Causal inference with observational data in a confounded world

**I prefer -** Coping with confounding variables for causal inferences in observational data.

***My opinion We should really redo all the figures as DAGs. Its confusing to have them as path diagrams and can further the conclusion about causal diagrams vs path models which arent equivalent.***

Jarrett E. K. Byrnes1 and Laura E. Dee2

1 - Department of Biology, University of Massachusetts Boston, Boston, MA 02125

2 - Department of Ecology and Evolutionary Biology, University of Colorado Boulder, Boulder, CO 80308-0334

**Code Repository:** <https://github.com/jebyrnes/ovb_yeah_you_know_me>

**App for one simulated dataset:** <https://shiny.umb.edu/shiny/users/jarrett.byrnes/shiny_ovb/>

**App for replicate simulations:** <https://shiny.umb.edu/shiny/users/jarrett.byrnes/ovb_sims/>

**Keywords:** omitted variable bias, econometrics, observational data, causality, correlation, panel regression

**Suggested reviewers:**

* \*Van Butsic
* \*
* Suchinta Arif
* Kathryn Cottingham
* Christoper Trisos
* Andrew Bell
* Joan Dudney
* Brian Inouye

**Abstract**

As ecology tackles progressively larger problems, we have begun to move beyond the scale at which we can conduct experiments to derive causal inferences. Experiments have long been seen as the gold standard for quantifying causal effects in ecological systems. Large-scale observational data, in contrast, has largely been seen to either provide a place to explore ideas derived from experiments or a source of interesting patterns to inspire randomized controlled experiments This avoidance of using observational data for causal conclusions arises from the valid fear of results of unmeasured confounding variables in observational analyses. Unmeasured confounders that influence both the causal variable of interest and effect can bias conclusions - a problem known as Omitted Variable Bias. This phenomenon is what leads to the old saying, “Correlation is not causation.” Many other scientific disciplines, however, cannot do experiments for reasons of ethics or feasibility; they have developed robust approaches for causal inference from observational data. Here we show how Ecologists can harness these tools to derive causal inference from observational data. We begin with the cornerstone of causal inference: using causal diagrams to identify potential known and unknown sources of confounding. We then use a motivating example of assessing the effects of warming on intertidal snails to discuss how ecologists currently handle observational survey data and inference - often incorrectly. We present alternative sampling designs and the statistical designs that make use of them, discuss how they work using the language of causal path diagrams, demonstrate how easily they can be applied to common ecological datasets, and finally how well they are able to overcome problems of unmeasured confounding variables. We present these tools in the hopes that others will pick them up and begin to apply them to their own science as an important complement to experiments for generating meaningful insights into ecological systems.

**Introduction**

As Ecology advances to tackle problems at scales from the continental to global, we are putting our theories to empirical test like never before – working at larger scales in space and time and with unprecedented streams of data. To address fundamental questions in Ecology with these data, we desire to answer questions about causal relationships - either to test basic theory at scale or inform conservation and resource management. Classically in ecology, understanding causal relationships between variables in nature has been the domain of experiments. Experiments, however, have limitations for generalizing to larger scales and contexts beyond study conditions. As Ecology seeks to address theory and application at scale, we must rapidly move beyond a scale where ideal randomized experiments are possible (reviewed in Kimmel *et al.* 2021), and instead must be able to seize the opportunity of new large-scale sources of observational data.

Our ability to test hypotheses about causal relationships in observational data is limited by two fundamental challenges. First, nature is complex! When we use observational data to attempt to answer causal questions, we face numerous **confounding variables** – variables correlated with the cause and the outcome of interest – that can lead to incorrect estimates of causal effects due. Leaving out confounding variables from an analysisleads to **bias** in our estimate of the relationship between the predictor and response; it is not guaranteed to be equal to its true value. Even when we know confounders to account for, collecting all data needed to account for each and every one is likely impossible. Second, as humans, we are limited by our ability to imagine how the different elements of complex ecological systems are related. Thinking through the entirety of the natural history of a system to build an analysis of observational data that will enable credible causal inferences is really hard. As a result, causal inference from observational data is often dismissed as impossible due to the potential for spurious correlations, leading to the common saying “correlation is not causation.” As a result, inferring causation from correlations centers on the problem of unmeasured confounding variables that influence both a causal variable and the response of interest (Fig. 1).

A white circle with a black background

Description automatically generated with low confidence

**Figure 1.** A response variable of interest (Y) is driven by both a measured variable (X) and an unmeasured variable (U) as well as other uncorrelated sources of residual variability (e). In one scenario (A), X and U are uncorrelated, and thus the lack of inclusion of U in a statistical model increases the standard error of the estimate of the effect of X on Y, but does not alter the causal identification of the statistical model. However, ff U also drives X (B) , or if U and X are driven by a common (C) the omission of U from a statistical model causes bias in the estimate of the effect of X on Y. The direction of the bias in the estimator depends on the effect of U on Y, with the result that the causal effects are of X on Y are either over- or under-estimated, masking or mimicking a causal effect.

The problem of excluding known but unmeasured, or unknown, confounding variables from an analysis creates what is known as **omitted variable bias** (Wooldridge 2015; Rinella *et al.* 2020). **Omitted variable bias** (OVB) could be positive or negative. Omitted confounding variables could occur because of missing measurements or due to failures of imagination – simply because we do not yet know confounding variables that are important. For example, one might measure plant communities to study competition, but not measure all the soil abiotic properties that drive all species due to financial or time constraints. Similarly, working with long-term survey data or in human impacted systems, missing data on confounding variables is common, such as when using historical measures of fish abundance to study the impacts of changes in biogenic habitat availability, without measurements of fishing pressure during the same time period. We have no way of knowing the direction or magnitude of the bias, because knowing all possible confounding variables and their relationships in a system is hard, if not impossible. Measuring, controlling for, or even knowing all potential confounding variables is nearly impossible in complex ecological systems (*reviewed in* Dee *et al.* 2023). In short, in observational data collection and analysis, we are always going to miss something, threatening the validity of our causal inferences.

*Do these challenges mean that we should not try to use observational data for causal inference?* We argue no. Rather than throwing up our hands, discounting, and abandoning the use of observational data for causal inference, we suggest that ecologists consider adopting techniques from other disciplines that cannot do experiments – often for logistical or ethical reasons. For instance, it is not ethical to make a person smoke cigarettes daily to test the causal effect of smoking on dementia (Hernan & Robins 2023); one can only manipulate curricula so far in an effort to understand educational outcomes. Thus, disciplines such as psychology, economics, education, epidemiology, sociology, computer science, and more have been building tools to handle OVB in the causal analysis from observational data for decades (Rubin 1974, 2005; Holland 1986; Robins 1989; Heckman 2000; Angrist & Pischke 2008; Pearl 2009; Imbens & Rubin 2015; Morgan & Winship 2015; Hernan & Robins 2023). Some of these tools were even at the center of the 2022 Nobel prize in Economics. As ecologists, we have a decades-long tradition of considering experiments as a gold standard for causal inference (Paine 1966; Reichman 1979; Lubchenco 1980; Carpenter *et al.* 1985; Power 1990; Benedetti-Cecchi & Cinelli 1997; Underwood *et al.* 1997; Silvertown *et al.* 2006; Gotelli & Ellison 2012; Kimmel *et al.* 2021). However, experiments also rely on assumptions (Kimmel et al), which can be hard to meet in the field, have artefacts (refs), and conditions that make them hard to generalize to natural ecosystems. Further, this reliance on the primacy of experiments has meant that the tools of other disciplines have been largely absent from the ecologist’s toolbox (but see Butsic et al. 2017, Rinella et al. 2020, Larsen et al Methods in Ecology & Evol, cite shipley SEM book?). Recently, though, there is a growing interest and use of causal inference in ecology for observational data, including by drawing on a diverse suite of methods from the field of casual inference (Arif & Macneil, 2023 Dee et al., 2023; Dudney et al., 2021; Grace & Irvine, 2020; Larsen, 2013; Larsen et al., 2019; Larsen & Noack, 2020; Macdonald & Mordecai, 2020; Simler-Williamson & Germino, 2022).

Here, we aim to provide a guide to readily available ways to cope with Omitted Variable Bias (OVB) for Ecologists. We begin by briefly describing the status quo for how ecologists most often deal with omitted variable bias and associated limitations. After, we review tools for identifying potential sources of omitted variable bias, building on foundation of using directed acyclic graphs that has become increasingly common in ecology (Arif & MacNeil 2023). We then outline sampling and statistical model designs for dealing with omitted variable bias. Most of these statistical model designs are underutilized, if not novel, for ecology. To illustrate problems with OVB and different ways to identify and address it, we present a motivating example that aims to quantify the causal effect of temperature on marine snail abundances. With this example, we demonstrate the conclusions that would be drawn from the typical approaches an ecologist might take with this data (e.g., random effects in a mixed model, *see* Bolker et al. 2009) – and why they fall short of dealing with OVB (i.e., have statistical bias) – compared to several other statistical model designs that can more adequately control for omitted variable. We then present results from simulation analyses showing that these designs – which have seen limited adoption in ecology – are more robust to OVB (unbiased). We provide guidance for choosing among these designs for different data contexts and questions. Our goal is to enable researchers to advance the field of Ecology at scale using observational data.

