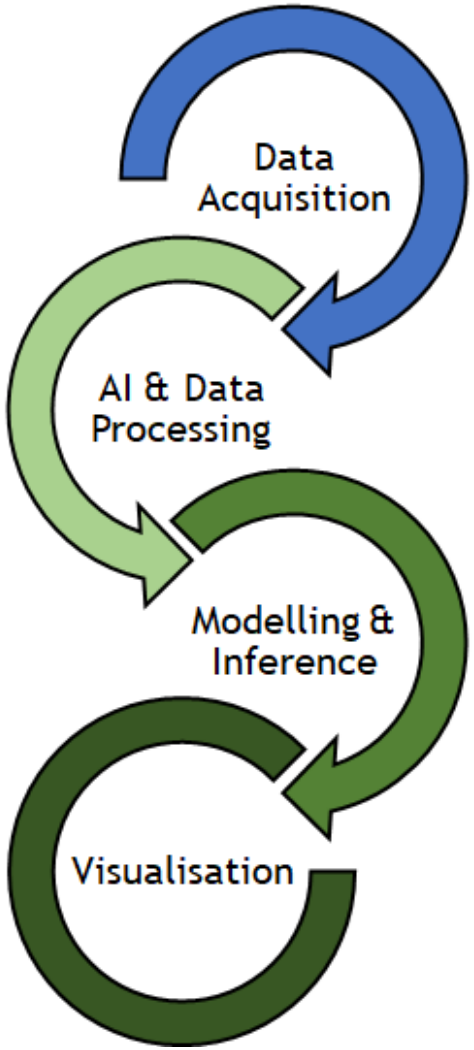
AI For The Environment, week 10

From AI to ecological models 2

Rory Gibb

People & Nature Lab / Centre for Biodiversity and Environment Research, UCL

From data collection to ecological models



BIOS0031 - Tech for Nature

BIOS0032- AI4Environment

Sessions 1-7

Sessions 8-9

Session 10

Week 10:

- Statistical modelling methods in ecology.
- Model selection.
- Autocorrelation in space, time and phylogeny.
- Spatial models.

Learning objectives for week 10

By the end of today you will…

- Understand key principles of statistical modelling in ecology.
- Understand common issues and biases affecting ecological data and how model design can address them.
- Understand the difference between descriptive, explanatory and predictive modelling tasks.
- Be familiar with basic principles of spatial models.

Structure of the day

Lecture 1 AM ~45min

- Linking ecological questions to models
- Widely-used modelling tools for ecological inference and prediction

Lecture 2 PM ~45min

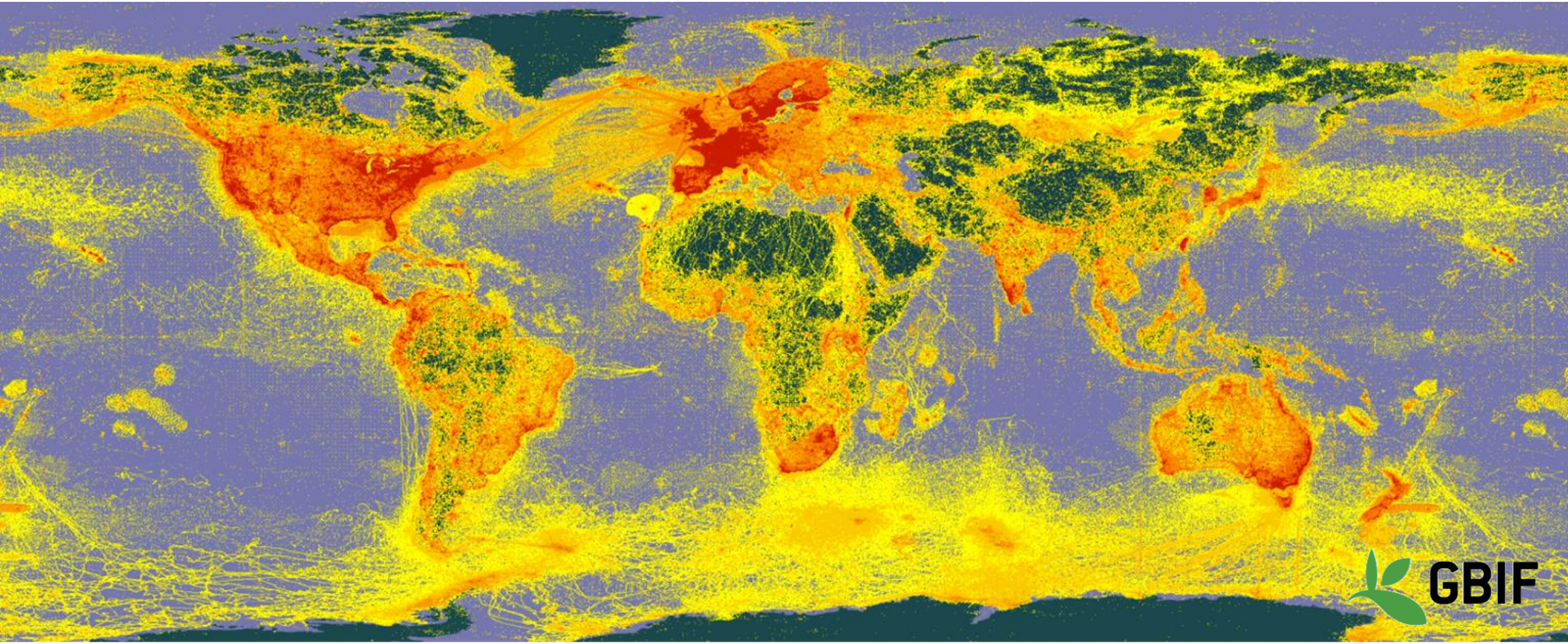
- Model selection and “what is a good model?”
- Autocorrelation and modelling phenomena in space and time

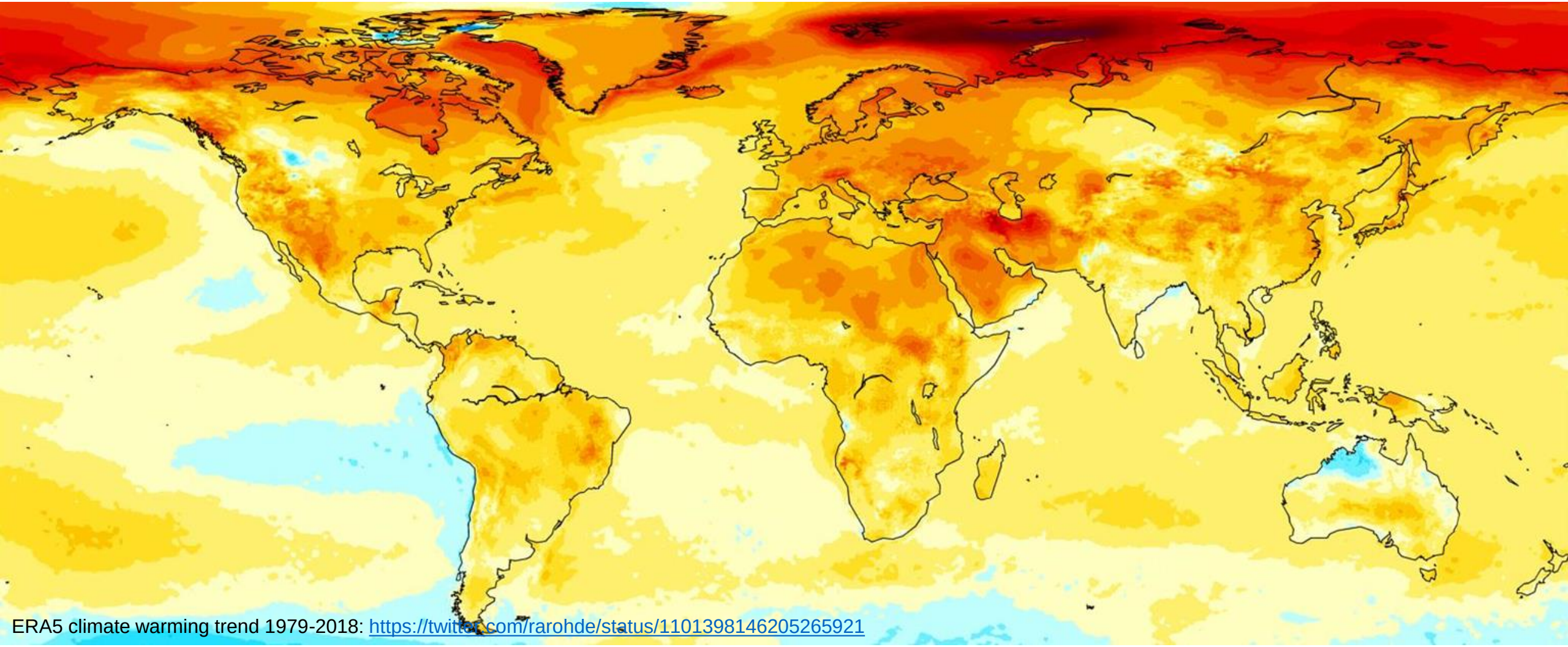
Workshop (all day)

