# The synthetic control method: Small T, small N Monte Carlo evidence and an application to the effects of privatizing probation services on revoke rates

Süß, Philipp

5 July 2016

# The synthetic control method: Small T, small N Monte Carlo evidence and an application to the effects of privatizing probation services on revoke rates

Philipp Süß*

July 5, 2016

**Abstract**

Relying on synthetic controls to estimate treatment effects recently gained popularity in applied econometrics. The small sample properties of the synthetic control estimator are however not sufficiently investigated and even the proofs of consistency impose a factor model and require either the pre-treatment period or the pre-treatment period and the size of the donor pool going to infinity $((T_0 \to \infty) \vee ((N - N_1) \to \infty \wedge T_0 \to \infty))$. Since applications often ignore the lack of statistical foundation in small samples, a "small T small N" Monte Carlo study covering standard econometric models like the ones from Differences-in-Differences setups, heterogeneous ADLX models with and without unit roots and random coefficient models is conducted. The results suggest that the estimator is frequently unbiased, that unit roots are problematic and that a main placebo test has good size and mediocre power properties. Furthermore, the synthetic control method is used to estimate the causal effect of outsourcing probation and parole services on revoke rates by exploiting a natural experiment in Germany. Results provide evidence against increases in revoke rates due to outsourcing.

**Keywords:** Synthetic control method, Finite sample properties, Unit roots, Probation, Parole, Privatization
**JEL Classification:** C13, C23, H11, K40

* Goethe University Frankfurt, Department of Applied Econometrics and International Economic Policy, Theodor-W.-Adorno-Platz 4, 60323 Frankfurt am Main, Germany Tel.: +49-69-798-34841; E-mail: Philipp.Suess@wiwi.uni-frankfurt.de

# 1 Introduction

The synthetic control method as a further advancement of the Differences-in-Differences methodology recently gained popularity in the applied and theoretical literature on treatment effect evaluation. Introduced by Abadie and Gardeazabal (2003), recent exemplary applications cover the estimation of the effect of the German reunification on West-German GDP by Abadie, Diamond and Hainmueller (2015) or Arizona's 2007 Legal Arizona Workers Act on the share of Hispanics without citizenship in Arizona's population by Bohn, Lofstrom and Raphael (2014). The main contributions toward the statistical foundation of the synthetic control method seem to be two consistency proofs. The first proof shows that if the true model has a factor structure and perfect pre-treatment fit, then consistency is achieved as the pre-treatment period approaches infinity ($T_0 \to \infty$). Again working under the assumption of a factor structure, the second proof shows that it is possible to relax the condition on the pre-treatment fit by assuming that the size of the donor pool and the pre-treatment period tend to infinity ($(N - N_1) \to \infty \wedge T_0 \to \infty$) (see Abadie, Diamond and Hainmueller 2010 and Gobillon and Magnac 2015). To my knowledge, the synthetic control estimator's finite sample properties are not sufficiently investigated, which places many recent working papers operating in small T, small N environments on shaky grounds.[1] Consequently, the first contribution of this paper is an extensive Monte Carlo study, which generally supports the use of synthetic controls in small T, small N environments, but indicates severe problems when independent unit roots are present. And since the synthetic control method does not provide standard errors to assess uncertainty I performed a size and power analysis of the main placebo test which suggests good size and mediocre power properties in small T, small N environments.

The application section of the paper focuses on the effects of privatizing probation and parole services in Germany. A holistic evaluation of the privatized probation and parole services in Baden-Wuerttemberg is provided by the Ministry of Justice Baden-Wuerttemberg (2014) and Dölling, Entorf and Hermann (2014). As part of the analysis Dölling, Entorf and Hermann (2014) analyzed the evolutions of the revoke rates (ratio of unsuccessful terminations to all terminations of probation and parole), which is one of the standard measures of performance of probation and parole services, by using mainly descriptive evidence and a standard aggregated Differences-in-Differences approach with two separate ad-hoc control groups. To reasses the conclusions concerning the effects of outsourcing on probation and parole violations this paper adds further German states to the control group and estimates the effect using the synthetic control method.

The remainder of the paper is structured as follows. Section 2 reviews the synthetic control group methodology and emphasizes shortcomings in its statistical foundation. Section 3 presents the results of a Monte Carlo study to assess the performance of the synthetic control estimator and the size and power properties of a frequently employed placebo test in a small T, small N

[1] A very recent working paper by Xu (2015) from 20th November, 2015 compared inter alia a generalization of the synthetic control estimator to the synthetic control estimator in a $T_0 = 15$ and $N = 40$ factor model setting and reports unbiasedness, given sufficient common support of the outcomes of treated and non-treated unit.

environment. Section 4 introduces to the privatization of probation and parole services, applies the synthetic control methodology to several revoke rates (revoke rates for juveniles, adults or revoke rates due to recidivism, etc.) with Baden-Wuerttemberg as the treatment group and performs several robustness checks, after which Section 5 concludes.

# 2 Synthetic control method and placebo tests

In their analysis of the economic cost of conflict for the Basque Country Abadie and Gardeazabal (2003) tackled the question what the economic situation of the Basque Country would have been without conflict. Instead of relying on a counterfactual economic situation based on a time trend or a single comparison unit, Abadie and Gardeazabal (2003) developed a method to construct the counterfactual economic situation of the Basque Country as a weighted average of other Spanish regions, which were not directly affected by terrorism, and termed this method "synthetic control method". Recently, Abadie (2015) provided a step by step introduction how to use the synthetic control method with the Stata program "synth". The following description of the method closely resembles the one provided in Abadie, Diamond and Hainmueller (2010). Assume we collected data for $N$ units over $T$ periods. Unit 1 was not treated up to $T_0$ and received treatment in period $T_0+1$. The remaining $N-1$ units did not receive treatment. Further, the main object of interest is the treatment effect for unit one in the treatment and post treatment period ($TE_1(t)$) defined by the difference of the corresponding potential outcomes ($Y_{1,d}(t)$):

$$TE_1(t) = Y_{1,1}(t) - Y_{1,0}(t),\ T_0+1 \leq t \leq T. \tag{1}$$

The core idea of the synthetic control group is to estimate the potential outcome without treatment at time $t$ for the treatment unit by a weighted average of the outcomes from the donor pool ($-i = \{2, ..., N\}$) given by:

$$\hat{Y}_{1,0}(t) = \sum_{j=2}^{N} w_j Y_j(t),\ T_0+1 \leq t \leq T. \tag{2}$$

To avoid extrapolation Abadie, Diamond and Hainmueller (2010) restrict the weights to lie in the unit interval and sum to one ($w_j \in [0,1]$ and $\sum_{j=2}^{N} w_j = 1$), which is in the spirit of the overlap condition for matching estimators. Therefore the support of the potential outcome estimator $\hat{Y}_{1,0}(t) = \sum_{j=2}^{N} w_j Y_j(t)$ without treatment is given by the convex hull of the donor pools observed outcomes $\{Y_2(t), ...., Y_N(t)\}$. In practice the weights which are collected in $W = [w_2, ..., w_N]'$ are estimated by minimizing the weighted euclidean distance between $k$ pre-intervention characteristics of the treated unit $X_{11}, ..., X_{1k}$ and the corresponding $k$ pre-intervention characteristics for the

members of the donor pool collected in the $k \times (N-1)$ matrix $X_0$. The optimal $W^*$ is hence given by:

