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

Jarrett E. K. Byrnes[1] and Laura E. Dee[2]

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/

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. To do so, we are working at larger scales in space and time and with unprecedented big data streams.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 providing answers and solutions for policy. 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 good randomized experiments are possible, and instead must be able to seize the opportunity of new large-scale sources of observational data.

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 our ecological systems are linked together. Second, armed with this understanding, the use of study designs and analytic techniques that can help us derive causal inference from observational data. These issues link together to 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. 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 potentially causal variable of interest is correlated with one or more other confounding variables that is not measured and thus is not in the statistical model of a system (Fig. 1). Unlike variables that do not influence both our causal variable of interest (Fig. 1A), these confounding variables (Fig. 1B) end up being included in the error term. By putting the confounding error 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. This makes the results of any statistical model invalid for causal inference. The model is not **causally identified**.

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. Indeed, measuring, controlling for, or even knowing all potential confounding variables is nearly impossible in complex ecological systems. In essence, in observational data analysis, we are always going to miss something. Period. It is a fact of life. 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 tus 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 variable)of interest. In observational studies, however, ecologists primarily attempt to deal with confounding variables by measuring the confounder of interest and including it in the model. Measuring a confounder, however, is frequently not possible, particularly for retrospective analyses. 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 Dee et al. 2021 In Review, Dudney et al. 2021, OTHERS?). 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 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. 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. 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. 2016 and Rinella et al. 2020 on OVB and instrumental variables). This difference could be due to Ecologists having few barriers to conducting experiments while other fields 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. We then present the typical approach in ecology to deal with unmeasured confounding variables, and thus omitted variable bias, and why it falls short. We then discuss study designs that, while omitted variables are still unmeasured, are ideal for analyses that can control for the influence of unmeasured confounding variables. We then review several robust statistical models that eliminate omitted variable bias, and provide guidance for choosing among them. As applied researchers, we have found that rather than creating confusion with complexity, these modeling approaches have clarified our own thinking about ecological systems. We hope that these relatively straightforward 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

One of the first tools in identifying and addressing omitted variable bias is knowing when and where OVB could be a problem for your analysis by making a causal diagram of the system. We recommend making this diagram, if possible, before designing an observational survey. Further, it should be a requirement before conducting an analysis from which one wants to make any causal conclusions. Making a causal diagram of your system - including both what one can and cannot measure - aids in determining where omitted variable problems could be lying in wait. 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).

We have already presented a causal diagram (Fig. 1), but let us take a moment to break down the elements of these diagrams. For the sake of simplicity, let us only consider causal diagrams with no feedbacks - so-called Directed Acyclic Graphs, or DAGs (Pearl 2009). 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 and Grace 2004) 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). 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.

For the variables and paths themselves, let us adopt the symbology common in Structural Equation Modeling (Bollen 1989), 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 - things that can be and are tangibly measured. We will represent these as terms within boxes - X and Y in figure 1. Second, unobserved and conceptual variables. We will present these as terms 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 model, you can determine whether you have an omitted variable bias problem and begin to determine solutions. What you are looking for is instances

where a driver that you are **not** interested in that 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 information to flow between your putative cause and 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 DATE) and will be causally valid (Fig. 2A). Variables that directly influence both a cause and effect of interest must be controlled for in order to produce proper causal inference. 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 estimand 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 in order to avoid omitted variable bias issues. 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., daggity Textor et al. 2011) 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, let us consider a system where both temperature and recruitment influence the abundance of snails in a marine benthic ecosystem (Fig. 3). In this system, we aim to study the causal relationship between temperature and snail abundance. Temperature influences metabolic and mortality rates, and we suspect fewer snails can survive in hotter sites. At the same time, as you make a causal diagram of your system, you realize 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 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 in order to prevent omitted variable bias from becoming a problem. Which one 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 would alleviate any known OVB *sensu stricto*.

