Evaluating natural experiments in ecology without a priori controls: the use of synthetic controls in assessments of remotely-sensed land-treatment effects.

# Abstract

The basic data requirements for synthetic control include (1) a known set of treated and un-treated units, (2) a known date of treatment intervention, and (3) a timeseries including both pre- and post-treatment outcomes.

# 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 (Hurlbert, 2004; Oksanen, 2001, 2004). For instance, it may be challenging to quantify the impact of a conservation program, introduction of a species, or biogeochemical addition when the unit of analysis is large, complex or poorly defined, such as a lake, sub-population of a species or an entire community.

In a management context, effective decision-making requires inference from past 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 learning is often limited by lack of monitoring data and the means to distinguish treatment effects from other confounding influences. 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 impossible to discriminate true treatment effects from coincident noise. While some management efforts do integrate experimental elements such as replication, randomization or basic controls into their design, the logistical cost of such designs make them rare in applications 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, Meng, & Kendall, 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, Katikireddi, Leyland, & Popham, 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, regression discontinuity, and instrumental variables (CITES).

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, Diamond, & Hainmueller, 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, Diamond and Hainueller (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 to more recent models implementing latent interactive fixed-effects regression (Xu 2017) or Bayesian structural timeseries models (Broderson et al. 2015). In some sense, the most basic implementation of the synthetic control approach is the classic “Difference in Differences” method (hereafter DiD, XXXX), whereby the average difference between control and treatment are compared before and after the intervention ( Xu ? Craig?). As each model formulation carries its own set of assumptions and strictures (eg. tolerance of missing data, assumption of parallel trajectories through time, ability to extrapolate, etc), different methods will likely have advantages and disadvantages in various scenarios.

## Common Approaches

Previous use of synthetic control in the environmental sciences has predominantly focused on determining the effectiveness of broad economic and social policies or events on social-ecological systems such as deforestation in the Amazon (Sills et al., 2015) However, we propose that this technique may be useful more broadly with ecological data, particularly in cases where the units of analysis are large, complex and lack replication or pre-meditated controls. In this study we examine the utility of synthetic controls for analyzing ecological events with timeseries of remote sensing imagery – e.g. data that is temporally and spatially extensive but also noisy and prone to confounding. Typical approaches to 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), or (b) use differencing techniques (e.g. DiD) which may over-simplify the contextual information provided by controls. For instance, imperfect matching between controls and treatment areas may produce bias if the controls respond differently to the same confounding factor, such as grassland and forest responding differently to the same climate anomaly.

In this study we first evaluate the performance of different methods of assessing treatment effects (timeseries-only, DiD, and synthetic control) using a simulated satellite timeseries of a spectral index. We include various sources of random and systematic confounding noise and examine how the signal-to-noise ratio, number of controls, and ecological mismatch between control and treatment pixels influence the ability of each method to identify a simple treatment effect representing vegetative disturbance followed by recovery. We then demonstrate the use of synthetic control 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 Methods): (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, 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` (Broderson et. al 2015). Although DiD and synthetic control are similar, they are qualitatively different enough and we hereafter consider DiD separate from ‘synthetic control methods’. We used default values for all functions, implemented in R (R Development Core Team, 2015).

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 sin-wave trends with amplitudes of 0.4 and 0.1, respectively, and baseline NDVI values of 0.X or 0.y (Vesserbelt et al. 2010). The treatment effect was modeled as an abrupt reduction in NDVI (-0.1) such as from a disturbance, followed by a linear recovery over two years. 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).

Since we were interested assessing treatment effects in the presence of a variety of potential confounding factors, we added three additional sources of systematic noise to simulated timeseries (fig. error): 1) random drops of 0.25 NDVI, corresponding to cloud contamination or sensor error in a satellite image; 2) a growing-season climate anomaly; and 3) signal drift over time as from vegetative dynamics. 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 (Appendix). 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’ pixel timeseries 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 the how the number of controls would affect accuracy of different methods. These pixels received the same set of confounders (climatic, satellite and drift) but separate realizations of random noise.

Different landscape patches are unlikely to respond to exogenous influences (e.g. climate) in the same way. To account for different 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 also 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 are likely to depend on the degree of underlying similarity between a treated unit and its controls. To assess the effects of potential mismatch between control and treated pixels on the accuracy of different methods, we generated three different scenarios (Fig. EvaluationExample): 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 methods). We assessed error as the point-wise absolute difference between the ‘true’ treatment effect and estimated treatment effect in the post-treatment time period. For methods which provided confidence intervals we also assessed whether estimated treatment effect intervals overlapped zero or contained the true treatment effect at each time point. Details for each method are supplied in APPENDIX and simulation code is hosted at XXXX.

