Comparative Politics and the Synthetic Control Method

Alberto Abadie   Harvard University and NBER
Alexis Diamond   International Finance Corporation
Jens Hainmueller   Stanford University

In recent years, a widespread consensus has emerged about the necessity of establishing bridges between quantitative and qualitative approaches to empirical research in political science. In this article, we discuss the use of the synthetic control method as a way to bridge the quantitative/qualitative divide in comparative politics. The synthetic control method provides a systematic way to choose comparison units in comparative case studies. This systematization opens the door to precise quantitative inference in small-sample comparative studies, without precluding the application of qualitative approaches. Borrowing the expression from Sidney Tarrow, the synthetic control method allows researchers to put "qualitative flesh on quantitative bones." We illustrate the main ideas behind the synthetic control method by estimating the economic impact of the 1990 German reunification on West Germany.

Starting with Alexis de Toqueville's *Democracy in America*, comparative case studies have become closely associated with empirical research in political science (Tarrow 2010). Comparative researchers base their studies on meticulous description and analysis of the characteristics of a small number of selected cases, as well as of their differences and similarities. By carefully studying a small number of cases, comparative researchers gather evidence at a level of granularity that is difficult if not impossible to incorporate in quantitative studies, which tend to focus on larger samples but employ much coarser descriptions of the sample units.[1] However, large-sample quantitative methods are sometimes adopted because they provide precise numerical results, which can be compared across studies, and because they are better adapted to traditional methods of statistical inference.

As a result of a recent and important methodological debate (Beck 2010; Brady and Collier 2004; George and Bennett 2005; King, Keohane, and Verba 1994; Ragin 1987; Tarrow 1995), a widespread consensus has emerged about the necessity of establishing bridges between the quantitative and the qualitative approaches to empirical research in political science. In particular, there have been calls for the development and use of quantitative methods that complement and facilitate qualitative analysis in comparative studies (Gerring 2007; Lieberman 2005; Sekhon 2004; Tarrow 1995, 2010). At the other end of the methodological spectrum, a recent strand of the quantitative literature is advocating for research designs that, like Mill's Method of Difference, carefully select comparison units to reduce biases in observational studies (Card and Krueger 1994; Rosenbaum 2005).

[1] See Lijphart (1971), Collier (1993), Mahoney and Rueschemeyer (2003), George and Bennett (2005), and Gerring (2004, 2007) for careful treatments of case study research in the social sciences.

DOI: 10.1111/ajps.12116

In this article, we discuss how synthetic control methods (Abadie, Diamond, and Hainmueller 2010; Abadie and Gardeazabal 2003) can be applied to complement and facilitate comparative case studies in political science. Following Mill's Method of Difference, we focus on a study design based on the comparison of outcomes between units representing the case of interest, defined by the occurrence of a specific event or intervention that is the object of the study, and otherwise similar but unaffected units. In this design, comparison units are intended to reproduce the counterfactual of the case of interest in the absence of the event or intervention under scrutiny.[2]

The selection of comparison units is a step of crucial importance in comparative case studies because using inappropriate comparisons may lead to erroneous conclusions. If comparison units are not sufficiently similar to the units representing the case of interest, then any difference in outcomes between these two sets of units may merely reflect disparities in their characteristics (Geddes 2003; George and Bennett 2005; King, Keohane, and Verba 1994). The synthetic control method provides a systematic way to choose comparison units in comparative case studies. Formalizing the way comparison units are chosen not only represents a way of systematizing comparative case studies (as advocated, among others, by King, Keohane, and Verba 1994), but it also has direct implications for inference. We demonstrate that the main barrier to quantitative inference in comparative studies comes not from the small-sample nature of the data, but from the absence of an explicit mechanism that determines how comparison units are selected. By carefully specifying how units are selected for the comparison group, the synthetic control method opens the door to the possibility of precise quantitative inference in comparative case studies, without precluding qualitative approaches to the same data set.

One distinctive feature of comparative political science is that the units of analysis are often aggregate entities, such as countries or regions, for which suitable single comparisons often do not exist (Collier 1993; George and Bennett 2005; Gerring 2007; Lijphart 1971). The synthetic control method is based on the premise that, when the units of analysis are a few aggregate entities, a combination of comparison units (which we term "synthetic control") often does a better job of reproducing the characteristics of the unit or units representing the case of interest than any single comparison unit alone. Motivated by this consideration, the comparison unit in the synthetic control method is selected as the weighted average of all potential comparison units that best resembles the characteristics of the case of interest.

Relative to regression-based comparative case studies, the synthetic control method has important advantages. Using a weighted average of units as a comparison precludes the type of model-dependent extrapolation that regression results are often based on (King and Zeng 2006). Below we show that a regression estimator can also be expressed as a weighted average of comparison units, with weights that sum to one. However, regression weights are not restricted to lie between zero and one, allowing extrapolation. Moreover, in contrast to regression analysis techniques, the synthetic control method makes explicit the contribution of each comparison unit to the counterfactual of interest. This allows researchers to use quantitative *and* qualitative techniques to analyze similarities and differences between the unit or units representing the case of interest and the synthetic control.

We apply the synthetic control method to estimate the economic impact of the 1990 German reunification on West Germany. This example illustrates the potential gains derived by using synthetic controls in comparative case studies. We show that no single country is able to closely approximate the values of economic growth predictors for West Germany before the reunification. However, a weighted average of a few OECD countries, namely, Austria, the United States, Japan, Switzerland, and the Netherlands, provides a very close approximation to West Germany prior to 1990. We also show that, in this example, a regression-based analysis relies on extrapolation to construct a comparison unit, whereas the synthetic control method avoids extrapolation. Our results suggest that reunification had a pronounced negative effect on West German income, with gross domestic product (GDP) per capita being reduced by about 1,600 U.S. dollars per year on average over the 1990–2003 period (approximately 8% of the 1990 baseline level). The findings are robust across a series of placebo studies and sensitivity checks.

In this section, we have briefly described the synthetic control method and motivated the use of this method in comparative case studies. Relative to small-sample studies, the synthetic control method helps in the selection of comparison cases and opens the door to quantitative inference. Relative to large-sample regression-based studies,

---

[2] See Fearon (1991) for an early discussion of the role of counterfactuals to assess causal hypotheses in political science. It is important, however, to recognize that comparative politics is "a river of many currents" (Hall 2003, 374), and researchers may have motivations for selecting cases other than the construction of counterfactuals (Collier and Mahoney 1996; Hall 2003). For example, researchers may select cases in order to examine causal mechanisms through within-case methods such as process tracing (George and Bennett 2005) or causal process observations (Collier, Mahoney, and Seawright 2004). We do not intend to criticize these approaches, as in our view our proposal is complementary to existing methods.

the synthetic control method avoids extrapolation biases and allows a more focused description and analysis of the similarities and differences between the case of interest and the comparison unit.

The rest of the article is organized as follows. The next section describes the synthetic control estimator, provides a formal comparison between this estimator and a conventional regression estimator, discusses inferential techniques, and provides recommendations for empirical practice. We then present an empirical application where we apply the synthetic control method to the study of the economic effects of the 1990 German reunification in West Germany. The last section concludes. Data sources for the empirical example are provided in an appendix.

Synthetic Control Method for Comparative Case Studies

Suppose that there is a sample of $J + 1$ units (e.g., countries) indexed by j, among whom unit $j = 1$ is the case of interest and units $j = 2$ to $j = J + 1$ are potential comparisons.[3] Borrowing from the medical literature, we will say that $j = 1$ is the "treated unit," that is, the unit exposed to the event or intervention of interest, and units $j = 2$ to $j = J + 1$ constitute the "donor pool," that is, a reservoir of potential comparison units. Studies of this type abound in political science (Gerring 2007; Tarrow 2010). Because comparison units are meant to approximate the counterfactual of the case of interest without the intervention, it is important to restrict the donor pool to units with outcomes that are thought to be driven by the same structural process as for the unit representing the case of interest and that were not subject to structural shocks to the outcome variable during the sample period of the study. In the application explored later in this article, we investigate the effects of the 1990, German reunification on the economic prosperity in West Germany. In this example, the case of interest is West Germany in 1990 and the set of potential comparisons is a sample of OECD countries.

