A generalized approach to modeling and estimating indirect effects in ecology

YANN CLOUGH

Agroecology, Department of Crop Sciences, Georg-August Universität Göttingen, Grisebachstrasse 6 D-37077 Göttingen, Germany

Abstract. The need to model and test hypotheses about complex ecological systems has led to a steady increase in use of path analytical techniques, which allow the modeling of multiple multivariate dependencies reflecting hypothesized causation and mechanisms. The aim is to achieve the estimation of direct, indirect, and total effects of one variable on another and to assess the adequacy of whole models. Path analytical techniques based on maximum likelihood currently used in ecology are rarely adequate for ecological data, which are often sparse, multi-level, and may contain nonlinear relationships as well as nonnormal response data such as counts or proportion data. Here I introduce a more flexible approach in the form of the joint application of hierarchical Bayes, Markov chain Monte Carlo algorithms, Shipley's d-sep test, and the potential outcomes framework to fit path models as well as to decompose and estimate effects. An example based on the direct and indirect interactions between ants, two insect herbivores, and a plant species demonstrates the implementation of these techniques, using freely available software.

Key words: BUGS; causal models; d-separation; indirect effects; mechanistic models; mediation; SEM; structural equation models; trait-mediated indirect effects.

INTRODUCTION

Univariate models have obvious limitations when it comes to fitting more complex models that include multiple relationships with direct and indirect effects (Wootton 2002, Grace 2006). In contrast, path analytical techniques (including mediation and structural equation models) are used to specify, estimate, and evaluate causal models outlining the hypothesized direct and indirect effects or “paths” between the input variable and the outcome variable (Grace 2006). The indirect effects are mediated by one or several other variables, which are both response and explanatory.

Modern approaches to path analysis use maximum likelihood estimation to compare the elements of the observed variance–covariance matrix to that expected given the specification of the model (Grace 2006; for a recent example, see Scherber et al. 2010). The significance of the discrepancy is evaluated by global fit measures. These methods have long been associated with constraints that make them difficult to apply to ecological data sets, such as the assumption of independence between data points, the assumption of multivariate normality, and the restriction to linear models. The sensitivity of global fit measures such as the \( \chi^2 \) statistic to small sample sizes (Bearden et al. 1982) can be expected to be a further issue in many ecological studies. While several commercial packages for modern path analysis have begun providing solutions to individual issues (e.g., Asparouhov and Muthén 2010), a more general approach to path modeling based on multiple univariate models was proposed by Shipley (2009). This approach is attractive as it allows the incorporation of models which are becoming standard in the analysis of ecological data sets, such as generalized linear models and mixed models, within path models. Shipley’s method allows to specify and test for the adequacy of path models, but not to estimate effect sizes.

Here I introduce a flexible approach to path analysis that allows direct and indirect effect decomposition in addition to path model specification and testing while avoiding the problems associated with maximum-likelihood-based methods. The parameters associated with the paths are estimated via simultaneous estimation of several univariate Bayesian hierarchical models with Markov chain Monte Carlo methods (MCMC; Gelman and Hill 2007). After testing for model adequacy following Shipley (2009), the joint posterior distributions of the parameters in the model are used in the potential outcomes framework (Pearl 2001, Imai et al. 2010a, b) for effect decomposition. A motivating example is provided by the evaluation of plant-mediated indirect interactions between herbivores (Ohgushi 2005) in an investigation of the effect of an aggressive, numerically and behaviorally dominant ant species on two herbivorous insects, where data are not only spatially grouped at two higher levels but also include binary as well as strongly overdispersed count and proportion data, which makes the use of existing methods difficult.
METHODS

Model specification.—As in classical path analysis, the model is specified graphically by linking variables with directed arrows representing hypothesized causal effects. The resulting path model should be non-recursive, i.e., specification of simultaneous feedbacks and non-directed correlations are not allowed. Here, each arrow represents a slope in a univariate model. A variable receiving an arrow is a response variable. In the case of binomial and binary models, the response is represented by several variables: the observed quantity (presence/absence for binary, number of “successes” for binominal distributions), the associated but unobserved incidence probability, and, in binomial models, the number of “trials”; these are not linked with slope parameter-associated arrows unless the number of “trials” is assumed to impact the probability of “successes.”

