

NBER WORKING PAPER SERIES

REGRESSION DISCONTINUITY IN TIME:
CONSIDERATIONS FOR EMPIRICAL APPLICATIONS

Catherine Hausman
David S. Rapson

Working Paper 23602
<http://www.nber.org/papers/w23602>

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2017

We thank Michael Anderson, Max Auffhammer, Severin Borenstein, Matias Cattaneo, Lucas Davis, Stephen Holland, Ryan Kellogg, Doug Miller, and Jeff Smith for helpful comments. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2017 by Catherine Hausman and David S. Rapson. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Regression Discontinuity in Time: Considerations for Empirical Applications
Catherine Hausman and David S. Rapson
NBER Working Paper No. 23602
July 2017
JEL No. C14,C21,C22,Q53

ABSTRACT

Recent empirical work in several economic fields, particularly environmental and energy economics, has adapted the regression discontinuity framework to applications where time is the running variable and treatment occurs at the moment of the discontinuity. In this guide for practitioners, we discuss several features of this "Regression Discontinuity in Time" framework that differ from the more standard cross-sectional RD. First, many applications (particularly in environmental economics) lack cross-sectional variation and are estimated using observations far from the cut-off. This is in stark contrast to a cross-sectional RD, which is conceptualized for an estimation bandwidth going to zero 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. Finally, tests for sorting or bunching near the discontinuity are often irrelevant, making the methodology 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 design.

Catherine Hausman
Gerald R. Ford School of Public Policy
University of Michigan
735 South State Street
Ann Arbor, MI 48109
and NBER
chausman@umich.edu

David S. Rapson
Department of Economics
University of California, Davis
One Shields Avenue
Davis, CA 95616
dsrapson@ucdavis.edu

1 Introduction

As noted by Lee and Lemieux (2010), the use of 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 treatment occurring at the moment of the discontinuity. Examples of this “regression discontinuity in time” (RDiT) include Anderson (2014); Auffhammer and Kellogg (2011); Bento et al. (2014); Burger et al. (2014); Busse et al. (2006, 2010); Davis (2008); Davis and Kahn (2010); Chen et al. (2009); Chen and Whalley (2012); De Paola et al. (2013); Gallego et al. (2013); Grainger and Costello (2014); Lang and Siler (2013). Papers using this “regression discontinuity in time” (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 that exceeds 700.¹ In this paper, we describe pitfalls that may affect RDiT designs and suggest that the name RDiT may imply more commonalities with cross-sectional RD designs than is actually true in practice. Our goal is not to provide any formal additions to econometric theory but rather to demonstrate to other practitioners, via simulations, possible issues that may affect empirical work.

Many – although not all – of the papers using the RDiT methodology are in the field of environmental and energy economics, where they have been used to estimate the treatment effect of various policies. 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. As a result, fewer local pollutants such as sulfur dioxide would be emitted. Suppose this regulation had no other impact on the electricity sector,

¹Citation counts were retrieved on Google Scholar on June 21, 2016.

other industry, people’s behavior, or economic activity more generally. The hypothetical RDiT researcher could then examine the impact of the power plant policy on ambient air quality, by estimating the change in sulfur dioxide 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 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 other potential time-varying confounders, but they must be 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 policy that required the device only for power plants of a size over a certain threshold (say, 1000 megawatts). While cities with smaller power plants (0 to 999 megawatts) may be different in many observable and unobservable ways from cities with larger power plants (1000 to 2000), the researcher applying the RD method would compare air quality after the policy change only for cities with power plants very close to the threshold (e.g., with size 990 to 1010). This research design would require a mass of power plants close to this threshold. In contrast to a difference-in-difference approach, it would not require pre-treatment data. In this typical RD, then, the researcher can observe many units near the cut-off, and other confounders are believed to change smoothly across the threshold that defines treatment.²

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. This is an entirely different source of identifying variation than is used in the canonical cross-sectional RD, rendering the traditional RD toolkit (for example, as described in Lee and Lemieux (2010)) unavailable to the researcher. Furthermore, the differences from a cross-sectional RD can lead to bias if ignored. We argue that some RDiT designs will drift further from

²As 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.

the experimental ideal than others, and that their strength and reliability are thus not homogenous. We describe our findings both qualitatively and via Monte Carlo simulations, which allow us to highlight some of the common pitfalls and identify potential solutions. We explore three main features of the RDiT setting.

First, there is a bias/precision trade-off in the context of RDiT. While this trade-off exists for many empirical strategies, increasing the sample size by growing T rather than N may make this particularly problematic for some RDiT applications. Identification in cross-sectional RD designs rests upon a conditional expectation as one approaches the cut-off. However, because the RDiT framework is frequently used in studies with no (or insufficient)³ cross-sectional identifying variation, the sample is too small for estimation as the bandwidth narrows around the cut-off. As a result, the RDiT researcher must rely on observations away from the cut-off in order to obtain sufficient power for precisely estimating the coefficient of interest. The use of observations remote (in time) from the discontinuity is a substantial conceptual departure from the identifying assumptions used in a cross-sectional RD, and 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 econometrics literature has developed ways to account for processes that are autoregressive, for instance, but these have not been applied in the RDiT context. We argue they are relevant for proper estimation and interpretation of short-run versus long-run effects.

Finally, an important test that is standard in cross-sectional RDs is often not applicable when time is the running variable. The McCrary (2008) density test allows practitioners (and consumers) of RD to assess the presence and nature of strategic behavior near the threshold, which can inject bias into the estimates. When the density of the running variable

³Some 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 .

is uniform (e.g. time), the test becomes irrelevant. As such, the researcher can evaluate endogenous movements around the threshold only indirectly. While the McCrary test does not allow the researcher to rule out *all* forms of manipulation of the treatment variable, it does substantially aid in evaluating the validity of an RD research design. In contrast, for the RDiT design, concerns about self-selection into and out of treatment as well as concerns about anticipation or avoidance behavior must be addressed on an ad hoc basis – much like when deploying other non-experimental methods.

Commenters have noted the similarity between RDiT design and the widely-used event study or interrupted time series methods. 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, widely-cited event studies on stock returns in environmental economics⁴ generally (1) do not use long time horizons, instead focusing on short windows around the event; (2) do not use single cross-sectional units, but rather a large sample of, e.g., firms; and (3) do not use high-order polynomial controls in time. The stock-returns event study framework forms a useful part of the empirical researcher’s toolbox, as it allows for empirical analysis in contexts where cross-sectional variation is limited. However, this paper is the first to clarify how the RDiT methodology as typically implemented differs from both the cross-sectional RD and the event study frameworks. It is also the first to call for RDiT to explicitly account for time-series properties of the data.

In what follows, we briefly review the canonical cross-sectional RD technology (Section 2), present results from Monte Carlo simulations that highlight some of the key insights we wish to convey (Section 3), and offer guidance for the empirical practitioner on ways to check the validity of the underlying assumptions and the robustness of the results (Section 4). One of our conclusions in Section 5 is that more research is needed to understand the behavior of the RDiT estimator.

⁴Examples include Hamilton (1995); Konar and Cohen (1997); Dasgupta et al. (2001).

2 RD Framework

In this section we briefly revisit 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). Readers who are familiar with these papers should feel free to skip to the next subsection, where our original contribution begins.

The appeal of the RD approach can be most clearly seen through the lens of the Neyman-Rubin potential outcomes framework (Rubin, 1974; Neyman, 1923 [1990]) and the intuitive similarity to the randomized controlled trial (RCT). The empiricist cannot observe an individual experimental subject (i) in more than one state of the world at once, rendering comparison of outcomes impossible under both receipt and non-receipt of treatment (which we denote Y_i^T and Y_i^C , respectively). The causal treatment effect is logically well-defined ($Y_i^T - Y_i^C$), but unobservable at the individual level. Instead, empiricists seek to estimate the average treatment effect (ATE) using some representative sample of the relevant population. Comparing average outcomes of subjects who receive treatment to that of subjects who do not receive treatment provides an estimate of the ATE. If the expected outcome of treated subject *had they not been treated* is equal to the expected outcome of non-treated subjects, then this estimate is free from selection bias.

