This report was prepared by Abt Associates using federal funding provided by the Bureau of Justice Statistics.

Document Title:	The NCRP Data as a Research Platform: Evaluation Design Considerations
Authors:	William Rhodes, Abt Associates Gerald Gaes, Abt Associates Ryan Kling, Abt Associates Jeremy Luallen, Abt Associates Tom Rich, Abt Associates
Document No.:	252677
Publication Date:	June 2021
Award No.:	This project was supported by award numbers 2010-BJ-CX-K067 and 2015-R2-CX-K135

Abstract:

A hypothetical evaluation question posits that a state introduced a reform intended to reduce incarceration for a targeted group of offenders. This paper discusses how the Bureau of Justice Statistics' National Corrections Reporting Program (NCRP) data might be used to investigate what that reform accomplished. Once a state introduces a reform, an evaluator can observe what happened following that introduction, but the evaluator cannot tell what would have happened had the state not introduced that reform. This paper is a discussion of selected quasi-experimental approaches that should be useful for dealing with the above evaluation question: pretest-posttest designs, difference-in-difference designs, difference-in-difference-in-differences designs, and synthetic control methods. While not an exhaustive list of evaluation strategies, this paper aims to emphasize the analysis of panel data derived from the NCRP.

Disclaimer

The Bureau of Justice Statistics funded this third-party report. It is not a BJS report and does not release official government statistics. The report is released to help inform interested parties of the research or analysis contained within and to encourage discussion. BJS has performed a limited review of the report to ensure the general accuracy of information and adherence to confidentiality and disclosure standards. Any statistics included in this report are not official BJS statistics unless they have been previously published in a BJS report. Any analysis, conclusions, or opinions expressed herein are those of the authors and do not necessarily represent the views, opinions, or policies of the Bureau of Justice Statistics or the U.S. Department of Justice.

This page intentionally left blank.

The NCRP Data as a Research Platform: Evaluation Design Considerations

June 2021

William Rhodes
Gerald Gaes
Ryan Kling
Jeremy Luallen
Tom Rich

Abt Associates
10 Fawcett St.
Cambridge, MA 02138

Contents

A hypothetical evaluation question posits that a state introduced a reform intended to reduce incarceration for a targeted group of offenders. This paper discusses how the Bureau of Justice Statistics' National Corrections Reporting Program (NCRP) data might be used to investigate what that reform accomplished.

From the modern framework of potential outcomes (Imbens & Rubin, 2015), evaluation always poses a missing data problem. Once a state introduces a reform, an evaluator can observe what happened following that introduction, but the evaluator cannot tell what would have happened had the state not introduced that reform. The counterfactual is *missing data*.

The solution to the missing value problem is to compare the outcome following implementation of the intervention with a selected counterfactual that presumably approximates what would have happened absent the intervention. With qualifications (Berk, 2005), evaluators usually feel confident about counterfactuals that are based on random assignment (Orr, 1999), but random assignment is impractical for large-scale prison reforms. The alternative to random assignment is quasi-experiments that exploit naturally occurring variation in what is sometimes called *observational data*. Quasi-experimental designs are tricky because they raise validity and reliability challenges.

This paper is a discussion of selected quasi-experimental approaches that should be useful for dealing with the above evaluation question: pretest-posttest designs, difference-in-difference designs, difference-in-difference-in-differences designs, and synthetic control methods. This is not an exhaustive list of evaluation strategies, but we intend to emphasize the analysis of panel data (defined below) derived from the NCRP. After examining these different evaluation approaches, we conclude that each of the approaches has merit and that a thoughtful evaluation would exploit the advantages of each.

As the argument advances, some definitions will be helpful.

Effect A treatment effect or just *effect* is what the state actually accomplished because of the reform. It might be defined as the reduction in the number or percentage of the targeted population appearing in state prison relative to the size or percentage of that targeted population that would have appeared absent the intervention.

Estimate The above is definitional. The evaluator's problem is to *estimate* the size of that effect by identifying an appropriate counterfactual using procedures described in many books concerned with evaluation (Angrist & Pischke, 2009; Cameron & Trivedi, 2005; Lee, 2005; Morgan & Winship, 2015; Rosenbaum, 2009).

Validity If the counterfactual does not provide a good comparison, we say that the evaluation design poses a *validity challenge*, meaning that even in a very large sample, the estimated effect would not approximate the real effect (Manski, 2007).

Reliability Even if the counterfactual is valid, the amount of information provided by the data may be so meagre that the estimated treatment effect is measured with great imprecision. When the sampling variance for the estimated treatment effect is large, we say that the evaluation design has little power or inadequate *reliability*.

The question facing us is: How we can use the NCRP to estimate effects that are both valid and reliable? Comparing the pre-implementation period with the post-implementation period within the same state raises validity concerns because changes might have happened without the intervention. Contrasting states that did and did not implement the intervention raises other validity concerns because differences might have occurred for reasons other than the intervention. Furthermore, reliability is challenging when performing state comparisons, because given a maximum of 50 states, sample sizes are small.

Without pretense of being either comprehensive or final, this paper walks through evaluation design considerations specific to the NCRP. We illustrate use of those designs using NCRP data from two states: Arizona and California. However, this paper does not provide an evaluation of policy interventions in either state; we merely use these two to demonstrate how an evaluation might be conducted. Arizona is a convenient choice because, to our knowledge, there has been no major policy intervention within the state during the period covered by the NCRP data. Using Arizona data, we would expect that a demonstration evaluation would find no effect of an imagined policy intervention. California is a convenient choice because its prison Realignment initiative toward the end of the data assembly period had a widely acknowledged effect on state prisons. Using California data, we would expect that a demonstration evaluation would identify the effect from that known intervention.

This paper has five principal parts. The first discusses how the NCRP can be arranged into panel data; this arrangement is especially useful for both description and evaluation. The second part introduces some terms, describes some data transformations, and discusses statistical methodology exclusive of evaluation methodology. The third part describes patterns in prison admissions and prison populations in Arizona and California. This description is background for the discussion of evaluation methodology, the focus of this paper, which appears in part four. Part five offers some concluding remarks.

1. The NCRP as Panel Data

Sponsored by the Bureau of Justice Statistics, the NCRP was redesigned beginning in 2010 to assemble prison term records and post-confinement community supervision term records provided by state authorities (Luallen et al., 2014).[1] A prison term record begins when an offender enters prison and ends when he or she leaves. The same offender may have multiple terms. The records are updated yearly for each currently participating state and have been collected retrospectively for some states that had not previously reported. Defined similarly, the assembly of post-confinement community supervision (PCCS) records is a recent expansion of the NCRP; post-confinement records are not considered further in this paper although we could apply analogous evaluation tools to PCCS.

The NCRP is designed to capture all prison terms that were active sometime during a window period beginning in 2000 and ending (as of the time of this paper) in 2014.[2] However, reporting patterns and data quality vary by state. For many states, reporting is complete starting in 2000 and their data are deemed to be sufficiently reliable so that the NCRP team could assemble prison term records for all reporting years. For other states the NCRP team either did not assemble term records at all or assembled term records beginning at some year after 2000. Prison terms have also been assembled for Federal prisons (as part of BJS's Federal Justice Statistics Program), but those Federal records are not yet part of the NCRP. Because of jurisdictional differences, it seems doubtful that Federal records would be useful counterfactuals for evaluating state interventions.

When assembling descriptive statistics, and when explaining patterns in prison usage, assembling the NCRP term file into *panel data* is helpful. In this paper, panel data comprise a cross-section of time-series aggregates.[3] *Cross-sections* are defined as states or frequently as offense combinations within a single state. *Time-series* are months although other time-series units might be useful. *Aggregates* are sums of units (such as admissions and prison stocks) or averages (such as average time-served). As an illustration, picture measures of the number of admissions (the aggregate) for violent crimes, property crimes and drug law violations within Arizona (the cross-sections) for every month between 2003 and 2012 (the time-series).

The analysis in this paper begins by using NCRP data from 2003 through 2012, a period during which 26 states have prison term records. The analysis eventually reduces this observation window because it turns out that most of the interesting trends happen after

[1] The redesign was intended to increase state participation, improve data quality, and increase the data's utility for research. Previous users of the NCRP might note that the prison-term-based record arrangement replaced the earlier reliance on unlinked admission and release records (A and B records) and stocks as of December 31 of the reporting year (D records). The current NCRP allows stocks to be known for any date within the observation window.

[2] As of year-end 2018, the NCRP team had built terms records thru 2017 for most states.

[3] Panel data might be expressed as individual units (terms in our application), in which case the individual units are the cross-section. For some purposes, analyzing the NCRP data at the individual level may be insightful, but this paper is concerned with analyzing aggregate units so it adopts a narrow definition for panel data.

2003, and by starting the observation window later, we can include additional states in the analysis.

2. Defining Variables and Statistical Methodology

This paper discusses evaluation methodology but preliminary to that discussion we define terms whose meaning might otherwise be ambiguous. We also discuss the regression specification that enters into the evaluation methodology. We do not discuss evaluation design per se in this section.

2.1 Terminology and Data Transformations

Three terms appear repeatedly in the rest of this paper.

- Offense seriousness: Correctional interventions frequently are targeted on a specific type of offense or offender. For this paper, we presume the intervention targets offense types defined by seriousness, and below we explain how we determined seriousness.
- Admissions: Some interventions are best characterized as altering the rate at which offenders enter prison.
- Stocks: Other interventions are characterized as altering the prevalence of offenders in prison.

Admissions and stocks are examined on a per capita basis, which requires some data manipulation, discussed below.

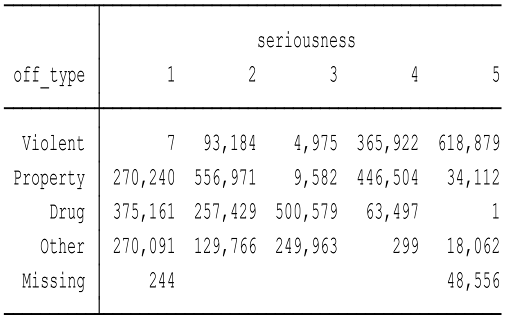
2.1.1 Offense Seriousness

Because correctional interventions often target offenses by seriousness, and because relative seriousness is not obvious from an offense name, this paper creates *offense seriousness* categories. Note that a useful definition of an offense category would depend on the intervention, so the seriousness categories used here are purely illustrative. For example, an intervention targeted on drug-law violators would dictate a different way of defining offense categories.

