Introduction to structural equation modelling

Frank Pennekamp

11/11/2020

Research questions

- Ecology is about the interactions between organisms and their environments.
- We often have ideas how things could be connected in ecological systems.
- To test these ideas, we need a way to dissect when they occurr for a reason versus randomness.
- We use statistics to understand, when connections are non-random.
- Research questions are about understanding cause and effect.

Why do we need to use SEM in Ecology

- System thinking
  - "Understanding the whole rather than the parts in isolation"
- SEM unites multiple variables in a single causal network -> simultaneous tests of multiple hypotheses.
- Causality is central to SEM: technique implicitly assumes that the relationships among variables represent causal links.

Why do we need to use SEM in Ecology

To better understand the relationship among variables in complex ecosystems:
- We often have multiple observed variables - We want to test and evaluate multivariate causal relationships - Test direct and indirect effects on assumed causal relationships - Incorporate observed and latent variables - Include interaction terms can test main effects and interaction effects

Why do we need to use SEM in Ecology

Two goals of SEM:
1) Understand the patterns of correlation/covariance among a set of variables.
2) Explain as much of their variance as possible with the model specified.

Thinking about the model

A SEM is usually specified based on theory to determine and validify a proposed causal process and/or model.

Consideration of variables:
- Supported by theory.
- Ecologically meaningful.
- Garbage in - garbage out concept (both data quality and ecologically meaningful).

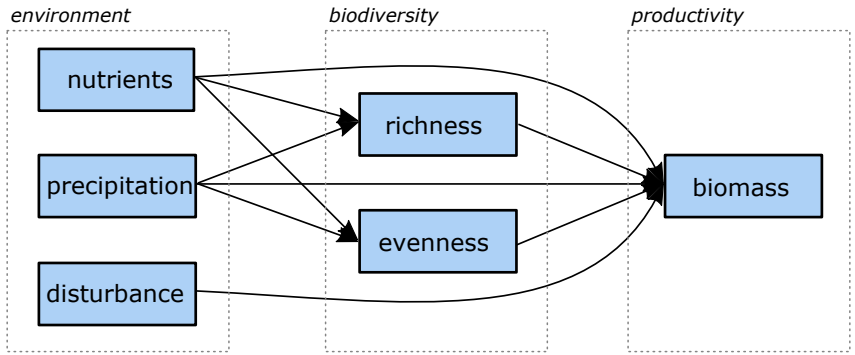
Setting arrow direction.
- Causal vs correlation.

Good practice: Make a table / graph of putative causal relationships before analysis.

Meta-model

- A metamodel summarizes the concept behind a model and links it to theory.

Productivity (biomass) is directly influence on the one hand by the environment (nutrients, disturbance and precipitation) and on the other hand by biodiversity (richness and evenness). Also some elements of the environment influence biodiversity and thus, have an additional indirect effect on productivity via biodiversity.



Modelling philosophy

- Two approaches: theory (hypothesis-driven) to exploratory (data-driven).
- Show as a continuum between two extremes.
- Importance of being aware about what the model is telling to us.
- Final discussion about what SEM provides and the importance of system thinking in ecology.

Introduction to the dataset

We will use an experimental dataset collected at the Cedar Creek Ecosystem Science Reserve to examine long-term consequences of human-driven environmental changes ecosystem responses to:

