Simulating for uncertainty with interrupted time series designs

Luke Miratrix (Harvard University)
with
Chloe Anderson, Brittany Henderson, Cindy Redcross,
and Erin Valentine (MDRC)

May 9, 2019

Abstract

Despite our best efforts, sometimes we are 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 no 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 or plausible 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. One way forward is to use 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.

Keywords: Neyman-Rubin causal model, ITS designs, criminal justice reform, posterior predictive check

1 Introduction

Interrupted Time Series (ITS) designs are classic and well-studied. Interrupted time series tend to occur when there is a governance body of some area, e.g. a county, that implements some sort of policy change at a particular point in time. The researcher is able to observe regular measures of some outcome of interest both for several time points before such a change as well as after. The research question is then whether there is any evidence that the policy has changed the course of the unit of interest. In particular, at the point of policy change, we might naturally expect various outcomes of interest to change, deviating from what they historically were. For example, if a school experiences a massive reorganization we would expect that measures of student wellbeing, such as rate of college-going or rate of graduation, could potentially change over time. In sharp contrast to regression discontinuity designs, however, we might not expect any treatment impact at the time of the policy change. It may take some amount of time for the policy to become fully implemented, and for the consequences of the policy to be felt. What makes these contexts particularly challenging is frequently there is only a single unit that received the policy change and/or no reasonable comparison units that did not receive such a change.

One area where we see these kind of reform efforts are in modern criminal justice reform, in particular pretrial reform. 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 (Zeng, 2018) to secure their release. Recently, several jurisdictions have sought to improve these judicial systems by attempting to build improved procedures to increase the rate of release for “low-risk” defendants; ideally, this would reduce negative impacts on these defendants as well as reduce load on jails. 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.

This is the context we use in this work. We primarily use data from a reform effort in Mecklenberg County, along with a simulated dataset loosely inspired by a reform effort in New Jersey.[1] The Mecklenberg evaluation, funded by the Laura and John Arnold Foundation (LJAF) and conducted by MDRC, is further described in Redcross et al. (2019). With such reform efforts, there are several primary outcomes of interest, of which we here examine two: the proportion of arrestees assigned monetary bail, and the total number of warrant arrests made (a warrant arrest here being an arrest with the possibility of detention). The methods described here were also applied to the other outcomes of interest (e.g., rates of new arrest, rates of failure to appear on giving court dates, time spent in jail, etc.) as described in Redcross et al. (2019).

[1] The simulated data were generated to illustrate the annual cyclic trend (a seasonality trend) in the total number of arrests found in an ongoing evaluation project of the reform effort in New Jersey. As this evaluation project is ongoing, we do not use any actual data or present any findings from this evaluation in this document.

The most classic 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, 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 auto-regressive model with linear trend to the pre-policy data, and then using that model to simulate 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 both the uncertainty of the linear model estimate as well as any autoregressive dependencies in the residuals by using a pseudo-Bayesian approach discussed in Gelman and Hill (2006) where we sample from a "posterior" of the parameters and then use that model to simulate the synthetic trajectory.

This approach is a straightforward extension of the standard ITS modeling approach, and can in large part be implemented using simple regression routines. We provide a simple R package, `simITS`, that implements the methods discussed and also provides code for generating plots and other descriptives needed for a full and transparent ITS analysis. This general approach can be extended in a variety of ways. First, we can easily extend the simulation approach to incorporate covariates to capture nonlinearities (in particular, cyclic trends in the outcome due to seasonality). We discuss how when doing so the seasonality variables also need to be included in the regression model in lagged form. Second, we discuss how to average, or smooth, the 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 rather than testing single points in isolation. Finally, one may want to control for time varying covariates such as the proportions of different types of criminal cases over time (in particular, whether individual arrests are warrant or summons arrests) to further control for potential confounders of apparent treatment impacts. We discuss how to achieve this by post-stratifying the individual observations within each time point, and re-weighting the overall sequence to account for changing mixtures of individual case type.

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 to assess uncertainty. Our method is also a parametric simulation approach, but also explicitly includes autoregressive dependencies and explicitly simulates the post-policy trajectories. We also discuss the estimands of interest more explicitly. Similarly, Brodersen et al. (2015) propose a fully Bayesian time-series approach, implemented with the `causalImpact` package, that relies on modeling a latent state space, that has a similar spirit to this work.

More broadly, ITS is a generally worse (in terms of strength of evidence typically provided) version of *Comparative* Interrupted Time Series (CITS) analyses, where the target treatment series has comparison units that are not treated. The idea here is that deviations from trend that are systematic, impacting all units, can be estimated by observing the control series and then removed. In some contexts, this can help provide further evidence that any deviations from trend are due to the impact of the policy change and not due to other systematic factors. These concerns are distinct from the core goal of the methodology discussed in this paper. In particular, our methods allow the researcher to assess whether there has been significant departures from the trend defined by the pre-policy series. Whether these departures are due to the policy change or other factors is not directly answered and must, without further evidence such as comparison series, be assumed.

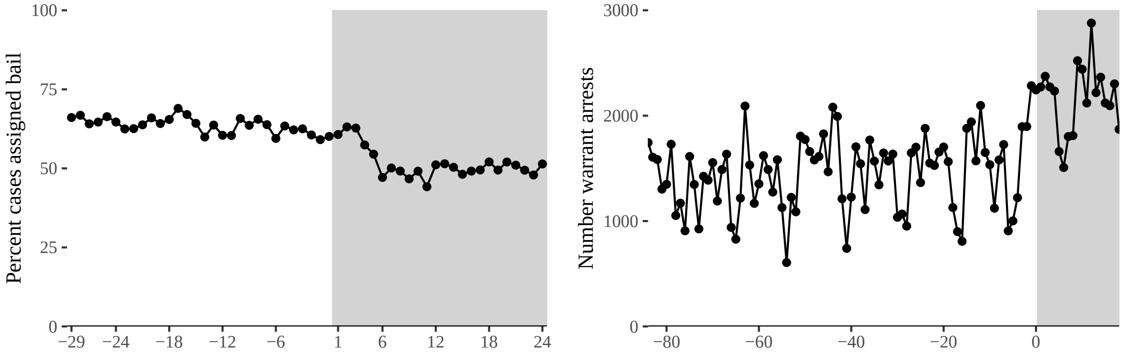
For an overview of CITS, consider Somers et al. (2013) and Jacob et al. (2016), who evaluate the CITS by comparing its finding to those of the more widely-accepted Regression Discontinuity Design, or Hallberg et al. (2018). For a detailed case study with CITS in the context of experimental trials, see Bloom et al. (2005); this approach has particular ties to ITS as they fit regressions to the sequence of paired differences. Interestingly, however, many treatments of CITS do not account for autoregressive structure, instead assuming the errors around the modeled linear trends are i.i.d., or are at least independent (this assumption is implied by the linear modeling with fixed effects approaches typically used). Alternatively, if there are multiple comparison units, a CITS analysis can solve the autoregressive issue by simply clustering the standard errors at the unit level, allowing for arbitrary structure.

ITS is also, of course, based on the idea of a time series. Classic time series methodology, e.g., ARIMA models, could account for linear trend by differencing the observations and then modeling the resulting differences as, ideally, a stationary time series. As this approach gets further away from classic linear modeling approaches more familiar with policy evaluators, we instead follow the ITS and CITS literature and focus primarily on improving the linear regression approach, borrowing from the idea of autoregression to improve model plausibility and consequent inference. For this alternative direction, however, see, e.g., Stoffer and Shumway (2006)

In this paper we first lay out the ITS problem and its classic treatment. We then describe the simulation procedure that allows for a simple autoregressive structure, illustrating with a real example taken from the Mechlenberg County evaluation. We then provide the extensions mentioned above (seasonality, smoothing, and covariate adjustment) in three sections. We offer some general cautions and concluding remarks at the end. We do not focus on the causal inference reasoning behind ITS, however. 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 et al. (2002).

2 Notation and Setup

We have a single treated unit. We observe this unit at several time points before treatment (e.g. a policy change) as well as for several time points after. For an example, consider



(a) Proportion of all arrests that assigned bail (or detention) in Mecklenberg.

(b) Number of warrant arrests made in hypothetical state

Figure 1: Two sample interrupted time series. The dark grey indicates the post-policy era. $t_0 = 0$ in these figures, with pre-policy time being non-positive. Right hand side shows evident seasonality. Left hand side suggests some auto-correlation which may be due to seasonality or other unknown factors.

Figure 1, showing two time series. The left shows the proportion of all arrests in Mecklenberg for each month for a period before and after a reform effort of switching to an improved risk assessment tool coupled with additional measures to change practice (Redcross et al., 2019). The right shows a hypothetical time series of the total number of warrant arrests each month before and after a major reform effort.

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 that the observed unit is stable; everything rests on the assumption that, absent any impact, our unit would evolve as it has been.

More precisely, we observe a sequence of months indexed by t with an outcome Y_t (e.g., in our running example, the proportion of arrests that get assigned bail) for each month. We have pre-policy data up to some point t_0, where the jurisdiction changed their policy in a manner which would plausibly have impact on the outcome moving forward. A core assumption is that all the months up to and including t_0 experienced no impact from the policy. We want to fit a model to these data to predict what would happen after t_0 had the policy change not taken place. Without loss of generality, we assume t are consecutive integers indicating regularly spaced points in time (e.g., in our circumstance, months) from t_{min} to t_{max}.