**How are ecologists coping with Omitted Variables Bias?**

Omitted variable bias is commonly dealt with in one of four ways in Ecology. The first is using randomized controlled experiments. In an ideal, randomized controlled experiment, the effect of confounding variables is eliminated (but see Kimmel *et al.* 2021 on why this can be difficult), because of the random assignment of treatments to units, so that the treatment and control groups have the same level of any confounders on average. However, randomized controlled experiments, particularly at scale, are not always feasible. Second, in observational studies, ecologists primarily attempt to deal with confounding variables by measuring the confounder and controlling for it in a model. As described above, measuring all confounders, however, is often impossible, particularly for retrospective analyses of existing data. Further, all ptential confounders in the system may not be known. Third, ecologists sometimes make causal claims rooted in their knowledge of the natural history of a system. These claims can be problematic due to a lack of transparency and potential for incorrect statements of effect size; even the knowledge of the most accomplished naturalist can have gaps in their understanding of a system. Finally, ecologists often qualify their results verbally to avoid making causal claims – even when their research focus is causal understanding, rather than description (but see Laubach *et al.* 2021 on causal aims and claims). This practice muddies the waters and can create confusion over whether an author claiming an association or implying causation while allowing themselves plausible deniability. We feel that given our current need to understand causal relationships from large-scale observational data sets, these solutions are not adequate, and can even lead to misleading inferences. Ecologists have an opportunity and, nay, obligation, to leverage (or at least consider) the solutions to Omitted Variable Bias in causal data analysis that other disciplines have been building for decades. This paper provides an entry point into several approaches and complements recent reviews of quasi-experimental methods (e.g., Antonakis *et al.* 2010; Kendall 2015; Butsic *et al.* 2017; Ferraro & Miranda 2017; Bell *et al.* 2018; Bellemare *et al.* 2020; Grace & Irvine 2020; Arif & MacNeil 2022; the appendices of Dee *et al.* 2023) by expanding on panel designs for accounting for OVB.

**A Problem of Omitted Snails {{I MOVED here BEFORE DAG SECTION}}**

To illustrate these empirical challenges and suite of potential solutions, we consider a marine benthic ecosystem, modeled after the Gulf of Maine, USA, where we aim to study the causal effect of temperature on snail abundance. Temperature influences ectotherm metabolic and mortality rates (some references), so we hypothesize snail mortality will increase as water temperature increases. However, total snail abundance is driven by both mortality and recruitment. Regional oceanography drives both water temperature and recruitment (Fig. 3). We assume that the researcher measured snail abundance and temperature at several sites, but not recruitment or any measurement of oceanography. Thus, recruitment and oceanography are unmeasured or so-called “unobserved” confounding variables that pose threats for bias. In addition to these confounding variables, what if our casual diagram in Figure 3 is wrong, and there are other lurking confounding variables? Then, even if oceanography were accounted for, omitted variable bias remains a real possibility – and the estimation of the effect of temperature on snails will be incorrect. Fortunately, we can control for both observed and unobserved confounding variables with appropriate sampling and statistical model designs.

## **A black background with a white circle Description automatically generated with low confidence**

**Figure 3.** A system describing the controls of snail abundance in the intertidal. Oceanography regulates both temperature and recruitment, both of which drive snail abundance. Temperature, however, is also driven by local influences as well. This could be variability in plot-level temperature within a site or site-level temperature variability between years uncorrelated with local oceanography and recruitment.

**​​Using DAGs to clarify our causal understanding and assumptions to ferret out Omitted Variables Bias**

Causal diagrams (forming Structural Causal Models, see Grace & Irvine 2020; Laubach *et al.* 2021; Arif & MacNeil 2023 for in depth introductions for Ecologists) are one of the first tools for identifying and addressing omitted variable bias (Pearl 1995; Pearl *et al.* 2016; Arif & MacNeil 2023). Causal diagrams, or Directed Acyclic Graphs (DAGs), visualize our understanding of and assumptions about causal relationships and confounding variables within a system. In doing so, DAGs transparently clarify the assumptions on which one relies for making causal claims about relationships inferred from observable data, including by ruling out potential confounding variables. If possible, we recommend making a diagram *before* data collection to inform which covariates might be confounding and should be measured if possible. However, due to feasibility constraints or when analyzing existing data, measuring all potential confounders might not be possible. Further, the data could have been collected for another purpose or question, so a set of confounders were deemed unimportant.

Critically, a causal diagram needs to include both measured and *unmeasured* confounding variables – and we argue should be a requirement before conducting an analysis from which one wants to make any causal conclusions. Finally, causal diagrams can also show what variables you should *not* include in an analyses, such as those that cause collider bias (for further discussion see Griffith *et al.* 2020 and McElreath McElreath 2020: Chapter 6).

**Box 1: A Brief Overview of the Elements of Directed Acyclic Graphs for Causal Analysis**

We briefly review the uses and the elements of causal diagrams, such as the one included in Figure 1, called Directed Acyclic Graphs (DAGs) (Grace & Irvine 2020; Laubach *et al.* 2021; Arif & MacNeil 2023). For the variables and implied causal relationships (as paths), we adopt a symbology to differentiate between observed and unobserved variables to reveal where confounding variables might lurk. First, observed variables, things that can be or have been measured, are represented as terms within boxes, as for X and Y in Figure 1. Second, our DAG in Figure 1 shows *unobserved* (unmeasured) variables contained in ellipses, such as the variable *U.* The error term is shown as *e*. In the case of e - the error term - this is a collection of different variables and influences outside of the scope of analysis. We need not know what they are precisely, but we can trace a causal path between them and other variables in our causal diagram. Finally, variables are connected by paths - i.e., arrows. The direction of these arrows represents a direct causal connection going in the direction the arrow is pointed. This is fundamentally different than the “=” sign in an equation, as there is applied directionality. If the value of a causal variable of interest changes (i.e., via manipulation), there will be a concomitant change in the response variable(s) it affects. If a response variable changes, say via direct manipulation, there will be no associated change in the causal variable of interest.

A common critique is that DAGs do not include feedbacks, to which we respond by asking the reader to think of their definition of causality. Here we adopt the Neyman-Rubin counterfactual causality framework (Rubin 1974, 2005; Holland 1986) where we recognize that cause temporarily precedes effect. Therefore, feedbacks can be handled by thinking about a system with a temporal lag (e.g., Larson *et al.* 2008). If an instantaneous feedback is truly present (albeit rare), or if a time-series of both the driver and response variable is not available, one will likely require other tools such as instrumental variables - something beyond the scope of this manuscript (but see Kendall 2015 for an excellent introduction).

In this article, we emphasize how thinking in terms of graphical causal models helps to determine both where omitted variable bias (OVB) might cause problems in analyses as well as in identifying solutions. As applied researchers, we have found that, rather than creating confusion with complexity, graphical causal models paired with robust statistical approaches for causal inferences have often clarified our own thinking about ecological systems.

After building a DAG, as described in Box 1, one can determine potential sources of omitted variable bias from variables influencing both the cause of interest and outcome that have not been observed in the system (e.g., confounding unobservable variables, *U* in Fig. 1B). Not controlling for this confounding variable opens a “back-door” for causal information to flow between your causal variable of interest and its response variable via an unassessed pathway (i.e., confounding). In the case of Figure 1B, *U* would be folded into a statistical model’s error term. The model’s error term and causal variable of interest would be correlated, leading to a violation of a core assumption of Gauss-Markov, known as an endogeneity problem (Antonakis *et al.* 2010; Abdallah *et al.* 2015). Endogeniety means that estimates of xxxxx will be wrong, or different than the true effect in magntiude or even sign. Including a variable in your analysis that blocks all paths between *X* and *Y* via *U* means that your ensuing analysis will satisfy the **back-door criterion** (Pearl 1995, Fig. 2A). Depending on the causal structure of a system and data, there may be many confounding but measured variables. Controlling for measured confounders, by including them as covariates in a statistical model, can “shut the back door” and enable causal identification when all confounding variables are observed and controlled for (see Fig 2. and caption for several examples). Without including such a confounding variable in a statistical model, this omitted variable will cause OVB.

A black background with white circles and white letters

Description automatically generated with low confidence

**Figure 2.** Examples of satisfying the back-door criterion with control variables. By including shaded observed variables, either U or W, in a statistical analysis of the effects of X on Y, omitted variable bias is controlled for the results have a causal interpretation. The relationship between the control variable and Y might (A and B) or might (C and D) not have a causal interpretation, depending on the structure of the system. Note, in (D), Q would have also served as an adequate control instead of W. R would have been a bad control.