## 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 XXXX ha treatment area (details in Karl et al. 2014 and Gillan et al. 2016). There is some contention around the effectiveness of such treatments, as well as potential erosion risks from increased exposure of bare ground following treatment.

Within a designated section of the broader treated area, three types brush-clearing methods were applied in distinct landscape patches (fig Map): (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 (wri.utah.gov).

We assessed treatment effects based on the Soil Adjusted Total Vegetative Index (SATVI ; XXXX ), which has been shown to accurately reflect total vegetative cover for arid regions. We calculated SATVI as,

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 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 XXXX).

For each pixel in the target areas, we identified a set of 100 control pixels, adapting methods from Nauman et al (20XX). Briefly, within a search radius of 3 km surrounding the perimeter of the treated area, buffered by 90 m, we first performed a ‘masking’ operation, removing from consideration any pixels known to be part of another treatment, 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 ( +/- 3 EC units) and fine sand fraction soil texture class to the focal treated pixel (Nauman, Ely, Miller, & Duniway, 2019). From this subset, we selected the 100 most-similar pixels in the control pool, using Gower distance (van der Loo, 2019) based on a suite of topo-edaphic variables (Appendix Variables). We estimated treatment effects using the same methods outlined in the simulation model exercise, for each pixel, setting the treatment date as June 1, 2009.

# Results

## Simulations

In simulations, the relative accuracies for the four different methods of determining treatment effects (timeseries only, DiD, gsynth and CausalImpact) 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 utilized data from the controls were superior to the baseline estimates from the timeseries-only method (BFAST), regardless of the number of controls available (Fig. ResultPanelError, top row). This emphasizes the importance of using well-matched controls for data with responses dynamically influenced by multiple potential environmental factors.

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 ResultPanelError, 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 consistent 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 ResultPanelError, bottom row). Both the DiD and gsynth methods were less accurate than baseline, and the DiD method performed worse with larger number of controls, again due to the naïve aggregation of controls for comparison. This result highlights again the important role of finding accurate matches between control and treatment populations. While the CausalImpact method was able to determine the treatment effect if given enough control data in this context, it is unclear the degree to which such inference could be achieved with non-simulated, poorly matched data. In the simulation, poorly matched controls responded to the same confounders (i.e. seasonality, clouds, trends) as the treated pixel, only to a different magnitude. This might not be the case with real data where a mismatched land-cover type might have no response to a confounder (e.g. an impervious surface) compared to the treated pixel.

In most cases, increases in signal-to-noise ratio ( effect size / s.d. of random noise) led to marginal reductions in error (Figure ResultPanelError), particularly after signal magnitude reached 10 – 50 % of the average variation in noise. 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 AppendixErrorConfounder).

Confidence envelopes for treatment effects revealed differences between methods, which varied by level of noise and control-mismatch (Figure PanelCI). The CausalImpact method was the most conservative (low sensitivity), especially when the magnitude of confounding was high, but had high specificity (avoiding false positives) when the treatment effect was negligible. 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. The Synthetic control method had overall high sensitivity and low specificity, possibly resulting from violations in the assumptions of the parametric bootstrap standard errors procedure which resulted in erroneously small confidence intervals. Predictive confidence intervals of the DiD model depended on the control population, with more heterogeneous controls leading to wide confidence intervals and vice-versa.

## Case Study

For the brush-clearing case study, estimated treatment effects were similar among all methods which included controls (Fig shayDistrosOverall), providing greater discrimination among treatment types than either the raw SATVI scores or BFAST. The pixel-level estimates were heterogeneous within treatment areas (Fig Map, panel D; Fig. shayDistrosOverall), with some pixels having greater treatment effects than others. This can be visualized in certain regions of the mastication treatment (Fig Map panel D), where small stands of brush were left intact or in peripheral rocky areas with little brush to begin with. Accounting for this within-treatment heterogeneity is likely to be important for estimating net treatment effects, necessitating some kind of masking or stratification for improved estimates.

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. Karl), 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. 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 in a management setting.