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. This clustering - e.g., site, year, block, subject, individual, etc. - forms the basis for statistical removal of any omitted variable bias. In essence, we will rely on the cluster or values of the predictor at the cluster level (see below) to shut the back door. The variability in our causal variable of interest at the sample level then provides grist for the mill of causal statistical analysis. These designs can be simple longitudinal or cross-sectional studies. Or something more exotic, depending on how the ways the omitted variable and the causal driver of interest vary across space and time. 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, consider nesting using sites as a cluster as our OVB solution in a simple cross-sectional design. If the omitted variable affects our response variable at the site level, then we assume a correlation between the omitted variable and the average level of our causal variable at the site level. Within the site, however, multiple samples will have different values of the causal variable of interest. For our snail example, cold sites will have high recruitment. Warm sites will have low recruitment. Within a site, however, plots will vary in temperature relative to the site average due to forces that do not covary with recruitment. Note, this is an assumption, and one that must be justified! Further, the amount of within-site variability in temperature is important in considering the power or shape of the posterior for any analysis. With this design, site or 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.

The effect of our omitted variable might vary at the same spatial scale as our causal variable of interest. In this case, panel design approaches for longitudinal analysis become more appropriate. Here, while the omitted variable varies by site (or other form of cluster) as does

the average level of our causal predictor of interest, we use variability in our predictor of interest across time to create causally valid conclusions, again utilizing site or average level of our causal predictor at a site over time to shut the back door in statistical models. In our snail example, cold sites might occur in areas of high recruitment while warm sites occur in areas of low recruitment. However, with water temperature varying over the years, we should be able to produce valid causal inferences.

The above two solutions work well for omitted variables that covary with a driver of interest across a single type of cluster - either space or time. We have thus far only talked about space, as it makes intuitive sense - cold sites could have high recruitment. Conceptually, this should extend easily to time. 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.

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. 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. If the omitted variable and the driver of interest vary at exactly 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 1953).

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 valid, but reduce the influence of variability of uncorrelated variables when estimating causal relationships.

Statistical Approaches to Coping with Omitted Variables

There are multiple, well-established statistical model frameworks for analyzing panel or clustered data. We emphasize the term *'frameworks*' 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 models have different costs and benefits and importantly assumptions. Below, we discuss how addressing omitted variables can be approached, differentiate between different models, and outline additional assumptions that must be met in order for them to be valid – i.e., yield an unbiased estimate of an effect. 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 & Pichke 2009; Ferraro and Miranda 2017; Dudney et al 2021) - something many Ecologists worry about!

We illustrate the different models using a common set of predictors (x), responses (y), and omitted variables (z) and our example of the snail system referenced in Figure 3 with different sites (i) sampled at multiple time points (j). For the sake of simplicity, let's 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 β_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, robust standard errors to adjust for heteroskedasticity or correlation between time points (REF) might also be recommended, but a full discussion is beyond the scope of this discussion.

What Ecologists Typically Do: Random Effects Models

Mixed models have been all the rage in ecology for the past decade (Bolker and other refs), and for good reason!. Originally drawn from ways to partition variation in experiments with subsamples (Yates, others), they are a powerful tool when applied to observational data to account for the types of hierarchical study designs such as those discussed here. By partitioning variation between different levels of sampling hierarchies, we are able to obtain better estimates of precision for coefficients (Gelman and Hill 2006). Further, mixed models have several other properties that have made them popular in Ecology. First, it accounts 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 XXXX as to how this works with respect to baseball statistics for a beautifully clear explanation). 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 (Wooldridge 2010, Bolker). This difference results in many benefits, but also introduces one new assumption not often considered - the Random Effects Assumption that our random effects do not correlate with our fixed effects (Woolridge 2010, Antonakis 2021). Let's see how it plays out. Were we to model our snail study using a mixed model, it would look like the following:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \delta_i + \epsilon_{ij}$$
$$\delta_i \sim \mathcal{N}(0, \sigma^2_{site})$$

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

$$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, y_{ij} is the abundance of snails at site i in year j, β_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!), β_1 is the effect of temperature x at site i in year j on snails, δ_i is the site-specific deviation at site i from our intercept due to random variation which follows a normal distribution and ϵ_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 framework, the random effects of 'site'' are assumed to be uncorrelated with temperature. This is due to how random effects are estimated - as a part of the error term of the model (Woolridge 2010). Indeed, if we were uninterested in modeling the site-level means, we could combine δ_i and ϵ_{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 Fig. 4. In essence, site effects here are site-level residuals drawn from a normal distribution. They represent all of the other abiotic and biotic forces happening at the site level, and all are assumed to be uncorrelated with temperature at the site level. However, given the information in Figure 3, we know that this is not accurate. 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 remote 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 - and likely is not well explored enough. We need a better solution.

