Optional ERIC Coversheet — Only for Use with U.S. Department of Education Grantee Submissions

This coversheet should be completed by grantees and added to the PDF of your submission if the information required in this form **is not included on the PDF to be submitted**.

INSTRUCTIONS

- Before beginning submission process, download this PDF coversheet if you will need to provide information not on the PDF.
- Fill in all fields—information in this form **must match** the information on the submitted PDF and add missing information.
- Attach completed coversheet to the PDF you will upload to ERIC [use Adobe Acrobat or other program to combine PDF files]—do not upload the coversheet as a separate document.
- Begin completing submission form at https://eric.ed.gov/submit/ and upload the full-text PDF with attached coversheet when indicated. Your full-text PDF will display in ERIC after the 12-month embargo period.

GRANTEE SUBMISSION REQUIRED FIELDS

Title of article, paper, or other content

Using Simulation to Analyze Interrupted Time Series Designs

All author name(s) and affiliations on PDF. If more than 6 names, ERIC will complete the list from the submitted PDF.

Last Name, First Name	Academic/Organizational Affiliation	ORCID ID
Miratrix, Luke	Harvard Graduate School of Education	0000-0002-0078-1906

Publication/Completion Date—(if *In Press,* enter year accepted or completed) published online May 28, 2022

Check type of content being submitted and complete one of the following in the box below:

☑ If article: Name of journal, volume, and issue number if available
□ If paper: Name of conference, date of conference, and place of conference
□ If book chapter: Title of book, page range, publisher name and location
□ If book: Publisher name and location
□ If dissertation: Name of institution, type of degree, and department granting degree

Evaluation Review, Vol. 46(6) 750–778

DOI or URL to published work (if available) https://journals-sagepub-com.ezp-prod1.hul.harvard.edu/doi/10

Acknowledgement of Funding— Grantees should check with their grant officer for the preferred wording to acknowledge funding. If the grant officer does not have a preference, grantees can use this suggested wording (adjust wording if multiple grants are to be acknowledged). Fill in Department of Education funding office, grant number, and name of grant recipient institution or organization.

"This work was supported by U.S. Department of Education **[Office name]** ______ through **[Grant number]** ______ to **Institution]** ______.The opinions expressed are those of the authors and do not represent views of the **[Office name]** ______ or the U.S. Department of Education.

Original Research Article

Evaluation Review
2022, Vol. 46(6) 750–778
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0193841X221101286
journals.sagepub.com/home/erx
SAGE

Using Simulation to Analyze Interrupted Time Series Designs

Luke W. Miratrix[1]

Abstract

We are sometimes forced to use the Interrupted Time Series (ITS) design as an identification strategy for potential policy change, such as when we only have a single treated unit and cannot obtain comparable controls. For example, with recent county- and state-wide criminal justice reform efforts, where judicial bodies have changed bail setting practices for everyone in their jurisdiction in order to reduce rates of pre-trial detention while maintaining court order and public safety, we have no natural and available comparison group other than the past. In these contexts, it is imperative to model pre-policy trends with a light touch, allowing for structures such as autoregressive departures from any pre-existing trend, in order to accurately and realistically assess the uncertainty of our projections. We aim to provide a methodological approach rooted in commonly understood and used modeling tools to achieve this. We quantify uncertainty with simulation, generating a distribution of plausible counterfactual trajectories to compare to the observed; this approach naturally allows for incorporating seasonality and other time-varying covariates, and provides confidence intervals along with point estimates for the potential impacts of policy change. We find simulation provides a natural framework to capture and show uncertainty in the ITS designs. It also allows for easy extensions such as nonparametric smoothing in order to handle multiple post-policy time points.

[1] Harvard Graduate School of Education, Cambridge, MA, USA

Corresponding Author:
Luke Miratrix, Department of Statistics, Harvard Graduate School of Education, 14 Appian Way, Cambridge, MA 02138-3752, USA.
Email: lmiratrix@g.harvard.edu

Keywords

Neyman-Rubin causal model, single unit case study analysis, ITS designs, criminal justice reform, posterior predictive checks, monte carlo simulations

Introduction

Currently, in the U.S., hundreds of thousands of people are incarcerated in local jails on any given day as they await resolution of their criminal case. These people have not been convicted, but are nonetheless incarcerated because, generally, they cannot afford to post monetary bail to secure their release (Zeng, 2018). Several jurisdictions have sought to improve these judicial systems building procedures to increase the rate of release for "low-risk" defendants. One general category of such reforms use risk assessment tools in early court proceedings, providing judges with information about the risk of a defendant, as measured by various characteristics such as previous criminal history, in order to improve judicial decision-making regarding what types of supervision or restrictions should be placed on defendants awaiting their case resolution.

This is the context for this work. We use data from two such reform efforts, one in Mecklenberg County, NC (Redcross et al., 2019), and one in the state of New Jersey (Golub et al., 2019). There are several primary outcomes of interest, of which we examine two: the proportion of arrestees assigned monetary bail, and the total number of warrant arrests made. We have three obstacles for rigorous evaluation. First, we have only a single treated unit (the county or state) in each case. Second, in both cases data on similar counties or states were not available due to reasons both pragmatic (the difficulty in collecting and marshalling such data in the first place was insurmountable given budget limitations) and structural (even basic elements of the data, such as the definitions of severity of cases, the criminal justice codes, or the management of cases in the judicial system, are not directly comparable across region). Third, we do not necessarily expect any impact of the policy right at the time of the policy change, as it may take time for the policy to become fully implemented, and for the consequences of the policy to be felt. More broadly, we are concerned that the impact of the policy itself may evolve over time, as the policy becomes institutionalized.

Given these three very serious limitations—a single treated unit, no available comparison units, and no sudden onset of treatment impact—how should an evaluator proceed? Causal impacts should be estimated via principled comparison of units experiencing some nominal treatment to those that did not. In this case, for example, the evaluator would ideally compare how a county undergoing a reform effort evolved over time to how some set of comparison counties deemed similar evolved. This could be done with a comparative interrupted time series, or difference-in-difference, or, perhaps, a synthetic control approach. But we have no comparison counties available. One might instead use something akin to a Regression Discontinuity Design, where time points just before the policy are compared to those just after; this unfortunately is also off the table due to potentially no sudden onset of treatment effects, and for wanting to estimate policy impacts at a sequence of time points into the future, to understand its evolution. Our three limitations render the usual tools for impact evaluation unavailable, forcing us to project what might have happened, based on historic data of the target region, as a means of generating a counterfactual. A common tool for this job is the Interrupted Time Series (ITS) (Ferron & Rendina-Gobioff, 2005).

This situation is far from ideal. An ITS approach is exposed to an extreme number of threats to validity: it is heavily model dependent (necessary for extrapolation), it assumes there are no time dependent shocks (events) that concur with treatment onset that themselves could impact the outcome, it is noisy due to the single unit of analysis, and the policy of interest is likely to be confounded with concurrent policy change efforts. We recommend instead, if at all possible, obtaining comparison units, multiple treated units, or some sort of plausible (quasi)random assignment. But failing this, ITS can be useful for contexts where the size of the impact would be large relative the uncertainty and size of likely biases present.

For an ITS design, we assume the researcher has observed regular measures of some outcome of interest both for several time points before such a change as well as after. The research question is whether there is any evidence that the policy has changed the course of the unit of interest. The researcher, implicitly or explicitly, builds a model based on history that forecasts what we would expect to happen post-policy. This is then compared to what actually happened. If the differences are large, and unlikely to be due to statistical fluctuations, we can tentatively claim that something—the policy or events concurrent to the policy—changed the course of our analyzed unit.

Perhaps the most used analytic approach for ITS is to fit a simple linear regression to the data, regressing the outcome of interest onto time and a series of dummy variables for each time point post-policy. The estimates of these dummy variables then provide impact estimates for each post-policy point. Unfortunately, even if the underlying linear trend were fundamentally sound, and all the assumptions discussed above were met, the deviations from trend are likely correlated and this correlation needs to be taken into account. Not doing so correctly will undermine any estimates of uncertainty by giving overly precise (too small) standard errors.

We propose to account for local dependencies by fitting an autoregressive model with linear trend to the pre-policy data, and then using that model to simulate, using a pseudo-Bayesian approach discussed in Gelman and Hill (2006) a distribution of plausible post-policy trajectories that we would expect if pre-policy trends continued unabated. By comparing this distribution to the observed post-policy trend, we can estimate impacts and test for the significance of impacts, given the set of rather stringent assumptions necessary for an ITS analysis. We can also calculate confidence intervals to assess ranges of impact. This simulation procedure takes into account the uncertainty of the linear model estimate, uncertainty in the measurement of the outcomes, and any autoregressive dependencies in the residuals.

