Received: 1 November 2024 | Accepted: 9 September 2025

DOI: 10.1111/1365-2664.70189

BRITISH ECOLOGICAL SOCIETY Journal of Applied Ecology

REVIEW

The use, misuse and opportunities for structural equation modelling (SEM) in wildlife ecology

Calebe Pereira Mendes[1] | Matthew Scott Luskin[2,3]

[1]Asian School of the Environment, Nanyang Technological University, Singapore, Singapore

[2]School of the Environment, University of Queensland, Brisbane, Queensland, Australia

[3]Centre for Biodiversity and Conservation Science, University of Queensland, Brisbane, Queensland, Australia

Correspondence
Matthew Scott Luskin
Email: mattluskin@gmail.com

Funding information
Nanyang Technological University; Australian Research Council, Discovery Early Career Researcher Award, Grant/Award Number: #DE210101440

Handling Editor: Chris Sutherland

Abstract

1. Technological advances in passive detectors like wildlife cameras and bioacoustics have immense potential to help community ecologists understand species interactions and food webs. However, while passive sensors are relatively easy to operate, the statistical tools needed to study complex causal pathways that constitute food webs have a steep learning curve.
2. Here, we review analytical tools for assessing ecological interactions using observational datasets. We motivate the use of path analysis and structural equation modelling (SEM) by showing how they build on generalised linear mixed modelling and relate them to hierarchical models that account for detectability. Using a simulated dataset of wildlife observations from cameras, we compare the functionality and performance of the four dominant SEM packages available in R.
3. The top performer was piecewiseSEM, with paths fit by the glmmTMB package, and zero-inflation accounted for by mediator variables. The brms package was the second-best performer. However, no SEM package possesses all the desirable functionalities, and they vary in terms of estimated bias, precision (standard errors) and computational time.
4. We also compare the use of discrete versus continuous distributions for data derived from passive detectors, finding that the zero-inflated negative binomial distribution (ZINB) is the most flexible and least biased for SEMs using count data.
5. *Synthesis and applications:* The integration of SEM with observational wildlife datasets from passive detectors can lead to a step change in the scope, richness and robustness of insights about food webs, ecosystem functioning and conservation, but accounting for detectability remains a key hurdle.

KEYWORDS
camera, causal networks, data analysis, passive detectors, path analysis, trail cameras, wildlife ecology

1 | INTRODUCTION

Community ecologists and conservationists seek to understand the relationships among animal species (e.g. predator–prey and competition), plant–animal interactions (e.g. herbivory, pollination and seed dispersal) and indirect effects of wildlife on ecosystem processes (e.g. carbon sequestration; Luskin et al., 2017; Moore et al., 2023; Peres et al., 2016; Sills et al., 2009). Quantifying species interactions is useful but rarely possible due to the lack of suitable multispecies datasets and causal modelling approaches. For example, quantifying species interactions is necessary to study trophic cascades, a dominant theme in ecology and conservation that describes how the changes to one trophic level (e.g. large carnivores) and the associated species interactions (e.g. predation) trigger impacts on other trophic levels (e.g. herbivores, vegetation communities; Estes et al., 2011). Similarly, network effects have become an important component of conservation planning and management, such as determining the order in which to manage invasive species (e.g. eradicating invasive cats before invasive rabbits to limit cats prey-switching to native threatened small mammals; McGregor et al., 2020). Linear models with a single response variable are common in ecology but are often inadequate for investigating species interactions. Quantifying changes in complex ecological systems requires statistics suitable for species interaction networks, such as path analysis and structural equation modelling (SEM), here grouped together as SEM unless otherwise stated. SEM is an integrated multivariate statistical approach that imposes user-defined causal relationships and tests these using observed correlations (Depaoli, 2021). SEM is increasingly being applied in ecological research and wildlife ecology (Figure 1), but its use, misuse and opportunities remain unclear or underdeveloped.

Two hurdles to developing a rich understanding of species interactions and ecological systems are (i) obtaining suitable multispecies and multi-covariate datasets and (ii) applying appropriate analytic approaches that yield intuitive and statistically robust results for multiple nodes in the network (with nodes being species or a habitat covariate). Ideal datasets should include abundance measurements for multiple species at multiple sites and/or time points, as well as a range of values for key response variables and covariates (e.g. predator abundances, fragment sizes). New technologies have enabled advances in the passive sampling of multiple wildlife species in their natural environments, using detectors such as camera traps, acoustic monitoring, metabarcoding, eDNA and others (Russo et al., 2023). The use of camera traps for animal ecology, for instance, has grown exponentially, with millions of records available in online databases (Bruce et al., 2025; Kays & Wikelski, 2023). Similarly, new technologies such as Artificial Intelligence are being used to facilitate the identification of species from photographic and sound records (Norouzzadeh et al., 2018), which in turn allow researchers to increase sampling, and advances in genetic techniques allow eDNA species detections through water, air and other mediums (Bohmann & Lynggaard, 2023). These sampling approaches produce detection history datasets denoting where and when each species was detected.

Traditional statistical techniques for detection histories include linear modelling and hierarchical detectability modelling (HDM), which usually have a single species response variable and are thus not suitable for networks. Techniques such as SEM and Bayesian networks are becoming increasingly popular options for using detection histories to assess relationships among multiple species (Eisenhauer et al., 2015; Fan et al., 2016). SEM networks can be intuitively visualised using path diagrams with nodes for response variables (species) that can affect other response variables (other species), and external covariates. The hypothesised causal paths can be direct or indirect.

Integrating observational wildlife datasets with SEM can lead to a step change in the richness and quality of insights about species interactions, food webs and ecosystem functioning. Notable examples of SEM in wildlife ecology and conservation include its use to evaluate the direct and indirect relationships among vertebrate species and the environments they inhabit (Fan et al., 2016) and to quantify the relative importance of predation, competition and invasive species on native animals and plants (Carreira et al., 2020; Cunningham et al., 2020; Dorresteijn et al., 2015; Moore et al., 2022). However, considering the powerful inferences, promising examples, and broad applicability of using SEM with wildlife datasets, there has been relatively little penetration into the field of wildlife ecology (Figure 1b). This may be due to legitimate concerns about violating model assumptions, namely limitations in using the appropriate distributions, accounting for nested sampling designs (i.e. repeated sampling occasions at a subset of sites), variable sampling effort per detector or site (e.g. some cameras active for 3 months, others for 1 or 2 months) and variable species detectability (Amir, Sovie, & Luskin, 2022). These statistical issues are not clearly addressed in the existing wildlife ecology literature for SEM nor in the SEM packages or available codebase.

Here, we assess the advantages and disadvantages of using SEM in wildlife ecology, as well as the options for implementation in available R packages. In the first section, we provide a concise overview of key statistical characteristics and challenges of wildlife count datasets (also detailed in the Supporting Information). In the second section, we explore the limitations of existing modelling approaches for assessing relationships among multiple species, namely GLMMs and HDMs. In the third section, we explore the applications of SEMs in wildlife ecology. In the fourth section, we guide the readers through the main R packages currently available to implement SEMs and discuss their advantages and limitations. In the final section, we compare the performance of SEM R packages using simulated wildlife count data.

We tested the following hypotheses:

1. Discrete versus continuous distributions—Observational wildlife datasets are generated through a Poisson process, which creates the detection records and often includes zero-inflation and overdispersion. We predicted that accounting for the discrete nature of wildlife counts using Poisson or negative binomial distributions that reflect the raw counts will improve SEMs' performance over continuous Gaussian or 'Student t' distributions that transform the response variable into relative abundance indices (RAIs; e.g. Moore et al., 2022).
2. Sample sizes—More intensive sampling increases statistical power, or the proportion of true relationships that can be recovered, and yields more accurate effect size estimations (Jackson, 2003). For SEMs, we predicted a stronger influence of having multiple

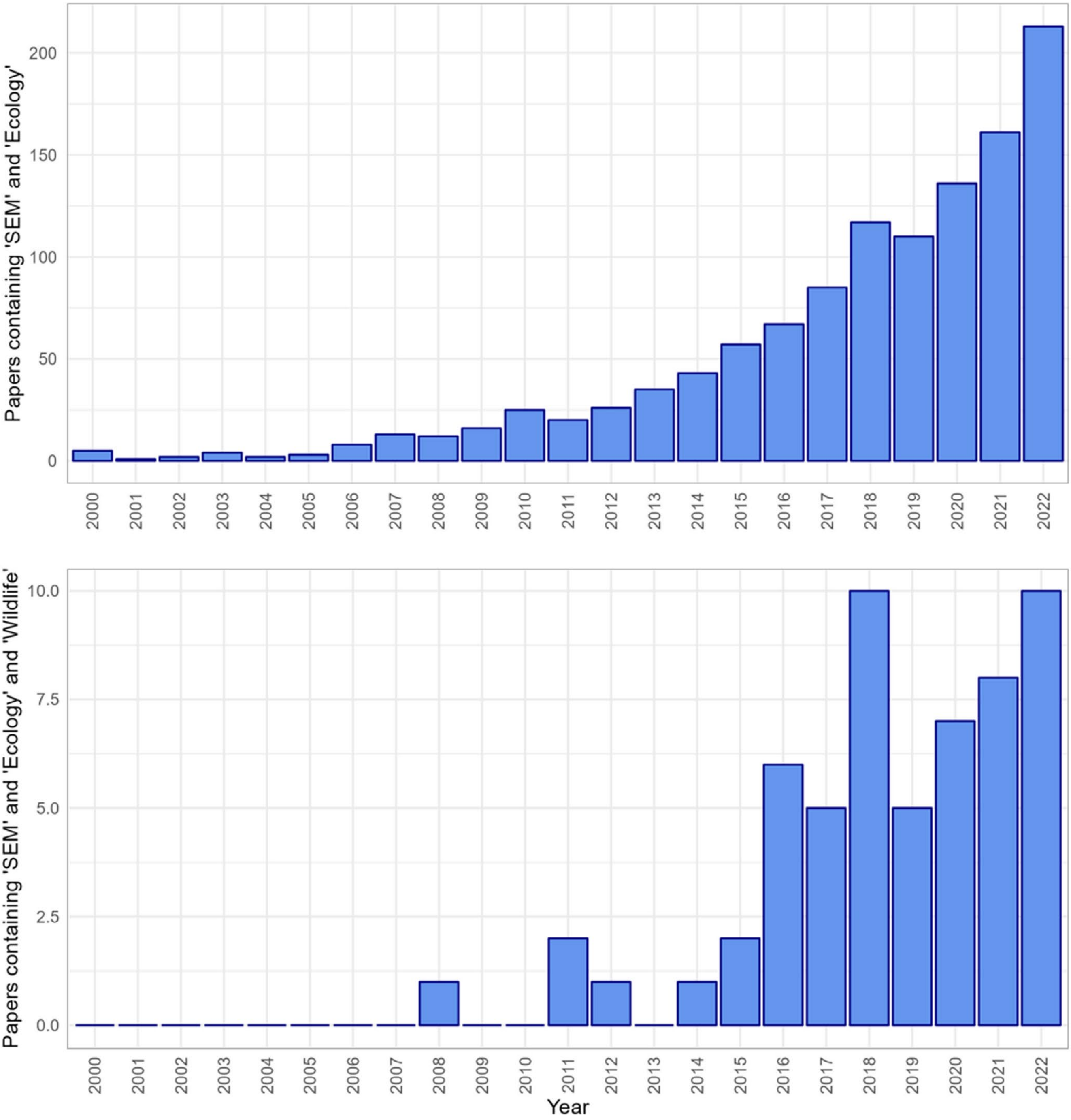

FIGURE 1 The number of published papers containing both the terms 'structural equation model' and 'ecology' (top) and with the terms 'structural equation model', 'ecology' and 'wildlife' (PubMed search using the 'rentrez' R package).

landscapes/communities—treated as the truly independent unit of observation and included as random effects—compared to increased sampling per survey (e.g. more detectors left active for longer) or multiple surveys from the same landscape.

