Journal of the American Statistical Association

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

A Framework for Synthetic Control Methods With High-Dimensional, Micro-Level Data: Evaluating a Neighborhood-Specific Crime Intervention

Michael W. Robbins, Jessica Saunders & Beau Kilmer

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2017, VOL. 112, NO. 517, 109–126, Applications and Case Studies
http://dx.doi.org/10.1080/01621459.2016.1213634

Taylor & Francis
Taylor & Francis Group

A Framework for Synthetic Control Methods With High-Dimensional, Micro-Level Data: Evaluating a Neighborhood-Specific Crime Intervention

Michael W. Robbins[a], Jessica Saunders[a,b], and Beau Kilmer[c]

[a]RAND Corporation, Pittsburgh, PA; [b]CNA Institute for Public Research, Arlington, VA; [c]RAND Corporation, Santa Monica, CA

ABSTRACT
The synthetic control method is an increasingly popular tool for analysis of program efficacy. Here, it is 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. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received January 2016
Revised June 2016

KEYWORDS
Calibration; Causal inferences; Crime interventions; Design effect; High-dimensional data; Survey analysis

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. When treatment is randomly assigned within an experimental design, treated cases can be compared to untreated cases, and since those cases are identical to the treated cases (aside from the assignment of treatment), one may conclude that any differences are the consequence of treatment. However, with nonexperimental data, the ability of an analyst to attribute an observed result as being a product of treatment is greatly hindered. For example, 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 aim to assess 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 (Hipple and McGarrell 2009; Kennedy 2009; Braga and Weisburd 2012), statistical analyses of its impact have encountered methodological problems that mainly stem from the lack of an appropriate comparison group (Draca, Machin, and Witt 2011; Corsaro et al. 2012; Saunders et al. 2015). That is, the treated area (in this case an overt drug market) is fundamentally different from other areas of the city.

Attempts to circumvent the shortcomings inherent in observational data typically involve the use of quasi-experimental methods. Specifically, the analyst attempts to determine the (hypothetical) post-intervention state of the treated units in the event that they had, in fact, not been treated. Consider, as an example, difference-in-differences approaches, which have been used previously to evaluate a DMI (e.g., Saunders et al. 2015). Under these techniques, one extrapolates the preintervention 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 and Gardeazabal 2003; Abadie, Diamond, and Hainmueller 2010). 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 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.

CONTACT Michael W. Robbins ✉ mrobbins@rand.org RAND Corporation, 4570 Fifth Ave #600, Pittsburgh, PA 15213.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

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 preintervention time periods. Sensitivity analyses and other comparisons 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.

The Roanoke 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 programs in situations where randomization is not possible.

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 presents application of other methods (e.g., difference-in-differences, the popular `Synth` algorithm, and propensity score techniques) to the Roanoke data and offers discussion of how our proposed method compares (in short, our method appears to provide the most defensible results). Section 5 concludes with discussion that details the advantages and disadvantages of synthetic control methods in our setting and emphasizes the policy implications of our findings regarding crime interventions.

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 $i \in (1, \dots, I)$ and that out of T total time periods measured, there are T_0 time periods measured prior to the intervention, which implies $t \in (1, \dots, T_0, T_0 + 1, \dots, T)$. 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 $j \in (1, \dots, J_0, J_0 + 1, \dots, J)$. Note that we do not have longitudinal measurements of the covariates in $\mathbf{R}_j$.

2.1. *Synthetic Control*

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

$$Y_{ijt} = Y_{ijt}(0) + \alpha_{ijt} D_{jt}, \tag{1}$$

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 α_{ijt}. For our purposes, it is sufficient to consider the effect of treatment when averaged across blocks within the treatment group:

$$\alpha^*_{it} = \frac{1}{J - J_0} \sum_{j=J_0+1}^{J} \alpha_{ijt}, \tag{2}$$

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

Approximation of the quantity in (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, $(w_1, \dots, w_{J_0})$ (with each block in the control group receiving its own nonnegative 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

$$\sum_{j=1}^{J_0} w_j Y_{ijt} = \sum_{j=J_0+1}^{J} Y_{ijt}, \tag{3}$$

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

$$\sum_{j=1}^{J_0} w_j = J - J_0, \tag{4}$$

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

$$\sum_{j=1}^{J_0} w_j \mathbf{R}_j = \sum_{j=J_0+1}^{J} \mathbf{R}_j, \tag{5}$$

which stipulates that the synthetic control has the same covariate values as the aggregated treatment group.

Given weights $(w_1, \dots, w_{J_0})$ that satisfy (3)–(5), the outcome values in the absence of treatment for the treated region at post-intervention time points can be approximated by $\widehat{Y}^*_{it}(0) = \sum_{j=1}^{J_0} w_j Y_{ijt}$. Therefore, we approximate the effect of the intervention across the treated region via

$$\widehat{\alpha}^*_{it} = \frac{1}{J - J_0} \left(\sum_{j=J_0+1}^{J} Y_{ijt} - \sum_{j=1}^{J_0} w_j Y_{ijt} \right), \tag{6}$$

for outcome i at time t where $t > T_0$. By design, $\widehat{\alpha}^*_{it}$ approximates the quantity in (2). Similarly, we suggest

$$\widehat{\alpha}^{**}_i = \frac{1}{T - T_0} \sum_{t=T_0+1}^{T} \widehat{\alpha}^*_{it}, \tag{7}$$

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

To establish validity of the methods outlined herein and to evaluate the potential for bias in (6) and (7) as estimators of the effect of treatment, we extend the framework of Abadie, Diamond, and Hainmueller (2010). Specifically, Abadie, Diamond, and Hainmueller (2010) assumed 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 $\mathbf{Y}_{jt}(0) = (Y_{1jt}(0), \dots, Y_{ijt}(0))'$ denote a length-I vector of outcomes for block j at time t, we assume

$$\mathbf{Y}_{jt}(0) = \boldsymbol{\delta}_t + \boldsymbol{\theta}_t \mathbf{R}_j + \boldsymbol{\lambda}_t \boldsymbol{\mu}_j + \boldsymbol{\varepsilon}_{jt}, \tag{8}$$

where $\boldsymbol{\delta}_t$ is a length-I vector of common factors, $\boldsymbol{\theta}_t$ is an $(I \times r)$ parameter matrix. Further, $\boldsymbol{\lambda}_t$ and $\boldsymbol{\mu}_j$ are $(I \times F)$ and $(F \times 1)$ factor matrices, respectively. Lastly, $\boldsymbol{\varepsilon}_{jt}$ is a length-I vector of transitory shocks with $\mathrm{E}[\boldsymbol{\varepsilon}_{jt}] = 0$ and $\mathrm{var}(\boldsymbol{\varepsilon}_{jt}) = \boldsymbol{\Sigma}_{jt}$.

Combining (1), (2), and (6), we see

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

where we assume that $\{w_j\}$ has been selected to satisfy (3)–(5). Using (8) and applying the arguments of Appendix B of Abadie, Diamond, and Hainmueller (2010), any potential bias in $\widehat{\alpha}^*_{it}$ can now be expressed via

$$\begin{aligned} &\mathrm{E}\left[\sum_{j=J_0+1}^{J} Y_{ijt}(0) - \sum_{j=1}^{J_0} w_j Y_{ijt}(0) \right] \\ &\quad = \mathrm{E}\left[\boldsymbol{\lambda}^*_{it} (\boldsymbol{\Lambda}'\boldsymbol{\Lambda})^{-1} \boldsymbol{\Lambda}' \sum_{j=1}^{J_0} w_j \boldsymbol{\varepsilon}^*_j \right], \end{aligned} \tag{9}$$

for $t > T_0$ where $\boldsymbol{\lambda}^*_{it}$ denotes the ith row of $\boldsymbol{\lambda}_t$, $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}'_1, \dots, \boldsymbol{\lambda}'_{T_0})'$ is a matrix with dimension $IT_0 \times F$, and $\boldsymbol{\varepsilon}^*_j = (\boldsymbol{\varepsilon}'_{j1}, \dots, \boldsymbol{\varepsilon}'_{jT_0})'$ is a length-IT_0 vector. Note that we assume $\boldsymbol{\Lambda}'\boldsymbol{\Lambda}$ is of full rank (which requires $IT_0 \geq F$). It does not necessarily hold that $\mathrm{E}[w_j \boldsymbol{\varepsilon}^*_j] = \mathbf{0}$ (as the weights are, in a sense, functions of the shocks), and therefore it is not clear that $\widehat{\alpha}^*_{it}$ is unbiased. However, a bound may be placed on the potential bias. Specifically, letting σ^2_{ijt} denote the ith diagonal element of $\boldsymbol{\Sigma}_{jt}$ and $\bar{\sigma}^2_i = \max_j \{T_0^{-1} \sum_{t=1}^{T_0} \sigma^2_{ijt}\}$, calculations show that (9) is bounded by a term that is proportional to $[\bar{\sigma}^2_i / (IT_0)]^{1/2}$—the denominator comes from the number of rows of $\boldsymbol{\Lambda}$. Hence, the potential for bias in $\widehat{\alpha}^*_{it}$ is mitigated if the number of outcomes (I), and/or the number of preintervention time periods (T_0) is large.

In addition illustrating the validity of the synthetic control method when data have been generated via a factor model akin to (8), Abadie, Diamond, and Hainmueller (2010) provided theory to validate the performance of the method when the outcome is generated via an autoregressive model. Likewise, we expect that the techniques outlined here will work for autoregressive outcomes; for brevity, details are not shown.

As an alternative to producing a single synthetic control for the combined treatment region, one could instead produce a synthetic control for each individual j block within the treated region. However, such an approach is not preferred (as described further in Section 5.1) due to a lack of efficiency and (often) feasibility.

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 preintervention behavior of the treated unit. Most of the subsequent applications of their procedure (e.g., Abadie, Diamond, and Hainmueller 2010; Billmeier and Nannicini 2013; Cavallo et al. 2013; Abadie, Diamond, and Hainmueller 2014; Bohn, Lofstrom, and Raphael 2014) 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 control weights are derived using calibration techniques (Deville and Särndal 1992; Särndal 2007).

We set

$$\mathbf{X}_j = (1, Y_{1j1}, \ldots, Y_{1jT_0}, Y_{2j1}, \ldots, Y_{2jT_0}, \ldots, Y_{Ij1}, \ldots, Y_{IjT_0}, \mathbf{R}_j')',$$

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

$$\sum_{j=1}^{J_0} w_j \mathbf{X}_j = \mathbf{t}_x. \quad (10)$$

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, Särndal, and Sautory (1993) and is labeled a generalized raking procedure. To calculate weights that may be used to obtain covariate balance in observational studies with binary treatment, Hainmueller (2012) proposed a method in the vein of the calibration procedure discussed here.

