**I think we have some pretty inconsistent terms used throughout and made a running list:**

**Casually valid – lets rethink that rephrasing… things aren’t causally valid, or not causally valid, it matters what one is willing to assume. Also its usually talked about as causal identification.**

causally valid results.

**Consistent terminology**

Study designs, designs – over “framework” -- Vs. models vs statistical models vs regressions

Study design vs estimation procedure or methods.

When OV is in the model, it’s not an omitted variable, so lets call it a cnfounding variable vs omitted variable

Analyses vs inferences

Causal diagrams – do they provide crtieria?

Causally valid – I think the correct term is causal identification or identified.

Estimator vs estimand.

**clustered sampling design – or is it hierarchical?**

I dont think we should discuss SEM symbols when introducing DAGs because it is conflating steps in the causal analyes and pertepetuating the misconception that just by doing an SEM ecologists are doing are casual model

# 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**

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 addressing causal relationships - either for testing of basic theory at scale or for informing 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 (reviwed in Kimmel et al 2021), and instead must be able to seize the opportunity of new large-scale sources of observational data.

{{this 2nd paragraph needs a look/rework at the organization and structure}}

Our ability to test causal hypotheses and uncover causal relationships in observational data is fundamentally limited by two things. First, we are limited by our ability to imagine how the different elements of complex ecological systems are linked together. Second, disentangling causal relationships in observational data poses challenges due to a myriad of known, and potentially unknown, confounding variables. ~~armed with this understanding, the use of study designs and analytic techniques that can help us derive causal inference from observational data~~. These issues create a single large problem - one that is often meant when one invokes the old chestnut “Correlation is not causation.” This problem is the absence of measurement of key factors in observational data sets creating what is known as **omitted variable bias** in our analyses [(Wooldridge 2015, Rinella et al. 2020)](https://www.zotero.org/google-docs/?IlGSYx). In short, by omitting variables that influence both a predictor and response of inference - known as a **confounding variable**- our estimate of the relationship between the predictor and response is biased away from its true value. This bias could be positive or negative, and we have no way of knowing its direction.

Omitted variable bias (OVB) occurs when a causal variable of interest and the outcome is correlated with one or more other confounding variables that is not measured and thus is not inclued and controlled for in statistical models of a system (Fig. 1). Unlike variables that do not influence both our causal variables of interest (Fig. 1A), these confounding variables (Fig. 1B) end up being included in the error term. If this variable is omitted from the model, and thus in the error term, we induce correlation between the causal variable of interest and the error term. This correlation creates statistical bias - the estimate of the parameter either being greater or less than its true value. Estimates reflect the joint influence of your causal variable of interest on the response and the confounding driver as well. If both drivers have the same sign, then the effect will be systematically overestimated. If they are of opposite sign, the effect will either be suppressed or, worse, look as if it is opposite in sign to its true influence. Regardless of the direction of bias, the effect of the correlated omitted driver will be attributed to your predictor of interest. Bias in estimates makes the results of any statistical model invalid for causal inference. The model is not **causally identified** [(Pearl 2009)](https://www.zotero.org/google-docs/?Mvs03c).

What might an omitted variable be? These missing measurements might be known factors or, perhaps more commonly, unknown factors of importance that we do not invoke due to a failure of our own imagination. For example…. Indeed, measuring, controlling for, or even knowing all potential confounding variables is nearly impossible in complex ecological systems. In essence, in observational data collection and analysis, we are always going to miss something. However, rather than to throw up our hands discounting and abandoning the use of observational data for causal inference because of this fact, there is an opportunity to we should rather work to understand the solutions to the grand problem of omitted variable bias in causal data analysis that other disciplines have been building for decades.

Omitted variable bias is commonly dealt with in two 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)](https://www.zotero.org/google-docs/?zJuy8W), they do not influence our estimate of causal effects of the treatment (or variable) of interest. In observational studies, however, 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. Alternatively, 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)](https://www.zotero.org/google-docs/?YTvZUJ). We feel, however, that given our current need to understand causal relationships from large-scale observational data sets, Ecologists have both an opportunity and, nay, obligation, to leverage or at least consider? the solutions to the grand problem of Omitted Variable Bias in causal data analysis that other disciplines have been building for decades.