- Analysing the drivers of species occurrence in the Masai Mara using generalized linear and generalized additive models

Morning

Principles of statistical modelling in ecology


GBIF



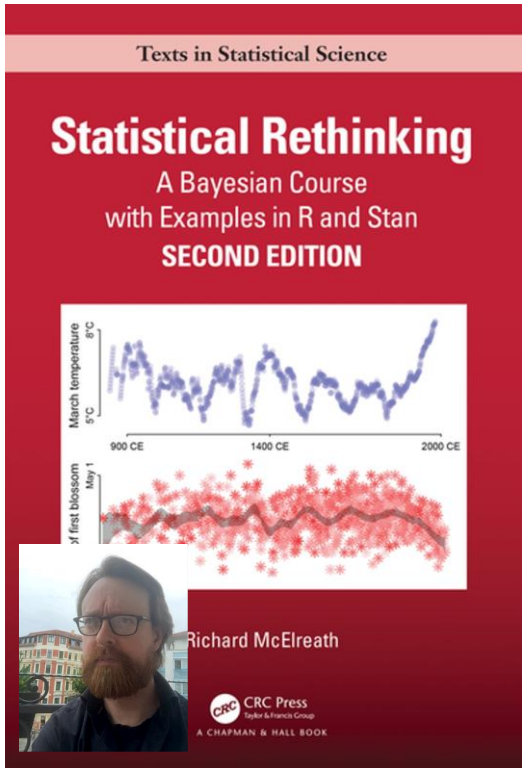
ERA5 climate warming trend 1979-2018: https://twitter.com/rarohde/status/1101398146205265921

How to *think* about modelling analyses

Research question → Data collection → Model design

- There is no “correct way” to model any data - a huge range of general and ecology-specific modelling tools and softwares are out there (many in R and Python)

- **Focus on the scientific question and on how your data were generated**



Richard McElreath -
Statistical Rethinking lectures
and “Science Before Statistics”

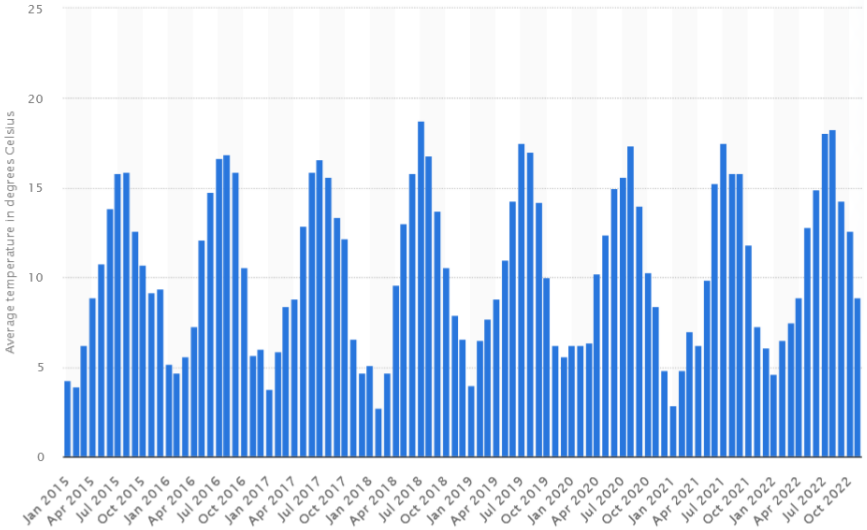
https://xcelab.net/rm/statistical-rethinking/

Three ecological questions

- **Descriptive (exploratory):** *“What winter weather variables are associated with springtime abundance of the Red Admiral butterfly?”*

- **Causal (hypothesis-led):** *“Do warmer temperatures between November-January lead to higher Red Admiral abundance in the following spring?”*

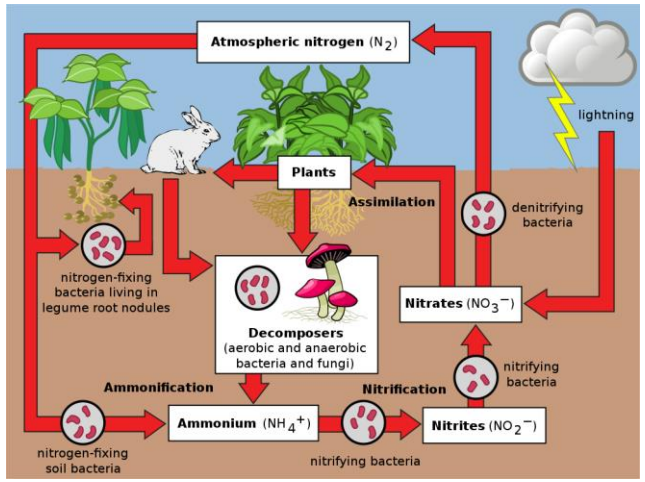
- **Predictive:** *“Can springtime Red Admiral abundance be accurately predicted based on winter weather variables?”*



What is a model?

What is a model?

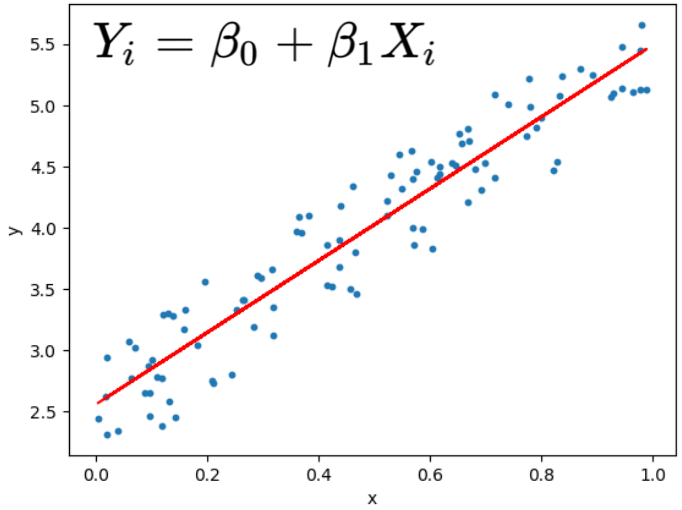
Conceptual



Conceptual representation of known/ hypothesised relationships in a system

Can make qualitative/ quantitative predictions about a system's expected behaviour

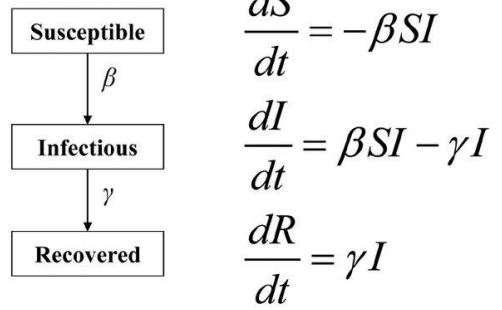
Statistical

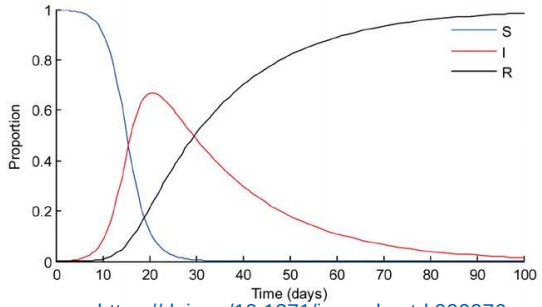


Estimation of parameters of interest through fitting to observed data

Examples: linear regression; GL(M)Ms/GAMs

Mathematical (mechanistic)





