# Oh S\*\*T! I forgot to measure that! Coping with omitted variable bias for the causal analysis of observational data

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

Figures: <https://docs.google.com/presentation/d/1m5eRq90xwpTpZ8sC3dH_URaKabePcn8oCFt-sEl_MgU/edit>

Code Repo: <https://github.com/jebyrnes/ovb_yeah_you_know_me>

Appendix: <https://htmlpreview.github.io/?https://github.com/jebyrnes/ovb_yeah_you_know_me/blob/master/markdown/models_and_ovb.html>

App for 1 sample: <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

**Introduction**

Experiments alone are no longer enough to generate the causal inferences needed by Ecology. 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. As Ecology seeks to address theory and application at scale, however, we 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 problems. First, nature is complex! When we use observational data, there are sure to be a number of lurking **confounding variables** – variables correlated with the cause of interest and the outcome – ready to introduce bias into our inferences. Even when we know confounders to account for, collecting all of the data needed to model each and every one might prove impossible. Second, we are human. As humans, we are limited by our ability to imagine how the different elements of complex ecological systems are linked together. It is hard to think through the entirety of the natural history of a system in order to build an analysis of observational data that will enable credible causal inferences. As a result,the entire endeavor is often deemed impossible, and causal inferences from observations are dismissed. This dismissal comes from our understanding of how the two above issues contribute to the spurious correlations and have led to the common saying “Correlation is not causation.”

This statement ultimately encompasses the central obstacle in deriving causal inference from observational data - the problem of unmeasured confounding variables that influence both a predictor and the response of interest (Fig. 1). Omitting confounding variables from an analysisleads to **bias**: an estimate of the relationship between the predictor and response is not guaranteed to be equal to its true value. 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** could be positive or negative. We have no way of knowing either its direction or its magnitude, because it is hard, and likely impossible, to know all possible confounding variables and their relationships in a system. We fully admit, measuring, controlling for, or even knowing all potential confounding variables is nearly impossible in complex ecological systems (reviewed in Dee et al., in press).

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. In essence, in observational data collection and analysis, we are always going to miss something – which threatens our causal inferences.

Does this mean that we should not try to derive causal inference from observational data? No. Rather than throwing up our hands, discounting and abandoning the use of observational data for causal inference, we suggest that we should adopt techniques from other disciplines that cannot do experiments, and yet have been building tools for causal analysis from observational data in the face of OVB for decades. OVB has been widely recognized to the point of obsession in fields such as psychology, economics, education, epidemiology, sociology, computer science, and more (Rubin 1974, 2005, Holland 1986, Robins 1989, Heckman 2000, Angrist and Pischke 2008, Pearl 2009, Imbens and Rubin 2015, Morgan and Winship 2015, Hernan and Robins 2023). These fields have developed a variety of solutions - some even at the center of the 2022 Nobel prize in Economics - that have been largely absent from the ecologist’s toolbox (but see Butsic et al. 2017, Rinella et al. 2020 and others on OVB and instrumental variables). This difference could be due to Ecologists having fewer barriers to conducting experiments while other fields (e.g., public health, economics, education, etc.) often cannot perform experiments for logistical or ethical reasons. It is not ethical to make a person smoke cigarettes daily to test the causal effect of smoking on dementia (Hernan and Robins 2023). One can only tweak curricula so far in an effort to understand educational outcomes. Yet, these disciplines have been tasked with causal inferences based on observational data in the face of confounding variables related to the outcomes and causal effects of interest.

Here, we aim to provide a guide to simple and readily available ways to cope with omitted variable bias for Ecologists. We begin by briefly describing the status quo for how ecologists deal with omitted variable bias and then quickly pivot to laying out tools for identifying potential sources of omitted variable bias using a foundation that has become increasingly common in ecology (Arif and MacNeil 2023) and then for dealing with omitted variable bias in statistical design. To illustrate problems with OVB and different ways to identify and address it, we present a motivating example using temperature as a predictor of marine snail abundances. With this example, we demonstrate the conclusions that would be drawn from the typical approaches ecologists take to this data (e.g., random effects in a mixed model Bolker et al. 2009) – and why they fall short of dealing with OVB – compared to several other statistical designs that address omitted variable bias. We then present results from simulation analyses showing the consistency of these designs that are more robust to OVB robust methods. We provide guidance for choosing among these designs for different data contexts and questions and discuss extensions to OVB problems under more complex (and yet realistic) OVB scenarios. Throughout all of this, we emphasize thinking in terms of graphical causal models to determine both where OVB might cause problems in analyses as well as 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. We hope that these techniques might enable other researchers to do more with less, as it were, and help advance the field of Ecology at scale.

**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. When treatments are perfectly randomized, and thus decoupled from other confounding influences (and often they are not - see Kimmel et al. 2021), they do not influence our estimate of causal effects of the treatment (or causal variable) of interest. Experimentation - particularly at scale - is not always feasible, however. Second, in observational studies, ecologists primarily attempt to deal with confounding variables by measuring the confounder and including it in the model. Measuring a confounder, however, is frequently not possible, particularly for retrospective analyses of existing data, and it is possible that confounders in the system are not even known. Third, ecologists can make causal claims rooted in their knowledge of the natural history of a system but provide no supporting evidence beyond this system-specific knowledge as to why their claim is causal. This can be problematic first for a simple lack of transparency. Further, even the knowledge of the most accomplished naturalist can have gaps in their understanding of a system, however. Finally, ecologists often qualify their results verbally in order to avoid making a causal claim - even when the goal of the analysis is causal understanding, rather than description of associations (but see Dudney et al. 2021). This creates muddy waters - is the author claiming a mere association, or implying causation while allowing themselves plausible deniability - particularly when it comes to appeasing hard-nosed reviewers? 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 both 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.

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

Causal diagrams (a.k.a. Structural Causal Models, see Grace and Irvine 2020, Laubach et al. 2021, Arif and MacNeil 2023 for in depth introductions) are one of the first tools for identifying and addressing omitted variable bias (Pearl 1995, Pearl et al. 2016, Arif and MacNeil 2023). Causal diagrams, or Directed Acyclic Graphs (DAGs) visualize our understanding of and assumptions about causal connections within a system. Said another way, a DAG is a visualization of the assumptions on which one relies for making causal claims from observable data. Making a DAG therefore helps determine potential confounding variables. If possible, we recommend making a diagram *before* designing an observational survey to inform which covariates might be confounding and should be measured. However, due to feasibility constraints or when analyzing existing data, measuring all potential confounders might not be possible. For example, the data could have been collected for another purpose or question, so a set of confounders were deemed unimportant. For this reason, 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 for both new and retrospective analyses can also show what variables you should include in an analyses, such as those that cause collider bias (for an excellent discussion of this topic beyond the scope of this manuscript, see McElreath McElreath 2020 Chapter 6 or , or Griffith et al. 2020 for examples in the analysis of Covid-19 risk factors).