Each prison term in the NCRP dataset is associated with a BJS offense code (assigned to the variable BJS_Offense_1 in the NCRP). Using data from all states reporting to the NCRP since 2000, we computed the mean time-served by individuals released from prison by offense code. (When computing time-served, we excluded the records for offenders who served fewer than 90 days because this exclusion allows us to adjust (imperfectly) for time-served following a revocation for a technical violation.) Using average time-served, we placed every offender into a unique quintile ranging from least to most serious offenses, i.e. from least to most time-served, on average. The quantiles define five ordered seriousness categories.

Based on prison admissions, table 1 shows the distribution of seriousness categories cross-tabulated with traditional generic offense groupings—violent, property, drug, other, and missing. Given the remarkable dispersion of seriousness across offense types, we question how informative generic offense types are for classifying data, but that is a topic for another time. We will use these offense seriousness categories in this paper.

Table 1: Tabulation of Generic Offense Categories and Ascribed Seriousness Categories



	seriousness				
off_type	1	2	3	4	5
Violent	7	93,184	4,975	365,922	618,879
Property	270,240	556,971	9,582	446,504	34,112
Drug	375,161	257,429	500,579	63,497	1
Other	270,091	129,766	249,963	299	18,062
Missing	244				48,556

Note that it is possible to include (or exclude) offense types and still classify by seriousness. For example, just select violent offenses and compute seriousness categories within that grouping. We suspect that this approach may place unwarranted weight on states having similar reporting conventions, but that too should be a topic for research.

Classifying offenses by seriousness using time-served as an objective measure has some appeal for understanding prison populations and comparing populations across states. Classification is especially useful for evaluation because reforms often target a specific seriousness category (especially the least serious crimes) for an intervention, suggesting that a counterfactual comes from comparing the targeted population with the next less serious crimes (which should not be affected by the intervention). This need for counterfactuals highlights the need for careful consideration of seriousness categories. Three considerations seem important:

1. Many interventions identify the targeted category using a combination of offense type and offender criminal history. The NCRP does not yet include any measure of criminal history although it is possible to develop a proxy measure suitable for many analyses.[4] Because our concern is with demonstration, we have not attempted to apply this proxy in our analysis.

[4] The NCRP data begin for most states in 2000, so if the analysis begins in 2003, it is possible to distinguish offenders who were released from prison during a three-year window before their current admission from offenders who lacked a previous criminal history so measured. This is a crude but presumably effective way to distinguish offenders based on criminal history. This paper does not demonstrate this application.

2. Useful evaluation requires careful thought about offense classification. For example, if the state targeted offenders convicted of drunk driving, the counterfactual might be other crimes that result in sentences roughly equivalent to the sentences for drunk driving. We employ the seriousness categories for demonstration, not because they are necessarily the best way to create counterfactuals for all evaluation questions, but because we are interested in demonstration.

3. Both random assignment and quasi-experiments require the evaluator to justify the *stable unit treatment evaluation assumption* (SUTVA). In the present context, SUTVA means that the effect of the intervention does not spill over into the counterfactual comparison. For purposes of discussion, we will maintain SUTVA, but a proper evaluation would carefully select the comparison subjects to make SUTVA most plausible.[5]

2.1.2 Admissions

When assembling data, we discarded admissions when the term lasted for 90 days or fewer. This choice is arguable but it eliminates short periods for revocations. The choice is also problematic in that we cannot tell time-served for those who enter within 90 days of the final observed date so, for a few states, there is a slight bias upward for admissions during the last 90 days of the observation window. (That is, when no other information is available, we assume all terms with unobserved releases last 91 days or longer.) This bias will not be serious for this paper because most of the states have reported 2013 data, and given 2013 data, we know when time-served lasted more than 90 days for terms commencing in 2012.

2.1.3 Stocks

Our definition of stocks is just releases minus admissions for a given month. This is really the *change in stocks*, but given the beginning stock in 2000, it is easy to compute cumulative stocks from changes in stocks.[6] For econometric analysis, dealing with changes in stocks (essentially a first difference) has more desirable statistical properties than dealing with cumulative stocks.

[5] SUTVA is most credible when interventions are rule driven, which we expect to be the case with most prison reforms. Morgan and Winship (2015) provide a helpful discussion of SUTVA and how to deal with violations. For example, suppose an intervention targeted drunk drivers but some offenders convicted of public intoxication (rather than drunk driving) are incidentally considered comparable and are released. The evaluator might drop public intoxication from the comparison group and contrast drunk driving with other offenses of comparable seriousness. Thoughtful consideration can mitigate or eliminate the SUTVA problem.

[6] As noted earlier, the NCRP include all terms that were active sometime during the observation window. This implies that an investigator can always construct the stock population on any date during that window by a cumulative tabulation over time of admissions minus releases.

2.1.4 Data Problems and Adjustments

The NCRP data have been matched with other data sources (Census data, FBI data, etc.) that provide general population (age, arrests, etc.) statistics on a yearly basis. However, to capture interventions that may have occurred during the year, we analyze prison statistics on a monthly basis, which causes problems requiring adjustments.

Arrests

For example, consider prison admissions during January of 2005. If we hypothesize that prison admissions are a function of arrests, we might regress admissions on arrests for 2005. The logical problem is that while the admissions by construction occurred in January 2005, about 11 of every 12 arrests during 2005 occurred after January (and this ignores the delay from arrest to conviction to incarceration), so the regression is misspecified.

Our approach is:

1. When analyzing year Y admissions in January, we use the weighted average of 11/12 year Y-1 arrests and 1/12 year Y arrests.
2. When analyzing year Y admissions in February, we use the weighted average of 10/12 year Y-1 arrests and 2/12 year Y arrests.
3. We make this adjustment progressively for other months.

This approach makes some strong assumptions about the lags between arrests and admissions, and a refined analysis is required to develop an empirically justified distributed lag structure.[7] We have not done that for this discussion.

Population

For many purposes, it is instructive to examine admissions per capita or stocks per capita, but the issue is "what should we use as population?" The current NCRP data report state population for the year, and we adopt an adjustment similar to that used for arrests to distribute population over time. However, this begs the question: Who is counted in the at-risk population? We adopted an expedient approach of using the male and female population

[7] Our assumption is that arrests during the current month and arrests during the previous 11 months contribute equally to admissions/stocks during the current month. An alternative would be to lag the effect of arrests. For example, the previous 12 months (not including the current month) might account equally for admissions/stocks. Or the previous months might have unequal weights so that arrests from 6 months in the past have greater weight than 1 month and 12 months in the past. Possibly the lag structure should extend longer than 12 months. Different lag structures are testable using the data to identify best fit but we have not done that here.

between 14 and 34; although 14 is too young for prison admissions, we are constrained by Census-reported age categories.[8]

Scaling by population facilitates cross-state comparisons by accounting for population growth. However, scaling can distort raw trends. For example, prison population may increase on a raw basis yet decrease on a per capita basis. Depending on the research question, scaling might be inappropriate.

2.1.5 Scaling for Visualization

Another form of scaling is important for visualization. For some of the analysis, our approach is to standardize change in stock by subtracting the mean change and dividing by the standard deviation. Because it places statistics on a standard basis, this scaling facilitates drawing comparisons by cross-sections. The application of this scaling will be obvious from the context because statistics will be centered on zero and have a standard deviation of one.

2.2 Regression Specifications

Our analyses are always based on regressions even when the analysis is motivated by description. We do not want to get too deeply into the details (which receive additional coverage in context) but:

1. To capture short-term patterns in trends, we use Fourier transformations that account for year and half-year cycles. To capture long-term trends, we use polynomials. Specifically, Fourier transformations use trigonometric functions (sine and cosine) to capture cycles that repeat every year and half year.[9] We do not know why these cycles occur, but we suspect they are related to court cycles and delays between conviction and prison admissions. The cycles do not much interest us, but accounting for them reduces residual variance so we can better see what does interest us. When we use Fourier transformations, we first test for whether the year and half-year effects are jointly statistically significant at $p < 0.05$. If not, we drop them from the analysis; otherwise, we retain both the year and half-year effects.

2. Polynomials are useful for modeling long-term trends, the patterns that do interest us. Time is always rescaled to run from 0 to 1 by dividing the months by 120, the total number of months in the observation window. This rescaling helps with interpretation

[8] The approach is expedient because older offenders are at risk of entering prison. An alternative approach would be to weight the age groupings according to the age of offenders entering prison. We have not taken that step in this paper.

[9] Fourier transformations are sometimes uses to capture cyclical behavior because a Fourier transformation can capture any repeated pattern with an arbitrary degree of precision. Our application requires four terms—a sine and cosine function that repeats on a yearly basis and a sine and cosine function that repeats on a half-year basis. Hence the regression shows four terms f1 through f4.

and does not alter the regression results.[10] When we use polynomials, we always start with a cubic. A polynomial based on a cubic includes time, time-squared and time-cubed. When we use a polynomial, we first test whether the cubed term is significant at $p < 0.05$. If it is not significant, we drop the cubed term and test for the squared term, and if that is not significant, we then test for the linear trend. If it is not significant, there is no trend.

Other variables are incorporated into the regression. Seeking to demonstrate techniques, we have not attempted to be comprehensive. The arrest variables (and sometimes lagged releases) enter into some of the regressions. Typically we perform a joint test for statistical significance, and if the variables are not statistically significant, we drop them (at $p < 0.05$).

Dependent variables are scaled by dividing by population and sometimes additional units (such as division by 100) to provide interpretable pictures. Regression parameters are difficult to interpret and we suggest examining them qualitatively (for direction) but ignoring them quantitatively (for magnitude). Because of collinearity, even qualitative interpretations can be uninformative, so the reader might treat collinear variables as just "adjusting" for past arrests; collinearity will not affect the joint explanatory power of even perfectly collinear variables.