Simulation also allows for several natural extensions. First, we can easily incorporate covariates to capture nonlinearities (in particular, seasonal trends). Second, we can average, or smooth, multiple months of potentially heterogeneous impacts typically found in such evaluations to better capture post-policy impacts in interpretable ways. This allows testing whether a *group* of post-policy time points differs statistically significantly from what would have occurred in the absence of an intervention. Simulation also makes statistical inference more explicit: we directly see that a “statistically significant” effect is one that would not likely occur as a natural extension of the pre-policy trend; we believe this puts the assumptions (in particular that of extrapolation) more firmly in the forefront of the analysis.

The idea for simulation for assessing uncertainty in these contexts is not new; see, for example, Zhang et al. (2009) who use a parametric bootstrapping approach. Our method is also a parametric simulation approach, but we explicitly include autoregressive dependencies and explicitly simulate post-policy trajectories. We also discuss the estimands of interest more explicitly. Relative to us, Brodersen et al. (2015) propose a more complex, fully Bayesian time-series approach, implemented with the *causalImpact* package, that relies on modeling a latent state space.

Instead of simulation, one could directly fit a linear regression model, in particular a generalized linear model with an autoregressive residual structure. We believe, however, that several of the above benefits of simulation would then be either less accessible or not possible. In particular, simulation allows us to easily summarize and visualize heterogeneous impacts post-policy, and in general is an approach that can enrich classic inference (King, Tomz, and Wittenberg, 2000). Simulation also allows for our smoothing approach, which can increase the power to detect an effect. Simulation also allows for direct hypothesis testing without placing any model on the form of the treatment impact itself (otherwise one would have to generate a theoretical prediction interval to compare to the observed data). All of this being said, without smoothing we would expect agreement between these approaches.

ITS is also, of course, based on the idea of a time series. Classic time series methodology, for example, ARIMA models, could account for linear trend by differencing the observations and then modeling the resulting differences as, ideally, a stationary time series. Selecting and fitting ARIMA models, however, typically require much denser time-series data (more time points) than is typically available in our context. Furthermore, as ARIMA approaches are quite distinct from the classic linear modeling approaches more familiar with policy evaluators, we believe they are less accessible as a tool for evaluation research. They also do not allow for, as far as we know, the smooth approach we present. Given these concerns, we build on the linear modeling approaches found in the public policy ITS literature. For more on the ARIMA direction, however, see Stoffer and Shumway (2006).

More broadly, ITS is a worse (in terms of strength of evidence) version of *Comparative* Interrupted Time Series (CITS) analyses, where we compare the target treatment series to those of comparison units that are not treated. For an overview of CITS, consider Somers et al. (2013) or Hallberg et al. (2018). Also see Jacob et al. (2016), who evaluate the CITS by comparing its findings to those utilizing the more widely-accepted Regression Discontinuity Design. For a detailed case study with CITS in the context of experimental trials, see Bloom et al. (2005); this example has ties to ITS as they fit regressions to the sequence of paired differences.

In terms of modeling, we also advocate for using *only* the pre-policy data in the modeling process, and then using the resultant model to extrapolate what would have happened, absent policy, to the post-policy era. The difference, then, between the observed and the imputed is the estimated change. Other modeling approaches jointly model both pre- and post-policy, which imposes a model (explicit or implicit) of the treatment effect or deviation itself. We argue from the potential outcomes view that this is not desirable, in that the post-policy observations potentially contaminated with unknown and likely time-varying treatment should not themselves influence assessment of a pre-policy trend. Of course, given a substantive model, one might instead choose to directly use these post-policy data. For example, in some ITS analyses that capture multiple points of treatment, one may elect to do a full model of the entire series across multiple interruptions. Our focus is on a single interruption.

In this paper we first lay out the ITS problem and its classic treatment, using the potential outcomes framework to make quantities such as the target impact parameters and necessary assumptions explicit. We then describe the simulation procedure that allows for a simple autoregressive structure, illustrating with an example taken from the Mecklenberg County evaluation. We then provide our two primary extensions mentioned above—seasonality and smoothing—in the following two sections We offer some general cautions and concluding remarks at the end. Our Online Only Supplemental Material gives further justification of the modeling choices we suggest, provides some mathematical derivations, and gives a brief overview of the accompanying publicly available R package, sim ITS, posted on the CRAN archive, that implements the methods discussed along with all routines needed to conduct a full and transparent ITS analysis. This sim ITS package overview walks through how to fit an interrupted time series to a given dataset using the approach we here present.

Throughout the paper, we do not focus on the causal inference reasoning behind ITS: our work focuses on assessing whether there was change; the question as to why requires additional work, and for that we refer the reader to sources such as Cook, Campbell, and Shadish (2002). Causal interpretation of an ITS finding can be quite fragile; see, for example, Baicker and Svoronos (2019). In our view, an ITS approach is a method of last resort: one should, if at all possible, generally locate comparison units for a more rigorous evaluation of treatment impact.

Notation and Setup

We have a single treated region. We observe this region at several time points before treatment (e.g., a policy change) as well as for several time points after. For example, consider Figure 1a and b, showing two time series: the proportion of all arrests in Mecklenberg for each month for a period before and after a reform effort (Redcross et al., 2019), and the total number of warrant arrests each month before and after a major reform effort in New Jersey (Golub et al., 2019).

Based on the trend of the unit before the policy change, we will extrapolate to determine what we would see post policy, had business continued as usual. For example, if we have observed a steady but slow increase in our outcome, we would project that steady but slow increase into the post-policy period. If what we actually observe deviates from that projected trend, we know that something has changed our system to cause this departure. The core assumption behind an Interrupted Time Series design is stability; everything rests on the assumption that, absent any impact, our unit would evolve as it has been.

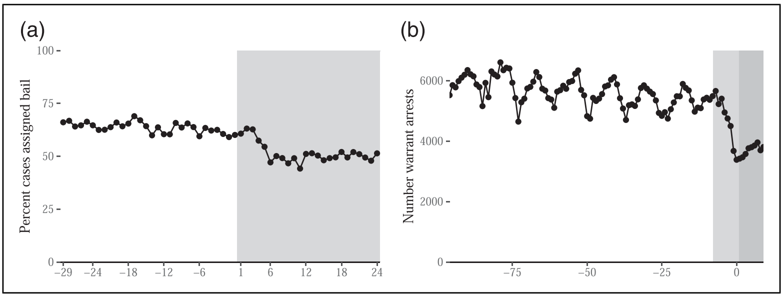

Figure 1. Two example Interrupted Time Series.
Comment: The dark gray indicates the post-policy era. Policy onset is $t_0 = 0$ in these figures, with pre-policy time being non-positive. Right side shows evident seasonality. Left side suggests some autocorrelation which may be due to seasonality or other unknown factors.

We borrow from the potential outcomes viewpoint (for an overview, see Imbens and Rubin (2015) or Rosenbaum (2009)) to make the above more precise. We have a single unit, and we can either treat it (invoke policy change) at time t_0, or not treat it at all. Let $Y_t(0)$, $t = t_{min},\ldots,t_{max}$, be the sequence of outcomes we would observe if we did not ever treat our unit.[1] Let the corresponding $Y_t(1)$ be the outcomes we would observe if we did treat the unit at t_0. We could allow $Y_t(1) \neq Y_t(0)$ for $t \leq t_0$ if we allowed anticipatory effects of treatment, that is, if the unit knows it will be treated it may change before the time of treatment. In this work, we make the further assumption, however, that there is *no anticipation of treatment*, that is, that $Y_t(1) = Y_t(0)$ for all $t \leq t_0$. In some cases, to achieve this assumption, one can move the point of treatment onset earlier, for example, to when a policy was initially being planned rather than its official adoption date. One could, in principle, model anticipation of treatment, but that would require having multiple units or a strong theoretical model for the impact. See, for example, Clark et al. (2008).

The impact of policy at a specified time t is then $\Delta_t \equiv Y_t(1) - Y_t(0)$. Our observed data consist of a single treated unit, so the $Y_t(1)$ are observed for all $t > t_0$. If we had the ability to estimate $Y_t(0)$ we could immediately estimate Δ_t. This converts our estimation problem to a missing data problem (Rubin, 2005): what are reasonable values for the $Y_t(0)$ for $t > t_0$? Within this framework, uncertainty around the difference is entirely dependent on uncertainty in our estimation of $Y_t(0)$.

ITS analysis estimates the $Y_t(0)$ by fitting a trend (i.e., model) to the pre-policy data and extrapolating to post-policy timepoints. We next discuss how this estimation is typically conducted and identify some problems with it. We then offer an augmented modeling approach with corresponding inference procedures.

