Ecological Applications, 31(3), 2021, e02264

Evaluating natural experiments in ecology: using synthetic controls in assessments of remotely sensed land treatments

Stephen E. Fick [1,2], Travis W. Nauman [1], Colby C. Brungard [2], and Michael C. Duniway [1,3]

[1]*US Geological Survey, Southwest Biological Science Center, Moab, Utah 84054 USA*
[2]*Department of Plant and Environmental Sciences, New Mexico State University, Las Cruces, New Mexico 88003 USA*

Citation: Fick, S. E., T. W. Nauman, C. C. Brungard, and M. C. Duniway. 2021. Evaluating natural experiments in ecology: using synthetic controls in assessments of remotely sensed land-treatments. Ecological Applications 31(3):e02264. 10.1002/eap.2264

Abstract. Many important ecological phenomena occur on large spatial scales and/or are unplanned and thus do not easily fit within analytical frameworks that rely on randomization, replication, and interspersed a priori controls for statistical comparison. Analyses of such large-scale, natural experiments are common in the health and econometrics literature, where techniques have been developed to derive insight from large, noisy observational data sets. Here, we apply a technique from this literature, synthetic control, to assess landscape change with remote sensing data. The basic data requirements for synthetic control include (1) a discrete set of treated and untreated units, (2) a known date of treatment intervention, and (3) time series response data that include both pre- and post-treatment outcomes for all units. Synthetic control generates a response metric for treated units relative to a no-action alternative based on prior relationships between treated and unexposed groups. Using simulations and a case study involving a large-scale brush-clearing management event, we show how synthetic control can intuitively infer treatment effect sizes from satellite data, even in the presence of confounding noise from climate anomalies, long-term vegetation dynamics, or sensor errors. We find that accuracy depends on the number and quality of potential control units, highlighting the importance of selecting appropriate control populations. Although we consider the synthetic control approach in the context of natural experiments with remote sensing data, we expect the methodology to have wider utility in ecology, particularly for systems with large, complex, and poorly replicated experimental units.

Key words: *causal analysis; time series; remote sensing; simulation; land treatments.*

Introduction

The problem

Many important ecological phenomena occur on large spatial scales or are unplanned and thus do not easily fit within analytical frameworks that rely on randomized, replicated, and interspersed a priori controls for statistical comparison. Analytical problems endemic to large-scale experiments and other ecological events, are well documented and have elicited lively debate (Oksanen 2001, 2004, Hurlbert 2004). For instance, manipulations of whole lakes, watersheds, islands, forests, or other large-scale ecosystems may be impossible to replicate, and therefore inappropriate for frequentist statistical approaches (Hurlbert 1984), but still worthy of formal assessment (Carpenter 1998). Other targets of manipulation may be complex or lack discrete boundaries (e.g., marine systems; Wernberg et al. 2012), making it difficult to identify suitable nearby analogues for comparison. As many traditional statistical approaches may be inappropriate for these types of data, there is a need for ways to efficiently derive quantitative insights about the effects of large-scale experiments, ecological events, or manipulations.

In a management or policy context, effective decision-making requires inference from past manipulations and ecological events as part of the adaptive management cycle (Williams 2011). Although historic management actions or "interventions" may be plentiful and widespread (Copeland et al. 2018), adaptive management is often limited by lack of monitoring data and the means to distinguish treatment effects from other confounding influences through controls and replication. For instance, the effectiveness of a rangeland planting may be ambiguous if subsequent recruitment was coincident with abnormally high precipitation and *natural* recruitment in the months following treatment. Without simultaneous monitoring of sites with similar ecological potential and ambient conditions, it is difficult to discriminate true treatment effects from coincident noise (Larsen et al. 2019). While some management efforts do integrate experimental elements such as replication,

Manuscript received 7 February 2020; revised 7 June 2020; accepted 16 August 2020. Corresponding Editor: David Schimel.
[3] E-mail: mduniway@usgs.gov

randomization or basic controls into their design (e.g., Karl et al. 2014, Bestelmeyer et al. 2019), the logistical cost of such designs make them rare in application settings. With the growing availability of large observational environmental data sets and spatially explicit records of management activities, there is both opportunity for new ecological insight and a simultaneous need for tools to effectively parse intervention effects from confounding signals.

Insights from social science

Analytical challenges related to large, poorly replicated, and uncontrolled phenomena are common in other disciplines including political science, public health, and economics, where quantifying the effects of policies or other events (economic "shocks," disease outbreaks) are critical for understanding large and complex systems (Larsen et al. 2019). In these disciplines, a host of analytical tools and methods have been developed to quantify the causal effects of a given event, despite the limitations imposed by small sample sizes, nonrandom exposure of experimental units, heterogenous confounders through time, and lack of a priori control groups (Craig et al. 2017). These techniques often place emphasis on identifying or generating proper comparisons among treated and untreated groups, such as the methods of propensity score matching (Dehejia and Wahba 2002), regression discontinuity (Imbens and Lemieux 2008), and difference in differences (Ashenfelter and Card 1985).

One relatively novel technique for causal analysis in the absence of predefined references is the "synthetic control" method, emerging from the econometrics literature (Abadie and Gardeazabal 2003, Abadie et al. 2010). This approach attempts to reconstruct what would have happened if a treatment had not occurred (a "counterfactual"), based on the pre-intervention relationship between the unit of interest and a population of unaffected units. It is particularly useful for cases with few treated units and a potentially large number of imperfectly matched control groups, such as when entire countries are the targets of analysis. For example, Abadie et al. (2015) estimate the effect of the German reunification in 1990 on the GDP of West Germany, using a weighted composite of countries sharing similar characteristics. They estimate that by 2003, West German GDP would have been almost 8% higher without reunification.

The synthetic control approach seeks to generate a composite counterfactual by functionally relating patterns in treated units to candidate controls using only data from the pre-treatment period, then extrapolating this function into the post-treatment period. A key innovation of synthetic control is the use of data-driven methods for selecting and weighting controls, an approach particularly suited for "wide" data sets with a large number of predictors (controls and/or covariates) relative to treated observations. While several different methods have been used "under the hood" to model this relationship (Kinn 2018), all methods share a set of general requirements about the data: (1) a known date of treatment intervention, (2) a known group of units not influenced by the treatment intervention, and (3) a time series spanning pre- and post-treatment event for all control and treated units. Given these constraints, a potentially large number of modeling approaches are available to generate such a synthetic control, each with their own set of assumptions and strictures (e.g., tolerance of missing data, assumption of parallel trajectories through time, ability to extrapolate, etc.), and each having advantages and disadvantages in ecological applications.

Synthetic control in ecology

The few previous uses of synthetic control in environmental contexts have predominantly focused on determining the effectiveness of policies or events on forest dynamics and socioeconomic outcomes (Sills et al. 2015, Jones 2018, Rana and Sills 2018, Rana and Miller 2019, Roopsind et al. 2019). However, we propose that this technique may be useful more broadly in ecology, particularly in cases where the units of analysis are large, complex, and lack replication or pre-meditated and well-matched controls. In this study, we examine the utility of synthetic controls for analyzing a hypothetical disturbance with time series of remote sensing imagery, i.e., data that are temporally and spatially extensive but also noisy and prone to confounding. Typical approaches for inferring effects from remote sensing data generally (1) use only the time series of treated pixels and thus ignore potentially useful contextual information from unaffected areas (Fiorella and Ripple 1993, Copeland et al. 2018, Monroe et al. 2020), or (2) use differencing techniques (e.g., Difference in Differences, hereafter DiD; Ashenfelter and Card 1985, Abadie 2005, Craig et al. 2017), which may require ad hoc decisions about controls and over-simplify the contextual information (Fig. 1; Malmstrom et al. 2009, Waller et al. 2018). For instance, imperfect matching between controls and treatment areas may produce bias if the controls respond differently to the same confounding factor, such as divergent responses to the same climate forcing among communities (Winkler et al. 2019). Reducing the need for exact matching between treatments and controls has been proposed to be a major advantage of the synthetic control approach (Craig et al. 2017).

Here we briefly review synthetic control methodology with attention to key assumptions, advantages, and best practices relevant for analyzing ecological data. We then evaluate the performance of several methods for assessing landscape-scale treatment effects (synthetic control, DiD, time series-only) using a simulated satellite time series of a spectral index (NDVI). We include various sources of random and systematic confounding noise and examine how the signal-to-noise ratio, available number of