We can borrow from the potential outcomes viewpoint 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.[2] Let the corresponding $Y_t(1)$ be the outcomes we would observe if we did treat the unit at t_0. In the most general case, we could allow $Y_t(1) \neq Y_t(0)$ for $t \leq t_0$ if we allowed anticipatory effects of treatment, i.e., 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*, i.e., 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 earlier, e.g., to when a policy was initially being planned rather than its official adoption date.

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 in effect converts our estimation problem to a missing data problem; see, e.g., Rubin (2005). Under the view, 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. Linear models are the most common, but other models are also possible. 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.

Remark. We here analyze at the group level by aggregating individual data within each month. We might imagine instead analyzing at the individual level, but this will bring in further complexity from, e.g., individuals being in multiple months (e.g., from multiple arrests in our context), and unknown correlation structure of individuals within a given month (Ferman and Pinto, 2015); aggregation avoids this. Furthermore, migration of individual 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. This does mean the results will be 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 the fewer data points is coupled with less residual noise for those points. See Angrist and Pischke (2008), Chapter 3, for more. For some further discussion on the ideas behind aggregation in time series contexts, also see Bloom et al. (2005), Appendix D in particular. For some dangers with aggregation if the number of units being aggregated changes significantly, see Ferman and Pinto (2015).

[2] Note the subscript here does not denote the unit, as is typically seen, but rather the time of observation for our single unit.

2.1 Classic ITS analysis

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

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

with $\epsilon_t \overset{\text{iid}}{\sim} N(0, \sigma^2)$ 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 $\hat{\beta}_0 + \hat{\beta}_1 t$. Under a homoskedasticity assumption, we can obtain standard errors and conduct inference for the estimated Δ_k, because we assume the variation from expected in the post-policy era is the same as for pre-policy. These standard errors will be driven by, and be no smaller than, $\hat{\sigma}$, the estimated residual standard deviation (see Appendix 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:

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

with

$$\epsilon_t \overset{\text{iid}}{\sim} N(0, \sigma_\epsilon^2)$$

Using this model, we then, for any point $t > t_0$ in the post-policy era, predict via extrapolation,

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

which results in an impact estimate at month t of

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

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 is a prediction model for predicting, via extrapolation, $Y_t(0)$ for all t of interest. 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 at all in the fitting process, some will instead put a structure on the post-policy impact as well, such as with

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

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 $t > t_0$. We advise against such models as the post-policy time points can now inform the estimated pre-policy trend. In particular, misspecification of the functional form of the treatment impact can distort the other parameter estimates.

All these models produce valid inference under the modeling assumptions, in particular the strong assumption of the 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 in the pre-policy period than extrapolation should be done only with extreme caution and skepticism. The *causal* interpretation of this approach, 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 these approaches, however, is that there may be seasonality effects or other shocks that operate in a window of time causing adjacent months to have similar outcomes beyond the underlying model. For example, the pattern of month-to-month averages in Mecklenberg (see Figure 1a) could contain local correlations of months around what is a generally linear trend (we discuss the case of clear nonlinear, cyclic trends such as shown in Figure 1b in Section 4 below).

Classic inference using these models is highly dependent not only on the linear trend assumption but also on the i.i.d. nature of the error term. 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 the model to allow for local dependence of these residuals. By allowing neighboring residuals to be correlated, we will better capture how the time series can "wander" from the linear trend, similar to a random walk.[3]

A simple approach for this 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. E.g., we can specify the residual of Model 2 to be

$$\epsilon_t = \rho\epsilon_{t-1} + \omega_t \text{ with } \omega_t \overset{\text{iid}}{\sim} N(0, \sigma^2). \tag{3}$$

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

[3] Depending on the frequency of the observations, this correlation may be increasingly more critical to model. For example, if we had daily measures, the autocorrelation could be quite strong. Yearly measures may be less of a concern, depending on context.

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

$$Y_t = \tilde{\beta}_0 + \tilde{\beta}_1 t + \tilde{\beta}_2 Y_{t-1} + \tilde{\epsilon}_t \text{ with } \tilde{\epsilon}_t \overset{\text{iid}}{\sim} N(0, \tilde{\sigma}^2) \tag{4}$$

to the pre-policy time points $t = 2, \ldots, t_0$, where t_0 is the last pre-policy month. The initial month has to be dropped in this case due to the need for each observation to have a lagged month. Up to how the parameters are interpreted, this model is equivalent to the lagged residual model. In particular, as the derivations in the appendix 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 with the residual $\tilde{\sigma}^2$ the same as the variance of the ω_t from Model 3. Therefore, we fit our lagged outcome model to get estimates for $(\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2, \tilde{\sigma}^2)$ and then, if desired, convert them to our lagged residual model parameters. See Appendix for some additional discussion.

Once this model is fit, we use it to extrapolate a reasonable counterfactual (assuming no policy change) prediction $Y_T(0)$ for any timepoint $T > t_0$ of interest. We discuss how to do this with simulation next.

3 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. For the classic analysis, the uncertainty and inference would then depend on the independence of residuals assumption that is typically implausible in these contexts. If we wish to incorporate autoregressive structure to address this independence assumption, it is less obvious how to use the model to form our counterfactual predictions. 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 $Y_t(0)$, but rather a $Y_t(1)$; any treatment impact in our lagged covariate 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.[4] 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. Subsequent sections discuss extensions to this overall approach.

[4] One could instead use maximum likelihood and some asymptotic approximations given the defined residual structure; we argue the parametric simulation approach we use provides a flexible alternative that does not rely on delta method expansions.

3.1 Extrapolating with known parameters

We initially assume that our 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. We assume the model given by Equation 4.

Using this model 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 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.

Our final prediction Y_T is a noisy prediction, however, because, for example, it could be high or low depending on the last residual draw. 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:

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

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

We can also examine our collection of simulations to get a plausible range of what we might see, under our modeling assumptions, in order to capture uncertainty. In particular, the middle 95% of our simulated $Y_T^{*(r)}$ forms a 95% prediction interval of what we would expect to see, 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 our model does not extrapolate to time T, suggesting 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 of how far we have drifted from the predicted trend (this is the quantity that could potentially be viewed as an impact). This prediction interval captures the uncertainty of the month-to-month variability of the observed trend. To see this, note how our autoregressive series gives a conditional prediction for each time point: given time $T - 1$, our prediction for time T is the structural component plus the autoregressive part of the residual. Under this view, write the final predicted outcome at time T as an average of the conditional predictions given time $T-1$:

$$\widehat{Y}_T = \frac{1}{R}\sum_{r=1}^{R}\left[\beta_0^{(r)} + \beta_1^{(r)}T + \beta_2^{(r)}Y_{T-1}^{*(r)}\right].$$

Even though there is no final $\epsilon_T^{(r)}$ in the above it is equivalent, up to simulation error, to the simple average of the simulated $Y_T^{*(r)}$ because the average of the $\epsilon_T^{*(r)}$ is 0. This shows that the variation in our predictions combines the variation in the prediction itself with the additional ϵ_T, which is the independent variability. Under this decomposition, the variation due to the ϵ_T is the variation of the observed series, and the remainder is the variation of the structural trend and autoregressive variation. All of this depends on correct model specification, in particular, the assumption that our observed post-policy series has the same month to month variation as our pre-policy series (i.e., homoskedasticity).

The major caveat to this process is we do not know the true 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.

3.2 Incorporating uncertainty in the parameters

So far we have assumed our parameters are known with certainty, but they themselves are estimated. If, for example, we estimate our rate of change to be too slow, than our extrapolations will steadily lag behind what they should be by more and more as t increases. 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 $\hat{\theta}$, pick 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 the 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 $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ 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).

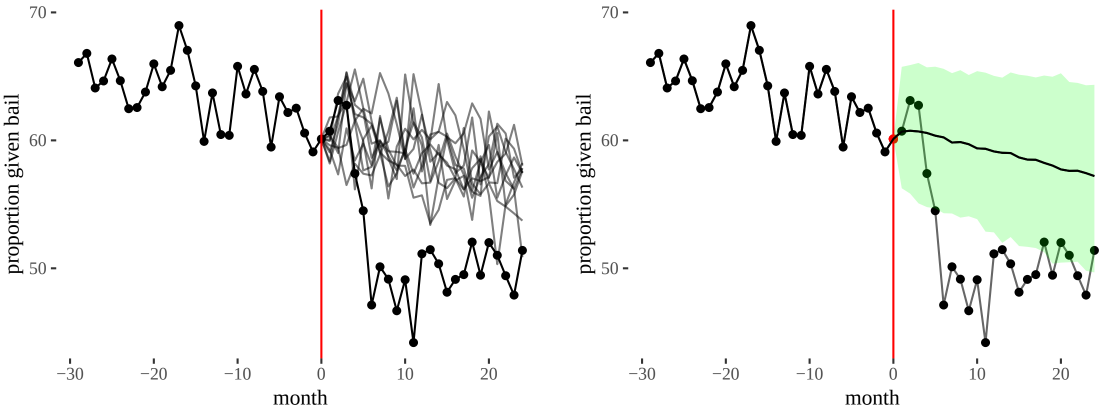
This approach is essentially Bayesian: the parameter draw step is similar to drawing a plausible value from a posterior distribution on the true θ (the implied prior here is implicitly a flat prior on the coefficients, roughly meaning that we are not differently preferring one value of θ over another). Under this view, the simulations constitute a posterior predictive distribution (see, e.g., Gelman et al. (1996)) for Y_T and the $\hat{Y}_T$ is the posterior mean predicted outcome given all the pre-policy data and the model. 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, e.g., 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 are also avoiding the asymptotic approximations used in maximum likelihood inference.

