



A multigroup extension to piecewise path analysis

JACOB C. DOUMA ^{1,2,†} AND BILL SHIPLEY ³

¹Centre for Crop Systems Analysis, Wageningen University, Droevendaalsesteeg 1, Wageningen 6708PB The Netherlands

²Laboratory of Entomology, Wageningen University, Droevendaalsesteeg 1, Wageningen 6708PB The Netherlands

³Département de Biologie, Université de Sherbrooke, Sherbrooke, Quebec J1K R1 Canada

Citation: Douma, J. C., and B. Shipley. 2021. A multigroup extension to piecewise path analysis. *Ecosphere* 12(5):e03502. 10.1002/ecs2.3502

Abstract. Path analysis allows one to test the consistency of data to hypothesized causal relationships between variables. Often, interest lies in how the hypothesized dependencies differ between groups. Multigroup comparisons can be made by imposing various constraints: constraints on the topology, the path coefficients, the residual variances, and more. To date, only classical path analysis and structural equation modeling can account for differences between groups. These techniques have assumptions that are often not appropriate for ecological studies. The d-sep test and the recently developed generalized chi-squared test relax many of these assumptions for path models that can be represented as directed acyclic graphs (DAGs), but are currently lacking a multigroup test. In this paper, we develop a multigroup extension to the d-sep test. Furthermore, we show how a recently developed generalized chi-squared test and AIC for DAGs can be used for multigroup testing. The approaches are illustrated by a worked example and implemented in the commonly used statistical package, R. Practical recommendations for multigroup modeling are made, and advantages and disadvantages of the multigroup d-sep and the chi-squared test are discussed.

Key words: d-separation; equality constraints; group comparison; hierarchical designs; nonnormal data; path analysis; structural causal modeling; structural equation modeling.

Received 19 June 2020; revised 17 December 2020; accepted 8 January 2021; final version received 25 February 2021. Corresponding Editor: Debra P. C. Peters.

Copyright: © 2021 The Authors. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

† **E-mail:** bob.douma@wur.nl

INTRODUCTION

Ecological studies often involve many measured variables that are interrelated in one way or another. The causal interdependencies between such variables can be investigated through path analysis, a type of structural equation modeling (SEM; Kline 2005, Grace 2006, Shipley 2016). This technique allows one to test, and potentially reject, a multivariate hypothesis concerning the underlying causal structure that generates the observed values of a multivariate observation. An important assumption of this technique is that all of the multivariate observations in the data set are generated

by the same causal structure (causal homogeneity), in terms of both its topology and the quantitative values of the free parameters linking the causal parent–child pairs. When this assumption does not hold, or when we wish to statistically test this assumption in order to determine whether (and where) the causal structure might differ between groups of observations, one must conduct a multigroup path analysis.

Group comparisons are an important part of statistical inference and involve testing if groups differ significantly from each other with respect to their mean, the degree of association between two variables, or their variance. Group

comparisons in a multivariate causal framework are currently restricted to classical path analysis and structural equation models (e.g., Kline 2005). Classical path analysis assumes that the data come from a multivariate normal distribution observations are independent of each other, and variables are linearly related (Kline 2012). Much effort has gone into relaxing these assumptions. For example, estimators have been developed that are quite robust to multivariate non-normality unless your data are strongly kurtotic and sample sizes are low (Shipley 2016). Furthermore, techniques are available to include simple curvilinear relationships and relatively simple hierarchical designs (Oberski 2014, Newsom 2015). Yet, standard SEM programs are quite limited in their ability to handle hierarchical designs and curvilinear relationships, and other methods are available that can more naturally accommodate nonlinearity, non-normality, small samples, and complicated hierarchical designs.

Besides classical path analysis, two additional methods are applicable to path models that only contain directed arrows, and do not contain feedback loops, that is, information flows in only one direction (a so-called directed acyclic graph, DAG), but multigroup extensions for these two additional methods do not yet exist. The first is the method of d-separation (Shipley 2000, 2009). Currently, multigroup comparison in the d-sep framework is only possible by either imposing no equality constraints on the groups or assuming all coefficients to be equal among groups (Shipley 2016). The second method is a generalization of the chi-square statistic that is used in classical maximum-likelihood SEM to DAGs whose piecewise functional relationships can be modeled in a maximum-likelihood context. In this paper, we present extensions to multiple group comparisons for both methods.