Given a distance metric $G(\cdot)$ that satisfies regularity conditions, the calibration 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 (10). To briefly explain in more detail, we solve for the set of weights $\{w_j\}$ and the vector of Lagrange multipliers $\boldsymbol{\xi}$ that minimize the objective function

$$\sum_{j=1}^{J_0} d_j G(w_j/d_j) - \boldsymbol{\xi}' \left(\sum_{j=1}^{J_0} w_j \mathbf{X}_j - \mathbf{t}_x \right).$$

Defining $g(x) = dG(x)/dx$ and $F(u) = g^{-1}(u)$, it follows that $w_j = d_j F(\mathbf{X}_j' \boldsymbol{\xi})$ for each j where $\boldsymbol{\xi}$ satisfies $\sum_{j=1}^{J_0} d_j F(\mathbf{X}_j'\boldsymbol{\xi})\mathbf{X}_j = \mathbf{t}_x$. In practice, Newton's method is used to extract the value of $\boldsymbol{\xi}$ that satisfies the latter formula. In lieu of an informed design, we use $d_j \propto 1$ for the initial weights. To ensure nonnegative weights, we use a truncated linear distance metric with no upper bound. Specifically, we fix $G(x) = (1/2)(x-1)^2$ for $x \geq 0$ and $G(x) = \infty$ otherwise, as proposed by Deville, Särndal, and Sautory (1993). This choice of metric has the added advantage of minimizing the Kish approximation of the design effect (commonly expressed as $n \sum w_i^2 / (\sum w_i)^2$ for samples of size n). For example, if we set $d_j = (J - J_0)/J_0$ (although the specific constant chosen will not influence the results) and assume that (4) is satisfied, we see $\sum_{j=1}^{J_0} d_j G(w_j/d_j) = (J - J_0)\{[J_0 \sum_{j=1}^{J_0} w_j^2 / (\sum_{j=1}^{J_0} w_j)^2] - 1\}/2$ whenever all $w_j \geq 0$. A distance metric akin to $G(x) = x \log(x)$, which is more likely to produce outlying weights than the truncated linear metric suggested here, is used within Hainmueller (2012).

Since numerical methods are only used to solve for $\boldsymbol{\xi}$, which is a vector with dimension equal to that of $\mathbf{t}_x$, 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 (10) (e.g., $\mathbf{t}_x$ does not fall within the convex hull of the $\mathbf{X}_j$ 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 (10). An upper bound can be imposed on the weights via the distance metric $G(x)$; however, bounding in this manner will increase the likelihood that an exact solution of the calibration equations is infeasible.

2.3. Test Statistics

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

$$\mathcal{H}_{0i} : \alpha_i^{**} = 0 \quad \text{against} \quad \mathcal{H}_{1i}^{-} : \alpha_i^{**} < 0, \quad (11)$$

for each outcome i, where $\alpha_i^{**} = \sum_{t=T_0+1}^{T} \alpha_{it}^{*}/(T - T_0)$ 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 supplementary materials. The alternative hypothesis in (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 $\widehat{\alpha}_i^{**}$ from (7) to approximate the treatment effect. However, to evaluate the hypotheses in (11), we must first determine the standard error of $\widehat{\alpha}_i^{**}$. Since 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.

Survey methods for standard error estimation

Survey methodologies can be used to develop approaches for standard error approximations that incorporate the design effect imposed by the weights. These methods mandate a linear model representation for the outcomes. For a specific outcome i, we fit the regression

$$Y_{ijt} = \beta_{it} + a_i^{**} D_{jt} + \epsilon_{ijt}, \quad (12)$$

where β_{it} is a fixed effect for time period t, D_{jt} is the treatment indicator seen in (1), ϵ_{ijt} is mean-zero error, and a_i^{**} 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 $j \in (J_0 + 1, \ldots, J)$.

The model in (12) is used because the estimate of a_i^{**} when found using WLS is equivalent to $\widehat{\alpha}_i^{**}$ in (7); mathematical calculations that illustrate this statement are found in the supplementary materials to this article. Hence, the fit of the regression, when calculated using the appropriate software (we use the function `svyglm()` in the R package `survey`), gives $\widehat{\alpha}_i^{**}$ in addition to $\widehat{\text{var}}(\widehat{\alpha}_i^{**})$, an approximation of its variance. This variance estimator is calculated using Taylor series linearization (Binder 1983).

Using these quantities, a statistic for testing the hypotheses in (11) is

$$\widehat{Z}_i^{**} = \frac{\widehat{\alpha}_i^{**}}{\sqrt{\widehat{\mathrm{var}}(\widehat{\alpha}_i^{**})}} \tag{13}$$

for each outcome. If $\mathcal{H}_{0i}$ and (12) hold, Z_i^{**} can be assumed to have been sampled from a standard normal distribution. We reject $\mathcal{H}_{0i}$ for small values of $\widehat{Z}_i^{**}$. The p-value of a test of the hypotheses in (11) is calculated via $\Phi(\widehat{Z}_i^{**})$, 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. The model in (12) 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 $\widehat{\alpha}_i^{**}$ can be monitored. Our objective is not to produce an accurate estimation of this variability (which would require that (12) hold), but instead to gauge how it is influenced by the structure of the weights. Robustness to the representation in (12) 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. Specifically, we wish to test the omnibus hypotheses

$$\mathcal{H}_0 : \alpha_1^{**} = \cdots = \alpha_I^{**} = 0$$

against

$$\mathcal{H}_1^- : \alpha_i^{**} \leq 0 \text{ for all } i \text{ with } \alpha_i^{**} < 0 \text{ for some } i.$$

Typically, omnibus hypotheses are tested using Wald statistics (or related quantities); however, these are not appropriate for one-sided alternatives. Another option is to aggregate the $\widehat{\alpha}_i^{**}$ for $i \in (1, \ldots, I)$; however, such a statistic would be dominated by components with larger variability. Therefore, we prefer to standardize the $\widehat{\alpha}_i^{**}$ prior to aggregation. That is, letting $\hat{\mathbf{a}} = (\widehat{\alpha}_1^{**}, \ldots, \widehat{\alpha}_I^{**})'$ and $\mathbf{C} = \mathrm{var}(\hat{\mathbf{a}})$, an estimate of which is denoted $\widehat{\mathbf{C}}$, the correctness of $\mathcal{H}_0$ with respect to $\mathcal{H}_1^-$ is evaluated using $(\widehat{\boldsymbol{\sigma}}^{-1/2})'\hat{\mathbf{a}}$, where $\widehat{\boldsymbol{\sigma}}$ is a vector that contains the diagonal of $\widehat{\mathbf{C}}$. (Note that $(\widehat{\boldsymbol{\sigma}}^{-1/2})'\hat{\mathbf{a}} = \sum_{i=1}^{I} \widehat{Z}_i^{**}$.) Further, it holds that $\mathrm{var}\{(\widehat{\boldsymbol{\sigma}}^{-1/2})'\hat{\mathbf{a}}\} = (\widehat{\boldsymbol{\sigma}}^{-1/2})'\widehat{\mathbf{C}}\widehat{\boldsymbol{\sigma}}^{-1/2}$. Consequentially, our test statistic for the omnibus hypotheses is

$$\widehat{Z}^{***} = \frac{(\widehat{\boldsymbol{\sigma}}^{-1/2})'\hat{\mathbf{a}}}{\sqrt{(\widehat{\boldsymbol{\sigma}}^{-1/2})'\widehat{\mathbf{C}}\widehat{\boldsymbol{\sigma}}^{-1/2}}}, \tag{14}$$

which is assumed to follow a standard normal distribution under $\mathcal{H}_0$. Statistics akin to $\widehat{Z}^{***}$ have been proposed previously for one-sided omnibus tests (e.g., Lachin 2014). Since, to our knowledge, no analytic expressions exist that may be used to approximate $\mathbf{C}$, our estimate $\widehat{\mathbf{C}}$ is calculated with a jackknife. We reject $\mathcal{H}_1^-$ for small values of $\widehat{Z}^{***}$.

Although not our focus here, we suggest the Wald statistic $\widehat{Z}^{\pm} = \hat{\mathbf{a}}'\widehat{\mathbf{C}}^{-1}\hat{\mathbf{a}}$, which has a χ^2 distribution with I degrees of freedom under $\mathcal{H}_0$, to address the two-sided alternative $\mathcal{H}_1^{\pm} : \alpha_i^{**} \neq 0$ for some i. Thus, $\mathcal{H}_1^{\pm}$ is rejected for large values of $\widehat{Z}^{\pm}$.

2.4. Generation of Placebo Groups Through Permutation

We have presented statistics for testing the hypotheses in (11) and have given algebraic expressions to approximate the sampling distribution of the treatment effect estimator of (7). However, these approximations are based upon restrictive model assumptions (e.g., the formulation in (12)) 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. 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, Diamond, and Hainmueller (2010) employed a placebo study wherein each of the (untreated) comparison areas that were available for construction of a synthetic control is used as a region. Placebo tests (a.k.a., falsification or refutability tests) have a rich history within the literature on program evaluation (e.g., DiNardo and Pischke 1997; Angrist and Krueger 1999; Auld and Grootendorst 2004). 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, Diamond, and Hainmueller (2010) is limited by the number of available comparison groups; they study between 19 and 38 valid placebo cases. This number is 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 (practically) 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 did hypothetically receive the intervention. Further, the observations for outcome i at time t within the kth reordering are given by the vector $(Y_{i1t}^{(k)}, \ldots, Y_{ijt}^{(k)})'$, where $(Y_{i(J_0+1)t}^{(k)}, \ldots, Y_{ijt}^{(k)})'$ corresponds to the placebo treatment region. Then, a vector of weights $(w_1^{(k)}, \ldots, w_{J_0}^{(k)})'$ is calculated (using the methods described earlier), which satisfies versions of (3)–(5) that use the $Y_{ijt}^{(k)}$ in place of Y_{ijt}.

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 kth placebo region and its respective synthetic control with weights $\{w_j^{(k)}\}$. If K total placebo regions have been sampled, in theory

$(Z^{(1)}, Z^{(2)} \dots, Z^{(K)})$ will encompass the sampling distribution of Z for sufficiently large K. Therefore, the p-value for Z may be given by

$$p = \frac{\{\#k : Z^{(k)} < Z\}}{K}. \tag{15}$$

This formula can be used to derive p-values for the statistics in Section 2.3 (by using, e.g., $Z = \widehat{Z}_i^{**}$ or $\widehat{Z}^{***}$).

One could use placebo groups to directly ascertain the sampling distribution of a treatment effect prior to any standardization (e.g., using $Z = \widehat{\alpha}_i^{**}$), which would circumvent the need for the model assumptions implicit in (12). 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 $\widehat{\alpha}_i^{**}$ using (13) prior to calculating a permuted p-value via (15). Further, deriving the distribution of $\widehat{Z}_i^{**}$ by using permuted placebo groups is expected to make the results robust to the assumptions required by (12). To summarize, we calculate p-values based on permuted placebo groups while using $\widehat{Z}_i^{**}$ instead of $\widehat{\alpha}_i^{**}$ so as to filter out the design effect yielded by the synthetic control weights.

2.5. *Confidence Intervals*

Rather than reporting the raw intervention effect $\widehat{\alpha}_i^{**}$ from (7), in practice it is often more informative to report the treatment effect as a percent change in the observed value of the outcome from the hypothetical outcome value that would have been observed in the absence of treatment. That is, we calculate the ratio

$$\widehat{\Delta}_i^{**} = \frac{\sum_{t=T_0+1}^{T} \left(\sum_{j=J_0+1}^{J} Y_{ijt} - \sum_{j=1}^{J_0} w_j Y_{ijt} \right)}{\sum_{t=T_0+1}^{T} \sum_{j=1}^{J_0} w_j Y_{ijt}}$$

for each outcome i, which is expressed in percentage terms when multiplied by 100. Letting $\widehat{\beta}_{it}$ denote the WLS estimate of β_{it} from (12), calculations (see the supplementary materials) show that

$$\bar{\beta}_i := \frac{1}{T - T_0} \sum_{t=T_0+1}^{T} \widehat{\beta}_{it} = \frac{1}{(J - J_0)(T - T_0)} \sum_{t=T_0+1}^{T} \sum_{j=1}^{J_0} w_j Y_{ijt}.$$

Using this and the previously stated fact that the WLS estimate of a_i^{**} in (12) is equivalent to $\widehat{\alpha}_i^{**}$, we see that

$$\widehat{\Delta}_i^{**} = \widehat{\alpha}_i^{**} / \bar{\beta}_i.$$

Our interest is in deriving a $100(1 - \gamma)\%$ confidence interval for $\widehat{\Delta}_i^{**}$. To ensure that the lower bound of this interval is not less than -1, we derive a confidence interval for $\hat{\delta}_i^{**} = \log(\widehat{\Delta}_i^{**} + 1)$ and then transform the bounds to obtain an interval for $\widehat{\Delta}_i^{**}$. Note that, as a consequence of the log transformation, the interval for $\widehat{\Delta}_i^{**}$ will not be symmetric.