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

Let us take a moment to break down the elements of a causal diagram, such as the one included in Figure 1, called Directed Acyclic Graphs, or DAGs (Grace and Irvine 2020, Laubach et al. 2021, Arif and MacNeil 2023).

For the variables and paths themselves, we adopt a symbology to differentiate between observed and unobserved variables in order to facilitate a connection to estimation techniques and reveal where problems might lie in a data set. First, observed variables - the variables that can be or have been tangibly measured. We will represent observed variables as terms within boxes as X and Y in figure 1. Second, our causal model in Figure 1 shows unobserved variables contained in ellipses. In the case of Z, this is a measurable quantity that has not been measured. 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.

We note that one might blanch at our use of DAGs and request that feedbacks be included in a causal model. To which we respond by asking the reader to think of their definition of causality. Here we adopt the Neyman-Rubin 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 (rare), or the only data available has a single time-step, 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).

After building a causal diagram, as described in Box 1, one can determine potential sources of omitted variable bias by looking for variables influencing both the cause of interest and outcome, but which have not been observed in the system (e.g., Z in Fig1B). 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. Including a variable in your analysis that blocks all paths between X and Y via Z means that your ensuing analysis will satisfy the **back-door criterion** (Pearl 1995, Fig. 2A). Depending on the causal structure of the system, there may be many variables that can “shut the back door” and ensure that an analysis will be causally identified (see Fig 2. and caption for several examples). Without including such a variable in a statistical model, the omitted, confounding variable will cause omitted variable bias.

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 wide variety of network analysis software for 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 statistical controls 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 handling omitted variable bias. It shows us where OVB might influence our modeled results but does not in and of itself provide a means for controlling for OVB if we do not have a control variable measured. Nor does a causal diagram help us in the face of unknown omitted 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 our statistical models with the data to produce causally identified results.

**A Problem of Omitted Snails**

To illustrate these empirical challenges and suite of potential solutions, we consider a marine benthic ecosystem, such as the Gulf of Maine, USA, where we aim to study the causal relationship between temperature and snail abundance. Let us say that both temperature and recruitment influence the abundance of snails as shown in (Fig. 3). As temperature influences metabolic and mortality rates, we hypothesize that fewer snails can survive in hotter sites. At the same time, the region’s oceanography determines temperature patterns and recruitment of new juvenile snails (Fig. 3). We assume that the researcher measured snail abundance and temperature at a number of sites, but not recruitment or any measurement of oceanography. This creates problems of omitted variable bias in any statistical analysis. Further, what if this diagram is wrong, and there are other lurking omitted variables? Then, even if oceanography were accounted for, omitted variable bias is a real possibility. Fortunately, we can deal with this issue and estimate the effect of temperature using appropriate sampling and statistical designs. As we will show below, using the proper sampling and statistical design based on the dynamics of temperature and oceanography in the system allow for the estimation of causal effect of temperature on snail abundance. If these methods are not used - even given additional modeled sources of variation in temperature or recruitment - then the estimation of the effect of temperature on snails will be incorrect.

## **Sampling Designs to cope with omitted variable bias**

There are multiple sampling designs that can address omitted variable bias. Which sampling design a researcher should use, and how to implement it, will depend on the way the omitted variable affects the causal variable and outcome of interest. These choices arise because each design makes assumptions about the spatial and temporal scale at which an omitted variable affects the system. We assume here that the researcher cannot, has not been able to, or does not know to measure an omitted variable. If they had, the inclusion of a covariate would eliminate the potential for omitted variable bias.

The key element of these designs is the **nesting** of measurements within a cluster such that the causal variable of interest varies across the smallest level of replication while the omitted variable varies at the cluster level. *Clustered data is often also referred to as a hierarchical or nested sampling. We use these terms interchangeably.* The use of a **cluster**, e.g., site, year, block, subject, individual, with multiple measurements taken per individual cluster (Fig. 4), can allow us to flexibly account for confounding variables, whether or not we know what they are. To explain how, let us consider these concepts with an eye towards our snail example, assuming that temperature and recruitment are inversely correlated, in the following study designs.

First, we consider a sampling design where plots are sampled within sites across an environmental gradient in a single year (Fig. 4A). This is a **cross-sectional design**, i.e., with multiple places/clusters sampled in a single year. When sites span large environmental gradients, there are many reasons that sites could differ (e.g., temperature, productivity, oceanographic conditions). If oceanography differs between sites, given that it affects both temperature and recruitment in our example (Fig. 3), we have a potential for confounded results if we only have one sample per site. We can still get a causal estimate of temperature, however, if, within a site, plots vary in their temperature relative to the site average due to factors that do not covary with recruitment. We can use this *within*-site variation in temperature to isolate temperature’s effects on recruitment from confounding effects of other variables that affect temperature and recruitment *between* sites.

This design will not produce causally valid coefficient estimates, however, if there is little to no site-level variation in the causal predictor of interest. If, however, there is temporal variation in temperature within sites, independent of fixed between-site differences in recruitment, we can use repeated sampling of sites to produce causally valid inference. Here we would use a **longitudinal design**, sometimes referred to as a **panel design**, where sites are sampled repeatedly through time, allowing us to capitalize on temporal variation decoupled from recruitment to remove the signal of our omitted variables (Fig. 4B).

These two sampling designs are generalizable beyond just the simple case presented in our example. One can imagine if temperature and recruitment varied temporally at a regional scale, but sites had additional variation in temperature, how to adapt the above designs. Further, if there were additional omitted variables that affected snail abundance at different levels of spatial or temporal hierarchy, we could incorporate more elaborate clustering procedures to produce data that could lead to causally identified statistical models.