Bayesian inference and MCMC sampling.—In the Bayesian modeling approach (Gelman et al. 2004, Gelman and Hill 2007), both data and the parameters are quantified as distributions. The posterior distribution of the parameters is obtained by updating the prior distribution with the information contained in the observed data. Stochastic techniques can be used to obtain a sample of values from the posterior. Inference is made on these samples. A method to conduct this sampling is MCMC (Markov chain Monte Carlo), in which a random walk over the probability distribution, or “chain,” is produced. The posterior distribution can be plotted using density plots, or summarized using point estimates, such as the mean or mode, and credible intervals. A 95% credible interval (CI) indicates that there is 95% chance the interval contains the true value of the parameter given the data and the model. The results thus obtained are valid for all sample sizes. However, inference can only be drawn if the chain has converged on the posterior distribution, which can be checked by running three chains, and assessing the mixing of the chains both graphically and by examining the R-hat statistic as a formal test of convergence. The part of the chain(s) produced before convergence is called the “burn-in” and should be discarded.

Goodness of fit.—Individual univariate models are assessed using the posterior predictive (PP) checking approach. Predicted values are generated as a byproduct of the MCMC sampling procedure, and a so-called PP P value is then computed by testing the null hypothesis that the observed value comes from the predictive distribution, using a test quantity called the discrepancy measure (Gelman et al. 2004). One example is the sum of squared Pearson residuals. Well-fitting models have Bayesian P values near 0.5, while values close to 0 or 1 suggest lack of fit. These P values cannot be used to compare alternative models (Gelman et al. 2004).

Testing for path model adequacy.—The path model is tested by using Shipley’s d-sep test (Shipley 2009), which I outline as follows (adapted from Shipley 2009):

1) List all pairs of variables that do not have an arrow between them.
2) For each pair list the set of other variables in the graph that are direct causes of one or the other variables in the pair. The pair of variables, along with this so-called “conditioning set,” define an independence claim and the full set of independence claims defines a “basis set.”
3) For each independence claim, formulate a model with one of the variables in the pair as a response, and the other variable in the pair, together with all variables in the conditioning set, as explanatory variables. Obtain the P value associated with the probability that the pair of variables is statistically independent conditional on the “conditioning set” using the posterior distribution of the slope for the second variable in the pair.
4) The probabilities calculated above are combined by computing the C value, where k is the number of claims tested, and pi the P value for the ith model (1 ≤ i ≤ k):

\[ C = -2 \sum_{i=1}^{k} \log(p_i). \] (1)

This value is compared to a \( \chi^2 \) distribution with 2k degrees of freedom. The path model is rejected if the P value is lower than a specified cut off point (e.g., \( P < 0.05 \)).

Estimating direct, indirect, and total effects.—Computing the path coefficients is straightforward only for linear models and normally distributed data (Yuan and MacKinnon 2009), where indirect effect estimates are obtained by multiplying the associated direct effects (the “calculus of coefficients” approach). A general non-parametric approach is provided by the potential outcomes framework (Pearl 2001) used in mediation analysis for calculating direct, indirect and total effects (Imai et al. 2010a, b). Specifically, it is applicable also to models with nonlinear relationships and discrete dependent variables. I exemplify the potential outcomes approach with the following model, where variable \( X \) is modeled as having a direct and an indirect effect via \( Y_1 \) on the outcome variable \( Y_2 \), such that, for each unit of observation,

\[ Y_{1i} = \alpha_1 + \beta_1 \times X_i \] (2)

\[ Y_{2i} = \alpha_2 + \beta_2 \times X_i + \beta_3 \times Y_{1i}. \] (3)