RATIONALE OF THE D-SEP TEST

The d-sep test is a general framework in which multivariate causal hypotheses can be tested by specifying how the variables are causally linked. The causal links between variables are specified in a directed acyclic graph (DAG; see Fig. 1a for an example) with the variables representing the nodes and the directional arrows representing the edges. From the DAG, it follows which

variables are causally dependent or independent of others, and how this causal dependence changes when physically holding some of the variables constant. If the observational data are generated by the underlying causal structure of the DAG, d-separation determines whether pairs of variables in the DAG will, or will not, be statistically independent upon conditioning in the statistical population (see for details Shipley 2016). Given this, the null probabilities (p_i) of independence associated with the k d-separation claims in the union basis set can be combined using the C-statistic and tested with the chi-square distribution (χ^2) whose degrees of freedom are $2k$:

$$C = -2 \sum_{i=1}^k \ln(p_i).$$

RATIONALE OF THE MULTIGROUP D-SEP TEST

A multigroup d-sep test tests whether the topology and/or the free parameter(s) of the model differs between groups. To do so, we construct for each group a multivariate causal hypothesis in the form of a DAG. The attribute defining group membership is not part of the multivariate causal hypothesis itself. Since the C-statistic follows a chi-square distribution with $2k$ degrees of freedom, and since the sum of N independent chi-square-distributed variables is itself chi-square-distributed with degrees of freedom equal to the sum of degrees of freedom of the N variables, it holds that:

$$\sum_{j=1}^N C_j \sim \chi^2_{(\sum_{j=1}^N 2k_j)}$$

Thus, the sum of the C-statistics and the sum of the degrees of freedom ($2k_j$) for each DAG result in an overall measure of agreement that can be used to assess the consistency of the data with the model.

From the above, it follows that the number of independence claims does not change with the number of constraints imposed on the parameters of the structural equations, except when changing the cause-effect structure of (one) the groups. However, by adding constraints on the free parameters, the statistical independence that was implied by the d-separation claim may no longer hold and this will negatively affect the C-statistic.

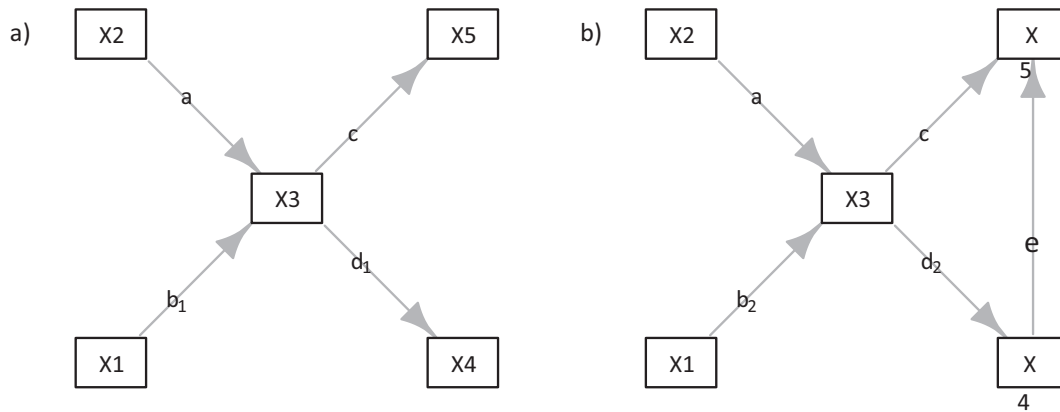


Fig. 1. Directed acyclic graph of two groups: (a) group 1 and (b) group 2. Boxes indicate the variables, and the arrows, the cause–effect relationships between the variables. The path coefficients are represented by letters. If both groups have the same letter without an index for group number, this implies that the path coefficients are equal across groups. For simplicity, the error variables are not shown.

The steps to fit and compare multigroup d-sep models are as follows:

- 1) Specify a causal hypothesis for each group in the form of a DAG. The DAGs can be different across groups provided that the causal relationships between at least two variables have the same direction in all groups.
- 2) Define the union basis set for each group.
- 3) Choose the appropriate model for the regressions that test the independence of a pair of variables in each d-separation claim. Apply constraints to the parameters of interest—either by setting them to a predefined value, or by constraining them to be equal across groups—and calculate the probability that the pair of variables in the basis set is not statistically associated. See Appendix S1 how various constraints can be applied in the software package R.
- 4) Collect the C-statistic per group and number of k independence claims in each group, and calculate the overall chi-square value and the degrees of freedom. See Appendix S1 for an illustration in R.
- 5) Repeat steps 1–4 with different parameters constraints or ancestor–descendant relationships. If multiple models are tested, one runs the risk of erroneously rejecting a model as being not consistent with the data. To control for this problem, one could apply a Bonferroni correction or the false discovery rate (Benjamini and Hochberg 1995). Note that some paths may not be part of the basis set, and thus, adding constraints to these paths will not affect the C-statistic.
- 6) After accepting the model as being consistent with the data, fit the regressions representing the individual paths. Some paths may not be part of the basis set. For these, fit different levels of group constraints and choose the most parsimonious model.