3. Bayesian versus frequentist—Bayesian modelling allows for more flexibility in terms of model specification. We predicted that using a Bayesian framework for SEMs to better fit the underlying system would reduce bias relative to frequentist frameworks. We also predicted Bayesian SEMs would have higher statistical power (i.e. the ability to detect an effect) because of the 'head-start' provided when there are informative priors (Depaoli, 2021).
4. Global versus local estimation—SEMs based on flexible local estimation, where each path can be estimated using different distributions, should more accurately reflect the true nature of the data and wield higher statistical power than the relatively rigid global estimation (Lefcheck et al., 2023).

Understanding and using SEM for ecology first requires understanding the component models and consideration of wildlife datasets (i.e. count data, zero-inflation and the Poisson and NB distributions). Second, testing SEM package performance would ideally be done using the same model structure (i.e. same distributions, random effects, etc.), but this is not possible because functionality varies among packages. This is a key limitation of the state of the science and our manuscript. Therefore, in this paper, we compare different SEM packages with slightly different models. We do this with a simulated camera trap dataset (based on real data) and assess the performance of SEM packages and their unique model structures in terms of their ability to recover known paths.

2 | TYPES OF WILDLIFE DATASETS AND STATISTICAL CONSIDERATIONS

Our review applies to a range of sampling techniques that produce wildlife counts, including direct observations (bird point counts) and passive detections from camera traps, acoustic detectors, metabarcoding and eDNA. Most of these approaches are rarely able to uniquely identify individuals, so binary detection/non-detection data are considered independent if observations of the same species occur after a rest period, often 30 min. These independent detections are aggregated into temporal windows (e.g. 1 day or week). The counts of detections per detector (site) per temporal window reflect local activity and likely abundance. A survey includes many detectors (e.g. 5–500 units) set across a landscape (5–5000 km^2) with spacing ideally matching the target species' home range. We focus on camera traps as the most common example of passive sensors to sample wildlife in natural conditions and which have the largest datasets currently available (Antunes et al., 2022; Brodie et al., 2023; Bruce et al., 2025; Lima et al., 2017; Mendes et al., 2024).

While this paper will move quickly into the details of path analyses and SEMs, we note three important concepts that permeate all analyses. First, the structure of the wildlife dataset dictates which analytical approaches can be used, with richer datasets (i.e. detection histories) enabling hierarchical modelling that accounts for imperfect detection (i.e., HDMs; see Table 1 with further discussion provided in the Supporting Information; see Table 2 for a glossary). Second, most wildlife detection data are counts created by a Poisson process with zero-inflation, meaning statistics requiring normal distributions are less suitable unless detection probability is very high (Figure 2). Third, it is often necessary to include random effects when there is uneven or nested sampling, and account for variable effort per detector or per survey. Counts are often 'overdispersed', meaning the variance in the count values is larger than expected from a normal Poisson distribution (often due to animal grouping and routine movement patterns). Overdispersion can violate traditional model assumptions (the Supporting Information provide details and solutions about zero-inflation and overdispersion for wildlife applications).

SEM paths (relationships) are generally fit in a similar framework as linear or generalised linear mixed models (hereafter just 'GLMMs') and therefore should be able to leverage the versatility of GLMMs. Namely, (i) the ability to accommodate both continuous and discrete data, such as external covariate with a continuous normal distribution (e.g. elevation or a plant diversity metric) or categorical variable (e.g., habitat type) affecting response variables with Poisson distributions (e.g. wildlife counts); (ii) weighting observations by their sampling effort using offsets as opposed to using continuous

TABLE 1 Common types of observational wildlife survey data and their use to infer species interactions. Details of statistical options and issues are provided in the Supporting Information text.

Dataset structure	Description	Option to infer species interactions	Availability, pros and cons
Detection histories	Matrix depicting each species' detection or non-detection per detector per sampling window. These are created from spreadsheets describing the sampling design (metadata with detector's location and period of operation) and the list of observations (e.g. all images in a camera trap study and their timestamps), the observation contents (species and number of individuals)	Hierarchical models that account for detection, including co-occurrence models, co-abundance models, multi-species co-occurrence models. Also can be used in generalised linear and mixed models (GLMMs), SEMs, diel activity studies and species accumulation curves	The dataset structure that contains the most amount of information. Rarely publicly available
Capture summaries	The total sampling effort and number of detections of each species per detector or per survey, often derived from detection histories	GLMMs, SEMs, species distribution modelling (SDMs) and joint SDMs (jSDMs)	Widely available in literature and data papers. Cannot be used for hierarchical models that account for detection
Species lists	Species detected at a camera or survey, which can be extracted from a detection history or capture summary. Lacks information about the number of detections	SDMs, jSDMs, range mapping, presence–absence	The most available of the three dataset types (e.g. on GBIF). Contains the least amount of information. Cannot be converted into the other dataset types

TABLE 2 Glossary of wildlife statistics and SEM terminology. Detection histories, capture summaries and species lists are described in Table 1 and zero-inflation and overdispersion are described in Figure 2.

Term (and alternatives)	Description
Estimator	A set of rules or an algorithm used to estimate the parameters in a model. Different estimators may be susceptible or robust against different statistical issues (distributions), which may lead to different estimated parameter values for the same path diagram analysed with different R packages
Factor analysis	Statistical technique used to reduce data dimensionality, reducing many variables into fewer factors. It can be used to estimate unmeasured variables by reducing multiple measured variables (i.e. manifest variables) in a single factor (i.e. the latent variable). When used as part of an SEM, it is called a 'measurement model'
Formative SEM	A SEM with latent variables (unmeasured) predicted by indicator covariates (measured), where the latent variable is generally treated as a construct
Global and local estimation	Global is the process in which the SEM algorithm fit the whole set of SEM paths in a single step. Local is a process in which the SEM algorithm fits each causal path independently, then combining those results to assemble the full SEM in a later step
Indicator or manifest variable	A variable that is measured (e.g. animal weight, elevation, NDVI)
Latent or factor variable	A variable that is not or cannot be directly measured but can be inferred/estimated by other measured manifest covariates (a process called a factor analysis) due to its correlation with several manifest variables
Measurement model	A type of SEM composed by a one latent variable and its indicator variables
Offsets	Within a regression, it is an algebraic trick that leverages the log-link function to convert counts into rates by considering the effort. This approach is advantageous in that it retains the difference in information captured among detectors with variable effort
Reflective SEM	A SEM with latent variables which are assumed to be the underlying cause of the correlation between indicator variables
Structural equation model	A model where multiple explanatory variables affect multiple response variables via direct and indirect causal pathways. In some academic circles, the term is used exclusively to describe models or parts of a model describing the paths involving latent variables
Path diagram	A diagram describing the causal paths between variables (also called a structural model)
Survey	A set of detectors (sensors like cameras or human observers as in bird point counts) within a single study area (e.g. 1–1000 km^2 area) sampling over continuous period (e.g. 1 month)
Weights	Within regressions, and especially meta-regressions, weights are used to reflect differences in the quality of information among observations. For example, you might have less confidence in a particular HDM abundance estimate as its standard error increases

relative abundance indices or 'RAI' (i.e. catch per unit effort); the latter leads to the equal weighting of observations despite variable sampling effort; (iii) zero-inflation and overdispersion, discussed in detail below and in the Supporting Information; and (iv) random effects to account for nested sampling designs. The practical integration of these statistics into SEMs is summarised below.

3 | STRUCTURAL EQUATION MODELLING (SEM)

SEMs range in complexity and begin with a simple causal network comprising as few as three observed variables (nodes). SEMs are referred to as a path analysis when there are no feedback loops or reciprocal relationships. More complex SEMs can include over 10 nodes and multiple unmeasured variables. Unlike classical correlative approaches (e.g. GLMMs, HDMs), SEMs propose and evaluate evidence for a causal relationship between one or more response variables, with explicit consideration of explanatory covariates (Hoyle, 2012). SEMs can be visualised using directed acyclic graphs (DAG) or 'path diagrams' with arrows showing the direction of the relationship, providing intuitive interpretations (Figure 3). Ideally, a DAG should depict the existing knowledge of causal relationships between variables and constructs of a study system, including effects derived from the data-generating process (Kunicki et al., 2023). For more information on DAG design, interpretation and its uses with SEMs, see (Kunicki et al., 2023; Pearl, 2019; Rohrer, 2018). SEMs offer considerable flexibility when it comes to the structure of causal networks being analysed, with direct and indirect effects among measured 'manifest' variables, unmeasured 'latent' variables and nonlinear relationships (Depaoli, 2021; Hoyle, 2012). The primary types of SEMs include path analysis (no latent variables), factor analysis (a single latent variable estimated through its correlation with multiple indicator variables) and structural regression (one or more latent variables as predictors of another latent or manifest variable). The presence of latent variables as predictors is technically what differentiates SEM from path analysis. However, we follow Grace et al. (2012) in using SEM as an umbrella term for path, factor and structural regression analyses.