Omitted variable bias is not a new problem in science nor statistics. OVB is widely recognized to the point of obsession in other fields. Fields such as psychology, economics, education, sociology, and more have been grappling with it for some time (REFS). 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 on OVB and instrumental variables)](https://www.zotero.org/google-docs/?UccdlG). This difference could be due to Ecologists having fewer barriers to conducting experiments while other fields (e.g., public health) often cannot perform experiments for logistical or ethical reasons. You cannot replicate a country. You cannot begin to imagine, let alone measure, all of the forces that shape whole economies. One can only tweak curricula so far in an effort to understand educational outcomes. Yet, these disciplines have been tasked with coming up with causal inferences based on observational data that surely has omitted variables confounded with predictors 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 laying out criteria for understanding when and where omitted variable bias could be important. 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 and simulations, we demonstrate the conclusions that would be drawn from the typical approaches ecologists take to this data - and why they fall short of dealing with OVB – compared several other statistical designs that eliminate omitted variable bias. We present results from simulation analyses showing the consistency of these designs that are more robust to OVB robust methods,. We also 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 causal models and transparency in their assumptions as a foundational principles for identifying when OVB might cause problems as well as determining solutions. As applied researchers, we have found that, rather than creating confusion with complexity, graphical causal models paired with robust estimation 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.

**​​Using DAGs to clarify our causal understanding and assumptions and ferret out OVB**

Causal diagrams are one of the first tools for identifying and addressing omitted variable bias by visualizing when and where OVB could be a problem for your inferences . If possible, we recommend making a diagram before designing an observational survey for data collection to inform which covariates to collect data on. However, we are increasing analyzing existing data – and thus it may not include measurements of every potential confounding variable for the question of interst; for instance, if the data was collected for another purpose or question. In these cases, we suggest that a causal diagram These causal diagrams should include both measured and *unmeasured* confounding variables. Further, it can also show what variables you should \*not\* be controlling for in order to produce causally identified results [(for an excellent discussion, see McElreath 2020 chapter 6 or, Griffith et al. 2020 for examples in the analysis of risk factors for Covid-19)](https://www.zotero.org/google-docs/?I9M91G).

Figure 1 presents a causal diagram. For the sake of simplicity, we consider causal diagrams with no feedbacks - so-called Directed Acyclic Graphs, or DAGs [(Pearl 2009)](https://www.zotero.org/google-docs/?fjHHri); also see a recent review in XXXX for ecologists. A DAG is a visualization of qualitative causal assumptions on which one relies for making causal claims from observable data. It is one of the most powerful tools we have in our arsenal to create sampling programs and analyses that will allow us to derive valid causal inferences from observational data. One might blanch at this and request that feedbacks be included but, what we term feedbacks can often be handled by thinking about a system with a temporal lag [(e.g, Larson et al. 2008)](https://www.zotero.org/google-docs/?dv4soW) or, if an instantaneous feedback is truly present, then one will likely require other tools such as instrumental variables - something beyond the scope of this manuscript [(but see Kendall 2015)](https://www.zotero.org/google-docs/?fzmUIJ). We note that, even with feedbacks (which we caution against unless necessary!) causal graphs will be able to elucidate when there are problems of omitted bias so that they can be properly fought against.

**Causal diagrams differ from path analyses or SEM.. they can be used to estimate the DAG but I think this is equating them too directly.**

For the variables and paths themselves, let us adopt the symbology common in Structural Equation Modeling [(Bollen 1989)](https://www.zotero.org/google-docs/?jGMFqo), as it provides a useful language for diagramming a system. There are others, but the core concepts of how we use them are fairly transportable between notations. First, we have observed variables - that can be and are tangibly measured. We will represent these as terms within boxes as X and Y in figure 1. Second, the DAG shows unobserved and conceptual variables, shown within ellipses. They might be latent variables that represent a wide swath of variables that are collected into a single concept. For example, both uncorrelated error (e) and the unmeasured variable (Z) in both panels of Figure 1. 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.

Once you build your causal diagram, you can determine whether you have an omitted variable bias problem and begin to troubleshoot. What you are looking for is instances where a driver that you are **not** interested in that it has an indirect effect mediated through the driver you are interested in (e.g., Z has an indirect effect on Y via X in Fig. 1B). Not controlling for this shared influence opens a “back-door” for confounding effects between a potential cause and an effect. 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)](https://www.zotero.org/google-docs/?xeP5v1) and will be causally identified (Fig. 2A). Variables that directly influence both a cause and effect of interest must be controlled (or the paths “blocked”) to isolate a causal effect from a confounding one. Neglecting them is the *prima facie* case of omitted variable bias. Missing this type of variable is the stuff of nightmares when presenting an analysis to colleagues or critical reviewers. This simple case is not the only way that omitted variable bias can cause problems, however (e.g., Fig. 2D).

Notably, the estimator of the relationship between the variable you use to shut the back door and your response of interest *might have no direct causal meaning* (e.g., Fig. 2C and 2D). A path from W to Y in these models would have no direct causal meaning, although it would allow us to estimate the causal relationship between X and Y. While this might seem odd, unless you are specifically interested in the relationship between that control variable and the response, it is not concerning; you must be aware of this fact when discussing your results, however.