Enter Econometric Fixed Effects Models

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 a variance component. There is a wealth of confusion around the term (Gelman and Hill 2006), and we hope to not add to this.

In the econometric sense, treating omitted variables as time-invariant site specific variables means that, if we wish to remove their influence, 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 model to eliminate it. Consider one mathematical description of the system.

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma z_i + \epsilon_i j$$

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})

This is an elegant solution where we can see the influence of the omitted variable removed through some simple algebra. Cluster robust standard errors are likely important here as well 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. We lose information, however, 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 β_1. Dummy variable coding allows site to be included as a fixed effect - in both senses of the term. Unlike “random effects,” 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 & Pichke 2009, Woolridge 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, it enables a causally identified estimate of the temperature effect, removing differences among sites that are otherwise confounding.

Again, we are switching the variation we are studying, now to deviations from site-level means. However, it is important to note that site-level differences in effects can readily be incorporated back into the model by interacting 'site' with X, to understand heterogeneity in the causal effect across sites. Doing so in this design does not require assumptions that the effect of X across sites follows a particular distribution as in many random effects designs.

This fixed effects technique does have two drawbacks. First, fixed effects for groups are inefficient compared to random effects. For each group, 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. Classically, 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 all 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 *group means*. For every cluster - site, year, region, etc. - 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):

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 \bar{x_i} + \delta_i + \epsilon_{ij}$$
$$\delta_i \sim \mathcal{N}(0, \sigma^2_{site})$$

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

$$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)$$

Here, $\beta_2 \bar{x_i}$ accounts for the effect of cluster-level correlated drivers. In Econometrics, this is known as a **Mundlak Device** (Mundlak 1978). For clarity, we term it a **Group Mean Covariate** model. We can see what this looks like graphically in Figure 5a. 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) 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 10C, what would snail abundance be if said plot was in a site with an average temperature of 5C versus 20C? 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 2021).

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 model 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 β_1 as the previous model - the effect of a one unit change in temperature on snails. The interpretation of β_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, β_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 to just conduct a cross-section 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 λ_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 β_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. 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 x_{2ij}, 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 λ_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 in order 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 in order 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 of the techniques as 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, wour 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 estimands, 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 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 the confounder. 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 a fixed interaction effect between our causal driver 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

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 \bar{x_i} + \beta_3 x_{ij} \bar{x_i} + \gamma_i x_{ij} + \delta_i + \epsilon_{ij}$$

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 followup 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 x_{2i} is a dummy variable for site to capture spatial omitted confounders, x_{3j} is a dummy variable for time to capture temporal omitted variables and x_{4ij} 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 in order 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 (REFS).

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 x_i 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

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma z_i + \sum \lambda_1 i x_i j + \sum \lambda_2 i x_i j^2 + \sum \lambda_3 i x_i j^3 + e_{ij}$$

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

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 (REFS).

Discussion

We hope that our introduction to thinking about statistical models with omitted variables using a causal diagram has shown that, through thinking carefully about biological systems, we have a solid set of 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 some day, 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. Be transparent in what models you are building and why. Be transparent in which assumptions you did and did not test. 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 but a starting point. There are many other methods out there for producing causally valid inference in the face of omitted variable bias. We recommend several recent reviews of instrumental variables approaches (REFS), quasi-experimental approaches (REFS), and are hopeful to see more on the emerging use of the front-door effect (Bellemare et al. 2019). 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 (DATE), McElreath's chapters on causal diagrams in Statistical Rethinking (DATE), Angrist and Pishke's Mostly Harmless Econometrics (DATE), Sloman's Causal Models (Date), and Pearl et al's Causal Inference in Statistics: A Primer (DATE). We also suggest Ecologist interrogate themselves about whether their experiments truly have causal interpretations, or interpretations outside of the context of experiments themselves given their design Kimmel et al. 2021). It is high time to critically interrogate how to get the cleanest causal inferences needed to grapple with our rapidly changing world in order 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 euiver of all Ecologists. It is time to address pressing applied and theoretical questions at scale with the amazing observational data sets now coming on line. 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 for helpful conversation and comments on early drafts of the manuscript.