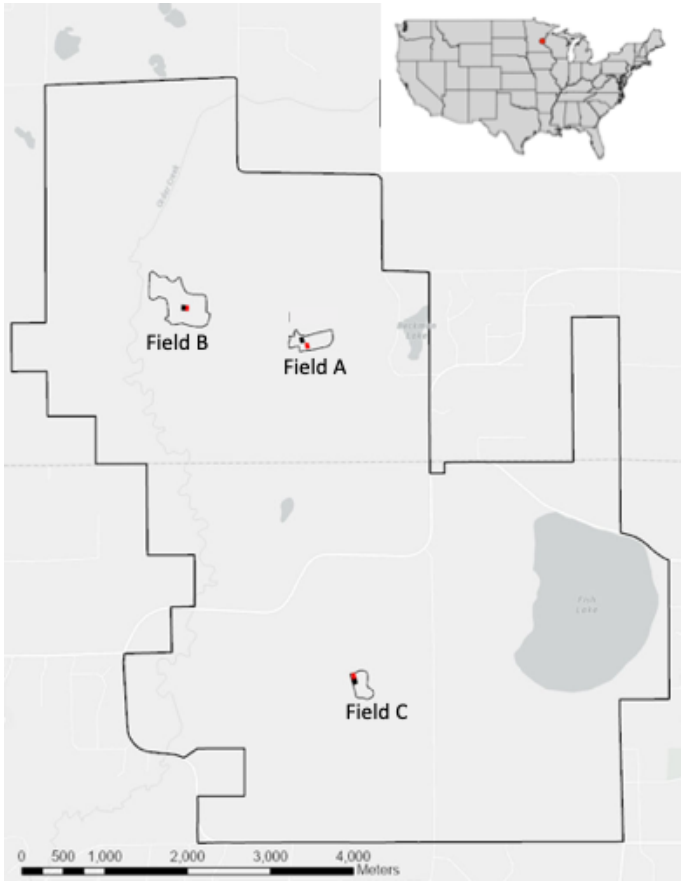
- Disturbance
- Nitrogen deposition
- Changes in precipitation

Introduction to the dataset

Understanding the recovery of a grassland for two decades following an intensive agricultural disturbance under ambient and elevated nutrient conditions.

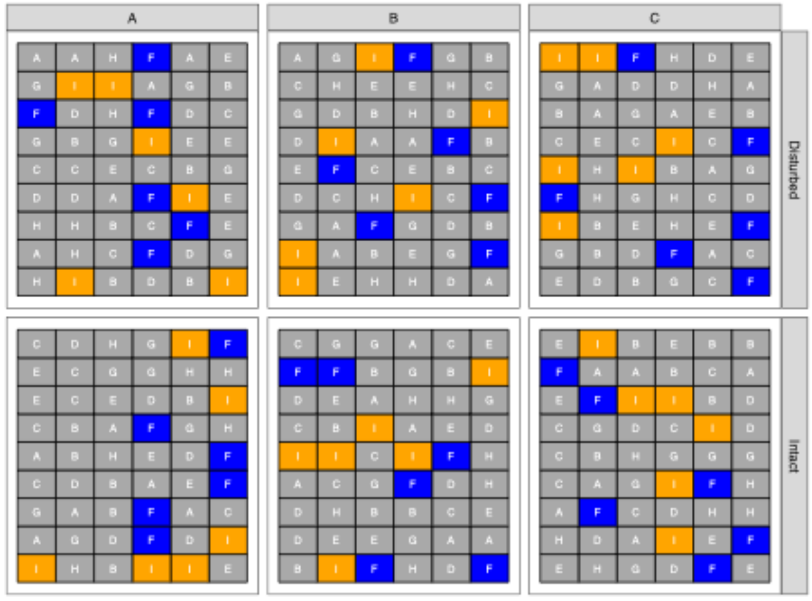
1) How has aboveground biomass changed as a function of disturbance (disking) and nutrient addition?
2) How are these effects mediated by diversity?

