Assessing the Causal Effect of Binary Interventions from Observational Panel Data with Few Treated Units

Pantelis Samartsidis, Shaun R. Seaman, Anne M. Presanis, Matthew Hickman and Daniela De Angelis

Abstract. Researchers are often challenged with assessing the impact of an intervention on an outcome of interest in situations where the intervention is nonrandomised, the intervention is only applied to one or few units, the intervention is binary, and outcome measurements are available at multiple time points. In this paper, we review existing methods for causal inference in these situations. We detail the assumptions underlying each method, emphasize connections between the different approaches and provide guidelines regarding their practical implementation. Several open problems are identified thus highlighting the need for future research.

Key words and phrases: Causal impact, causal inference, difference-in-differences, intervention evaluation, latent factor models, panel data, synthetic controls.

1. INTRODUCTION

Evaluation of the causal effect of an intervention (e.g., a newly introduced policy, a novel experimental

Pantelis Samartsidis is Investigator Statistician, MRC Biostatistics Unit, Cambridge Institute of Public Health, Forvie Site, Robinson Way, Cambridge Biomedical Campus, CB2 0SR, United Kingdom (e-mail: pantelis.samartsidis@mrc-bsu.cam.ac.uk). Shaun R. Seaman is Senior Research Associate, MRC Biostatistics Unit, Cambridge Institute of Public Health, Forvie Site, Robinson Way, Cambridge Biomedical Campus, CB2 0SR, United Kingdom (e-mail: shaun.seaman@mrc-bsu.cam.ac.uk). Anne M. Presanis is Senior Investigator Statistician, MRC Biostatistics Unit, Cambridge Institute of Public Health, Forvie Site, Robinson Way, Cambridge Biomedical Campus, CB2 0SR, United Kingdom (e-mail: anne.presanis@mrc-bsu.cam.ac.uk). Matthew Hickman is Professor, MRC Biostatistics Unit, Cambridge Institute of Public Health, Forvie Site, Robinson Way, Cambridge Biomedical Campus, CB2 0SR, United Kingdom (e-mail: matthew.hickman@bristol.ac.uk). Daniela De Angelis is Programme Leader, Population Health Sciences, Bristol Medical School, University of Bristol, Bristol BS8 2PS, United Kingdom (e-mail: daniela.deangelis@mrc-bsu.cam.ac.uk).

practice or an unexpected event) on an outcome of interest is a problem frequently encountered in several fields of scientific research. These include economics (Angrist and Pischke, 2009; Imbens and Wooldridge, 2009), epidemiology and public health (Rothman and Greenland, 2005; Glass et al., 2013), management (Antonakis et al., 2010), marketing (Rubin and Waterman, 2006; Varian, 2016) and political sciences (Keele, 2015).

Researchers are often interested in assessing the impact of an intervention (occasionally referred to as treatment henceforth) in situations where: (i) the data are observational, that is, the allocation of the sample units to the intervention and control groups is not randomised, but instead determined by factors that confound the association between the indicator of intervention and the outcome of interest; (ii) the intervention is binary, that is, sample units cannot receive the interventions at varying intensities; (iii) only one or a small number of units are treated; and (iv) at each of a set of time points, before and after the time at which the intervention is introduced, the outcome is measured on every sampled unit, thus giving rise to panel data.

Several statistical methods for causal inference in this setting have been developed to account for the special characteristics of the data: the presence of (likely

unobserved) confounders, the existence of temporal trends in the outcome and the limited number of sample units to which the intervention is given. In this paper, we review the existing literature, motivated by recent methodological developments and the increasing application of these methods to real-life problems. Since the existing literature comes from a wide range of research disciplines, our focus is on unifying the various methods under a common terminology and notation, appropriate to a statistical audience. Further, we draw connections between various methods and point out issues related to their practical implementation. Finally, we suggest some possible directions for future research.

We focus on four classes of methods: *difference-in-differences*, *latent factor models*, *synthetic control-type methods* and the *causal impact* method. Excluded from our review are propensity score methods (Rosenbaum and Rubin, 1983) (see Austin, 2011, for a recent review), because the small number of treated units does not allow accurate estimation of the parameters of a propensity score model, and the interrupted time series method (Lopez Bernal, Cummins and Gasparrini, 2016), because it does not use data on the units that do not receive the intervention.

This manuscript is structured as follows. In Section 2, we define notation, describe the causal framework underlying the methods and introduce the illustrative example. Section 3 presents the four classes of methods. Section 4 is about quantification of uncertainty and hypothesis testing. Section 5 discusses issues related to practical implementation. Sections 6 and 7 contain an applications to real data and a discussion, respectively. Finally, in Section 8 we highlight some remaining problems in the field.

2. PRELIMINARIES

2.1 Notation

Let $i = 1, \ldots, n$ index the entities (e.g., hospitals or general practices) for which the outcome of interest is observed: henceforth, we refer to these entities as units. For unit i we have measurements $\boldsymbol{y}_{i\cdot} = (y_{i1}, \ldots, y_{iT})^\top$, where t indexes time. We let $\boldsymbol{y}_{\cdot t} = (y_{1t}, \ldots, y_{nt})^\top$ denote the vector containing the set of n observations at time t and $\boldsymbol{y}_{i,t_1:t_2} = (y_{it_1}, \ldots, y_{it_2})^\top$ denote the measurements on unit i from time t_1 to time t_2 ($t_1 \leq t_2$). Throughout, we assume that y_{it} is univariate.

Let $d_{it} = 1$ if unit i receives the intervention at or before time t, and $d_{it} = 0$ otherwise. Let $\boldsymbol{d}_i =$ $(d_{i1}, \ldots, d_{iT})^\top$. Of the n units, the first n_1 remain untreated for the entire study period. We call these the *controls*. For the n_2 *treated* units, there is a time T_1 ($1 < T_1 < T$) immediately after which the intervention is applied. We assume that all treated units receive the intervention at the same time. Hence $d_{it} = 1$ if $i > n_1$ and $t > T_1$, and $d_{it} = 0$ otherwise. We make this assumption to simplify the notation, but the methods we describe can be easily extended to allow for different treatment times. The number of post-intervention observation times is denoted by $T_2 = T - T_1$. For each unit and time, we may also observe a set of K covariates $\boldsymbol{x}_{it} = (x_{it1}, \ldots, x_{itK})^\top$.

2.2 Potential Outcomes

We adopt the *potential outcomes* framework (Rubin, 1974, 1990), also known as the *Rubin causal model* (Holland, 1986, RCM). Under this model, for each treated unit ($i > n_1$) and post-intervention time ($t > T_1$) there are two potential outcomes, $y_{it}^{(0)}$ and $y_{it}^{(1)}$: $y_{it}^{(0)}$ represents the outcome that would be observed if intervention were not applied, and $y_{it}^{(1)}$ is the outcome that would be observed if the intervention were applied. We only observe $y_{it}^{(1)}$, that is, $y_{it} = y_{it}^{(1)}$. For the control units ($i \leq n_1$) at any time t and the treated units ($i > n_1$) at a pre-intervention time ($t \leq T_1$), we only define $y_{it}^{(0)}$ and $y_{it}^{(0)} = y_{it}$ is observed.

The RCM allows the effect of intervention unit i ($i > n_1$) at time t ($t > T_1$) to be expressed as $\tau_{it} = y_{it}^{(1)} - y_{it}^{(0)}$. Estimation of τ_{it} is complicated by the fact that $y_{it}^{(0)}$ is not observed. In order to estimate τ_{it} from the observed data, it is necessary to make identifying assumptions (Morgan and Winship, 2007, Keele and Minozzi, 2013). For the methods considered in this paper, these assumptions allow the unobserved *counterfactual* outcomes $y_{it}^{(0)}$ of treated units in the post-intervention period (i.e., for $i > n_1$ and $t > T_1$) to be predicted using the observed outcomes on control and treated units. Denote these predictions as $\hat{y}_{it}^{(0)}$. The intervention effect τ_{it} can then be estimated as $\hat{\tau}_{it} = y_{it} - \hat{y}_{it}^{(0)}$.

2.3 An Illustrative Example

As our illustrative example we use data from Abadie, Diamond and Hainmueller (2015) who investigated the effect that West Germany's reunification with East Germany in 1990 had on the economic growth of the former. To do so, they compared West Germany's annual per-capita GDP (the outcome variable) to its counterfactual GDP (i.e., its GDP had reunification not taken
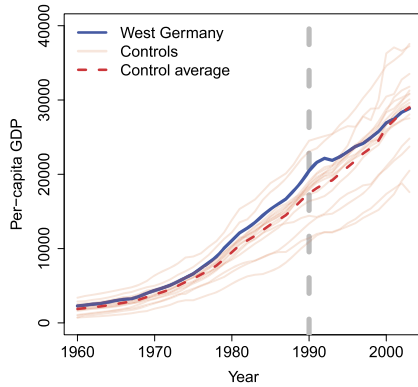
FIG. 1. *Time series plot of the German reunification data. The values on the y-axis represent per-capita GDP measured in U.S. dollars. West Germany's per-capita GDP is shown in blue; the data on control units (16 other OECD countries) are shown in light red; the dashed red line represents the average GDP of the control units. The dashed gray line indicates* 1990, *the year of reunification.*

place), which they predicted based on annual per-capita GDP data from $n_1 = 16$ member countries of the Organisation for Economic Co-operation and Development (OECD) (none of which underwent reunification, and so are 'control units'). The authors used data from 1960–2003 and hence there are $T_1 = 30$ pre-intervention and $T_2 = 13$ post-intervention time points. Figure 1 shows the time-series of the outcome on all 17 units. In Section 6, we analyse this dataset using the methods reviewed in this article.

3. ESTIMATION METHODS

In this section, we review four classes of methods for predicting the counterfactual treatment-free outcomes $y_{it}^{(0)}$ of the treated units at post-intervention times, needed to calculate $\hat{\tau}_{it}$. Here we focus on the intuition and the assumptions underlying each method, and report results on theoretical properties of unbiasedness and consistency in Appendix A. For full technical details of each approach, the reader is directed to the original publications.

3.1 Difference-in-Differences

Early works (Ashenfelter, 1978; Ashenfelter and Card, 1985; Card and Krueger, 1994) used so-called difference-in-differences (DID) models to compare two time periods (pre- versus post-intervention). The identifying assumption in DID models is that the average outcomes of control and treated units in the absence of an intervention would follow *parallel trends* over time (Abadie, 2005).

Figure 2 is a graphical representation of the basic DID method for a single control and single treated unit.
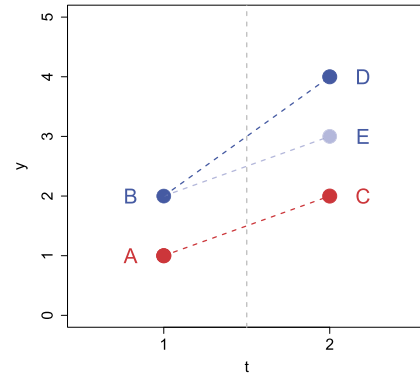

FIG. 2. *Graphical illustration of the difference-in-differences method.*

The four points A–D on the graph represent the control (A) and treated (B) units at $t = 1$, and the control (C) and treated (D) units at $t = 2$. Under the parallel trends assumption, the difference between the outcome of the treated unit and of the control unit would be constant over time in the absence of intervention. The counterfactual outcome for the treated unit at the post-intervention time can then be predicted as point E in Figure 2. Letting y_A, y_B, y_C, y_D and y_E denote the y-values corresponding to the points A, B, C, D and E in Figure 2, respectively, the estimated effect of the intervention is

$$\hat{\tau}_{22} = y_D - y_E$$
$$= y_D - \{y_C + \{y_B - y_A\}\}$$
$$= \{y_D - y_C\} - \{y_B - y_A\},$$

that is, the difference (after versus before) of the differences between the two units. The same method can be used when multiple time points and multiple control units are available.

