in the cross-sectional RD framework above), the researcher observes air quality near power plants that are treated in some periods and untreated in others. Among the identification challenges in such a setting are that (1) there is no cross-sectional variation in treatment, so no difference-in-differences is possible;[10] and (2) there are many confounders, such as changes in weather and emissions from other industries, that may change at the same time as the policy change.

In contrast to the time-series examples provided above, it is possible to imagine an RD context where time is the running variable but the framework approaches the as-good-as-random interpretation of a cross-sectional RD. We would exclude such an empirical context from the RDiT class. Consider Ito (2015), which uses time as the running variable but relies on asymptotics in N. The time threshold in that study determines eligibility for an electricity rebate program. Customers who initiated service after the threshold date were ineligible, while those before were eligible. This threshold date was set retroactively, long after it had passed. To return to the power plant example, a comparable RDiT would be if coal plants built after the year 2001 were required to install emissions controls, and this policy were announced and took effect in the year 2010. For Ito's design, there is a credible claim to exogeneity of service initiation date with respect to the threshold (no sorting behavior), a substantial mass of customers on either side of the threshold (asymptotics in N), and the ability to deploy the standard toolkit of cross-sectional RD diagnostics. This design differs markedly from studies that rely on a time threshold and identify based on asymptotics in T (rather than N).

[10]RDiT can also provide value in settings where difference-in-differences is possible but limited in some way. For instance, Auffhammer and Kellogg (2011) use a difference-in-difference framework as part of their analysis, but they additionally use the RDiT approach because they are interested in heterogeneous treatment effects across space. Some papers also compare RDiT and difference-in-difference estimates, such as Chen and Whalley (2012) and Gallego et al. (2013)—this comparison can be useful where an untreated unit exists but its validity as a control is in doubt.

1.3 Conceptual Differences Between RD and RDiT

Comparing the cross-sectional RD with the RDiT, several conceptual differences emerge. The first is one of interpretation. The previous sections described a "local randomization" interpretation of the RD design in which, within a small neighborhood around the threshold, treatment status can be thought of as essentially akin to a roll of a die. As noted by Jacob et al. (2012), while some researchers have focused on local randomization, others have emphasized instead the RD as characterized by the discontinuity at a threshold. The randomization characterization is emphasized by, for instance, Lee (2008) and Lee and Lemieux (2010), whereas the discontinuity characterization is emphasized by Hahn et al. (2001). In part, the distinction is about whether the RD framework is closer to a randomized experiment than are other quasi-experimental frameworks, a matter of some debate (Shadish et al., 2002). The RDiT framework in the power plant example clearly accords with the "discontinuity at a threshold" interpretation of RD designs (where the discontinuity is at a moment in time), but it is less clear that it accords with the "local randomization" interpretation of RD.[11] The running variable is time itself—and time cannot be thought of as randomly assigned within a neighborhood around a threshold.[12] To the extent that the RD framework is simply another quasi-experimental framework (one that uses a discontinuity), RDiT is conceptually similar. To the extent that RD is closer to a randomized trial (i.e., uses a local randomization), RDiT is conceptually distinct.

A second, and more important, difference between RD and RDiT is the way in a which the sample size is allowed to grow. The standard RD is identified in the N dimension, allowing the researcher to grow the sample size even as the bandwidth shrinks arbitrarily around the threshold. In contrast, RDiT is typically identified using variation in the T dimension—of

[11] A brief discussion of RD in time is given in Lee and Lemieux (2010), motivated by the large number of studies that use age as the running variable. The focus of that discussion is on the "inevitability" of treatment.

[12] The threshold itself (the particular date the treatment is implemented) is also not generally randomly assigned; rather it is chosen by policymakers and announced in advance.

the fourteen papers we cite, only three use cross-sectional variation in the treatment date.[13] When using a time series of air quality data with a single treatment date, a researcher cannot grow the sample size while simultaneously shrinking the bandwidth (the proximity to the threshold). Rather than simply a matter of interpretation, the inability to grow the sample size near a threshold poses operational problems for the empirical researcher. RDiT papers tend to rely on observations quite far (in time) from the threshold, leaving the researcher more vulnerable to bias from unobservable confounders. The majority of the RDiT papers we examine use a sample size of at least two years, and several use eight years or more—a period over which many unobservable changes could occur.

A related difference between RD and RDiT is the need to include control variables. In the cleanest cross-sectional RD, since treatment is as good as random within a narrow bandwidth, few controls are needed. Researchers include a control for the running variable itself, so that the necessary assumption is simply that the association between the running variable and the potential outcomes is smooth. Other covariates can be included to reduce noise and increase precision, but they are not needed for an unbiased estimate (Lee and Lemieux, 2010). In contrast, in the RDiT setting, unobservables correlated with the running variable may have discontinuous impacts on the potential outcome. For instance, supposing that the power plant policy were implemented on a Monday and the researcher observed air pollution at the daily level, it is hard to imagine that the potential outcome evolves smoothly from the weekend to the beginning of the work week. Of the fourteen papers we cite, all but one include discontinuous controls such as day-of-week or weekend effects. The impact of such covariates is in part what prevents the researcher from using a very narrow window: one could not simply compare the day before the threshold to the day after. As a result, covariates in the RDiT context may need to be included as controls to prevent bias, rather than simply to improve precision.

Finally, the assumptions needed for inference are different from those in the cross-sectional

[13]Some of the papers have multiple cross-sectional sites, but with a common treatment date. Others have multiple treatment dates, but run the RDiT specifications on site-specific time-series.

RD. In particular, the errors are likely to exhibit persistence. We refer the reader to the large time-series literature for how to deal with serially correlated errors.

Two additional observations are worth making with regard to inference. First, for a local linear RDiT framework, small-sample inference may be necessary. This differs from the typical implementation of the cross-sectional RD, where the researcher may have a large sample even close to the threshold. Second, some series (such as commodity prices) will contain unit roots and as such require different procedures for inference than are found in cross-sectional RD designs.

2 Monte Carlo Simulations

In this section, we describe Monte Carlo simulations that allow us to examine and decompose threats to identification in the RDiT and to test potential remedies. Using real-world pollution monitoring data as a base, we impose simulated treatment effects. Essentially, we are simulating the impact of a policy such as the one described previously, in which power plants are required to install an emissions control device. We then attempt to retrieve the policy impact using RDiT approaches. In addition to results using real-world pollution monitoring data (with a simulated treatment), we present results using simulated data in which we control the data-generating process completely. While the former dataset has properties similar to data used in many existing RDiT applications, the latter allows us to examine how our results might generalize to other settings.

For the results using air quality data, we use the daily ozone and weather data from Auffhammer and Kellogg (2011) (AK). We drop monitors in California, since those are subject to the sharply discontinuous treatment found in AK. It is important to note that the remaining monitors may also face confounding treatments, such as (unobserved to the researcher) environmental regulations. However, these are the sorts of confounders that the RD approach is meant to control for. We restrict the sample to those monitors open for the

entire sample period and with no two-month gaps in coverage, leaving us with 108 monitors. The AK data contain daily observations of the maximum measured ozone concentration. As is standard in the literature, we use the log of ozone concentrations. Control variables include daily minimum and maximum temperature and daily rain and snowfall totals. Table A1 in the Appendix gives summary statistics.

The first set of Monte Carlo simulations is aimed at demonstrating how the RDiT method recovers the true causal treatment effect when the estimating equation is correctly specified. We begin by constructing a simulated treatment effect that begins at a known start date (using ten randomly selected start dates).[14] Then for each monitor's data series $x_{i,t}$, and using $\beta = -0.2$, we construct our outcome variable $y_{i,t} \equiv x_{i,t} + \beta \cdot \mathbb{1}\{t \geq t_{start}\}$. A representative figure of both the true pollution data series $x_{i,t}$ and the constructed outcome variable $y_{i,t}$ is given in the Appendix (Figure A1).

We evaluate the RDiT method for our constructed outcome variable. We assume the researcher knows the true start date and wishes to uncover the treatment effect β. We run the usual RDiT regressions, both using a polynomial approach and a local linear approach.[15] For the polynomial specifications, we follow Davis (2008) and use an eight-year window around the treatment start date. We show results using both a global polynomial for the entire sample period, and using separate polynomials for the pre- and post-periods.[16] As is standard in the air pollution policy evaluation literature, we control in all specifications for seasonality with month effects and day of week effects and for weather using cubic functions of minimum daily temperature, maximum daily temperature, rainfall, and snowfall.

Ambient air quality data can be highly variable (see e.g., Figure A1 in the Appendix), so

[14] We exclude potential start dates that do not have four years of data each for the pre- and post- periods. A full list of the treatment start dates is given in the Appendix.

[15] For the local linear specifications, we use a rectangular kernel and 30 days of data in the pre- and post-period—so the specification is an OLS regression with a linear time trend using only observations within one month of the treatment start date.

