Methods in Ecology and Evolution

DR ASHLEY LARSEN (Orcid ID : 0000-0001-7491-9245)

Article type : Review

Handling editor: Professor Robert B. O'Hara

Causal Analysis in Control-Impact Ecological Studies with Observational Data

Ashley E. Larsen[a,*], Kyle Meng[a,b], Bruce E. Kendall[a]

Affiliations: [a]Bren School of Environmental Science & Management, University of California, Santa Barbara, [b]Department of Economics, University of California, Santa Barbara.

*Corresponding author, larsen@bren.ucsb.edu

Running headline: Causality in observational control-impact studies

Abstract

1: Randomized experiments have long been the gold standard in determining causal effects in ecological control-impact studies. However, it may be difficult to address many ecologically and policy-relevant control-impact questions-such as the effect of forest fragmentation or protected areas on biodiversity-through experimental manipulation due to scale, costs and ethical considerations. Yet, ecologists may still draw causal insights in observational control-impact settings by exploiting research designs that approximate the experimental ideal.

2: Here we review the challenges of making causal inference in non-experimental control-impact scenarios as well as a suite of statistical tools specifically designed to overcome such challenges. These tools are widely used in fields where experimental research is more limited (i.e., medicine, economics), and could be applied by ecologists across numerous sub-disciplines.

3: Using hypothetical examples, we discuss why bias is likely to plague observational control-impact studies in ways that do not surface with experimental manipulations, why bias is generally the barrier to causal inference, and different methods to overcome this bias.

4: Satellite-, survey- and citizen-science data hold great potential for advancing key questions in ecology that would otherwise be prohibitive to pursue experimentally. However, to harness such data to understand causal impacts of land, environmental and policy changes, we must expand our toolset such that we can improve inference and more confidently advance ecological understanding and science-informed policy.

Keywords: before-after-control-impact, causal analyses, econometrics

Introduction

The methodological gold standard in ecology, as in many scientific disciplines, is the randomized control trial, also known as the control-impact experiment. The random assignment of subjects (or sites) into treatment (or impact) and control groups with pre-determined treatment levels has been used to uncover innumerable fundamental findings in ecology. For example, common garden experiments seek to compare the effect of a fixed treatment (fertilizer supplements, fungal inoculations, predator exclusions, genetic strains) using two or more groups that are otherwise exposed to the same environmental conditions. Since we cannot observe the exact same site both treated and not treated simultaneously, we must compare between sites to identify the effect of treatment. The key to valid comparison is to assign treatments to sites at random. In such randomized experiments, only the treatment should differ systematically between treatment subjects and control subjects; this allows researchers to interpret the average difference between treatment and control groups as the average causal effect of treatment at the population-level.

Ecologists are increasingly interested in taking advantage of survey, remote-sensing and citizen-science data to address ecologically and policy-relevant questions in systems that do not easily lend themselves to experimental manipulation. For example, the placement of protected areas is rarely under the control of the researcher and they are generally not randomly placed on a landscape. In such cases, how can one identify the causal effect of protected areas on the abundance of, say, an economically or ecologically important species? To do so, the researcher must overcome the fundamental challenge present in non-experimental settings: the inability of researchers to have full control over treatment assignment (i.e. protected and not protected sites), which opens up the possibility that the outside forces that influence observed treatment are doing so in a non-random manner. Naively applying regression, anova or other statistical approaches without accounting for the

non-experimental nature of observational data can lead to inappropriate conclusions due to overlooked bias from improper comparisons between areas chosen and not chosen for treatment. In other words, the common mantra of "correlation does not imply causation" applies. However, not all is lost. Ecologists can establish causal inference with observational data in a control-impact framework if we incorporate careful research design and rigorous statistical approaches expressly designed for the purpose.

Here we discuss the challenge and promise of inferring causality from non-experimental data in control-impacts studies. We begin by discussing frameworks for causal inference. We then expand on the nature of why observational data present specific challenges not encountered in randomized experiments, which provides paths forward. To that end, we review several statistical approaches often associated with econometrics that can potentially overcome bias in control-impact analyses with observational data. To the extent possible, we seek to build intuition rather than to delve into the technical details. We use hypothetical examples to do so, since few real data sets are amenable to all methods discussed and the true population parameter is indiscernible in real data.

Frameworks of Causal Inference

In control-impact studies, causal inference is achieved through explicit comparison across units that are treated and units that serve as controls. In such settings, the key concept is that of a counterfactual: what would outcomes for the treated units look like in the absence of the treatment? If control units differ from treated units in the absence of the treatment, then a causal interpretation is not feasible.

There are several different frameworks for conceptualizing causal relationships in order to facilitate causal inference. Two of the most well-known are Pearl's structural causal model (SCM; Pearl 2000, 2010) and Rubin's potential outcomes model (PO; Rubin 2005).

SCM is a powerful framework for assessing causal relationships between variables. It integrates nonlinear structural equation modeling (SEM), graphical representation of causal pathways, and potential outcomes analysis (Pearl 2010). SEM, first developed in the early decades of the 20th century (Wright 1921), has been used in ecological systems to generate and test complex hypotheses about direct and indirect species interactions and system processes (Grace *et al.* 2010; Fan *et al.* 2016). SCM extended SEM to more flexible distributional assumptions, and links the equations embodied in the causal diagram (or directed acyclic graph, DAG) to the concept of counterfactuals.