Introduction to the dataset



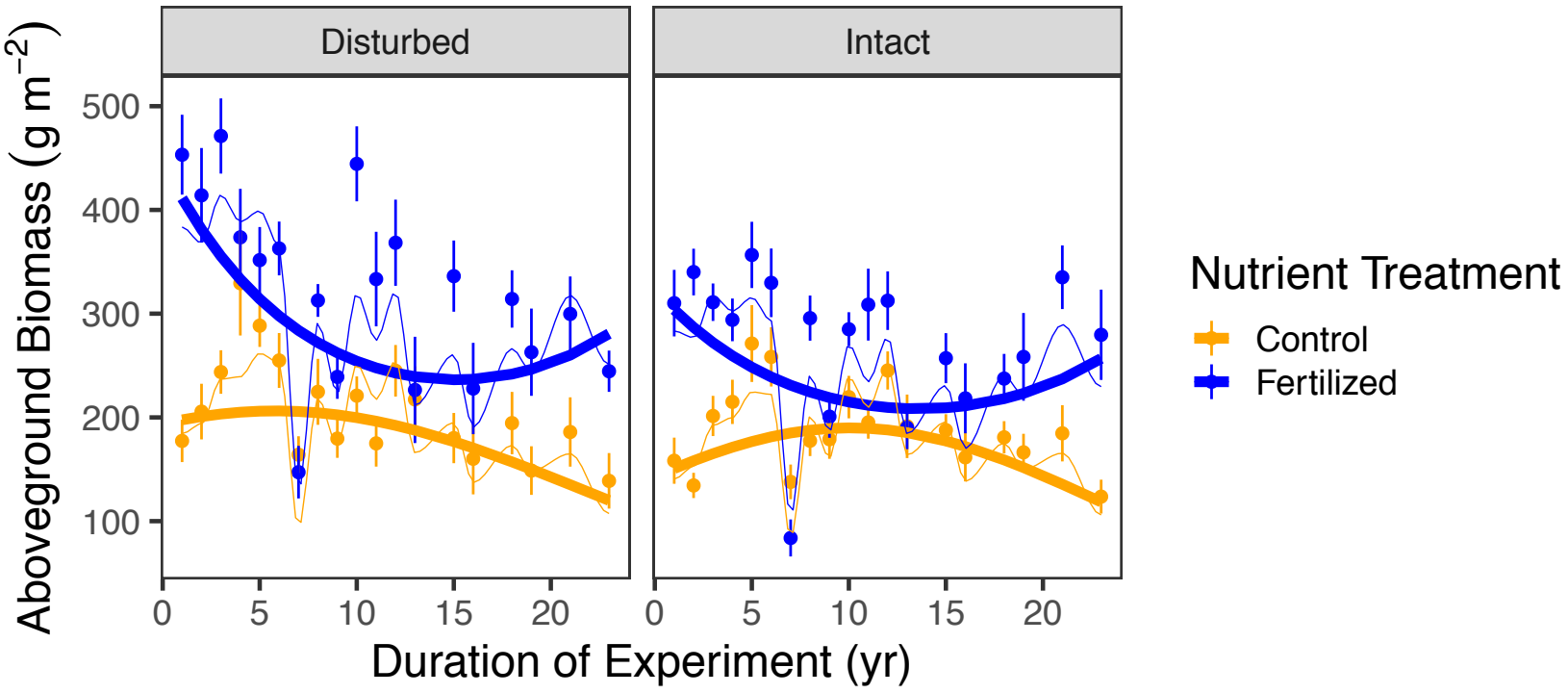
Map showing location of the study site (Cedar Creek Ecosystem Science Reserve), the location of the three study fields within the reserve, and location of the 35 x 55 m intact (black) and disturbed (red) plots within each field.

Introduction to the dataset



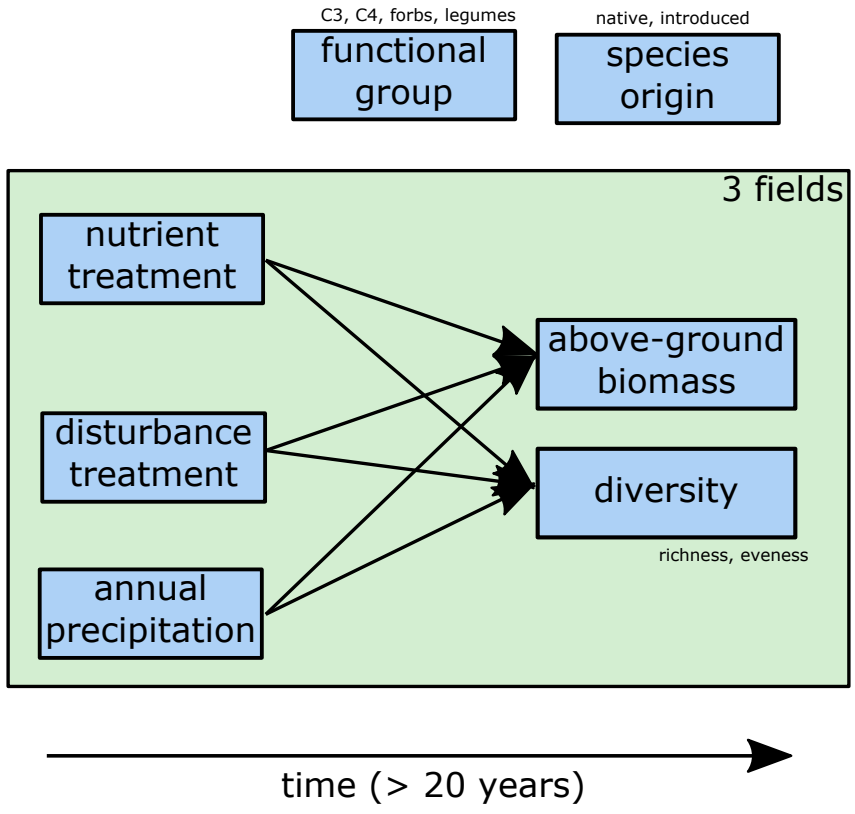
Location of the 4 x 4 m nutrient treatment plots within each 35 x 55 m Intact or Disturbed plot within each of three fields (A, B, and C). Letters indicate the nutrient treatments, and the colored plots are treatments that are the focus of the analyses presented here: Control (orange) and 9.5 g N m-2 yr-1 (blue).