A commonly used solution to adjust for the effect of covariates $\boldsymbol{x}_{it}$[1] is to specify a parametric linear DID model for the observed outcome y_{it} (Angrist and Pischke, 2009, Jones and Rice, 2011)

$$(3.1) \quad \begin{aligned} y_{it} &= y_{it}^{(0)} + \tau_{it} d_{it}, \\ y_{it}^{(0)} &= \boldsymbol{x}_{it}^\top \boldsymbol{\theta} + \kappa_i + \mu_t + \varepsilon_{it}, \end{aligned}$$

where $\boldsymbol{\theta}$ is a vector of regression coefficients, κ_i is an (unknown) fixed effect of unit i, μ_t allows for temporal trends and ε_{it} are the zero-mean error terms which

---

[1]Blundell et al. (2004) and Abadie (2005) propose alternative DID estimators that can account for the effect of covariates. However, these methods are not suitable when only a small number of units are treated and hence are not reviewed in this article.

are independent of d_{js}, $\boldsymbol{x}_{js}$, κ_j, μ_s for all i, j, t, s. Lagged outcomes and/or transformations of $\boldsymbol{x}_{it}$ can be included as extra covariates in the linear DID model (Jones and Rice, 2011). The parameters of the linear DID model can be estimated by ordinary least squares (OLS) regression (see Angrist and Pischke, 2009, pp. 167, for details). Let $\hat{\tau}_{it}^{\text{DID}}$ denote the resulting estimate of τ_{it}.

The linear DID model (3.1) makes very strong assumptions regarding the data generating mechanism. The term κ_i in (3.1) allows expected counterfactual treatment-free outcomes to be higher (or) lower in the treated units than in control units, even after adjusting for observed covariates $\boldsymbol{x}_{it}$. Hence, κ_i can represent an unobserved confounder. However, equation (3.1) assumes that the effect of this possible confounder on the outcome is constant over time. Similarly, the term μ_t in (3.1) can only account for temporal trends that are common to both treated and control units.

Although the linear DID specification (3.1) is often preferred in practice due to its simple interpretation and implementation, there exist other methods that build on the parallel trends idea. Athey and Imbens (2006) relax the linearity assumption of (3.1), allowing the outcome y_{it} to be a more general (nonlinear) function of the unobserved characteristics of unit i. However, it is difficult to implement their method when there are more than two time points and $\boldsymbol{x}_{it}$ is high-dimensional. Another method based on parallel trends is the triple differences method (Atanasov and Black, 2016, Wing, Simon and Bello-Gomez, 2018), which uses two groups of control units. For example, when the treated group consists of male employees of a company, then the control group can be either the female employees of the same company, or the male employees of a different company. In such situations, the triple differences method can use the second control group to correct for biases caused by the violation of the assumption of parallel trends between the outcomes of the treated group and the control group.

There are many examples of the use of DID models. Ashenfelter (1978) and Ashenfelter and Card (1985) investigate the effect of training programs on worker earnings. Card (1990) assesses the impact that the Mariel Boatlift, a mass migration of Cuban citizens to Miami in 1980, had on the city's labour market, using four other cities as controls. Card and Krueger (1994) estimate the effect that the increase of the minimum salary had on employment rates in New Jersey's fast-food industry in 1992, using fast-food restaurants located in Pennsylvania as the control group. See

Galiani, Gertler and Schargrodsky (2005), Branas et al. (2011), King et al. (2013) for recent works applications of the DID approach.

3.2 Latent Factor Models

In the linear DID model of equation (3.1), there is one unit-specific term, κ_i, and this can represent a single unobserved confounder whose effect on the outcome is constant over time. In the following latent factor model (LFM), κ_i is replaced by $\boldsymbol{\lambda}_i \boldsymbol{f}_t$

$$
\begin{aligned}
y_{it} &= y_{it}^{(0)} + \tau_{it} d_{it}, \\
y_{it}^{(0)} &= \boldsymbol{x}_{it}^\top \boldsymbol{\theta} + \boldsymbol{\lambda}_i^\top \boldsymbol{f}_t + \varepsilon_{it},
\end{aligned}
\tag{3.2}
$$

where $\boldsymbol{f}_t = (f_{1t}, \ldots, f_{Jt})^\top$ are J time-varying factors, $\boldsymbol{\lambda}_i = (\lambda_{i1}, \ldots, \lambda_{iJ})^\top$ are unit-specific factor loadings, and ε_{it} are the zero-mean errors which are independent of d_{js}, $\boldsymbol{x}_{js}$, $\boldsymbol{\lambda}_j$, $\boldsymbol{f}_s$ for all i, j, t, s. When $\boldsymbol{f}_t = (1, \mu_t)^\top$ and $\boldsymbol{\lambda}_i = (\kappa_i, 1)^\top$, the second line in (3.2) reduces to the second line in (3.1). So, the linear DID model is a special case of the LFM. Just as κ_i in the linear DID model can represent a single unobserved confounder, $\boldsymbol{\lambda}_i$ can represent J unobserved confounders, whose effect on the outcome varies with time and is described by $\boldsymbol{f}_t$. Hence, the LFM (3.2) relaxes the DID assumption that the average outcomes of control and treated units follow parallel trends. In econometrics, $\boldsymbol{f}_t$ is interpreted as a 'shock' that affects all units at time t and $\boldsymbol{\lambda}_i$ represents the response of unit i to these shocks (Bai, 2009).

Xu (2017) proposes a three-step estimation procedure for predicting counterfactual treatment-free outcomes using the LFM model. In the first step, observations on control units are used to estimate $\boldsymbol{\theta}$, $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_T$ and $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_{n_1}$ through the iterative procedure of Bai (2009) that minimises $\sum_{i=1}^{n_1} \sum_{t=1}^{T} (y_{it} - \hat{y}_{it}^{(0)})^2$, the mean squared error (MSE) between the observations y_{it} and the corresponding predicted values $\hat{y}_{it}^{(0)} = \boldsymbol{x}_{it}^\top \hat{\boldsymbol{\theta}} + \hat{\boldsymbol{\lambda}}_i^\top \hat{\boldsymbol{f}}_t$. In the second step, the estimated factor loadings for the treated units, $\hat{\boldsymbol{\lambda}}_i$ ($i > n_1$), are obtained conditional on the parameter estimates obtained in the first step by minimising the MSE between y_{it} and $\hat{y}_{it}^{(0)}$ for the treated units in the pre-intervention period. Finally, the third step involves estimating the intervention effects τ_{it} as $\hat{\tau}_{it}^{\text{XU}} = y_{it} - \hat{y}_{it}^{(0)}$.

Several authors have proposed alternative methods for predicting counterfactuals using the LFM. These include Ahn, Lee and Schmidt (2013), Gobillon and Magnac (2016), Chan and Kwok (2016) and Athey et al. (2017). We have focused on the method of Xu

because, to the best of our knowledge, it is the only one for which an R package has been developed.

Gobillon and Magnac (2016) and Xu (2017) describe applications of the LFM to real data. Gobillon and Magnac (2016) estimate the effect on unemployment rates of a French program offering tax reliefs to companies that hired at least 20% of their personnel from the local labour force. Xu (2017) evaluates the impact of Election Day Registration (EDR), a law that enables eligible citizens to register on site when they arrive at the voting centre, on voter turnout in the US. For more applications of the LFM see Kim and Oka (2014) and Sanso-Navarro, Sanz-Gracia and Vera-Cabello (2018).

3.3 Synthetic Control-Type Approaches

The original synthetic controls method (SCM) was developed by Abadie and Gardeazabal (2003) and Abadie, Diamond and Hainmueller (2010) and can only be applied to one treated unit at a time. The idea behind the SCM is to find weights $\boldsymbol{w} = (w_1, \ldots, w_{n_1})^\top$ for the control units such that the weighted average of the controls' outcomes best predicts (in terms of MSE) the outcome of the treated unit during the pre-intervention period, and then use the weights to estimate the counterfactual treatment-free outcomes in the post-intervention period. The set of weights $\boldsymbol{w}$ minimises

$$\sqrt{(\boldsymbol{y}_{n_1+1,1:T_1} - \boldsymbol{Y}^{\mathrm{c}}_{1:T_1}\boldsymbol{w})^\top \boldsymbol{V}(\boldsymbol{y}_{n_1+1,1:T_1} - \boldsymbol{Y}^{\mathrm{c}}_{1:T_1}\boldsymbol{w})},$$

(3.3)

subject to the constraints

$$(3.4) \qquad \sum_{i=1}^{n_1} w_i = 1 \quad \text{and} \quad w_i \geq 0,$$

where $\boldsymbol{Y}^{\mathrm{c}}_{1:T_1}$ is the $T_1 \times n_1$ matrix with ith column $\boldsymbol{y}_{i,1:T_1}$ and $\boldsymbol{V}$ is a $T_1 \times T_1$ symmetric, positive semidefinite matrix reflecting the importance given to the different pre-intervention time points (Abadie and Gardeazabal, 2003). The predicted counterfactual of the treated unit is

$$(3.5) \qquad \hat{\boldsymbol{y}}^{(0)}_{n_1+1,(T_1+1):T} = \boldsymbol{Y}^{\mathrm{c}}_{(T_1+1):T}\boldsymbol{w},$$

where $\boldsymbol{Y}^{\mathrm{c}}_{(T_1+1):T}$ is defined analogously to $\boldsymbol{Y}^{\mathrm{c}}_{1:T_1}$. The estimated intervention effect at times t ($t > T_1$) is then $\hat{\tau}^{\mathrm{SCM}}_{n_1+1,t} = y_{n_1+1,t} - \hat{y}^{(0)}_{n_1+1,t}$.

It is also possible to use the covariates, by replacing $\boldsymbol{y}_{i,1:T_1}$ with $\boldsymbol{z}_{i,1:T_1} = (\boldsymbol{y}^\top_{i,1:T_1}, \boldsymbol{x}^\top_{i,1:T_1})^\top$ in equation (3.3). Abadie, Diamond and Hainmueller (2010)

suggest that instead of using the full data $\boldsymbol{z}_{i,1:T_1}$, it may be reasonable to consider only a few summaries, such as the mean outcome $\frac{1}{T_1}\sum_{t=1}^{T_1} y_{it}$ in the pre-intervention period, and the corresponding means of the covariates, that is, to replace $\boldsymbol{z}_{i,1:T_1}$ by $(\frac{1}{T_1}\sum_{t=1}^{T_1} y_{it}, \frac{1}{T_1}\sum_{t=1}^{T_1} \boldsymbol{x}^\top_{it})^\top$. Such reduction of the dimensionality of $\boldsymbol{z}_{i,1:T_1}$ might be necessary in applications with $T_1 \gg n_1$ in order to reduce computation time.