In contrast, the PO framework is based on a notion of causality which places an emphasis on what researchers can and cannot observe, and an emphasis on isolating the effect of usually a single explanatory variable of interest (i.e. treatment variable) on a single outcome rather than on disentangling complex relationships within a network. Thus, PO is a particularly amenable framework for conceptualizing randomized and non-randomized control-impact studies. A specific insight illustrated by the PO framework is that causal interpretations are stymied by the fundamental truth that a subject cannot be both treated and not treated simultaneously (Holland 1986). As we will see below, randomization allows for the estimation of an average treatment effect in the population, while the absence of randomization requires additional understanding of the data-generating mechanism to develop a credible comparison. While part of the richness of SCM is the development of a new mathematical language describing causal relationships without reliance on probability math, it is not our goal to summarize this for ecologists. We point the interested reader to Pearl (2010). Our goals are to first illustrate why statistical bias presents a particular challenge for observational studies and then introduce some practical tools from econometrics to improve causal inference in observational control-impact studies. As such, we build on the potential outcomes framework as a simple way to relate to ecology's

foundations in randomized experiments. Nonetheless, bias can also be described using the mathematical and graphical representations of SCM, which we include in our illustrations. Finally, we emphasize that by employing the specific control-impact notion of causality, this review will not cover the notion of causality found in coupled dynamical systems, such as those pertaining to models of coupled predator-prey interactions. That notion of causality, sometimes referred to as "Granger" causality (Granger 1969) in time-series econometrics and recently advanced for nonlinear dynamical settings (Sugihara *et al.* 2012), examines how several interacting time series variables may be coupled over time. Because our aim is to inform research in control-impact studies, this review will exclude this dynamical notion of causality.

Potential Outcomes Framework

To be concrete, take for example, a study that is interested in estimating the effect of forest thinning (e.g. through US Forest Service Collaborative Forest Landscape Restoration Program) on songbird abundance. Which forest stands are chosen for thinning treatment is not under the manipulation of the researcher where treatment and control sites could be assigned randomly at precisely known levels of treatment. Rather, as is common with observational data, the decision of where to thin is likely determined by a tangle of possibly unknown or unobserved environmental (e.g. climate, soil), social (e.g. land values) and policy factors that cannot be manipulated by the researcher. As such, here and throughout we model the treatment as a random, rather than fixed, variable. The implication of this distinction will become clear later on (see *Treatment as a Random Variable* below).

The US Forest Service's priorities often include improving ecosystem function and reducing fire risk, and thus we can imagine that more degraded sites or sites closer to human habitation are more likely to be given resources than intact forests far from the Wildland-

Urban Interface. In that case, a survey of songbird abundance across thinned and unthinned sites is likely to find lower mean songbird abundance in thinned sites. A similar result might occur if there were different levels of thinning treatments based on proximity to surrounding development or fire risk. In both scenarios, we would be remiss to conclude that thinning reduces songbird abundance based on a simple comparison of means because sites chosen for treatment (or sites chosen for higher levels of treatment) differed systematically from those not chosen (or those chosen for lower levels of treatment). This systematic difference between the sites assigned treatment (or different treatment levels) results in inaccurate estimation of the effect of treatment. More formally, the mean or expected value of the estimated effect of the treatment, $E[\hat{\beta}]$, is different from the true value, β. This is known as statistical bias. The challenge is therefore overcoming bias stemming from non-random treatment assignment so we can isolate the effect of the treatment on bird abundance.

For simplicity we start by formalizing the above scenario with a binary treatment (thinned, unthinned forest stands). For any site, there are two outcomes that can potentially be observed—songbird abundance if the site was selected for the thinning treatment and songbird abundance if the site was not selected for the thinning treatment. Formally,

$$Potential\ outcome = \begin{cases} Y_{1i}\ if\ T_i = 1 \\ Y_{0i}\ if\ T_i = 0 \end{cases}, \#(1)$$

where Y_{0i} is songbird abundance in site *i* had that site not been chosen for treatment ($T_i = 0$), and Y_{1i} is songbird abundance in site *i* had it been chosen ($T_i = 1$)[1]. The observed outcome Y_i can be related to the potential outcomes by,

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})T_i. \#(2)$$

[1] The formal notation for potential outcomes was introduced by Neyman (1923, translated and reprinted in 1990) in the context of randomized experiments. It wasn't until the work of Rubin (1974) that the potential outcomes framework was considered for observational data settings. The term "Rubin Causal Model" first appears in Holland (1986).

The causal effect of thinning for site *i* is $Y_{1i} - Y_{0i}$. For many empirical applications, the question of interest, or estimand, is the population average treatment effect (ATE). Let *E[]* represent the expectation operator, or the population mean of a random variable. By the law of large numbers, the sample mean converges to the population mean so *E[]* can also be thought of as the sample average in very large samples. The ATE can be written as

$$\beta = E[Y_{1i}] - E[Y_{0i}] = \left(\frac{1}{N}\right)\sum_{i=1}^{N}(Y_{1i} - Y_{0i}) \#(3)$$

where N is the population size. β is the causal effect we would like to be able to estimate if it were possible to observe, for every site *i*, its outcome both when it is thinned (Y_{1i}) and when it is not thinned (Y_{0i}). Since this is impossible, we must learn about the effect of forest thinning through comparisons across untreated units that can serve as valid counterfactuals.

If we took the simple observed differences in mean songbird abundance between treated and untreated sites, we may capture more than we intended. The simple difference in means between sites that were and were not treated is equivalent to

$$E[Y_i|T_i = 1] - E[Y_i|T_i = 0]$$

$$= E[Y_{1i}|T_i = 1] - E[Y_{0i}|T_i = 0]$$

$$= \underbrace{E[Y_{1i}|T_i = 1] - E[Y_{0i}|T_i = 1]}_{\text{Average treatment effect on the treated (ATT)}} + \underbrace{E[Y_{0i}|T_i = 1] - E[Y_{0i}|T_i = 0]}_{\text{Selection bias}}. \#(4)$$

The first composite term on the right-hand side of equation (4) represents the average effect of treatment on sites that were thinned ("average treatment on the treated", ATT). The second term captures the systematic difference between sites that are and are not treated in the absence of treatment (e.g., if the thinning program was cancelled after site selection but before thinning occurred, would average bird abundance differ between selected and not selected sites?). Thus, the second term captures the "selection" bias stemming from non-random treatment assignment. Selection bias would arise if sites chosen for thinning were

less isolated or otherwise in less pristine condition than sites not chosen. In that situation the estimated effect of thinning would capture both the true effect of the thinning treatment on bird abundance and the pre-treatment difference in site quality. Quasi-experimental approaches including BACI designs seek to remove selection bias so we can isolate the causal effect of the treatment from observed differences in outcomes between treatment and control groups.