Remark. Estimation involves uncertainty, and when fitting a lagged variable we have a range of possible coefficients for ρ that could include values larger than 1 or less than 0. This can cause difficulties; in particular, if the uncertainty on the coefficients carries the coefficient for the prior Y to more than 1, those associated projections will compound exponentially and be nonsensical. This happens when there is little model stability in the fitting of the model (e.g., with few months of data, in particular), or if there are large nonlinearities in the pre-policy which make the estimated ρ coefficient large to compensate. If the coefficient is negative, the predictions can oscillate, again in a nonsensical manner. If either happens only in the extreme draws of the posterior, there is no major concern as the prediction intervals will trim these extremes. If they are more frequent, the confidence intervals will give overly wide ranges that signal the model fitting issues.

3.3 Case Study: Mecklenberg County

Mecklenberg instituted a series of reforms including changing their pre-trial risk assessment tool to a tool called 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 not appear at future court hearings or break further laws while awaiting trial,



(a) Ten extrapolated series.

(b) Envelope from 10,000 simulations.

Figure 2: Results of Mecklenberg analysis. At left ten sample simulated series along with observed data. At right the overall envelope of plausible series given pre-policy data. We see that 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.

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 4 to the Mecklenberg data displayed on Figure 2. Our estimated coefficients are $\widehat{\beta_0} = 45$, $\widehat{\beta_1} = -0.12$, and $\widehat{\beta_2} = 0.26$. The lagged term ($\hat{\beta}_2$) 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 there may actually be some small amount of autocorrelation. 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. In other words, the failure to find significant correlation does not demonstrate that it is not present; it could be nonsignificant due to a power issue.

Using this 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. Pre-policy trends do not tend to fall as far as what actually occurred. Interestingly, we see that the observed outcomes for the first four months after the policy change are still potentially following the pre-policy trend; the departure is only really significant at month 5 and 6. At this point, actual bail mostly levels off at the reduced rate of around 50%. Patterns such as these raise important issues of how to ascribe the change: was this drop at month 5 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.

Also important is that impact is the difference of the *projected* trend and the actual. 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 2 years post policy we cannot rule out an absence of impact: perhaps the policy forced the change earlier than it might have been, but we may have reached those bail levels regardless.

But then again, 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 wider uncertainty in later months is only due to estimation error, and is still dependent on the model being correctly specified. This includes the assumption that the pre-policy process would have continued unabated in the absence of the particular policy change implemented. 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 such analyses, the first two of which can be quantified: (1) parameter estimation error for the model, (2) the natural variation due to month-to-month changes and associated auto-regression from the last time point in the pre-intervention period, and (3) model specification.

4 Including seasonality effects

In some jurisdictions, when a person is arrested the arresting officer can elect to serve a summons, where the officer gives the arrestee a court date for a future appearance and then lets them go home, or serve a warrant, which could result in detention until a hearing is made to determine whether the defendant should be given bail or otherwise be supervised before trial. One aspect of a policy rooted in risk assessment might be to change policing behavior towards only giving warrants for more serious offenses. An outcome of interest that assess this might be the total number of warrant arrests made.

Unfortunately, the (synthetic) series of total number of warrant arrests pictured on Figure 1b shows a strong periodic trend across the years.[5] In particular, we see reduced number

[5] These synthetic data are inspired by seen trends in actual evaluations; we do not use actual data due to not having finalized those evaluations at this time.

of arrests when it is winter, and more in summer. Such trends are not uncommon when looking at counts of cases; these seasonal cycles are likely due to factors such as increased time spent indoors and away from the public eye during the colder winter months.

We could fit our linear model to capture the underlying trend, and hope that the clear and consequent autocorrelation would capture uncertainty correctly. However, a simple autocorrelation model on the residuals would miss the cyclic nature of our trend, which means we have clear model misspecification and which, in this case, would result in substantial loss of power (we examine this further below). We instead extend our linear model to model the periodic trend by introducing additional covariates such as indicator variables for the four seasons. We then augment the model to preserve the autoregressive element to 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. We list a few such approaches next. 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 additional 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} as 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 in our hypothetical state (temperature tends to be correlated with public/viewable criminal activity and thus number of arrests). 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:

$$Y_t = \beta_0 + \beta_1 t + \beta_2 Temp_t + \beta_3 Q_{2t} + \beta_4 Q_{3t} + \beta_5 Q_{4t} + \epsilon_t. \tag{5}$$

One potential concern with the dummy variable approach is the resulting curve will be a step function rather than a smooth curve, and the steps are at pre-specified points and are not data driven. Alternatively, therefore, we can fit a sinusoidal trend by building two covariates that correspond to the sin and cos 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. One model using this approach would be

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

Different coefficient values for ρ_1 and ρ_2 control where the peaks and valleys of this trend are. To illustrate these four fitting approaches, see Figure 3, which shows simple fits (without any lagged variables) of these models to the pre-policy data. Of the four models, the model with both quarter and temperature has the best pre-policy fit, with an estimated residual standard deviation of 295 vs. above 400 for the other models.

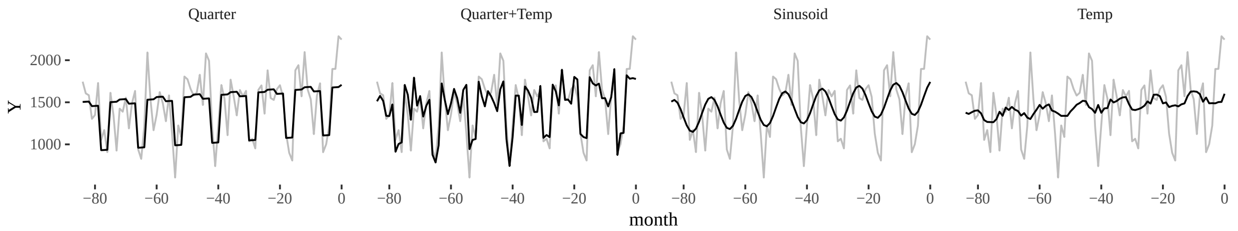


Figure 3: Four seasonality models for number of warrant arrests. Black is predicted outcome, grey is actual. Only pre-policy series shown; models fit without lagged covariates.

4.1 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 periodic structure which we need to remove. The autoregression should be on the deviation from trend only. The fix, however, is relatively straightforward: we also include lagged values of the covariates. We next discuss why this is the case.

First take the general structural model with autoregressive residuals of

$$Y_t(0) = f_\beta(X_t) + \epsilon_t,$$

with $f_\beta(X_t)$ being a model of covariates X_t (where X_t is a vector of potentially time-varying covariates including t itself) indexed by some parameter vector β. The $f_\beta(X_t)$ is the structural aspect of our model and our residuals are then $\epsilon_t = Y_t - f_\beta(X_t)$. We assume that once we remove the structure we have a stationary[6] autoregressive process given by Equation 3 on the residuals. To connect to the above, we have heretofore assumed that $f_\beta(X_t) = \beta_0 + \beta_1 t$ with $\theta \equiv (\beta_0, \beta_1)$.

With this more general model, using Equation 3 and the consequent $\epsilon_{t-1} = Y_{t-1} - f_\beta(X_{t-1})$, we have

$$\begin{aligned} Y_t &= f_\beta(X_t) - \rho\epsilon_{t-1} + \omega_t \\ &= f_\beta(X_t) + \rho\left[Y_{t-1} - f_\beta(X_{t-1})\right] + \omega_t \\ &= f_\beta(X_t) - \rho f_\beta(X_{t-1}) + \rho Y_{t-1} + \omega_t. \end{aligned} \tag{6}$$

Now consider the case when f_β is a linear model, with $f_\beta(X_t) = X_t'\beta$. For example, for Model 5 we would have $X_t = (1, t, Q_{t2}, Q_{t3}, Q_{t4}, Temp_t)$. Plugging $X_t'\beta$ in to the more general Equation 6 gives

$$\begin{aligned} Y_t &= X_t'\beta - \rho X_{t-1}'\beta + \rho Y_{t-1} + \omega_t \\ &= X_t'\beta - X_{t-1}'\beta_\ell + \rho Y_{t-1} + \omega_t, \end{aligned}$$

[6] "Stationary" means the autoregressive structure is constant across time, i.e., that the auto-correlation remains the same.

with $\beta_\ell = -\beta\rho$. The second line above shows that, if we do not insist on keeping the structure of the same β in both the X_t and the X_{t-1} terms in the above, we can simply regress Y_t on X_t, X_{t-1} and Y_{t-1}, dropping the constraint of $\beta_\ell = -\beta\rho$.

There is a small technical caveat: the lagged covariates can frequently be collinear with the contemporaneous covariates, thus 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 $X_{t,2}$.[7] This colinearity is easily resolved, however: simply drop any collinear columns (in particular the intercept and time variables), allowing the parameters to estimate the combined influence of both the primary observation and the structural component of the lagged outcome due to that variable.