\( Y_1 \) is called the mediator or mediating variable. I write the observed outcomes of \( Y_1 \) and \( Y_2 \) as \( Y_{1i}(X_i) \) and \( Y_{2i}(X_i, Y_{1i}) \). The effect decomposition is done for a specified change to the input variable, from a reference level \( X_{ref} \) to a treatment level \( X_{treat} \) (Imai et al. 2010b). Unless dictated by the study design, these levels are arbitrary. In the ant–herbivores example (see Example), I define \( X_{ref} \) and \( X_{treat} \) as absence and presence of the ant, respectively. The formulation of an indirect effect corresponds to the change that would occur to the
outcome if one changes the \( Y_1 \), from the value that would be realized under the reference condition, \( Y_1(X_{\text{ref}}) \), to the value that would be observed under the treatment condition, \( Y_1(X_{\text{treat}}) \), while holding the treatment status at \( X_{\text{ref}} \). Similarly, the direct effect is formulated as the change to the outcome that would occur if one changes \( X_i \) from \( X_{\text{ref}} \) to \( X_{\text{treat}} \), while holding \( Y_1 \) at the value it would have under \( X_{\text{ref}} \). I write, for each unit \( i \), the total, direct and indirect effects \( \tau_i \), \( \delta_i \), and \( \zeta_i \) as

\[
\begin{align*}
\tau_i &= Y_2(X_{\text{treat}}, Y_1(X_{\text{treat}})) - Y_2(X_{\text{ref}}, Y_1(X_{\text{ref}})) \\
\delta_i &= Y_2(X_{\text{treat}}, Y_1(X_{\text{treat}})) - Y_2(X_{\text{ref}}, Y_1(X_{\text{ref}})) \\
\zeta_i &= Y_2(X_{\text{treat}}, Y_1(X_{\text{treat}})) - Y_2(X_{\text{treat}}, Y_1(X_{\text{ref}})).
\end{align*}
\]

We also have

\[
\tau_i = \delta_i(X_{\text{treat}}) + \zeta_i(X_{\text{treat}}).
\]

My aim is to calculate the average, or expected effects \( \bar{\tau} \), \( \bar{\delta} \), and \( \bar{\zeta} \). A theorem for the identification of average causal effects was published and an algorithm to compute estimates of these effects proposed by Imai et al. (2010a, b):

1) Fit models for the observed outcome and mediator variables.
2) Simulate \( M \) sets of model parameters from their (multivariate) sampling distribution. In the Bayesian MCMC estimation these are the samples of the joint posterior distribution.
3) Repeat the following three steps for each \( m \) (\( 1 \leq m \leq M \)):
   a) for each \( i \) simulate \( Q \) potential values of the mediator, in this example once for \( Y_1(X_{\text{ref}}) \) and once for \( Y_1(X_{\text{treat}}) \). The simulation is necessary only for discrete mediators, such as in the Poisson or logistic regression, where the expected values are generally continuous, while the realized values are discrete.
   b) For each combination of \( X_{\text{ref}}, X_{\text{treat}}, Y_1(X_{\text{ref}}) \) and \( Y_1(X_{\text{treat}}) \), and for each combination of \( i \) and \( q \) (\( 1 \leq q \leq Q \)) simulate the potential outcomes of \( Y_2 \) given the simulated values of the mediator, thus obtaining samples of the distribution of the two components of the right-hand side of Eqs. 4, 5, and 6.
   c) Calculate the \( M \) samples of the distribution of the average effects by averaging across the \( Q \) simulations and the \( n \) observations as exemplified here for the indirect effects,

\[
\delta = \frac{1}{nQ} \sum_{i=1}^{n} \sum_{q=1}^{Q} (Y_2(X_{\text{treat}}, Y_1(X_{\text{treat}})) - Y_2(X_{\text{treat}}, Y_1(X_{\text{ref}}))).
\]

4) Compute point estimates and confidence intervals based on the \( M \) samples of the direct, indirect and total average effects.

The theorem proposed by Imai et al. (2010a) holds for single-level models under sequential ignorability, meaning that conditional on the observed pretreatment covariates, the treatment is independent of all potential values of the outcome and mediating variables, and that the observed mediator is independent of all potential outcomes given the observed treatment and pretreatment covariates. The estimates of effects have been shown to be robust to confounding by variables correlated with both treatment and outcome only if the confounding effect is weak (Imai et al. 2010a). Likely confounders should thus be recorded and included in the model (see Discussion). This approach can be extended to the case where the outcome is affected only by indirect effects, or to cases with multiple mediating variables. If one assumes no interference between individuals within groups, i.e., the outcomes of one individual do not depend on the value of the treatment and mediating variables for another individual, this approach can also be used for multilevel models (VanderWeele 2010).

