WILEY ECOLOGICAL APPLICATIONS

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

Journal:	*Ecological Applications*
Manuscript ID	EAP20-0072
Wiley - Manuscript type:	Articles
Date Submitted by the Author:	07-Feb-2020
Complete List of Authors:	Fick, Stephen; US Geological Survey Southwest Biological Science Center Nauman, Travis; US Geological Survey Southwest Biological Science Center Brungard, Colby; New Mexico State University Duniway, Michael; US Geological Survey, Southwest Biological Science Center
Substantive Area:	Methods/Instrumentation/Software < Statistics and Modeling < Theory < Substantive Area, Spatial Statistics and Spatial Modeling < Statistics and Modeling < Theory < Substantive Area, Management < Substantive Area, Remote Sensing < Methodology < Substantive Area
Organism:	
Habitat:	
Geographic Area:	
Additional Keywords:	
Abstract:	Many important ecological phenomena occur on large spatial scales and/or are unplanned and thus do not easily fit within analytical frameworks which 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 relatively sophisticated techniques have been developed to derive insight from large, noisy observational datasets. 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 un-treated units, (2) a known date of treatment intervention, and (3) timeseries response data that includes 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—even in the absence of priori controls. 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. While we found the synthetic control approach useful for interpreting natural experiments with remote sensing data, and we expect the methodology to have wider utility in ecology, particularly for systems with large, complex, and poorly replicated experimental units, such as conservation districts, communities, and populations.

Evaluating natural experiments in ecology: using synthetic controls in assessments of remotely-sensed land-treatment effects.

Stephen E. Fick[1,2*]
Travis W. Nauman[1]
Colby C. Brungard[2]
Michael C. Duniway[1]
[1]US Geological Survey, Southwest Biological Science Center, Moab, UT
[2]New Mexico State University, Department of Plant and Environmental Sciences, Las Cruces, NM

* corresponding author: sfick@usgs.gov

Abstract

Many important ecological phenomena occur on large spatial scales and/or are unplanned and thus do not easily fit within analytical frameworks which 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 relatively sophisticated techniques have been developed to derive insight from large, noisy observational datasets. 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 un-treated units, (2) a known date of treatment intervention, and (3) timeseries response data that includes 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—even in the absence of priori controls. 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. While we found the synthetic control approach useful for interpreting natural experiments with remote sensing data, and we expect the methodology to have wider utility in ecology, particularly for systems with large, complex, and poorly replicated experimental units, such as conservation districts, communities, and populations.

Keywords

Causal Analysis, Timeseries, 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 which 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, Hurlbert 2004, Oksanen 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 (eg. 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 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. 2017), 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, 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 datasets 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, non-random 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 pre-defined references is the '*synthetic control*' method, emerging from the political science literature (Abadie et al. 2010). This approach attempts to reconstruct what would have happened (a 'counterfactual') had a treatment not occurred, based on the pre-intervention relationship between the unit of interest and a population of unaffected units. It is particularly useful for cases with a relatively small 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. While several methods have been proposed to model this relationship, 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 timeseries spanning pre- and post- treatment event for all control and treated units. Common methods include the original formulation proposed by Abadie et al. (2010) which generates a counterfactual from a weighted average of control units, and more recent models implementing latent interactive fixed-effects regression (Xu 2017) or Bayesian structural timeseries models (Brodersen et al. 2015). In some sense, the most basic implementation of the synthetic control approach is the classic "Difference in Differences" method (hereafter DiD) whereby the average difference between control and treatment are compared before and after the intervention (Ashenfelter and Card 1985, Abadie 2005, Craig et al. 2017). As each model formulation carries its own set of assumptions and strictures (e.g. tolerance of missing data, assumption of parallel trajectories through time, ability to extrapolate, etc.), different methods will likely have advantages and disadvantages in ecological applications.

Synthetic Control in Ecology

The few previous uses of synthetic control in environmental contexts has predominantly focused on determining the effectiveness of policies or events on forest dynamics and socio-economic outcomes (Sills et al. 2015, Jones 2018, Rana and Sills 2018, Rana and Miller 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 timeseries of remote sensing imagery – i.e. data that is temporally and spatially extensive but also noisy and prone to confounding. Typical approaches for inferring effects from remote sensing data generally (a) use only the timeseries of treated pixels and thus ignore potentially useful contextual information from unaffected areas (Copeland et al. 2019, Fiorella and Ripple, 1993, Monroe et al. 2020), or (b) use differencing techniques (e.g. DiD) which may over-simplify the contextual information provided by controls (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).

In this study we evaluate the performance of several methods for assessing landscape-scale treatment effects (timeseries-only, DiD, and synthetic control) using a simulated satellite timeseries 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 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 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

Simulation modeling

We examined three approaches for estimating landscape-scale treatment effects using simulated remote sensing data (Table 1): (1) a timeseries-only method which does not consider controls (BFAST; Vesserbelt et al. 2010); (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: (a) A linear interactive fixed effects model with latent confounders using the R package `gsynth` (Xu 2017), and (b) A Bayesian structural timeseries model using the R package `CausalImpact` (Brodersen et. al 2015). Gsynth generates counterfactuals by first estimating a set of time varying latent factors (essentially unobserved confounders) for which each unit has a specific coefficient, or 'loading', using data from the control population (Bai 2009). The loadings for treated units are then estimated and used to predict in the post treatment period. The number of latent factors is a key parameter determining the flexibility of model and is estimated via cross-validation. The model in CausalImpact uses a state-space or hidden Markov chain framework in which the data generating process is divided into a 'state' equation that represents the temporal evolution of a latent process and a 'observation' equation which 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).

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 Development Core Team 2015). It is important to note that the timeseries-only method used here, BFAST, is commonly used for changepoint detection (i.e. without a priori knowledge about the date of an intervention), and we use it here simply as a coarse baseline or 'null hypothesis' for estimating trends without considering controls.

We generated simulated 16-day NDVI timeseries data following the approach of Vesserbelt et al. (2010) by additively combining an NDVI signal from a hypothetical treatment with various sources of noise (Fig Example). 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 (Vesserbelt et al. 2010). The treatment effect was modeled as an abrupt reduction in NDVI (-0.3) such as from a large disturbance (e.g. fire or clearing), followed by a linear recovery over four years (Fig 1, 'Treatment' panel). Following Vesserbelt et al (2010) we added random Gaussian noise, systematically controlling the variance of this noise among simulations (s.d. = 0.1, 0.2, …, 0.7; Figure 1, 'Noise' panel).

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 timeseries: 1) random drops of 0.25 NDVI, corresponding to cloud contamination or sensor error in a satellite image (Fig. 1, panel 'Satellite'); 2) a growing-season climate anomaly resulting in increased or decreased production (Fig. 1, panel 'Climate'); and 3) signal drift over time as from vegetative dynamics (Fig. 1, 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 April 20, 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. This coefficient was determined by adding `one` to a value drawn from a zero-mean gaussian distribution (sd = .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 simulations, control pixels from different landscape types (forest or grassland) were designed to respond to the same confounders only at a different magnitude. To assess the effects of potential mismatch between control and treated pixels on the accuracy of different methods, we generated three different scenarios (Fig. 2): 1) All control pixels are of the same landscape type (forest or grassland) as the treated pixel (mismatch = 0); 2) Fifty percent 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).

For each combination of conditions (landscape type, control mismatch, number of controls, random noise level) we generated 1000 simulated timeseries and obtained treatment effect estimates for all methods (Table 1). We assessed errors as the absolute difference between the 'true' simulated treatment effect ('Treatment' in Fig 1) and treatment effects estimated by various methods, at each point in the post-treatment time period for each simulation. For methods which provided confidence intervals we also assessed sensitivity by counting whether estimated treatment effect intervals overlapped zero at each point in the post-treatment timeseries. Details for each method are supplied in APPENDIX and simulation code is hosted at https://github.com/fickse/ssim.