The choice of matrix $\boldsymbol{V}$ can either be based on a subjective judgement of the relative importance of the variables in $\boldsymbol{y}_{i,1:T_1}$ or $\boldsymbol{z}_{i,1:T_1}$ or be determined through a data-driven approach. For example, Abadie and Gardeazabal (2003) and Abadie, Diamond and Hainmueller (2010) choose $\boldsymbol{V}$ as the positive definite diagonal matrix that minimises the MSE between the observed outcomes $\boldsymbol{y}_{i,1:T_1}$ ($i = 1, \ldots, T_1$) and estimated outcomes $\hat{\boldsymbol{y}}_{i,1:T_1} = \boldsymbol{Y}^{\mathrm{c}}_{1:T_1}\boldsymbol{w}(\boldsymbol{V})$ in the pre-intervention period, where $\boldsymbol{w}(\boldsymbol{V})$ is the solution to (3.3) for a fixed $\boldsymbol{V}$.

The SCM makes no assumptions regarding the data generating mechanism. The method has strong links with the matching literature, where the outcome of each treated individual is compared to the outcomes of controls with similar covariate values (Rosenbaum, 2002; Stuart, 2010). However, it is more general in the sense that a good match is sought by weighted averaging of the controls. The SCM also relates to the method of analogues used for time-series prediction. The difference is that in the method of analogues there is only one time-series and 'controls' are simply earlier segments of the time-series; for more details see Viboud et al. (2003).

There have been several proposed extensions of the SCM. To allow for multiple treated units, Kreif et al. (2016) apply the SCM to the averaged vector outcome $\bar{\boldsymbol{y}}^{\mathrm{tr}} = (\frac{1}{n_2}\sum_{i=n_1+1}^{n} y_{i1}, \ldots, \frac{1}{n_2}\sum_{i=n_1+1}^{n} y_{iT})^\top$ of the treated units. Acemoglu et al. (2016) assume that the intervention effects $\boldsymbol{\tau}_{n_1+1}, \ldots, \boldsymbol{\tau}_n$ are equal and estimate the common effect at time t ($t > T_1$) as the weighted average $\sum_{i=n_1+1}^{n} q_i^{-1}\hat{\tau}_{it} / \sum_{i=n_1+1}^{n} q_i^{-1}$, where $\hat{\tau}_{it}$ ($i > n_1$) is obtained by applying the original algorithm to just the data on treated unit i and the control units $1, \ldots, n_1$, and

$$q_i = \sqrt{T_1^{-1}(\boldsymbol{y}_{i,1:T_1} - \hat{\boldsymbol{y}}^{(0)}_{i,1:T_1})^\top(\boldsymbol{y}_{i,1:T_1} - \hat{\boldsymbol{y}}^{(0)}_{i,1:T_1})}.$$

Their stated rationale for using weights q_i^{-1} is that units with good fit in the pre-intervention period should be more reliable for estimating the common intervention effect and hence receive higher weights.

Hsiao, Ching and Wan (2012), henceforth HCW, and Doudchenko and Imbens (2016), henceforth DI, extend the SCM by adding a time-constant intercept term to the SCM estimator and removing the constraints on the weights. The intercept is necessary when the outcome of the treated unit is systematically (over time) higher or lower than the outcomes of the controls units and hence there exists no set of weights that can provide a good fit for $y_{n_1+1,t}$ in the pre-intervention period. The removal of the constraints on the weights is useful, for example, when there exist control units with outcomes that are negatively correlated with the outcomes on the treated unit. HCW suggest estimating $y_{n_1+1,t}^{(0)}$ $(t > T_1)$ as $\hat{y}_{n_1+1,t}^{(0)} = \beta_0 + \sum_{i=1}^{n_1} \beta_i y_{it}$, where $\beta_0, \ldots, \beta_{n_1}$ are the OLS coefficient estimates of the regression of $\boldsymbol{y}_{n_1+1,1:T_1}$ on $\boldsymbol{y}_{1,1:T_1}, \ldots, \boldsymbol{y}_{n_1,1:T_1}$, that is, they minimise

$$(3.6) \quad \begin{aligned} &(\boldsymbol{y}_{n_1+1,1:T_1} - \beta_0 \boldsymbol{1} - \boldsymbol{Y}_{1:T_1}^{c} \boldsymbol{\beta})^{\top} \\ &\cdot (\boldsymbol{y}_{n_1+1,1:T_1} - \beta_0 \boldsymbol{1} - \boldsymbol{Y}_{1:T_1}^{c} \boldsymbol{\beta}), \end{aligned}$$

where $\boldsymbol{1}$ denotes a T_1-vector of ones, β_0 is the intercept and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{n_1})^{\top}$. Amjad, Shah and Shen (2018) also remove the constraint on the weights and suggest that, before estimating these weights, the data on the control outcomes $\boldsymbol{Y}_{1:T_1}^{c}$ should be de-noised.

Ben-Michael, Feller and Rothstein (2018) introduced the augmented SCM. First, the SCM is applied and weights $w_1, \ldots, w_{n_1}$ obtained. Second, a model (e.g., a LFM) for the untreated outcomes $y_i^{(0)}$ of all $n_1 + 1$ units is fitted to all the outcomes of the untreated units and the pre-intervention outcomes of the untreated unit. If $\tilde{y}_{it}^{(0)}$ $(i = 1, \ldots, n_1+1; t > T_1)$ denote the predicted untreated outcomes from this model, then $\sum_{i=1}^{n_1} w_i y_{it} - \tilde{y}_{n_1+1,t}^{(0)}$ is an estimate of the bias of the SCM estimator. The augmented SCM estimator of the counterfactual $y_{n_1+1}^{(0)}$ equals the original SCM estimate plus this estimated bias. They argue that this method is particularly useful when the SCM method provides a poor fit in the pre-intervention period.

Hazlett and Xu (2018) estimate the weights using a kernel transformation of the pre-intervention outcomes. This is done to ensure that higher-order features of the outcomes (authors mention, e.g., volatility and variance) are taken into account when estimating the weights. Using simulated examples, they showed that their approach can eliminate biases that occur if the untransformed outcomes are used to estimate the weights.

Several recent works utilise synthetic control-type approaches for estimating the effects of an intervention. These include Cavallo et al. (2013), who examine the effect of large-scale natural disasters on

gross domestic product, and Ryan et al. (2016), who investigate the impact that UK's Quality and Outcomes Framework, a pay-for-performance scheme in primary health, had on population mortality. For more applications of synthetic control-type methods, see Billmeier and Nannicini (2013), Fujiki and Hsiao (2015), Saunders et al. (2015) and Aytuğ et al. (2017).

3.4 Causal Impact

The causal impact method (CIM) was introduced by Brodersen et al. (2015) and can only be applied to a single treated unit at a time. A Bayesian model is assumed for the outcome of the treated unit. This model includes a time-series component that relates the outcome of the treated unit at time t to previous outcomes on the same unit, and a regression component that uses the outcomes on control units as covariates. Specifically,

$$y_{n_1+1,t}^{(0)} = \beta_{0t} + \sum_{i=1}^{n_1} \beta_i y_{it} + \varepsilon_t \quad (t = 1, \ldots, T),$$

$$(3.7) \quad \beta_{0,t+1} = \beta_{0,t} + \delta_t + \eta_t,$$

$$\delta_{t+1} = \delta_t + \zeta_t,$$

with mutually independent $\varepsilon_{it} \sim \mathrm{N}(0, \sigma_\varepsilon^2)$, $\eta_t \sim \mathrm{N}(0, \sigma_\eta^2)$ and $\zeta_t \sim \mathrm{N}(0, \sigma_\zeta^2)$, and priors for $\beta_{00}, \delta_0, \beta_1, \ldots, \beta_{n_1}, \sigma_\varepsilon^2, \sigma_\eta^2$ and σ_ζ^2. In equations (3.7), the component β_{0t} induces temporal correlation in the outcome, the regression component $\sum_{i=1}^{n_1} \beta_i y_{it}$ relates $y_{n_1+1,t}^{(0)}$ to measurements from control units, and the error component ε_t accounts for unexplained variability. More complex models can be adopted (Brodersen et al., 2015), for example, by adding a seasonal component.

The model (3.7) is fitted to the observed data, $y_{n_1+1,1}^{(0)}, \ldots, y_{n_1+1,T_1}^{(0)}$, treating the counterfactuals $y_{n_1+1,T_1+1}^{(0)}, \ldots, y_{n_1+1,T}^{(0)}$ as unobserved random variables. Independent, improper, uniform priors are used for $\tau_{n_1+1,T_1+1}, \ldots, \tau_{n_1+1,T}$. Then, L samples $y_{n_1+1,t}^{(0,l)}$ $(l = 1, \ldots, L)$ are drawn from the resulting posterior predictive distribution of the counterfactual outcome $y_{n_1+1,t}^{(0)}$ $(t > T_1)$, thus providing samples $y_{n_1+1,t} - y_{n_1+1,t}^{(0,l)}$ from the posterior distribution of $\tau_{n_1+1,t}$. Typically, this would be done using a Markov chain Monte Carlo algorithm. A point estimate $\hat{\tau}_{n_1+1,t}^{\mathrm{CIM}}$ for the causal effect $\tau_{n_1+1,t}$ at time t $(t > T_1)$ is then given by its posterior mean.

Bruhn et al. (2017) use the CIM to assess the impact of pneumococcal conjugate vaccines on pneumonia-related hospitalisations using hospitalisations from

other diseases as the control time-series. de Vocht et al. (2017) evaluate the benefits of stricter alcohol licensing policies on alcohol-related hospitalisations in several areas, control areas being other areas where these policies were not implemented. See also de Vocht (2016), González and Hosoda (2016), Vizzotti et al. (2016) for other applications of the CIM.

4. QUANTIFICATION OF UNCERTAINTY & HYPOTHESIS TESTING

We now describe approaches to estimating standard errors and testing the null hypothesis that $\tau_{it} = 0$, or, for Bayesian methods, estimating the posterior distribution of τ_{it}.

DID. If it is assumed that the errors ε_{it} in the linear DID are mutually independent and homoscedastic, variance estimates for the OLS estimates of τ_{it} ($i > n_1$, $t > T_1$) are easy to obtain. These represent the variance of $\hat{\tau}_{it}^{\text{DID}}$ over repeated samples of the errors ε_{it} holding $(\boldsymbol{x}_{11}^{\top}, \ldots, \boldsymbol{x}_{1T}^{\top}, \kappa_1, \boldsymbol{d}_1^{\top}, \ldots, \boldsymbol{x}_{n1}^{\top}, \ldots, \boldsymbol{x}_{nT}^{\top}, \kappa_n, \boldsymbol{d}_n^{\top}, \mu_1, \ldots, \mu_T)^{\top}$ fixed. A Wald test for $\tau_{it} = 0$ can then be performed. However, the assumption that the errors ε_{it} are mutually independent may not be plausible. Bertrand, Duflo and Mullainathan (2004) show that when, as is likely, the errors $\varepsilon_{i1}, \ldots, \varepsilon_{iT}$ are serially correlated, the variance estimator for $\hat{\tau}_{it}$ is biased downwards and type-I error rates are inflated, and they describe methods to deal with this. Standard errors can also be underestimated if there are correlations due to units being grouped (e.g., hospitals within the same county); Donald and Lang (2007) discuss possible solutions.