# Discussion

* Importance of controls
* Importance of Matching
* Expectations for other data, improvements
* Method specifics
* Broader implications
* Choice of good matching controls important
  + In some cases algorithms may be able to learn from mismatched pixels, but only with sufficient data
* Simple methods may be sufficient for some types of data
  + Where variation is high and effects more subtle, synthetic control a good match
* Bfast a bit of a straw-man
  + Mostly used for change detection
  + Our purposeful inclusion of confounders was designed to throw this off.
* Gsynth issues with selection of R\*
  + But, can accommodate multiple treated pixels at once, treated at different times
* CausalImpact best of methods tried
  + Downside: no missing data
* Results still noisy when using every satellite image – a cloud-masking algorithm might improve point-wise estimates.
* Length of pre-treatment timeseries not investigated
* Overall benefits of Synthetic Control
* Applications in ecology

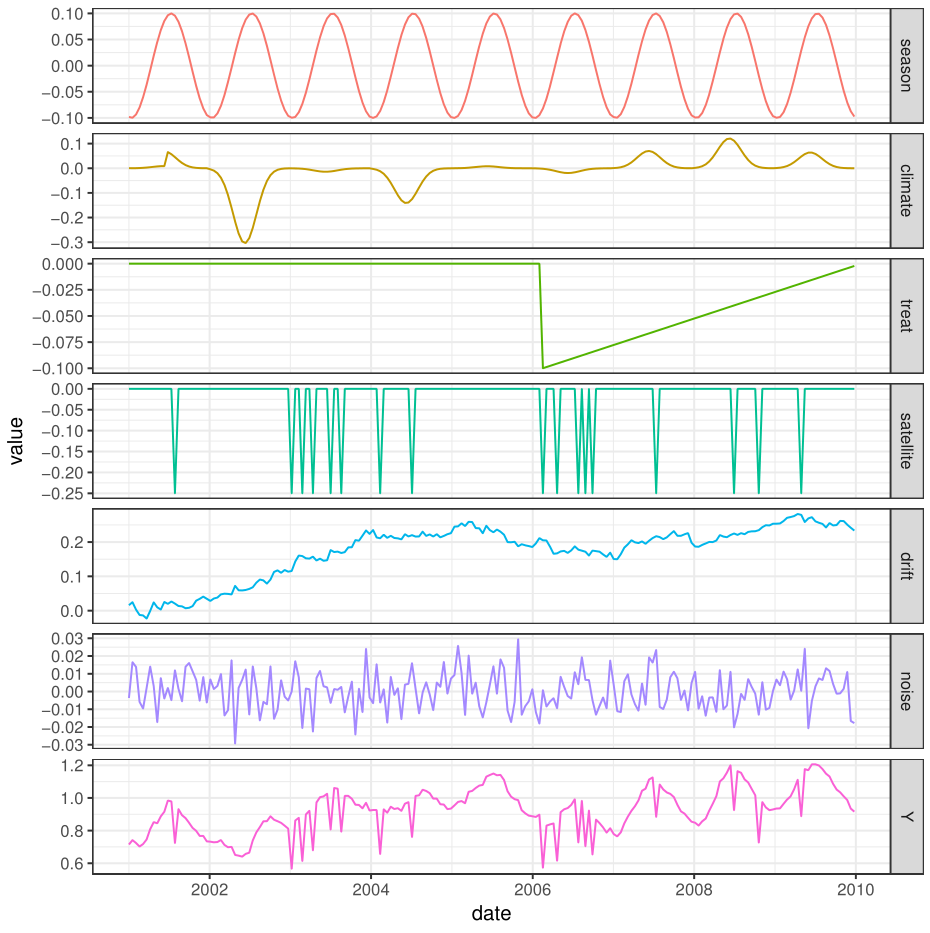
# Tables

#### Table Methods

|  |  |  |  |
| --- | --- | --- | --- |
| 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 from a linear model:  Yit = Xit + Ci + Lt + Eit |  |
| 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 | Broderson et al. 2018 |

# Figures

### Figure SimulationExample

Example of a simulated NDVI timeseries for a forest (Y) composed by adding various trends and sources of random noise.

