A Framework for Synthetic Control Methods with High-Dimensional, Micro-Level Data

Evaluating a Neighborhood-Specific Crime Intervention

Michael Robbins, Jessica Saunders, and Beau Kilmer

RAND Justice, Infrastructure, and Environment

WR-1080-NIJ
February 2015
A FRAMEWORK FOR SYNTHETIC CONTROL METHODS WITH HIGH-DIMENSIONAL, MICRO-LEVEL DATA: EVALUATING A NEIGHBORHOOD-SPECIFIC CRIME INTERVENTION

MICHAEL W. ROBBINS\textsuperscript{a}, JESSICA SAUNDERS\textsuperscript{b} AND BEAU KILMER\textsuperscript{b}

Synthetic control methods are an increasingly popular tool for analysis of economic programs. Here, they are applied to a neighborhood-specific crime intervention in Roanoke, VA, and several novel contributions are made to the synthetic control toolkit. We examine high-dimensional data at a granular level (the treated area has several cases, a large number of untreated comparison cases, and multiple outcome measures). Calibration is used to develop weights that exactly match the synthetic control to the treated region across several outcomes and time periods. Further, we illustrate the importance of adjusting the estimated effect of treatment for the design effect implicit within the weights. A permutation procedure is proposed wherein countless placebo areas can be constructed, enabling estimation of $p$-values under a robust set of assumptions. An omnibus statistic is introduced that is used to jointly test for the presence of an intervention effect across multiple outcomes and post-intervention time periods. Analyses indicate that the Roanoke crime intervention did decrease crime levels, but the estimated effect of the intervention is not as statistically significant as it would have been had less rigorous approaches been used.

Keywords: Synthetic control methods, High-dimensional data, Survey analysis, Calibration, Design effect, Causal inferences, Crime interventions.

1. INTRODUCTION

Efforts to draw causal inferences from the study of a treatment or intervention while using observational data suffer a singular inevitable deficiency: a lack of an experimental design. This shortcoming greatly hinders an analyst’s ability to attribute an observed result as being a consequence of treatment. Attempts to circumvent this shortcoming involve the determination of the (hypothetical) post-intervention state of the treated units in the event that they had, in fact, not been treated. Under difference-in-differences approaches, this is done by extrapolating the pre-intervention state of the treated units onto post-intervention time points via a temporal trend that is determined using untreated (control) units. Difference-in-differences methods are underpinned by assumptions that are at times unrealistic, such as the assumption of a parallel trend (and that the only shock to the system of treated cases within the observed time frame was the intervention).

A generalization of the difference-in-differences approach that provides results that have a more palpable interpretation is synthetic control methodology (Abadie et al., 2010; Abadie and Gardeazabal, 2003). Therein, an untreated version of the treated case(s) (i.e., a synthetic control) is created using a weighted combination of untreated cases. Via comparison of the treated units to their...
M. ROBBINS, J. SAUNDERS AND B. KILMER

respective synthetic control, the analyst can paint a clear visualization of the effect of the intervention. The primary setting for applications of synthetic control methods involve a single treated case with multiple untreated cases for comparison (where all cases have been measured across several time periods before and after the intervention). The relative dearth of data in such settings complicates efforts to a) develop a synthetic control that matches the treated case, b) precisely estimate the effect of treatment, c) gauge the statistical significance of that effect, and d) jointly incorporate multiple outcome variables.

Although micro-level data measured across a large number of dimensions are becoming increasingly commonplace in numerous scientific fields, synthetic control methods are not currently equipped (neither computationally nor methodologically) to handle such data. Here, we enhance the synthetic control toolbox for the purpose of addressing this deficiency—in doing so we also address the shortcomings mentioned earlier. Specifically, the use of high-dimensional, micro-level data within synthetic control methods makes two primary contributions to the program evaluation literature: 1) the creation of a synthetic comparison that can match across multiple covariates and outcomes in an efficient manner and is flexible to be applied across different levels of aggregation and units, and 2) the enabling of statistical assessment jointly across several outcome variables and follow-up periods.

To expound, micro-level data measurements enable incorporation of multiple treated cases with a plethora of untreated cases for comparison. Consequentially, we frame synthetic control methods within the context of survey analysis—doing so permits exploitation of a vast pool of analytical tools (the benefits of which will be illustrated in detail). The use of micro-level data is shown to facilitate the simultaneous analysis of several outcomes and pre-intervention time periods. Sensitivity analyses are used to illustrate that misleading results may be yielded if a synthetic control is created while omitting outcomes or if data have been aggregated to a higher level than necessary (e.g., using state-level data when county-level measurements are available).

We also propose an omnibus test that detects a treatment effect jointly across multiple outcomes and post-intervention time periods. Such a test allows the analyst to control for multiple comparisons. That is, researchers traditionally compare findings across dependent variables, noting which ones are significantly impacted and which ones are not. This new test allows us to go beyond this limited approach and determine if the intervention impacts a group of variables.

This study is motivated by the need to evaluate a neighborhood-based intervention designed to close an overt drug market. Using data at the level of a census block, we evaluate the effectiveness of a Drug Market Intervention (DMI) implemented in the Hurt Park neighborhood of Roanoke, Virginia in late 2011. While DMI has received a great deal of recognition for being an effective strategy (Braga and Weisburd, 2012; Hipple and McGarrell, 2009; Kennedy, 2009), statistical analyses of its impact have encountered methodological problems that mainly stem from the lack of an appropriate comparison group (Corsaro et al., 2012).
SYNTHETIC CONTROL METHODS WITH MICRO-LEVEL DATA

2012; Draca et al., 2011; Saunders et al., 2015b). The DMI data are ideal for illustration of the utility of the procedures we propose. Specifically, the data have sufficient granularity, multiple equally relevant outcomes, and a structure that cannot be captured by commonly used techniques (which we illustrate through comparisons made later). Nonetheless, the methodology introduced here can be used for evaluating a multitude of economic programs in situations where randomization is not possible.1

The rest of this article proceeds as follows. Section 2 reviews traditional synthetic control approaches and presents our methodological innovations. Specifically, we introduce calibration as a tool for calculating synthetic control weights and develop a framework gauging the statistical significance of estimators of a treatment effect in micro-level, high-dimensional settings. Also, we propose a method for approximating an estimator’s sampling distribution by generating permuted placebo groups. Section 3 describes the DMI in greater detail and illustrates findings from application of the proposed methods to the Roanoke crime intervention, and Section 4 illustrates sensitivity analyses that utilize data at a more aggregated level and with truncated dimensionality. Section 5 concludes with discussion that details the advantages and disadvantages of synthetic control methods in our setting emphasizes the policy implications of our findings regarding crime interventions. An appendix presents generalized hypotheses and relates statistical methods for testing them.

2. METHODOLOGY

We begin with the introduction of some notation that is used throughout. Let \( Y_{ijt} \) denote the observed value of outcome \( i \) in block \( j \) at time \( t \). Further, let \( \mathbf{R}_j \) denote a length-\( r \) vector of covariates for block \( j \). We assume that there are a total of \( I \) separate outcomes measured so that \( i \in (1, \ldots, I) \) and that out of \( T \) total time periods measured, there are \( T_0 \) time periods measured prior to the intervention, which implies \( t \in (1, \ldots, T_0, T_0 + 1, \ldots, T) \). Similarly, the control group (i.e., blocks outside of the region that received the intervention) consists of \( J_0 \) blocks out of \( J \) total blocks across both the treatment region and the control group. Hence, after indexing blocks within the control group first, we use \( j \in (1, \ldots, J_0, J_0 + 1, \ldots, J) \). Note that we do not have longitudinal measurements of the covariates in \( \mathbf{R}_j \).

1 As an example, consider enterprise zone programs, which are targeted economic revitalization efforts that provide tax breaks and economic development incentives to troubled urban areas. Research into their effectiveness remains mostly inconclusive due to the types of methodological challenges we address in this paper (Boarnet, 2001). For instance, the programs operate in differently sized geographical areas and may impact a variety of economic outcomes (looking at only one outcome could be misleading). Bondonio and Greenbaum (2007) evaluate enterprise zones using a model that combines propensity scores (Rosenbaum and Rubin, 1983) and fixed effects with data at the level of a U.S. Postal ZIP code. Their approach is subject to restrictive model formulations and difficulties with jointly incorporating and evaluating multiple outcome measures—our method addresses these drawbacks.
2.1. Synthetic Control

When outlined in our context, the paradigm of Abadie et al. (2010) for synthetic control methods stipulates that each observed outcome $Y_{ijt}$ has the representation

\[ Y_{ijt} = Y_{ijt}(0) + \alpha_{ijt}D_{jt}, \]

where $D_{jt}$ is a treatment indicator that is unity only if block $j$ has received the treatment at time $t$ and is zero otherwise. Further, $Y_{ijt}(0)$ is a (sometimes latent) quantity indicating the outcome measurement in the absence of treatment; the underlying model structure imposed upon $Y_{ijt}(0)$ is described later. In the presence of treatment, the observed outcome is $Y_{ijt} = Y_{ijt}(1) := Y_{ijt}(0) + \alpha_{ijt}$.

Therefore, our interest is in determination (or at least approximation) of the treatment effect given by $\alpha_{ijt}$. For our purposes, it is sufficient to consider the effect of treatment when averaged across blocks within the treatment group:

\[ \alpha_{it}^* = \frac{1}{J - J_0} \sum_{j = J_0 + 1}^{J} \alpha_{ijt}, \]

for $i \in (1, \ldots, I)$ and $t \in (T_0 + 1, \ldots, T)$.