Remark. This model fitting process relaxes some of the structure of the parameters. In particular, we no longer enforce $\beta_\ell = -\beta\rho$. However, given the consistency of estimation for linear regression, we immediately have that as n increases our OLS approach will converge on the correct parameterization, giving overall consistency even if we drop this constraint. Interestingly, as a model check, given the relaxation one could compare $-\hat{\beta}\hat{\rho}$ to $\hat{\beta}_\ell$. They should be the same, up to estimation error.

To further examine why we have the lagged covariates in our model, write the above as

$$Y_t = (X_t - \rho X_{t-1})'\beta + \rho Y_{t-1} + \omega_t.$$

This formulation suggests that our regression is, in effect, regressing our outcome onto the differences of our covariates (including any linear time term or intercept) with the lagged covariates scaled by our unknown ρ parameter. We have this differencing component because the lagged Y_{t-1} also includes those structural components and we would not want them to be counted twice. Overall, the lagged covariates allow us to estimate and then subtract out the lagged structural part of the Y_{t-1}, leaving the lagged residual as desired.

Alternate estimation strategies. We use the lagged outcome model and OLS due to its simplicity, and to take advantage of the easy ability to simulate from the fitted model. Other options are possible, which we briefly acknowledge and outline next.

First, there is *Feasible Generalized Least Squares (FGLS).* Here we would explicitly model the residual structure as AR1 and use feasible generalized least squares by first estimating the model using simple OLS, and then estimating the implied residual matrix with the empirical residuals. While entirely viable, we did not use this approach as it is unclear how to incorporate uncertainty of the estimated residual covariance matrix in the simulation. We also wanted to take full advantage of the `sim` function from the `arm` package; we therefore

[7] This colinearity is why the simple lagged outcome model does not have an extra term beyond the lagged outcome itself.

instead use OLS in order to leverage existing packages to obtain the pseudo-posteriors. That being said, FGLS coupled with classic maximum likelihood estimation would provide asymptotic confidence intervals based on the normal approximation.

One could also employ, *iterative model fitting.* In particular, if we had the ϵ_{t-1} we could use them as covariates in our regression instead of the Y_{t-1}. This motivates first estimating them using a model without lagged covariates, and then using those estimates in a second run of our model. First fit $Y_t = f(X_t) + \epsilon_t$ using OLS. Then calculate $\hat{D}_t = Y_t - \hat{Y}_t$ with $\hat{Y}_t = \hat{f}(X_t)$. Then refit a model of

$$Y_t^{(2)} = f(X_t) + \rho \hat{D}_{t-1} + \omega_t$$

and calculate new differences

$$\epsilon_t^{(2)} = Y_t - \hat{Y}_t^{(2)} = Y_t - \hat{f}^{(2)}(X_t).$$

Repeat until convergence. Under this approach to get a model fit, it is unclear how to simulate to get prediction uncertainty, or how to otherwise codify uncertainty of our model parameters correctly.

Finally, one could simply specify the model and fit it using a Bayesian model-fitting package such as `stan` (Carpenter et al., 2017). This would be functionally equivalent to the above, although there would no longer be need to incorporate lagged covariates or outcomes as the model could directly work with the latent residuals. One would also have to explicitly choose priors, and this approach may feel less accessible to many practitioners than our approach rooted in classic linear regression.

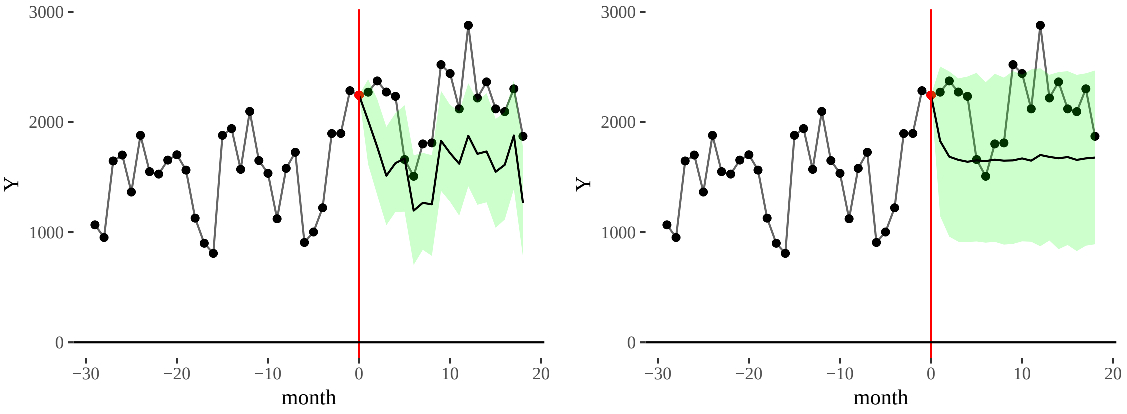
4.2 Case study: Number of warrant arrests

To illustrate our simulation extrapolation using a model with a seasonality component, we fit the best-fitting Model 5 with the autoregressive residual model of $\epsilon_t = \rho\epsilon_{t-1} + \omega_t$. As before, we cannot directly fit this model with ordinary least squares due to the unobserved residuals. We, following the above, extend the model to include the lagged outcome and lagged covariates. This gives

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

Using this temperature and quarter model, now with lagged covariates and outcomes, we generate the predictive envelope following the process described above. See results on Figure 7a.

By comparison, if we had not included a seasonality model and instead simply fit our simple linear trend model, we get Figure 4b. The raw model shows a stronger autocorrelation



(a) Projection with seasonality.

(b) Projection without seasonality

Figure 4: Prediction envelopes for number of warrant arrests. Time period (x-axis) truncated to show more detail of model fitting in post-policy era. Left shows seasonality model, right shows simple linear model with no seasonality.

(estimated as 0.63 vs. 0.50), because points near each other are correlated due to the periodic trend around the base linear model, while 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 and thus we end up with a conservative inference.

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 imprecisely estimated autocorrelation with resulting simulated series having some with large degree of autocorrelation, to capture the overall uncertainty. Determining when to fit the more complex model vs. 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 unstable and did not increase precision.

5 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. Consider, for example, Figure 5a: only the middle months are outside the envelope. In this section we discuss some more formal inferential strategies, 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. We first talk about inference, and improving power with simple averaging, and then move to strategies for the estimation and display of the entire curve of impacts post-policy.

5.1 Inference

To introduce 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.

In this view, our observed outcome Y_T^{obs} is our observed value of the test statistic of outcome at time T: we compare this observed value to the simulated values that capture what our model says is possible. If the observed value is outside the central range of these simulated values (which we consider our reference distribution), we reject the null that the pre-policy trend continued unabaited (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, which is a potentially highly variable single point Y_T^{obs}, to a distribution of simulated values Y_T^* that all themselves could be quite variable. In particular, 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. This is akin to a posterior predictive check 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, i.e., conclude that something changed our trajectory from what we would have expected. One can calculate nominal p-values by calculating the proportion of time the observed test statistic is exceeded by the reference distribution. These are called *posterior predictive p-values*, and do not necessarily have strictly valid frequentist properties, but they are argued to generally be conservative (Meng et al., 1994). Also see Robins et al. (2000).

To be specific, take our observed series $Y = (Y_1, \dots, Y_T)$ 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 $Y_{t_0+1}, \dots, Y_T$). 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 can also calculate nominal p-values using the quantile q of our observed t^{obs}, with $p = \min(q, 1 - q)$ (for a two-sided test).

5.2 Smoothing

In investigating a place-based initiative we generally want to understand the evolution of the impact over time as the policy continues to evolve. For example, as we saw with the Mecklenberg results on Figure 2b, it appears as if the policy had 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 use the simple averaging method from above, and did not look at the overall graph of impacts, we would lose this nuance. But the graph is noisey, 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, usually without imposing a global structural model to ensure local structural variation can come through. Smoothing is generally nonparametric, and can be done with splines, taking averages within a sliding window, or using loess (Locally Weighted Smoothing).

We can easily use smoothing coupled with our inferential approach above. In particular, we can smooth each of the simulated time series using a specific (pre-specified) smoothing method. We then compare the distribution of these smoothed time series to the actual time series smoothed in the same way. Under our null hypothesis, the smoothed observed trajectory should be exchangeable with any of the smoothed simulated trajectories. Now our smoothed estimate at a given timepoint T is our test statistic, and the distribution of smoothed estimates of our simulated series at that same time point serve as a prediction distribution of what values we might expect if the policy had no impact. This should have greater power: we are now examining the overall trend in the neighborhood of T, which potentially increases precision as idiosyncratic monthly variation will get averaged out.

To implement this procedure, we generate our collection of series as before. We then smooth each series using a fixed process, such as a loess smoother with a specified bandwidth.

Finally, we smooth our observed series using the exact same method. We then compare the observed smoothed series to the distribution of smoothed series. One caveat is that, if there is a sudden change in policy impact, if we are smoothing across t_0 we can cause the smoothed line of our observed series to artificially deviate from the *pre-policy* trend since the post-policy points will be included in the local average near the policy change. Similarly, the pre-policy timepoints near t_0 will drag the smoothed post-policy timepoints near t_0 towards their values, potentially masking impacts. To avoid this, one could choose to smooth the post-policy series only, not including any pre-policy points in the smoothing process; if this is done, then it needs to be done for both the distribution of 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.

