**Using structural equation modeling to evaluate management of pest 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 must contend with the larger ecological networks in which practices take place. Controls reducing pest or invasive species can have non-target effects, while introducing beneficial species can have unforeseen consequences in food webs. Historically, applied researchers have relied 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, with the piecewiseSEM package being amongst the most popular, due in part to its ability to build directly on generalized linear models already written in R. The piecewiseSEM package is extremely useful when leveraged in applied work, but there are key caveats researchers must address before beginning analysis or drawing conclusions. We outline the utility of piecewise structural equation (path analysis) and suggest key recommendations on its application in ecological management scenarios. Of note, we include suggestions for directly modeling treatments which exclude beneficial natural enemies or predators, and resolving more complex experimental designs in managed ecosystems.

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

**Introduction**

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. Research in the fields of restoration ecology, conservation, ecological pest management and sustainability all rely on experimental actions which are then evaluated for their ability to cause desired changes when scaled. However, applied ecologists must cope with issues like non-target effects, confounding intermediate steps, or opposing indirect effects (Wootton 1994). While not a panacea, structural equation modeling (SEM) has become an important tool supplementing generalized linear models to resolve issues that take an ‘interaction network’ perspective to ecological management of pest or beneficial species (Wei et al. 2013). In multiple fields where interaction networks are considered, SEM has been adopted in psychology (Senn et al. 2004), genetics (Valente et al. 2013), and agriculture (Liere et al. 2015).

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 (Fisher 1919, Parolini 2015)(citation from fischer). Due to the availability of statistical software like JMP (SAS) or open-source programming languages 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 (Kain et al. 2015)(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).

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). Other packages include Lavaan (citation needed) and x, y. The focus of this paper is on piecewiseSEM as it has been applied in multiple systems with success by the authors of this paper.

**Examples 1**

**Part 1: Problems path analysis or structural equation modeling can solve**

*Correlation of predictor variables.* 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).

*Modeling intermediate steps.* 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.

*Modeling non-target effects.* 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.

*Modeling opposing effects running counter to treatments.* 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.

*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.

*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 built on a quantitative foundation, and that the strength of direct and indirect effects, as well as their statistical significance, are incorporated into diagrams.

**Examples 2**

**Part 2: Applied problems path analysis may not be best suited for**

*Exploratory analyses.* 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).

*Networks where cause and effect relationships are not yet established*. 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??).

*Systems in which bidirectional feedbacks dominate* 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). For larger networks of bidirectional interactions that make hypotheses about connectivity or evenness, bipartite networks would be the appropriate technique.

*Experiments with higher-order interaction terms.* 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.

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.

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.

*Experiments with blocks or different treatments with non-additive effects.* An alternative is 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.

*Datasets in which high-level variable selection is necessary.* 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 interaction network. If researchers are still at the exploratory phase and looking to rule out or account for continuous variables like daily max temps, elevation, year, etc., traditional multivariate tools like principle components may be more informative. Model selection is indeed provided in some frameworks (see paper that I said half scooped us).

*Interaction networks with more than ten nodes.* 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 to reduce clutter in the interaction network diagram.

**Conclusions**

A core recommendation for the ecologist is to consider your hypotheses and predictions before running a single statistical test. Restraint prevents unnecessary digging that could be misleading or time-consuming. We suggest the same strategy is used for path analysis. Consider the drawing out of a path diagram, with expected structure, directions, and relative effect sizes as critical steps before even specifying the first structural equation model. This approach is highlighted in the three worked examples with PEMV, forest ants, and mulch treatments (See appendix 1). Data exploration with simple linear models is warranted, but should be limited to evaluating main treatment effects. In our workflow we used the a priori path diagram to write out generalized linear models or generalized linear mixed models. If an *a priori* path diagram fails the test of directed separation (P < 0.05), the response variable with the lowed significance value is added in a single step, then the path analysis is rerun. Once a putative path model is accepted, terms included in the a priori model are removed starting with a single term with the highest P-value, and the change in AIC is noted. This process is repeated until the AIC no longer improves or all pathways determined as experimentally necessary are the only ones retained. The resulting *a posteri* model is used for interpretation, and a comparison of the *a priori* model is used to make qualitative conclusions.

Modern statistical programming methods emphasize the use of algorithmic approaches when large datasets include highly colinear data. In these cases, prediction is more relevant than hypothesis testing. Here we suggest that path analysis and structural equation modeling are viable alternatives to black-box approaches, especially in the context of testing hypotheses regarding management of pest and beneficial species. The machine-learning family of methods are powerful tools for forecasting and prediction, particularly for habitat suitability of invasive species (PNAS giant hornet paper) or dispersal behavior for species of conservation concern. Path analysis offers a stricter framework that requires core assumptions of causality to be met, but often experimenters are able to meet these assumptions in management contexts. In particular, we find this tool to be useful when modeling biotic interactions in which data cannot easily be collected *via* remote sensing. Techniques like random forest have been criticised as problematic for understanding the mechanistic basis for changes, and in some cases are eschewed by ecologists even if traditional hypotheses testing approaches are less robust (citation DEFINITELY needed for this Rob).

The piecewiseSEM library and its associated publication has been cited 2500 times (accession date, March 6, 2023), increasing linearly since release in 2016.