Case Study

We demonstrate the use of synthetic control for inferring management intervention effects without *a priori* controls in the context of a brush-clearing treatment which 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 (details in Karl et al. 2014 and Gillan et al. 2016). There is some contention around the long-term effectiveness of such treatments, as well as potential erosion risks from increased exposure of bare ground following treatment (Archer et al. 2011, Gillan et al. 2016).

Within a designated section of the broader treated area, three types of brush-clearing methods were applied in distinct zones (fig 5): (1) Mechanical tree mastication, leaving debris scattered throughout, (2) Lopping followed by burning piled debris and (3) Lopping followed by broadcast burn of scattered debris. A fourth area was used as a control and monitored for pre- and post-treatment surface cover as with the other areas (Karl et al. 2014). We obtained rough outlines of the treated and control areas from the Utah watershed Restoration Initiative dataset (https://wri.utah.gov/wri/).

We assessed treatment effects based on the Soil Adjusted Total Vegetative Index (SATVI ; Marsett et al. 2006), which has been shown to accurately reflect total vegetative cover in the region of the case study (Poitras et al. 2018) . We calculated SATVI as:

$$SATVI = 1.9 * \frac{SWIR1 - RED}{SWIR1 + RED + 0.9} - \frac{SWIR2}{2}$$

using a timeseries of images from the Landsat archive from 1984 to 2018. Since single sensors do not span the entire timeseries, we used Landsat 5 for years between 1984 and 2011, Landsat 7 for 2012, and

Landsat 8 for 2013 to 2018. As the synthetic control method in theory automatically accounts for satellite-derived noise shared among pixels, we were interested in performance of methods without recalibrating different Landsat products to a standard reflectance or subjecting images to cloud-masking algorithms. We used tier-1 surface reflectance products from all satellites, compiled using google earth engine (Gorelick et al. 2017).

For each pixel in the target areas, we identified a set of 100 control pixels, adapting methods from Nauman and Duniway (2016). Briefly, within a search radius of 3 km surrounding the perimeter of the treated area, we first performed a 'masking' operation, removing from consideration any pixels known to be part of another treatment or within 90 m of the focal treatment boundary, those disturbed by infrastructure (roads, oil and gas development), or those belonging to a non-analogous landscape cover-class (e.g. agriculture, urban, water) according to the National Landcover Dataset (NLCD 2011). We then narrowed candidate pixels to those with similar salinity (+- 5%) measured as saturated paste soil electrical conductivity and particle size in the control section classification (Soil Survery Staff 2010) to the focal treated pixel (Nauman et al. 2019). Restricting candidate reference locations by salinity and soil textural class was found to be an important step in this process, given the outsized role these variables play in determining ecological potential in these arid contexts (Nauman and Duniway 2016). From this subset, we selected the 100 most-similar pixels in the pool of control pixels, using Gower's distance (van der Loo 2019) based on a suite of topo-edaphic variables (Table A1). We estimated treatment effects using the same methods outlined in the simulation model exercise, for each pixel, considering all observations after June 1, 2009 as post-treatment.

Results

Simulations

In simulations, absolute point-wise errors for the different methods of determining treatment effects (timeseries 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 which included controls were superior to the baseline estimates from the timeseries-only method (BFAST), regardless of the number of controls available (Fig. 3, top row).

As more mismatched pixels were introduced to the control population, accuracy depended more on the number of controls available, with larger number of controls generally improving estimates for the synthetic control methods (Fig 3, middle row). The CausalImpact synthetic control method needed only 5 controls to achieve estimates superior to baseline, while gsynth required between 5 and 50. Unlike the synthetic control methods, DiD was generally less accurate than the timeseries-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 timeseries-only method, and only with many controls (Fig 3, 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 / s.d. of random noise) led to marginal reductions in error (Figure 3), 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 (Figure A2).

Confidence envelopes for treatment effects revealed differences between methods, which varied by level of noise and control-mismatch (Figure 4). The CausalImpact method was the most conservative (low sensitivity), especially when the magnitude of confounding was high (Fig 4). Even when the signal-to-noise ratio was high and confounding relatively low, approximately 50% of the true effects were determined to be significantly different from zero. Both DiD and gsynth method tended to have smaller confidence intervals, which resulted in more frequent 'significant' treatment effects but also erroneously significant effects when the treatment effect was negligible (fig 4). The prediction intervals for both DiD and gsynth did appear to reflect greater uncertainty in cases where the control populations were perfectly mismatched to treatment (Fig 4 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 2 right column).

Case Study

For the brush-clearing case study, estimated treatment effects were similar among all methods which included controls (Fig 6), all of which providing greater discrimination among treatment types than BFAST. The pixel-level estimates were heterogeneous within treatment areas (Fig 5, panel D; Fig. 6), with some pixels having greater treatment effects than others. This can be visualized in certain regions of the mastication treatment (Fig 5 panel D), where small stands of brush were left intact or in peripheral rocky areas with little brush to begin with.

Remote sensing estimates for treatment effects using SATVI, a proxy for ground cover, followed the same general ordering as ground cover change reported in Karl et al. (2014, fig. 6), with broadcast burn (B) having the greatest overall drop in SATVI, followed by pile burn (P) and then mastication (M).

However, the increase in ground cover for the mastication treatment (M) observed by Karl et al. (2014) was not detected in this exercise, perhaps indicating that SATVI was not sensitive to the increased litter derived from slash debris or that debris did not compensate for reduced canopy cover. The broadcast burn area was also associated with greater wind and water erosion than the pile burn and control areas, as reported in Gillan et al. (2016), indicating that simple satellite-derived assessments may be useful for indicating relative functional treatment effects.

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 both the simulations and the case study, methods which incorporated data from some kind of properly-matched, untreated group were more accurate at estimating ‘true’ treatment effects than methods which relied on timeseries alone (Fig 3, Fig 6). For data with many potential confounding variables, such as remote sensing timeseries, controls provide an intuitive baseline to remove these effects. In the simulations, relatively large (but not unreasonably so; Vesserbelt 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 dynamic landscapes subject to greater confounding (Reed et al. 1994). However, in the brush clearing case study, where the landscape is dominated by perennial tree and shrub species, use of the CausalImpact method helped discriminate slight variations in treatment effects, compared to other methods, suggesting that at least some confounding noise was removed (Fig 6). Patterns among treatments in particular become more apparent when visualizing cumulative values (Fig. 7).

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 accurately estimate the treatment effect given enough control data in simulations, likely in part because it explicitly includes a seasonality component in its model (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. grassland). Our results highlight the important role of finding accurate matches between control and treatment populations, a common challenge in observational studies in both the physical and social sciences. The further development and implementation of reliable, automated techniques for finding spatial comparisons across ecological contexts is needed (Nauman and Duniway 2016).

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. 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 (R2 for the CausalImpact Bayesian Structural Timeseries model and mean squared prediction error for gsynth). 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. confidence intervals not crossing zero) occurring at a maximum of roughly 50% of the pointwise instances with very high signal-to-noise ratios (Fig. 4). By contrast, the gsynth method typically had tighter confidence bands (e.g. fig 2) even in cases where predictions were obviously poor due to mismatched controls (Fig. 4). These overly-narrow confidence bands may be an artifact of violated assumptions of the parametric standard error estimates used in simulations (Xu 2017).