[16] The order of the polynomial is chosen by the Bayesian information criterion (BIC). The global polynomial specification allows up to order nine. When the polynomial is separated into a pre-period polynomial and post-period polynomial, we use a BIC selection over all nine possible combinations of polynomials of order one, two, and three.

controls are important for absorbing noise. A difficulty with a local linear specification using only a few weeks of observations is that controls can be difficult to include. If the treatment happens to begin on a Monday, a local linear specification needs to separately identify the "Monday effect" from the treatment effect of interest. We propose an alternative to the standard local linear specification. We use a two-step procedure, which we refer to hereafter as "augmented local linear." First, the impacts of weather and seasonality controls are estimated, and the residuals are saved, using an eight-year data window. Then, a local linear specification is estimated using just the residuals for dates that are within a narrow bandwidth, such as 30 days in the pre-period and 30 days in the post-period. The researcher can retrieve consistent estimates of standard errors by implementing a bootstrapping procedure that allows first-stage variance to be reflected in the second stage.

Additionally, to compare the RDiT approach using high frequency data to a simple pre/post analysis where high frequency data are not available, we collapse the data to monthly averages and run a regression that controls for the weather and seasonality variables as well as a linear trend. We run a separate regression for each monitor. For all results, we show the average and its standard deviation across the ten randomly selected start dates and 108 monitors.[17]

We also show Monte Carlo simulations for a dataset in which the counterfactual is simply an error term that follows a standard normal distribution (the realization of a Gaussian white noise process). The length of the observed data series is the same as for the air quality data, and we show results for 1000 iterations. We again construct a treatment effect beginning at the known start date: $y_{i,t} \equiv \beta \cdot \mathbb{1}\{t \geq t_{start}\} + \varepsilon_{i,t}$, where now $\varepsilon_{i,t} \sim N(0,1)$.

As can be seen in Table 2, the RDiT performs well when the treatment start date is precisely known by the researcher and the treatment effect is constant over time. All five columns show estimates close to the true treatment effect, although with varying degrees of

[17]We report the standard deviation of the estimated treatment effect across all monitors and start dates to give a sense of the variation in the estimated treatment effect. The standard error of the mean of the estimates would naturally be smaller.

Table 2: RD in Time Estimates for Simulated Treatment, $\beta = -0.2$

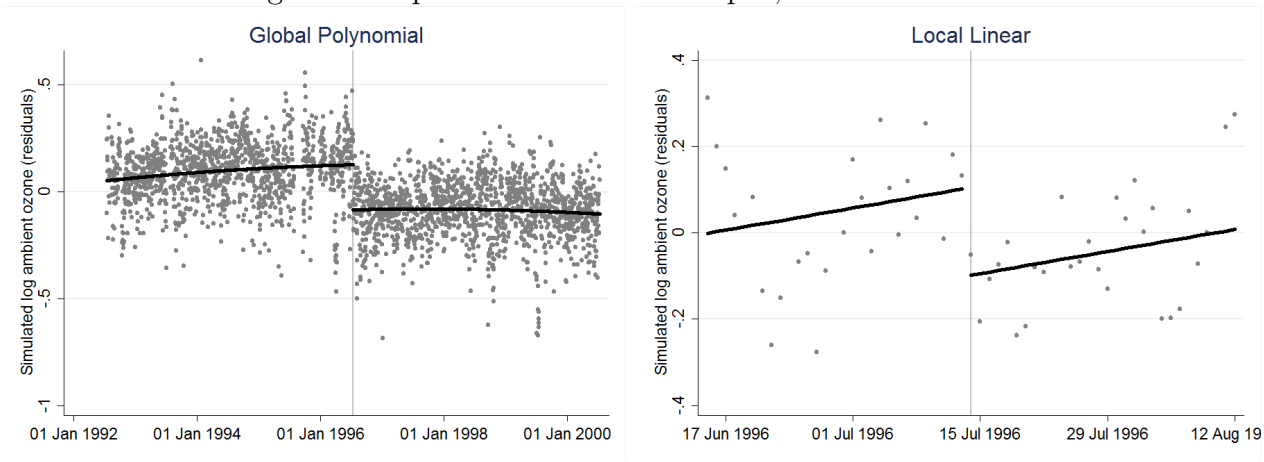
	(1) BIC-chosen Global Polynomial	(2) BIC-chosen Separate Polynomials	(3) Local Linear	(4) Augmented Local Linear	(5) Pre/Post, Monthly Observations
			A. Air Quality Data		
Estimated Beta	-0.20	-0.21	-0.19	-0.18	-0.19
	(0.11)	(0.11)	(0.29)	(0.26)	(0.10)
Observations	2822	2822	59	59	96
	(72)	(72)	(3.5)	(3.5)	(1.0)
Polynomial Order	4.0	1.7, 1.6	-	-	-
	(2.9)	(0.8), (0.8)			
			B. Simulated Data		
Estimated Beta	-0.20	-0.20	-0.19	-0.19	-0.19
	(0.07)	(0.08)	(0.51)	(0.51)	(0.08)
Observations	2921	2921	61	61	96
	-	-	-	-	-
Polynomial Order	1.0	1.0, 1.0	-	-	-
	(0.05)	(0.03), (0.06)			

Note: Each column of this table reports the mean treatment effect coefficient and number of observations across 2,080 regressions. The standard deviation of the estimate and of the number of observations is in parentheses. Panel A uses ambient air quality data with a simulated treatment effect, while Panel B uses white noise with a simulated treatment effect. For both panels, the true treatment effect is -0.2. The empirical specification varies across the columns; for details see text. Differences in the sample size across regressions arise because of missing data points in the air quality data series. Note the observation count in Column 4 is for the second estimation step; the first step uses the same number of observations as the polynomial specifications in Columns 1 and 2.

precision. The global polynomial results are more precise than the local linear, a standard finding in the RD literature. In Panel A (using air quality data), the augmented local linear, which strips out the effects of controls using a much longer data series, is less noisy than the traditional local linear, as intended. In Panel B, which uses noise data, there are no controls needed, so the local linear and augmented local linear results are identical. The pre/post results are similar to the polynomial and local linear results, indicating that little has been gained by the use of high-frequency controls.

Figure 1 plots these results at a representative monitor, where "representative" is defined as having an estimated beta close to the average estimated beta. For space considerations, only the global polynomial and local linear approaches are plotted. Figure A2 in the Appendix shows plots for the other specifications. The left-hand panel shows the global polynomial estimation; the right-hand panel shows a local linear specification with a bandwidth

Figure 1: Representative RDiT Graphs, No Confounders



Note: These figures plot regression discontinuity in time estimates for a single representative monitor, with separate panels for two different empirical specifications. Both panels shows residuals after controls have been removed (grey circles) and a line fitted to those residuals (black line). The left-hand panel uses a global polynomial approach, and the right-hand panel uses a local linear approach with 30 days of observations on either side of the threshold. Plots for additional specifications (separate pre/post polynomials; an augmented local linear approach; and a collapse to monthly averages) are provided in the Appendix.

of 30 days before and 30 days after the treatment date.[18]

Table 2, Figure 1, and Figure A2 demonstrate how the RDiT method (whether implemented with a polynomial control, a local linear specification, or our augmented local linear specification) recovers the causal effect of interest when the estimating equation is correctly specified. Next, we examine how different implementations of the RDiT method perform when the estimating equation is not correctly specified.

3 Sources of Potential Bias

3.1 Time-Varying Treatment Effects

The RDiT dataset frequently has little or no cross-sectional variation. A study might analyze outcomes relating to a single cross-sectional unit, or evaluate separate RDiT specifications for each individual cross-sectional unit. In such cases, the researcher has two choices for increasing the sample size: (1) increasing the frequency of data ("infill" or "fixed-domain"

[18]While the BIC-chosen polynomial order happens to be low for this monitor, on average the BIC-chosen order is 4 for the global specification (see Table 2).

asymptotics) or (2) expanding the time window ("increasing domain" asymptotics). In practice, RDiT practitioners have done both by using, for instance, daily data across several years. Both approaches have limitations. Increasing the frequency of the data will not add much power if the data are serially correlated. Expanding the time window increases the probability of bias, as data are added far from the threshold. This problem is not unique to RDiT; the RD literature has documented the bias/precision trade-off as the sample size increases away from the threshold. However, the problem could be more severe for an RDiT application because there is frequently only one cross-sectional unit.

Consider a time series of daily pollution levels, which will exhibit substantial noise from seasonality and weather—as mentioned in Section 1.2, a common RDiT application is the evaluation of a policy impacting ambient air quality. To absorb the seasonality and weather effects, the researcher will want to include several years of data, necessitating inclusion of observations far from the threshold date when the treatment begins. Two assumptions are then needed. First, the model must be correctly specified. Any potential confounders such as other policy changes, weather events, or changes in pollution from other sources must be either controlled for directly or be sufficiently well-absorbed by the global polynomial approximation. Of the papers we cite, all but one include controls (such as weather) in addition to time trends. Second, the researcher must correctly specify the treatment effect. In particular, she must take a stand on whether the treatment effect is smooth and constant throughout the post-period, or whether it varies (and, if so, how). Importantly, these two assumptions may interact: the polynomial must be specified such that it is uncorrelated with any unobserved (or mis-specified) variation in the treatment effect.

Of particular interest in environmental applications are time-varying treatment effects. A time-varying treatment effect violates the assumption (untestable in the RDiT framework) that the researcher has correctly specified the treatment variable. Again, this potential problem is not unique to RDiT—researchers using cross-sectional RD must address the possibility of heterogeneous treatment effects that vary with the running variable. However,

the long sample window frequently used in RDiT (to aid with power) make this problem more relevant than in many cross-sectional RD settings. Recall that the majority of the papers we examine use a sample window of multiple years.