Causal diagrams allow us to detect a broader class of cases that must be accounted for in analyses with multiple predictor variables to avoid omitted variable bias. Many drivers in a system can influence both a cause and effect while lacking a direct connection to one or both (Fig. 2B-D). Without a causal diagram, it can be difficult to understand whether the influence of these variables must be controlled for somehow. 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)](https://www.zotero.org/google-docs/?Kt2R9B) to find open back-doors that need to be controlled for in order to eliminate omitted variable bias.

In short, causal diagrams help visualize the assumptions and potential sources of omitted variable bias for a given analysis. They are a vital tool in any causal analysis of observational data. This is not to say that if a researcher has a causal diagram in hand their analysis is guaranteed to be correct. If their hypothesized causal diagram is wrong, their analysis might still be incorrect and adjusting for known omitted variables still might be insufficient. Indeed, “All models are wrong, but some are useful,” (Box 1976) just as “All experiments are right, but some are useful,” (J.J. Stachowicz pers. com. 2006).

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 of these issues, we must consider the design of our observational studies (if possible) and how we build our statistical models with the data these studies produce.

**A Problem of Omitted Snails**

To illustrate these empirical challenges and suite of potential solutions, we consider a system where both temperature and recruitment influence the abundance of snails in a marine benthic ecosystem (Fig. 3), such as the Gulf of Maine, USA. In this system, we aim to study the causal relationship between temperature and snail abundance. Temperature influences metabolic and mortality rates, and we e fewer snails can survive in hotter sites (REF?). At the same time, making a causal diagram of this system reveals that the same oceanographic influences that shape temperature also shape recruitment of new juvenile snails (Fig. 3). Let’s say you have measured both snail abundance and temperature at a number of sites, but not recruitment. Were there to be no other driver of either recruitment or temperature, this would be an intractable problem. We can estimate the effect of temperature under two different scenarios, however, using appropriate designs. If drivers of variability in temperature at the within-site scale in the case of a **cross-sectional study** (sampling sites at a single time point) with multiple plots sampled per site as a **clustered sampling design** for proper estimation. Or, we can estimate the effect if there is variability in temperature across years in the case of a **longitudinal study** (repeated sampling sites or other units over time, also known as **‘panel data**’ in other fields) where we take one or more measurements per site (e.g., we might want to employ a clustered design for other reasons). As we will show below, these scenarios allow for the estimation of causal temperature effects, provided proper sampling designs and methods are used. If these methods are not used - even given additional sources of variation in temperature or recruitment - then the estimation of the effect of temperature on snails will be incorrect.

Depending on how temperature and recruitment are correlated, statistical estimates of the effect of temperature on snail abundance will be **biased**. If they have the same sign of effect, then estimates of the temperature effect will be too high. If they are opposite in sign, estimates will be biased towards zero or even have the wrong sign. If one has an effect and the other does not, your model could produce a false positive. This will occur no matter how many other covariates you measure and include if those covariates are not part of the confounding pathway. Further, while in this example we will consider recruitment as the only other omitted variable in the system, it is of course possible that other oceanographically driven omitted variables also play a role in regulating snail abundance. Regardless, there are still ways to resolve your omitted variable bias problem.

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

There are multiple study designs that researchers can use to address omitted variable bias. Which design a researcher should use, and how to implement it, will depend on the way the omitted variable affects the response variable of interest, as each makes assumptions about how the 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, as otherwise inclusion of a covariate could eliminate the OVB.

There are multiple, well-established statistical designs for analyzing panel or clustered data. We emphasize the term *‘designs*’ over *‘methods,’*  because one could implement these models using different methodological/estimation approaches (e.g., linear models, as part of Structural Equation Models, Bayesian techniques). 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)](https://www.zotero.org/google-docs/?pV3bIR) – something many Ecologists worry about.

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 (Figure X), can allow us to flexibly account for confounding variables, whether or not they are measured. 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 data context where plots are sampled within sites across a environmental gradient in a single year (Figure 4X). This is a cross-sectional dataset, 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). When a variable affects both temperature and recruitment, it could be confounding if omitted from the regression. their (i.e., a colder or warmer site) We can use this within-site variation in temperature to isolate its effects on recruitment from confounding effects of other variables that affect temperature and recruitment *across* sites (or ‘between’ site differences). Said another way, the variation between sites is likely more subject to confounding variables than the variation *within* sites.

~~Including site as a categorical variable in our regression or an average If t site-level temperature can then be used to shut the back door on the recruitment effect or other drivers that covaries with temperature at the site level.~~

Important confounding variables could vary at the same spatial scale (e.g., at plot-level in Figure 4) as our causal variable of interest. For instance, i,, andses by plot and year. Alternatively, if recruitment is uniform across space, but cold years have high recruitment and warm years have low recruitment in our snail example, we can use a temporal version of the spatial designs above.

