**I suggest we include a table with a glossary**

* Bias
* **Confounding variables**

Introduction

We have all been there. As a PhD student, you complete a field season’s worth of work observing your study system, measuring as much as was humanly possible for you and your research assistants, only to come in front of your committee as one member sagely raises an eyebrow in looking at your analysis and suggest, “Why didn’t you measure that? It could confound your entire story!” As a Post Doc, you are handed an amazing set to work with to make your career, but you swiftly notice it is missing the key variable is needed to properly resolve the story you want to tell. As a PI, you are ready to move from small-scale experiments to testing whether your theories are applicable across large spatial and temporal scales - but measurement of key elements of the system are simply not possible outside of small confined environments. Or, more vexingly, any attempt to do so is called ‘ambitious’ by funding agencies, leading to proposals being rejected.

As Ecology advances to tackle problems at scales from the continental to global, we are putting our theories to empirical test like never before – at larger scales in space and time, with unprecedent data streams. Our ability to test causal hypotheses and uncover causal connections between ?? is limited by two things. First, our ability to imagine how the different elements of our ecological systems of interest are linked together (I.e., our conceptual or theoretical model of how the system works). Second, armed with this understanding, the use of proper analysis designs that can help us derive causal inference from observational data. BUILD THIS UP – the challenge - HERE SINCE I SUGGEST CUTTING THE FIRST BIT: in the absence of key factors. We are always going to miss something. Period. Rather than to throw up our hands and abandon observational designs for causal inference because of this fact, it is better to try and understand what are the solutions to the grand problem of **omitted variable bias**.

Omitted variable bias occurs when a predictor of interest is correlated with other confounding variables that you have not measured and thus is not in your model of a system.

* Bias – differentiated from other issues like noise, sampling variability, and it arises from confounding variables. Misattributing the effect of Z to X on Y
* Randomized Experiments – when perfectly randomized – removes this issue of OVB vs Observational Designs

Omitted variable bias is not a new problem. Fields such as psychology, econometrics, education, sociology, and more have been grappling with it for some time (REFS). These are fields that 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 are 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 forms of coping with omitted variable bias. We begin by laying out criteria for understanding when and where omitted variable bias could be important. We first present the typical approach in ecology to deal with confounding variables, and thus omitted variables bias. We then discuss study designs that, while omitted variables are still unmeasured, are ideal for analyses that can eliminate confounding variables. We then review several robust techniques to model data with omitted variables, and provide guidelines for choosing among them. As applied researchers, we have found that these guidelines have clarified our own thinking about the analysis of 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.

**Example – new section heading**

To illustrate these empirical challenges and suite of potential solutions, we use a system where both temperature and recruitment influence the abundance of snails in a marine benthic ecosystem (Fig. 1). In this system, we aim to study the causal relationship between xxxxxx and yyyyyy.

Introduce measured versus confounding variables in this example:

Temperature influences metabolic and mortality rates. At the same time, the same oceanographic influences that shape temperature also shape recruitment of new juvenile snails. You go out and measure both snail abundance and temperature at a number of sites, but not recruitment.

**Confounding Variables and Bias**

In the context of the example - Defining Bias. Defining Confounding Variables. Differentiating from Inference and Sampling Variability – this is not a “signal to noise” issue!!!

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. If one has an effect and the other does not, your model could produce a false positive. That is because recruitment is omitted from your model. In the more general case (Fig. 2), for any predictor X, any unmeasured variable that either causes both X and a response Y, or whose ultimate cause influences X while it influences Y - is an omitted variable that can bias your estimates of the causal influence of X on Y.

**Typical approaches in Ecology**

*Typical approach*

These designs require foreknowledge (or at least for assumptions) about where omitted variable bias could become problematic. As with any study design, they can of course inflate the cost of a study, change sample allocation and thus cause other issues that could reduce power, or wreak havoc in other ways. Even when using them, we highly recommend looking at sample allocation in order to maximize the ability to detect signals of key predictors in the face of sources of environmental variability that do not influence results via the backdoor (SCOTT REFERENCES AND THE LIKE).

The most obvious solution is to design one’s study to incorporate confounding variables. This does not mean measuring every single variable that is correlated with both the predictor and response. Rather, a researcher can use their causal diagram wisely in order to determine a (hopefully) small suite of variables through which the influences of any omitted variable flows. Consider the graph in Figure XXXX.