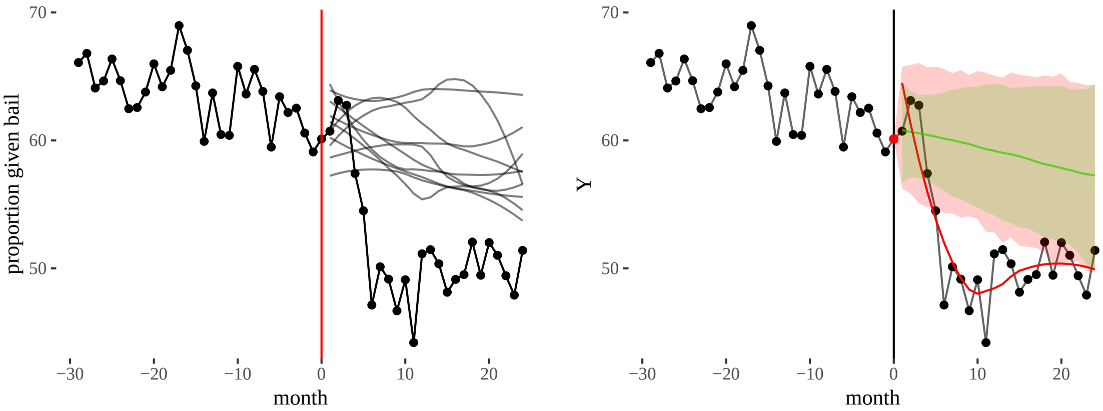
5.3 Mecklenberg County, continued

We continue our Mecklenberg example discussed above by showing how to improve power using averaging and loess smoothing. We first compare the average rate of bail setting over the first 1.5 years of the policy to what we believe would have happened, had pre-policy trends continued unabated. The difference in these two quantities is the average impact, which is an overall summary measure of interest to policy makers. We first calculate the observed average of the first 18 months of bail. We then calculate, for each generated series, the average rate of simulated bailsetting over the first 18 months; these calculations give a distribution of what average bailsetting rates we would have expected, absent any change in trend.

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 would therefore conclude that something changed the pre-policy trajectory so we are seeing lower rates of bail-setting than we would have expected. We can also take the difference to get estimated impacts. Here we would give a 95% confidence interval (technically a credible interval) for the true average impact being in $(-3\%, -12\%)$. To get a point estimate for the impact, we would average the simulated averages and take the difference. Here we predict bail setting of 59%, and thus we have an estimated impact of -7%.

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

We can also use loess smoothing to smooth the post-treatment trajectory. We first smooth our observed smoothed 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 therefore fit the same smoother to each of our simulated series, ignoring the pre-policy points there as well. Results are on



(a) Ten extrapolated smoothed series

(b) 10,000 smoothed simulations

Figure 5: Results of Mecklenberg analysis (with smoothing). Left shows how smoothed trajectories have less variability than the raw series did. Right compares smoothed envelope with envelope without smoothing. We see less variability. The heavy line denotes smoothed observed trend to be compared to envelope and counterfactual trend.

Figure 5. On the left we see 10 smoothed trajectories in the post-policy period. On the right we have our 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 arguable 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 the loess smoother being more variable at the endpoints.

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.

5.4 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 in terms of what outcome we would expect at a given time point, given the overall patterns.

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. We, for each time series (observed or simulated) smooth as follows:

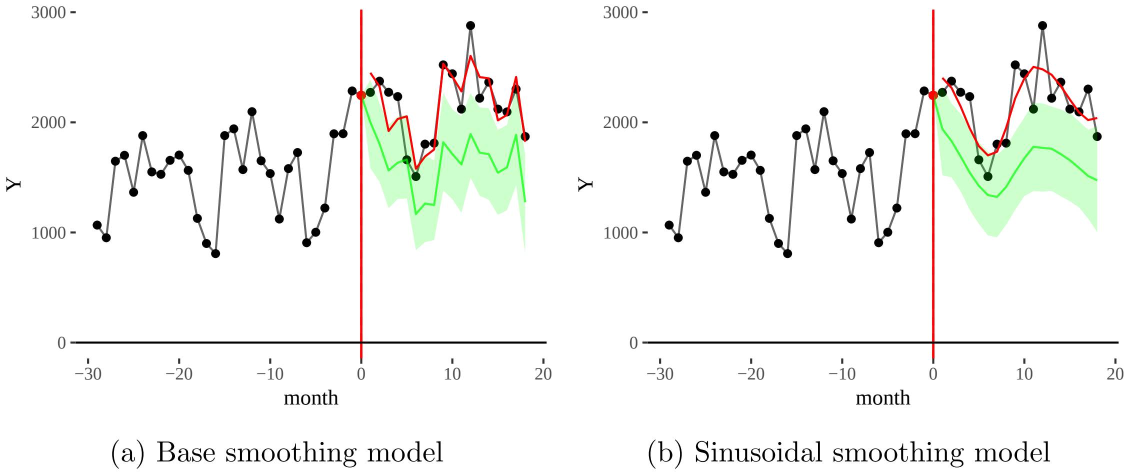
1. Fit a seasonality model to the data. This is not the original seasonality model, but a new model. There is no need for lagged outcomes or uncertainty estimation in this model. As before we can choose to fit to post-policy data only, all data, or pre-policy data.
2. Predict all the outcomes given our 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 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 still maintains the overall structure. In particular, any estimated seasonality component will not get smoothed and so if there is a strong cycle it will not be removed by the smoothing process.

If we repeat this for each simulated series, we get a null distribution of what distribution of smoothed series we would see if the model were correct. We then do a final step of fitting our observed data with the same smoothing model, and compare this series to the distribution of predicted series.

5.5 New Jersey, continued

To see smoothing with a seasonality model in practice, we extend our analysis of warrant arrests with a temperature and quarter seasonality model. We compare two methods of smoothing. In both, we use the same model for extrapolation, but we will calculate our residuals using two different models. We first fit the base model with quarter and temperature, and then for comparison fit the sinusoidal model without temperature. In either case we first fitting our non-lagged model to each synthesized series and then smooth the residuals with a loess smoother. Our second 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 6. At left we see that the month-to-month variation predicted by the temperature changes is not smoothed, and so we have a more jagged sequence. On the right, these have been smoothed over to show underlying structure. We see that there are increased numbers of warrant arrests, and the difference appears fairly constant over time. This structure is a summarization of the larger gross trends, however, and we should acknowledge that there is more serious bias-variance tradeoff at play.



(a) Base smoothing model

(b) Sinusoidal smoothing model

Figure 6: Prediction envelopes for number of warrant arrests using smoothing. Time period and y-axes truncated to show more detail of model fitting in post-policy era.

Remark. For smoothing, it is important to fit the same model to the same range of data; otherwise However, if there is a large treatment impact, the model fit to the full observed series could be misspecified. This could give an odd smoothed series for the observed data. But if the model fitting process is held to be the same, then comparing the observed series to the reference distribution is still valid for testing. For estimation, however, one might be more cautious. Overall, we recommend selecting a smoother that is not overly dependent on the pre-policy patterns, but instead naturally fits to the observed data. In particular, we suggest fitting the seasonality model to the post-policy data only.

The key idea is that the working model used to smooth the data is not needed to be a correctly specified model, in terms of testing. It is purely to extract any seasonal structure so it does not get over-smoothed.

6 Adjusting for time-varying covariates

The montly outcomes in our examples are summaries of individual data, and if the composition of the individual data are changing we might see changes in the aggregate summaries due to factors other than the policy reform being investigated. For example, in our context, the charges associated with an arrest fall in different gross categories, and these categories tend to be treated differently due to having generally different levels of severity. For example, in Mecklenberg, charges fall into Traffic, Misdemeanor, and Felony and, as shown in the left of Figure 7, the overall rates of misdemeanors is falling from before the policy shift to well after. This means the mix of cases (as captured by being misdemeanor, felony, or traffic related) at each month is changing, and so post-policy we have proportionally more felony cases, which are typically more severe and would likely involve greater rates of detention. Thus our outcome of interest, percent bail, may be being impacted by this changing case mix.

We therefore want a sensitivity check to test whether our overall findings are being partially or fully driven by this change in case composition, rather than change in how cases are being treated by the judicial system. In other words, we want to adjust for case composition in our estimation process.

Post-stratification is one approach for handling this type of problem when one has access to the individual data composing the overall outcome. With post-stratification, we, instead of calculating the simple average outcome for a month, reweight the individual cases so that the proportion of each case type matches some canonical distribution. For example, in Mecklenberg, 33% of the charges are felony, 54% misdemeanor, and 13% traffic across all our post-policy months. Therefore, for each month, we calculate an adjusted rate of bail by first reweighting the cases in that month to match these percentages before averaging to obtain the outcome. (This is the same as first calculating the proportion of cases given bail for each of the three categories, and then averaging these proportions with the weights of 33%, 54% and 13%.) After doing this for each month, any differences in our adjusted outcomes across months can be ascribed to how the cases are being treated differently rather than being due to different types of cases, as captured by the covariates.