We assume that the sample is a balanced panel, that is, a longitudinal data set where all units are observed at the same time periods, $t = 1, \ldots, T$.[4] We also assume that

the sample includes a positive number of preintervention periods, T_0, as well as a positive number of postintervention periods, T_1, with $T = T_0 + T_1$. Unit 1 is exposed to the intervention of interest (the "treatment") during periods $T_0 + 1, \ldots, T$, and the intervention has no effect during the pretreatment period $1, \ldots, T_0$. The goal of the study is to measure the effect of the intervention of interest on some postintervention outcome.

As stated above, the preintervention characteristics of the treated unit can often be much more accurately approximated by a combination of untreated units than by any single untreated unit. We define a synthetic control as a weighted average of the units in the donor pool. That is, a synthetic control can be represented by a $(J \times 1)$ vector of weights $W = (w_2, \ldots, w_{J+1})'$, with $0 \le w_j \le 1$ for $j = 2, \ldots J$ and $w_2 + \cdots + w_{J+1} = 1$. Choosing a particular value for W is equivalent to choosing a synthetic control. Following Mill's Method of Difference, we propose selecting the value of W such that the characteristics of the treated unit are best resembled by the characteristics of the synthetic control. Let X_1 be a $(k \times 1)$ vector containing the values of the preintervention characteristics of the treated unit that we aim to match as closely as possible, and let X_0 be the $k \times J$ matrix collecting the values of the same variables for the units in the donor pool. The preintervention characteristics in X_1 and X_0 may include preintervention values of the outcome variable.

The difference between the preintervention characteristics of the treated unit and a synthetic control is given by the vector $X_1 - X_0 W$. We select the synthetic control, W^*, that minimizes the size of this difference. This can be operationalized in the following manner. For $m = 1, \ldots, k$, let X_{1m} be the value of the m-th variable for the treated unit and let X_{0m} be a $1 \times J$ vector containing the values of the m-th variable for the units in the donor pool. Abadie and Gardeazabal (2003) and Abadie, Diamond, and Hainmueller (2010) choose W^* as the value of W that minimizes:

$$\sum_{m=1}^{k} v_m (X_{1m} - X_{0m}W)^2, \quad (1)$$

where v_m is a weight that reflects the relative importance that we assign to the m-th variable when we measure the discrepancy between X_1 and $X_0 W$.[5] It is of crucial

---

[3]For expositional simplicity, we focus on the case where only one unit is exposed to the event or intervention of interest. This is done without a loss of generality. In cases where multiple units are affected by the event of interest, our method can be applied to each affected unit separately or to an aggregate of all affected units.

[4]This is typically the case in political science applications, where sample units are large administrative entities like nation-states

or regions, for which data are periodically collected by statistical agencies. We do not require, however, that the sample periods are equidistant in time.

[5]More formally, let $\| \cdot \|$ be a norm or seminorm in $\mathbb{R}^k$. One example is the Euclidean norm, defined as $\|u\| = \sqrt{u'u}$ for any $(k \times 1)$ vector u. For any positive semidefinite $(k \times k)$ matrix, V, $\|u\| = \sqrt{u'Vu}$ defines a seminorm. The synthetic control

importance that synthetic controls closely reproduce the values that variables with a large predictive power on the outcome of interest take for the unit affected by the intervention. Accordingly, those variables should be assigned large v_m weights. In the empirical application below, we apply a cross-validation method to choose v_m.

Let Y_{jt} be the outcome of unit j at time t. In addition, let Y_1 be a $(T_1 \times 1)$ vector collecting the postintervention values of the outcome for the treated unit. That is, $Y_1 = (Y_{1\,T_0+1}, \ldots, Y_{1\,T})'$. Similarly, let Y_0 be a $(T_1 \times J)$ matrix, where column j contains the postintervention values of the outcome for unit $j + 1$. The synthetic control estimator of the effect of the treatment is given by the comparison of postintervention outcomes between the treated unit, which is exposed to the intervention, and the synthetic control, which is not exposed to the intervention, $Y_1 - Y_0 W^*$. That is, for a postintervention period t (with $t \geq T_0$), the synthetic control estimator of the effect of the treatment is given by the comparison between the outcome for the treated unit and the outcome for the synthetic control at that period:

$$Y_{1\,t} - \sum_{j=2}^{J+1} w_j^* Y_{jt}.$$

The matching variables in X_0 and X_1 are meant to be predictors of postintervention outcomes. These predictors are themselves not affected by the intervention. Critics of Mill's Method of Differences rightfully point out that the applicability of the method may be limited by the presence of unmeasured factors affecting the outcome variable as well as by heterogeneity in the effects of observed and unobserved factors. However, using a linear factor model, Abadie, Diamond, and Hainmueller (2010) argue that if the number of preintervention periods in the data is large, matching on preintervention outcomes (i.e., on the preintervention counterparts of Y_0 and Y_1) helps control for unobserved factors and for the heterogeneity of the effect of the observed and unobserved factors on the outcome of interest. The intuition of this result is straightforward: Only units that are alike in both observed and unobserved determinants of the outcome variable as well as in the effect of those determinants on the outcome variable should produce similar trajectories of the out-

come variable over extended periods of time. Once it has been established that the unit representing the case of interest and the synthetic control unit have similar behavior over extended periods of time prior to the intervention, a discrepancy in the outcome variable following the intervention is interpreted as produced by the intervention itself.[6]

Computation of synthetic controls can be done using freely available software scripts that we have written for R, MATLAB, and Stata.[7]

Comparison to Regression

Constructing a synthetic comparison as a linear combination of the untreated units with coefficients that sum to one may appear unusual. However, here we show that a regression-based approach also uses a linear combination of the untreated units with coefficients that sum to one as a comparison, albeit implicitly. In contrast to the synthetic control method, the regression approach does not restrict the coefficients of the linear combination that define the comparison unit to be between zero and one, therefore allowing extrapolation outside the support of the data.

The proof of this assertion is as follows. A regression-based counterfactual of the outcome for the treated unit in the absence of the treatment is given by the $(T_1 \times 1)$ vector $\widehat{B}' X_1$, where $\widehat{B} = (X_0 X_0')^{-1} X_0 Y_0'$ is the $(k \times T_1)$ matrix of regression coefficients of Y_0 on X_0.[8] As a result, the regression-based estimate of the counterfactual of interest is equal to $Y_0 W^{\text{reg}}$, where $W^{\text{reg}} = X_0'(X_0 X_0')^{-1} X_1$. Let ι be a $(J \times 1)$ vector of ones. The sum of the regression weights is $\iota' W^{\text{reg}}$. Assume that, as usual, the regression includes an intercept, so the first row of X_0 is a vector of ones.[9] Then $(X_0 X_0')^{-1} X_0 \iota$ is a $(k \times 1)$ vector with the first element equal to one and all the rest equal to zero. The

---

$W^* = (w_2^*, \ldots, w_{J+1}^*)'$ is selected to minimize $\| X_1 - X_0 W \|$, subject to $0 \leq w_j \leq 1$ for $j = 2, \ldots J$ and $w_2 + \cdots + w_{J+1} = 1$. Typically, V is selected to weight covariates in accordance to their predictive power on the outcome (see Abadie, Diamond, and Hainmueller 2010; Abadie and Gardeazabal 2003). If V is diagonal with main diagonal equal to $(v_1, \ldots, v_k)$, then W^* is equal to the value of W that minimizes Equation (1). Because W^* is invariant to scale changes in $(v_1, \ldots, v_k)$, these weights can always be normalized to sum to one.

[6]In this respect, the synthetic control method combines the synchronic and diachronic approaches outlined in Lijphart (1971). As pointed out by Gerring (2007), this approach is close in spirit to comparative historical analysis methods (Mahoney and Rueschemeyer 2003; Pierson and Skocpol 2002).

[7]In Stata, the script can be installed by typing `ssc install synth, replace,` which downloads the software from the Statistical Software Components (SSC) archive. In R, the software is available as the `Synth` package from the Comprehensive R Archive Network at http://CRAN.R-project.org/package=Synth. The R package is also described in detail in Abadie, Diamond, and Hainmueller (2011). MATLAB code is available on the authors' websites.

[8]That is, each column r of the matrix $\widehat{B}$ contains the regression coefficients of the outcome variable at period $t = T_0 + r$ on X_0.