LFM. Xu (2017) uses parametric bootstrap to obtain confidence intervals for $\hat{\tau}_{it}^{\text{XU}}$ and p-values, assuming that $\varepsilon_{1t}, \ldots, \varepsilon_{nt}$ are independent and homoscedastic at each individual time t. Repeated sampling here is of the errors ε_{it} holding $(\boldsymbol{x}_{11}^{\top}, \ldots, \boldsymbol{x}_{1T}^{\top}, \boldsymbol{\lambda}_1^{\top}, \boldsymbol{d}_1^{\top}, \ldots, \boldsymbol{x}_{n1}^{\top}, \ldots, \boldsymbol{x}_{nT}^{\top}, \boldsymbol{\lambda}_n^{\top}, \boldsymbol{d}_n^{\top}, \boldsymbol{f}_1^{\top}, \ldots, \boldsymbol{f}_T^{\top})^{\top}$ fixed. Li (2018) derive the asymptotic distribution (as $T_1, T_2 \to \infty$) of the average effect $\sum_{t=T_1+1}^{T} \hat{\tau}_{it}^{\text{XU}}$ for the ith treated unit.

Synthetic control-type approaches. Abadie, Diamond and Hainmueller (2010, 2015) argue that traditional statistical inference is difficult in this setting, unless one is prepared to assume that the unit that received the intervention was chosen at random. Under that assumption, a standard permutation test would provide a valid p-value for the null hypothesis that treatment would have no effect on any of the units (i.e., $\tau_{it} = 0$ for all $i = 1, \ldots, n_1 + 1$). Abadie, Diamond and Hainmueller propose using a very similar test even

in settings where the intervention is not randomly assigned and called this a 'placebo test'. They argue that such a test provides an alternative mode of inference, saying that our confidence that a large treatment effect estimate truly reflects the effect of the intervention would be undermined if similarly large effect estimates were obtained when the treatment labels of the units were permuted.

More specifically, Abadie, Diamond and Hainmueller (2010, 2015) compare $\hat{\tau}_{n_1+1,t}^{\text{SCM}}$ to $\hat{\tau}_{1t}^{\text{SCM}}, \ldots, \hat{\tau}_{n_1 t}^{\text{SCM}}$, the estimated effects considering each of the control units in turn as though it had been the treated unit, and using the remaining $n_1 - 1$ controls to estimate the weights, at each post-intervention time. Their test statistic r_i is

$$(4.1) \quad r_i = \frac{T_1 (\boldsymbol{y}_{i,T_1+1:T} - \hat{\boldsymbol{y}}_{i,T_1+1:T})^{\top} (\boldsymbol{y}_{i,T_1+1:T} - \hat{\boldsymbol{y}}_{i,T_1+1:T})}{T_2 (\boldsymbol{y}_{i,1:T_1} - \hat{\boldsymbol{y}}_{i,1:T_1})^{\top} (\boldsymbol{y}_{i,1:T_1} - \hat{\boldsymbol{y}}_{i,1:T_1})},$$

that is, the ratio of post- to pre-intervention MSE between the observed and predicted outcomes. The predicted counterfactual of control unit i ($i \leq n_1$) is obtained by applying the SC method to that unit, using the remaining $n_1 - 1$ controls to find weights. Their intuition is that under the null hypothesis the predictive ability of the SCM should be similar in the two periods and thus the ratio r_{n_1+1} close to 1. Hence, a value of r_{n_1+1} that lies in the tail of the empirical distribution of $r_1, \ldots, r_{n_1+1}$ can be viewed as evidence for a nonzero intervention effect. Firpo and Possebom (2018) investigate the impact that the choice of the test statistic has on the results of Abadie, Diamond and Hainmueller's test, finding that r_i outperformed alternative statistics that they considered in several performance measures.

Firpo and Possebom (2018) propose a generalisation of Abadie, Diamond and Hainmueller's placebo test. Rather than giving equal weight to all possible permutations of the treatment labels when calculating the p-value, they make the weights depend on a sensitivity parameter ϕ. The reasoning is that even if the unit that received the intervention had actually been chosen at random, some units might have been more likely to be chosen, thus making some permutations of the treatment labels more probable than others. Firpo and Possebom (2018) vary the value of ϕ and assess how robust to this value is the conclusion of a treatment effect (or lack thereof).

Amjad, Shah and Shen (2018) take an empirical Bayes approach to test the hypothesis that $\tau_{n_1+1,t} = 0$. They assume that $\boldsymbol{y}_{n_1+1,1:T} \sim \mathcal{N}(\tilde{\boldsymbol{Y}}_{1:T_1}^{\text{c}} \boldsymbol{w}, \sigma^2 \boldsymbol{I})$, where $\tilde{\boldsymbol{Y}}_{1:T_1}^{\text{c}}$ are the de-noised control outcomes obtained via singular value thresholding, and the weights $\boldsymbol{w}$ have

a $\mathcal{N}(0, \sigma_{\boldsymbol{w}}^2 \boldsymbol{I})$ prior distribution, for some value of $\sigma_{\boldsymbol{w}}^2$. The posterior distribution of $\boldsymbol{w}$ can be used to calculate the posterior predictive distribution of $y_{n_1+1,t}^{(0)}$ $(t > T_1)$. Let a and b be the 97.5th and 2.5th centiles of this distribution. The 95% posterior credible interval for $\tau_{n_1+1,t}$ is $(y_{n_1+1,t} - a, y_{n_1+1,t} - b)$.

Abadie, Diamond and Hainmueller (2010) also consider a variant of their placebo test in which the time of the intervention, rather than the unit that receives intervention, is changed. They do not, however, propose this as being a way to calculate a p-value.

In their application, HCW fit an autoregressive model to the estimated intervention effects $\hat{\tau}_{n_1+1,T_1+1}^{\mathrm{HCW}}$, $\ldots, \hat{\tau}_{n_1+1,T}^{\mathrm{HCW}}$. They then test the null hypothesis that the mean of these effects, which they refer to as the long-run intervention effect, equals zero. In their implementation of the HCW method, Gardeazabal and Vega-Bayo (2017) use a test that is equivalent to the test proposed by Abadie, Diamond and Hainmueller (2010, 2015) for the SCM, to test if $\tau_{it} = 0$ for all i and $t > T_1$. As pointed out by one of the referees, an intuitive approach to obtain confidence intervals for $\hat{\tau}_{n_1+1,t}^{\mathrm{HCW}}$ would be to use the bootstrap. Finally, Li and Bell (2017) derive the asymptotic distribution (as $T_1, T_2 \to \infty$) of the average effect $\sum_{t=T_1+1}^{T} \hat{\tau}_{n_1+1,t}^{\mathrm{HCW}}$.

CIM. For the CIM, a 95% posterior credible interval for $\tau_{n_1+1,t}$ can be calculated as $(y_{n_1+1,t} - a, y_{n_1+1,t} - b)$, where a and b denote, respectively, the 97.5th and 2.5th centiles of the posterior predictive distribution of the counterfactual $y_{n_1+1,t}^{(0)}$.

All methods. Recently, there has been work building upon the end-of-sample stability test (Andrews, 2003). For a single treated unit, the idea is that under the hypothesis of no intervention effect, the process $y_{n_1+1,t} - \hat{y}_{n_1+1,t}^{(0)}$ $(t = 1, \ldots, T)$ is stationary. Chernozhukov, Wüthrich and Zhu (2017) propose a permutation procedure to test the stationarity of this process and show that their approach gives valid inference for several methods including DID, LFM and SCM. Hahn and Shi (2017) apply the same idea to the SCM. Both note that confidence sets for the intervention effect can be obtained by statistic inversion.

5. IMPLEMENTATION ISSUES

In this section, we discuss issues related to the practical implementation of the methods presented in Section 3: model choice and diagnostic checks.

5.1 Model Choice

Choosing the control units. When implementing the SCM, HCW, DI and CIM, it may be desirable to exclude some of the potential control units. Using all potential controls might result in nonunique causal effect estimates when there are more such controls than pre-intervention time points. Moreover, standard errors of estimates can be reduced by discarding controls whose outcomes are not related to the outcome of the treated unit.

HCW develop a two-stage approach to exclude potential controls. For each $\ell = 1, \ldots, n_1$ they implement their method $\binom{n_1}{\ell}$ times, where each time they use a different subset of size ℓ of the control units. For each ℓ they choose the subset that maximises the regression R^2 and thus obtain n_1 candidate models. They recommend choosing one of these n_1 models according to a model selection criterion such as the AIC. An alternative approach was suggested by Li and Bell (2017), who use the least absolute shrinkage and selection operator (LASSO) to select controls.

DI exclude potential controls by encouraging some of the weights β_i to shrink towards (or even equal to) zero. This achieved by including the penalty term

$$(5.1) \qquad \rho \left(\frac{1-\delta}{2} \sum_{i=1}^{n_1} \beta_i^2 + \phi \sum_{i=1}^{n_1} |\beta_i| \right),$$

in the objective function (3.6), where ρ and ϕ are penalty parameters. For CIM Brodersen et al. (2015) induce sparsity on the vector $(\beta_1, \ldots, \beta_{n_1})^\top$ that describes the dependence on controls by using a spike-and-slab prior.

Choosing the covariates. An issue that may arise when implementing the linear DID and the LFM methodology of Xu (2017) is the choice of covariates to include. Exclusion of potential covariates may be desirable for the same reasons that one might exclude control units. For the linear DID model, covariates, which may include lagged outcomes and interactions of lagged outcomes with the covariates, may be selected by imposing sparsity on the regression coefficient vector, using, for example, the LASSO. For the LFM, one can use the factor-lasso approach of Hansen and Liao (2019).

When implementing the SCM, users need to decide which variables (pre-intervention outcomes, covariates or summaries of these) to use to determine the weights. Ferman, Pinto and Possebom (2016) demonstrate that the estimated counterfactual may differ depending on which variables are used. Dube and Zipperer (2015)

develop the following approach for selecting among K sets of variables. First, for every set k ($k = 1, \ldots, K$) they apply the SCM to the data on every control unit in turn, and calculate the predicted outcomes $\hat{y}_{it}^{(k)}$ ($i = 1, \ldots, n_1$) based on the estimated weights. Then, they choose the set k^* that minimises the mean (over control units) MSE between observed and predicted outcomes $\hat{y}_{it}^{(k)}$ in the post-intervention period. An alternative approach when T_1 is large is to split the pre-intervention data into a training dataset, to which the SCM is applied using different sets of variables, and a validation dataset, which is used to assess which set has the best predictive performance.

Other issues. Some of the methods for estimating the parameters of the LFM (Gobillon and Magnac, 2016; Chan and Kwok, 2016, Xu, 2017) require that the number of factors J be chosen. The usual approach is to fit the LFM for various values of J and determine the optimal J using cross-validation. An alternative for choosing J is to use the procedures developed by Bai and Ng (2002). However, these approaches provide estimates of standard errors that do not account for the uncertainty about J.

For the CIM, practitioners need to decide what dynamical components to include in the counterfactual model. Similar methods to those used for choosing variables for the SCM can be used. An alternative is to fit several models and use the one that achieves the optimal trade-off between accuracy (the difference $y_{n_1+1,t} - \hat{y}_{n_1+1,t}$) and precision (the length of the credible interval for $y_{n_1+1,t}$) in the pre-intervention period. In small datasets the inferences provided by the CIM impact method will be sensitive to the choice of prior distributions. Therefore, these specifications should ideally be determined based on expert opinion.

5.2 Diagnostics

All the methods described in this article make assumptions about the counterfactual outcomes of the treated units in the post-intervention period. Since these outcomes cannot be observed, it is never possible to test the full set of assumptions. Nonetheless, it is sometimes possible to assess the validity of a subset of these assumptions using data from the pre-intervention period.

When no covariates are used and $T_1 > 1$, an informal check of the parallel trends assumption of DID methods can be conducted by plotting the average outcomes of control and treated units in the pre-intervention period (Keele and Minozzi, 2013): an approximately constant (over time) distance between the two lines suggests that parallel trends is plausible. The SCM should

not be used when the outcome of the treated unit lies outside the convex hull of the outcomes of controls units. This can be checked by plotting the time-series of the outcome on all the units.