This attribution does rely on the assumption, however, that the measured covariates being adjusted for capture all the case differences that both are changing across time and are associated with the outcome. For example, if the proportion of misdemeanors is going down, but the (unobserved) severity of the remaining misdemeanors is going up, then a change in outcome may still be connected to changing case mix as captured by this unobserved severity. That being said, using post-stratification as an adjustment technique can help explore whether it is plausible that observed potential impacts are simply due to change in case mix, and provide a good way of conducting a sensitivity check on one's results. This approach has ties to matching or propensity score reweighting. In particular, we refer to template matching (Silber et al., 2014), where the mixture of patients within each of a collection of hospitals are reweighted to match a reference distribution to ease comparability across the hospitals. We here would match across time.

6.1 Reweighting for post-stratification

The first step for reweighting is to identify the target case mix that we wish to reweight each month to match. We recommend doing this via identifying a time period of interest (e.g., the year following the policy change) and calculating the proportion of cases of each type in that period, storing them as a vector of proportions π^*. For Mecklenberg, to illustrate, this is

$$\pi^* = (\pi^{F*}, \pi^{M*}, \pi^{T*}) = (33\%, 54\%, 13\%).$$

The next step depends on the type of outcome being considered. There are two core types of outcome: total counts and mean outcomes. The former is, for example, total number of cases that had an FTA. The second is, for example, the average days spent in jail for a case, or the proportion of cases that were given bail. (The proportion of cases assigned bail can be thought of as the average of a 0/1 indicator variable for getting assigned bail.)

To adjust the outcomes based on the proportions of cases in each subgroup, we, for each month t, first calculate the outcome for each subgroup. Call these Y_t^s, with s being in our case F, M, or T. This could be total count, average outcome in the subgroup, or the proportion of the subgroup with a given outcome.

Next we re-weight the observed overall outcome (there is no modeling here) depending on type of outcome. For means and proportions, we calculate

$$Y_t^* = \sum_s \pi^{s*} Y_t^s.$$

For count, we calculate, if we let N_t be total number of cases at month t and N_t^s total number of cases in subgroup s at month t,

$$Y_t^* = N_t \sum_s \pi^{s*} \frac{Y_t^s}{N_t^s}$$

(This formula comes from first changing our outcome to the mean outcome in the group, getting the estimated mean for the whole month using the group weights, and then scaling back to raw count.)

Use the above formula to adjust all months, both pre- and post-policy. This gives an adjusted time series where we have controlled for the strata considered. This series could diverge from the raw series, if the proportions are changing and if the average outcomes Y_t^s differ across s.

We now, at this point, simply fit our normal ITS model on the adjusted series. This is effectively weighted regression, where we have re-weighted units at the individual level. (It is a bit odd in appearance in that we do not model the individual level units but instead aggregate.)

Remark. Other adjustment approaches are also possible. One alternate method of post-stratification would be to divide the cases into subgroups and fit several distinct ITS models, one for each subgroup. Unfortunately, especially given our approach to handling seasonality and the auto-regressive structure, we would want to account for possible dependence between the subgroups (e.g., they all may have higher or lower outcomes in a given month in some correlated fashion). How to do this when each subgroup is fit separately is unclear. This is why we instead reweigh each month by the subgroup proportions and then fit the ITS model to the resulting combined series.

One could alternatively implement a version of the above scheme with a single modeling step where we include covariates such as proportion of cases in the three categories, and then allowing for trend by covariate interactions. In particular, one would add interaction terms between the intercept and the time covariates and the proportion in each group to the model. We would then post-process the estimated coefficients after. This approach is most similar to "controlling" for variables in a regression. We discuss this approach in more detail in the supplement.

The simple version we presented above is to adjust the raw data by aggregating the data with weights, and then conduct the analysis on a single aggregated dataset. More complex versions are to in effect fit individual models to the subgroups, and then aggregate the model results. This could in principle be more powerful as we are using the proportion of cases in a given month as a covariate to predict variation, which could lead to smaller standard errors. Unless these proportions are highly predictive of outcome, the gains from this will likely be minimal. We leave exploring this approach more fully to future work.

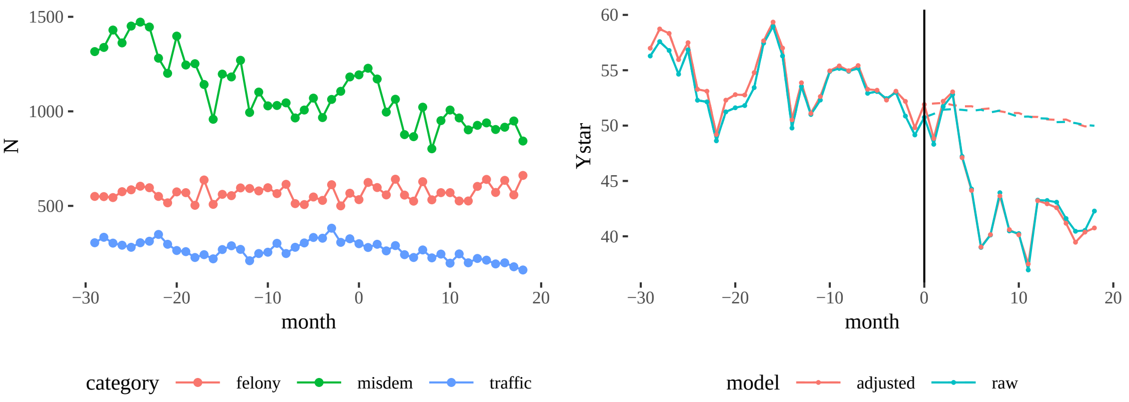
6.2 Case study: A sensitivity check for Mecklenberg

We now formally use post-stratification to investigate whether our original results are sensitive to the changing case mix of cases in Mecklenberg over time. To do this we first calculate what proportion of the charges were in each class post-policy (we select post-policy to capture current trends rather than historical, but this is a design decision up to the analyst). As stated above, we find 33% of the charges are felony, 54% misdemeanor, and 13% traffic. We then reweight each month pre- and post-policy, getting an adjusted rate of bail setting as the weighted average of the rates within each of our three groups. Finally, we use our standard lagged outcome analysis on these reweighted totals to get our predicted trend line. The results are on Figure 7b; note how both our counterfactual extrapolation and the observed trend are impacted by the reweighting. That being said, we see little change in the overall impacts, and therefore our substantive results remain the same. There is no evidence that the results are substantially driven by the change in case mix, as measured by category of case.

7 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 easy or good comparison regions available.

Our modeling framework captures uncertainty in a flexible manner that allows for dependencies in outcomes from month to month, something that the more naïve approaches lack. We here study ITS designs with many pre-policy timepoints; with fewer timepoints,



(a) Number of cases of each charge type.

(b) Adjusted impact on bail setting

Figure 7: Adjusting the impact on bail setting by charge type in Mecklenberg. At left we see the number of misdemeanors falling at a non-linear rate; this causes the mix of cases to be changing in a complex way over time. At right our adjust rate of bail setting and associated impact analysis when we reweight each month to have a canonical mix of cases.

however, estimating the autoregressive component of the model is generally be much more difficult. In maximum likelihood approaches, it is known that this can cause bias and poor coverage when there are only 5 or so observations (see, e.g., St. Clair et al. (2016)). We leave whether simulation instead of asymptotic approximation would help in these short ITS designs to future work.

Our framework, while useful, also has some areas that call for development. In particular, especially due to explicitly taking into account autocorrelation, this approach does not easily lend itself to straightforward 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 repeatidly 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. Developing this is an important next step for this analytic approach.

The modeling itself could also potentially be extended. 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, 2015). 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.

There are also further 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 preposition that a found change was caused by the policy shift.

Appendix A: A few small lemmas

We here provide the small lemmas and derivations that complement the text.

Standard errors for classic OLS-based ITS

Lemma 1. *Under the classic OLS approach, the standard errors for the $\widehat{\Delta}_k$, $k > t_0$ are*

$$\widehat{SE}\left[\widehat{\Delta}_k\right] = \left(1 + S_{00} + 2kS_{10} + k^2 S_{11}\right)^{1/2} \hat{\sigma} \geq 1,$$

with the S_{00}, S_{11} and S_{01} being the elements of the variance-covariance matrix one would obtain for the coefficients of a simple intercept-slope regression of the outcome on the prepolicy data only.

Proof: The design matrix X of our regression consists of a column of 1s, a column with the time values $1, \ldots, T$, and one column for each post-policy timepoint, where the column for time point k has all 0s except a single 1 at row k. Given this, $X'X$ is a $(2+K) \times (2+K)$ block matrix:

$$X'X = \left[\begin{array}{cccccc} T & \sum_t t & 1 & 1 & \cdots & 1 \\ & \sum_t t^2 & t_0+1 & t_0+2 & \cdots & t_0+K \\ & & 1 & 0 & \cdots & 0 \\ & & 0 & 1 & \cdots & 0 \\ . & . & . & . & \cdots & . \\ & & & & \cdots & 1 \end{array}\right] = \left[\begin{array}{cc} A & B \\ B' & I \end{array}\right],$$

where K is the number of post-policy timepoints (and the number of our different Δ_k we are estimating). Note the bottom-right block is a $K \times K$ identity matrix. A corresponds to what we would get from the simple linear regression of Y on the months $1, \ldots, T$ with an intercept.