Under the stable unit treatment value assumption (SUTVA), a well-designed experiment solves selection concerns by randomly assigning subjects into control or treatment status (as in a RCT). This is generally implemented by assigning 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. Phrased differently, there exists a discontinuity at $\nu = 0.5$ such that subjects with draws (from a continuous distribution) on different sides of the threshold receive interventions that are discretely different.

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). Treatment assignment in the RD context occurs as a result of the proximity of some treatment assignment variable⁵ (X) with respect to a threshold, or discontinuity, c . That is, if $X > c$, the subject is treated, and if $X < c$, the subject is not. Typically, the mass of subjects is dense in the neighborhood of the threshold, allowing estimation via a cross-sectional comparison of subjects “just above” and “just below” the threshold. A compelling RD design thus shares an important feature of the distribution of ν in the RCT: in the neighborhood of the discontinuity, whether $X > c$ or $X < c$ is “as good as random.” If true, RD designs can be analyzed, tested and trusted like randomized controlled trials.

The focus on the local randomization interpretation of the RD design is somewhat controversial in the literature. 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, also a controversial issue (Shadish et al., 2002).

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 internally valid estimate of the causal local average treatment effect.

⁵Also referred to as the “forcing” or “running” variable.

2.1 RD in Time Framework

There is nothing inherently wrong with using time as the forcing variable in a regression discontinuity design. Our focus in this paper is the growing number of papers (and, indeed, the vast majority of RDs that use time as the forcing variable in the environmental economics literature today) that rely on time-series variation for identification. By way of contrast, consider Ito (2015), which uses time as the forcing variable but relies on asymptotics in N . The time threshold in that study determined eligibility for an electricity rebate program. Customers who initiated service after the threshold date were ineligible, while those before were eligible. Furthermore, the eligibility date was set long after it had passed. As such, there is a credible claim to exogeneity of initiation date with respect to the threshold (no strategic 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 differs markedly from studies that rely on a time discontinuity, but identify based on asymptotics in T (rather than N).

The most common 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)).⁶ A common application in the energy and environmental economics literature is the evaluation of policies impacting air quality (Auffhammer and Kellogg, 2011; Bento et al., 2014; Chen and Whalley, 2012; Davis, 2008; Gallego et al., 2013). Recall the example in which a federal policy takes effect on a particular date, requiring more stringent pollution abatement from all firms in a polluting industry. Suppose the researcher does not observe emissions from individual firms, but rather ambient air quality within a city. The researcher might then wish to estimate the impact of the policy on ambient concentrations of sulfur dioxide, a pollutant emitted by coal-fired power plants. Among

⁶Some 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 .

the identification challenges in such a setting are that (1) there is no cross-sectional variation in treatment, so no difference-in-difference is possible;⁷ and (2) there are many confounders, such as changes in weather and emissions from other industries, that may change at the threshold.

In this set-up, the researcher knows the date c of a policy change, and thus assumes that for all $t > c$, the unit is treated, and for all $t < c$, the unit is not. The RDiT framework thus clearly accords with the “discontinuity at a threshold” interpretation of RD designs, but it is less clear that it accords with the “local randomization” interpretation of RD.⁸ The running variable is time itself, and thus cannot be thought of as randomly assigned. The threshold (the date the treatment is implemented) is also not typically randomly assigned; rather it is chosen by policymakers and announced in advance. This is in contrast to the canonical test-scores cross-sectional RD, for which whether an individual falls just above or just below the cut-off can be thought of “as good as random.”

Identification in the RDiT setting is different from the standard cross-sectional RD. 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. A necessary assumption required for identification in any RD is that unobserved determinants of the outcome variable Y are continuous with respect to the forcing variable. For RDiT, there will almost always be covariates that are discontinuous in time, such as day-of-week effects. These covariates need to be included as controls, whereas in many cross-sectional RD frameworks these covariates are used only to improve precision (Lee and Lemieux, 2010).

⁷Auffhammer and Kellogg (2011) are able to use a difference-in-difference framework as part of their analysis. They additionally use the RDiT approach because they are interested in heterogeneous treatment effects across space. This demonstrates that RDiT could provide value even in settings where a difference-in-difference procedure is possible. Some papers also compare RDiT estimates and differences-in-differences estimates, such as Chen and Whalley (2012) and Gallego et al. (2013). This can be useful where an untreated unit exists but its validity as a control is in doubt.

⁸A 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.

The continuity assumption is weaker than, for instance, what is required for identification in a pre/post analysis, which assumes that unobserved determinants of Y are not correlated with time. As such, if this is the only assumption required, an RDiT may be identified even when a pre/post analysis would not be. In the case of the power plant example, there could be many confounding impacts on sulfur dioxide concentrations: there is seasonality in electricity demand that leads to seasonality in coal-fired power plant emissions; electricity demand similarly follows a specific pattern over days of the week; exogenous weather shocks impact both electricity demand (and therefore generation) and also atmospheric chemistry relating to the pollutant; etc. As such, a simple pre/post analysis that ignored confounding factors that evolve over time would be noisy at best, and possibly biased. The justification a researcher might provide for using the RDiT method is that by using daily data over a multiple-year horizon, controls could be included to account for weather, seasonality, etc. The inclusion of a flexible time trend would then require only an assumption that nothing changes discontinuously across the threshold date $t = c$, so that the impact of the regulation – local to the date of the policy change – can be identified.

Unfortunately, assuming continuity of unobservables is likely to be insufficient for identification in a typical deployment of the RDiT. A researcher using RDiT for the hypothetical power plant example would need to also rule out additional challenges to identification. For instance, she would want to verify that there are no anticipation effects close to the threshold. She would also want to consider the possibility of time-varying treatment effects – power plant owners, or electricity consumers, might adapt their behavior over time, causing the long-run policy effect to be different from the short-run policy effect. The literature has generally considered this to be simply a matter of whether RDiT measures a local average treatment effect,⁹ but below we show that current methods may not be reliable in this scenario. Finally, the potential for auto-regression in time-series data could pose a threat to identification, as we discuss below.

⁹Here and throughout, we refer to a “local” average treatment effect meaning local in time (“temporal LATE”), not local in space.

The assumptions needed for inference are also different from those in the cross-sectional 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 RD framework (which typically uses a small bandwidth), small-sample inference may be necessary. This differs from the cross-sectional RD framework, where the researcher may have a large sample even local to the cut-off. 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.

2.2 Comparison to Other Methods

The RDiT is, of course, closely related to other designs, such as pre/post analysis, interrupted time series, and event studies.¹⁰ In fact, the program evaluation literature has noted the similarity between interrupted time series and RD (Shadish et al., 2002). One difference between RDiT and some applications of these methods is the use in RDiT of high-frequency data, which allows the researcher to either incorporate flexible controls in time or to use a small sample window.¹¹ As noted by Davis (2008), the RDiT can have advantages over a pre/post analysis using OLS, which must assume there are no unobservables correlated with time. In contrast, the RDiT can allow for unobservables impacting the outcome variable, so long as they do not change discontinuously at the threshold, and so long as the relationship can be controlled for nonparametrically. An additional benefit of RDiT, discussed in Auffhammer and Kellogg (2011), is the possibility of uncovering spatially heterogeneous treatment effects. We now describe three conceptual differences between RDiT and the cross sectional RD.