Without a causal diagram, understanding how and when to control for these confounding variables is difficult. With a diagram in hand, it can either be visually obvious or one can utilize a variety of software for analyzing DAGs (e.g., Textor *et al.* 2016) to find “open back-doors” that need to be controlled for in order to eliminate omitted variable bias. One can evaluate multiple options and find the most efficient (in terms of sampling effort) variable or variables to measure to control for omitted variable bias. Perhaps, most importantly, one can justify their choice of control variables with a DAG, making their assumptions about how a system works before an analysis transparent to readers of their work in the literature.

A causal diagram is, therefore, the first step on the way for identifying omitted variable bias. On their own, however, they do not in and of themselves provide a means for statistically controlling for OVB, particularly if we have not measured the confounding variable. Nor does a causal diagram help us in the face of *unknown* confounding variables that we have failed to imagine as part of our system. To address both issues, we must consider the sampling design of our observational studies and how we build statistical models with the data to produce causally identified estimates; both reviewed next.

## **Sampling Designs that enable statistical methods to cope with omitted variable bias**

Multiple sampling designs for data collection enable the use of statistical model designs that can address omitted variable bias from confounding variables that vary across space, time, or both. A key feature in the sampling designs is the **nesting** of multiple observations within a cluster or group (e.g. site), such that the causal variable of interest varies across the level of replicates while the confounder varies at the cluster level (Fig 4). *Clustered data is often also referred to as a hierarchical or nested sampling design. We use these terms interchangeably.* Using our snail and temperature example , we outline these sampling designs and discuss how they generate different source of variation in space and time that enable use to use statistical model designs that deal with confounders.

Nested sampling designs can take several forms and generate difference types of variation to study. First, a sampling design could include plots that are sampled within sites across an environmental gradient in a single year (Fig. 4A) -- or a **cross-sectional design** with multiple places (plots) within a site, but sampled at only one point in time. When sites span large environmental gradients, sites differ for many reasons (e.g., temperature, productivity, oceanographic conditions) that affect both temperature and recruitment (Fig. 3) and thus confound this causal relationship of interest. However, with data collected from a cross-sectional sampling design, we can use *within*-site variation in temperature to isolate its effect on recruitment rather than the variation *between* sites, which has a lot of confounding variation.

A second sampling design samples , generating **longitudinal data or** **panel data**. ~~Thus, we do not need to rely on within-site variation as a signal of our causal variable of interest.~~ Longitudinal data enables many approaches to remove the effect of a wide range of confounding variables that vary across sites (i.e., the confounded, between-site variation), reviewed in Dee et al 2023. In the discussion, we consider other types of sampling designs.

A picture containing screenshot, square, rectangle, design

Description automatically generated

**Figure 4.** Visualexamples of clustered/nested/hierarchical study designs with n = 3 per site. In these studies, sites are distributed along a coastline with a corresponding thermal gradient. However, they leverage spatial or temporal clusters to take advantage of temperature variation within sites in a cross-sectional study design (A) or temperature variation across time for a longitudinal (panel) design (B). Open squares are sites. Closed squares are plots within sites. Size of square is proportional to temperature.

## **Statistical Model Designs to Coping with Omitted Variables**

With data and a DAG in hand, there are multiple, well-established statistical designs for analyzing clustered data to handle omitted confounders for causal analysis. We emphasize the term *‘designs*’ over *‘methods,’* because one could implement these model designs using different estimation approaches (e.g., linear regression, Generalized Linear Models, as a part of Structural Equation Models, or with Bayesian techniques). These statistical designs have different costs and benefits, and they differ in their assumptions required for interpreting an estimate as a causal effect. Further, most of the following designs allows us to flexibly control for confounding variables that are both known and unknown (see Angrist & Pischke 2008; Ferraro & Miranda 2017; Dudney *et al.* 2021) – something many Ecologists worry about. Thus, w

We illustrate the different designs using a common set of terms for causal variables of interest *(x*; e.g. local temperature), responses (*y*; e.g. snail counts), and confounding variables (u; e.g. something concrete, which are in the error term e) in a regression, applied to our example of the snail system in Figure 3. Our example includes data from different sites (i) sampled either at multiple time points (j) in panel data or in multiple plots (j) in the case of a cross-sectional data. For the sake of simplicity, we assume a linear model form with normally distributed error (), although the framework will work for generalized linear models as well. If we had measured all of the the confounding variables, we could fit the following model with them included and get a causally identified estimate of the effect of x on y.

(1)

Our goal is to estimate (the effect of temperature on snail abundance). If the confounder was not measured in the design above, and instead a researcher tried to fit the following equation,

(2),

our causal inference about would be incorrect. This is because would now be included in the error term, inducing a correlation between our error and causal variable of interest. This **endogeneity problem** violates the assumptions of the Gauss-Markov theorem and its extensions (Wooldridge 2015) and is what leads to the problem of omitted variable bias.

## *What Ecologists Typically Do: Random or Mixed Effects Models*

Mixed effects models have been popular in ecology for the past two decades (for a useful review, see Bolker et al. 2009, Schielzeth and Nakagawa 2012, Harrison et al. 2018). Originally used to partition variation in heritability between different relatives (Fisher 1919), **random effects --** the effects of clusters in data assumed to come from a random distribution (but see Gelman & Hill 2006 on the linguistic difficulties of defining random effects) -- quickly became a mainstay in the partitioning of variation in randomized experiments with subsamples taken within clusters (Cochran 1937; Eisenhart 1947). They have become a standard part of the toolbox for analyzing ecological experiments (Schielzeth & Nakagawa 2012), and are frequently used when analyzing observational data in ecology.

Random effects account for clustering in our data in the error structure of the model (Gelman & Hill 2006; Bolker *et al.* 2009). ~~This could of course be done with clusters as a~~ **~~fixed effects~~** ~~(i.e., recoding clusters to 0/1 dummy variables – done automatically in most statistical software), (see Box 4 and Cameron and Miller 2015).~~ Random effects, however, have the added second benefit of efficiency, costing fewer degrees of freedom to estimate as we assume all cluster means follow from a distribution (i.e., estimating a grand mean and variance) rather than estimating a separate coefficient for each cluster mean. With this can come an improvement in the estimates of *precision* for coefficient estimates for our causal variable of interest (Gelman & Hill 2006) relative to fixed effects. This can contrast to changes in precision from fixed effects for clusters with cluster robust standard errors which make no assumptions about the distribution of cluster and their means, and fewer assumptions about the homogeneity of residuals between clusters (see Box 4, and Oshchepkov & Shirokanova 2022 for an excellent comparison between mixed models and cluster robust standard errors). Further, as random effects are assumed to be drawn from a common distribution, they have benefits for analyses of unbalanced samples as well as regularizing of cluster means (i.e., drawing them towards the grand mean, see Efron & Morris 1975).

For these reasons, Ecologists conducting a study akin to our snail-temperature example would likely gravitate towards a mixed model to account for site-to-site variability in snail abundances (e.g., the models in Duffy *et al.* 2016 as an example - these are colleagues whom we do not wish to single out, but rather show an example), such as using the following mixed effects model design:

(3)

Here, yij is the abundance of snails at site *i* in year *j*, is the abundance of snails if the temperature was 0 (you might want to center your temperatures to make this the abundance of snails at the mean temperature!), is the effect of temperature x at site i in year j on snails, is the site-specific deviation at site i from our intercept due to random variation which follows a normal distribution and is the residual variability for snail abundance at site *i* in year *j.*

*What assumptions is a random effects design making when it comes to omitted variables bias?*

So why does the above model not control for omitted confounders via its site effect? Why do mixed effects designs produce incorrect results in the face of omitted confounders (i.e., a statistically biased estimate of the causal effect)? To understand this problem, it is key to remember, that when we model random effects, we are not modeling group means *per se*. Rather, we are modeling correlation in our error structure due to clustering in our data (Bolker *et al.* 2009; Wooldridge 2010; Schielzeth & Nakagawa 2012). The coefficient estimates of the causal variable of interest are unaffected by including or not including a random effect (we recommend you try this with any demo data set you have lying around). This difference – modeling error instead of modeling means *per se* – results in many of the above benefits, but also introduces one new assumption not often considered – which we call the**Random Effects Assumption** – that the random effects do not correlate with any covariates in the regression (Wooldridge 2010; Antonakis *et al.* 2021). In the above model, while site is incorporated into here, the statistical model design is not causally identified due to the violation of the **Random Effects Assumption** and will produce biased estimates of .

Here, the random effects of ‘site’’ are part of the error term and assumed to be uncorrelated with temperature for an unbiased estimate (Wooldridge 2010; Schielzeth & Nakagawa 2012). However, our causal variable of interest varies at the site level in a way that is confounded with other drivers, so is not independent of our values. This correlation is a problem and violates the basic assumptions of Ordinary Least Squares estimation and any extension.