A key assumption

Regardless of whether treatment is randomly assigned, deriving causal inference based on counterfactuals invokes the assumption that there is no treatment spillover or interference between sites. This is known as the Stable Unit Treatment Value Assumption (SUTVA; Rubin 1980; 2005). SUTVA also assumes there are not different versions of the same treatment. This would be violated if, for example, some sites are only treated on paper, but action never happens on the ground.

SUTVA is required for potential outcomes to be well defined and is built into the potential outcomes definition in equation (1). However, one can envision conditions in ecological systems that violate SUTVA. For example, if population growth in a non-treated site is so high that there is net dispersal away from the site and into a treatment site, there would be treatment spillover, which would obfuscate the effect of the treatment alone. Treatment spillover would generally occur with spatial dependence between outcomes, where treatment of one site *caused* higher abundance at a nearby site. However, spatial correlation of the standard errors (a common feature of ecological data) would not violate SUTVA.

At first glance, SUTVA seems overly restrictive. However, studies can often be designed such that SUTVA is reasonable. For example, researchers can aggregate to larger units (e.g. individual to population, patch to landscape; Imbens & Wooldridge 2009). Lack of

interference between observations underlies many statistical analyses trying to ascertain treatment effects in randomized trials as well as observation studies. If one is to relax SUTVA, additional information is needed to specify the exact extent and intensity of interactions across individuals (e.g. Deschenes and Meng, 2018). This is an active area of research (e.g. Manski 2013).

Randomized Experiments

If we are willing to make the SUTVA, causal inference becomes a problem associated with assignment of treatment. If treatment status, T_i, is independent of potential outcomes as it theoretically would be in a random experiment, the second composite term of equation (4) drops out since $E[Y_{0i}|T_i = 0] = E[Y_{0i}|T_i = 1]$. Further, the conditional expectation simplifies to the unconditional expectation in the first term, $E[Y_{1i}|T_i = 1] - E[Y_{0i}|T_i = 1] = E[Y_{1i}] - E[Y_{0i}]$ because potential outcomes are independent of treatment status ($Y_{1i}, Y_{0i} \perp T_i$, where $\perp$ denotes statistical independence). Thus, the simple difference in population means, the left-hand side of equation (4), is equal to ATE, equation (3), if treatment status is randomly assigned. This highlights why experimental manipulations are the gold standard for causal inference. Replacing the population means with the corresponding sample analogs results in a consistent estimate of the ATE.

In observational analyses, we must remove selection bias associated with non-random assignment of treatment as bias precludes the identification of causal relationships. How we do so depends on what we know about how treatment is assigned and whether we can observe relevant covariates that determine treatment assignment. Below we transition from potential outcomes to regression, and from there to different regression-based methods for deriving causality for treatment selection based on observable and unobservable

characteristics. See SI for example code and table 1 for a summary of data requirements and key assumptions for each method.

Regression Analysis

Equation (2) can be rewritten in terms of a regression model. To build intuition in the most straightforward manner, we omit covariates for now. For simplicity, we also assume that treated sites respond the same way to thinning (i.e. constant treatment effects) and the model is linear in parameters. In this case, we can write equation (2) as,

$$Y_i = \alpha + \beta T_i + \varepsilon_i, \#(5)$$

where $\alpha = E[Y_{0i}]$, $\beta = Y_{1i} - Y_{0i}$ is the treatment effect, and ε_i is the site-specific random error term.

Evaluating equation (5) for treated and untreated sites yields,

$$E[Y_i|T_i = 1] - E[Y_i|T_i = 0] =$$

$$= (\alpha + \beta + E[\varepsilon_i|\mathrm{T}_i = 1]) - (\alpha + E[\varepsilon_i|\mathrm{T}_i = 0])\#(6)$$

$$= \beta + E[\varepsilon_i|\mathrm{T}_i = 1] - E[\varepsilon_i|\mathrm{T}_i = 0]\#(7)$$

This illustrates that the bias that prevents us from isolating the causal effect (β) from the simple difference in the treatment and control sites ($E[Y_i|T_i = 1] - E[Y_i|T_i = 0]$) stems from a correlation of the treatment with the error term. In other words, if the site-specific, random error term were not related to treatment status, $E[\varepsilon_i|T_i = 1] = E[\varepsilon_i|T_i = 0]$, the average treatment effect, β, is all that remains. Though we used the population regression for ease of illustration, by the law of large numbers, the sample regression coefficients are a consistent estimate of the population coefficients.

Treatment as a Random Variable

It is worth noting that throughout, we have been considering the treatment as a random, rather than a fixed, variable. This distinction, which is less essential in the context of randomized experiments, is the basis for why bias may arise in observational data settings.

In theory, a randomized experiment enables the researcher to fully manipulate which units are assigned to treatment or control, and for non-binary treatments, to determine the specific levels of treatment. The ability to fully manipulate treatment means that the researcher may be willing to assume, as Sokal & Rohlf (2012) describe in their seminal Biometry text (p 475), "the independent variable *X* is measured without error. We therefore say that the *X*'s are "fixed," which means that whereas the dependent variable *Y* is a random variable, *X* does not vary at random, but rather is under the control of the investigator". If *X* is assumed to be fixed, the correlation between the treatment variable and the error that we have been discussing at length is zero, by assumption[2]. This point is not often emphasized because in a perfectly executed randomized experiment, treatment (as a random variable) is uncorrelated with the errors anyway. Of course, in practice, assuming *X* is obtained without error may not hold due to naturally occurring variation, and randomization may not inherently provide bias-free estimates if randomization is incomplete (e.g. due to unknown individual variation in study units).