Introduction to the dataset



Effect of soil disturbance (disking) and nutrient enrichment on live, aboveground plant biomass. Colors indicate nutrient addition treatment: Control and NPK+ (all nutrients plus 9.5 g N m-2 yr-1).

Question of interest

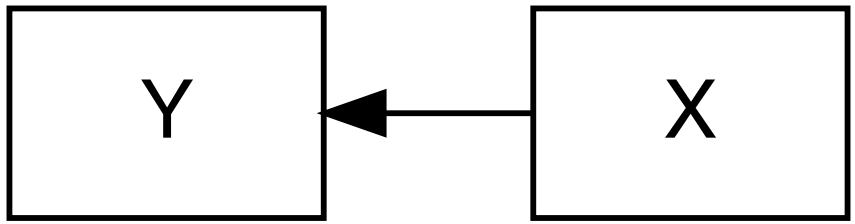


Present one research question based on the data. This question only NEEDS to include 2 variables. Biomass -> richness and evenness

Graphical representation

Draw a model by hand

Directed acyclic graphs:



General notation:
- Variables are nodes.
- Arrows are causal relationships.

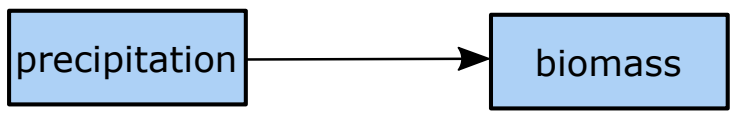
Causality

- What is a causal relationship?
- “Correlation does not imply causation”
- Causation indicates a relation between two variables in which one variable is affected by another.
- The arrow above indicates a causal relationship.
- Explain the concept and discuss with students the topic.
- Show the difference between direct effects and correlation together with the graphical representation.

Simple statistical model

Present the bivariate model. Include the equation and a plot. It is important that students understand that all they know can be defined with a graphical representation.