Notes for Application

In the case study, we observed significant amounts of heterogeneity in estimated treatment effects, both within treated areas and through time (Fig. 5). 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 this study we also used raw 16-day SATVI timeseries 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 5 Panel F). In practice, an additional step of cloud masking, aggregating to a broader temporal scale or implementing low-pass filtering on the timeseries may help improve results. In this example we also used only the pre-treatment SATVI control pixel timeseries for modeling the relationship between treatment and controls, but potentially any number of other time-varying predictors could also be included (but see Ferman et al. 2017), including other remotely sensed indices or climate data.

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) which parse regions by ecological potential (e.g. NRCS Ecologcal 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). Our approach in this study essentially systematizes this search for suitable comparison sites by integrating information both about ecological potential (Soils and topography) and ecological state (remotely-sensed timeseries of vegetation). Thus, this approach may be seen as a quantitative and scalable framework for conducting a common activity which 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 datasets (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 timeseries 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 datasets in ecology.

Acknowledgements

This research was conducted with support from the US National Resources Conservation Service and the US Geological Survey ecosystems mission area.

Literature Cited

Abadie, A. 2005. Semiparametric difference-in-differences estimators. The 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.

Archer, S., K. W. Davies, T. E. Fulbright, K. C. McDaniel, B. P. Wilcox, K. Predick, and D. Briske. 2011. Brush management as a rangeland conservation strategy: a critical evaluation. Conservation benefits of rangeland practices: assessment, recommendations, and knowledge gaps'.(Ed. DD Briske) pp:105–170.

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.

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

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.

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

Carpenter, S. R. 1998. The need for large-scale experiments to assess and predict the response of ecosystems to perturbation. Pages 287–312 Successes, limitations, and frontiers in ecosystem science. Springer.

Copeland, S. M., S. M. Munson, D. S. Pilliod, J. L. Welty, J. B. Bradford, and B. J. Butterfield. 2017. 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.

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. 2017. Cherry picking with synthetic controls. Munich Personal RePEc Archive 80970.

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.

Gorelick, N., M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore. 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment 202:18–27.

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.

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.

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.

van der Loo, M. 2019. gower: Gower’s Distance. R package version 0.2.1.

Malmstrom, C. M., H. S. Butterfield, C. Barber, B. Dieter, R. Harrison, J. Qi, D. Riaño, A. Schrotenboer, S. Stone, C. J. Stoner, and others. 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.

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.

Nauman, T. W., C. P. Ely, M. P. Miller, and M. C. Duniway. 2019. Salinity yield modeling of the Upper Colorado River Basin using 30-meter resolution soil maps and random forests. Water Resources Research.

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.

Poitras, T. B., M. L. Villarreal, E. K. Waller, T. W. Nauman, M. E. Miller, and M. C. Duniway. 2018. Identifying optimal remotely-sensed variables for ecosystem monitoring in Colorado Plateau drylands. Journal of Arid Environments 153:76–87.

R Development Core Team. 2015. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

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.

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., D. Herrera, A. J. Kirkpatrick, A. Brandão, R. Dickson, S. Hall, S. Pattanayak, D. Shoch, M. Vedoveto, L. Young, and A. Pfaff. 2015. Estimating the Impacts of Local Policy Innovation: The Synthetic Control Method Applied to Tropical Deforestation. PLoS ONE 10.

Soil Survery Staff. 2010. Keys to Soil Taxonomy. 11th Edition. United States Department of Agriculture, Natural Resources Conservation Service, Washington, DC.

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

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.

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.

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

Tables

Table 1

Method	Approach	Method	Citation
Timeseries-only	Timeseries decomposed into seasonal, trend, and noise components.	Trend component estimated with iterative breakpoint detection and piecewise linear regression.	Verbesselt et al. 2010
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: $Y_{it} = C_i + A_t + B * D_{it} + E_{it}$ Where Y_{it} is the response for pixel *i* at time *t*, C is a vector of individual effects, A is a vector time effects, B is the treatment effect with D_{it} a 0/1 dummy variable indicating treatment and error as E_{it}	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.	Xu 2017
		Bayesian structural timeseries model with local linear trend, seasonality and linear regression components in the 'process' part of the model.	Broderson et al. 2018

Figure Legends

Figure 1

Example of a simulated NDVI timeseries 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 timeseries and estimated by various methods.

Figure 2

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: The same simulated NDVI signal (red) with differing control pixels (grey). Bottom rows: Estimated treatment effect (solid line), actual treatment effect (dashed line), and confidence intervals (shading) for different methods. The treatment occurs in February 2006 and is indicated by a vertical dotted line.

Figure 3

Simulation results for average absolute 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 broken down by number of controls available (columns) and degree of mismatch between the control population and treated pixels (rows; top = no mismatch, middle = equally well and poorly matched, bottom = total mismatch).

Figure 4

Power curves showing the percentage of time each method identified significant treatment effects (confidence interval excluding zero). Rows represent level of mismatch between treated and reference pixels, columns represent pointwise absolute magnitude of net confounder effect (season + climate + drift + satellite) and signal-to-noise ratio is represented on the x-axis of each panel. Data shown are from simulations with greater than five controls.

Figure 5

Shay Mesa pinion-juniper clearing case study. A and B: Treated Site before and after brush treatments. Treatments include (M) mastication, (P) pile burning, (B) broadcast burning and (C) control. Panel C: demonstration of pixel-matching algorithm modifying Nauman and Duniway (2016). One hundred control pixels were selected for each treated pixel from a narrowed pool of candidates with similar topographic and soil properties. Panel D: estimated median per-pixel treatment effect using the CausalImpact method for the 2010 growing season (Mar – Nov) in units of SATVI * 1000. Panel E depicts the raw Satvi timeseries for the treated (red) and control (grey) pixels. Panel F depicts point-wise estimated treatment effects and trendline using the Bayesian structural timeseries in CausalImpact. Panel G depicts cumulative treatment effects, analogous to exposure of bare ground integrated over time.

Figure 6

Effect of brush clearing treatments on treated areas by assessment method. Treatments include control (C), mastication (M), pile-burn (P) and broadcast burn (B). Top Half: Distributions represent all estimated pixel-wise median effects of treatments on SATVI (x axis = SATVI * 1000) between March and October