Yet, in observational data there is a clear distinction with regard to the treatment variable. By definition, treatment (e.g. location and extent of deforestation, protected areas, hunting pressure etc.) is determined by "outside" and potentially unknown forces that are beyond a researcher's control. Treating explanatory variables as random variables acknowledges the possibility of a correlation between the treatment variable and the

[2] Mathematically, this stems from the "exogeneity assumption" required for unbiased estimators. Exogeneity implies zero correlation between the treatment and the true model error, $E[T_i \varepsilon_i] = 0$. If treatment is considered fixed, it can be removed from the expectation such that $E[T_i \varepsilon_i] = T_i * E[\varepsilon_i]$. Since the latter term equals zero by assumption, assuming treatment is fixed implicitly assumes away any potential correlation between the explanatory variables and the error term, and thus the possibility of many forms of statistical bias.

unmodeled determinants of the outcome (i.e. model errors), and thus various sources of bias that preclude causal interpretations of correlations. We next discuss these sources of bias before turning to various research designs that potentially enable causal inference with observational data.

Sources of Bias

Bias implies that the expected value of the sample estimator does not reflect the true population parameter, $E[\hat{\beta}] \neq \beta$ (Fig. 1a). While the correlation between the hypothetical model errors and treatment ($E[T_i \varepsilon_i] \neq 0$) is broadly referred to as **endogeneity bias**, there are a couple of specific scenarios that are widely observed in observational studies.

Any covariate that is excluded from the model ends up in the error term. Thus, any variable that is correlated with the treatment and drives the outcome would result in a correlation between the errors and the treatment if not explicitly included in the model. For example, if forest stand age was correlated with the treatment (e.g. thinning) and bird abundance (e.g. through habitat availability), omitting forest age as a covariate would induce a correlation between the errors and the treatment and result in a biased estimator of the effect of thinning on bird abundance due to the **selection bias** problem illustrated earlier (which is also referred to as omitted variable bias and can be illustrated via a DAG, fig. 2). This contrasts with variables that drive the outcome but are not correlated with the treatment. Failing to control for these variables adds noise (i.e. increases the standard error of the parameter estimate) but does bias regression coefficients.

The second major source of endogeneity bias occurs when there is a feedback between the outcome variable back to explanatory variables, known as **reverse causality**. In other words, if thinned sites were chosen to avoid areas with high bird abundance, then abundance drives thinning and thinning drives abundance. In this case, it is impossible to

estimate either directional relationship without addressing the feedback because of the induced correlation between the errors and the treatment going in either direction (bird abundance → thinning, thinning → bird abundance).

Lastly, a persistent challenge for observational studies is the presence of **measurement error** in the explanatory variables. While measurement error of the outcome variable results in noise, it does not cause bias unless the measurement error is correlated with the explanatory variables. In contrast, measurement error in the explanatory variables causes what is known as Classical Errors-in-Variables, which biases the slope estimates towards zero.

Methodological Approaches

This section details five empirical approaches that, under different statistical assumptions, enable causal interpretations when examining observational data.

1. Difference-in-Difference (DiD): In the absence of experimental manipulation, it is difficult to parse apart the effect of the treatment from background changes in environmental conditions. Luckily, many survey data sources are collected over multiple years. When "panel" (or "longitudinal") data are available, the analyst can sometimes leverage repeated observations over time to address bias due to omitted, time invariant confounders.

Like BACI paired (Stewart-Oaten et al. 1986), DiD is a paired design where treatment and control sites are observed at the same time before and after the treatment occurs (Angrist & Pischke 2009). We introduce the basic DiD despite its similarities to BACI to introduce readers to another methodological literature and as an entryway to the panel data models discussed below.

With repeated observations of the same groups over time a DiD is estimated using the below model,

$$Y_{igt} = \alpha + \delta_1 treat_g + \delta_2 after_t + \beta\left(treat_g * after_t\right) + \varepsilon_{igt} \#(8)$$

where *i* denotes an individual observation, *g* denotes group, and *t* denotes the time period. Here "treat" is a dummy variable that is equal to one for sites that eventually received treatment (treatment group) and "after" is a dummy variable that is equal to one "after" the treatment occurs. By conditioning on these dummy variables in an ordinary least squares (OLS) framework, the average differences between treatment and control (before treatment) and average differences between pre-treatment control sites and post-treatment control sites are removed. Thus, the coefficient on the interaction term, β, indicates the change in outcome due to the treatment after differencing away persistent difference between groups and shared time trends. Normality of the errors is not required for OLS to be unbiased. While the basic model could be estimated with a repeated measure ANOVA if normality of the errors is assumed, a regression approach is advantageous with complex models, missing or unbalanced data, and when assuming normality or homoscedasticity of the errors is overly restrictive.

The simplest setup is when outcomes are observed in two periods for both groups where one group's treatment status changes from the first period to the next. However, the fundamental assumption of DiD (and other BACI designs) is that if not for the treatment, the two groups would have parallel time trends (Angrist & Pischke 2009). As an indirect test of this assumption, one can see if there are common time trends across groups before the treatment by using additional pre-treatment time periods, when available. DiD can be extended to include covariates, different timing of treatment ("staggered" DiD) and an additional control group ("triple difference").

2. Within-estimator Panel Data Model: The within-estimator panel data model is a generalization of DiD models to multiple groups and time periods.

Let us say we are again interested in song bird abundance, but this time as a function of forest fragmentation. With repeated observation of the same sites over time, we can exploit year-to-year deviations from the mean forest fragmentation of a site to estimate how fragmentation affects bird abundance, under certain conditions, even if we do not have measurements of all the covariates.

The within-estimator (also called the least-squares dummy variable model) is often and confusingly termed a "fixed effects" panel data model, but we continue with "within-estimator" to avoid confusion with "fixed effects", as defined in biostatistics (i.e. a non-random variable). The within-estimator model could be represented as follows,

$$Y_{it} = \alpha + \beta Fragmentation_{it} + c_i + \gamma_t + \varepsilon_{it} \#(9)$$

where Y_{it} indicates bird abundance in site *i* and time *t*, α is the intercept, β is the coefficient of interest, and ε_{it} is the random error term. As elsewhere in this manuscript, we ignore covariates for notational convenience.