First, we present an approach that uses normal approximations derived from fit of the model in (12). Specifically, Taylor series linearization (i.e., the multivariate delta method) is used to approximate the variance of $\hat{\delta}_i^{**}$. Letting $f(\alpha, \beta_1, \dots, \beta_m) = \log(\alpha/\bar{\beta} + 1)$ where $\bar{\beta} = \sum_{t=1}^{m} \beta_t / m$, it holds that

$$\begin{aligned} &\nabla f(\alpha, \beta_1, \dots, \beta_m) \\ &\quad = \left(\frac{1}{\alpha + \bar{\beta}}, -\frac{\alpha}{m\bar{\beta}(\alpha + \bar{\beta})}, \dots, -\frac{\alpha}{m\bar{\beta}(\alpha + \bar{\beta})} \right). \end{aligned}$$

Define $\widehat{\boldsymbol{B}}_i = \nabla f(\widehat{\alpha}_i^{**}, \widehat{\beta}_{i1}, \dots, \widehat{\beta}_{im})$. If we let $\widehat{\boldsymbol{\Sigma}}_i = \widehat{\text{var}}\{(\widehat{\alpha}_i^{**}, \widehat{\beta}_{i1}, \dots, \widehat{\beta}_{im})'\}$, which is provided with the output from `svyglm()` when (12) is fit, the variance of $\hat{\delta}_i^{**}$ is approximated using

$$\hat{\upsilon}_i := \widehat{\text{var}}(\hat{\delta}_i^{**}) = \widehat{\boldsymbol{B}}_i \widehat{\boldsymbol{\Sigma}}_i \widehat{\boldsymbol{B}}_i'.$$

A $100(1 - \gamma)\%$ confidence interval for $\hat{\delta}_i^{**}$ is given by $(\hat{\delta}_i^{**} - z_{1-\gamma/2}\hat{\upsilon}_i, \hat{\delta}_i^{**} - z_{\gamma/2}\hat{\upsilon}_i)$ where $z_\nu = \Phi^{-1}(\nu)$. Furthermore, an analogous confidence interval for $\widehat{\Delta}_i^{**}$ is given by $\{\kappa_{i,1-\gamma/2}(\widehat{\Delta}_i^{**} + 1) - 1, \kappa_{i,\gamma/2}(\widehat{\Delta}_i^{**} + 1) - 1\}$, where $\kappa_{i,\nu} = \exp(-z_\nu \hat{\upsilon}_i)$.

The above theory can be modified to develop confidence intervals that are calculated using the permutation groups described in Section 2.4. Let $\widehat{\Delta}_i^{(k)}$, $\hat{\delta}_i^{(k)}$, and $\hat{\upsilon}_i^{(k)}$ denote the values of $\widehat{\Delta}_i^{**}$, $\hat{\delta}_i^{**}$, and $\hat{\upsilon}_i$, respectively, found using the kth placebo grouping. Letting $\delta_i^{**} = E[\hat{\delta}_i^{**}]$, we assume that the set $\boldsymbol{\Xi}_i = (z_i^{(1)}, \dots, z_i^{(K)})$, where $z_i^{(k)} = \hat{\delta}_i^{(k)} / \hat{\upsilon}_i^{(k)}$, gives a reasonable scope of the sampling distribution of $z_i^{**} = (\hat{\delta}_i^{**} - \delta_i^{**}) / \hat{\upsilon}_i^{**}$ for general values of δ_i^{**} (since under H_0, $\delta_i^{**} = 0$). Therefore, letting $\tilde{z}_\nu$ denote the 100νth percentile of the empirical distribution given by the set $\boldsymbol{\Xi}_i$, we derive a $100(1 - \gamma)\%$ confidence interval for $\widehat{\Delta}_i^{**}$ via $\{\tilde{\kappa}_{i,1-\gamma/2}(\widehat{\Delta}_i^{**} + 1) - 1, \tilde{\kappa}_{i,\gamma/2}(\widehat{\Delta}_i^{**} + 1) - 1\}$, where $\tilde{\kappa}_{i,\nu} = \exp(-\tilde{z}_\nu \hat{\upsilon}_i)$.

3. Roanoke Crime Intervention

In this section, we consider the efficacy of an intervention to close overt drug markets, which bring 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 (Reuter and MacCoun 1992; Harocopos and Haugh 2005). 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 incarcerated, 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, Soole, and Rombouts 2006).

The Drug Market Intervention (DMI, see, Kennedy and Wong 2009; McGarrell, Corsaro, and Brunson 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 (e.g., Kennedy, Piehl, and

Braga 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, the team identifies sellers involved in the overt market, and then the police make undercover purchases from all of the identified dealers to build credible cases to prosecute the offenders. 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. Concurrently, the community is encouraged to take back their neighborhoods and prevent the reemergence of the drug market.

The market disruption and reduction in associated crime and disorder that stem from the DMI are maintained by stronger neighborhood institutions and more positive police-community relations and cooperation (Kennedy 2009; Saunders et al. 2016). That is, the DMI program is grounded in procedural justice—law enforcement clearly articulates that they are committed to community safety and not primarily there to incarcerate community members, and then follow through with their promises to help improve neighborhood quality of life (Hipple and McGarrell 2009; NNSC 2015). In summary, the program seeks to reduce crime and disorder through multiple mechanisms: (1) Incapacitating violent drug dealers; (2) Establishing a clear and credible deterrence threat against all nonviolent drug dealers (and their potential replacements) to stop all market activity at once; (3) Improving police effectiveness and response while building positive relationships with the community to elicit their help in preventing market reemergence; and (4) Building community resiliency that can effectively exert positive community social norms (Saunders et al. 2016).

Peer-reviewed evaluations of DMI find that it can reduce crime and drug activity around the target markets in some locations; however, the methods that have been used to estimate program impact have suffered from limitations preventing strong causal inference. The first DMI in High Point, North Carolina has been evaluated multiple times using several strong quasi-experimental methods (Corsaro et al. 2012; Saunders et al. 2015); however, the rest of the sites have not been rigorously evaluated. While DMI has been implemented across dozens of sites, there are only publicly available evaluations for five of them, which generally support its effectiveness. The DMI in Nashville, TN was found to reduce crime using time series analysis (Corsaro, Brunson, and McGarrell 2013); three additional sites experienced crime reductions associated with DMI, although the methods do not allow for any causal attribution (Winston-Salem, NC: Frabutt et al. 2009; Providence, RI: Kennedy and Wong 2009, and Rockford, IL: Corsaro and McGarrell 2009); and it was not associated with any changes in crime in Peoria, IL (Corsaro and Brunson 2013).

This study specifically focuses on the DMI that was implemented in the Hurt Park neighborhood of Roanoke, Virginia in late 2011. The Roanoke DMI took place in the Hurt Park neighborhood, an area of 2785 residents that are predominantly African-American with 80% 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 intervention team, comprised members from law enforcement, prosecution, social services, and the community, identified 15 active drug dealers, and determined that 10 were violent and should be arrested and prosecuted, while 5 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, the police increased their presence in the neighborhood and held numerous meetings and events to address the negative narrative between law enforcement and the community. Community members reported that police community relations were improved and the DMI shut down the overt market (Saunders et al. 2016). An in-depth description of the Roanoke DMI is available an evaluation of implementation fidelity in Saunders et al. (2016).

3.1. Data

We have longitudinal data measurements at the census block level taken at quarterly intervals that extend from 3 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 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 Reporting procedure (Lejins 1966; Gove, Hughes, and Geerken 1985). 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 aggravated 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, Hughes, and Geerken 1985).

The aggregated outcomes monitored here include `i_any_crime`, a sum of all reported crimes (both Part I

Table 1. Variables used in creation of synthetic control weights. Crime counts (the last two columns) are aggregated across the three-year preintervention period.

	Measure	Frequency	Description	Hurt Park	All blocks (scaled)
Outcomes	`i_rape`	36-month	Nº rapes reported	8	2.63
	`i_robbery`	18-month	Nº robberies reported	13	5.03
	`i_aggassault`	6-month	Nº aggravated assault reported	25	10.25
	`i_burglary`	3-month	Nº burglaries reported	62	25.88
	`i_larceny`	3-month	Nº larcenies reported	107	99.96
	`i_cartheft`	3-month	Nº car thefts reported	11	8.81
	`i_arson`	12-month	Nº arsons reported	5	1.11
	`i_simpassault`	3-month	Nº simple assaults reported	187	83.14
	`i_drugs`	3-month	Nº drug crimes reported	170	38.94
	`i_weapons`	12-month	Nº weapons crimes reported	36	6.72
Aggregated outcomes	`i_violent`	3-month	Total nº violent crimes reported	238	102.4
	`i_property`	3-month	Total nº property crimes reported	180	134.7
	`i_any_crime`	3-month	Total nº of crimes reported	1599	793.1
Covariates	`TotalPop`	Baseline	Nº of residents	910	910.0
	`black`	Baseline	Nº African American residents	684	259.0
	`hispanic`	Baseline	Nº Hispanic residents	56	50.13
	`white`	Baseline	Nº white residents	123	563.1
	`males_1521`	Baseline	Nº male residents aged 15-21	62	36.51
	`households`	Baseline	Nº households	351	400.6
	`vacant_units`	Baseline	Nº vacant housing units	73	44.46
	`female`	Baseline	Nº female headed households	148	69.25
	`renter`	Baseline	Nº households occupied by renters	226	181.1

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 aside from the individual crimes listed in Table 1; 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 1 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 $T_0 = 12$ preintervention time periods as well as covariates and the intercept, the synthetic control is created by matching across 129 variables.

3.2. Main Results

We applied the methodology introduced in Section 2.2 to create synthetic control weights for the Hurt Park treatment region in addition to 1000 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 23 of the 1000 placebo regions (and does not occur at all if the dimensionality of the constraint vector $\mathbf{t}_x$ is reduced slightly); these 23 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 3.18, and 18 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 (10) is not preferable to the large design effect that comes from unbounded weights. Further, of the 3535 blocks assigned a weight, only 119 are given a weight with a value greater than 0.