https://doi.org/10.1371/journal.pntd.0000761

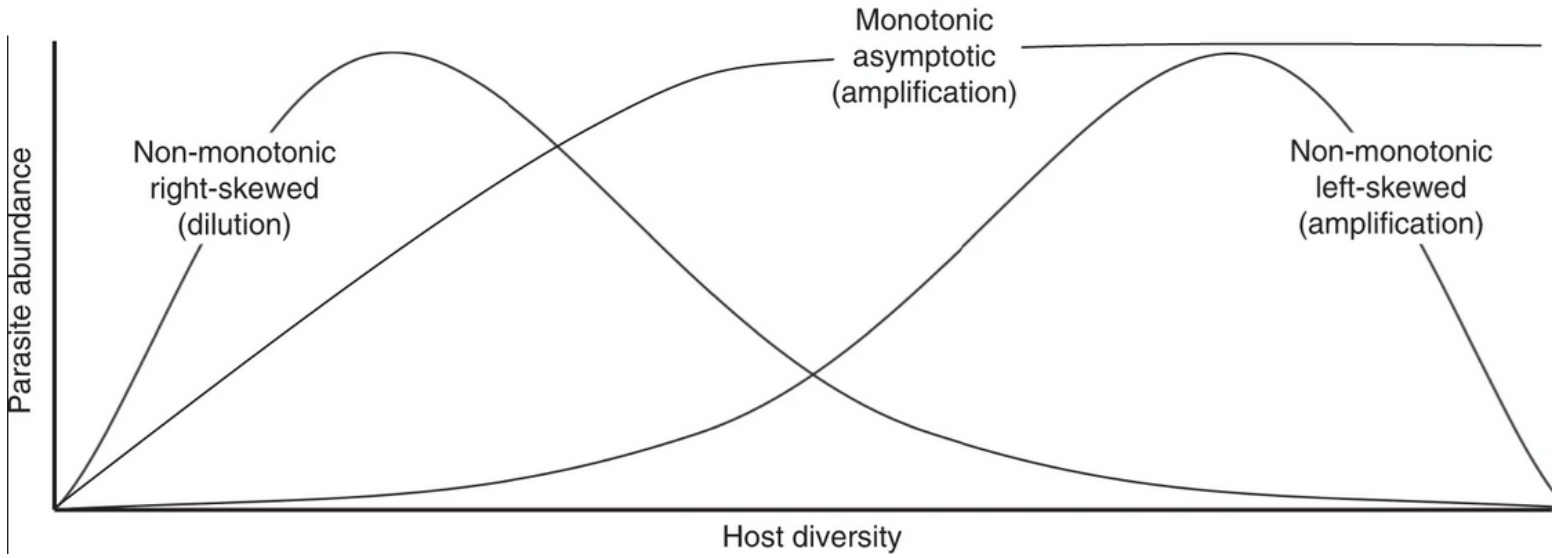
Mathematical description of a system used to explore and predict system behaviour

Examples: SIR model (disease dynamics); Lotka-Volterra (population dynamics)

What is a model?

From conceptual to statistical: the dilution effect

- “Dilution effect” theory proposes that parasite prevalence increases as local species diversity is lost, vs. “amplification effects” which propose the opposite.
  (Rohr et al. 2020, Nat Eco Evo)
- Either might be possible depending on the system, and results have often been conflicting.
- Conceptual model outlines 3 hypotheses for the nonlinear shape of this relationship - these can be challenged with data!



https://doi.org/10.1038/s41467-019-13049-w OPEN

Measuring the shape of the biodiversity-disease relationship across systems reveals new findings and key gaps

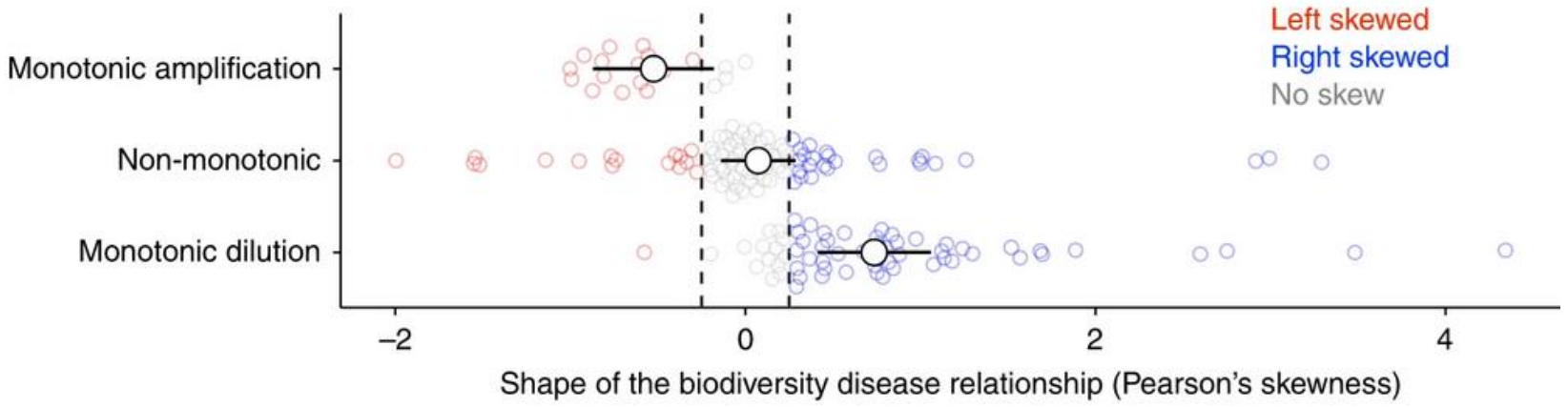
Fletcher W. Halliday[1*] & Jason R. Rohr[2]

nature COMMUNICATIONS

From conceptual to statistical: the dilution effect

- Meta-analysis of the shape of the diversity-disease relationship across 205 studies - do real-world observations follow the predictions of the conceptual model?

- Systems with dilution effects are significantly right skewed, and vice versa, as predicted by the conceptual model - the analysis estimates *parameters* to describe this shape.



https://doi.org/10.1038/s41467-019-13049-w OPEN

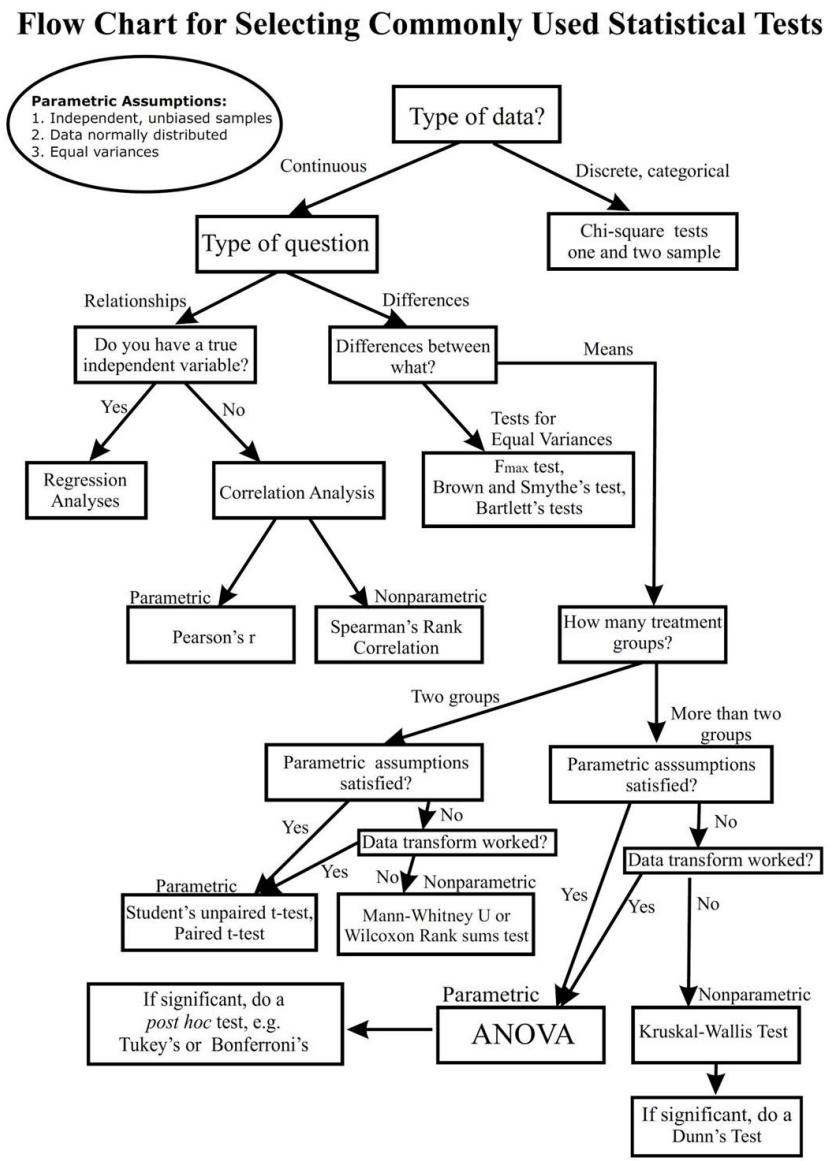
Measuring the shape of the biodiversity-disease relationship across systems reveals new findings and key gaps