Approximation of the quantity in (2.2) requires enumeration of the aggregated outcomes for the treated regions in the absence of treatment at post-intervention time points: $Y_{ij}^*(0) = \sum_{j = J_0 + 1}^{J} Y_{ijt}(0)$. This term cannot be observed; however, it may be approximated through construction of a synthetic control group mimics the hypothetical behavior of the treatment group in the absence of treatment. Specifically, we aim to calculate a set of weights, $(w_1, \ldots, w_J)$ (with each block in the control group receiving its own non-negative weight), so that for every outcome $i$ and time period $t$, the weighted blocks in the control group aggregate to their respective totals across the blocks within the treatment group. Specifically, the weights satisfy

\[ \sum_{j = 1}^{J_0} w_j Y_{ijt} = \sum_{j = J_0 + 1}^{J} Y_{ijt}, \]

for each combination of outcomes and time periods that has $i \in (1, \ldots, I)$ and $t \in (1, \ldots, T_0)$. We also impose

\[ \sum_{j = 1}^{J_0} w_j = J - J_0, \]

which implies that the synthetic weights sum to the number of blocks within the treatment group, and that

\[ \sum_{j = 1}^{J_0} w_j R_j = \sum_{j = J_0 + 1}^{J} R_j, \]
which stipulates that the synthetic control has the same covariate values as the aggregated treatment group.

Given weights \((w_1, \ldots, w_J_0)\) that satisfy (2.3)-(2.5), the outcome values in the absence of treatment for the treated region at post-intervention time points can be approximated by \(\hat{Y}_{it}^*(0) = \sum_{j=1}^{J_0} w_j Y_{ijt}\). Therefore, we approximate the effect of the intervention across the treated region via

\[
\hat{\alpha}_{it}^* = \frac{1}{J - J_0} \left( \sum_{j=J_0+1}^{J} Y_{ijt} - \sum_{j=1}^{J_0} w_j Y_{ijt} \right),
\]

for outcome \(i\) at time \(t\) where \(t > T_0\). By design, \(\hat{\alpha}_{it}^*\) approximates the quantity in (2.2). Similarly, we suggest

\[
\hat{\alpha}_{i}^{**} = \frac{1}{T - T_0} \sum_{t=T_0+1}^{T} \hat{\alpha}_{it}^*,
\]

to estimate the average post-intervention treatment effect on outcome \(i\).

Our first concern is whether or not the quantities in (2.6) and (2.7) are unbiased estimators of the effect of treatment. If \(\{w_j^*\}\) denotes weights that satisfy a version of (2.3) that replaces \(Y_{ijt}\) with its (unknown) conditional mean, straightforward calculations would illustrate the unbiasedness of (2.6) and (2.7) when calculated using \(\{w_j^*\}\) in place of \(\{w_j\}\). Evaluating bias in quantities that utilize \(\{w_j\}\) requires a greater degree of rigor. Abadie et al. (2010) assume that \(Y_{ijt}(0)\) is derived linearly via a factor model with mean-zero shocks—we make similar assumptions here, although additional bookkeeping is needed to account for multivariate response. Letting \(Y_{jt}(0) = (Y_{1jt}(0), \ldots, Y_{Ijt}(0))'\) denote a length-\(I\) vector of outcomes for block \(j\) at time \(t\), we assume

\[
Y_{jt}(0) = \delta_t + \theta_t R_j + \lambda_t \mu_j + \epsilon_{jt},
\]

where \(\delta_t\) is a length-\(I\) vector of common factors, \(\theta_t\) is an \((I \times r)\) parameter matrix. Further, \(\lambda_t\) and \(\mu_t\) are \((I \times F)\) and \((F \times 1)\) factor matrices, respectively. Lastly, \(\epsilon_{jt}\) is a length-\(I\) vector of transitory shocks with \(\text{E}[\epsilon_{jt}] = 0\) and \(\text{Var}(\epsilon_{jt}) = \Sigma_{jt}\).

Combining (2.1), (2.2) and (2.6), we see

\[
\hat{\alpha}_{it}^* - \alpha_{it}^* = \frac{1}{J - J_0} \left( \sum_{j=J_0+1}^{J} Y_{ijt}(0) - \sum_{j=1}^{J_0} w_j Y_{ijt}(0) \right),
\]

where we assume that \(\{w_j\}\) has been selected to satisfy (2.3)-(2.5). Using (2.8) and applying the arguments of Abadie et al. (2010), any potential bias in \(\hat{\alpha}_{it}^*\) can now be expressed via

\[
\text{E} \left[ \sum_{j=J_0+1}^{J} Y_{ijt}(0) - \sum_{j=1}^{J_0} w_j Y_{ijt}(0) \right] = \text{E} \left[ \lambda_{it}' (A'A)^{-1} A' \sum_{j=1}^{J_0} w_j \epsilon_{jt}^* \right],
\]
for $t > T_0$ where $X_{it}^*$ denotes the $i$th row of $X_t$, $X = (X_1^T, \ldots, X_{T_0}^T)^T$ is a matrix with dimension $(IT_0 \times F)$, and $\epsilon_{jt}^* = (\epsilon_{j1}^T, \ldots, \epsilon_{jT_0}^T)^T$ is a length-$IT_0$ vector. The potential for bias arises from the fact that one may not assume that $E[w_i \epsilon_{jt}^*] = 0$ (as the weights are, in a sense, functions of the shocks). However, a bound may be placed on the bias. Specifically, letting $\sigma_{ijt}^2$ denote the $i$th diagonal element of $\Sigma_{jt}$ and $\bar{\sigma}_i^2 = \max_j \{T_0^{-1} \sum_{t=1}^{T_0} \sigma_{ijt}^2\}$, calculations show that (2.9) is bounded by a term that is proportional to $[\bar{\sigma}_i^2/(IT_0)]^{1/2}$ — the denominator comes from the number of rows of $A$. Hence, the potential for bias in $\hat{\alpha}_{jt}^*$ is mitigated if the number of outcomes ($I$), and/or the number of pre-intervention time periods ($T_0$) is large.

### 2.2. Calibration of weights

The analyses of Abadie and Gardeazabal (2003), wherein the original framework for synthetic control methods was proposed, are performed over large regions leaving limited data units available for the construction of a synthetic control. Even though their study is restricted to analysis of a single outcome variable with an intervention applied to a single case, they are unable to build a weighted control that exactly matches the pre-intervention behavior of the treated unit. Most of the subsequent applications of their procedure (e.g., Abadie et al., 2010, 2014; Billmeier and Nannicini, 2013; Bohn et al., 2014; Cavallo et al., 2013) follow a similar framework. However, our data are recorded at a much more granular level with thousands of untreated cases available. Therefore, we consider options for construction of a synthetic control that exactly matches the treated region with respect to observed characteristics.

Specifically, we exploit methods commonly used in analysis of surveys. Synthetic weights are derived using calibration techniques (Deville and Särndal, 1992; Särndal, 2007). We set

\[
X_j = (1, Y_{1j1}, \ldots, Y_{1jT_0}, Y_{2j1}, \ldots, Y_{2jT_0}, \ldots, Y_{I1j1}, \ldots, Y_{IjT_0}, R_j^T)^T,
\]

that is, a vector of all outcomes at all pre-intervention time points (with an intercept term and covariates) for block $j$. The target totals for the treatment region are given by $t_x = \sum_{j=J_0}^{J_0} X_j$. Given initial values of the weights, the calibration process finds values of $w_j$ for $j = 1, \ldots, J_0$ that satisfy a set calibration equations given by

\[
(2.10) \quad \sum_{j=1}^{J_0} w_j X_j = t_x.
\]

The weights are calculated in this manner using the function `calibrate` (with `calfun = 'raking'`) within the `survey` package in R (Lumley, 2004, 2011). The algorithm used therein is outlined in Deville et al. (1993) and is labeled
a generalized raking procedure. Given a distance metric \( G(\cdot) \) that satisfies regularity conditions, the procedure solves for the set \( \{ w_j \} \) that minimizes the distance between the calibrated weights and corresponding initial weight values \( \{ d_j \} \) (specifically, the quantity \( \sum_{j=1}^{J_0} d_j G(w_j/d_j) \) is minimized) subject to the constraints imposed by (2.10). To briefly explain in more detail, we solve for the set of weights \( \{ w_j \} \) and the vector of Lagrange multipliers \( \xi \) that minimize the objective function

\[
\sum_{j=1}^{J_0} d_j G(w_j/d_j) - \xi' \left( \sum_{j=1}^{J_0} w_j X_j - t_x \right).
\]

Defining \( g(x) = dG(x)/dx \) and \( F(u) = g^{-1}(u) \), it follows that \( w_j = d_j F(X'_j \xi) \) for each \( j \) where \( \xi \) satisfies \( \sum_{j=1}^{J_0} d_j F(X'_j \xi) X_j = t_x \). In practice, Newton’s method is used to extract the value of \( \xi \) that satisfies the latter formula. In lieu of an informed design, we use \( d_j = (J - J_0)/J_0 \) for each \( j \) as initial weights. As a distance metric, \( G(x) = x \log x - x + 1 \) is used, which ensures non-negative weights.

Since numerical methods are only utilized to solve for \( \xi \), which is a vector with dimension equal to that of \( t_x \), generalized raking is a more efficient algorithm than one which optimizes over a parameter space that has dimension equal to \( J_0 \), the number of synthetic control weights. However, a key drawback of this procedure is as follows. In circumstances where the algorithm is unable to determine weights that offer exact satisfaction of (2.10) (e.g., \( t_x \) does not fall within the convex hull of the \( X_j \) for \( j = 1, \ldots, J_0 \)), the algorithm will not necessarily return weights that still have practical utility (e.g., the weights may diverge from their targeted values). In such instances, one would need to use a different procedure to find the set of weights that most closely satisfy (2.10). Large (or perhaps divergent) weight values can be prevented through the use of the distance metric defined in the so-called logit method of Deville et al. (1993) by placing bounds on the weight values. However, bounded weights do not provide an exact solution to (2.10) in our application; therefore, we do not consider them further.