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 `i_any_crime`, `i_drugs` and `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 1) 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 preintervention levels of `i_any_crime` and `i_drugs` (this is the case for all variables listed in Table 1 that are used at 3-month frequencies). An exact match between treatment and synthetic control in each quarter is not obtained for `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 1 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

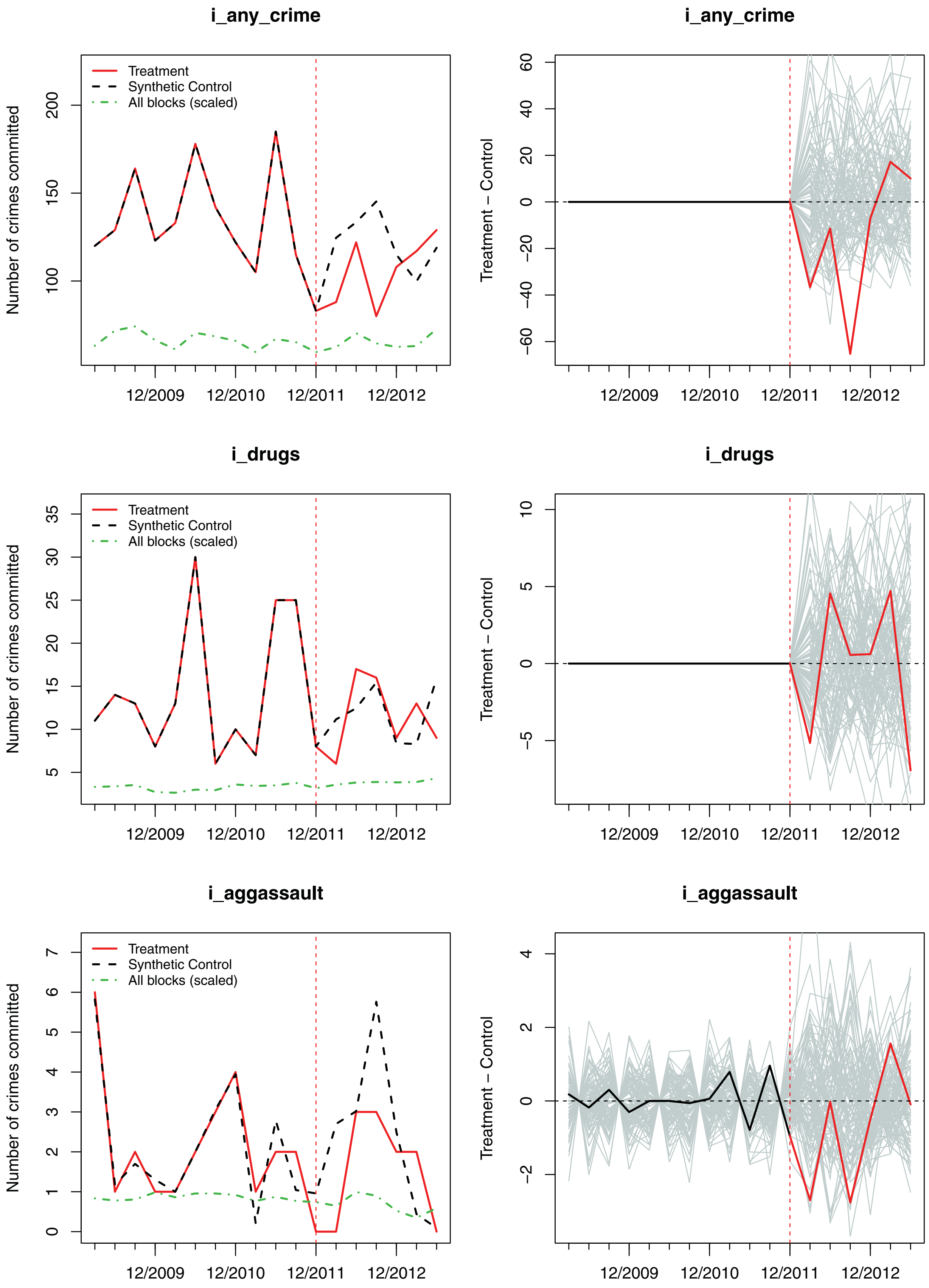

Figure 1. Crime levels for the treatment area, its synthetic control, and all of Roanoke (left panels) and the difference in crime levels between the treatment area and its synthetic control with placebo groups plotted in gray (right panels). The dashed red lines indicate the time of the intervention.

Table 2. Estimates of the effect of the drug market intervention in Roanoke, VA, and *p*-values (the final three columns) of tests for the presence of an intervention effect.

		No. of crimes			90% Confidence interval		$\widehat{\alpha}_i^{**}$	Standardized	
		Hurt Park	Synthetic Hurt Park	Percent change	Norm.	Perm.	Perm.	Norm.	Perm.
6 months post intervention	i_rape	0	0.44	−100.0	(−100.0, N/A)	(−100.0, N/A)	0.350	0.081	0.394
	i_robbery	1	0.84	18.4	(−83.1, 731.3)	(−86.1, 163.9)	0.621	0.555	0.617
	i_aggassault	3	5.72	−47.5	(−81.7, 50.2)	(−83.1, 11.0)	0.075	0.129	0.179
	i_burglary	6	18.40	−67.4	(−87.7, −13.3)	(−85.5, −15.4)	0.001	0.048	0.083
	i_larceny	19	26.92	−29.4	(−56.8, 15.4)	(−52.3, 8.8)	0.097	0.122	0.124
	i_cartheft	1	2.15	−53.6	(−92.7, 194.9)	(−93.8, 20.0)	0.214	0.222	0.268
	i_arson	2	0.03	5758	(679.5, 43920)	(384.2, 5758)	0.988	0.919	0.988
	i_simpassault	23	27.14	−15.3	(−47.9, 37.9)	(−46.0, 29.0)	0.240	0.285	0.240
	i_drugs	23	23.60	−2.5	(−49.6, 88.5)	(−51.1, 97.5)	0.477	0.474	0.496
	i_weapons	1	4.51	−77.8	(−96.2, 28.6)	(−97.2, −54.8)	0.000	0.040	0.202
	i_violent	29	34.18	−15.1	(−47.0, 35.8)	(−45.4, 23.7)	0.209	0.279	0.227
	i_property	26	47.47	−45.2	(−65.7, -12.4)	(−62.5, −19.2)	0.002	0.028	0.028
	i_any_crime	210	258.01	−18.6	(−37.5, 5.9)	(−31.3, −1.3)	0.056	0.099	0.052
	Omnibus	—	—	—	—	—	0.005	0.066	0.278
12 months post intervention	i_rape	2	2.08	−3.7	(−84.7, 504.8)	(−89.6, 107.9)	0.470	0.486	0.480
	i_robbery	2	3.20	−37.4	(−84.6, 154.3)	(−86.1, 64.5)	0.248	0.283	0.339
	i_aggassault	8	13.98	−42.8	(−72.6, 19.7)	(−75.0, 6.3)	0.008	0.109	0.150
	i_burglary	11	30.79	−64.3	(−82.4, −27.3)	(−81.1, −24.6)	0.000	0.012	0.038
	i_larceny	30	51.56	−41.8	(−61.5, −12.2)	(−60.1, −12.9)	0.014	0.020	0.035
	i_cartheft	3	6.00	−50.0	(−83.1, 48.6)	(−84.8, 33.4)	0.129	0.126	0.205
	i_arson	2	2.06	−3.1	(−81.4, 404.5)	(−87.1, 27.2)	0.565	0.488	0.578
	i_simpassault	46	53.17	−13.5	(−39.4, 23.5)	(−41.0, 22.2)	0.196	0.249	0.223
	i_drugs	48	47.43	1.2	(−34.0, 55.2)	(−37.4, 60.3)	0.507	0.518	0.493
	i_weapons	2	6.73	−70.3	(−91.7, 6.0)	(−94.2, −29.1)	0.005	0.033	0.083
	i_violent	62	74.48	−16.8	(−39.7, 14.9)	(−40.9, 12.7)	0.104	0.170	0.149
	i_property	44	88.35	−50.2	(−65.0, −29.2)	(−63.4, −31.0)	0.001	0.001	0.005
	i_any_crime	398	518.31	−23.2	(−35.6, −8.5)	(−32.5, −11.8)	0.004	0.006	0.001
	Omnibus	—	—	—	—	—	0.001	0.002	0.044
18 months post intervention	i_rape	2	5.08	−60.6	(−94.3, 172.7)	(−95.8, 32.8)	0.028	0.205	0.266
	i_robbery	3	5.05	−40.6	(−81.4, 89.9)	(−83.2, 52.6)	0.166	0.224	0.282
	i_aggassault	10	14.51	−31.1	(−65.2, 36.6)	(−69.3, 30.5)	0.062	0.187	0.211
	i_burglary	15	36.56	−59.0	(−77.5, −25.4)	(−75.1, −19.2)	0.001	0.010	0.041
	i_larceny	45	67.75	−33.6	(−53.9, −4.4)	(−53.5, −1.7)	0.048	0.034	0.055
	i_cartheft	4	7.14	−44.0	(−78.3, 44.2)	(−77.9, 36.2)	0.139	0.136	0.185
	i_arson	3	2.06	45.4	(−67.8, 556.2)	(−76.4, 103.8)	0.825	0.660	0.706
	i_simpassault	73	73.88	−1.2	(−26.7, 33.1)	(−30.0, 34.7)	0.449	0.474	0.451
	i_drugs	70	71.65	−2.3	(−31.0, 38.3)	(−35.9, 44.2)	0.435	0.456	0.448
	i_weapons	4	8.04	−50.3	(−80.5, 26.9)	(−84.5, 14.7)	0.054	0.088	0.145
	i_violent	93	100.60	−7.6	(−29.5, 21.3)	(−31.7, 21.4)	0.275	0.316	0.282
	i_property	64	111.46	−42.6	(−57.9, −21.6)	(−57.1, −20.5)	0.002	0.002	0.002
	i_any_crime	644	737.00	−12.6	(−24.6, 1.2)	(−22.9, −0.0)	0.080	0.065	0.062
	Omnibus	—	—	—	—	—	0.011	0.000	0.026

within WLS (as described in Section 2.3) for the treatment region have a design effect of 30.4, whereas the median corresponding design effect for the placebo regions is 15.8 (with a 90th percentile of 18.2). The large magnitude of all of these design effects is due to an abundance of weights being calculated as zero. Furthermore, the synthetic control estimator $\widehat{\alpha}_i^{**}$ likely has more variability when calculated for the treatment region than when calculated for the placebo 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.

For each outcome, we report the estimated effect of the intervention as a percent change (as described in Section 2.5)—these figures are computed by comparing post-DMI crime levels with the corresponding levels observed within the synthetic control. We provide 90% confidence intervals for the estimated percent changes that are found using normal approximations (Norm.) and permuted placebo groups (Perm.). In theory, if the upper bound of the 90% confidence interval is less than zero, one should reject the hypothesis that there is no intervention effect in favor of the hypothesis that the intervention reduced crime levels (a lower-tailed, one-sided hypothesis test) at the 5% significance level. To further inform our inferences, we calculate *p*-values in three different manners to assess the effect of the intervention on pertinent outcomes. Specifically, using the raw value of $\widehat{\alpha}_i^{**}$ as our statistic, we report a *p*-value found with the permuted placebo groups (Perm.). Likewise, using the version of $\widehat{\alpha}_i^{**}$ when standardized with the survey methodological approach of (13), we report *p*-values found using a standard normal sampling distribution (Norm.) and permuted placebo groups (Perm.). In each case, the omnibus test is calculated across each of the nonaggregated outcomes listed in Table 1. Results are provided in Table 2 for each outcome and for several choices of T, the maximum time period considered. For context, the table also provides the number of crimes observed in the treatment area (Hurt Park) and its synthetic control.

It appears that the normal approximation and the permutation method yield similar approximations of the sampling distribution of the standardized synthetic control estimator (i.e., compare the *p*-values for "Norm." and "Perm." under the standardized approach). However, there are some concerns that the normal approximation may not work well for crimes that are not prevalent (e.g., i_rape, i_robbery, i_cartheft, i_arson and i_weapons)—this is not surprising given that the validity of normal approximations for count data tend to be dependent upon the expected magnitude of the counts (as opposed to just sample size, e.g.). For the nonprevalent outcomes, the p-value found using the normal approximation tends to be smaller than the p-value found with placebo methods, although this difference is negligible in some cases. This issue appears to transfuse into the omnibus statistic (the components of which are several of the nonprevalent crime outcomes). Similarly, the confidence intervals shown in Table 2 are usually concordant with the respective p-values (i.e., when zero lies outside the interval, it is usually the case that the respective p-value is less than 0.05), but lack of consistency in this respect seems to occur only when outcomes have low numbers of crimes. Likewise, the two methods of calculating intervals only provide differing results for outcomes with low counts. In cases where intervals and p-values are not in agreement, the hypothesis tests should be used for inference as these are directly designed to assess the inferential question of interest (i.e., is there an effect of the intervention?) and require minimal assumptions.