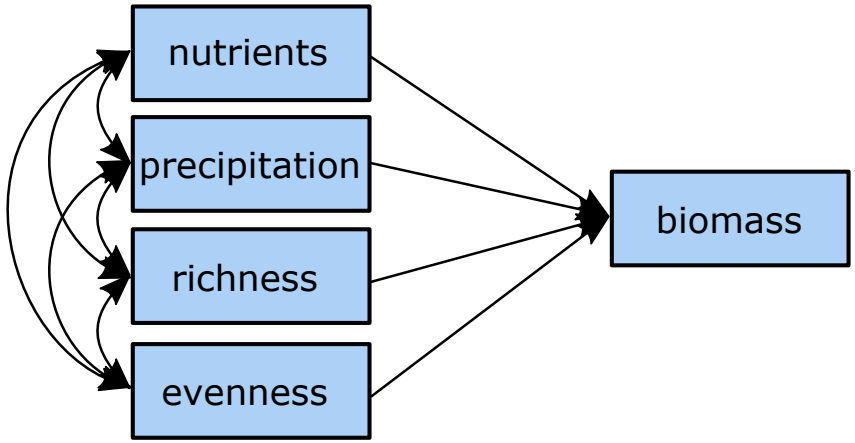
`lm(y~x1)`



Multiple independent variables

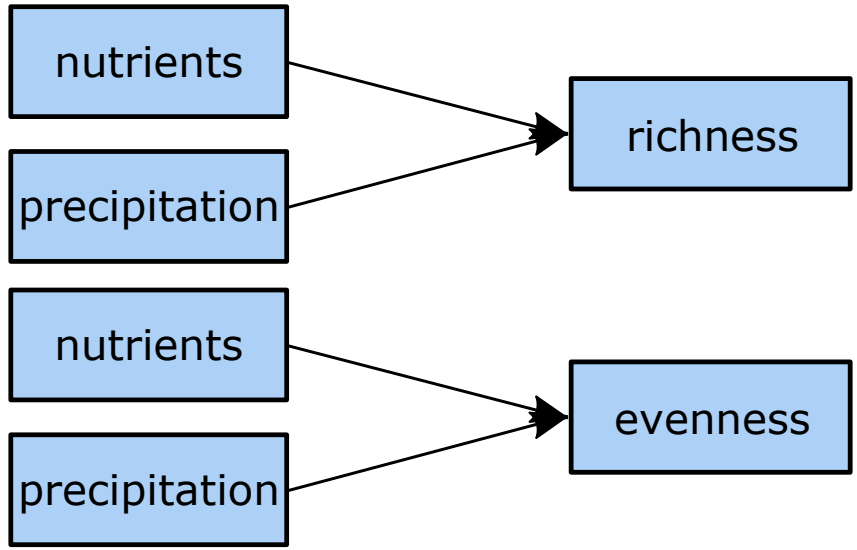
- Discuss with students other variables that may affect B.
- Introduce a new variable and the question that it answers.
- Add nutrient variable.
- Show the graphical model (X1 -> Y, X2 -> Y).
- Include the equation and the two partial plots.
- Be clear this is a multiple regression model.

```
lm(y~x1+x2+x3)
```



Indirect effect

- All relationships in the previous models are direct.
  - Directional relation between two variables, e.g., independent and dependent variables.
- Indirect effect is the effect of an independent variable on a dependent variable through one or more intervening or mediating variables.
- Include partial plots for both to illustrate the indirect effect.

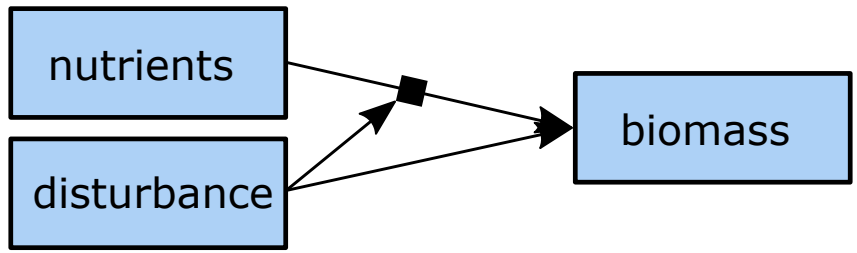


Mediation

- Tests whether a particular variable has a mediating effect.
- Often used to test underlying mechanisms.

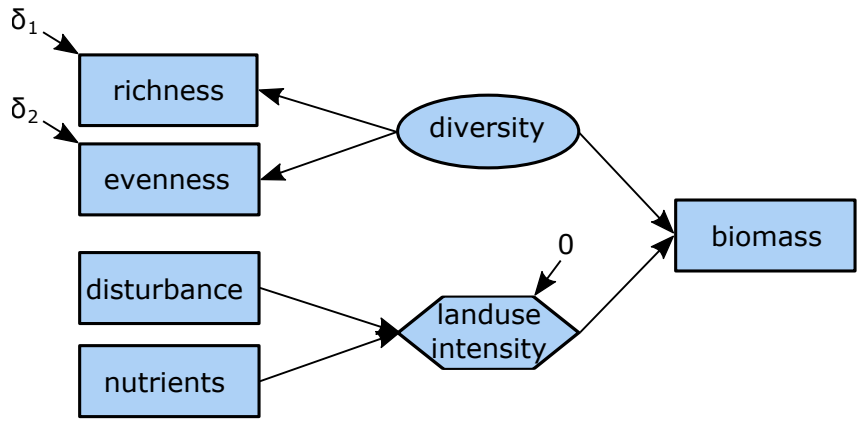
Interaction questions (Moderation)