A picture containing circle, moon, astronomy

Description automatically generated

**Figure 5.** The system assumed underlying a mixed model (A) versus the system as it is (B). Note that a mixed model does not account for the correlation between site and temperature, nor does it separate out the non-site drivers of temperature. Instead, the effect of temperature on snails is confounded by any correlated site-level drivers that correlate with temperature at the site level.

We can see more clearly how a mixed model would violate the random effects assumption using a path diagram in Figure 5a. In essence, site effects here are site-level residuals drawn from a normal distribution. They represent all other abiotic and biotic forces happening at the site level, but they also assume all are uncorrelated with temperature at the site level. However, given the information in Figure 3, we know that this is not accurate and the key assumption for an unbiased estimator is violated. If we were to take a step back and think about our goals of analysis and our causal understanding, again representing unmeasured quantities in ellipses, what we actually have is more like Figure 5b. Here, while a random site effect would be wonderful in terms of all the benefits discussed above, we would need to remove the effects of site-level confounders to use it – which is not done with the mixed model design above, as shown in Figure 5a. This example illustrates the difficulty in satisfying the *Random Effects Assumption.* More generally*,* we posit that satisfying this assumption is often quite difficult in Ecology – particular in observational data that spans environmental gradients – yet how badly this assumption is violated the is not well explored or acknowledged widely enough. We need a solution that does not produce biased results due to violating assumptions.

*Enter the Econometric Fixed Effects Design*

The Fixed Effects Design represents a familiar starting point for many ecologists who are used to using categorical variables in ANOVA and ANCOVA (e.g., Gotelli & Ellison 2012). Here we use Fixed Effect in two senses of the phrase. The first is the use of the term “fixed effect” is drawn from the econometrics literature, where it refers to attributes of a system (e.g., site, plot, or year) that vary by cluster (i.e., a within cluster intercept). In Ecology, we typically think of this as a categorical predictor representing site, block, or other descriptor of how our data is clustered. In our snail example, this attribute would be the site-level time-invariant effect of recruitment. We also use “fixed effect” as is typically done in ecology, where the term often refers to the coefficient estimates for variables that are not part of the error term – such as a random effect. We acknowledge that there are many uses and definitions of “fixed effect”, leading to a wealth of confusion with different uses of the term across fields (see Gelman & Hill 2006). We hope to not add to the confusion here.

In the econometric sense, treating omitted variables as cluster-invariant variables means that, if we wish to remove them as sources of confounding variation, we can use a bit of algebra known as the **within transformation (bell cite)** or **fixed effects estimator (woolridge cite)**. We illustrate using the following equation

(4)

where is our casual variable of interest, and the error term is composed of idiosyncratic (random error) and unobserved confounding variables, , which represent confounding differences across sites *i* . Left in the error term, will lead to omitted variable bias because is correlated with both and . Instead, if wet transform the data by subtracting this average value at each site across all years, we remove the effects of all site-level omitted variables contained in .

(6)

Using simple algebra, we have removed the confounding influence of time invariant, confounding variables for each site, whether they were observed or not. This approach can be implemented as above by making a deviation from means transformation (Fig. 6A) or by using a model design with a categorical or so-called dummy variable (known as an econometric fixed effect) for each cluster (Fig. 6B). This kind of model design – familiar as an ANCOVA-style model design to many Ecologists – will control for omitted variable bias and produce identical results to the preceding model for (Angrist & Pischke 2008; Wooldridge 2010). For clarity, we can write this model either incorporating the dummy variables (x2i) and site effect () or with just the site effect alone – the means model notation (Gelman & Hill 2006). Note that unlike random effects in a mixed model design, is not constrained to be drawn from any predefined distribution. INTRO TO THIS EQUATiON? PUT IT IN LINE WHEN INTRoDUCED ABoVE?

(7)

Returning to our example, with site as a fixed effect as in equation X, we can control for different sites having different levels of recruitment or other omitted variables correlated with temperature – whether those confounding variables were measured or not. Thus, this design allows us to relax the strong assumption that all confounding variables are observed and measured to interpret B1 as reflective of a causal relationship (e.g., see Figure 6b which shows we are estimating the effect of temperature controlling for recruitment or any other site covariates). We note that cluster robust standard errors are likely important for both approaches for inference (see Box 4 and Cameron & Miller 2015; Abadie *et al.* 2017). For ecological examples using this design, see Ratcliffe et al. XXX, and Dee et al (2023).

The two-way fixed effect design has some drawbacks, despite its simplicity and its strength in controlling for both observed and unobserved confounding variables. First, while fixed effect estimators make much weaker assumptions about confounding variables, these estimators are inefficient compared to random effects. For each fixed effect (each site in our example), we estimate a coefficient for each and thus are estimating many more parameters. Thus, we need a larger sample size to achieve the same level of precision of our estimate using fixed effects versus random effects, presenting a bias-variance trade-off (REF). However, in the case of omitted variable bias, this framework is still preferable over a random effects for causal questions, given the strong assumptions it requires to be an unbiased estimator. Second, we lose information about between-site variation, including gradients between sites that may be of interest. These gradients, while confounded with other variables, could be the focus of some research questions which cannot be easily addressed using fixed effect model designs.

A black background with white circles

Description automatically generated with low confidence

**Figure 6.** Path model representations of different statistical models handling omitted variables in the text.

## *Group Means for efficiency, inference, fun, and profit*

In order to study between-site variation and mitigate the loss of efficiency from the two-way fixed effect design, we can step into the world of ***correlated* random effects designs**. The foundation of correlated random effects model designs is leveraging *group means of XXXXX* to eliminate the effects of confounding variables. For every cluster – e.g., each site, year, or region, researchers calculate a group mean of XXXXXXX to include as a predictor. These group means control for the confounding effects of xxxxxxxxx and estimates a coefficient for between cluster gradients as well. We can then use a random effect for cluster, i.e. make this a hierarchical mixed model, as it represents variation due to cluster after the effect of the confounder is accounted for. By using a random effect for cluster while spending one degree of freedom on our group mean term, we can gain significantly in efficiency over the fixed effects model. In Econometrics, this model design is known as a **Mundlak Device** (Mundlak 1978) and has many extensions (e.g., Wooldridge 2021). For clarity, we term it a **Group Mean Covariate** design.

Correlated random effects designs come in two forms – a group mean covariate model design and a group-mean centered model design (shown in Figure 6 c and 6d respectively). For the group mean covariate model design, consider the following equation:

(8)

where accounts for the effect of cluster-level confounders and is a random effect for site??? XXX . We can see what this looks like as a DAG in Figure 6c. From this diagram, we see that the site mean temperature is controlled for in estimating the temperature effect. The mean temperature of a site is estimated while controlling for each measured temperature.

The site mean temperature coefficient, called a **contextual effect** (Antonakis *et al.* 2021) in the Group Mean Covariate design, shows how changing the mean temperature of a site – and all properties that correlate with site mean temperature – would affect snail abundance were the temperature within a plot to stay the same. For example, *if our plot was 10 degrees C, what would snail abundance be if said plot was in a site with an average temperature of 5 degrees C versus 20 degrees C*? If the contextual effect is 0, then we can conclude that a simple mixed model would have sufficed and that omitted variable bias was not a problem in this particular analysis (Antonakis *et al.* 2021).

The above statistical model design will run into problems, however, if the correlation between our causal variable of interest and its cluster-level mean is too strong for the estimator being used (e.g., *r* > 0.85 for OLS REF??). To solve this, we can use a design that removes this correlation by looking at cluster-level anomalies in the causal variable of interest relative to its cluster mean. We accomplish this with a **Group Mean Centering** design which subtracts the cluster-level mean OF THE VARIABLE FROM THE VARIABLE? from the causal variable of interest as follows

(9).

Equation 9 decomposes our causal variable of interest into between- and within-cluster terms. FThe site mean temperature term is the between-site effect, and the anomaly from the site mean term is the within-site temperature effect. Figure 6d shows the DAG for this design and the similarities and key differences with the previous designs. In particular, the interpretation of is different than in the Group Mean Covariate design. is now a **between estimator** of the combined effect of moving across gradients in temperature and correlated drivers between the sites. If = , omitted variables are not meaningfully influencing snail abundances; both our between and within site differences are due solely to temperature or multiple confounders have cancelled one another out.

The Group Mean Covariate, Group Mean Centered, and Fixed Effects designs all differ in structure but they will yield the same point estimates under most conditions, and they all rely on within-site variation in temperature (see simulations below and Wooldridge 2010). Table 1 shows that these designs yield the same estimates of the temperature effect on snails, – are compared to the random effects design which is XXXXX. Thus, one might ask: *which statistical model design should I use*? This decision depends on the structure and size of one’s data (e.g., how many coefficients do you you have the power to etimate given your sample size) and the question of interest (e.g., are you interested in between-site differences?) . For example, do you have many sites and are only interested in the causal effect of temperature? Fixed effects design. Do you want to know how plot-level snail abundance would change if the average site temperature changes, but plot temperature stays the same? Group Mean Covariate design. Do you want to understand the effects of temperature while examining the net effect of many forces shaping between-site gradients? Group Mean Centered design. Each design can further be extended to cases where the magnitude of the causal variable of interest’s effect is moderated by the level of confounding variables (i.e., an interaction or “heterogeneous” effect, see Box 2).