[9]It is easy to extend the proof to the more general case where the unit vector, ι, belongs to the subspace of $\mathbb{R}^{J+1}$ spanned by the rows of $[X_1 \ X_0]$.

reason is that $(X_0 X_0')^{-1} X_0 \iota$ is the vector of coefficients of the regression of ι on X_0. Because ι is a vector of ones and because the first row of X_0 is also a vector of ones, the only nonzero coefficient of this regression is the intercept, which takes a value equal to one. This implies that $\iota' W^{\text{reg}} = \iota' X_0'(X_0 X_0')^{-1} X_1 = 1$ (because the first element of X_1 is equal to one).

That is, the regression estimator is a weighting estimator with weights that sum to one. However, regression weights are unrestricted and may take on negative values or values greater than one. As a result, estimates of counterfactuals based on linear regression may extrapolate beyond the support of comparison units. Even if the characteristics of the case of interest cannot be approximated using a weighted average of the characteristics of the potential controls, the regression weights extrapolate to produce a perfect fit. In more technical terms, even if X_1 is far from the convex hull of the columns of X_0, regression weights extrapolate to produce $X_0 W^{\text{reg}} = X_0 X_0'(X_0 X_0')^{-1} X_1 = X_1$.

Regression extrapolation can be detected if the weights W^{reg} are explicitly calculated, because it results in weights outside the $[0, 1]$ interval. We do not know, however, of any previous article that explicitly computes regression weights, and we are also unaware of previous results interpreting regressions as weighting estimators with weights that sum to one. Because regression weights are not calculated in practice, the extent of extrapolation produced by regression techniques is typically hidden from the analyst. In the empirical section below, we provide a comparison between the unit synthetic control weights and the regression weights for the German reunification example. For that example, we show that the regression-based counterfactual relies on extrapolation. Extrapolation is, however, unnecessary in the context of our German reunification example. We show that there exists a synthetic control that closely fits the values of the characteristics of the units and that does not extrapolate outside of the support of the data.[10]

---

[10]While using weights that sum to one and fall in the $[0, 1]$ interval prevents extrapolation biases, interpolation biases may be severe in some cases, especially if the donor pool contains units with characteristics that are very different from those of the unit representing the case of interest. To reduce interpolation biases, we recommend restricting the donor pool to units that are similar to the one representing the case of interest. In addition, the $\|X_1 - X_0 W\|$ objective function for the weights could be complemented with penalty terms that incorporate the discrepancies in characteristics between the unit representing the case of interest and the units with positive weights in the synthetic control. Such penalty terms can also be useful to select a synthetic control when minimization of $\|X_1 - X_0 W\|$ has multiple solutions, because X_1 falls in the convex hull of the columns of X_0.

Inference with the Synthetic Control Method

The use of statistical inference in comparative case studies is difficult because of the small-sample nature of the data, the absence of randomization, and the fact that probabilistic sampling is not employed to select sample units. These limitations complicate the application of traditional approaches to statistical inference.[11] However, by systematizing the process of estimating the counterfactual of interest, the synthetic control method enables researchers to conduct a wide array of falsification exercises, which we term "placebo studies," that provide the building blocks for an alternative mode of qualitative and quantitative inference. This alternative model of inference is based on the premise that our confidence that a particular synthetic control estimate reflects the impact of the intervention under scrutiny would be severely undermined if we obtained estimated effects of similar or even greater magnitudes in cases where the intervention did not take place.

Suppose, for example, that the synthetic control method estimates a sizable effect for a certain intervention of interest. Our confidence about the validity of this result would dissipate if the synthetic control method also estimated large effects when applied to dates when the intervention did not occur (Heckman and Hotz 1989). We refer to these falsification exercises as "in-time placebos." These tests are feasible if there are available data for a sufficiently large number of time periods when no structural shocks to the outcome variable occurred. In the example of the next section, we consider the effect of the 1990 German reunification on per capita GDP in West Germany. The German reunification occurred in 1990, but we have data from 1960 and thus can test whether the method produces large estimated effects when applied to dates before the reunification. If we find estimated effects that are of similar or larger magnitude than the one estimated for the 1990 reunification, our confidence that the effect estimated for the 1990 reunification is attributable to reunification itself would greatly diminish. In that case, the placebo studies would suggest that synthetic controls do not provide good predictions of the trajectory of the outcome in West Germany in periods when the reunification did not occur. Conversely, in the empirical section below, we find a very large effect for the 1990 German reunification, but no effect at all when we artificially reassign the reunification period in our data to a date before 1990.

---

[11]See Rubin (1990) for a description of the different modes of statistical inference for causal effects.

Another way to conduct placebo studies is to reassign the intervention not in time, but to members of the donor pool. We refer to these tests as "in-space placebos." Here the premise is that our confidence that a sizable synthetic control estimate reflects the effect of the intervention would disappear if similar or larger estimates arose when the intervention is artificially reassigned to units not directly exposed to the intervention.

A particular implementation of this idea consists of applying the synthetic control method to estimate placebo effects for every potential control unit in the donor pool. This creates a distribution of placebo effects against which we can then evaluate the effect estimated for the unit that represents the case of interest. Our confidence that a large synthetic control estimate reflects the effect of the intervention would be undermined if the magnitude of the estimated effect fell well inside the distribution of placebo effects. As in traditional statistical inference, a quantitative comparison between the distribution of placebo effects and the synthetic control estimate can be operationalized through the use of p-values. In this context, a p-value can be constructed by estimating in-space placebo effects for each unit in the sample and then calculating the fraction of such effects greater than or equal to the effect estimated for the treated unit. Notice that this inferential exercise reduces to classical randomization inference when the intervention is randomized (Rosenbaum 2005). In the absence of randomization, the p-value still has an interpretation as the probability of obtaining an estimate at least as large as the one obtained for the unit representing the case of interest when the intervention is reassigned at random in the data set. Notice that the inferential methods outlined in this section do not produce confidence intervals or posterior distributions, and that the inferential exercises (and associated p-values) are restricted to the question of whether or not the estimated effect of the actual intervention is large relative to the distribution of placebo effects. In the empirical section below, we show that the synthetic control estimate for West Germany is very large relative to the distribution of placebo estimates for the countries in the donor pool.

Limitations of the Approach and Some Recommendations for Empirical Practice

The synthetic control method facilitates comparative case studies in instances when no single untreated unit provides a good comparison for the unit affected by the treatment or event of interest. This is often the case when the treatment affects large aggregates like regions or countries, so a limited number of untreated units are available.

In this section, we discuss requirements and limitations of the method and provide recommendations for empirical practice.

Constructing a donor pool of comparison units requires some care. First, units affected by the event or intervention of interest or by events of a similar nature should be excluded from the donor pool. In addition, units that may have suffered large idiosyncratic shocks to the outcome of interest during the study period should also be excluded if such shocks would have not affected the treated unit in the absence of the treatment. Finally, to avoid interpolation biases, it is important to restrict the donor pool to units with characteristics similar to the treated unit. Another reason to restrict the size of the donor pool and consider only units similar to the treated unit is to avoid overfitting. Overfitting arises when the characteristics of the unit affected by the intervention or event of interest are artificially matched by combining idiosyncratic variations in a large sample of unaffected units. The risk of overfitting motivates our adoption of the cross-validation techniques applied in the empirical section below.

The applicability of the method requires a sizable number of preintervention periods. The reason is that the credibility of a synthetic control depends upon how well it tracks the treated unit's characteristics and outcomes over an extended period of time prior to the treatment. We do not recommend using this method when the pretreatment fit is poor or the number of pretreatment periods is small.[12] A sizable number of postintervention periods may also be required in cases when the effect of the intervention emerges gradually after the intervention or changes over time.

The Economic Cost of the 1990 German Reunification

In this section, we apply the synthetic control method to estimate the impact of the 1990 German reunification, one of the most significant political events in postwar Europe. After the fall of the Berlin Wall on November 9, 1989, the German Democratic Republic and the Federal Republic of Germany officially reunified on October 3, 1990. At that time, per capita GDP in West Germany was about three times higher than in East Germany (Lipschitz and McDonald 1990). Given the large income disparity, the integration of both countries after 45 years of

---

[12]See Abadie, Diamond, and Hainmueller (2010) for a related discussion.

separation called for political and economic adjustments of unprecedented complexity and scale. The 1990 German reunification therefore provides an excellent case study to examine the economic consequences of political integration. As Burda and Hunt (2001, 1) put it, "it is difficult to find a more dramatic episode of economic dislocation in peacetime during the twentieth century than that associated with the reunification of Germany."