3. Descriptive Statistics

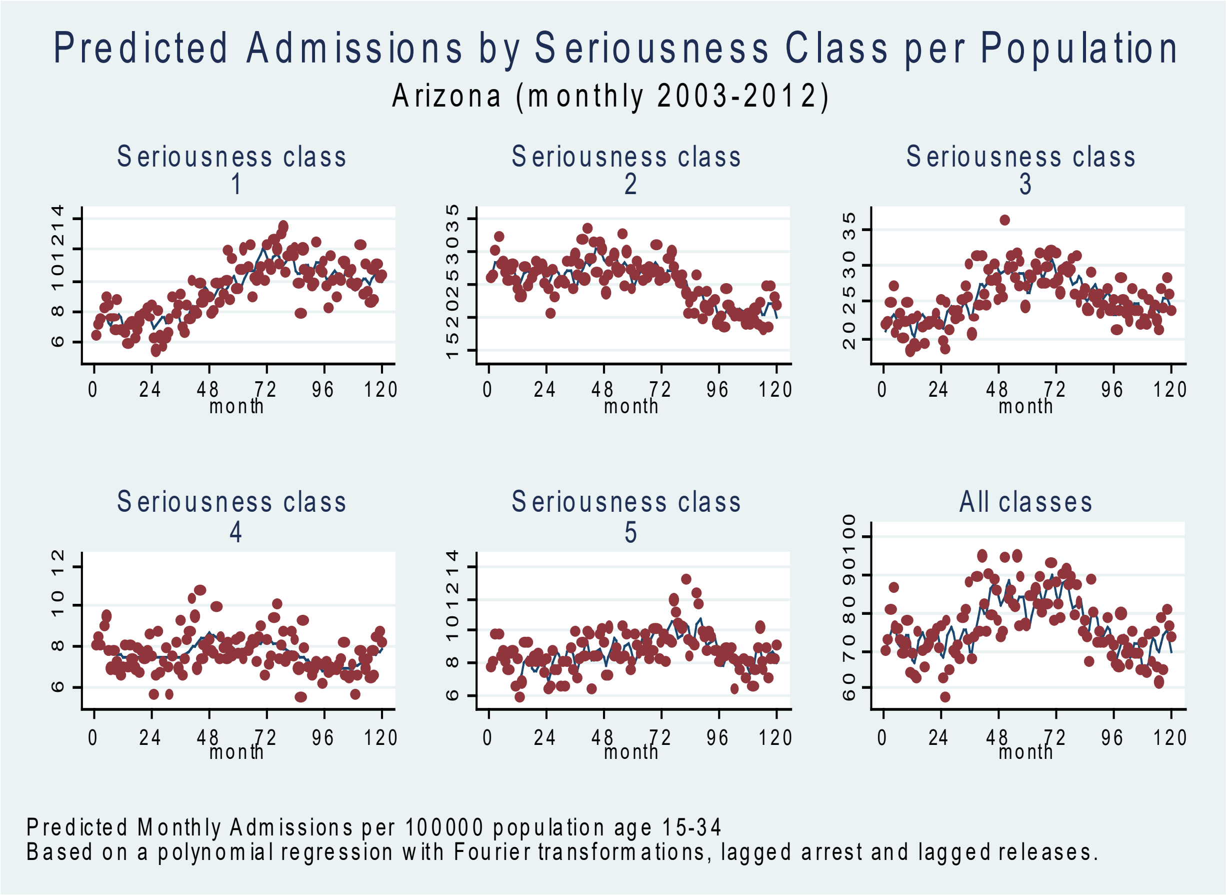
Descriptive analysis is a useful starting point. We show figures summarizing long-term trends in two states that have reported to the NCRP since 2003. The purpose of presenting descriptive trends is simply to illustrate considerable fluctuation in stocks (prison stock) and flows (admissions) over short periods of time. These fluctuations complicate evaluation because, when short-term changes occur naturally, interrupted time-series are unreliable for forming counterfactuals. Given this limited purpose, we only show trends for Arizona and California, two states that are the focus when this paper turns from description to illustrating approaches to evaluation.

Figure 1 shows Arizona admissions per 100,000 residents between 16 and 34, in total and broken down by offense seriousness category. The figure has six panels corresponding to the five seriousness classes and all classes rolled together. The dots are actual data. Table 2 shows regression results. If the cycles were insignificant, then the curve would be smooth. Therefore, by just looking at the figure, we can tell that the Fourier transformations are statistically significant except for seriousness class 4. The cycles might exist for seriousness class 4, but power is insufficient to detect the pattern. Regardless, unless the line is flat (perhaps with cyclical perturbations), we can tell whether the polynomial is statistically

[10] The regressions used here are invariant (except for scale effects) to linear transformations. Although a polynomial may seem nonlinear, it is actually linear in its arguments, which is sufficient for the invariance properties to hold.

significant. A sharp eye can even tell which degree of the polynomial is statistically significant. There are strong seasonal and long-term trends in Arizona.

Figure 1: Trends in Prison Admissions per Capita in Arizona



Predicted Monthly Admissions per 100000 population age 15-34
Based on a polynomial regression with Fourier transformations, lagged arrest and lagged releases.

Because the figures are adequately descriptive, the regression parameters (table 2) are relatively uninteresting. The polynomials are captured by the **T, Tsq** and **Tq** terms. The Fourier transformations are captured by the **f1** through **f4** terms. If parameters appear in the table, then the polynomial/cycles are statistically significant at $p < 0.05$.[11] That is, the table indicates the degree of the polynomial used to estimate the regression and whether the Fourier transformations entered the regression. The table shows that past arrests are important for explaining admissions; the arrest variables would not appear in the table if they were not jointly significant. Lagged releases are typically not statistically significant. Except for seriousness class 4 admissions, the R^2 gives an impression of substantial change in admissions per capita over time. This is a context where R^2 tells us little. If there are no cyclical patterns and no trend, then the R^2 would be near zero. An R^2 of zero does not mean

[11] The table also shows which specific parameters are statistically significant, but the significance of individual parameters should be of little interest. Joint tests are most interesting but not shown in the table.

that we have explained nothing; on the contrary, we have explained much—namely, there is no discernable trend.

Table 2: Arizona Polynomial Regression

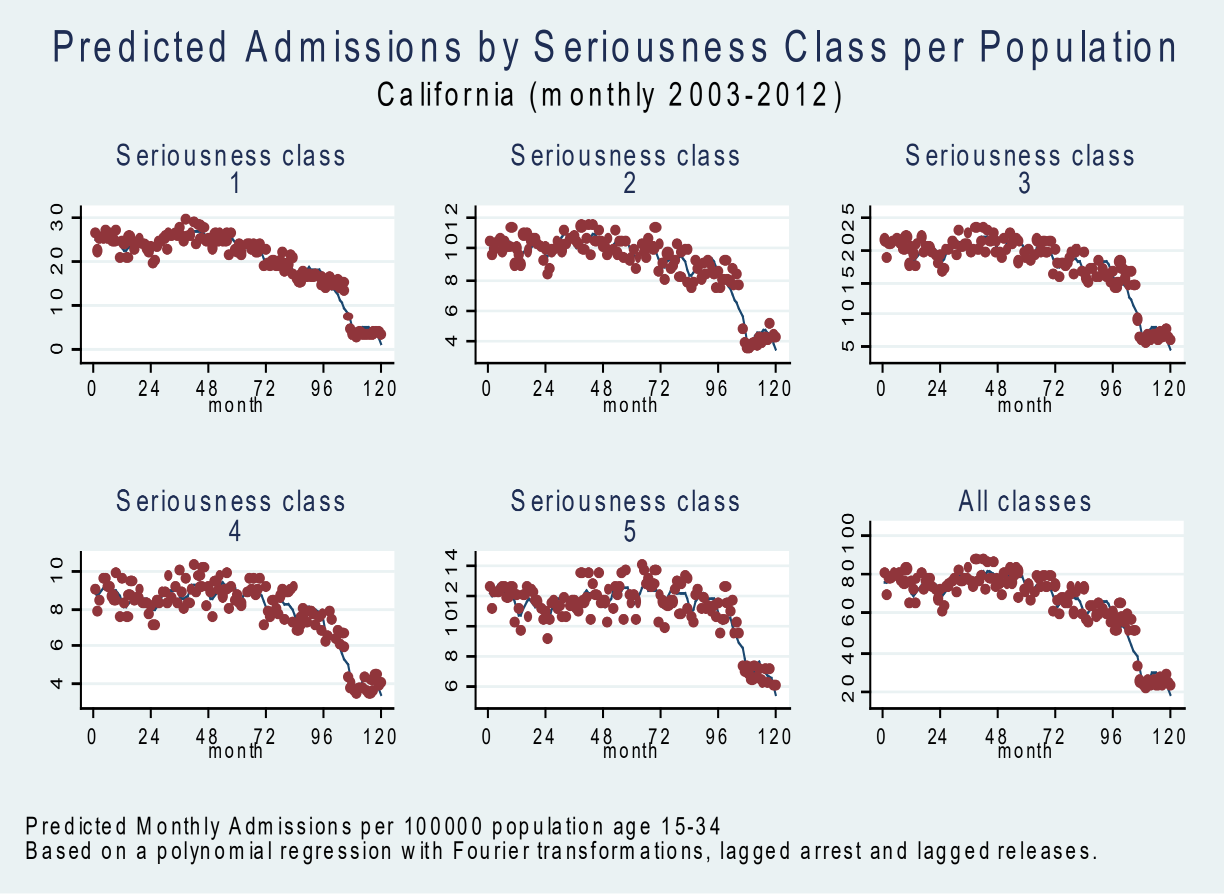
	admissions1	admissions2	admissions3	admissions4	admissions5	total
T	22.361**	20.800**	43.494**	5.807**	13.074**	107.886**
Tsq	-21.040**	-27.090**	-43.617**	-5.653**	-11.562**	-111.826**
viol	2,028.386	12,336.912**	11,082.520**	3,967.658**	2,264.606*	32,503.767**
prop	123.841	-1,393.047**	-1,340.379**	-397.512**	166.634	-2,931.792**
drug	-1,002.431**	-1,058.146**	-1,651.783**	-320.935*	-100.615	-4,629.464**
f1	-0.129	-0.605*	-0.403		-0.218	-1.373
f2	-0.207	0.061	0.063		0.101	0.263
f3	-0.367**	-0.873**	-1.059**		-0.543**	-3.105**
f4	-0.023	0.398	0.267		-0.096	0.507
lagged_releases5					0.054**	
_cons	11.830	12.824	25.060	1.406	-14.172	50.566
R2	0.71	0.70	0.63	0.23	0.47	0.58
N	120	120	120	120	108	120

* p<0.05; ** p<0.01

For present purposes, the story behind the trends in Arizona is simple. There are short-term fluctuations and long-term reversals in trends. If we attempted to evaluate a policy intervention in Arizona, these short-term fluctuations and long-term shifts would raise validity concerns. We return to this point later.