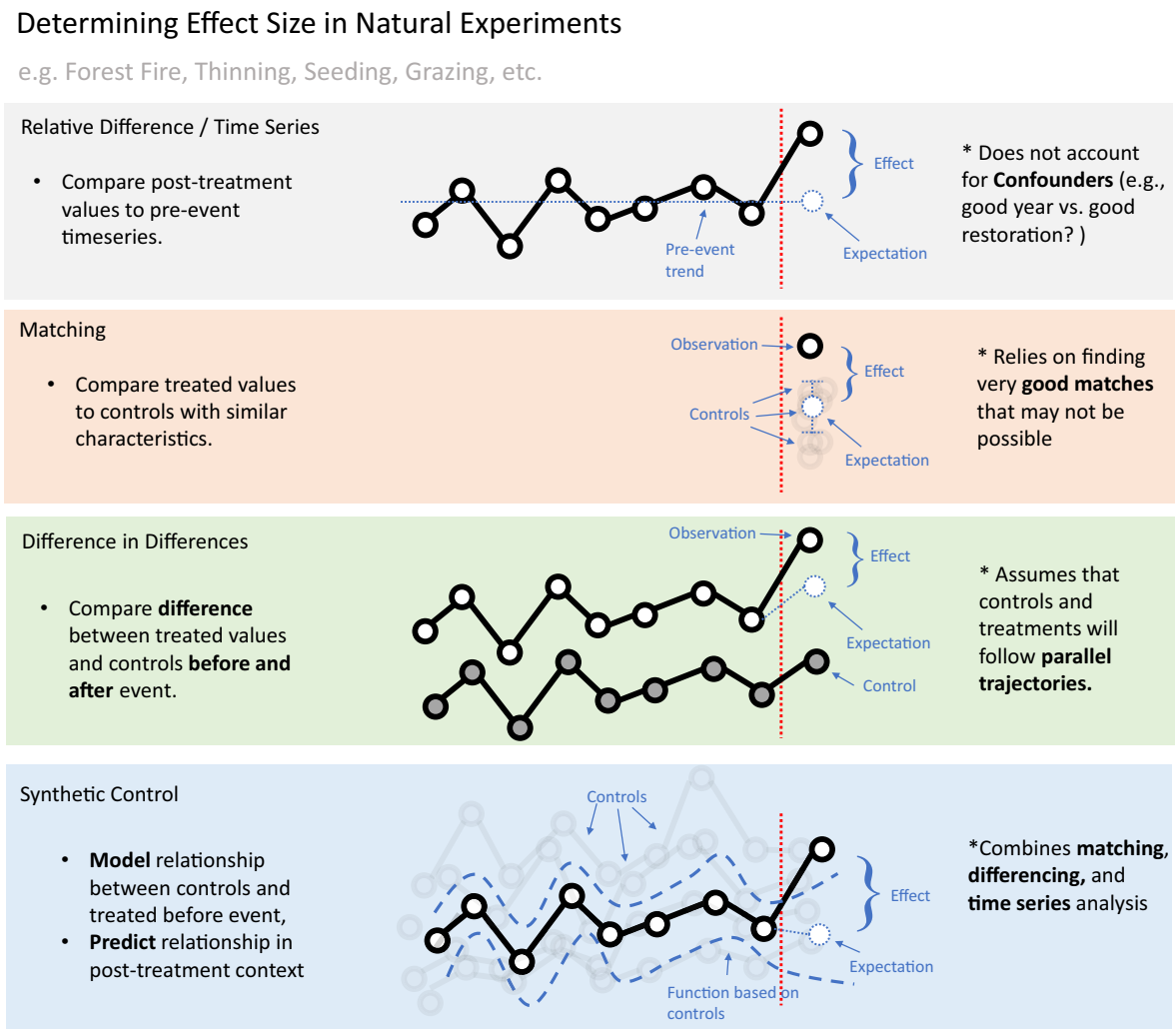

FIG. 1. General strategies for inferring treatment effects with observational data lacking randomized controls.

reference pixels, and ecological mismatch between reference and treatment pixels influence the ability of each method to identify a simple treatment effect representing vegetative disturbance followed by recovery. We hypothesized that synthetic controls would more accurately detect "true" treatment responses in the face of confounding random noise, and imperfect matching between controls and treatment, but that these effects would be contingent on the number and quality of controls available. We then demonstrate the use of synthetic control and other methods using a case study involving a brush-clearing treatment in southeastern Utah.

METHODS

Synthetic control and other causal inference methods

Synthetic control comes from a family of "quasi-experimental" analysis methods originating in the social sciences (Campbell and Stanley 1963). These approaches tackle the fundamental problem of causal inference – that one cannot observe both the treated and untreated outcome for the same individual (Holland 1986) – while using purely observational data without randomized controls. In the social sciences, this challenge has been formalized in the "counterfactual model" or "potential outcomes framework," a set of definitions and assumptions built around making valid causal inference (Rubin 1974, Morgan and Winship 2015). While all quasi-experimental methods in this framework share some key assumptions, such as the lack of contagion in treatment effects between treated and control units (i.e., Rubin 1974) approaches tend to be differentiated based on the types of assumptions made about relationships in the data – such as whether controls and treated units would respond similarly to the same treatment based on the fact they are similar in other, observable ways (i.e., "selection on the observables"); or whether controls and treated

units would be expected to follow similar paths in the absence of a treatment (i.e., "parallel paths"). Following these assumptions, analysis focuses around generating appropriate comparisons for the treated units (Fig. 1).

Time series methods.—One intuitive, but basic way to generate a comparison for a treated observation is to use its pre-treatment state or trajectory as a contrast, particularly if outcomes are in a time series and there are no suitable ways to select controls. For comparing pre- and post-treatment trends, a large number of time series analysis methods are available, but one particularly straightforward implementation is Interrupted Time Series (ITS) analysis (Bhaskaran et al. 2013). The most basic ITS model is a segmented regression that includes a term for pre-treatment trend, post-treatment trend, and a step-adjustment for the period immediately following treatment

$$Y_t = B_0 + B_1 D_t + B_2 T + B_3 T D_t + E_t$$

where B_0 is the intercept, B_1 is the additive effect of treatment, D_t is a binary indicator of whether treatment has occured at time t, B_2 is the effect of time T, B_3 is the effect of treatment on slope, and E_t is random error. Other covariates may be needed to account for seasonality or nonlinearity (Kontopantelis et al. 2015), as well as any serial autocorrelation in residuals that violates modeling assumptions and interpretation of standard errors (Bernal et al. 2017). Regardless of the functional form of this time-series-only modeling approach, strong assumptions are necessary, including that trends would remain the same in the post-treatment period as in the pre-treatment period, and that any difference in trend is related to the treatment. If these assumptions are unrealistic (as often the case in ecological time series for example), controls are needed to parse treatment effects from confounding.

Matching.—Including information from controls is an obvious way to remove confounding in post-treatment outcomes, but naive comparisons between treated and untreated populations can be misleading if systematic differences exist between control and treatment groups either in likelihood of selection for treatment, or responsiveness to treatment. For instance, assessing the impacts of wildfire across systems may be complicated by the fact that some communities are more prone to fire or burn more intensely than others. Such treatment "selection bias" is a major challenge for natural experiments, and much work has been done in the social sciences developing methods to explicitly account for this bias via the process of "matching" treated and untreated units based on their observed characteristics ("conditioning on the observables"; Angrist and Pischke 2008). A major assumption is that if underlying differences in treatment assignment or responsiveness are associated with observable characteristics (e.g., flammability of biomass), comparisons between treatments and controls are valid if adjusted or matched by these properties (the Conditional Independence Assumption; Morgan and Winship 2015).

A wide variety of matching approaches have been developed (Abadie and Imbens 2006), but one very common approach in the social sciences is the use of propensity scores, or predicted probabilities of treatment assignment based on a logistic multiple regression on observed covariates (Athey and Imbens 2017). Propensity scores are useful because they reduce the dimensionality of comparisons between controls and treatments to a single axis, and serve as a basis for weighting or stratifying observations. For land treatment events where the pattern of treatment assignment is known, propensity scores or other matching methods may be more applicable as a sorting method to align treated and untreated observations with similar characteristics. In cases with low or no replication in treated units, multivariate distance metrics may be more appropriate.

Difference in differences.—Matching techniques typically evaluate post-treatment outcomes only and make strong assumptions about the ability of observed characteristics to make good matches (Fig. 1). An alternative is difference in differences (DiD), in which the average difference between control and treatment are compared before and after an intervention (Ashenfelter and Card 1985, Abadie 2005, Craig et al. 2017). DiD does not rely on observable characteristics to guide comparisons, and thus may be useful for cases where such data are unavailable or unreliably associated with selection bias, provided both pre- and post-intervention outcome data are available. DiD has been widely used in the social sciences since the 1990s, starting with an influential analysis of the effects of Cuban migration to Miami on the labor markets resulting from the Mariel boat lift (Card 1990). A basic implementation of DiD in a "two-factor" regression context follows the form

$$Y_{it} = C_i + A_t + B \times D_{it} + E_{it}$$

where Y_{it} is the response for individual i at time t, C_i is an individual-level effect, A_t is time-point effect, D_{it} is a binary indicator of whether individual I has been treated at time t, E_{it} is the random error for individual i at time t and B is the overall treatment effect. One strong assumption of DiD is that both treated and untreated units follow the same trajectory in the absence of treatment, known as the "parallel trends" assumption. A key advantage of this method is that if the parallel trends assumption is met, the full set of explanatory covariates are not needed, but rather biases between control and treatment groups are accounted for by comparing differences across time points (also known as "selection on unobservables"; Morgan and Winship 2015).

Synthetic control.—In DiD, treatment outcomes are compared to a single control outcome or average of

outcomes. Synthetic control extends DiD by including an automated matching step to weight controls and generate a composite counterfactual for treated units over time for comparison. The method was developed for analyzing the effects of economic and political events on aggregated geopolitical entities (states, regions, countries), where there was need for (1) ways to estimate treatment effects for individual units, since many economic statistics of interest are only reported at the aggregate level and (2) a transparent, data-driven procedure for selecting comparison groups from a potentially large pool of candidates (Abadie et al. 2010). Since its initial formulation, this method has been used widely in the social sciences largely in policy evaluation contexts (see Sills et al. 2015) but also in economics, for instance in a reanalysis of the Mariel Boatlift data (Peri and Yasenov 2019). Much of the popularity of the method is attributed to its intuitive improvement over methods that generate a counterfactual using a coarse average or single control unit (Athey and Imbens 2017).

The original formulation of synthetic control proposed by Abadie and Gardeazabal (2003) and Abadie et al. (2010) generates counterfactual estimates based on a weighted sum of control values, the weights being selected based on an optimization scheme for data in the pre-treatment period. By using non-negative weights between 0 and 1, this method constrains the counterfactual predictions to the range of control unit values, which conservatively prevents extrapolation but may generate inferior estimates if controls differ significantly in magnitude from the treatment (Athey and Imbens 2017). One criticism of the method is that there are no analytical uncertainty metrics for counterfactual estimates; however, the authors propose a robustness check by iteratively removing weighted controls and refitting the model to examine effects on counterfactuals (Abadie et al. 2015). Furthermore, software implementations of this method have been critiqued for providing inconsistent weight selection (by cross-validation; Klößner et al. 2018) and inducing problems when all pre-treatment outcome data are included as criteria for selecting weights (Kaul et al. 2015). The original weighting method also assumes that weights remain constant in the post-treatment phase.