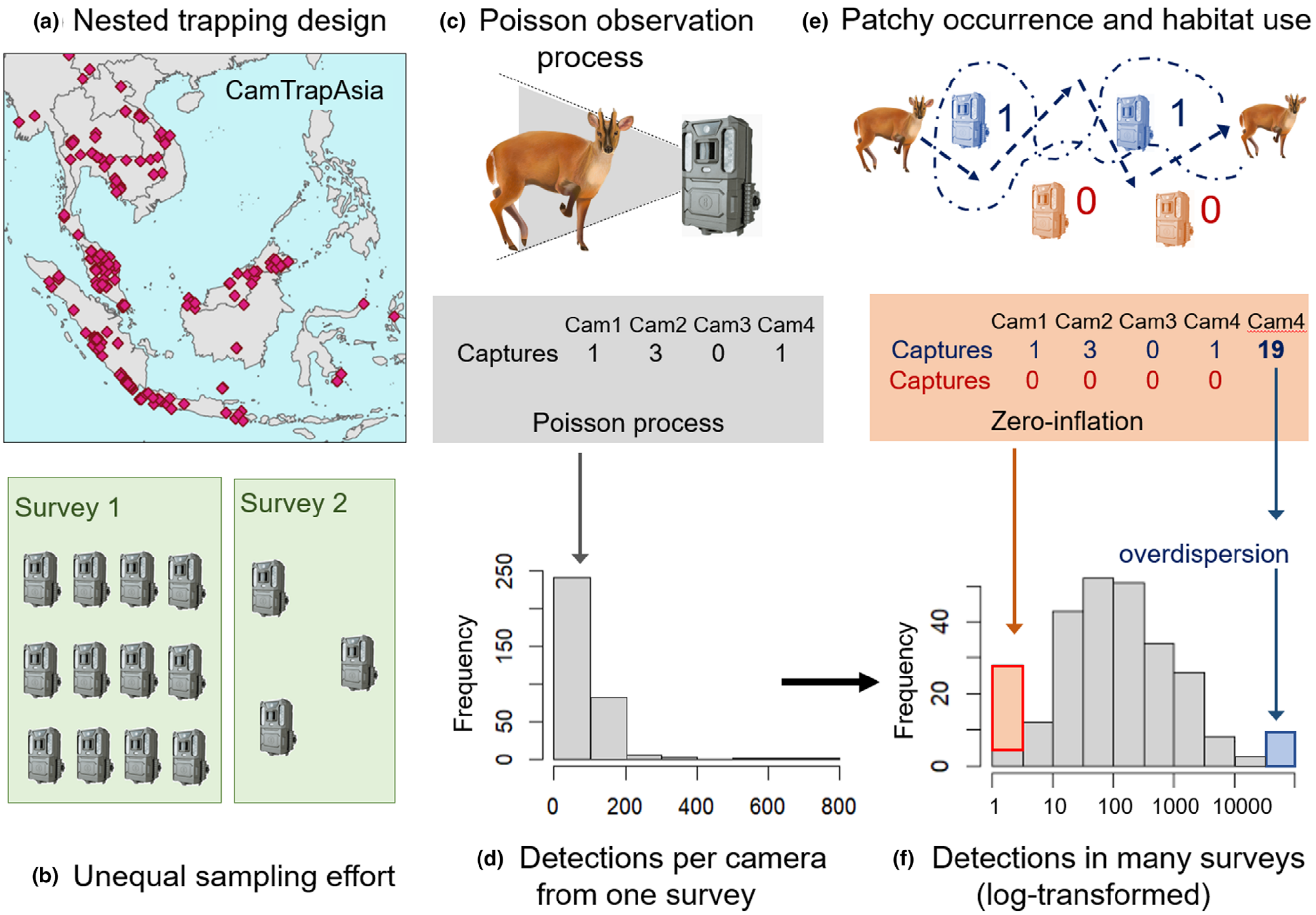

FIGURE 2 Statistical complexities that are common in wildlife survey datasets. (a) Nested sampling designs are common, such as when there are multiple surveys of the same area (e.g. a famous national park). (b) Different objectives or resources lead to variable sampling effort. This also happens within a single survey when a detector fails before the survey is finished. (c, d) Capture data are generated by a Poisson process and, therefore, not normally distributed. (e, f) Patchy habitat use leads to zero-inflation and overdispersion. Zero-inflation refers to when there are excess zeros unaccounted for by the distribution used to describe it. In wildlife ecology, excess zeros are often caused by a process different from the process driving the non-zero values (e.g. the occasional non-detection at some detectors of an elusive species that is present versus excess zeros at all detectors for an locally extirpated species). Overdispersion refers to when the data contain excessive variance, exceeding that allowed under its assumed distribution. In wildlife ecology, for example, this may be caused by one detector unknowingly set by a animal's den or nest, leading to exceptionally high numbers of detections at a small subset of detectors.

SEMs are a powerful approach to analysing observational wildlife data because they can model complex causal structures inherent to most ecological networks. Namely, for our purposes, they can estimate asymmetric interactions among three or more species. This overcomes key limitations with other approaches to study species interactions, including GLMMs that are limited to a single species as the response variable, co-abundance HDMs that are limited to species pairs, multi-species co-occurrence HDMs that are limited to symmetric species interactions and co-occurrence and joint species distribution models (jSDMs) where the response variable (probability of occurrence or presence) is unsuitable for inferring interactions (Amir, Sovie, & Luskin, 2022; Blanchet et al., 2020; Tobler et al., 2019). Given the aspiration of developing a causal understandings of food webs (Grace, 2024), the structure of the SEM causal network must have solid theoretical foundations that describe the directionality and mechanisms underpinning the relationships tested (Depaoli, 2021).

3.1 | SEMs with detection histories

SEMs and HDMs could be employed together to address the limitations of each technique, in principle. For example, deriving detection-corrected occupancy or abundance in HDMs and then using these as inputs for an SEM to assess the relationships among species. An advantage of this approach is that HDMs produce continuous values as inputs for the SEMs while accounting for variable sampling. However, one complication is propagating errors to avoid overfitting in the SEMs. This requires an approach similar to a meta-regression that weights observations by their precision (often the inverse standard error). It is also unclear the stage at which explanatory variables should be included as covariates: (i) in the HDMs to produce more accurate abundance estimates to input into the SEMs, (ii) only in the SEMs to account for shared direct and indirect effects of the covariate on multiple species (to avoid conflation with species correlations) or (iii) both?

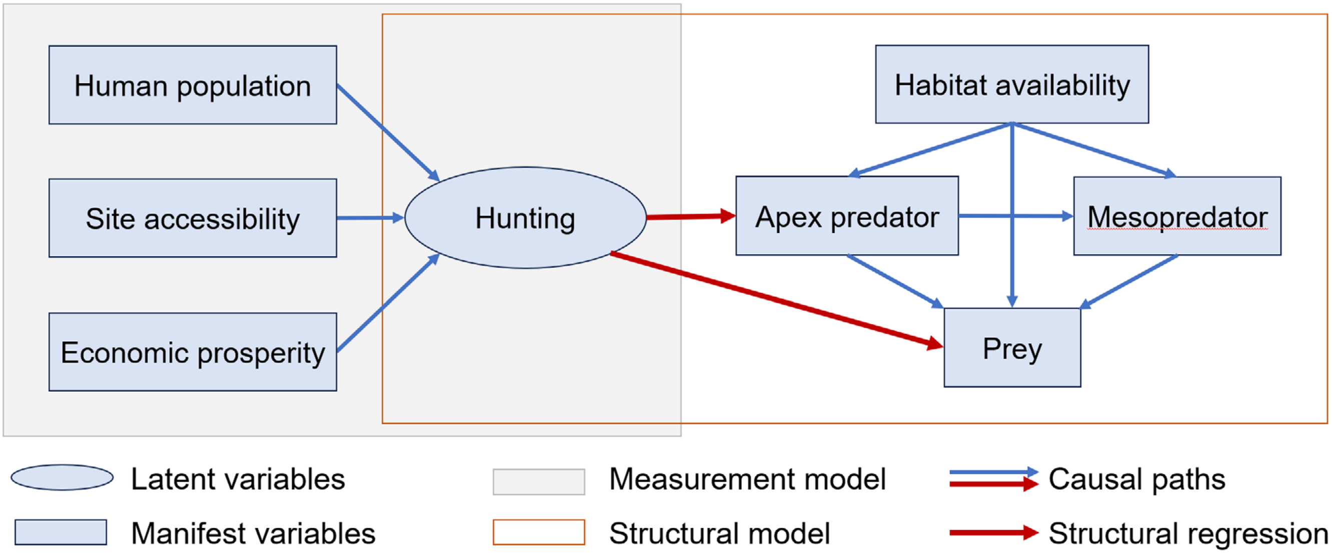

FIGURE 3 A path diagram showing an SEM with theorised causal relationships between wildlife functional groups as response variables (e.g. apex predators, mesopredators and herbivore prey), measured environmental covariates (e.g. habitat availability) and unmeasured latent variables (e.g. hunting). Here, hunting is inferred from three measured indicator variables (human population, site accessibility and economic prosperity). Note that causal paths between two nodes (e.g. apex predator and prey) may be direct (e.g. apex predator ➔ prey) or indirect (e.g. apex predators ➔ mesopredators ➔ prey). Causal paths are called structural regressions when a latent variable is used as a predictor for another node.

3.2 | SEMs with capture summaries

Capture summaries are the raw independent counts of a species per camera or per survey, or the capture rates (referred to as relative abundance index or 'RAI' in camera studies). Two counts of the same species at the same site are usually considered independent if they occur >1 hour apart). SEMs can be constructed from capture summaries of raw counts and one can control for differing sampling effort per sensor or per survey using offsets (available in a subset of R packages). As this approach does not consider species' detection probabilities, the SEMs assess variations in species activity rather than the true abundance, limiting population-level inferences (similar to GLMMs). However, SEMs using species capture rates can utilise large volumes of freely available data (see Antunes et al., 2022; Lima et al., 2017; Mendes et al., 2024). Including more surveys is useful because SEMs are data-hungry, and this demand scales exponentially with model complexity (Grace et al., 2012; Kline, 2023).

Options for constructing SEMs from capture rates (continuous RAI, as opposed to counts) face several issues. Namely, most SEM software options require the response variables to follow a multivariate normal distribution (i.e. normally distributed in all its dimensions). This is frequently violated in wildlife studies where count data often have zero-inflation and overdispersion (Depaoli, 2021; Figure 2). There are some partial workarounds discussed in the piecewiseSEM and glmmTMB literature (Brooks et al., 2017; Lefcheck et al., 2023). Appropriately accounting for zeros is especially important in an SEM framework for analysing multiple species simultaneously, since the non-detection of target species in sites where others are present provides valuable information (e.g. spatial avoidance of a predator or competitive exclusion). The simple use of zero-inflated variables, such as Poisson and negative binomial, may produce bias if endogenous variables predict the zero-inflated variable without incorporating the zero-inflation process. One possible solution is to account for zero-inflation as an element in the model structure using intermediate variables (Jiang et al., 2023). Another option is to use robust estimators that remain unbiased even with limited deviation from normality. For example, lavaan is a popular R package that offers SEM algorithms with 12 different estimators, including weighted least squares, diagonally weighted least squares and a series of maximum likelihood estimators with robust standard errors, which can handle limited zero-inflated and overdispersed (Gana & Broc, 2019). Finally, SEMs with hurdle distribution can also be used for handling zero-inflation in continuous variables.

Another possible source of bias when using RAIs is that observations are weighted equally despite differences in sampling effort. Likewise, packages that allow the use of weights for continuous data observations—such as lavaan—do not permit it to be used together with random effects to account for nested sampling (Rosseel et al., 2024). To attribute an adequate weighting for each observation according to the associated effort, one can use generalised SEMs with raw counts and discrete distributions such as Poisson or negative binomial, and account for sampling effort as an explanatory variable or as an offset (similar to GLMMs; Lefcheck et al., 2023). The use of offsets is only possible for count data; however, to our knowledge, Stata is the only user-friendly SEM software that includes offsets (StataCorp, 2023). An advantage of incorporating effort as a covariate is its relative ease of implementation. However, the model interpretation is more complex and may require counterfactual predictions, such as when path estimates are shown via a model prediction where effort is set to a standardised value (McElreath, 2020).

Note that not all user-friendly SEM software supports counterfactual predictions, as is the case with the popular PiecewiseSEM package (Table 2). Prediction using Bayesian packages, such as brms, is generally straightforward since the model produces a posterior distribution.

3.3 | Dealing with the unmeasurable (latent variables)

Community ecology and conservation studies often focus on phenomena that are challenging to measure directly. For example, hunting impacts wildlife but is difficult to measure directly due to its clandestine nature. Yet, hunting can be indirectly inferred through various indicators (i.e. proxies), such as accessibility via road networks and surrounding population density (Brodie et al., 2023). Similarly, indices of anthropogenic impact, like the Human Footprint Index and the Forest Landscape Integrity Index, integrate measurements of multiple human activities to represent an unmeasured vague construct (Grantham et al., 2020; Venter et al., 2016). Despite their utility, proxies often fail to accurately reflect the unmeasured variables and constructs they represent (Seltzer, 2021).