2.3. Test Statistics

To evaluate the presence of an intervention effect, we consider tests of

\[
H_{0i} : \alpha_{i}^{**} = 0 \quad \text{against} \quad H_{1i} : \alpha_{i}^{**} < 0,
\]

for each outcome \( i \), where \( \alpha_{i}^{**} = \sum_{t=T_0+1}^{T} \alpha_i^t / (T - T_0) \) is the average post-intervention treatment effect for outcome \( i \). Methods that enable estimation of a unique treatment effect at each time period are discussed in the Appendix. The alternative hypothesis in (2.11) is one-sided since the result of interest is a decrease in crime as a result of intervention; however, each of the statistics introduced here can be modified for a two-sided alternative.
Upon estimation of weights for the synthetic control, we can use $\hat{\alpha}_i^{**}$ from (2.7) to approximate the treatment effect. However, in order to evaluate the hypotheses in (2.11), we must first determine the standard error of $\hat{\alpha}_i^{**}$. As we are placing synthetic control procedures in the context of survey methodologies, we note that there is a design effect (in the nomenclature of Kish, 1965) inherent in the weights $\{w_j\}$. Therefore, the incorporation of $\{w_j\}$ into estimation of a treatment effect will result in increased standard error of estimators. We consider methods for standard error approximation that will incorporate this design effect.

**Naïve approach to standard error estimation**

We first present a naïve approximation for standard error that represents a rudimentary effort to incorporate the added variability supplied by the weights for synthetic control. Assume that $Y_{ijt}$ has temporally constant mean $\mu_{ij}$ and variance $\sigma^2_{ij}$ for each $i$ and $j$. Then, $\hat{\alpha}_i^{**}$ will observe

$$E[\hat{\alpha}_i^{**}] = \frac{1}{J - J_0} \left( \sum_{j=J_0+1}^{J} \mu_{ij} - \sum_{j=1}^{J_0} w_{ij} \mu_{ij} \right) \approx 0,$$

and

$$\text{Var}(\hat{\alpha}_i^{**}) = \frac{1}{(J - J_0)^2(T - T_0)} \left( \sum_{j=J_0+1}^{J} \sigma^2_{ij} + \sum_{j=1}^{J_0} w_{ij}^2 \sigma^2_{ij} \right) \approx \frac{1}{(J - J_0)^2(T - T_0)} \left( \sum_{j=J_0+1}^{J} s^2_{ij} + \sum_{j=1}^{J_0} w_{ij}^2 s^2_{ij} \right) =: \text{Var}(\hat{\alpha}_i^{**}),$$

where $s^2_{ij}$ is the sample variance of $(Y_{ij1}, \ldots, Y_{ijT_0})$ and where the elements of $\{w_j\}$ are treated as fixed constants. Under the stated assumptions and given a sufficient sample size, a test statistic defined by

$$Z_i^{**} = \frac{\hat{\alpha}_i^{**}}{\sqrt{\text{Var}(\hat{\alpha}_i^{**})}}$$

should approximately have a standard normal sampling distribution under $H_{0i}$. We reject $H_{0i}$ for small values of $Z_i^{**}$. The $p$-value of a test of the hypotheses in (2.11) is calculated via $\Phi(Z_i^{**})$, where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution; the same expression is used to calculate the $p$-value of any test statistic that has a standard normal sampling distribution. For a test of the two-sided alternative, $H_{1i}^\ell : \alpha_i^{**} \neq 0$, we use $(Z_i^{**})^2$ as a test statistic.

There are several flaws inherent in the statistic defined above. First, its underlying model does not enable temporal shifts in the mean and variance sequence of
an outcome under the null hypothesis. Secondly, although the statistic’s construction incorporates the design effect implicit within the synthetic control weights, it is not necessarily appropriate to treat the weights (which, as mentioned earlier, are functions of the observed data to a degree) as fixed constants.

**Survey methods for standard error estimation**

We exploit survey methodologies in order to develop more sophisticated approaches for standard error approximations. These methods mandate a linear model representation for the outcomes. For a specific outcome $i$, we fit the regression

$$Y_{ijt} = \beta_{it} + a_{it}^* D_{jt} + \epsilon_{ijt},$$

where $\beta_{it}$ is a fixed effect for time period $t$, $D_{jt}$ is the treatment indicator seen in (2.1), $\epsilon_{ijt}$ is mean-zero error, and $a_{it}^*$ is the coefficient of interest. The model is applied separately for each outcome, and we isolate to post-intervention measurements (i.e., we restrict to $t \in (T_0 + 1, \ldots, T)$). The above model is fit using weighted least squares (WLS); the weights used for WLS are the synthetic control weights for $j \in (1, \ldots, J_0)$, and we set $w_j = 1$ for $j \in (J_0 + 1, \ldots, J)$. Although we omit the relevant calculations, the model in (2.13) is used because the estimate of $a_{it}^*$ when found using WLS is equivalent to $\hat{\alpha}_{it}^{**}$ in (2.7). Therefore, the fit of the regression, when calculated using the appropriate software (we use the function `svyglm()` in the R package `survey`), also gives $\hat{\text{Var}}(\hat{\alpha}_{it}^{**})$, an approximation of the variance of $\hat{\alpha}_{it}^{**}$. This variance estimator is calculated using Taylor series linearization (Binder, 1983).

Using these quantities, a statistic for testing the hypotheses in (2.11) is

$$Z_{i}^{**} = \frac{\hat{\alpha}_{it}^{**}}{\sqrt{\hat{\text{Var}}(\hat{\alpha}_{it}^{**})}}$$

for each outcome. If $H_{0i}$ and (2.13) hold, $Z_{i}^{**}$ can be assumed to have been sampled from a standard normal distribution. The model in (2.13) is not introduced for the purpose of imposing new restrictions on the behavior of our data; this formula serves as a channel through which the variability in $\hat{\alpha}_{it}^{**}$ can be monitored. Our objective is not to produce an accurate estimation of this variability (which would require that (2.13) hold), but instead to gauge how it is influenced by the structure of the weights. Robustness to the representation in (2.13) is procured through the permutation schemes described later.

**Omnibus tests**

It may be desirable to develop a statistic that tests for a treatment effect simultaneously across multiple outcomes. It is not sufficient to apply the above methods to an aggregated outcome, as this will not necessarily give equal emphasis to potential treatment effects within each outcome. Therefore, for a test
of the omnibus hypotheses
\[ H_0 : \alpha_{1}^{**} = \cdots = \alpha_{I}^{**} = 0 \]
against
\[ H_{i}^{-} : \alpha_{i}^{**} \leq 0 \text{ for all } i \text{ with } \alpha_{i}^{**} < 0 \text{ for some } i, \]
we suggest the statistic
\[ \hat{Z}^{**} = \sum_{i=1}^{I} \hat{Z}_{i}^{**}. \]

An analogous quantity can be stated based on the \( \tilde{Z}_{i}^{**} \). We do not attempt to derive an expression for the sampling distribution of \( \hat{Z}^{**} \) here; it is instead approximated through permutation methods described later. To address the two-sided alternative, \( H_{i}^{\pm} : \alpha_{i}^{**} \neq 0 \text{ for some } i \), we suggest \( \hat{Z}^{\pm} = \sum_{i=1}^{I} (\hat{Z}_{i}^{**})^2 \).

Our omnibus statistic does not adjust for correlation across the outcomes in the vein of a standard Wald statistic (i.e., through utilization of a full variance/covariance matrix). Although it is possible to develop a Wald-type omnibus measure for our setting (note, however, that approximation of the covariances between the \( \hat{\alpha}_{i}^{**} \) while incorporating the synthetic control weights is not trivial), the outcomes in our application are not correlated to the extent needed to warrant exploration of such techniques. For our purposes, standardization of each \( \hat{\alpha}_{i}^{**} \) within \( \hat{Z}^{**} \) is sufficient for ensuring that each outcome receives equal emphasis in omnibus evaluations.

2.4. Difference-in-Differences

As a basis of comparison for the synthetic control methods studied here, we consider a difference-in-differences (DiD) approach. DiD models have become the primary tool for exploration of intervention effects in observational settings akin to those considered here. DiD models were applied in a similar setting by Saunders et al. (2014).

The DiD specification employed herein, which is similar in construction to (2.13), models the mean sequence for outcome \( i \) via
\[ \log(\mu_{ijt}) = \beta_{it} + \gamma_{ij} + \alpha_{i, did} D_{jt} + \eta_{i}^{*} \mathbf{R}_{j}, \]
where \( j \in (1, \ldots, J) \) and \( t \in (1, \ldots, T) \). Further, \( \mu_{ijt} = E[Y_{ijt}] \), the \( \beta_{it} \) are fixed effects for each time period, the \( \gamma_{ij} \) are block-level fixed effects, and \( D_{jt} \) is the treatment indicator seen in (2.1). The model is fit separately for each outcome \( i \). Further, \( \mathbf{R}_{j}^{*} = \log(\mathbf{R}_{j} + 1) \) — since the mean sequence is modeled on the log scale, the covariates are also measured on the log scale (one is added prior to applying
the logarithm in order to ensure a computationally feasible transformation)—
and $\eta_i$ is a vector of corresponding regression coefficients. Lastly, $\alpha_{i,did}$ is the
coefficient of interest which gives the treatment effect. When standardized in the
vein of (2.14), the estimate of $\alpha_{i,did}$ is assumed to have approximately a standard
normal distribution.