While most references to synthetic control in the literature refer explicitly to the weighting method developed by Abadie et al. (2010), other methods apply a similar approach of building data-driven composite counterfactuals, albeit with different modeling frameworks. One such implementation of synthetic control designed specifically to relax the assumption of parallel trends between control and treatment groups through time is the "generalized" synthetic control method developed by Xu (2017). This method generates counterfactuals by first estimating a set of time-varying latent factors (essentially unobserved confounders) for which each unit has a specific affinity, or "loading," using data from the control population (Bai 2009). The loadings for treated units are then estimated and used to predict outcomes in the post-treatment period. The general functional form of generalized synthetic control is

$$Y_{it} = \delta_{it} D_{it} + x_{it}\beta + \lambda_i f_t + \varepsilon_{it}$$

where Y_{it} is the observation i at time t, D_{it} is a binary indicator of treatment and δ_{it} is the treatment effect, x_{it} are covariates and β their corresponding coefficients, λ_i are the loadings for unit i on factors f_t, and ϵ_{it} is Gaussian random variation. The factors f are estimated using control data, while λ_i are fit using a least-squares procedure (Xu 2017). An important assumption of this method is that the latent factors apply to all units (though heterogeneously via different loadings), and that there is sufficient overlap or support between loadings for treated units and controls, violations of which could lead to unreasonable extrapolations. The number of latent factors is a key parameter determining model flexibility and is estimated via cross-validation in the software implementation of the method gsynth in R. While the generalized synthetic control tends to require more pretreatment data than traditional synthetic control and has stricter assumptions, the method has the added advantages of estimating treatment effects for many distinct units simultaneously, allowing for missing data and asynchronous treatment implementation dates, and a bootstrapped semi-parametric standard error estimation (Xu 2017).

Another implementation of synthetic control uses a Bayesian structural time series framework to model and predict the temporal evolution of a counterfactual observation including information from controls as well as temporal dynamics of the treated unit (Brodersen et al. 2015). This implementation uses a state-space or hidden Markov model framework in which the data generating process is divided into a "state" equation that represents the temporal evolution of a latent process and an "observation" equation that relates the how the state is realized by observed data. The state equation integrates several sub-models, including a local linear trend, seasonality, and regression component using values of controls as predictors and a spike-and-slab prior for variable selection (Brodersen et al. 2015). The observation equation follows the general form

$$Y_t = \alpha_t + x_t\beta + \varepsilon_t$$

and the state equation

$$\alpha_{t+1} = \mu_{t+1} + \gamma_{t+1} + \eta_t$$

where Y_t is the treated unit response at time t, α is the state component, x is a vector of control values, β are regression coefficients, ϵ is random error, μ_{t+1} is a local linear trend (composed itself of both immediate and long-term processes), γ is the seasonality component, and η_t is white noise. The spike-and-slab prior is an

effective method for variable selection in a Bayesian context whereby coefficients are either included or excluded based on a probability mass centered on "0" (the "spike"), and then regularized by a Gaussian-shaped prior if included in the model (the "slab"; Ishwaran and Rao 2005). In this case, the spike-and-slab prior is used to identify which controls to include and their weights, which helps avoid overfitting (Brodersen et al. 2015). In the software implementation of this method, only equally spaced, time-varying covariates (i.e., controls) are accepted, with no missing data (although treated values may be missing).

Simulation modeling

We examined several quasi-experimental approaches for estimating landscape-scale events (disturbances or management activities with a distinct time of initiation) using simulated remote-sensing data (Table 1): (1) Interrupted Time Series (ITS), which does not consider controls (Bhaskaran et al. 2013); (2) traditional "Difference in Difference" (DiD), where pre-treatment and post-treatment differences between control and treated pixels are compared using a linear two-way factor model (Ashenfelter and Card 1985, Larsen et al. 2019); and (3) Synthetic Control, in which treatment effects are estimated against an expectation based on the pre-treatment relationship between control pixels and treated pixels. We implemented two formulations of synthetic control: (1) a linear interactive fixed effects model with latent confounders using the R package gsynth (Xu 2017) and (2) a Bayesian structural time series model using the R package CausalImpact (Brodersen et al. 2015).

Although DiD and synthetic control are similar, they are often considered separately in the literature, and we hereafter consider DiD distinct from synthetic control methods. We used default values for all functions, and simulations and analyses were implemented in R (R Core Team 2019). It is important to note that the time-series-only method used here, ITS, can be heavily customized in application settings, and we use a very simplified version as a coarse baseline for estimating trends without considering controls.

We generated simulated 16-d NDVI time-series data following the approach of Verbesselt et al. (2010) by additively combining an NDVI signal from a hypothetical treatment with various sources of noise (Fig. 1). Pixels were modeled either as grassland or forest pixel types, with a corresponding seasonal sine-wave trends with amplitudes of 0.4 and 0.1, respectively, and baseline NDVI values of 0.6 or 0.8 (Verbesselt et al. 2010). The treatment effect was modeled as an abrupt reduction in NDVI (−0.1) such as from a large disturbance (e.g., fire or clearing), followed by a linear recovery over 4 yr (Fig. 2, "Treatment" panel). To estimate false-positive error rates, we also included a placebo treatment, with a treatment effect of 0 after the treatment date. Following Verbesselt et al (2010), we added random Gaussian noise, systematically controlling the variance of this noise among simulations (SD = 0.1, 0.2, . . ., 0.7; Fig. 2, "Noise" panel).

Table 1. Summary of causal inference methods used in simulations.

Method	Approach	Method	Key Assumptions	Citation
Interrupted Time Series (ITS)	trend of treated compared before and after treatment	Trend component estimated with segmented linear regression: $Y_t = B_0 + B_1 D_t + B_2 T + B_3 T D_t + \gamma_1 T + \gamma_2 T$ where B_0 is the intercept, B_1 is the additive effect of treatment, D_t is a binary indicator of treatment, B_2 is the effect of time T, B_3 is the effect of treatment on slope, and γ_1 and γ_2 are sine and cosine of $2\pi \times$ Time divided by a period of 365 d (Fourier terms for seasonality).	Post-treatment trajectory is exclusively related to treatment, after accounting for seasonality and other variables.	Bernal et al. (2017)
Difference in Difference	pre-treatment differences between control and treated compared to post-treatment differences.	Applied treatment effect estimated by subtracting individual and time-period effects in a linear "two-way fixed effects" model: $Y_{it} = C_i + A_t + B \times D_{it} + E_{it}$ where Y_{it} is the response for pixel i at time t, C represents individual effects, A represent time effects, and B is the treatment effect with D_{it} a 0/1 dummy variable indicating treatment and error as E_{it}	Parallel trends: that controls and treated would follow same trend, in the absence of treatment.	Ashenfelter and Card (1985)
Synthetic Control	treatment values compared to prediction from functional relation between control and treatment, before exposure	Interactive factor model with latent variables selected by cross-validation.	Latent time-varying factors influence both treated and control units.	Xu (2017)
		Bayesian structural time series model with local linear trend, seasonality and linear regression components in the "process" part of the model.	Parallel trends	Brodersen et al. (2015)

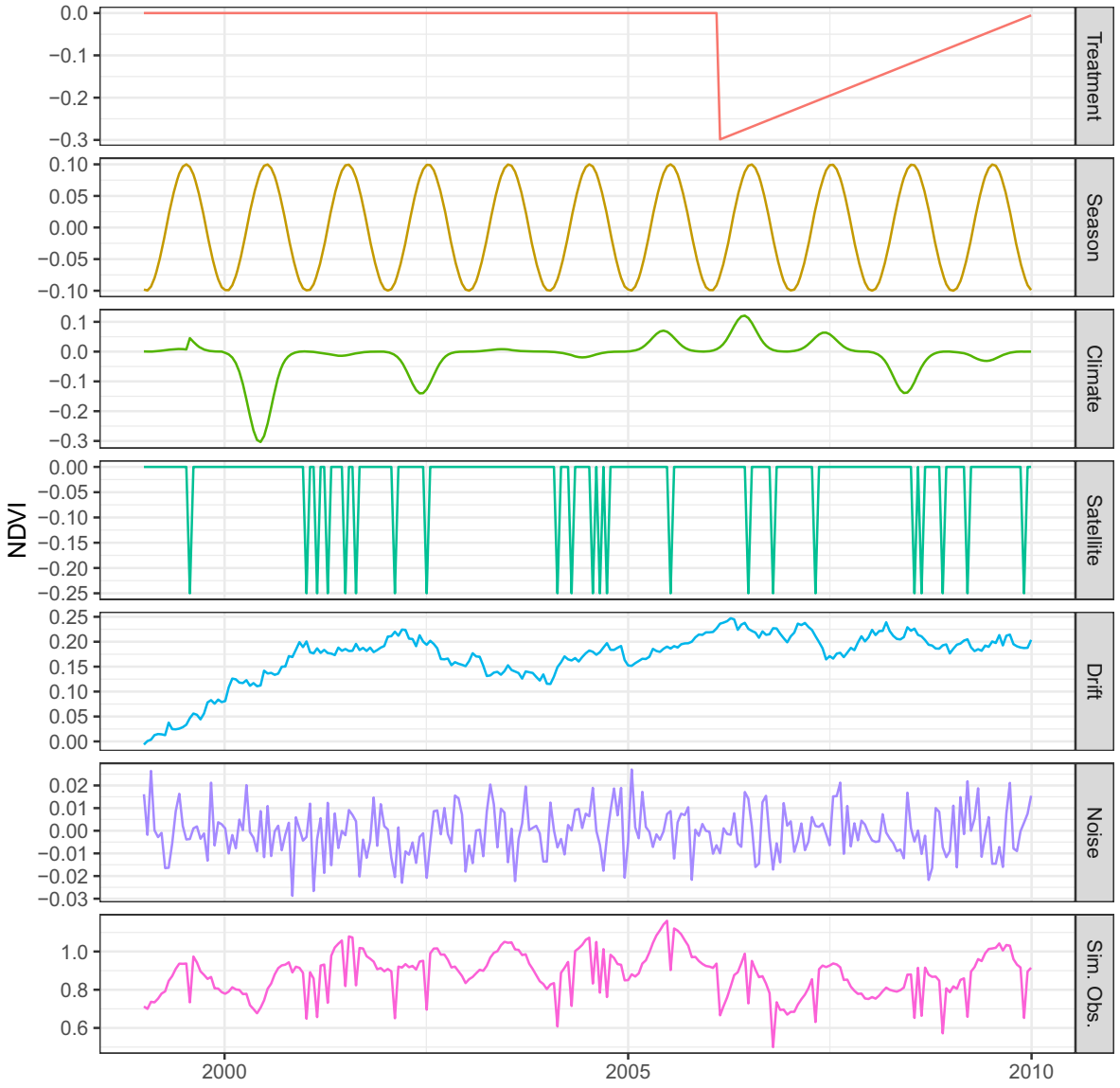

FIG. 2. Example of a simulated NDVI time series for a forest (Sim. Obs.) composed by adding various trends and sources of random noise. "Treatment" represents the hypothetical "true" disturbance and recovery trajectory added to the simulated remote-sensing time series and estimated by various methods.