**Example**

*Study system.* I investigated the effect of an aggressive, numerically and behaviorally dominant ant species, *Philidris cf. cordata*, on two herbivorous insects feeding on cacao pods, the fruit of the tropical tree *Theobroma cacao* in Indonesia. The herbivores are the mirid bug *Helopeltis sulawesi* and the pod borer *Comonomophora crameterella*, a gracillarid moth (Appendix A: Fig. A1). The study comprised 43 plots with one fertilized and one unfertilized subplot each, with five focal trees per subplot. On the 430 trees the presence or absence of *P. cf. cordata* and the incidence of herbivore damage were recorded every two weeks for six months. I expected that *P. cf. cordata* would decrease the incidence of both herbivores on the trees it occupies via predation or predator avoidance (Fig. 1A). However, preliminary analyses suggested a positive association between the ant and the pod borer. This lead to a revised hypothesis (Fig. 1B) in which *P. cf. cordata* has negative direct effects on both herbivores, with the incidence of *C. crameterella* being negatively affected by the incidence of *H. sulawesi*. Avoidance of pods with feeding marks of *H. sulawesi* can occur because the feeding activity by *H. sulawesi* temporally precedes the oviposition of *C. crameterella*. Thus, *P. cf. cordata* may have an indirect positive effect on *C. crameterella* reducing pre-emptive competition by *H. sulawesi*. In addition, several ecological covariates, available at plot, subplot or tree scale were hypothesized to affect the presence of the ant as well as herbivore incidence (see Appendix A). My goal was to compute the direct and indirect effects of the ant on the herbivores and compare the model with the indirect interaction with the model with only direct interactions, while taking into account the variable type (discrete, continuous), the
Fig. 1. Alternative conceptual models of the effects of the dolichoderine ant *Philidris cf. cordata* on two herbivores of *Theobroma cacao*, the mirid bug *Helopeltis sulawesi* and the gracillarid moth *Conopomorpha cramerella*. Plus and minus signs indicate a positive or negative effect, respectively. (A) Original model based only on predation or nonlethal effects of the ant on the herbivores. (B) Model including the hypothesized positive indirect effect of the ant on *C. cramerella*, mediated by the predation or nonlethal effects of the ant on *H. sulawesi*, and the avoidance of mirid feeding damage by *C. cramerella*.

Spatial design of the survey (trees nested in subplots nested in plots), and the covariates.

Model formulation.—I fitted two models, one without the interaction between herbivores (model A), and one including the interaction (model B) (Fig. 2). The trees are indexed by $i$ ($1 \leq i \leq 430$), the subplot by $j$ ($1 \leq j \leq 86$) and the plots by $k$ ($1 \leq k \leq 43$). Variables used in the models are presence/absence of *P. cf. cordata* ($Pc$, 0 = absent, 1 = present), number of harvested pods ($N_{podh}$), number of pods damaged by *H. sulawesi* ($Hs$), number of pods damaged by *C. cramerella* ($Cc$), and the unobserved probabilities leading to the discrete outcomes in the logistic models ($Pc$, $Pc$, $Pc$, $Hs$, and $Pc$). The covariates are application of urea ($N_{fert}$, 0 = control, 1 = fertilized), mean annual temperature ($temp$, °C) and age of the cacao trees in a plot (age, in years). In the following, I detail the structure of model B (with interaction), which is illustrated in Fig. 2. Intercepts are denoted by $\alpha$, slopes by $\beta$, the suffix “.hat” means “estimate of the mean.” To model the number of harvested pods I use a hierarchical, overdispersed Poisson model:

$$N_{podh} \sim \text{Poisson}(N_{podh}.\hat{hat})$$

$$\log(N_{podh}.\hat{hat}) = \alpha_{N_{podh}} + \epsilon_{N_{podh}}$$

with

$$\epsilon_{N_{podh}} \sim \text{Normal}(0, \sigma_{N_{podh}})$$

$$\alpha_{N_{podh}} \sim \text{Normal}(\alpha_{N_{podh}.\hat{hat}}, \sigma_{N_{podh}.\text{subplot}})$$

$$\alpha_{N_{podh}.\hat{hat}} = \alpha_{N_{podh}} + \beta_{N_{fert} \cdot N_{podh}} \times N_{fert}$$

$$\alpha_{N_{podh},k} \sim \text{Normal}(\alpha_{N_{podh}.\hat{hat},k}, \sigma_{N_{podh}.\text{plot}})$$

$$\alpha_{N_{podh}.\hat{hat},k} = \alpha_{N_{podh},k} + \beta_{temp \cdot N_{podh}} \times temp_k.$$