The fit provided for the outcomes on treated units in the pre-intervention period can be used as a diagnostic check. Intuitively, if a model is not predictive of the outcome in the pre-intervention period, it is less likely to provide good predictions for the counterfactuals in the post-intervention period. Goodness-of-fit can be assessed using the MSE between the observed and predicted values. However, in order for a good pre-intervention fit to be reassuring, one needs to establish that it does not occur due to overfitting, as can be the case *e.g.* for the SCM when $n_1 > T_1$. For the methods that provide fitted values for the outcomes of control units, that is, the linear DID model and the method of Xu (2017), one can further use the fit over the post-intervention period for these units as a diagnostic tool.

Finally, for both the linear DID model and the LFM method of Xu (2017), extrapolation biases may occur when the covariates (and loadings for the LFM) of treated and control units do not share a common support. In order to exclude the possibility of such biases, it suffices to ensure that the characteristics of the treated units are not extreme compared to the characteristics of control units. When a small number of covariates (and factors) is used, one can visually compare the two groups for each covariate (and loading) in turn. If this is not feasible, methods for multivariate outlier detection (e.g., Filzmoser, Maronna and Werner, 2008) can be used to identify treated units with extreme characteristics.

6. APPLICATION: EFFECT OF GERMAN REUNIFICATION ON GDP

In this section, we demonstrate the use of the methods we have described by analysing the data introduced in Section 2.3. The dataset is publicly available.[2] We omit the available covariates because they might have been affected by the reunification.

For the DID and SCM, some of the diagnostic checks described in Section 5.2 do not require implementing these methods and therefore we started by carrying out these tests. Figure 4 of Appendix B shows the difference between West Germany's GDP, $y_{17,t}$, and the average GDP in the control countries, $\frac{1}{16} \sum_{i=1}^{16} y_{it}$, over the pre-reunification period. The difference has a clear

---

[2]http://dx.doi.org/10.7910/DVN/24714

increasing trend suggesting that the parallel trends assumption does not hold, so the linear DID model is not appropriate for this application. As we see from Figure 1, the outcome of the treated unit lies in the convex hull of the outcomes of control units so this provides no evidence that the SCM should not be used.

We only implement methods for which (to the best of our knowledge) R (R Core Team, 2016) software exists. The linear DID method can be implemented using any linear regression function (*e.g.*, *lm*). For the remaining methods, we used the packages specifically developed for these methods: *gsynth* for the LFM; *Synth* (Abadie, Diamond and Hainmueller, 2011) for the SCM; *pamp* (Vega-Bayo, 2015) for the HCW method; and *CausalImpact* for the CIM. The code we used for our real data analysis is available online.[3]

We fitted the linear DID model (3.1). For the method of Xu (2017) we set $f_{1t} = 1$ for all t and $\lambda_{i2} = 1$ for all i in order to have time and country fixed effects, respectively. The total number of latent factors was set via cross-validation. For the SCM, we estimated the weights using the whole vector of outcomes $y_{i,1:T_1}$ in the pre-intervention period (rather than summaries of the outcomes). The HCW method was implemented using all control countries and pre-intervention time points. Finally, for the CIM we fitted the model of equation (3.7) but without the term δ_t because we found that inclusion of this term did not improve the fit and led to substantially wider credible intervals for the causal effect of interest. The prior distributions for all model parameters were set to the software defaults. We fitted the linear DID method for illustration purposes even though DID should not be used here.

Before examining the causal estimates, we performed the remaining diagnostic checks. Figure 3 shows the difference between the actual and estimated counterfactual West German GDP, $y_{17,\cdot} - \hat{y}_{17,\cdot}$ for the entire study period. We see that all methods except for the linear DID almost perfectly reproduce West Germany's GDP before reunification. Thus, the pre-intervention goodness-of-fit diagnostic provides no indication against any of the methods except for linear DID. The estimated factor loadings for the 17 countries in the dataset are shown in Table 2 of Appendix B. The estimated loadings for West Germany are not extreme compared to the estimated loadings of the control countries, hence suggesting that the predicted counterfactual is not obtained by extrapolation.

Overall we see that the only method that fails our diagnostic checks is the linear DID.

Figure 3 reveals that the other four methods provide similar estimates of the causal effect. In particular, the difference between the observed and counterfactual outcomes is positive during the first three years after 1989, suggesting that reunification initially had a positive impact on West Germany's GDP. Abadie, Diamond and Hainmueller (2015) attribute this to a 'demand boom'. The estimated impact reduces thereafter, and is negative for all four methods in year 2003. The estimated average reduction in annual GDP over the period 1990–2003 due to the reunification (we also show DID for completeness) is shown in Table 1.

Figure 3 presents 95% intervals for the LFM of Xu (2017) and the CIM. These exclude zero in all years after 1993 thus suggesting a significant intervention effect. The placebo test of no intervention effect in any of the years 1990–2003 described by Abadie, Diamond and Hainmueller (2010, 2015) is also suggestive of a nonzero intervention effect. In particular, the r statistic defined in equation (4.1) is $r_{17} = 30.72$ for West Germany, larger than all the r_i values obtained for the 16 control countries. We further implemented this test with the HCW method. The rank of the r statistic for West Germany is 16, that is, there is only one country whose r statistic is higher. Table 3 in Appendix B shows the r statistics obtained by applying the SCM and HCW methods. Overall, taking into consideration all tests conducted, we conclude that there is evidence that reunification had a negative long-term impact on West Germany's per-capita GDP, although it may have had a positive short-term impact.

7. DISCUSSION

7.1 Connections Between Methods

There are several ways in which the methods described in this paper relate to one another.

First, for the case of a single treated unit and no covariates, most of them propose counterfactual estimators of the form $\hat{y}_{n_1+1,t}^{(0)} = \alpha_t + \sum_{i=1}^{n_1} \beta_i y_{it}$ ($t > T_1$) with α_t and β_i being estimated using the data from the pre-intervention period (i.e., $t \leq T_1$). For the DID method, the parallel trends assumption implies that $\alpha_t = \frac{1}{T_1} \sum_{s=1}^{T_1} (y_{n_1+1,s} - \frac{1}{n_1} \sum_{i=1}^{n_1} y_{is})$ for all $t > T_1$ and $\beta_i = \frac{1}{n}$ for all $i \leq n_1$ (Chernozhukov, Wüthrich and Zhu, 2017). The SCM assumes $\alpha_t = 0$ for all t and requires that $\beta_1, \ldots, \beta_n$ are nonnegative and sum to one. The HCW and DI methods impose the constraint that
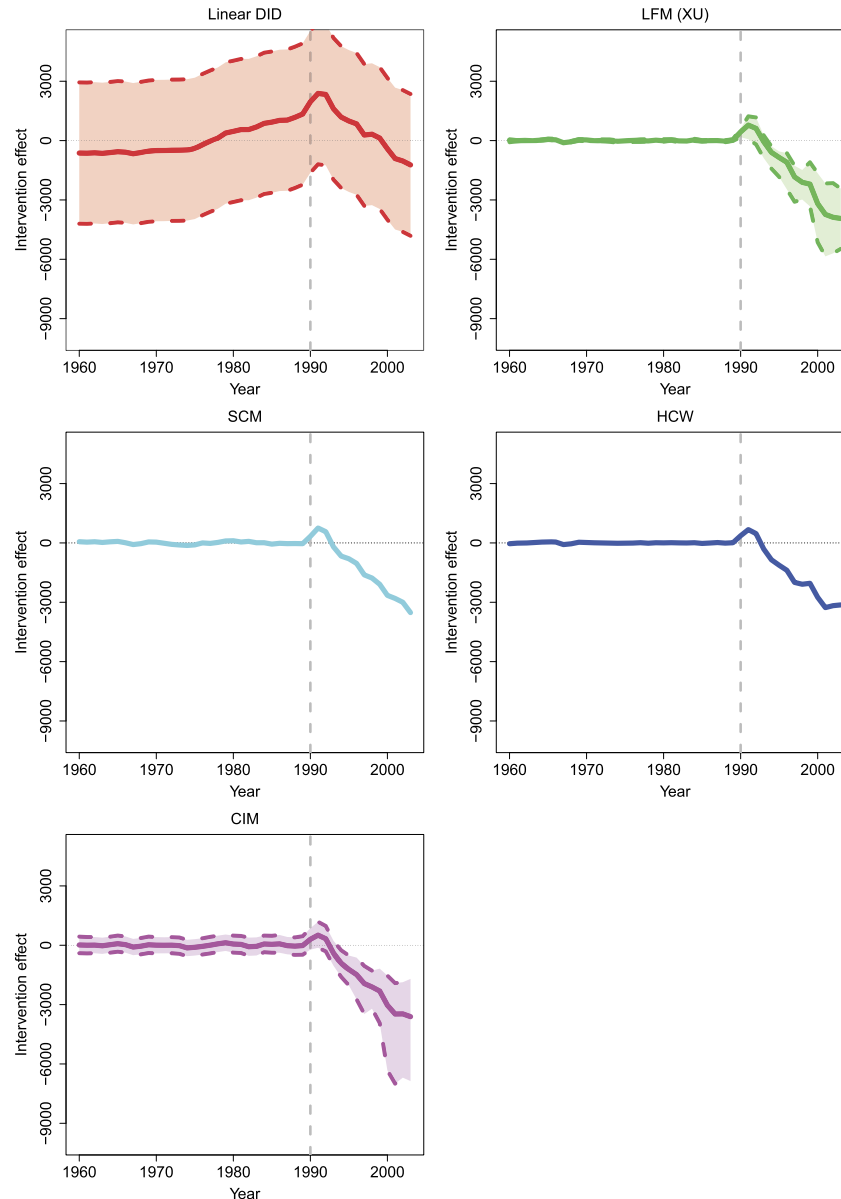
FIG. 3.   *Annual estimates of the effect of the German reunification on West Germany's per-capita GDP obtained using the linear DID model* (*red*), *the LFM* (*green*), *the SCM method* (*light blue*), *the HCW method* (*blue*) *and the CIM* (*purple*). *The dashed lines* (*when applicable*) *represent the* 95% *confidence/credible intervals.*

the intercept is constant over time, that is, $\alpha_t = \alpha$ for all t. Finally, the CIM assumes that α_t obeys a time-series model (e.g., a random walk model). When there are covariates these similarities break down because methods account for covariates in a different way.

Second, most of the methods relate to the LFM. We have already seen that the linear DID model (3.1) is a special case of the LFM (3.2). As discussed in Appendix A, the SCM and HCW estimators are asymptotically unbiased when the true data-generating mechanism obeys a LFM. We expect that due to their similar-

ities with the HCW estimator just explained, both DI and CIM estimators will be unbiased under the same LFM.

7.2 Recommendations for Implementation

None of the methods is universally superior to the others. Extensive simulation experiments comparing the relative performance of a subset of them have been conducted by multiple authors including Gobillon and Magnac (2016), O'Neill et al. (2016), Gardeazabal and Vega-Bayo (2017), Xu (2017) and Kinn (2018). They

TABLE 1
Average (over the period 1990–2003) reduction in West Germany's annual per capita GDP, as estimated by the five methods. All values are in United States dollars

Method	GDP decrease
Linear DID	−604
LFM (XU)	1546
SCM	1322
HCW	1473
CIM	1629

all find settings in which one of the methods outperforms the others. However, the findings from these simulation studies may not generalise to other data generating mechanisms. Practitioners should choose the method to apply on the basis of the characteristics of the dataset and, in particular the values, of n_1, T_1 and their ratio n_1/T_1.