Fletcher W. Halliday[1*] & Jason R. Rohr[2]

nature COMMUNICATIONS

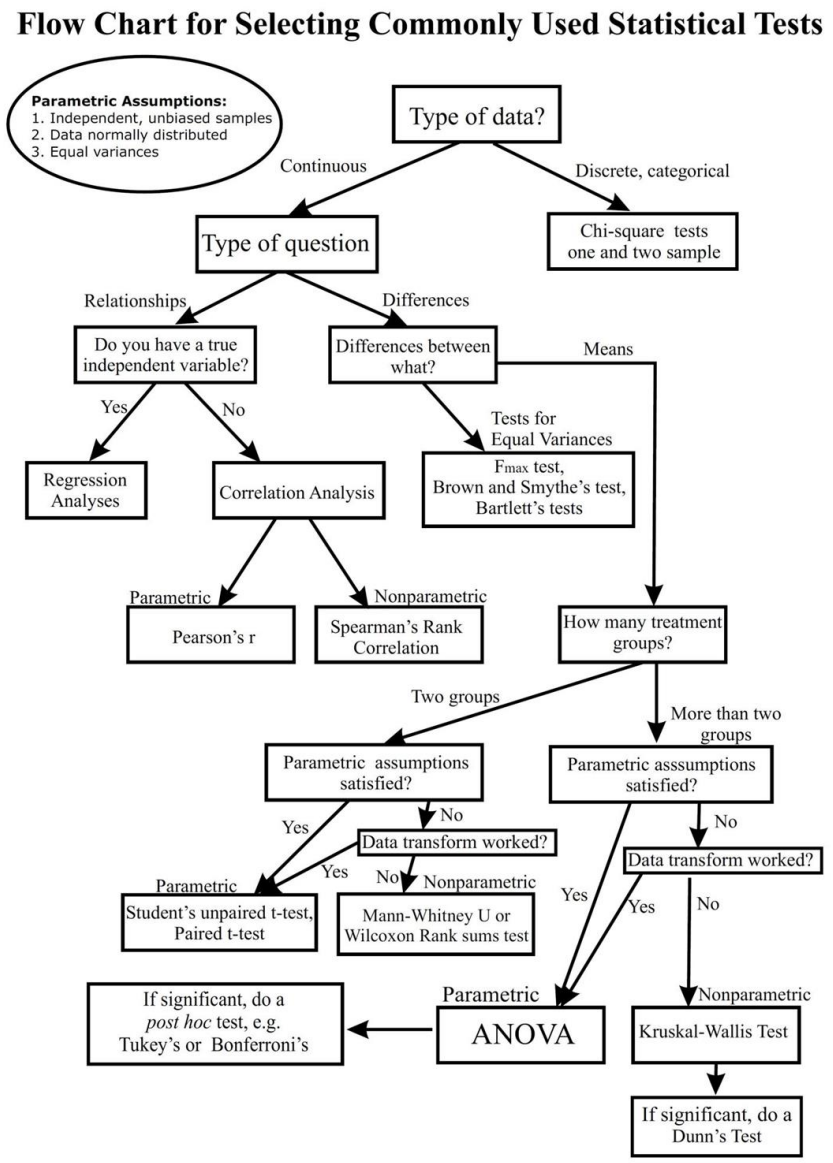
What is our conceptual model of the world?

The ‘cookbook’ approach to teaching statistics



https://twitter.com/TrevorABranch/status/1082135735757504512

The ‘cookbook’ approach to teaching statistics



This approach ignores the most important things that help us ask the right question!

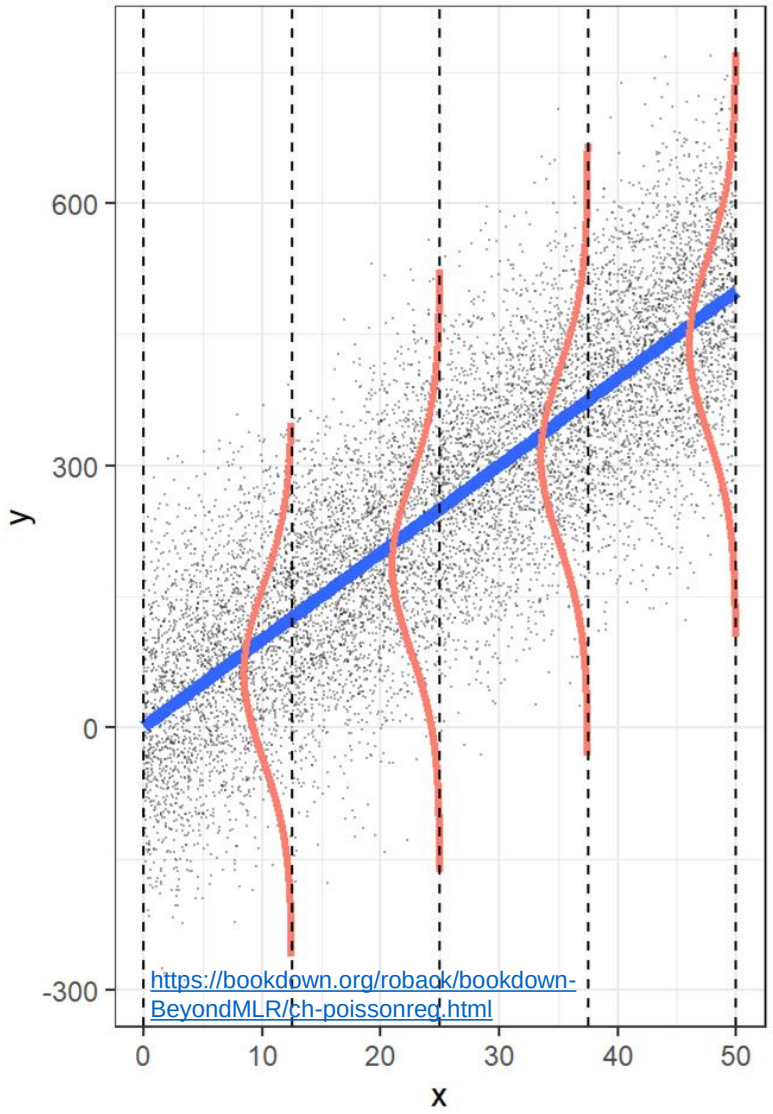
(1) Our scientific understanding of the system (the specific conceptual model)

(2) How the data were generated

https://twitter.com/TrevorABranch/status/1082135735757504512

Building blocks of a statistical model

Linear regression



https://bookdown.org/robaok/bookdown-BeyondMLR/ch-poissonreg.html

General formula for linear regression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$$

Intercept

Slope(s) for covariates X

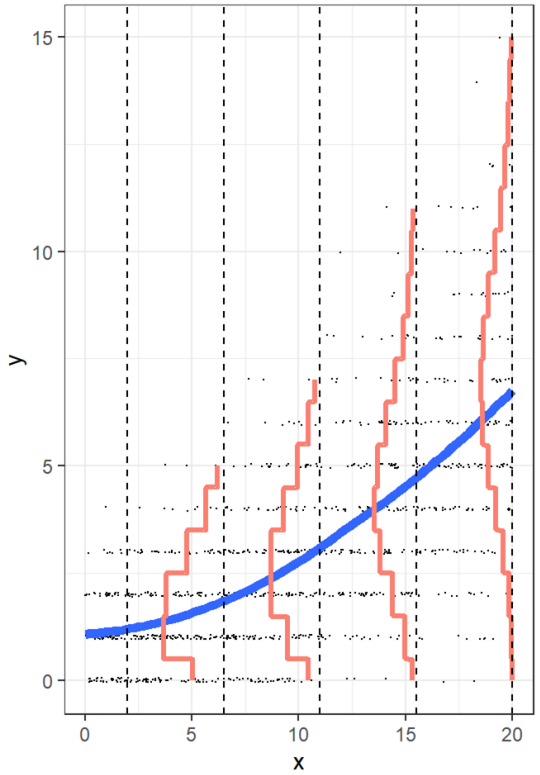
Residual error - *assumed independent, normally distributed and homoscedastic conditional on the model*

Generalised linear models

- Likelihood and link function relax assumptions around linearity and error distribution
- Still assume observations are independent conditional on the model!
- **Poisson regression**: count data, Poisson likelihood *(error variance equal to expected value)*, log link function *(assume exponential relationship between X and Y)*

Poisson

(e.g. butterfly counts)



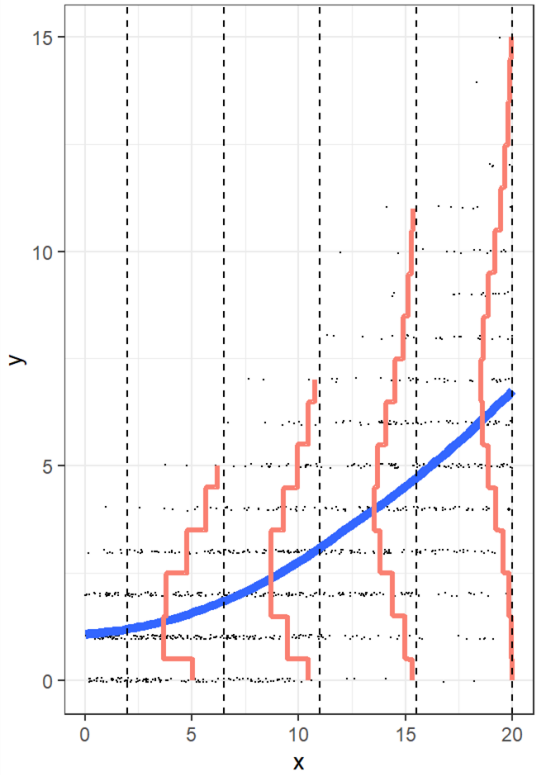
Likelihood function
(describes error distribution) → $Y_i \sim Pois(\lambda_i)$