Here c_i represents unobserved heterogeneity that is unique to each site *i* but time invariant over the study period (e.g. climate, soil quality) and γ_t represents unobserved heterogeneity that is unique to each year (e.g. weather, technology) that is shared by all sites. If either c_i or γ_t is ignored, it ends up in the error term, potentially creating endogeneity as described above. Ecologists are familiar with using site or year random effects in mixed effects models. Random effects models, such as random intercept models, assume that the unobserved site- or year-specific heterogeneity is uncorrelated with the treatment (Wooldridge 2002). In many cases this is a strong assumption. For example, climate, soil quality, proximity to urban centers are all likely to be correlated with fragmentation. If these variables were measured and included directly, there would be no issue. However, if they are

not, a site random effect would not avoid omitted variable bias because, although the correlation of observations at the same site is modeled, the correlation between covariate (fragmentation) and the error term is not removed. Instead, the within-estimator can be used. The effect of the within-estimator is that observations are differenced from their site-specific mean and thus identified by "within" site (or year) variation. If the site-specific (time-specific) unobserved heterogeneity is correlated with fragmentation does not matter because it is effectively removed from the model in the differencing. In the case where the site-specific (time-specific) heterogeneity was indeed uncorrelated with the covariates (the random effects assumption), the within-estimator would remain unbiased but would be less statistically efficient, or in other words have a larger variance, than the random effects estimator (Fig. 1). However, if the site-specific (time-specific) heterogeneity was correlated with the observed covariates, only the within-estimator model would remain unbiased. Though we only discuss site and year above, the same logic and applies to other group characteristics as well. We point the reader to Larsen & Noack (2017) for an example of using the within-estimator to understand how crop diversity affects agricultural pesticide use, after controlling for year-specific, crop-specific and region-specific unobserved heterogeneity.

3. Instrumental Variables: The within-estimator requires panel data and generally does not solve reverse causality bias (Table S1; for an exception see Larsen *et al.* 2014). However, the instrumental variables (IV) approach can jointly solve selection bias, measurement error, and reverse causality, provided certain assumptions are met. To isolate causal effects of a treatment on an outcome, the IV approach requires the researcher to select an "instrument" that (1) is sufficiently correlated with the endogenous treatment variable and (2) does not affect other determinants of the outcome (i.e. does not belong in the main regression). These two assumptions ensure that the variation in the treatment variable driven by the instrumental

variable is also uncorrelated with other determinants of the outcome, thus removing the source of endogeneity bias.

As an illustration of how IV works, consider predator-prey relationships which are classic examples of reverse causality as predator abundance drives prey abundance, but the reverse is also true (Kendall 2015). If we were, for example, interested in estimating the effect of wolf abundance on moose abundance using a linear regression, our linear coefficients may instead capture the reverse effect. To estimate the effect of wolf on moose abundance, we need to sever the reverse causality pathway by isolating a driver of wolf abundance that has no direct effect on moose abundance. One possible instrument would be the prevalence of canine distemper, which drives wolf abundance, but should not affect moose abundance (except through changes in wolf abundance). Note, we are assuming here that this predator-prey system is not closely coupled. If it were closely coupled such that there were offset boom-and-bust cycles, our estimates of the causal effect using cross-sectional data at any point in time would fail to capture the cyclical nature of the relationship (e.g. Sugihara *et al*. 2012).

Turning to how an IV approach would work in this setting, we can use the exogenous change in wolf abundance due to canine distemper to estimate the effect of wolf abundance on moose abundance. Conceptually, an IV approach occurs over a two-stage regression process. The first stage regression relates canine distemper prevalence to wolf abundance via,

$$PredAbundance_i = \delta + \gamma Distemper_i + u_i . \#(10)$$

In the second stage regression, moose abundance is then regressed on the wolf abundance predicted by canine distemper from the first stage,

$$PreyAbundance_i = \alpha + \beta \widehat{PredAbundance_i} + \varepsilon_i \#(11)$$

$$= \alpha + \beta\left(\hat{\delta} + \hat{\gamma} Distemper_i\right) + \varepsilon_i \, . \#(12)$$

As equations 10-12 show, the variation in wolf abundance used to estimate the effect on moose abundance comes only from canine distemper. Provided that canine distemper is not correlated with other drivers of moose abundance, contained in the error term ε_i, then an IV model estimates a causal effect.

In practice, the IV approach entails two further details. First, IV is usually implemented with two-stage least squares, where equations 10 and 11 are jointly estimated. This is to account for sampling variability in the predicted endogenous variable. Second, as a diagnostic of whether the instrumental variable is strongly correlated with the endogenous variable, one often examines variants of the F-statistic from the first-stage regression in equation 10. Such tests reveal whether there is a "weak instrument" problem, the presence of which introduces a bias in the IV estimate that can be as large as the endogeneity bias in the initial linear regression model (Bound *et al.* 1995). For a more in-depth discussion of IV in an ecological context, we direct the reader to Kendall (2015). For an ecological application which uses the IV approach to the effect of forest fragmentation on Lyme disease incidence, we direct the reader to MacDonald *et al.* (2018).

4. Regression Discontinuity: In some settings, the assignment of treatment may depend on an arbitrary rule arising from policy or institutional features. Modifying our earlier land-use example, let's say forest stands were eligible for thinning if they were within 15 km of at least one developed area and were at least 3 ha in size. As is often the case with such cutoff rules, both the 15 km distance and 3 ha size criteria may have been arbitrarily specified by some policy. However, it may not be desirable to implement a difference-in-difference method if finding control units that satisfy these criteria requires a researcher to expand the data setting into places that are unlikely to be similar. For example, a forest stand in Minnesota is unlikely to be a valid control for a forest parcel in California even if both have

the same distance to a developed area and size. Similarly, using instrumental variables may not be feasible in some cases due to a lack of a satisfactory instrument.