$$W^* = \underset{W}{argmin} \sum_{m=1}^{k} v_m (X_{1m} - X_{0m}W)^2 \tag{3}$$

The importance weights $v_m$ correspond to the importance of the pre-intervention characteristic $m$. They can be chosen by the researcher or estimated to minimize squared pre-intervention differences between outcomes for the treatment and the synthetic control group (see Abadie 2015). The pre-intervention characteristics used in the literature are a mixture of pre-intervention outcome predictors, functions like averages of those and pre-intervention outcomes or averages. Assuming stable data-generating processes for the units, which are functions of observed ($Z$) and unobserved variables ($U$), $Y_j = g_j(Z_j, U_j)$, it seems beneficial to at least include some pre-intervention outcomes since the weights then take into account dependence in the unobservables $U_j$ and $U_1$ like common shocks (see the comment by Abadie, Diamond and Hainmueller 2015, p. 498). To my knowledge no consensus regarding the choice of pre-intervention characteristics has been found yet.[2] The statistical foundation of the method, which passed a refereeing process, is given in Abadie, Diamond and Hainmueller (2010) and Gobillon and Magnac (2015). Abadie, Diamond and Hainmueller (2010) assume a factor model and perfect fit in the pre-intervention period and show that the bias is smaller than an expression which goes to zero as the pre-treatment period approaches infinity (which implies that $\underset{T_0 \to \infty}{lim} \left( E(\hat{Y}_{1,0}(t) - Y_{1,0}(t)) \right) = 0$ ). Gobillon and Magnac (2015) establish a sufficient condition, given non-perfect fit in pre-intervention period, that additionally requires the size of the control group tending to infinity ($(N-1) \to \infty$). Further, it should be mentioned that assuming AR processes with time-varying but cross-sectionally homogeneous parameters Abadie, Diamond and Hainmueller (2010) show unbiasedness for arbitrary $T_0$, assuming perfect pre-treatment fit.[3] Many applications (frequently in working paper status) lack a perfect fit in the pre-intervention period or have small $N$, small $T$ or even both, which places the analysis on shaky statistical grounds, if one opts for the assumption that laws of motion are heterogeneous. Further, the estimation procedure does not provide standard errors. To substitute for the lack of an uncertainty measure several placebo tests are frequently conducted in applications. Abadie, Diamond and Hainmueller (2015) present placebo tests which are based on:

- a reassignment of the treatment to a period before the actual treatment
- estimating treatment effects for non-treated units and a systematic comparison of those

[2] Just for completeness there is a working paper by Kaul et al. (2015), which makes the strong claim not to use all pre-intervention outcomes as economic predictors. The main argument is not very detailed and in my understanding not fully accurate. No supporting Monte Carlo evidence is provided.

[3] Furthermore, a recent working paper discusses the properties of the synthetic control method given a stationary VAR structure as data-generating processes (potentially stationary after transformation - see Carvalho, Masini, Medeiros 2014). To me it seems that they do not restrict the weights (the $W$ estimator is based on linear projection) hence it is not entirely clear if results carry over. Their Monte Carlo study is based on small $N$ and $T = 100$, which is not comparable to the setting in this paper.

treatment effect estimates to the one for the treated unit

- changes in the composition of the control group.

The performance of those placebo tests with respect to wrongly indicating a treatment effect when the null hypothesis of no treatment effect is true (type 1 error; size) or indicating a treatment effect when there truly is a treatment effect (1-(type 2 error); power) appears insufficiently investigated. The placebo test of reassigning the treatment to a period before the actual treatment is mainly based on visual inspection and therefore not immediately suitable for size and power investigations (see the appendix for an example). The second placebo test which estimates treatment effects for non-treated units is based on ranks and therefore formal and hence part of the Monte Carlo study. The third placebo test: changes in the composition of the control group is not considered. As the performance of the synthetic control estimator in small T, small N settings is largely unknown and the properties of corresponding placebo tests are not sufficiently investigated the following section performs a Monte Carlo study using various types of data generating process in a small $N$ and small $T_0$ environment, which may reduce scepticism concerning the results of the application.

# 3 Monte Carlo study

The Monte Carlo study can be broken down into two parts. In the first part, the main concerns are unbiasedness of the estimator and its root mean squared error (RMSE). Those properties are investigated by varying the size of the donor pool ($N-1$), the length of the pre-treatment period ($T_0$), the share of noise in the process and the treatment effect strength across five data generating processes (DGP). The second part attempts to shed some light on the actual size and power properties of the most systematic placebo test, which is based on the estimation of treatment effects of non-treated units at the time of treatment and a systematic comparison of those treatment effect estimates to the one of the treated unit. One would expect that the treatment effect estimate of the treated unit is higher in absolute terms than the treatment effect estimates for the non-treated unit. In detail the placebo test works as follows: First, the treatment effect is estimated by the synthetic control estimator for unit two (which did not receive treatment) with unit one, three up to unit $N$ in the donor pool. Then the ratio of the root mean square prediction errors (RMSPE) in the post- and pre-treatment period is calculated and saved. This is done for all $N$ units. Then the $N$ units are ranked according to their RMSPE ratio (see the rank formula in appendix 1). If the method worked well, then one would expect that the ratio of the RMSPE is greater for the treated unit than for the untreated units, i.e. that the treated unit is ranked first. To see this, note that the ratio for the treated unit becomes large if the pre-intervention fit is good and the treatment effect is sizeable and correctly estimated. For the non-treated units

a large ratio indicates that the method would have spuriously "found" a treatment effect.[4] The Monte Carlo study is performed in Matlab using code for the synthetic control method provided by Hainmueller.[5] The code was slightly adapted to allow for $N > T_0$ using Matlabs quadratic programming routine. The pre-intervention characteristics are chosen to be the pre-treatment outcomes ($X_1 = [Y_{11}, ..., Y_{1T_0}]$).

## 3.1 Data generating processes

The section below lists the data generating processes under consideration. To focus on the pure model characteristics the error terms $\epsilon_{it}$ are assumed to be independent over time and cross sectional units. The pre-treatment period is half as long as the overall period ($T_0 = T/2$).

**i) Differences-in-Differences:** The first data generating process under consideration is an additive model for the potential outcomes frequently assumed in Difference-in-Difference setups. The data generating process is given by:

$$y_{it} = \alpha_i + \gamma_t + \delta D_{it} + \sigma\epsilon_{it} \tag{4}$$

with $D_{it} = I[i = 1] \cdot I[t > T_0]$. Since the DGP is a factor model econometric theory suggests consistent estimates for $T_0 \to \infty$. For fixed $T_0$ theoretical results are not available. The individual specific effects $\alpha_i$ and the common time shocks $\gamma_t$ are draws from a pseudo random number generator imitating the uniform distribution on $(0, 1)$. $\epsilon_{it}$ is a draw from a standard normal distribution.

**ii) Stationary heterogeneous ADLX:** The second data generating process is given by a stationary heterogeneous ADLX model. Each unit has its own parameters and the data generating process is given by:

$$y_{it} = \alpha_i y_{i,t-1} + \beta_i x_{it} + \gamma_t + \theta_i + \delta D_{it} + \sigma\epsilon_{it} \tag{5}$$

with $D_{it} = I[i = 1] \cdot I[t > T_0]$ and $\alpha_i \in (0, 1) \forall i$ . Specifically $\alpha_i$, $\beta_i$, $\theta_i$ and $\gamma_t$ are drawn from a uniform distribution on $(0, 1)$, whereas $\epsilon_{it}$ and $x_{it}$ are drawn from a standard normal distribution. The process is initialized with a draw from a standard normal distribution ($y_{i0} \sim N(0, 1)$).

**iii) Non-stationary heterogeneous ADLX:** The third data generating process is given by a non-stationary heterogeneous ADLX model. Each unit has its own parameters and the data

[4] As expected the placebo test is not a wonder weapon. It is not indicative of biased estimates. Results not presented in the paper indicate weak correlation between the rank of the treated unit and the absolute bias as well as similar bias given the treated attains the highest rank and the bias unconditional of the treated rank.