Our data itself consists of a collection of observed arrests for each observed month. Our region-level potential outcomes, depicted on Figure 1a and b, are aggregates, aggregating individual data within each month, although we could work to fit the multilevel structure instead. An individual-level analysis would bring in further complexity from, for example, individuals being in multiple months (e.g., from multiple arrests), and unknown correlation structure of individuals within a given month; aggregation avoids this. Furthermore, migration of individuals into and out of the policy region could further exacerbate the difficulties with individual trend approaches. The aggregation avoids these problems by focusing on the "health" of the policy unit rather than the impact on individuals. Results are then regarding changes at the larger unit level, which can impact interpretation. That being said, without strong individual-level predictors, aggregation will surprisingly not have a high cost in power; the variation in the month-to-month averages is a reflection of individual variation (as well as shared month shocks) and so while we have fewer month-level data points, we also have less residual noise for those points. See Angrist and Pischke (2008), Chapter 3, for a more detailed discussion of this principle. For further discussion on aggregation see Bloom et al., (2005). For some dangers with aggregation if the number of units being aggregated changes significantly, see Ferman and Pinto (2019).

Classic ITS Analysis

In a classic ITS analysis one would fit the simple linear regression model of

$$Y_t = \beta_0 + \beta_1 t + \sum\nolimits_{k=t_0+1}^{t_{max}} \Delta_k \mathbf{1}_{\{t=k\}} + \epsilon_t \tag{1}$$

with $\epsilon_t \sim N(0,\sigma^2)$ and the $\mathbf{1}_{\{t=k\}}$ 0/1 indicators of whether $t = k$ for each post-policy time point k. This model will perfectly fit all post-policy months, meaning the estimates of β_0 and β_1 will only depend on pre-policy months. The $\widehat{\Delta}_k$ are then the specific impact estimates for each month k, capturing the departure of Y_t from the projected $\widehat{\beta}_0 + \widehat{\beta}_1 t$. Under a homoskedasticity assumption, we can obtain standard errors and conduct inference for the estimated Δ_k, because we assume the variation post-policy is the same as pre-policy. These standard errors will be driven by, and be no smaller than, $\widehat{\sigma}$, the estimated residual standard deviation (see Part A of the Online Only Supplemental Material for derivation).

Nearly equivalent to the above, one can simply fit the model to the pre-policy data only, dropping the post-policy dummy variables

$$Y_t(0) = \beta_0 + \beta_1 t + \epsilon_t \tag{2}$$

We then, for any point $t > t_0$ in the post-policy era, predict via extrapolation

$$\widehat{Y}_t(0) = \widehat{\beta}_0 + \widehat{\beta}_1 t$$

which results in an impact estimate at month t of

$$\widehat{\Delta}_t = Y_t^{obs} - \widehat{Y}_t(0)$$

These point estimates will be identical to the $\widehat{\Delta}_t$ from Model 1. However, Model 2 makes the connection to the potential outcomes framework most clear: our model predicts, via extrapolation, $Y_t(0)$ for all $t > t_0$. We fit our model to pre-policy data, data unaffected by the policy (by assumption), and then use our fitted model to impute (predict) the missing $Y_t(0)$ for $t > t_0$.

By contrast, instead of not using post-policy data in the fitting process, some will instead put a structure on the post-policy impact, such as with

$$Y_t = \beta_0 + \beta_1 t + \delta_0 \mathbf{1}_{\{t > t_0\}} + \delta_1 \mathbf{1}_{\{t > t_0\}}(t - t_0 - 1) + \epsilon_t \tag{3}$$

with $\mathbf{1}_{\{t>t_0\}}$ being a 0/1 indicator of t being after t_0, the end of the pre-policy era. Now the parameters δ_0 and δ_1 form a model of effects for the impact (in this case the impact begins at size δ_0 and grows by δ_1 each month, and $\Delta_t = \delta_0 + \delta_1(t - t_0 - 1)$ for $t > t_0$. This goes against the idea of imputing the missing $Y_t(0)$, however, as such a model allows the post-policy data, and policy impact, to inform the estimated residual variance.

Regression models produce valid inference under their modeling assumptions, in particular the strong assumption of the counterfactual linear trend continuing into the post-policy period. As a model check, the linear trend can be assessed in the pre-policy period; if there are strong deviations pre-policy, then extrapolation should be done with skepticism. The *causal* interpretation, however, relies on any found deviation being only explainable by the policy change; it is a substantive question whether there were other factors or changes that happened concurrently or after the policy reform, producing changes in outcomes that should not be ascribed to the policy.

One concern with simple regression is that there may be effects that operate in windows of time causing adjacent months to have similar outcomes beyond the underlying model. For example, the pattern of month-to-month averages in Mecklenberg (Figure 1a) could contain local correlations of months around what is a generally linear trend (we discuss the case of cyclic seasonal trends such as shown in Figure 1b).

If we do not model temporal dependence, we are assuming that, other than the underlying linear trend, there is no dependence between months beyond the explicit model. For example, if month t were surprisingly high, this would not imply any other month, such as month $t + 1$, would have any particular value. To produce more principled inference, we therefore extend Model 2 to allow for neighboring residuals to be correlated. This better captures how the time series can "wander" from the linear trend.

A simple approach is to model local dependence using an "AR1" model that uses the residual in the prior time period as a predictor of the residual of the next. For example, we can specify the residual of Model 2 to be

$$\epsilon_t = \rho\epsilon_{t-1} + \omega_t \; with \; \omega_t \sim N\left(0,\sigma^2\right) \tag{4}$$

The parameter ρ governs how much autocorrelation we have. If $\rho = 0$ the residuals are in fact independent. Higher values of ρ means deviations from trend tend to be similar, month-to-month. A $\rho > 1$ would mean a successive observation would be some percent larger than the last, in expectation, and thus the series would exponentially move away from the trend line; we therefore require $\rho < 1$.

An easy way of fitting such a model is to fit the lagged *outcome* model of

$$Y_t = \tilde{\beta}_0 + \tilde{\beta}_1 t + \tilde{\beta}_2 Y_{t-1} + \tilde{\epsilon}_t \text{ with } \tilde{\epsilon}_t \sim N\left(0, \tilde{\sigma}^2\right) \tag{5}$$

to the pre-policy time points $t = t_{min} + 1, \ldots, t_0$. The initial month has to be dropped as it has no lagged month. Up to how the parameters are interpreted, this model is equivalent to the lagged residual model. In particular, as the derivations in Part A of the Online Only Supplemental Materials show, we have $\rho = \tilde{\beta}_2$, $\beta_1 = \tilde{\beta}_1/(1-\tilde{\beta}_2)$ and $\beta_0 = \tilde{\beta}_0/(1-\rho) - \tilde{\beta}_1 \rho/(1-\rho)^2$. The residuals in the lagged outcome model are, under our residual autoregressive model, again independent, corresponding to the ω_t from Model 4. See Online Supplement A for additional discussion.

This model is a form of a finite distributed lag model; one could imagine including multiple lagged outcomes in the model if there were data availability. Other autoregressive models could also be used here, such as a moving average (MA) approach. We do not use the MA approach as we focus on modeling autoregression with lagged outcomes, and MA relies on dependencies of the latent residuals. One could instead use generalized least squares to specify the residual dependencies (see, e.g., Fox & Weisberg, 2018, for a practical overview in R). One could also use an ARIMA package fit to the prior data and forecast the counterfactual trajectory (see, e.g., Tashman, 2000, for an overview of forecasting); because this does not give a set of simulated trajectories, it would prevent the smoothing approach we later propose. Also, because we do not have ample time points, tuning such ARIMA models with measures of out-of-sample forecasting assessment, which is the purpose of the extrapolation approach, would be difficult.

In our approach, we fit a model to the pre-policy data only; the post-policy data (which could have arbitrary form in terms of trend and variation, depending on the impact of the policy) is set aside. Once this model is fit, we use it to extrapolate a reasonable counterfactual prediction of $Y(0)$ for any timepoint $T > t_0$ of interest. In the next section, we discuss how to do this with simulation.

Extrapolating Pre-policy Trends via Simulation

Impacts are estimated by extrapolating the pre-policy model to a post-policy timepoint, $T > t_0$, of interest. It not obvious how to use the model to form counterfactual predictions when using autoregressive structure. In particular, for $T > t_0 + 1$, if the treatment has impacted point $T - 1$, we cannot use the observed Y_{T-1} as our lagged covariate for our prediction because Y_{T-1} is not an observed $Y_{T-1}(0)$, but rather a $Y_{T-1}(1)$; any treatment impact in our lagged outcome will contaminate our imputation of $Y_T(0)$. Second, assessing uncertainty for a point T dependent on prior points is, mathematically, not entirely transparent. We therefore assess uncertainty and form predictions via simulation.[2] In the next subsection, we first consider the case where we are willing to assume the lagged model is correct and we knew with certainty the parameters $\tilde{\beta}_0$,$\tilde{\beta}_1$,$\tilde{\beta}_2$, and $\tilde{\sigma}^2$ of our lagged model. This case is not quite valid since we do not know these parameter values and so our uncertainty is not fully captured; we include it for clarity of exposition. We then, in the following subsection, extend to our actual proposed method that incorporates the additional uncertainty of these parameters.