Many studies have examined the consequences of the reunification (see Burda and Hunt 2001; Heilemann and Rappen 2000; and Sinn 2002 for reviews), but most of them focus on the consequences for (the former) East Germany and the convergence between the East and West German economy. The question of what the economic costs are for West Germany has received less attention in the literature. And while many argue that the economic impact of reunification on the West German economy has been negative, the magnitude of this effect remains unknown (e.g., Canova and Ravn 2000; Heilemann and Rappen 2000; Meinhardt et al. 1995). In particular, to our knowledge no study has rigorously estimated the impact of reunification on West German per capita GDP using a comparative case study analysis. Here we fill this gap and construct a synthetic West Germany as a weighted average of other advanced industrialized countries chosen to resemble the values of economic growth predictors for West Germany prior to the reunification. The synthetic West Germany is meant to replicate the (counterfactual) per capita GDP trend that West Germany would have experienced in the absence of the 1990 reunification. We then estimate the effect of the reunification by comparing the actual (with reunification) and counterfactual (without reunification) per capita GDP series for West Germany.[13]

Data and Sample

We use annual country-level panel data for the period 1960–2003. The German reunification occurred in 1990, giving a preintervention period of 30 years. Our sample period ends in 2003 because a roughly decadelong period after the reunification seems like a reasonable limit on the span of plausible prediction. Recall that the synthetic West Germany is constructed as a weighted average of potential control countries in the donor pool. Our donor pool includes a sample of 16 OECD member countries: Aus-

tralia, Austria, Belgium, Denmark, France, Greece, Italy, Japan, the Netherlands, New Zealand, Norway, Portugal, Spain, Switzerland, the United Kingdom, and the United States.[14]

We provide a list of all variables used in the analysis in the data appendix, along with data sources. The outcome variable, Y_{jt}, is the real per capita GDP in country j at time t. GDP is Purchasing Power Parity (PPP)-adjusted and measured in 2002 U.S. dollars (USD, hereafter). For the pre-reunification characteristics in X_1 and X_0, we rely on a standard set of economic growth predictors: per capita GDP, inflation rate, industry share of value added, investment rate, schooling, and a measure of trade openness (see the appendix for details). For each variable, we checked that the German data refer exclusively to the territory of the former West Germany.[15] We experimented with a wide set of additional growth predictors, but their inclusion did not change our results substantively.

Constructing a Synthetic Version of West Germany

Using the techniques described in the methodological section above, we construct a synthetic West Germany with weights chosen so that the resulting synthetic West Germany best reproduces the values of the predictors of per capita GDP in West Germany in the prereunification period. We use a cross-validation technique to choose the weights v_m in Equation (1). We first divide the pretreatment years into a training period from 1971 to 1980 and a validation period from 1981 to 1990. Next, using predictors measured in the training period, we select the weights v_m such that the resulting synthetic control minimizes the root mean square prediction error

---

[13] Additionally, one could also try to estimate the effect of reunification on East Germany. However, concerns about the quality of the official East German statistics before the German reunification render this a questionable endeavor. See Lipschitz and McDonald (1990).

[14] To construct this sample, we started with the 23 OECD member countries in 1990 (excluding West Germany). We first excluded Luxembourg and Iceland because of their small size and because of the peculiarities of their economies. We also excluded Turkey, which had in 1990 a level of per capita GDP well below the other countries in the sample. We finally excluded Canada, Finland, Sweden, and Ireland because these countries were affected by profound structural shocks during the sample period. Ireland experienced a rapid Celtic Tiger expansion period in the 1990s. Canada, Finland, and Sweden experienced profound financial and fiscal crises at the beginning of the 1990s. It is important to note, however, that when included in the sample, these four countries obtain zero weights in the synthetic control for West Germany. Therefore, our main results are *identical* whether or not we exclude these countries.

[15] For that purpose, when necessary, our data set was supplemented with data from the German Federal Statistical Office (*Statistisches Bundesamt*).

TABLE 1 Synthetic and Regression Weights for West Germany

Country	Synthetic Control Weight	Regression Weight	Country	Synthetic Control Weight	Regression Weight
Australia	0	0.12	Netherlands	0.09	0.14
Austria	0.42	0.26	New Zealand	0	0.12
Belgium	0	0	Norway	0	0.04
Denmark	0	0.08	Portugal	0	−0.08
France	0	0.04	Spain	0	−0.01
Greece	0	−0.09	Switzerland	0.11	0.05
Italy	0	−0.05	United Kingdom	0	0.06
Japan	0.16	0.19	United States	0.22	0.13

Notes: The synthetic weight is the country weight assigned by the synthetic control method. The regression weight is the weight assigned by linear regression. See text for details.

(RMSPE) over the validation period.[16] Intuitively, the cross-validation technique selects the weights v_m that minimize out-of-sample prediction errors. Finally, we use the set of v_m weights selected in the previous step and predictor data measured in 1981–90 to estimate a synthetic control for West Germany. The v_m weights chosen by the cross-validation indicate that the most important predictors (in order from highest to lowest weight) are GDP per capita (0.442), investment rate (0.245), trade openness (0.134), schooling (0.107), inflation rate (0.072), and industry share (0.001).[17] We estimate the effect of the German reunification on per capita GDP in West Germany as the difference in per capita GDP levels between West Germany and its synthetic counterpart in the years following the reunification.

Table 1 shows the weights of each country in the synthetic version of West Germany. The synthetic West Germany is a weighted average of Austria, the United States, Japan, Switzerland, and the Netherlands with weights decreasing in this order. All other countries in the donor pool obtain zero weights. As a comparison, Table 1 also

[16]The RMSPE measures lack of fit between the path of the outcome variable for any particular country and its synthetic counterpart. The pre-1990 RMSPE for West Germany is defined as

$$\text{RMSPE} = \left(\frac{1}{T_0} \sum_{t=1}^{T_0} \left(Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt} \right)^2 \right)^{1/2}.$$

The RMSPE can be analogously defined for other countries or time periods.

[17]Our results are robust to alternative procedures to choose v_m. In particular, Abadie and Gardeazabal (2003) and Abadie, Diamond, and Hainmueller (2010) chose v_m so that the resulting synthetic control best approximates the preintervention path of the outcome variable. For the German reunification example, this way to choose v_m produces results that are almost identical to the results that we obtain using the cross-validation technique used in this article.

TABLE 2 Economic Growth Predictor Means before German Reunification

	West Germany	Synthetic West Germany	OECD Sample
GDP per capita	15808.9	15802.2	8021.1
Trade openness	56.8	56.9	31.9
Inflation rate	2.6	3.5	7.4
Industry share	34.5	34.4	34.2
Schooling	55.5	55.2	44.1
Investment rate	27.0	27.0	25.9

Notes: GDP per capita, inflation rate, trade openness, and industry share are averaged for the 1981–90 period. Investment rate and schooling are averaged for the 1980–85 period. The last column reports a population-weighted average for the 16 OECD countries in the donor pool.

reports the weights that regression analysis employs implicitly when applied to the same data (these weights are backed out using the formulas presented in the methodological section above). By construction, both sets of weights sum to one. The two sets of weights show some similarities. For example, Austria receives the highest weight in both approaches. Overall, however, the weights are very different. For example, the regression weights for Japan and Austria are much closer in values than their synthetic control counterparts. Moreover, regression assigns negative weights to four of the 16 control units in the donor pool: Greece (–0.09), Italy (–0.05), Portugal (–0.08), and Spain (–0.01). As discussed previously, negative weights indicate that regression relies on extrapolation.