All but three of the RDiT papers we examine assume a constant treatment effect in their main specification. Some discuss short-run versus long-run effects qualitatively, and a few examine the possibility of time-varying treatment effects using difference-in-differences. The RDiT method by itself does not allow for direct tests of time-varying treatment; either the researcher must assume how the treatment effect evolves over the sample window, or control units are required.

To demonstrate the impact of a time-varying treatment effect, we next return to the Monte Carlo simulations. We assume the treatment effect is not constant, and that this is not known to the researcher. Suppose the treatment effect lasts a given number of days $t_{length} \in [1, n]$, where n is the number of days in the post-period. After t_{length} days, the true treatment effect goes immediately to zero.[19] Assume that, since the researcher does not know the treatment effect varies over time, she models a standard RDiT with a single dummy for the entire post-treatment period (recall she accurately knows the start date). For this framework, we again estimate a global polynomial with BIC-chosen order and a local linear regression.

Table 3 shows results for a treatment length of one year.[20] The results are sensitive to the specification. The local linear specification performs well, since the treatment effect is constant within the window studied. However, the BIC-chosen global polynomial gives an estimate that could be either too large or too small. In Panel A, the estimated treatment effect is substantially larger than the true beta, while in Panel B the estimated effect is somewhat smaller than the true effect. The estimate in Panel B would accord with intuition that the specification is giving an estimate between the true short-run (-0.2) and long-run

[19] In the Appendix, we consider a smooth decay process, which may accord with adaptive behavior in general equilibrium.

[20] In the Appendix, we show results for a treatment length of one month.

Table 3: Treatment Sharply Decays after One Year, $\beta_{initial} = -0.2$ and $\beta_{long-run} = 0$

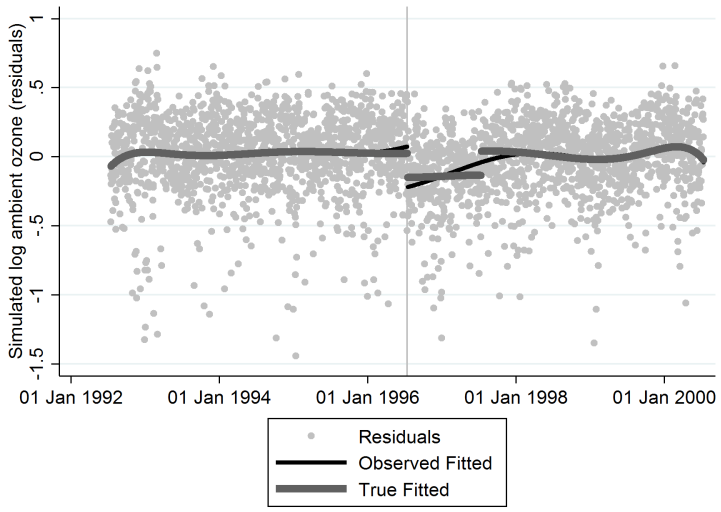
	BIC-chosen Global Polynomial	BIC-chosen Separate Polynomials	Local Linear	Augmented Local Linear	Pre/Post, Monthly Observations
			A. Air Quality Data		
Estimated Beta	-0.27	-0.29	-0.19	-0.18	-0.15
	(0.11)	(0.12)	(0.29)	(0.26)	(0.11)
Observations	2822	2822	59	59	96
	(72)	(72)	(3.5)	(3.5)	(1.0)
Polynomial Order	5.9	1.7, 2.1	-	-	-
	(2.3)	(0.8), (0.6)			
			B. Simulated Data		
Estimated Beta	-0.16	-0.18	-0.19	-0.19	-0.15
	(0.08)	(0.10)	(0.51)	(0.51)	(0.08)
Observations	2921	2921	61	61	96
	-	-	-	-	-
Polynomial Order	1.1	1.0, 1.1	-	-	-
	(0.33)	(0.03), (0.32)			

Note: Each column of this table reports the mean treatment effect coefficient and number of observations across 2,080 regressions—see Table 2 for details. In contrast to Table 2, the true treatment effect of -0.2 lasts only one year, then sharply drops to 0.

(0.0) effects. However, this intuition does not hold in Panel A. When the treatment lasts for one year with the simulated air quality data, the BIC-chosen polynomial gives an estimate that is larger (-0.27) than the true immediate effect (-0.2). One explanation is that the polynomial in time adjusts its shape. For the constant-treatment results in Table 2, the average polynomial order selected is 4. In contrast, for the time-varying treatment results in Table 3, the BIC-selected order is 6. Related concerns regarding the use of high-order polynomial controls have been raised by Gelman and Imbens (2017).

To see the overfitting of the polynomial, Figure 2 shows a monitor for which the estimated treatment effect (discontinuity in the thin black line) is larger than the true effect (discontinuity in the thick grey line). Because of the noise in the series, this confounder is not apparent in the residuals to which the polynomial is fitted. While around half of the RDiT papers we examine mention the possibility of a difference between short-run and long-run treatment effects, we have not seen discussion of the possibility of bias (from, e.g., the polynomial overfitting). However, it appears possible for the estimate to exhibit bias under a time-varying treatment effect, rather than approximating *either* the short-run or

Figure 2: Global Polynomial Can Overfit when Treatment Sharply Decays



Note: This figure plots the RDiT estimate for a representative monitor, with an estimated treatment effect similar to the average in Panel A of Table 3. The grey circles plot the residuals (after removing covariates) of log ambient air quality with a simulated treatment effect of -0.2 lasting one year. The thick grey line shows the polynomial a researcher would fit when (correctly) modeling the temporary nature of the treatment. The thin black line shows the polynomial a researcher would fit when (incorrectly) modeling the treatment effect as constant. Four outliers have been used for estimation but dropped from the plot.

the long-run effect.

As can be seen by comparing Tables 3 and A2, the estimated effect also varies by the treatment length. To further explore the impact of time-varying treatment effects, we show plots in the Appendix of the estimated effect for varying treatment lengths and polynomial orders (Figures A3, A4, and A5). We also explore a smooth decay process (Table A3). With the global polynomial approach, we estimate effects both substantially smaller than and larger than the true initial effect, depending on the decay process. Since the long-run effect is zero, estimates larger (in absolute value) than the short-run effect are not bounded by the initial and permanent effects. What these simulations make clear is that given a time-varying treatment effect, the global polynomial RDiT estimate may not simply be a weighted average of the short-run and long-run effects, nor is it simply the initial impact local (in time) to the threshold.

Finally, we have motivated these Monte Carlo simulations with the presence of time-

varying treatment effects. However, the same Monte Carlos could be used to understand the impact of unobservables correlated with time. For instance, a setting with a constant treatment effect combined with a sharply discontinuous unobservable of equal magnitude (and opposite sign) is observationally equivalent to the sharply discontinuous time-varying treatment effect modeled above. That is, it is easy to imagine settings where unobservables lead to the polynomial overfitting, even with a constant treatment effect.

3.2 Autoregression

The second potential pitfall we examine in RDiT implementation relates to its use of time-series data, which are likely to exhibit serial dependence. The first implication of this serial dependence relates to inference. If there is serial dependence in the residuals, standard errors must account for it. The existing RDiT literature has generally addressed this by using clustered standard errors. The second implication is that autoregression in the dependent variable (even after accounting for serial correlation in the exogenous variables and the residuals) will impact estimation of short-run versus long-run effects. Local air pollution, for instance, can dissipate in minutes or days or weeks, depending on the pollutant and local atmospheric conditions (MacDonell et al., 2014).

This slow dissipation introduces dynamic effects not generally considered in the RDiT literature (the two exceptions of which we are aware are Chen and Whalley (2012) and Lang and Siler (2013)). Consider the power plant policy described above, and suppose that the regulation will decrease emissions by β. The researcher observes daily pollution levels in the city for several years before and after the policy change. Assume that plant owners comply on the first date of the policy and there are no other confounders. Finally, suppose that the pollutant of interest dissipates at a rate of $1 - \alpha$ each day, implying that α remains from one day to the next. Then on the first day of the new policy, the treatment effect on air pollution levels will be a decrease of β. On the second day, however, the treatment effect will be a combination of lower emissions plus lower pollution left over from the previous day:

$\beta + \beta \cdot \alpha$. The long-run effect will be $\frac{\beta}{1-\alpha}$.

The magnitude of the estimated treatment effect will depend on the specification, and on the extent to which identification is being achieved from a discontinuity (i.e. as the bandwidth shrinks towards the threshold) or from time-series variation outside of the neighborhood of the threshold. If the study included only the day immediately before and the day immediately after the policy change, the estimated effect would be the short-run effect. With a longer window, it is not clear what will be recovered. With a longer window, the researcher could include the lagged dependent variable in estimation, then recover both the short-run and long-run effects if the regression is properly specified. However, if the researcher omits the lagged dependent variable, bias arises from a similar source as that of the time-varying treatment effects problem described above. Overfitting of the global polynomial may arise. The degree to which these dynamics will matter in practice depends on how large the true autoregressive coefficient is, and on researcher choices about bandwidth and specification. High-frequency data, while allowing for more power, may be more likely to exhibit qualitatively important autoregression. The median frequency of the data used in the papers we examine is daily, and eleven of the fourteen papers use daily or hourly data.