[5] http://web.stanford.edu/~jhain/synthpage.html last accessed 7th. Sept, 2015. The code has no option for the importance weights $v_m$ in formula (3). They seem to be chosen to minimize squared pre-intervention differences between outcomes for the treatment and the synthetic control group.

generating process is given by:

$$y_{it} = y_{i,t-1} + \beta_i x_{it} + \gamma_t + \theta_i + \delta D_{it} + \sigma\epsilon_{it} \tag{6}$$

with $D_{it} = I[i = 1] \cdot I[t > T_0]$ . The random variables $\beta_i$, $\theta_i$ and $\gamma_t$ are drawn from a uniform distribution on $(0, 1)$ and $\epsilon_{it}$ and $x_{it}$ are drawn from a standard normal distribution. The process is initialized with a draw from a standard normal distribution ($y_{i0} \sim N(0, 1)$).

**iv) Random coefficient model:** The fourth data generating process is given by a random coefficient model:

$$y_{it} = \beta_{it} x_{it} + \delta D_{it} + \sigma\epsilon_{it} \tag{7}$$

with $D_{it} = I[i = 1] \cdot I[t > T_0]$ and $\beta_{it}$ being a serially and cross-sectionally dependent random variable. The law of motion for $\beta_{it}$ is given by $\beta_{it} = .5\beta_{i,t-1} + \bar{\beta} + \nu_{it}$. $\bar{\beta}$ is the mean of five draws from a uniform distribution on $(0, 1)$. $\beta_{it}$ is initialized with a draw from a standard normal distribution ($\beta_{i,0} \sim N(0, 1)$). $\nu_{it}, \epsilon_{it}, x_{it}$ are draws from standard normal random variables.

**v) White noise:** The fifth data generating process is a cross-sectionally independent white noise process given by:

$$y_{it} = \sigma\epsilon_{it} + \delta D_{it} \tag{8}$$

with $D_{it} = I[i = 1] \cdot I[t > T_0]$ and $\epsilon_{it}$ being standard normal.

The magnitude of the treatment effect may be interpreted relative to the remaining variation in the process (for example for the White noise process $\delta/\sigma$). Further, it seems plausible to relate the (average) bias to the size of the (average) treatment effect $\frac{(T-T_0)\sum_{t=T_0+1}^{T} Bias(t)}{(T-T_0)\sum_{t=T_0+1}^{T} TE(t)}$.

The main hypothesis are:

**H1:** The estimator is unbiased across setups.

**H2:** The estimators RMSE is decreasing in $T_0$ and $N$ and increasing in the error variance.

**H3:** The actual size of the placebo test is $1/N$.

**H4:** The power of the placebo test is increasing in $T_0$, $N$ and $\delta$.

## 3.2 Results

The following table presents the results with respect to the bias and root mean squared error of the synthetic control estimator.

Table 1: Bias and RMSE of the estimator

| DGP | T=8,N=8, $\delta = 0.5$ Bias (RMSE) | T=8,N=16, $\delta = 0.5$ Bias (RMSE) | T=16,N=8, $\delta = 0.5$ Bias (RMSE) | T=16,N=16, $\delta = 0.5$ Bias (RMSE) | T=8,N=8, $\delta = 5$ Bias (RMSE) | T=8,N=16, $\delta = 5$ Bias (RMSE) | T=16,N=8, $\delta = 5$ Bias (RMSE) | T=16,N=16 $\delta = 5$ Bias (RMSE) |
|---|---|---|---|---|---|---|---|---|
| Differences-in-Differences; low | 0.03 | -0.00 | -0.04 | 0.01 | 0.03 | -0.02 | -0.01 | -0.01 |
| | (1.28) | (1.28) | (1.25) | (1.20) | (1.27) | (1.27) | (1.23) | (1.21) |
| Differences-in-Differences; high | 0.05 | -0.14 | -0.04 | 0.05 | 0.05 | 0.08 | -0.04 | 0.03 |
| | (2.52) | (2.46) | (2.45) | (2.33) | (2.50) | (2.50) | (2.45) | (2.33) |
| Stationary heterogeneous ADLX; low | 0.02 | -0.18 | 0.25 | -0.18 | 0.22 | -0.45 | 0.05 | -0.55 |
| | (2.36) | (2.35) | (3.16) | (2.75) | (3.87) | (3.54) | (4.08) | (3.84) |
| Stationary heterogeneous ADLX; high | -0.02 | -0.29 | 0.23 | -0.15 | 0.47 | 0.01 | 0.48 | -0.38 |
| | (3.60) | (3.43) | (4.36) | (4.19) | (4.77) | (4.45) | (5.45) | (5.32) |
| Non-stationary heterogeneous ADLX; low | 0.26 | -0.01 | 0.04 | 0.40 | 2.68 | 2.34 | 2.93 | 2.76 |
| | (2.90) | (2.62) | (4.36) | (3.58) | (4.27) | (3.76) | (5.35) | (4.76) |
| Non-stationary heterogeneous ADLX; high | 0.17 | 0.08 | 0.11 | 0.61 | 2.49 | 2.39 | 3.39 | 2.95 |
| | (4.72) | (4.42) | (7.10) | (5.78) | (5.76) | (5.27) | (7.47) | (6.79) |
| Random coefficient model; low | 0.02 | 0.06 | -0.02 | 0.02 | -0.02 | -0.01 | 0.00 | -0.05 |
| | (2.24) | (2.21) | (2.25) | (2.18) | (2.24) | (2.24) | (2.22) | (2.19) |
| Random coefficient model; high | -0.08 | -0.02 | 0.04 | 0.04 | -0.11 | -0.02 | 0.03 | 0.00 |
| | (3.10) | (2.98) | (3.10) | (2.95) | (3.09) | (3.11) | (3.04) | (2.89) |
| White noise; low | -0.01 | 0.07 | 0.00 | 0.00 | 0.02 | 0.00 | 0.01 | -0.01 |
| | (1.28) | (1.23) | (1.21) | (1.16) | (1.22) | (1.23) | (1.23) | (1.18) |
| White noise; high | -0.02 | 0.07 | -0.05 | -0.02 | -0.08 | 0.00 | 0.02 | 0.08 |
| | (2.43) | (2.41) | (2.44) | (2.32) | (2.50) | (2.40) | (2.43) | (2.34) |

The first number displayed in each cell refers to the bias. The second number displayed in parentheses refers to the RMSE. "Low" refers to a "low" error variance, which was chosen to be one ($var(\sigma\epsilon_{it}) = 1$). The "high" indicates an error variance of four ($var(\sigma\epsilon_{it}) = 4$). The number of repetitions was chosen to be 500 and the bias and RMSE formulas are presented in appendix 1. The length of $T_0 = T/2$.

**Results concerning unbiasedness and RMSE**

Unbiasedness

The $Bias_t = E(\hat{TE}(t) - TE(t))$ is approximated by $Bias_t = 500^{-1} \sum_{r=1}^{500} (\hat{TE}(r,t) - TE(r,t))$ which may be justified by the law of large numbers. Further if not mentioned differently "bias" refers to the average bias over time ($Bias = (T - T_0)^{-1} \sum_{t=T_0+1}^{T} Bias_t$). For details the reader may consult the formulas in appendix 1. The actual implementation of 500 repetitions trades off accuracy and computing time of the 120 setups. One robustness check with 2000 repetitions yielded very similar results as the ones with 500 repetitions. The main results concerning unbiasedness are listed below.

- The synthetic control estimator appears essentially unbiased for all $(T_0, N, \delta, \sigma)$ combinations under consideration given non-autoregressive data generating processes.