Table 2 compares the prereunification characteristics of West Germany to those of the synthetic West Germany, and also to those of a population-weighted average of the 16 OECD countries in the donor pool. Overall, the results in Table 2 suggest that the synthetic West Germany
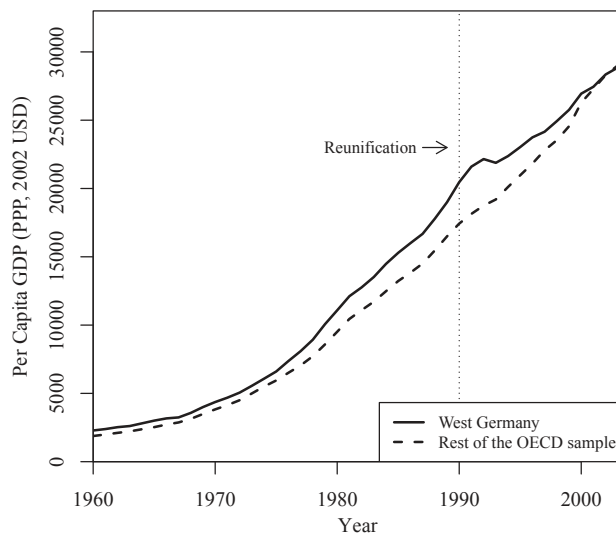
FIGURE 1 **Trends in per Capita GDP: West Germany versus Rest of the OECD Sample**

FIGURE 2 **Trends in per Capita GDP: West Germany versus Synthetic West Germany**

provides a much better comparison for West Germany than the average of our sample of other OECD countries. The synthetic West Germany is very similar to the actual West Germany in terms of pre-1990 per capita GDP, trade openness, schooling, investment rate, and industry share. Compared to the average of the OECD countries, the synthetic West Germany also matches West Germany much closer on the inflation rate. Because West Germany had the lowest inflation rate in the sample during the prereunification years, this variable cannot be perfectly fitted using a combination of the comparison countries. Figure 1 shows that before the German reunification, West Germany and the OECD average experienced different paths in per capita GDP. However, in the next section, we will show that a synthetic control can accurately reproduce the pre-1990 per capita GDP path for West Germany.

One of the central points of this article is that the synthetic control method provides the qualitative researcher with a quantitative tool to select or validate comparison units. In our analysis, Austria, the United States, Japan, Switzerland, and the Netherlands emerge, in this order, as potential comparisons to West Germany. Regression analysis fails to provide such a list. In a regression analysis, typically all units contribute to the regression fit, and the contribution of units with large positive regression weights may be counterbalanced by the contributions of units with negative weights. In this example, the synthetic control involves a combination of five countries. Below we show how researchers can construct, if desired, synthetic controls that use a smaller number of countries.
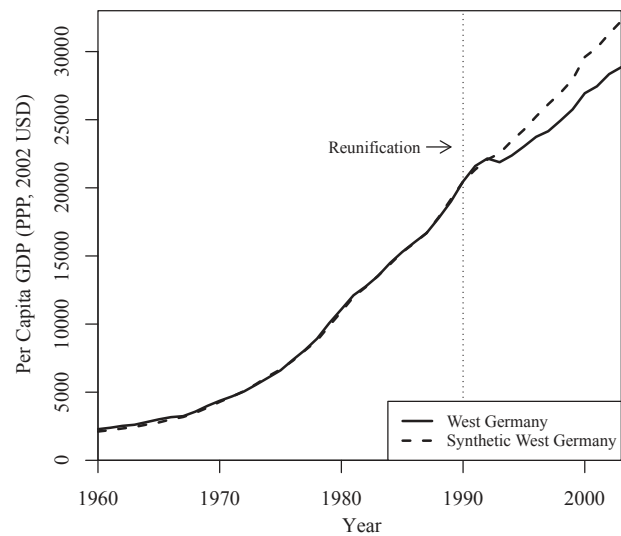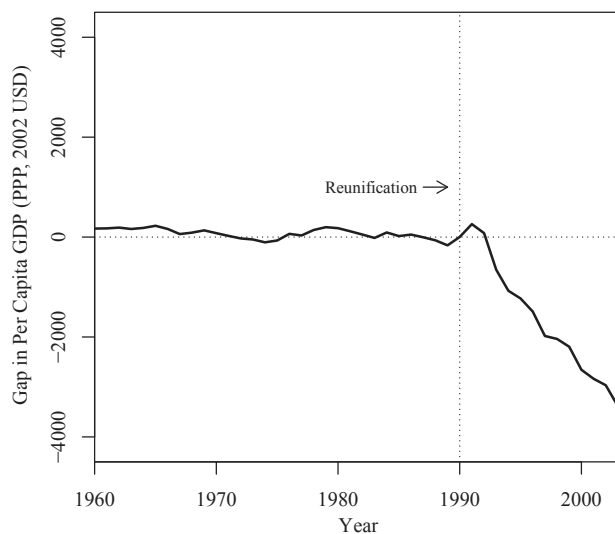
The Effect of the 1990 Reunification

Figure 2 displays the per capita GDP trajectory of West Germany and its synthetic counterpart for the 1960–2003 period. The synthetic West Germany almost exactly reproduces the per capita GDP for West Germany during the entire prereunification period. This close fit for the prereunification per capita GDP and the close fit that we obtain for the GDP predictors in Table 2 demonstrate that there exists a combination of other industrialized countries that reproduces the economic attributes of West Germany before the reunification. That is, it is possible to closely reproduce economic characteristics of West Germany before the 1990 reunification without extrapolating outside of the support of the data for the donor pool.

Our estimate of the effect of the German reunification on per capita GDP in West Germany is given by the difference between the actual West Germany and its synthetic version, visualized in Figure 3. We estimate that the German reunification did not have much of an effect on West German per capita GDP in the first two years immediately following reunification. In this initial period, per capita GDP in the synthetic West Germany is even slightly lower than in the actual West Germany, which is broadly in line with arguments about an initial demand boom (see, e.g., Meinhardt et al. 1995). From 1992 onward, however, the two lines diverge substantially. While per capita GDP growth decelerates in West Germany, for the synthetic West Germany per capita GDP keeps ascending at a pace similar to that of the prereunification period.

FIGURE 3 **Per Capita GDP Gap between West Germany and Synthetic West Germany**

FIGURE 4 **Placebo Reunification 1975–Trends in per Capita GDP: West Germany versus Synthetic West Germany**

The difference between the two series continues to grow towards the end of the sample period. Thus, our results suggest a pronounced negative effect of the reunification on West German income. We find that over the entire 1990–2003 period, per capita GDP was reduced by about 1,600 USD per year on average, which amounts to approximately 8% of the 1990 baseline level. In 2003, per capita GDP in the synthetic West Germany is estimated to be about 12% higher than in the actual West Germany.