To explore how important autoregression might be in the air quality setting, we estimate the autoregressive AR(1) parameters for six pollutants at air quality monitors in the U.S., using daily data and controlling flexibly for seasonality and weather (Tables A5 and A6).[21] Overall, we obtain estimates of the AR(1) parameter of 0.3 to 0.5, varying by pollutant.[22]

To examine how RDiT behaves in the presence of autoregression, we simulate an AR(1) process with varying degrees of dependence. The process is given by the following equation:

$$y_{i,t} \equiv \alpha \cdot y_{i,t-1} + \beta \cdot \mathbb{1}\{t \geq t_{start}\} + \varepsilon_{i,t}$$

where the error component is standard normal. We fix $\beta = -0.2$ but allow α to vary from

[21] We use gridded daily weather data from Roberts and Schlenker (2009).

[22] It is possible some of this estimated autoregression is driven by serially correlated unobservables; see Appendix.

Table 4: Dropping AR(1) Term, $\beta_{initial} = -0.2$ and $\beta_{long-run} = \frac{-0.2}{1-\alpha}$

			BIC-chosen Global Polynomial	BIC-chosen Separate Polynomials	Local Linear	Augmented Local Linear	Pre/Post, Monthly Observations
AR(1) parameter α	Implied long-run treatment effect						
0.1	-0.22	Estimated beta	-0.23	-0.22	-0.24	-0.24	-0.21
			(0.08)	(0.09)	(0.55)	(0.55)	(0.08)
0.3	-0.29	Estimated beta	-0.29	-0.29	-0.30	-0.30	-0.28
			(0.11)	(0.12)	(0.68)	(0.68)	(0.10)
0.5	-0.40	Estimated beta	-0.41	-0.40	-0.38	-0.38	-0.39
			(0.16)	(0.20)	(0.90)	(0.90)	(0.15)
0.7	-0.67	Estimated beta	-0.68	-0.67	-0.52	-0.52	-0.65
			(0.38)	(0.42)	(1.34)	(1.34)	(0.24)
0.9	-2.00	Estimated beta	-1.91	-1.90	-0.65	-0.65	-1.96
			(1.37)	(1.38)	(2.40)	(2.40)	(0.73)

Note: Each column of this table reports the mean treatment effect coefficient and number of observations across regressions using 1,000 simulated datasets. The true data-generating process follows an AR(1) process with a true initial treatment effect of -0.2, and therefore a long-run treatment effect of $-0.2/(1-\alpha)$. The estimation procedure does not include a lagged dependent variable; as such, the table shows the impact of incorrectly ignoring a dynamic process.

0.1 to 0.9 in increments of 0.2. Table 4 shows the estimated treatment effect when the autoregressive component is excluded from the regression.[23]

In general, the estimated effect approximates the long-run, rather than the short-run, effect—perhaps because the treatment effect approaches the long-run impact after only a few days. In this context, the long-run treatment effect may be more policy-relevant. The policy-maker is likely to be more interested in the *stock* of pollutants to which the population is exposed than to the *flow* of pollutants on any given day, so the retrieved effect may be of more policy relevance than the immediate effect. However, this aspect of the estimates in Table 4 highlights that identification of the treatment effect in an RDiT setting is not achieved narrowly at the threshold, but rather relies on observations away from the threshold.

To uncover the true dynamic process, the researcher would need to include the lagged dependent variable in her regression. These estimates are provided in the Appendix (Table A7). As expected, the BIC-chosen global polynomial gives accurate estimates $\hat{\beta}$ and $\hat{\alpha}$, allowing the researcher to retrieve both the true short-run and long-run effects. The local

[23]For these data, note the specifications in the local linear and augmented local linear columns are in practice identical, since there are no controls to include the first step of the augmented local linear. In contrast, if an AR(1) term is included in the first step, the two estimation procedures will not be identical; see Appendix.

linear also performs well, although with noise (as expected) and a downward bias on $\hat{\alpha}$ (typical for autoregression estimation in small samples). Note that when the data are collapsed to the monthly level, these two effects are not separable and the estimated treatment effect is closer to the long-run effect.

In this section we have explored how an autoregressive process ought to be considered by the researcher.[24] However, we have not yet discussed a closely-related phenomenon: serial correlation in ε_{it}. The concentration of local pollutants may exhibit time dependence either due to its own natural decay rate, or due to the persistence of factors that augment or dissipate it. For example, ozone is created as a function of ambient air temperature, which exhibits its own time dependence. The researcher needs to think carefully about serial dependence in both ε_{it} and possibly y_{it} (via AR1). If there is only serial dependence in ε_{it}, and not autoregression, then including a lagged-dependent variable would be a mis-specification. In short, time dependence in the data generating process may occur via different channels. When deploying the RDiT approach, decisions that the researcher makes may affect both the probability that the retrieved treatment effect estimate is biased, and also its proper interpretation.

3.3 Sorting and Anticipation Effects

Finally, in a cross-sectional RD, a density test (e.g. McCrary (2008)) is a key check for sorting behavior. It is generally used to rule out selection into or out of treatment, thus making it unnecessary to further control for it. When time is the running variable, however, it is generally not possible to test for such behavior around the threshold. While the researcher can check for discontinuities in other covariates at the threshold, and for discontinuities in the outcome variable at other thresholds, the researcher cannot check for discontinuities in

[24] Of course, the dispersion of pollutants in the atmosphere cannot be completely modeled by the inclusion of a lagged dependent variable. The dispersion is a function of weather (for instance, wind speeds and the existence of a thermal inversion) and the properties of the specific chemical compound of interest. What these simplified AR(1) simulations are meant to show is that dynamics in the endogenous variable introduce a type of time-varying treatment effect, which the researcher should address explicitly.

the conditional density of the running variable. That the density of the running variable (time) is uniform renders such tests logically irrelevant.

Consider the hypothetical power plant example. In this RD, the outcome variable would be air quality in a city, and the unit of observation would be a daily pollution monitor reading in that city. One could imagine a type of sorting in which power plants change their behavior to avoid the policy or to preemptively comply. For instance, some of the plants could decide to install the emissions control device early. Thus the plant would be treated in the pre-period—this is analogous to the test scores example of a canonical RD in which an untreated student successfully changes her behavior in order to be treated. However, in the RDiT case there is an important difference relating to the researcher's ability to identify this behavior. With a cross-sectional RD, one can test for these effects (via the McCrary test, for example), but in RDiT with time-series data, it is untestable.

As a result, estimates retrieved from RDiT are of a *compound* effect: the causal treatment effect of interest (i.e. what we try to retrieve from, say, a randomized controlled trial) *and* any unobserved sorting/anticipation/adaptation/avoidance effects that may exist but cannot be tested for. The extent to which the results should be interpreted solely as the causal treatment effect of interest depend on the researcher's ability to make a compelling case that the sorting effects are not present. Furthermore, the absence of these effects is a necessary but insufficient condition for identification, as we discussed above.

4 Recommendations and Conclusion

We have demonstrated that the RDiT differs in conceptualization and in implementation from the cross-sectional RD. Table 5 summarizes pitfalls that the RDiT researcher may encounter. Empirical researchers wishing to use RDiT ought to expose their assumptions to many opportunities to fail. Below, we offer a checklist along the lines of Lee and Lemieux (2010) applied to the context of RDiT. Recall that we have identified three main features of

Table 5: Summary of Concerns for RDiT Practitioners

Concern	Intuition
Unobservables correlated with time	Covariates are more important than in many cross-sectional RDs. Even with covariates included, bias is possible—for instance, a global polynomial control may overfit.
Time-varying treatment effects	Mis-specification of the treatment effect will lead to bias, particularly when using a global polynomial control.
Autoregressive properties	Short- and long-run treatment effects will differ. Unless the nature of autoregression is known, estimates could approximate the short- or long-run effects, or neither.
Selection and strategic behavior	The running variable follows a uniform distribution across the discontinuity. It is thus impossible to test for sorting/selection around the threshold (e.g. the McCrary test).

the RDiT setting that may induce bias: time-varying treatment effects, autoregression, and selection into or out of treatment. So, in addition to the standard cross-sectional RD diagnostics, researchers using RDiT should make every effort to test for and eliminate potential bias from those sources. Our recommendations focus on these issues. While there is some overlap with the suggestions from Lee and Lemieux (2010), the recommendations serve a different purpose in the context of time-series data. Additional suggestions from Lee and Lemieux (2010) should also be deployed whenever relevant.

1. Plot the raw data or residuals after removing covariates (e.g., weather). Overlay the various polynomial and local linear controls. If results differ across alternate time trends, it may be a sign of time-varying treatment effects.

2. Present several robustness checks. When using a global polynomial, recognize the possibility of overfitting. Show robustness to polynomial order and robustness to alternative local linear bandwidths.[25]

3. Placebo tests. Estimate parallel RDs on nearby geographical areas that were not subject to the treatment, and using other dates.

4. Plot a parallel RD estimated on control variables (e.g. weather, economic activity) to demonstrate continuity.

[25]In addition to multiple columns for different regression specifications, a plot of estimated treatment effects across bandwidths, as in Figure 18 of Lee and Lemieux (2010), is informative.

5. Estimate a "donut" RD (removing observations near the threshold; see Barreca et al. (2011)) to mitigate concerns about short-run selection/anticipation/avoidance effects.