- For the stationary autoregressive processes there appears to be a small sample bias, which is small in magnitude and not proportionally increasing with the treatment effect. The most extreme absolute bias of -0.55 corresponds to 6% of the average treatment effect.[6] A robustness check for the stationary heterogeneous ADLX setup with $T = 16$, $N = 16$, $\delta = 5$ and low idiosyncratic variance based on 2000 repetitions yielded a bias of -0.61, which supports the validity of the $R = 500$ results.

- For the non-stationary autoregressive processes the bias seems to be substantial. The most extreme absolute bias is 3.39 and corresponds to 15% of the average treatment effect. In the T=8, N=8, $\delta = 5$ setting the bias of 2.68 corresponds to 21% of the average treatment effect.[7] This indicates some spuriousness in the regression, which is a very common finding in the econometrics literature for independent unit root processes and deserves further investigation.

- Results not displayed based on the T=16, N=16, $\delta = 5$ setups do not indicate a clear pattern of the magnitude of the bias depending on whether one estimates the bias at $T_0 + 1, T_0 + 2, ..., T$. This indicates that the method is similarly (un)biased shortly after treatment and further apart. This is however due to the idealized Monte Carlo setup, with no confounding further treatments for the treatment group or the synthetic control. In reality however the likelihood of further treatments affecting outcomes increases over time. The RMSE is however higher in periods further apart from the initial treatment.

Hypothesis 1, the unbiasedness across all setups, seems to hold only for the static processes and is strongly rejected for the independent unit root case.

[6] The treatment effect formulas are recursive and outlined in the appendix. The expected value of the autoregressive parameter is 0.5. The T=16, $\delta = 5$ setting corresponds to an average treatment effect of 8.75.

[7] The average treatment effect in the non-stationary autoregressive setup is given by $\delta \frac{(T-T_0)\cdot(T-T_0+1)}{2\cdot(T-T_0)}$.

Root mean squared error

Given the synthetic control estimator is unbiased the RMSE provides its standard deviation. Given bias the RMSE can be perceived as a penalty term which equally weighs the squared bias and the variance. Generally the RMSE is seen as an operationalization of the estimators overall performance (high RMSE indicates bad performance). Further if not mentioned differently "RMSE" refers to the average RMSE (averaged over the treatment period - see formula in appendix 1).

*The effects of $\sigma$ and $N$ on RMSE*

- $\sigma \uparrow \rightarrow RMSE \uparrow$: The estimators overall performance is decreasing in the variance of idiosyncratic shocks, across all setups.
- $N \uparrow \rightarrow RMSE \downarrow$: Except in two cases the overall performance is non-decreasing in the size of the donor pool $N$. This result is intuitive, since higher $N$ increases the likelihood that the treatment groups pre-treatment outcomes belong to the convex hull of the donor pools pre-treatment periods outcomes ($P(y_1 \in Co(\{y_2, ..., y_N\}))$ with $y_i = [y_{i1}, ..., y_{iT_0}]'$ is weakly increasing in $N$).

*The effects of $T$ and $T_0$ on RMSE*

- $T \uparrow \rightarrow RMSE \downarrow$ (Non-autoregressive models): For the non-autoregressive models the RMSE is generally decreasing with $T$.
- $T \uparrow \rightarrow RMSE \uparrow$ (Autoregressive models): For the autoregressive models the RMSE is generally increasing with $T$. By construction of the DGP, if $T$ increases the pre-treatment period and the post treatment period increases. If $T_0$ increases the estimator has more observations to fit the weights and identifies the optimal convex combination more precisely. This should lead to a lower RMSE. But if the autoregressive parameter $\alpha_1$ is different from $\sum_{i=2}^{N} w_i \alpha_i$ errors tend to propagate and the difference between the estimated and true treatment effect generally increases over time. This increases the RMSE.
- $T_0 \uparrow, \bar{T} \rightarrow RMSE$: To isolate the effect of solely increasing $T_0$ I compare the RMSE of the estimator at $T_0 + 1$ for all $N = 16$, $\sigma = 1$, $\delta = 5$ setups varying $T_0 \in \{4, 8\}$. Unexpectedly the results divide again into the non-autoregressive and the autoregressive processes. For the non-autoregressive processes a larger pre-treatment period (as always local, given a small sample setting) decreases the RMSE. For the autoregressive models surprisingly a local increase from $T_0 = 4$ to $T_0 = 8$ leads to a higher RMSE (for the non-stationary case the increase is more substantial). There are several potential explanations. First the estimator may behave non-monotonically, more exactly one may suspect that the increase in the RMSE is only local and further increases in $T_0$ would lead to lower root mean squared errors. Second,

as indicated above since $P(y_1 \in Co(\{y_2, ..., y_N\}))$ decreases with $T_0$ this effect may dominate the gain in precision in identifying the optimal linear combination for autoregressive models. Finally, the result could be by chance and deserves further investigation.

*Further results concerning RMSE*

- Evolution of RMSE over $T_0 + 1, ..., T$: Results not displayed based on the $T = 16$, $N = 16$, $\delta = 5$ setups show that for the non-autoregressive models the root mean squared error is very similar across time ($RMSE_{T_0+1} \approx ... \approx RMSE_T$). For the autoregressive models the root mean squared error is substantially and monotonically increasing over time (for the stationary model the RMSE is roughly 50% higher in $T$ than in $T_0 + 1$ , and for the non-stationary model the increase is roughly between 50% (low variance) and 90% (high variance).

- Comparison across DGPs: A simple comparison across DGPs indicates that the estimator performs better in the Difference-in-Difference and the white noise setup than in the random coefficient model. For the autoregressive models the RMSE are highest. It should however be clear that the results most likely depend on the exact parametrization of the data-generating process. Identifying deep parameters which drive the RMSE would be of interest, but is not attempted.

The joint hypothesis 2, that the estimators RMSE is decreasing in $T_0$ and $N$ and increasing in the error variance, seems only partially correct. The RMSE is increasing in the error variance and decreasing in the size of the donor pool for all data generating process. For static models the RMSE is decreasing with the length of the pre-treatment period, whereas for the autoregressive models the RMSE seems to increase (at least locally).

**Results of the placebo test**

Table 2: Size and Power of the placebo test. $\delta \in \{0, 0.5, 5\}$

| DGP \ Share (Rank $\leq \frac{1}{8} \cdot N$) | T=8, N=8 | T=8, N=16 | T=16, N=8 | T=16, N=16 |
|---|---|---|---|---|
| Differences-in-Differences; low | (0.12,0.13,0.37) | (0.13,0.13,0.19) | (0.14,0.16,0.89) | (0.12,0.13,0.92) |
| Differences-in-Differences; high | (0.12,0.14,0.24) | (0.14,0.13,0.17) | (0.12,0.13,0.64) | (0.13,0.12,0.62) |
| Stationary het. ADLX; low | (0.13,0.13,0.04) | (0.16,0.11,0.01) | (0.12,0.15,0.46) | (0.13,0.14,0.37) |
| Stationary het. ADLX; high | (0.15,0.15,0.12) | (0.13,0.14,0.04) | (0.13,0.13,0.46) | (0.14,0.14,0.37) |
| Non-stationary het. ADLX; low | (0.16,0.15,0.08) | (0.15,0.12,0.00) | (0.12,0.17,0.69) | (0.11,0.18,0.72) |
| Non-stationary het. ADLX; high | (0.14,0.15,0.15) | (0.13,0.13,0.05) | (0.13,0.17,0.66) | (0.11,0.16,0.75) |
| Random coefficient model; low | (0.13,0.13,0.24) | (0.12,0.11,0.21) | (0.12,0.14,0.56) | (0.14,0.13,0.51) |
| Random coefficient model; high | (0.10,0.13,0.23) | (0.12,0.12,0.15) | (0.13,0.14,0.52) | (0.09,0.13,0.46) |
| White noise; low | (0.15,0.15,0.41) | (0.12,0.14,0.20) | (0.13,0.15,0.91) | (0.11,0.15,0.92) |
| White noise; high | (0.12,0.13,0.25) | (0.12,0.10,0.19) | (0.12,0.15,0.66) | (0.11,0.13,0.64) |