Since the outcomes of interest are count variables, the above model is fit using
negative binomial regression (without weighting observations); this is why the
log-scale mean sequence $\log(\mu_{ijt})$ is modeled. See Saunders et al. (2014) for fur-
ther description of analogous DiD models. Other discrepancies between (2.13)
and (2.16) are explained thusly. Since the pre-intervention levels of outcomes
and covariates are matched precisely between the treatment area and its syn-
thetic control, the model in (2.13) does not need to incorporate data observed
prior to the intervention, and likewise block-level fixed effects are unnecessary
in (2.13). Further, least squares is used when fitting (2.13) so as to ensure that
the estimated treatment effect equals $\hat{\alpha}_i^{**}$ as expressed in (2.7). Distributional
misspecification within (2.13) is handled through the use of permutation in cal-
culation of standard errors.

Our setting enables more precise estimation of treatment effects via DiD ap-
proaches than settings that utilize a single treated unit (e.g., Abadie et al., 2010;
Billmeier and Nannicini, 2013; Cavallo et al., 2013). However, it has been ob-
served that DiD models underestimate standard error (Bertrand et al., 2004).
Synthetic control methods enable less restrictive assumptions due to underlying
models akin to (2.8). Further, synthetic controls provide a modus for simulta-
neous incorporation of multiple outcomes. Hence, synthetic control methods are
preferred here. If the model in (2.16) and its underlying assumptions are indeed
valid, DiD methods may be preferable since they do not incorporate a design
effect (and associated loss in precision) that results from weighting observations.

2.5. Generation of Placebo Groups through Permutation

We have presented statistics for testing the hypotheses in (2.11) and have given
algebraic expressions to approximate the sampling distribution of the treatment
effect estimator of (2.7). However, these approximations are based upon restric-
tive model assumptions (e.g., the formulation in (2.13)) and do not account for
complex aspects of the process used to calculate the statistics (e.g., generation
of a synthetic control region); we are concerned that the approximations may not
adequately incorporate all variability inherent in the statistics. Further, we have
not developed an expression for the sampling distribution of omnibus statistics.
Hence, we explore resampling techniques as a mechanism for garnering a more
robust scope of the sampling distributions.

To gauge the statistical significance of their results, Abadie et al. (2010) em-
ploy a placebo study wherein each of the (untreated) comparison areas that were
available for construction of a synthetic control is utilized as a placebo region.
Placebo tests (a.k.a., falsification or refutability tests) have a rich history within
the literature on program evaluation (e.g., Angrist and Krueger, 1999; Auld and Grootendorst, 2004; DiNardo and Pischke, 1997). In the guise of synthetic control methods, these tests involve the creation of a synthetic control group with respective weights for each placebo region. Further, an estimate of the hypothetical treatment effect is calculated for each placebo region based on a comparison to its synthetic control. The resulting estimates are assumed to provide a reasonable scope of the sampling distribution (under the hypothesis of a null effect) for an analogous estimate of the effect of the intervention in the true treatment region. However, the placebo test of Abadie et al. (2010) is limited by the number of available comparison groups; they study between 19 and 38 valid placebo cases. This number is far too small to enable a researcher to garner a sufficient understanding of the extremities of the sampling distribution. Thereby, efforts to report \( p \)-values are incapacitated (since precise approximations of \( p \)-values cannot be provided).

Our setting contains thousands of valid comparison areas at the block level; this enables a robust accounting of the sampling distribution of our test statistics through placebo methods. Specifically, we employ a permutation technique that can be used to randomly generate any desired number of placebo regions using the available data. The \( J \) total blocks are randomly reordered—the first \( J_0 \) blocks of the reordered data denote the comparison blocks (used for building synthetic control) that hypothetically did not receive treatment, and the final \( J - J_0 \) blocks of the reordered data denote the placebo region that hypothetically receive the intervention. Further, the observations for outcome \( i \) at time \( t \) within the \( k^{th} \) reordering are given by the vector \((Y_{i1}^{(k)}, \ldots, Y_{iJ}^{(k)})^\prime\), where \((Y_{i,J_0+1,t}, \ldots, Y_{iJ,t})^\prime\) corresponds to the placebo treatment region. Then, a vector of weights \((w_1^{(k)}, \ldots, w_{J_0}^{(k)})^\prime\) is calculated (using the methods described earlier) which satisfies versions of (2.3)–(2.5) that use the \( Y_{ij}^{(k)} \) in place of \( Y_{ij,t} \).

Let \( Z \) denote a generic test statistic calculated using the observed data from the treatment region (prior to any permuting) and corresponding synthetic control with weights \( \{w_j\} \), and let \( Z^{(k)} \) denote a version of \( Z \) calculated using the \( k^{th} \) placebo region and its respective synthetic control with weights \( \{w_j^{(k)}\} \). If \( K \) total placebo regions have been sampled, in theory \((Z^{(1)}, Z^{(2)}, \ldots, Z^{(K)})\) will encompass the sampling distribution of \( Z \) for sufficiently large \( K \). Therefore, the \( p \)-value for \( Z \) may be given by

\[
(2.17) \quad p = \frac{\#k : Z^{(k)} < Z}{K}.
\]

This formula can be used to derive \( p \)-values for the statistics in Section 2.3 (by using, for example, \( Z = Z^{**}_t \)). Note that (2.17) is the only mechanism presented here that may be used for approximating the \( p \)-value of the omnibus statistic \( Z^{***} \).

One could use placebo groups to directly ascertain the sampling distribution
of a treatment effect prior to any standardization (e.g., using \( Z = \hat{\alpha}_t^{**} \)), which would circumvent the need for the model assumptions implicit in (2.13). However, the placebo areas here represent a random assortment of blocks, whereas the treatment area is likely more structured. That is, it is conceivable that the weights for treatment area will have a larger design effect on average than corresponding weights for the placebo areas. As a manner of guarding against biases that result from such a discrepancy, we prefer to standardize the treatment effect estimator \( \hat{\alpha}_t^{**} \) using (2.12) or (2.14) prior to calculating a permuted \( p \)-value via (2.17).

Further, deriving the distribution of \( \tilde{Z}_t^{**} \) by using permuted placebo groups is expected to make the results robust to the assumptions required by (2.13). To summarize, we calculate \( p \)-values based on permuted placebo groups while using \( \tilde{Z}_t^{**} \) instead of \( \hat{\alpha}_t^{**} \) so as to filter out the design effect yielded by the synthetic control weights.

### 3. Roanoke Crime Intervention

In this section, we consider the efficacy of an intervention to close overt drug markets that brings together buyers and sellers of drugs at set times in geographically well-defined areas. Overt markets facilitate the sale and use of drugs and can pose threats to public health and safety (Harocopos and Haugh, 2005; Reuter and MacCoun, 1992). Participants in these markets sometimes engage in violence, and the markets can also have other negative effects on the quality of life for nearby residents, including noise, vandalism, burglary, prostitution, traffic congestion, panhandling, and disorderly conduct (Baumer et al., 1998; Blumstein and Rosenfeld, 1998; Weisburd and Mazerolle, 2000). Efforts to disrupt street-level drug operations are notoriously difficult because, even if arrested, dealers are often quickly replaced and traditional law enforcement responses have the potential to exacerbate already tenuous police-citizen relations (Caulkins, 1993; Kleiman, 1997; Mazerolle et al., 2006). Traditional enforcement methods often challenge the delicate relationship between communities and the criminal justice system, particularly when the involvement of law enforcement exacerbates long-standing mistrust between police and community members (Caulkins, 1993; Frydl et al., 2004; Tyler, 2004; Weitzer et al., 2006).

The Drug Market Intervention (DMI, c.f., Kennedy and Wong, 2009; McGarrell et al., 2010) is a problem-solving program where actors in the criminal justice system and the community work together to address an overt drug market. DMI was designed in response to criticism regarding aggressive police tactics that were seen as unfair and racially motivated (Kennedy, 2009) and was modeled on previous focused deterrence programs (Kennedy et al., 1996) It is a collaboration between law enforcement, prosecutors, the community, and social service providers, providing a holistic model to disrupting overt drug markets. As part of a DMI, police officers (and sometimes community members) identify sellers involved in the overt market, make undercover buys, and help build credible cases to prosecute the offenders engaged in drug sales. Police and prosecutors
arrest and prosecute those dealers who are deemed to be violent and dangerous. The remaining dealers are publicly presented with a second chance and told their cases will be prosecuted if they continue to deal drugs in the overt market. Concurrently, the community is encouraged to take back their neighborhoods and prevent the reemergence of the overt market. The market disruption and reduction in associated crime and disorder are maintained by stronger neighborhood institutions and more positive police-community relations and cooperation (Kennedy, 2009; Saunders et al., 2015a).

This study specifically focuses on the DMI that was implement in the Hurt Park neighborhood of Roanoke, Virginia in late 2011. The Roanoke DMI took place in the Hurt Park neighborhood, an area of 2,785 residents that are predominantly African-American with 80 percent of households earning incomes under $35,000 per year. The neighborhood has a long history of drugs and violence crime with a 6 by 7 block area that the police and community characterized as an overt drug market. The DMI team, comprised members from law enforcement, prosecution, social services, and the community, identified 15 active drug dealers, and determined that ten were violent and should be arrested and prosecuted, while five were deemed nonviolent and given an opportunity to avoid incarceration. The nonviolent dealers were brought to a call-in in December 2011 where they were shown the evidence against them and offered a second chance if they participated in social services. After the call-in, police increased their presence in Hurt Park and worked to mend relationships with the community. A complete description of the DMI is available in Saunders et al. (2015b).

We have longitudinal data measurements at the census block level taken at quarterly intervals that extend from three years prior to the intervention to 18 months following the intervention. Several covariates are also measured (though not longitudinally) for each block. We also monitor multiple aggregated outcome measures. Specific outcomes, aggregated outcomes, and covariates are given in Table 1. The table also lists the baseline levels for each variable within the treatment region (see the column labeled “Hurt Park”) and all blocks in Roanoke (“All blocks”). Baseline levels of crime variables denote the total number of crimes occurring during during the 36-month period prior to the intervention, whereas baseline levels for the covariates are not aggregated across time (our data contain a single baseline value for covariates). To enable Hurt Park to be compared to all of Roanoke on a per capita basis, baseline levels for all blocks are scaled by a constant that equals the number of residents in Hurt Park divided by the total number of residents across all blocks. We see that Hurt Park experiences more criminal activity per capita than the rest of Roanoke. Specifically, Hurt Park accounts for 0.9% of the population of Roanoke, but reports 1.9% of the total crimes and 4.1% of the total drug crimes in the city.