6. Test for the presence of autoregression in the outcome variable of interest using pre-intervention data. If it is present, consider including the lagged dependent variable as a regressor, and consult the time-series literature for additional options.

7. Consider deploying our "augmented local linear" methodology to increase power of the local linear specification. This two-step procedure uses the full sample to identify important regressors (e.g. temperature and various fixed effects), then estimates the conditioned second stage on a smaller sample bandwidth.[26] This procedure eliminates the need for a global polynomial and the overfitting concerns that accompany its use.

Many of these strategies have been deployed by some of the papers we cite, although not in a comprehensive way. Table 6 displays the proportion of the papers using these strategies. It is worth noting that frequently the robustness check is mentioned but results are not shown; we include these instances in our counts. Passing these diagnostic tests is necessary but insufficient for identification, so consumers of RDiT should evaluate the evidence as they would any quasi-experimental paper. There is no set of tests that, if passed, indicate that the desired causal effect has been retrieved. Instead, one must assess the preponderance of evidence, keeping in mind that certain features in the data-generating process will lead to different potential biases (or require stronger assumptions).

In this paper we have articulated reasons why using time as the running variable in a regression discontinuity design, an increasingly popular empirical strategy, is worth closer examination and a deeper understanding. RDiT requires assumptions for identification that are often strong and inherently untestable. When $N = 1$, RDiT approaches should be thought of as being more like event studies or pre-post analyses than like randomized controlled trials. Papers should state clearly the extent to which identification relies upon observations

[26] Low power may be unavoidable in the augmented local linear specification. It is important to recognize that this is simply a feature of an RDiT that has little or no cross-sectional variation in the treatment status.

Table 6: Robustness Checks Used in the Literature

Check	Proportion of Papers
Plot of data	0.79
Robustness: bandwidth or polynomial order	0.79
Placebo	0.29
RD on continuous controls (e.g. weather, economic activity)	0.36
Donut RD	0.14
Test for autoregression	0.14
Augmented local linear	0.14

Note: We count the proportion of papers conducting each check, of the fourteen papers listed in Table 1. We include checks mentioned by the authors but not shown. It is worth noting that while 79 percent of the papers conduct some sort of robustness check on either the bandwidth or the polynomial order, only 43 percent conduct checks on both the bandwidth and the polynomial order. Also, several papers conduct sensitivity to a polynomial of e.g. order 7 versus 8 but not a lower order; both higher-order specifications may overfit.

far away from the threshold, and should explore whatever options are available to mitigate concerns about autoregression and dynamic behavior. While regression discontinuity designs have been a tremendously valuable addition to the suite of identification strategies, we must also recognize that the RDiT framework strays in important ways from the experimental ideal.

References

Anderson, Michael L, "Subways, Strikes, and Slowdowns: The Impacts of Public Transit on Traffic Congestion," *American Economic Review*, 2014, *104* (9), 2763–2796.

Auffhammer, Maximilian and Ryan Kellogg, "Clearing the Air? The Effects of Gasoline Content Regulation on Air Quality," *American Economic Review*, 2011, *101* (6), 2687–2722.

Barreca, Alan, Melanie Guldi, Jason Lindo, and Glen Waddell, "Saving Babies? Revisiting the Effect of Very Low Birth Weight Classification," *Quarterly Journal of Economics*, 2011, *126* (4), 2117–2123.

Bento, Antonio, Daniel Kaffine, Kevin Roth, and Matthew Zaragoza-Watkins, "The Effects of Regulation in the Presence of Multiple Unpriced Externalities: Evidence from the Transportation Sector," *American Economic Journal: Economic Policy*, 2014, *6* (3), 1–29.

Burger, Nicholas E, Daniel T Kaffine, and Bo Yu, "Did California's Hand-Held Cell Phone Ban Reduce Accidents?," *Transportation Research Part A*, 2014, *66* (1), 162–172.

Busse, Meghan, Jorge Silva-Risso, and Florian Zettelmeyer, "$1,000 Cash Back: The Pass-Through of Auto Manufacturer Promotions," *American Economic Review*, 2006, *96* (4), 1253–1270.

Busse, Meghan R, Duncan I Simester, and Florian Zettelmeyer, ""The Best Price You'll Ever Get": The 2005 Employee Discount Pricing Promotions in the U.S. Automobile Industry," *Marketing Science*, 2010, *29* (2), 268–290.

Chen, Xinlei, George John, Julie M Hays, Arthur V Hill, and Susan E Geurs, "Learning from a Service Guarantee Quasi-Experiment," *Marketing Research*, 2009, *46* (5), 584–596.

Chen, Yihsu and Alexander Whalley, "Green Infrastructure: The Effects of Urban Rail Transit on Air Quality," *American Economic Journal: Economic Policy*, 2012, *4* (1).

Dasgupta, Susmita, Benoit Laplante, and Nlandu Mamingi, "Pollution and Capital Markets in Developing Countries," *Journal of Environmental Economics and Management*, 2001, *42*, 310–335.

Davis, Lucas W, "The Effect of Driving Restrictions on Air Quality in Mexico City," *Journal of Political Economy*, 2008, *116* (1), 38–81.

_ **and Matthew E Kahn**, "International Trade in Used Vehicles: The Environmental Consequences of NAFTA," *American Economic Journal: Economic Policy*, 2010, *2* (4), 58–82.

DePaola, Maria, Vincenzo Scoppa, and Mariatiziana Falcone, "The Deterrent Effects of the Penalty Points System for Driving Offences: A Regression Discontinuity Approach," *Empirical Economics*, 2013, *45*, 965–985.

Gallego, Francisco, Juan-Pablo Montero, and Christian Salas, "The Effect of Transport Policies on Car Use: Evidence from Latin American Cities," *Journal of Public Economics*, 2013, *107*, 47–62.

Gelman, Andrew and Guido Imbens, "Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs," *Journal of Business and Economic Statistics*, 2017.

Grainger, Corbett and Christopher Costello, "Capitalizing Property Rights Insecurity in Natural Resource Assets," *Journal of Environmental Economics and Management*, 2014, *67*, 224–240.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw, "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, 2001, *69* (1), 201–209.

Hamilton, James T, "Pollution as News: Media and Stock Market Reactions to the Toxics Release Inventory Data," *Journal of Environmental Economics and Management*, 1995, *28*, 98–113.

Imbens, Guido W and Thomas Lemieux, "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics*, 2008, *142*, 615–635.

Ito, Koichiro, "Asymmetric Incentives in Subsidies: Evidence from a Large-Scale Electricity Rebate Program," *American Economic Journal: Economic Policy*, 2015, *7* (3), 209–237.

Jacob, Robin, Pei Zhu, Marie-Andree Somers, and Howard Bloom, "A Practical Guide to Regression Discontinuity," *MDRC Working Paper*, 2012.

Konar, Shameek and Mark A. Cohen, "Information As Regulation: The Effect of Community Right to Know Laws on Toxic Emissions," *Journal of Environmental Economics and Management*, 1997, *32*, 109–124.

Lang, Corey and Matthew Siler, "Engineering Estimates versus Impact Evaluation of Energy Efficiency Projects: Regression Discontinuity Evidence from a Case Study," *Energy Policy*, 2013, *61*, 360–370.

Lee, David S, "Randomized Experiments from Non-Random Selection in U.S. House Elections," *Journal of Econometrics*, 2008, *142*, 675–697.

Lee, Davis S and Thomas Lemieux, "Regression Discontinuity Designs in Economics," *Journal of Economic Literature*, 2010, *48*, 281–355.

MacDonell, Margaret, Michelle Raymond, David Wyker, Molly Finster, Young-Soo Chang, Thomas Raymond, Bianca Temple, Marcienne Scofield, Dena Vallano, Emily Snyder, and Ron. Williams, "Mobile Sensors and Applications for Air Pollutants," Accessed from https://cfpub.epa.gov/si/si_public_record_report.cfm?dirEntryId=273979 2014. U.S.

Environmental Protection Agency, Washington, DC, EPA/600/R-14/051 (NTIS PB2014 105955).

McCrary, Justin, "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test," *Journal of Econometrics*, 2008, *142*, 698–714.

Roberts, Wolfram and Michael J Schlenker, "Nonlinear Temperature Effects Indicate Severe Damages to U.S. Crop Yields under Climate Change," *Proceedings of the National Academy of Sciences*, 2009, *106* (37), 15594–15598.

Rubin, Donald B., "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies," *Journal of Educational Psychology*, 1974, *66* (5), 688–701.

Shadish, William R, Thomas D Cook, and Donald T Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Boston: Houghton Mifflin Company, 2002.

Splawa-Neyman, Jerzy, "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9," *Statistical Science*, 1923 [1990], *5* (4), 465–472. Trans. D. M. Dabrowska and T. P. Speed.

Online Appendix: Regression Discontinuity in Time

Catherine Hausman and David S. Rapson

In this Appendix, we provide additional tables and figures.

Figure A1 shows the air quality data used in the Monte Carlo simulations for a representative monitor. The true logged ambient ozone concentration (daily maximum, in parts per million) is shown in grey. The impact of the simulated treatment effect is shown in black, for the post-period. Table A1 provides summary statistics for the air quality data used in the Monte Carlo simulations. The data are from Auffhammer and Kellogg (2011).