In case the union basis sets differ across groups (Fig. 1) and constraints are imposed on the parameters, the following procedure should be used. Check whether the d-sep claim for a given pair of variables is different across groups. If so, add the variables to the d-sep claim that are direct ancestors of this pair in any of the other groups. Because the path from those ancestors to the pair of variables is lacking in the group of interest, its coefficient should be forced to zero (see Table 1).

RATIONALE OF MULTIGROUP χ^2_{ML} TEST

In classical (covariance) SEM, the consistency of the hypothesized causal structure with the data is tested by measuring the deviance of the model-implied covariance matrix to the observed covariance matrix. If the causal structure is underlying the data, any difference between the

Table 1. The basis set of d-separation claims and the relationships to test for independence that follow from the DAGs of two groups.

Claim number	Group 1		Group 2	
	d-sep claims	Relationship to test	d-sep claims	Relationship to test
1	$X_1 \perp\!\!\!\perp X_2 \mid \{\}$	$X_1 \sim (z_1, z_2)X_2$	$X_1 \perp\!\!\!\perp X_2 \mid \{\}$	$X_1 \sim (z_1, z_2)X_2$
2	$X_5 \perp\!\!\!\perp X_2 \mid \{X_3\}$	$X_5 \sim (c, c)X_3 + (0, e)X_4 + (z_1, z_2)X_2$	$X_5 \perp\!\!\!\perp X_2 \mid \{X_3, X_4\}$	$X_5 \sim (c, c)X_3 + (0, e)X_4 + (z_1, z_2)X_2$
3	$X_4 \perp\!\!\!\perp X_2 \mid \{X_3\}$	$X_4 \sim (d_1, d_2)X_3 + (z_1, z_2)X_2$	$X_4 \perp\!\!\!\perp X_2 \mid \{X_3\}$	$X_4 \sim (d_1, d_2)X_3 + (z_1, z_2)X_2$
4	$X_1 \perp\!\!\!\perp X_5 \mid \{X_3\}$	$X_5 \sim (c, c)X_3 + (0, e)X_4 + (z_1, z_2)X_1$	$X_1 \perp\!\!\!\perp X_5 \mid \{X_3, X_4\}$	$X_5 \sim (c, c)X_3 + (0, e)X_4 + (z_1, z_2)X_1$
5	$X_1 \perp\!\!\!\perp X_4 \mid \{X_3\}$	$X_4 \sim (d_1, d_2)X_3 + (z_1, z_2)X_1$	$X_1 \perp\!\!\!\perp X_4 \mid \{X_3\}$	$X_4 \sim (d_1, d_2)X_3 + (z_1, z_2)X_1$
6	$X_5 \perp\!\!\!\perp X_4 \mid \{X_3\}$	$X_5 \sim (c, c)X_3 + (z_1, e)X_4$		

Notes: The claims of independence are tested by regressing the pair of variables (e.g., X_1 and X_4) against each other while holding the causal parents constant. The regression is written as $X_4 \sim (d_1, d_2)X_3 + (z_1, z_2)X_1$ with the path coefficients indicated by letters. A letter with subscript implies the regression coefficients are different across groups. If the data are consistent with the model, the regression coefficients z_1 and z_2 are not significantly different from zero. When imposing parameter constraints across groups, the d-sep claim may need to be expanded by variables that are causal ancestors of the pair of interest in of the other groups. For example, to test the independence of X_5 and X_2 given X_3 in group 1 while constraining the effect of X_3 on X_5 between groups, the effect of X_4 on X_5 in group 2 needs to be taken into account.

observed and model-implied covariance matrix would be due to random sampling variation, and the maximum-likelihood chi-squared statistic (X_{ML}^2) will asymptotically follow a χ^2 distribution with appropriate degrees of freedom. In this procedure, the parameters of the model-implied covariance matrix are optimized simultaneously and are therefore referred as global estimation ($X_{G,ML}^2$).