## *What a Difference Differencing Makes*

Our examples thus far have focused on confounding variables that are unobserved and vary across space (i.e., between sites). We have not discussed omitted confounding variables that differ across time. In the case of omitted confounders varying solely across time and not space (e.g., sites vary randomly in recruitment across space, but year-to-year regional variation in recruitment is correlated with year-to-year regional variation in temperature), we can extend the frameworks presented above, adding years as we did for sites as clusters. If time-varying confounders are uniform across sites (i.e., are additive with spatial confounders), then we can use a fixed effect of time and fixed effect of space (a two-way fixed effect design, Wooldridge 2021) or a site-average of predictors and a time-average of predictors (a Two-Way Mundlak model design; Wooldridge 2021). If, however, temporal confounders differ by site, we need a more general solution. If omitted confounders vary spatiotemporally, we can extend the framework further using the same principles (see Box 3 and Dee *et al.* 2023). Fortunately, the first and second difference statistical model design provide easy solutions that deal with both spatial confounders and site-varying temporal confounders.

To illustrate these approaches, consider that, in addition to site-level oceanographic recruitment effects, the abundance of snails is influenced by coastal development over time at each site (Fig. 7A). However, rates of development are not the same across all sites. As such, separating the effect of local coastal development from the effect of local temperature variability on snail abundance is hard. We can see this in a small modification to the dynamics of our system from eq. 1:

(10).

Here is a site-specific trend in snails over time (j). Due to this trend, if there is also a temporal trend in temperature (e.g., climate change), our estimation of would again be contaminated and our estimate would no longer be causally identified if we did not measure site-level development. We fit a model with a fixed effect of site to account for spatial confounders and then a site by time effect to account for the site-level trends, but this can make the number of parameters blow up quickly (i.e., twice the number of parameters as sites). On the surface, this appears to be a difficult problem to tease apart while keeping the number of estimates parameters low for short datasets with many sites.

A picture containing black, circle, darkness, astronomical object

Description automatically generated

**Figure 7.** Path model representations of different models handling spatial and temporal omitted variables.

Fortunately, there is a simple solution related to the fixed effects transformation: temporal differencing. For each time point in our data, if we subtract the previous time point, we produce a model evaluating the relationship between change in our response variable versus change in our causal variable of interest. Like the fixed effects transformation, the confounding effects of site-level omitted variables that have a temporal trend are therefore eliminated. After the transformation, remains as a term to be estimated with a fixed effect (i.e., using a dummy variable for site, as before, will recover an estimate of the trend for each site). This approach has the added benefit of sweeping up other unknown site-level trends into our estimate of . Our first difference model design, represented as a path diagram in Figure 7B, translates to the following with means model notation:

(11)

Here estimates the effect of temperature as before with estimating the site-level trend of other drivers. Note that confounders at the site level, above, are removed algebraically in this design. If there is no temporal trend in temperature, and as such there is no correlation with other site-level trends, we *could* use random effects for the site term. We caution, however, that this adds back the random effects assumption which is unlikely to be met. Note that if the time between sampling events is unequal across sites, we can divide change by time between samples to model change per unit time. Finally, if we are uninterested in site-specific trends, we can calculate the second difference which eliminates the need to estimate . Note that in this second differences statistical model design model is estimating the relationship between acceleration in change in temperature and acceleration in change in snails.

Using either temporal differencing design has several advantages. We again remove the effect of omitted confounders at the site level. We also remove the effects of temporal confounders at the cluster level that have similar trends to our causal variable of interest. Thus, our estimate of a temperature effect is again causally identified. As we are handling two potential forms of omitted variable bias, our analysis would be more robust to omitted variable bias from unknown confounders. The main drawback of these approaches is the reduced sample sizes; we lose observations from one or two time periods. This reduction in the sample size reduces power and can lead to less precise standard errors, especially in the case of the second difference design. However, this reduction in sample size could be counterbalanced by robustness to violates of assumptions regarding both spatial *and* temporal confounders as well as gains in efficiency from estimating fewer parameters as compared to a fixed effect approach using dummy variables coupled with also estimating separate site-level trends. If site-level trends are nonlinear, however, a fixed effect approach or a correlated random effects model to accommodate spatial confounders coupled with nonlinear site-level trends using either polynomials or even Generalized Additive Models might be more appropriate (for a polynomial example to deal with space and time see Dee *et al.* 2016a). Regardless, temporal differencing can be incredibly useful in cases of both spatial and temporal omitted confounders: a situation all too common in many real ecological systems.

## **Box 2: A Difficult Slope: Omitted Variables that Cause Variation in the Magnitude of the Causal Effect**

Frequently, an omitted confounder does not merely contaminate our estimate of a causal effect but can also lead to model misspecification in the form of missed heterogeneity in the causal effect. This occurs when the causal effect of our variable of interest depends on the level of the confounder itself (i.e., it modifies the causal effect – an interaction effect). In our example, consider that thermal effects in our snail system might depend on levels of recruitment because dense aggregations of intertidal organisms are often better at retaining water and thus resisting desiccation or other forms of thermal stress (e.g., Silliman *et al.* 2011). This is problematic if we have not measured recruitment. In a naive mixed model, we might incorporate this heterogeneity as a random slope. As before, however, the random effects assumption is violated, so a random effects estimator will be biased. To deal with the problem of omitted variable bias here, however, we present two solutions. First, we can use a fixed effects design and include an interaction term between the site dummy variable and our causal variable of interest, allowing us to estimate site-specific temperature effects. Given that we now have site-level slopes, the number of parameters can blow up, leading to this approach being highly inefficient and not advisable for small sample sizes. Rather, we can use correlated random effects approaches with an interaction between the group mean and our causal variable of interest. For example, for a group mean covariate (i.e. Mundlak device) design, we would use the following equation:

This design allows us to examine how site-level confounders – known and unknown – can lead to variation in the effect of our causal variable of interest. It could also show that they have no effect if the estimand for is not different from 0. We could use a similar model for the group mean centered design if deemed appropriate. If we recognized that the magnitude of the temperature effect varied with other non-confounded covariates, we could even use a random slope. In general, models with interactions representing moderators can provide powerful insights into both the effect of the causal driver of interest as well as how those effects vary.

## **Comparison of Approaches**

To demonstrate the utility the preceding solutions, and the consequences of not using them, we used a simulation model based on a longitudinal study of snail populations at multiple sites based on Figure 3. We provide results from 100 simulated data sets with the same initial parameters. Interested users can see the code in Appendix A or can download and run it themselves using the markdown code provided at https://github.com/jebyrnes/ovb\_yeah\_you\_know\_me. Further, for a more interactive exploration, see the web applications written using R Shiny provided as Appendix B (for a single simulated run) and Appendix C (with 100 or more replicate simulations exploring aggregate properties). For the purposes of this manuscript, we simulated the system in Figure 3 where:

* We sample sites over 10 years.
* The Oceanography variable has a mean of 0 and a SD of 1.
* Site temperature is calculated as twice the oceanography variable and then transformed to have a mean of 15C.
* Site recruitment is -2 multiplied by the oceanography variable and then transformed to have a mean of 10 individuals per plot.
* There is additional random variation between sites with a mean of 0 and SD of 1 (not shown in Fig. 3).
* Within a site, the temperature varies over time according to a normal distribution with a mean of 1.
* There is a 1:1 relationship between temperature and snail abundance and recruitment and snails.
* Other non-correlated drivers in the system influence snail abundance with a mean influence of 0 and a SD of 1.

We then analyzed this data using all the statistical designs described above, compared to naive models with no site effect. We also included group mean covariate and group mean centered models without a random effect to demonstrate what a random effect in these models is doing.

A picture containing text, screenshot, diagram, design

Description automatically generated

**Figure 8.** Distribution of point estimates of temperature effects from different models.

Broadly, our simulations show that the point estimates from the random effects (RE) model - what ecologists typically do - is consistently biased in these simulations and well-below the estimates from the other designs and true effect size (Fig. 8, Table 1). Further, not only is the estimated coefficient of the RE model always biased compared to other estimators in our simulations, it is more often within 2SE of 0 than all other models. In most simulations, the 95% confidence intervals of the RE model do not contain the true value of the temperature effect (Table 1).

