Title

Using structural equation modeling to understand indirect effects in horticulture and integrated pest & pollinator management: (or 10 things I hate about path analysis)

Abstract

Agricultural productivity is determined by both direct and indirect effects on yields. Agricultural scientists can evaluate the direct impacts of new strategies using multiple regression or generalized linear models. 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 appreciated in community ecology as an important tool for quantifying and evaluating indirect effects. Relatively user-friendly R packages have been developed that require only basic knowledge of R commands and linear models to implement, but they have not yet been extensively adopted in agricultural research. In this paper, we provide several examples from real agroecological experiments conducted in the Pacific Northwest crops, and demonstrate the usefulness of path analysis and how to implement it. Importantly, we demonstrate that important biological inferences would be otherwise obfuscated had path analysis not been utilized as a statistical tool. We urge other researchers to attempt this approach and highlight several examples, such as quantification of biological control and crop yield, where path analysis should be considered a default tool for the evaluation of experimental outcomes.

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). Within the field of agriculture, horticultural models and integrated pest management may benefit the most from understanding both direct and indirect effects. In both of these subfields of agriculture, trials are applied to entire plants to improve yield and reduce the impacts of pests. However, there are often multiple intermediate steps involving changes to plants or insects before the desired outcome is achieved.

Most agricultural scientists are trained in so-called traditional statistical techniques involving hypothesis testing of single variables, applying univariate statistics including ANOVA and Regression. Due to the availability of statistical software, a broader extension of univariate statistics are now frequently used (Bolker et al. 2009), including Generalized Linear Models (GLMs) and Generalized Linear Mixed Models (GLMMs). Within the field of agroecology, the most commonly used R package for GLMMs is cited over (xxx) times. This technique is flexible and makes it relatively straightforward to build sophisticated linear models in which multiple predictor variables can be included in the same model (GLMM model construction citation). However, one of the primary assumptions of these univariate approaches is non-independence of predictors. Models assume that Factor A and Factor B do not directly influence each other, and each independently are related to changes in Response 1 (Fig 1). 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).

In specific instances in which Factor A influences Factor B, a stepwise approach can be used to model a series of linear interactions. One common technique is path analysis (a type of structural equation modeling). 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 applied agriculture 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. In hand-pollination trials, increasing the number of flowers pollinated can lead to more fruit. Conversely, trees will produce smaller fruits to offset the cost of producing more fruit, leading to lower yield. Depending on how the model is specified, one can get two seemingly contradictory answers to the question of whether treatments increased yield.
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.

Challenges when applied to agricultural or entomological data (ten limitation of path analysis as its currently employed).

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>