Finally, while it seems difficult - and painfully realistic - some omitted variables could vary by both space and time. The sampling designs for coping with spatio-temporal omitted variables are based on the same principles as above. For example, consider our snail scenario. Assume cold sites have higher recruitment than warm, but, at the same time within a site, years that are colder also have higher recruitment than those that are warm. This spatio-temporal omitted variable can be dealt with as long as the omitted variable works at the site-year level and there is variability within a site within a year for the causal variable of interest. One can then observe plots within a site over time to ultimately control for OVB. If the omitted variable and the driver of interest vary at the same scale, it might not be possible to control for OVB for this particular driver, and a researcher would have to fall back on measuring the omitted variable or conducting experiments. Otherwise, “nothing to be done” (Beckett 1954).

Each of the above sampling designs flows naturally from the intuition about a system generated from a causal diagram. This is why building a causal diagram at the outset of designing an observational study is key. Regardless, even without a causal diagram in hand, creating observational study designs using cross-sectional, panel, or combined approaches as a matter of course will enable better estimates of causal effects. Combining these designs with others, such as the classic stratified random sampling design or others (Stevens and Olsen 2004, Robertson et al. 2013, Grafström and Lundström 2013, Foster et al. 2018, Kermorvant et al. 2019), will allow for the analyses that are not only causally identified, but also reduce the influence of variability of uncorrelated variables when estimating causal relationships.

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

Once the appropriate sampling design has been implemented, there are multiple, well-established statistical designs for analyzing clustered data (i.e., panel data) data. We emphasize the term *‘designs*’ over *‘methods,’* because one could implement these models using different estimation approaches (e.g., linear models, as part of Structural Equation Models, Bayesian techniques, etc.). These different designs have different costs and benefits -- and differ in their assumptions required for interpreting an estimate as a causal relationship. We believe these models are a key advance worth considering for ecologists. Further, each of the models we outline allow us to flexibly control for confounding variables that are both known and unknown (see Angrist and Pischke 2008, Ferraro and Miranda 2017, Dudney et al. 2021) – something many Ecologists worry about.

We illustrate the different designs using a common set of terms for predictors (x), responses (y), and confounding variables (z) in a regression, applied to our example of the snail system in Figure 3 with different sites (i) sampled either at multiple time points (j) in a panel design or in multiple plots (j) in the case of a cross-sectional design. For the sake of simplicity, we assume a linear model form with normally distributed error (e) such that if we had measured the confounder, we could fit the following model and get a causally identified estimate of the effect of x on y.

(1)

$y**\_{**ij**}** = **\beta\_0** **+** **\beta\_1** x**\_{**ij**}** **+** **\gamma** z**\_**i **+** e**\_{**ij**}$**

This can of course be extended to generalized linear modeling frameworks.

Our goal is to estimate $\beta\_1$. Note, going forward, for some statistical designs where we replicate within sites across time, we will also include an index for replicates within a site (k). If the confounder was not measured in the model above, and instead a researcher tried to fit the following model:

y_{ij}=\beta_0+\beta_1x_{ij}+e_{ij} (2)

(2)

$y**\_{**ij**}** = **\beta\_0** **+** **\beta\_1** x**\_{**ij**}** **+** e**\_{**ij**}$**

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** violated 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 model designs have been all the rage in ecology for the past two decades due to several desirable features (for some useful reviews 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 and Hill 2006 on the linguistic difficulties of defining random effects) - quickly became a mainstay in the partitioning of variation in experiments with subsamples within clusters (Cochran 1937, Eisenhart 1947). They have become a standard part of the toolbox for analyzing ecological experiments (Schielzeth and Nakagawa 2012).

In observational data, random effects allow us to account for clustering in our data. This could of course be done with clusters as a **fixed effects** (i.e., recoding clusters to 0/1 using a contrast structure), although in such a case one might need to use cluster robust standard errors to accommodate for clustered error structures ((see Box 4 and Cameron and Miller 2015). Random effects, however, have the added second benefit of efficiency - they cost fewer degrees of freedom to estimate as we assume all cluster means follow from a distribution. With this can come an improvement in the estimates of *precision* for coefficient estimates (Gelman and Hill 2006) relative to fixed effects. This can contrast to changes in precision from cluster robust standard errors (see Oshchepkov and Shirokanova 2022 for an excellent comparison between mixed models and CSRE). 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 (see an excellent discussion by Efron and 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 model statistical design:

(3)

$$y\_{ij} = \beta\_0 + \beta\_1 x\_{ij} + \delta\_i + \epsilon\_{ij} \\

\delta\_i \sim \mathcal{N}(0, \sigma^2\_{site}) \\ \\

\epsilon\_{ij} \sim \mathcal{N}(0, \sigma^2)$$

Here, yij is the abundance of snails at site *i* in year *j*, $\beta\_0$ 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!), $\beta\_1$ is the effect of temperature x at site i in year j on snails, $\delta\_i$ is the site-specific deviation at site i from our intercept due to random variation which follows a normal distribution and $\epsilon\_ij$ is the residual variability for snail abundance at site *i* in year *j.* While site is incorporated into $\delta\_i$ here, the statistical design is not causally identified due to the violation of the **Random Effects Assumption** and will produce biased estimates of $\beta\_1$.

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

So why will the design above produce incorrect results (i.e., a statistically biased estimate of the causal effect)? 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 and 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 our random effects do not correlate with our fixed effects (Wooldridge 2010, Antonakis et al. 2021).

Here, the random effects of ‘site’’ are assumed to be uncorrelated with temperature for an unbiased estimate. This is due to how random effects are estimated - as a part of the error term of the model (Wooldridge 2010, Schielzeth and Nakagawa 2012). Indeed, if we were uninterested in modeling the site-level means, we could combine $\delta\_i$ and $\epsilon\_{ij}$ into $u\_{ij} = \delta\_i + \epsilon\_{ij}$ and estimate the model with ordinary least squares as long as there was no omitted variable bias. Given that our causal variable of interest varies at the site level in a way that is confounded with other drivers, however, we can easily see that $u\_{ij}$ is not independent of our $x\_ij$ values. This is a problem and violates the basic assumptions of Ordinary Least Squares estimation and any extension.

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 something more like Figure 5b. Here we can see that 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 in order to use it. That is not what we have done with the mixed model design above. We have fit data to the causal diagram show in Figure 5a. This example illustrates the difficulty in satisfying the *Random Effects Assumption; i*ndeed, satisfying this assumption is often quite difficult in Ecology – particular in observational data that spans environmental gradients - and how badly we violate the assumption is not well explored or acknowledged widely enough. We need a better solution.