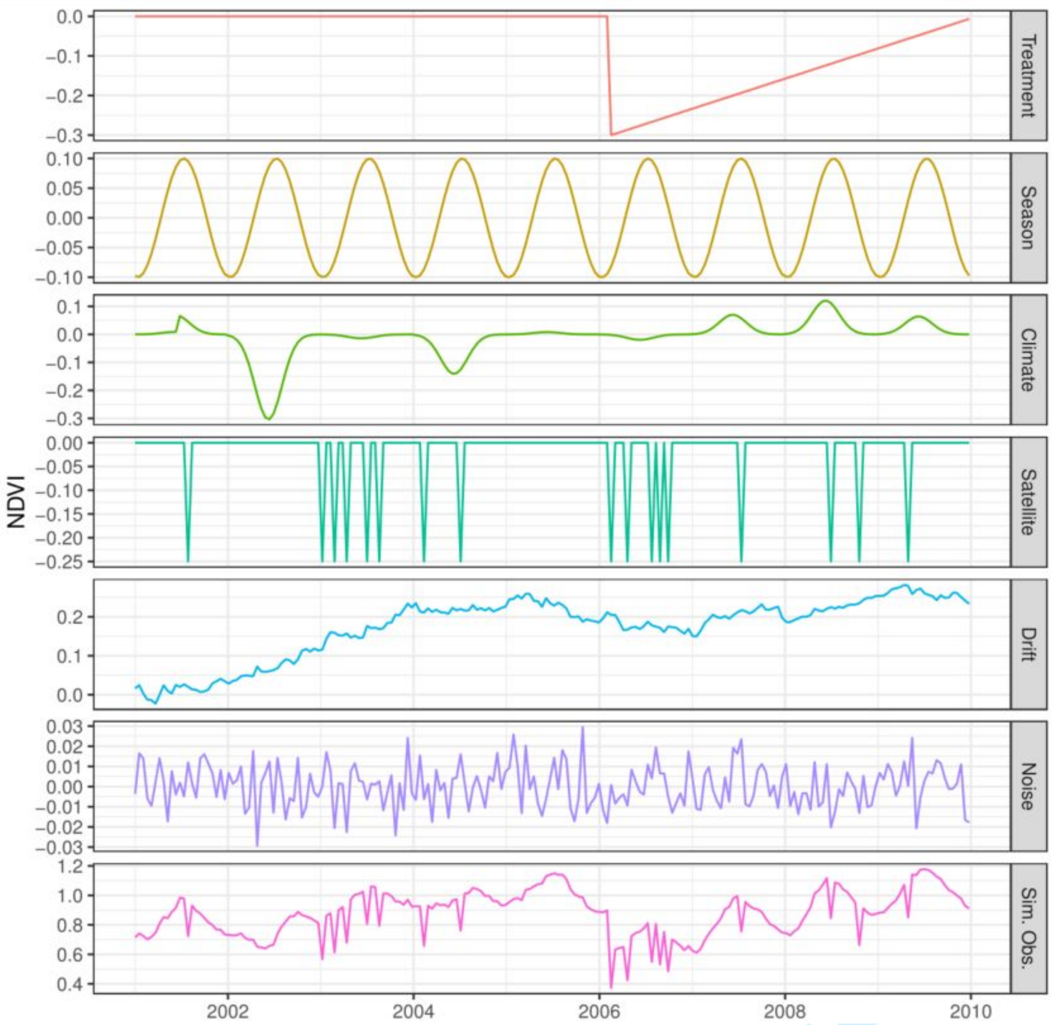
2010, one year after implementation. Colors indicate population quantiles. Bottom Half: Percent change in repeated line-point intercept cover values before and after treatments at Shay Mesa, from Karl et al. 2014.

Figure 7

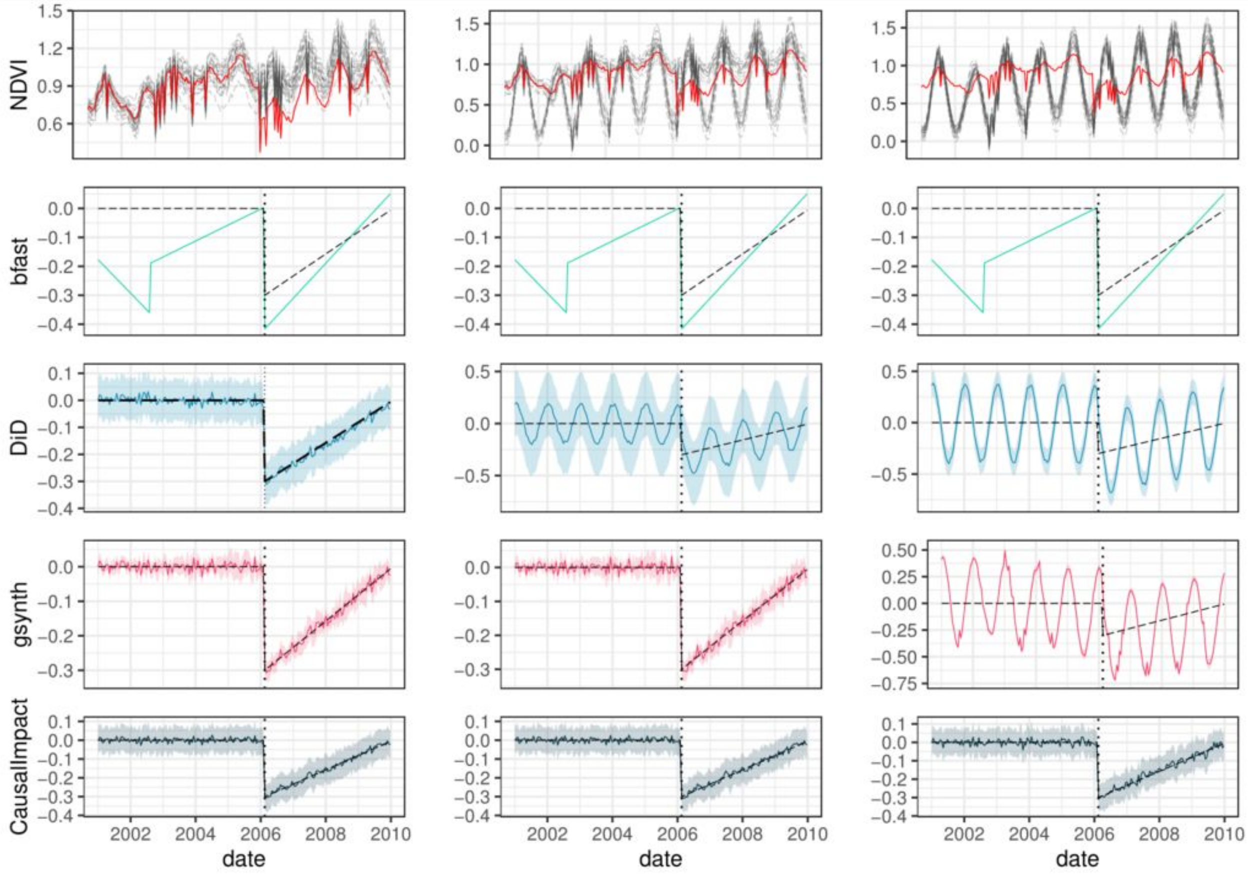
Cumulative treatment effects using Causal Impact. Small lines indicate individual pixel trajectories and thick lines represent trends by treatment-type.