Continuing, the p-values for the standardized synthetic control estimator are noticeably larger (although still statistically significant at the 5% level for many outcomes) than those yielded by the raw treatment effect ($\widehat{\alpha}_i^{**}$). In summary, the standardized synthetic control estimator (which is 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. Comparisons of the procedure outlined here to a wider array of existing techniques are provided in Section 4.

We next discuss substantive findings. From Table 2, we see that when the treatment region is compared to its synthetic control, nearly all of the outcomes observe a decrease in levels following the intervention. Specifically, the percent change for i_any_crime was −18.6% with a p-value of 0.052 under our preferred specification at 6 months following the intervention. These results are largely driven by a reduction in property crimes (−45.2%; $p = 0.028$), not violent crimes (−15.1%, $p = 0.227$). Perhaps surprisingly, none of the specifications identify an effect on drug crimes (−2.5%, $p = 0.496$). While it may be the case that intervention really did not reduce drug market activity in Hurt Park, it could also be the case that the intervention 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 2 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 (−50.2%, $p = 0.005$) and any crimes (−23.2%, $p = 0.001$), 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: −42.6%, $p = 0.002$; any crimes: −12.6%, $p = 0.062$). This suggests that some of the DMI effect may dissipate over time.

Furthermore, each panel in Table 2 presents the p-value for an omnibus test, which tells us whether the intervention had some sort of effect across all the nonaggregated measures we test (i.e., excluding i_violent, i_property, and i_any_crime). Using our preferred specification, the p-value is not statistically significant at 6 months (0.278), but it is marginally significant at 12 months (0.044) and at 18 months (0.026). Unlike the any_crime measure, which is driven largely 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 the vein of several extant studies (e.g., Reppetto 1976; Clarke and Weisburd 1994; Telep et al. 2014), we performed analyses that assessed the possibility that the DMI in Hurt Park displaced crimes to different regions and/or observed a diffusion of crime control benefits to areas adjacent to Hurt Park. Complete details are found in the supplementary materials to this article. In summary, we find some evidence of diffusion of benefits to adjacent areas, as well as potential displacement of drug crime (but not other types of crimes) to a separate overt drug market.

3.3. Sensitivity Analyses

To assess the importance of multivariate modeling of outcomes, 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 27.5% (with a p-values of 0.188 and 0.237 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 31.9% ($p = 0.054$ and 0.134), and at 18 months post-intervention, the drop is 37.4% ($p = 0.010$ and 0.063). Note that the main analyses never indicated more than 2.5% ($p = 0.496$) decrease in drug crimes.

Figure 2 provides analogs 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 preintervention 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

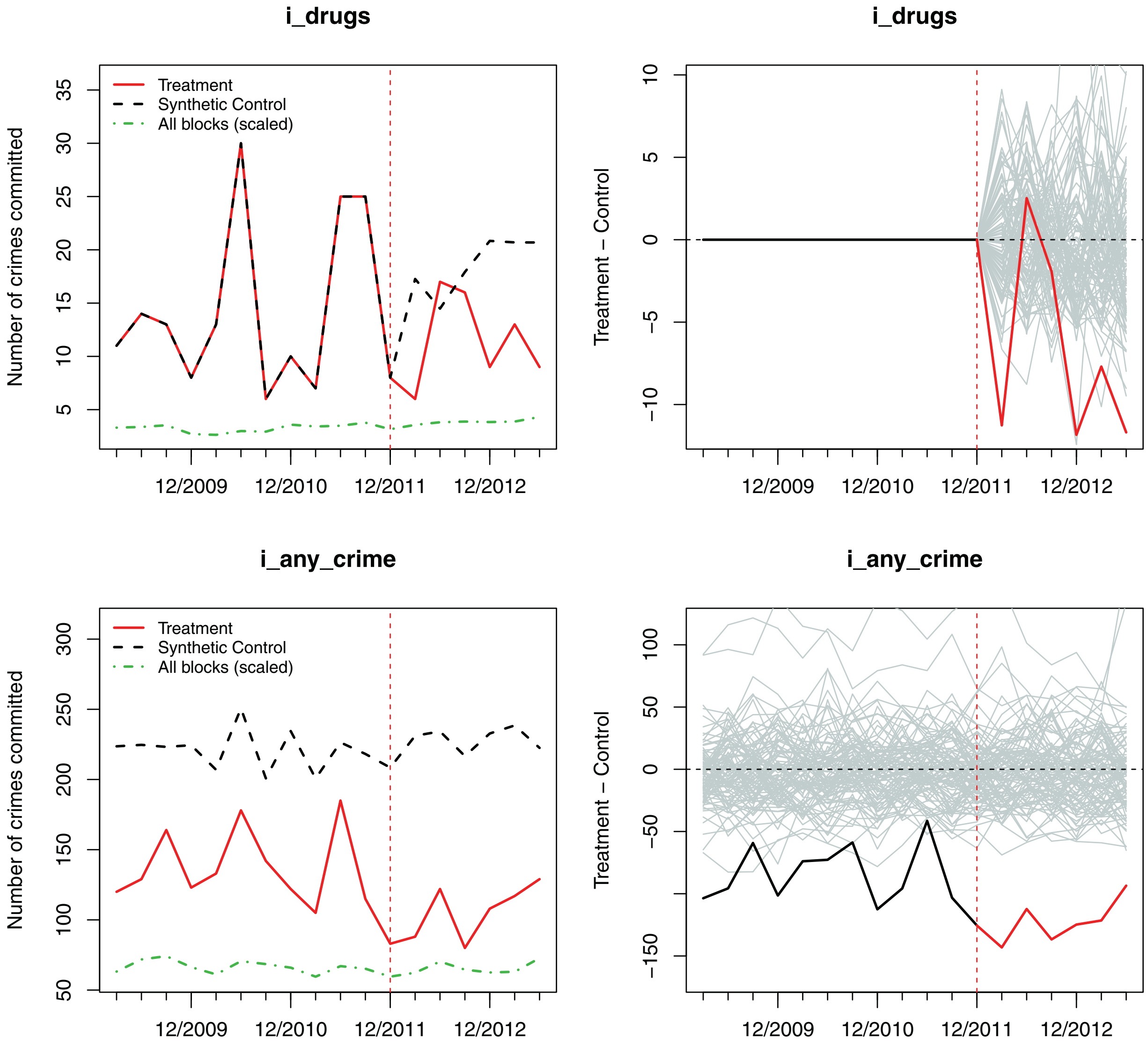

Figure 2. Findings of a sensitivity analysis in which the synthetic control is built by matching only drug crimes and covariates to corresponding preintervention levels for the treatment group. Otherwise, plots are analogous to those seen in Figure 1.

for estimating the effect of the intervention on drug crime levels. To summarize, this analysis illustrates the superiority of a synthetic control region that is built using a wide array of outcomes.

4. Comparison of Methods

In this section, we compare the sensitivity of our findings regarding the efficacy of the Roanoke DMI to the choice of method used. We consider a variety of extant methods, which are outlined as follows.

4.1. *Difference-in-Differences*

As our first comparison, 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. (2015).

The DiD specification employed herein, which is similar in construction to (12), models the mean sequence for outcome i via

$$\log(\mu_{ijt}) = \beta_{it} + \gamma_{ij} + \alpha_{i,\text{did}} D_{jt} + \boldsymbol{\eta}_i' \mathbf{R}_j^*, \qquad (16)$$

where $j \in (1, \dots, J)$ and $t \in (1, \dots, T)$. Further, $\mu_{ijt} = \mathrm{E}[Y_{ijt}]$, the β_{it} are fixed effects for each time period, the γ_{ij} are block-level fixed effects, and D_{jt} is the treatment indicator seen in (1). The model is fit separately for each outcome i. Additionally, $\mathbf{R}_j^* = \log(\mathbf{R}_j + 1)$—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 to ensure a computationally feasible transformation)—and $\boldsymbol{\eta}_i$ is a vector of corresponding regression coefficients. Lastly, $\alpha_{i,\text{did}}$ is the coefficient of interest, which gives the treatment effect. When standardized, the estimate of $\alpha_{i,\text{did}}$ is assumed to have approximately a standard normal distribution. As an omnibus statistic, we use $\sum_{i=1}^{I} \{\widehat{\alpha}_{i,\text{did}} / \text{var}(\widehat{\alpha}_{i,\text{did}})^{-1/2}\}$, the sampling distribution of which is approximated using the aforementioned permutation methods. The treatment effect may be expressed as a percent change by calculating $100\{\exp(\widehat{\alpha}_{i,\text{did}}) - 1\}$; this gives the percent change in the expected outcome of a treatment unit from its hypothetical expectation had it instead been a control case.

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. (2015) for further description of analogous DiD models. Other discrepancies between (12) and (16) are explained thusly. Since the preintervention levels of outcomes and covariates are matched precisely between the treatment area and its synthetic control, the model in (12) does not need to incorporate data observed prior to the intervention, and likewise block-level fixed effects are unnecessary in (12). Further, least squares is used when fitting (12) so as to ensure that the estimated treatment effect equals $\widehat{\alpha}_i^{**}$ as expressed in (7). Distributional misspecification within (12) is handled through the use of permutation in calculation of standard errors.

Our setting enables more precise estimation of treatment effects via DiD approaches than settings that use a single treated unit (e.g., Abadie, Diamond, and Hainmueller 2010; Billmeier and Nannicini 2013; Cavallo et al. 2013). However, it has been observed that DiD models underestimate standard error (Bertrand, Duflo, and Mullainathan 2004). Synthetic control methods enable less restrictive assumptions due to underlying models akin to (8). For instance, the parallel trend assumption used by difference-in-differences models effectively imposes that $\boldsymbol{\lambda}_1 = \cdots = \boldsymbol{\lambda}_T$ within (8).

4.2. `Synth`

Next, we aggregate 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, Diamond, and Hainmueller (2010). Thus, the `Synth` algorithm (Abadie, Diamond, and Hainmueller 2011) may be used 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). The synthetic control that is created with the weights from `Synth` is used to estimate the treatment effect for each outcome; an omnibus statistic is also calculated by summing the estimates for each outcome. Significance levels for each outcome's treatment effect estimator and for the omnibus statistic are found using placebo methods.

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 to use per capita data (however, many of these blocks had reported crimes, which means it would be imprudent to drop them).

We do not offer comparisons to `Synth` for the purpose of assessing its efficacy. `Synth` was not designed for our setting (micro-level data) and is considered the gold-standard for the setting in which it was purposed. It is studied here so that the comparisons provided may illustrate the advantages of using a more granular level of data when possible.

4.3. Propensity Scores

Propensity scores (Rosenbaum and Rubin 1983), which are commonly used to balance treatment and control groups across a variety of characteristics, also have utility in our setting. Letting $\mathbf{Y}_j^*(0) = (\mathbf{Y}_{j1}(0)', \ldots, \mathbf{Y}_{jT_0}(0)')'$ denote a length-IT_0 vector of preintervention outcomes, where $\mathbf{Y}_{jt}(0)$ is as defined within (8), propensity scores are defined here as $\pi_j = P(j \in \mathcal{T} | \mathbf{Y}_j^*(0), \mathbf{R}_j)$, for $j \in (1, \ldots, J)$, where $\mathcal{T} = (J_0 + 1, \ldots, J)$ denotes the set of blocks contained in the treatment region. Since the objective is to weight the control cases so that, when aggregated, they resemble the aggregated treatment cases, we use average treatment effect on the treated weights (see, e.g., Imbens 2004). These are calculated by setting $w_j = 1$ for $j \in \mathcal{T}$ and $w_j = \hat{\pi}_j/(1 - \hat{\pi}_j)$ for $j \notin \mathcal{T}$, where $\{\hat{\pi}_j\}$ denotes an estimated version of $\{\pi_j\}$.