In these case, it is particularly helpful to have longitudinal data (i.e., panel data) where we sample each unit (plots) within a cluster (sites) through time.

We next discuss ways to exploit within site (cluster) variation across space and time to deal with confounding variables and omitted variables bias, that can be done in either cross-sectional or longitudinal data when observations (samples) are nested in space and/or time.

~~The above two solutions work well for omitted variables that covary with a driver of interest across a single type of cluster -~~

While it seems difficult - and painfully realistic - that some omitted variables could vary by space and time, the strategies for coping with this type of problem are the same as above and quite flexible. From our snail example, assume cold sites have higher recruitment than warm, but, at the same time within a site, years that are colder 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-year for the driver of interest. **One can then observe plots within a site over time in order to ultimately control for OVB.**

We recognize that one or more omitted variables might have influences at different levels of clustering. Some might vary by time, some by space, and some at different levels of each. This could require a clever design for proper levels of nesting, or even using different types of clustering for different sources of OVB. 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 that use nested designs as a matter of course will enable better estimates of causal effects. Combining these techniques with others, such as the classic stratified random sampling design or others (SCOTT REFERENCES AND THE LIKE), will allow for the analyses that are not only causally identified?, but reduce the influence of variability of uncorrelated variables when estimating causal relationships.

Next, we discuss how addressing omitted variables can be approached, differentiate between different models, and outline additional assumptions that must be met for them to be valid – i.e., yield an unbiased estimate of an effect.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 at multiple time points (j). For the sake of simplicity, we assume a linear model form with normally distributed error (e) such that

y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma z_i + e_{ij}

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

which can of course be extended to generalized linear modeling frameworks. Our goal is to estimate $\beta\_1$. Note, for some models, we will also assume replicates within a site (k). Extensions to cross-sectional sampling designs will either be discussed or are easily related to the examples and models below (i.e., replace time points with replicate plots within each site at a single time point). For some models, we recommend use clustered robust standard errors to flexibly handle heteroskedasticity or correlation between time points [(Cameron and Miller 2015, Abadie et al. 2017)](https://www.zotero.org/google-docs/?q6TKdE) as well as allow for arbitrary correlation structures within clusters. 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 a comprehensive review in [(Cameron and Miller 2015, Abadie et al. 2017)](https://www.zotero.org/google-docs/?q6TKdE).

## *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 [(Bolker et al. 2009, Schielzeth and Nakagawa 2012, Harrison et al. 2018)](https://www.zotero.org/google-docs/?17mZet). Originally used to partition variation in heritability between different relatives [(Fisher 1919)](https://www.zotero.org/google-docs/?r7eHpT), random effects quickly became a mainstay in the partitioning of variation in experiments with subsamples within clusters [(e.g., Cochran 1937, Eisenhart 1947)](https://www.zotero.org/google-docs/?HaV6pc) and have a standard part of the arsenal of analysis of ecological experiments [(e.g., Schielzeth and Nakagawa 2012)](https://www.zotero.org/google-docs/?ChF2Vd). They can also account for the types of hierarchical study designs such as those discussed here. By partitioning variation between different levels of sampling hierarchies, they can improve the *precision* for coefficient estimates (Gelman and Hill 2006). Further, mixed models have several other properties that have made them popular in Ecology. First, they account for non-independence of measurements. This could be done with clusters as a fixed effect, but, random effects have the added second benefit of efficiency - they cost fewer degrees of freedom to estimate (REF) as we assume all cluster means follow from a distribution. Because of this assumption, random effects have a third benefit of recognizing that different data points from different clusters are not from wholly different populations. Cluster means do not have to be estimated as if there is no other information in the data about their possible values (see McElreath 2020 for a discussion of models with retrograde amnesia). This property enables a model to share information between clusters, aiding in the estimation of cluster means in unbalanced designs. It also creates shrinkage of cluster means towards a grand mean, as we have more information than is just contained in the sample of that cluster alone. This is a feature of the technique [(see an excellent discussion by Efron and Morris 1975 as to how this works with respect to baseball statistics for a beautifully clear explanation)](https://www.zotero.org/google-docs/?pDHqWp). For these reasons, Ecologists conducting a study akin to our snail-temperature study would likely gravitate towards a mixed model to account for site-to-site variability in snail abundances.

