Model selection isn't causal inference

Suchinta Arif[1] and M. Aaron MacNeil[1]

[1]Dalhousie University

February 9, 2022

Abstract

Ecologists often rely on observational data to understand causal relationships. Although observational causal inference methodologies exist, model selection based on information criterion (e.g., AIC) remains a common approach used to understand ecological relationships. However, such approaches are meant for predictive inference and is not appropriate for drawing causal conclusions. Here, we highlight the distinction between predictive and causal inference and show how model selection techniques can lead to biased causal estimates. Instead, we encourage ecologists to apply the backdoor criterion, a graphical rule that can be used to determine causal relationships across observational studies.

Model selection isn't causal inference

Suchinta Arif (suchinta.arif@dal.ca) *[1], Aaron MacNeil[1](a.macneil@dal.ca)

*Corresponding author

[1]Dalhousie University, Department of Biology

Life Sciences Building,

1355 Oxford St. B3H 3Z1,

Halifax, Nova Scotia, Canada

Running title: Model selection isn't causal inference

Keywords: causal inference, cause-and-effect, model selection, directed acyclic graphs (DAGs), back-door criterion

Statement of authorship: SA conceived the idea for this viewpoint and led the drafting of the manuscript. All authors edited the manuscript and collaboratively discussed their perspectives throughout writing the manuscript.

Data accessibility statement: No new data were used in this manuscript.

Type of Article: Viewpoint; Counts: Abstract words (94), main text words (1797), references (35), figures (1), tables (1)

Abstract:

Ecologists often rely on observational data to understand causal relationships. Although observational causal inference methodologies exist, model selection based on information criterion (e.g., AIC) remains a common approach used to understand ecological relationships. However, such approaches are meant for predictive inference and is not appropriate for drawing causal conclusions. Here, we highlight the distinction between predictive and causal inference and show how model selection techniques can lead to biased causal estimates.

Instead, we encourage ecologists to apply the backdoor criterion, a graphical rule that can be used to determine causal relationships across observational studies.

As ecologists, we are often interested in answering causal questions about human impacts on the natural world, such as the effect of climate-induced bleaching events on coral reef ecosystems (e.g., Graham et al. 2015), the impact of deforestation on biodiversity (e.g., Brook et al. 2003), or the effect of conservation and management responses on restoring ecosystem services (e.g., Sala et al. 2018). Often, randomized controlled experiments are unfeasible, and ecologists instead rely on observational data to answer fundamental causal questions in ecology (MacNeil, 2008). Recently, new advances in technology such as remote-sensing and animal-borne sensors, as well as increased availability of citizen science and electronic data have further increased opportunities to answer causal questions from observational data (Sagarin and Pauchard 2010).

In recent years, researchers have advocated for the increased application of causal inference in ecology for answering cause and effect relationships from observational data (e.g., Larsen et al. 2019; Laubach et al. 2021) but these approaches have yet to be widely adopted. Instead, drawing causal conclusions from observational data is typically taboo, with Pearson's oft-cited "correlation doesn't equal causation" used to block attempts to do so (Glymour 2009). This misconception – that causality cannot be inferred using observational data – has resulted in a culture where ecologists dependent on observational data for understanding causal relationships avoid explicitly acknowledging the causal goal of research projects and instead use coded language that implies causality without explicitly saying so (Hernan 2018; Arif et al. 2021).

A common strategy used to understand ecological relationships is to apply model selection, using information metrics such as Akaike's information criterion (AIC; Akaike 1973). Such approaches select the 'best' model among a candidate set and subsequently make inferences from parameters that are of ecological interest within the winning model. Often, these inferences are tied up with causal language, implying that having selected the best model, one can proceed to using causal language in reference to it (Table 1). However, model selection is not a valid method for inferring causal relationships – rather, these techniques aim to select the best model for predicting a response variable of interest. For example, AIC approximates a model's out of sample predictive accuracy, using only within-sample data (Akaike 1973). Although numerous model selection criteria exist (e.g., BIC, Schwarz et al. 1978; DIC; Spiegelhalter et al. 2002; WAIC, Watanabe 2013; LOO-CV; Vehtari et al. 2017), they are all used to compare models based on predictive accuracy (McElreath 2020; Laubach et al. 2021; Tredennick et al. 2021). Thus, model selection is appropriate for predictive inference (i.e., which model best predicts Y?), which is fundamentally distinct from causal inference (i.e., what is the effect of X on Y?).