Classic OLS gives our standard errors for our coefficients as the diagonals of $(X'X)^{-1}\hat{\sigma}$. We next calculate $(X'X)^{-1}$. The inverse $(X'X)^{-1}$ will give another symmetric matrix defined by the block matrix

$$(X'X)^{-1} = \left[\begin{array}{cc} (A - BI_K^{-1}B')^{-1} & -A^{-1}B(I_K - B'A^{-1}B)^{-1} \\ & (I_K - B'A^{-1}B)^{-1} \end{array} \right].$$

The standard errors for our impact estimates are governed by the bottom-right corner of the above. We can simplify the bottom-right corner by using the Woodbury identity of

$$(a + cbc')^{-1} = a^{-1} - a^{-1}c(b^{-1} + c'a^{-1}c)^{-1}c'a^{-1}$$

and $(-M)^{-1} = -M^{-1}$ to get

$$\begin{aligned} (I_K - B'A^{-1}B)^{-1} &= I_K^{-1} - I_K^{-1}B'\left[-A + B(I_K)B'\right]^{-1} BI_K^{-1} \\ &= I_K - B'\left[-A + BB'\right]^{-1} B \\ &= I_K + B'\left[A - BB'\right]^{-1} B. \end{aligned}$$

We can then simplify the second term further. Let X_{pre} be the entries in the first $T - k$ rows and the first two columns of the design matrix. This is the design matrix of the simple regression on pre-policy units only. Then we have $A - BB' = X_{pre}'X_{pre}$. To see this note that BB' has the form

$$BB' = \left[\begin{array}{cc} k & \sum_{j=1}^{k} T - k + 1 \\ & \sum_{j=1}^{k} (T - k + 1)^2 \end{array} \right].$$

Thus, subtracting BB' from A simply takes off the last elements of the sums. Therefore $(A - BB')^{-1} = (X_{pre}'X_{pre})^{-1} = S$, the variance-covariance matrix for the coefficients of our simple pre-policy regression. This gives

$$(X'X)^{-1} = \left[\begin{array}{cc} (X_{pre}'X_{pre})^{-1} & -A^{-1}B(I_K - B'A^{-1}B)^{-1} \\ & I_K + B'(X_{pre}'X_{pre})^{-1}B \end{array} \right].$$

Finally, take the r^{th} column of B, which corresponds to the r^{th} post-policy period at time $t_0 + r$. This column is $(1, t_0 + r)$. We then have

$$\left(B'(A - BB')^{-1}B\right)_{rr} = \left(B'SB\right)_{rr} = S_{00} + 2(t_0 + r)S_{10} + (t_0 + r)^2 S_{11} \geq 0. \qquad (7)$$

The inequality is because $B'SB$ will have a nonnegative diagonal as $v'Sv \geq 0$ for any vector v due to S being positive semi-definite. Since this bottom right corner of the variance-covariance matrix is the above plus the identity matrix, we finally have our result.

□

This lemma shows how our fundamental error is due to the initial 1 in the sum. The remaining terms, in particular the $(t_0+r)^2 S_{11}$ term, correspond to the standard errors in the intercept and slope estimated on pre-policy data being extrapolated. In particular, note how as r increases, these terms grow quadratically in the variance and linearly in the standard error. This is due to uncertainty in the slope causing increasingly large levels of extrapolated uncertainty.

Connection of lagged outcome model to residual dependence

Our original goal was to allow for dependencies of the residuals in our linear model, but we used lagged outcomes instead. These are the same thing, with different interpretation of the parameters in the model. To see that a lagged model on Y is the same as a model with lagged dependent residuals, take our simple regression model (Equation 2) with its lagged residual model (Equation 3). Then

$$\epsilon_{t-1} = Y_{t-1} - \beta_0 - \beta_1(t-1)$$

and

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 t + [\rho\epsilon_{t-1} + \omega_t] \\ &= \beta_0 + \beta_1 t + \rho\left[Y_{t-1} - \beta_0 - \beta_1(t-1)\right] + \omega_t \\ &= \beta_0 + \rho(\beta_1 - \beta_0) + \beta_1(1-\rho)t + \rho Y_{t-1} + \omega_t \\ &= \tilde{\beta}_0 + \tilde{\beta}_1 t + \tilde{\beta}_2 Y_{t-1} + \omega_t. \end{aligned}$$

This shows $\tilde{\beta}_0 = \beta_0 + \rho(\beta_1 - \beta_0)$, $\tilde{\beta}_1 = \beta_1(1-\rho)$, and $\tilde{\beta}_2 = \rho$. Fitting our lagged outcome model will give estimates for $(\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2)$. We can then convert them to our target (β_0, β_1, ρ); it is simply a different parameterization. Some algebra gives the conversions to the residual model as

$$\begin{aligned} \rho &= \tilde{\beta}_2 \\ \beta_1 &= \frac{1}{1-\rho}\tilde{\beta}_1 \\ \beta_0 &= \frac{1}{1-\rho}\tilde{\beta}_0 - \frac{\rho}{(1-\rho)^2}\tilde{\beta}_1. \end{aligned}$$

The estimated residual variance is the variance of the ω_t.

Properties of the residuals. For our lagged residual model, we have $\mathbb{E}\left[\epsilon_t\right] = 0$ and, for any given t,

$$\begin{aligned} var(\epsilon_t) &= \mathbb{E}\left[(\epsilon_t - 0)^2\right] \\ &= \mathbb{E}\left[\rho^2\epsilon_{t-1}^2\right] + 2\rho\,\mathbb{E}\left[\epsilon_{t-1}\omega_t\right] + \mathbb{E}\left[\omega_t^2\right] \\ &= \rho^2\sigma_\epsilon^2 + 0 + \sigma^2. \end{aligned}$$

This gives

$$var(\epsilon_t) = \frac{1}{1-\rho^2}\sigma^2.$$

We also have

$$cov(\epsilon_t, \epsilon_{t-1}) = \mathbb{E}\left[\epsilon_t \epsilon_{t-1}\right] = \mathbb{E}\left[(\rho\epsilon_{t-1} + \omega_t)\epsilon_{t-1}\right] = \rho\sigma^2_\epsilon$$

meaning that the residuals have correlation ρ.

Appendix B: Post stratification with interacted models

In this section we discuss an alternative form of poststratification where each subgroup gets its own model, which we allow for by having interaction terms. Unlike the simpler approach in the main text, this approach could be extended to allow for completely different time trends for the different subgroups and could be used to improve the face validity of model if there were substantive reasons to believe such variation were present.

To illustrate, we take a context where we want to adjust for two types of cases, Drug and Violent. Consider one of our models, indexed by V for violent crime cases only:

$$Y^V_t = f^V_\beta(X_t) + \epsilon^V_t$$

with $\epsilon^V_t = \rho\epsilon^V_{t-1} + \omega^V_t$. The X_t are the same for both drug and violent crimes (i.e., we do not have crime-type specific covariates). The drug model is the same as above, but with D instead of V to index.

For any month t we have, letting π^D_t being the proportion of cases in month t in the drug classification, and $\pi^V_t = 1 - \pi^D_t$,

$$Y_t = \pi^D_t Y^D_t + \pi^V_t Y^V_t,$$

i.e., our overall outcome is a weighted average of the component subgroup outcomes. This expands into

$$\begin{aligned} Y_t &= \pi^D_t Y^D_t + \pi^V_t Y^V_t \\ &= \pi^D_t \left[f^D_\beta(X_t) + \epsilon^D_t\right] + \pi^V_t \left[f^V_\beta(X_t) + \epsilon^V_t\right] \\ &= \pi^D_t f^D_\beta(X_t) + \pi^V_t f^V_\beta(X_t) + \pi^D_t \epsilon^D_t + \pi^V_t \epsilon^V_t \\ &= \pi^D_t f^D_\beta(X_t) + \pi^V_t f^V_\beta(X_t) + \tilde{\epsilon}_t \end{aligned}$$

This looks like our original ITS model, but with more parameters and an extra set of covariates, i.e., the π^s_t, $s \in \{V, D\}$. The $\tilde{\epsilon}_t$ are our combined residuals; below we show how we can leave them aggregated given some assumptions.

Let our covariate vector be $X_t = (1, t, S_t)$ with S_t representing the seasonality covariates. If we assume a linear model of $f_\beta(X_t) = X_t'\beta$ for each case type we then also have a parameter vector $\beta = (\beta_0, \beta_t, \beta_S)$. Plugging into the above gives

$$\begin{aligned} Y_t &= \pi_t^D X_t'\beta^D + \pi_t^V X_t'\beta^V + \tilde{\epsilon}_t \\ &= \beta_0^D\pi_t^D + \beta_t^D\pi_t^D t + \pi_t^D S_t'\beta_S^D + \beta_0^V\pi_t^V + \beta_t^V\pi_t^V t + \pi_t^V S_t'\beta_S^D + \tilde{\epsilon}_t \end{aligned}$$

The above is a fully interacted model with interactions between proportion in the groups and the covariates (including time and intercept). One could identify a baseline group and re-write as

$$Y_t = \beta_0^V + (\beta_0^D - \beta_0^V)\pi_t^D + \beta_t^V t + (\beta_t^D - \beta_t^V)\pi_t^D t + \dots$$

but this makes subsequent post-processing of the model more difficult.