Each of the three number displayed in each cell corresponds to the share of the repetitions, where the treated unit is ranked first (N=8) or first or second (N=16) (indicating the highest or second highest post- to pre-RMSE ratio). The first number refers to the results of the no treatment setting ($\delta$=0) and can be regarded as an assessment of the actual size of the placebo test. Number two and three indicate the power of the test for medium sized ($\delta$=0.5) and strong treatment effects ($\delta$=5). "Low" refers to a "low" error variance, which was chosen to be one (var($\sigma\epsilon_{it}$)=1). "High" indicates a error variance of four (var($\sigma\epsilon_{it}$)=4). The number of repetitions was chosen to be 500. The rank formula is provided in appendix 1 and $T_0$=T/2 .

Size: In all setups the nominal size is $\frac{1}{8} = 0.125$. The first number in the cells of Table 2 corresponds to the setup with true null hypothesis: $H_0 : TE_t = 0 \; \forall t > T_0$. Generally speaking the actual size seems to be close to the nominal size for most setups, which supports hypothesis 3. No clear pattern concerning the slight size distortions in some setups is apparent.

Power: The second and third number in each cell correspond to setups where the alternative hypothesis is true. The second number refers to the local alternative with $\delta$=0.5, the third number to the local alternative with $\delta$=5. For a small treatment effect ($\delta$=0.5) the rejection rates are close to the nominal size. To put this into perspective note that simulated t-tests based on the least squares dummy variable estimator (LSDV) with common time effects in the T=8, N=8, $\delta$=0.5 Differences-in-Differences setting with high error variance reject the null-hypothesis in only 10% of the cases, given nominal size of 12.5% (critical value=1.535). For the same parameters and a White noise DGP, the t-test based on simple pooled OLS regressions rejects in 18% of the cases. This indicates that the placebo test does perform similarly bad as regression based t-tests for small samples (N·T=64) and small treatment effects. For a high treatment effect ($\delta$=5 ) and T=16, the power of the test ranges from low (.37) to high (.92). Again to put the results into perspective OLS based t-tests in the T=16, N=16, $\delta = 5$ White noise setting were simulated. The rejection rate was

100% given a critical value of 1.535 irrespective of the error variance. The corresponding t-tests for the Differences-in-Differences DGP based on the LSDV estimator produced rejection rates close to one. This indicates that the placebo test performs worse than a t-test for high treatment effects and medium sized samples (N·T=256). The fact that the power for the non-stationary process is much higher than for the stationary process may be explained by the more substantial upward bias of the treatment effect estimator in the non-stationary setup. For a high treatment effect and T=8 the power ranges from very low to medium. For the autoregressive models the T=8, N=16 results are again puzzling. In this case the test has virtually no power.[8] Hypothesis 4, that the power of the placebo test is increasing in $T_0$, $N$ and $\delta$, seems only partially true. The power is increasing in $T_0$ and $\delta$, but is not uniformly increasing in $N$.

**Conclusion of the Monte Carlo study**

A generally promising result is the unbiasedness in the non-autoregressive models and the small bias in the stationary autoregressive model. Nevertheless, researchers should be aware of the substantial RMSEs. The lack of standard error estimates to operationalize uncertainty about the estimates is only partially compensated by the placebo test, which has generally good size and mediocre power properties for T=16. For non-stationary processes (more exactly independent unit roots) the results are worse. The estimator is substantially biased and the RMSE high. As a preliminary recommendation researchers may apply panel unit root tests like the one in Im, Pesaran and Shin (2003), which is implemented in Stata, or the one proposed in Pesaran (2007) to receive an indication of the presence of a unit root. However, the power properties of these tests for processes with substantial persistence ($\alpha_i \approx 0.9$) are weak in small samples (T=10, N=10) and the test may be non-informative.[9] Acquiring non-sample information on the process is often useful, but may be hard to come by.[10]

After finding that unit roots pose a problem for the synthetic control estimator, which is otherwise at most slightly biased, the application in the next section will apply panel unit root tests and the synthetic control method to estimate the effects of outsourcing probation and parole services on revoke rates. Additionally the low power results of the Monte Carlo study indicate that we cannot expect the placebo test to indicate an effect of the privatization on the revoke rates unless the effect is large.

[8] Further, for these setups the distribution of the share of ranks seems closer to a normal distribution, than to a uniform distribution or left skewed distribution, which frequently appear for the $H_0$ and $\delta$=5 setups.

[9] I thank Mehdi Housseinchouak for highlighting this point.

[10] Given one is convinced of the (independent) unit root nature of the processes or attempts another robustness check against unit roots, a further thought based on analogizing from the spurious regression phenomenon is the possibility to first difference the outcomes in the pre-treatment period, estimate the weights and then use those weights to estimate the treatment effects. If the restricted linear projection used by the synthetic control estimator is close to an unrestricted linear projection one would expect that the bias is reduced by the differencing strategy. If the processes would however be cointegrated instead of having independent unit roots I would guess that differencing removes useful information to pin down the appropriate weights.

# 4 Application to revoke rates after outsourcing probation services

According to the federal statistical office around 146,000 persons in West-Germany (without Hamburg) were on parole or probation in 2011. Around 60,000 probations or paroles were terminated, whereof 15,000 of the probations or paroles were revoked either because of new crimes or other misconduct. On 01/01/2007 one German state (Baden-Wuerttemberg) administered a transition from public to private probation and parole services, by outsourcing the entire responsibility for the service to the Neustart gGmbH.[11]

Internationally, private probation services are not uncommon. Hardly surprising is the fact that private organizations are responsible for probation services in several US states. Recently England and Wales outsourced probation services and reorganized probation for medium and low risk offenders by establishing 21 Community Rehabilitation Companies, which were sold by the state to private firms and compensation will be realized on a payment by results basis (Ministry of Justice UK 2013; Robinson et al. 2016). A "proclaimed aim of this process (involvement of the private sector) is to end up with services that "work" in reducing re-offending and promoting rehabilitation and desistance" (Deering et al. 2014, p. 235). This section aims at answering the question if the outsourcing in Germany reduced parole violations like recidivism.

Unlike the effect of privatization of prisons the literature on the effect of private probation services seems not as voluminous. Some work is done on the evolution of criminal justice identities in the process of privatization (Deering et al. 2014 or Robinson et al. 2016) or the comparison of standards concerning private probation services in the US (Schloss and Alarid 2007). Some further articles relating to privatization and probation can be found in the Probation Journal (among others Teague 2011). Except for the publication of Dölling, Entorf and Hermann (2014), I am not aware of a similar quantitative analysis of the effect of outsourcing probation services on revoke rates.[12]

A commonly used performance measure is the ratio of terminations of probation or parole without success to total terminations within a year, which is called *revoke rate*. Several modifications of this measure listed below are common and employed in the analysis. Furthermore, Table 3 provides summary statistics for the treatment group and the average of the donor pool in the years 2006 and 2007, based on the Rechtspflegestatistik of the federal statistical office.

[11] From 01/01/2005 to 31/12/2006 the Neustart gGmbH was responsible for the parole and probation services in Stuttgart and Tuebingen as part of a Pilot project. This fact is neglected in the analysis. Stuttgart and Tuebingen represent roughly 6.5% of the population in Baden Wuerttemberg.