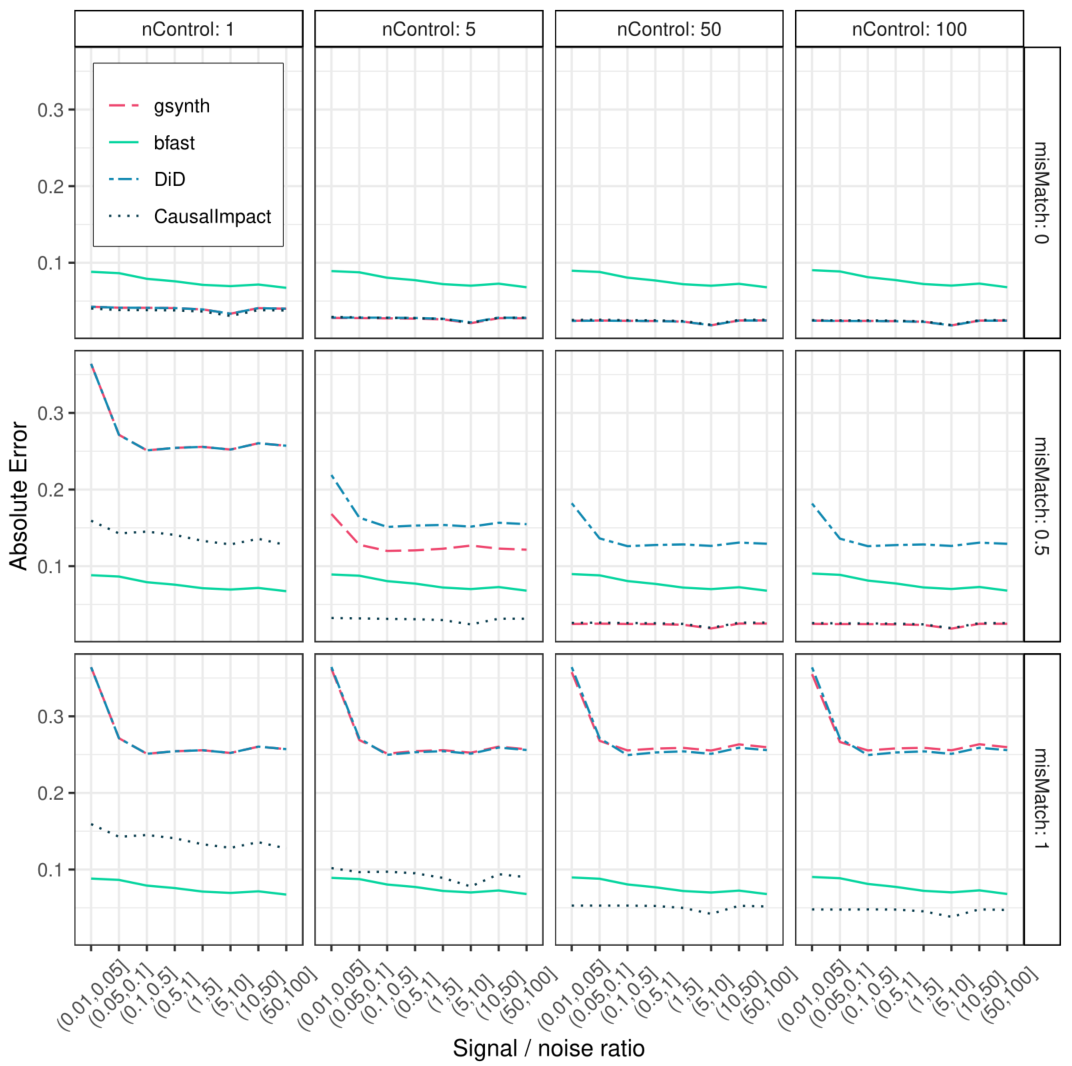
### Figure EvaluationExample

Example treatment effect estimates for different methods when controls are well matched (left column, mismatch = 0), and both 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 ResultPanelError

Simulation results for absolute error in estimated treatment effect as a function of signal-to-noise ratio ( treatment effect magnitude / sd random noise). Results broken down by number of controls available (columns) and degree of mismatch between the landscape type of controls and treated pixels ( 0 = no mismatch, 1 = total mismatch).



### Figure PanelCI

Power curves showing empirical frequency of concluding that an effect is different from zero, by method, level of mismatch between treated and reference pixels (rows), pointwise absolute magnitude of net confound (columns; season + climate + drift + satellite) and signal-to-noise ratio (x-axis). Showing data from simulations with greater than five controls.