In the above our residual term can be simplified if we assume the autoregressive correlation ρ is shared across the groups:

$$\begin{aligned} \tilde{\epsilon}_t &= \pi_t^D\epsilon_t^D + \pi_t^V\epsilon_t^V \\ &= \pi_t^D\rho\epsilon_{t-1}^D + \pi_t^D\omega_t^D + \pi_t^V\rho\epsilon_{t-1}^V + \pi_t^V\omega_t^V \\ &= \left(\pi_t^D\rho\epsilon_{t-1}^D + \pi_t^V\rho\epsilon_{t-1}^V\right) + \left(\pi_t^D\omega_t^D + \pi_t^V\omega_t^V\right) \\ &= \rho\left(\pi_t^D\epsilon_{t-1}^D + \pi_t^V\epsilon_{t-1}^V\right) + \tilde{\omega}_t \\ &\approx \rho\tilde{\epsilon}_{t-1} + \tilde{\omega}_t \end{aligned}$$

The last line is an approximation which depends on how similar π_t^D is to π_{t-1}^D (i.e., if the proportions are changing slowly across time, this approximation is good). The error in this approximation is

$$\begin{aligned} error &= \left[\pi_t^D\epsilon_{t-1}^D + \pi_t^V\epsilon_{t-1}^V\right] - \left[\pi_{t-t}^D\epsilon_{t-1}^D + \pi_{t-1}^V\epsilon_{t-1}^V\right] \\ &= \epsilon_{t-1}^D(\pi_t^D - \pi_{t-1}^D) + \epsilon_{t-1}^V\left[(1 - \pi_t^D) - (1 - \pi_{t-1}^D)\right] \\ &= (\epsilon_{t-1}^D - \epsilon_{t-1}^V)(\pi_t^D - \pi_{t-1}^D). \end{aligned}$$

In particular, if the residuals for the groups are correlated then the error is likely to be smaller.

Counts as outcomes. Note that the above is for outcomes which are means or proportions, such as proportion of cases that result in an FTA or average days in jail pre-disposition. If the outcome is counts, then we no longer have a weighted average but instead have to simply sum the outcomes. We advocate using the mean outcomes instead, potentially rescaling back up to counts at the end.

7.1 Fitting the model

The above shows that we, up to a small error in the residuals, can implement post-stratification by fitting our autoregressive model with our observed outcomes regressed onto the interacted covariates.

This means we have a lot of coefficients, which increases uncertainty. This may be worth it if it were believed the different groups had different structure. That being said, if it were believed the impact of some of the covariates were similar for the groups, those coefficients could be pooled. For example, the impact of seasonality S_t could be pooled giving

$$Y_t = \beta_0^D \pi_t^D + \beta_t^D \pi_t^D t + \beta_0^V \pi_t^V + \beta_t^V \pi_t^V t + S_t' \beta_S + \tilde{\epsilon}_t.$$

This immediately comes from, if we assume $\beta_S^D = \beta_S^V$,

$$\pi^D S_t' \beta_S^D + \pi^V S_t' \beta_S^V = S_t' \beta_S.$$

This model would then be transformed to a lagged outcome model as described in the main text. I.e., we will actually fit, if $\tilde{X}_t$ are the covariates not including 1 and t,

$$Y_t = (\pi_t^D X_t)' \beta^D + (\pi_t^V X_t)' \beta^V + (\pi_{t-1}^D \tilde{X}_{t-1})' \beta_\ell^D + (\pi_{t-1}^V \tilde{X}_{t-1})' \beta_\ell^V + \rho Y_{t-1} + \omega_t.$$

If we pool seasonality we get two linear trends and a common seasonality model, with a final regression of

$$Y_t = \beta_0^D \pi_t^D + \beta_1^D \pi_t^D t + \beta_0^V \pi_t^V + \beta_1^V \pi_t^V t + S_t' \beta^S + S_{t-1}' \beta_\ell^S + \rho Y_{t-1} + \omega_t.$$

7.2 Implementation

Once we have the ability to model outcomes for each of our subgroups (in the above, all these come from a single model), we can aggregate to get adjusted outcomes that rebalance the subgroups so they have constant relative weight across time. We first need to calculate the overall weights as we did with the reweighing method presented in the main text, and then we take a weighted average of the subgroup results.

In particular, our linear model is a function of time, the seasonality covariates, and the proportions of units in each subgroup. To get our adjusted series, we simply use the original time and seasonality covariates, but fix the proportions to π^*.

For example, to generate our simulated series, we would iteratively calculate

$$\hat{Y}_t = \beta_0^D \pi^{D*} + \beta_1^D \pi^{D*} t + \beta_0^V \pi^{V*} + \beta_1^V \pi^{V*} t + S_t' \hat{\beta}^S + S_{t-1}' \hat{\beta}_\ell^S + \hat{\rho} Y_{t-1} + \omega_t^*$$

where the ω_t^* are simulated as with the non-poststratified case. Everything else follows.

To estimate the final impacts and make the relevant plots we would, as usual, look at the distribution of simulated series, and compare to the (now adjusted) observed outcomes. Differences would suggest impacts after controlling for the variables forming the post-stratification groups.

Appendix C: R Package overview

We provide an R package, `simITS`, to implement the methods described in this paper. The documentation with the package is more comprehensive than this appendix, and includes a vignette walking through parts of the Mecklenberg and New Jersey analyses, but we provide a brief overview of the New Jersey analysis here as well.

To fit a seasonality model, first specify the functional form of the model:

```
my.model =  make.fit.season.model( ~ temperature + Q2 + Q3 + Q4 )
```

This creates a model that can then be fit to data; the named variables are all assumed to be in the dataset we will eventually analyze. The following code fits and displays our model to the pre-policy data only (`Y` is the outcome, stored as another column in our data), and does not include lagged covariates:

```
mod = my.model( dat = filter( newjersey, month <= 0 ), "Y", lagless = TRUE )
summary( mod )
```

In the above, `mod` is the result of simple linear regression, and the `summary()` call will print out the estimated coefficients and overall R^2.

To conduct the simulation, first add the lagged covariates needed (the package will automatically extract the needed covariates given a specified model) and then make the call to process the outcome of interest:

```
newjersey = add.lagged.covariates( newjersey,
                                   outcomename = "Y",
                                   covariates = my.model )
envelope = process.outcome.model( "Y", newjersey, t0 = 0, R = 1000,
                                  summarize = TRUE, smooth = FALSE,
                                  fit.model = my.model )
```

If there are multiple outcomes, simply change the outcome name in the call above. Loess smoothing can be done by changing the flag for `smooth` to `TRUE`.

We can make our plot as follows (this method uses the `ggplot` plotting environment):

```
plt <- make.envelope.graph( envelope, t0 = 0 )
plt
```

See the package documentation for further specifications and details of the resulting objects returned from these primary method calls.

References

Angrist, J. D. and J.-S. Pischke (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Bloom, H., J. A. Riccio, and N. Verma (2005, March). Promiting work in public housing: the effectiveness of Jobs-Plus. pp. 1–292.

Brodersen, K. H., F. Gallusser, J. Koehler, N. Remy, and S. L. Scott (2015). Inferring causal impact using bayesian structural time-series models. *Annals of Applied Statistics 9*, 247–274.

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell (2017). Stan: A probabilistic programming language. *Journal of statistical software 76*(1).

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

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

Ferman, B. and C. Pinto (2015). Inference in differences-in-differences with few treated groups and heteroskedasticity. *MIT Press*.

Gelman, A. and J. Hill (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gelman, A., X.-L. Meng, and H. Stern (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, 733–760.

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)*, 83–100.

Hallberg, K., R. Williams, A. Swanlund, and J. Eno (2018, April). Short Comparative Interrupted Time Series Using Aggregate School-Level Data in Education Research. *Educational Researcher 47*(5), 295–306.

Jacob, R., M.-A. Somers, P. Zhu, and H. Bloom (2016, June). The Validity of the Comparative Interrupted Time Series Design for Evaluating the Effect of School-Level Interventions. *Evaluation Review 40*(3), 167–198.

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

Redcross, C., B. Henderson, E. Valentine, and L. Miratrix (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., A. van der Vaart, and V. Ventura (2000). Asymptotic distribution of p values in composite null models. *Journal of the American Statistical Association 95*(452), 1143–1156.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applies statistician. *The Annals of Statistics*, 1151–1172.

Rubin, D. B. (2005, March). Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association 100*(469), 322–331.

Silber, J. H., P. R. Rosenbaum, R. N. Ross, J. M. Ludwig, W. Wang, B. A. Niknam, N. Mukherjee, P. A. Saynisch, O. Even-Shoshan, R. R. Kelz, and L. A. Fleisher (2014, March). Template Matching for Auditing Hospital Cost and Quality. *Health Services Research 49*(5), 1446–1474.

Somers, M.-A., P. Zhu, R. Jacob, P. E. Jacob, and H. Bloom (2013, September). The Validity and Precision of the Comparative Interrupted Time Series Design and the Difference-in-Difference Design in Educational Evaluation. pp. 1–152.

St. Clair, T., K. Hallberg, and T. D. Cook (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.

Stoffer, D. S. and R. H. Shumway (2006). Time series analysis and its applications: With r examples.

Zhang, F., A. K. Wagner, S. B. Soumerai, and D. Ross-Degnan (2009, February). Methods for estimating confidence intervals in interrupted time series analyses of health interventions. *Journal of Clinical Epidemiology 62*(2), 143–148.