*Enter the Econometric Fixed Effects Design*

The Fixed Effects Design represents a great and familiar starting point for many ecologists. 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 on panel models, where it refers to the effect of a cluster-varying attribute of the system (i.e., a within cluster intercept). 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 of predictor variables that are estimated directly, rather than as part of the error term. 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 (Gelman and 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** or **fixed effects transformation** to create a model that can be fit simply. Given that the recruitment effect in our example is time invariant, we can transform the data before modeling in order to eliminate it. Consider the following mathematical description of the system.

(4)

y\_{ij}= \beta\_0 + \beta\_1 x\_{ij} + \gamma z\_i + \epsilon\_{ij}

We can average this equation over all time points at each site to get the following

(5)

\bar{y\_i} = \beta\_0 + \beta\_1 \bar{x\_i} + \gamma z\_i + \bar{\epsilon\_i}

If we subtract this average value at each site across the all years, as shown above, we cancel out the site-level omitted variables.

(6)

y\_{ij} - \bar{y\_i} = \beta\_1 (x\_{ij} - \bar{x\_i}) + (\epsilon\_{ij} - \bar{\epsilon\_i})

Using simple algebra, we have removed the confounding influence of omitted variables, as all site effects have been removed. Thus, a simple model with snail deviation from site mean as a response and temperature deviation from site mean, as seen in Figure 6a, will prove sufficient. The transformation has removed any paths from site or site-correlated drivers to the response. We note that cluster robust standard errors are likely important here for inference (see Box 4 and Cameron and Miller 2015, Abadie et al. 2017) as data points within a site are still likely to be more similar to one another than between sites. For an excellent ecological example of using this technique, see Dudney et al. (2021).

The fixed effect transformation does have some drawbacks, despite its simplicity and its strength in controlling for both observed and unobserved confounding variables. For one, we lose information about site-level abundances. Further, we cannot use this model for predictive inference outside of the sites observed.

To solve these problems, we can use a design where cluster membership (site in our snail example) is included as a categorical variable. This kind of design - familiar as an ANCOVA-style model design to many Ecologists - will produce identical results to the preceding model for $\beta\_1$. We note that categorical variables are turned into **Dummy Variables** by software (i.e., for replicate i, the value for cluster j is 1 if it is in that cluster, otherwise it is 0). This dummy variable coding allows site to be included as a fixed effect - in both senses of the term. Unlike random effects in a mixed model design, econometric fixed effects designs are not constrained to be drawn from any predefined distribution nor do they refer to a single “fixed" estimated effect for a predictor variable across all units here. A dummy (or categorical) variable is estimated directly in the regression resulting in an estimate for each unit – i.e., in our example site.

(7)

$$y\_{ij} = \beta\_1 x\_{1ij} + \sum\alpha\_i x\_{2i} + \epsilon\_{ij} $$

where is our variable of interest and is the fixed effect at site i, and is 0 or 1 - a dummy variable that is 1 if the site is i and 0 if it is not. Including a site-level fixed effect is essentially controlling for the average “level” of variable per site or subtracting off a site level mean for each variable. This control makes this design equivalent to the within transformation design - and has the same effect in controlling for omitted variable bias (Angrist and Pischke 2008, Wooldridge 2010).

Returning to our example, with site as a fixed effect, we are able to control for different sites having different levels of recruitment or other omitted variables correlated with temperature. Hence, using econometric fixed effects (i.e., dummy 0s and 1s) enables a causally identified estimate of the temperature effect, removing differences among sites that are otherwise confounding. We can represent this in a causal diagram in Figure 5b with site as a variable where we control for correlation between site and temperature. This makes it clearer that we are estimating the effect of temperature controlling for recruitment or any other site covariates.

The econometric fixed effects design has two main drawbacks. First, fixed effects estimators are inefficient compared to random effects - meaning that we need a larger sample size to achieve the same level of precision of our estimates. This is because, for each group/fixed effect (site in our example), we have a corresponding column of dummy 1/0 variables in the model. Thus, we are estimating many more parameters. However, in the case of omitted variable bias, this framework is still preferable over the random effects model as it produces an unbiased causally identified parameter estimate. Second, we lose information about relationships between sites. While the parameter estimate for the temperature effect is causally valid, it is based on variation in temperature within a site. We have coefficients for individual sites, but, if an investigator is interested in gradients between sites (e.g., sites are along a thermal gradient in this example), this approach does not allow for any inference about the effects of these gradients - and other drivers correlated with them - between sites. This lack can be problematic if a researcher wishes to use the model for prediction as well as causal inference – such as predicting snail abundances at new sites not included in our initial study.

## *Group Means For Efficiency, Inference, Fun, and Profit*

To solve the above problems, we can step into the world of **correlated random effects model designs**. In this style of design, we again assume that our confounding variables that correlate with our causal variable of interest vary at the cluster - in this case site - level. The foundation of correlated random effects model designs is leveraging *group means* to get a causally identified estimate of the effect of our predictor of interest. For every cluster – e.g., site, year, region, etc. - researchers calculate a group mean to include as a predictor in the statistical model for this design. This hierarchical predictor variable – a predictor at the cluster level – now acts to control for omitted variables that vary between sites and correlate with our causal variable of interest. Because of this, we can use hierarchical mixed effects statistical models that include a random effect of site and no longer violate the random effects assumption. Further, the coefficient for our causal variable of interest is now estimated while controlling for cluster-level confounding variables - whether they are known or not. This statistical design approach is useful as it allows us to derive causally valid inference about our driver, study the effects of gradients between sites that are correlated with our causal variable of interest via group means, and learn about the variation between sites that is not correlated with our causal variable of interest via the random effects efficiently. As we have random effects, tt also means that our causally identified models can be used in a prediction or forecast setting for sites not included in our analysis.