**Table 1.** Summary results of simulations. Mean and SD of point estimates of temperature effects from different models in the first two columns. Fraction of simulated runs where the mean +/- 2 SE of the temperature effect either overlapped 0 (i.e., high likelihood of committing a type II error) or did not contain the true effect of temperature in the final columns.

| **Model Type** | **Mean Estimate** | **SD Estimate** | **Fraction Sims where 95% CI Contains 0** | **Fraction Sims where 95% CI does Not Contain 1** |
| --- | --- | --- | --- | --- |
| Naive | 0.231 | 0.165 | 0.56 | 0.99 |
| RE | 0.640 | 0.232 | 0.08 | 0.54 |
| FE Using Mean Differencing | 0.985 | 0.215 | 0.00 | 0.05 |
| FE with Dummy Variables | 0.985 | 0.215 | 0.00 | 0.05 |
| Group Mean Covariate | 0.985 | 0.215 | 0.00 | 0.05 |
| Group Mean Centered | 0.985 | 0.215 | 0.00 | 0.05 |
| Group Mean Covariate, no RE | 0.985 | 0.215 | 0.01 | 0.04 |
| Group Mean Centered, no RE | 0.985 | 0.215 | 0.01 | 0.04 |
| First Differences | 0.971 | 0.259 | 0.01 | 0.12 |

Additional explorations show that, in line with the benefits of random effects in mixed models, a site-level random effect is crucial when either the study design is unbalanced or there is site-level variation that is uncorrelated with temperature (Appendix A). If our simulation has no site-level variation other than temperature and our confounder, a random effect does not improve model’s ability to estimate the effect of our causal variable of interest. This assumption is unrealistic for most real data sets. As such, we highlight the need for a site level random effect with either of these two designs. In general, we urge researchers to incorporate random effects or robust standard errors as needed to accommodate clustering in the error, per the study design, recognizing the tradeoffs of using both and appropriate context (see Oshchepkov & Shirokanova 2022 for an excellent review).

## **Box 3: Reality Bites: Coping with spatiotemporal omitted confounders**

Spatiotemporal confounding variables – those that are site (or plot) specific and vary through time – pose challenges, and the solutions require more thoughtful study and statistical model design. To illustrate, we consider a scenario where recruitment, a confounding variable related to both snail abundance and temperature, is not static through time but instead varies by site and year (as in a realistic case). For example, sites that experience strong cold-water pulses in a year also experience unusually snail high recruitment in those same years due to oceanographic drivers. The sampling designs for coping with spatiotemporal omitted variables are based on the same principles as cross-sectional and longitudinal sampling, only now we combine the two.

With longitudinal data that includes multiple plots sampled within a site through time, we can flexibly control for this sort of spatiotemporal confounding at the site level by extending the two-way fixed effect designs discussed above. We can add a site-by-time fixed effect, , to our model, in addition to a fixed effect of plot, , where k is a fixed plot within site resampled over time (see below for a discussion of fixed versus re-randomized plots). This produces the following means model:

From this equation, we can see that captures time invariant plot-level confounding effects while captures the effects of spatiotemporal omitted variables at the site by time level. Note, there could be additional spatial or temporal only confounders. This model design sweeps their effects onto the spatiotemporal term.

In small datasets, the above model design can consume degrees of freedom rapidly. For this reason, we can instead use the more efficient correlated random effects model design (e.g., a variation on the Two-way Mundlak model design *sensu* Wooldridge 2021) using site-year means () and plot means () for the entire survey to control for spatiotemporal and plot confounding respectively:

Here the and terms are random effects for plot and unique site-time combinations respectively.

When sampling to handle spatiotemporal confounders, should plots within sites over time be permanent or randomly placed each year? The above models assume permanent plots, so we can eliminate confounding variables at the plot-level that is time invariant over the study period. For this reason, permanent plots help us cope with within-site OVB issues and have higher power to detect change over time (Urquhart & Kincaid 1999). Logistically, however, permanent plots within sites might not be possible. As such, the above models can be modified to drop plot effects; however, they would then assume that there are no confounding differences across plots and could have lower power to detect effects of drivers. We emphasize that the choice of fixed or random plot placement with these designs is a balancing act, however, as fixed plots can lead to a lower sample size due to logistical considerations in many environments, and direct readers to other explorations of this topic (see Gomes 2022 for an excellent elaboration). Finally, without a nested data structure – e.g., plots within sites resampled over years – we cannot include a site by year effect as in the above models. We can attempt to use site-level polynomials or GAMs to mimic site-by-time effects (for a polynomial example, see Dee *et al.* 2016a), but, this requires knowledge of how the confounder varies at sites over time and extensive testing for robustness to these assumptions. In the many cases where this is not possible or inadvisable given the likelihood of creating incorrect causal inference, without multiple plots per site over time, “nothing to be done” (Beckett 1954).

In general, we urge caution when dealing with spatiotemporal omitted variables, and careful use of causal diagrams to ensure that we are controlling for a confounder at the relevant spatiotemporal scale. This topic is one that that deserves far more exploration in Ecology. More from other disciplines on this tricky class of problem and approaches can be found from literature outside of the scope of this paper (e.g., Ferraro & Hanauer 2014; Athey & Imbens 2017; Oster 2019).

**Discussion**

We hope that our introduction to statistical and sampling designs to address the problem of omitted variable bias and causal inference from observational data has shown you that through thinking carefully about biological systems, we can draw on a solid set of existing methods for causal inferences from observational data. At the core is building an *a priori* causal model of how a system works, and then applying that to the statistical and sampling design that you will use in answering your causal questions. The techniques for addressing omitted variable bias are well within the standard statistical toolbox of most modern ecologists. The results, as seen in at least this toy example, can be profound for our ability to understand biological systems. It is time to begin using these tools to address some of the most pressing questions in the study of nature.

Further, we hope that Ecologists can see the above concepts as part of a generalizable approach to handling confounding variables using clustered data (sometimes from a suite of tools called panel models in other fields). While we have talked of sites and years, the same concepts apply to studies with cohort effects, individual effects, or other lower levels of clustering as well as to larger-scale studies with not just sites and years but regions and decades. The general suite of approaches remains the same, and potential confounding variables at these different scales can be identified in initial causal diagrams. Further, we show examples using linear models, but these approaches can be implemented as generalized linear models or with other functional forms and interactions. Similarly, cross-sectional and longitudinal sampling designs are generalizable beyond just the simple case presented in our example; for instance, one could adapt the above designs if temperature and recruitment varied temporally at a regional rather than site scale (e.g., sampling plots within a single site or sampling many – both over many years to leverage spatial variation in temperature) or for spatiotemporal designs (see Box 3). Combining these designs with others, such as the classic stratified random sampling design (Stevens & Olsen 2004; Grafström & Lundström 2013; Robertson *et al.* 2013; Foster *et al.* 2018; Kermorvant *et al.* 2019), will allow for the analyses that can improve causal identification and reduce the influence of variability of uncorrelated variables when estimating causal relationships. How to design a study to fully account for confounders, however, will hinge on a causal diagram of the system.

**Box 4: Clustered Robust Standard Errors: An Underutilized Tool in Ecology**

{need **broader intro to non-independence and structure in the errors and why its important to consider}}** In light of that, we recommend the use of clustered robust standard errors above. Clustered robust standard errors are not commonly used in Ecology (but see examples in Dee *et al.* 2016b; Dudney *et al.* 2021) despite being a way to continue to use Ordinary Least Squares and then flexibly apply a post-hoc adjustment to accommodate clustered data, heteroskedasticity, correlation between time points, and other arbitrary correlation structures within the data (Cameron & Miller 2015; Abadie *et al.* 2017). While random effects, autocorrelation structures in statistical models, and more, can address some of the same issues in the design of a model, robust standard errors often provide a simpler solution allowing researchers to not have to make more assumptions about the structure of their data that they are not interested in. There are tradeoffs, however, and as multiple already widely known techniques cover similar ground, we recommend looking at comparisons of approaches such as (Oshchepkov & Shirokanova 2022). A full discussion or review of robust standard errors is beyond the scope of this discussion, but we refer applied researchers to the documentation for the ‘*sandwich’* package in R and to comprehensive reviews (e.g., Cameron & Miller 2015; Abadie *et al.* 2017).

The important thing is to be transparent in how we deal with problems of omitted confounding variables. What are the assumptions they are making to interpret an effect as causal or not? If you are using mixed models, do you meet the random effects assumption? Have you evaluated your residuals to determine if you need to implement robust standard errors? Why did you include some covariates and not others? Do you have a DAG or a conceptual model of your system that might help a reader understand your thought process? Putting these types of results in even a brief sentence – if not a figure or full breakdown in a manuscript supplement (e.g., see Dee *et al*. 2023) – will go far in terms of making your analyses more transparent and better able to be built upon to advance science. Accepting that our models and knowledge are imperfect and that someday, someone will come along with a different one that will produce different conclusions and yield new insights.