[12] A draft of the section in Ministry of Justice Baden Wuerttemberg (2014) and Dölling, Entorf and Hermann (2014) and the analysis were performed by the author of this paper. Slightly oversimplifying the findings of the section in the report the treatment effect estimates were generally found to be positive indicating slightly higher revoke rates, due to the Neustart gGmbH (however no direct causal claim is made).

Table 3: Revoke Rates

| Type of revoke rate | Purpose | BW 2006 | BW 2007 | Donors av. 2006 | Donors av. 2007 |
|---|---|---|---|---|---|
| Total revoke rate | Performance measure for all clients of the parole and probation services | 0.196 | 0.179 | 0.243 | 0.239 |
| Total revoke rate for the general criminal law | Performance measure specific to those convicted under general criminal law | 0.223 | 0.206 | 0.280 | 0.270 |
| Total revoke rate for juvenile criminal law | Performance measure specific to those convicted under juvenile criminal law | 0.134 | 0.115 | 0.159 | 0.162 |
| Revoke rate based on new crimes for the general criminal law | Performance measure based on new crimes specific to those convicted under general criminal law | 0.178 | 0.165 | 0.223 | 0.211 |
| Revoke rate based on new crimes for juvenile criminal law | Performance measure based on new crimes specific to those convicted under juvenile criminal law | 0.092 | 0.084 | 0.105 | 0.103 |

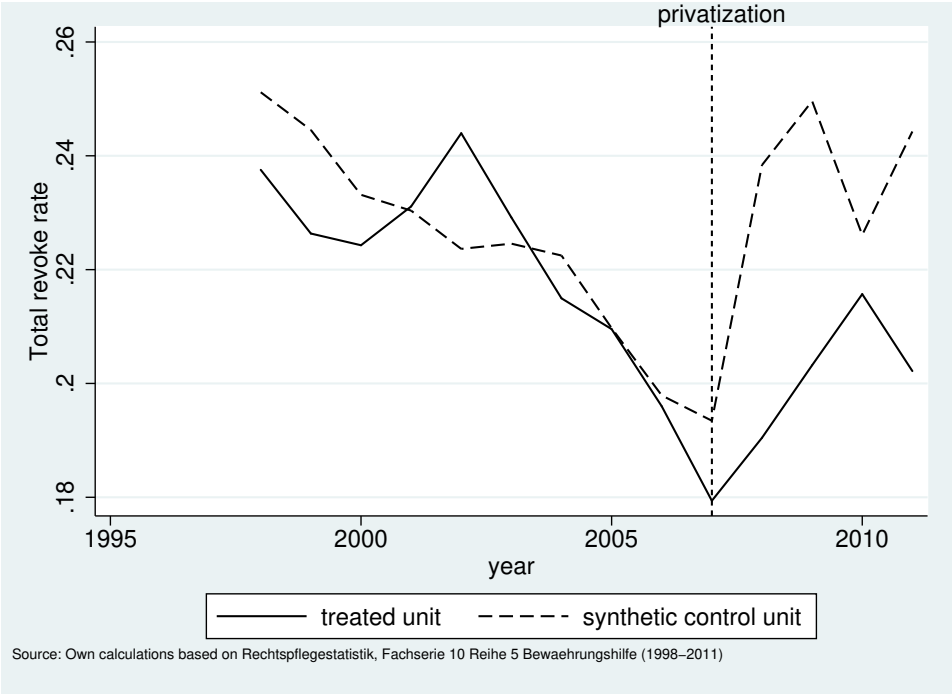
Source: Own calculations on the basis of the Rechtspflegestatistik, Fachserie 10 Reihe 5 "Bewährungshilfe" (1998-2011) provided by the federal statistical office. Precise definitions are provided in appendix (2). Several data problems arise. For some states no data is available (mainly East German states) and for some states data is partially missing. The remaining donor pool includes: Bavaria, Bremen, Hesse, Lower Saxony, Northrhine-Westphalia, Rhineland Palatine, Saarland, Mecklenburg-Western Pomerania. Donors av. 2006 represents the cross sectional average of the corresponding revoke rate in 2006 for the eight donor states.

## 4.1 Results concerning the overall revoke rate

To perform the analysis I used the Stata synth program provided by Hainmueller using the default settings with the pre-intervention outcomes from 1998 to 2006 as predictor variables.[13] As discussed above using pre-intervention outcomes is motivated by potential dependence between unobservables in treatment and control group. The following graph depicts the main results of the synthetic control estimates for the total revoke rate.

[13]The program was downloaded from http://web.stanford.edu/~jhain/synthpage.html on 7th Sept. 2015.

Figure 1: Development of the true and synthetic total revoke rate from 1998 to 2011




The pre-treatment fit is mediocre. Differences between the treated unit and its synthetic control in the pre-treatment period are up to two percentage points. The treatment effect estimates range from -1 to -4.8 percentage points. Its mean is -3.22 percentage points per year.[14] For an economist the composition of the synthetic control group as roughly one fifth Hesse one fourth Saarland and the remainder as Mecklenburg Western-Pommerania is rather puzzling.[15] The economic prosperity of the regions is diverse, but potentially other driving factors of total revoke rates are similar.[16]

To check for unit roots in the total revoke rates the Im-Pesaran-Shin test implemented in Stata, which allows for heterogeneous slopes was performed. It rejects the null that all panels contain unit roots just at the 10% level (p-value 9.23 %). This can be regarded as some evidence against the unit root problem outlined in the Monte Carlo section. Given that Im, Pesaran and Shin (2003) demonstrates a rather low power of the test in a N=10, T=10 setting with $\alpha_i = 0.9$, the p-value of 0.09 may be regarded as evidence for rather low persistence, which supports the use of the synthetic control method.

Reassigning the treatment to 2003 or 2005 delivers treatment effect estimates in 2003 or 2005 which are close to zero (for details see appendix 3) and Baden-Wuerttemberg has the second highest

[14] Using the Matlab m-file results in a very similar mean treatment effect of -0.0319, which can be seen as evidence that the results are not driven by the choice of the optimizer.

[15] Using the Matlab m-file yields similar weights, except for the inclusion of Rhineland Palatine with a weight of 2.6%.

[16] The revoke rate of Baden-Wuerttemberg is generally at the lower end of the distribution, frequently placing second to third lowest. Mechanically, since weights for the synthetic control group are restricted to [0,1] the weights for Mecklenburg Western-Pommerania and Saarland are explained by their low revoke rates.