The outcome variables most commonly used in crime evaluations generally come from administrative crime data collected by police departments through their records management systems in accordance with the FBI’s Uniform Crime
SYNTHETIC CONTROL METHODS WITH MICRO-LEVEL DATA

TABLE I

<table>
<thead>
<tr>
<th>Measure</th>
<th>Frequency</th>
<th>Description</th>
<th>Hurt Park</th>
<th>All blocks</th>
</tr>
</thead>
<tbody>
<tr>
<td>i Rape</td>
<td>36-month</td>
<td>N rapes reported</td>
<td>8</td>
<td>2.63</td>
</tr>
<tr>
<td>i Robbery</td>
<td>18-month</td>
<td>N robberies reported</td>
<td>13</td>
<td>5.03</td>
</tr>
<tr>
<td>i Aggravated Assault</td>
<td>6-month</td>
<td>N aggravated assault reported</td>
<td>25</td>
<td>10.25</td>
</tr>
<tr>
<td>i Burglary</td>
<td>3-month</td>
<td>N burglaries reported</td>
<td>62</td>
<td>25.88</td>
</tr>
<tr>
<td>i Larceny</td>
<td>3-month</td>
<td>N larcenies reported</td>
<td>107</td>
<td>99.96</td>
</tr>
<tr>
<td>i Car Theft</td>
<td>3-month</td>
<td>N car thefts reported</td>
<td>11</td>
<td>8.81</td>
</tr>
<tr>
<td>i Arson</td>
<td>12-month</td>
<td>N arsons reported</td>
<td>5</td>
<td>1.11</td>
</tr>
<tr>
<td>i Simple Assault</td>
<td>3-month</td>
<td>N simple assaults reported</td>
<td>187</td>
<td>83.14</td>
</tr>
<tr>
<td>i Drugs</td>
<td>3-month</td>
<td>N drug crimes reported</td>
<td>170</td>
<td>38.94</td>
</tr>
<tr>
<td>i Weapons</td>
<td>12-month</td>
<td>N weapons crimes reported</td>
<td>36</td>
<td>6.72</td>
</tr>
</tbody>
</table>

Outcomes Aggregated

<table>
<thead>
<tr>
<th>Measure</th>
<th>Frequency</th>
<th>Description</th>
<th>n violent Crime reported</th>
<th>n Property Crime reported</th>
<th>n Any Crime reported</th>
</tr>
</thead>
<tbody>
<tr>
<td>i Violent</td>
<td>3-month</td>
<td>Total n violent crimes reported</td>
<td>238</td>
<td>102.4</td>
<td></td>
</tr>
<tr>
<td>i Property</td>
<td>3-month</td>
<td>Total n property crimes reported</td>
<td>180</td>
<td>134.7</td>
<td></td>
</tr>
<tr>
<td>i Any Crime</td>
<td>3-month</td>
<td>Total n crimes reported</td>
<td>1599</td>
<td>793.1</td>
<td></td>
</tr>
</tbody>
</table>

Covariates

<table>
<thead>
<tr>
<th>Measure</th>
<th>Baseline</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TotalPop</td>
<td>Baseline</td>
<td>N of residents</td>
</tr>
<tr>
<td>Black</td>
<td>Baseline</td>
<td>N African American residents</td>
</tr>
<tr>
<td>Hispanic</td>
<td>Baseline</td>
<td>N Hispanic residents</td>
</tr>
<tr>
<td>White</td>
<td>Baseline</td>
<td>N white residents</td>
</tr>
<tr>
<td>Males 15-21</td>
<td>Baseline</td>
<td>N male residents aged 15-21</td>
</tr>
<tr>
<td>Households</td>
<td>Baseline</td>
<td>N households</td>
</tr>
<tr>
<td>Vacant Units</td>
<td>Baseline</td>
<td>N vacant housing units</td>
</tr>
<tr>
<td>Female</td>
<td>Baseline</td>
<td>N female headed households</td>
</tr>
<tr>
<td>Renter</td>
<td>Baseline</td>
<td>N households occupied by renters</td>
</tr>
</tbody>
</table>

Reporting procedure (Gove et al., 1985; Lejins, 1966). The FBI has divided crimes in Part I and Part II Offenses, which are further classified as violent, property, or “other”, and has jurisdictions report their crime reports according to a set of common definitions. Part I crimes are homicide, rape, robbery, and assault, burglary, larceny, motor vehicle theft, and arson; all other crimes, including drug use and sales, are classified as Part II crimes (Law Enforcement Support Section, and Crime Statistics Management Unit, 2013). We also use indicators of weapons charges, drug charges, and simple assaults, as these outcomes have been examined in prior program evaluations; however, they are considered less reliable measures since they are Part II crimes (Gove et al., 1985).

The aggregated outcomes monitored here include i Any Crime, a sum of all reported crimes (both Part I and Part II), i Property, which includes all Part I property crimes (e.g., burglary, larceny, motor vehicle theft, and arson), and i Violent, which includes all Part I violent crimes (homicide, rape, robbery, and assault). The rest of the crime measures reflect only those crimes falling under that particular category (e.g., i Drugs are all drug-related crimes summed across all subcategories of drug offenses). The i Property, i Violent and i Any Crime variables include crimes that are not otherwise analyzed here; therefore, they are included as outcomes in the process used to create synthetic control weights.

Not all outcomes are included at the quarterly level within the process to create weights for the synthetic control. Some crimes occur infrequently and must be aggregated to larger time intervals (a proliferation of zeros across outcomes often makes calculation of nonnegative weights that match the treatment and synthetic control areas mathematically infeasible). Table I also indicates the
temporal frequency at which each outcome is measured when used in calculation of synthetic control weights. Frequencies were chosen so as to use the minimal amount of temporal aggregation while maintaining the feasibility of a synthetic control region that meets constraints.

In all, we have data measured across \( J = 3601 \) blocks; the intervention area contains 66 blocks (thus, \( J_0 = 3535 \)). Including the \( I = 13 \) outcomes across which treatment and control are matched for at most \( T_0 = 12 \) pre-intervention time periods as well as covariates and the intercept, the synthetic control is created by matching across 129 variables.

We applied the methodology introduced in Section 2.2 to create synthetic control weights for the Hurt Park treatment region in addition to 1,000 permuted placebo regions. Under rare circumstances, weights could not be created that exactly match a placebo region to its respective synthetic control. This occurred in 39 of the 1,000 placebo regions (and does not occur at all if the dimensionality of the constraint vector \( \mathbf{t} \) is reduced slightly); these 39 regions were deleted from further use. As is common for generalized raking procedures, large weights are present. Focusing our attention on the weights for the Hurt Park treatment region, the largest weight assigned to a single block is 3.57, and 16 blocks are given a weight greater than one. We considered bounding and/or trimming weights to control outlying values; however, we feel that the potential for bias that is induced by the resulting loss of satisfaction of the conditions in (2.10) is not preferable to the large design effect that comes from unbounded weights. Further, of the 3535 blocks assigned a weight, only 338 are given a weight with a value greater than 0.0001 (these blocks account for 99.94% of the total weight).

Our first set of results is presented in Figure 1. The figure shows outcome levels for the treatment region and its synthetic control region for the \( \text{i\_any\_crime} \), \( \text{i\_drugs} \) and \( \text{i\_aggassault} \) outcomes. Crime levels for Roanoke as a whole (which are also plotted after being scaled in the same manner as analogous results in Table I) indicate small amounts of seasonality but do not show further temporal trends that may inform crime patterns within the treatment region. Additionally, Figure 1 includes plots of the difference in outcome measurements between the treatment region and its synthetic control overlaid on the corresponding differences for the first one hundred placebo regions.

Figure 1 illustrates that the synthetic Hurt Park region exactly matches the observed Hurt Park with respect to pre-intervention levels of \( \text{i\_any\_crime} \) and \( \text{i\_drugs} \) (this is the case for all variables listed in Table I that are used at 3-month frequencies). An exact match between treatment and synthetic control in each quarter is not obtained for \( \text{i\_aggassault} \) since it is used at 6-month intervals when calculating weights. Similarly, an exact match is not obtained for any variable listed in Table I that is aggregated beyond a quarterly frequency.

We note that the weights used within the synthetic control for the treatment region appear to have a larger design effect than corresponding weights for the placebo groups—this is likely a consequence of the Hurt Park being a more structured region than the randomly selected placebo groups. Specifically, using
the Kish approximation (as mentioned earlier), the weights used within WLS (as described in Section 2.3) for the treatment region have a design effect of 31.3, whereas the median corresponding design effect for the placebo regions is 16.7 (with a 90th percentile of 19.3). The large magnitude of all of these design effects is due an abundance of weights being calculated as (nearly) zero. Furthermore, the synthetic control estimator $\hat{\alpha}_{i}^{**}$ likely has more variability when calculated for the treatment region than when calculated for the control areas—it is necessary to use a statistic that is standardized with a variance term that incorporates the design effect of the synthetic control weights.