SEMs provide a more robust alternative to metrics for difficult-to-measure variables through factor analysis (multiple measured covariates informing an unmeasured variable), also referred to as the measurement model(s) portion of the SEM (Depaoli, 2021). The unmeasured phenomena (referred to as a 'latent variable') is estimated using the covariance between multiple measured variables (referred to as 'manifest variables'; Fan et al., 2016). The latent causal structure can be reflective (i.e. the manifest variables are caused by the latent variable) or formative (i.e. the manifest variables cause the latent variable). The classic SEM with factor analysis estimating latent variables is called covariance-based SEM (CB-SEM). It relies on the covariance between manifest variables to estimate the latent variable, assuming that it is the underlying cause of the indicators (i.e. reflective causal structure). While CB-SEMs traditionally used a multivariate normal distribution, which limits their application in ecological studies where data rarely conform to normality, there are now robust estimators and generalised algorithms that have expanded their applicability to include count data, categorical variables and ordinal variables as manifest variables (Boker et al., 2011; Depaoli, 2021; Rosseel, 2012).

Alternatively, the variance-based SEM (VB-SEM) estimates paths based on the variance rather than the covariance of the variables. Among VB-SEMs, the Partial Least Squares SEM (PLS-SEM) is the most commonly used approach (Hair, Hult, Ringle, & Sarstedt, 2021). The PLS-SEM does not assume any specific data distribution, making it particularly advantageous for analysing biological data, including capture summaries (Hair, Hult, Ringle, & Sarstedt, 2021). PLS-SEMs also offer flexibility in the causal relationship between latent and manifest variables, allowing for both reflective and formative causal structures, a choice that must be guided by the theoretical framework that describes the system (Hair, Hult, Ringle, & Sarstedt, 2021). PLS-SEMs are also computationally efficient and require less data to fit compared to CB-SEMs (Hair, Hult, Ringle, & Sarstedt, 2021). However, a drawback of PLS-SEM is that it has an inherent bias, even though this is sometimes considered within acceptable limits (Dijkstra & Henseler, 2015; Hair, Hult, Ringle, & Sarstedt, 2021).

4 | SEM SOFTWARE AVAILABILITY AND LIMITATIONS

This section covers the features, limitations and the main R packages available for SEM (Table 3). We focus on packages that offer user-friendly functions for fitting SEMs, such as the widely used lavaan and piecewiseSEM packages (Lefcheck et al., 2023; Rosseel, 2012) and exclude software and R packages that require the user to specify the mathematical formulations, such as Rstan (Stan Development Team, 2023) and Jags (Plummer et al., 2023). Proprietary software, such as LISREL (Cziráky, 2004), Mplus (Muthén & Muthén, 2007) and Stata (StataCorp, 2023) will not be covered (see Byrne, 2012; Sakaria et al., 2023), nor will open libraries in other programming languages beyond R, like the semopy2 library for Python (Meshcheryakov et al., 2021). The list of features presented here is not exhaustive, and most of these packages are continually being developed, with new features being added regularly. Some features may not have been explicitly mentioned or may have gone unnoticed due to variations in SEM terminology used in SEM across different packages.

4.1 | Lavaan and blavaan

The lavaan package is a popular option for SEMs in R because of its straightforward syntax and efficient algorithm (Rosseel, 2012). Lavaan primarily fits a classic single covariance matrix SEM algorithm (CB-SEMs) and allows for both reflective and formative causal structures for latent variables. This is known as 'global estimation' and assumes that a common data-generating process for the covariates and that they share a common distribution (e.g. multivariate normality; Hoyle, 2012). However, lavaan offers various estimators that can handle limited deviations from normality, such as robust maximum likelihood methods, weighted least squares methods and bootstrap-based estimators (Gana & Broc, 2019). Lavaan also allows for the inclusion of categorical and ordinal variables, as well as random effects (referred to as 'clusters' in lavaan's terminology). It should be noted that lavaan assumes all non-categorical variables to be continuous (Gana & Broc, 2019; Rosseel, 2012). However, given that a Poisson distribution with a high rate (λ) approximates a normal distribution (Forbes et al., 2010), as long as the count data mean is relatively high (>5), it can be utilised in lavaan with a robust estimator (Gana & Broc, 2019). Lavaan's key limitation for analysing observational wildlife datasets is that it does not support zero-inflated distributions. Another limitation is the lack of offsets, which leaves effort to be included as either a covariate or using RAI as the response variable. Notably, using RAI overweights low-effort observations

TABLE 3 Features and limitations of popular SEM R packages.

Feature	Lavaan	Blavaan	PiecewiseSEM	SEMinR	BRMS
SEM type	CB-SEM	Bayesian, CB-SEM	Local estimation	PLS-SEM & CB-SEM	Bayesian, local estimation
Handles count data	No+	No+	Yes	Not applicable	Yes
Offset term	No	No	No	No	Yes
Weights	Yes*	No	No	No	Yes
Zero-inflated distributions	No+	No+	Yes	Not applicable	Yes
Random effects	Yes*	Yes	Yes	No	Yes
Latent variables	Yes	Yes	No	Yes	Yes
Latent causal structure	Reflective and formative	Reflective and formative	Not applicable	Reflective and formative	Formative
Counterfactual predictions	Yes	Yes	No	Yes	Yes
Model selection	AIC, BIC	WAIC, LOO, BFA	AIC	AIC, BIC	WAIC LOO

Note: The algorithms used by the packages can be covariance-based (CB-SEM) or variance-based (VB-SEM). Local estimations mean that the algorithm estimates each path individually and combines the paths in a later step, in contrast to the more common global estimation approach, where all relationships are estimated synchronously. The latent causal structure can be reflective (i.e. the manifest variables are caused by the latent variable) or formative (i.e. the manifest variables cause the latent variable). Partially implemented features are marked with *, while absent features that are included with alternative methods are marked with a +. Lavaan does not allow the simultaneous use of weights and random effects (the latter termed clusters in those packages' terminology). AIC and BIC are Akaike and Bayesian information criteria, respectively; BFA is Bayes factor approximation, WAIC is widely applicable (or Watanabe–Akaike) information criterion, and LOO is leave-one-out cross-validation. Offsets in the piecewiseSEM are permitted by the package and its dependent regression package (glmmTBM), but the summary functions fail, an issue that may be solved in future versions.

(as described in detail for continuous GLMMs in the Supporting Information). Lavaan does allow for counterfactual predictions, enabling interpretations of SEMs with the effort covariate set to a constant value.

Blavaan is a Bayesian alternative to lavaan that accommodates smaller sample sizes due to the use of priors that constrain the parameter space during estimation, in conjunction with the Markov chain Monte Carlo algorithm (Depaoli, 2021; Merkle & Rosseel, 2018). Blavaan was designed to use the same syntax as lavaan, incorporating a few additional functions to handle Bayesian hyperparameters. Similar to lavaan, blavaan can be used to model error covariances, control for causal bias and perform sensitivity analyses. A drawback of blavaan is the increased computation time and the requirement for users to specify the priors, which can be a complex task within an SEM framework (Depaoli, 2021).

4.2 | PiecewiseSEM

Some variance-based SEM (VB-SEM) algorithms employ local estimation, wherein each path is estimated individually through linear regressions before combining them into an SEM (Lefcheck et al., 2023). Local estimation provides greater flexibility in the model specification than the classic SEM global estimation technique because the data used for each path can have a different distribution. This flexibility is particularly useful in ecological studies where distributions vary. For instance, in camera trap studies, the capture counts of a target species may follow a Poisson distribution, while a second endogenous variable, such as vegetation density or elevation, may follow a normal Gaussian distribution. Local estimation SEM allows Poisson and Gaussian regression paths to be fitted within the same model (Lefcheck et al., 2023).

The piecewiseSEM package is a popular choice for conducting local estimation SEM in R. Unlike other SEM packages, piecewiseSEM combines multiple separate regression paths that are fit using dependent packages into a single SEM, which is then evaluated using tests of directed separation (Lefcheck et al., 2023). Consequently, the coding syntax may vary depending on the regression package used to fit the paths. For example, paths with count data following Poisson distributions can be fit using the lme4 package (Bates et al., 2015), while zero-inflated and overdispersed count data paths can be fit using the glmmTMB package with a ZINB distribution (Brooks et al., 2017). Random effect variables can also be incorporated during the path regression fitting (Bates et al., 2015; Brooks et al., 2017). PiecewiseSEM is a suitable option for studies with small sample sizes; however, exceedingly small sample sizes should also be avoided, as the lack of statistical power can cause the d-separation test to fail, resulting in Type II errors (Lefcheck et al., 2023).

The main limitation of the piecewiseSEM package is that it is designed for path analysis and does not allow latent variables. PiecewiseSEM does not include functions for making predictions of the fitted models. It also does not supply the standardised coefficients from certain non-Gaussian distributions such as zero-inflated Poisson and zero-inflated negative binomial (but see (Grace et al., 2018) for binary distribution). The package appears able to handle offsets, yet the current version contains a coding error preventing the model summary from being extracted when offsets are included.

An important consideration is that each regression path must satisfy the statistical assumptions of the regression algorithm used to fit the path. Failure to meet these assumptions may lead to convergence problems or yield biased path estimates, which can impact the final SEM results and interpretation. Given these drawbacks, to create intuitive outputs from mixed distribution SEMs, one should first check the individual path regressions. If all the individual regression models seem reliable, the paths can be combined into an SEM using piecewiseSEM.

4.3 | SEMinR

The package SEMinR allows for fitting SEMs using either covariance-based algorithms or variance-based Partial Least Squares algorithms, with a focus on the latter (Hair, Hult, Ringle, Sarstedt, Danks, & Ray, 2021). For ecological studies, the package offers great flexibility since PLS-SEM algorithms have no premises regarding the data distribution. However, the package currently lacks the option for estimating random variables (i.e. multilevel mixed-model SEMs), which limits its applicability when there is nested sampling. Another issue that requires attention is the tendency of PLS-SEMs to underestimate structural model relationships (i.e. relationships between construct elements). In contrast, overestimating measurement model relationships (i.e. reflective or formative relationships between manifest variables and the constructs they represent) is a phenomenon sometimes referred to as the PLS-SEM bias (Hair, Hult, Ringle, Sarstedt, Danks, & Ray, 2021). For complex models that include multiple latent variables, the overestimated structural relationships and the underestimated measurement relationships can balance each other, resulting in overall accurate models (Hair, Hult, Ringle, Sarstedt, Danks, & Ray, 2021). However, ecological models in the current literature rely heavily on manifest variables, with few or no latent variables, and as a result, PLS-SEM may underestimate relationships.

4.4 | brms

The brms package fits multilevel regression models and can be utilised as a Bayesian alternative to the piecewiseSEM package (Bürkner, 2017). The inclusion of priors and local estimation allows brms to conduct SEMs with small sample sizes. The brms syntax also allows for more flexibility to specify priors than blavaan's syntax. The path regressions in brms can accommodate count data, zero-inflated distributions, random variables, offset terms and latent variables. Since brms was not specifically designed for SEM, it can also be used for a variety of regression techniques.