Extrapolating With Known Parameters

We initially assume the model of equation (5) with parameters $\theta = (\tilde{\beta}_0,\tilde{\beta}_1,\tilde{\beta}_2,\tilde{\sigma}^2)$ of the pre-policy model are known. We also have observed Y_{t_0}, the last point in the pre-policy era.

Using this, we can simulate Y_{t_0+1} by drawing a new $\epsilon^*_{t_0+1} \sim N(0,\tilde{\sigma}^2)$ and calculating

$$Y^*_{t_0+1} = \tilde{\beta}_0 + \tilde{\beta}_1(t_0 + 1) + \tilde{\beta}_2 Y_{t_0} + \epsilon^*_{t_0+1}$$

This simulated outcome is a plausible post-policy (untreated) outcome, given our model. We can then simulate an outcome for $t_0 + 2$ using $Y^*_{t_0+1}$, drawing a new $\epsilon^*_{t_0+2}$ and adding up the components just as for $t_0 + 1$. Our second simulated outcome depends on our first. If our first is elevated due to a positive residual, our second will also be elevated. We then simulated our third, using the second, and continue in this manner until we reach T, and are left with a prediction for Y_T. By this point we have generated an entire sequence of plausible outcomes, given our model. Furthermore, this simulation process has fully captured the autoregressive structure. We could also extend to have non-normal errors, which might be relevant if in a context with extreme and rare events that we wanted to take into account. See, for example, Mohtadi and Weber (2021).

Our final prediction Y_T is a noisy prediction: it could be high or low depending on the residual draws. This noise is the key to capturing uncertainty. Both to get a more precise prediction and also to model the prediction uncertainty, we repeat the simulation process many times, for each iteration beginning at t_0 and Y_{t_0} and simulating a new time series. We then calculate the average of these series to get our final prediction

$$\widehat{Y}_T(0) = \frac{1}{R} \sum\nolimits_{r=1}^{R} Y_T^{*(r)}$$

where R is the total number of simulated series and r indexes these simulated series.

For inference, the middle 95% of our simulated $Y_T^{*(r)}$ forms a 95% prediction interval of what we would expect to see, $Y_T(0)$, had the pre-policy trend continued. If what we actually see, $Y_T^{obs} = Y_T(1)$, lies outside of this interval, we have evidence that something happened to change our model. This would be evidence of an impact of either the policy change or some other event within the system.

We can subtract the prediction interval from the observed Y_T to obtain a prediction interval for the deviation from the predicted trend (this is the quantity that could potentially be viewed as an impact). This prediction interval correctly captures the month-to-month variability of the observed trend; see Online Only Supplemental Material, Part A.

The major caveat to this process is we do not know the true parameters θ; we instead have an estimate $\widehat{\theta}$. If we simply plug in $\widehat{\theta}$ our inference will be overly optimistic as we have not taken uncertainty in the estimation of the parameters themselves into account; we do that next.

Incorporating Uncertainty in the Parameters

To capture parameter uncertainty we use a method rooted in Bayesian thinking and taken from Gelman and Hill (2006). It also has ties to the parametric bootstrap (see, e.g., Davison, 1997). The idea is this: instead of using $\widehat{\theta}$, draw a random vector of parameters θ^* for our model given our observed pre-policy data. This randomly drawn vector of parameters is itself a plausibly true value, just as we were drawing plausibly true values for the Y_t, above. We then simulate a sequence of Y_t^* using the simulation process described above but with θ^* (and still starting at Y_{t_0}) to get a plausibly true prediction conditional on the parameters. This two-step process captures the uncertainty in model estimation as well as uncertainty in extrapolation due to the autoregressive structure and residual error. The distribution of the Y_T^* over repeated iterations gives an overall predictive distribution that is integrated over both these components.

To get our distribution of plausible θ^*, we use the (estimated) standard errors from the original model fitting process. In particular, we draw a random $\beta^* = (\beta_0^*, \beta_1^*, \beta_2^*)$ vector from a multivariate normal centered at $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2)$ with a variance-covariance matrix based on the estimated variance-covariance matrix from the linear model fitting procedure (the σ^{2*} term is handled similarly). This is implemented using the *sim()* function in the R package *arm*. The *arm* package was written specifically for this form of uncertainty quantification, and is the companion package to Gelman and Hill (2006).

This approach is essentially Bayesian: the parameter draw step is similar to drawing a plausible value from a posterior distribution on the true θ (the implied prior here is implicitly a flat prior on the coefficients, roughly meaning that we are not differently preferring one value of θ over another). Under this view, the simulations constitute a posterior predictive distribution for Y_T and the $\widehat{Y}_T$ is the posterior mean predicted outcome given all the pre-policy data and the model (see Gelman, Meng, and Stern (1996) for a discussion of posterior predictive distributions). Further, under this view, the final prediction interval can be interpreted as a posterior predictive interval for $Y_T(0)$. Imputing missing potential outcomes in this way follows the approaches discussed in, for example, Rubin (2005). Regardless, the core feature of this approach is that we end up with a range of plausible values for $Y_T(0)$ that incorporate the natural variation in the data as well as uncertainty about the parameters of our model.

The validity of the range of plausible values depends on the model being correctly specified. We believe this approach to uncertainty quantification renders model dependency more transparent (salient) than a classic maximum likelihood analysis or regression approaches. For example, we here see more explicitly the importance of the correct specification of the initial linear trend and the homoskedasticity assumption. We are not making more or different assumptions than the classic approaches with autoregressive specifications, but rather are making the identical assumptions more explicit. We do avoid some of the asymptotic approximations used in maximum likelihood inference.

Case Study: Mecklenberg County and the Proportion of Cases Assigned Bail

Mecklenberg instituted a series of reforms including changing their pre-trial risk assessment tool to a tool called the Public Safety Assessment (the PSA). These reforms were designed to reduce the negative impacts on arrestees while maintaining public safety; the goal is to identify and release those defendants unlikely to fail to appear at future court hearings or break laws while awaiting trial, while imposing monitoring on the remainder. One outcome of interest in evaluating the effectiveness of this program is the rate of bail setting (what proportion of cases resulted in the assignment of bail or outright detention) as compared to outright release. See Redcross et al. (2019) for further discussion.

To investigate this, we fit equation (5) to the Mecklenberg data displayed on Figure 1a. Our estimated coefficients are $\widehat{\beta}_0 = 45$, $\widehat{\beta}_1 = -0.12$, and $\widehat{\beta}_2 = 0.26$. The lagged outcome term $(\widehat{\beta}_2)$ is not significantly different from 0. We see that the pre-policy trend does appear roughly linear. The lack of significance of our autoregression term suggests that there is little autocorrelation after the linear trend is accounted for, but keeping it in our simulation incorporates the additional uncertainty that even a small amount could bring. Dropping the lagged term from our model would be imposing the assumption of independence, which, given substantive knowledge of seasonality effects on criminal and policing behavior, is not tenable. The failure to find a significant correlation could be a power issue.

Using our model we can generate trajectories starting at $t_0 = 0, Y_0 = 60.1$. Ten such extrapolations are on Figure 2a. We generate 10,000 such extrapolations based on 10,000 draws of possible parameters θ, and summarize by, for each time point, taking the middle 95% range of values. We plot these as an envelope on Figure 2b.

Overall we see evidence of a reduction of the use of bail beginning a few months after the policy onset. Pre-policy trends do not indicate a decline as far as what actually occurred. The observed outcomes for the first four months after the policy change are, however, still potentially following the pre-policy trend; the departure is only really significant at month five and six. At this point, actual bail mostly levels off at the reduced rate of around 50%. Nonlinear patterns of impact such as these raise important issues of how to ascribe the change: was this drop at month five due to the policy shift, or due to some subsequent intervention that may or may not have been part of the policy? In this case, there is some qualitative evidence that Mecklenberg continued to reinforce their policy change with additional trainings of court agents, which could have caused this delayed impact.