As described in the text, we randomly select ten treatment start dates for the Monte Carlo simulations using air quality data. The dates selected are as follows:

09 Aug 1993
15 Jul 1994
26 Oct 1994
21 Jan 1995
05 Oct 1995
14 Jul 1996
17 Jul 1997
15 Aug 1999
27 Oct 2000
04 Dec 2002

Figure A2 shows RDiT plots for the specifications used throughout the paper. All six panels shows residuals after controls have been removed (grey circles) and a line fitted to those residuals (black line). Panel A uses a global polynomial approach and Panel B uses separate polynomials in the pre- and post-periods. Panel C uses a local linear approach with 30 days of observations on either side of the threshold. Panels D and E use a two-step augmented local linear approach, where controls are estimated and residuals saved using the sample as a whole (Panel D) while the treatment effect is estimated with just 30 days of observations on either side of the threshold (Panel E). In Panel F, observations are collapsed to monthly averages before the regression is estimated. All regressions include weather and seasonality controls.

The next set of figures and tables examine results under time-varying treatment effects. The main text shows (Table 3) the average estimated treatment effect when the treatment sharply decays after one year, for five different empirical specifications. In this Appendix, we examine how this estimate varies with the length of the treatment. Table A2 is identical to Table 3, but for a sharp decay of one month, rather than one year. As can be seen here, whether the estimated treatment effect is too small or too large depends on both the true treatment length and the specification chosen by the researcher.

Next, we focus on the global polynomial specification shown in Column 1 of Table 3. Figure A3 shows the mean estimated treatment effect for different treatment lengths, when the global polynomial is of order 1. For comparison, it also shows the true immediate effect of -0.2 and the true long-run effect of 0. The estimated effect is close to the true effect for a long-lasting treatment (one that lasts the entire post period). For a very short-lived treatment, the estimated effect is bounded by the short-run and long-run effects. However, for a range of treatment lengths, the estimated effect is not bounded by the short-run and long-run effects; that is, it is not a weighted average of the treatment effect across time. Rather, the estimate is larger (in absolute terms) than either the short-run or long-run effect.

Next, Figure A4 shows a similar plot for 9 different polynomial orders. The estimate depends on the polynomial order, as can be seen in this figure. Finally, Figure A5 shows the estimate for the BIC-chosen polynomial.

The assumption that the treatment sharply decays may be unrealistic. We next consider instead a smooth decay of the treatment effect, which one might expect with, for instance, adaption behavior in a general equilibrium framework. To approximate this smooth decay, we use the generalized logistic function to represent an S-shaped decay path. (We do not use the simpler exponential decay function, because it initially decays sharply, and here we

aim to demonstrate how the RDiT behaves under smooth decay.) This function is given by:

$$A + \frac{K - A}{(1 + Q \cdot \exp(-B * (date - M)))^{1/v}} \tag{1}$$

We fix A, the lower asymptote, at 0; K, the upper asymptote, at 1; v at 0.5 and Q at 0.01. We vary the parameters B and M, which together affect how quickly the treatment decays. Figure A6 shows four representative combinations of B and M. In simulations, we allow the B parameter to vary from -0.005 to -0.015, and the M parameter to vary from 0 to 1000. We only include decay functions for which the true treatment value has decayed by at least 90 percent by the end date, for ease of comparison with the sharp decay parameterization. This leaves us with a total of 23 parameter combinations.

Table A3 shows the mean estimated treatment effect for the different parameterizations of the smooth decay function, when the researcher uses a global polynomial framework with a BIC-chosen polynomial. The estimate depends on the shape of the decay function, with some estimates lying between the true short-run and long-run effects, and with other estimates larger (in absolute value) than either the short-run or long-run effect.

As another form of time-varying treatment, we next consider a smooth phase-in of the treatment. Here we assume that the treatment ramps up linearly for a given period, then is constant for the rest of the post-period. Again, the researcher does not know that the treatment effect varies over time, and she models a standard RDiT with a post-treatment dummy. Table A4 shows results for a seven-day linear ramp-up. Here, the local linear regression estimates something approximating the short-run effect, while the global polynomial estimates are closer to the long-run effect, although this may be specific to the parameterization and data we use.

Finally, we show additional results relating to the autoregressive processes described in the main text. Table A5 shows estimated autoregressive parameters for six pollutants in the U.S. For each pollutant, we obtain daily monitor readings for 2004-2005 from the EPA

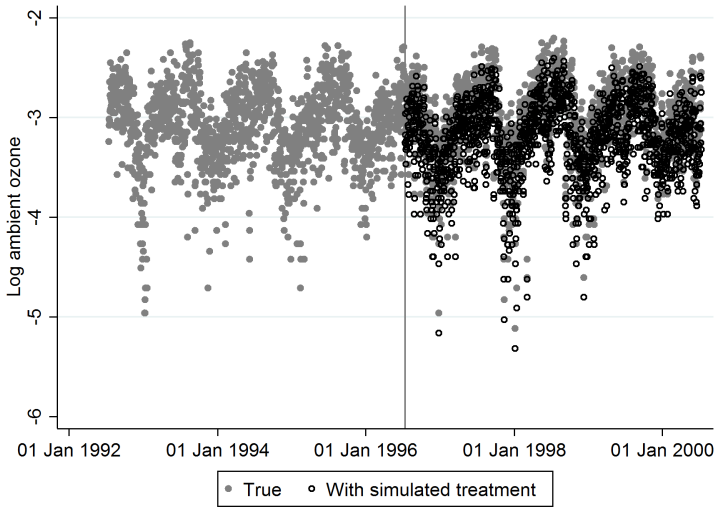
(https://aqs.epa.gov/aqsweb/documents/data_mart_welcome.html). We obtain temperature and windspeed data from the same source. For each monitor-pollutant pair, we regress the daily pollution level on the lagged pollution level and a set of weather and seasonality controls. These include a cubic function of the daily mean temperature, a cubic function of windspeed, the interaction of mean temperature and windspeed, day of week effects, month effects, and a cubic time trend. Standard errors are clustered by sample week. Overall, we obtain estimates of the AR(1) parameter of 0.3 to 0.5, varying by pollutant.[27]

Unfortunately, the EPA data do not include precipitation. To verify that results are not sensitive to the inclusion of precipitation, we match the air quality data to gridded daily weather data from Roberts and Schlenker (2009). We limit the number of states used for this sample as the matching procedure is computationally burdensome. We use the six states with the greatest number of PM2.5 monitors: California, Florida, North Carolina, Ohio, Pennsylvania, and Texas. We again include windspeed from the EPA as a control. Where windspeed is missing, we use a daily state-level average. Controls include a cubic function of the daily mean temperature, a cubic function of windspeed, a cubic function of precipitation, the interaction of mean temperature and windspeed, the interaction of temperature and precipitation, day of week effects, month effects, and a cubic time trend. Standard errors are clustered by sample week. Results, shown in Table A6, are very similar to the results presented in the previous table.

While the main text (Table 4) showed the impact of incorrectly dropping an autoregressive term, Table A7 shows estimation results when the AR(1) term is included. In general, this performs better (as expected) than dropping the AR(1) term, since both short-run and long-run effects can be calculated. The local linear (Column 3) gives estimates of the autoregressive term that are too small, as is common with short time series, but the augmented local linear (Column 4) performs better.

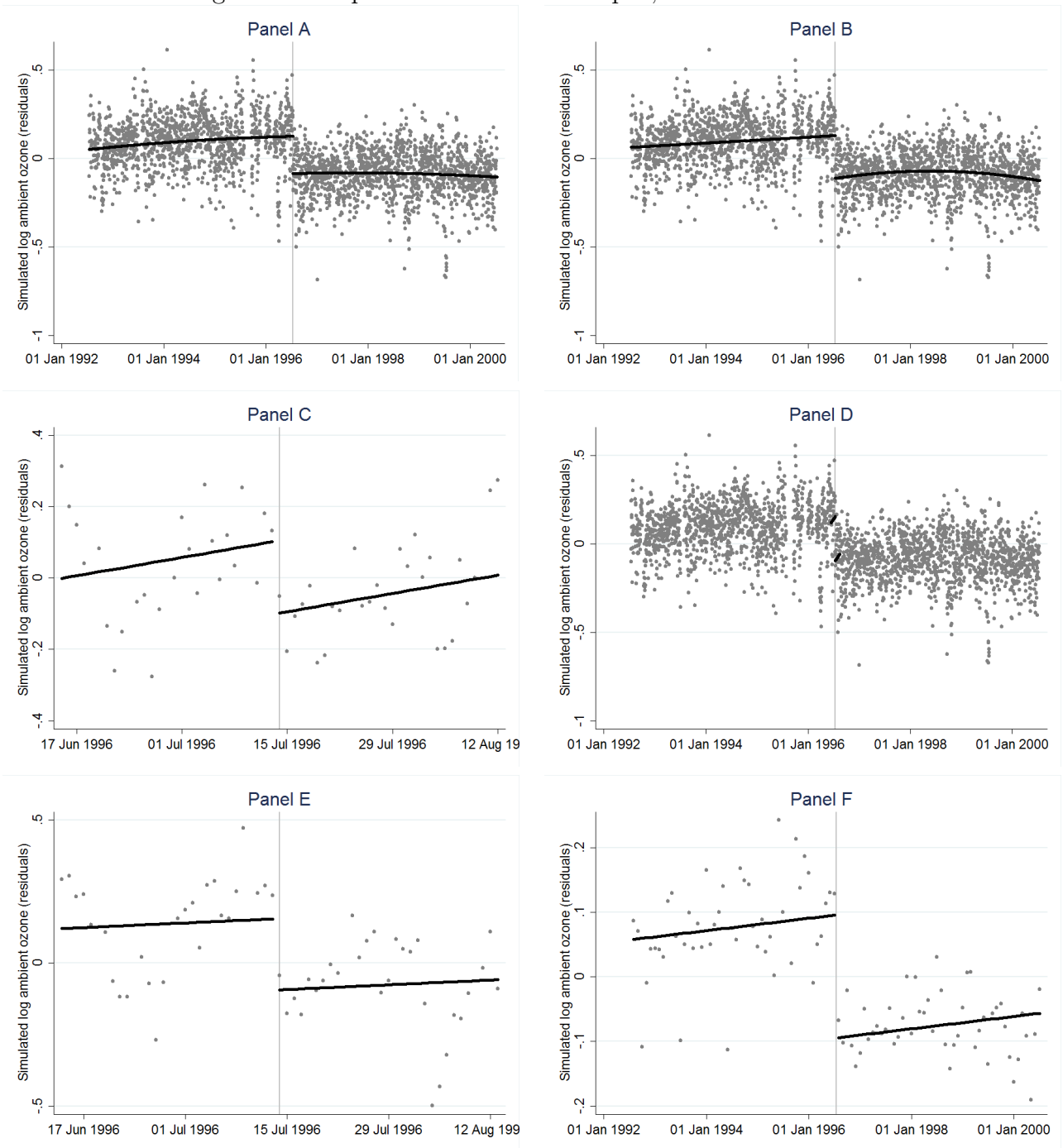
[27]The coefficient in the ozone equation is surprising, given ozone's short half-life; it is possible that serially correlated unobservables are driving that estimate.