Propensity score methodologies require an ignorability assumption; in our context, this states that $f(\mathbf{Y}_{jt}(0)|\mathbf{Y}_j^*(0), \mathbf{R}_j, j \in \mathcal{T}) = f(\mathbf{Y}_{jt}(0)|\mathbf{Y}_j^*(0), \mathbf{R}_j)$ for each $t > T_0$, where $f(\cdot)$ is a density function. Although such an assumption is perhaps more congenial with the autoregressive model posited by Abadie et al. (2010) (wherein the synthetic control method was also shown to be valid), one may show that the model in (8) satisfies the ignorability assumption. From (8), we see it holds that

$$\mathbf{Y}_j^*(0) = \boldsymbol{\Delta} + \boldsymbol{\Theta}\mathbf{R}_j + \boldsymbol{\Lambda}\boldsymbol{\mu}_j + \boldsymbol{\varepsilon}_j^*,$$

where $\boldsymbol{\Delta} = (\boldsymbol{\delta}_1', \ldots, \boldsymbol{\delta}_{T_0}')'$ is a length-IT_0 vector, $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1', \ldots, \boldsymbol{\Theta}_{T_0}')'$ and $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}_1', \ldots, \boldsymbol{\lambda}_{T_0}')'$ are matrices with dimension $IT_0 \times F$, and $\boldsymbol{\varepsilon}_j^* = (\boldsymbol{\varepsilon}_{j1}', \ldots, \boldsymbol{\varepsilon}_{jT_0}')'$ is a length-IT_0 vector of mean-zero errors. Combining the (8) with the formula above, we see

$$\mathbf{Y}_{jt}(0) = (\boldsymbol{\delta}_t - \mathbf{A}\boldsymbol{\Delta}) + (\boldsymbol{\theta}_t - \mathbf{A}\boldsymbol{\Theta})\mathbf{R}_j + \mathbf{A}\mathbf{Y}_j^*(0) + (\boldsymbol{\varepsilon}_{jt} - \mathbf{A}\boldsymbol{\varepsilon}_j^*),$$

for $t > T_0$ where $\mathbf{A} = \boldsymbol{\lambda}_t(\boldsymbol{\Lambda}'\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}'$ is an $I \times IT_0$ matrix. Therefore, the stochastic component of $f(\mathbf{Y}_{jt}(0)|\mathbf{Y}_j^*(0), \mathbf{R}_j)$ is represented by $\boldsymbol{\varepsilon}_{jt} - \mathbf{A}\boldsymbol{\varepsilon}_j^*$, which is independent and identically distributed across all j for a fixed $t > T_0$.

A variety of methods for estimation of propensity scores have been proposed in the literature. Note that estimation is complicated in our application due to the dimensionality of our data (e.g., the number of variables across which we must balance is nearly double the number of cases that are in the treatment region). We consider three methods for estimation of $\{\pi_j\}$. First, propensity scores are estimated using a standard logistic regression model—this method is straightforward but perhaps lacks the necessary rigor. Second, we employ generalized boosted regression models (GBM) as proposed by Friedman (2001, 2002) and implemented in the `twang` package (Ridgeway et al. 2014). GBM is designed to capture nonlinearities and other nuances that are overlooked by logistic regression. Lastly, we consider estimation through the covariate balancing propensity scores (CBPS) procedure (Imai and Ratkovic 2014) as implemented within Fong et al. (2015). Since CBPS is designed to

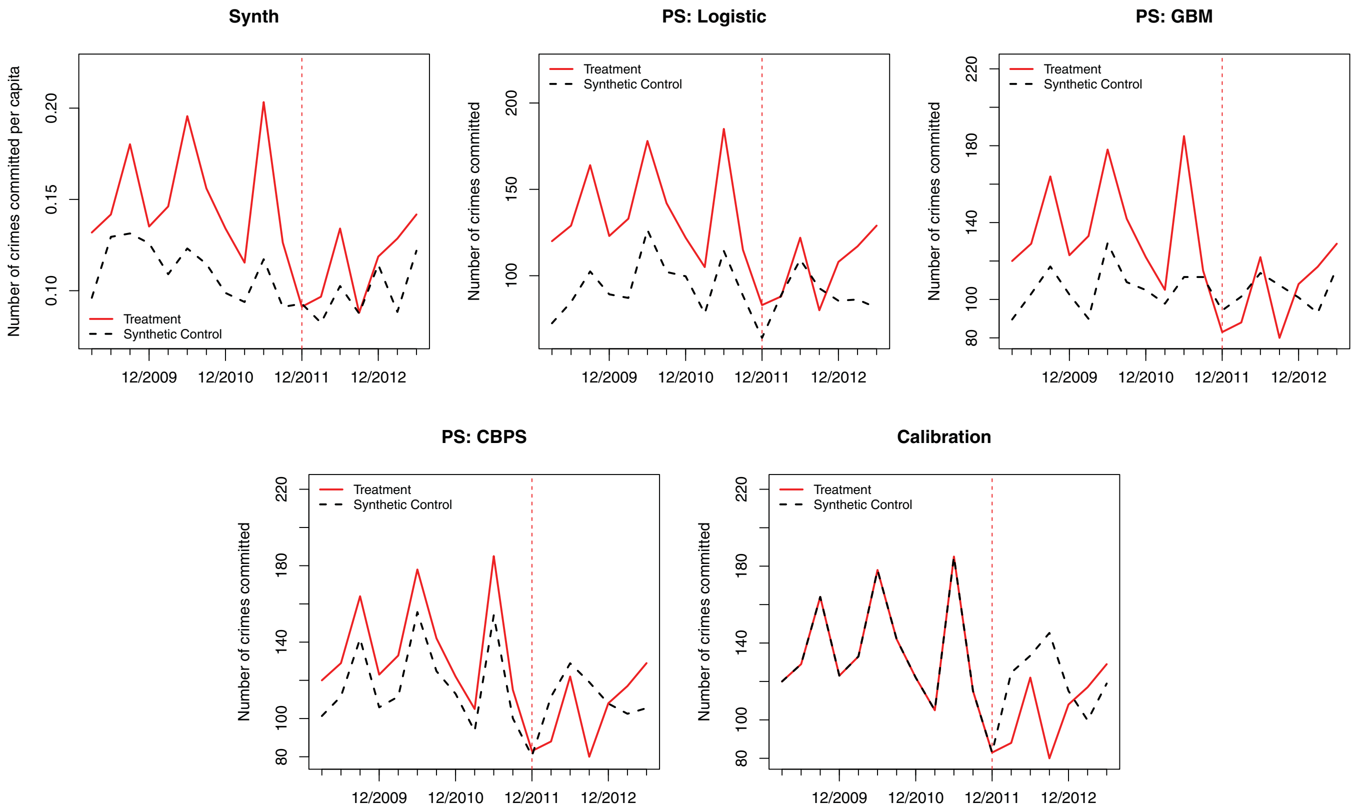

Figure 3. Plots of crime levels for `i_any_crime` in Hurt Park and its synthetic control when weights are found using `Synth`, propensity scores calculated with three different methods, and calibration (our proposed procedure).

optimize covariate balance, it should produce results that most closely resemble those which are seen from weights that balance treatment and control through calibration as proposed here. To enable comparisons of crime counts when aggregated across the treatment area (as opposed to examining per-block averages), all sets of propensity score weights for the control cases are rescaled so that they sum to the number of blocks in the treatment area.

4.4. *Results of Comparisons*

In this subsection, we show results of application of the aforementioned comparative methods (DiD, `Synth`, and propensity scores found using logistic regression, GBM and CBPS) when applied to the Roanoke DMI data. With DiD, each outcome is modeled separately; for the remaining methods, we attempt to balance the treatment and control regions jointly across the variables indicated in Table 1. Since each of these methods (aside from DiD) enables a manner of calculating weights for the control blocks, a synthetic control region is established using these procedures and we can produce time series plots akin to those provided in Figure 1 (wherein the treatment region is compared to its synthetic control); Figure 3 provides such plots for the `i_any_crime` outcome. Note that, as expected, the more complex methods provide better balance. However, the synthetic Hurt Park neighborhood for each of the alternative methods understates the crime levels prior to the intervention, and when viewed in contrast results found using calibration (which is our proposed method and which is also shown in Figure 3 to facilitate comparisons), we see that the synthetic control for each of these comparison methods likely also understates crime levels post-intervention (which would induce bias into the treatment effect estimate). Results for the other outcomes (not shown) are similar to what is seen in Figure 3. We note that the balance given by each of the weight-based comparison methods can be greatly improved by analyzing each outcome separately. However, such analyses will be subject to omitted variable biases of the form illustrated in Section 3.3.

For more rigorous comparisons, we provide the estimate of the percent change in crime levels caused by the DMI, as well as p-values found using the normal approximation and the permutation method (as applied using the standardized version of the treatment effect estimator $\widehat{\alpha}_i^{**}$), in Table 3. Note that when weights are used (for `Synth` or propensity scores), analysis may proceed in the same manner as outlined in Section 2 (however, there is not a normal approximation that may be used to estimate a p-value when `Synth` is applied). To facilitate comparisons, the table also shows results when weights are found using the proposed technique (calibration); such findings are copied from Table 2.

From Table 3, we see that in several cases, the difference-in-differences approach indicates a more substantial effect of treatment that than which is estimated by our proposed method (although DiD is subject to omitted variable biases akin to those described in Section 3.3). As implied by Figure 3, the remaining methods consistently yield smaller estimates of the treatment effect than our proposed method (this is likely due to the other methods constructing synthetic controls that understate crime levels prior to the intervention). The estimates of the treatment effect (and their respective levels of statistical significance) found using the more complicated manners of

Table 3. Estimates of the effect of the drug market intervention in Roanoke, VA (see the columns labeled “% Chg.”) and *p*-values for the significance of that effect (see the columns labeled “Norm.” and “Perm.”) when found using six different methods.