In such settings, a researcher may exploit the arbitrary nature of the cutoff rule. Here, one can try to compare stands above 3 ha in size that are just less than 15 km from a developed area (treatment) with similarly sized stands that are just more than 15 km from a developed area (control). Alternatively, for all parcels that are less than 15 km from a developed area, one can compare stands that are just above 3 ha in size (treatment) with those that are just below 3 ha (control). Such comparisons implement the regression discontinuity (RD) design. Specifically, the RD method exploits a discontinuity in treatment assignment around some threshold value of a "forcing" variable, which in our example would be either distance to a developed area or parcel size.

The key statistical assumption for the RD method to be valid is that only the probability of receiving the treatment jumps discontinuously as the forcing variable crosses the threshold. All other factors that determine the outcome must be continuous around the threshold. That is, going back to our example, only thinning eligibility changes at the 15 km distance threshold so that any outcome differences across the threshold can be attributed solely to thinning eligibility. Under these conditions, the RD method estimates the local average treatment effect only for the subpopulation close to the threshold. In practice, this means that the RD method is very data demanding, and requires a sufficient density of observations within narrow bandwidths around the threshold of the forcing variable. Interested readers can learn more about this issue and many other RD implementation considerations in Lee and Lemieux (2010).

5. Propensity score. Finally, in some settings, it may be argued that a researcher can observe all known determinants of an outcome that is correlated with the treatment of interest. In that case, known as "selection on observables", simply controlling for those covariates in a

standard regression setting would enable a causal interpretation. However, for many ecosystems, the list of covariates may number in the hundreds, with possible combinations of covariates observed for a treated unit not appearing for a control unit.

Propensity scores avoid this high-dimensionality problem by matching or weighting the probability that a site receives treatment based on a function of observable characteristics. The propensity score is the probability a site receives treatment given its baseline characteristics, $p(X_i) = Pr(T_i = 1| X_i)$ where $0 < p(X_i) < 1$. It follows from the treatment ignorability assumption that $T_i \perp (Y_{0i}, Y_{1i})|p(X_i)$ (Rosenbaum & Rubin 1983). Thus, conditional on the propensity score, treatment is independent of potential outcomes. Rosenbaum & Rubin (1983) also show that treatment and control observations with the same value of the propensity score balance in the distribution of baseline characteristics.

Propensity scores are estimated using a regression model for binary outcome variables (e.g. logit or probit) where probability of treatment is estimated as a function of baseline characteristics with highly flexible functional form. The specification should balance the distribution of baseline characteristics across the distribution of propensity scores.

There are several ways propensity scores can be used including matching on propensity scores, inverse probability weighting the estimator, using propensity scores in a weighted regression, and using propensity scores as a covariate adjustment in linear regressions. A thorough discussion of different methods can be found elsewhere (Austin 2011). We simulate propensity score matching and propensity scores as a covariate adjustment in a linear regression (SI), and point the reader to Pearson *et al.* (2016) for an ecological application focused on agricultural land cover and aquatic ecosystem impacts.

Discussion

A multitude of environmental and ecological challenges facing natural systems in the coming decades can be informed by observational data. Leveraging the data-rich landscape of the twenty-first century for impact studies necessitates incorporating statistical tools specifically developed for disentangling causal relationships in the absence of randomized experiments. Here we discussed how observational data differ from experimental data, why this difference is of crucial statistical importance, and introduced some assumptions and approaches that can be used to recover a causal interpretation of treatment effects in the absence of randomly assigned treatment.

In particular, we emphasized the fundamental importance of zero correlation between the covariate of interest and a model's error term. The presence of such a correlation leads to what is known as endogeneity bias and thus, incorrect coefficient estimates. Though we avoided discussing specific estimation methods, all common regression methods (ordinary least squares, maximum likelihood, generalized least squares, etc.) will generally produce biased estimates of the causal effect in the presence of endogeneity bias.

The symptoms of endogeneity bias can present as spatial or temporal autocorrelation in the residuals. However, if autocorrelation is due to omitted variables that are spatially or temporally correlated (e.g. climate, soil quality) and correlated with the treatment variable, methods that only adjust for autocorrelation of the errors will fail to produce unbiased slope estimates for the treatment of interest. Similarly adding random effects of site or year may not reduce bias. If site characteristics are correlated with the covariate of interest, random effects estimators will remain biased. Rather, recognizing and applying methods to overcome the underlying source of endogeneity bias are fundamental to reliable point estimates.

This paper's main contribution is to provide basic intuition for developing causal inference using observational data for different types of control-impact analyses. We necessarily could not provide a full treatment of such approaches, nor comprehensive treatment of causality in all observational settings. For instance, our maintained assumption throughout this manuscript that a random sample could be drawn from the population (at least in the cross-section dimension; Wooldridge 2002), extends to more complicated sampling designs such as stratified or clustered sampling (Wooldridge 2002). Further, we ignored concerns regarding the efficiency of estimators. Lastly, our focus on control-impact analyses does not include all notions of causality relevant to ecologists. In particular, while many of the methods discussed can be extended to nonlinear models where the marginal effect of the treatment variable is not constant over its entire range (e.g. logistic regressions), we excluded discussion of dynamic notions of causality involving coupled variables (e.g. Granger 1969; Sugihara *et al.* 2012). For coupled systems such as coupled predator-prey cycles, the methods discussed here would misspecify the nature of relationship as such systems cycle among positive, negative and neutral correlation between predator and prey. As observational data expand to provide sufficiently expansive species-specific time series observations, dynamic forms of causality will become increasingly relevant.