To demonstrate this distinction, the directed acyclic graph (DAG) in Figure 1 shows the causal structure of a hypothetical ecological system. DAGs can be used to visualize causal relationships, where variables (nodes) are connected to each other via directed arrows, pointing from cause to effect (Elwert 2013). For example, forestry effects species Y both directly (there is a directed arrow between them) and indirectly, via the directed arrow from forestry to species A and from species A to species Y (Fig 1). To illustrate the difference between model selection and causal inference we created a simulated dataset that matches the linear causal structure of this DAG, setting the total (i.e., direct and indirect) causal effect of forestry on species Y to -0.75 (Appendix S1). We further specified candidate linear regression models that included all possible covariate combinations where species Y is a response. Using our simulated data and our candidate models, both AIC and BIC selected a 'best' model where forestry, species A, human gravity, climate, and invasive species Z were included as covariates (Appendix S1). However, interpreting the coefficients of this model can provide biased causal estimates. For example, the effect of forestry on species richness is shown to be -0.36 [-0.38, -0.33], instead of -0.75 (Appendix S1).

In this scenario, there are two statistical biases at play. The first is overcontrol bias, which occurs when the inclusion of intermediate variables along a causal pathway removes the indirect causal effect between predictor and response (Cinelli et al. 2021). Here, the inclusion of the intermediate variable species A removes the indirect effect between forestry and species Y. Second, the inclusion of invasive species Z as

a covariate leads to collider bias, which can result from adjusting for a variable that is caused by both predictor and response (Cinelli et al. 2021). Here, the inclusion of invasive species Z induces an additional, but non-causal, association between forestry and species Y.

It is worth noting that although the true predictive model (i.e., the data-generating model for species Y, where all direct predictor variables – human gravity, species A, forestry, and climate were included as covariates) was included as a candidate model, both AIC and BIC selected a more complex model with invasive species Z as a covariate. Here, even though invasive species Z is not a predictor variable for species Y, its statistical (non-causal) association with species Y increased predictive accuracy, resulting in better out of sample predictive accuracy. Indeed non-causal associations including collider bias and reverse causation has been shown to increase predictive accuracy (e.g., Luque-Fernandez et al. 2019; Griffith et al. 2020). Thus, a model selected based on predictive accuracy should not be assumed to be causally accurate.

A more subtle point is that even if a model captures the data generating process for a response variable of interest, it may not be appropriate for answering specific causal queries. For example, if we want to know the total effect of forestry on species Y, a model with all direct predictor variables – human gravity, species A, forestry, and climate – included as covariates, returns a causal estimate of -0.21[-0.23, -0.18] instead of -0.75 (Appendix S1). Here, the inclusion of species A as a covariate leads to overcontrol bias between forestry and species Y, removing this indirect effect. As well, this model cannot be used to determine the causal estimates of other distal drivers, such as climate or fire. Ultimately, causal models must be built based on the specific causal question at hand, as well as through the careful consideration of the overall causal structure, including how different predictor variables may be related to one another.

The Backdoor Criterion: Covariate Selection for Causal Inference

As a contrast from model selection approaches, a causal inference methodology that has recently emerged in ecology is Judea Pearl's structural causal model (SCM; Pearl 2009). This framework uses DAGs to visualize researchers' assumptions about the causal structure of a system or process under study. Once a DAG has been created, a graphical rule known as the backdoor criterion can be applied to determine the covariates required to answer a causal question from observational data.

Conceptually, the backdoor criterion instructs us to block all non-causal paths between a predictor and response variable of interest, while leaving all causal pathways open. Graphically, this translates to blocking all backdoor paths between a predictor and response variable. Backdoor paths are sequences of nodes and arrows with an arrow pointing into both the predictor and response variable of interest; if left open, they can lead to non-causal associations between variables of interest. To block a backdoor path, we can either (1) adjust for an intermediate arrow-emitting variable or (2) not adjust for a variable with two incoming arrows (i.e., a collider variable: X).

For example, given our DAG in Fig 1, to determine the total effect of forestry on species Y, there are four backdoor paths that must be blocked:

1. Species Y Climate Forestry
2. Species Y Climate Fire Species A Species Y
3. Species Y Species A Fire Climate Forestry
4. Species Y Human Gravity Forestry

The first three backdoor paths can each be blocked by adjusting for the intermediate arrow-emitting variable climate. The fourth backdoor path can be blocked by adjusting for the intermediate arrow-emitting variable human gravity. Therefore, to determine the total effect of forestry on species Y, we must adjust for climate and forestry. Following covariate selection, researchers can determine the appropriate statistical analysis, given their data. It is important to note that DAGs and the backdoor criterion are compatible with both linear and non-parametric approaches (Pearl 2009; Elwert 2013). As our simulated data was created using linear relationships, we have chosen a linear regression model, setting species Y as our response, forestry as our predictor, and including climate and forestry as controls. This model returned an accurate total

causal estimate of -0.75[-0.77, -0.73] (Appendix S1). The application of the backdoor criterion can become increasingly complex with larger DAGs and as such, tools such as 'dagitty' (www.dagitty.net; instructions within site) can help in composing DAGs and specifying causal questions, which will subsequently identify required backdoor adjustment sets.