It is key to remember, however, that when we model random effects, we are no longer modeling group means. 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)](https://www.zotero.org/google-docs/?DkQbgA). This difference results in many 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)](https://www.zotero.org/google-docs/?UR0Cvx). We demonstrate how violating this assumption plays out for our snail study using a mixed model specified as:

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)


$$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 i*n 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.*

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

In this mixed model design, 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)](https://www.zotero.org/google-docs/?3uXRNW). 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. It is immediately apparent, however, that $u\_{ij}$ is not independent of our $x\_ij$ - likely resulting in bizarre plots between predicted and residual values.

We can see more clearly how a mixed model would violate the random effects assumption using a path diagram in Figure 4. 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 the statistical modeling problem at hand, again representing unmeasured quantities in ellipses, what we actually have is something more like Figure 4b. Here we can see that while a random site effect would be wonderful in terms of efficiency, if we could somehow remove the correlated omitted variable elements, this is not the model we are fitting with a standard mixed model above. Indeed, satisfying the random effects assumption is often quite difficult in Ecology – particular in observational data that spans environmental gradients - and likely is not well explored or acknowledge widely enough. We need a better solution.

*Enter Econometric Fixed Effects Models -* I LEFT OFF HERE ON DETAILED EDITS SO I CAN SEND TO YOU AND HIGHER LEVEL COMMENTS FROM HERE ON

## *{broader intro needed on why we would use this approach; feel free to pull and modify some text from the NutNet paper}*

~~If Random Effects are not the answer, then we can turn to fixed effects. We refer to fixed effects in two senses of the word.~~ 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 time-invariant attribute of the system. In our snail example, this would be the site-level time-invariant effect of recruitment. We also use it 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. There are many uses and definitions of this term, leading to wealth of confusion around it and different uses of the term fixed effects across fields [(Gelman and Hill 2006)](https://www.zotero.org/google-docs/?7TR7NM), and we hope to not add to it.

In the econometric sense, treating omitted variables as time-invariant and site-specific 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**. Given that the recruitment effect in our example is time invariant, we can transform the model to eliminate it. Consider the following mathematical description of the system.

y_{ij}= \beta_0 + \beta_1 x_{ij} + \gamma z_i + \epsilon_ij

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

\bar{y_i} = \beta_0 + \beta_1 \bar{x_i} + \gamma z_i + \bar{\epsilon_i}

\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.

y_{ij} - \bar{y_i} =  \beta_1 (x_{ij} - \bar{x_i}) + (\epsilon_{ij} - \bar{\epsilon_i})

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

Using simple algebra, we remove 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 5a, 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. 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 controlling for temperature. Further, we cannot use this model for predictive inference.

To solve these problems, we can use a model where a study unit (site in our snail example) is included as categorical or dummy variables*.* This kind of model - familiar as an ANCOVA model for Ecologists - will produce identical results to the preceding model for $\beta\_1$. Dummy variable coding allows site to be included as a fixed effect - in both senses of the term. Unlike in a mixed or random effects designs, econometric fixed effects 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.

y_{ij}  = \beta_1 x_{1ij} + \sum\alpha_i x_{2i} + \epsilon_{ij}

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

where $x\_{1ij}$ is our variable of interest and $ \alpha\_{i}$ is the fixed effect, estimated as a unique intercept per site, and $x\_{2i}$ 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 removing the average “level” of variable per site, or subtracting off a site level mean for each variable - equivalent to the within transformation model - and has the same effect in controlling for omitted variable bias [(Angrist and Pischke 2008, Wooldridge 2010)](https://www.zotero.org/google-docs/?EPiE9r).

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* site.

It is important to note, however, that site-level differences in effects can be incorporated back into the model by interacting ‘site’ with temperature, to understand heterogeneity in the causal effect across sites presuming there are multiple measures of temperature per site over time. Doing so in this design does not require assumptions that the effect of temperature across sites follows a particular distribution, as assumed by many random effects designs. Narrow ranges of variation per site, however, could cause problems in the ability to detect such an interaction.

The econometric fixed effects design does have two drawbacks. First, fixed effects estimators are inefficient compared to random effect estimators. For each group/fixed effect (site in our example), we get a corresponding column of dummy 1/0 variables in the model. 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 causally valid results. Second, we lose information about relationships between sites. While the estimand 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 can be problematic particularly with respect to prediction of new values - such as predicting snail abundances at new sites not included in our initial study, for example.

Further, the question begs, in the absence of information, how can one tell whether to use a fixed or random effects model and whether the random effects assumption is being violated? While a fixed effects model will always be safer, there are formal tests. In principle, one can use a **Hausman test** which looks at the difference between the RE and FE model coefficient of interest scaled by the difference in their standard errors. This test makes assumptions of a large sample size and the denominator can be 0, however, making it not ideal in many situations. For a better test, we need models that incorporate site random effects, but control for omitted variable bias.