The above described method can be generalized to distributions other than normal, nonlinear relationships and hierarchical data (Shipley and Douma 2020) because the joint probability distribution of n variables represented by a DAG can be decomposed into n univariate probability distributions through a local Markov decomposition. Thus, the joint log likelihood of n variables can be obtained by summing the univariate log likelihoods, both in the hypothesized model and in the saturated model. The saturated model is a path model that assumes no missing arrows between the variables. If the hypothesized causal structure could have generated the data, then twice the difference between the likelihood of both models will asymptotically follow a chi-squared distribution with the degrees of freedom equal to the difference in number of free parameters between the hypothesized model and the saturated model (Wilks 1938). When variables are normally distributed and linearly related, this gives identical results as the classical covariance path analysis. Because the likelihood of both

models is calculated from the local regressions, we refer to this X_{ML}^2 statistic as $X_{L,ML}^2$.

Within this framework, one can test, identical to classical SEM, the consistency of multigroup path models by constraining some coefficients to be equal across groups, and a saturated model that assumes that all estimated coefficients to be group-specific. For a multigroup $X_{L,ML}^2$, take the following steps:

1. Specify a causal hypothesis for each group in the form of a DAG.
2. Choose the appropriate model for the regressions representing the paths in the DAG (i.e., not in the d-separation claims). Apply constraints to the parameters of interest—either by setting them to a predefined value, or by constraining them to be equal across groups—and calculate the sum of the log likelihoods of the variables.
3. Calculate the log likelihood of the saturated model.
4. Use the $X_{L,ML}^2$ test to test whether the hypothesized path model is significantly different from the saturated model. We refer to Shipley and Douma (2020) for details and to Appendix S1 for an multigroup example in R. Because a multigroup model is comparing constraints on the fixed effects, models should be fitted by maximum likelihood and not, in case of mixed-effects models, restricted maximum likelihood (Zuur et al. 2009).

MODEL SELECTION ON MULTIGROUP MODELS

After having fitted several models that are each judged to be consistent with the data based both on the above procedure and on pre-existing causal knowledge, one may wish to select the most parsimonious model. This can be done through model selection criteria such as Akaike's information criterion (AIC; Anderson 2008) or the likelihood ratio test that assesses the fit of nested models. Both take the likelihood of the joint probability distribution of the hypothesized path model as input (see *Rationale of multigroup X^2_{ML} test*). The AIC of the full model can be obtained by summing the AICs the piecewise structural equations. The model with lowest overall AIC is the most parsimonious model.

The C-statistic resulting from the d-sep test cannot be used for model selection in this multigroup context for two reasons. First, the number of degrees of freedom does not increase with increasing group constraints because the number of independence claims in the union basis set does not change when imposing group constraints. Second, depending on the topology, some paths whose multigroup structure is of interest may not be part of the union basis set. Hence, testing for group differences on these path coefficients will neither affect the degrees of freedom, nor (or only slightly) affect the C-statistic. This implies that the AIC method based on the C-statistic (Shipley 2013) cannot be used for model selection in this multigroup context.

PRACTICAL SUGGESTIONS FOR MODEL INFERENCE

When the hypotheses concerning the multigroup structure are specified a priori and are few in number, then the appropriate test is the change in the $X^2_{L,ML}$ statistic. For post hoc testing, adjustments to the significance level need to be made. Alternatively, one could use information metrics such as AIC or BIC statistics when conducting post hoc comparisons between competing multigroup models while insuring that the models identified by the AIC or BIC statistic are not rejected by the data. It may ease the multigroup analysis to increase the number of constraints in a progressive fashion (Bollen 1989, Grace 2006). First, assume the same causal

structure among groups. Next, successively impose constraints on the functional form of the regressions, the path coefficients, the intercepts, and the variances using a saturated model that is least constrained. See Appendix S1 for how to impose various constraints in R.

Multigroup comparisons should be performed on unstandardized values (Grace 2006) since, when the standard deviation of the values differ among groups, the standardized path coefficient will become different as well. When constraining a path coefficient to be similar across groups, our ability to detect significant differences between them will be affected by the number of observations in each group, like in ANOVA type of analyses.

If the data in the groups have some nesting or multilevel structure that does not, itself, define the group structure (e.g., males/females in sites), mixed-effects models can be applied. Multigroup SEM could also be used for model testing. Split the data in a training set and test set and check whether the test set is consistent with the model fitted on the training set.

CASE STUDY

Bieber et al. (2018) applied path analysis to study the effect of age on the timing and duration of hibernation of the Edible Dormouse (*Glis glis*). They developed a path model for males and females separately and showed that, both for males and for females, age affects the onset of hibernation differently along two different paths: (1) Increasing age advances hibernation onset, and (2) increasing age delays hibernation onset through increasing the probability of reproduction, and the probability of reproduction in turn delays hibernation onset.