The DID method can be used for any n_1 and T_1. As explained in Sections 3.1 and 7.1, the DID method arguably requires the strongest assumptions. As a result, it may provide more precise estimates of the intervention effects compared to the other methods. However, these estimates might be severely biased when the parallel trends assumption does not hold (see simulation studies by O'Neill et al., 2016 and Gobillon and Magnac, 2016). This occurred in our application (Section 6), where the DID estimate of the average reunification effect had opposite sign compared to all the other estimates. Hence, it is essential to test the plausibility of parallel trends in the pre-intervention period before applying DID to a dataset. This is easy when there are no covariates.

The LFM can be used for any value of n_1/T_1. However, both n_1 and T_1 should be at least moderate in size in order to accurately estimate the factors and loadings, respectively. For example, for the asymptotic unbiasedness property of Xu's method (see Appendix A) to be relevant to a finite sample, they recommend $T_1 > 10$ and $n_1 > 40$.

Synthetic control approaches are mostly suited for applications where T_1 is large. This is required to accurately estimate the relationships between the outcome of the treated unit and the outcomes of control units. When $n_1 \geq T_1$ regularisation is required because the number of parameters exceeds the number of observations.[4] Regularisation is possible for both the HCW

and DI estimators, as described in Section 5.1, but not for the SCM. The SCM should not be used when the outcome of the treated unit does not lie in the convex hull of the outcomes of control units.

The CIM is similar in spirit to synthetic control approaches and also requires large T_1. Because of its time-series component it can work even in cases when the outcome of treated unit is not correlated to the outcomes of control units. However, it requires larger T_1 than synthetic control-type approaches to estimate the additional time-series parameters. In practice, the value of T_1 required will depend on the complexity of the time-series model. When $n_1 > T_1$ regularisation can be achieved via a spike-and-slab prior on the regression coefficients.

In applications where certain covariates are known to be highly predictive of the outcome, it is preferable to use the linear DID or LFM. This is because they use the covariates of control units and therefore can estimate the regression parameters of the predictive covariates with higher precision compared to the HCW, DI and CIM.[5] This can in turn lead to more precise estimates of the counterfactuals. Covariates that are potentially affected by the intervention should not be included when using any of the methods except the SCM, because the treatment-free values of the covariate are not observed in the post-intervention period. The SCM can use these covariates in the pre-intervention period to estimate the weights.

There will be applications where more than one method is appropriate. This is to be expected considering their connections explained in Section 7.1. For example, the SCM, HCW, DI and CIM estimators are all well-suited when T_1 is large and there are few control units. In such cases, users might choose any of these methods. However, it is still worth applying the remaining methods in order to check that conflicting results are not obtained. Methods that perform poorly on diagnostic checks or are based on assumptions that seem unrealistic for the dataset of interest should not be considered. Even within the same method a sensitivity analysis is recommended. This can be carried out by implementing the method using different model specifications as explained in Section 5.1. Ideally, results obtained from the different models should not conflict. For the SCM, HCW, DI and CIM, one can re-implement these methods excluding control units that

---

[4]Synthetic control approaches regress the outcome of the treated unit on the outcomes of the control units. Therefore we can think of $y_{1:n,t}$ as a single data point (observation) in a regression model.

[5]Although one might argue that for the HCW, DI and CIM, the effect of covariates is taken into account through the outcomes of control units which the covariates affect.

received large coefficients (or weights for the SCM) in the first implementation, to provide reassurance that results are not driven by a single control unit.

7.3 Connections with Matching

Our review does not cover matching methods even though some forms of matching are suitable for application in the setting that we are investigating. This is because we view the SCM as the best suited matching method in this setting: by using data on all control units it attempts to construct an exact match for the treated unit.[6]

However, matching can be used prior to applying the methods described in this paper, to restrict the pool of controls to those with similar characteristics to the treated units. This approach has been adopted for the DID (O'Neill et al., 2016), LFM (Gobillon and Magnac, 2016) and CIM (Schmitt, Tull and Atwater, 2018). For the DID method, Ryan et al. (2018) showed that matching can reduce biases that occur when the parallel trends assumption is violated. For a detailed overview of the matching literature in the context of causal inference with observational data, see Rosenbaum (2002), Chapter 10, or Stuart (2010). See Imai, Kim and Wang (2018) for matching techniques for time-series data.

8. PROPOSALS FOR FUTURE RESEARCH

There remain several open problems. Most existing methods do not fully account for autocorrelation in the outcome of the treated unit measured over time. In particular, the treatment effect estimates obtained by any of the methods except for the CIM are invariant to permutation of the time labels in the pre-intervention period. There may be potential gains in efficiency by extending these methods to account for structure over time.

The SCM, HCW, DI and CIM assume a linear relationship between the outcome of the treated unit and the outcomes of control units but this is a strong assumption. Carvalho, Masini and Medeiros (2018) account for nonlinear relationships by regressing $y_{n_1+1,t}$ on transformations of the outcomes of control units but it is hard to choose which transformations to use. Therefore, it would be worth estimating the relationship between $y_{n_1+1,t}$ and the outcomes of the control

units nonparametrically using, for example, machine learning techniques.

The methods we have described are designed to be applied to a single outcome. In the majority of applications there are several outcomes that may be affected by the intervention. For example, in the case study of Section 2.3 we have considered per-capita GDP but there are alternative indexes, such as the unemployment rate, which we could instead be examined. Modelling of all outcomes jointly may provide a more precise estimate of the causal effect of intervention on any one of them. Although Robbins, Saunders and Kilmer (2017) provide an extension of the SCM method for multiple outcomes, the other methods could also benefit from being extended to handle multiple outcomes.

Another possible direction for future research is to develop models that take into account geographic location of units. In many applications, one might expect the outcomes on units with spatial proximity to be correlated. It would be useful to develop models that incorporate these correlations. Lopes, Salazar and Gamerman (2008) present a Bayesian LFM that models the correlation between the loadings of any two units as a function of the distance between these units. Their model could be used to estimate intervention effects with minor modifications.

We will investigate some of these problems in our future work.

APPENDIX A: UNBIASEDNESS AND CONSISTENCY

Unbiased or asymptotically unbiased estimates of τ_{it} can be obtained with all four methods described in this review. For each one of them, we now describe the sampling framework and the main assumptions for the unbiasedness to hold. For ease of exposition, we choose not to list some technical regularity conditions required for the results presented to hold; readers can refer to the original publications for these.

DID. For the linear DID estimator, we make use of some well-known results for OLS regression, see e.g. Wooldridge (2013). If the DID model of equation (3.1) holds then $\hat{\tau}_{it}^{\text{DID}}$ is unbiased, that is,

$$\mathbb{E}[\hat{\tau}_{it}^{\text{DID}}] = \tau_{it},$$

where the expectation is taken with respect to the conditional distribution of $\boldsymbol{\varepsilon} = (\varepsilon_{11}, \ldots, \varepsilon_{1T}, \ldots, \varepsilon_{n1}, \ldots, \varepsilon_{nT})^\top$ given $\boldsymbol{S}_n$, where $\boldsymbol{S}_n = (\boldsymbol{x}_{11}^\top, \ldots, \boldsymbol{x}_{1T}^\top, \kappa_1, \boldsymbol{d}_1^\top, \ldots, \boldsymbol{x}_{n1}^\top, \ldots, \boldsymbol{x}_{nT}^\top, \kappa_n, \boldsymbol{d}_n^\top)^\top$. That is, $\boldsymbol{S}_n$ is common to the repeated samples but the errors $\boldsymbol{\varepsilon}$ differ in each repeated sample.

---

[6]HCW, DI and CIM also attempt to construct an exact match for the treated units. However, these methods may rely on extrapolation, which is not done in matching approaches.

LFM. Xu (2017) studies the properties of $\hat{\tau}_{it}^{\mathrm{XU}}$. If the LFM of equation (3.2) holds then under some regularity conditions (which include weak serial correlation of the error terms within each unit) $\hat{\tau}_{it}^{\mathrm{XU}}$ is asymptotically unbiased, that is,

$$\mathbb{E}[\hat{\tau}_{it}^{\mathrm{XU}}] \to \tau_{it}$$

as $n_1 \to \infty$ and $T_1 \to \infty$, where the expectation is taken with respect to the conditional distribution of $\boldsymbol{\varepsilon}$ given S_n, where $S_n = (x_{11}^{\top}, \ldots, x_{1T}^{\top}, \lambda_1^{\top}, d_1^{\top}, \ldots, x_{n1}^{\top}, \ldots, x_{nT}^{\top}, \lambda_n^{\top}, d_n^{\top}, f_1^{\top}, \ldots, f_T^{\top})^{\top}$. Intuitively, we require that both n_1 and T_1 are large in order to accurately estimate factors at each post-intervention time point and loadings for the treated units, respectively, which we need in order to predict the counterfactual outcomes.

Synthetic control-type approaches. These methods do not assume a generative model, but rather exploit linear relationships between the data on the treated and control units in order to construct a counterfactual. Such relationships may arise from various data-generating mechanisms. Hence, their unbiasedness properties can be studied under any of these.

Assume that the LFM

$$(A.1) \quad \begin{aligned} y_{it} &= y_{it}^{(0)} + \tau_{it} d_{it}, \\ y_{it}^{(0)} &= \mu_t + x_i^{\top} \theta_t + \lambda_i^{\top} f_t + \varepsilon_{it}, \end{aligned}$$

holds, the error terms ε_{it} have zero mean given S_n and $\varepsilon_{it} \perp\!\!\!\perp \varepsilon_{js}$ given S_n ($i \neq j$ and $t \neq s$), where μ_t are time fixed-effects, $x_i = (x_{i1}, \ldots, x_{iK})^{\top}$ are time-invariant covariates and $S_n = (x_1^{\top}, \lambda_1^{\top}, d_1^{\top}, \ldots, x_n^{\top}, \lambda_n^{\top}, d_n^{\top}, f_1^{\top}, \ldots, f_T^{\top}, \mu_1, \ldots, \mu_t)^{\top}$. Abadie, Diamond and Hainmueller (2010) show that if there exist $\varpi_1, \ldots, \varpi_{n_1}$ such that

$$(A.2) \quad \begin{aligned} \lambda_{n_1+1} &= \sum_{i=1}^{n_1} \varpi_i \lambda_i, \\ x_{n_1+1} &= \sum_{i=1}^{n_1} \varpi_i x_i, \end{aligned}$$

then under some regularity conditions $\hat{\tau}_{n_1+1,t}^{\mathrm{SC}}$ is asymptotically unbiased, that is,

$$\mathbb{E}[\hat{\tau}_{n_1+1,t}^{\mathrm{SC}}] \to \tau_{n_1+1,t}$$

as $T_1 \to \infty$, where the expectation is taken with respect to the conditional distribution of $\boldsymbol{\varepsilon}$ given S_n. The conditions (A.2) imply that both observed (x_{n_1+1}) and unobserved (λ_{n_1+1}) characteristics of the treated unit lie in the convex hull of the characteristics of control units, thus allowing interpolation. When this is not true (i.e., when such $\varpi_1, \ldots, \varpi_{n_1}$ do not exist, thus forcing extrapolation to be used), the SCM estimator will be generally biased (Gobillon and Magnac, 2016; Ferman and Pinto, 2016).