5 | ASSESSING DIFFERENT SEM PACKAGE PERFORMANCE USING A SIMULATED DATASET

5.1 | Data simulation

In this section, we compare the common options using a simulated dataset that is grounded in reality. The simulated dataset was based on camera trap sampling in Southeast Asian rainforest landscapes, where the conservation of numerous threatened wildlife species and food webs is impacted by humans (e.g. hunting) and non-forest land cover, such as agriculture and oil palm plantations (Carr et al., 2023; Decœur et al., 2023; Dehaudt et al., 2022; Dunn et al., 2022; Hendry et al., 2023; Honda et al., 2023; Lee et al., 2024; Luskin, Arnold, et al., 2023; Mendes et al., 2023; Nursamsi et al., 2023). Prior research suggests that the key covariate nodes in these food webs are forest cover, agriculture and humans, with wildlife grouped into coarse guilds (Amir, Moore, et al., 2022; Luskin et al., 2017). We choose to separate abundant native wild boars (*Sus scrofa*) from other herbivores due to their notable and strong food web impacts (Figure 4; Luskin et al., 2017, 2019; Luskin, Johnson, et al., 2021; Lamperty et al., 2023). This is also especially relevant for conservation given the outbreak of African Swine Fever that has driven the recent collapse of Asia's wild pig populations (Luskin, Meijaard, et al., 2021; Luskin, Moore, et al., 2023). Wild boars and other herbivores are also affected by land use, such as habitat patch size and configuration, and the presence of humans and anthropogenic food subsidies, particularly because they crop raid in oil palm plantations (Brearley et al., 2024; Ke & Luskin, 2019; Luskin et al., 2014, 2017; Moore et al., 2023). Land use can induce direct changes to wildlife populations (e.g. predators preferring habitat edges and herbivores avoiding edges) and ecosystem properties (e.g. more light and higher plant biomass in edges provides improved forage for herbivores), plus additive or synergistic indirect effects via species interactions (e.g. much higher predation and much lower herbivory in edges; Bogoni et al., 2022; Gardner et al., 2019; Gaynor et al., 2018; Mendes et al., 2016; Orrock & Danielson, 2005; Prugh et al., 2009).

Our simulated dataset is guided by the CamTrapAsia database for the region (Mendes et al., 2024), which contains 239 wildlife camera surveys across 80 landscapes. The simulated dataset contains interactions between three simulated guilds and species (i.e. 'wild boars', 'other herbivores' and 'predators') and three environmental covariates (i.e. 'forest cover', 'oil palm cover' and 'human density'). To simulate overdispersion, the species capture counts were simulated using a negative binomial distribution with more variation than allowed by a Poisson distribution. For each landscape, the simulated capture counts (detections) received a random intercept drawn from a common normal distribution (i.e. random effects). To keep the simulated coefficients within a realistic range, we fit three GLMMs with the CamTrapAsia database and used their coefficients as parameters for the simulation. The resulting structure of the simulated dataset contains four significant structural relationships (oil_palm and wild_boars, wild_boars and herbivores, human_activity and herbivores and human_activity and predators), with the true known coefficients for paths set to 0.5, 0.0005, −0.47 and −0.19, respectively. The simulated coefficients include a significant positive value, significant negative values and a significant near-zero value (Figure 4). Latent variables were not added to the SEM structure because it could not be fit to the proposed model structure in any of the four R packages assessed. To the best of our knowledge, the packages brms and SEMinR cannot use latent variables as predictor variables, while lavaan allows it but struggles with model convergence during our simulations, possibly due to zero-inflation and overdispersion. Forest cover was not significant in the preliminary GLMMs and thus was simulated as random values drawn from a normal distribution. Zero-inflation was simulated by randomly selecting a portion of the dataset (0%, 30% and 50%) and setting the simulated capture counts to zeros. The zero-inflation was added independently for each variable; that is, the zero-inflation of one variable does not affect the zero-inflation of the other variables. The code used to generate the simulated dataset, along with further information about the simulated parameters, is available in the Supporting Information (Code S1). The ecological implications of the SEMs will not be discussed here since they are not relevant to the simulation experiment. Our goal was to compare the rates at which each SEM package was able to recover true known relationships, as well as the fidelity of the path estimate values and variance.

5.2 | SEM fitting

The simulated dataset was analysed using four R packages: piecewiseSEM, lavaan, SEMinR and brms. We ran SEMs with both discrete count data and continuous RAIs. The details of all approaches tested are listed in Table 3 (see also Supporting Information Code S1). Both real and simulated captures did not have equal mean and variance, so we used the ZINB distribution instead of the ZIP when testing the packages brms, piecewiseSEM and SEMinR. Lavaan does not allow for ZIP and ZINB distributions. Therefore, we attempt to address the zero-inflation for continuous distributions in the lavaan_RAI analysis using 'MLM' maximum likelihood estimation with robust standard errors and a Satorra–Bentler scaled test statistic. The Student's t-distribution was used in the brms analysis with RAI data, but not for piecewiseSEM because the models had convergence problems. For lavaan approaches with count data, we logged the counts to approach normality and used the robust estimator 'MLM', as lavaan treats all numeric values as continuous data (Rosseel, 2012). For all RAI tests, we used the transformed log(RAI+1) to approach normality while retaining zeros. We included landscape as a random effect for all approaches with the exception of SEMinR, which did not support it. The dataset simulations and SEM analyses were repeated 1000 times, except for brms, which was repeated 200 times due to its long computational time. The analysis was performed using the R statistical environment version 4.4.2 (R Core Team, 2024).

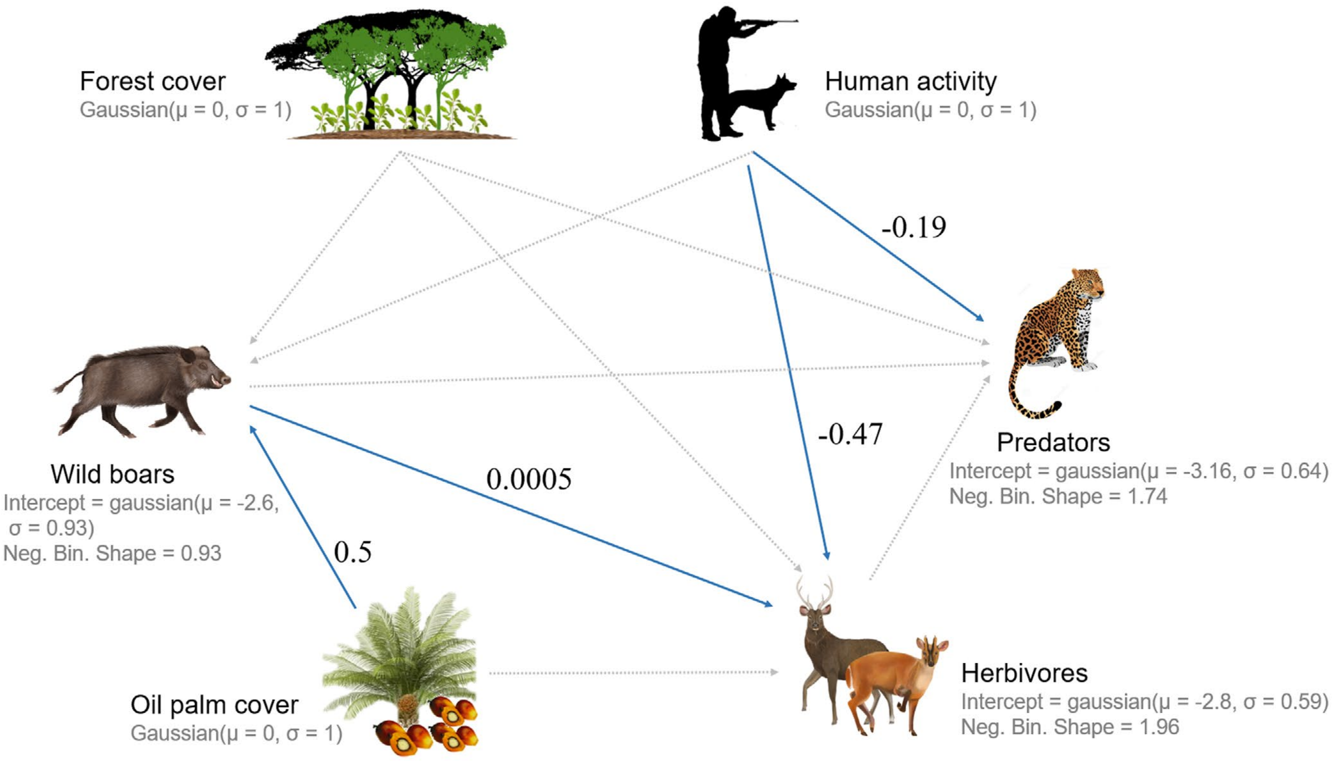

FIGURE 4 Path diagram of the causal structure of the test dataset, which was simulated using evidence from real camera trapping in Asia. The path coefficient between wild boars and herbivores has a small value relative to other paths because both wild boars and herbivores are measured as counts, while oil palm cover and human activities have a mean of 0 and a standard deviation of 1. This happens because it is not possible to scale inputs prior to SEMs without violating their distributions (e.g. continuous scaled distributions were inadequate to account for zero-inflation and overdispersion). Details about the study system are provided in Luskin et al. (2017).

5.3 | Assessing performance

We sought SEM approaches that could recover known coefficients with little bias and high precision (i.e. small standard errors). We ranked the performance of the six SEM approaches based on their statistical power, measured by whether the 95% confidence interval of the estimate included the known path value and the percentage of simulated paths that were successfully recovered. We assessed the accuracy of the estimates by the deviation of estimated coefficients from the known true value and assessed precision based on standard errors. For piecewise SEM and lavaan, we also extracted a statistical significance metric for the whole model. We repeated the analyses, varying zero-inflation from 0%, 30% and 50%, while keeping all other parameters constant. We also measured the effect of sample size by repeating the simulations for databases with 200, 400, 600, 800, 1000 and 1200 surveys. Finally, we measured the effect of random variables on the SEM approaches by repeating the simulations with datasets containing 10, 20, 40, 80 and 120 landscapes. The deviations from the true coefficient produced by the SEM approaches along the 1000 repetitions were compared using the Kolmogorov–Smirnov test (KS test) and the Anderson–Darling k-sample test (AD test).

6 | RESULTS: SEM PERFORMANCE ON SIMULATED ECOLOGICAL DATA

The three accurate approaches used local estimation; piecewiseSEM_ZI.med_counts, piecewiseSEM_ZI_counts and piecewiseSEM_counts produced similar path coefficients (AD test comparing the three approaches; AD = 1.47, p-value = 0.623) with an average deviation from the true coefficients of 0.146 (SD = 0.217), 0.150 (SD = 0.222) and 0.152 (SD = 0.223), respectively (Figure 5a, Table 3; Figures S2 and S3). Despite losing some accuracy in the presence of zero-inflation, piecewiseSEM_ZI.med_counts and piecewiseSEM_counts remain the most accurate approaches with 30% zero-inflation, while piecewiseSEM_ZI_counts's accuracy is severely impacted. With 50% zero-inflation, piecewiseSEM_ZI_counts fails to converge, while the first two approaches still score among the best options. The approaches brms_RAI, brms_ZI.med_RAI, piecewiseSEM_RAI and piecewiseSEM_ZI.med_RAI ranked 4th–7th in accuracy with an average deviation from the true coefficients of 0.159 (SD = 0.175), 0.164 (SD = 0.187), 0.167 (SD = 0.193) and 0.171 (SD = 0.284), respectively. The path estimates of these four approaches are statistically similar (AD test comparing the four approaches; AD = 2.77, *p*-value = 0.48) but differ from those of the three most accurate approaches (AD test comparing the seven approaches; AD = 240.8, *p*-value <0.001).