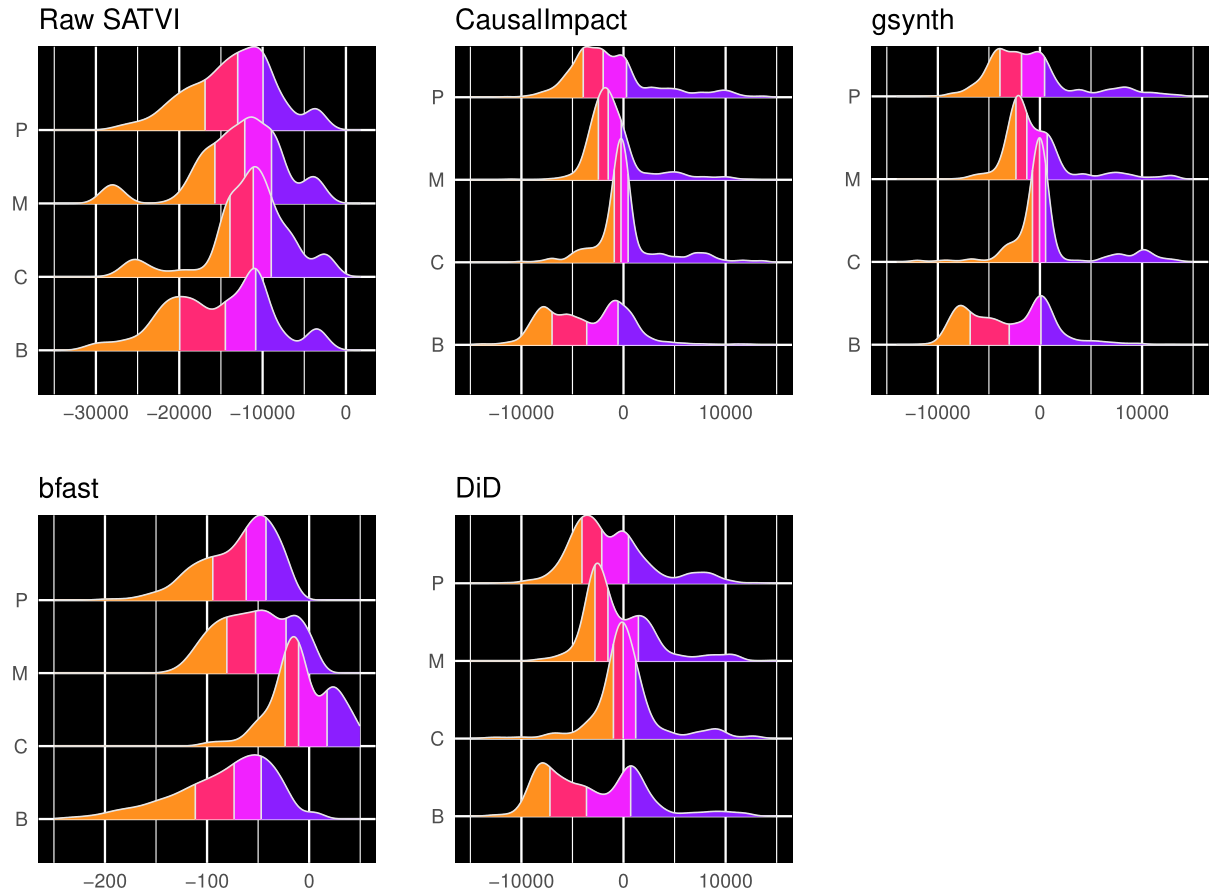
### Figure Map

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 following Nauman et al. (20XX). One hundred control pixels were selected for each treated pixel from a narrowed pool of candidates with similar topographic and edaphic properties. Panel D: estimated median per-pixel treatment effect for the 2010 growing season (Mar – Nov) in units of SATVI \* 1000. Bottom Right: Example timeseries analysis for the pixel in panel C, using the Bayesian structural timeseries in CausalImpact. 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. Panel G depicts cumulative treatment effects, analogous to exposure of bare ground integrated over time.



### Figure shayDistrosOverall

Effect of brush clearing treatments on treated areas, by assessment method. Distributions represent all estimated pixel-wise effects of treatments on SATVI (x axis = SATVI \* 1000) between March and November 2010, one year after implementation. Treatments include control (C), mastication (M), pile-burn (P) and broadcast burn (B).



### Figure Karl

Percent change in repeated line-point intercept cover values before and after treatments at Shay Mesa, from Karl et al. 2014. Treatments include controls (C), mastication (M), pile burn (P) and broadcast burn (B).