¹⁰These three terms are sometimes used interchangeably.

¹¹Event study papers from the finance literature do, however, make use of high frequency data; typically they use a small sample window and thus are closely related to the local linear methods we describe below.

2.2.1 Time-Series Variation and Time-Varying Treatment Effects

In a standard cross-sectional RD, the gold standard study would have a “large-enough” sample in the neighborhood of the cut-off. This would approximate the limit of the conditional expectation of the outcome variable just-below and just-above the cut-off. In an RDiT, the researcher frequently has little or no cross-sectional variation. For instance, a study might analyze outcomes relating to a single cross-sectional unit, or evaluate separate RDiT specifications for each individual cross-sectional unit (e.g. Auffhammer and Kellogg (2011); Chen and Whalley (2012); Davis (2008); Davis and Kahn (2010); Lang and Siler (2013)). In such cases, the researcher has two choices for increasing the sample size: (1) increasing the frequency of data (“infill” or “fixed-domain” 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 variation if the data are serially correlated. Expanding the time window increases the probability of bias, as data are added away from the cut-off. 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 cut-off. However, the problem could be more severe for an RDiT application, and it is likely to be an important consideration when there is 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 2.1, 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. However, this moves the researcher away from the cut-off date when the treatment begins. Two assumptions are then needed: (1) that the model is 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. And, (2) the researcher must assume the correct specification of 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). Note that 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. Below, we give examples from Monte Carlo simulations in which these conditions are met, and examples in which they are not.

Of particular interest in environmental applications are time-varying treatment effects. The assumption violated with time-varying treatment effects is that the researcher has correctly specified the treatment variable. Again, this 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) may make this problem more relevant than in many cross-sectional RD settings. In addition, as with sorting behavior, this assumption is not testable in the RDiT framework. To the extent that RDiT relies on observations far from the cut-off, the researcher must consider whether the temporal LATE that she might uncover is a policy-relevant parameter. If the long-run effect is of greater importance, an RDiT study may be inappropriate.

Beyond the question of whether the short-run temporal LATE is different from a longer-run effect, a question arises as to how well the existing RDiT estimation procedures can uncover the true LATE. In particular, the use of higher-order polynomial controls poses issues when the treatment effect varies over time, as we show in Monte Carlo simulations. It is possible that the polynomial could attenuate the estimated treatment effect much like a model that is oversaturated with fixed effects (see Fisher et al. (2012)). Moreover, we show below a setting with time-varying treatment effects where the global polynomial control overfits the data and produces an estimated treatment effect that is too large, rather than attenuated. While multiple papers have discussed possible time-varying treatment effects (Anderson, 2014; Auffhammer and Kellogg, 2011; Bento et al., 2014; Burger et al., 2014; Chen and Whalley, 2012; Davis, 2008; Gallego et al., 2013) we have not seen any discussion

of the possibility in this context that a global polynomial could overfit the data.

The researcher could instead use a specification in which the treatment effect is explicitly allowed to vary over time. This may be preferred whenever adaptation effects are plausible, as it avoids issues related to the global polynomial (e.g. overfitting). However, it poses additional challenges for the researcher: lack of power, and an inability to identify whether the estimated change in the treatment effect comes from such adaptation effects or from unobservable confounders.

2.2.2 Autoregression

Time-series data are likely to exhibit serial dependence. The first implication 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, which to our knowledge has not been addressed in the literature, is that if there is autoregression in the dependent variable (even after accounting for serial correlation in the exogenous variables and the residuals), this 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). To the extent that this impacts estimation, it is relevant for much of the existing RDIT literature. To quantify the impact of autoregression in this setting, we estimate the autoregressive parameters for six pollutants at air quality monitors in the U.S., using daily data. Table 1 shows the estimated autoregressive parameters, after controlling flexibly for seasonality and weather (details in the Appendix). Overall, we obtain estimates of the AR(1) parameter of 0.3 to 0.5, varying by pollutant.¹²

This serial dependence introduces dynamic effects not generally estimated 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 the

¹²The coefficient in the ozone equation is surprising, given ozone's short half-life; it is possible that serially correlated unobservables are driving that estimate.

Table 1: Estimated Autoregressive Parameters

	CO	NO2	Ozone	PM10	PM2.5	SO2
Lagged pollution	0.42 (0.05)	0.35 (0.04)	0.41 (0.05)	0.33 (0.06)	0.51 (0.06)	0.30 (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 coefficients, standard errors, number of observations, and R2 values 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. The Appendix shows robustness to alternative specifications, such as including precipitation.

regulation will decrease emissions by $x\%$. Suppose the researcher observes daily pollution levels in the city for several years before and after the policy change. Assume the plant owners comply on the first date of the policy and there are no other confounders. Finally, suppose the pollutant of interest dissipates at a rate of $1 - y\%$ each day, implying that $y\%$ 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 $x\%$. On the second day, however, the treatment effect will be a combination of lower emissions that day and lower pollution left over from the previous day: the total treatment effect on air pollution levels will be a decrease of $x\% + x\% \cdot y\%$. The long-run effect will be $\frac{x\%}{1-y\%}$.

The magnitude of the estimated treatment effect will depend on whether the researcher accounts for this autoregression, and on the extent to which identification is being achieved from the discontinuity (i.e. as the bandwidth shrinks towards the threshold) or from time-series variation outside of the neighborhood of the discontinuity. 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. In particular, over-fitting 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 research choices about bandwidth and specification. High-frequency data, while allowing for more power, may be more likely to exhibit qualitatively-important autoregression. Below, we give examples from Monte Carlo simulations to illustrate these points.

2.2.3 Sorting Behavior and Anticipation Effects

Finally, if the intervention may alter behavior in the neighborhood of the discontinuity, we would like to be able to test for this and, if necessary, control for it. In a cross-sectional RD, a density test such as the McCrary (2008) test is a key check for strategic behavior or selection. It is generally used to rule out these confounding factors, thus making it unnecessary to control for them. When time is the running variable, however, it is generally not possible to test for strategic behavior or selection 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 the conditional density of the forcing variable. That the density of the forcing variable (time) is uniform renders such tests logically irrelevant.

Consider the hypothetical power plant example. In this RD design, 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 data on regional averages, 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.

3 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 test potential remedies. We use the Monte Carlo simulations to examine the potential pitfalls of relying on time-series variation in the presence of constant and non-constant treatment effects, as well as to determine the importance of correctly specifying any autoregressive process in the dependent variable. Using real-world pollution monitoring data as a base, we impose simulated treatment effects with “known” magnitude, timing, and persistence properties. We then attempt to retrieve them using RDiT approaches. This procedure offers a transparent and compelling laboratory to test our main hypotheses. In addition to results using pollution monitoring data, 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.

3.1 Data

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 power plant emission restrictions, or other (unobserved to the researcher) environmental policies. 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. This leaves us with 108 monitors. The AK data contain daily observations of the maximum measured ozone concentration. As is standard in the literature, we take the log of the dependent variable and then interpret the coefficient on our policy dummy as the approximation of a percentage change. Control variables include daily minimum and maximum temperature and daily rain and snowfall totals. Table A1 in the Appendix gives summary statistics. The mean ozone concentration for this time period was 0.05 parts per million (ppm). For comparison, the EPA standard for ozone over this time period ranged from a daily maximum of 0.12 ppm to an eight hour standard of 0.08 ppm (Auffhammer and Kellogg, 2011).

3.2 Results

For the first set of Monte Carlo simulations, we construct a simulated treatment effect that begins at a known start date (using ten randomly selected start dates).¹³ Then for each monitor’s data series $x_{i,t}$, and using $\beta = -0.2$, we construct our outcome variable $y_{i,t}$:

$$y_{i,t} \equiv x_{i,t} + \beta \cdot \mathbb{1}\{t \geq t_{start}\} \quad (1)$$

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), for ozone concentrations at a monitor in Louisiana.