Figures

Figure 1



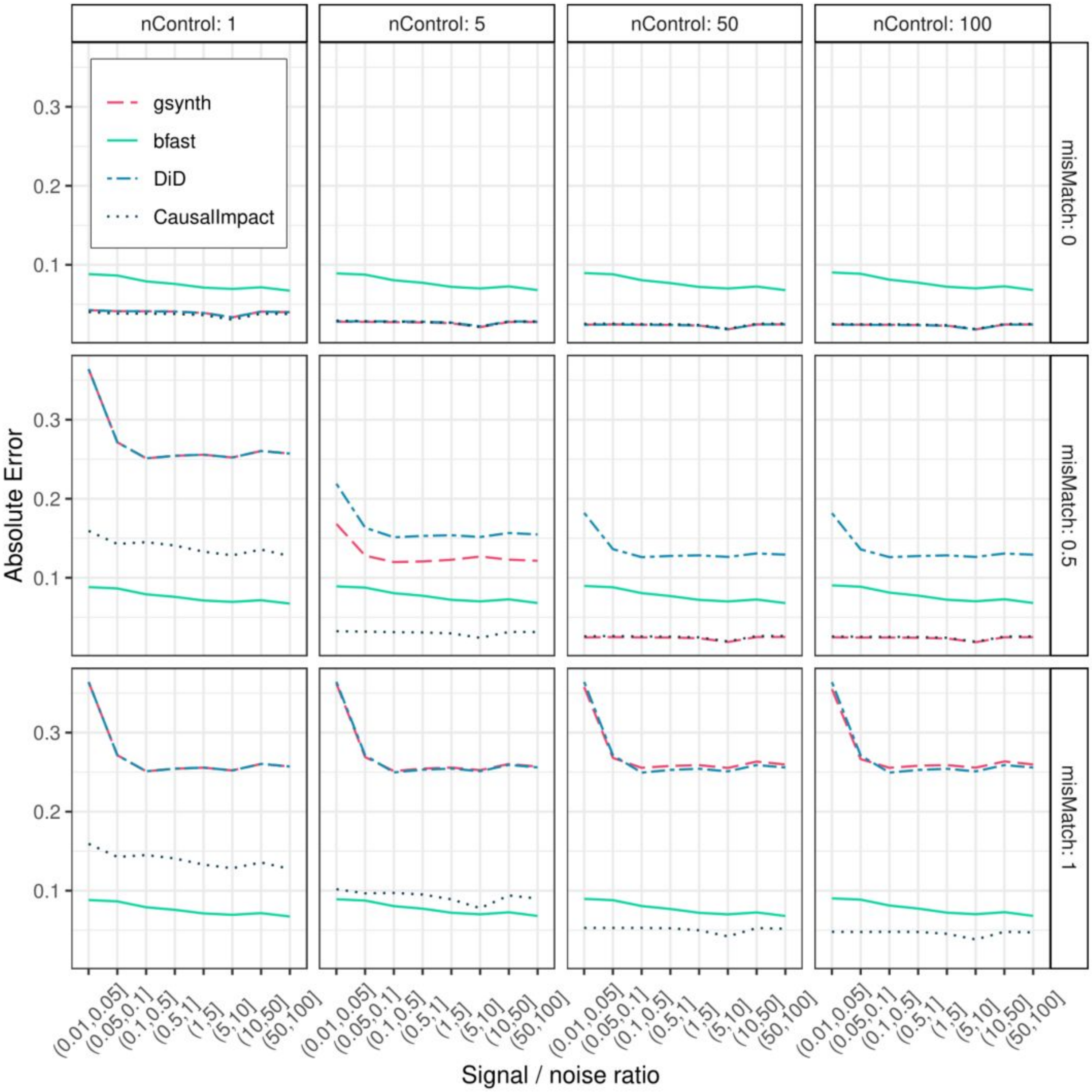
574 Figure 2



575

576

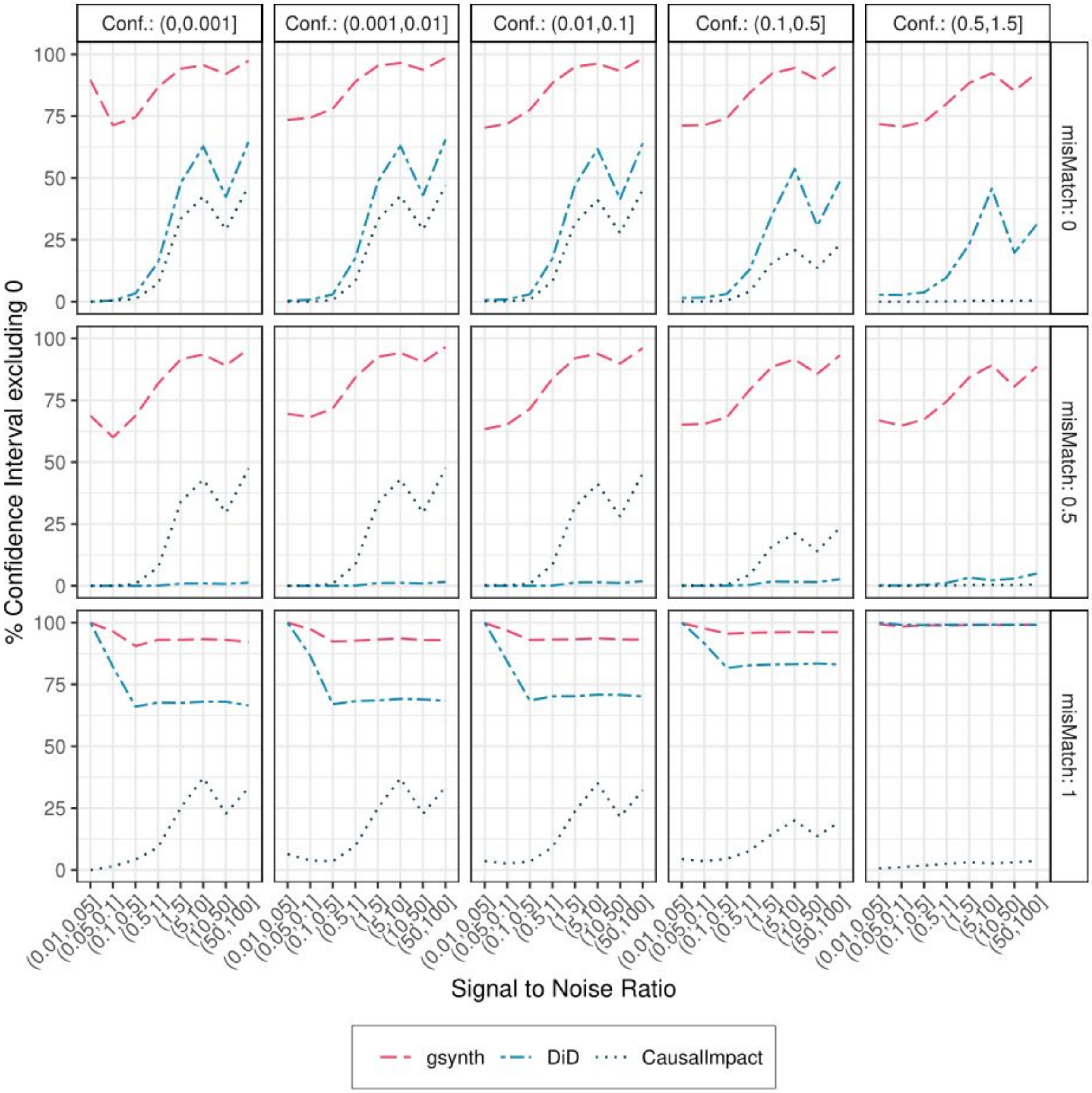
577 Figure 3



578

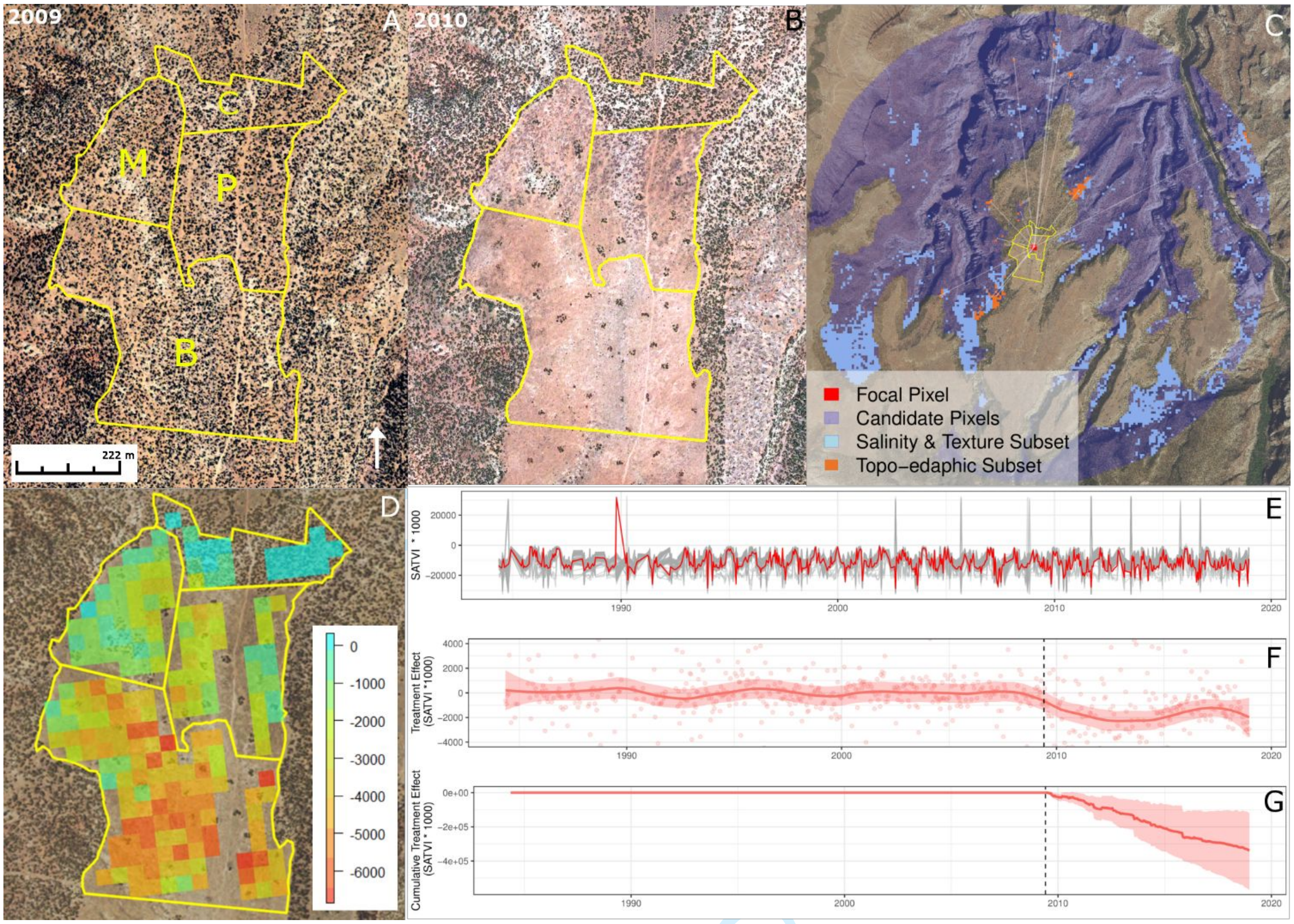
579

580 Figure 4



581

Figure 5


2009
A
C
M
P
B
222 m
2010
B
C
Focal Pixel
Candidate Pixels
Salinity & Texture Subset
Topo−edaphic Subset
D
0
-1000