Because reproduction was measured differently in males and females, the authors developed separate path models for each group. However, as reproduction was measured as a binary variable (yes/no) in both sexes, it is possible to test whether age impacts the reproduction, hibernation onset, and hibernation duration in a similar way for males and females. For this reason, we employed a multigroup path model with each group defined by sex. We tested a slightly modified path model compared with the one tested in the original paper. Specifically, we (1)

left out the variable individual quality since this was the maximum age that an animal reached and it is thus not independent from age itself, and (2) we removed the variable hibernation end, since the end of the hibernation is fully determined by the hibernation start and its duration. Finally, (3) we added a path from age to body mass (i.e., body mass increases with increasing age) to get a model that was consistent with the data.

To account for the additional dependencies caused by repeated measurements per individual, year-to-year variation, and variation caused by diet, we added all three variables as crossed random effects. Reproductive success was modeled with a binomial distribution, all other variables with a normal distribution. As the packages in R that allow crossed random effects (lme4, Bates et al. 2015, and glmmTMB, Brooks et al. 2017) do not include options to model heterogeneous residual variance among groups, it was assumed that the residual variation did not differ between groups. These data consisted of 289 observations belonging to 75 individuals consisting of 30 females and 45 males and sampled over three different diets and 11 yr. We refer to Data S1 for more details on the analysis.

The similarities between males and females were tested with the d-sep test and with the generalized $X^2_{L,ML}$ test. The list of independence claims is given in Table 2. For the d-sep test, we used restricted maximum-likelihood (REML) estimation to obtain the null probabilities of the independence claims in the d-sep test since

REML provides better estimates for the fixed effects. For the $X^2_{L,ML}$ test, we used maximum (i.e., not restricted)-likelihood estimation in order to compare the hypothesized model with the saturated equivalent since fixed effects in mixed-effects models should be compared by maximum likelihood (Zuur et al. 2009). We compared four models that varied in the level of constraints: (1) fixed path coefficients, but free intercepts of start of hibernation and hibernation duration across groups; (2) free path coefficients and free intercepts of start of hibernation and hibernation duration across groups; (3) fixed path coefficients and fixed intercepts of start of hibernation and hibernation duration across groups; and (4) as 3, but only the intercept start of hibernation constrained across groups.

The resulting C-statistic and $X^2_{L,ML}$ statistics of the four models are shown in Table 3, and the path coefficients of the most parsimonious model of this series are shown in Fig. 2. The chi-squared $X^2_{L,ML}$ test and the d-sep test agreed that the only model of this set that is not consistent with the data is the one in which all path coefficients and the intercepts of start of hibernation and hibernation duration are constrained to be equal across sexes. Two of the three remaining models were essentially equally parsimonious (ΔAIC 0.2) according to the AIC statistic: (1) model 2 (AIC = 8,381.5), in which the path coefficients are equal between sexes but the intercepts are free, and (2) model 4 (AIC = 8,381.7), in which the path coefficients are equal between sexes, as is the intercept of the start of hibernation. The BIC statistics gave a slightly stronger preference for model 4 (BIC = 8,488.1) compared with model 1 (BIC = 8,491.5). As the AIC and BIC are sample statistics, and thus may change with the sample at hand, we used bootstrapping to assess the uncertainty in model ranks. As repeated measurements were made on the same animals over time and to preserve the within animal correlation between variables, 500 bootstrapped datasets were constructed by randomly selecting 75 animals with replacement after which the four path models were fitted. Across the 500 bootstrapped datasets, model 4 was selected most often, although the difference with model 1 was smaller than 2 in 82% of the cases (20% with BIC). The model allowing free path coefficients between

Table 2. d-sep claims from the union basis set to test for consistency of the path model with the data (x , y , and z).

Claim number	d-sep claims	Relationship to test
1	Bodymass \perp Reproduction {Age}	Bodymass $\sim (a_1, a_2)Age + (x_1, x_2)Repro$
2	Bodymass \perp Hibstart {Age, Repro}	Bodymass $\sim (b_1, b_2)Age + (c_1, c_2)Repro + (y_1, y_2)Hibstart$
3	Hibduration \perp Repro {Age, Bodymass, Hibstart}	Hibduration $\sim (d_1, d_2)Age + (e_1, e_2)Bodymass + (f_1, f_2)Hibstart + (z_1, z_2)Repro$

Note: No constraints are put on the path coefficients in the specified d-sep claims.