One valid concern in the context of this study is the potential existence of spillover effects. In particular, it is possible that the German reunification had effects on per capita GDP in countries other than Germany. Notice, however, that the limited number of units in the synthetic control allows the evaluation of the existence and direction of potential biases created by spillover effects. For example, if the German reunification had negative spillover effects on the per capita GDP of the countries included in the synthetic control, then the synthetic control would provide an underestimate of the counterfactual per capita GDP trajectory for West Germany in the absence of the reunification and, therefore, an underestimate of the negative effect of the reunification on per capita GDP in West Germany. On the other hand, if the German reunification had positive effects in the economies included in the synthetic control, this would exacerbate the negative effect of the synthetic control estimates. Notice also that spillover effects on countries not included in the synthetic control do not affect the synthetic control estimates.
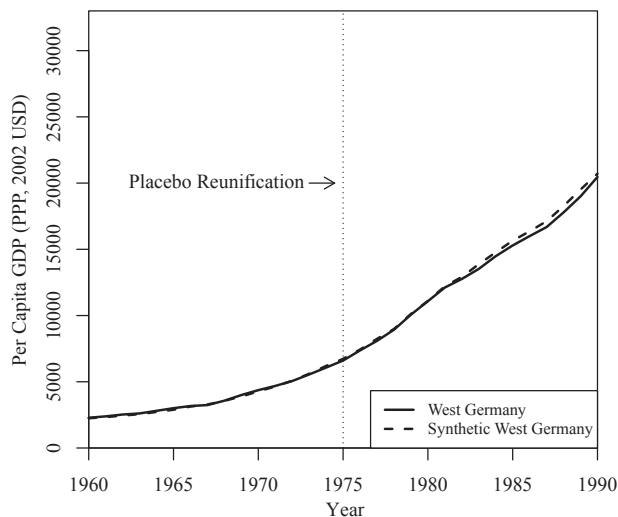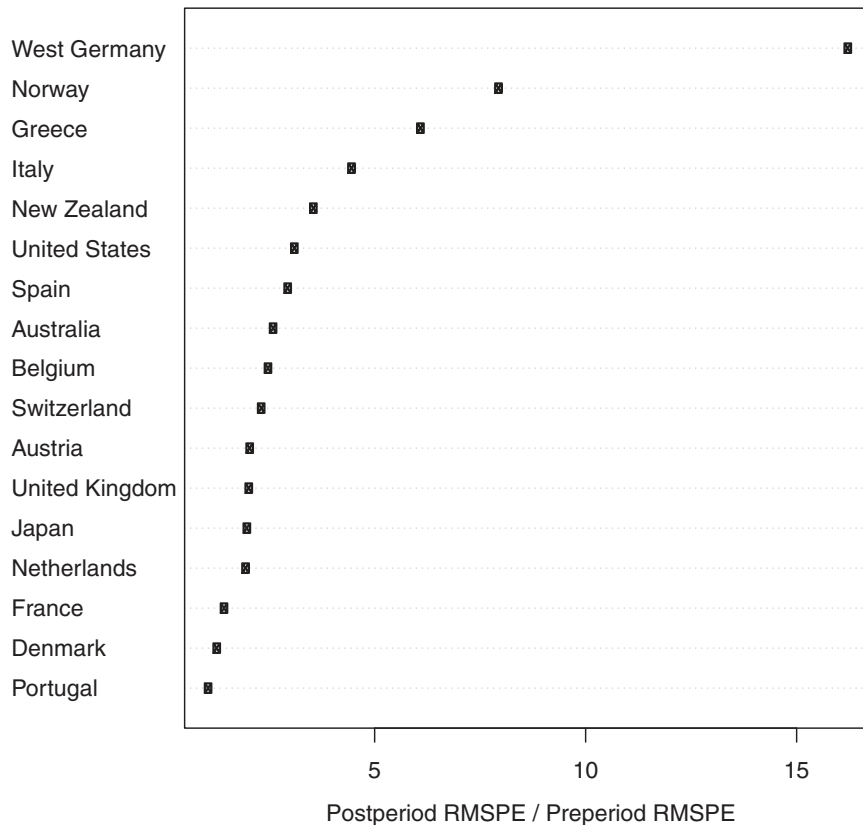
Placebo Studies

To evaluate the credibility of our results, we conduct placebo studies where the treatment of interest is reassigned in the data to a year other than 1990 or to countries different from West Germany. We first compare the reunification effect estimated above for West Germany to a placebo effect obtained after reassigning the German reunification in our data to a period before the reunification actually took place. A large placebo estimate would undermine our confidence that the results in Figure 2 are indeed indicative of the economic cost of reunification and not merely driven by lack of predictive power.

To conduct this placebo study, we rerun the model for the case when reunification is reassigned to the middle of the pretreatment period in the year 1975, about 15 years earlier than reunification actually occurred. We use the same out-of-sample validation technique to compute the synthetic control, and we lag the predictors variables accordingly for the training and validation period. Figure 4 displays the results of this "in-time placebo" study. The synthetic West Germany almost exactly reproduces the evolution of per capita GDP in the actual West Germany for the 1960–1975 period. Most importantly, the per capita GDP trajectories of West Germany and its synthetic counterpart do not diverge considerably during the 1975–1990 period. That is, in contrast to the actual 1990 German reunification, our 1975 placebo reunification has no perceivable effect. This suggests that the gap estimated in Figure 2 reflects the impact of the German

FIGURE 5 Ratio of Postreunification RMSPE to Prereunification RMSPE: West Germany and Control Countries

reunification and not a potential lack of predictive power of the synthetic control.[18]

An alternative way to conduct placebo studies is to reassign the treatment in the data to a comparison unit. In this way, we can obtain synthetic control estimates for countries that did not experience the event of interest. Applying this idea to each country in the donor pool allows us to compare the estimated effect of the German reunification on West Germany to the distribution of placebo effects obtained for other countries. We will deem the effect of the German reunification on West Germany significant if the estimated effect for West Germany is unusually large relative to the distribution of placebo effects.

Figure 5 reports the ratios between the post-1990 RMSPE and the pre-1990 RMSPE for West Germany and for all the countries in the donor pool. Recall that RM-

SPE measures the magnitude of the gap in the outcome variable of interest between each country and its synthetic counterpart. A large postintervention RMSPE is not indicative of a large effect of the intervention if the synthetic control does not closely reproduce the outcome of interest prior to the intervention. That is, a large postintervention RMSPE is not indicative of a large effect of the intervention if the preintervention RMSPE is also large. For each country, we divide the postreunification RMSPE by its pre-reunification RMSPE.[19] In Figure 5, West Germany clearly stands out as the country with the highest RMSPE ratio. For West Germany, the postreunification gap is about 16 times larger than the prereunification gap. If one were to pick a country at random from the sample, the chances of obtaining a ratio as high as this one would be $1/17 \simeq 0.059$.

---

[18]We have computed similar in-time placebo studies where we reassign in our data the German reunification to the years 1970 and 1980, respectively, and the results are similar to the results for 1975 shown here.

[19]By taking the ratio between the post-1990 RMSPE, and the pre-1990 RMSPE, we avoid having to discard countries with pre-1990 per capita GDP values that cannot be approximated with a synthetic control. See Abadie, Diamond, and Hainmueller (2010).
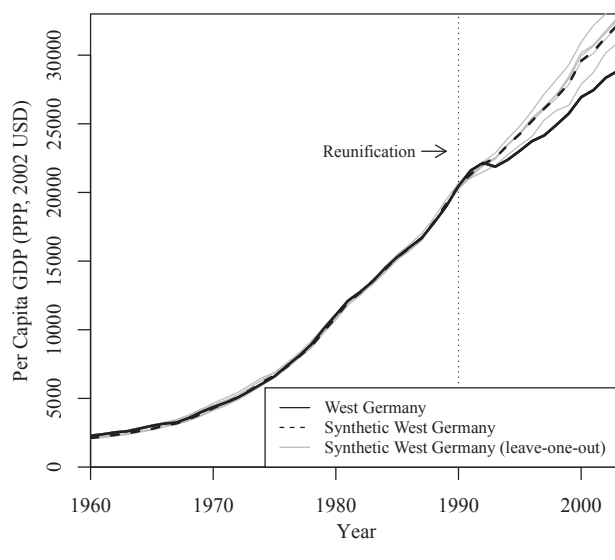
Robustness Test

In this section, we run a robustness check to test the sensitivity of our main results to changes in the country weights, W^*. Recall from Table 1 that the synthetic West Germany is estimated as a weighted average of Austria, the United States, Japan, Switzerland, and the Netherlands, with weights decreasing in this order. Here we iteratively reestimate the baseline model to construct a synthetic West Germany omitting in each iteration one of the countries that received a positive weight in Table 1. By excluding countries that received a positive weight we sacrifice some goodness of fit, but this sensitivity check allows us to evaluate to what extent our results are driven by any particular control country.

Figure 6 displays the results and reproduces Figure 2 (solid and dashed black lines) while also incorporating the leave-one-out estimates (gray lines). This figure shows that the results of the previous analysis are fairly robust to the exclusion of any particular country from our sample of comparison countries. The leave-one-out synthetic control that shows the smallest effect of the reunification is the one that excludes the United States. Even this estimate is fairly large in substantive terms: Per capita GDP over the 1990–2003 period is reduced by about 630 USD per year on average, approximately 3% of the 1990 baseline level. In 2003, per capita GDP in this synthetic West Germany is estimated to be about 7% higher than in the actual West Germany. The other leave-one-out synthetic controls show either a very similar or slightly

larger effect compared to the results of the previous analysis.

Reducing the Number of Units in a Synthetic Control

Recall again that the synthetic West Germany in Figure 2 is a weighted average of five control countries: Austria, the United States, Japan, Switzerland, and the Netherlands. Comparative researchers, however, typically choose a very small number of cases, with the aim of meticulously describing and analyzing the characteristics and outcomes of each of those cases. As a result, in many instances, comparative researchers may favor sparse synthetic controls, that is, synthetic controls that involve a small number of comparison countries. Reducing the number of units in the synthetic control may, nonetheless, impact the extent to which the synthetic control is able to fit the characteristics of the unit of interest. In this section, we examine the trade-off between sparsity and goodness of fit in the choice of the number of units that contribute to the synthetic control for West Germany.