Link function
(describes shape of relationship between X and Y) → $\log(\lambda_i) = \beta_0 + \beta_1 X_{i1}$

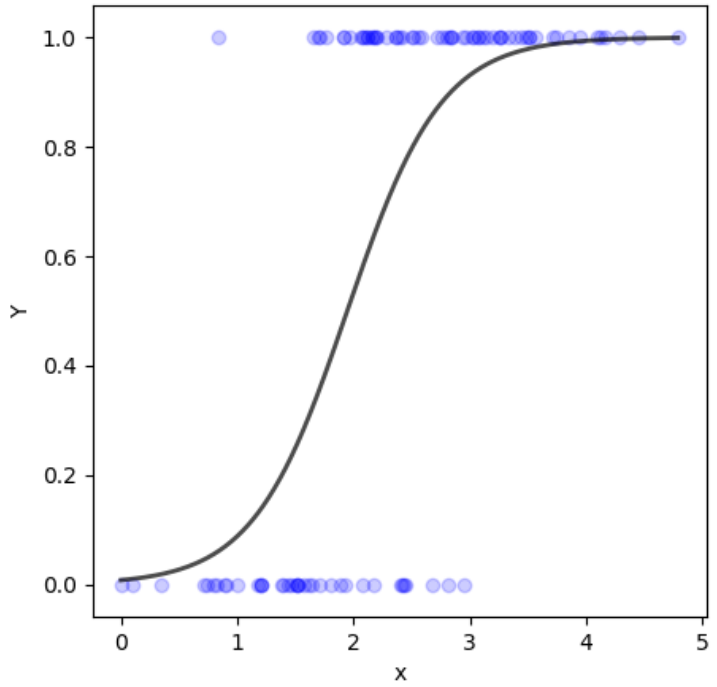
Generalised linear models

- Likelihood and link function relax assumptions around linearity and error distribution
- Still assume observations are independent conditional on the model!
- **Poisson regression**: count data, Poisson likelihood *(error variance equal to expected value)*, log link function *(assume exponential relationship between X and Y)*
- **Logistic regression**: binary outcome (1/0), binomial likelihood *(probability of success)*, logit link *(assume linear relationship between X and log odds of Y)*

Poisson
(e.g. butterfly counts)



Logistic (binomial)
(e.g. hatching success)



Likelihood function
(describes error distribution) →

$$Y_i \sim Pois(\lambda_i)$$

Link function
(describes shape of relationship between X and Y) →

$$\log(\lambda_i) = \beta_0 + \beta_1 X_{i1}$$

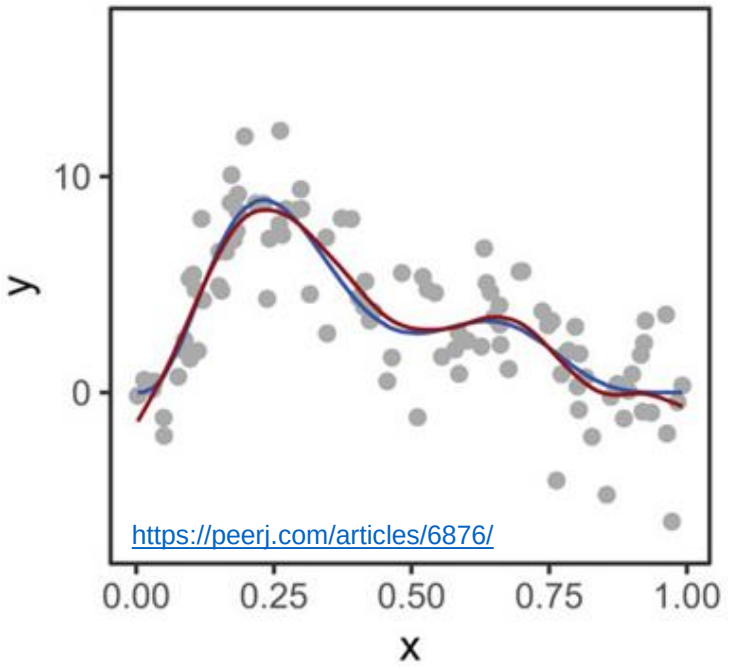
$$Y_i \sim Binom(p_i)$$

$$\log(\frac{p_i}{1 - p_i}) = \beta_0 + \beta_1 X_{i1}$$

Extensions to nonlinearity, multilevel data, space and time

Nonlinear

(e.g. generalised additive models)



$$Y_i = f(\eta_i) = \beta_0 + f(X_{i1})$$

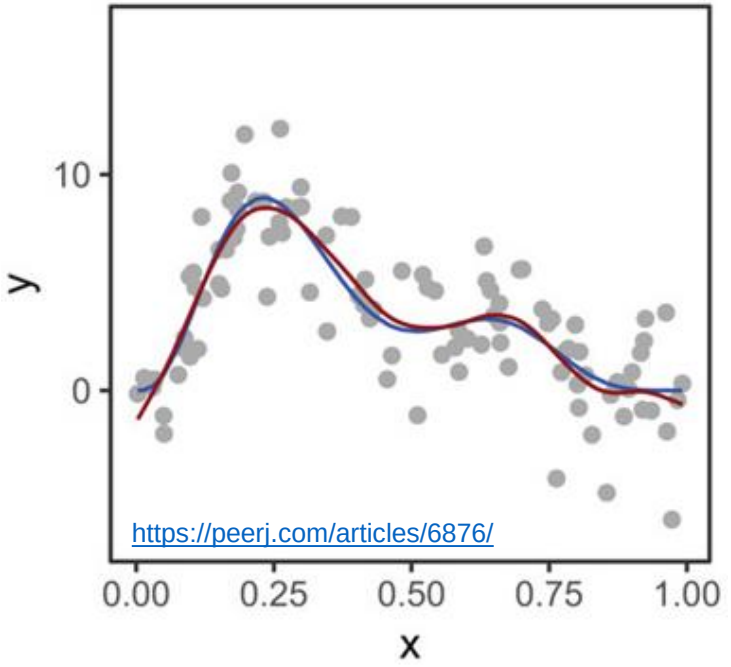
Nonlinear function of covariates X

(e.g. penalised splines)

Extensions to nonlinearity, multilevel data, space and time

Nonlinear

(e.g. generalised additive models)

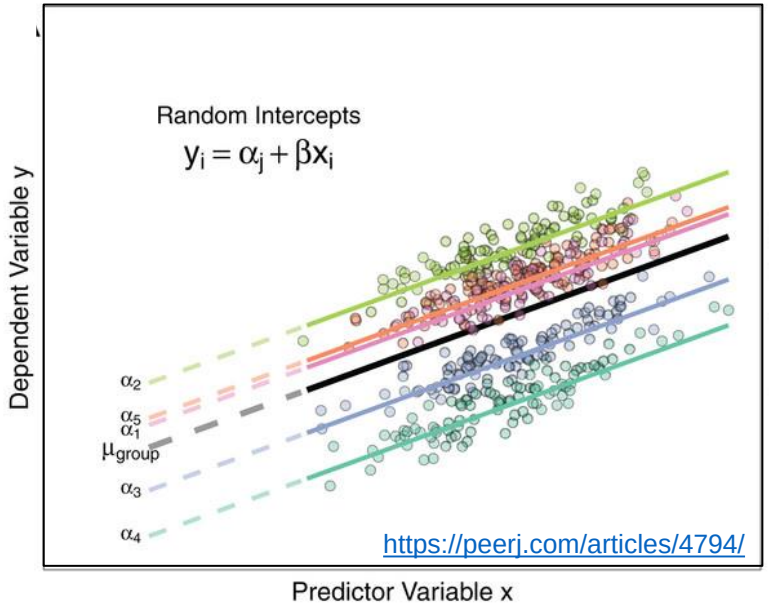


$$Y_i = f(\eta_i) = \beta_0 + f(X_{i1})$$

Nonlinear function of covariates X
(e.g. penalised splines)

Multilevel (mixed effects)

(e.g. GLMMs; random slopes or intercepts)