Polynomials can give distorted impressions when admission practices suddenly shift. California (Figure 2) illustrates this. California had been experiencing a decrease in prison population per capita before it changed its admission practices (called Realignment) to make greater use of county jails. The polynomial suggests a downward trend that really has abated by the last year of data, but the polynomial does not show that subsequent abatement. (A higher degree polynomial might be helpful, but probably a spline recognizing the known break in California admissions would be more helpful.) Notice the high R^2; these occur because of the precipitous drop in admissions, not because the regressions really explain better in California than in Arizona.

Figure 2: Trends in Prison Admissions per Capita in California



Predicted Monthly Admissions per 100000 population age 15-34
Based on a polynomial regression with Fourier transformations, lagged arrest and lagged releases.

Table 3: California Polynomial Regression

	admissions1	admissions2	admissions3	admissions4	admissions5	total
T	-58.610**	-28.160**	-59.884**	-19.684**	-31.378**	-197.717**
Tsq	130.626**	66.244**	141.910**	48.809**	81.263**	468.851**
Tq	-88.539**	-39.915**	-90.954**	-29.415**	-52.312**	-301.135**
viol	59.711	-374.048	-469.230	604.787	259.377	80.597
prop	-1,529.135*	289.447	-297.389	-104.952	-13.903	-1,655.931
drug	1,832.610**	755.396**	1,473.717**	567.860**	571.489**	5,201.072**
f1	-0.997**	-0.465**	-1.042**	-0.391**	-0.526**	-3.421**
f2	-0.153	-0.087	0.005	-0.053	0.035	-0.254
f3	-0.324	-0.149	-0.337	-0.115	-0.287*	-1.211
f4	-0.045	0.062	-0.070	-0.102	-0.100	-0.254
_cons	5.349	-8.119	-3.623	-10.672	-4.225	-21.290
R2	0.95	0.89	0.91	0.88	0.81	0.92
N	120	120	120	120	120	120

* p<0.05; ** p<0.01

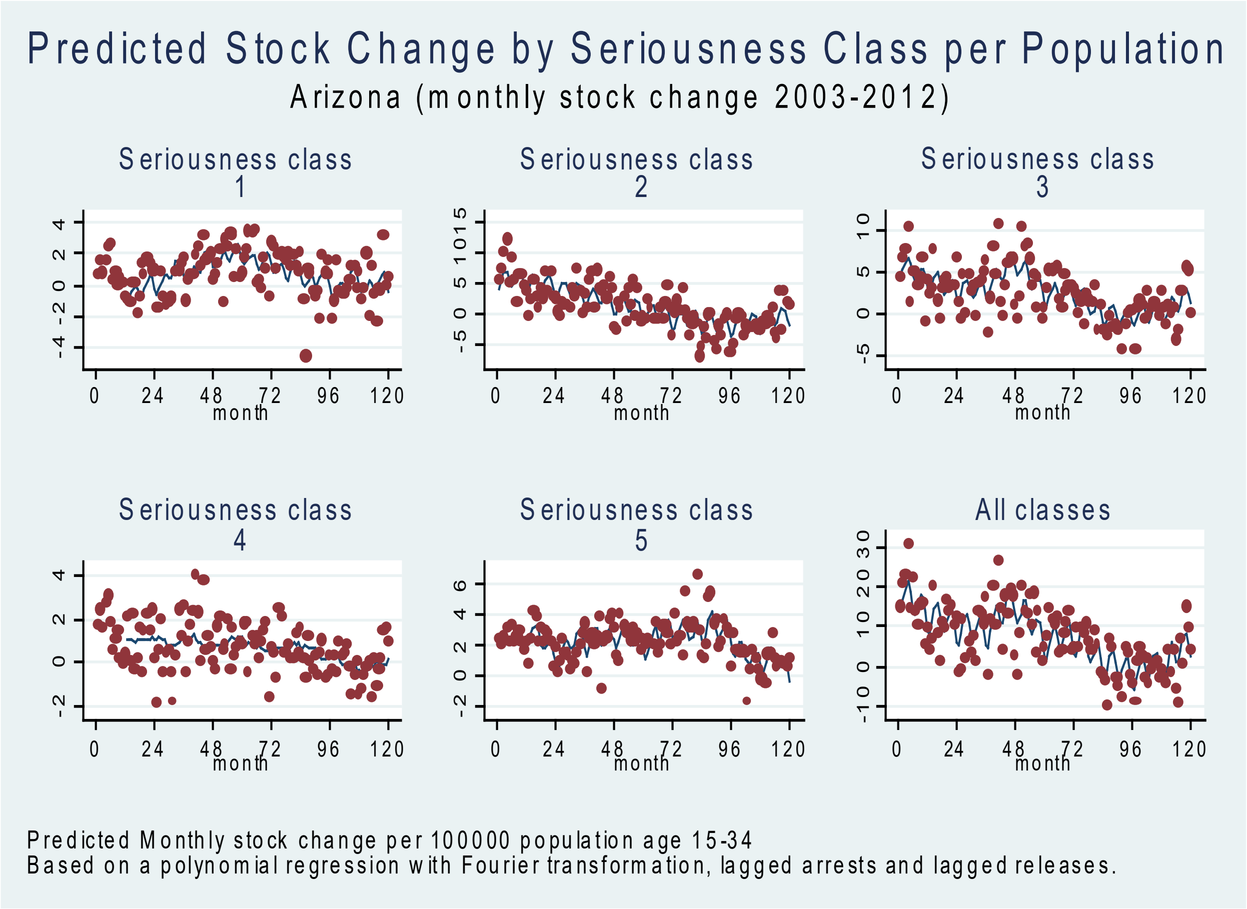
California offers a useful contrast to Arizona. In Arizona, the figure shows short-term fluctuations and long-term reversals in trends; by assumption, made for purposes of this discussion, neither could be attributed to a statewide intervention. If we had attempted to evaluate an intervention, these naturally occurring changes would raise validity issues. In

California, we know that the state correctional system underwent a profound policy change, shifting offenders from state prisons to county jails. Interestingly, a cynical evaluator could point out that the post-intervention trends appear to be an extension of pre-existing trends. Descriptive statistics provide an inadequate platform for evaluation.

Additional descriptive analysis comes from examining the monthly change in stocks beginning in the first month (i.e., January 2003) and ending in December 2012. Monthly change—the first difference of the cumulative change—is more useful for understanding trends because it more clearly relates changes to covariates. That is, if we wanted to analyze changes in stocks, serial correlation would be severe, so we would take first differences to reduce the serial correlation. That step is taken here.

Figure 3 shows actual data (the dots) and predictions (the lines) for Arizona. The change in the stock of prisoners is the difference between admissions and releases in each month, so in theory this new figure might tell us something different than did its admissions counterpart, but in fact the story does not much change. As before, we see fluctuations in the change in the stock, cycles and long-term shifts in trends. Basing an evaluation on an interrupted time-series would be tenuous.

Figure 3: Trends in per Capita Changes in Stocks in Arizona



Predicted Monthly stock change per 100000 population age 15-34
Based on a polynomial regression with Fourier transformation, lagged arrests and lagged releases.

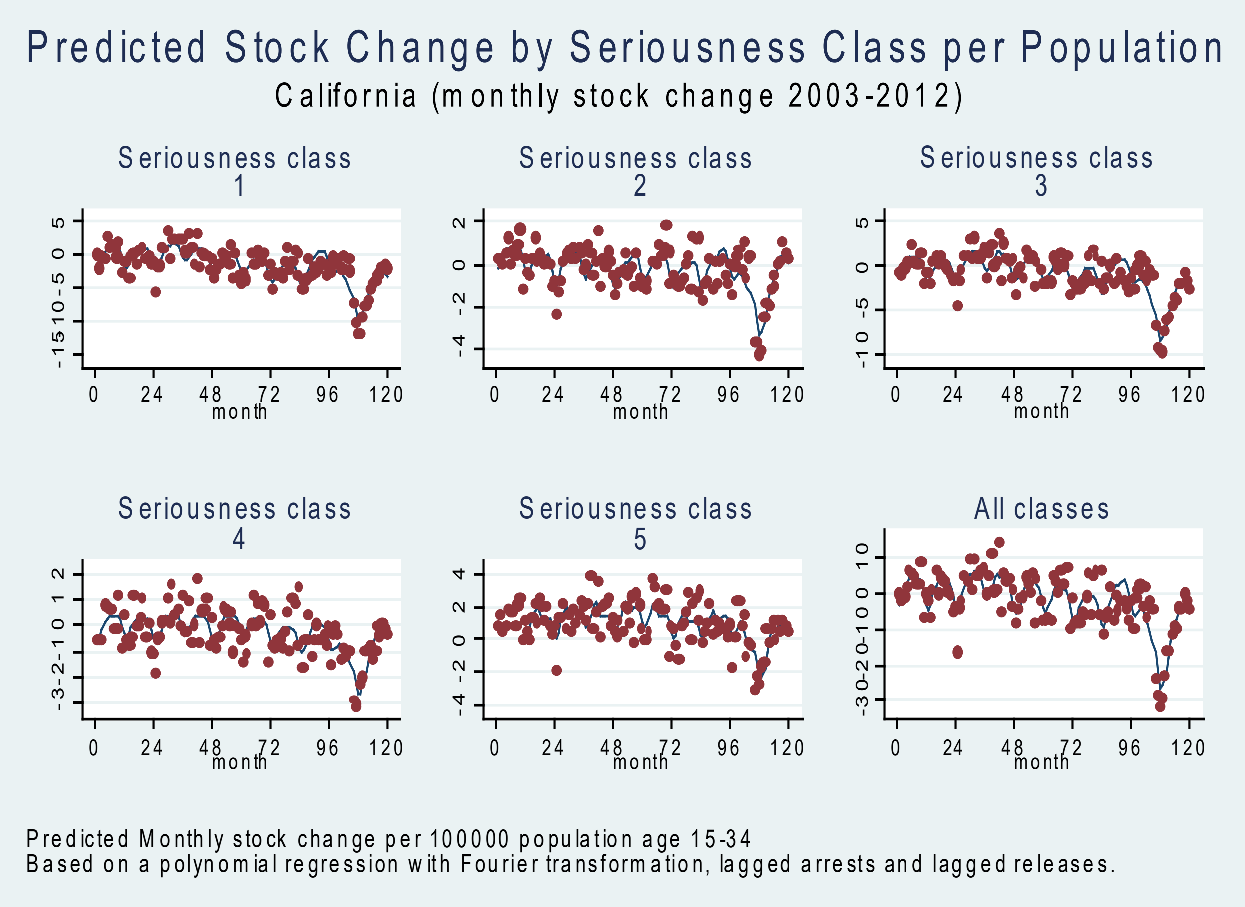
Table 4: Regression Results for Trends in per Capita Changes in Stocks in Arizona