In this system, one could control for all of the confounding variables in determining a relationship between x1 and y simply by including z1. Alternately, including x2, z3, and z2 would also be sufficient. This is a far cry from including everything in the diagram. By realizing the small suite of variables a researcher needs to sample for a specific question, the problem of study design or justification to skeptical reviewers becomes far less daunting.

In many cases, sampling more predictors and responses might not be possible, despite knowledge that other variables are important from *a priori* causal diagrams. For example, consider investigations using repurposed data sets or if additional measurements are too expensive. There are a wide variety of solutions to this problem, each utilizing some sort of grouping structure where groups are a stand-in for omitted variables. In the *Statistical Approaches* section, we’ll discuss how those groups should be modeled. But, in general, designs that recognize at what level these additional omitted variables influence the system can be used with care to accommodate omitted variable bias. These can be classic stratified random sampling designs or various modifications (SCOTT PAPERS), accommodating omitted variables that vary at the group-level. These SRDs can be either spatial, for purely cross-sectional data, or temporal, for omitted variables that might vary through time. Similarly, longitudinal sampling designs - sampling the same plots or sites over time - allow for researchers to adjust for omitted variables that covary with site. Combinations of multiple group-types are also possible. For example, taking multiple replicates from multiple sites that are resampled over time can enable a researcher to accommodate both spatial and temporal omitted variables. This grouping approach can extend further to different types of groups and can be adapted into designs with variable groups, etc (e.g., Lebo and Weber 2015).

The *Selection on observables assumption – or ‘satisfying the back-door criterion’*

**The risk: unobserved confounding variables - Simulations**

When this assumption isn’t met, then …

**Using DAGs to clarify our causal understanding and assumptions**

*Directed Acyclic Causal Graph (DAG):* A DAG is a visualization of qualitative causal assumptions on which one relies for making causal claims from observable data (*1*). See Section S2 for more information and the relationship between a DAG and a “path diagram.”

DAG Figure. Diagram of causal connections in an example system. Variables with boxes around them are measured variables. Variables with ellipses around them are unmeasured or ‘unobserved’ variables. e is for additional sources of variability uncorrelated with other drivers. Note that drivers of Y shown in e that are uncorrelated with the X of interest are not a problem for bias, and more-so to do with reducing noise in predictions of Y.

If a researcher is concerned about whether their model suffers from unobserved variable bias, constructing a causal diagram is the swiftest way to determine if there is an obvious problem. This is not to say they will always be correct - hypothesized causal diagrams can be incorrect. Therefore, adjusting for known omitted variables might still be insufficient. Nevertheless, they should provide a simple means to address the worries of a researcher late in the research process who has suddenly has to contend with the possibility that they did not measure what could be an important variable.

DAGs help visualize assumptions and potential sources of bias from confounding variables. They should identify confounders!

This could help identify what to measure and control for in analyses, as outlined above, or help us choose and use more flexible analysis designs to eliminate

**References for DAGs in more complex settings too – dynamics, interference..**

## Simple Solutions From Other Fields: Designs to cope with unobserved variable bias

There are multiple study designs that researchers can use in order to prevent omitted variable bias from becoming a problem – that don’t require measuring, knowing, and controlling for every possible confounding variable. These designs require foreknowledge (or at least for assumptions) about where omitted variable bias could become problematic. As with any study design, they can of course inflate the cost of a study, change sample allocation and thus cause other issues that could reduce power, or wreak havoc in other ways. Even when using them, we highly recommend looking at sample allocation in order to maximize the ability to detect signals of key predictors in the face of sources of environmental variability that do not influence results via the backdoor (SCOTT REFERENCES AND THE LIKE).

The most obvious solution is to design one’s study to incorporate confounding variables. This does not mean measuring every single variable that is correlated with both the predictor and response. Rather, a researcher can use their causal diagram wisely in order to determine a (hopefully) small suite of variables through which the influences of any omitted variable flows. Consider the graph in Figure XXXX.

In this system, one could control for all of the confounding variables in determining a relationship between x1 and y simply by including z1. Alternately, including x2, z3, and z2 would also be sufficient. This is a far cry from including everything in the diagram. By realizing the small suite of variables a researcher needs to sample for a specific question, the problem of study design or justification to skeptical reviewers becomes far less daunting.