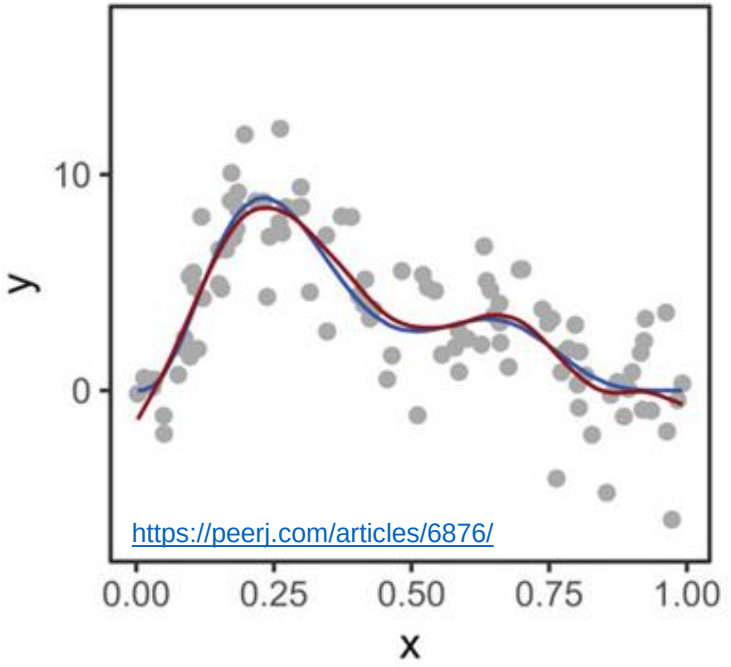
$$Y_i = f(\eta_i) = \beta_0 + \beta_1 X_{i1} + \alpha_{s(i)}$$

Random intercept for study site *S*
(intercepts vary to account for hierarchical structure; pools information across sites)

Extensions to nonlinearity, multilevel data, space and time

Nonlinear

(e.g. generalised additive models)

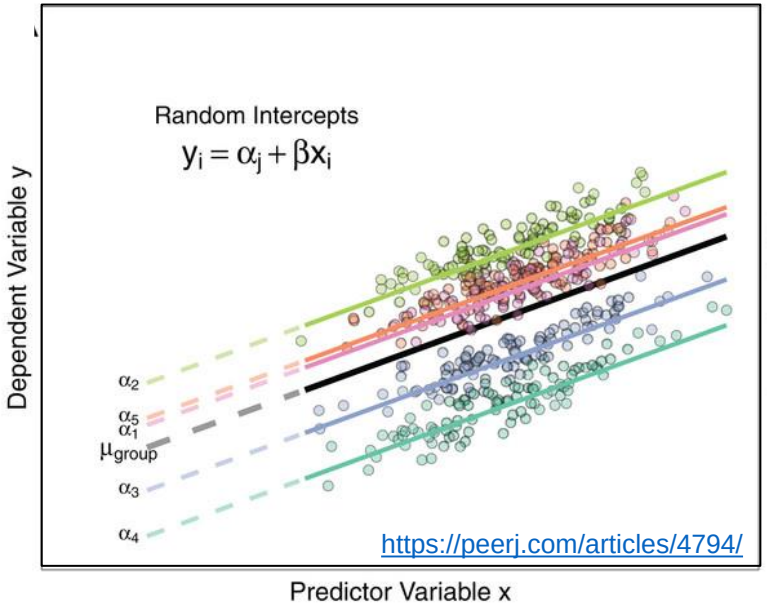


$$Y_i = f(\eta_i) = \beta_0 + f(X_{i1})$$

Nonlinear function of covariates X
(e.g. penalised splines)

Multilevel (mixed effects)

(e.g. GLMMs; random slopes or intercepts)

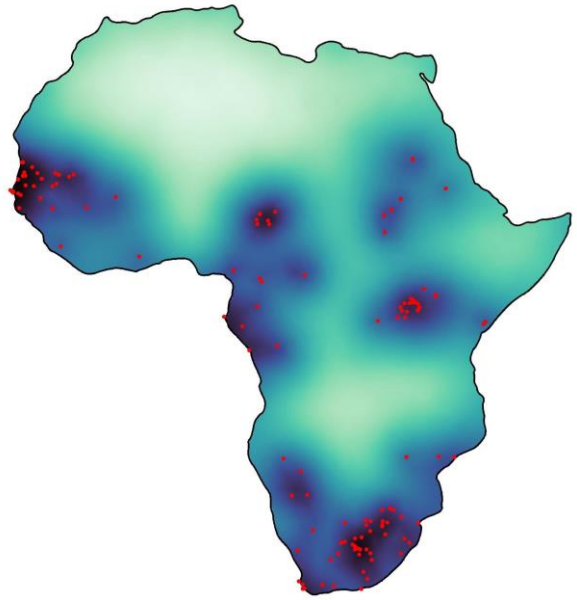


$$Y_i = f(\eta_i) = \beta_0 + \beta_1 X_{i1} + \alpha_{s(i)}$$

Random intercept for study site *S*
(intercepts vary to account for hierarchical structure; pools information across sites)

Spatial/temporal

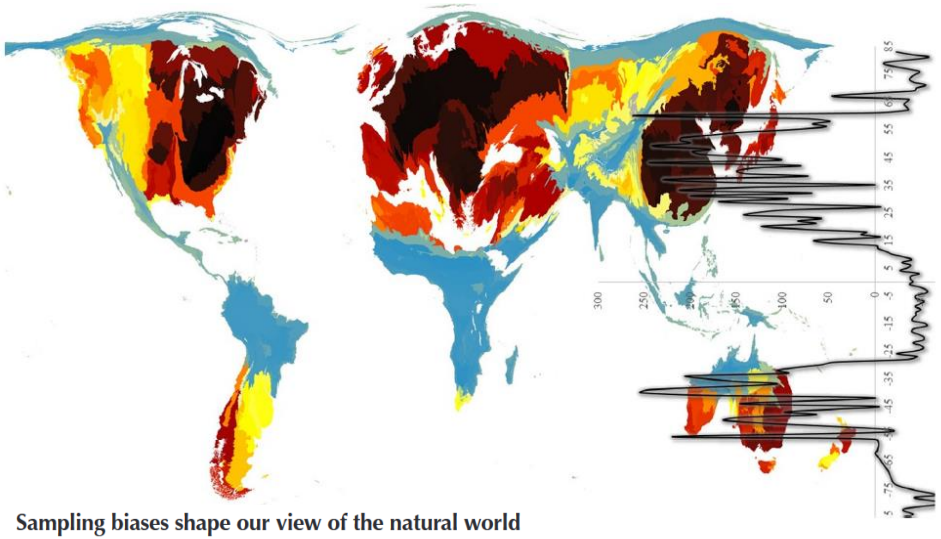
(e.g. conditional autoregressive; ARIMA; Gaussian processes)



$$Y_i = f(\eta_i) = \beta_0 + \beta_1 X_{i1} + s_i$$

Spatially or temporally-structured effect
(observations that are closer together are more closely related)

Real worlds are noisy (and so are data)



Sampling biases shape our view of the natural world

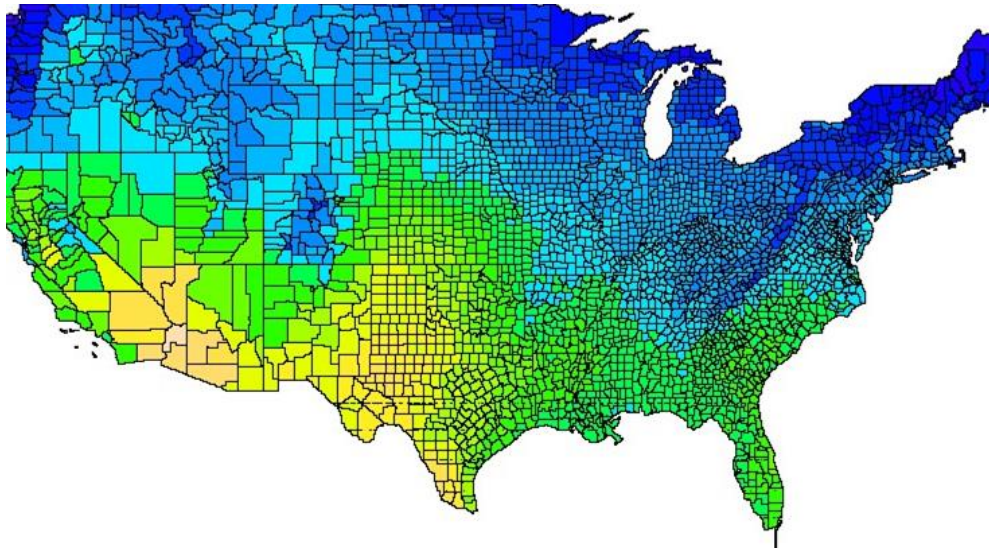
Alice C. Hughes, Michael C. Orr, Keping Ma, Mark J. Costello, John Waller, Pieter Provoost, Qinmin Yang, Chaodong Zhu and Huijie Qiao