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

To solve the above problems of efficiency and inference, we can step into the world of **correlated random effects models**. In these models, we again assume that our omitted variables that correlate with our predictor of interest vary at the study unit - in this case site - level. In these hierarchical models, we include a random effect of site but we also include a term that soaks up the variability from our omitted variables that is correlated with our predictor of interest. This 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 driver of interest, and learn about the variation between sites that is not correlated with our driver of interest.

The foundation of these approaches is leveraging *group means*. For every cluster – e.g., site, year, region - researchers calculate a group mean to include as a predictor in the model. This hierarchical predictor now acts to control for omitted variables that vary between sites and correlate with our driver of interest. The coefficient for our predictor of interest is now estimated while controlling for cluster-level correlated drivers. Consider the following model, first proposed by Mundlak [(1978)](https://www.zotero.org/google-docs/?KYkOLA) and further developed by Woolridge (REF).

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)


$$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 correlated drivers. In Econometrics, this is known as a **Mundlak Device** [(Mundlak 1978)](https://www.zotero.org/google-docs/?VEUOLc). For clarity, we term it a **Group Mean Covariate** model. We can see what this looks like graphically in Figure 5c. 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 interpretation of the site mean temperature coefficient, called a **contextual effect** [(Antonakis et al. 2021)](https://www.zotero.org/google-docs/?6DA6IC) 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 is not a problem in this particular analysis [(Antonakis et al. 2021)](https://www.zotero.org/google-docs/?eSpjat) .

The above model 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. To solve this, we can build a cleaner model that removes this correlation by looking at cluster-level anomalies. We can accomplish this with **group mean centering** where we subtract the cluster level mean from a given predictor. decomposes our predictor of interest into a between and within term. Now, the site mean temperature term would take on the meaning of a between site effect, and a group mean centered term would take on the meaning of a within-site temperature effect. We can see this in the following model:

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)


$$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 5d. You can see the similarities - and the key differences - with the Mundlak device and the previous fixed effect model. This model should produce the same estimate for $\beta\_1$ as the previous desig- the effect of a one unit change in temperature on snails. The interpretation of $\beta\_2$ is different than in the Mundlak device, however. It now provides a **between estimator** of the combined effect of gradients in temperature and correlated drivers between the sites. This is often a more useful estimand for ecologist. If $\beta\_1 = \beta\_1$, we might conclude, *tentatively*, that omitted variables are not influencing snails and both our between and within site differences are due solely to temperature.

While the group mean covariate, group-mean centered, and Fixed Effects models all differ in structure, they ultimately are all equivalent when it comes to estimating the temperature effect, $\beta\_1$, as they use within-site variation in temperature. As such, all three should produce similar estimates. Which model you use depends on the structure of your data (e.g., how many coefficients do you feel comfortable estimating with a fixed effects approach 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 model. Do you want to know how plot-level snail abundance would change if the average site temperature changes but plot temperature stayed the same? Group mean covariate model. Do you want to understand the effects of both within and between-site gradients? Group mean centered model. Note, one can work with a model and simulation to answer any of these questions with any of these models, but model choice 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 either vary across space (e.g., fixed effects approach). We have not discussed omitted variables that differ across time. Fortunately, the general framework above can be extended to these cases in a manner that showcases a more general underlying approach to omitted variables in all manner of situations that could be found in ecological systems.

First, while we have discussed how to handle site-specific temporal trends in omitted variables through differencing, what if your omitted confounding variables are temporal in nature. For example, in our snail system, consider a case where recruitment was uniform across sites but varied by year in a manner correlated with regional temperature. One simple approach might be conducting a cross-sectional study. But, if we suspect there are other site-specific omitted variables, or that recruitment could interact somehow with temperature, as discussed below, a cross-sectional study alone will not be sufficient. Fortunately, the causal diagram for such a scenario would not differ from Figure 3 save that, instead of site as our cluster that collects omitted variables, it would instead be year. We could then use year just as we have used site in any of the above approached - correlated random effects models, fixed effects models, etc. If there were indeed other omitted variables that varied by site, we could handle these just as before.

The world is rarely that simple, however. For panel designs, even if there is a spatial omitted variable, such as recruitment, temporal trends in a driver of interest at the site level can often covary with other site-level trends. These trends need not be uniform across sites, but instead can be site specific. Consider a small modification to the dynamics of our system:

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

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

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 $\beta\_1$ could again be contaminated. We could see this in a causal graph if the local variation was, say, coastal development increasing over time. This would have come out in a causal diagram such as that seen in Figure 6A. Here we see how time influences coastal development which influences temperature. Local temperature variability is also influenced by time. 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 before. The solution is differencing. For each time point of 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 predictor. Like the fixed effects transformation, site-level fixed omitted variables drop out. However, our temporal trend remains as a site-specific effect that we can accommodate 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 6B.