The approaches based on the packages lavaan and SEMinR performed poorly compared to piecewiseSEM and brms, with relatively high deviations from the true coefficient and low statistical power, especially in the presence of zero-inflation (Figure 5b). The exception is the approach lavaan_ZI.med_RAI, which has the second-best statistical power under 50% zero-inflation while keeping a deviation from the true coefficient of 0.247 (SD = 0.316), statistically similar to brms_ZI.med_RAI and piecewiseSEM_ZI.med_RAI (AD test comparing the three approaches; AD = 3.4, *p*-value = 0.1). In the absence of zero-inflation, lavaan_ZI.med_RAI ranks ninth in accuracy, with a deviation from the true coefficient of 0.184 (SD = 0.213). The SEMinR_counts was the least accurate approach in the absence of zero-inflation and also had the weakest statistical power.

Increasing the sample size (number of cameras and surveys) reduced the deviation from the true coefficient for most approaches (Figure 5c). For example, increasing the sample size from 200 to 1200 reduces the deviation from the true coefficient from 0.164 to 0.057 for the piecewiseSEM_ZI.med_counts approach (KS test; *D* = 0.34, *p*-value < 0.001). The only exceptions are the approaches based on the package SEMinR, which become slightly less accurate with increased sample size, and the approaches using the brms package with count data, which seem to vary randomly. Increasing the sample size also increases statistical power, as demonstrated by the brms_ZI approach.med_RAI, for example, recovering 42.5% of the simulated paths with 200 samples a 91% of the simulated paths with 1200 samples (Figure 5d).

Increasing the number of landscapes (i.e. levels of the random effect variable) resulted in small improvements in accuracy (Figure 5e) except laavan with count data (KS test comparing lavaan_counts approach with 10 and 120 landscapes: *D* = 0.03, *p*-value = 0.24; KS test comparing lavaan_counts approach with 10 and 120 landscapes: *D* = 0.04, *p*-value = 0.14). Approaches based on the brms package with count data appear to have an optimal number of levels to minimise deviation from the true coefficient (20 for brms_ZI_counts and brms_ZI.med_counts, 40 for brms_counts). Finally, approaches based on the piecewiseSEM package suffer from reductions in statistical power as the number of random variables increases. In contrast, approaches based on the lavaan and SEMinR packages exhibit an increase in statistical power.

7 | DISCUSSION

This review sought to explore how SEM can be used to infer species interactions in ways that overcome key limitations with alternative methods, namely that: (i) GLMMs that are limited to single species and do not account for detectability, (ii) hierarchical detection models (HDM) do account for detectability but are limited to single species, (iii) co-abundance HDMs that are limited to species pairs and (iv) multispecies co-occurrence HDMs are limited to symmetric species interactions and not suitable for tri-trophic questions or complex food webs. SEMs are suitable for inferring interactions among three or more species; however, when using raw data (counts or relative abundance indices), a key limitation of SEMs is that they do not account for detectability. SEMs may address the detectability issue using a two-step process, first deriving detection-corrected occupancy or abundance from HDMs and then using those as inputs for SEMs. However, that two-step approach must propagate error,

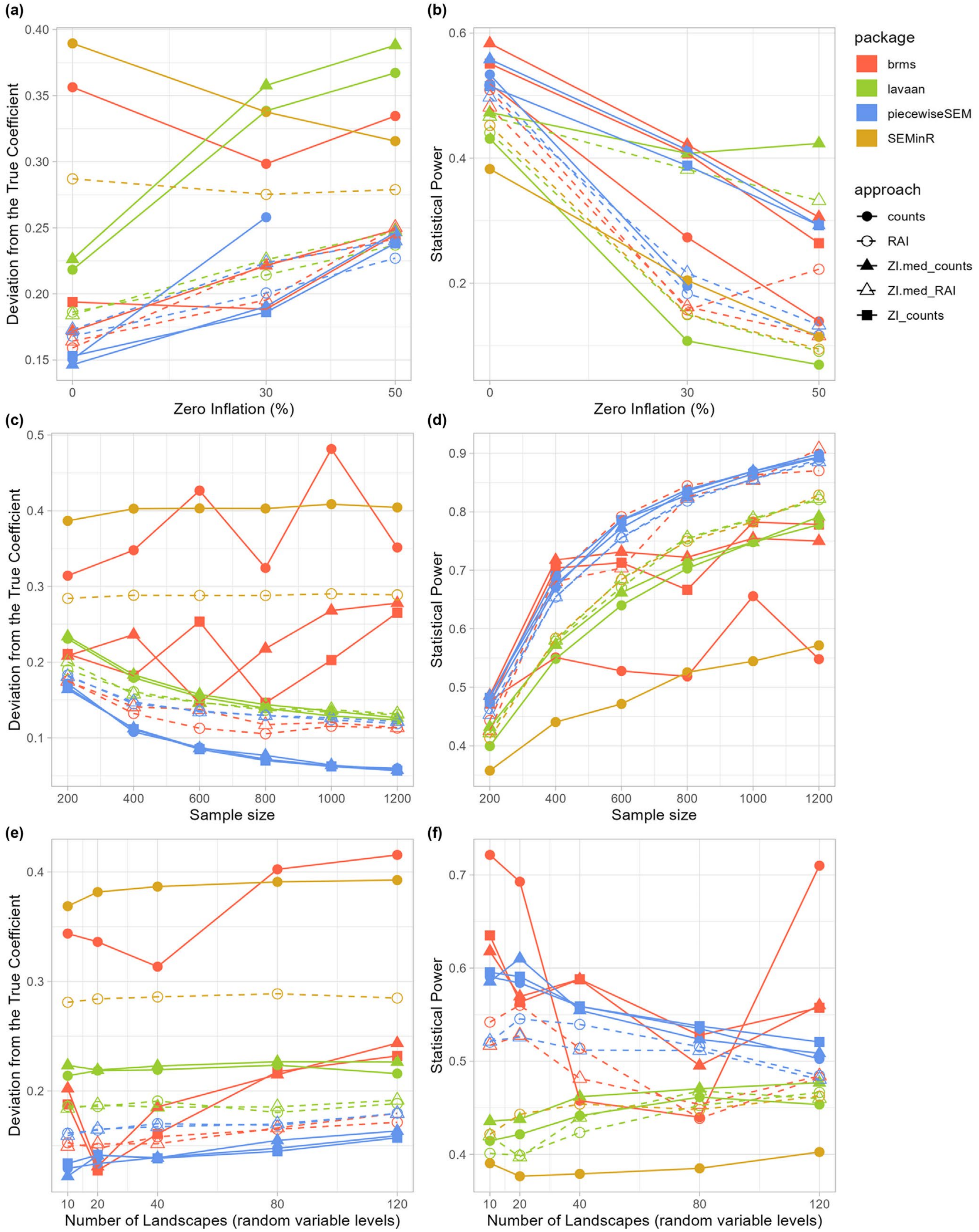

FIGURE 5 The ability of SEMs to recover true coefficient values in the simulated data containing varying amounts of (a) zero-inflation, (b) sampling effort or size and (c) landscape replicates (i.e. nestedness). Deviation from the true value is measured as the distance from zero. The statistical power of the SEM approaches is measured as the proportion of paths correctly recovered ($p < 0.05$) under varying degrees of (d) zero-inflation, (e) sample size (i.e. number of simulated camera surveys in the dataset, which can be replicates from the same landscape), and (f) landscape replicates.

which requires meta-analytic SEMs whose suitability has not been evaluated for ecological applications (Cheung, 2015). The piecewiseSEM can propagate error using weights, but this package does not allow for latent variables (and thus is limited to path analyses). It also remains unclear where and when covariates should be incorporated in a two-step process. While there is no silver bullet, some approaches vastly outperform others.

The review and simulated experiment demonstrate that SEMs vary in their functionality and performance, with one piecewiseSEM model formulation performing best for path analyses (no latent variables) with count data. Specifically, piecewiseSEM with paths fit using glmmTBM with a zero-inflated negative binomial (ZINB), mediator variables to account for zero-inflation and including effort as a covariate. However, the use of counts incurs the assumption of constant detectability or the need to interpret species interactions as effects on activity (rather than occupancy or abundance). We also note that piecewiseSEMs are evaluated using a 'test of directed separation' (TDS), which may result in a Type II error when sample sizes are small, and thus still require a relatively large dataset to produce reliable results that can be verified.

If, for some reason, RAI is preferable to count data in the given study, the package brms offers the best overall accuracy and statistical power, followed by approaches using piecewiseSEM. The lavaan package performed relatively poorly compared to piecewiseSEM and brms, possibly due to its global estimation algorithm. SEMinR performed poorly in our analysis, yet we are not convinced that its poor performance has any connection with the issue known as the 'PLS bias' (Hair, Hult, Ringle, Sarstedt, Danks, & Ray, 2021; Rönkkö & Evermann, 2013). Nevertheless, PLS bias is reported to be more common in models with simple structures and no latent variables (Hair, Hult, Ringle, & Sarstedt, 2021; Hair, Hult, Ringle, Sarstedt, Danks, & Ray, 2021). Despite the statistical hurdles to accommodate the nature of summarised camera trap data, the models used in the current manuscript have relatively simple structures for an SEM.

Increasing sample sizes enhances the statistical power of SEMs, although each causal path has different sample size requirements, both for local estimation (piecewiseSEM, brms) and global estimation (lavaan, SEMinR). For example, while employing the approach piecewiseSEM_ZI.med_counts with 240 observations, the path for wild boar ~ herbivores was successfully recovered in 98.9% of the 1000 repetitions, while the path for human ~ carnivores was only recovered in 19.2% of the repetitions. At 1200 observations, the path for human ~ carnivores was recovered in 58% of the repetitions, despite all other paths being recovered with over 99% success. This suggests that the 'N:q rule of thumb' of 20 samples per free parameter (Jackson, 2003) may be inadequate for detecting all causal paths reliably. Especially for complex ecological systems, as SEM's minimum sample size requirements are influenced by many factors, including the statistical distribution(s) used, the amount of noise in response and predictor variables, model complexity and strength of the causal paths under study (Fan et al., 2016; Grace et al., 2012; Kline, 2023). Power analyses to estimate the minimal sample sizes are highly recommended.

In summary, while no single R package encompasses all the SEM tools ecologists desire, piecewiseSEM and brms stand out by offering a flexible set of tools for fitting path analyses and SEMs with latent variables, respectively. Addressing statistical complexities like zero-inflation, overdispersion and nestedness is possible in SEMs. However, the specific method depends on the R package and influences how the data should be incorporated (as counts or capture rates). Although there is a need for further development of SEM tools for ecological studies, the primary obstacles hindering the widespread use of SEM in ecology are accounting for detectability and the availability of large datasets. Therefore, we advocate for enhanced collaboration within the scientific community to promote data sharing, thereby improving data accessibility and advancing the progress of ecological research (Bruce et al., 2025).

AUTHOR CONTRIBUTIONS

Calebe Pereira Mendes and Matthew Scott Luskin conceived the ideas and designed the methodology. Calebe Pereira Mendes analysed the data. Both authors wrote the manuscript and gave approval for publication.