Sampling biases shape our understanding

(historical and evolving biases in survey effort confound our knowledge of pattern and process in ecology)

Dependency in space and time

(observations closer together are more closely related)

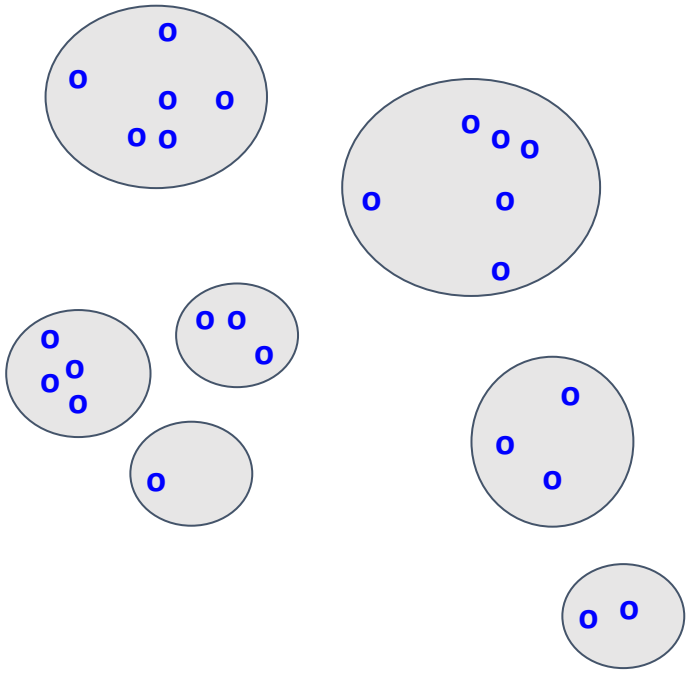


Observation is imperfect

(ecological data often contain a mixture of true and false zeroes, and we often do not measure key variables acting on the system)

Mixed-effects (multilevel/hierarchical) models

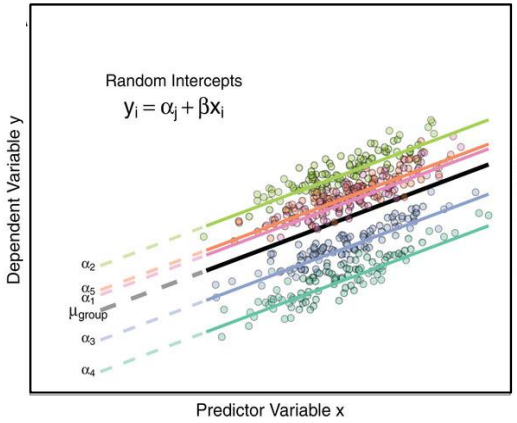
Ecological study of bird abundance across an island archipelago



O = Abundance record of bird species X
Sampled across 7 islands *s* (s = 1…7)

- Data often contain some clustered/nested structure (e.g. space, time, phylogeny, individual, population)
- **Observations within each cluster may be non-independent -** account for this by allowing intercepts/slopes to vary across levels (*"random effects"*)

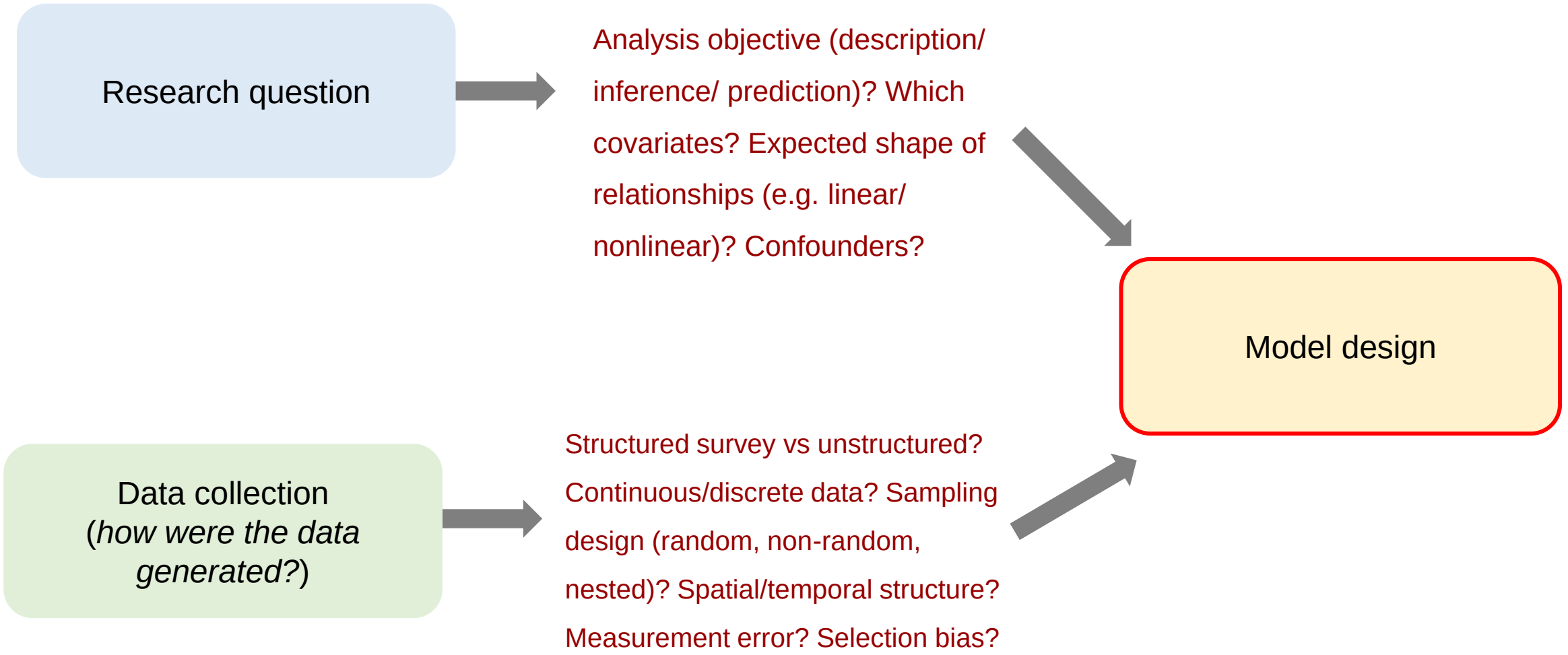
$$Y_i \sim Pois(\lambda_i)$$
$$\log(\lambda_i) = \beta_0 + \beta_1 X_1 + \alpha_s$$
$$\alpha_s \sim N(0, \sigma)$$



Island-level intercepts are modelled as a population described by a normal distribution with variance σ (how variable between islands?)

A model within a model - hence "hierarchical"!
(σ is a '*hyperparameter*')

From question, to data, to model design



Using scientific understanding to inform model design

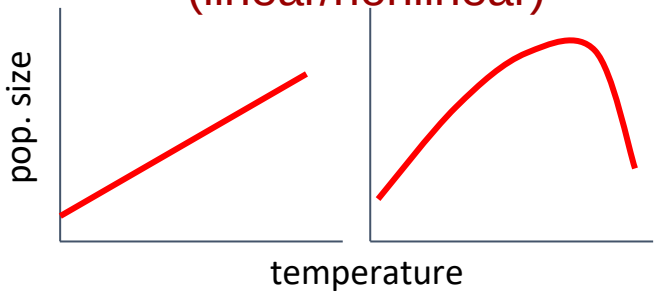
"What winter weather variables are associated with springtime abundance of the Red Admiral butterfly?"

What variables?

(temperature, precip, humidity…?)

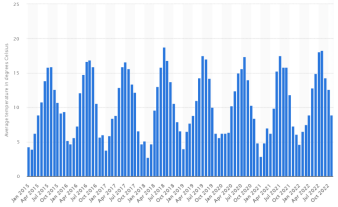
What shape of relationships do we expect?

(linear/nonlinear)



What timescale of climate - multi-month averages?

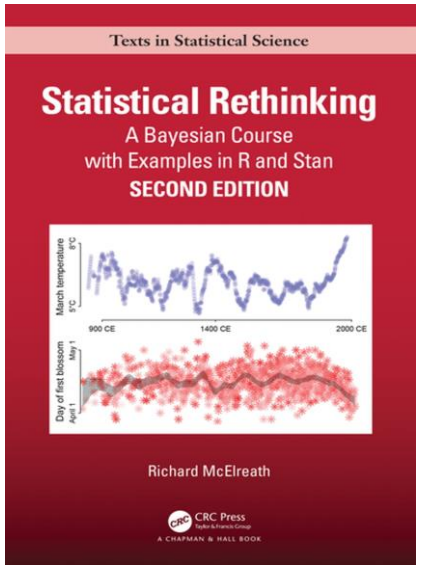
Transient extremes?



What other factors might be impacting abundance, or interacting with weather?