We calculate $p$-values for four different test statistics calculated for pertinent outcomes (for three of the statistics, $p$-values are calculated using a stand normal approximation in addition to an approximation based on permuted placebo groups). The statistics are outlined as follows:

- Diff.-in-Diff. – The difference-in-differences approach of Section 2.4; $p$-values are approximated with a standard normal sampling distribution (Norm.) and permuted placebo groups (Perm.). This method does not incorporate the synthetic control.
- $\hat{\alpha}_{i}^{**}$ – The raw value of $\hat{\alpha}_{i}^{**}$, the synthetic control estimator in (2.7); $p$-values are approximated with permuted placebo groups (Perm.).
- Naïve – The statistic $\hat{\alpha}_{i}^{**}$ when standardized via the naïve approach in (2.12); $p$-values are approximated using a standard normal sampling distribution (Norm.) and permuted placebo groups (Perm.).
- Survey – The statistic $\hat{\alpha}_{i}^{**}$ when standardized via survey methodological approach of (2.14); $p$-values are approximated using a standard normal sampling distribution (Norm.) and permuted placebo groups (Perm.).

The statistics above are calculated for each of the outcomes listed in Table I. An omnibus statistic in the vein of (2.15) is also calculated for each method that utilizes placebo groups (no valid algebraic approximations for sampling distribution of omnibus statistics are outlined here). The omnibus test is calculated across each of the (non-aggregated) outcomes listed in Table I.

Results of the tests mentioned above are provided in Table II for several choices of $T$, the maximum time period considered. Therein, $p$-values for each of the statistics listed above is provided for each outcome. Note that a $p$-value for the omnibus measure cannot be calculated using standard normal assumptions. To give context to the results, the table also lists post-DMI changes in the levels of the various outcomes (see the column labeled “Chng. (%)”). The percentage change figures were computed by comparing post-DMI crime levels with the corresponding levels observed within the synthetic control. These percent changes are replete with statistical uncertainty, which we measure here through the host of testing procedures. The findings show that when the treatment region is compared to its synthetic control, nearly all of the outcomes observe a decrease in levels following the intervention.

The findings in Table II inform our understanding of the performance of synthetic control methods. First, it appears that the normal approximation and the
TABLE II

p-Values of tests for a drug market intervention effect in Roanoke, VA.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DiD-In-Diff</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 Months Post Intervention</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1_rape</td>
<td>0.500 0.395</td>
<td>0.273</td>
<td>0.374 0.362</td>
<td>0.220</td>
<td>0.226</td>
<td>0.024 0.094</td>
</tr>
<tr>
<td>1_robbery</td>
<td>0.306 0.001</td>
<td>0.670</td>
<td>0.547 0.639</td>
<td>0.103</td>
<td>0.011</td>
<td>0.097 0.287</td>
</tr>
<tr>
<td>1_aggassault</td>
<td>0.212 0.151</td>
<td>0.105 0.190</td>
<td>0.144</td>
<td>0.046</td>
<td>0.081 0.051</td>
<td></td>
</tr>
<tr>
<td>1_burglary</td>
<td>0.103 0.081</td>
<td>0.000 0.002</td>
<td>0.223</td>
<td>0.025</td>
<td>0.055 0.203</td>
<td></td>
</tr>
<tr>
<td>1_larceny</td>
<td>0.144 0.051</td>
<td>0.012 0.000</td>
<td>0.984</td>
<td>0.574</td>
<td>0.959 0.740</td>
<td></td>
</tr>
<tr>
<td>1_cartheft</td>
<td>0.249 0.003</td>
<td>0.001 0.000</td>
<td>0.039</td>
<td>0.086</td>
<td>0.259 0.103</td>
<td></td>
</tr>
<tr>
<td>1_drugs</td>
<td>0.028 0.004</td>
<td>0.246 0.315</td>
<td>0.984</td>
<td>0.594</td>
<td>0.917 0.790</td>
<td></td>
</tr>
<tr>
<td>1_weapons</td>
<td>0.028 0.004</td>
<td>0.246 0.315</td>
<td>0.984</td>
<td>0.594</td>
<td>0.917 0.790</td>
<td></td>
</tr>
<tr>
<td>1_violent</td>
<td>0.045 0.024</td>
<td>0.000 0.000</td>
<td>0.039</td>
<td>0.086</td>
<td>0.259 0.103</td>
<td></td>
</tr>
<tr>
<td>1_property</td>
<td>0.002 0.173</td>
<td>0.040 0.147</td>
<td>0.039</td>
<td>0.086</td>
<td>0.259 0.103</td>
<td></td>
</tr>
<tr>
<td>1_anycrime</td>
<td>0.002 0.173</td>
<td>0.040 0.147</td>
<td>0.039</td>
<td>0.086</td>
<td>0.259 0.103</td>
<td></td>
</tr>
<tr>
<td>Omnibus</td>
<td>0.075 0.012</td>
<td>0.000 0.000</td>
<td>0.045</td>
<td>0.024</td>
<td>0.000 0.000</td>
<td></td>
</tr>
<tr>
<td>12 Months Post Intervention</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1_rape</td>
<td>0.595 0.589</td>
<td>0.450 0.488</td>
<td>0.220 0.109</td>
<td>0.401 0.396</td>
<td>0.011 0.072</td>
<td></td>
</tr>
<tr>
<td>1_robbery</td>
<td>0.220 0.109</td>
<td>0.293 0.381</td>
<td>0.306 0.367</td>
<td>0.401 0.396</td>
<td>0.011 0.072</td>
<td></td>
</tr>
<tr>
<td>1_aggassault</td>
<td>0.220 0.109</td>
<td>0.293 0.381</td>
<td>0.306 0.367</td>
<td>0.401 0.396</td>
<td>0.011 0.072</td>
<td></td>
</tr>
<tr>
<td>1_burglary</td>
<td>0.064 0.060</td>
<td>0.000 0.001</td>
<td>0.118 0.138</td>
<td>0.007 0.028</td>
<td>0.012 0.002</td>
<td></td>
</tr>
<tr>
<td>1_larceny</td>
<td>0.007 0.028</td>
<td>0.012 0.002</td>
<td>0.118 0.138</td>
<td>0.007 0.028</td>
<td>0.012 0.002</td>
<td></td>
</tr>
<tr>
<td>1_cartheft</td>
<td>0.165 0.155</td>
<td>0.133 0.130</td>
<td>0.137 0.182</td>
<td>0.165 0.155</td>
<td>0.133 0.130</td>
<td></td>
</tr>
<tr>
<td>1_drugs</td>
<td>0.020 0.076</td>
<td>0.175 0.259</td>
<td>0.232 0.197</td>
<td>0.069 0.040</td>
<td>0.147 0.146</td>
<td></td>
</tr>
<tr>
<td>1_weapons</td>
<td>0.020 0.076</td>
<td>0.175 0.259</td>
<td>0.232 0.197</td>
<td>0.069 0.040</td>
<td>0.147 0.146</td>
<td></td>
</tr>
<tr>
<td>1_violent</td>
<td>0.046 0.000</td>
<td>0.000 0.000</td>
<td>0.206 0.061</td>
<td>0.011 0.034</td>
<td>0.173 0.136</td>
<td></td>
</tr>
<tr>
<td>1_property</td>
<td>0.001 0.101</td>
<td>0.000 0.000</td>
<td>0.167 0.136</td>
<td>0.001 0.101</td>
<td>0.000 0.000</td>
<td></td>
</tr>
<tr>
<td>1_anycrime</td>
<td>0.000 0.113</td>
<td>0.002 0.036</td>
<td>0.008 0.004</td>
<td>0.000 0.113</td>
<td>0.002 0.036</td>
<td></td>
</tr>
<tr>
<td>Omnibus</td>
<td>0.054 0.000</td>
<td>0.010 0.001</td>
<td>0.037 0.004</td>
<td>0.054 0.000</td>
<td>0.010 0.001</td>
<td></td>
</tr>
<tr>
<td>18 Months Post Intervention</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1_rape</td>
<td>0.326 0.226</td>
<td>0.024 0.096</td>
<td>0.201 0.213</td>
<td>0.180 0.163</td>
<td>0.216 0.329</td>
<td></td>
</tr>
<tr>
<td>1_robbery</td>
<td>0.180 0.163</td>
<td>0.216 0.329</td>
<td>0.230 0.312</td>
<td>0.180 0.163</td>
<td>0.216 0.329</td>
<td></td>
</tr>
<tr>
<td>1_aggassault</td>
<td>0.360 0.366</td>
<td>0.054 0.173</td>
<td>0.196 0.201</td>
<td>0.360 0.366</td>
<td>0.054 0.173</td>
<td></td>
</tr>
<tr>
<td>1_burglary</td>
<td>0.048 0.051</td>
<td>0.001 0.003</td>
<td>0.011 0.032</td>
<td>0.048 0.051</td>
<td>0.001 0.003</td>
<td></td>
</tr>
<tr>
<td>1_larceny</td>
<td>0.003 0.332</td>
<td>0.033 0.002</td>
<td>0.029 0.030</td>
<td>0.003 0.332</td>
<td>0.033 0.002</td>
<td></td>
</tr>
<tr>
<td>1_cartheft</td>
<td>0.129 0.123</td>
<td>0.145 0.162</td>
<td>0.209 0.244</td>
<td>0.129 0.123</td>
<td>0.145 0.162</td>
<td></td>
</tr>
<tr>
<td>1_drugs</td>
<td>0.934 0.926</td>
<td>0.835 0.693</td>
<td>0.496 0.576</td>
<td>0.934 0.926</td>
<td>0.835 0.693</td>
<td></td>
</tr>
<tr>
<td>1_weapons</td>
<td>0.011 0.112</td>
<td>0.371 0.448</td>
<td>0.314 0.261</td>
<td>0.011 0.112</td>
<td>0.371 0.448</td>
<td></td>
</tr>
<tr>
<td>1_violent</td>
<td>0.000 0.135</td>
<td>0.000 0.000</td>
<td>0.002 0.003</td>
<td>0.000 0.135</td>
<td>0.000 0.000</td>
<td></td>
</tr>
<tr>
<td>1_property</td>
<td>0.000 0.208</td>
<td>0.053 0.121</td>
<td>0.031 0.017</td>
<td>0.000 0.208</td>
<td>0.053 0.121</td>
<td></td>
</tr>
<tr>
<td>Omnibus</td>
<td>0.084 0.008</td>
<td>0.034 0.005</td>
<td>0.056 0.006</td>
<td>0.084 0.008</td>
<td>0.034 0.005</td>
<td></td>
</tr>
</tbody>
</table>