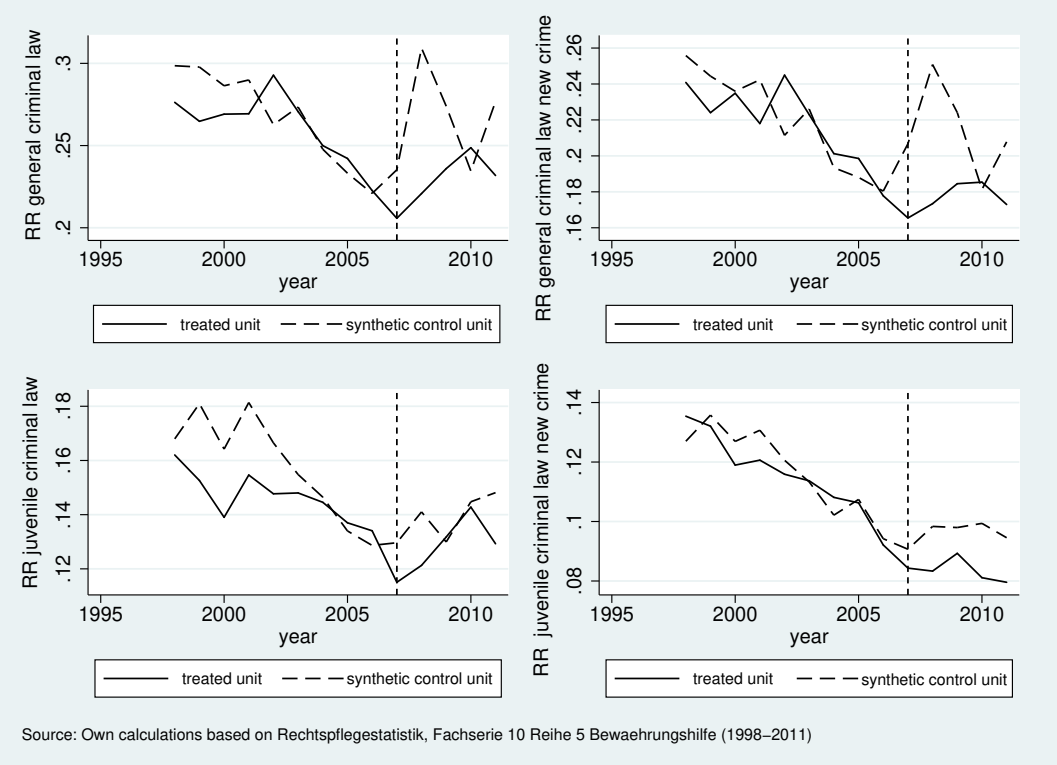
RMSE ratio (3.22). Only Bremen has a higher RMSE ratio (3.48). This corresponds to a p-value of 2/9 ($\approx 0.22$) under the null hypothesis of no treatment effect.[17]

A conservative interpretation of the synthetic control results is that an increase of the total revoke rate due to outsourcing probation and parole services is quite unlikely. Focusing solely on the point estimates the synthetic control methodology does indicate a reduction in the total revoke rate, whereas the ad-hoc choice of Bavaria or West-Germany (excluding Baden-Wuerttemberg and Hamburg) in Dölling, Entorf and Hermann (2014) suggests slight increases in the total revoke rate due to the outsourcing. Merging the evidence suggests that strong increases in the total revoke rate due to the outsourcing seem rather unlikely in the Baden-Wuerttemberg - Neustart gGmbH case.

## 4.2 Juvenile, general criminal law and crime specific results

Figure 2 depicts the synthetic control estimates for the total revoke rates for the general criminal law and the revoke rate for the general criminal law due to new crimes of the probationer (and parolee) in the upper half and the corresponding revoke rates for the juvenile criminals in the lower half.

Figure 2: Crime specific and general revoke rates for juvenile and general criminal law




[17] I implemented the placebo test in Matlab.

The composition of the synthetic control group, the likelihood of unit roots and the RMSE ratio rank varies across revoke categories and are summarized in Table 4.

**Table 4: Composition of the synthetic control group, unit roots and RMSE-ratio**

| Category | Synthetic control group | p-value IPS unit root test | RMSE ratio rank |
|---|---|---|---|
| Total revoke rate for the general criminal law | Bremen 2%; Hesse 46%; Saarland 50% | 6.5% | 3 |
| Revoke rate based on new crimes for the general criminal law | Bavaria 25%; Hesse 28%; Saarland 47% | 30.4% | 3 |
| Total revoke rate for juvenile criminal law | Bremen 15%; Rhineland-Palatine 21%; Saarland 41%; Mecklenburg-Western Pomerania 23% | 0.3% | 9 |
| Revoke rate based on new crimes for juvenile criminal law | Bavaria 45%; Hesse 9%; Rhineland-Palatine 21%; Saarland 18%; Mecklenburg-Western Pomerania 7% | 34.8% | 6 |

All graphs indicate a negative treatment effect (averaged over time), which suggests lower revoke rates due to the outsourcing of probation and parole services. The pre-treatment fit seems acceptable for the general criminal law and the revoke rate due to new crimes for those convicted after juvenile criminal law. The relatively low ranks indicate high uncertainty about the estimated reduction in the revoke rates due to the Neustart gGmbH. However as emphasized in the Monte Carlo section the power of the RMSE ratio "test" is rather low for medium sized treatment effects. For the revoke rates based on new crimes, evidence against unit roots is weak and results may be spurious. But as discussed the high p-values may merely reflect a low power of the unit root test.

# 5 Conclusion

Theoretical or Monte Carlo based results on the small sample properties of the synthetic control estimator are scarce and limited to factor models and homogeneous stationary AR processes. Consequently, this paper performed a small sample Monte Carlo analysis using data generating processes based on factor and random coefficient models as well as stationary and non-stationary heterogeneous ADLX processes. The simulations give rise to the conclusion that the synthetic control estimator is unbiased for the non-autoregressive processes. For stationary heterogeneous autoregressive processes the bias is small relative to the treatment effect, but independent unit roots pose a problem for the synthetic control estimator and induce a substantial bias. One preliminary recommendation for applied researchers is to perform panel unit root tests to assess the time series properties of the outcome variable. It should though be noted that such tests frequently have low power in small sample settings. The size properties of the main placebo test are good and the power properties are mediocre. If treatment effects are rather small in size it is worth noting that the placebo test is unlikely to detect the presence of a treatment effect.

After performing the Monte Carlo study the synthetic control methodology was applied to estimate the effects of outsourcing probation and parole services in Germany on revoke rates. Overall, the findings of the application section provide evidence against the hypothesis that the outsourcing increased the revoke rates.

Several interesting research questions remain unanswered. First, one could investigate whether the minor small sample bias for heterogeneous stationary autoregressive processes is driven by the occurrence of highly persistent series. Since the autoregressive parameters were drawn from a uniform distribution on (0,1) this could indeed be the case. Second, further research is needed to find out how the synthetic control estimator behaves if the time series are cointegrated instead of having independent unit roots and how well a strategy of first differencing the outcomes and then applying the synthetic control estimator works. And third, it is not clear how well the estimator performs if researchers opt for a sequential procedure by first applying a panel unit root test and then analyze the data only if the presence of unit roots is rejected at some pre-specified significance level.

# References


[1] Abadie, A. 2015. Using Synthetic Controls to Evaluate an International Strategic Positioning Program in Uruguay: Feasibility, Data Requirements, and Methodological Aspects, Study commissioned by the Inter-American Development Bank. http://idbdocs.iadb.org/wsdocs/getdocument.aspx?docnum=37312864, last accessed 11/01/2016.

[2] 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(490): 493–505.

[3] Abadie, A., Diamond, A., and Hainmueller, J. 2015. Comparative Politics and the Synthetic Control Method. *American Journal of Political Science* 59(2): 495–510.

[4] Abadie, A., and Gardeazabal, J. 2003. The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review* 93(1): 112–132.

[5] Bohn, S., Lofstrom, M., and Raphael, S. 2014. Did the 2007 Legal Arizona Workers Act reduce the state's unauthorized immigrant population? *Review of Economics and Statistics* 96(2): 258-269.

[6] Carvalho, C. V., Masini, R., and Medeiros, M. C. 2014. Synthetic Control Theory and Inference for Stationary Processes. NBER-NSF Time Series Conference, Federal reserve Bank of St. Louis 2014.

[7] Deering, J., Feilzer M., and Holmes, T. 2014. The transition from public to private in probation: Values and attitudes of managers in the private sector. *Probation Journal* 61(3): 234-250.

[8] Dölling, D., Hermann, D., and Entorf, H. 2014. Evaluation der Bewährungs- und Gerichtshilfe sowie des Täter-Opfer-Ausgleichs in Baden Württemberg - Abschlussbericht. *http://www.uni-heidelberg.de/institute/fak2/krimi/Evaluation%20der%20BWH_GH_TOA.pdf*, last accessed 11/01/2016.