\Delta y_{ij} =  \beta_1 \Delta x_{1ij} + \sum \lambda_i x_{2ij} + \Delta \epsilon_{ij} 

\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 with respect to the non-temperature slope of change and change in temperature. For many studies investigating human-driven changes as their predictors of interest, this could be inadvisable. Note that if there are no temporal trends that vary by site, we can remove the site fixed effect to increase model efficiency and use cluster-robust standard errors (REF). 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 model, as represented by a causal diagram in 6C, has the advantage of estimating far fewer parameters if we have many sites, and thus could prove more efficient.

Taking either of these approaches has several advantages. We again are removing the effect of omitted site-level variables. We are also removing any 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 model is making fewer assumptions. Further, this approach shifts the type variation we are studying. Now, the researcher is estimating how *change* in a driver corresponds to *change* in a response. Said another way, researchers are no longer evaluating the relationship between a driver of interest and a response, controlling for unobserved site-level drivers but instead 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, as we lose observations from one or two time steps of data points. Loss of observations could make this approach have lower power (i.e., noisier stand errors). This can be even more evident in the case of the second difference approach, 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.

There are at least two possible solutions to these problems. To retain all of our data but still use an approach that eliminates a linear time varying omitted variable, we can transform our data in a manner akin to the within transformation. Rather than subtracting the mean of snails and temperature, however, we regress time on both snails and temperature at each site individually. We then analyze the relationship between the residuals of snails and temperature to estimate the effect of temperature after having removed the signal of any site-level temporal trends that are confounded with site-level trends in temperature. This approach makes a strong assumption, however, that variation should be ascribed to omitted temporal variables before our driver of interest, however, and can result in incorrect inference if this is not a valid assumption. To handle the irregular sampling issue, models could be modified to incorporate a time since the last sample, and have that variable interact with the driver of interest to calculate a rate of change standardized for differing sample intervals.

## **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 above. 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 replicate simulation runs exploring aggregate properties). For the purposes of this manuscript, we simulate 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 techniques described above, as well as using naive models with no site effect as well as group mean covariate and group mean centered models without a random effect. Broadly, our simulations show that the point estimate of the RE model is downward biased in these simulations compared to any other estimate (Fig. 7,8, Table 1.). Further, not only is the estimated coefficient of the RE model always lower than the other estimators, but, it more often is within 2SE of 0-- and frequently does not contain the true value of the temperature effect (Table 2). Additional explorations show that, with respect to incorporating a site random effect in group mean covariate or centered models, while this does not make a difference with respect to the temperature coefficient when the study design is balanced, it does affect results if the design is unbalanced and there is site-level variation that is uncorrelated with temperature (Appendix A). Again, 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.

**Further Extensions**

## *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 it. 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. 8, REF). In a naive mixed model, we would 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)$$

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, for a fixed effects dummy variable approach, we can incorporate an interaction effect between our causal variable of interest and site. Given that we now have 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 Mundlak device

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

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}

or a similar model for the group mean centered approach. Note, \gamma\_i might not be needed in this model if the omitted variable is the only cause of variation in the temperature effect.

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.

We caution, however, that if the relationship between the driver of interest and the outcome is nonlinear (e.g., a Generalized Linear Model with a non-identity link function), an interaction effect might not be appropriate. Consider if the relationship between temperature and snail abundance was exponential (e.g., we used a Poisson glm with a log link). While it might be tempting to let group centered temperature and site mean temperature interact, as plots with overall higher temperatures would seem to have a greater change per unit of temperature change than those with lower temperatures, the log link itself takes care of this problem. On a linear scale, the effect is additive. This is a minor concern, but it is one that users of generalized linear models should be aware of in their analyses.

## *Reality Bites: Coping with spatio-temporal omitted variables*

Spatio-temporal omitted variables can be extremely challenging, and the solutions can require more thoughtful study design. Consider that 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. If there is variability within a site in temperature and we have multiple plots sampled across multiple sites each year, we can cope with this sort of spatio-temporal omitted variable in ways that echo the types of models already seen above. For example, we can use a fixed effects approach 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}

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

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. Still, 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 Mundlak device approach analogous to the fixed effect version, although we note that terms capturing spatial and temporal omitted confounders might 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}

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

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 provide 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 (REFS FROM GOMON PAPER). 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)](https://www.zotero.org/google-docs/?4PAY0k).

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 spatio-temporal 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 fixed site effect would represent the slope of a site-specific temporal slope.

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

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

Where xi is a 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 (GAMs) (Wood et al. DATE). 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.

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. For more on this tricky class of problem and approaches outside of the scope of this paper (see Ferraro & Hauner, Athey and Imbens, Oster).