The model for *P. cf. cordata* is specified as a hierarchical logistic model. Convergence diagnostics suggested joint estimation of subplot and plot-scale variance in presence/absence of *P. cf. cordata* was difficult, so subplot-level variance was dropped, resulting in a rapidly converging overall model, but without affecting the estimates of other parameters in the model:

$$Pc_j \sim \text{Binary}(Pc_{\hat{hat}})$$

$$\text{logit}(Pc_{\hat{hat}}) = \alpha_{Pc} + \epsilon_{Pc}$$

$$Pc_{\hat{hat}} = \alpha_{Pc} + \beta_{N_{fert} \cdot Pc} \times N_{fert}$$

$$Pc_{\hat{hat},k} \sim \text{Normal}(\alpha_{Pc,\hat{hat},k}, \sigma_{Pc,\text{plot}})$$

$$Pc_{\hat{hat},k} = \alpha_{Pc,\hat{hat},k} + \beta_{temp \cdot Pc} \times temp_k + \beta_{age \cdot Pc} \times age_k.$$
The original model without the interaction (model A) is identical except the model formulation for the probability of occurrence of *C. cramerella*, which is hypothesized not to be affected by *H. sulawesi*: \( \logit(P_{Cc}) = \alpha_{Cc} + \beta_{Pc_{Cc}} \times P_{Cc} + \epsilon_{Cc} \)

with

\[ \epsilon_{Cc} \sim \text{Normal}(0, \sigma_{Cc}) \]

\[ \alpha_{Cc} \sim \text{Normal}(\alpha_{Cc, hat}, \sigma_{Cc, subplot}) \]

\[ \alpha_{Cc, hat} = \alpha_{Cc} + \beta_{Nfert_{Cc}} \times Nfert_{j} \]

\[ \alpha_{Cc, k} \sim \text{Normal}(\alpha_{Cc, hat, k}, \sigma_{Cc, plot}) \]

\[ \alpha_{Cc, hat, k} = \alpha_{Cc} + \beta_{temp_{Cc}} \times temp_{k} \]

The data, and BUGS and R code can be found in the Supplement.

**Results.**—While the posterior predictive \( P \) values based on discrepancy (sum of squared Pearson residuals) in data observed and simulated given the model suggest the specification of the individual multi-level models is appropriate for both model A and B (all \( P \) values between 0.37 and 0.53 for model A, and between 0.36 and 0.53 for model B; Appendix B: Figs. B1 and F2).
B2), Shipley’s d-sep test shows model A is not adequate (C = 50.907; k = 8; \( P(\chi^2_k) < 0.0001 \)), while model B, which includes the interaction between the two herbivores, is adequate (C = 12.636; k = 7; \( P(\chi^2_k) = 0.555 \)). The details of Shipley’s test as applied to model A and B are provided in Appendix C. The posterior distributions of the total, direct and indirect effects of the ant on C. cramerella based on model B are shown in Fig. 3 whereas the mean of the posterior of all other slopes are shown in Fig. 2 and Appendix D: Table D1. The 95% CI for the estimates of the (logit-scale) effect of the ant P. cf. cordata on the probability of occurrence of C. cramerella includes zero (posterior mean \( \beta_{PC_{Cc}} \) = 0.123, 95% CI = –0.079, 0.335; Fig. 2) whereas the effect of the presence of the proportion of pods affected by H. sulawesi on the probability of occurrence of C. cramerella is negative, with the 95% CI not including zero (posterior mean \( \beta_{HC_{Cc}} \) = –2.797, 95% CI = –3.160, –2.436; Fig. 2). The estimate for the effect of the ant P. cf. cordata on the probability of occurrence of H. sulawesi (posterior mean \( \beta_{PC_{Hs}} \) = –0.526, 95% CI = –0.932, –0.092; Fig. 2) is strongly negative and its 95% CI does not include zero. The mean average direct effect in terms of percentage pods affected by C. cramerella is small and its 95% CI includes zero (mean = 2.55, 95% CI = –1.67, 6.80; Fig. 3). The mean of the average indirect effect is larger and its 95% CI does not include zero (mean = 6.15, 95% CI = 1.00, 11.13; Fig. 3). In other words this effect may account on average for up to 12% incidence by C. cramerella. A mean 78% of the total effect (95% CI = 24, 142; Fig. 3) is accounted for by the indirect effect. Note that the distribution of the relative indirect effects can include values below 0 and above 100 given the possibility of sign differences between indirect and direct effects in the posterior sample (Fig. 3).