Correlated random effects model designs come in two forms. First, consider the following model:

(8)

$$y\_{ij} = \beta\_0 + \beta\_1 x\_{ij} + \beta\_2 \bar{x\_{i}} + \delta\_i + \epsilon\_{ij} \\

\delta\_i \sim \mathcal{N}(0, \sigma^2\_{site}) \\ \\

\epsilon\_{ij} \sim \mathcal{N}(0, \sigma^2) $$

where $\beta\_2 \bar{x\_{i}}$ accounts for the effect of cluster-level confounders. In Econometrics, this 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. 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 site mean temperature effect itself 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 design will run into problems, however, in a data set where the correlation between our predictor of interest and its cluster-level mean is too strong for the estimator being used (e.g., r > 0.85 for OLS). To solve this, we can use a model 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. Here we subtract the cluster level mean from the causal variable of interest. This approach decomposes our causal variable of interest into between and within cluster terms. The site mean temperature term is the between site effect, and the anomaly from the site mean term is the within-site temperature effect. We can see this in the following model:

(9)

$$y\_{ij} = \beta\_0 + \beta\_1 (x\_{ij}-\bar{x\_{i}}) + \beta\_2 \bar{x\_{i}} + \delta\_i + \epsilon\_{ij} \\

\delta\_i \sim \mathcal{N}(0, \sigma^2\_{site}) \\ \\

\epsilon\_{ij} \sim \mathcal{N}(0, \sigma^2) $$

The DAG for this design in Figure 6d. You can see the similarities - and the key differences - with previous designs. In particular, the interpretation of $\beta\_2$ is different than in the Group Mean Covariate design. $\beta\_2$ is now a **between estimator** of the combined effect of moving across gradients in temperature and correlated drivers between the sites. This is often a more useful estimand for interpretation ecologists. If $\beta\_2 = \beta\_1$, we might conclude, *tentatively*, that omitted variables are not influencing snails and both our between and within site differences are due solely to temperature.

Most importantly, while the Group Mean Covariate design, Group Mean Centered design, and Fixed Effects design all differ in structure, they are equivalent when it comes to estimating the temperature effect, – the effect of a one unit change in temperature on snails. This is because they rely on within-site variation in temperature. All three should produce similar estimates. Which design you use depends on the structure of your data (e.g., how many coefficients do you feel comfortable estimating with a Fixed Effects design given your sample size) as well as what answers you want to derive from the non-causal terms. Do you just want site means? 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 both within and between-site gradients? Group Mean Centered design. Each design can even be extended to cases where the magnitude of the causal variable of interest’s effect depends on the level of confounding variables (i.e., an interaction effect, see Box 2). The choice of statistical design will dictate which answers are most readily available to a researcher.

## *What a Difference Differencing Makes*

Our examples thus far have focused on omitted confounding variables that 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 use the same framework as above, swapping years for sites as clusters. If omitted confounders vary spatiotemporally, we can extend the framework further using the same principles (see Box 3). What if, however, site-level omitted temporal trends in other drivers covaried with site-level temporal trends in our causal variable of interest? Worse, what if the slope of these trends varied by site?

Consider the following example (Fig. 7A). Not only does site-level temperature covary with recruitment due to the confounding influence of oceanography, but, the abundance of snails is also influenced by coastal development over time at each site. Rates of development are not the same across all sites, and as such it is extremely difficult to separate out local coastal development’s signal from the signal of local temperature variation if a panel design is being used. To see this mathematically, consider a small modification to the dynamics of our system from eq. 1:

(10)

$y\_{ij} = \beta\_0 + \beta\_1 x\_{ij} + \gamma z\_i + \lambda\_i j + e\_{ij} **$**

Here $\lambda\_i$ is a site-specific trend in snails over time. Due to this trend, if there is also a temporal trend in temperature, our estimation of $\beta\_1$ would again be contaminated and our estimate would no longer be causally identified. On the surface, this appears to be a difficult problem to tease apart.

Fortunately, there is a simple solution for this case, and it is related to the fixed effects transformation. The solution is 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, site-level fixed omitted variables drop out. Our site-specific temporal trend remains, and we can accommodate it using dummy variables as before. This site-specific coefficient multiplied by the dummy variable, here x2ij, now represents the linear rate of change at this site that is not related to temperature, and we estimate the effect of change in temperature on change in snails controlling for other linear trends at the site level, as seen in Figure 7B.

(11)

\Delta y\_{ij} = \beta\_1 \Delta x\_{1ij} + \sum \lambda\_i x\_{2ij} + \Delta \epsilon\_{ij}

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. Further, for many studies investigating human-driven changes as their predictors of interest, assuming there is no temporal trend in drivers is often difficult to justify.

If we are uninterested in site specific trends, we can also calculate the second difference - e.g.$ \Delta^2 y\_{ij} = \Delta y\_{ij} - \Delta y\_{i,j-1}$ which eliminates $\lambda\_i$. This design (see Fig. 7C) has the advantage of estimating far fewer parameters if we have many sites, and thus could prove more efficient, although cluster robust standard errors might prove important (see Box 4). Note that $\beta\_1$ in this model is estimating the effects of acceleration in change in temperature. This estimand still establishes a causal connection, but requires more thought in interpretation. It will also require clarity of explanation, as many reviewers can and will miss the subtlety (we speak from experience).

Using either temporal differencing design has several advantages. We again remove the effect of omitted site-level variables. We also remove the effects of site-specific trends that could reflect more dynamic site-level omitted variables. Thus, our estimate of a temperature effect is again causally identified. Indeed, as we are handling two potential forms of omitted variable bias, our analysis would be more robust to omitted variable bias from unknown confounders. Further, this approach shifts the type variation we are studying. We are asking **how *change* in a driver corresponds to *change* in a response.** For the second difference model, we are examining how the **acceleration of a driver corresponds to the acceleration of a response**.

The main drawback of these approaches is the reduced sample sizes. We lose observations from one or two time steps of data. Loss of observations could lead to lower power (i.e., noisier stand errors). This can be even more problematic in the case of the second difference design, although the loss of data could be counterbalanced by the gain in efficiency from estimating fewer parameters. Further, both models assume equal time between sampling events, although this can be fixed by dividing by time between samples to standardize or including an interaction effect with time since the last sample. Despite the sample-size drawback, however, the temporal differencing approach can be incredibly useful in cases of both spatial and temporal omitted confounders - a situation all too common in many real ecological systems.

## **Comparison of Approaches**

To demonstrate the utility and consequences of the preceding solutions, 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 Shiny provided as Appendix B (for a single simulated run) and C (for 100 or more replicate simulation runs 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 of the techniques described above, as well as using naive models with no site effect. We also included group mean covariate and group mean centered models without a random effect in order to demonstrate what a random effect in these models is doing.