[9] Gobillon, L., and Magnac, T. 2015. Regional Policy Evaluation: Interactive Fixed Effects and Synthetic Controls. *Review of Economics and Statistics* (forthcoming).

[10] Kaul, A., Klößner, S., Pfeifer, G., Schieler M. 2015. Synthetic Control Methods: Never Use All Pre-Intervention Outcomes as Economic Predictors. http://www.oekonometrie.uni-saarland.de/papers/SCM_Predictors.pdf Version: May 27 2015, last accessed 11/01/2016.

[11] Ministry of Justice Baden Württemberg. 2014. Evaluation der Bewährungs- und Gerichtshilfe sowie des Täter-Opfer-Ausgleichs in Baden Württemberg. Justizministerium Baden Württemberg 2014.

[12] Ministry of Justice United Kingdom. 2013. Transforming Rehabilitation: A revolution in the way we manage offenders. *https://www.gov.uk/government/uploads/system/uploads/attachment_ data/file/228580/8517.pdf*, last accessed 11/01/2016.

[13] Im, K. S., Pesaran, M. H., and Shin, Y. 2003. Testing for unit roots in heterogeneous panels. *Journal of Econometrics* 115(1): 53-74.

[14] Pesaran, M. H. 2007. A simple panel unit root test in the presence of cross-section dependence. *Journal of Applied Econometrics* 22(2): 265-312.

[15] Robinson, G., Burke, L., and Millings, M. 2016. Criminal justice identitites in transition: The case of devolved probation services in England and Wales. *British Journal of Criminology* 56(1): 161-178.

[16] Schloss, C. S., and Alarid, L. F. 2007. Standards in the Privatization of Probation Services: A statutory Analysis. *Criminal Justice Review* 32(3): 233-245.

[17] Teague, M. 2011. Probation in America: Armed, private and unaffordable? *Probation Journal* 58(4): 317-332.

[18] Xu, Y. 2015. Generalized Synthetic Control Method for Causal Inference with Time-Series Cross-Sectional Data. MIT Political Science Department Research Paper No. 2015-1.

# Appendix

**(1) Formulas**

The following formulas correspond to the verbal descriptions in the Monte Carlo section ($R$ denotes the number of repetitions in the Monte Carlo simulation):

$$TreatmentEffect = E(y_{1,r,t}(D = [1, ..., 1]) - y_{1,r,t}(D = [0, ..., 0])) \tag{9}$$

$$Bias = R^{-1} \sum_{r=1}^{R} \left( (T - T_0)^{-1} \sum_{t=T_0+1}^{T} (Y_{1,1}(t,r) - \hat{Y}_{1,0}(t,r)) - [E(y_{1,r,t}(D=[1,...,1]) - y_{1,r,t}(D=[0,...,0]))] \right) \tag{10}$$

$$RMSE = \sqrt{(R \cdot (T - T_0))^{-1} \sum_{r=1}^{R} \sum_{t=T_0+1}^{T} ((Y_{1,1}(t,r) - \hat{Y}_{1,0}(t,r)) - [E(y_{1,r,t}(D=[1,...,1]) - y_{1,r,t}(D=[0,...,0])])^2} \tag{11}$$

$$Rank = Rank \left( \frac{\sqrt{(T - T_0)^{-1} \sum_{t=T_0+1}^{T} (Y_{1,1}(t) - \hat{Y}_{1,0}(t))^2}}{\sqrt{(T_0)^{-1} \sum_{t=1}^{T_0} (Y_{1,0}(t) - \hat{Y}_{1,0}(t))^2}} \right) \tag{12}$$

Formula (12) provides the rank of unit 1 based on the RMSPE in the post-treatment period divided by the RMSPE in the pre-treatment period. The treatment effect expressed in parameters of the data generating processes differs across the static and dynamic data generating processes. The treatment dummy can be written as a $1 \times T$ vector consisting of $T_0$ zeros and $T - T_0$ ones $[0, ..., 0, 1, ..., 1]$. We can define the treatment indicator in the post treatment period as $D_{post} = [1, ..., 1]$. The corresponding thought experiment would be receiving treatment in each of the post-treatment periods. For the static models (Differences-in-Differences, random coefficient model, white noise) the treatment effect given by

$$E(y_{i,t}(D = [1, ..., 1]) - y_{i,t}(D = [0, ..., 0]) = \delta \ \ \forall t \geq T - T_0. \tag{13}$$

For the dynamic models the treatment effect is more complicated, as it explicitly depends on time and the model parameters:

$$E(y_{i,t}(D = [1, ..., 1]) - y_{i,t}(D = [0, ..., 0]) = I[t \geq T - T_0] \cdot \sum_{j=T-T_0}^{t} \alpha_i^{j-(T-T_0)} \delta \tag{14}$$

**(2) Revoke rate definitions**

The following definitions indicate the categories used from the Rechtspflegestatistik, Fachserie 10 Reihe 5 "Bewährungshilfe" table 3.2. All revoke rates are ratios of the number of certain terminations without success to the number of terminations in an appropriate base category. For illustrative purposes suppose there are two categories of terminations without success (a and b) and one category which reflects the appropriate base (c), then the revoke rate can be expressed as follows: Revoke rate $= {}^{(\#Termination_a + \#Termination_b)}/_{\#Terminations_c}$.

- Total revoke rate: The total revoke rate is the sum over the following categories: *Beendete Unterstellungen unter Bewährungsaufsicht nach allgemeinem und Jugendstrafrecht durch Widerruf nur oder auch wegen neuer Straftat oder aus sonstigen Gründen* and the *Verhängung der Jugendstrafe nur oder auch wegen neuer Straftat oder aus sonstigen Gründen* divided by *Beendete Unterstellungen insgesamt.*

- Total revoke rate for the general criminal law: The total revoke rate for the general criminal law is given by: *Beendete Unterstellungen unter Bewährungsaufsicht nach allgemeinem Strafrecht durch Widerruf nur oder auch wegen neuer Straftat oder aus sonstigen Gründen* divided by *Beendete Unterstellungen nach allgemeinem Strafrecht.*

- Total revoke rate for juvenile criminal law: The total revoke rate for juvenile criminal law is the sum over the following categories: *Beendete Unterstellungen unter Bewährungsaufsicht nach Jugendstrafrecht durch Widerruf nur oder auch wegen neuer Straftat oder aus sonstigen Gründen* and the *Verhängung der Jugendstrafe nur oder auch wegen neuer Straftat oder aus sonstigen Gründen* divided by *Beendete Unterstellungen nach Jugendstrafrecht.*

- Revoke rate based on new crimes for the general criminal law: The revoke rate based on new crimes for the general criminal law is given by: *Beendete Unterstellungen unter Bewährungsaufsicht nach allgemeinem Strafrecht durch Widerruf nur oder auch wegen neuer Straftat* divided by *Beendete Unterstellungen nach allgemeinem Strafrecht.*

- Revoke rate based on new crimes for juvenile criminal law: The revoke rate based on new crimes for the general criminal law is the sum over the following categories: *Beendete Unterstellungen unter Bewährungsaufsicht nach Jugendstrafrecht durch Widerruf nur oder auch wegen neuer Straftat* and *Verhängung der Jugendstrafe nur oder auch wegen neuer Straftat* divided by *Beendete Unterstellungen nach Jugendstrafrecht.*

**(3) Further robustness checks**

As stated in the main text there appears no immediate treatment effect in 2003 or 2005 (see Figure 3). This may be perceived as evidence against spuriousness of the results with treatment in 2007.

Figure 3: Placebo test in time for the total revoke rate