		Diff.-in-Diff.			Synth		PS: Logistic			PS: GBM			PS: CBPS			Calibration		
		% Chg.	Norm.	Perm.	% Chg.	Perm.	% Chg.	Norm.	Perm.	% Chg.	Norm.	Perm.	% Chg.	Norm.	Perm.	% Chg.	Norm.	Perm.
6 months post intervention	i_rape	−100.0	0.500	0.395	−100.0	0.107	−100.0	0.043	0.362	−100.0	0.004	0.424	−100.0	0.053	0.390	−100.0	0.081	0.394
	i_robbery	−43.3	0.306	0.001	−46.6	0.146	11.6	0.540	0.518	−44.3	0.253	0.359	−21.4	0.405	0.482	18.4	0.555	0.617
	i_aggassault	−42.2	0.212	0.151	−30.6	0.146	−14.1	0.398	0.301	−19.7	0.344	0.247	−46.4	0.119	0.134	−47.5	0.129	0.179
	i_burglary	−43.0	0.103	0.081	−11.3	0.359	−51.5	0.127	0.082	−38.9	0.105	0.057	−53.3	0.117	0.122	−67.4	0.048	0.083
	i_larceny	−24.8	0.144	0.466	24.5	0.670	−17.0	0.255	0.069	−10.9	0.309	0.240	−26.6	0.130	0.103	−29.4	0.122	0.124
	i_cartheft	−55.5	0.223	0.025	−38.2	0.126	−30.8	0.347	0.266	−63.4	0.077	0.144	−54.5	0.208	0.237	−53.6	0.222	0.268
	i_arson	430.5	0.984	0.974	275.7	0.981	1047	0.902	0.978	1055	0.903	0.981	1649	0.910	0.981	5758	0.919	0.988
	i_simpassault	−35.5	0.039	0.086	−21.3	0.165	−5.0	0.425	0.121	−0.9	0.486	0.193	−17.2	0.236	0.174	−15.3	0.285	0.240
	i_drugs	3.3	0.561	0.609	88.5	0.883	7.8	0.574	0.346	20.1	0.683	0.513	−0.1	0.499	0.435	−2.5	0.474	0.496
	i_weapons	−69.1	0.131	0.000	−25.4	0.204	−73.5	0.064	0.207	−55.8	0.129	0.258	−74.2	0.040	0.195	−77.8	0.040	0.202
	i_violent	−36.0	0.028	0.064	−20.5	0.165	−1.3	0.480	0.138	−2.5	0.459	0.148	−17.7	0.222	0.152	−15.1	0.279	0.227
	i_property	−32.6	0.045	0.249	9.9	0.524	−29.2	0.118	0.024	−23.3	0.080	0.039	−36.5	0.054	0.040	−45.2	0.028	0.028
	i_any_crime	−29.7	0.002	0.173	24.8	0.728	6.0	0.654	0.155	−2.4	0.425	0.084	−12.6	0.173	0.069	−18.6	0.099	0.052
	Omnibus	—	—	0.075	—	0.495	—	0.128	0.200	—	0.039	0.431	—	0.058	0.121	—	0.066	0.278
12 months post intervention	i_rape	12.0	0.595	0.589	−6.1	0.359	48.0	0.623	0.468	54.9	0.637	0.555	9.5	0.532	0.486	−3.7	0.486	0.480
	i_robbery	−44.9	0.220	0.109	−31.9	0.175	43.5	0.661	0.546	−44.7	0.162	0.122	−15.5	0.410	0.437	−37.4	0.283	0.339
	i_aggassault	−12.6	0.401	0.396	2.7	0.641	−24.2	0.257	0.148	6.0	0.561	0.348	−40.0	0.100	0.108	−42.8	0.109	0.150
	i_burglary	−39.7	0.064	0.060	−17.0	0.214	−50.5	0.053	0.037	−34.5	0.082	0.046	−55.1	0.028	0.048	−64.3	0.012	0.038
	i_larceny	−39.9	0.007	0.287	−5.6	0.495	−27.1	0.067	0.014	−27.9	0.024	0.039	−40.5	0.011	0.008	−41.8	0.020	0.035
	i_cartheft	−45.6	0.165	0.155	−34.6	0.117	−17.5	0.371	0.241	−42.6	0.123	0.087	−39.0	0.201	0.215	−50.0	0.126	0.205
	i_arson	92.2	0.821	0.836	23.6	0.854	242.7	0.838	0.890	196.7	0.823	0.842	239.0	0.833	0.910	−3.1	0.488	0.578
	i_simpassault	−32.0	0.020	0.076	−17.9	0.194	10.3	0.683	0.211	0.8	0.517	0.178	−5.4	0.389	0.272	−13.5	0.249	0.223
	i_drugs	1.7	0.603	0.570	44.6	0.874	28.0	0.820	0.431	28.5	0.841	0.601	12.6	0.679	0.515	1.2	0.518	0.493
	i_weapons	−71.2	0.046	0.000	−46.9	0.078	−64.0	0.050	0.063	−58.1	0.038	0.099	−67.1	0.027	0.065	−70.3	0.033	0.083
	i_violent	−25.5	0.040	0.132	−12.1	0.204	11.6	0.718	0.200	5.4	0.621	0.209	−7.1	0.340	0.216	−16.8	0.170	0.149
	i_property	−41.8	0.001	0.101	−11.3	0.359	−34.3	0.020	0.004	−30.9	0.005	0.002	−44.9	0.002	0.001	−50.2	0.001	0.005
	i_any_crime	−31.4	0.000	0.113	13.1	0.592	5.8	0.711	0.116	−6.1	0.237	0.030	−14.8	0.047	0.016	−23.2	0.006	0.001
	Omnibus	—	—	0.054	—	0.379	—	0.286	0.121	—	0.080	0.267	—	0.037	0.053	—	0.002	0.044
18 months post intervention	i_rape	−31.7	0.326	0.226	−25.4	0.243	−32.1	0.343	0.223	−9.9	0.457	0.322	−42.3	0.281	0.277	−60.6	0.205	0.266
	i_robbery	−43.3	0.180	0.163	−6.7	0.291	58.6	0.733	0.594	−43.4	0.126	0.123	−1.3	0.492	0.485	−40.6	0.224	0.282
	i_aggassault	−14.7	0.360	0.366	−4.5	0.350	−20.9	0.269	0.130	5.6	0.564	0.336	−35.3	0.112	0.121	−31.1	0.187	0.211
	i_burglary	−37.8	0.048	0.051	−6.0	0.282	−41.2	0.074	0.034	−29.4	0.092	0.036	−48.7	0.029	0.047	−59.0	0.010	0.041
	i_larceny	−37.9	0.003	0.332	8.7	0.573	−17.0	0.163	0.029	−23.4	0.044	0.082	−32.9	0.020	0.025	−33.6	0.034	0.055
	i_cartheft	−45.6	0.129	0.123	−46.8	0.039	−19.6	0.332	0.195	−45.9	0.080	0.071	−38.4	0.168	0.202	−44.0	0.136	0.185
	i_arson	143.9	0.934	0.926	85.5	0.942	406.9	0.916	0.968	305.6	0.903	0.940	408.1	0.914	0.968	45.4	0.660	0.706
	i_simpassault	−27.9	0.019	0.112	−8.0	0.233	20.8	0.856	0.299	10.4	0.733	0.274	3.0	0.570	0.401	−1.2	0.474	0.451
	i_drugs	1.5	0.577	0.584	41.2	0.854	34.5	0.918	0.512	28.4	0.900	0.637	14.6	0.752	0.596	−2.3	0.456	0.448
	i_weapons	−60.6	0.038	0.006	−47.1	0.039	−42.8	0.127	0.083	−44.5	0.068	0.117	−49.7	0.066	0.094	−50.3	0.088	0.145
	i_violent	−24.2	0.025	0.135	−4.4	0.252	18.0	0.849	0.272	10.4	0.754	0.263	−1.0	0.474	0.295	−7.6	0.316	0.282
	i_property	−40.1	0.000	0.135	−1.4	0.437	−24.5	0.056	0.006	−26.8	0.010	0.013	−37.7	0.003	0.003	−42.6	0.002	0.002
	i_any_crime	−25.8	0.000	0.208	18.5	0.660	18.4	0.975	0.387	1.7	0.592	0.117	−4.6	0.281	0.132	−12.6	0.065	0.062
	Omnibus	—	—	0.084	—	0.524	—	0.445	0.141	—	0.148	0.214	—	0.010	0.087	—	0.000	0.026

estimating weights (e.g., GBM and CBPS) tends to more closely approximate those seen with calibration.

Discrepancies in the p-value yielded by the normal approximation and the permutation methods are much more noticeable for the comparison methods than they are for the proposed 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, Duflo, and Mullainathan 2004), whereas for the other comparison methods, the permutation methods yield smaller p-values than those seen with their respective normal approximation. This likely an additional consequence of the synthetic control for these methods systematically understating crime levels preintervention; this issue pervades into the placebo groups and thereby the permutation method naturally corrects some of this problem.

In addition to offering the most defensible findings, calibration is by far the most computationally efficient of the methods outlined in this section (with the exception of logistic regression). For instance, calibration can estimate synthetic control weights on the dataset described here in under 2 sec of computing time, whereas CBPS requires upwards 10 min to calculate a single set of weights (the use of this method becomes impractical if one hopes to apply the permutation procedure). These computations were executed on a Windows machine with a 2.90 GHz processor and 8 GB of memory.

5. Discussion

5.1. Method

Our study illustrates the advantages of using 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, preintervention 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. Furthermore, our comparisons implicate that when weights are calculated using propensity score methods, a biased estimate of the treatment effect may result; this is due to the inability of propensity score weights to adequately balance treatment and control in a setting with the dimensionality described here.

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.

As noted previously, a drawback of the proposed approach is the assumption that a feasible solution to (10) exists. Unfortunately, the feasibility of a solution is data-dependent. Therefore, we are limited in terms of the guidance we can provide as to understanding circumstances when a solution will and will not be feasible. Obviously, feasibility is more likely in settings with a larger number of untreated cases and fewer constraints. However, we also find that (even if holding the total number of cases constant), feasibility is more easily obtained if the treated area consists of a larger number of cases. Therefore, the previously mentioned strategy of finding a synthetic control for each individual data unit within the treated region is highly likely to be infeasible (such a strategy involves the calculation of several sets of weights, each with $J - J_0 = 1$). Furthermore, this strategy will become computationally impractical if a `Synth`-type procedure is used to circumvent infeasible constraints.

When an exact solution to the calibration equations is not feasible, we recommend the use of quadratic programming (e.g., Boyd and Vandenberghe 2004) to find weights that satisfy a prespecified subset of the constraints and minimize the imbalance across the remaining constraints. We prefer such a technique to propensity scores since one can impose exact balance on the average (across all preintervention time periods) of a given outcome and thereby avoid issues seen in Figure 3.

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 (Sherman et al. 2002; Saunders et al. 2015). The DMI, an example of such a program, has been lauded as an effective program despite the lack of rigor in evaluations of it and the fact that the majority of the evidence comes from the city where it was designed (e.g., Kennedy and Wong 2009; Corsaro et al. 2011, 2012; Saunders et al. 2015).

The results of this evaluation find that the Roanoke DMI significantly reduced crime in the target area at the same magnitude as in the first implementation in High Point, NC (Saunders et al. 2015). Further, our findings were derived using methods that are more rigorous and defensible than previous evaluations, and we consider a wider array of crime types than evaluated in previous studies. In our evaluation, we saw that the DMI did not have an effect on drug crime, but did reduce most of the other types of crime, which is in line with earlier findings (Corsaro et al. 2011; Saunders et al. 2015). Additionally, our displacement/diffusion analyses yielded interesting findings. Specifically, we observed that the intervention effect diffused to areas surrounding the treatment area, and we saw evidence of a displacement of drug crimes to a separate overt drug market. Similar findings have not been previously noted in the literature on the DMI.

Our findings 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 (Weisburd and Eck 2004; Braga and Weisburd 2010, 2011). 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.

Supplementary Materials

Web Appendix: This appendix contains calculations for WLS parameter estimation, derivation of test statistics under alternative model specifications, and crime diffusion/displacement analyses.