The approaches we present here are surely not a panacea and require assumption for causal inference, as do any approach, including experiments (Kimmel et al 2021). These include assumptions shared with experiments: i.e., SUTVA – or the stable unit treatment value assumption -- which includes no interference and no multiple versions of treatment (reviewed in Kimmel et al 2021). These statistical designs we present herein also include assumptions of linear additive effects (Imai and Kim 2020) and that effects are homogeneous across units and time periods (i.e., no dynamic treatment effects, see ([Goodman-Bacon 2021](https://bookdown.org/mike/data_analysis/two-way-fixed-effects.html" \l "ref-goodman2021difference); [De Chaisemartin and d’Haultfoeuille 2020](https://bookdown.org/mike/data_analysis/two-way-fixed-effects.html" \l "ref-de2020two); [Sun and Abraham 2021](https://bookdown.org/mike/data_analysis/two-way-fixed-effects.html" \l "ref-sun2021estimating); Callaway & Santa Ana ). Related to this latter assumption, each on these methods rely on the parallel trends assumption, which assumes that the outcomes through time of treated units – if they had not been treated – would be the parallel to that of the untreated group, conditional on observed covariates. This assumption has come under scrunity lately (reviewed in Roth et al 2023), particularly when treatments are staggered - or the treatment (or change in the causal variable of interest) happen at different points in time across units. This is not a problem with only two time periods (REFS); but in multiple time periods, recent methodological research finds bias in the estimators reviewed herein, in the presence of staggered treatments assignment – i.e., the “treatment” or change in a causal variable happens at different time points for different units -- because it complicates how the regression weights different observations of control versus treated groups (for details, see ([Goodman-Bacon 2021](https://bookdown.org/mike/data_analysis/two-way-fixed-effects.html" \l "ref-goodman2021difference); [De Chaisemartin and d’Haultfoeuille 2020](https://bookdown.org/mike/data_analysis/two-way-fixed-effects.html" \l "ref-de2020two); [Sun and Abraham 2021](https://bookdown.org/mike/data_analysis/two-way-fixed-effects.html" \l "ref-sun2021estimating); [Borusyak, Jaravel, and Spiess 2021](https://bookdown.org/mike/data_analysis/two-way-fixed-effects.html" \l "ref-borusyak2021revisiting)) . However, recent approaches relax these assumptions in a rapidly evolving literature on difference-in-difference and generalizations for two-way fixed effects (reviewed in Roth et al. 2023), including for heterogeneous treatment effects, non-linear cases, and continuous treatments as we have in our example (Callaway et al. 2021) . For instance, see Imai and Kim (2019) for extensions to allow for dynamic causal effects (REF); XXXXXXX on XXXXXXX; and [Borusyak, Jaravel, and Spiess 2021](https://bookdown.org/mike/data_analysis/two-way-fixed-effects.html" \l "ref-borusyak2021revisiting) and De Chaisemartin & d’Haultfoeuille with extensions for heterogeneous causal effects and staggered treatments. Many of these extensions are also readily available in R packages such as the ‘did’ package and ….. (see Roth et al. 2023 for more extensive list). Thus, we suggest to use these approaches reviewed herein in concert with sensitivity tests (REF ROSEMBAUM, Oster, Altonji) and other designs that make different assumptions to probe robustness of results (see Dee et al 2023 for example in ecology.

*{{*

*This is very useful - https://bcallaway11.github.io/did/articles/multi-period-did.html*

A screenshot of a document

Description automatically generated

}}

Finally, we emphasize that this paper provides an entry point into a broader interdisciplinary literature on causal inference in observational data and longitudinal data analysis (see also books on panel regression methods). Indeed, other quasi-experimental methods, such as instrumental variables and regression discontinuity designs, can be used to eliminate omitted variable bias when their assumptions are met (reviewed in (Butsic *et al.* 2017) (Angrist *et al.* 1996; Kendall 2015; Grace 2021) Dee et al 2023), and as can thoughtful uses of the front-door criterion (Bellemare *et al.* 2020) Pearl. We urge ecologists, long grounded in experiments, to open up to writings in Econometrics, Epidemiology, Computer Science, Public Health, and other disciplines that cannot always do field experiments. Embracing this transdisciplinary approach will enable us to increase our toolkit and breadth of knowledge of the tremendous advances in causal inference. As an incomplete -- and one day out of date -- set of starting points for further reading, we recommend Cunningham’s Causal Inference: The Mixtape (2021), McElreath’s chapters on causal diagrams in Statistical Rethinking (2020), Angrist and Pishke’s Mostly Harmless Econometrics (Angrist & Pischke 2008), Sloman’s Causal Models (2005), and Pearl’s Causal Inference in Statistics: A Primer (Pearl *et al.* 2016). We also suggest Ecologists interrogate the assumptions and interpretations of their experiments (Kimmel *et al.* 2021). Given how an experiment was designed and run, are its results causally valid with respect to the purported mechanism? It is high time to critically interrogate how to get the robust causal inferences needed to grapple with our rapidly changing world to learn how to mitigate, acclimate, and adapt at scale.

## **Conclusion**

“Correlation does not equal causation” rings in many of our heads from our Biostatistics 101 courses. One main driver of this message is the specter of Omitted Variable Bias from unmeasured confounding variables, which has impeded the use of observational data for causal inference in Ecology for much of its history. However, a large tool kit can help developed largely in other disciplines that have vetted and adopted frameworks for robust and transparent causal inference can help. To translate and spur their adoption, this review, with worked examples and code, offers an introduction to these methods to serve as a new arrow in the quiver of Ecologists. With a growing and unprecedented volume of observational data, and appropriate tools to analyze it in hand, we can expand our ability to address pressing applied and theoretical questions at scale.

## **Acknowledgements**

We thank the NCEAS LTER working group Scaling-up productivity responses to changes in biodiversity for initiating the conversations and feedback that led to this paper, supported by the NSF Long-Term Ecological Research (LTER) Network Communications Office and DEB-1545288. This work was partially supported by the National Science Foundation as part of the PIE-LTER Program (award #1637630) and the Stone Living Lab to J.B.; and NSF OCE # 2049360 to L.E.D. We thank S. Elmendorf, I. Rosenthal, R. Stevenson, A. Carter, and the UMB Stats Snack for helpful conversation and comments on early drafts of the manuscript.

**References**

Abadie, A., Athey, S., Imbens, G.W. & Wooldridge, J. (2017). *When Should You Adjust Standard Errors for Clustering?* (Working Paper No. 24003). Working Paper Series. National Bureau of Economic Research.

Abdallah, W., Goergen, M. & O’Sullivan, N. (2015). Endogeneity: How Failure to Correct for it can Cause Wrong Inferences and Some Remedies. *Br. J. Manag.*, 26, 791–804.

Angrist, J.D., Imbens, G.W. & Rubin, D.B. (1996). Identification of Causal Effects Using Instrumental Variables. *J. Am. Stat. Assoc.*, 29.

Angrist, J.D. & Pischke, J.-S. (2008). Mostly harmless econometrics. In: *Mostly Harmless Econometrics*. Princeton university press.

Antonakis, J., Bastardoz, N. & Rönkkö, M. (2021). On Ignoring the Random Effects Assumption in Multilevel Models: Review, Critique, and Recommendations. *Organ. Res. Methods*, 24, 443–483.

Antonakis, J., Bendahan, S., Jacquart, P. & Lalive, R. (2010). On making causal claims: A review and recommendations. *Leadersh. Q.*, Leadership Quarterly Yearly Review, 21, 1086–1120.

Arif, S. & MacNeil, M.A. (2022). Utilizing causal diagrams across quasi-experimental approaches. *Ecosphere*, 13, e4009.

Arif, S. & MacNeil, M.A. (2023). Applying the structural causal model framework for observational causal inference in ecology. *Ecol. Monogr.*, 93, e1554.

Athey, S. & Imbens, G.W. (2017). The State of Applied Econometrics: Causality and Policy Evaluation. *J. Econ. Perspect.*, 31, 3–32.

Beckett, S. (1954). *Waiting for Godot: tragicomedy in 2 acts*. Evergreen book. Grove Press, New York.

Bell, A., Fairbrother, M. & Jones, K. (2018). Fixed and random effects models: making an informed choice. *Qual. Quant.*, 55, 117.

Bellemare, M.F., Bloem, J.R. & Wexler, N. (2020). The Paper of How: Estimating Treatment Effects Using the Front-Door Criterion.

Benedetti-Cecchi, L. & Cinelli, F. (1997). Confounding in field experiments: direct and indirect effects of artifacts due to the manipulation of limpets and macroalgae. *J. Exp. Mar. Biol. Ecol.*, 209, 171–184.

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H., *et al.* (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends Ecol. Evol.*, 24, 127–135.

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 Appl. Ecol.*, 19, 1–10.

Cameron, A.C. & Miller, D.L. (2015). A Practitioner’s Guide to Cluster-Robust Inference. *J. Hum. Resour.*, 50, 317–372.

Carpenter, S.R., Kitchell, J.F. & Hodgson, J.R. (1985). Cascading Trophic Interactions and Lake Productivity. *BioScience*, 35, 634–639.