Since we were interested in assessing treatment effects in the presence of a variety of potential confounding factors, we added three additional sources of systematic noise to simulated time series: (1) random drops of 0.25 NDVI, corresponding to cloud contamination or sensor error in a satellite image (Fig. 2, panel "Satellite"); (2) a growing-season climate anomaly resulting in increased or decreased production (e.g., Wang et al. 2003; Fig. 2, panel "Climate"); and (3) signal drift over time as from vegetative dynamics (e.g., Anyamba and Tucker 2005; Fig. 2, panel "Drift"). The probability of a satellite/cloud error was set at 5%. The climate anomaly was added as a symmetric Gaussian function centered around 20 April, with the magnitude drawn from a Gaussian distribution (SD = 0.1). We introduced a small amount of serial correlation in climate anomalies to account for multi-year climate trends using a low-pass filter (R function "filter" with 1 lagged forecast error). Vegetation drift was simulated by a random Gaussian walk with a standard deviation of 0.05.

For each simulation, we also generated a set of "control" pixels, which did not include the treatment effect. We set the number of control pixels in a simulation to either 1, 5, 10, 50, or 100 to observe how the number of controls would affect the accuracy of different methods. These pixels received the same set of confounders

(climatic, satellite, and drift) but separate realizations of random noise.

Different parts of a landscape are likely to have heterogenous responses to a similar exogenous influence (e.g., climate). To account for differing sensitivities to confounding factors among pixels, the signals for confounding variables were multiplied by a pixel-specific coefficient before being added to the overall NDVI response. Coefficients were drawn from a Gaussian distribution with a mean of 1 and standard deviation of 0.25. Since sensitivity to confounders might also vary through time, confounders were multiplied by a similar coefficient with a random Gaussian coefficient (1 + SD = 0.05) for each pixel at each time point.

The accuracy of synthetic control and other differencing methods is likely to depend on the degree of underlying similarity between a treated unit and its controls. In an application setting, a large number potential controls might need to be screened using matching techniques to find control pixels with similar properties (Abadie et al. 2010) but perfect matching is unlikely. To assess the effects of potential mismatch between control and treated pixels during an initial matching step on the accuracy of different methods, we generated three different scenarios (Fig. 3): (1) All control pixels are of the same landscape type (forest or grassland) as the treated pixel (mismatch = 0); (2) 50% of the control pixels are of a *different* landscape type (mismatch = 0.5), or (3) all of the control pixels are of a *different* landscape type (mismatch = 1).

Finally, a unique aspect of the remote sensing data for causal analysis is the potential for spatial dependence among control pixels. To explore how high levels of spatial autocorrelation might influence estimates, we induced a pseudo-spatial correlation structure in control values for a subset of simulations (all permutations with noise level fixed at 0.01). For each simulation, a 20 × 20 pixel Gaussian random field was generated using an exponential distance decay function with a sill of 0.25, and a range of 10, using the R package gstat (Gräler et al. 2016). Controls were then randomly assigned to pixels from this simulated "image," and the degree of autocorrelation was assessed using Moran's *I*. Random field values were then multiplied with either net confounders (climate + satellite + vegetation dynamics), random noise, or the raw response profile of the controls

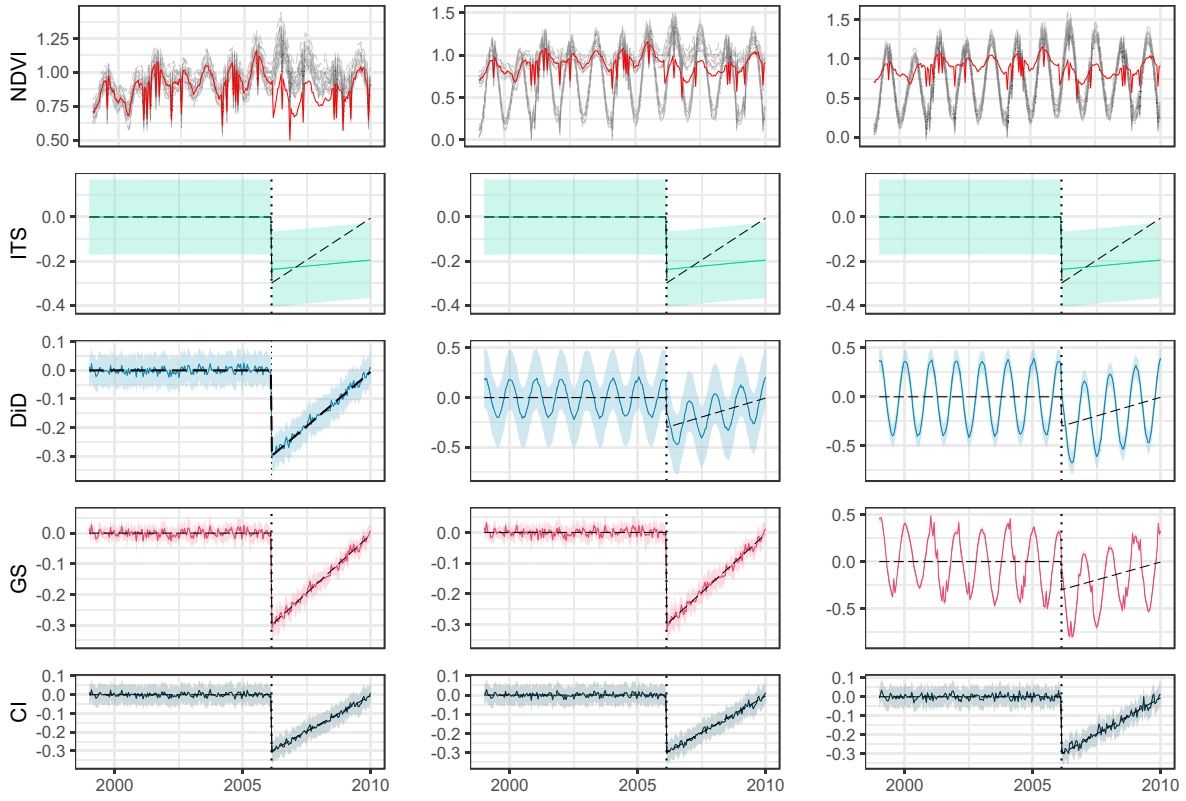

FIG. 3. Example treatment effect estimates for different methods when controls are well matched (left column, mismatch = 0), equally well and poorly matched (center column, mismatch = 0.5), or poorly matched (right column, mismatch = 1). Top row shows the same simulated NDVI signal (red) with differing control pixels (gray). Bottom rows show estimated treatment effect (solid line), actual treatment effect (dashed line), and confidence/credible intervals (shading) for different methods. Methods included Causal impact (CI), gsynth (GS), Difference in Differences (DD), and Interrupted Time Series (ITS). The treatment occurs in February 2006 and is indicated by a vertical dotted line. Treatment effect in simulations was 0.1 NDVI, but exaggerated here (0.3 NDVI) to improve contrast.

to induce pseudo-spatial structure. For each simulation, the correlation structure remained static through time.

For each combination of conditions (landscape type, control mismatch, number of controls, random noise level) we generated 1,000 simulated time series and obtained treatment effect estimates for all methods (Table 1). We assessed errors as the difference between the "true" simulated treatment effect ("Treatment" in Fig. 2) and treatment effects estimated by various methods, at each point in the post-treatment time period for each simulation. For methods that provided confidence or credible intervals, we also assessed sensitivity by counting whether the upper bounds of the estimated treatment effect intervals excluded zero at each point in the post-treatment time series. Simulation code is hosted on Zenodo; see Data Availability.

Case study

We briefly demonstrate use-case of synthetic control for inferring management intervention effects without a priori controls in the context of a brush-clearing treatment that occurred in southeastern Utah, USA in 2009. The Shay Mesa Restoration Project was designed to reduce fuel loads and improve wildlife habitat by removing pinion (*Pinus edulis*) and juniper (*Juniperus osteosperma*) trees over a 750-ha treatment area, using various brush-clearing methods (details in Karl et al. 2014 and Gillan et al. 2016). We assessed treatment effects based on a bimonthly time series of Landsat-based 30-m resolution Soil Adjusted Total Vegetative Index (SATVI; Marsett et al. 2006), comparing pixels in the treated areas to their synthetic controls from pixels in surrounding areas. For each treated pixel, we narrowed candidate pool of control pixels with a matching algorithm developed by Nauman and Duniway (2016). Further details available in Appendix S1: Section S1.

Results

Simulations

In simulations, absolute point-wise errors for the different methods of determining treatment effects (time-series-only, DiD, synthetic control) were largely contingent on both data availability (i.e., the number of controls available) and data quality (the degree of mismatch between controls and treatments). When controls were well matched with the treatment pixel, all methods that included controls were superior to the baseline estimates from the time-series-only method (ITS), regardless of the number of controls available (Fig. 4, top row).

As more mismatched pixels were introduced to the control population, accuracy depended more on the number of controls available, with a larger number of controls generally improving estimates for the synthetic control methods (Fig. 4, middle row). The CausalImpact synthetic control method needed only 5 controls to achieve estimates superior to ITS, while gsynth required between 5 and 50. Unlike the synthetic control methods, DiD was generally less accurate than the time-series-only method, likely stemming from its naïve aggregation of all controls, resulting in bias.

