**Using structural equation modeling to understand indirect effects in the management of pests and beneficial species**

Robert E. Clark1,2\*, Emily Rampone1, Robert Orpet3, David W. Crowder1

1 Department of Entomology, Washington State University, Pullman, Washington, USA

2 EcoData Technology, Plantsville, Connecticut, USA

3 Tree Fruit Research and Extension Center, Washington State University, Wenatchee, Washington, USA

\* Corresponding author: [robert.e.clark@wsu.edu](mailto:robert.e.clark@wsu.edu)

**Acknowledgements** Data presented in this publication was funded by USDA-NIFA award 2017-07156, USDA-NIFA award 2017-67013-26537, and NSF Doctoral Dissertation Improvement Grant DEB-1404177. Support for R. Clark was provided by USDA-NIFA award 2022-67013-36422. We thank the Crowder Lab 2023 modeling group at WSU Entomology for helpful feedback on early versions of this manuscript.

**Abstract**

Management of pests and beneficials species in applied ecology wrestles with the impacts of new strategies. Controls reducing pest or invasive species can have non-target effects, while introducing beneficial species can have unforeseen consequences in food webs. Ultimately applied ecologists rely on univariate or multivariate statistics to evaluate how treatments will play out. However, quantification of direct vs. indirect effects among a network of interactions cannot be achieved with this approach. Path analysis, a type of structural equation modeling, has been increasingly leveraged for its ability to model indirect effects while still facilitating hypothesis testing. Multiple R packages have been developed which directly build on generalized linear models to implement, but the strengths and weaknesses of this approach are difficult to evaluate. In this paper, we provide several examples from applied ecological work on beneficial and pest insects, demonstrating the usefulness of path analysis and how to avoid common pitfalls.

**Keywords** Ecological pest management, beneficial species, environmental management, path analysis, structural equation modeling, generalized linear mixed models

**Introduction**

Indirect effects are common in biological systems, and structural equation modeling (SEM) has become an important tool in the life sciences to compare the strength of direct and indirect effects. SEM has been adopted in psychology (Senn et al. 2004), genetics (Valente et al. 2013), ecology (Wei et al. 2013), and agriculture (Liere et al. 2015). In managed ecosystems, such as farms, timber stands, or conservation plots, treatments are often implemented to reduce the impacts of pests or invasive species or restore functions of beneficial species. However, applied ecologists often must cope with intermediate steps involving changes before the desired outcome is achieved.

Most experimenters leverage univariate linear models, an approach which is an outgrowth of the original development of ANOVA and the F-statistic from Fischer’s agricultural trials (citation from fischer). Due to the availability of statistical software like JMP (SAS) or open-source programming languiages like R a broader extension of univariate statistics are now frequently used (Bolker et al. 2009). Generalized Linear Models (GLMs) and Generalized Linear Mixed Models (GLMMs) are ubiquitous with experimental work (citation for how many times a common R package is used), and sophisticated linear models in which multiple predictor variables can be included in the same model (GLMM model construction citation).

One of assumption of these univariate linear models is non-independence of predictors (need a classic citation here from glm papers or even a stats textbook). For example, models with a single response variable, like plant growth, assume that light and temperature are independent. It is assumed that these two independent variables do not directly influence each other, but in fact they are fundamentally intertwined bottom-up effects (any citation on plant productivity would work here). If conditions arise in which Factor A and Factor B influence each other or are similarly correlated with Response 1, model interpretations can be misleading. For example, Factor A and Factor B may appear to have weaker effect sizes on Response 1. Also, in cases in which both Factor A and Factor B have poor confidence in rejecting the null hypothesis, there can be a mistaken acceptance of the null hypothesis (Type II error).

Path analysis (SEM) is developed in the ecological literature and is now widely adopted in the fields of food webs and species interactions. More specifically, piecewise structural equation modeling was developed in a series of papers by Shipley (Shipley et al. 2002, etc). Piecewise SEM is often employed in environmental biology via the piecewiseSEM() package developed in 2012 for the R programming environment (Lefcheck 2012).

Use cases for SEM in scenarios include the following:

1. Intermediate mechanisms. The use of biological control agents does not directly impact yield, but instead indirectly increases yield by suppressing. Correct evaluation of the efficiency of biological control requires measuring both the impact on pest populations on host plants as well as changes to yield.
2. Unintended indirect effects compared to main direct effects. Sometimes there can be unintended effects of treatments that mask the primary effect of treatment. For example, the use of some insecticides can lead to secondary outbreaks of mites. Tracking these mite numbers and impacts on yield can be used to determine if this unintended outcome offsets the benefit of treatments.
3. Opposing indirect effects known *a priori*. An experimental treatment impacts yield, but there are multiple intermediate steps in which the treatment ultimately impacts yield, and those effects can be opposing.
4. Evaluation of larger network of interactions: In larger-scale experiments, a large amount of data can be collected, and the primary task is reducing this complexity so management decisions can be made. Path analysis can be used as a model-pruning tool to evaluate which pathways can be removed from a network of variables and still retain some explanatory power.
5. Illustration of system complexity: Visualized structural equation models are powerful pedagogical tools for researchers. Tables and raw data cannot tell a story, but a well-designed interaction network can. By using SEM, one can ensure that the interaction network is build on a quantitative foundation, and that the strength of direct and indirect effects, as well as their statistical significance, are incorporated into diagrams.

Practical limitation 1: Direct effects need to be known ahead of time.