-2000
-3000
-4000
-5000
-6000
E
SATVI * 1000
20000
0
−20000
1990
2000
2010
2020
F
Treatment Effect (SATVI *1000)
4000
2000
0
−2000
−4000
1990
2000
2010
2020
G
Cumulative Treatment Effect (SATVI *1000)
0e+00
-2e+05
-4e+05
1990
2000
2010
2020

585 Figure 6

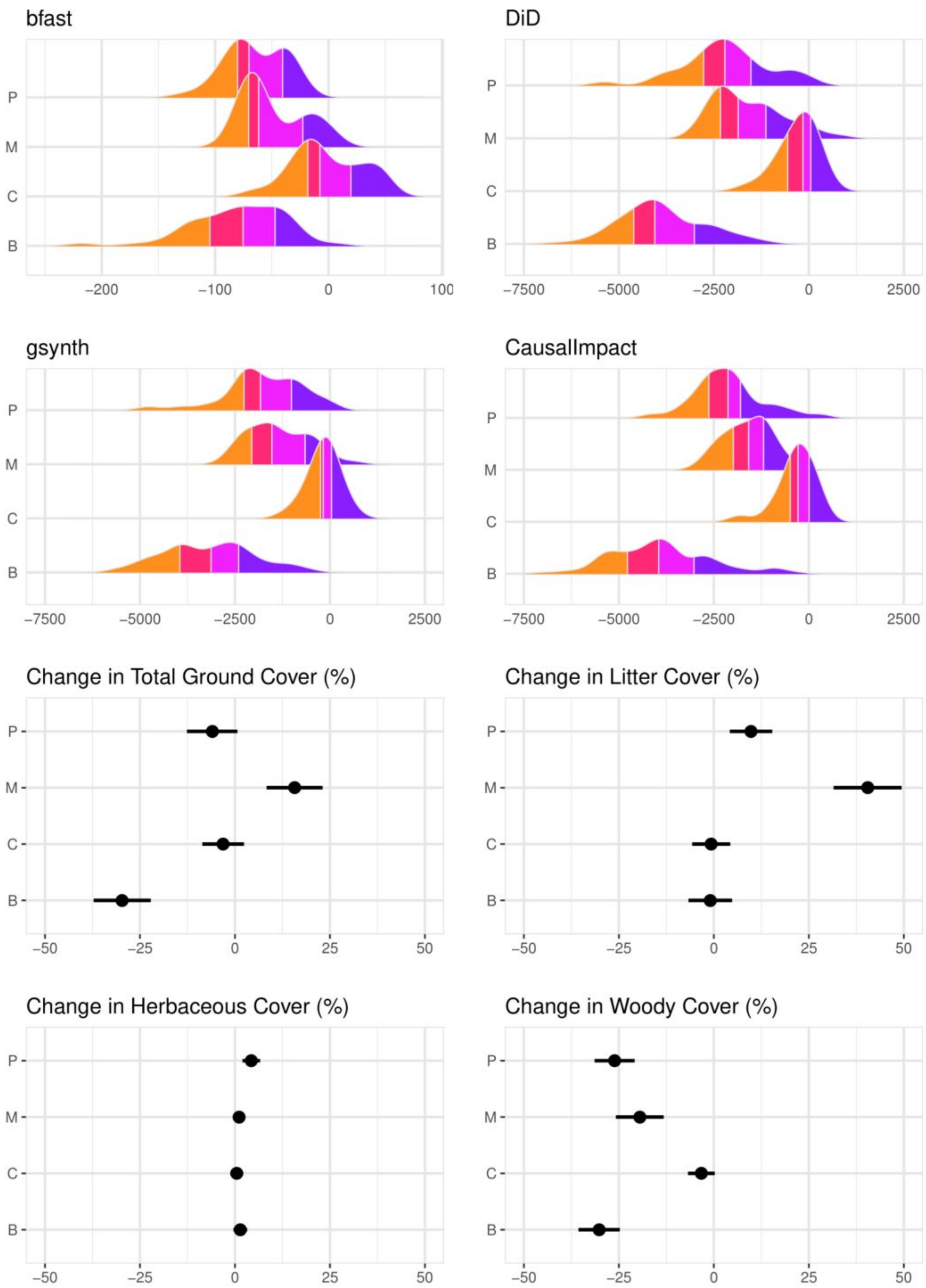
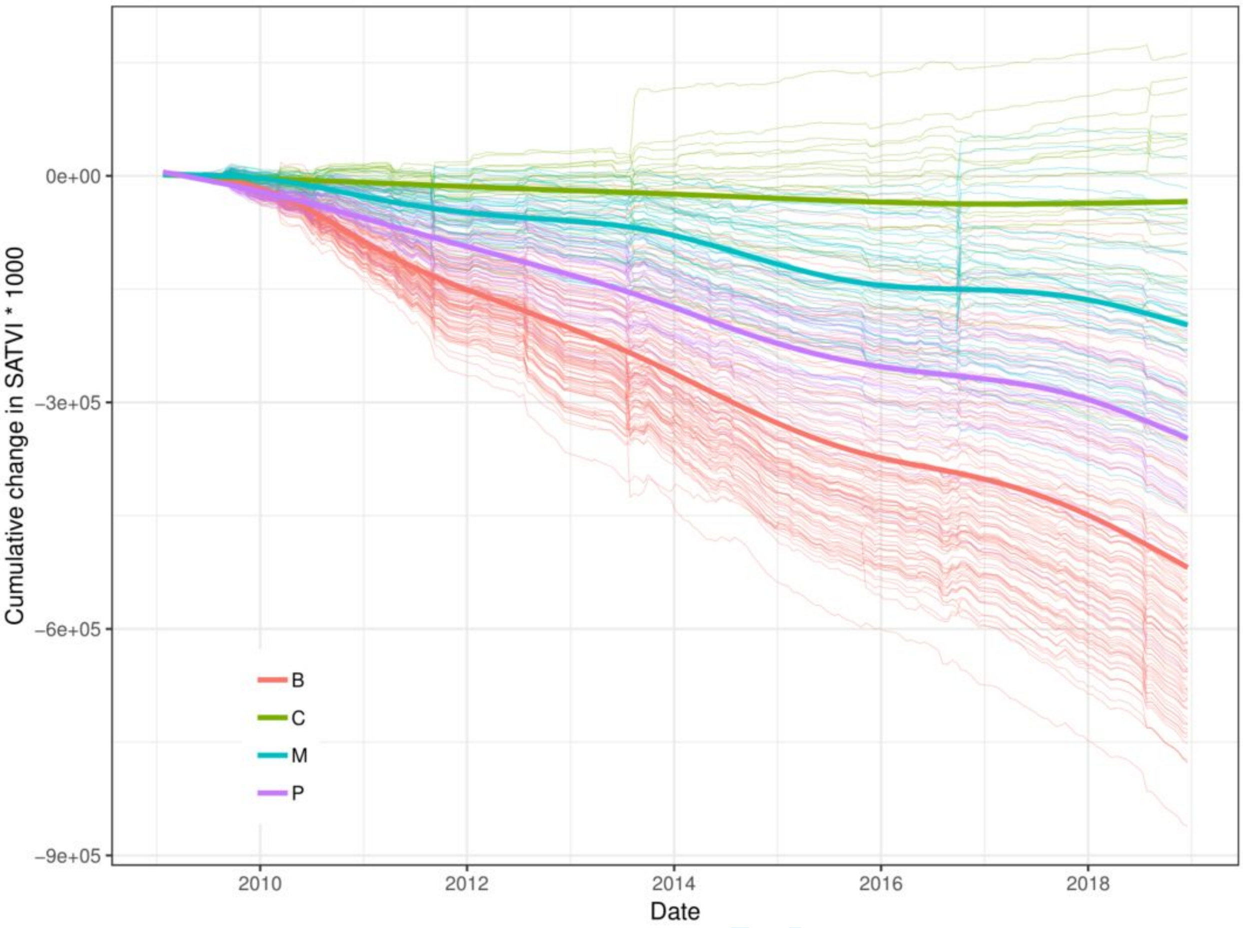