When all control pixels were poorly matched to the treated pixel, only the CausalImpact method outperformed the baseline time series-only method, and only with many controls (Fig. 4, bottom row). Poorly matched controls resulted in both DiD and gsynth methods being less accurate than baseline, and the DiD method performed worse with larger numbers of poorly matched controls, again due to the naïve aggregation of controls for comparison.

In most cases, increases in signal-to-noise ratio (effect size/SD of random noise) led to marginal reductions in error (Fig. 4), particularly after signal magnitude reached 10–50% of the average variation in the random noise component. The absolute magnitude of the combined confounder signal also contributed to error, but only when imperfect matches between controls and treated pixels were present.

Confidence envelopes for treatment effects revealed differences between methods, which varied by level of noise and control mismatch (Fig. 5). The CausalImpact method was the most conservative (low sensitivity), especially when the signal-to-noise ratio was low (Fig. 5). Even when the signal-to-noise ratio was high, only ~25% of the true effects were determined to be significantly different from zero. Both DiD and gsynth tended to have smaller confidence envelopes, which resulted in more frequent "significant" treatment effects but also erroneously significant effects for the placebo (Fig. 5). Gsynth's sensitivity, but also its false-positive rate tended to increase with more controls (Fig. 5 top row). Confidence envelopes for both DiD and gsynth did not appear to reflect greater uncertainty when control populations were completely mismatched to treatment (Fig. 5 bottom row), remaining relatively narrow. This may have been driven by the inability of either method to account for the differing seasonal signal in the control populations (e.g., Fig. 3 right column).

Error for control-based methods generally increased with confounding, and this effect was heightened in cases where controls were mismatched (Fig. 6). Average errors tended to be negative, meaning on average, predicted effect sizes tended to be greater than true effects (Fig. 6). Bias also tended to increase with mismatch (Fig. 6), but mostly for DiD and gsynth methods. The CausalImpact synthetic control method tended to have lower bias across simulations.

Simulated autocorrelation tended to have negligible effects on bias and error for the methods that used controls, except for when autocorrelation was applied to each control pixel's response trajectory (Appendix S1: Fig. S1, S2). In this case, predictions tended to improve with greater levels of autocorrelation, particularly for the gsynth method in conditions of extremely poor

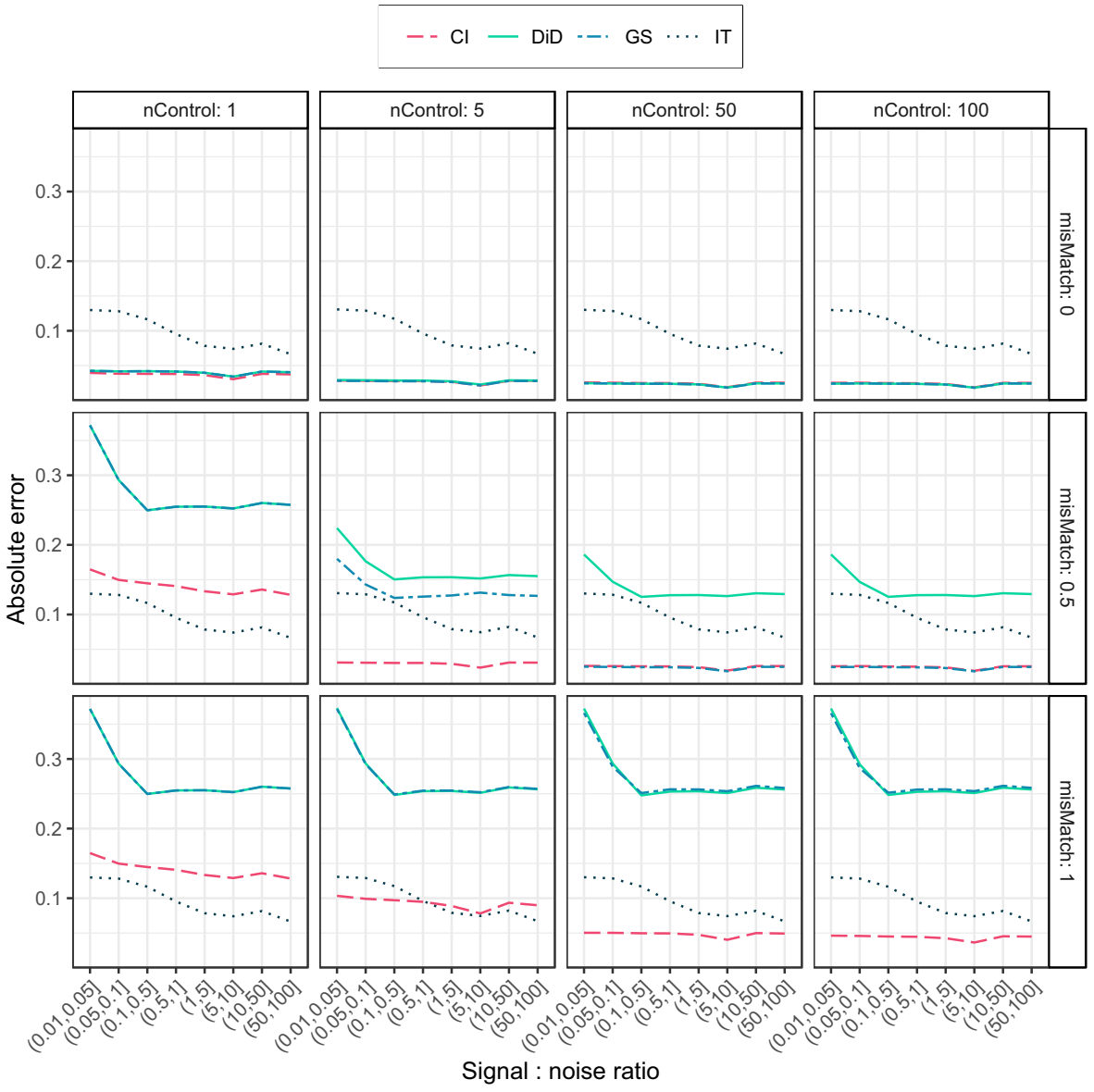

FIG. 4. Simulation results for average absolute error in estimated treatment effect (in NDVI) as a function of signal-to-noise ratio. Signal-to-noise ratio (x-axis of all panels) was calculated by dividing the magnitude of a simulated treatment effect at a given point in time by the prescribed standard deviation in random noise for that simulation. Results are broken down by number of controls available (nControl; columns) and degree of mismatch between the control population and treated pixels (top row, no mismatch; middle row, equally well and poorly matched; bottom row, total mismatch). Methods included CausalImpact (CI), gsynth (GS), Difference in Differences (DiD), and Interrupted Time Series (ITS).

matching (mismatch = 1). This may be due to effective "shrinkage" or reduction in noise induced by the autocorrelation procedure.

Case study

An example implementation of synthetic control for determining different brush-clearing treatment effects using CausalImpact (Brodersen et al. 2015) is presented in Fig. 7. There was clear differentiation among treatment areas in terms of cumulative effect size derived from synthetic control (Fig. 7F) but point-wise estimates were significantly noisier (Fig. 7E).

DISCUSSION

Controls are important

On a basic level, our study highlights the value of using controls when estimating the effects of large-scale ecological interventions, particularly with noisy data from satellites. In the simulations, methods that