Table 3. Model fit of four competing path models with different levels of group constraints

No.	Model description	No. d-sep claims	p_1	p_2	C_{overall}	p_{overall}	$X^2_{L,ML}$ (df)	$p(X^2_{L,ML})$	k	AIC	BIC
1	All path coefficients constrained across groups	2 * 3	0.52	0.24	13.22	0.35	10.78 (13)	0.63	30	8,381.5	8,491.5
2	No path coefficients constrained across groups	2 * 3	0.34	0.13	16.7	0.16	8.40 (6)	0.21	37	8,393.1	8,528.8
3	All path coefficients and intercepts start of hibernation and hibernation duration constrained across groups	2 * 3	0.22	8.13e-07	46.88	4.89e-6	54.56 (15)	2.12e-06	28	8,421.3	8,523.9
4	All path coefficients and intercept of start of hibernation constrained across groups	2 * 3	0.23	0.19	16.85	0.16	13.04 (14)	0.52	29	8,381.7	8,488.1

Notes: The group-wide C -statistic (C_{overall}), and the overall p -value (p_{overall}) and the p -values of each group are indicated (p_1 and p_2 , respectively). k is the number of parameters used in the structural equations. The $X^2_{L,ML}$ statistic results from twice the difference in log likelihood between the hypothesized model and the saturated model with a df difference in degrees of freedom between the hypothesized and the saturated model; $p(X^2_{L,ML})$ refers to the p -value associated with the $X^2_{L,ML}$ statistic. Model 1 and 4 were most parsimonious according to the AIC statistic, while model 4 was most parsimonious according to the BIC statistic.

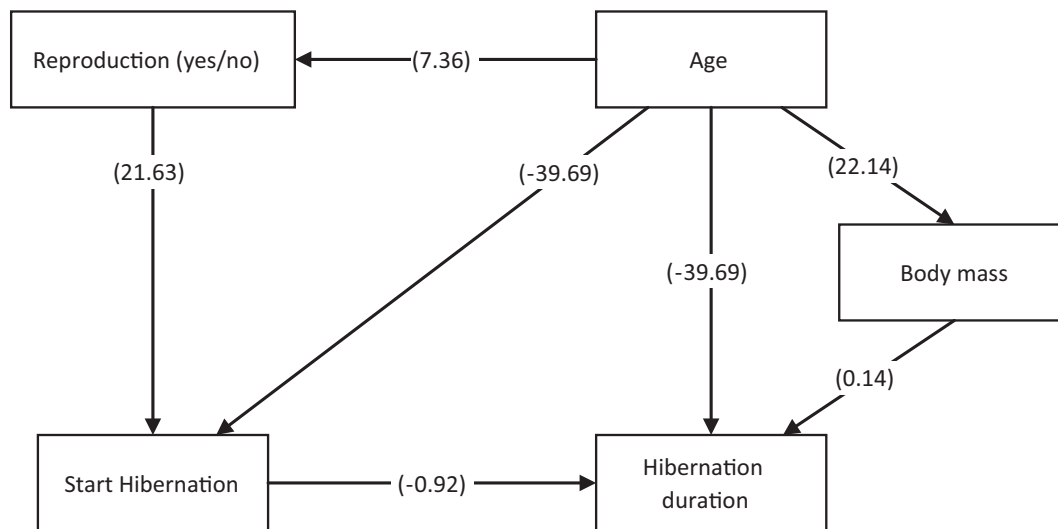


Fig. 2. Multigroup path model fitted to data presented in Bieber et al. (2018) relating age of an individual to its hibernation characteristics. A path model with path coefficients similar for males and females appeared to be consistent with data and most parsimonious according to AIC. See model statistics in Table 3. The variables in path model are: age of *Glis glis* (log-transformed), reproduction represents whether the animal is sexually active (binary yes/no), start of hibernation (Julian day), hibernation duration (days), and body mass prior to hibernation (grams).

sexes was in 75% (99% BIC) cases worse than >8 AIC points compared with model 4.

Importantly, the study was set out to test how sex affected the relationship of age on the chance of reproduction, hibernation onset, and the

hibernation duration, and with the multigroup analysis, we showed that these effects did not differ significantly between sexes. This is a conclusion that could not have been reached without an explicit multigroup analysis.

Even though this analysis was based on a rather large sample size (289 observations), we tested the robustness of this conclusion. We did so by calculating the power (i.e., the probability of rejecting the hypothesis of no difference between sexes at a level of 0.05) assuming that the estimated effect sizes of the three regressions are the true effect sizes. We simulated 500 times from the three fitted univariate models and tested how often \log_age significantly interacted with sex on the three aforementioned variables. The power of the estimated effect sizes was, respectively, 0.12, 0.12, and 0.08. To obtain a power of 80%, the effect of age on the three variables had to differ substantially between the sex varying from a 50% to a 200% difference between sexes. This suggests that the power of this study was too low to be able to reject the null hypothesis.