	stock1	stock2	stock3	stock4	stock5	total
T	14.193**	-1.579	-6.990**	-3.136**	-18.081**	-16.063**
Tsq	-9.446**	-31.970			35.856**	
lagged_releases1	-0.037**					
f1	-0.119	-0.558	-0.106		-0.232	-1.033
f2	-0.317	0.072	0.124		-0.006	0.243
f3	-0.474**	-1.039**	-1.161**		-0.680**	-3.542**
f4	0.002	0.961**	0.130		-0.055	1.072
Tq		27.635*			-25.672**	
viol			6,570.743**			21,662.546**
prop			-1,421.503**			-3,479.537**
drug			-547.111*			-1,446.532**
lagged_releases4				0.047**		
lagged_releases5					0.089**	
_cons	1.856	5.772**	16.311	-3.164*	-3.442*	12.771
R2	0.31	0.60	0.45	0.16	0.46	0.60
N	108	120	120	108	108	120

* p<0.05; ** p<0.01

Figure 4 is the counterpart to figure 2 for California. Although the story might have been different from that told by admissions in California, in fact the story is quite similar. We see fluctuations and cycles but no large interruptions in the trend except for the drastic drop in stocks following California's policy intervention. Late in the period, the change in stocks has hovered around zero, much as it had during the years prior to the intervention. Even the cynical evaluator, identified earlier, might find this abrupt change immediately after the intervention compelling; still, it would be helpful to have a formal test. This concern brings us to the transition between descriptive statistics and inferential statistics used for evaluation, the topic of the next section.

Figure 4: Trends in per Capita Stocks in California



Predicted Monthly stock change per 100000 population age 15-34
Based on a polynomial regression with Fourier transformation, lagged arrests and lagged releases.

Table 5: Regression Results for Trends in Per Capita Stocks in California

	stock1	stock2	stock3	stock4	stock5	total
T	-71.510**	-25.467**	-52.572**	-9.449**	-14.066**	-172.254**
Tsq	144.079**	52.120**	110.051**	18.662**	26.834**	352.982**
Tq	-51.795**	-13.774*	-38.823**			-108.210**
viol	-114.764	1,265.110**	870.502	1,230.239**	1,413.594*	4,728.302
prop	1,939.548**	728.416**	1,211.545*	328.098	295.366	4,514.428*
drug	2,439.854**	927.589**	1,972.270**	582.595**	919.159**	6,672.004**
f1	-1.474**	-0.573**	-1.315**	-0.444**	-0.560**	-4.367**
f2	-0.084	-0.048	0.203	0.002	0.179	0.238
f3	-0.381	-0.266*	-0.280	-0.142	-0.384**	-1.419*
f4	0.044	0.064	0.113	0.009	-0.087	0.168
lagged_releases5					0.004	
_cons	-90.733**	-50.439**	-79.699**	-35.503**	-48.842**	-297.535**
R2	0.69	0.51	0.67	0.49	0.45	0.65
N	120	120	120	120	108	120

* p<0.05; ** p<0.01

4. Evaluation

The rudiments of evaluation appear in the discussion above (that is, our eyes can detect patterns), but formal designs are required to meet validity and reliability challenges. We discuss four evaluation designs: interrupted time-series; difference-in-differences; difference-in-difference-in-differences; and synthetic control methods. Throughout this discussion, the motivational illustration is that a state decides to reduce its prison population for the least serious offenders. This policy shift occurs at a defined point in time, although we might assume that the intervention takes time to reach full implementation so the full effect is lagged.

Let:

S_{ijk} This is the stock of offenders from seriousness category *i* at time *j* in state *k*.

s_{ijk} This is the change in the stock from seriousness category *i* at time *j* in state *k*.

$$s_{ijk} = S_{ijk} - S_{i(j-1)k}$$

M_j This is the month, typically parameterized to run from 0 to 1 by dividing months by the number of months in the observation window as described above. When drawing figures, to assist the reader, we revert to using the months rather than transformed version of months.

These are all variables that we used above when presenting descriptive statistics.

The discussion of *design* in the remainder of this section is progressive. That is, the interrupted time-series is the least useful and the synthetic estimation is arguably the most useful, but they actually have much in common, so value comes from building more sophisticated approaches onto the less sophisticated approaches. As the term is used here, an approach is more sophisticated if it raises fewer validity concerns.

Although we derived the descriptive statistics from 2003–2012, based on the descriptive statistics we doubt that such a long time-series is useful for evaluation because perturbations and reversals in trend that occur early in the time-series are likely uninformative about interventions that occur later in the time-series. Consequently, in the following demonstration, we will abbreviate the time-series. This has the additional advantage of allowing us to expand the number of states under study.

As a road map of the following subsections, for Arizona we imagine an intervention that happened exactly two years before the end of the NCRP time-series. In fact, there was no intervention on that date, so we would not expect to observe an effect. We then discuss using an interrupted time-series (section 4.1), a difference-in-differences design (section 4.2) and a difference-in-difference-in-differences design (section 4.3) to “evaluate” this imagined intervention. The point is that the least rigorous design can lead to spurious conclusions and the more sophisticated designs are more believable. For California, a real intervention occurred toward the end of the time-series, purposefully substituting confinement in county jails for confinement is state prisons. We use the synthetic case control method to detect the consequences of that policy change (section 4.4).

4.1 Interrupted Time-Series

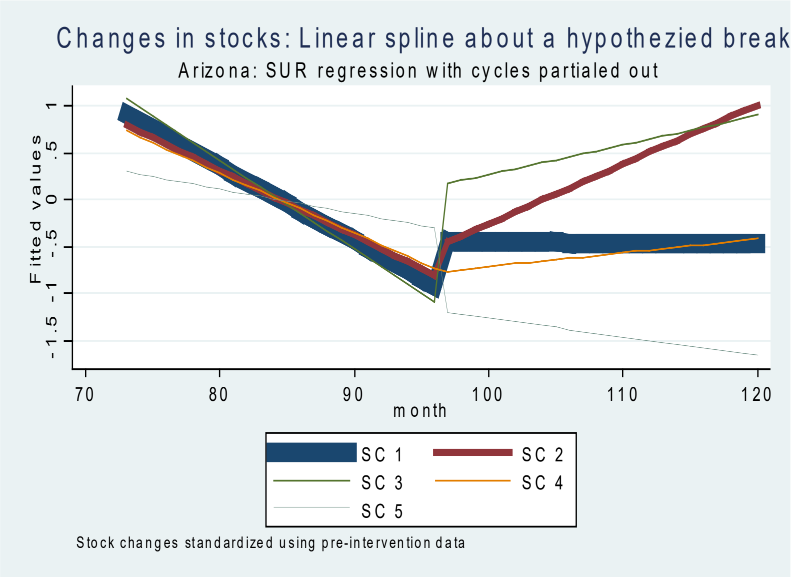
For Arizona we hypothesize a break in a trend on January 1, 2011, only for the least serious offenses, which are assumed to be the target of the intervention. An approach to an interruptive time-series is to assume that trends are linear or nearly linear immediately to the left and immediately to the right of the break.[12] The "treatment effect" is the shift in the regression lines at the intervention point. The typical application selects a *bandwidth* (of time) that is clustered about the intervention point. Without more discussion, we limit the analysis to two years before the intervention and two years after the intervention. Given yearly cycles, bandwidths should always be specified as years. In practice we would test alternative bandwidths, but this testing is not important for this demonstration.

We have standardized the stock by subtracting the mean change and dividing by the standard deviation for the pre-intervention period. Without standardization, difference-in-differences and difference-in-difference-in-differences comparisons are difficult to discern. With standardization, statistics are centered near zero and have a standard deviation near one regardless of the original scale.

Using the Arizona data, we fit a linear model in time to the left and a linear model in time to the right of the intervention point. This model also includes cycles, and they are very important over this short interval, but we will not show them because they dominate the picture. See figure 5. It shows the predictions, based on the linear model after removing (partialing out) the cycles, for all five offense seriousness categories (SC 1 through SC 5), but current attention is just on the first offense seriousness category (SC 1).

[12] Although the point is arguable, some evaluators treat an interrupted time-series as being a regression discontinuity design (Imbens & Lemieux, 2007). From the RDD perspective, the estimated treatment effect is most valid when it is estimated immediately about the break point using local linear regressions. The RDD—and hence the interrupted time-series—has less appeal when the impact of an intervention materializes over a lengthy period, one of the points made in this paper. Within a criminal justice context, some of these issues are discussed in Rhodes and Jalbert (2013).