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 RD regressions, both using a global polynomial approach (with the order of the

¹³We 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.

polynomial chosen by BIC) and a local linear approach.¹⁴ For the global polynomial, we follow Davis (2008) and use an eight-year window around the treatment start date. As is standard in the air pollution policy evaluation literature, we control for seasonality with month effects and day of week effects and for weather using cubic functions of minimum daily temperature, maximum daily temperature, precipitation, and snowfall.

Ambient air quality data can be highly variable (see e.g., Figure A1 in the Appendix), so controls have proven important in empirical applications 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. For instance, 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. As such, 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 (e.g., 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. A separate regression is run for each monitor. For all results, we show averages across the ten randomly selected start dates and 108 monitors.

We also show Monte Carlo simulations for a dataset in which the counterfactual data is simply an error term that follows a standard normal distribution (the realization of a

¹⁴For 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 using only observations within one month of the treatment start date.

Gaussian white noise process). The length of the observed data 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 estimate close to the true treatment effect, although with varying degrees of precision.¹⁵ The global polynomial results are more precise than the local linear, a standard finding in the RD literature. The augmented local linear, which strips out the effects of controls using a much longer data series, is substantially less noisy than the traditional local linear, as expected. The global polynomial results are similar to the pre/post results.

Figure 1 plots these results at a representative monitor, where “representative” is defined as having an estimated beta close to the average estimated beta. Panel A shows the global polynomial estimation; Panel B shows estimation using separate polynomials for the pre- and post- periods.¹⁶ Panel C shows a local linear specification with a bandwidth of 30 days before and 30 days after the treatment date. Panel D shows the augmented local linear specification with the estimation window used in the first step; Panel E zooms in on the subsample used in the second step. Panel F shows the pre/post specification using monthly data.

3.2.1 Time-Varying Treatment Effects

Next, 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. (In the next section, we consider a smooth decay process, which may

¹⁵We 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.

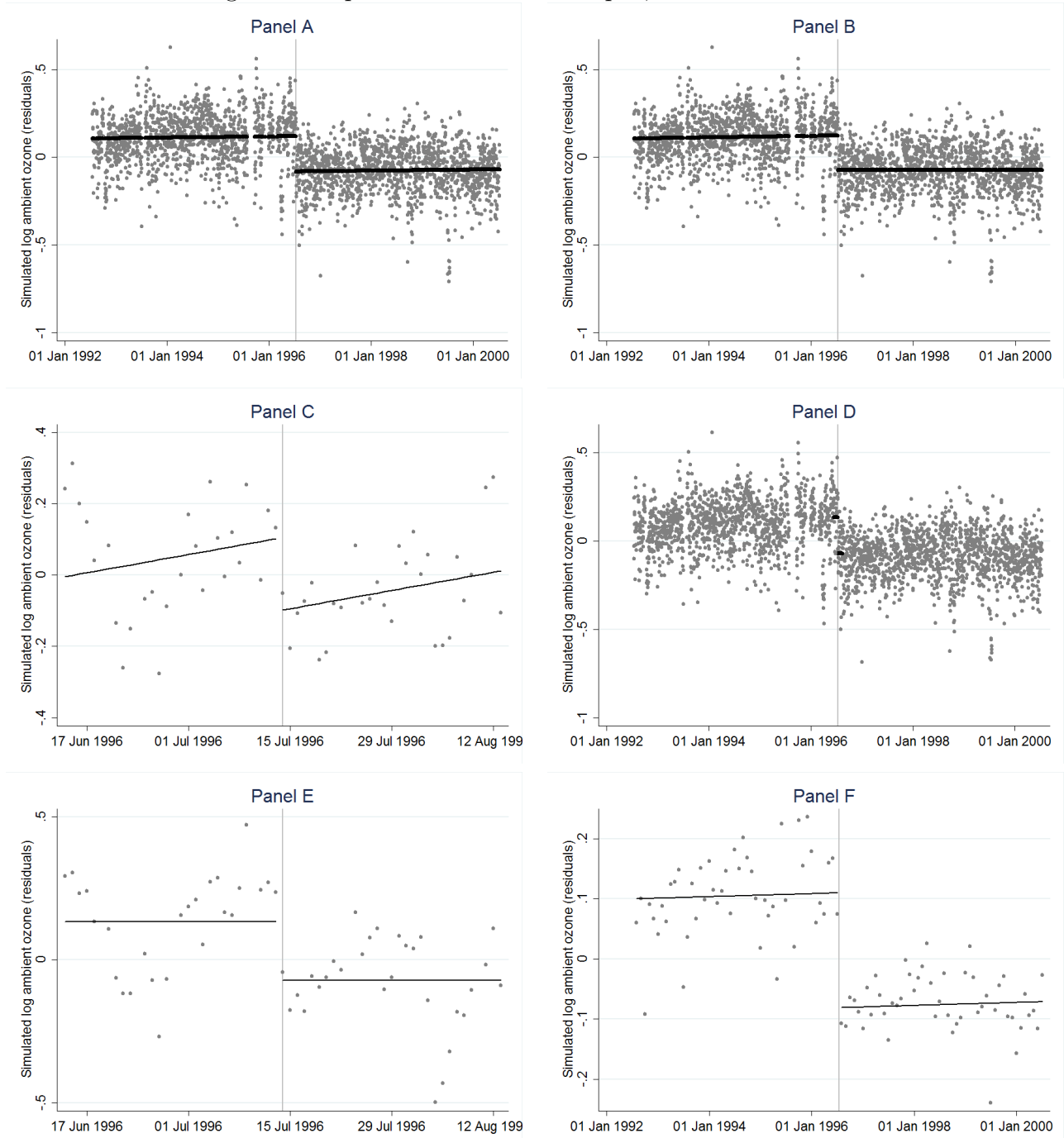
¹⁶While the BIC-chosen polynomial order happens to be low for this monitor for both Panels A and B, on average the BIC-chosen order is 4 for the global specification and 1.6 to 1.7 for the separate polynomials (see Table 2).

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
<u>A. Air Quality Data</u>					
Estimated Beta	-0.20	-0.21	-0.19	-0.19	-0.19
S.D.	(0.11)	(0.11)	(0.29)	(0.14)	(0.10)
Observations	2822	2822	59	2822	96
	(72)	(72)	(3.5)	(72)	(1.0)
Polynomial Order	4.0	1.7, 1.6	-	-	-
	(2.9)	(0.8), (0.8)			
<u>B. Simulated Data</u>					
Estimated Beta	-0.20	-0.20	-0.19	-0.20	-0.19
S.D.	(0.07)	(0.08)	(0.51)	(0.26)	(0.08)
Observations	2921	2921	61	2921	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. 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. Column 1 uses a global polynomial approach, and Column 2 uses separate polynomials in the pre- and post-periods. Both display the mean order of the BIC-selected polynomial. Column 3 uses a local linear approach with 30 days of observations on either side of the cut-off. Column 4 uses a two-step augmented local linear approach, where controls are estimated and residuals saved using eight years of data, while the treatment effect is estimated with just 30 days of observations on either side of the cut-off. In Column 5, observations are collapsed to monthly averages before the regression is estimated. All regressions include weather and seasonality controls, as described in the text. Differences in the sample size across regressions arise because of missing data points in the air quality data series.

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 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 cut-off. 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 cut-off (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.