Assume the following variant of the LFM:

$$(A.3) \quad y_{it}^{(0)} = \kappa_i + \lambda_i^{\top} f_t + \varepsilon_{it},$$

where κ_i are unit fixed effects and ε_{it} are zero-mean, homoscedastic error terms which are independent of f_s for all t, s and independent of d_{js} for all $i \neq j$. Hsiao, Ching and Wan (2012) prove that if there exist $\gamma_1, \ldots, \gamma_{n_1}$ such that

$$(A.4) \quad \lambda_{n_1+1,j} = \sum_{i=1}^{n_1} \gamma_i \lambda_{ij}$$

is true for every $j = 1, \ldots, J$ (along with some technical conditions), then $\hat{\tau}_{n_1+1,t}^{\mathrm{HCW}}$ is unbiased, that is,

$$\mathbb{E}[\hat{\tau}_{n_1+1,t}^{\mathrm{HCW}}] = \tau_{n_1+1,t},$$

where the expectation is taken with respect to the conditional distribution of $\boldsymbol{\varepsilon}$ given S_n, where $S_n = (\kappa_1, \lambda_1^{\top}, d_1^{\top}, \ldots, \kappa_n, \lambda_n^{\top}, d_n^{\top}, f_1^{\top}, \ldots, f_T^{\top})^{\top}$.

CIM. If the CIM of equations (3.7) is the true data-generating model and the prior on the vector of model parameters $\boldsymbol{\vartheta} = (\beta_1, \ldots, \beta_{n_1}, \sigma_\varepsilon^2, \sigma_\eta^2, \sigma_\zeta^2)^{\top}$ assigns nonzero probability to its true value, then the posterior distribution of $\boldsymbol{\vartheta}$ will converge to a point mass on its true value as $T_1 \to \infty$. Consequently, the posterior mean of $y_{n_1+1,t}^{(0)}$ will converge to its true value, and so $\hat{\tau}_{n_1+1,t}^{\mathrm{CIM}}$ is an asymptotically unbiased estimate of of $\tau_{n_1+1,t}$ (as $T_1 \to \infty$).

The above results concern (asymptotic) unbiasedness. Consistent estimation of τ_{it} is not feasible, unless it is assumed that $\tau_{it} = \tau_i$ or $\tau_{it} = \tau_t$, that is, that the unit-specific treatment effects are the same at all post-intervention times or (when $n > n_1 + 1$) are the same at each time for all treated units. This is because, regardless of how many units and time-points there are, $y_{it}^{(1)}$ is only measured once for each $i > n_1$ and $t > T_1$. It is not uncommon to assume that $\tau_{it} = \tau$ (e.g., Angrist and Pischke, 2009; Gobillon and Magnac, 2016). When this is done, existing results for the linear DID model (e.g., Wooldridge, 2013) and the LFM method of Xu (2017) (e.g., Bai, 2009) imply consistency of the estimator of τ when either of those methods are used. These results, though, require some additional technical assumptions to hold. Alternatively, a looser structure could be imposed on τ_{it}. For example, HCW assume that $\tau_{n_1+1,T_1+1}, \ldots, \tau_{n_1+1,T}$, is an auto-regressive moving-average process. This enables the mean of this process to be consistently estimated.

TABLE 2
Factor loadings for the 17 countries, as obtained by fitting the LFM of Xu (2017) to the West German reunification data

Country	Factor									
	1	2	3	4	5	6	7	8	9	10
West Germany	−0.75	−0.16	0.29	0.26	−0.15	−0.12	−0.80	0.35	−0.42	−1.42
Australia	−0.14	0.47	0.91	−0.55	1.71	−0.62	−0.34	−0.64	−0.25	0.93
Austria	−0.50	−0.52	−0.17	0.18	−0.69	−1.06	0.26	0.45	−2.28	0.21
Belgium	−0.22	−0.46	0.10	1.02	−0.44	−0.32	0.98	1.61	−0.43	−0.20
Denmark	−0.34	0.24	0.21	−1.34	0.09	−1.01	−0.93	−0.25	1.24	0.02
France	−0.14	−0.33	0.05	0.11	−0.45	−0.18	−0.95	1.12	−0.19	−0.92
Greece	1.95	−0.02	1.52	1.10	−0.55	−0.88	0.17	−1.99	−0.01	−1.54
Italy	−0.03	−0.72	−0.62	−0.65	−0.39	−0.26	−0.49	0.48	−0.85	−1.14
Japan	−0.33	−1.18	−2.04	−0.30	−0.07	−0.53	2.30	−1.28	1.15	0.12
Netherlands	−0.45	0.31	0.02	1.30	0.07	−1.68	−0.83	1.08	1.99	0.74
New Zealand	1.32	−0.12	1.40	−2.35	−0.26	0.35	1.24	1.10	0.14	0.96
Norway	−1.60	2.55	0.29	−0.15	−2.04	0.83	0.53	−0.70	0.08	−0.05
Portugal	1.73	0.49	−2.18	−0.58	−0.53	0.92	−1.74	−0.17	0.02	0.14
Spain	0.98	0.85	−0.44	1.57	0.74	0.54	0.37	−0.16	−0.93	2.16
Switzerland	−0.71	−2.19	1.10	0.70	−0.71	2.23	−0.77	−0.49	0.77	0.77
UK	0.10	0.89	−0.10	0.49	1.92	1.58	0.88	1.04	0.61	−1.92
USA	−1.63	−0.24	−0.05	−0.55	1.58	0.07	−0.68	−1.20	−1.05	−0.28

APPENDIX B: REAL DATA SUPPLEMENTARY ANALYSIS

In this section, we provide supplementary material for the data analysis of Section 6. Figure 4 shows the parallel trends diagnostic check described in Section 5.2. The estimated factor loadings obtained by applying the LFM method of Xu (2017) are shown in Table 2. Finally, the *r* statistics obtained by applying the empirical test of Abadie, Diamond and Hainmueller (2015) with the SCM and HCW methods are shown in Table 3.
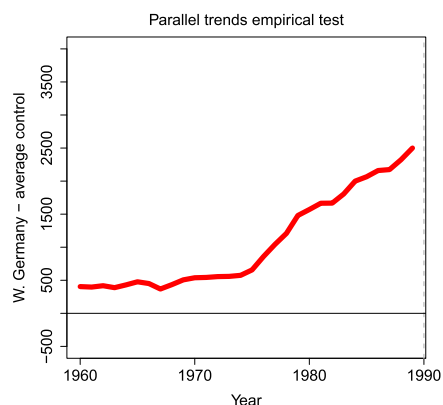

FIG. 4. *Difference over time between West Germany's per capita GDP and the average of control countries. Rather than being constant, the difference increases over time thus suggesting that the DID parallel trends assumption might not be plausible.*

TABLE 3
r statistics obtained by applying the empirical test of Abadie, Diamond and Hainmueller (2015) to the German reunification data, for both the SCM and HCW methods

Country	Method	
	SCM	HCW
West Germany	30.72	71.53
Australia	6.86	16.83
Austria	4.24	45.04
Belgium	4.52	16.90
Denmark	6.25	21.35
France	8.00	52.07
Greece	7.83	18.72
Italy	20.48	46.72
Japan	4.87	29.73
Netherlands	20.44	36.83
New Zealand	5.16	16.99
Norway	13.78	77.36
Portugal	0.70	57.69
Spain	7.53	14.32
Switzerland	2.36	29.75
UK	6.94	30.27
USA	5.97	42.55

MC_UU_00002/10 (SRS) and MC_UU_00002/11 (DDA, AMP), Public Health England (DDA), and NIHR PGfAR RP-PG-0616-20008 (EPIToPe, PS, MH, DDA). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

REFERENCES

ABADIE, A. (2005). Semiparametric difference-in-differences estimators. *Rev. Econ. Stud.* **72** 1–19. MR2116973

ABADIE, A., DIAMOND, A. and HAINMUELLER, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *J. Amer. Statist. Assoc.* **105** 493–505. MR2759929

ABADIE, A., DIAMOND, A. and HAINMUELLER, J. (2011). Synth: An R package for synthetic control methods in comparative case studies. *Journal of Statistical Software* **42**.

ABADIE, A., DIAMOND, A. and HAINMUELLER, J. (2015). Comparative politics and the synthetic control method. *Amer. J. Polit. Sci.* **59** 495–510.

ABADIE, A. and GARDEAZABAL, J. (2003). The economic costs of conflict: A case study of the Basque country. *Am. Econ. Rev.* **93** 113–132.

ACEMOGLU, D., JOHNSON, S., KERMANI, A., KWAK, J. and MITTON, T. (2016). The value of connections in turbulent times: Evidence from the United States. *J. Financ. Econ.* **121** 368–391.

AHN, S. C., LEE, Y. H. and SCHMIDT, P. (2013). Panel data models with multiple time-varying individual effects. *J. Econometrics* **174** 1–14. MR3036957

AMJAD, M., SHAH, D. and SHEN, D. (2018). Robust synthetic control. *J. Mach. Learn. Res.* **19** Paper No. 22, 51. MR3862429

ANDREWS, D. W. K. (2003). End-of-sample instability tests. *Econometrica* **71** 1661–1694. MR2015416

ANGRIST, J. D. and PISCHKE, J.-S. (2009). *Mostly Harmless Econometrics*: *An Empiricist's Companion*. Princeton Univ. Press, Princeton, NJ.

ANTONAKIS, J., BENDAHAN, S., JACQUART, P. and LALIVE, R. (2010). On making causal claims: A review and recommendations. *Leadersh. Q.* **21** 1086–1120.

ASHENFELTER, O. (1978). Estimating the effect of training programs on earnings. *Rev. Econ. Stat.* **60** 47–57.

ASHENFELTER, O. and CARD, D. (1985). Using the longitudinal structure of earnings to estimate the effect of training programs. *Rev. Econ. Stat.* **67** 648–660.

ATANASOV, V. and BLACK, B. (2016). Shock-based causal inference in corporate finance and accounting research. *Critical Finance Review* **5** 207–304.

ATHEY, S. and IMBENS, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica* **74** 431–497. MR2207397

ATHEY, S., BAYATI, M., DOUDCHENKO, N., IMBENS, G. and KHOSRAVI, K. (2017). Matrix completion methods for causal panel data models. Preprint. Available at arXiv:1710.10251.

AUSTIN, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivar. Behav. Res.* **46** 399–424.

AYTUĞ, H., KÜTÜK, M. M., ODUNCU, A. and TOGAN, S. (2017). Twenty years of the EU–Turkey customs union: A synthetic control method analysis. *J. Common Mark. Stud.* **55** 419–431.

BAI, J. (2009). Panel data models with interactive fixed effects. *Econometrica* **77** 1229–1279. MR2547073

BAI, J. and NG, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191–221. MR1926259

BEN-MICHAEL, E., FELLER, A. and ROTHSTEIN, J. (2018). The augmented synthetic control method. Preprint. Available at arXiv:1811.04170.

BERTRAND, M., DUFLO, E. and MULLAINATHAN, S. (2004). How much should we trust differences-in-differences estimates? *Q. J. Econ.* **119** 249–275.

BILLMEIER, A. and NANNICINI, T. (2013). Assessing economic liberalization episodes: A synthetic control approach. *Rev. Econ. Stat.* **95** 983–1001.

BLUNDELL, R., DIAS, M. C., MEGHIR, C. and VAN REENEN, J. (2004). Evaluating the employment impact of a mandatory job search program. *J. Eur. Econ. Assoc.* **2** 569–606.