Broadly, our simulations show that the point estimate of the RE model - what ecologists typically do - is downward biased in these simulations compared to any other estimate (Fig. 8,9, Table 1). Further, not only is the estimated coefficient of the RE model always biased compared to other estimators in our simulations, but, it more often is within 2SE of 0 than all other models. Analysis of the model under frequentist null hypothesis testing or examination of the 95% confidence interval of the coefficient would lead to type II error. Frequently the 95% confidence interval does not contain the true value of the temperature effect in the RE model as well (Table 2). 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 performance. This is an unrealistic assumption for real data sets, however. 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 as well as the questions they can versus cannot answer (Oshchepkov and Shirokanova 2022)).

## ***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, the causal effect of our variable of interest might depend on the level of the confounder itself. This interaction effect between the predictor of interest and the confounder can have profound consequences if ignored. Consider that thermal effects in our snail system might depend on levels of recruitment - dense aggregations of intertidal organisms are often better at retaining water and thus resisting desiccation or other forms of thermal stress (Fig. 10??, REF). In a naive mixed model, we might incorporate this into a random slope.

$$y\_{ij} = \beta\_0 + \beta\_1 x\_{ij} + \gamma\_i x\_{ij} + \delta\_i + \epsilon\_{ij} \\ \\

\gamma\_i \sim \mathcal{N}(0, \sigma^2\_{site \;slope}) \\ \\

\delta\_i \sim \mathcal{N}(0, \sigma^2\_{site}) \\ \\

\epsilon\_{ij} \sim \mathcal{N}(0, \sigma^2)$$

As before, however, the random effects assumption is violated, making this approach inappropriate for analysis. To rectify the problem of omitted variable bias properly here, however, we have two solutions. First, we can use a fixed effects design. In a fixed effects design, the site effect dummy variable would interact with our causal driver of interest, allowing us to estimate site-specific slopes. 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 driver of interest. For example, for a group mean covariance (i.e. Mundlak device) design, we could use the following model design:

$$y\_{ij} = \beta\_0 + \beta\_1 x\_{ij} + \beta\_2 \bar{x\_i} + \beta\_3 x\_{ij} \bar{x\_i} + \gamma\_i x\_{ij} + \delta\_i + \epsilon\_{ij}$$

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 $\beta\_3$ is not different from 0. We could use a similar model for the group mean centered design if deemed appropriate.

Models with interactions can provide powerful insights into both the effect of the causal driver of interest as well as how those effects vary given ambient conditions. For example, consider a group mean centered model with an interaction effect for our snail-recruitment system. We can ask if the effect of a temperature anomaly differs in warm versus cool sites - something which can prompt follow-up investigations as to what are the underlying differences that correlate with the thermal gradient that could cause temperature anomaly to have differing effects in different sites.

## **Box 3: Reality Bites - Coping with spatiotemporal omitted variables**

Spatiotemporal omitted variables can be extremely challenging, and the solutions can require more thoughtful study and statistical design. Consider a scenario where recruitment is not static through time. Rather, it is correlated with temperature in both space and time. For example, sites that experience strong cold-water pulses in a year also experience unusually high recruitment in those same years. This is on top of colder sites already having higher recruitment. If there is both variability within a site in temperature and we have multiple plots sampled across multiple sites each year, we can cope with this sort of spatiotemporal omitted variable problem in ways that echo the types of models already seen above. For example, we can use a fixed effects design as in the following model with plot within site and time designated as k:

$$y\_{ijk} = \beta\_1 x\_{1ijk} + \sum\alpha\_i x\_{2i} + \sum\lambda\_j x\_{3j} + \sum\nu\_{ij} x\_{4ij} + \epsilon\_{ijk} $$

Here x2i is a dummy variable for site to capture spatial omitted confounders, x3j is a dummy variable for time to capture temporal omitted variables and x4ij is a dummy variable that combines site and time in order to capture spatio-temporal omitted variables. It is possible that the first two are not needed and only the spatio-temporal fixed effect is necessary, which would increase efficiency. Regardless, this style of model can consume degrees of freedom rapidly. For this reason, a more efficient correlated random effects approach can also be used. Here is a model using the group mean covariate desigWe n analogous to the fixed effect version, although we note that terms capturing spatial and temporal omitted confounders might again not be necessary (and, indeed, if they are 0, then we can conclude that OVB is unimportant at these levels).

$$y\_{ijk} = \beta\_0 + \beta\_1 x\_{ijk} + \beta\_2 \bar{x\_i} + \beta\_3 \bar{x\_j} + \beta\_4 \bar{x\_i} \bar{x\_j} + \delta\_i + \delta\_j + \delta\_{ij} + \epsilon\_{ijk} $$

Here the \delta terms are random effects for site, time, and site:time, although, again, some of these could be unnecessary depending on relevant sources of residual variation.

The implementation of the multiple plots within sites sampled over time brings up one additional issue that could be relevant for omitted variable bias - the question of whether plots should be fixed or randomized each time a site is sampled. There are some practical logistical considerations here - it might not be possible to permanently mark or otherwise revisit plots. As such, the above model should be adequate with the assumption that plots are re-randomized at each sampling interval. Fixed plots, however, provide two advantages. First, with respect to omitted variable bias, we know that variation in a driver within a site likely correlates with many other within-site drivers. For example, cooler plots within a site in the intertidal might happen to be shaded by a nearby boulder. With fixed sites, we can add a plot effect into our models to potentially cope with plot-level OVB. Second, for other time series models, fixed plots have the advantage of greater power to detect change, as, even with a random effect, we can remove variation due to plot from our residual error term for hypothesis tests (Urquhart and Kincaid 1999). We emphasize that it 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 jumping off point).