Permutation method yield similar approximations of the sampling distribution of the synthetic control estimator (i.e., compare the p-values for “Norm.” and “Perm.” under either the naïve or survey approach). However, this is not the case for the difference-in-differences method. For DiD, the normal approximation commonly yields smaller p-values than those seen with the permutation technique (this is in line with the observations of Bertrand et al., 2004). Additionally, p-values found using the raw treatment effect (\( \hat{\alpha}_i^{*\ast} \)) are slightly lower on average than those found using the naïve method, which seemingly implies the naïve method accomplishes the objective of incorporating added variability. However, the p-values for the survey methods are noticeably larger (although still
statistically significant at the 5% level for many outcomes) than those yielded by the raw treatment effect. Also, all methods indicate that there is a statistically significant effect of the intervention on the any_crime measure. Unlike the any_crime measure, which is driven largely by frequently occurring crimes, the omnibus statistic can gauge the presence of an intervention effect simultaneously across several outcomes while giving equal emphasis to each outcome. In summary, the statistic based on survey adjustments when used with permuted placebo groups is the most exhaustive of the procedures described herein— it is designed to incorporate the most sources of variability and is the most robust to model assumptions. Therefore, it is our preferred specification.

We next discuss substantive findings. Each panel in Table II presents the p-value for an omnibus test which tells us whether the intervention had some sort of effect across all the non-aggregated measures we test (i.e., excluding violent, property, and any crime). Using our preferred specification, the p-value is not statistically significant at 6 months (0.189), but it is at 12 months (0.037) and it is marginally significant at 18 months (0.056).

Table II also includes the percent change and p-values for each of the crimes used in the omnibus test as well as aggregate categories violent, property, and any crime. The percent change for any crime was -18.6% with a p-value of 0.039 under our preferred specification at six months following the intervention. These results are largely driven by a reduction in property crimes (-45.9%; \( p = 0.026, 0.016 \)), not violent crimes (-13.3%, \( p = 0.248 \)). Perhaps surprisingly, none of the specifications identify an effect on drug crimes (-2.2%; \( p = 0.472 \)). While it may be the case that DMI really did not reduce drug market activity in Hurt Park, it could also be the case that the DMI increased the probability that residents called the police to report drug crime or the police paid more attention to it (thus potentially off-setting any actual reduction).

The second and third panels of Table II present parallel values for 12 and 18 month post-intervention, respectively (results are cumulative in that the second panel incorporates all time periods in the year following the intervention). The absolute value of the effect sizes get slightly larger from 6 to 12 months for property (-50.6%, \( p = 0.002 \)) and any crimes (-22.9%, \( p = 0.004 \)), and are much more statistically significant. The absolute value of the effect sizes get smaller from 12 to 18 months, albeit still substantively and statistically significant (property: -43.1%, \( p = 0.003 \); any crimes: -12.7%, \( p = 0.017 \)). This suggests that some of the DMI effect may dissipate over time.

4. Sensitivity Analyses

Several analyses were performed to examine the sensitivity of our findings to a wide array of initial conditions. For example, we varied the choice of weights to initialize the calibration algorithm and altered the depth of variables and/or data-level used within the calculation of weights. We also applied the procedure while bounding and/or trimming weights (although when doing so, we were
no longer able to develop a synthetic control that exactly matches treatment to control). Generally speaking, results similar to those provided by the main analyses were seen across the sensitivity analyses; however, in certain cases, noticeable distinctions were evident. Specifically (and in line with the theme of this article), we present an analysis wherein a synthetic control is created using a truncated set of outcomes and a separate analysis that uses data aggregated to the neighborhood (as opposed to block) level.

First, we consider an analysis that involved separating out each outcome and building a synthetic control by matching to only the specific outcome variable (e.g., i_burglary, i_larceny, i_drugs, etc.) and the covariates. Of particular interest are the results when this is done for i_drugs. Our main analyses found no effect of the intervention on drug crimes; however, a synthetic control is built by only matching to quarterly drug crimes and covariates, we see the appearance of strong decrease in drug crimes (when compared to the synthetic control region) following the intervention. Specifically, at 6 months following the intervention, this sensitivity analysis indicated a 28.4% (with a p-values of 0.230 and 0.280 when using the survey methods statistic with the normal and permutation approximations, respectively) drop in drug crimes from the levels seen in its synthetic control. At 12 months following the intervention, the decrease in drug crime is estimated to be 35.4% ($p = 0.066$ and $0.117$), and at 18 months post-intervention, the drop is 43.4% ($p = 0.012$ and $0.073$). Note that the main analyses never indicated more than 2.2% ($p = 0.478$ and $0.472$) decrease in drug crimes.

Figure 2 provides analogues of the plots seen in Figure 1 for this sensitivity analysis. The figure implies that drug crime levels for the synthetic control are much higher following the intervention than they were for the synthetic control used in the main analysis (i.e., compare to Figure 1). However, examination of crime rates across the synthetic control calculated using only drug crimes provides insight as to which results are more trustworthy. Figure 2 shows levels for i_any_crime across this synthetic control region. It is seen that in the pre-intervention period, crime levels in the synthetic control are much higher than in the treatment region. Therefore, it is not surprising that drug crimes in this synthetic control region are at high levels following the intervention. Note that we would not use weights based on drug crime levels to estimate the treatment effect in a separate variable such as i_any_crime; we include the latter variable in Figure 2 to show why these weights are suboptimal for estimating the effect of the intervention on drug crime levels. To summarize, this sensitivity analysis illustrates clearly the superiority of a synthetic control region that is built using a wide array of outcomes.

To perform an additional sensitivity analysis, we aggregated data from the block level to the neighborhood level. This is done by first pooling all blocks within the treatment region into a single treated case (i.e., overt drug market). We also aggregate the comparison regions into 109 disjoint (and exhaustive) segments that are of similar geographic size to the treated neighborhood. This
revised setting is analogous to the original setting for synthetic control methods as outlined by Abadie and Gardeazabal (2003) and Abadie et al. (2010). Therefore, we employ the Synth algorithm (Abadie et al., 2011) in calculation of synthetic control weights (the calibration procedure used previously is not optimal here since an exact match between the treatment and control is not possible with the meso-level DMI data). We examine all outcome variables and covariates on a per capita basis by dividing by TotalPop for each neighborhood. Our proposed method was not applied to per capita data since 1) when the treatment region matches the synthetic control on raw values of all variables (including TotalPop), the treatment region will also match the synthetic control in per capita terms, and 2) multiple blocks have TotalPop equal to zero, meaning they would have to be discarded in order to use per capita data (however, many of these blocks had reported crimes, which means it would be imprudent to drop them).

First, we use Synth to create a synthetic control while matching across all variables (in accordance with the scheme outlined in Table I). In this case, there are simply not enough donor units to build a synthetic control that reasonably matches the treated neighborhood across all variables. For example, the value of i\_any\_crime for treated and synthetic control regions across all observed time periods is shown in the left-hand plot of Figure 3; the synthetic control systematically reports fewer crimes than the treated neighborhood. We then applied Synth to each outcome individually to see if improved matching can be obtained. The right-hand plot in Figure 3 shows the values of i\_any\_crime for the treatment and synthetic control regions when only i\_any\_crime and covariates input into the Synth procedure. An improved (although still imprecise) match is obtained; however, matching on a single outcome while using Synth will induce the same potential for bias that was seen in the sensitivity analysis described above which used our algorithm.

5. DISCUSSION

5.1. Method

Our study illustrates the advantages of utilizing high-dimensional, micro-level data in the context of synthetic control methods. Synthetic control methods with a large $J$ enable the analyst to frame the problem in the context of survey analysis and to tap into the associated methodologies. Specific advantages of using micro-level data, wherein there are multiple treated cases (so that the cumulative treatment effect is the result of interest) with a large number of donor units, that have been illustrated from this study are:

- The precision of estimators of the effect of treatment is improved.
- One can jointly incorporate a large number of variables (in terms of distinct outcomes, pre-intervention time periods and covariates). Failure to incorporate a robust set of outcomes may result in omitted variable biases.
• It is often possible to develop a synthetic control area that exactly matches the treated region across several variables.
• When an exact match can be made, algorithms for calibration of weights can be used that are computationally efficient. Algorithms that are typically used to calculate weights for synthetic control can be computationally burdensome.
• One can develop a (nearly) infinite number of placebo areas via permutation techniques. This allows precise estimation of \( p \)-values for statistics that test for a treatment effect. Further, it enables development of an omnibus test for the presence of a treatment effect across multiple outcomes.
• Reasonable approximations of sampling distributions of synthetic control-based estimators for the effect of treatment can be made without using placebo regions (i.e., through a normal distribution).

Several of the above items overcome drawbacks of difference-in-differences approaches. For instance, standard DiD models cannot completely account for multiple outcomes, and normal approximations involving difference-in-differences estimators are unreliable.

We have also illustrated that one should be cautious about certain aspects of synthetic control methods when using placebo groups that have been permuted in the manner described herein. Specifically, the weights that correspond to the treatment region may have a systematically larger design effect than the corresponding weights for the placebo regions. Therefore, prior to calculating permutation-based \( p \)-values, one should adjust the estimators of a treatment effect using a quantity that incorporates the design effect.