Figure 5: An Interrupted Time-Series for Changes in Per Capita Prison Stocks in Arizona



Focusing our attention on seriousness class 1, the visual impression is that the stock increased at the time of the imaginary intervention (i.e. after 24 months) and that the previously decreasing trend reversed its course. Because there was no actual intervention, we expected to see a continuation of the pre-24-month trend. In fact, the jump after 24 months is not statistically significant, but the reversal in the trend is highly significant ($p = 0.02$); based on these results alone, we would falsely conclude that our imaginary intervention changed the trend. In fact, looking across all five seriousness classes, the jump is significant ($p = 0.001$) for one class and the reversal in trends is significant at $p < 0.05$ for two seriousness classes and insignificant at $p < 0.06$ for a third. These results illustrate that resting evaluation on an interrupted time-series is treacherous and raises validity concerns, causing us to recommend against using an interrupted time-series to evaluate policy interventions intended to affect populations in state prisons.[13]

4.2 Difference-in-Differences

A problem with the interrupted time-series is that the post-intervention period may differ from the pre-intervention period for reasons that have nothing to do with the intervention. One way to strengthen the inference about treatment effectiveness is to presume that, absent an effective intervention, whatever changes occur during the post-intervention period would

[13] An evaluator might choose to use a polynomial instead of local linear regressions, but this approach raises validity issues. When the regression is nonlinear in the vicinity of the break point, distinguishing between naturally occurring nonlinearity and nonlinearity induced by a true intervention is tenuous. We concede that other evaluators may prefer a nonlinear regression nevertheless, and rather than argue the point, we just emphasize that an interrupted time-series raises difficult problems of interpretation.

affect both seriousness class 1 and seriousness class 2 crimes in approximately the same way. This implies that we should compare the difference in trends for seriousness class 1 and seriousness class 2 crimes and only reject the null of no treatment effect when the break/trend for seriousness class 1 crimes differs from the break/trend for seriousness class 2 crimes. A similar logic might be employed to contrast seriousness class 1 and seriousness class 3 crimes. This type of comparison is an application of a difference-in-differences (DD) design. Note that this approach depends critically on being correct about SUTVA so in a real application evaluators would be especially careful about choosing the cross sections.[14]

There is a trick to deriving the standard error for the test statistic because the time-series are not independent. We have used a linear seemingly unrelated regression (SUR) to estimate covariances. Variances are unaffected because, for each seriousness class, the right-hand-side variables are the same.

We compare the break in the time-series for seriousness class 1 with the break in the time-series for seriousness class 2 and find no statistically significant difference. We compare the break for SC 1 with the break for SC 3 and again find no significant break. We compare the break for SC 1 with the average of SC 2 and SC 3 and again find no statistically significant difference. Using these same contrasts for the post-intervention trends, we find no statistically significant differences. The DD design provides more satisfying results both because we fail to reject the null (which we know is correct) and because the logic of a DD is more compelling than the logic of an interrupted time-series.

4.2.1 An Alternative Approach

Although the DD framework specified above is familiar, an alternative that uses essentially the same identification strategy may be better. The alternative uses a ratio:

$$r_{ijk} = \frac{S_{ijk}}{\sum_{i \neq 1} S_{ijk}}$$

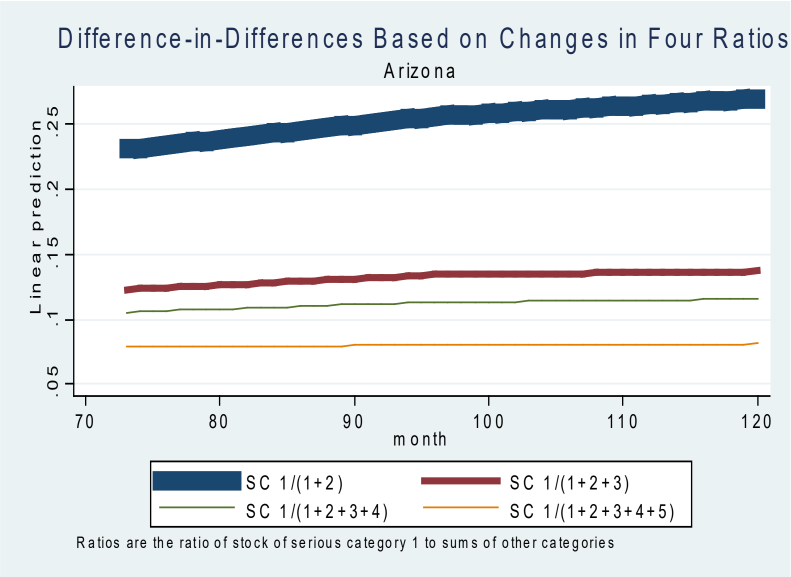
The numerator is stock for the seriousness class that is targeted by the intervention. The denominator is some combination of seriousness classes that are not targeted for the intervention. As before, the denominator should probably be restricted to seriousness classes that are similar to the seriousness class of interest.

We can substitute the ratio into the same regression framework used earlier for the interrupted time-series. Because we have not taken a first difference, autocorrelation is a problem, and consequently we have introduced a Prais-Winston transformation to adjust the

[14] Often interventions are rule driven, such as: release drug-law violators convicted of minor trafficking offenses. A suitable comparison group would be offenders convicted of low-level property crimes or minor assaults. The most desirable comparison group depends on the context so our choice of seriousness classes is only for illustration.

regression for autocorrelation. Figure 6 shows four ratios, over an abbreviated observation window, for Arizona. The highest curve shows the ratio of class 1 seriousness offenders to the sum of class 1 and class 2 seriousness offenders. The lowest curve shows the ratio of class 1 seriousness offenders to the sum of all offenders. Visual inspection of the figure suggests no strong sharp breaks at 96 months. The evidence is less compelling regarding trends, and in fact, the trends are statistically different during the hypothetical intervention period (when no intervention in fact occurred) for two of the four contrasts. However, there is no statistically significant change in the trends for SC 1/(SC 1 + SC 2) or for SC1/(SC1 + SC2 + SC 3); these are the contrast that seem most justified because SC 4 and SC 5 crimes are very different than SC 1 crimes. Even if we decide to place some emphasis on ratios that are statistically significant, we note that the size of the effect is not substantively large, so effects might be statistically significant but not substantively important. Given that the more proximate seriousness classes are the most informative for SC 1, we put more faith in the comparisons for these first two ratios.

Figure 6: An Application of a Difference-in-Differences Estimator for Changes in Prison Stocks in Arizona



The difference-in-differences approach does not eliminate validity challenges and a rigorous evaluation might more carefully construct and examine the contrasts. The simple point made in the above two figures is that a difference-in-difference design greatly reduces validity challenges that arose in the interrupted time-series approach. A reader might think of the difference-in-difference approach as starting with an interrupted time-series (which is an obvious element of the DD) and improving the credibility of the inference.

4.2.2 Multi-State Considerations

Noteworthy, the DD estimates in the above two sections are *state-specific*. For some evaluations, several states have implemented the interventions at about the same time, and we might be interested in comparing effects across states or in combining effects to get an average effect. Because the estimate from one state is independent of the estimate from another state, we can combine estimates using an approach generically known as meta-analysis. There are many sources explaining meta-analysis (Borenstein et al., 2009) but the simple analytics, suitable for present purposes, appear in accessible sources (Borenstein et al., 2010; Rhodes, 2012).

Basically if we derive treatment effects for N states, then we can average across those N states to derive a composite treatment effect. We will not discuss the details, but a meta-analysis approach leads to an average (or, as appropriate, a weighted average) with standard error that depend on (among other things) whether the chosen estimator is a fixed-effect or a random-effect estimator.

When variations on a general policy initiative are implemented across many states, the meta-analysis might attempt to explain the size of treatment effects using specific program components as explanatory variables. Technically, this approach is a straightforward regression problem (Rhodes, 2012). Practically, however, inferences are limited by the number of states that have adopted the initiative and the diversity of initiative practices across those states. Rather than identifying an average treatment effect (a fixed-effect model), it may be useful to identify the variance in treatment effects across settings (a random-effect model) even it that variance cannot be explained due to insufficient data.

The difference-in-differences designed matched with meta-analysis appears to be a strong design for evaluating correctional interventions that are targeted on a specific group of offenders. We recommend combining DD and meta-analysis as the basic approach to dealing with state-level correctional interventions. However, this paper discusses an additional approach—a difference-in-difference-in-differences design—that may be suitable in some circumstances.

4.3 Difference-in-Difference-in-Differences

Although the DD framework appears to provide a useful basis for evaluation, we might strengthen that inferential framework using a difference-in-difference-in-differences design, hereafter DDD. The logic is that we compare the changes in slope in the state that implemented the intervention (Arizona, here) with the changes in slope for states that did not implement the intervention.[15] This is a DDD design because the first difference is within state (using the ratio approach) and the second difference is across state.

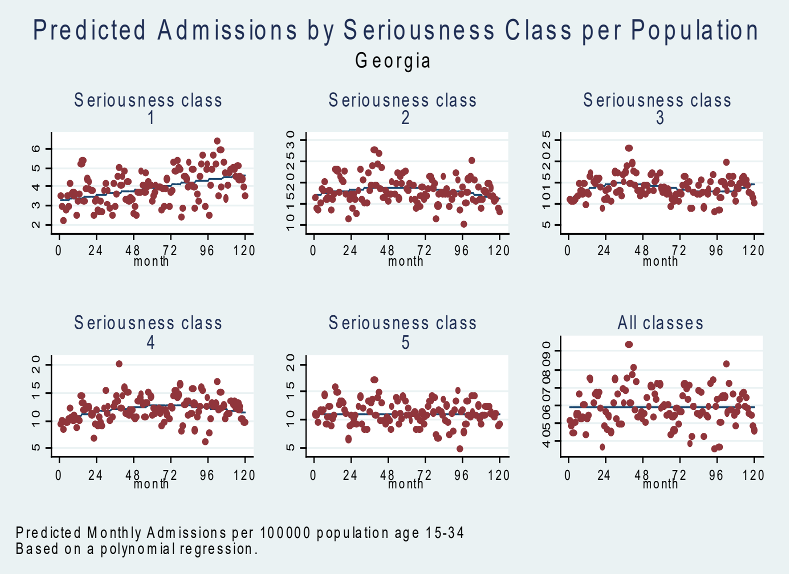
[15] This is one way of testing the SUTVA. Returning to the earlier example, suppose the intervention affects drunk driving but that offenders convicted of public intoxication might be treated similarly, perhaps because they were actually charged with drunk driving but entered a plea to public intoxication. We would expect to see a different trend for drunken driving in the state that implemented the intervention than in the state that did not implement the intervention. If there was no spillover, we would not expect to see the same contrast for public intoxication.