Without a nested data structure - e.g., plots within sites resampled over years - we cannot include a site by year effect as above. We only have a single measure per site and year. There would be no variation left to study! We still have some options, however, although they can be more *ad hoc*. As we are considering spatiotemporal confounders, if we can build structure in our model that accommodates site-specific variation in our confounding variable. In the differencing section above, we discussed that, after differencing, a dummy variable site effect would represent the slope of a site-specific temporal slope. We can use this here and build in that trend explicitly.

y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma z_i  + \sum \lambda_i x_{i} j + e_{ij}

$ y**\_{**ij**}** = **\beta\_0** **+** **\beta\_1** x**\_{**ij**}** **+** **\gamma** x**\_**i + \sum **\lambda\_i** x**\_{**i**}** j **+** e**\_{**ij**}$**

OLD AND I THINK WRONG $ y**\_{**ij**}** = **\beta\_0** **+** **\beta\_1** x**\_{**ij**}** **+** **\gamma** z**\_**i + \sum **\lambda\_i** x**\_{**i**}** j **+** e**\_{**ij**}$**

Where xi is a the dummy variable for site. To accommodate spatio-temporal variation, however, we will need additional nonlinear terms that enable, for example, sites to have individual nonlinear trajectories without eating up all of the degrees of freedom from time. For example

y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma z_i  + \sum \lambda_{1i} x_{i} j +  \sum \lambda_{2i} x_{i} j^2 +  \sum \lambda_{3i} x_{i} j^3 + e_{ij}

$y**\_{**ij**}** = **\beta\_0** **+** **\beta\_1** x**\_{**ij**}** **+** **\gamma** z**\_**i + \sum **\lambda\_{1i}** x**\_{**i**}** j + \sum **\lambda\_{2i}** x**\_{**i**}** j^2 + \sum **\lambda\_{3i}** x**\_{**i**}** j^3 **+** e**\_{**ij**}$**

allows for a cubic fit trend that differs by site. For a practical example, Dee et al. (2016) examined the effects of biodiversity on fisheries yields using Large Marine Ecosystems (LMEs) as spatial units of replication followed through time; they controlled for spatio-temporal omitted variables via squared temporal trends that varied by LME using squared per-LME trends as well as LME fixed effects for intercepts in addition to multiple observed confounders. Similar approaches can likely be taken with site-specific Generalized Additive Models (Wood 2017). Smoothing terms in GAMs, however, are fit in the same manner as random effects, leading to concerns about violating the random effects assumption. Residuals from site-specific GAM effects could be an alternate way to handle spatio-temporal OVB, however, by assigning all variation to the GAM, we risk throwing out some of the signal of casual drivers.

In general, we urge caution when dealing with spatio-temporal omitted variables, and careful use of causal diagrams to ensure that we are controlling for a confounder without throwing out the signal of a real driver. We admit the manuscript has gotten a bit long at this point, but, spatio-temporal confounding variables are very real and very problematic. It is an area that deserves far more exploration in Ecology. For more from other disciplines on this tricky class of problem and approaches outside of the scope of this paper (see Ferraro & Hauner, Athey and Imbens, Oster).

**Box 4: Robust Standard Errors - an Underutilized Tool in Ecology**

For some models above, we recommend the use of clustered robust standard errors. Robust standard errors are not commonly used in Ecology (but see examples in Dee et al. 2016, Dudney et al. 2021) despite being a way to continue to use Ordinary Least Squares and then flexibly apply a post-hoc adjustment SEs to accommodate clustered data, heteroskedasticity, correlation between time points, and other arbitrary correlation structures within the data (Cameron and 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 and 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 and Miller 2015, Abadie et al. 2017).

**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, dear reader, that through thinking carefully about biological systems, we can draw on a solid set of existing methods to produce causally valid 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 one 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, in developing an understanding of how to accommodate OVB from spatial or temporal confounders, Ecologists can see the above concepts as part of a generalizable approach to handling confounding variables. Many types of clusters in a study could have omitted variables lurking around the corner. With a large enough sample size and astute study design, we can structure our research to accommodate multiple suites of confounding omitted variables occurring at multiple levels of clustering. While we have talked of sites and years, consider small-scale studies with cohort effects, individual effects, or lower levels. Consider larger-scale studies with not just sites and years but regions and decades. The framework remains the same, and the potential sources of OVB should reveal themselves through initial causal diagrams.

The approaches we present here are surely not a panacea. Model misspecification can lead to overconfidence that some omitted variable bias problems have been accounted for by these methods when, in truth, they have not. In particular, not fully reckoning with the way omitted confounding variables correlate with our causal variables of interest can produce models that are subtly misspecified - such as thinking that a confounding variable only varies in space, when it varies in both space and time. Moreover, while these methods might aid in accounting for known unknowns, we should always be humble in the face of unknown unknowns. If we are honest with ourselves, there is no full protection from these, other than attempting to ground our work in the blend of theory and natural history that is required for a truly insightful analysis. Accepting that our models are not perfect and that someday, someone will come along with a different one that will produce different conclusions and yield new insights is the cost of doing science. We must embrace creative failure rather than be paralyzed by it.

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 path diagram - even a brief verbal one - 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 - will go far in terms of making your analyses more useful and, to be frank, more robust to a cranky reviewer.

We also emphasize that this paper is a starting point. There are many other methods for producing causal inference in the face of omitted variable bias. We recommend several recent reviews of instrumental variables approaches (Angrist et al. 1996, Kendall 2015, Grace 2021), quasi-experimental approaches (Butsic et al. 2017), and are hopeful to see more on the emerging use of the front-door criterion (Bellemare et al. 2020). We urge ecologists, long grounded in experiments being the gold standard for causality, to open up to writings in Econometrics, Sociology, AI, and other disciplines that cannot always do clean experiments (if they can conduct experiments at all). Embracing this transdisciplinary approach will enable us to increase our breadth of knowledge in the tremendous advances in causal inference. As an incomplete (and one day out of date) set of starting points for the curious, 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 and Pischke 2008), Sloman’s Causal Models (2005), and Pearl et al.’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 cleanest causal inferences needed to grapple with our rapidly changing world to learn how to mitigate, acclimate, and adapt at scale.

## **Conclusions**

The specter of Omitted Variable Bias from unmeasured confounding variables has stymied the use of observational data for causal inference in Ecology for much of its history. “Correlation does not equal causation,” rings in many of our heads from our Biostatistics 101 courses. We have all been there – realizing that an omitted variable might be wreaking havoc with an analysis of hard-won data, feeling the frustration of knowing there is something crucial that you will not be able to measure, or watching a key instrument go up in smoke limiting just what data you are able to collect. We want this guide to serve as a new arrow in the quiver of all Ecologists. It is time to address pressing applied and theoretical questions at scale with the amazing observational data sets now available. It is time to look to other disciplines that have gone through similar bouts of soul-searching about how to derive causal inference from real-world data in an honest and transparent manner. Rather than sweep the problem under the rug and lose valuable knowledge, we hope that you, dear reader, can now move forward with confidence. We look forward to the new insights that these techniques will help you generate.