Nevertheless, many global environmental challenges of today and tomorrow will take the form of control-impact studies, where treatment evaluation is of primary interest. It is for those questions that a focus on unbiased statistical estimates of the treatment effect will be invaluable for addressing important ecological questions. Though we relied on hypothetical examples to streamline discussion, these methods discussed herein are not entirely new to ecologists. We point the reader to Gross & Rosenheim (2011), Bonds *et al.* (2012), Larsen (2013), Larsen and Noack (2017), and MacDonald *et al.* (2018) for empirical ecological studies using these methods, to Kendall (2015) and Butsic *et al.* (2017) for additional

methodological discussion aimed at the ecology audience, and to Wooldridge (2002) or Angrist & Pischke (2009) for advanced and introductory texts, respectively, on econometric methods. Ecologists have a strong tradition of causal inference in experimental research. Here we encourage a similarly strong interest in causality in observational control-impact studies such that we can better leverage novel data sources to inform ecological understanding and environmental policy.

Acknowledgements: This contribution was improved from discussions with C. Kremen, A. Liu, A. MacDonald, & M. Wilber, as well as from comments by the editor and three anonymous reviewers. Any errors are our own.

Data Accessibility: No data were used in this publication.

Author contributions: AEL conceived of the study, AEL & BK wrote simulation code, AEL, KM & BK wrote the manuscript.

References

Angrist, J.D. (1998). Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica* **66**, 249–288.

Angrist, J.D. & Pischke, J.S. (2009). Mostly harmless econometrics. Princeton University Press, Princeton, New Jersey.

Austin, P.C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, **46**, 399–424.

Bonds, M.H., Dobson, A.P. & Keenan, D.C. (2012). Disease ecology, biodiversity, and the latitudinal gradient in income. *Plos Biology*, **10**, e1001456.

Bound, J., Jaeger, D.A. & Baker, R.M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, **90**, 443–450.

Butsic, V., Lewis, D.J., Radeloff, V.C., Baumann, M. & Kuemmerle, T. (2017). Quasi-experimental methods enable stronger inferences from observational data in ecology. *Basic and Applied Ecology*, **19**, 1–10.

Deschenes, O. & Meng, K.C. (2018). "Quasi-experimental methods in environmental economics: Opportunities and challenges," Handbook of Environmental Economics, Vol3., forthcoming.

Grace, J.B., Anderson, T.M., Olff, H. & Scheiner, S.M. (2010). On the specification of structural equation models for ecological systems. *Ecological Monographs*, **80**, 67–87.

Granger, C.W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. Econometrica **37**, 424-438.

Gross, K. & Rosenheim, J.A. (2011). Quantifying secondary pest outbreaks in cotton and their monetary cost with causal-inference statistics. *Ecological Applications*, **21**, 2770–2780.

Holland, P.W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, **81**, 945–960.

Imbens, G.W. & Wooldridge, J.M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* **47**, 5–86.

Kendall, B.E. (2015). A statistical symphony: Instrumental variables reveal causality and control measurement error. Pp. 149-167 in G.A. Fox, S. Negrete-Yankelevich, and V.J. Sosa, eds., *Ecological Statistics: Contemporary Theory and Application*. Oxford University Press, Oxford, UK.

Larsen, A.E. (2013). Agricultural landscape simplification does not consistently drive insecticide use. *Proceedings of the National Academy of Sciences*, **110**, 15330–15335.

Larsen, A.E. & Noack, F. (2017). Identifying the landscape drivers of agricultural insecticide use leveraging evidence from 100,000 fields. *Proceedings of the National Academy of Sciences*, **114**, 5473–5478.

Larsen, A.E., MacDonald, A.J. & Plantinga, A.J. (2014). Lyme disease risk influences human settlement in the wildland-urban interface: Evidence from a longitudinal analysis of counties in the northeastern United States. *American Journal of Tropical Medicine and Hygiene*, **91**, 747–755.

Lee, D.S. & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, **48**, 281–355.

MacDonald, A.J., Larsen, A.E., & Plantinga, A.J. (2018). Missing the people for the trees: Identifying coupled natural-human system feedbacks driving the ecology of Lyme disease. *Journal of Applied Ecology*, **17**, 267–11.

Manski, C.F. (2013). Identification of treatment response with social interactions. *The Econometrics Journal*, **16**, S1–S23.

Neyman, J. (1923, 1990) "On the application of probability theory to agricultrual experiments. Essay on principles. Section 9," translated in *Statistical Science*, (with discussion), **5**, 465-480.

Pearl, J. (2000). Causality: Models, Reasoning, and Inference. Cambridge University Press, New York.

Pearl, J. (2010). An introduction to causal inference. *The international journal of biostatistics*, **6**, 1-59.

Pearson, C.E., Ormerod, S.J., Symondson, W.O.C., & Vaughan, I.P. (2016). Resolving large-scale pressures on species and ecosystems: propensity modelling identified agricultural effects on streams. *Journal of Applied Ecology*, **53**, 408-417.

Rosenbaum, P.R. & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrica*, **70**,41-55.

Rubin, D.B. (1974). Estimating causal effects of treatment in randomized and non-randomized studies. *Journal of Educational Psychology*, **66**, 688-701.

Rubin, D.B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, **75**, 591–593.

Rubin, D.B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association*, **100**, 322–331.

Stewart-Oaten, A., Murdoch, W.W., & Parker, K.R. (1986). Environmental impact assessment: 'Pseudoreplication' in time? *Ecology*, **67**, 929–940.

Sokal, R.R. & Rohlf, F.J. (2012). Biometry, 4th edn. W.H. Freeman and Company, New York, New York.

Sugihara, G., May, R., Ye, H., Hsieh, C. H., Deyle, E., Fogarty, M., & Munch, S. (2012). Detecting causality in complex ecosystems. *Science*, **338**, 496–500.

Van Butsic, Lewis, D.J., Radeloff, V.C., Baumann, M., & Kuemmerle, T. (2017). Quasi-experimental methods enable stronger inferences from observational data in ecology. *Basic and Applied Ecology*, **19**, 1–10.

Wooldridge, J.M. (2002). Econometric Analysis of Cross Section and Panel Data, 1st edn. MIT Press, Cambridge, Massachusetts.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, **20**, 557–585.