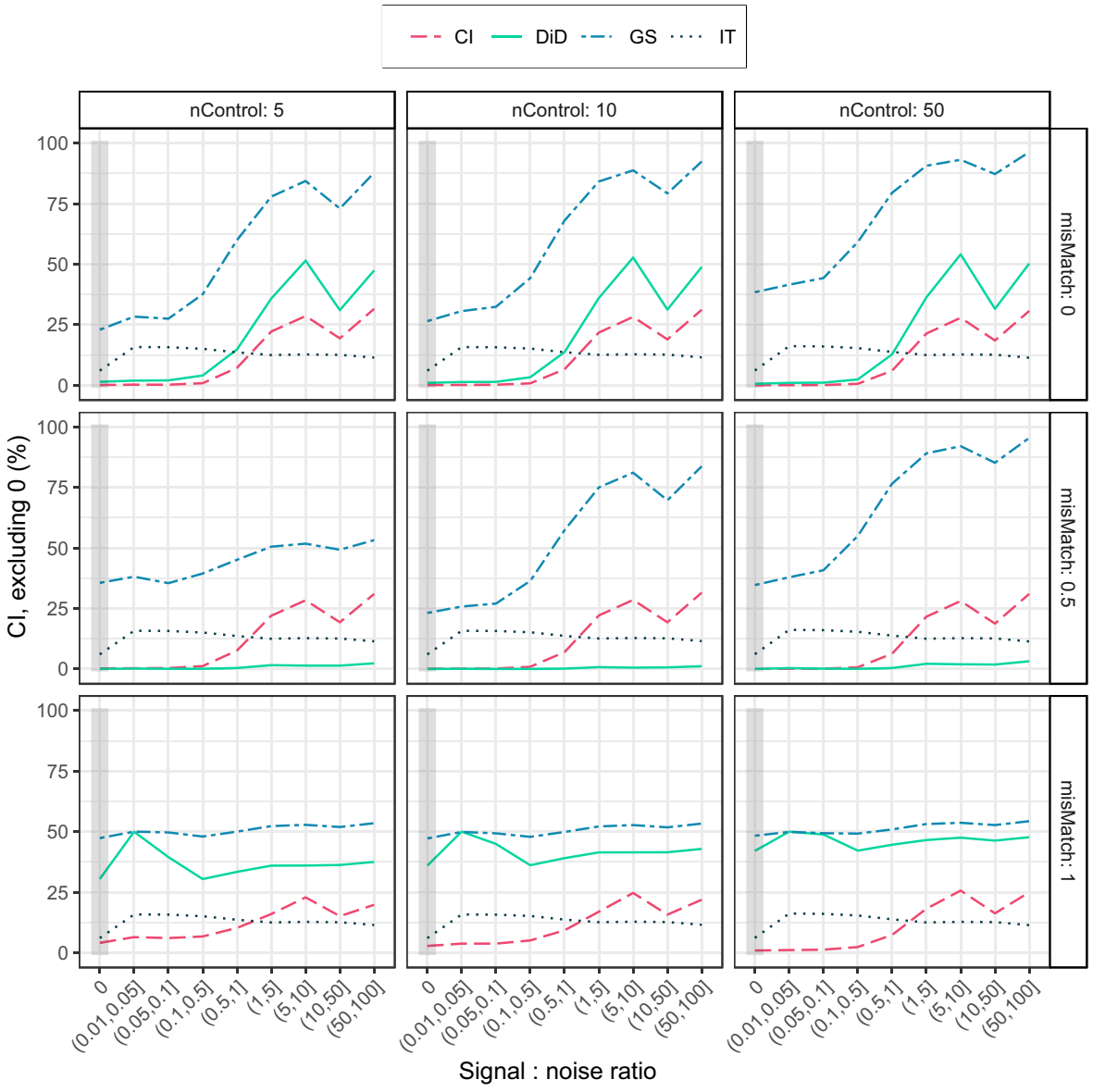

FIG. 5. Power curves + placebo results showing the percentage of time each method considered a treatment effect significant (confidence/credible interval upper bound excluding zero). Rows represent level of mismatch between treated and reference pixels, columns represent differing numbers of controls, and the point-wise absolute magnitude of net confounder effect (season + climate + drift + satellite) is represented on the x-axis of each panel. Placebo results, where no treatment was included after the hypothetical treatment date highlighted with shaded regions. Methods included CausalImpact (CI), gsynth (GS), Difference in Differences (DiD), and Interrupted Time Series (ITS).

incorporated data from some kind of properly matched, untreated group were more accurate at estimating "true" treatment effects than methods that relied on time series alone (Fig. 4). For data with many potential confounding variables, such as remote sensing time series, controls provide an intuitive baseline to remove these effects. In the simulations, relatively large (but not unreasonably so; Verbesselt et al 2010) confounders were intentionally included as proof of concept. In actual remotely sensed data, the strength of confounding will likely depend on ecological context, with more heterogeneous landscapes subject to greater confounding (Reed et al. 1994).

Matching is important

While post hoc controls were useful for estimating treatment effects, simulations showed that improperly matched controls could be counter-productive, depending on the availability of data and the method used to infer effects. The CausalImpact method was able to

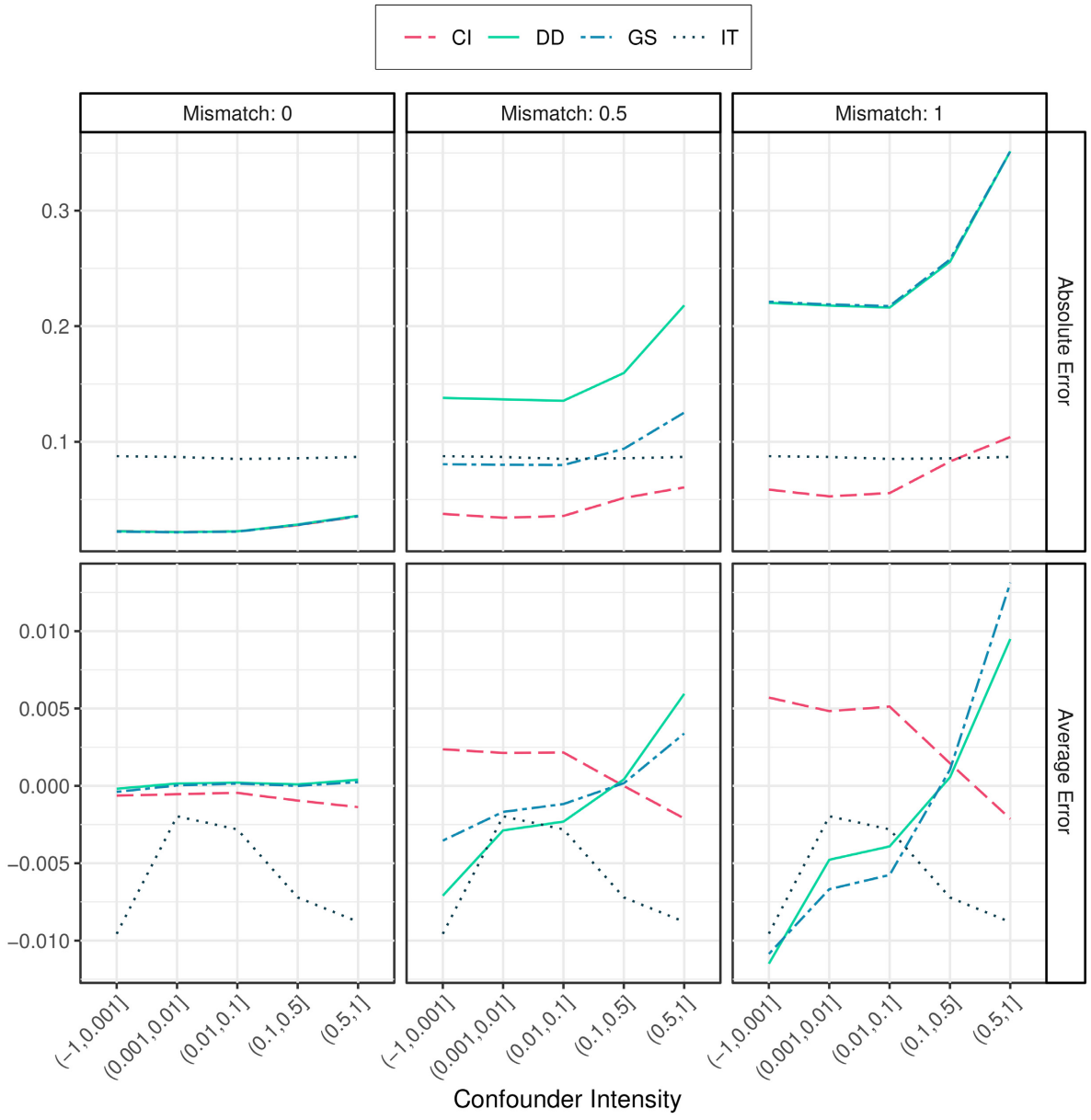

FIG. 6. Absolute error and bias (average predicted treatment effect minus true treatment effect) as a function of point-wise confounder magnitude (net displacement from satellite, climate, and drift) and control mismatch (misMatch), by method. Methods included CausalImpact (CI), gsynth (GS), Difference in Differences (DiD), and Interrupted Time Series (ITS).

accurately estimate the treatment effect given enough control data in simulations, likely in part because it explicitly includes a seasonality component in its model and it implements an efficient control selection method in the spike-and-slab prior (Brodersen et al. 2015; see Fig. 2). It is unclear the degree to which such inference could be achieved with non-simulated, poorly matched data. In the simulation, poorly matched controls were designed to respond to the same confounders (i.e., seasonality, clouds, trends) as the treated pixel, only at a different magnitude. This might not be the case with real data where a mismatched land-cover type might have a qualitatively different response to a confounder compared to the treated pixel (e.g., an irrigated field vs. rainfed grasslands). Ultimately, some level of underlying correspondence must exist between control and treatment observations for control-based methods to be valid, highlighting the important role of finding accurate matches between control and treatment populations. The further development and implementation of reliable, automated techniques for finding spatial comparisons across ecological contexts is needed in an application context (Nauman and Duniway 2016). Nevertheless, our results suggest that under certain conditions synthetic