In order to investigate this trade-off, we construct synthetic controls for West Germany, allowing only combinations of four, three, two, and a single control country, respectively.[20] Table 3 shows the countries and weights for the sparse synthetic controls. For this example, the countries contributing to the sparse versions of the synthetic control for West Germany are subsets of the set of five countries contributing to the synthetic control in the baseline specification.[21] Austria retains the largest weight in all instances, whereas the United States, Japan, and Switzerland are second, third, and fourth in terms of their synthetic control weights.

Table 4 compares economic growth predictors of West Germany, synthetic West Germany, the sparse versions of synthetic West Germany, and the OECD sample. This table documents the sacrifice in terms of goodness of fit resulting from a reduction in the number of countries, l, allowed to contribute to the synthetic control.

---

[20]More precisely, for $l = 4, 3, 2, 1$, and for all possible combinations of l control countries, we choose the one that produces the synthetic control unit that minimizes the loss defined in Equation (1). To reduce computational complexity, we used the weights v_m obtained for the baseline model instead of attempting to recalculate these weights for the many different combinations of l countries.

[21]Also, each subset of countries in Table 3 is nested by any other subset containing a larger number of countries. This is not imposed in our analysis, as we consider all possible combinations of countries among the 16 countries in the donor pool, and it may not necessarily be the case for other applications.

TABLE 3 Synthetic Weights from Combinations of Control Countries

Synthetic Combination	Countries and *W*-Weights				
Five control countries	Austria	USA	Japan	Switzerland	Netherlands
	0.42	0.22	0.16	0.11	0.09
Four control countries	Austria	USA	Japan	Switzerland	
	0.56	0.22	0.12	0.10	
Three control countries	Austria	USA	Japan		
	0.59	0.26	0.15		
Two control countries	Austria	USA			
	0.76	0.24			
One control country	Austria				
	1				

Notes: Countries and *W*-weights for synthetic control are constructed from the best-fitting combination of five, four, three, and two countries, as well as one country. See text for details.

TABLE 4 Economic Growth Predictor Means before the German Reunification for Combinations of Control Countries

	West Germany	Synthetic West Germany Number of countries in synthetic control					OECD Sample
		5	4	3	2	1	
GDP per capita	15808.9	15802.2	15800.9	15492.9	15580.9	14817.0	8021.1
Trade openness	56.8	56.9	55.9	52.5	61.5	74.6	31.9
Inflation rate	2.6	3.5	3.6	3.6	3.8	3.5	7.4
Industry share	34.5	34.4	34.6	34.8	34.3	35.5	34.2
Schooling	55.5	55.2	57.6	57.7	60.7	60.9	44.1
Investment rate	27.0	27.0	27.2	26.8	25.6	26.6	25.9

Notes: GDP per capita, inflation rate, and trade openness are averaged for the 1981–1990 period. Industry share is averaged for the 1981–1990 period. Investment rate and schooling are averaged for the 1980–1985 period.

Overall, relative to the baseline synthetic control with five countries, the decline in goodness of fit is moderate for $l = 4, 3, 2$. The "matching"case of $l = 1$, where the synthetic West Germany is Austria, produces a much worse goodness of fit relative to $l > 1$, with substantial discrepancies in the per capita GDP and trade openness variables. However, even the matching case, with $l = 1$, represents a large improvement in terms of goodness of fit relative to the comparison unit consisting of the population-weighted average of the OECD sample.

Figure 7 shows the per capita GDP path for West Germany and the sparse synthetic controls with $l = 4, 3, 2, 1$. With the exception of the matching case ($l = 1$), the sparse synthetic controls in Figure 7 produce results that are very similar to the baseline result in Figure 2. However, using a single country, Austria, as a comparison provides a much poorer fit to the pre-1990 per capita GDP path

for West Germany.[22] This analysis illustrates the potential gains from using combinations of countries rather than single countries as comparison cases in comparative research.[23]

[22]For example, at the time of the German reunification, the per capita GDP in Austria was about 7% lower than that of West Germany.

[23]The comparison between Austria and West Germany in the lower right panel of Figure 7 comes close to a difference-in-differences design, with Austria having a lower per capita GDP than West Germany during the pre-1990 period, but a fairly similar trend in this variable. In this comparison, we see that the Austrian per capita GDP surpassed the West German per capita GDP a few years after the reunification, even when it had been consistently lower for at least 30 years prior to the reunification. Therefore, this comparison also suggests a negative effect of the German reunification in West Germany. See Abadie, Diamond, and Hainmueller (2010) for a discussion of the relationship between the synthetic control and the difference-in-differences estimator.

FIGURE 7 Per Capita GDP Gaps between West Germany and Sparse Synthetic Controls
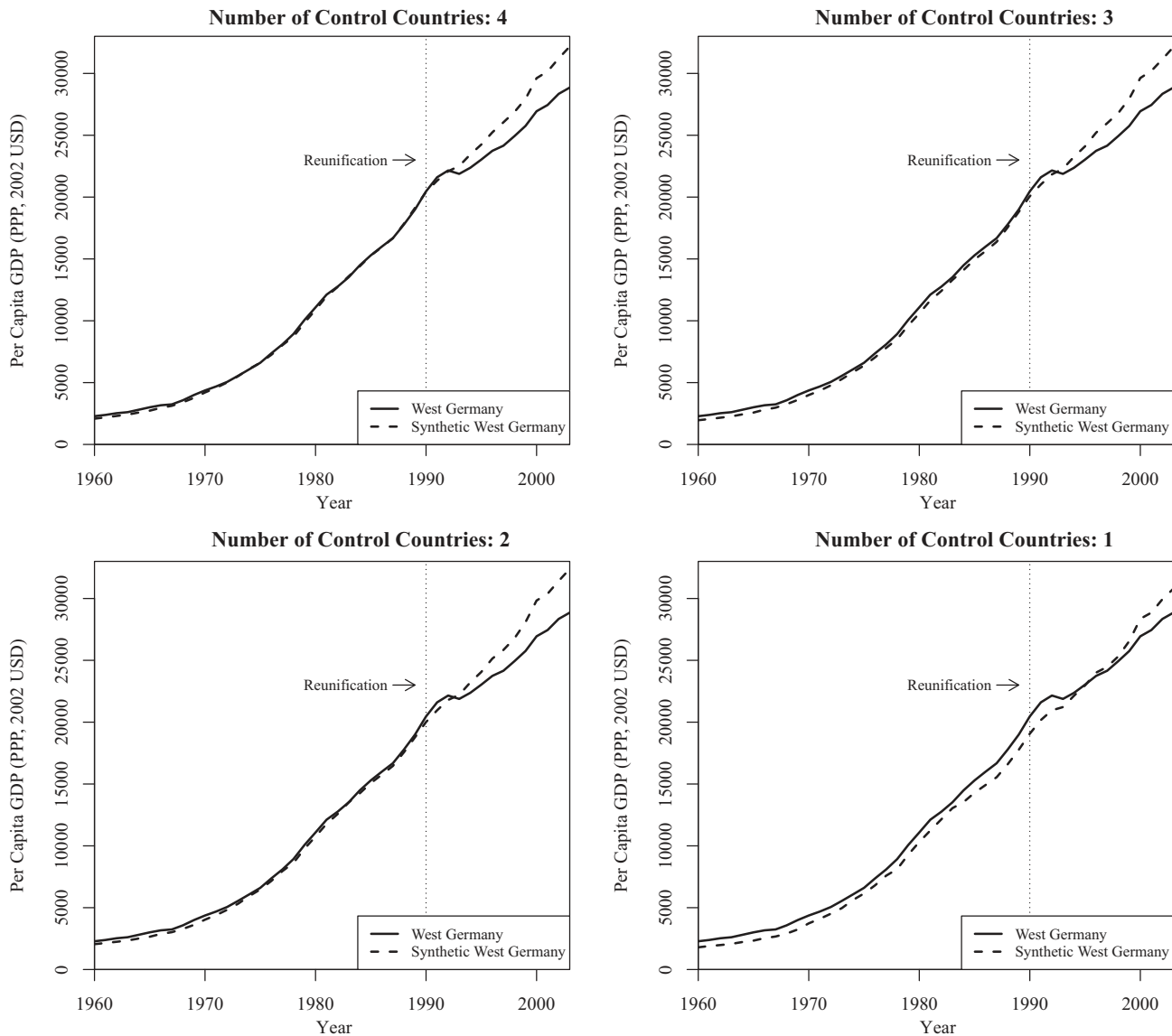
Conclusion