## **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. This work was partially supported by the National Science Foundation as part of the PIE-LTER Program (award #1637630) and the Stone Living Lab. We thank 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., S. Athey, G. W. Imbens, and J. Wooldridge. 2017. When Should You Adjust Standard Errors for Clustering? Working Paper, National Bureau of Economic Research.

Angrist, J. D., G. W. Imbens, and D. B. Rubin. 1996. Identification of Causal Effects Using Instrumental Variables. Journal of the American Statistical Association:29.

Angrist, J. D., and J.-S. Pischke. 2008. Mostly harmless econometrics. Page Mostly Harmless Econometrics. Princeton university press.

Antonakis, J., N. Bastardoz, and M. Rönkkö. 2021. On Ignoring the Random Effects Assumption in Multilevel Models: Review, Critique, and Recommendations. Organizational Research Methods 24:443–483.

Arif, S., and M. A. MacNeil. 2023. Applying the structural causal model framework for observational causal inference in ecology. Ecological Monographs 93:e1554.

Beckett, S. 1954. Waiting for Godot: tragicomedy in 2 acts. Grove Press, New York.

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

Bolker, B. M., M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J.-S. S. White. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology & Evolution 24:127–135.

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

Cameron, A. C., and D. L. Miller. 2015. A Practitioner’s Guide to Cluster-Robust Inference. Journal of Human Resources 50:317–372.

Cochran, W. G. 1937. Problems arising in the analysis of a series of similar experiments. Supplement to the Journal of the Royal Statistical Society 4:102–118.

Cunningham, S. 2021. Causal inference. Page Causal Inference. Yale University Press.

Dee, L. E., S. J. Miller, L. E. Peavey, D. Bradley, R. R. Gentry, R. Startz, S. D. Gaines, and S. E. Lester. 2016. Functional diversity of catch mitigates negative effects of temperature variability on fisheries yields. Proceedings of the Royal Society B: Biological Sciences 283:20161435.

Dudney, J., C. E. Willing, A. J. Das, A. M. Latimer, J. C. B. Nesmith, and J. J. Battles. 2021. Nonlinear shifts in infectious rust disease due to climate change. Nature Communications 12:5102.

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

Efron, B., and C. Morris. 1975. Data Analysis Using Stein’s Estimator and its Generalizations. Journal of the American Statistical Association 70:311–319.

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

Ferraro, P. J., and J. J. Miranda. 2017. Panel Data Designs and Estimators as Substitutes for Randomized Controlled Trials in the Evaluation of Public Programs. Journal of the Association of Environmental and Resource Economists 4:281–317.

Fisher, R. A. 1919. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. Earth and Environmental Science Transactions of The Royal Society of Edinburgh 52:399–433.

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

Gelman, A., and J. Hill. 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.

Grace, J. B. 2021. Instrumental variable methods in structural equation models. Methods in Ecology and Evolution 12:1148–1157.

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

Grafström, A., and N. Lundström. 2013. Why Well Spread Probability Samples Are Balanced. Open Journal of Statistics 3:36–41.

Griffith, G. J., T. T. Morris, M. J. Tudball, A. Herbert, G. Mancano, L. Pike, G. C. Sharp, J. Sterne, T. M. Palmer, G. Davey Smith, K. Tilling, L. Zuccolo, N. M. Davies, and G. Hemani. 2020. Collider bias undermines our understanding of COVID-19 disease risk and severity. Nature Communications 11:5749.

Harrison, X. A., L. Donaldson, M. E. Correa-Cano, J. Evans, D. N. Fisher, C. E. D. Goodwin, B. S. Robinson, D. J. Hodgson, and R. Inger. 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\*. The Quarterly Journal of Economics 115:45–97.

Hernan, M. A., and J. M. Robins. 2023. Causal Inference: What If. CRC Press, Boca Raton.

Holland, P. W. 1986. Statistics and Causal Inference. Journal of the American Statistical Association 81:945–960.

Imbens, G. W., and D. B. Rubin. 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., F. D’Amico, N. Bru, N. Caill-Milly, and B. Robertson. 2019. Spatially balanced sampling designs for environmental surveys. Environmental Monitoring and Assessment 191:524.

Kimmel, K., L. E. Dee, M. L. Avolio, and P. J. Ferraro. 2021. Causal assumptions and causal inference in ecological experiments. Trends in Ecology & Evolution 36:1141–1152.

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

Laubach, Z. M., E. J. Murray, K. L. Hoke, R. J. Safran, and W. Perng. 2021. A biologist’s guide to model selection and causal inference. Proceedings of the Royal Society B: Biological Sciences 288:20202815.

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

Morgan, S. L., and C. Winship. 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., and A. Shirokanova. 2022. Bridging the gap between multilevel modeling and economic methods. Social Science Research 104:102689.

Pearl, J. 1995. Causal Diagrams for Empirical Research. Biometrika 82:669–688.

Pearl, J. 2009. Causality. Cambridge university press.

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

Rinella, M. J., D. J. Strong, and L. T. Vermeire. 2020. Omitted variable bias in studies of plant interactions. Ecology 101:e03020.

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

Robins, J. 1989. The control of confounding by intermediate variables. Statistics in Medicine 8:679–701.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66:688–701.

Rubin, D. B. 2005. Causal Inference Using Potential Outcomes. Journal of the American Statistical Association 100:322–331.

Schielzeth, H., and S. Nakagawa. 2012. Nested by design: model fitting and interpretation in a mixed model era. Methods in Ecology and Evolution 4:14–24.

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

Stevens, D. L., and A. R. Olsen. 2004. Spatially Balanced Sampling of Natural Resources. Journal of the American Statistical Association 99:262–278.

Textor, J., B. van der Zander, M. S. Gilthorpe, M. Liśkiewicz, and G. T. Ellison. 2016. Robust causal inference using directed acyclic graphs: the R package ‘dagitty.’ International Journal of Epidemiology 45:1887–1894.

Urquhart, N. S., and T. M. Kincaid. 1999. Designs for Detecting Trend from Repeated Surveys of Ecological Resources. Journal of Agricultural, Biological, and Environmental Statistics 4:404–414.

Wood, S. N. 2017. Generalized Additive Models: An Introduction with R, Second Edition. Second edition. Chapman and Hall/CRC, New York.

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, August 17. Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators. SSRN Scholarly Paper, Rochester, NY.