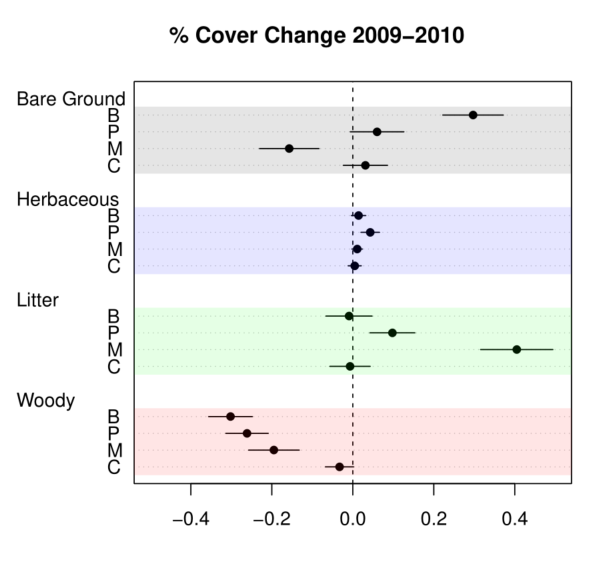
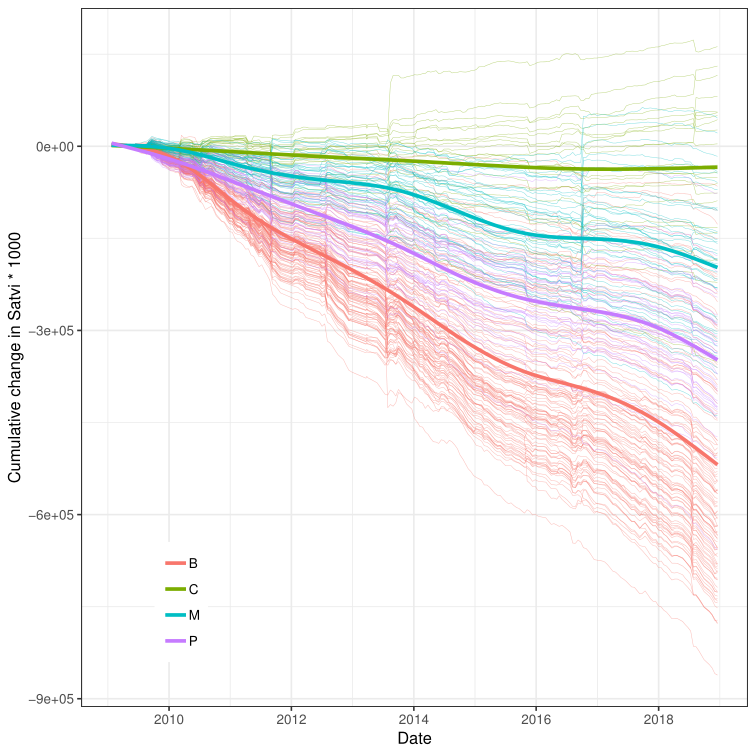