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
<u>A. Air Quality Data</u>					
Estimated Beta	-0.27	-0.29	-0.19	-0.19	-0.15
S.D.	(0.11)	(0.12)	(0.29)	(0.14)	(0.11)
Observations	2822	2822	59	2822	96
	(72)	(72)	(3.5)	(72)	(1.0)
Polynomial Order	5.8	1.7, 2.1	-	-	-
	(2.3)	(0.8), (0.6)			
Treatment Length	365	365	365	365	365
<u>B. Simulated Data</u>					
Estimated Beta	-0.17	-0.18	-0.19	-0.21	-0.15
S.D.	(0.08)	(0.10)	(0.51)	(0.26)	(0.08)
Observations	2921	2921	61	2921	96
	-	-	-	-	-
Polynomial Order	1.1	1.0, 1.1	-	-	-
	(0.33)	(0.03), (0.32)			
Treatment Length	365	365	365	365	365

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.

accord with adaptive behavior in general equilibrium.) 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.¹⁷

Table 3 shows that the results are sensitive to the specification and the underlying data.¹⁸ The local linear specification again 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

¹⁷In the Appendix, we show results for a treatment length of one month.

¹⁸As Table A2 in the Appendix shows, results are also sensitive to the length of the treatment.

the true effect. The estimate in Panel B would accord with intuition that the specification is giving an attenuated estimate. 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.

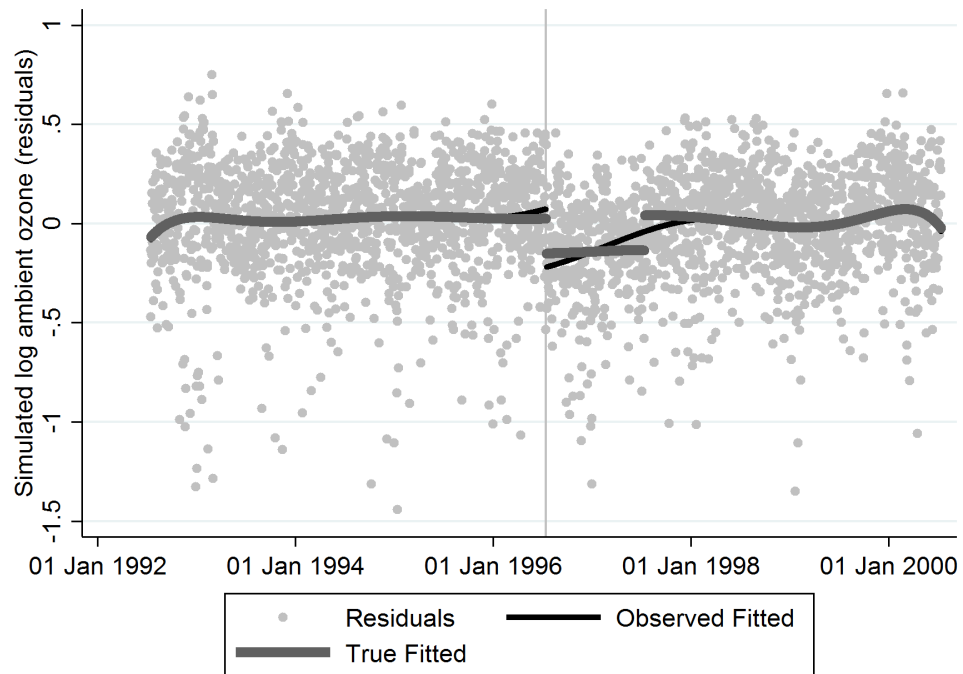
To see the overfitting of the polynomial, Figure 2 shows a monitor for which the estimated treatment effect (discontinuity in the thin black line) is larger than the true effect (discontinuity in the thick grey line). Because of the noise in the series, this confounder is not apparent in the residuals to which the polynomial is fitted. While the existing RDIT literature has discussed the possibility of a difference between short-run and long-run treatment effects, we have not seen it address the possibility that the polynomial overfits. We believe that this could be an issue of bias, separate from the question of local average treatment effects.

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. Overall, the direction and magnitude of the difference between the estimated and true effect depend on the underlying data and the treatment window used by the researcher. However, what these simulations make clear is that given a time-varying treatment effect, the RDIT estimate may not simply be a weighted average of the short-run and long-run effects.

The assumption that the treatment sharply decays may be unrealistic. 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. This function is given by:

$$A + \frac{K - A}{1 + Q \cdot \exp(-B * (date - M))^{1/v}} \quad (2)$$

Figure 2: Representative RDiT Graph, Global Polynomial when Treatment Sharply Decays



Note: This figure plots the regression discontinuity in time estimate for a single representative monitor, with an estimated treatment effect qualitatively similar to the average in Panel A of Table 3. The grey circles plot the residuals (after removing weather and seasonality controls, see text) of log ambient air quality with a simulated treatment effect of -0.2 lasting one year. The thick grey line shows the resulting polynomial a researcher would fit when (correctly) modeling the temporary nature of the treatment. The thin black line shows the resulting polynomial a researcher would fit when (incorrectly) modeling the treatment effect as constant in the post-period. Four outliers (with residuals greater than 1.5 in absolute value) have been used for estimation but dropped from the plot.

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. In the Appendix (Figure A5), we show 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 30 parameter combinations.

Table 4 shows the estimated treatment effect as it varies across the B and M parameters. The estimates are generally close to the true short-run effect, although whether they are larger or smaller than the true short-run effect depends on the decay process and the underlying data. Additionally, while Table 4 uses the BIC-chosen polynomial order, we also explore (in the Appendix) how the polynomial order selected impacts the estimated treatment effect. Overall, the potential for problems in the estimate depends on the shape of the decay function, how quickly the effect decays, and the RDiT specification.

3.2.2 Autoregression

We next simulate an AR(1) process, with varying degrees of dependence. This would occur, for instance, in an ambient air quality setting, where pollution does not immediately dissipate. The process is given by the following equation:

$$y_{i,t} \equiv \alpha \cdot y_{i,t-1} + \beta \cdot \mathbb{1}\{t \geq t_{start}\} + \varepsilon_{i,t} \quad (3)$$

where the error component is standard normal. We allow for a burn-in period of approximately 20,000 observations. We fix $\beta = -0.2$ but allow α to vary from 0.1 to 0.9 in increments of 0.2.¹⁹ Table 5 shows the estimated treatment effect when the autoregressive component is not included in the regression. Here the true treatment effect grows with time;

¹⁹As described in 2.2.2, we estimate historical average AR(1) parameters ranging from 0.3-0.5 for daily data.

Table 4: 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 (0.11)	-0.21 (0.11)	-0.21 (0.11)	-0.21 (0.11)	-0.21 (0.11)	-0.21 (0.11)
-0.0075	-0.22 (0.11)	-0.21 (0.11)	-0.21 (0.11)	-0.21 (0.11)	-0.21 (0.11)	-0.21 (0.11)
-0.01	-0.23 (0.11)	-0.22 (0.11)	-0.21 (0.11)	-0.21 (0.11)	-0.21 (0.11)	-0.20 (0.11)
-0.0125	-0.23 (0.12)	-0.23 (0.11)	-0.21 (0.11)	-0.21 (0.11)	-0.21 (0.11)	-0.20 (0.11)
-0.015	-0.22 (0.12)	-0.24 (0.11)	-0.21 (0.12)	-0.20 (0.11)	-0.21 (0.11)	-0.20 (0.11)
B. Simulated Data						
	M parameter					
B parameter	0	200	400	600	800	1000
-0.005	-0.23 (0.07)	-0.24 (0.08)	-0.24 (0.08)	-0.23 (0.07)	-0.22 (0.07)	-0.21 (0.07)
-0.0075	-0.19 (0.08)	-0.23 (0.08)	-0.25 (0.08)	-0.25 (0.08)	-0.23 (0.08)	-0.21 (0.08)
-0.01	-0.16 (0.08)	-0.21 (0.08)	-0.25 (0.07)	-0.26 (0.08)	-0.25 (0.08)	-0.22 (0.08)
-0.0125	-0.13 (0.08)	-0.20 (0.08)	-0.24 (0.08)	-0.26 (0.08)	-0.26 (0.08)	-0.23 (0.08)
-0.015	-0.11 (0.08)	-0.19 (0.08)	-0.23 (0.08)	-0.26 (0.08)	-0.26 (0.08)	-0.24 (0.08)

Note: Each cell of this table reports the mean treatment effect coefficient and number of observations across 2,080 regressions. 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 A5.