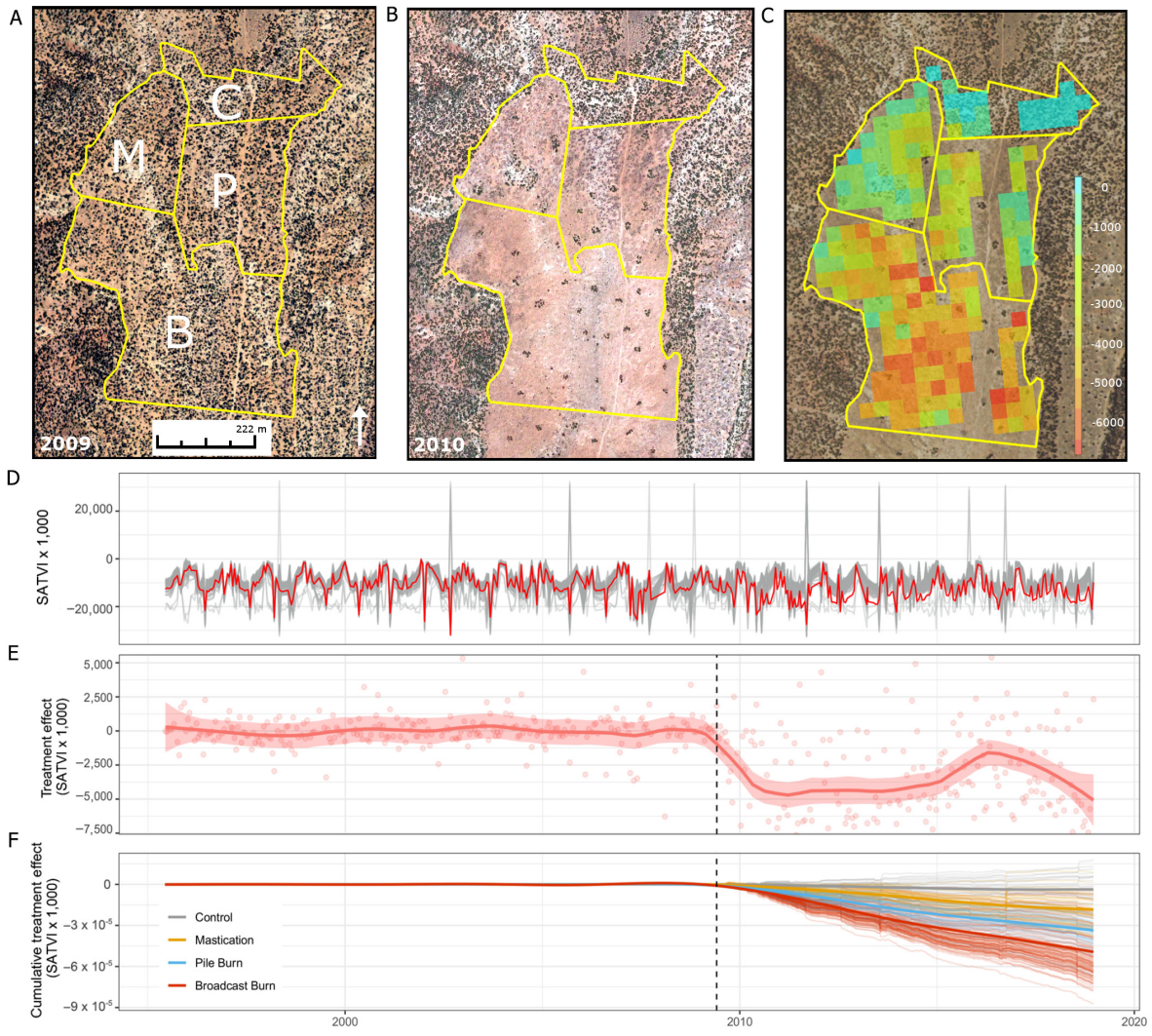

FIG. 7. Example implementation of synthetic control for a pinion–juniper clearing in southeast Utah (details of study in Appendix S1: Section S1). (A and B) Treated site before and after brush treatments (C, Control; P, Pile Burn; M, Mastication; B, Burn). (C) Estimated median per-pixel treatment effect heatmap using the CausalImpact method for the 2010 growing season (March–November) in terms of Soil Adjusted Total Vegetative Index (SATVI) × 1,000. (D) Raw SATVI time series for a treated pixel (red) and its controls (gray). (E) Point-wise estimated treatment effects and trend line using the Bayesian structural time series in "CausalImpact." (F) Pixel-wise cumulative treatment effects over time, analogous to exposure of bare ground integrated over time.

control can relax some of the need for direct correspondence between controls and treated units by flexibly "learning" the shared patterns in the data.

Method specific details

While the synthetic control approach may be useful for a wide variety of ecological data, specific implementations and models may have distinct advantages in different contexts. For instance, if treatments and controls are well matched and unlikely to violate the parallel trajectories assumption, simple DiD implementations may be sufficient, as they were in simulations with no treatment–control mismatch (Fig. 4, top row). Some variation of DiD is probably the most common approach for remote sensing applications seeking to infer landscape change currently (e.g., forest regeneration, grazing impacts, etc.). However, parallel trends assumptions are often violated in real data (Abadie 2005, Xu 2017), and more sophisticated models may be able to flexibly learn relationships before treatment happens and extend these relationships through time after the treatment happens.

In this study, we investigated two such synthetic control approaches (gsynth and CausalImpact), but in theory any function-fitting method may be used. While our study found the CausalImpact method to be generally most accurate at predicting "true" treatment effects across simulations conditions, advantages of the gsynth method include its ability to generate counterfactuals for multiple treated units simultaneously, and its robustness

to missing data. One consideration for both methods is selecting the degree of flexibility used in model fitting, which includes the number of potential latent variables (r) for gsynth and the inclusion of time-varying regression coefficients for CausalImpact. In both cases, high flexibility may lead to overfitting and biased predictions for counterfactuals (Brodersen et al. 2015, Xu 2017). Use of validation metrics can further help evaluate reliability of different model formulations (R^2 for the CausalImpact Bayesian Structural Time series model and mean squared prediction error for gsynth). Validating the accuracy of synthetic control predictions against data withheld from the pretreatment period is another recommended practice to sense-check model fit. It is important to note that we used default values for each method in the simulations to avoid bias, but in practice fine-tuning the settings of a selected method is recommended.

The levels of uncertainty reported by various methods may also be important to consider for use of synthetic controls in applications. CausalImpact generally had conservative estimates, with "significant" treatment effects (e.g., credible intervals not crossing zero) occurring at a maximum of roughly 50% of the point-wise instances with very high signal-to-noise ratios (Fig. 5). By contrast, the gsynth method typically had tighter confidence bands (e.g., Fig. 3) even in cases where predictions were obviously poor due to mismatched controls (Fig. 5). These overly narrow confidence bands in gysnth may be an artifact of violated assumptions of the parametric standard error estimates used in simulations (Xu 2017).

Notes for application

In application settings, there may be observed significant amounts of heterogeneity in estimated treatment effects, both within treated areas and through time (Fig. 7). Without accounting for such within-treatment heterogeneity, aggregations across space to estimate net effects may discount important variation in treatment response and potentially bias conclusions. Rather, this variation may be used to extend insight about fine-scale environmental controls on treatments or expected spatial variance in treatment efficiency by ecological context. Masking out unresponsive areas (e.g., rocky areas unlikely to change) or stratifying responses by environment may be necessary for modeling general responses. In the case study, we also used raw 16-d SATVI time series as our response variable of interest, without implementing any cloud masking. In aggregate, predicted effects showed clear trends but point-wise estimates remained noisy (Fig. 6 E). In practice, an additional step of cloud masking, aggregating to a broader temporal scale or implementing low-pass filtering on the time series may help improve results.

In this example, we also used only the pre-treatment SATVI control pixel time series for modeling the relationship between treatment and controls, but potentially any number of other predictors could also be included (but see Ferman et al. 2020), including other remotely sensed indices or climate data. Currently, the CausalImpact method only accepts time-varying covariates; however, it has been argued that control outcomes are more important than covariates for generating synthetic controls (Doudchenko and Imbens 2016, Athey and Imbens 2017).

One potentially useful feature of the synthetic control that we did not explore is the fact that control weights or coefficients are available to be examined and interpreted. For instance, the generalized synthetic control method (gsynth) provides estimates of time-varying latent factors, which may have real-world interpretations given contextual knowledge (Xu 2017). Abadie et al. (2015) generate a ranked list of countries that are given large weights in approximating synthetic west Germany and speculate about the underlying reasons for this. In an implementation of synthetic control for landscape treatment effects, visualization and interpretation of these weights may be helpful for understanding spatial relationships in data and generating hypothesis for drivers of ecological change.

Broader implications

In deciding how to manage ecological systems, one often looks to examples of similar sites or situations to gauge the range of expected behavior resulting from an action. The power of this inference typically depends on how well the comparison sites are representative of the location of interest, both in terms of ecological potential and ecological state at the time of intervention. The frequent need for these types of comparisons has led to many landscape classification systems (Salley et al. 2016) that parse regions by ecological potential (e.g., NRCS Ecological Site Descriptions; Duniway et al. 2010) and describe the range of ecological conditions expected given that potential (e.g., State and transition models; Bestelmeyer et al. 2004). The synthetic control approach has the potential to essentially systematize this search for suitable comparison sites by integrating information about ecological state (remotely sensed time series of vegetation) with ecological potential (soils and topography), especially if the pool of candidate controls initially screened based on environmental data (e.g., Nauman and Duniway 2016). Thus, this approach may be considered a quantitative and scalable framework for conducting a common activity that is often conducted on an ad hoc basis.

With the burgeoning availability of ecological data from remote-sensing imagery, sensor and monitoring networks, and crowd-sourced data, there is new opportunity for ecological insight but also a growing need for methods to make sense of large, noisy, observational data sets (e.g., Copeland et al. 2018). The synthetic control framework is particularly well suited for this kind of data in that it generates intuitive interpretations of treatment effects without relying on many of the formal strictures of experimental design. For instance, synthetic

control can provide a quantitative estimate for the response to a "no-action alternative," commonly included in environmental analysis (e.g., NEPA; Steinemann 2001). Furthermore, sophisticated versions of synthetic control methods can be easily implemented in open-source software environments, flexibly learn from multiple types of time series data, and provide robust estimates of uncertainty. In this study, we show how synthetic control can be used in the context of quantifying the effects of landscape-scale ecological events using remote sensing data. However, we believe that these techniques developed in the disciples of political science and econometrics can be helpful for a wide variety of questions and data sets in ecology.

Acknowledgments

This research was conducted with support from the U.S. National Resources Conservation Service and the US Geological Survey Ecosystems Mission Area. Any use of trade, product or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Literature Cited

Abadie, A. 2005. Semiparametric difference-in-differences estimators. Review of Economic Studies 72:1–19.

Abadie, A., A. Diamond, and J. Hainmueller. 2010. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. Journal of the American statistical Association 105:493–505.

Abadie, A., A. Diamond, and J. Hainmueller. 2015. Comparative politics and the synthetic control method. American Journal of Political Science 59:495–510.

Abadie, A., and J. Gardeazabal. 2003. The economic costs of conflict: A case study of the Basque Country. American Economic Review 93:113–132.

Abadie, A., and G. W. Imbens. 2006. Large sample properties of matching estimators for average treatment effects. Econometrica 74:235–267.

Angrist, J. D., and J.-S. Pischke. 2008. Mostly harmless econometrics: An empiricist's companion. Princeton University Press, Princeton, New Jersey, USA.

Anyamba, A., and C. J. Tucker. 2005. Analysis of Sahelian vegetation dynamics using NOAA-AVHRR NDVI data from 1981–2003. Journal of Arid Environments 63:596–614.

Ashenfelter, O. C., and D. Card. 1985. Using the longitudinal structure of earnings to estimate the effect of training programs. Review of Economics and Statistics 67:648–660.

Athey, S., and G. W. Imbens. 2017. The state of applied econometrics: Causality and policy evaluation. Journal of Economic Perspectives 31:3–32.

Bai, J. 2009. Panel data models with interactive fixed effects. Econometrica 77:1229–1279.

Bernal, J. L., S. Cummins, and A. Gasparrini. 2017. Interrupted time series regression for the evaluation of public health interventions: a tutorial. International Journal of Epidemiology 46:348–355.