Distinction From Model Selection

Covariate selection using the backdoor criterion is fundamentally distinct from information-based model selection techniques. The backdoor criterion is based on counterfactual reasoning, equating observational distributions to what would be expected under a randomized control experiment (Pearl 2009). Unlike model selection, the backdoor criterion was specifically created to answer cause and effect relationships from observational data. Further, whereas model selection relies on the data to determine the best model, the backdoor criterion uses domain knowledge, above all else, to determine the best causal model for a given causal query. The use of DAGs and the subsequent application of the backdoor criterion allows ecologists to move away from an automated approach of model selection to one that empowers ecologists to think critically about the cause-and-effect relationships in their study system. The use of DAGs also facilitates open critique of causal assumptions therefore their causal conclusions, which in turn can lead to productive scientific debate that deepens our understanding of ecological phenomena (e.g., see Schoolmaster Jr. et al. 2020; rebuttal by Grace et al. 2021; and reply by Schoolmaster Jr. et al. 2021).

Currently, DAGs and the backdoor criterion are significantly less utilized than predictive model selection techniques for understanding causal relationships in ecology. Thus far, the backdoor criterion has been applied to understand the causes of species level trait covariation (Cronin and Schoolmaster Jr. 2018), biodiversity-ecosystem function correlations (Schoolmaster Jr. et al. 2020), and causal drivers of coral-algal regime shifts (Arif et al. 2021). As these varied examples demonstrate, the backdoor criterion can be widely applicable for understanding ecological causal relationships. Increasing its use across ecological studies will require a shift in culture toward openly discussing causality. While predictive model selection techniques can play an important role in developing good statistical models, they should not be conflated with causal inference (Laubach et al. 2021). Ultimately, ecologists must start to rely on valid causal inference methods to answer fundamental causal questions in observational ecology.

Literature Cited:

Arif, S., Graham, N., Wilson, S., MacNeil, A. (2021) Causal drivers of climate-mediated coral reef regime shifts. *Ecosphere* , in press (accepted).

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory,* Ed(s). Petrov, B.N., Csaki, B.F. Academiai Kiado: Budapest, pp 267-281.

Brook, B., Sodhi, N., Ng, P. (2003). Catastrophic extinctions follow deforestation in Singapore. *Nature* , 424, 420-423.

Chinn, S.M., Kilgo, J.C., Vukovich, M.A., Beasley, J.C. (2021). Influence of intrinsic and

extrinsic attributes on neonate survival in an invasive large mammal.*Sci Rep* , 11, 11033.

https://doi.org/10.1038/s41598-021-90495-x

Cinelli, C., Forney, A., Pearl, J. (2021). Crash course in good and bad controls. Technical Report R-493.

Elwert, F. (2013). Graphical causal models. In Handbook of causal analysis for social research, Ed. Morgan, S.L. Springer, Dordrecht, pp 245-273.

Grace, J. B., M. Loreau, and B. Schmid. (2021). A graphical causal model for resolving species identity effects and biodiversity–ecosystem function correlation: comment. *Ecology* , 0, e03378. https://doi.org/10.1002/ecy.3378

Graham, N., Jennings, S., MacNeil, A., Mouillot, D., Wilson, S. (2015). Predicting climate-driven regime shifts versus rebound potential in coral reefs. *Nature* , 518, 94-97.

Glymour, C. (2009). Causation and Statistical Inference. In *The Oxford Handbook of Causation* , Ed(s). Beebee, H., Hitchcock, C., Menzies, P. Oxford University Press, New York, USA.

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

Hernan, M. (2018). The C-Word: Scientific euphemisms do not improve causal inference from observational data. *Am J Public Health* , 108(5), 616-619.

Larsen, A., Meng, K., Kendall, B. (2019). Causal analysis in control-impact ecological studies with observational data. *Methods Ecol Evol* , 10, 924-934.

Laubach, Z., Murray, E., Hoke, K., Safran, R., Perng, W. 2021. A biologist's guide to model selection and causal inference. *Proc Royal Soc B* , 288(1943), 20202815. Doi: 10.1098/rspb.2020.2815.

Lu, X., Chen, X., Zhao, X., Lv, D., Zhang, Y. (2021). Assessing the impact of land surface temperature on urban net primary productivity increment based on geographically weighted regression model. *Sci Rep* , 11, 22282. https://doi.org/10.1038/s41598-021-01757-7

Luque-Fernandez, M., Schomaker, M., Refondo-Sanchez, D., Perez, M., Vaidya, A., Schnitzer, M. (2019). Educational note: paradoxical collider effect in the analysis of non-communicable disease epidemiological data: a reproducible illustration and web application. *Int J Epidemeiol* , 1(48), 640-653.

MacNeil, A. (2008). Making empirical progress in observational ecology. Environ Conserv, 35(3), 193-196.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* . CRC Press, Boca Raton, USA.