A notable disadvantage of the procedure that we have outlined is that calibration (i.e., generalized raking) is not preferable for developing weights in the event that an exact match between treatment and synthetic control is infeasible. Further, there is little guidance that can be offered to help the analyst predict when such circumstances may arise. As the number of variables used for matching treatment to control increases, the likelihood of an exact match being infeasible also increases (whereas this likelihood decreases as the number of cases available for matching increases). Our method also suffers the same drawbacks commonly encountered when using observational data to draw causal inferences (e.g., indisputable determination of causal relationship cannot be reached without a randomized design). An interesting extension of the procedure discussed here involves development of a separate synthetic control for each treat unit (e.g., block). Such an approach is much more likely to encounter computational issues (e.g., computing time and the likelihood of being unable to match treatment with control) and was therefore not explored here.

5.2. Drug Market Application

While neighborhood problem-solving approaches to crime reductions have been shown to be promising crime control strategies, most evaluations suffer
from significant methodological shortcomings, with the selection of the appropriate comparison group and statistical modeling strategy being most problematic (Saunders et al., 2014; Sherman et al., 2002). The DMI, an example of such a program, has been lauded as an effective program, despite the fact that the majority of the evidence comes from the city where it was designed (e.g., Corsaro et al., 2011, 2012; Kennedy and Wong, 2009; Saunders et al., 2014). Our application of this procedure to administrative crime data from Roanoke, VA generates results suggesting the approach can substantively and significantly reduce crime in other neighborhoods, especially property crimes. The results of evaluation contribute to the growing body of evidence that the more focused and specific the strategies of the police, and the more tailored to the problems they seek to address, the more effective the police will be in controlling crime and disorder (Braga and Weisburd, 2010, 2011; Weisburd and Eck, 2004). Future work could apply this procedure to administrative data in places where DMI, or other problem-solving crime control programs have been conducted and not been evaluated.

APPENDIX A: MODELS WITH INTERACTED TREATMENT EFFECTS

All the statistics discussed in Section 2.3 describe estimation of the treatment effect $\alpha_i^{**}$ which is aggregated effect across $t \in (T_0 + 1, \ldots, T)$ for outcome $i$ (this is akin to the imposition that the effect of treatment is the same at each post-intervention time point for outcome $i$). Here, we consider development of test statistics akin to those described in Section 2.3 that enable a unique treatment effect at each time period. Specifically, consider the following linear relationship:

$$Y_{ijt} = \beta_{it} + \alpha_{it}^{**} D_{jt} + \epsilon_{ijt},$$

where the $\beta_{it}$ are fixed effects for the time period and where the $\alpha_{it}^{**}$ for $t \in (T_0 + 1, \ldots, T)$ are coefficients that correspond to an interaction of the time-based fixed effects and the treatment indicator $D_{jt}$.

Let $\alpha_i^* = (\alpha_{i,T_0+1}, \ldots, \alpha_{iT})'$ represent the vector of treatment effects for the intervention area where $\alpha_{it}^{**}$ is defined in (2.2). Further, let $\hat{\alpha}_i^* = (\hat{\alpha}_{i,T_0+1}, \ldots, \hat{\alpha}_{iT})'$ denote the estimated version of $\alpha_i^*$ when (A.1) is fit using weighted least squares (with weights as described in Section 2.3) and let $\hat{\Sigma}_i = \text{Var}(\hat{\alpha}_i^*)$; both $\hat{\alpha}_i^*$ and $\hat{\Sigma}_i$ can be calculated using the survey package in R. Note that the elements of $\hat{\alpha}_i^*$ equal $\hat{\alpha}_{it}^{**}$ as seen in (2.6). Denote the diagonal elements of $\hat{\Sigma}_i$ via the vector $\hat{\sigma}_i = (\hat{\sigma}_{i,T_0+1}, \ldots, \hat{\sigma}_{iT})'$ so that $\hat{\sigma}_{iT}^2 = \text{Var}(\hat{\alpha}_{iT})$. Further, let $\hat{D}_i$ denote a matrix with $\hat{\sigma}_i$ along the diagonal and zeros on the off diagonals. Note that each $\hat{\alpha}_{it}^{**}$ when calculated in the manner described here is equivalent to the expression seen in (2.6).

**Tests of treatment effects at single time periods**

To test for the presence of a treatment effect at a single post-intervention time point, we consider $H_{0it} : \alpha_{it}^{**} = 0$ against $H_{1it} : \alpha_{it}^{**} < 0$, for a specific $t \in (T_0 + 1, \ldots, T)$ where $\alpha_{it}^{**}$ is defined in (2.2). For a test statistic of these hypotheses, we use $\hat{Z}_{it}^2 = \hat{\alpha}_{it}^{**} / \sqrt{\hat{\sigma}_{it}}$, which has a standard normal distribution under $H_{0it}$. To test $H_{0it}$ against the two-sided alternative $H_{1it} : \alpha_{it}^{**} \neq 0$, we suggest $(\hat{Z}_{it})^2$ as a test statistic, which will have a $\chi^2$ distribution with one degree of freedom if $H_{0it}$ holds.

**Tests of treatment effects for single outcomes**

In testing for a treatment effect across one outcome while allowing for each post-intervention time period to have a unique treatment effect, we consider the following one-sided hypotheses...
for a specific $i$:

$$H_{0i} : \alpha_i^{*} = \cdots = \alpha_i^{* T} = 0$$

against

$$H_{1i}^- : \alpha_i^{* t} \leq 0 \text{ for all } t > T_0 \text{ with } \alpha_i^{* t} < 0 \text{ for some } t > T_0.$$ 

We suggest $Z_{i, \text{MSE}}^* = \sum_{t=T_0+1}^{T} \hat{Z}_{i, t}^*/\hat{v}_t$, where $\hat{v}_t^2 = 1' \hat{D}_t^{-1/2} \hat{\Sigma}_t \hat{D}_t^{-1/2} 1$, for testing the above hypotheses. We use $1$ to denote a vector of ones with appropriate length. It follows that $Z_{i, \text{MSE}}^*$ will have a standard normal sampling distribution under $H_{0i}$. In practice, we expect that $Z_{i, \text{MSE}}^*$ will offer results that are more or less equivalent to those given by $Z_i^*$ of (2.14).

Although they do not formally define hypothesis tests, the statistic used by Abadie et al. (2010) (post-intervention MSE) is analogous to $Z_{i, \text{MSE}}^* = \sum_{t=T_0+1}^{T} (\hat{g}_{it}^*)^2$ in our setup. This statistic could be used to test $H_{0i}$ against the two-sided alternative $H_{1i}^+ : \alpha_i^{* t} \neq 0$ for some $t$. We will not derive the sampling distribution of $Z_{i, \text{MSE}}^*$. Instead, we suggest the related Wald-type quantity

$$\hat{Z}_{i}^\ddagger = (\hat{\alpha}_i^*)' (\hat{\Sigma}_i)^{-1} \hat{\alpha}_i^*,$$

which has a $\chi^2$ distribution with $T - T_0$ degrees of freedom under $H_{0i}$.

**Other omnibus tests**

We now consider the following omnibus hypotheses with a one-side alternative (which allow for different treatment effects for each $t$) that consider the treatment effect $\alpha_i^*$ across all relevant values if $i$ and $t$:

$$H_{0i} : \alpha_i^{* t} = \cdots = \alpha_i^{* T} = 0 \text{ for each } i$$

against

$$H_{1i}^- : \alpha_i^{* t} \leq 0 \text{ for each } i \text{ and } t > T_0 \text{ with } \alpha_i^{* t} < 0 \text{ for some } i \text{ and } t > T_0.$$ 

To test the above, we propose $Z_{i, \text{MSE}}^{**} = T^{-1/2} \sum_{i=1}^{I} \hat{Z}_{i}^\ddagger$. When using the two-sided alternative $H_{0i}^\ddagger : \alpha_i^{* t} \neq 0$ for some $i$ and $t > T_0$, we propose $Z_{i, \text{MSE}}^{***} = T^{-1/2} \sum_{i=1}^{I} \hat{Z}_{i}^\ddagger$. Placebo methods are suggested as a means of deriving the sample distribution of these quantities.

In a final set of omnibus hypotheses, we examine the treatment effect across all outcomes for a specific time period. That is, for a fixed $t$ we consider

$$H_{0i,t} : \alpha_i^{* t} = 0 \text{ for each } i$$

against

$$H_{1i,t}^- : \alpha_i^{* t} \leq 0 \text{ for all } i \text{ with } \alpha_i^{* t} < 0 \text{ for some } i.$$ 

As a test statistic for the above, we suggest $\hat{Z}_{i,t}^\ddagger = \sum_{i=1}^{I} \hat{Z}_{i,t}^\ddagger$, the limit distribution of which is best derived using placebo methods. A test of $H_{0i,t}$ against $H_{1i,t}^- : \alpha_i^{* t} \neq 0$ for some $i$ can be executed using $\hat{Z}_{i,t}^\ddagger = \sum_{i=1}^{I} (\hat{Z}_{i,t}^\ddagger)^2$.

**ACKNOWLEDGEMENTS**

This work was made possible by Award Number 2010-DJ-BX-1672 from the National Institute of Justice, Office of Justice Programs. The authors thank David Powell and Greg Ridgeway for helpful comments. The opinions, findings, and conclusions or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the Department of Justice.
REFERENCES


Figure 1.— Crime levels for the treatment area, its synthetic control, and all of Roanoke (left panels) and the difference in crime levels between the treatment area and its synthetic control with placebo groups plotted in gray (right panels). The dashed red lines indicate the time of the intervention.
Figure 2.— Findings of a sensitivity analysis in which the synthetic control is built by matching only drug crimes and covariates to corresponding pre-intervention levels for the treatment group. Otherwise, plots are analogous to those seen in Figure 1.
Figure 3.— Plots of treatment versus synthetic control for \texttt{i\_any\_crime} when the Synth algorithm is used while matching to all variables (left) and while matching to only \texttt{i\_any\_crime} and covariates (right).