The nominal impact is the difference of the *projected* trend and the actual, which means the change in the overall level of an outcome does not necessarily mean there is a measured impact. In this case, for example, we see the overall linear pre-policy trend projecting a steadily decline of bail assignment. This means that at around two years post policy we cannot rule out an absence

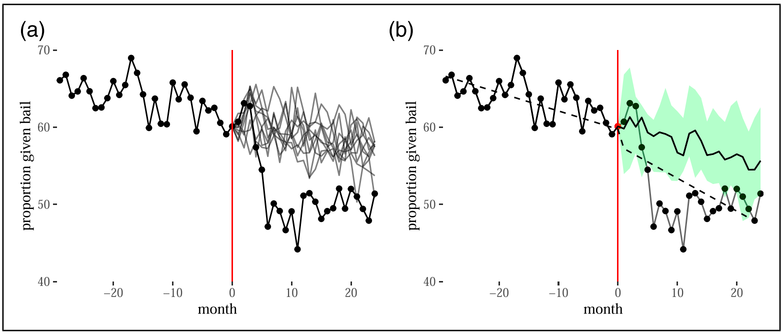

Figure 2. Results of Mecklenburg analysis.

Comment: At left ten sample simulated series along with observed data. Y-axis truncated to 40 to 70%. At right the overall envelope of plausible series given pre-policy data. For many post-policy months the proportion of cases assigned bail is not in the range of likely bail rates, suggesting that there was a more rapid decline of bail-setting after the policy change than expected given the slow decline of the pre-policy trend. Dashed line shows contrast fitting approach of equation (3) fit using generalized least squares with AR(1) residuals; it does not well account for the nonlinear post-policy series.

of impact: those bail levels may have been reached regardless, considering the pre-policy declining trend, but at a later time than with the policy change.

By contrast, a classic regression or generalized least squares approach might be to fit a regression equation on the full data, specifically modeling the post-policy trend rather than focusing on imputation of the counterfactual. In particular, we could fit equation (3) under an assumption of *i.i.d.* errors (classic OLS). This gives an initial significant impact at policy onset of $\delta_0 = -5.02$, and a marginally significant growth in impact of $\delta_1 = -0.20$; these spurious significance findings are driven by the overly small standard errors that come from falsely assuming residual independence. This approach is simply wrong. We can instead fit equation (3) using generalized least squares and specifying an AR(1) residual structure. In this case, we estimate an impact curve of $\Delta_t = \delta_0 + \delta_1(t - t_0 - 1) = -2.09 - 0.19(t - t_0 - 1)$, with neither coefficient being significantly different from zero. This GLS fit model is presented as the dashed line Figure 2b; it enforces a linear treatment impact and does not allow for treatment patterns such as later onset of treatment. By fitting to pre-policy data only, by contrast, we can remain entirely agnostic as to the pattern of effects. In particular, we argue the linear model misses some nuance; in particular, we do not see the period of months 6 through 18 as having a significant reduction in bail setting due to the linearity of the model. One could use generalized least squares just on the pre-policy data and forecast to individual time points, but that would preclude smoothing, which we discuss below, unless we used the model to simulate trajectories.

Regardless of modeling approach, the further out an extrapolation the greater our dependance on the model being correctly specified, both statistically and as a representation of a dynamic and complex system. The statistical model can extrapolate assuming the general model fit to pre-policy, but the assumption that these trends would continue indefinitely becomes substantively less plausible the further away from the transition we go. The greater uncertainty in later months is only due to estimation error, and is dependent on the assumption that the pre-policy process would have continued unabated in the absence of the policy change. In particular, we cannot know if alternate measures would have been taken had the policy not been imposed or if the system would have naturally reached some change point given the dynamics.

Overall, there are three sources of uncertainty to attend to in an ITS analyses, with only the first two quantifiable: (1) parameter estimation error for the model, (2) the natural variation due to month-to-month changes and associated auto-regression, and (3) model specification.

Seasonality Effects

In New Jersey, when a person is arrested the arresting officer can (1) serve a summons, where the officer gives the arrestee a court date for a future appearance and then sends them home, or (2) serve a warrant, which could result in detention until the resolution of the case. One consequence of a policy rooted in risk assessment might be to change policing behavior towards only giving warrants for the more serious offenses. An outcome of interest that assesses this is the total number of warrant arrests made.

Counts can be more difficult to model than proportions. Figure 1b shows a strong periodic trend across the years, with reduced number of arrests when it is winter, and more in summer. In fact, average temperature in a month (a good proxy for season) is found to predict total arrests quite strongly; see scatterplot in Online Supplement, Part C. These seasonal cycles are likely due to factors such as increased time spent indoors during the colder winter months, which could both reduce the true level of crime and the chance of arrest. Both these factors would reduce arrest count.

Fitting a simple autocorrelation model would miss the cyclic nature of our trend, which means we have clear model misspecification and which, in this case, results in substantial loss of power (as we show below). We instead extend our linear model to model the periodic trend. The autoregressive element would then allow local departures from the overall seasonality model, just as we had local departures from the linear model above.

There are several ways one might capture a periodic seasonality structure with linear regression. A simple approach is to include dummy variables for the four seasons. The following model, for example, has the first quarter as a baseline, has three offsets for the other three quarters, and also allows an overall linear trend

$$Y_t = \beta_0 + \beta_1 t + \gamma_2 Q_{2t} + \gamma_3 Q_{3t} + \gamma_4 Q_{4t} + \epsilon_t$$

with Q_{2t},Q_{3t},and Q_{4t} $0/1$ indicators for being in the 2nd, 3rd, and 4th quarters of the year. A second approach is to use a covariate that is predictive of outcome and is itself periodic, such as, in our case, monthly average temperature in the region

$$Y_t = \beta_0 + \beta_1 t + \beta_2 Temp_t + \epsilon_t$$

where $Temp_t$ is a measure of average temperature for month t. The periodic nature of our data is then driven by the periodic nature of our time-varying covariate. These general approaches can easily be combined

$$Y_t = \beta_0 + \beta_1 t + \beta_2 Temp_t + \beta_3 Q_{2t} + \beta_4 Q_{3t} + \beta_5 Q_{4t} + \epsilon_t \qquad (6)$$

One potential concern with seasonal dummy variables is the resulting curve will be a step function rather than a smooth curve, with steps at pre-specified points that are not data driven. We could alternatively fit a sinusoidal trend by building two covariates that correspond to the sine and cosine of the month (rescaled to have a yearly period). Linear combinations of these two covariates allow for sinusoidal curves that can be smoothly shifted left or right. For example

$$Y_t = \beta_0 + \beta_1 t + \rho_1 sin(2\pi t/12) + \rho_2 cos(2\pi t/12) + \epsilon_t$$

Different coefficient values for ρ_1 and ρ_2 control where the peaks and valleys of this trend are.

The four fitting approaches are shown in Online Supplement, Part C. Of the four models, the model with both quarter and temperature has the best pre-policy fit, with an estimated residual standard deviation of 192 compared to 250 and above for the other models.

Seasonality With Autoregressive Residuals

Once a seasonality model is selected, we again are faced with how to fit the autoregressive residual structure in a simple way that also lends itself to simulation. We cannot simply include the lagged outcome, as this lagged outcome includes the lagged periodic structure. We therefore include the lagged values of the covariates used to model seasonality along with the outcome; this subtracts out the lagged structural component of the trend, resulting in a corrected model that puts the autoregression solely on the residuals. See the Online Supplemental Material, Part B for a derivation of this result, along with some alternative estimation strategies.

For example, for Model 6 we would have $X_t = (1, t, Q_{2t}, Q_{3t}, Q_{4t}, Temp_t)$. We would then fit

$$Y_t = X_t'\beta - X_{t-1}'\beta_\ell + \rho Y_{t-1} + \omega_t$$

with β our primary trend and β_ℓ our lagged "anti-trend" (generally $\beta \approx \beta_\ell$, with an exact equality if we fully believe our lagged model). There is a small technical caveat: the lagged covariates can frequently be collinear with the contemporaneous covariates, producing an overall design matrix that is not full rank. For example, if we include a linear time component by including the covariate $X_{t,2} = t$ as one of the columns of our design matrix, the design matrix with our lagged covariate of $X_{t,k} = t - 1$ will clearly be fully collinear with $X_{t,2}$.[3] This can also happen with periodic covariates such as $X_{t,k} = sin(at)$. This colinearity is easily resolved: simply drop collinear columns (in particular the intercept and time variables), allowing the remaining parameters to estimate the combined influence of both the primary observation and the structural component of the lagged outcome.

Case Study: New Jersey and the Number of Warrant Arrests

We next analyze the data on warrant arrests shown on Figure 1b with our seasonality model. We fit Model 6 with the autoregressive residual model of $\epsilon_t = \rho \epsilon_{t-1} + \omega_t$. We set $t_0 = -8$ due to evidence that there was some preparatory restructuring and changes made in advance of the policy's official start date to ensure a smooth launch; by setting $t_0 = -8$ we increase the plausibility of our no anticipation assumption. We, following the above, extend our model to include the lagged outcome and lagged covariates, giving

$$\begin{aligned} Y_t = \beta_0 + \beta_1 t + \beta_2 Temp_t + \beta_3 Q_{2,t} + \beta_4 Q_{3,t} + \beta_5 Q_{4,t} \\ + \beta_6 Temp_{t-1} + \beta_7 Q_{2,t-1} + \beta_8 Q_{3,t-1} + \beta_9 Q_{4,t-1} + \rho Y_{t-1} + \omega_t \end{aligned}$$

We then generate the predictive envelope on Figure 3a by following the process described above.