Figure A1: Air Quality Data with and without Simulated Treatment



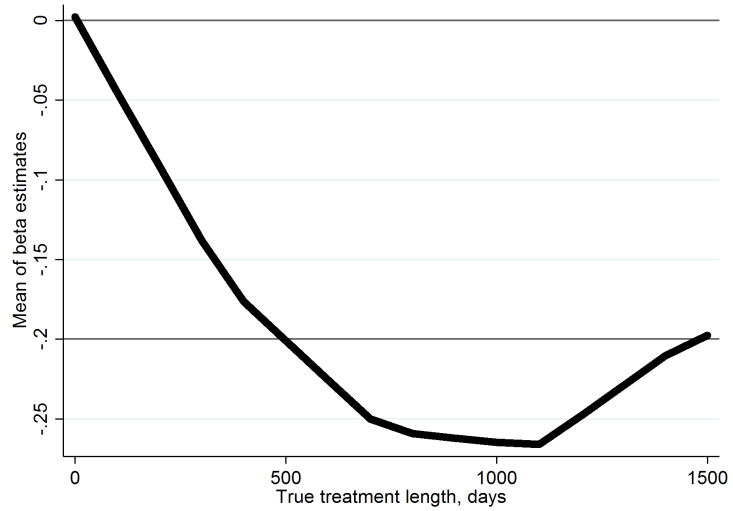
Note: This figure plots the true air quality data for a representative monitor (in Louisiana) in grey, and the simulated impact (in black) of a treatment effect of -0.2.

Figure A2: Representative RDiT Graphs, No Confounders



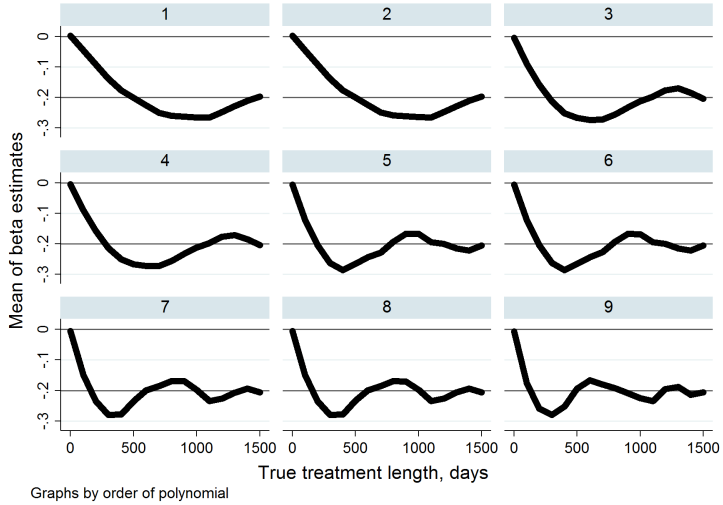
Note: These figures plot regression discontinuity in time estimates for a single representative monitor, with separate panels for different empirical specifications. All six panels shows residuals after controls have been removed (grey circles) and a line fitted to those residuals (black line). Panel A uses a global polynomial approach and Panel B uses separate polynomials in the pre- and post-periods (for this particular monitor, the BIC-chosen polynomial is of low order and thus appears approximately horizontal. Across all monitors, the average polynomial chosen is given in Table 2). Panel C uses a local linear approach with 30 days of observations on either side of the threshold. Panels D and E use a two-step augmented local linear approach, where controls are estimated and residuals saved using the sample as a whole (Panel D) while the treatment effect is estimated with just 30 days of observations on either side of the threshold (Panel E). In Panel F, observations are collapsed to monthly averages before the regression is estimated. All regressions include weather and seasonality controls, as described in the text.

Figure A3: Mean Estimated Treatment Effect, RDiT Estimate using Polynomial Order 1, When Treatment Sharply Decays



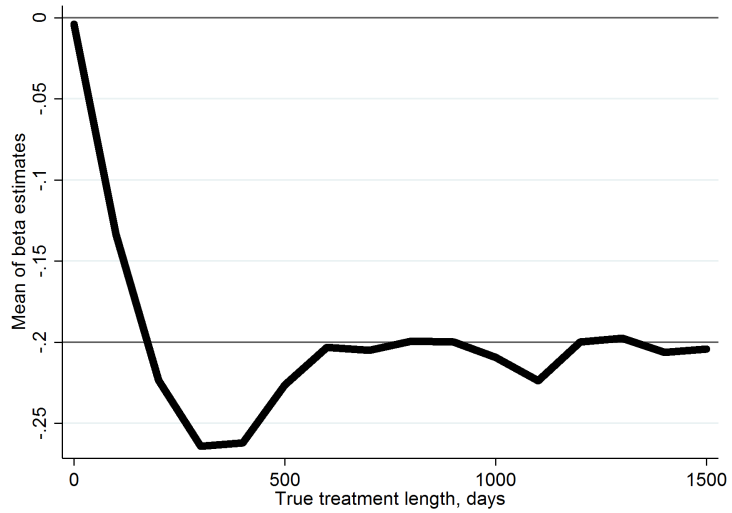
Note: This figure plots the mean estimated treatment effect as the true treatment length varies (across the x-axis). The treatment effect is estimated using an RDiT framework with global polynomial of order 1, and the researcher is assumed to not know that the true treatment decays after x days.

Figure A4: Mean Estimated Treatment Effect, RDiT Estimate using Polynomial Orders 1 through 9, When Treatment Sharply Decays



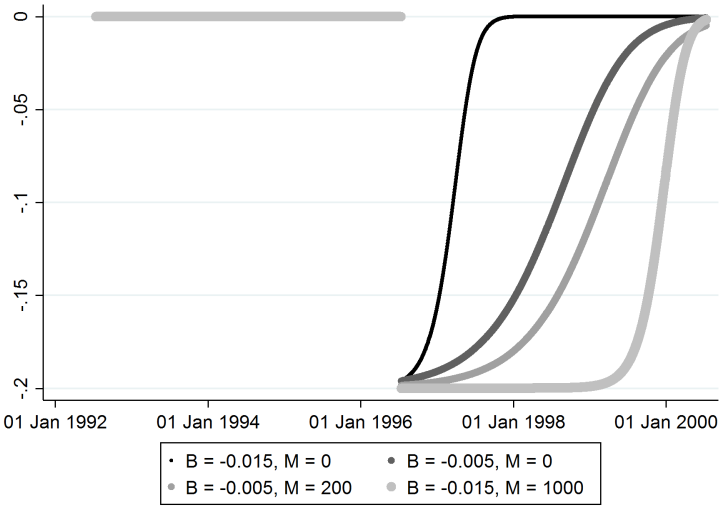
Note: This figure plots the mean estimated treatment effect as the true treatment length varies (across the x-axis). The treatment effect is estimated using an RDiT framework with global polynomial of varying order (across the panels), and the researcher is assumed to not know that the true treatment decays after x days.

Figure A5: Mean Estimated Treatment Effect, RDiT Estimate using BIC-Chosen Polynomial Order, When Treatment Sharply Decays



Note: This figure plots the mean estimated treatment effect as the true treatment length varies (across the x-axis). The treatment effect is estimated using an RDiT framework with global polynomial with BIC-chosen order, and the researcher is assumed to not know that the true treatment decays after x days.

Figure A6: Generalized Logistic Decay Functions



Note: This figure plots four logistic decay functions, using four of the combinations of parameters B and M for which the true treatment value has decayed by at least 90 percent by the end date (see text for details).