Bestelmeyer, B. T., L. M. Burkett, L. Lister, J. R. Brown, and R. L. Schooley. 2019. Collaborative approaches to strengthen the role of science in rangeland conservation. Rangelands 41:218–226.

Bestelmeyer, B. T., J. E. Herrick, J. R. Brown, D. A. Trujillo, and K. M. Havstad. 2004. Land management in the american southwest: a state-and-transition approach to ecosystem complexity. Environmental Management 34:38–51.

Bhaskaran, K., A. Gasparrini, S. Hajat, L. Smeeth, and B. Armstrong. 2013. Time series regression studies in environmental epidemiology. International Journal of Epidemiology 42:1187–1195.

Brodersen, K. H., F. Gallusser, J. Koehler, N. Remy, and S. L. Scott. 2015. Inferring causal impact using Bayesian structural time-series models. Annals of Applied Statistics 9:247–274.

Campbell, D. T., and J. C Stanley. 1963. Experimental and quasi-experimental designs for research. Houghton Mifflin, Boston, Massachusetts, USA.

Card, D. 1990. The impact of the Mariel boatlift on the Miami labor market. ILR Review 43:245–257.

Carpenter, S. R. 1998. The need for large-scale experiments to assess and predict the response of ecosystems to perturbation. Pages 287–312 *in* M. L. Pace, and P. M. Groffman, editors. Successes, limitations, and frontiers in ecosystem science. Springer, Millbrook, New York, USA.

Copeland, S. M., S. M. Munson, D. S. Pilliod, J. L. Welty, J. B. Bradford, and B. J. Butterfield. 2018. Long-term trends in restoration and associated land treatments in the southwestern United States. Restoration Ecology 26:311–322.

Craig, P., S. V. Katikireddi, A. Leyland, and F. Popham. 2017. Natural experiments: an overview of methods, approaches, and contributions to public health intervention research. Annual Review of Public Health 38:39–56.

Dehejia, R. H., and S. Wahba. 2002. Propensity score-matching methods for nonexperimental causal studies. Review of Economics and Statistics 84:151–161.

Doudchenko, N., and G. W. Imbens. 2016. Balancing, regression, difference-in-differences and synthetic control methods: a synthesis. National Bureau of Economic Research. No. w22791. https://www.nber.org/papers/w22791

Duniway, M. C., B. T. Bestelmeyer, and A. Tugel. 2010. Soil processes and properties that distinguish ecological sites and states. Rangelands 32:9–15.

Ferman, B., C. Pinto, and V. Possebom. 2020. Cherry picking with synthetic controls. Journal of Policy Analysis and Management 39: 510–532.

Fick, S. T., C. B. Nauman, and M. Duniway. 2020. Evaluating natural experiments in ecology: using synthetic controls in assessments of remotely-sensed land-treatments. Dryad. https://doi.org/10.5061/dryad.1jwstqjt5.

Fiorella, M., and W. J. Ripple. 1993. Analysis of conifer forest regeneration using Landsat Thematic Mapper data. Photogrammetric Engineering and Remote Sensing 59: 1383–1388.

Gillan, J. K., J. W. Karl, N. N. Barger, A. Elaksher, and M. C. Duniway. 2016. Spatially explicit rangeland erosion monitoring using high-resolution digital aerial imagery. Rangeland Ecology & Management 69:95–107.

Gräler, B., E. Pebesma, and G. Heuvelink. 2016. Spatio-Temporal Interpolation using gstat. R Journal 8:204–218.

Holland, P. W. 1986. Statistics and causal inference. Journal of the American statistical Association 81:945–960.

Hurlbert, S. H. 1984. Pseudoreplication and the design of ecological field experiments. Ecological Monographs 54:187–211.

Hurlbert, S. H. 2004. On misinterpretations of pseudoreplication and related matters: a reply to Oksanen. Oikos 104:591–597.

Imbens, G. W., and T. Lemieux. 2008. Regression discontinuity designs: A guide to practice. Journal of Econometrics 142:615–635.

Ishwaran, H., and J. S. Rao. 2005. Spike and slab variable selection: frequentist and Bayesian strategies. Annals of Statistics 33:730–773.

Jones, B. A. 2018. Forest-attacking invasive species and infant health: evidence from the invasive emerald ash borer. Ecological Economics 154:282–293.

Karl, J. W., J. K. Gillan, N. N. Barger, J. E. Herrick, and M. C. Duniway. 2014. Interpretation of high-resolution imagery for detecting vegetation cover composition change after fuels reduction treatments in woodlands. Ecological Indicators 45:570–578.

Kaul, A., S. Klößner, G. Pfeifer, and M. Schieler. 2015. Synthetic control methods: Never use all pre-intervention outcomes together with covariates. Munich Personal RePEc Archive. No. 83790. https://mpra.ub.uni-muenchen.de/83790/

Kinn, D.2018. Synthetic control methods and big data. arXiv preprint arXiv:1803.00096.

Klößner, S., A. Kaul, G. Pfeifer, and M. Schieler. 2018. Comparative politics and the synthetic control method revisited: A note on Abadie et al. (2015). Swiss Journal of Economics and Statistics 154:11.

Kontopantelis, E., T. Doran, D. A. Springate, I. Buchan, and D. Reeves. 2015. Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis. The BMJ 350:h2750.

Larsen, A. E., K. Meng, and B. E. Kendall. 2019. Causal analysis in control-impact ecological studies with observational data. Methods in Ecology and Evolution 10:924–939.

Malmstrom, C. M., H. S. Butterfield, C. Barber, B. Dieter, R. Harrison, J. Qi, D. Riaño, A. Schrotenboer, S. Stone, and C. J. Stoner. 2009. Using remote sensing to evaluate the influence of grassland restoration activities on ecosystem forage provisioning services. Restoration Ecology 17:526–538.

Marsett, R. C., J. Qi, P. Heilman, S. H. Biedenbender, M. C. Watson, S. Amer, M. Weltz, D. Goodrich, and R. Marsett. 2006. Remote sensing for grassland management in the arid southwest. Rangeland Ecology & Management 59:530–540.

Monroe, A. P., C. L. Aldridge, M. S. O'Donnell, D. J. Manier, C. G. Homer, and P. J. Anderson. 2020. Using remote sensing products to predict recovery of vegetation across space and time following energy development. Ecological Indicators 110:105872.

Morgan, S. L., and C. Winship. 2015. Counterfactuals and causal inference. Cambridge University Press, Cambridge, UK.

Nauman, T. W., and M. C. Duniway. 2016. The automated reference toolset: a soil-geomorphic ecological potential matching algorithm. Soil Science Society of America Journal 80:1317–1328.

Oksanen, L. 2001. Logic of experiments in ecology: is pseudoreplication a pseudoissue? Oikos 94:27–38.

Oksanen, L. 2004. The devil lies in details: reply to Stuart Hurlbert. Oikos 104:598–605.

Peri, G., and V. Yasenov. 2019. The labor market effects of a refugee wave synthetic control method meets the mariel boatlift. Journal of Human Resources 54:267–309.

R Core Team 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. www.R-project.org

Rana, P., and D. C. Miller. 2019. Explaining long-term outcome trajectories in social–ecological systems. PLoS ONE 14: e0215230.

Rana, P., and E. Sills. 2018. Does certification change the trajectory of tree cover in working forests in the tropics? An application of the synthetic control method of impact evaluation. Forests 9:98.

Reed, B. C., J. F. Brown, D. VanderZee, T. R. Loveland, J. W. Merchant, and D. O. Ohlen. 1994. Measuring phenological variability from satellite imagery. Journal of Vegetation Science 5:703–714.

Roopsind, A., B. Sohngen, and J. Brandt. 2019. Evidence that a national REDD+ program reduces tree cover loss and carbon emissions in a high forest cover, low deforestation country. Proceedings of the National Academy of Sciences USA 116:24492–24499.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66:688.

Salley, S. W., C. J. Talbot, and J. R. Brown. 2016. The natural resources conservation service land resource hierarchy and ecological sites. Soil Science Society of America Journal 80:1–9.

Sills, E. O., et al. 2015. Estimating the impacts of local policy innovation: the synthetic control method applied to tropical deforestation. PLoS ONE 10:e0132590.

Steinemann, A. 2001. Improving alternatives for environmental impact assessment. Environmental Impact Assessment Review 21:3–21.

Verbesselt, J., R. Hyndman, G. Newnham, and D. Culvenor. 2010. Detecting trend and seasonal changes in satellite image time series. Remote sensing of Environment 114:106–115.

Waller, E. K., M. L. Villarreal, T. B. Poitras, T. W. Nauman, and M. C. Duniway. 2018. Landsat time series analysis of fractional plant cover changes on abandoned energy development sites. International Journal of Applied Earth Observation and Geoinformation 73:407–419.

Wang, J., P. M. Rich, and K. P. Price. 2003. Temporal responses of NDVI to precipitation and temperature in the central Great Plains, USA. International Journal of Remote Sensing 24:2345–2364.

Wernberg, T., D. A. Smale, and M. S. Thomsen. 2012. A decade of climate change experiments on marine organisms: procedures, patterns and problems. Global Change Biology 18:1491–1498.

Williams, B. K. 2011. Adaptive management of natural resources—framework and issues. Journal of Environmental Management 92:1346–1353.

Winkler, D. E., J. Belnap, D. Hoover, S. C. Reed, and M. C. Duniway. 2019. Shrub persistence and increased grass mortality in response to drought in dryland systems. Global Change Biology 25:3121–3135.

Xu, Y. 2017. Generalized synthetic control method: Causal inference with interactive fixed effects models. Political Analysis 25:57–76.

Supporting Information

Additional supporting information may be found online at: http://onlinelibrary.wiley.com/doi/10.1002/eap.2264/full

Data Availability

Data are available from the Dryad Digital Repository (Fick et al. 2020): https://doi.org/10.5061/dryad.1jwstqjt5. Simulation code is available on Zenodo: https://doi.org/10.5281/zenodo.4274935