- Present another question involving a new variable with an interactive effect (X3 -> (Y -> Z)).
- Show the graphical representation.
- Explanation of interactive effects.
- Include partial plots to illustrate the interaction effect.
- Make clear that this is the same as an interaction in traditional regression models.



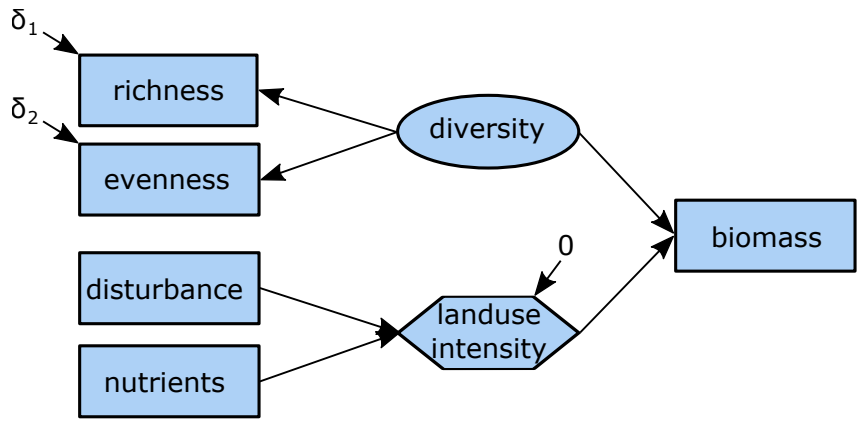
Latent variables

Until now everything was about paths but modelling also involves variables. A variable that is not directly measured is a latent variable (examples).



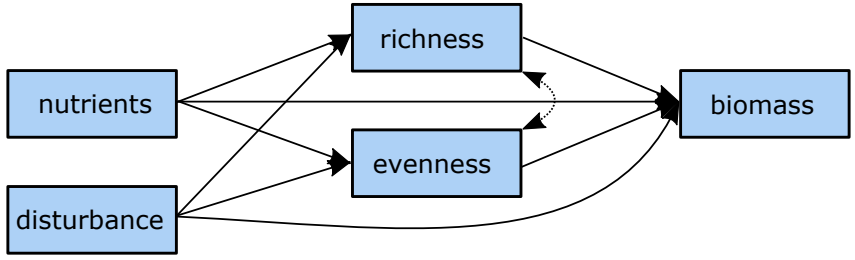
Composite variables

Composite variables specify the influences of collections of other variables (examples)



System level approach

- Show a very complex model including multiple types of variables, arrows (direct and correlations) for the study system.
- The complex model is built based on submodels, each one addressing specific questions.
- We need a framework to analyze this together -> SEM
- Discuss about complexity in ecology and if SEM is useful to answer your questions ("We need to do things as simple as possible, but no simpler").



Assumptions of SEM with lavaan

- Normality
- Global estimation[1]
- Directed acyclic relationships[2]
- Linear relationships[3]
- Backdoor criterion

[1] Local estimation possible.

[2] Causal loops possible.

[3] Nonlinear relationships possible.

SEM workflow

1) Review the relevant theory and research literature to support model specification
2) Specify a model (e.g., diagram, equations)
3) Determine model identification (e.g., if unique values can be found for parameter estimation; the number of degrees of freedom, df, for model testing is positive)
4) Select measures for the variables represented in the model
5) Collect data
6) Conduct preliminary descriptive statistical analysis (e.g., scaling, missing data, collinearity issues, outlier detection)
7) Estimate parameters in the model
8) Assess model fit
9) Re-specify the model if meaningful
10) Interpret and present results visually