There is a widespread consensus among political methodologists about the necessity to integrate and exploit complementarities between qualitative and quantitative tools for empirical research in political science. However, some of the efforts in this direction have been denounced by qualitative methodologists as attempts to impose quantitative templates on qualitative research that disregard or do not make use of the many genuine advantages of qualitative research (Brady and Collier 2004; George and Bennett 2005). The synthetic control method discussed in this article falls in between the qualitative and quan-

titative methodologies and provides a potentially useful tool for researchers of both traditions. On the one hand, the synthetic control method provides a systematic way to select comparison units in quantitative comparative case studies. In this way, as in Card and Krueger (1994) and Rosenbaum (2005), the synthetic control method brings to quantitative studies the careful selection of cases that is done in qualitative analysis. In addition, by explicitly specifying the set of units that are used for comparison, the method does not preclude but facilitates detailed qualitative analysis and comparison between the case of interest and the set of comparison units selected by the method. That is, the synthetic control method can be used to guide

the selection of comparison units in qualitative studies, allowing what Tarrow (1995) calls "qualitative inference with quantitative bones."

Appendix

The data sources employed for the application are as follows:

- GDP per Capita (PPP, 2002 USD). Source: OECD National Accounts (retrieved via the OECD Health Database). Data for West Germany was obtained from Statistisches Bundesamt 2005 (Arbeitskreis "Volkswirtschaftliche Gesamtrechnungen der Länder") and converted using PPP monetary conversion factors (retrieved from the OECD Health Database).
- Investment Rate: Ratio of real domestic investment (private plus public) to real GDP. The data are reported in five-year averages. Source: Barro, Robert Joseph, and Jong-wha Lee. 1994. "Data Set for a Panel of 138 Countries." Available at `http://www.nber.org/pub/barro.lee/`.
- Schooling: Percentage of secondary school attained in the total population aged 25 and older. The data are reported in five-year increments. Source: Barro, Robert Joseph, and Jong-wha Lee. 2000. "International Data on Educational Attainment: Updates and Implications." *CID Working Paper No. 42, April 2000 – Human Capital Updated Files*.
- Industry: industry share of value added. Source: World Bank WDI Database 2005 and Statistisches Bundesamt 2005.
- Inflation: annual percentage change in consumer prices (base year 1995). Source: World Development Indicators Database 2005 and Statistisches Bundesamt 2005.
- Trade Openness: Export plus imports as percentage of GDP. Source: World Bank: World Development Indicators CD-ROM 2000.

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 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.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2011. "Synth: An R Package for Synthetic Control Methods in Comparative Case." *Journal of Statistical Software* 42(13): 1–17.

Abadie, Alberto, and Javier Gardeazabal. 2003. "The Economic Costs of Conflict: A Case Study of the Basque Country." *American Economic Review* 93(1): 112–32.

Beck, Nathaniel. 2010. "Causal Process 'Observation': Oxymoron or (Fine) Old Wine." *Political Analysis* 18(4): 499–505.

Brady, Henry E., and David Collier, eds. 2004. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, MD: Rowman & Littlefield.

Burda, Michael C., and Jennifer Hunt. 2001. "From Reunification to Economic Integration: Productivity and the Labor Market in Eastern Germany." *Brookings Papers on Economic Activity* 2: 1–92.

Canova, Fabio, and Morten O. Ravn. 2000. "The Macroeconomic Effects of German Unification: Real Adjustments and the Welfare State." *Review of Economic Dynamics* 3: 423–460.

Card, David, and Alan B. Krueger. 1994. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania." *American Economic Review* 84(4): 772–93.

Collier, David. 1993. "The Comparative Method." In *Political Science: The State of the Discipline II*, ed. Ada W. Finifter. Washington, DC: American Political Science Association, 105–19.

Collier, David and James Mahoney. 1996. "Insights and pitfalls: selection bias in qualitative research." World Politics. 49(1): 56–91.

Collier, David, James Mahoney, and Jason Seawright. 2004." Claiming Too Much: Warnings about Selection Bias." In *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, Eds, ed. Henry E. Brady and David Collier. Lanham, MD: Rowman & Littlefield, 85–102.

Fearon, James D. 1991. "Counterfactuals and Hypothesis Testing in Political Science." World Politics 43(2): 169–195.

Geddes, Barbara. 2003. *Paradigms and Sand Castles: Theory Building and Research Design in Comparative Politics*. Ann Arbor:University of Michigan Press.

George, Alexander L., and Andrew Bennett. 2005. *Case Studies and Theory Development in the Social Sciences*. Cambridge MA: MIT Press.

Gerring, John. 2004. "What Is a Case Study and What Is It Good For?" *American Political Science Review* 98(2): 341–54.

Gerring, John. 2007. *Case Study Research: Principles and Practices*. Cambridge: Cambridge University Press.

Hall, Peter. 2003. "Aligning Ontology and Methodology in Comparative Politics." In *Comparative Historical Analysis in the Social Sciences*, ed. James Mahoney and Dietrich Rueschemeyer. Cambridge: Cambridge University Press, 373-429.

Heckman, James, and V. Joseph Hotz. 1989."Choosing among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training." *Journal of the American Statistical Association* 84(408): 862–74.

Heilemann, Ullrich, and Hermann Rappen. 2000. "Zehn Jahre Deutsche Einheit: Bestandsaufnahme und Perspektiven." *RWI-Papiere Essen* 67.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research.* Princeton, NJ: Princeton University Press.

King, Gary, and Langche Zeng. 2006. "The Dangers of Extreme Counterfactuals." *Political Analysis* 14(2): 131–59.

Lieberman, Evan. 2005. "Nested Analysis as a Mixed-Method Strategy for Comparative Research." *American Political Science Review* 99(3): 435–52.

Lijphart, Arend. 1971. "Comparative Politics and the Comparative Method." *American Political Science Review* 65(3): 682–93.

Lipschitz, Leslie, and Donogh McDonald. 1990. "*German Unification: Economic Issues.*" International Monetary Fund Occasional Paper No 75.

Mahoney, James, and Dietrich Rueschemeyer. 2003. *Comparative Historical Analysis in the Social Sciences.* Cambridge: Cambridge University Press.

Meinhardt, Volker, Bernhard Seidel, Frank Stile, and Dieter Teichmann. 1995. "Transferleistungen in die neuen Bundesländer und deren wirtschaftliche Konsequenzen." *Deutsches Institut für Wirtschaftsforschung, Sonderheft 154.*

Pierson, Paul, and Theda Skocpol. 2002. "Historical Institutionalism in Contemporary Political Science." In *Political Science: The State of the Discipline*, ed. Ira Katznelson and Helen Millner. New York: W. W. Norton: 693–721.

Ragin, Charles C. 1987. *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies.* Berkeley: University of California Press.

Rosenbaum, Paul R. 2005. "Heterogeneity and Causality: Unit Heterogeneity and Design Sensitivity in Observational Studies." *The American Statistician* 59(2): 147-52.

Rubin, Donald B. 1990. "Formal Modes of Statistical Inference for Causal Effects." *Journal of Statistical Planning and Inference* 25: 279–92.

Sinn, Hans-Werner. 2002. "Germany's Economic Unification: An Assessment after Ten Years." *Review of International Economics* 10(1): 113–28.

Sekhon, Jasjeet S. 2004. "Quality Meets Quantity: Case Studies, Conditional Probability, and Counterfactuals." *Perspectives on Politics* 2(2): 281–93.

Tarrow, Sidney. 1995. "Bridging the Quantitative-Qualitative Divide in Political Science." *American Political Science Review* 89(2): 471–74.

Tarrow, Sidney. 2010. "The Strategy of Paired Comparison: Toward a Theory of Practice." *Comparative Political Studies* 43(2): 230–59.