*Contrasting these approaches with random effects: Random effects make strong assumptions*

**First differences – good entry point. Easy to see how this is done in an equation.**

**e.g., pasting from the NutNet SM (see equation 2)**

**Then two-way fixed effects with dummy variables**

*Fixed Effect:* A “fixed effect” refers to a time-invariant attribute of the system; e.g., a plot-level fixed effect is an attribute of the plot that does not change over the study period, such as topography or historical patterns of use. This use of the term “fixed effect” differs from how the term is typically used in ecology, where the term often refers to the coefficient estimates of explanatory variables in mixed (multi-level) modeling. In our study, the fixed effect is not part of the error term, as it would be in mixed (multi-level) modeling context. It is a parameter to be estimated (i.e., it is assumed to be fixed and estimable, rather than assumed to have a distribution). To read more, see (*3*).

**Group mean, Mundlak etc.**

Something on implementation?

**Other solutions**

* Front door

*Quasi experimental approaches*

* – briefly mention and can cite Larsen et al MEE

## A Front-door solution

Perhaps one of the least implemented but most powerful solution is building a model to achieve the so-called Front-Door criterion (PEARL REF). Simply put, if you have an open back door, if there is a variable that mediates the relationship between a purported cause and effect and is not influenced by anything other than the cause, then we can establish a link between the cause and effect as well as estimate it’s net effect size by looking at the change in the mediator due to its cause and the corresponding change in the response due to the change in the mediator. This is naturally done in Structural Equation Modeling (Bollen 1989), for example.

To clarify the front-door criterion, consider an example. A sewage plant is suspected of causing mortality of soft sediment organisms on a bay. However, all of the impacts are in the nearshore. Control sites are far away, and have different abiotic regimes - temperature, depth, recruitment, etc. However, the sewage plant puts out sludge. Thus, an attempt to look at the relationship between distance from sewage plant outfall and, say, infaunal species richness would be hopelessly contaminated by a number of open backdoors. However, if at least one of the impacts is via deposition of sludge, then we can estimate a) the relationship between distance from plant and depth of sludge on the benthos and b) depth of sludge and infaunal species richness. If both relationships are different from zero, then there is an impact of the sewage plant. Further, as sludge depth is correlated with the myriad of other impacts that we have not measured, if we estimate the effect of one unit change in distance on sludge, and the corresponding change in that number of units of sludge on species richness, we have a net estimate of the sewage plant on species richness.

**Discussion:**

Inference vs bias. [Trade-offs between bias and variance ]

Trade-offs between within- and between-estimators

* Sensitivity tests (Oster … ; Altonji), Dee et al example from NutNet.

R2 is not a valid way to assess

Other sources of bias and confounding variables

* Dynamic confounders
* Measurement error
* Reverse causality
* Colliders

Something on implementation? Or SM ?

**Discussion**

We aim to provide an introduction to thinking about statistical models with observed variable using a causal diagram has shown that, through thinking carefully about biological systems, it is use relatively simple techniques to reduce the problem of observed variable bias. The techniques for reducing observed 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.

The approach we present here is surely not a panacea. Model mispecification might 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 variable correlate with our observed variables of interest can produce models that are subtly misspeciefied - 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, some one 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.

Moreover, we feel that the general approach to diagramming systems before attacking them with multiple regression-style approaches not only aids in the omitted variable bias problem, but it serves to reduce bias from multiple other sources (e.g., from interference or so-called neighborhood effects, see Ogburn and Vanderwheele **Causal Diagrams for Interference )**. Fitting all variables as predictors without careful thought as to the causal structure of a system can produce many different types of errors of inference (see McElreath Chapter XXX and Pearl XXX for review). While we have long sought for the right statistical technique to determine what variables we should shoehorn into our multiple regression analyses - be it saturated models, stepwise approaches, massive multi-model fits and AIC tables with model averaging, or more arcane statistical incantations - we here show the paramount importance of thinking careful about a system before trying to divine its secrets. It is a technique that will serve every scientist in every field well.

**Conclusions**

We hope that this guide serves to provide a ready arrow in the quiver of all scientists - particularly those just beginning to learn how to think about asking and answering questions of biological systems. 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. Rather than sweep the problem under the rug, we hope that you can now move forward with confidence in coping with what is an undeniably common problem. We look forward to the new insights that these techniques will help you generate.