Figure Cumulative

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



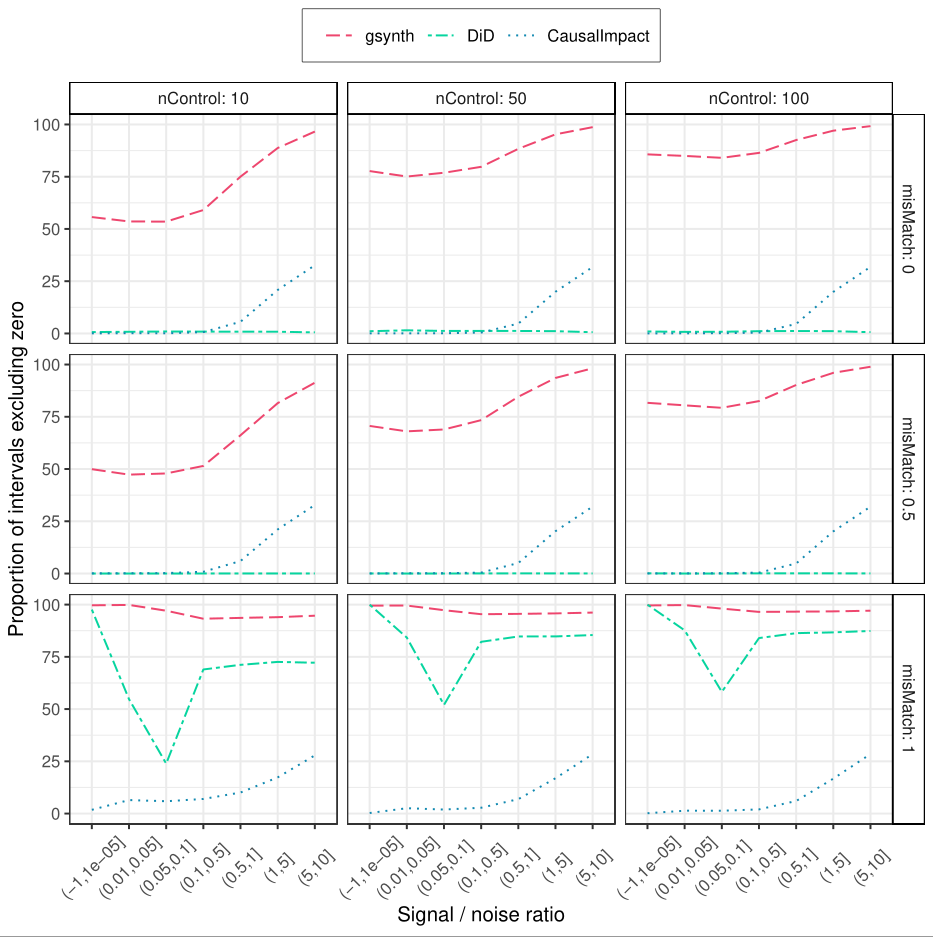
# Appendix

Serial Correlation

Generation of serial correlation in climate anomalies was accomplished by filtering a random normal series with the r function `filter`, using 1 lagged forecast error (argument `filter` = c(0,0,1)).

### Figure AppendixPanelSensitivity

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 AppendixErrorConfounder

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 Raster Data

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Function | Dataset(s) | Preparation Notes |
| SATVI | response | USGS Landsat 5 (years 1984-2011) , 7 (2012), and 8 (2013-2018) Tier 1 Surface Reflectance. |  |
| Roads | mask | TIGER ( ) | Interstates and major roads buffered to 60 m, local roads buffered to 30 m |
| Land Cover | mask | NLCD 2011 | Masked and buffered by 1 pixel (30 m) all water(11), snow (12), developed(21-24), and cultivated(81,82) pixels |
| Fires | mask | MTBS (), Utah BLM () | 30m buffer |
| Disturbances | mask | LandFire disturbance grids (LF 1.4.0) | Masked any pixel with non-zero disturbance value between 1999 and 2016 |
| Other Land Treatments | mask | Utah WRI, LTDL, NRCS traxx | Any treatment area marked as completed, 30m buffer |
| Elevation | Reference matching |  |  |
| Slope | Reference matching |  |  |
| Southness | Reference matching |  |  |
| Eastness | Reference matching |  |  |
| PCURV | Reference matching |  |  |
| TCURV | Reference matching |  |  |
| Relative Height | Reference matching |  | Used separate variables including neighborhoods of 1, 32, 128 pixels |
| RELMNHT | Reference matching |  | Used separate variables including neighborhoods of 1, 32, 128 pixels |
| MRRTF | Reference matching |  |  |
| MRVBF | Reference matching |  |  |
| Topographic Wetness Index | Reference matching |  |  |
| Calog\_10 | Reference matching |  |  |
| LFELEMS | Reference matching |  |  |
| Soil EC | Reference matching |  |  |

# Removed Text

* Methods with controls performed better than timeseries-only if controls were well matched with treated pixels
* When poorly matched pixels were included in the control population, increased data (n. Controls) improved the synthetic control estimates, but not DID
* When all control pixels were poorly matched, only Causal Impact was better than the timeseries-only approach, and only with sufficient controls.
* Causal impact confidence intervals most conservative, grew with larger confounder influence
* Why DID weird relative to signal to error

* Wide variability in a polygon through time
  + Some treatment effects clearly weaker than others within same area
  + All methods with controls better and similar
  + Higher precision for S.C. for example with the control distribution
  + Need to be careful with aggregations in space and time
* Order of SATVI responses matched cover responses
  + Did not pick up overall positive cover for mechanical treatment
    - Neither did SAVI or NDVI
  + Cumulative exposure of bare ground may be useful metric
  + Relative nature of Synthetic control may buffer against measurement errors for example in fractional cover estimation.