DISCUSSION

This paper presents two methods for multigroup analysis for piecewise path models that can be represented by directed acyclic graphs. The first method that we proposed is an extension of the d-sep method, while the second method is an application of the generalized $X_{L,ML}^2$ statistic of Shipley and Douma (2020). Having two methods that allows one to formally test for differences between path models of different groups without requiring the restricted assumptions of classical multigroup SEM is a big step forward, and the case study illustrates its utility. Until now, studies considering multiple groups in path models could either resort to classical SEM with the risk of violating assumptions, or use the d-sep test and test only two extremes: a separate path model for every group or one path model for all groups; examples are shown in Ogilvie et al. (2017), Barel et al. (2018), Theodorou et al. (2016), Vaz et al. (2019), Lefcheck et al. (2018), and Gow et al. (2019).

The advantages and disadvantages of the multigroup generalized $X_{L,ML}^2$ test compared with the multigroup d-sep test are similar to its advantages with respect to standard path models (Shipley and Douma 2020). However, one benefit of the generalized $X_{L,ML}^2$ test, that is, it simultaneously tests the topology and estimates the path coefficients, has particular advantages in the

multigroup context. First, as the possible number of models to test is large in the multigroup context, it makes the procedure less prone to error compared with the d-sep test where the constraints must be imposed both on the union basis set and on the piecewise structural equations. Second, not all paths in the multigroup d-sep test are necessarily part of the union basis set, and thus, constraints imposed on these paths are not reflected in the C-statistic. One should therefore carefully check which path constraints are not tested in the basis set, and these should be tested separately. Third, the degrees of freedom in the $X_{L,ML}^2$ test follows directly from the number of constraints. This is not the case in the d-sep test where the degrees of freedom is independent of the number of constraints applied, except when all parameters are assumed to be equal across groups.

Given the large number of constraints that one could impose on a multigroup path model in a post hoc analysis, one runs the risk of (over)fitting the path model to the data. How to select the optimal model is an active field of research (Lam and Bacchus 1994, Liu et al. 2012, Lubke and Campbell 2016, Lin et al. 2017, Lubke et al. 2017), and we refer to Grace (2020) for a more detailed discussion. Cross-validation, that is, the testing of the model to an independent dataset or to a subset of the data that was left out during the fitting, is needed and the preferred way to test the validity of a model (Preacher and Merkle 2012). When using AIC and BIC, one has to bear in mind that these are sample statistics, and thus, they may produce different model rankings when the same model would have been fitted to another data sample (Preacher and Merkle 2012, Preacher et al. 2013).

To test the robustness of our conclusion of no difference between sexes, we used two techniques. We used bootstrapping to calculate the variability in AIC and BIC statistics (see Preacher and Merkle 2012 for a detailed discussion), which confirmed that model 1 and model 4 do not differ greatly and thus cannot be distinguished from each other, while the model with no constraints on sex was almost never selected as best model. To further strengthen our conclusions of no difference between sexes, we performed a power analysis on the variables that were considered the core

of the study aim. This showed that, for the current sample size, the estimated power was quite low, and substantially larger effect sizes would have been needed to obtain a significant interaction between sex of age on the chance of reproduction, start of hibernation, or hibernation duration.

ACKNOWLEDGMENTS

This work was supported by NWO Earth and Life Sciences (NWO-ALW) through a VENI Grant (Project No. 863.14.018 to JCD). We thank an anonymous reviewer for constructive comments. JCD conceived the methodology and wrote the first draft. BS contributed to developing the methodology and the writing of the manuscript.