bfast
P
M
C
B
−200
−100
0
100
DiD
P
M
C
B
−7500
−5000
−2500
0
2500
gsynth
P
M
C
B
−7500
−5000
−2500
0
2500
CausalImpact
P
M
C
B
−7500
−5000
−2500
0
2500
Change in Total Ground Cover (%)
P
M
C
B
−50
−25
0
25
50
Change in Litter Cover (%)
P
M
C
B
−50
−25
0
25
50
Change in Herbaceous Cover (%)
P
M
C
B
−50
−25
0
25
50
Change in Woody Cover (%)
P
M
C
B
−50
−25
0
25
50

Figure 7


Cumulative change in SATVI * 1000
0e+00
−3e+05
−6e+05
−9e+05
2010
2012
2014
2016
2018
Date
B
C
M
P

Appendix

Figure A1

Power curves showing empirical frequency of concluding that an effect is different from zero, by method and level of mismatch between treated and reference pixels.

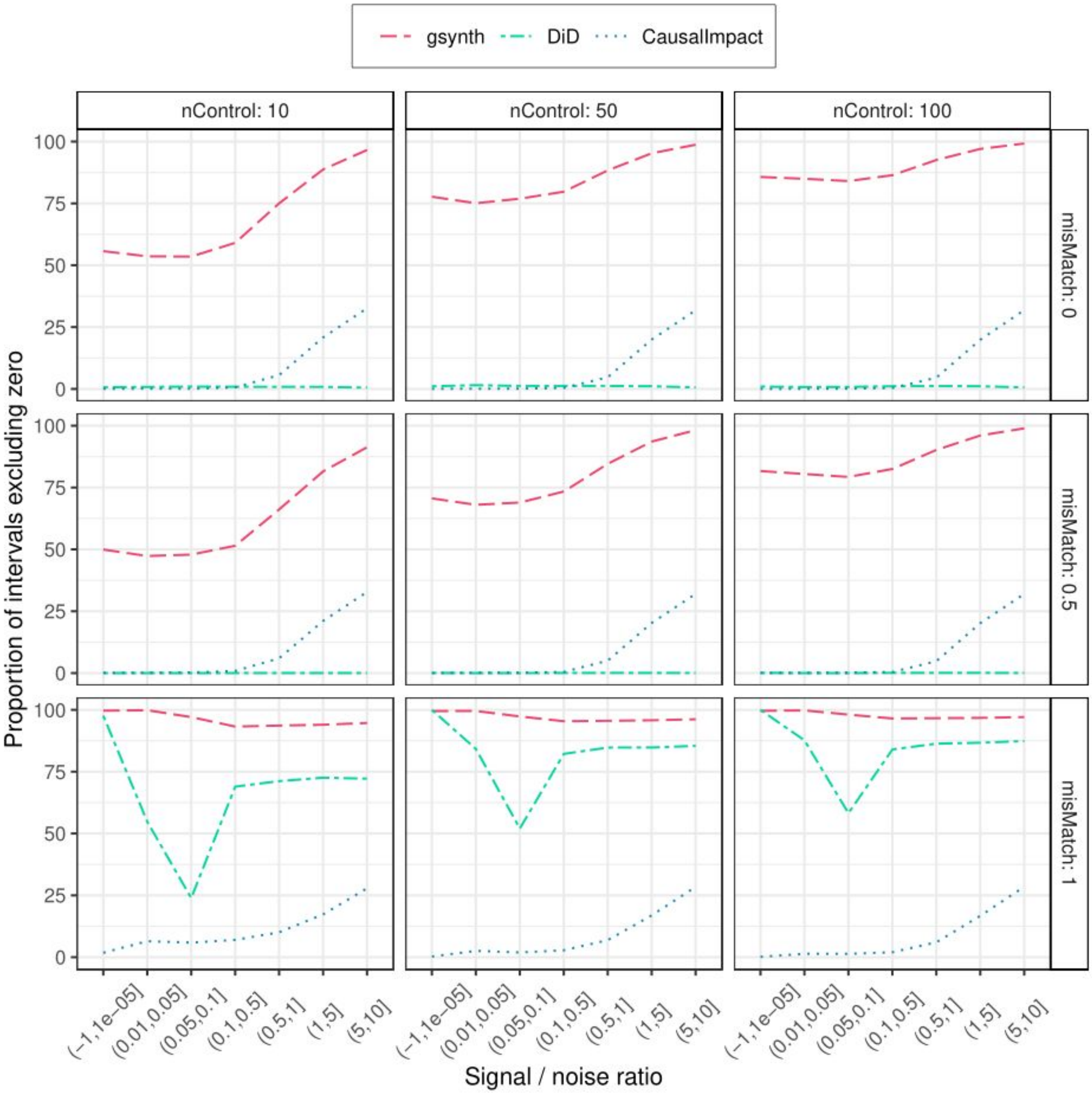


Figure A2

Absolute point-wise error by method and magnitude of confounder (columns), mismatch (rows), and signal-to-error ration (x axis). Data shown for simulations with greater than 5 controls.