ACKNOWLEDGEMENTS

We thank Jarrett Byrnes, Jun Ying Lim, Mike Chueng, the Ecological Cascades Lab at UQ and three anonymous reviewers for their helpful comments on the manuscript.

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest.

DATA AVAILABILITY STATEMENT

All the data are archived in Zenodo https://doi.org/10.5281/zenodo.17122491 (Mendes & Luskin, 2025), https://doi.org/10.5281/zenodo.10780971 (Mendes & Luskin, 2024) and GitHub https://github.com/CalebePMendes/SEM_methods_manuscript.git.

ORCID

Calebe Pereira Mendes https://orcid.org/0000-0003-1323-3287
Matthew Scott Luskin https://orcid.org/0000-0002-5236-7096

REFERENCES

Amir, Z., Moore, J. H., Negret, P. J., & Luskin, M. S. (2022). Megafauna extinctions produce idiosyncratic Anthropocene assemblages. *Science Advances*, *8*, eabq2307.

Amir, Z., Sovie, A., & Luskin, M. S. (2022). Inferring predator–prey interactions from camera traps: A Bayesian co-abundance modeling approach. *Ecology and Evolution*, *12*, e9627.

Antunes, A. C., Montanarin, A., Gräbin, D. M., dos Santos Monteiro, E. C., de Pinho, F. F., Alvarenga, G. C., Ahumada, J., Wallace, R. B., Ramalho, E. E., Barnett, A. P. A., Bager, A., Lopes, A. M. C., Keuroghlian, A., Giroux, A., Herrera, A. M., de Almeida Correa, A. P., Meiga, A. Y., de Almeida Jácomo, A. T., de Barros Barban, A., … Ribeiro, M. C. (2022). AMAZONIA CAMTRAP: A data set of mammal, bird, and reptile species recorded with camera traps in the Amazon forest. *Ecology*, *103*, e3738.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48.

Blanchet, F. G., Cazelles, K., & Gravel, D. (2020). Co-occurrence is not evidence of ecological interactions. *Ecology Letters*, *23*, 1050–1063.

Bogoni, J. A., Ferraz, K. M. P. M. B., & Peres, C. A. (2022). Continental-scale local extinctions in mammal assemblages are synergistically induced by habitat loss and hunting pressure. *Biological Conservation*, *272*, 109635.

Bohmann, K., & Lynggaard, C. (2023). Transforming terrestrial biodiversity surveys using airborne eDNA. *Trends in Ecology & Evolution*, *38*, 119–121.

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., Spies, J., Estabrook, R., Kenny, S., Bates, T., Mehta, P., & Fox, J. (2011). OpenMx: An open source extended structural equation modeling framework. *Psychometrika*, *76*, 306–317.

Brearley, F. Q., Song, H., Tripathi, B. M., Dong, K., Zin, N. M., Abdul Rachman, A. R., Ickes, K., Adams, J. M., & Luskin, M. S. (2024). Wild pigs mediate far-reaching agricultural impacts on tropical forest soil microbial communities. *Forest Ecology and Management*, *572*, 122320.

Brodie, J. F., Mohd-Azlan, J., Chen, C., Wearn, O. R., Deith, M. C. M., Ball, J. G. C., Slade, E. M., Burslem, D. F. R. P., Teoh, S. W., Williams, P. J., Nguyen, A., Moore, J. H., Goetz, S. J., Burns, P., Jantz, P., Hakkenberg, C. R., Kaszta, Z. M., Cushman, S., Coomes, D., … Luskin, M. S. (2023). Landscape-scale benefits of protected areas for tropical biodiversity. *Nature*, *620*, 807–812.

Brooks, M., Kristensen, K., van Benthem, K., Magnusson, A., Berg, C., Nielsen, A., Skaug, H., Maechler, M., & Bolker, B. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*, 378–400.

Bruce, T., Amir, Z., Allen, B. L., Alting, B. F., Amos, M., Augusteyn, J., Ballard, G., Behrendorff, L. M., Bell, K., Bengsen, A. J., Bennett, A., Benshemesh, J. S., Bentley, J., Blackmore, C. J., Boscarino-Gaetano, R., Bourke, L. A., Brewster, R., Brook, B. W., Broughton, C., … Luskin, M. S. (2025). Large-scale and long-term wildlife research and monitoring using camera traps: a continental synthesis. *Biological Reviews*, 100, 530–555. Portico. https://doi.org/10.1111/brv.13152

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*, 1–28.

Byrne, B. M. (2012). Choosing SEM computer software: Snapshots of LISREL, EQS, AMOS, and Mplus (pp. 307–324).

Carr, E., Amir, Z., Mendes, C. P., Moore, J. H., Nursamsi, I., & Luskin, M. S. (2023). The highs and lows of serow (*Capricornis sumatraensis*): Multi-scale habitat associations inform large mammal conservation strategies in the face of synergistic threats of deforestation, hunting, and climate change. *Raffles Bulletin of Zoology*, *71*, 400–416.

Carreira, D. C., Brodie, J. F., Mendes, C. P., Ferraz, K. M. P. M. B., & Galetti, M. (2020). A question of size and fear: Competition and predation risk perception among frugivores and predators. *Journal of Mammalogy*, *101*, 648–657.

Cheung, M. W.-L. (2015). metaSEM: An R package for meta-analysis using structural equation modeling. *Frontiers in Psychology*, *5*, 1521.

Cunningham, C. X., Johnson, C. N., & Jones, M. E. (2020). A native apex predator limits an invasive mesopredator and protects native prey: Tasmanian devils protecting bandicoots from cats. *Ecology Letters*, *23*, 711–721.

Cziráky, D. (2004). LISREL 8.54: A program for structural equation modelling with latent variables. *Journal of Applied Econometrics*, *19*, 135–141.

Decœur, H., Amir, Z., Mendes, C. P., Moore, J. H., & Luskin, M. S. (2023). Mid-sized felids threatened by habitat degradation in Southeast Asia. *Biological Conservation*, *283*, 110103.

Dehaudt, B., Amir, Z., Decoeur, H., Gibson, L., Mendes, C., Moore, J. H., Nursamsi, I., Sovie, A., & Luskin, M. S. (2022). Common palm civets *Paradoxurus hermaphroditus* are positively associated with humans and forest degradation with implications for seed dispersal and zoonotic diseases. *Journal of Animal Ecology*, *91*, 794–804.

Depaoli, S. (2021). *Bayesian structural equation modeling*. The Guilford Press.

Dijkstra, T. K., & Henseler, J. (2015). Consistent partial least squares path modeling. *MIS Quarterly*, *39*, 297–316.

Dorresteijn, I., Schultner, J., Nimmo, D., Fischer, J., Hanspach, J., Kuemmerle, T., Kehoe, L., & Ritchie, E. (2015). Incorporating anthropogenic effects into trophic ecology: Predator—Prey interactions in a human-dominated landscape. *Proceedings of the Royal Society - Biological Sciences/the Royal Society*, *282*, 1–8.

Dunn, A., Amir, Z., Decoeur, H., Dehaudt, B., Nursamsi, I., Mendes, C., Moore, J. H., Negret, P. J., Sovie, A., & Luskin, M. S. (2022). The ecology of the banded civet (*Hemigalus derbyanus*) in Southeast Asia with implications for mesopredator release, zoonotic diseases, and conservation. *Ecology and Evolution*, *12*, e8852.

Eisenhauer, N., Bowker, M. A., Grace, J. B., & Powell, J. R. (2015). From patterns to causal understanding: Structural equation modeling (SEM) in soil ecology. *Pedobiologia*, *58*, 65–72.

Estes, J. A., Terborgh, J., Brashares, J. S., Power, M. E., Berger, J., Bond, W. J., Carpenter, S. R., Essington, T. E., Holt, R. D., Jackson, J. B. C., Marquis, R. J., Oksanen, L., Oksanen, T., Paine, R. T., Pikitch, E. K., Ripple, W. J., Sandin, S. A., Scheffer, M., Schoener, T. W., … Wardle, D. A. (2011). Trophic downgrading of planet earth. *Science*, *333*, 301–306.

Fan, Y., Chen, J., Shirkey, G., John, R., Wu, S. R., Park, H., & Shao, C. (2016). Applications of structural equation modeling (SEM) in ecological studies: An updated review. *Ecological Processes*, *5*, 19.

Forbes, C., Evans, M., Hastings, N., & Peacock, B. (2010). *Statistical distributions* (4th ed.). Wiley, New Jersey.

Gana, K., & Broc, G. (2019). *Structural equation modeling with lavaan*. John Wiley & Sons.

Gardner, C. J., Bicknell, J. E., Baldwin-Cantello, W., Struebig, M. J., & Davies, Z. G. (2019). Quantifying the impacts of defaunation on natural forest regeneration in a global meta-analysis. *Nature Communications*, *10*, 4590.

Gaynor, K. M., Hojnowski, C. E., Carter, N. H., & Brashares, J. S. (2018). The influence of human disturbance on wildlife nocturnality. *Science*, *360*, 1232–1235.

Grace, J. B. (2024). An integrative paradigm for building causal knowledge. *Ecological Monographs*, *94*, e1628.

Grace, J. B., Johnson, D. J., Lefcheck, J. S., & Byrnes, J. E. K. (2018). Quantifying relative importance: Computing standardized effects in models with binary outcomes. *Ecosphere*, *9*, e02283.

Grace, J. B., Schoolmaster, D. R., Jr., Guntenspergen, G. R., Little, A. M., Mitchell, B. R., Miller, K. M., & Schweiger, E. W. (2012). Guidelines for a graph-theoretic implementation of structural equation modeling. *Ecosphere*, *3*, art73.

Grantham, H. S., Duncan, A., Evans, T. D., Jones, K. R., Beyer, H. L., Schuster, R., Walston, J., Ray, J. C., Robinson, J. G., Callow, M., Clements, T., Costa, H. M., DeGemmis, A., Elsen, P. R., Ervin, J., Franco, P., Goldman, E., Goetz, S., Hansen, A., … Watson, J. E. M. (2020). Anthropogenic modification of forests means only 40% of remaining forests have high ecosystem integrity. *Nature Communications*, *11*, 5978.

Hair, J., Hult, G. T. M., Ringle, C., Sarstedt, M., Danks, N., & Ray, S. (2021). Partial least squares structural equation modeling (PLS-SEM) using R: A workbook.

Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2021). *A primer on partial least squares structural equation modeling (PLS-SEM)* (3rd ed.). SAGE Publications.

Hendry, A., Amir, Z., Decoeur, H., Mendes, C. P., Moore, J. H., Sovie, A., & Luskin, M. S. (2023). Marbled cats in Southeast Asia: Are diurnal and semi-arboreal felids at greater risk from human disturbances? *Ecosphere*, *14*, e4338.

Honda, A., Amir, Z., Mendes, C. P., Moore, J. H., & Luskin, M. S. (2023). Binturong ecology and conservation in pristine, fragmented and degraded tropical forests. *Oryx*, *58*, 1–10.

Hoyle, R. H. (2012). *Handbook of structural equation modeling*. The Guilford Press.

Jackson, D. L. (2003). Revisiting sample size and number of parameter estimates: Some support for the N:q hypothesis. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*, 128–141.