BRANAS, C. C., CHENEY, R. A., MACDONALD, J. M., TAM, V. W., JACKSON, T. D. and TEN HAVE, T. R. (2011). A difference-in-differences analysis of health, safety, and greening vacant urban space. *Am. J. Epidemiol.* **174** 1296–1306.

BRODERSEN, K. H., GALLUSSER, F., KOEHLER, J., REMY, N. and SCOTT, S. L. (2015). Inferring causal impact using Bayesian structural time-series models. *Ann. Appl. Stat.* **9** 247–274. MR3341115

BRUHN, C. A., HETTERICH, S., SCHUCK-PAIM, C., KÜRÜM, E., TAYLOR, R. J., LUSTIG, R., SHAPIRO, E. D., WARREN, J. L., SIMONSEN, L. et al. (2017). Estimating the population-level impact of vaccines using synthetic controls. *Proc. Natl. Acad. Sci. USA* **114** 1524–1529.

CARD, D. (1990). The impact of the Mariel boatlift on the Miami labor market. *Ind. Labor Relat. Rev.* **43** 245–257.

CARD, D. and KRUEGER, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *Am. Econ. Rev.* **84** 772–793.

CARVALHO, C., MASINI, R. and MEDEIROS, M. C. (2018). ArCo: An artificial counterfactual approach for high-dimensional panel time-series data. *J. Econometrics* **207** 352–380. MR3868187

CAVALLO, E., GALIANI, S., NOY, I. and PANTANO, J. (2013). Catastrophic natural disasters and economic growth. *Rev. Econ. Stat.* **95** 1549–1561.

CHAN, M. K. and KWOK, S. (2016). Policy evaluation with interactive fixed effects. Preprint. Available at https://ideas.repec.org/p/syd/wpaper/2016-11.html.

CHERNOZHUKOV, V., WÜTHRICH, K. and ZHU, Y. (2017). An exact and robust conformal inference method for counterfactual and synthetic controls. Preprint. Available at arXiv:1712.09089.

DE VOCHT, F. (2016). Inferring the 1985–2014 impact of mobile phone use on selected brain cancer subtypes using Bayesian structural time series and synthetic controls. *Environ. Int.* **97** 100–107.

DE VOCHT, F., TILLING, K., PLIAKAS, T., ANGUS, C., EGAN, M., BRENNAN, A., CAMPBELL, R. and HICKMAN, M. (2017). The intervention effect of local alcohol licensing policies on hospital admission and crime: A natural experiment

using a novel Bayesian synthetic time-series method. *J. Epidemiol. Community Health* **71** 912–918.

DONALD, S. G. and LANG, K. (2007). Inference with difference-in-differences and other panel data. *Rev. Econ. Stat.* **89** 221–233.

DOUDCHENKO, N. and IMBENS, G. W. (2016). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Preprint. Available at arXiv:1610.07748.

DUBE, A. and ZIPPERER, B. (2015). Pooling multiple case studies using synthetic controls: An application to minimum wage policies.

FERMAN, B. and PINTO, C. (2016). Revisiting the synthetic control estimator.

FERMAN, B., PINTO, C. and POSSEBOM, V. (2016). Cherry picking with synthetic controls.

FILZMOSER, P., MARONNA, R. and WERNER, M. (2008). Outlier identification in high dimensions. *Comput. Statist. Data Anal.* **52** 1694–1711. MR2422764

FIRPO, S. and POSSEBOM, V. (2018). Synthetic control method: Inference, sensitivity analysis and confidence sets. *J. Causal Inference* **6**.

FUJIKI, H. and HSIAO, C. (2015). Disentangling the effects of multiple treatments—measuring the net economic impact of the 1995 great Hanshin-Awaji earthquake. *J. Econometrics* **186** 66–73. MR3321526

GALIANI, S., GERTLER, P. and SCHARGRODSKY, E. (2005). Water for life: The impact of the privatization of water services on child mortality. *J. Polit. Econ.* **113** 83–120.

GARDEAZABAL, J. and VEGA-BAYO, A. (2017). An empirical comparison between the synthetic control method and Hsiao *et al.*'s panel data approach to program evaluation. *J. Appl. Econometrics* **32** 983–1002. MR3689015

GLASS, T. A., GOODMAN, S. N., HERNÁN, M. A. and SAMET, J. M. (2013). Causal inference in public health. *Annu. Rev. Public Health* **34** 61–75.

GOBILLON, L. and MAGNAC, T. (2016). Regional policy evaluation: Interactive fixed effects and synthetic controls. *Rev. Econ. Stat.* **98** 535–551.

GONZÁLEZ, R. and HOSODA, E. B. (2016). Environmental impact of aircraft emissions and aviation fuel tax in Japan. *J. Air Transp. Manag.* **57** 234–240.

HAHN, J. and SHI, R. (2017). Synthetic control and inference. *Econometrics* **5** 52.

HANSEN, C. and LIAO, Y. (2019). The factor-lasso and *k*-step bootstrap approach for inference in high-dimensional economic applications. *Econometric Theory* **35** 465–509. MR3947253

HAZLETT, C. and XU, Y. (2018). Trajectory balancing: A general reweighting approach to causal inference with time-series cross-sectional data.

HOLLAND, P. W. (1986). Statistics and causal inference. *J. Amer. Statist. Assoc.* **81** 945–970. MR0867618

HSIAO, C., CHING, H. S. and WAN, S. K. (2012). A panel data approach for program evaluation: Measuring the benefits of political and economic integration of Hong Kong with mainland China. *J. Appl. Econometrics* **27** 705–740. MR2949893

IMAI, K., KIM, I. S. and WANG, E. (2018). Matching methods for causal inference with time-series cross-section data.

IMBENS, G. W. and WOOLDRIDGE, J. M. (2009). Recent developments in the econometrics of program evaluation. *J. Econ. Lit.* **47** 5–86.

JONES, A. M. and RICE, N. (2011). Econometric evaluation of health policies. In *The Oxford Handbook of Health Economics* (S. Glied and P. C. Smith, eds.). *Oxford Handbooks* 890–923. Oxford Univ. Press, Oxford.

KEELE, L. (2015). The statistics of causal inference: A view from political methodology. *Polit. Anal.* **23** 313–335.

KEELE, L. and MINOZZI, W. (2013). How much is Minnesota like Wisconsin? Assumptions and counterfactuals in causal inference with observational data. *Polit. Anal.* **21** 193–216.

KIM, D. and OKA, T. (2014). Divorce law reforms and divorce rates in the USA: An interactive fixed-effects approach. *J. Appl. Econometrics* **29** 231–245. MR3233244

KING, M., ESSICK, C., BEARMAN, P. and ROSS, J. S. (2013). Medical school gift restriction policies and physician prescribing of newly marketed psychotropic medications: Difference-in-differences analysis. *BMJ* **346** f264.

KINN, D. (2018). Synthetic control methods and big data. Preprint. Available at arXiv:1803.00096.

KREIF, N., GRIEVE, R., HANGARTNER, D., TURNER, A. J., NIKOLOVA, S. and SUTTON, M. (2016). Examination of the synthetic control method for evaluating health policies with multiple treated units. *Health Econ.* **25** 1514–1528.

LI, K. (2018). Inference for factor model based average treatment effects.

LI, K. T. and BELL, D. R. (2017). Estimation of average treatment effects with panel data: Asymptotic theory and implementation. *J. Econometrics* **197** 65–75. MR3598646

LOPES, H. F., SALAZAR, E. and GAMERMAN, D. (2008). Spatial dynamic factor analysis. *Bayesian Anal.* **3** 759–792. MR2469799

LOPEZ BERNAL, J., CUMMINS, S. and GASPARRINI, A. (2016). Interrupted time series regression for the evaluation of public health interventions: A tutorial. *Int. J. Epidemiol.* **46** 348–355.

MORGAN, S. L. and WINSHIP, C. (2007). *Counterfactuals and Causal Inference*: *Methods and Principles for Social Research. Analytical Methods for Social Research*. Cambridge Univ. Press, Cambridge.

O'NEILL, S., KREIF, N., GRIEVE, R., SUTTON, M. and SEKHON, J. S. (2016). Estimating causal effects: Considering three alternatives to difference-in-differences estimation. *Health Serv. Outcomes Res. Methodol.* **16** 1–21.

R CORE TEAM (2016). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

ROBBINS, M. W., SAUNDERS, J. and KILMER, B. (2017). A framework for synthetic control methods with high-dimensional, micro-level data: Evaluating a neighborhood-specific crime intervention. *J. Amer. Statist. Assoc.* **112** 109–126. MR3646556

ROSENBAUM, P. R. (2002). *Observational Studies*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR1899138

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974

ROTHMAN, K. J. and GREENLAND, S. (2005). Causation and causal inference in epidemiology. *Am. J. Publ. Health* **95** S144–S150.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688.

RUBIN, D. B. (1990). Formal mode of statistical inference for causal effects. *J. Statist. Plann. Inference* **25** 279–292.

RUBIN, D. B. and WATERMAN, R. P. (2006). Estimating the causal effects of marketing interventions using propensity score methodology. *Statist. Sci.* **21** 206–222. MR2324079

RYAN, A. M., KRINSKY, S., KONTOPANTELIS, E. and DORAN, T. (2016). Long-term evidence for the effect of pay-for-performance in primary care on mortality in the UK: A population study. *Lancet* **388** 268–274.

RYAN, A. M., KONTOPANTELIS, E., LINDEN, A. and BURGESS, J. F. JR. (2018). Now trending: Coping with non-parallel trends in difference-in-differences analysis. *Stat. Methods Med. Res.* DOI:10.1177/0962280218814570.

SANSO-NAVARRO, M., SANZ-GRACIA, F. and VERA-CABELLO, M. (2018). The demographic impact of terrorism: Evidence from municipalities in the Basque Country and Navarre. *Regional Studies* **0** 1–11.

SAUNDERS, J., LUNDBERG, R., BRAGA, A. A., RIDGEWAY, G. and MILES, J. (2015). A synthetic control approach to evaluating place-based crime interventions. *J. Quant. Criminol.* **31** 413–434.

SCHMITT, E., TULL, C. and ATWATER, P. (2018). Extending Bayesian structural time-series estimates of causal impact to many-household conservation initiatives. *Ann. Appl. Stat.* **12** 2517–2539. MR3875710

STUART, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statist. Sci.* **25** 1–21. MR2741812

VARIAN, H. R. (2016). Causal inference in economics and marketing. *Proc. Natl. Acad. Sci. USA* **113** 7310–7315.

VEGA-BAYO, A. (2015). An R package for the panel approach method for program evaluation. *R J.* **7**.

VIBOUD, C., BOËLLE, P.-Y., CARRAT, F., VALLERON, A.-J. and FLAHAULT, A. (2003). Prediction of the spread of influenza epidemics by the method of analogues. *Am. J. Epidemiol.* **158** 996–1006.

VIZZOTTI, C., JUAREZ, M. V., BERGEL, E., ROMANIN, V., CALIFANO, G., SAGRADINI, S., RANCAÑO, C., AQUINO, A., LIBSTER, R. et al. (2016). Impact of a maternal immunization program against pertussis in a developing country. *Vaccine* **34** 6223–6228.

WING, C., SIMON, K. and BELLO-GOMEZ, R. A. (2018). Designing difference in difference studies: Best practices for public health policy research. *Annu. Rev. Public Health* **39** 453–469.

WOOLDRIDGE, J. M. (2013). *Introductory Econometrics*: *A Modern Approach*. Cengage Learning, Mason.

XU, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Polit. Anal.* **25** 57–76.