By comparison, if we had not included a seasonality model and instead simply fit our simple linear trend model, we get Figure 3b. The model without seasonality has more autocorrelation (estimated as 0.77 vs 0.69) because points near each other are correlated due to the periodic trend around the base linear model. The seasonality model captures and removes these dependencies. This autocorrelation allows for large deviations from trend in the simulated extrapolated series, and thus we see a large confidence envelope. In general, without the seasonality model we are not able to take advantage of the seasonal structure of the data, but the autoregressive element does capture that there is local dependence, resulting in a conservative inference.

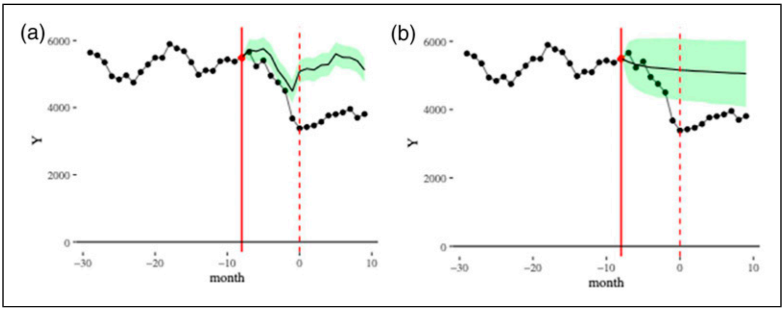

Figure 3. Prediction envelopes for number of warrant arrests in New Jersey.
Comment: Time period (*x*-axis) truncated to show more detail of model fitting in post-policy era. (a) shows seasonality model, (b) shows model with no seasonality. See raw data on Figure 1b.

One might ask whether Mecklenberg should also be fit with a seasonality component. Generally, to reliably estimate a seasonality structure, we would need several cycles of the seasons; Mecklenberg is "too short" to ascertain that structure. In this case, we rely on the estimated autocorrelation to capture the overall uncertainty. Determining when to fit the more complex model versus not is an important area for future work, but we found that with eight years of data, and a clear seasonal trend, the seasonality model was easily estimable. For Mecklenberg, however, seasonality models were quite unstable.

Inference and Smoothing

Reading the envelope graphs from the above analyses can be somewhat confusing as there are multiple post-policy months with some of them having observed outcomes lying outside of the predictive envelope and others not. In this section, we discuss inference more formally and also discuss how to increase power by averaging the outcomes of post-policy months together. For this averaging we can either average a fixed range of months, or use methods akin to a sliding window by nonparametrically smoothing the observed trends to account for month-to-month variation. This sliding window approach is appealing in that we can display an entire curve of impacts post-policy, which allows for a more nuanced interpretation of how a policy may have evolved over time.

Inference

Consider the null hypothesis of there being no change in the pre-policy trend (and that we have correct model specification). In this case, our simulated series are all plausible forecasting series, given the pre-policy data. For any given point $T > t_0$, we can therefore examine the distribution of simulated values at T to see how much variability we would see under the null hypothesis. These are Monte Carlo tests (Kroese et al., 2011), similar in spirit to permutation tests (e.g., Pesarin & Salmaso, 2010).

Monte Carlo tests compare a test statistic (a function of the observed data) to a reference distribution of what that statistic would tend to look like if the null hypothesis were true. In the simplest case for testing for a deviation from expected at time T, we use our observed outcome Y_T^{obs} as the observed value of a test statistic: we compare this observed value to the set of simulated values (this is our reference distribution) that capture what our model says is possible. If the observed value is outside the central range of these simulated values, we reject the null that the pre-policy trend continued unabated (again assuming the pre-policy model is correct). We could do this for each $T > t_0$.

While reasonable and sound, there are two concerns: first, we have a multiple testing issue. If the series is long enough, we are bound to find some points outside their respective predictive ranges simply due to random fluctuation. Second, we have a power issue. We are comparing our test statistic, a potentially highly variable single point Y_T^{obs}, to a distribution of simulated values Y_T^* that all themselves could be quite variable. If the policy caused a modest reduction in Y_t^{obs} for all $t > t_0$, it is possible that no individual $T > t_0$ would look significantly reduced when examined in isolation.

As a contrast to testing a specific point in time, we might instead test for a systematic and sustained shift in the outcomes over a range of times post-policy. In order to test a larger sequence of time points, we need to combine our observed data into some sort of average and compare that *average* to the distribution of averages we would have likely seen under the null.

The simplest approach to do this is to simply average all the outcomes in a pre-specified range of months post-policy. We then compare this simple average to the distribution of simple averages calculated from the distribution of plausible trajectories. The key point is once we have our distribution of plausible trajectories, we can test our null hypothesis by comparing a summary statistic of our outcome to the distribution of that same summary statistic calculated on our trajectories. To be specific, take our observed series $Y = (Y_1, \ldots, Y_T)$ and calculate our summary $t^{obs} = t(Y)$, where $t(\cdot)$ a function that takes our data and summarizes it in some way (e.g., by calculating the average of $Y_{t_0+1}, \ldots, Y_T$). Next, for each simulated series $Y^{*(r)}$, calculate $t^{*(r)} = t(Y^{*(r)})$, and then calculate the $\alpha/2$ and $1 - \alpha/2$ quantiles $t_{(\alpha/2)}$ and $t_{(1-\alpha/2)}$ of these $t^{*(r)}$. Our prediction interval of what value of the summary statistic we would expect to see is then $CI = (t_{(\alpha/2)}, t_{(1-\alpha/2)})$. If $t^{obs} \notin CI$, we reject our null hypothesis. We calculate nominal p-values using the percentile q of our observed t^{obs}, with $p = min(q, 1 - q)$ (for a two-sided test).

Testing in this way is akin to posterior predictive checks of model fit (Rubin, 1984; Guttman, 1967): we want to know if the model fit to pre-policy data fits our post-policy observed data. If it does not, we reject the model, that is, we conclude that something changed our trajectory. Importantly, this testing requires no assumptions on the form of the treatment impact; the reference distribution is entirely driven by pre-policy data and the null hypothesis.

Our p-values are called *posterior predictive p-values*, and do not necessarily have strictly valid frequentist properties, but they are argued to generally be conservative (Meng, 1994). Also see Robins, Vaart, and Ventura (2000). A rejection via this approach is evidence that the assumed model is incompatible with the observed data: this could be due to a shift in trend, or be due to model misspecification with regard to the pre-policy data itself. The simulation approach makes model dependencies for the $Y_t(0)$ explicit (while making no assumptions on the $Y_t(1)$, which we view as an advantage over many other approaches). These assumptions need to be specifically acknowledged and grappled with in any ITS analysis, regardless of estimation approach.

Smoothing

In investigating a place-based initiative we generally want to understand the evolution of the impact over time. For example, with Mecklenberg, it appears as if the policy induced a large reduction in the rate of bail setting a few months into the post-policy period, with that level of bail setting generally sustained over time. If we only use the simple averaging method from above, and did not look at the overall graph of impacts, we would lose this nuance. But the raw graph is noisy, making trends somewhat difficult to discern. We therefore might want to smooth the trend in the graph to, as much as possible, remove month-to-month variation. Smoothing is when one locally summarizes a trend to remove some variation, ideally without imposing a global structural so local structure is preserved (Cleveland, 1993). Smoothing is generally nonparametric, and can be done with splines, averaging within a sliding window, or using loess (Locally Weighted Smoothing) (Cleveland (1981), but see Cleveland (1993)). Smoothing can make communication with various stakeholders easier, as it removes random variation that may draw one's attention if not removed; see, for example, discussions in Starling et al. (2019).

We can easily use smoothing coupled with our inferential approach above. In particular, we smooth each simulated time series using a specific (pre-specified) method. We then compare the distribution of these smoothed time series to the actual time series smoothed in exactly the same way. Under our null hypothesis, the smoothed observed trajectory should be exchangeable with any of the smoothed simulated trajectories. Our smoothed estimate at a given timepoint T is now our test statistic, and the distribution of smoothed estimates of our simulated series our prediction distribution of what values we might have expected. This should have greater power: we are now examining the overall trend in the neighborhood of T, potentially increasing precision as idiosyncratic monthly variation gets averaged out.