**CONCLUSIONS**

By using state of the art generalized multilevel models, I could show that the inclusion of indirect interactions between herbivores is necessary to adequately model the impact of a dolichoderine ant on herbivores of T. cacao. Using the potential outcomes framework I could estimate direct and indirect effects of the ant on the herbivore C. cramerella, and show that the latter benefits from a reduction by the ant in the incidence of another herbivore, H. sulawesi.

The approach I present shares some inherent dangers with path analysis techniques presently used in ecology. The validity of the causal interpretation of a model and the resulting effect decomposition is contingent on a well-specified model and the inclusion of important confounding variables (Grace 2006). This can generally not be proven, since there are an infinite number of the latter. Therefore, the ecological soundness and plausibility of the models compared requires the diligence of the analyst. For this reason, emphasis is put here on (1) the comparison of plausible alternative models rooted in theory and prior knowledge and (2) iterative processes in which modeling is alternated with data collection leading to increasingly refined hypothetical models. In the example used in this paper, the unexpected positive relationship between the ant P. cf. cordata and C. cramerella in a model encompassing only direct effects was difficult to explain biologically. This prompted a choice experiment in which gravid females of C. cramerella were offered pods with and without damage by H. sulawesi. A very strong preference for undamaged pods (Wielgoss et al. 2012) supported a mechanistic basis for the effect of H. sulawesi on C. cramerella and therefore further justified the inclusion of this effect in model B. This does not of course exclude the possibility that other paths have been omitted. For example, P. cf. cordata tends pseudococcids, which could affect the two herbivores studied via host-plant quality. This may explain that while its 95% CI contains zero, the mean of the direct effect on the ant on C. cramerella is still positive in the model including an indirect interaction between the two herbivores.

Several other recent publications have addressed related methods. Multilevel analyses for path analysis and structural equation models (SEM) are being developed elsewhere (Asparouhov and Muthén 2010),
but rely on commercial software and are not yet well documented. Finally Imai et al. (2010b) implemented the potential outcomes approach for single-level, single-mediator models in an R package.

The generality of the methods jointly applied in this study, namely hierarchical Bayesian modeling, the d-sep test and the potential outcomes framework, is an advantage over path analytical techniques presently used in ecology. The flexibility to bring together hierarchical data from different distribution types is a major benefit here, as has been recognized in the context of univariate multilevel models (Gelman and Hill 2007). A limitation is the difficulty to model recursive paths such as simultaneous feedbacks and non-directed correlations, which could be circumvented by making models dynamic. The method can further be extended to combine models of process and measurement error, a subject well covered in hierarchical Bayes literature written for ecologists (e.g., Clark and Gelfand 2006).

ACKNOWLEDGMENTS
I thank A. C. Wielgoss, Iqbal, and Team Kakao for field work. Comments by C. Dormann, B. Shipley, and an anonymous reviewer greatly improved previous versions of the manuscript. Part of the work was supported by DFG Grant SFB 552-STORMA and a DAAD Grant to A. C. Wielgoss.

LITERATURE CITED

SUPPLEMENTAL MATERIAL
Appendix A
Illustrations and details of the study system with a supporting figure and text (Ecological Archives E093-160-A1).

Appendix B
Posterior predictive diagnostic plots for the univariate models within path models A and B (Ecological Archives E093-160-A2).

Appendix C
Details of the d-sep procedure as applied to models A and B (Ecological Archives E093-160-A3).

Appendix D
A table summarizing the posterior distributions for parameters of model B (Ecological Archives E093-160-A4).

Supplement
The data as well as the R and WinBUGS code to conduct the analyses (Ecological Archives E093-160-A5).