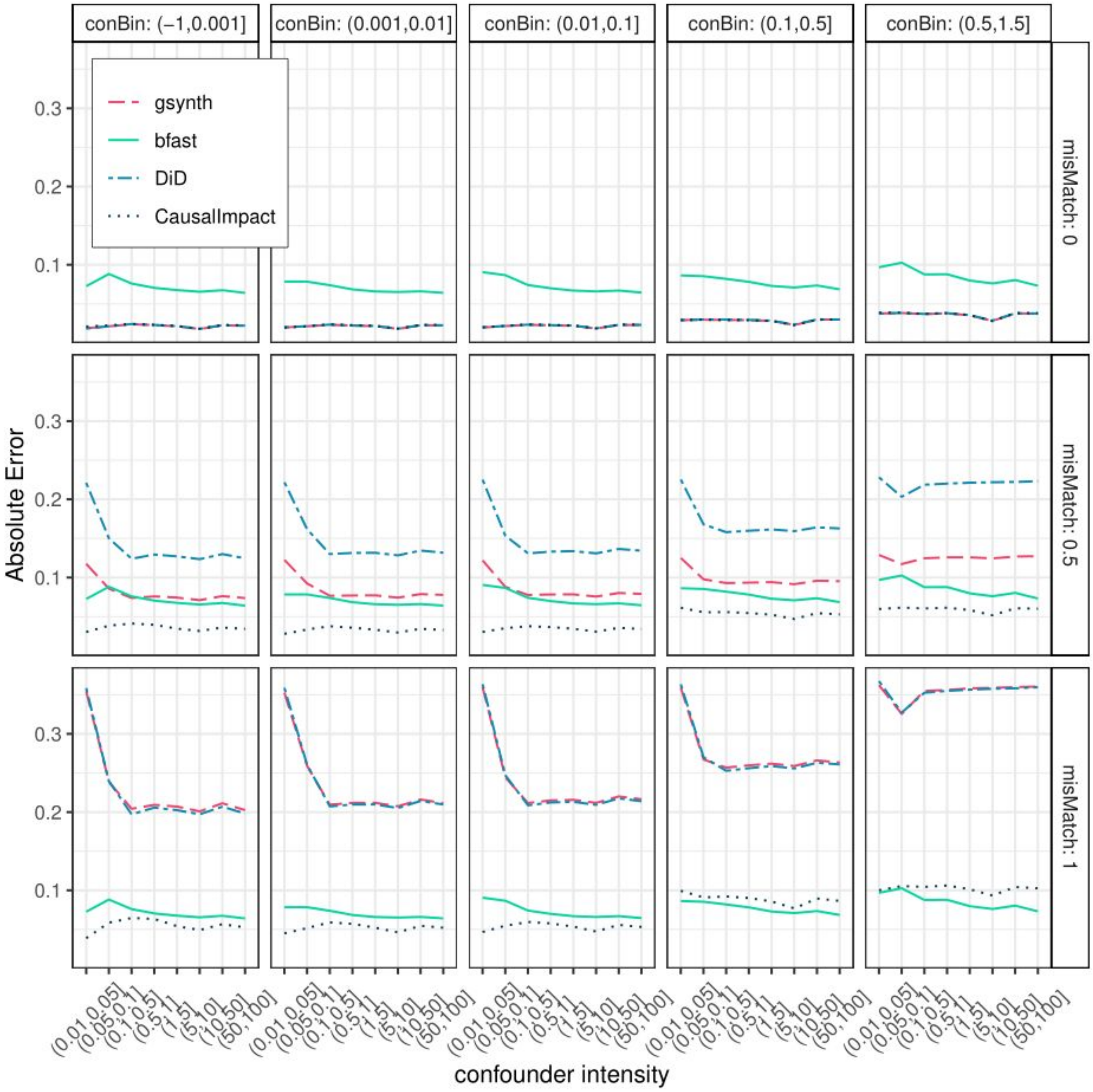


Table A1

Variable	Usage	Source(s)	Preparation Notes
SATVI	response	USGS Landsat 5 (years 1984-2011) , 7 (2012), and 8 (2013-2018) Tier 1 Surface Reflectance. From Google Earth Engine (Gorelick et al. 2017)	Calculated as: $SATVI = 1.9 * \frac{SWIR1 - RED}{SWIR1 + RED + 0.9} - \frac{SWIR2}{2}$
Roads	mask	US Census Bureau TIGER primary and secondary roads, 2018. <https://www2.census.gov/geo/tiger/TIGER2018PLtest/ROADS/>	Interstates and major roads buffered to 60 m, local roads buffered to 30 m
Land Cover	mask	NLCD 2011 Land Cover (CONUS). <https://www.mrlc.gov/data>	Masked and buffered by 1 pixel (30 m) all water(11), snow (12), developed(21-24), and cultivated(81,82) pixels
Fires	mask	Monitoring Trends in Burn Severity (MTBS) fire perimeters (https://www.mtbs.gov/)	Masked with 30m buffer
Disturbances	mask	LandFire LF 1.4.0 disturbance grids. <https://www.landfire.gov/getdata.php>	Masked any pixel with non-zero disturbance value between 1999 and 2016
Other Land Treatments	mask	Utah Watershed Restoration Initiative (wri.utah.gov)	Masked all completed treatment perimeters using a 30m buffer
Elevation	matching	National Elevation Dataset, 1-arc second, meters	elevation in meters
Slope	matching	National Elevation Dataset, 1-arc second, meters	Slope gradient in degrees
Southness	matching	National Elevation Dataset, 1-arc second, meters	index from 1 to -1 of how northwest (1) or southeast (-1) a site faces
Eastness	matching	National Elevation Dataset, 1-arc second, meters	index from 1 to -1 of how south (1) or north (-1) a site faces
PCURV	matching	National Elevation Dataset, 1-arc second, meters	curvature parallel to the slope direction
TCURV	matching	National Elevation Dataset, 1-arc second, meters	curvature perpendicular to the slope direction
Relative Height	matching	National Elevation Dataset, 1-arc second, meters	Height of cell above the local minimum elevation. Included separate variables including local neighborhoods of 1, 32, 128 pixels
RELMNHT	matching	National Elevation Dataset, 1-arc second, meters	Height of cell above the local mean elevation. Used separate variables including neighborhoods of 1, 32, 128 pixels
MRRTF	matching	National Elevation Dataset, 1-arc second, meters	multiple resolution ridgetop flatness index
MRVBF	matching	National Elevation Dataset, 1-arc second, meters	multiple resolution valley bottom flatness index
Topographic Wetness Index	matching	National Elevation Dataset, 1-arc second, meters	Topographic wetness index (TWI) from topmodel in SAGA GIS.
Calog_10	matching	National Elevation Dataset, 1-arc second, meters	Upslope contributing area in log$_{10}$ units
LFELEMS	matching	National Elevation Dataset, 1-arc second, meters	Landform classification system using DEM: landform elements

Soil EC	Edaphic matching	Nauman et al. (20XX)	Soil electrical conductivity (dS/m) averaged from 0 to 60 cm, saturated paste method
Soil Particle Size	Edaphic matching	Nauman et al. (20XX)	Soil particle size class (family level of US soil taxonomy) raster map

U.S. Census Bureau, 2018. TIGER/Line Shapefiles (machine- readable data files).

https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.2018.html