Cochran, W.G. (1937). Problems arising in the analysis of a series of similar experiments. *Suppl. J. R. Stat. Soc.*, 4, 102–118.

Cunningham, S. (2021). Causal inference. In: *Causal Inference*. Yale University Press.

Dee, L.E., Ferraro, P.J., Severen, C.N., Kimmel, K.A., Borer, E.T., Byrnes, J.E.K., *et al.* (2023). Clarifying the effect of biodiversity on productivity in natural ecosystems with longitudinal data and methods for causal inference. *Nat. Commun.*, 14, 2607.

Dee, L.E., Miller, S.J., Peavey, L.E., Bradley, D., Gentry, R.R., Startz, R., *et al.* (2016a). Functional diversity of catch mitigates negative effects of temperature variability on fisheries yields. *Proc. R. Soc. B Biol. Sci.*, 283, 20161435.

Dee, L.E., Miller, S.J., Peavey, L.E., Bradley, D., Gentry, R.R., Startz, R., *et al.* (2016b). Functional diversity of catch mitigates negative effects of temperature variability on fisheries yields. *Proc. R. Soc. B Biol. Sci.*, 283, 20161435.

Dudney, J., Willing, C.E., Das, A.J., Latimer, A.M., Nesmith, J.C.B. & Battles, J.J. (2021). Nonlinear shifts in infectious rust disease due to climate change. *Nat. Commun.*, 12, 5102.

Duffy, J.E., Lefcheck, J.S., Stuart-Smith, R.D., Navarrete, S.A. & Edgar, G.J. (2016). Biodiversity enhances reef fish biomass and resistance to climate change. *PNAS*, 113, 6230–6235.

Efron, B. & Morris, C. (1975). Data Analysis Using Stein’s Estimator and its Generalizations. *J. Am. Stat. Assoc.*, 70, 311–319.

Eisenhart, C. (1947). The Assumptions Underlying the Analysis of Variance. *Biometrics*, 3, 1–21.

Ferraro, P.J. & Hanauer, M.M. (2014). Advances in Measuring the Environmental and Social Impacts of Environmental Programs. *Annu. Rev. Environ. Resour.*, 39, 495–517.

Ferraro, P.J. & Miranda, J.J. (2017). Panel Data Designs and Estimators as Substitutes for Randomized Controlled Trials in the Evaluation of Public Programs. *J. Assoc. Environ. Resour. Econ.*, 4, 281–317.

Fisher, R.A. (1919). XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Earth Environ. Sci. Trans. R. Soc. Edinb.*, 52, 399–433.

Foster, S., Monk, J., Lawrence, E., Hayes, K., Hosack, G. & Przeslawski, R. (2018). Statistical considerations for monitoring and sampling.

Gelman, A. & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gomes, D.G.E. (2022). Should I use fixed effects or random effects when I have fewer than five levels of a grouping factor in a mixed-effects model? *PeerJ*, 10, e12794.

Gotelli, N.J. & Ellison, A.M. (2012). *A Primer of Ecological Statistics*. Second Edition. Oxford University Press, Oxford, New York.

Grace, J.B. (2021). Instrumental variable methods in structural equation models. *Methods Ecol. Evol.*, 12, 1148–1157.

Grace, J.B. & Irvine, K.M. (2020). Scientist’s guide to developing explanatory statistical models using causal analysis principles. *Ecology*, 101.

Grafström, A. & Lundström, N. (2013). Why Well Spread Probability Samples Are Balanced. *Open J. Stat.*, 3, 36–41.

Griffith, G.J., Morris, T.T., Tudball, M.J., Herbert, A., Mancano, G., Pike, L., *et al.* (2020). Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nat. Commun.*, 11, 5749.

Harrison, X.A., Donaldson, L., Correa-Cano, M.E., Evans, J., Fisher, D.N., Goodwin, C.E.D., *et al.* (2018). A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ*, 6, e4794.

Heckman, J.J. (2000). Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective\*. *Q. J. Econ.*, 115, 45–97.

Hernan, M.A. & Robins, J.M. (2023). *Causal Inference: What If*. CRC Press, Boca Raton.

Holland, P.W. (1986). Statistics and Causal Inference. *J. Am. Stat. Assoc.*, 81, 945–960.

Imbens, G.W. & Rubin, D.B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, Cambridge.

Kendall, B.E. (2015). *A statistical symphony: instrumental variables reveal causality and control measurement error*.

Kermorvant, C., D’Amico, F., Bru, N., Caill-Milly, N. & Robertson, B. (2019). Spatially balanced sampling designs for environmental surveys. *Environ. Monit. Assess.*, 191, 524.

Kimmel, K., Dee, L.E., Avolio, M.L. & Ferraro, P.J. (2021). Causal assumptions and causal inference in ecological experiments. *Trends Ecol. Evol.*, 36, 1141–1152.

Larson, D., Grace, J. & Larson, J. (2008). Long-term dynamics of leafy spurge (*Euphorbia esula*) and its biocontrol agent, flea beetles in the genus Aphthona. *Biol. Control*, 47, 250–256.

Laubach, Z.M., Murray, E.J., Hoke, K.L., Safran, R.J. & Perng, W. (2021). A biologist’s guide to model selection and causal inference. *Proc. R. Soc. B Biol. Sci.*, 288, 20202815.

Lubchenco, J. (1980). Algal Zonation in the New England Rocky Intertidal Community: An Experimental Analysis. *Ecology*, 61, 333–344.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC.

Morgan, S.L. & Winship, C. (2015). *Counterfactuals and Causal Inference*. Cambridge University Press.

Mundlak, Y. (1978). On the Pooling of Time Series and Cross Section Data. *Econometrica*, 46, 69–85.

Oshchepkov, A. & Shirokanova, A. (2022). Bridging the gap between multilevel modeling and economic methods. *Soc. Sci. Res.*, 104, 102689.

Oster, E. (2019). Unobservable Selection and Coefficient Stability: Theory and Evidence. *J. Bus. Econ. Stat.*, 37, 187–204.

Paine, R.T. (1966). Food web compexity and species diversity. *Am. Nat.*, 100, 65–75.

Pearl, J. (1995). Causal Diagrams for Empirical Research. *Biometrika*, 82, 669–688.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pearl, J., Glymour, M. & Jewell, N.P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.

Power, M.E. (1990). Effects of Fish in River Food Webs. *Science*, 250, 811–814.

Reichman, O.J. (1979). Desert Granivore Foraging and Its Impact on Seed Densities and Distributions. *Ecology*, 60, 1086–1092.

Rinella, M.J., Strong, D.J. & Vermeire, L.T. (2020). Omitted variable bias in studies of plant interactions. *Ecology*, 101, e03020.

Robertson, B.L., Brown, J.A., McDonald, T. & Jaksons, P. (2013). BAS: Balanced Acceptance Sampling of Natural Resources. *Biometrics*, 69, 776–784.

Robins, J. (1989). The control of confounding by intermediate variables. *Stat. Med.*, 8, 679–701.

Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, 66, 688–701.

Rubin, D.B. (2005). Causal Inference Using Potential Outcomes. *J. Am. Stat. Assoc.*, 100, 322–331.

Schielzeth, H. & Nakagawa, S. (2012). Nested by design: model fitting and interpretation in a mixed model era. *Methods Ecol. Evol.*, 4, 14–24.

Silliman, B.R., Bertness, M.D., Altieri, A.H., Griffin, J.N., Bazterrica, M.C., Hidalgo, F.J., *et al.* (2011). Whole-community facilitation regulates biodiversity on Patagonian rocky shores. *PloS One*, 6, e24502.

Silvertown, J., Poulton, P., Johnston, E., Edwards, G., Heard, M. & Biss, P.M. (2006). The Park Grass Experiment 1856–2006: its contribution to ecology. *J. Ecol.*, 94, 801–814.

Sloman, S. (2005). *Causal models: How people think about the world and its alternatives*. Oxford University Press.

Stevens, D.L. & Olsen, A.R. (2004). Spatially Balanced Sampling of Natural Resources. *J. Am. Stat. Assoc.*, 99, 262–278.

Textor, J., van der Zander, B., Gilthorpe, M.S., Liśkiewicz, M. & Ellison, G.T. (2016). Robust causal inference using directed acyclic graphs: the R package ‘dagitty.’ *Int. J. Epidemiol.*, 45, 1887–1894.

Underwood, A.J., Underwood, A.L., Underwood, A.J. & Wnderwood, A.J. (1997). *Experiments in ecology: their logical design and interpretation using analysis of variance*. Cambridge university press.

Urquhart, N.S. & Kincaid, T.M. (1999). Designs for Detecting Trend from Repeated Surveys of Ecological Resources. *J. Agric. Biol. Environ. Stat.*, 4, 404–414.

Wooldridge, J.M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

Wooldridge, J.M. (2015). *Introductory econometrics: A modern approach*. Cengage learning.

Wooldridge, J.M. (2021). Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators.