Properties of data on ecological communities

Bert van der Veen

Department of Mathematical Sciences, NTNU

The goal of this presentation

Instill basic thinking about study design and data properties.
It may be a bit boring.

Gathering data

You go out, register species at multiple sites

Figure 1: Geir-Harald Strand / NIBIO

Gathering data

You go out, register species at multiple sites

Figure 1: Geir-Harald Strand / NIBIO

What does community data look like?

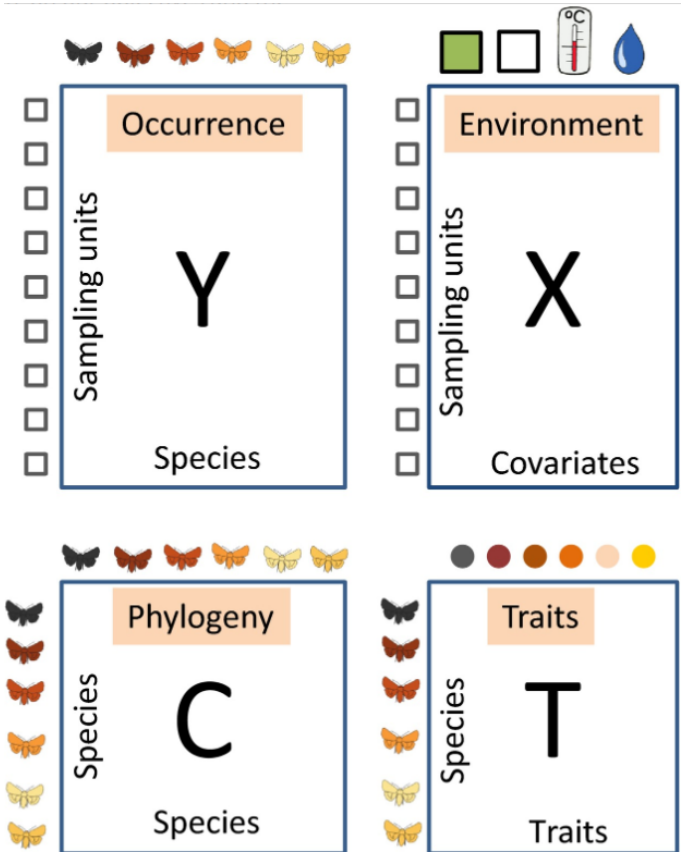
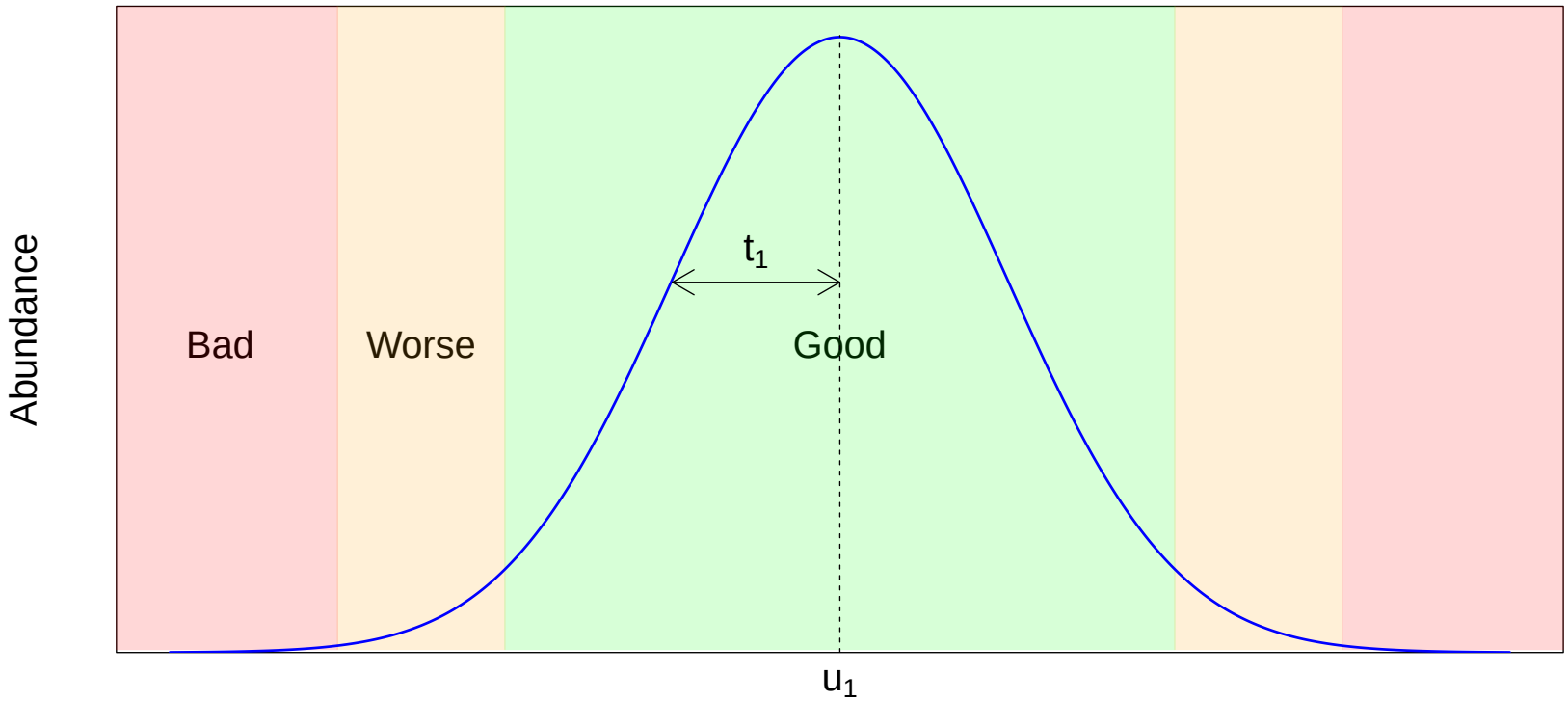


Figure 2: Ovaskainen et al. (2017) Fig. 3

Goal: analysis of sparse data

Shelford's law of tolerance (1931) tells us:

- There are specialist and generalist species
- So, most species occur at few places



Getting results

You have got your data, and are ready to do some ecology!

Answering ecological questions starts at a good study design.

In the absence of a good study design, or due to inherent properties of the process under study, a model helps.

The purpose of modeling is to get a good answer to our questions.

Process-based thinking

I will reiterate this multiple times, but:

1. There is a sampling process
2. There is an ecological process

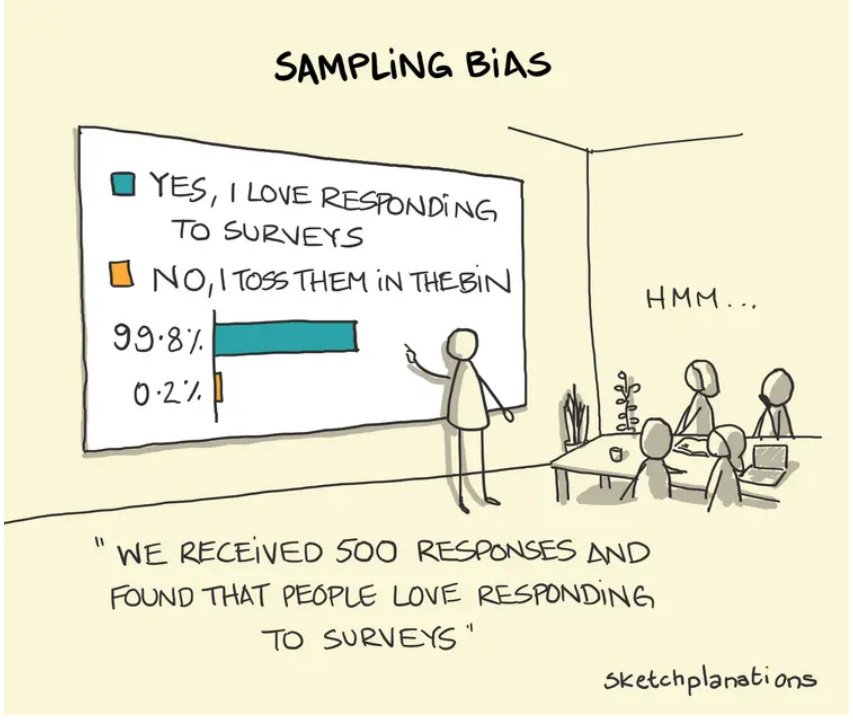
Our data is the result of both, our primary interest is the latter

Minimizing impact of the sampling process

We can minimize the effects of sampling by considering its effects *a-priori*

There are many sampling designs in community ecology

- Opportunistic (eek)
- Random sampling
- Systematic sampling
- Stratified sampling
- Stratified-random sampling
- Adaptive sampling
- Cluster sampling
- Paired sampling

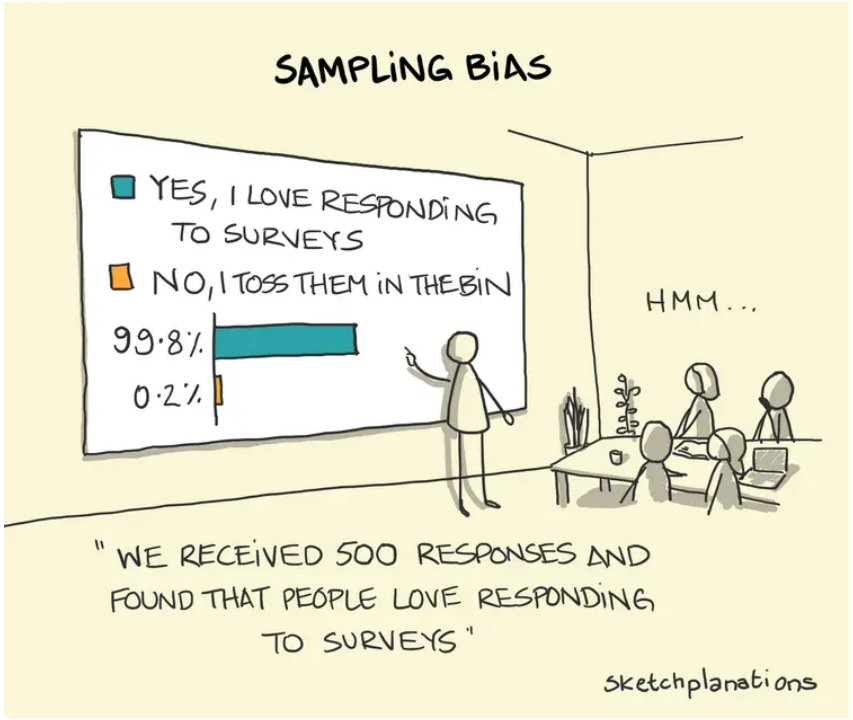


Minimizing impact of the sampling process

We can minimize the effects of sampling by considering its effects *a-priori*

There are many sampling designs in community ecology

- Opportunistic (eek)
- Random sampling
- Systematic sampling
- Stratified sampling
- Stratified-random sampling
- Adaptive sampling
- Cluster sampling
- Paired sampling



Sampling design affects our sample size, and the ecological results. It needs to be taken into account during analysis.

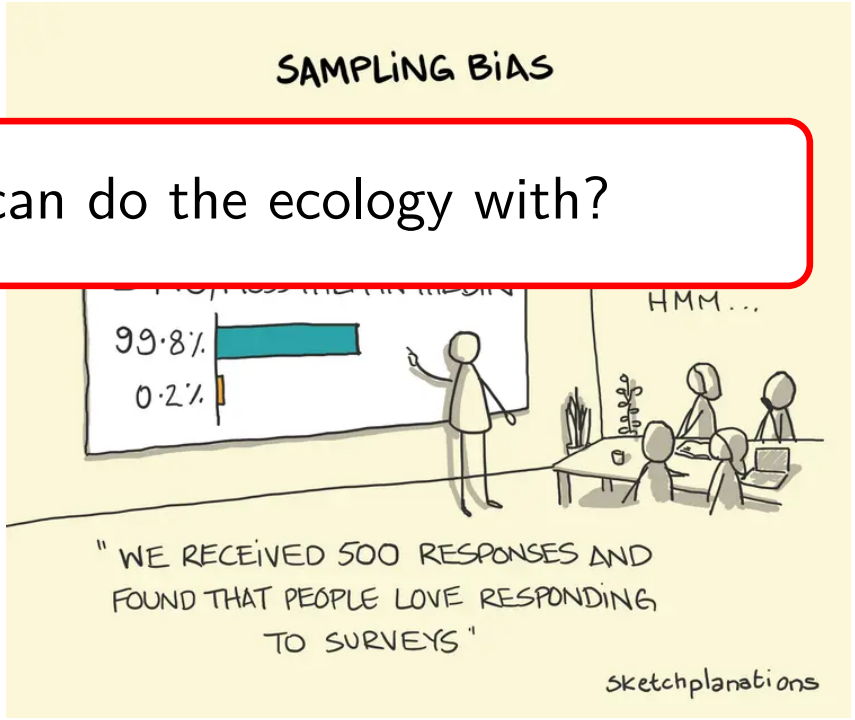
Minimizing impact of the sampling process

We can minimize the effects of sampling by considering its effects *a-priori*

There are many sampling designs in community ecology

Does it give data that you can do the ecology with?

- Systematic sampling
- Stratified sampling
- Stratified-random sampling
- Adaptive sampling
- Cluster sampling
- Paired sampling



Sampling design affects our sample size, and the ecological results. It needs to be taken into account during analysis.

Aspects of how we sample

Statistically, the important thing is that samples are **independent**

- Space
- Time
- Species
- Hierarchical designs

Can mess that up (and often do), consequence: biased results and inflated Type error

Preferential sampling

"I want to survey community A"

or

"I sample on an elevation gradient"

a) You have predefined your community; the predefinition affects your results
b) You have predefined your environment; the predefinition affects your results

Preferential sampling

"I want to survey community A"

or

"I sample on an elevation gradient"

a) You have predefined your community; the predefinition affects your results
b) You have predefined your environment; the predefinition affects your results

You self-limited the scope of your study, self-selected results for diversity, composition, environment, and so on.

Preferential sampling

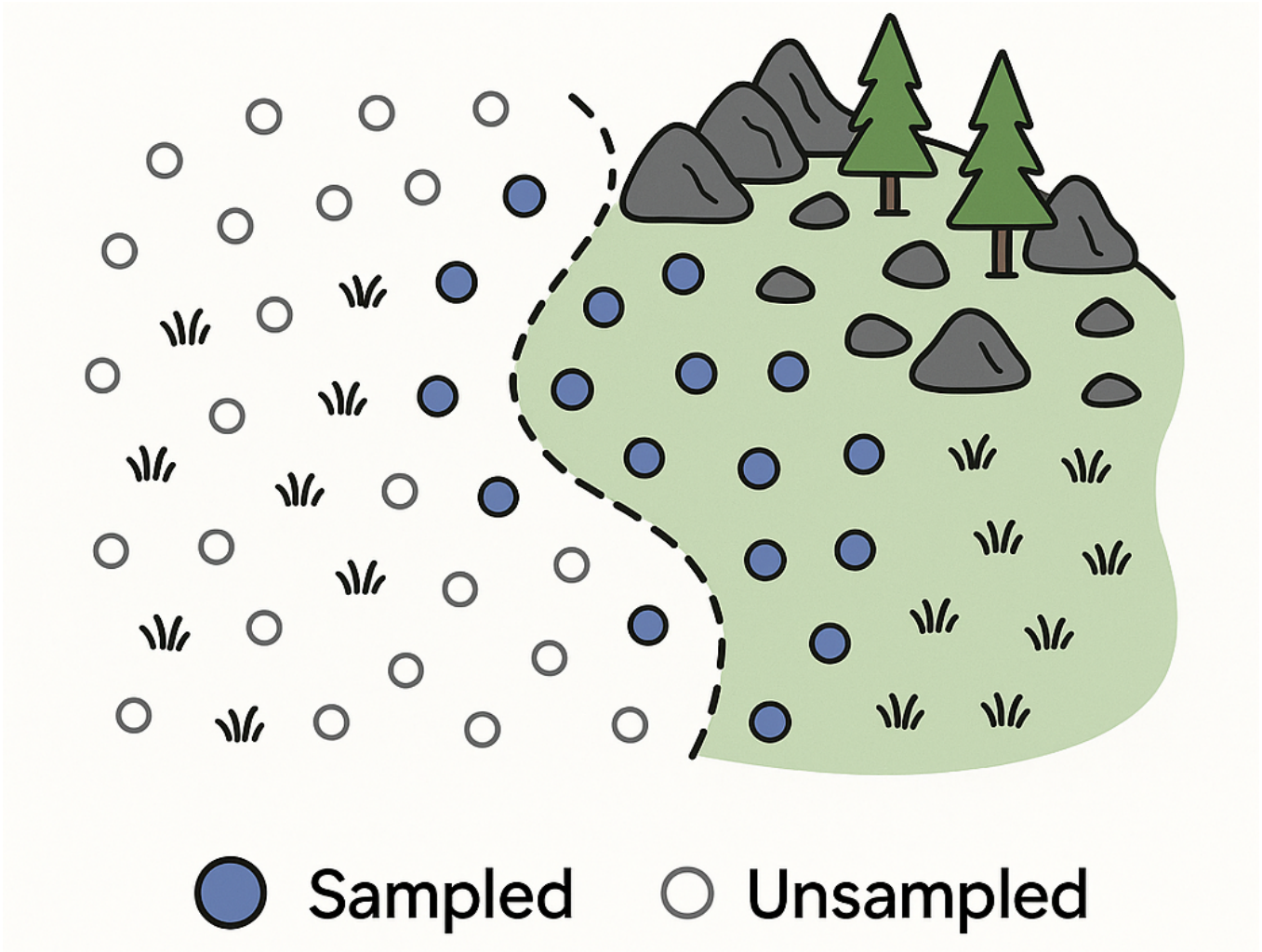


Figure 3: Thanks chatGPT

Detection bias

This one is not often covered, but certain species are harder to sample (identify or find) than others.

- Not considering it: you assume perfect detection
- Plants are easier than moving things
- Plants or flower are seasonal
- Pollinators fly at particular conditions
- Insects have different life stages (some easier to detect)
- Some people are better at finding things

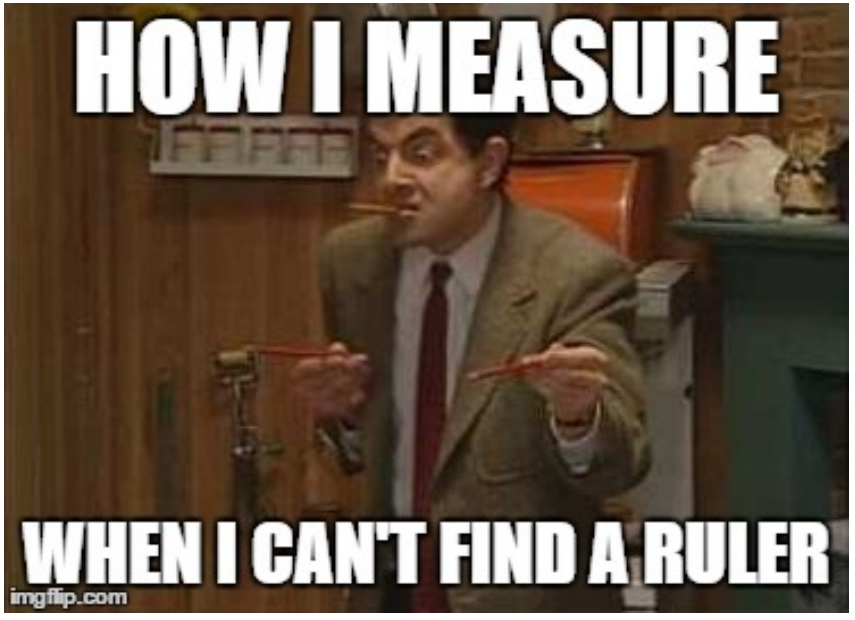
Classification error

Classification mistakes introduce error: we confuse a species with another.

Exacerbated if you have multiple observers.

Measurement error

- Measuring the environment (elevation, moisture)
- In responses of species (counts or cover data)
- eDNA data
- Can be due to faulty instrument calibration
- Experiment gone wrong
- Sample omitted, miswritten, or wrongly entered in excel



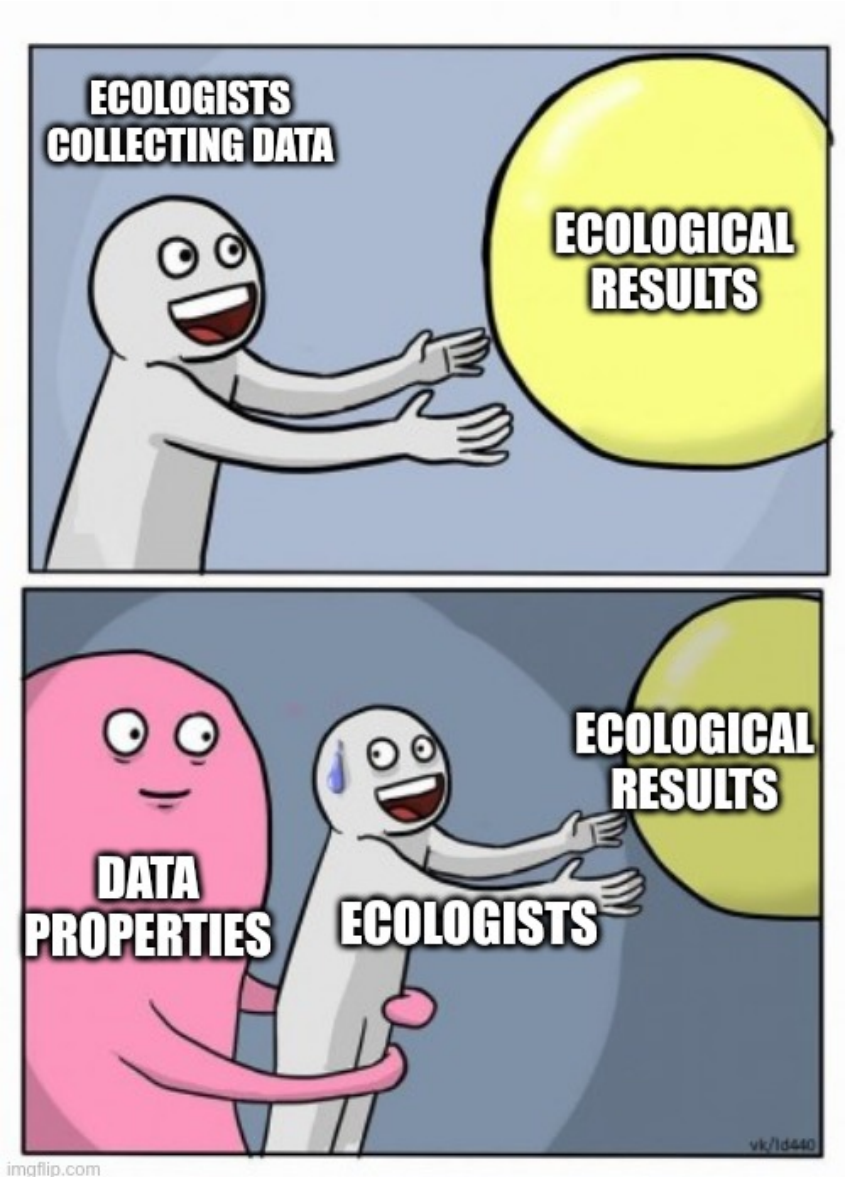
To do ecology

We need to carefully consider how our data is sampled and what our analysis needs to accommodate.

Data properties

Let's say you collected your data without too much sampling error.

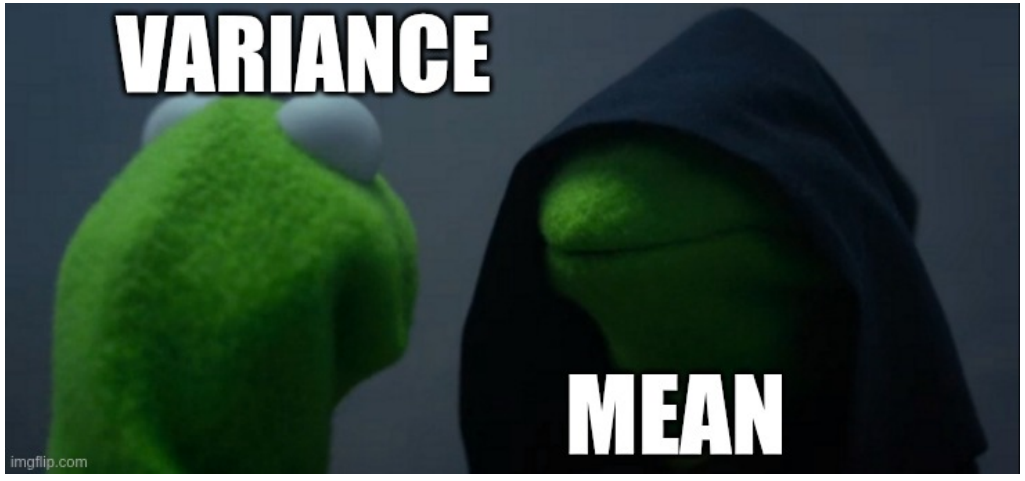
Data of ecological communities has various common properties that tend to get in the way.



Strong mean-variance relationships

Unless your data come from a normal distribution, the variance depends on the mean

- Ecological often have strong mean-variance relationships
- This will much up your results if not accommodated
- For example, overdispersion biases results



Sample size

Field work is hard, takes time, costs money.

- Community ecological studies often have low samples
- And are noisy
- Combined with strong mean-variance relations this causes issues
- Studies are overpowered and lack information
- Drawing conclusions is sometimes not possible
- Can largely be avoided with power analysis

Dimensionality

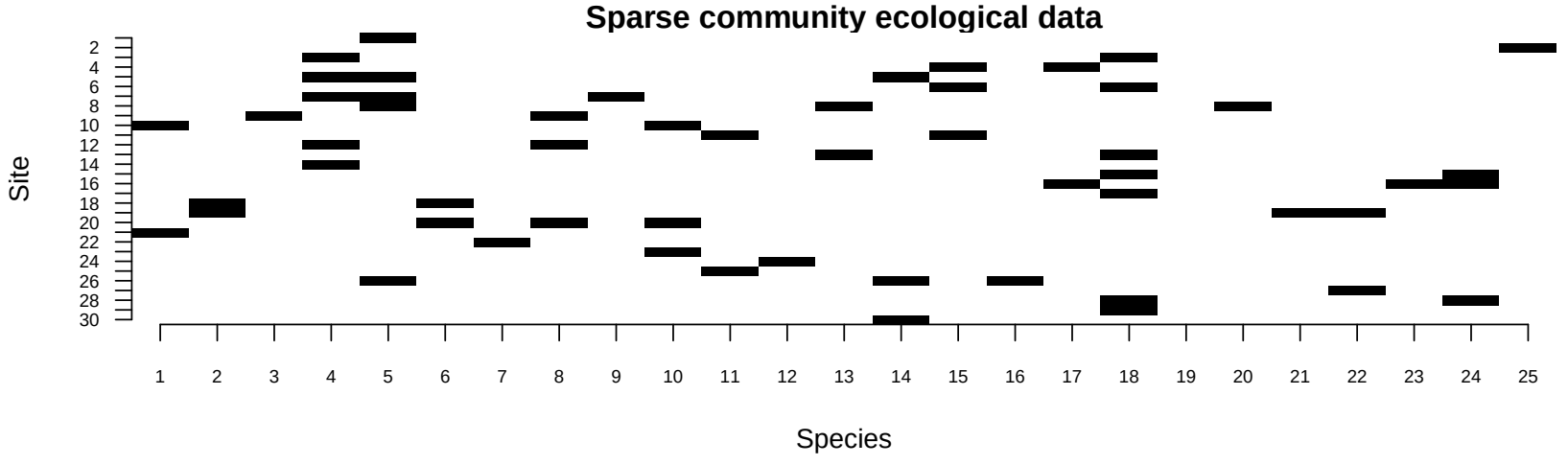
There are often many species in the data; sieving through results is difficult, and analysis can be computationally intensive.

At the same time, data are sparse.

Sparsity: zero-inflation

Zero-inflation is not immediately an issue.

- Zero-inflation is only until after applying an analysis
- Sometimes, covariates can account for it
- For example; sampling an elevation gradient and some plants only occur in the lowlands
- If we cannot explain it, we need to accommodate it



Non-linearity

This is a big one in community ecology.

Compositionality

Some data are inherently composition; magnitude of counts are meaningless. Not addressing it biases results; library depth for example.

- Overdispersion
- Bias
- Inflated Type I error
- Bad ordinations

Microbiome Datasets Are Compositional: And This Is Not Optional

Gregory B. Gloor[1*] Jean M. Macklaim[1] Vera Pawlowsky-Glahn[2] Juan J. Egozcue[3]

[1] Department of Biochemistry, University of Western Ontario, London, ON, Canada
[2] Departments of Computer Science, Applied Mathematics, and Statistics, Universitat de Girona, Girona, Spain
[3] Department of Applied Mathematics, Universitat Politècnica de Catalunya, Barcelona, Spain

Models

Traditional methods of analysis in community ecology are not good at dealing with many of these issues.

Forum | Open Access

The central role of mean-variance relationships in the analysis of multivariate abundance data: a response to Roberts (2017)

David I. Warton, Francis K. C. Hui

First published: 26 July 2017 | https://doi.org/10.1111/2041-210X.12843 | Citations: 28

Models to the rescue

Plant Ecol (2015) 216:669–682
DOI 10.1007/s11258-014-0366-3

Model-based thinking for community ecology

David I. Warton · Scott D. Foster ·
Glenn De'ath · Jakub Stoklosa · Piers K. Dunstan

General attitude

This too I will reiterate a few times:

We adjust the model, not the data

Summary

- Most of these can be accommodated with contemporary methods
- Not all of them in the `gllvm` package
- Choose the appropriate model, not the software you like
- We will explore some in the workshop
- Some are more difficult to accommodate than others
- Sparsity, sample size issues, and misclassification are tough
- Many issues do not show in traditional methods
- Models will be more honest to you

So what do we do with them?

- Detection bias: repeatedly sample and estimate detection probability
- Classification error: difficult to deal with, could estimate confusion
- Measurement error: estimate it
- Nested designs: random effects
- Space, time: spatial/temporal random effects
- Strong mean-variance: residual diagnostics and adjust distribution
- Sample size: regularisation, dimension reduction, simplify model
- Dimensionality: dimension reduction
- Zero-inflation: add covariates or change distribution
- Non-linearity: adjust the model to accommodate
- Compositionality: offset or intercepts

Ecological implications

If you do not address it, don't trust your results nuanced details apply

But, sometimes things are safe to ignore consult an expert

Ecological implications

If you do not address it, don't trust your results nuanced details apply
But, sometimes things are safe to ignore consult an expert

There of course additional ecological considerations for interpretation.

For example:

- Systems not in equilibrium
- High turnover or beta diversity
- Multicollinearity in the environment
- Rare species; how do we deal with them?

And so on.