Table A1: Air Quality Data Summary Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
Ozone	682,130	0.046	0.019	0.000	0.250
Daily Min. Temperature	682,130	51	18	-30	96
Daily Max. Temperature	682,130	72	19	-14	121
Daily Mean Temperature	682,130	62	18	-21	105
Rain	682,121	0.11	0.37	0.00	18.40
Snow	681,654	0.38	4.19	0.00	336.94

Note: This table describes the air quality data used for the Monte Carlo simulations. Observations are by day (t=6,574) and monitoring station (n=108). Coverage is 1989 to 2006 for the U.S., excepting California. Ozone is the daily maximum concentration, in parts per million. Temperature is in degrees Fahrenheit. Precipitation is in inches.

Table A2: Treatment Sharply Decays after One Month, $\beta_{initial} = -0.2$ and $\beta_{long-run} = 0$

	BIC-chosen Global Polynomial	BIC-chosen Separate Polynomials	Local Linear	Augmented Local Linear	Pre/Post, Monthly Observations
			A. Air Quality Data		
Estimated Beta	-0.05	-0.05	-0.20	-0.19	<0.01
	(0.11)	(0.12)	(0.29)	(0.26)	(0.11)
Observations	2822	2822	59	59	96
	(72)	(72)	(3.5)	(3.5)	(1.0)
Polynomial Order	4.1	1.7, 1.7	-	-	-
	(2.9)	(0.8), (0.8)			
			B. Simulated Data		
Estimated Beta	-0.02	-0.02	-0.20	-0.20	<0.01
	(0.07)	(0.08)	(0.51)	(0.51)	(0.08)
Observations	2921	2921	61	61	96
	-	-	-	-	-
Polynomial Order	1.0	1.0, 1.0	-	-	-
	(0.05)	(0.03), (0.09)			

Note: Each column of this table reports the mean treatment effect coefficient and number of observations across 2,080 regressions—see Table 2 for details. In contrast to Table 2, the true treatment effect of -0.2 lasts only one month, then sharply drops to 0.

Table A3: Treatment Smoothly Decays, $\beta_{initial} = -0.2$ and $\beta_{long-run} = 0$

A. Air Quality Data

	M parameter					
B parameter	0	200	400	600	800	1000
-0.005	-0.21	-0.21				
	(0.11)	(0.11)				
-0.0075	-0.22	-0.21	-0.21	-0.21		
	(0.11)	(0.11)	(0.11)	(0.11)		
-0.01	-0.23	-0.22	-0.21	-0.21	-0.21	
	(0.11)	(0.11)	(0.11)	(0.11)	(0.11)	
-0.0125	-0.23	-0.23	-0.21	-0.21	-0.21	-0.20
	(0.12)	(0.11)	(0.11)	(0.11)	(0.11)	(0.11)
-0.015	-0.22	-0.24	-0.21	-0.20	-0.21	-0.20
	(0.12)	(0.11)	(0.12)	(0.11)	(0.11)	(0.11)

B. Simulated Data

	M parameter					
B parameter	0	200	400	600	800	1000
-0.005	-0.23	-0.24				
	(0.08)	(0.08)				
-0.0075	-0.19	-0.23	-0.25	-0.25		
	(0.08)	(0.08)	(0.08)	(0.08)		
-0.01	-0.16	-0.21	-0.25	-0.26	-0.25	
	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	
-0.0125	-0.13	-0.20	-0.24	-0.26	-0.26	-0.23
	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)
-0.015	-0.11	-0.19	-0.23	-0.26	-0.26	-0.24
	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)

Note: Each cell of this table reports the mean treatment effect coefficient across 1,080 regressions (Panel A) or 1,000 regressions (Panel B). The specification is identical to that of Column 1 in Table 2: a global polynomial using eight years of data and controlling for weather and seasonality. In contrast to Table 2, the true treatment effect of -0.2 smoothly decays towards 0, following a generalized logistic decay function. The B and M parameters govern the speed and shape of this decay; see Figure A6. We only include parameter combinations for which the true treatment value has decayed by at least 90 percent by the end date, for ease of comparison with the sharp decay parameterization.

Table A4: Treatment Has 7-Day Linear Ramp-Up, $\beta_{long-run} = -0.2$

	BIC-chosen Global Polynomial	BIC-chosen Separate Polynomials	Local Linear	Augmented Local Linear	Pre/Post, Monthly Observations
			A. Air Quality Data		
Estimated Beta	-0.20	-0.20	-0.11	-0.11	-0.19
	(0.11)	(0.11)	(0.29)	(0.26)	(0.10)
Observations	2822	2822	59	59	96
	(72)	(72)	(3.5)	(3.5)	(1.0)
Polynomial Order	4.0	1.7, 1.6	-	-	-
	(2.9)	(0.8), (0.8)			
			B. Simulated Data		
Estimated Beta	-0.20	-0.20	-0.12	-0.12	-0.19
	(0.07)	(0.08)	(0.51)	(0.51)	(0.08)
Observations	2921	2921	61	61	96
	-	-	-	-	-
Polynomial Order	1.0	1.0, 1.0	-	-	-
	(0.05)	(0.03), (0.06)			

Note: Each column of this table reports the mean treatment effect coefficient and number of observations across 2,080 regressions—see Table 2 for details. In contrast to Table 2, the true treatment effect is assumed to ramp-up in a linear way, beginning at a known start date. Following the linear phase-in, the treatment is a constant -0.2.

Table A5: Estimated Autoregressive Parameters

	CO	NO2	Ozone	PM10	PM2.5	SO2
Lagged pollution	0.42	0.35	0.41	0.33	0.51	0.30
	(0.05)	(0.04)	(0.05)	(0.06)	(0.06)	(0.06)
Monitors	101	148	280	28	29	101
Observations per monitor	584	581	573	450	490	562
R2	0.61	0.67	0.71	0.58	0.61	0.40

Note: This table reports the mean autoregressive coefficient, the mean standard error of that coefficient, the mean number of observations, and the mean R2 value from 687 individual regressions at the level of a monitor by pollutant, using daily data. Dates are 2004-2005. Controls include a cubic function of the daily mean temperature, a cubic function of windspeed, the interaction of mean temperature and windspeed, day of week effects, month effects, and a cubic time trend. Standard errors are clustered by sample week.

Table A6: Robustness Check: Estimated Autoregressive Parameters

	CO	NO2	Ozone	PM10	PM2.5	SO2
Lagged pollution	0.42	0.35	0.48	0.29	0.57	0.37
	(0.05)	(0.04)	(0.04)	(0.06)	(0.07)	(0.06)
Monitors	82	123	194	24	23	48
Observations per monitor	624	647	654	531	499	624
R2	0.62	0.69	0.74	0.52	0.64	0.46

Note: This table reports the mean autoregressive coefficient, the mean standard error of that coefficient, the mean number of observations, and the mean R2 value from 494 individual regressions at the level of a monitor by pollutant, using daily data. Dates are 2004-2005. For computational ease, only monitors from six states are included: California, Florida, North Carolina, Ohio, Pennsylvania, and Texas (the six states with the greatest number of PM2.5 monitors). Controls include a cubic function of the daily mean temperature, a cubic function of windspeed, a cubic function of precipitation, the interaction of mean temperature and windspeed, the interaction of temperature and precipitation, day of week effects, month effects, and a cubic time trend. Where windspeed data are missing, a state-level daily average has been used. Standard errors are clustered by sample week.

Table A7: Including AR(1) Term

AR(1) parameter α		BIC-chosen Global Polynomial	BIC-chosen Separate Polynomials	Local Linear	Augmented Local Linear	Pre/Post, Monthly Observations
0.1	Estimated beta	-0.20	-0.20	-0.24	-0.22	-0.22
		(0.07)	(0.08)	(0.53)	(0.49)	(0.09)
	Estimated alpha	0.10	0.10	0.04	0.11	-0.03
		(0.02)	(0.02)	(0.13)	(0.02)	(0.10)
0.3	Estimated beta	-0.20	-0.20	-0.25	-0.22	-0.28
		(0.07)	(0.08)	(0.55)	(0.49)	(0.11)
	Estimated alpha	0.30	0.30	0.23	0.31	-0.02
		(0.02)	(0.02)	(0.13)	(0.02)	(0.10)
0.5	Estimated beta	-0.20	-0.20	-0.26	-0.22	-0.39
		(0.07)	(0.08)	(0.58)	(0.49)	(0.15)
	Estimated alpha	0.50	0.50	0.41	0.51	-0.01
		(0.02)	(0.02)	(0.12)	(0.02)	(0.10)
0.7	Estimated beta	-0.20	-0.20	-0.29	-0.22	-0.64
		(0.08)	(0.08)	(0.64)	(0.49)	(0.25)
	Estimated alpha	0.70	0.70	0.59	0.72	0.02
		(0.01)	(0.01)	(0.11)	(0.01)	(0.10)
0.9	Estimated beta	-0.21	-0.21	-0.32	-0.22	-1.65
		(0.08)	(0.08)	(0.73)	(0.49)	(0.64)
	Estimated alpha	0.90	0.90	0.76	0.92	0.16
		(0.01)	(0.01)	(0.10)	(0.01)	(0.10)

Note: Each column of this table reports the mean treatment effect coefficient and number of observations across 1,000 regressions, matching Table 4. In contrast to Table 4, the researcher correctly models the AR(1) process.