Table 5: Dropping AR(1) Term, $\beta_{initial} = -0.2$ and $\beta_{long-run} = \frac{-0.2}{1-\alpha}$

AR(1) parameter α	Implied long-run treatment effect		BIC-chosen Global Polynomial	BIC-chosen Separate Polynomials	Local Linear	Augmented Local Linear	Pre/Post, Monthly Observations
0.1	-0.22	Estimated beta (S.D.)	-0.22 (0.08)	-0.22 (0.08)	-0.21 (0.55)	-0.22 (0.04)	-0.21 (0.08)
0.3	-0.29	Estimated beta (S.D.)	-0.29 (0.11)	-0.29 (0.13)	-0.26 (0.69)	-0.29 (0.05)	-0.28 (0.10)
0.5	-0.40	Estimated beta (S.D.)	-0.40 (0.17)	-0.41 (0.21)	-0.33 (0.90)	-0.40 (0.07)	-0.39 (0.14)
0.7	-0.67	Estimated beta (S.D.)	-0.68 (0.39)	-0.65 (0.42)	-0.46 (1.34)	-0.67 (0.12)	-0.65 (0.24)
0.9	-2.00	Estimated beta (S.D.)	-1.85 (1.37)	-1.86 (1.37)	-0.53 (2.38)	-2.00 (0.35)	-1.96 (0.70)

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.

it is accordingly somewhat akin to the time-varying treatment described above. While the treatment effect is equal to β on day 1, the accumulated effect on day 2 is equal to $\beta + \beta \cdot \alpha$, etc. The long-run accumulated effect is $\frac{\beta}{1-\alpha}$.

In general, the estimated effect approximates the long-run, rather than the short-run, effect. In this context, the long-run treatment effect may be more policy-relevant, since the treatment effect will approach this long-run impact after only a few days. However, this aspect of the estimates in Table 5 highlights that identification of the treatment effect in an RDIT setting is not achieved narrowly at the cut-off, but rather relies on observations away from the discontinuity itself. We examine the extent to which the estimate varies with the sample window length, finding that the length of the sample matters less than the degree of auto-regression.²⁰ Some sample window lengths (in particular, samples much shorter than the eight year windows we show in Table 5) yield estimated treatment effects that are larger (in absolute terms) than the true long-run effect. This would be consistent with the results

²⁰Full results available upon request.

shown in Table 3, in which a time-varying treatment unknown to the researcher leads to estimates that are too large.

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 A5). As expected, the BIC-chosen global polynomial gives accurate estimates $\hat{\beta}$ and $\hat{\alpha}$, allowing the researcher to give both the true short-run and long-run effects. The local 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.

The results do not invalidate the literature using RDiT in the context of air pollution. First, the accumulated effect may be of more policy relevance than the immediate effect. 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. Thus papers that fail to incorporate the dynamics may be approximating the treatment effect that is of greater relevance. Moreover, we do not claim that the dispersion of pollutants in the atmosphere can be completely modeled by the inclusion of a lagged dependent variable. Rather, 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 may wish to model more explicitly.

In this section we have explored how an autoregressive process ought to be considered by the researcher. 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 misspecification. 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.

4 Recommendations and Conclusions

The empirical researcher wishing to use RDiT in an $N = 1$ setting ought to expose her assumptions to many opportunities to fail. Below, we offer a checklist along the lines of what is suggested in Lee and Lemieux (2010), but that applies in the context of RDiT. Recall that we have identified three main features of the RDiT setting that may induce bias: time-varying treatment effects, autocorrelation, and manipulation of behavior near the threshold (e.g. selection, anticipation, avoidance). 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), they serve a different purpose in the context of time-series data. Additional suggestions from Lee and Lemieux (2010) should also be deployed whenever relevant (e.g. presenting the main RD plot using binned local averages on the full support of the active sample).

1. Present several robustness checks:
 - (a) Global polynomial versus local linear.
 - (b) Order of polynomial. Overlay lines of best fit (lowess) corresponding to various choices of polynomial time control onto the main RD plot, or simply present a table reporting the treatment effect for each polynomial order.

- (c) Bandwidth. Present a graph of the treatment effect estimated using a variety of bandwidths (see Figure 18 in Lee and Lemieux (2010)).
2. Placebo tests. Estimate a parallel RD:
 - (a) On nearby geographical areas that weren't subject to the treatment.
 - (b) Using other dates.
 3. Plot a parallel RD estimated on control variables to demonstrate continuity.
 4. Estimate a “donut” RD (see Barreca et al. (2011)) to mitigate concerns about short-run selection/anticipation/avoidance effects.
 5. Test for the presence of autocorrelation 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.
 6. Compare results from an event study graph (i.e., with period-specific dummy variables) to the various polynomial and local linear control specifications. If results differ, it may be a sign of time-varying treatment effects.
 7. Consider deploying our “augmented local linear” methodology in order 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. This 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. 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 DGP will lead to different potential biases (or require stronger assumptions).

In this paper we have articulated reasons why using time as the forcing 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 far away from the discontinuity, and should explore whatever options are available to mitigate concerns about autocorrelation and dynamic behavior. While “regression discontinuity” designs have been a tremendously valuable addition to the suite of valid identification strategies, we must also recognize that there is substantial heterogeneity in the credibility of results that arise from the various approaches under the RD umbrella.

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.

- Fisher, Anthony, Michael Hanemann, Michael Roberts, and Wolfram Schlenker**, “The Economic Impacts of Climate Change: Evidence from Agricultural Output and Random Fluctuations in Weather: Comment,” *American Economic Review*, 2012, 102 (7), 3749–3760.
- 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.
- 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.

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.

Paola, Maria De, 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.

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.

Appendix

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

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 A2 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 A3 shows a similar plot for 9 different polynomial orders. The range over which the bias is in this direction depends on the polynomial order, as can be seen in this figure. Finally, Figure A4 shows the estimate for the BIC-chosen polynomial.

Similarly, the main text shows (Table 4) the mean estimated treatment effect for different parameterizations of a smooth decay function. Figure A5 shows four representative combinations of parameters for this logistic decay function. Then, Figure A6 shows the estimated treatment effect across various values of the B parameter (the horizontal axis), for different orders of polynomial (the 9 panels). As with the sharp decay, the estimate of the treatment effect varies with the order of the polynomial. Figure A7 shows the estimated treatment effect for the BIC-chosen polynomial across values of the B parameter (horizontal axis). For these data, the BIC-chosen polynomial is close to the true short-run effect of -0.2, no matter how quickly the effect decays.

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 A3 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.