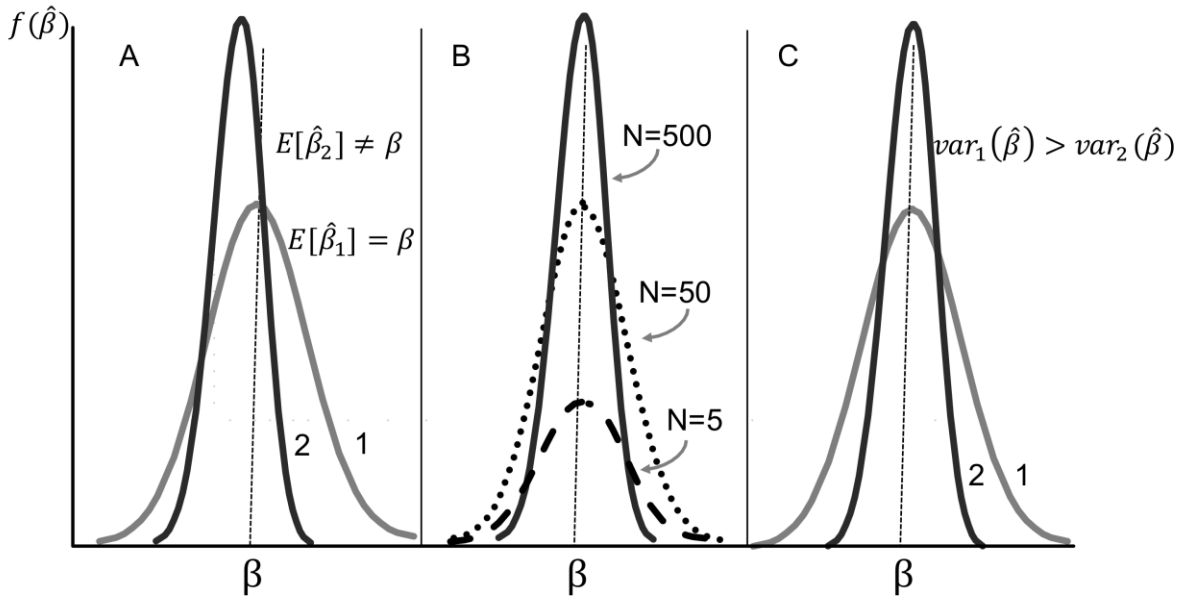
Table 1. Data requirements and key assumptions of different methodology discussed.

Method	Addresses	Situation	Data requirements	Key Assumptions
Difference-in-difference	Selection bias stemming from which group gets treatment.	Time trends and group specific averages differ between treatment and control groups.	At least two periods of data, before and after, observed for both a treatment and control group.	Parallel time trends between the treatment and control group prior to treatment.
Within-estimator	Selection bias stemming from unobserved or not included variables that are correlated with the covariate of interest and the outcome.	Time shocks shared by all observations (time dummies), time-invariant characteristics unique to individual observations or groups (individual, group dummies)	Panel data where covariates of interest and outcome variable vary over time and/or within individuals (i.e. within the dummy variable group(s)).	Strict exogeneity.
Instrumental Variables	Reverse causality. Can also be used to address other endogeneity bias.	There exists a feedback between the magnitude of outcome variable and the treatment variable	Requires an "instrumental" variable that is correlated with the endogeneously determined treatment variable, but otherwise does not drive the outcome.	Instrument is "relevant" (i.e. correlated with endogeneous variable) and uncorrelated with the errors.
Propensity Scores	Selection bias, if selection is determined by observable characteristics.	Reduces the high dimensionality problem associated with including all variables that could determine treatment vs control status.	Data on variables that determine selection into treatment and control groups.	Treatment ignorability assumption. Common support between treatment and control groups. Additional assumptions depending on how p-scores are used.
Regression Discontinuity	Selection bias	Discrete treatment assignment as a function of some threshold in a "forcing" variable.	Because treatment is assumed to be as good as random only near the threshold, there needs to be sufficient mass of data within narrow bandwidths of the forcing variable on either side of the threshold.	Assignment of treatment is as good as random across the threshold of the forcing variable. Units are unable to sort across the threshold.

Figure Legends

Figure 1. Properties of Linear Estimators. The desirable properties of linear estimators are that the estimator is unbiased (A,1), consistent (B) and efficient (C). Unbiasedness is a finite sample property. An estimator is unbiased, if the average (or expected value) of the sampling distribution is equal to the true parameter value (B, gray line). If there is a correlation between the model errors and treatment variables, the estimator will generally be biased (A,2). Consistency, like unbiasedness, is related to identification of the true relationship (i.e. the frequency distribution of estimated coefficients is centered on the true value, β) . However, consistency is an asymptotic property. We focus on unbiasedness, which is most relevant to finite samples, however, instrumental variables, due to its two step process, is a consistent but biased estimator. Efficiency is related to the spread of the distribution of the estimator. An efficient estimator has the minimum variance of all estimators in its class of estimators (e.g. linear estimators).

Figure 2. Causal diagram or Directed Acyclic Graph. Nodes represent variables, arrows represent possible causal effects in the direction of the arrow (a drives b, a $\rightarrow$ b), bi-directional arcs represent possible confounding relationships, and solid and dashed lines represent observed and unobserved variables, respectively. Importantly, causal assumptions are represented by the lack of connections, thus (A) assumes model 1 is correct, that there is no omitted variable confounding the estimate of the causal effect of thinning. If there was and it was unobserved (B), estimating model 1 would produce biased estimates of the effect of thinning on bird abundance due to the correlation between the errors (which include the unobserved confounding variable) and the treatment. If the researcher knew and could measure the confounding variable (C), the researcher could find unbiased estimate for the effect of thinning on bird abundance by modeling it explicitly; estimating model 2 rather than model 1.



Model 1: $Birds_i = \alpha + \beta Thinning_i + \varepsilon_i$

Model 2: $Birds_i = \alpha + \beta Thinning_i + \gamma Isolation_i + \varepsilon_i$