One caveat is that if we smooth across t_0 we can cause the smoothed line of our observed series to artificially deviate *pre-policy* since the post-policy points will be included in the local average near the policy change. Similarly, the pre-policy timepoints near t_0 can drag the smoothed post-policy timepoints near t_0 towards their values, potentially masking impacts. To avoid this, one can smooth the post-policy series only, not including any pre-policy points; if this is done, then it needs to be done for both the simulated series as well as the observed series. The key is to implement the same process on all series, simulated and observed, to maintain the validity of the comparison.

Mecklenberg County, continued. We continue our Mecklenberg example by showing how to improve power using both averaging and loess smoothing. We initially average the outcomes for the initial 18 months after t_0. In our data, we observe an average bail rate of 52%. The middle 95% prediction interval of the averages of our simulated series ranges from 55% to 64%. We therefore conclude that something changed the pre-policy trajectory so we are seeing lower average rates of bail-setting than we would have expected. If we take the difference to get estimated impacts we obtain a 95% confidence interval (technically a credible interval) for the true average impact of $(-3\%, -12\%)$. To get a point estimate for the average impact, we average the simulated averages, predicting an average bail setting of 59% and an estimated reduction of 7 percentage points.

If we look at a tighter range of months (which we would ideally have prespecified) of 6–18 months, we observe an average of 49%, a corresponding prediction interval of 54%–64%, and a slightly larger estimated impact of between 5 pp and 15 pp. Choice of summary measure can matter as they will differently weight what are often quite heterogeneous impacts across time.

We also use loess smoothing to smooth the post-treatment trajectory. We first smooth our observed series with a loess smoother fit to the post-policy data only to avoid any influence of pre-policy points on our resulting line. We then fit the same smoother to each of our simulated series, ignoring the prepolicy points there as well. Results are on Figure 4a and b. Figure 4a shows 10 smoothed trajectories in the post-policy period. Figure 4b shows the envelope based on these trajectories, along with the smoothed observed line and, in the background, the original envelope without smoothing. The smoothed observed curve is arguably easier to read than the raw data. We also see precision gains from the smoothing process, which stabilizes the estimation. Also note the wider envelope at far left; this is due to loess smoothers being more variable at endpoints.

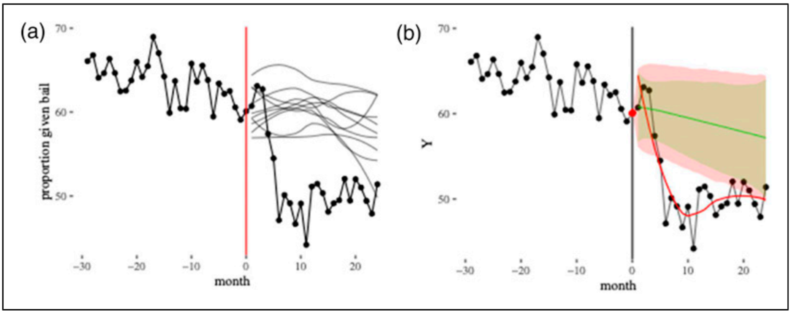

Figure 4. Results of Mecklenberg analysis (with smoothing).
Comment: 4a shows how smoothed trajectories have less variability than the raw series did. 4b compares smoothed envelope with envelope without smoothing. We see less variability. The red line denotes the smoothed observed trend to be compared to the envelope and counterfactual predicted trend.

Smoothing does require specifying a tuning parameter of how much to smooth. For loess, for example, we essentially specify what fraction of the data should be used to calculate the smoothed outcome at each time point. If we smooth a lot, then local variation in the structure will be removed, but the lines will be more stable. If we smooth little, then we do not really average local points, and thus our variance will remain high. This is a bias-variance tradeoff in the estimation and visualization.

Smoothing with Seasonality

When the model has a seasonality component causing oscillation, a simple loess smoother might dampen the oscillations, creating a smoothed series that is more flat than the data. This not only looks odd, but can be deceptive. But, as discussed above, we can smooth in any fashion we choose, as long as we smooth our observed data in the same way as the simulated. This allows for the following multi-step smoothing approach that smooths the residual variation around the structural component of a seasonality model. For each time series (observed or simulated), smooth as follows:

1. Fit a working seasonality model to the data. This is not the original seasonality model, but a new model. There is no need for lagged outcomes or uncertainty estimation in this model. As with loess smothing, we can choose to fit to post-policy data only, pre-policy data only, or all data.
2. Predict all the outcomes given our fit seasonality model.
3. Calculate the residuals by subtracting the predicted outcomes from the actual outcomes.
4. Fit a loess smoother (or some other smoother) to the residuals (again choosing whether to focus on post-policy only or on all data).
5. Add the smoothed residuals back to the predictions to get a final smoothed curve.

This process strips the estimated approximation of the structural component from the series and sets it aside to prevent it from being smoothed or averaged out. Step (5) puts it back so our final series maintains the overall structure. In particular, any estimated seasonality component will not get smoothed out. The key idea is that our smoothing model does not need to be correctly specified; it is purely to set aside any seasonal structure so it does not get over-smoothed.

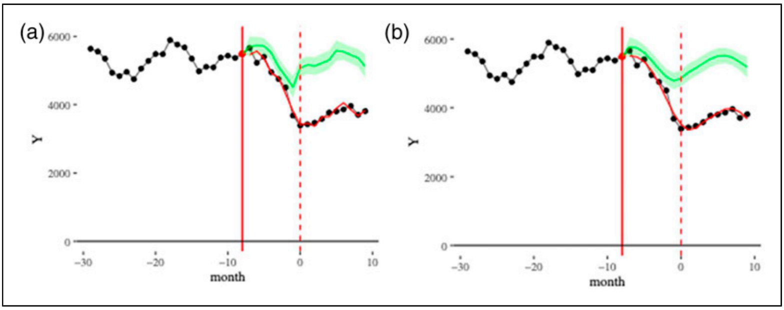

Figure 5. Prediction envelopes for number of warrant arrests using smoothing.
Comment: Time period and *y*-axes truncated to show more detail.

New Jersey, Continued

To demonstrate smoothing with a seasonality model we extend our analysis of warrant arrests. We compare two methods of smoothing. In the first, we extrapolate with the base model with quarter and temperature, but no lag, and in the second we extrapolate with a newly fitted sinusoidal model without temperature. The same quarter, temperature, and lag base model is used for extrapolation in both cases. Our second smoothing model intentionally smooths away month-to-month variability due to fluctuating temperature in both our simulated and observed series, even though we use the temperature to fit and extrapolate our data to obtain our predictive series before smoothing. The results are on Figure 5a and b. The left has preserved the month-to-month variation predicted by the temperature changes, giving a more jagged sequence. The right, by contrast, is smoother, showing underlying structure more clearly.

If we do not fit the same base model to the same range of data for both the observed and simulated series, smoothing in the observed series could cause different distortions than in the simulated series. This could create systematic differences between observed and simulated even if the null of no interruption were true. Further, if there were a large initial treatment impact, a model fit to the full observed series could be misspecified. This in turn could give an odd smoothed series for the observed data. Regardless, as long as the model fitting process is held to be the same, then comparing the observed series to the reference distribution of simulated series is valid for testing. We recommend selecting a smoother that is not overly dependent on the pre-policy patterns, but instead naturally fits to the observed post-policy data. In particular, we suggest fitting the seasonality model to the post-policy data only.

Conclusion

We have demonstrated a simple modeling (linear regression with lagged outcomes and covariates) and simulation framework for capturing uncertainty for Interrupted Time Series designs. These designs often appear when attempting to assess the impact of a policy change on a single region of interest when there are no good comparison regions available.

Our modeling framework allows for the incorporation of seasonality models and of smoothing in a straightforward manner. It also naturally allows for incorporation of autoregressive structure to better account for overall uncertainty. Simulation makes the dependency on assumptions much more explicit, and also renders more clear the consequent fragility of the overall inference. Finally, we argue that this approach naturally lends itself to generating clear visualization of impacts and transparent reporting of results. One could instead use direct model fitting, although this would prohibit the smoothing approach and itself require post-processing of the model output to estimate a series of impacts for each post-policy month. Alternative modeling approaches, such as fitting ARIMA models to the pre-treatment data and using forecasting methods to identify the counterfactual, might be used in richer data contexts that allow for modeling more complex autocorrelation structures. We promote simulation as being direct, interpretable, and a natural continuation of the CITS, ITS, and regression approaches that are most familiar to policy analysts.

In this work, we have examined ITS designs with a moderate number of pre-policy timepoints; with fewer timepoints estimating the autoregressive component of the model will generally be much more difficult. In maximum likelihood approaches, it is known that autoregressive models can be biased and have poor coverage when there are only five or so observations (see, e.g., St Clair, Hallberg, and Cook (2016)). We leave whether simulation, simulation which specifically incorporates the uncertainty in the estimated lagged coefficients, would help in these short ITS designs to future work.

