**Reading Notes: Suchinta Arif and M. Aaron MacNeil (2022), Predictive models aren't for causal inference, Ecology Letters, 25:1741–1745.**

**Uwe Ehret**

**Summary**

A key challenge in ecology is to infer causal relationships among ecosystem subsystems, as controlled experiments are unfeasible, observational data are few, and the number of involved variates is high. In this context, the key message of this paper is that ecological model selection (maximizing predictive accuracy plus some complexity control such as the Akaike information criterion AIC ) does not necessarily select models which are optimal for causal inference, i.e. to infer from the model cause-and-effect relationships among ecosystem subsystems. Statistical biases (non-causal associations between variables of interest) that might occur in such an approach include confounding bias (a predictor that effects both the predictor and target variable of interest is not controlled), overcontrol bias (an intermediate variable between the predictor and target of interest is controlled) and collider bias (a variable influenced by both the predictor and target of interest is controlled).

The authors illustrate this with an synthetic ecosystem model expressed by a directed acyclic graph (DAG), which, when optimized for AIC and BIC do not identify the correct causal relationships among the DAG components. These biases also occur if ecosystems are modelled by ML approaches. The authors argue that if causal inference is the goal, methods directly developed for causal inference from observational data such as the Structural Causal Model framework (SCM) or the backdoor criterion should be employed. The latter provides a formal means for isolation causal effects from observations data by removing all non-causal paths between a predictor and a target, and with this removes the above-mentioned statistical biases.

**How this might be useful**

* Do models optimized for causal inference provide better predictive out-of-sample performance when trained on small samples than those optimized for predictive accuracy? In other words, does causality provide a useful regularization?

**Open questions**

* The authors only describe methods for DAGs, but ecosystems more often than not include feedback. How to identify causal relations in such systems?