The construction of a path diagram and completion of a path analysis all assume the researcher is able to make strong assertions about cause-and-effect. For example, the addition of natural enemies will reduce the abundance of a prey herbivore. Clark et al. 2016 makes the assertion that a predator exclusion by way of a tanglefoot barrier on tree saplings directly reduce the abundance of predatory ants. Following this, these predatory ants are expected to have a negative impact on the abundance of caterpillars, an established prey item (citation on cc eating caterpillars).

The requirement of known cause-and-effect relationships & the direction of interactions make path analysis a poor tool for data exploration. For studies in complex insect or plant communities, hypothesis may revolve around questions of community partitioning. For example “are pollinator communities different in urban environments compared to conventional rural farms.” In this case, ordination techniques are superior as they do not require assumptions about the interactions among members of this community or even if they directly impact pollination in host plants. Path analysis would be a suitable technique once several key pathways, like a dominant pollinator and its known effect on seed production, are established (are there any pollinator path analysis examples??).

Practical limitation 2: While reciprocal interactions may be of interest to the researcher, path analysis does not accommodate bidirectional pathways.

Feedbacks are common in ecological networks. For example, visitation by nectar-feeding insects can induce more nectar production, which can lead to increased pollinator recruitment (citation needed). Similarly, vector-borne pathogens can weaken the induced defenses of host plants, increasing the fitness and reproduction of the vector herbivore. In systems where higher vector densities increase transmission, a similar feedback loop occurs until the host dies (citation from pea aphids).

To resolve this limitation, path analysis cannot be employed for any sort of hypothesis tests. For larger networks of bidirectional interactions that make hypotheses about connectivity or evenness, bipartite networks would be the appropriate technique.

Practical limitation 3

Categorical variables and higher-order interaction terms affect interpretability. There is rarely embedded in R packages preventing a user form specifying a predictor variable with a large number of classes. For example, a researcher may intend to examine an organic pest control practice with 6 different protocols compared to a control. In this case the model may be approached by comparing these 6 protocols to a control. However, this means drawing 6 lines to additional nodes, and if indirect effects are being modeled, the outcome quickly becomes a tangled mess.

Figure x: 6 nodes going to a single point with cascading indirect effects

Since, in multiple regression, an interaction term is just specifying a larger number of regression parameters, the same issue arises. For example, if a research intends to see how a treatment with 2 levels interactions with two further treatments (2x2x2). Now eight linkages are being analyzed including any cascading impacts.

Figure x: similar outcome

These limitations don’t preclude analysis, but instead present a barrier for interpretability. Effective models convey a clear message and shouldn’t be harder to interpret than the raw data. [Ok Rob is getting to philosophical here, but perhaps there is a citation on the limits of data visualization in complex ecological systems]. Alternatives include

(1) partitioning the analysis into smaller models. In the case of comparing 4 locations, each location could be its own path model.

(2) relying on glms and glmms to analyze the terminal variable of interest. If there are still unresolved concerns about indirect effects a smaller dataset can be used to create a path model.

Limitation 4

Path analysis still is susceptible to overfitting because the component models are glms and glmms. Often this is caused by adding too many predictor variables in hopes of accounting for all possible contingencies. This is a classic problem in the multiple regression family of analyses (citation from stats textbook).

Similarly, path analysis can’t overcome limitations of low replication or pseudocopulation. Good experimental design is still necessary. [Example from only 2 sites of predator release from rorpet]

It is our experience that path analysis is a tool to be considered far into the process of understanding the key environmental variables driving the structure of an interation network. If researchers are still at the exploratory phase and looking to rule out or account for continous variables like daily max temps, elevation, year, etc., traditional multivariate tools like principle components may be more informative.

Limitation 6

Like limitation 4, this is a practical limitation rather than a statistical one. Effective visualization of path diagrams is time consuming and not easily automated programmatically. While there are tools for drawing network diagrams in R and Markup languages, they are generally useful once the analysis is complete. Automatically updating path model visualizations is difficult if the overall layout changes significantly. The problem is further exacerbated if the model has to go through many interactions of peer review or client feedback.

Predefining multiple hypothetical diagrams that can be easily modified can be effective. However, we highlight a few conventions that may aid with cluttered figures but also prevent the amount of fiddling required to finalize a path diagram.

In Clark et al. 2019, non-significant paths were dropped from the analysis

Glossary section:

Path analysis

Structural equation modeling

Autocorrelation

Fixed effects

Random effects

Interaction networks

Tests of directed separation

Path coefficients

Model pruning

References:

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H., White, J.-S.S., 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology & Evolution 24, 127–135. <https://doi.org/10.1016/j.tree.2008.10.008>

Liere, H., Kim, T.N., Werling, B.P., Meehan, T.D., Landis, D.A., Gratton, C., 2015. Trophic cascades in agricultural landscapes: indirect effects of landscape composition on crop yield. Ecological Applications 25, 652–661. <https://doi.org/10.1890/14-0570.1>

Senn, T.E., Espy, K.A., Kaufmann, P.M., 2004. Using Path Analysis to Understand Executive Function Organization in Preschool Children. Developmental Neuropsychology 26, 445–464. <https://doi.org/10.1207/s15326942dn2601_5>

Valente, B.D., Rosa, G.J.M., Gianola, D., Wu, X.-L., Weigel, K., 2013. Is Structural Equation Modeling Advantageous for the Genetic Improvement of Multiple Traits? Genetics 194, 561–572. <https://doi.org/10.1534/genetics.113.151209>

Wei, C., Yu, Q., Bai, E., Lü, X., Li, Q., Xia, J., Kardol, P., Liang, W., Wang, Z., Han, X., 2013. Nitrogen deposition weakens plant-microbe interactions in grassland ecosystems. Glob Change Biol 19, 3688–3697. <https://doi.org/10.1111/gcb.12348>