Millard, J., Outhwaite, C.L., Kinnersley, R., Freeman, R., Gregory, R., Adedoja, O. *et al* . (2021). Global effects of land-use intensity on local pollinator biodiversity. *Nat Commun* , 12, 2902.

Montano-Centellas, F., Loiselle, B., Tingley, M. (2020). Ecological drivers of avian community assembly along a tropical elevation gradient.*Ecography* , 44(4), 574-588.

Morton, O., Scheffers, B.R., Haugaasen, T., Edwards, D. 2021. Impacts of wildlife trade on terrestrial biodiversity. *Nat Ecol Evol* , 5, 540-548.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference* . 2nd Edition. Cambridge University Press; Cambridge, UK.

Rode, K., Regehr, E., Bromaghin, J., Wilson, R., Martin, M., Crawford, J., Quakenbush, L. (2021). Seal body condition and atmospheric circulation patterns influence polar bear body condition, recruitment, and feeding ecology in the Chukchi Sea. *Glob Chang Biol* , 27(12), 2684-2701.

Safaie, A., Silbiger, N.J., McClanahan, T.R., Pawlak, G., Barshis, D.K., Hench, J.L. *et al* . (2018). High frequency temperature variability reduces the risk of coral bleaching. *Nat Commun* , 9, 1671.

Sagarin, R., Pauchard, A. (2010). Observational approaches in ecology open new ground in a changing world. Front Ecol Environ, 8(7), 379-386.

Sala, E., Giakoumi, S. (2018). No-take marine reserves are the most effective protected areas in the ocean. *ICES J Mar Sci* , 75(3), 1166-1168.

Schoolmaster Jr., D., Zirbel, C., and Cronin, P. (2020). A graphical causal model for resolving species identity effects and biodiversity–ecosystem function correlations. *Ecology* , 101(8): e03070.

Schoolmaster Jr., D., Zirbel, C., and Cronin, P. (2021). A graphical causal model for resolving species identity effects and biodiversity-ecosystem function correlations: Reply. *Ecology* , 0, e03593. doi:10.1002/ecy.3593.

Schwarz, Gideon E. (1978). Estimating the dimension of a model. *Ann Stat* , 6(2), 461-464.

Sinnot-Armstrong, M., Donoghue, M., Jetz, W. (2021). Dispersers and environment drive global variation in fruit colour syndromes. *Ecol Lett* , 24(7), 1387-1399.

Spiegelhalter, D., Best, N., Carlin, B., van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *J R Stat Soc Ser B Methodol* , 64(4), 583-639.

Teixeira, H., Montade, V., Salmona, J., Metzger, J., Bremond, L., Kasper, T. (2021). Past environmental changes affected lemur population dynamics prior to human impact in Madagascar. 2021. *Commun Biol* , 4, 1084.

Tredennick, A., Hooker, G., Ellner, S., Adler, P. (2021). A practical guide to selecting models for exploration, inference and prediction in ecology. *Ecology* , 102(6), e03336.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat Comput* , 27(5), 1413-1432.

Watanabe, S. 2013. A widely applicable Bayesian information criterion. *J Mach Learn Res* , 14, 867-897.

Walker, M., Uribasterra, M., Asher, V., Getz, W., Ryan, S., Ponciano, J., Blackburn, J. 2021. Factors influencing scavenger guilds and scavenging efficiency in Southwestern Montana. *Sci Rep* , 11, 4254.

Table 1. A sample of recent observational ecological studies that have used model selection techniques to answer causal questions, where results are communicated using causal language (e.g., effect, driver, influence).

Paper	Causal Question
Millard et al. 2021	What are the global effects of land-use intensity on local pollination biodiversity?
Lu et al. 2021	What is the impact of land surface temperature on urban net primary productivity?
Safaie et al. 2021	How does high frequency temperature variability effect the risk of coral bleaching?
Morton et al. 2021	What is the impact of wildlife trade on terrestrial biodiversity?
Chinn et al. 2021	What is the influence of intrinsic and extrinsic attributes on neonate survival in wild pigs?
Montano-Centellas et al. 2021	What are the ecological drivers of avian community assembly along a tropical elevation gr
Rode et al. 2021	What are the combined effects of sea ice, seal body condition and atmospheric circulation
Sinnott-Armstrong et al. 2021.	What are the biotic and abiotic drivers of fruit colour syndrome?
Walker et al. 2021	What factors influence scavenger guilds and scavenging efficacy in Southwestern Montana?
Teixeira et al. 2021.	How did past environmental changes (prior to human impact) effect lemur population dyn

Hosted file

`image1.emf` available at https://authorea.com/users/381917/articles/555837-model-selection-isn-t-causal-inference

Figure 1. A directed acyclic graph (DAG) representing the causal structure of a hypothetical ecological system.