There are a number of ways to specify a model, but they all have a flaw: What other states should be used in the comparison? This question used to receive little attention in econometrics. Recently it has been receiving widespread attention (Abadie et al., 2010, 2015; Imbens & Wooldridge, 2009; Wooldridge, 2007).

The basic problem is that statistical testing assumes that the state or states used in the DDD comparison are in fact appropriate comparisons, so that measurement error comes exclusively from time-series fluctuations. In fact, if there are differences across states that are not taken into account by matching states, then an additional level of uncertainty—that attributable to selecting the comparison states—is incorrectly ignored by the analysis.

Rather than performing a mock evaluation with cross-state comparisons, we use descriptive statistics from Georgia to illustrate the potential danger of a cross-state comparison. Compare figure 7 (Georgia) with figure 1 (Arizona). The presumption is that neither state had implemented major interventions to alter prison admissions. The logic of a difference-in-difference (or difference-in-difference-in-differences) methodology is that, for Georgia to be a useful counterfactual, both states should show comparable trends absent any interventions, but clearly the comparison shows that equivalency to be erroneous. Any comparison of trends in Arizona and Georgia would fail to capture the uncertainly of selecting a state (Georgia) for purposes of comparison (with Arizona). Simply put, Georgia is a poor counterfactual but how would an evaluator know that?

Figure 7: Trends in Prison Admissions per Capita in Georgia



Predicted Monthly Admissions per 100000 population age 15-34
Based on a polynomial regression.

A DDD methodology has the potential to improve the validity of inferences otherwise based on a DD methodology, but not necessarily if we lack a principled basis for selecting one or more comparison states. We turn to that issue next. We caution that the synthetic control methodology is emerging in the evaluation literature.

4.4 Synthetic Control Methodology

We illustrate using California because we know California implemented a major reform (Realignment) to reduce its prison population, and we might think that this reform would change the ratio of seriousness class 1 offenders to the sum of seriousness class 1 and 2 offenders. Note that California did not target its intervention to emphasize seriousness class 1 offenders over seriousness class 2 offenders, although this seems like an interesting research question, and it nicely illustrates the synthetic control methodology (Abadie et al., 2010, 2015). We do not claim, however, that this is a serious evaluation.

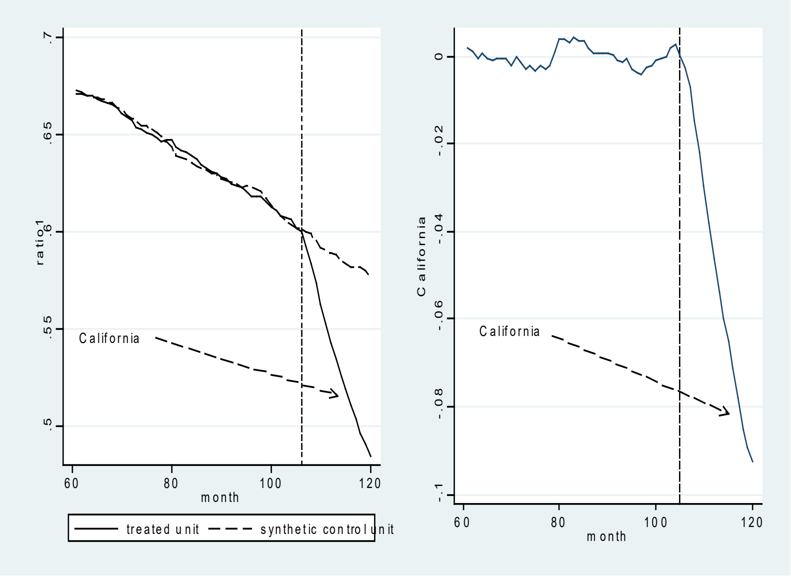
Still thinking about DDD, and still using the DD ratio, we face two problems: (1) What states should be included in the comparison, and (2) What statistical test is appropriate for analysis? The synthetic control methodology answers both questions.

First we treat the ratio SC 1/(SC 1 + SC 2) as the variable of interest. The left-hand panel of figure 8 shows that prior to the California intervention (the broken vertical line) this ratio had been decreasing steadily and that after the intervention the ratio fell precipitously. From a DD perspective, this is fairly strong evidence that Realignment has worked to reduce the proportion of offenders in California prisons for relatively minor crimes (SC 1). The broken line shows the trend for the synthetic cohort, identified as a cluster of states that experienced trends much like those experienced in California *prior to the intervention*.[16] After the intervention, the trend in the synthetic cohort states continued its fairly linear pattern. The fact that the post-intervention trend in California departs from the post-intervention trend in the synthetic cohort suggests that California successfully reduced the proportion of SC 1 offenders to the sum of SC 1 and SC 2 offenders. We have used a DDD perspective to strengthen the evidence from the DD perspective. However, we have not yet provided a statistical test.

To understand the statistical test, first perform a mental calculation. Looking at the left-hand panel, subtract the ratio for California from the ratio for the synthetic cohort. Graph that difference into the panel on the right. Prior to the intervention, the difference is near zero, so the line on the right-hand panel is flat until the intervention. Thereafter the line becomes increasingly negative.

[16] We refer readers elsewhere (Abadie et al., 2010, 2014) for a technical explanation of identifying the synthetic control group. Intuitively, the synthetic control group comprises other states that have trends similar to those experienced in California and have explanatory variables (such as arrests per capita) that have similar values. Members of the synthetic control group are weighted by relevance so some states receive higher weights than do others. Many states receive a weight of zero, meaning they are excluded from the synthetic control group.

Figure 8: California and Synthetic Cohort: SC 1/(SC 1 + SC 2)



Next, translate the right-hand panel from figure 8 to become the left-hand panel in figure 9. Then, repeat the exercise applied to California to every other state; for each state imagine the counterpart to the left-hand panel, and draw that imagined counterpart into the right-hand panel.[17] The right-hand panel looks like a ball of yarn, but what is important is that the trend for California forms a lower boundary for the cluster of lines. Thirty states entered the analysis, so by chance, under the null hypothesis of no treatment effect, California would provide the lower boundary in this figure with a probability of 1/30. Thus we reject the null of no treatment effect in California with a probability of 1/30 = 0.0333.

[17] Some states are so unique that they lack a synthetic cohort. They are excluded from the analysis so that the 33 states that entered the original analysis have been reduced to 30 states.

Figure 9: Test Statistic for Trend in SC 1/(SC 1 + SC 2)

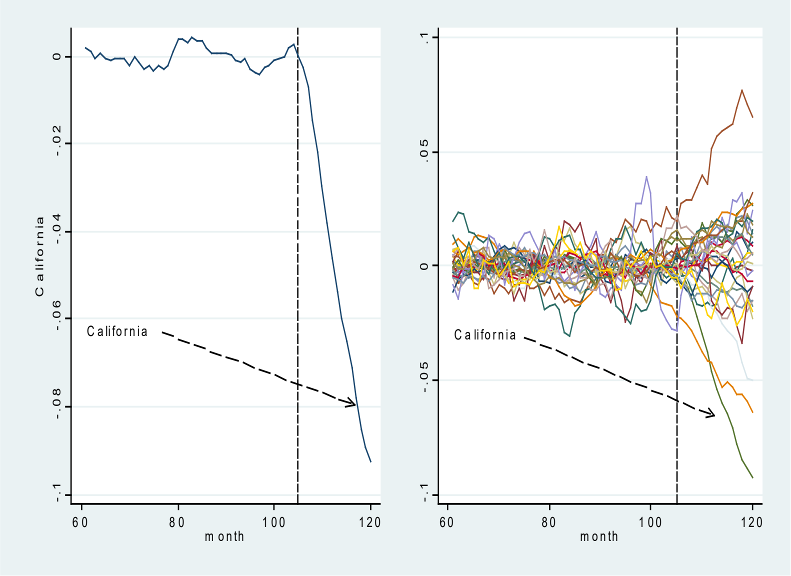
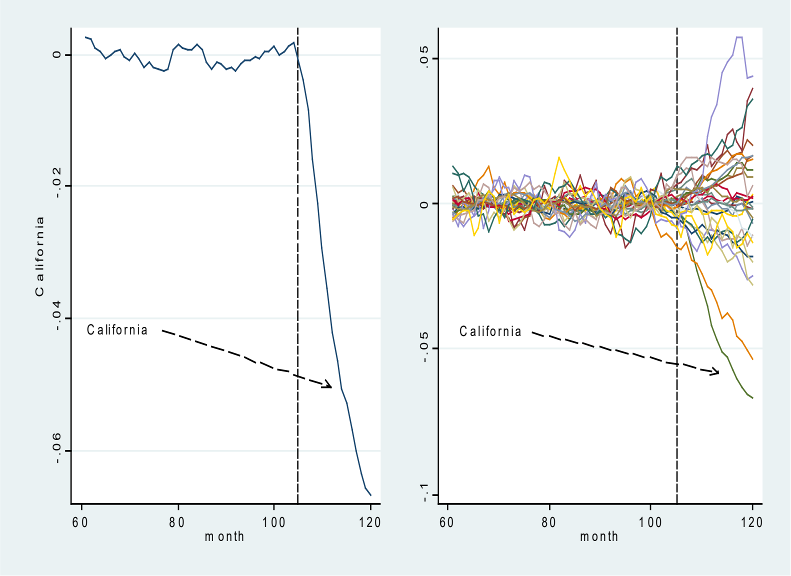


Figure 10 provides a different measure for examining the trend: the ratio of SC 1 to the sum of SC 1, SC 2 and SC 3. The impression is not much changed. From a DD perspective, illustrated by the left-hand panel, we have strong evidence that California's Realignment has altered the composition of its prison population in the intended direction. California has decreased the ratio of the least serious offenders (as judged by offense seriousness) relative to other low seriousness offenders. From the right-hand panel, we have evidence that this change is not spurious, because it has not occurred in other states.