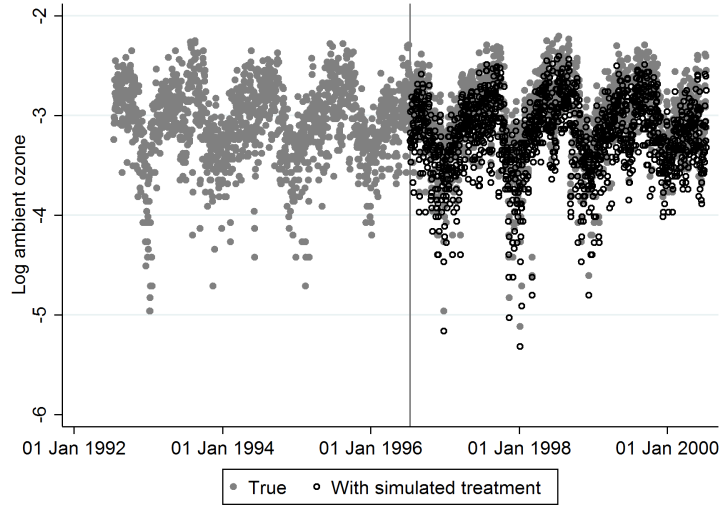
Finally, we show additional results relating to the autoregressive processes described in the main text. Table 1 in the main text shows estimated autoregressive parameters for six

pollutants in the U.S. For each pollutant, we obtained daily monitor readings for 2004-2005 from the EPA (https://aqs.epa.gov/aqsweb/documents/data_mart_welcome.html). We obtained temperature and windspeed data from the same source. For each monitor-pollutant pair, we then regressed 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.

Unfortunately, the EPA data do not include precipitation. To verify that results are not sensitive to the inclusion of precipitation, we matched 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 PM_{2.5} 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 A4, are very similar to the results presented in the main text.

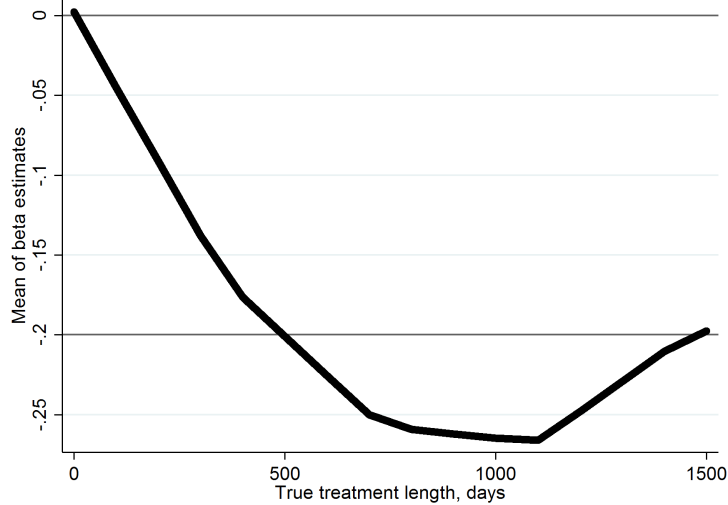
While the main text (Table 5) showed the impact of incorrectly dropping an autoregressive term, Table A5 shows estimation results when the AR(1) term is included. In general, this performs better (as expected), since both short-run and long-run effects can be calculated. The local linear (Column 3) gives 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.

Figure A1: Air Quality Data with and without Simulated Treatment



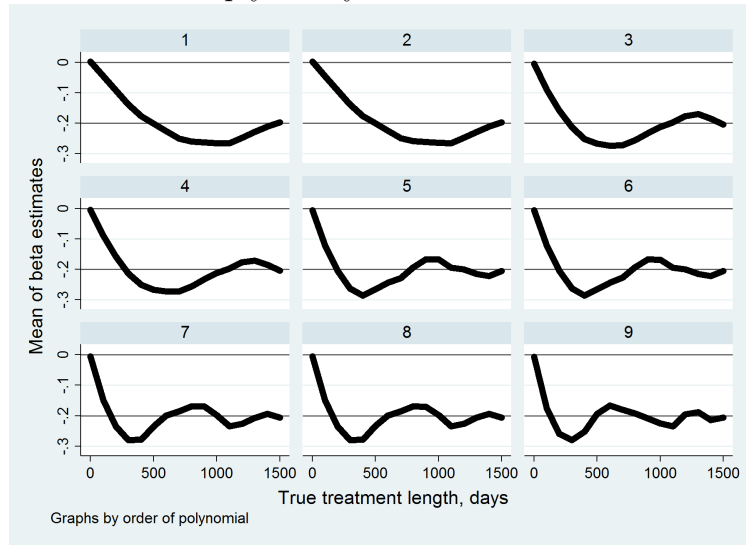
Note: This figure plots the true air quality data for a representative monitor in grey, and the simulated impact (in black) of a treatment effect of -0.2.

Figure A2: Mean Estimated Treatment Effect, RDiT Estimate using Polynomial Order 1, When Treatment Sharply Decays



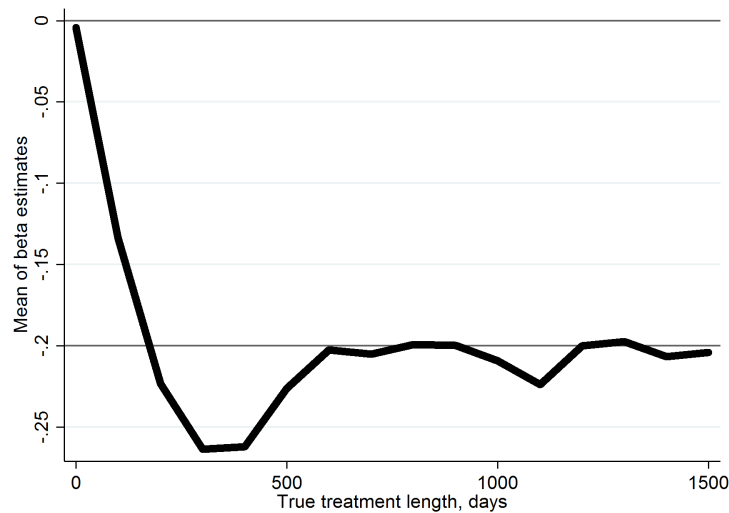
Note: This figure plots the mean estimated treatment effect as the true treatment length varies (across the x -axis). The treatment effect is estimated using an RDiT method with global polynomial of order 1, and the researcher is assumed to not know that the true treatment decays after x days.

Figure A3: Mean Estimated Treatment Effect, RDiT Estimate using Polynomial Orders 1 through 9, When Treatment Sharply Decays



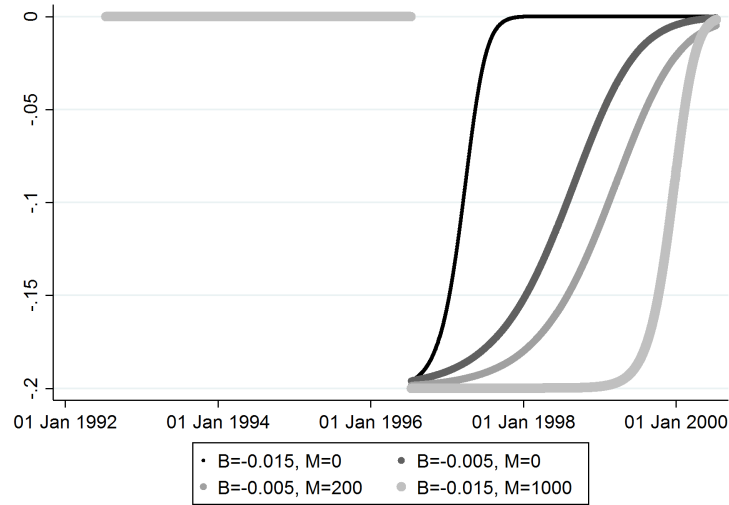
Note: This figure plots the mean estimated treatment effect as the true treatment length varies (across the x -axis). The treatment effect is estimated using an RDiT method with global polynomial of varying order (across the panels), and the researcher is assumed to not know that the true treatment decays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

Figure A4: Mean Estimated Treatment Effect, RDiT Estimate using BIC-Chosen Polynomial Order, When Treatment Sharply Decays