LITERATURE CITED

- Anderson, D. R. 2008. Model based inference in the life sciences. A primer on evidence. Springer-Verlag New York, New York, USA.
- Barel, J. M., T. W. Kuyper, W. de Boer, J. C. Douma, and G. B. De Deyn. 2018. Legacy effects of diversity in space and time driven by winter cover crop biomass and nitrogen concentration. *Journal of Applied Ecology* 55:299–310.
- Bates, D., M. Mächler, B. Bolker, and S. Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67:48.
- Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57:289–300.
- Bieber, C., C. Turbill, and T. Ruf. 2018. Effects of aging on timing of hibernation and reproduction. *Scientific Reports* 8:13881.
- Bollen, K. A. 1989. Structural equations with latent variables. John Wiley and Sons, New York, New York, USA.
- Brooks, M. E., et al. 2017. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal* 9:378–400.
- Gow, E. A., et al. 2019. Effects of spring migration distance on tree swallow reproductive success within and among flyways. *Frontiers in Ecology and Evolution* 7:380.
- Grace, J. B. 2006. Structural equation modeling and natural systems. Cambridge University Press, New York, New York, USA.
- Grace, J. B. 2020. A ‘Weight of Evidence’ approach to evaluating structural equation models. *One Ecosystem* 5:e50452.
- Kline, R. B. 2005. Principles and practice of structural equation modelling. Second edition. The Guilford Press, New York, New York, USA.
- Kline, R. B. 2012. Assumptions in structural equation modeling. Pages 111–125 in R. H. Hoyle, editor. *Handbook of structural equation modeling*. Guilford Press, New York, New York, USA.
- Lam, W., and F. Bacchus. 1994. Learning Bayesian belief networks: an approach based on the mdl principle. *Computational Intelligence* 10:269–293.
- Lefcheck, J. S., et al. 2018. Long-term nutrient reductions lead to the unprecedented recovery of a temperate coastal region. *Proceedings of the National Academy of Sciences of the United States of America* 115:3658–3662.
- Lin, L.-C., P.-H. Huang, and L.-J. Weng. 2017. Selecting path models in SEM: a comparison of model selection criteria. *Structural Equation Modeling: A Multidisciplinary Journal* 24:855–869.
- Liu, Z., B. Malone, and C. Yuan. 2012. Empirical evaluation of scoring functions for Bayesian network model selection. *BMC Bioinformatics* 13 (Suppl):15.
- Lubke, G. H., and I. Campbell. 2016. Inference based on the best-fitting model can contribute to the replication crisis: assessing model selection uncertainty using a bootstrap approach. *Structural Equation Modeling: A Multidisciplinary Journal* 23:479–490.
- Lubke, G. H., I. Campbell, D. McArtor, P. Miller, J. Luningham, and S. M. van den Berg. 2017. Assessing model selection uncertainty using a bootstrap approach: an update. *Structural Equation Modeling: A Multidisciplinary Journal* 24:230–245.
- Newsom, J. T. 2015. Longitudinal structural equation modeling - a comprehensive introduction. Taylor & Francis, New York, New York, USA.
- Oberski, D. 2014. lavaan.survey: an R Package for Complex Survey Analysis of Structural Equation Models. 2014 57:27.
- Ogilvie, J. E., S. R. Griffin, Z. J. Gezon, B. D. Inouye, N. Underwood, D. W. Inouye, and R. E. Irwin. 2017. Interannual bumble bee abundance is driven by indirect climate effects on floral resource phenology. *Ecology Letters* 20:1507–1515.
- Preacher, K. J., and E. C. Merkle. 2012. The problem of model selection uncertainty in structural equation modeling. *Psychological Methods* 17:1–14.
- Preacher, K. J., G. Zhang, C. Kim, and G. Mels. 2013. Choosing the optimal number of factors in exploratory factor analysis: a model selection perspective. *Multivariate Behavioral Research* 48:28–56.

- Shipley, B. 2000. A new inferential test for path models based on directed acyclic graphs. *Structural Equation Modeling: A Multidisciplinary Journal* 7:206–218.
- Shipley, B. 2009. Confirmatory path analysis in a generalized multilevel context. *Ecology* 90:363–368.
- Shipley, B. 2013. The AIC model selection method applied to path analytic models compared using a d-separation test. *Ecology* 94:560–564.
- Shipley, B. 2016. *Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference with R*. Cambridge University Press, Cambridge, UK.
- Shipley, B., and J. C. Douma. 2020. Generalized AIC and chi-squared statistics for path models consistent with directed acyclic graphs. *Ecology* 101: e02960.
- Theodorou, P., R. Radzevičiūtė, J. Settele, O. Schweiger, T. E. Murray, and R. J. Paxton. 2016. Pollination services enhanced with urbanization despite increasing pollinator parasitism. *Proceedings of the Royal Society B: Biological Sciences* 283:20160561.
- Vaz, P. G., M. N. Bugalho, J. M. Fedriani, M. Branco, X. Lecomte, C. Nogueira, and M. C. Caldeira. 2019. Unravelling associations between tree-seedling performance, herbivory, competition, and facilitation in high nature value farmlands. *Journal of Environmental Management* 232: 1066–1074.
- Wilks, S. S. 1938. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics* 9:60–62.
- Zuur, A. F., E. N. Ieno, N. Walker, A. A. Saveliev, and G. M. Smith. 2009. *Mixed effects models and extensions in ecology with R*. Springer-Verlag New York, New York, USA.

DATA AVAILABILITY

Code is available from Zenodo: <https://doi.org/10.5281/zenodo.4562002>

SUPPORTING INFORMATION

Additional Supporting Information may be found online at: <http://onlinelibrary.wiley.com/doi/10.1002/ecs2.3502/full>