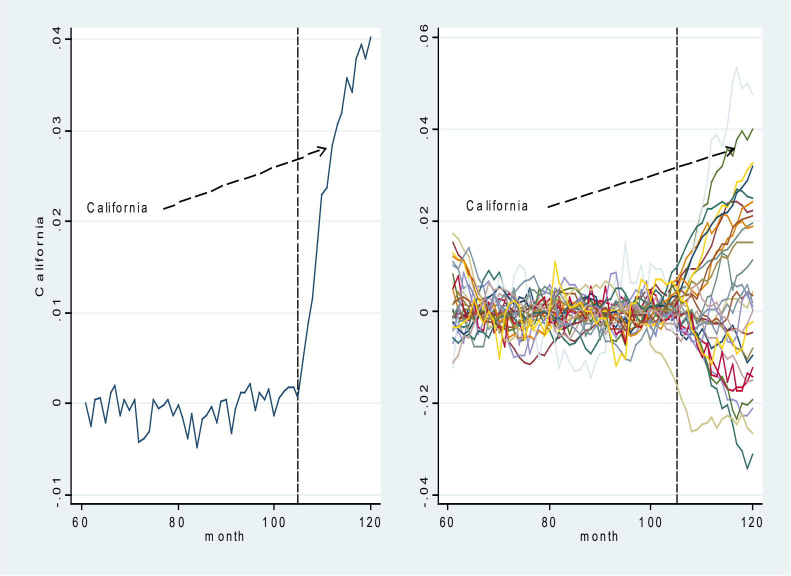
Figure 10: Test Statistics for Trend in SC 1/(SC 1 + SC 2 + SC 3)



The next figure is the counterpart to figure 10 but the ratio represented is SC 5/(SC 3 + SC 4 + SC 5). This tests the null that California has increased the proportion of the most serious offense classes relative to other relatively serious offense classes. One other state actually forms the upper boundary for the ball of twine, so the effect is statistically significant at 0.066.

Another possible null is that California simultaneously decreased the use of incarceration for SC 1 relative to SC 1 and SC 2 and increased the use of incarceration for SC 5 relative to SC 4 and SC 5. What we find is that California has reduced the use of prison for SC 1 (compared with SC 1 plus SC 2) by more than any other state and California has simultaneously increased the use of prison for SC 5 (compared with SC 4 plus SC 5) by more than every other state except one. These two trends are independent, so it is highly unlikely that California could have accomplished these simultaneous changes by chance.

Figure 11: Test Statistics for Trends in SC 5/(SC 3 + SC 4 + SC5)



Although we have rejected an interrupted time-series as a useful evaluation design, we have not reached a conclusion that a DDD is superior to a DD. The synthetic cohort approach is recent and we feel uncomfortable that not enough experience has accumulated to adopt the synthetic control method as the principal evaluation method for statewide prison interventions. Furthermore, when it has been applied, the synthetic control approach has assumed that one state (California, above) has implemented the intervention while other states have not. Authorities have suggested how to deal with multiple states (Abadie et al., 2010), but when several states have adopted an intervention the usefulness of the suggestions is not so obvious.

Our recommendation is using the DD and DDD in conjunction to strengthen conclusions in the face of potential validity challenges. Because prison reforms typically target a specific prison population defined by offense type (seriousness) and offender type (criminal history), it is practical to identify a within-state counterfactual of offenses that are slightly less serious (hence not a target for the intervention) and offenses that are slightly more serious (and hence not a target for the intervention). Many evaluators would argue that this is a relatively strong basis for estimating a treatment effect provided that SUTVA is met.

Nevertheless, a known deficiency of a DD framework is that pre-intervention trends may not portend post-intervention trends in the absence of the intervention. Using the logic of an interrupted time-series framework by limiting the bandwidth is helpful for dealing with this validity challenge, but we have suggested another procedure, namely using a DDD framework to test whether the trends in the state being evaluated differ substantially from the trends in other states. Not all states may offer good comparisons, so using statistical tests, we have applied the synthetic estimation framework to select states.

If the DD-estimated effect is not statistically significant or substantively meaningful, we probably halt the investigation. If it is significant/substantively meaningful, we then apply DDD through the synthetic estimation approach. However, it seems unreasonably conservative to put heavily reliance on statistical significance from the synthetic comparison approach. After all, if we require both tests (the DD and the DDD), under the null the probability of rejecting the null is no longer 0.05, but rather, 0.05x0.05 = 0.0025. Clearly the test is too conservative.

We are unsure of an optimal test, but it seems sensible to mix quantitative and qualitative tests. The quantitative test is based on the DD. As already noted, if we fail to reject the null, then testing ceases. If we reject the null, the qualitative test is based on the DDD. To pass the qualitative test, we would expect California to fall near the lower or upper envelope of the multiple curves, but requiring California to form the envelope (the only way to achieve $p < 0.05$ given 30 states in the study) seems too severe.

5. Conclusions

This paper has discussed approaches to evaluating state-level reforms intended to reduce the use of prison for selected classes of offenders. Evaluation is difficult because random assignment is impractical and evaluation requires other approaches. Alternative approaches face validity and reliability challenges because it is difficult to identify suitable counterfactuals, and when they are identified, sample sizes are small.

We believe that interrupted time-series are poor designs that can lead to spurious findings, sometimes causing evaluators to reject interventions that are beneficial and sometimes causing evaluators to accept interventions that are ineffective. When the intervention targets a class of offenders, then a class of similar offenders within the same state may be a suitable counterfactual. This is the logic of a difference-in-difference design. Some additional rigor may be gained by augmenting the difference-in-differences with a difference-in-difference-in-differences approach, comparing trends across states. The problem is to identify suitable states for comparison and to identify statistical tests that recognize the small sample involved in the comparison. Synthetic control may provide a useful approach.

We have skirted or only briefly mentioned important issues. One issue is identifying the counterfactuals. We based the counterfactuals exclusively on five offense seriousness classes, but this is probably inadequate for many evaluation questions. As already mentioned, most prison reforms target certain offenses and offenders, and the counterfactual should be built around those types. Another issue is that states use their prisons in different capacities. For example, some states may frequently send offenders convicted of domestic assault to prison; other states may do so rarely. If offenders convicted of domestic assault are not part of the targeted group, it seems inappropriate to include them in any analysis that makes cross-state comparisons. This is just to say that an evaluator must think carefully about appropriate counterfactuals, and the choice of a counterfactual will hang on the evaluation question.

Especially when drawing cross-state comparisons, an evaluator needs to consider what other interventions are occurring. For convenience, our illustrations assume that no other interventions were occurring, but that assumption was for convenience and a serious evaluation would carefully determine its truth. California was an extreme choice; no other state, to our knowledge, has imposed such a strong change on its prison system during the same time frame. This will not always be the case, however.

Finally, the discussion has concerned prison population composition, but this is not the only type of question that might be posed and answered using NCRP data. The NCRP is especially useful for studying recidivism defined as returning to prison in the same state. (The NCRP team is working on linking NCRP data across states so over time the definition will be expanded.) Questions about recidivism are equally amenable to the research designs posed here.

References

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Associatio n, 105*(490), 493-505. https://www.jstor.org/stable/29747059

Abadie, A., Diamond, A., & Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science, 59*(2), 495-510. https://doi.org/10.1111/ajps.12116

Angrist, J., & Pischke, J. (2009). *Mostly harmless econometrics: An empiricist's companion.* Princeton University Press.

Berk, R. A. (2005). Randomized experiments as the bronze standard. *Journal of Experimental Criminology, 1*(4), 417-433. https://doi.org/10.1007/s11292-005-3538-2

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis.* John Wiley & Sons. https://onlinelibrary.wiley.com/doi/book/10.1002/9780470743386

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1*(2) 97-111. https://doi.org/10.1002/jrsm.12

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications.* Cambridge University Press.

Imbens, G. W., & Lemieux, T. (2007). *Regression discontinuity designs: A guide to practice* [Working paper 13039]. National Bureau of Economic Research. https://www.nber.org/papers/w13039.pdf

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

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature, 47*(1), 5-86. https://doi.org/10.1257/jel.47.1.5

Lee, M.-J. (2005). *Micro-econometrics for policy, program and treatment effects.* Oxford University Press.

Luallen, J., Rhodes, W., Gaes, G., Kling, R., & Rich, T. (2014). *A description of computing code used to identify correctional terms and histories.* Abt Associates Inc.

Manski, C. F. (2007). *Identification for prediction and decision.* Harvard University Press.

Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference: Methods and principals for social research* (2nd ed.). Cambridge University Press.

Orr, L. L. (1999). *Social experiments: Evaluating public policy programs with experimental methods.* Sage Publications.

Rhodes, W. (2012). Meta-analysis: An introduction using regression models. *Evaluation Review, 36*(1), 24-71. https://doi.org/10.1177%2F0193841X12442673

Rhodes, W., & Jalbert, S. K. (2013). Regression discontiuity design in criminal justice evaluation: An introduction and illustration. *Evaluation Review*, *37*(3-4), 239-273. https://doi.org/10.1177%2F0193841X14523004

Rosenbaum, P. R. (2009). *Design of observational studies.* Springer.

Wooldridge, J. (2007). *What's new in econometrics? Lecture 10 difference-in-differences estimation*. National Bureau of Economic Research. https://www.nber.org/WNE/Slides7-31-07/slides_10_diffindiffs.pdf