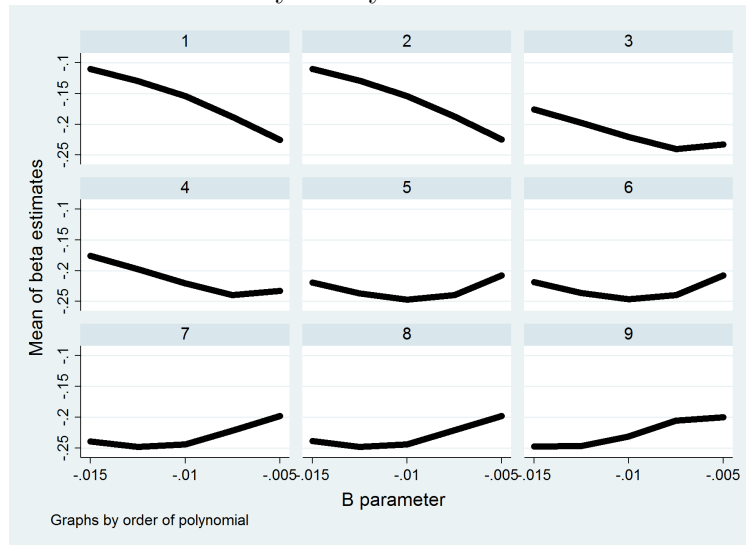
Note: This figure plots the mean estimated treatment effect as the true treatment length varies (across the x -axis). The treatment effect is estimated using an RDiT method with global polynomial with BIC-chosen order, and the researcher is assumed to not know that the true treatment decays after x days.

Figure A5: Generalized Logistic Decay Functions



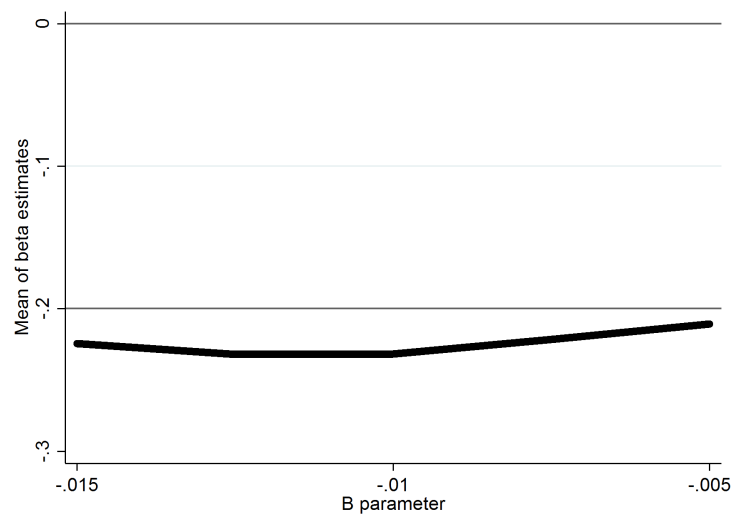
Note: This figure plots four logistic decay functions, using two sets each of parameters B and M (see text for details).

Figure A6: Mean Estimated Treatment Effect, RDiT Estimate using Polynomial Orders 1 through 9, When Treatment Smoothly 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 method with global polynomial of varying order (across the panels), and the researcher is assumed to not know that the true treatment decays following a smooth process (generalized logistic decay; see text for details).

Figure A7: Mean Estimated Treatment Effect, RDiT Estimate using BIC-Chosen Polynomial Order, When Treatment Smoothly 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 method with global polynomial of BIC-chosen order, and the researcher is assumed to not know that the true treatment decays following a smooth process (generalized logistic decay; see text for details)

Table A1: Air Quality Data Summary Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
Ozone	693,835	0.046	0.019	0.000	0.250
Daily Min. Temperature	693,835	51	18	-30	96
Daily Max. Temperature	693,835	72	19	-14	121
Daily Mean Temperature	693,835	62	18	-21	105

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.

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
<u>A. Air Quality Data</u>					
Estimated Beta	-0.05	-0.05	-0.20	-0.18	<0.01
S.D.	(0.11)	(0.12)	(0.29)	(0.14)	(0.11)
Observations	2822	2822	59	2822	96
	(72)	(72)	(3.5)	(72)	(1.0)
Polynomial Order	4.1	1.7, 1.7	-	-	-
	(2.9)	(0.8), (0.8)			
Treatment Length	30	30	30	30	30
<u>B. Simulated Data</u>					
Estimated Beta	-0.02	-0.02	-0.20	-0.20	<0.01
S.D.	(0.07)	(0.08)	(0.51)	(0.26)	(0.08)
Observations	2921	2921	61	2921	96
	-	-	-	-	-
Polynomial Order	1.0	1.0, 1.0	-	-	-
	(0.05)	(0.03), (0.09)			
Treatment Length	30	30	30	30	30

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 Has 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
<u>A. Air Quality Data</u>					
Estimated Beta	-0.20	-0.20	-0.11	-0.19	-0.19
S.D.	(0.11)	(0.11)	(0.29)	(0.12)	(0.10)
Observations	2822	2822	59	2822	96
	(72)	(72)	(3.5)	(72)	(1.0)
Polynomial Order	4.0	1.7, 1.6	-	-	-
	(2.9)	(0.8), (0.8)			
Phase-In Length	7	7	7	7	7
<u>B. Simulated Data</u>					
Estimated Beta	-0.20	-0.20	-0.12	-0.19	-0.19
S.D.	(0.07)	(0.08)	(0.51)	(0.16)	(0.08)
Observations	2921	2921	61	2921	96
	-	-	-	-	-
Polynomial Order	1.0	1.0, 1.0	-	-	-
	(0.05)	(0.03), (0.06)			
Phase-In Length	7	7	7	7	7

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 A4: 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 coefficients, standard errors, number of observations, and R2 values 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 A5: Including AR(1) Term

		BIC-chosen Global Polynomial	BIC-chosen Separate Polynomials	Local Linear	Augmented Local Linear	Pre/Post, Monthly Observations
AR(1) parameter α						
0.1	Estimated beta	-0.20	-0.20	-0.20	-0.20	-0.22
	(S.D.)	(0.07)	(0.07)	(0.54)	(0.04)	(0.08)
	Estimated alpha	0.10	0.10	0.04	0.10	-0.03
	(S.D.)	(0.02)	(0.02)	(0.12)	(0.02)	(0.10)
0.3	Estimated beta	-0.20	-0.20	-0.20	-0.20	-0.28
	(S.D.)	(0.07)	(0.07)	(0.55)	(0.04)	(0.11)
	Estimated alpha	0.30	0.30	0.23	0.30	-0.02
	(S.D.)	(0.02)	(0.02)	(0.12)	(0.02)	(0.10)
0.5	Estimated beta	-0.20	-0.20	-0.21	-0.20	-0.39
	(S.D.)	(0.07)	(0.07)	(0.58)	(0.04)	(0.15)
	Estimated alpha	0.50	0.50	0.41	0.50	-0.01
	(S.D.)	(0.02)	(0.02)	(0.12)	(0.02)	(0.10)
0.7	Estimated beta	-0.20	-0.20	-0.23	-0.20	-0.64
	(S.D.)	(0.07)	(0.07)	(0.62)	(0.04)	(0.24)
	Estimated alpha	0.70	0.70	0.59	0.70	0.01
	(S.D.)	(0.01)	(0.01)	(0.11)	(0.01)	(0.10)
0.9	Estimated beta	-0.21	-0.21	-0.26	-0.21	-1.65
	(S.D.)	(0.08)	(0.08)	(0.71)	(0.04)	(0.62)
	Estimated alpha	0.90	0.90	0.77	0.90	0.16
	(S.D.)	(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 5. In contrast to Table 5, the researcher correctly models the AR(1) process.