Jiang, M., Lee, S., O'Malley, A. J., Stern, Y., & Li, Z. (2023). A novel causal mediation analysis approach for zero-inflated mediators. *Statistics in Medicine*, *42*, 2061–2081.

Kays, R., & Wikelski, M. (2023). The internet of animals: What it is, what it could be. *Trends in Ecology & Evolution*, *38*, 859–869.

Ke, A., & Luskin, M. S. (2019). Integrating disparate occurrence reports to map data-poor species ranges and occupancy: A case study of the Vulnerable bearded pig *Sus barbatus*. *Oryx*, *53*, 377–387.

Kline, R. B. (2023). *Principles and practice of structural equation modeling* (5th ed.). The Guilford Press.

Kunicki, Z. J., Smith, M. L., & Murray, E. J. (2023). A primer on structural equation model diagrams and directed acyclic graphs: When and how to use each in psychological and epidemiological research. *Advances in Methods and Practices in Psychological Science*, *6*, 25152459231156085.

Lamperty, T., Chiok, W. X., Khoo, M. D. Y., Amir, Z., Baker, N., Chua, M. A. H., Chung, Y. F., Chua, Y. K., Koh, J. J.-M., Lee, B. P. Y.-H., Lum, S. K. Y., Mendes, C. P., Ngiam, J., ODempsey, A., Png, K. G. C., Sovie, A. R., Tan, L., Teo, R., Thomas, N., … Luskin, M. S. (2023). Rewilding in Southeast Asia: Singapore as a case study. *Conservation Science and Practice*, *5*, e12899.

Lee, S. X. T., Amir, Z., Moore, J. H., Gaynor, K. M., & Luskin, M. S. (2024). Effects of human disturbances on wildlife behaviour and consequences for predator-prey overlap in Southeast Asia. *Nature Communications*, *15*, 1521.

Lefcheck, J., Byrnes, J. E., & Grace, J. B. (2023). piecewiseSEM: Piecewise structural equation modeling. Version 2.3.0. (R package).

Lima, F., Beca, G., Muylaert, R., Jenkins, C., Perilli, M., Paschoal, A., Massara, R., Paglia, A., Chiarello, A., Graipel, M., Cherem, J., Regolin, A., Oliveira-Santos, L., Brocardo, C., Paviolo, A., Di Bitetti, M., Moraes, L., Lopes Rocha, F., Fusco-Costa, R., & Galetti, M. (2017). ATLANTIC-CAMTRAPS: A dataset of medium and large terrestrial mammal communities in the Atlantic Forest of South America. *Ecology*, *98*, 2979.

Luskin, M., Moore, J., Mendes, C., Davies, S., Nasardim, M., & Manabu, O. (2023). The mass mortality of Asia's native pigs induced by African Swine Fever. *Wildlife Letters*, *1*, 8–14.

Luskin, M. S., Arnold, L., Sovie, A., Amir, Z., Chua, M. A. H., Dehaudt, B., Dunn, A., Nursamsi, I., Moore, J. H., & Mendes, C. P. (2023). Mesopredators in forest edges. *Wildlife Letters*, *1*, 107–118.

Luskin, M. S., Brashares, J. S., Ickes, K., Sun, I. F., Fletcher, C., Wright, S. J., & Potts, M. D. (2017). Cross-boundary subsidy cascades from oil palm degrade distant tropical forests. *Nature Communications*, *8*, 2231.

Luskin, M. S., Christina, E. D., Kelley, L. C., & Potts, M. D. (2014). Modern hunting practices and wild meat trade in the oil palm plantation-dominated landscapes of Sumatra, Indonesia. *Human Ecology*, *42*, 35–45.

Luskin, M. S., Ickes, K., Yao, T. L., & Davies, S. J. (2019). Wildlife differentially affect tree and liana regeneration in a tropical forest: An 18-year study of experimental terrestrial defaunation versus artificially abundant herbivores. *Journal of Applied Ecology*, *56*, 1379–1388.

Luskin, M. S., Johnson, D. J., Ickes, K., Yao, T. L., & Davies, S. J. (2021). Wildlife disturbances as a source of conspecific negative density-dependent mortality in tropical trees. *Proceedings of the Royal Society B: Biological Sciences*, *288*, 20210001.

Luskin, M. S., Meijaard, E., Surya, S., Sheherazade, S., Walzer, C., & Linkie, M. (2021). African swine fever threatens Southeast Asia's 11 endemic wild pig species. *Conservation Letters*, *14*, e12784.

McElreath, R. (2020). *Statistical rethinking. A Bayesian course with examples in R and STAN*. Chapman and Hall/CRC.

McGregor, H., Moseby, K., Johnson, C. N., & Legge, S. (2020). The short-term response of feral cats to rabbit population decline: Are alternative native prey more at risk? *Biological Invasions*, *22*, 799–811.

Mendes, C., & Luskin, M.S. (2024). CamTrapAsia: A dataset of tropical forest vertebrate communities from 239 camera trapping studies [data set]. *Zenodo*. https://doi.org/10.5281/zenodo.10780971

Mendes, C. P., Albert, W. R., Amir, Z., Ancrenaz, M., Ash, E., Azhar, B., Bernard, H., Brodie, J., Bruce, T., Carr, E., Clements, G. R., Davies, G., Deere, N. J., Dinata, Y., Donnelly, C. A., Duangchantrasiri, S., Fredriksson, G., Goossens, B., Granados, A., … Luskin, M. S. (2024). CamTrapAsia: A dataset of tropical forest vertebrate communities from 239 camera trapping studies. *Ecology*, *105*, e4299.

Mendes, C. P., Liu, X., Amir, Z., Moore, J. H., & Luskin, M. S. (2023). A multi-scale synthesis of mousedeer habitat associations in Southeast Asia reveals declining abundance but few extirpations in fragments and edges. *Austral Ecology*, *49*, e13470.

Mendes, C. P., & Luskin, M. (2025). Supplementary code for "The use, misuse, and opportunities for structural equation modelling (SEM) in wildlife ecology". *Zenodo*. https://doi.org/10.5281/zenodo.17122491

Mendes, C. P., Ribeiro, M. C., & Galetti, M. (2016). Patch size, shape and edge distance influence seed predation on a palm species in the Atlantic forest. *Ecography*, *39*, 465–475.

Merkle, E. C., & Rosseel, Y. (2018). blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software*, *85*, 1–30.

Meshcheryakov, G., Igolkina, A. A., & Samsonova, M. G. (2021). semopy 2: A structural equation modeling package with random effects in python. *ArXiv*, abs/2106.01140.

Moore, J., Palmeirim, A., Peres, C., Ngoprasert, D., & Gibson, L. (2022). Invasive rat drives complete collapse of native small mammal communities in insular forest fragments. *Current Biology*, *32*, 2997–3004.

Moore, J. H., Gibson, L., Amir, Z., Chanthorn, W., Ahmad, A. H., Jansen, P. A., Mendes, C. P., Onuma, M., Peres, C. A., & Luskin, M. S. (2023). The rise of hyperabundant native generalists threatens both humans and nature. *Biological Reviews*, *98*, 1–10.

Muthén, L. K., & Muthén, B. O. (2007). *Mplus user's guide* (8th ed.). Mplus.

Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C., & Clune, J. (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences of the United States of America*, *115*, E5716–E5725.

Nursamsi, I., Amir, Z., Decoeur, H., Moore, J. H., & Luskin, M. S. (2023). Sunda pangolins show inconsistent responses to disturbances across multiple scales. *Wildlife Letters*, *1*, 59–70.

Orrock, J. L., & Danielson, B. J. (2005). Patch shape, connectivity, and foraging by Oldfield mice (*Peromyscus polionotus*). *Journal of Mammalogy*, *86*, 569–575.

Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, *62*, 54–60.

Peres, C. A., Emilio, T., Schietti, J., Desmoulière, S. J., & Levi, T. (2016). Dispersal limitation induces long-term biomass collapse in overhunted Amazonian forests. *Proceedings of the National Academy of Sciences of the United States of America*, *113*, 892–897.

Plummer, M., Stukalov, A., & Denwood, M. (2023). rjags: Bayesian graphical models using MCMC. R package.

Prugh, L. R., Stoner, C. J., Epps, C. W., Bean, W. T., Ripple, W. J., Laliberte, A. S., & Brashares, J. S. (2009). The rise of the mesopredator. *Bioscience*, *59*, 779–791.

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, *1*, 27–42.

Rönkkö, M., & Evermann, J. (2013). A critical examination of common beliefs about partial least squares path modeling. *Organizational Research Methods*, *16*, 425–448.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36.

Rosseel, Y., Jorgensen, T. D., Wilde, L. D., Oberski, D., Byrnes, J., Vanbrabant, L., Savalei, V., Merkle, E., Hallquist, M., Rhemtulla, M., Katsikatsou, M., Barendse, M., Rockwood, N., Scharf, F., Du, H., Jamil, H., & Classe, F. (2024). Package 'lavaan'. Latent variable analysis. CRAN.

Russo, N. J., Davies, A. B., Blakey, R. V., Ordway, E. M., & Smith, T. B. (2023). Feedback loops between 3D vegetation structure and ecological functions of animals. *Ecology Letters*, *26*, 1597–1613.

Sakaria, D., Maat, S. M., & Mohd Matore, M. E. E. (2023). Examining the optimal choice of SEM statistical software packages for sustainable mathematics education: A systematic review. *Sustainability*, *15*, 3209.

Seltzer, R. G. N. (2021). The perilous use of proxy variables. *Evaluation & the Health Professions*, *44*, 428–435.

Sills, J., Brodie, J. F., & Gibbs, H. K. (2009). Bushmeat hunting as climate threat. *Science*, *326*, 364–365.

Stan Development Team. (2023). RStan: The R interface to Stan. R package.

StataCorp. (2023). *Stata: Release 18. Statistical software*. StataCorp LLC.

Tobler, M. W., Kéry, M., Hui, F. K. C., Guillera-Arroita, G., Knaus, P., & Sattler, T. (2019). Joint species distribution models with species correlations and imperfect detection. *Ecology*, *100*, e02754.

Venter, O., Sanderson, E. W., Magrach, A., Allan, J. R., Beher, J., Jones, K. R., Possingham, H. P., Laurance, W. F., Wood, P., Fekete, B. M., Levy, M. A., & Watson, J. E. M. (2016). Sixteen years of change in the global terrestrial human footprint and implications for biodiversity conservation. *Nature Communications*, *7*, 12558.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Figure S1. Histogram showing that RAIs are frequently present a log-normal distribution. The data were extracted from the CamTrapAsia dataset (Mendes et al., 2024). The *p*-values of Shapiro–Wilk tests (SW) are also displayed.

Figure S2. Differences between the estimated coefficients and the true value. For descriptions of each model, see the main text in Table 2. The models marked with 'ZI.med' in the name uses the approach outlined by Jiang et al. (2025) wherein a mediator variable is used to handle the data zero-inflation.

Figure S3. Standard error of the estimated coefficients for all SEM approaches tested.

Text S1. Types and uses of observational wildlife datasets.

Text S2. Common statistical complexities in capture summary datasets.

How to cite this article: Mendes, C. P., & Luskin, M. S. (2025). The use, misuse and opportunities for structural equation modelling (SEM) in wildlife ecology. *Journal of Applied Ecology*, *00*, 1–17. https://doi.org/10.1111/1365-2664.70189