Acknowledgments

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

Abadie, A., Diamond, A., and Hainmueller, J. (2010), "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statistical Association*, 105, 493–505. [109,110,111,113,121]

——— (2011), "Synth: An R Package for Synthetic Control Methods in Comparative Case Studies," *Journal of Statistical Software*, 42, 1–17. [121]

——— (2014), "Comparative Politics and the Synthetic Control Method," *American Journal of Political Science*, 59, 495–510. [111]

Abadie, A., and Gardeazabal, J. (2003), "The Economic Costs of Conflict: A Case Study of the Basque Country," *American Economic Review*, 93, 113–132. [109,111,121]

Angrist, J. D., and Krueger, A. B. (1999), "Empirical Strategies in Labor Economics," in *Handbook of Labor Economics* (Vol. 3), eds. O. C. Ashenfelter & D. Card, pp. 1277–1366. Amsterdam, The Netherlands: Elsevier, North Holland. [113]

Auld, M. C., and Grootendorst, P. (2004), "An Empirical Analysis of Milk Addiction," *Journal of Health Economics*, 23, 1117–1133. [113]

Baumer, E., Lauritsen, J. L., Rosenfeld, R., and Wright, R. (1998), "The Influence of Crack Cocaine on Robbery, Burglary, and Homicide Rates: A Cross-City, Longitudinal Analysis," *Journal of Research in Crime and Delinquency*, 35, 316–340. [114]

Bertrand, M., Duflo, E., and Mullainathan, S. (2004), "How Much Should We Trust Differences-in-Differences Estimates?" *The Quarterly Journal of Economics*, 119, 249–275. [121,124]

Billmeier, A., and Nannicini, T. (2013), "Assessing Economic Liberalization Episodes: A Synthetic Control Approach," *The Review of Economics and Statistics*, 95, 983–1001. [111,121]

Binder, D. A. (1983), "On the Variances of Asymptotically Normal Estimators From Complex Surveys," *International Statistical Review / Revue Internationale de Statistique*, 51, 279–292. [113]

Blumstein, A., and Rosenfeld, R. (1998), "Explaining Recent Trends in US Homicide Rates," *Journal of Criminal Law and Criminology*, 88, 1175–1216. [114]

Bohn, S., Lofstrom, M., and Raphael, S. (2014), "Did the 2007 Legal Arizona Workers Act Reduce the State's Unauthorized Immigrant Population?" *The Review of Economics and Statistics*, 96, 258–269. [111]

Boyd, S., and Vandenberghe, L. (2004), *Convex Optimization*, Cambridge, UK: Cambridge University Press. [124]

Braga, A., and Weisburd, D. (2012), "The Effects of "Pulling Levers" Focused Deterrence Strategies on Crime," *Campbell Systematic Reviews*, 8, doi: 10.4073/csr.2012.6. [109]

Braga, A. A., and Weisburd, D. L. (2010), "Editors' Introduction: Empirical Evidence on the Relevance of Place in Criminology," *Journal of Quantitative Criminology*, 26, 1–6. [125]

——— (2011), "The Effects of Focused Deterrence Strategies on Crime: A Systematic Review and Meta-Analysis of the Empirical Evidence," *Journal of Research in Crime and Delinquency*, 49, 323–358. [125]

Caulkins, J. P. (1993), "Local Drug Markets' Response to Focused Police Enforcement," *Operations Research*, 41, 848–863. [114]

Cavallo, E., Galiani, S., Noy, I., and Pantano, J. (2013), "Catastrophic Natural Disasters and Economic Growth," *The Review of Economics and Statistics*, 95, 1549–1561. [111,121]

Clarke, R. V., and Weisburd, D. (1994), "Diffusion of Crime Control Benefits: Observations on the Reverse of Displacement," *Crime Prevention Studies*, 2, 165–184. [119]

Corsaro, N., Brunson, R., Gau, J., and Oldham, C. (2011), *The Peoria Pulling Levers Drug Market Intervention: A Review of Program Process, Changes in Perceptions, and Crime Impact*, Washington, DC: Bureau of Justice Assistance, Office of Justice Programs, U.S. Department of Justice. [125]

Corsaro, N., and Brunson, R. K. (2013), "Are Suppression and Deterrence Mechanisms Enough? Examining the 'Pulling Levers' Drug Market Intervention Strategy in Peoria, Illinois, USA," *International Journal of Drug Policy*, 24, 115–121. [115]

Corsaro, N., Brunson, R. K., and McGarrell, E. F. (2013), "The Peoria Pulling Levers Drug Market Intervention: A Review of Program Process, Changes in Perception, and Crime Impact," *Crime & Delinquency*, 59, 1085–1107. [115]

Corsaro, N., Hunt, E. D., Hipple, N. K., and McGarrell, E. F. (2012), "The Impact of Drug Market Pulling Levers Policing on Neighborhood Violence," *Criminology & Public Policy*, 11, 167–199. [109,115,125]

Corsaro, N., and McGarrell, E. (2009), "An Evaluation of the Nashville Drug Market Initiative (DMI) Pulling Levers Strategy," Technical Report, Michigan State University, School of Criminal Justice, East Lansing, MI. Available at *https://nnscommunities.org/uploads/NashvilleEvaluation.pdf* [115]

Deville, J.-C., and Särndal, C.-E. (1992), "Calibration Estimators in Survey Sampling," *Journal of the American Statistical Association*, 87, 376–382. [111]

Deville, J.-C., Särndal, C.-E., and Sautory, O. (1993), "Generalized Raking Procedures in Survey Sampling," *Journal of the American Statistical Association*, 88, 1013–1020. [112]

DiNardo, J. E., and Pischke, J.-S. (1997), "The Returns to Computer Use Revisited: Have Pencils Changed the Wage Structure too?" *The Quarterly Journal of Economics*, 112, 291–303. [113]

Draca, M., Machin, S., and Witt, R. (2011), "Panic on the Streets of London: Police, crime, and the July 2005 Terror Attacks," *The American Economic Review*, 101, 2157–2181. [109]

Fong, C., Ratkovic, M., Hazlett, C., and Imai, K. (2015), *CBPS: Covariate Balancing Propensity Score, R Package Version 0.10*. Available at *http://CRAN.R-project.org/package=CBPS* [121]

Frabutt, J. M., Shelton, T. L., Di Luca, K. L., Harvey, L. K., and Hefner, M. K. (2009), "A Collaborative Approach to Eliminating Street Drug Markets Through Focused Deterrence," Technical Report, National Institute of Justice, Washington, DC. Available at *https://www.ncjrs.gov/pdffiles1/nij/grants/239242.pdf* [115]

Friedman, J. H. (2001), "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics*, 29, 1189–1232. [121]

——— (2002), "Stochastic Gradient Boosting," *Computational Statistics & Data Analysis*, 38, 367–378. [121]

Gove, W. R., Hughes, M., and Geerken, M. (1985), "Are Uniform Crime Reports a Valid Indicator of the Index Crimes? An Affirmative Answer With Minor Qualifications," *Criminology*, 23, 451–502. [115]

Hainmueller, J. (2012), "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies," *Political Analysis*, 20, 25–46. [112]

Harocopos, A., and Haugh, M. (2005), *Drug Dealing in Open-Air Markets*, Washington, DC: U.S. Department of Justice, Office of Community Oriented Policing Services, Center for Problem-Oriented Policing. [114]

Hipple, N., and McGarrell, E. (2009), *Bureau of Justice Assistance Drug Market Intervention Implementation Guide and Lessons Learned*, East Lansing, MI: Michigan State University, School of Criminal Justice. [109,115]

Imai, K., and Ratkovic, M. (2014), "Covariate Balancing Propensity Score," *Journal of the Royal Statistical Society*, Series B, 76, 243–263. [121]

Imbens, G. W. (2004), "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *Review of Economics and Statistics*, 86, 4–29. [121]

Kennedy, D. (2009), "Drugs, Race and Common Ground: Reflections on the High Point Intervention," *NIJ Journal*, 262, 12–17. [109,114,115]

Kennedy, D. M., Piehl, A. M., and Braga, A. A. (1996), "Youth Violence in Boston: Gun Markets, Serious Youth Offenders, and a Use-Reduction Strategy," *Law and Contemporary Problems*, 59, 147–196. [115]

Kennedy, D. M., and Wong, S. (2009), *The High Point Drug Intervention Strategy*, Washington, DC: U.S. Department of Justice Office of Community Oriented Policing Services. [114,115,125]

Kish, L. (1965), *Survey Sampling*, New York: Wiley. [112]

Kleiman, M. A. (1997), "The Problem of Replacement and the Logic of Drug Law Enforcement," *Drug Policy Analysis Bulletin*, 3, 8–10. [114]

Lachin, J. M. (2014), "Applications of the Wei-Lachin Multivariate One-Sided Test for Multiple Outcomes on Possibly Different Scales," *PloS one*, 9, e108784. [113]

Law Enforcement Support Section, and Crime Statistics Management Unit (2013), *Summary Reporting System (SRS) User Manual*, Washington, D.C.: U.S. Department of Justice, Federal Bureau of Investigation, Criminal Justice Information Services Division. [115]

Lejins, P. P. (1966), "Uniform Crime Reports," *Michigan Law Review*, 64, 1011–1030. [115]

Lumley, T. (2004), "Analysis of Complex Survey Samples," *Journal of Statistical Software*, 9, 1–19. [112]

——— (2011), *Complex Surveys: A Guide to Analysis Using R* (Vol. 565), New York: John Wiley & Sons. [112]

Mazerolle, L., Soole, D. W., and Rombouts, S. (2006), "Street-Level Drug Law Enforcement: A Meta-Analytical Review," *Journal of Experimental Criminology*, 2, 409–435. [114]

McGarrell, E. F., Corsaro, N., and Brunson, R. K. (2010), "The Drug Market Intervention Approach to Overt Drug Markets," *Journal of Criminal Justice and Security*, 12, 397–407. [114]

NNSC (2015), "Group Violence Intervention: An Implementation Guide," Technical Report NCJ 244850, National Network for Safe Communities, U.S. Department of Justice, Office of Community Oriented Policing Services, Washington, D.C. [115]

Reppetto, T. A. (1976), "Crime Prevention and the Displacement Phenomenon," *Crime & Delinquency*, 22, 166–177. [119]

Reuter, P. H., and MacCoun, R. J. (1992), *Street Drug Markets in Inner-City Neighbourhoods*, Santa Monica, CA: RAND Corporation. [114]

Ridgeway, G., McCaffrey, D., Morral, A., Griffin, B. A., and Burgette, L. (2014), *Twang: Toolkit for Weighting and Analysis of Nonequivalent Groups, R Package Version 1.4-0*. Available at *http://CRAN.R-project.org/package=twang* [121]

Rosenbaum, P. R., and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55. [121]

Särndal, C.-E. (2007), "The Calibration Approach in Survey Theory and Practice," *Survey Methodology*, 33, 99–119. [111]

Saunders, J., Lundberg, R., Braga, A. A., Ridgeway, G., and Miles, J. (2015), "A Synthetic Control Approach to Evaluating Place-Based Crime Interventions," *Journal of Quantitative Criminology*, 31, 413–434. [109,115,120,124,125]

Saunders, J., Ober, A., Barnes-Proby, D., and Brunson, R. (2016), "Police Legitimacy and Disrupting Overt Drug Markets," *Policing: An International Journal of Policy and Practice*, 39, 667–679. [115]

Saunders, J., Ober, A., Kilmer, B., and Greathouse, S. (2016), *A Community-Based Focused Deterrence Approach to Closing Overt Drug Markets*, Santa Monica, CA: RAND Corporation. [115]

Sherman, L. W., MacKenzie, D. L., Farrington, D. P., and Welsh, B. C. (eds.) (2002), *Evidence-Based Crime Prevention*, London: Routledge. [124]

Telep, C. W., Weisburd, D., Gill, C. E., Vitter, Z., and Teichman, D. (2014), "Displacement of Crime and Diffusion of Crime Control Benefits in Large-Scale Geographic Areas: A Systematic Review," *Journal of Experimental Criminology*, 10, 515–548. [119]

Weisburd, D., and Eck, J. E. (2004), "What Can Police do to Reduce Crime, Disorder, and Fear?" *The Annals of the American Academy of Political and Social Science*, 593, 42–65. [125]

Weisburd, D., and Mazerolle, L. G. (2000), "Crime and Disorder in Drug Hot Spots: Implications for Theory and Practice in Policing," *Police Quarterly*, 3, 331–349. [114]