Finally, …

**Discussion**

We hope that our introduction to statistical designs to identify and address omitted variables using a causal diagram has shown that, through thinking carefully about biological systems, we can draw on a solid set of existing methods - from study design to analytic techniques - for coping with omitted variable bias and producing causally valid inferences from observational data. The techniques for reducing omitted variable bias are well within the standard statistical toolbox of most modern ecologists. And the results, as seen in at least this one toy example, can be profound for our ability to understand biological systems.

Further, we hope that in coming to understand the models presented here for dealing with OVB due to spatial or temporal confounders, Ecologists are able to see that this is a highly generalizable approach. Many types of clusters in a study could have omitted variables lurking around the corner. With a large enough sample size, however, models can be structured to accommodate multiple different types of clusters representing different suites of omitted variables quite simply as long as they are additive. 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 variables correlate with our observed variables of interest can produce models that are subtly misspecified - such as thinking that an omitted 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 what models you are building and why – and in the assumptions they are making in order to interpret an effect as causal or not. ~~If you are using mixed models, did you evaluate the random effects assumption? How?~~ 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 a 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 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)](https://www.zotero.org/google-docs/?z1LjK3), quasi-experimental approaches [(Butsic et al. 2017)](https://www.zotero.org/google-docs/?gCY86g), and are hopeful to see more on the emerging use of the front-door criterion? [(Bellemare et al. in press)](https://www.zotero.org/google-docs/?OEt7wo). We urge ecologists, long grounded in experiments as 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) to begin to increase their breadth of knowledge about how these fields have produced tremendous advances using variation in the world around them. 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)](https://www.zotero.org/google-docs/?KyTNkG), McElreath’s chapters on causal diagrams in Statistical Rethinking [(2020)](https://www.zotero.org/google-docs/?tlBkog), Angrist and Pishke’s Mostly Harmless Econometrics [(2008)](https://www.zotero.org/google-docs/?Vsoxu1), Sloman’s Causal Models [(2005)](https://www.zotero.org/google-docs/?L3RdjP), and Pearl et al’s Causal Inference in Statistics: A Primer [(2016)](https://www.zotero.org/google-docs/?ZWBboX). We also suggest Ecologists interrogate the assumptions and interpretations of their experiments [(Kimmel et al. 2021)](https://www.zotero.org/google-docs/?EcjmPr). 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). We thank S. Miller, I. Rosenthal, R. Stevenson, A. Carter, and the UMB Stats Snack for helpful conversation and comments on early drafts of the manuscript.

**—-------**

**Cut Text**

*Comparisons Show When to Ditch Random Effects* **~ I think this works better in the comparison section**

In many of the models above, we continue to include a random effect of site. For the group mean covariate, group mean centered, or differenced models, do we need a random site effect? It is possible that there is no additional site-level variation that generates correlation in the error of plots within sites. Possible. But unlikely. When modeling clustered data, using a ordinary least squares or other approaches and not accounting for clustering will produce incorrect standard errors and can lead to false conclusions (Primo et al. 2007). The structure of such a model is inherently pseudo-replicated. This does not mean, however, that a random effect *must* be included. Rather, we have two options. First, yes, random effects for the cluster level variable. The second, and less used option in Ecology and Evolutionary Biology, is a correction for clustered standard errors (REFS). Post-hoc clustered standard error corrections allow for the accommodation of a wide variety of violations of the I.I.D. assumption of traditional ordinary least squares, and have the added advantage over random effects of making no distributional assumptions about between cluster variance. Moreover, as these corrections are applied post-hoc, they do not affect point estimation of any coefficients in the model and, as we are not modeling the random effects structure, models are more likely to converge - a frequent frustration when complicated mixed model structures are invoked. For those unfamiliar with clustered standard errors, there is a large and deep literature on them (LAURA GIMME SOME REFS!) from which we suggest XXX as a starting point. For the purposes of the models discussed here, we would DO WHAT I DONT KNOW HELP CLUSTERED robust STANDARD ERRORS SCARE ME which can, for example, be implements in R with the XXX package (REF, see Appendix A).

This is not to say that we should simply drop mixed models in favor of clustered errors. If a researcher is interested in cluster specific means (e.g., the actual site specific intercepts themselves), making a comparison between site-level and plot-level variance components, building simulations that incorporate between-site variance components, or other questions specific to a mixed model structure, they are the way to go. Further, mixed models provide more efficient and accurate estimation in the face of unbalanced data - a frequent bugbear in ecological data. In cases where the only interest is the estimation of causal effects, clustered standard errors might prove a simpler approach. Both should produce equivalent answers with respect to the causal effects.One or the other *must* be used rather than omitting consideration of clustering altogether which will produce incorrect answers.