This approach could also be extended to power calculations. Minimal detectable effect size (MDES) and power depend on several factors: the number of cases per month, the month-to-month variability beyond natural variation due to the cases, the number of months of pre-policy data, and the desired window of predicting impacts after the policy implementation. Each of these can heavily influence the ability to detect effects. One way forward is to again turn to simulation. In particular, given specific parameterized values for the factors listed above, one could repeatedly simulate a dataset, and then analyze that dataset using the above simulation approach as an inner step. For each initially simulated dataset we would then record the width of the simulated extrapolations. The average width of these prediction intervals at each time point could then be tied to MDES.

Finally, the modeling itself could also potentially be extended and enriched to better capture some data contexts. For example, if the number of individual cases changed substantially over the course of a series, we might want to let our residual error be a function of sample size to capture differing levels of precision (see, e.g., Ferman and Pinto (2019). One approach would be to regress residual size onto number of cases, giving an intercept and slope which would represent core month-to-month variability and within-month variability, and use this decomposed variation in the autoregressive model.

With ITS, there are some concerns with interpretation, in particular in the case of a dynamic system. For example, if the impacts in early post-policy months are creating a feedback loop (e.g., changing patterns in detention causing changes in the patterns of new charges) then the mix of individual cases constituting the overall region may be changing as a result of the policy change. This further underscores that interpreting impacts has to occur at the region level, which naturally takes these changes into account. In particular, a reduction of bail rates could potentially be due to the policy changing the cases themselves, rather than be due to changes in how cases are being handled. Ideally we thus should focus on measures that are of interest when viewed at the aggregate level.

And finally, fundamentally, we note that all that this type of analysis can show us, using this method or any other, is that the trend has changed in a surprising way. Why it did so, the statistics cannot answer. The researcher in the end must turn toward substance matter knowledge and argument to defend the proposition that a found change was caused by the policy shift.

Acknowledgments

I gratefully acknowledge the productive and stimulating collaboration on this project with MDRC, in particular Chloe Anderson, Brit Henderson, Cindy Redcross, and Erin Valentine. This work is in part generously funded by Arnold Ventures, and the Institute for Education Sciences. The opinions expressed are those of the authors and do not represent views of the funders. Further thanks to the useful feedback and commentary from Howard Bloom, Rich Hendra, and the Miratrix C.A.R.E.S. Lab.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by MDRC through a grant from Arnold Ventures The work was also supported by the U.S.

Department of Education, Institute for Education Sciences, through Grant R305D200010.

ORCID iD

Luke Miratrix https://orcid.org/0000-0002-0078-1906

Supplemental Material

Supplemental Material for this article is available online.

Notes

1. The subscript here does not denote the unit, as is typically seen, but rather the time of observation for our single unit.
2. One could instead use maximum likelihood and asymptotic approximations given the defined residual structure; we argue the parametric simulation approach we use provides a flexible and easily extendible alternative.
3. This colinearity is why the simple lagged linear trend model does not have an extra term beyond the lagged outcome itself.

References

Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Baicker, K., & Svoronos, T. (2019). *Testing the validity of the single interrupted time series design*. Technical report. National Bureau of Economic Research.

Bloom, H. S., Riccio, J. A., Verma, N., & Walter, J. (2005). *Promoting work in public housing. The effectiveness of jobs-plus: Final report*. Technical report. MDRC.

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., & Scott, S. L. (2015). Inferring causal impact using Bayesian structural time-series models. *Annals of Applied Statistics*, *9*(1), 247–274. https://doi.org/10.1214/14-aoas788

Clark, A. E., Diener, E., Georgellis, Y., & Lucas, R. E. (2008). Lags and leads in life satisfaction: A test of the baseline hypothesis. *The Economic Journal*, *118*(529), F222–F243. https://doi.org/10.1111/j.1468-0297.2008.02150.x

Cleveland, W. S. (1981). Lowess: A program for smoothing scatterplots by robust locally weighted regression. *American Statistician*, *35*(1), 54. https://doi.org/10.2307/2683591

Cleveland, W. S. (1993). *Visualizing data*. Hobart Press.

Cook, T. D., Campbell, D. T., & Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Davison, A. C. (1997). *Bootstrap methods and their application* (Volume 1). Cambridge University Press.

Ferman, B., & Pinto, C. (2019). Inference in differences-in-differences with few treated groups and heteroskedasticity. *Review of Economics and Statistics*, *101*(3), 452–467. https://doi.org/10.1162/rest_a_00759

Ferron, J., & Rendina-Gobioff, G. (2005). *"Interrupted Time Series Design"*. Encyclopedia of Statistics in Behavioral Science, American Cancer Society.

Fox, J., & Weisberg, S. (2018). *An R companion to applied regression*. Sage publications.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multi-level/hierarchical models*. Cambridge University Press.

Gelman, A., Meng, X. -L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*(4), 733-760.

Golub, C. A., Redcross, C., Valentine, E., & Miratrix, L. (2019). *Evaluation of pretrial justice system reforms that use the public safety assessment: Effects of New Jersey's criminal justice reform*. Technical report. MDRC.

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society: Series B (Methodological)*, *29*(1), 83–100. https://doi.org/10.1111/j.2517-6161.1967.tb00676.x

Hallberg, K., Williams, R., Swanlund, A., & Eno, J. (2018). Short comparative interrupted time series using aggregate school-level data in education research. *Educational Researcher*, *47*(5), 295–306. https://doi.org/10.3102/0013189x18769302

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press.

Jacob, R., Somers, M. -A., Zhu, P., & Bloom, H. (2016). The validity of the comparative interrupted time series design for evaluating the effect of school-level interventions. *Evaluation Review*, *40*(3), 167–198. https://doi.org/10.1177/0193841x16663414

King, G., Tomz, M., & Wittenberg, J. (2000). Making the most of statistical analyses: Improving interpretation and presentation. *American Journal of Political Science*, *44*(2), 347–361. https://doi.org/10.2307/2669316

Kroese, D. P., Taimre, T., & Botev, Z. I. (2011). *Handbook of Monte Carlo methods*. Wiley Series in Probability and Statistics, John Wiley and Sons.

Meng, X. -L. (1994). Posterior predictive p-values. *The Annals of Statistics*, *22*(3), 1142–1160.

Mohtadi, H., & Weber, B. S. (2021). Catastrophe and rational policy: Case of national security. *Economic Inquiry*, *59*(1), 140–161. https://doi.org/10.1111/ecin.12925

Pesarin, F., & Salmaso, L. (2010). *Permutation tests for complex data* (pp. 1–450). Wiley.

Redcross, C., Henderson, B., Valentine, E., & Miratrix, L. (2019). *Evaluation of pretrial justice system reforms that use the public safety assessment: Effects in Mecklenburg County, North Carolina*. Technical report. MDRC.

Robins, J. M., van der Vaart, A., & Ventura, V. (2000). Asymptotic distribution of p values in composite null models. *Journal of the American Statistical Association*, *95*(452), 1143–1156. https://doi.org/10.1080/01621459.2000.10474310

Rosenbaum, P. R. (2009). *Design of observational studies*. Springer Science & Business Media.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, *12*(4), 1151–1172. https://doi.org/10.1214/aos/1176346785

Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association*, *100*(469), 322–331. https://doi.org/10.1198/016214504000001880

Somers, M.-A., Zhu, P., Jacob, R., Jacob, P. E., & Bloom, H. (2013). *The validity and precision of the comparative interrupted time series design and the difference-in-difference design in educational evaluation*. Technical report. MDRC.

Starling, J. E., Murray, J. S., Lohr, P. A., Aiken, A. R. A., Carvalho, C. M., & Scott, J. G. (2019). Targeted smooth Bayesian causal forests: An analysis of heterogeneous treatment effects for simultaneous versus interval medical abortion regimens over gestation. *arXiv:1905.09405*.

St Clair, T., Hallberg, K., & Cook, T. D. (2016). The validity and precision of the comparative interrupted time-series design: Three within-study comparisons. *Journal of Educational and Behavioral Statistics*, *41*(3), 269–299. https://doi.org/10.3102/1076998616636854

Stoffer, D. S., & Shumway, R. H. (2006). *Time series analysis and its applications: With R examples* (2nd ed.). Springer.

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, *16*(4), 437–450. https://doi.org/10.1016/s0169-2070(00)00065-0

Zeng, Z. (2018). *Jail inmates in 2016*. Technical report. Bureau of Justice Statistics.

Zhang, F., Wagner, A. K., Soumerai, S. B., & Ross-Degnan, D. (2009). Methods for estimating confidence intervals in interrupted time series analyses of health interventions. *Journal of Clinical Epidemiology*, *62*(2), 143–148. https://doi.org/10.1016/j.jclinepi.2008.08.007