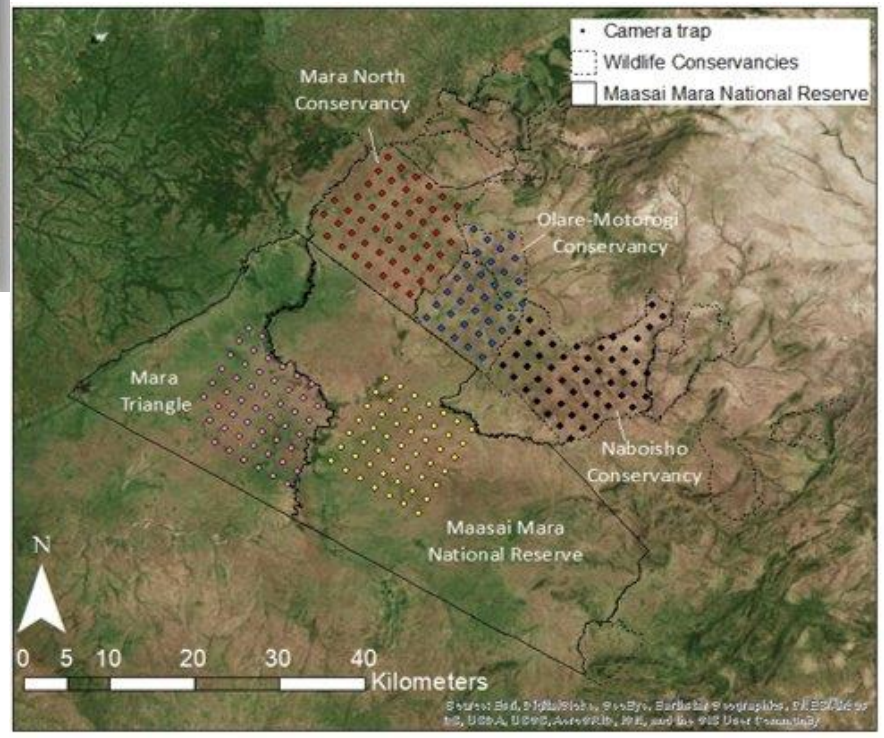
Models are powerful but unaware

- Statistical models are powerful machines for helping us to understand and predict the world, but lack insight.
- The model only knows what you tell it!



https://xcelab.net/rm/statistical-rethinking/

Workshop: effects of livestock on hare occurrence



Cape hare (*Lepus capensis*)

https://www.inaturalist.org/guide_taxa/271513

Afternoon

Ecological inference and prediction in space and time

- Model selection and evaluation: What is a “good” model?
- Autocorrelation
- Brief introduction to spatial models

Model selection and evaluation

What is a “good” model?

- Usually there are multiple potential models we could fit to a given dataset, e.g.
    - Representing alternative hypotheses
    - Different combinations of covariates (*“should covariate X be included?”*)

- **How do we select the “best” or most appropriate model to fit to the data?**

- A major issue for observational studies, where we may need include covariates to statistically adjust for other factors acting on the system.

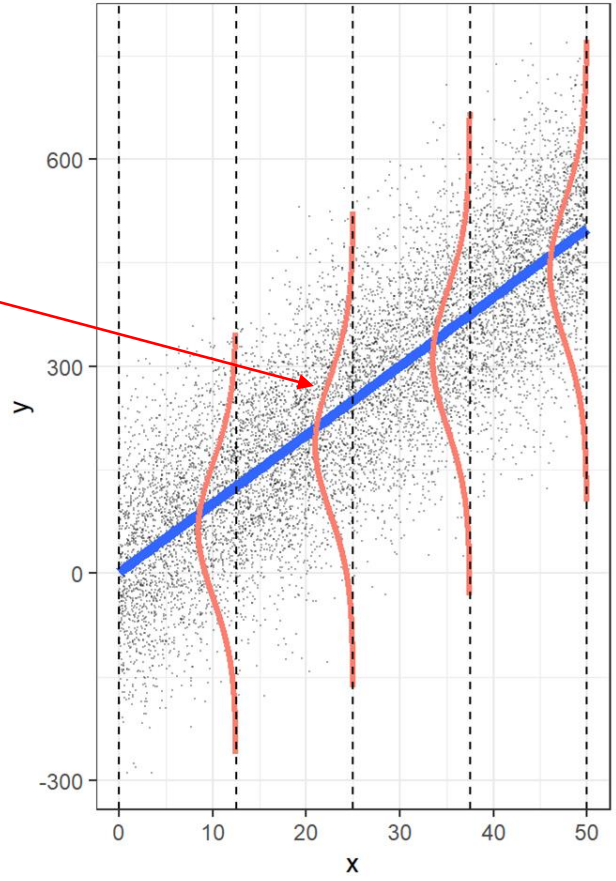
M1: *Abundance ~ Temperature*
M2: *Abundance ~ Rainfall*
M3: *Abundance ~ Temperature + Rainfall*

?

Model selection

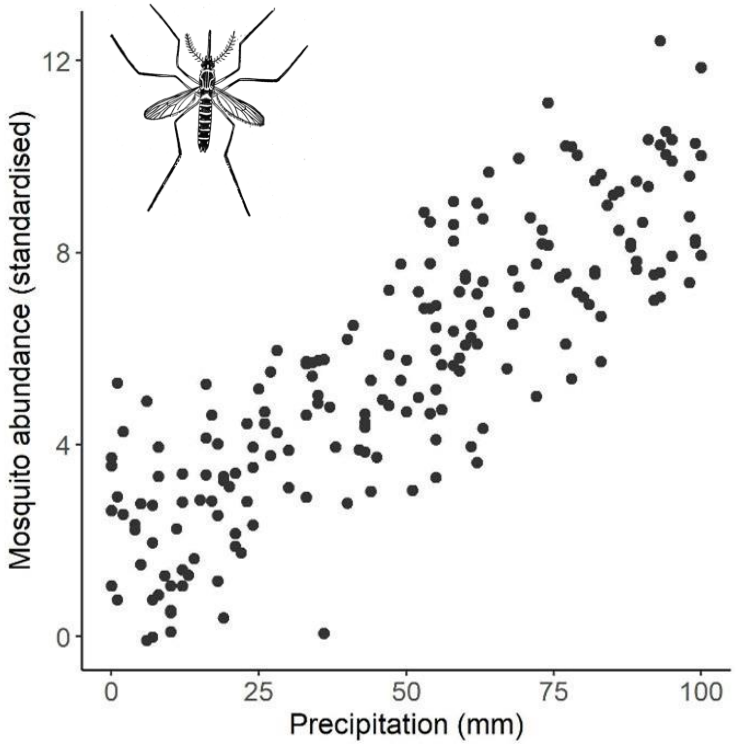
One simple definition: **the better model explains more of the observed variation in the data.**

(i.e. reduces residual error)

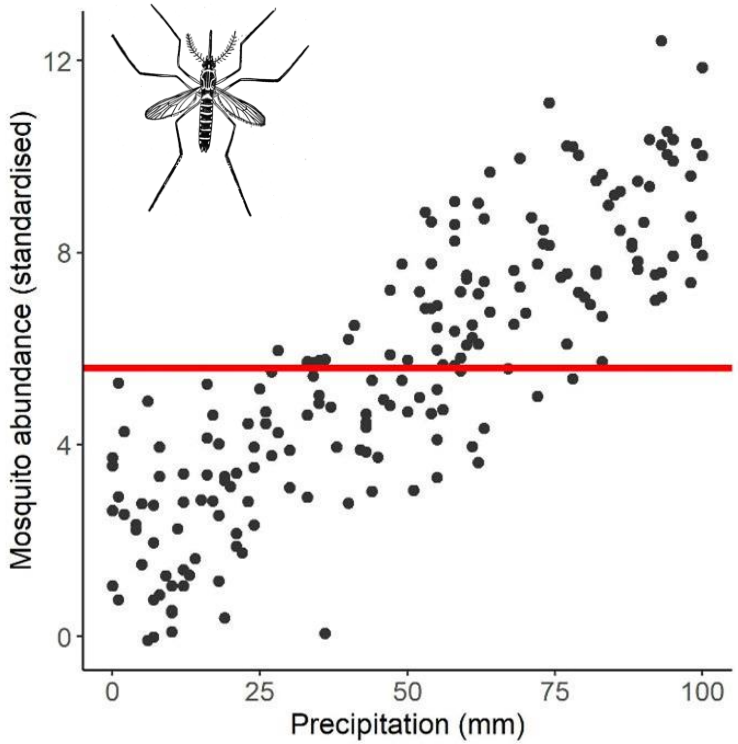


The basis of **goodness-of-fit** statistics such as **likelihood ratio tests**, **R^2** and **deviance** – all compare how well the data are explained by candidate models.

Model selection



Model selection

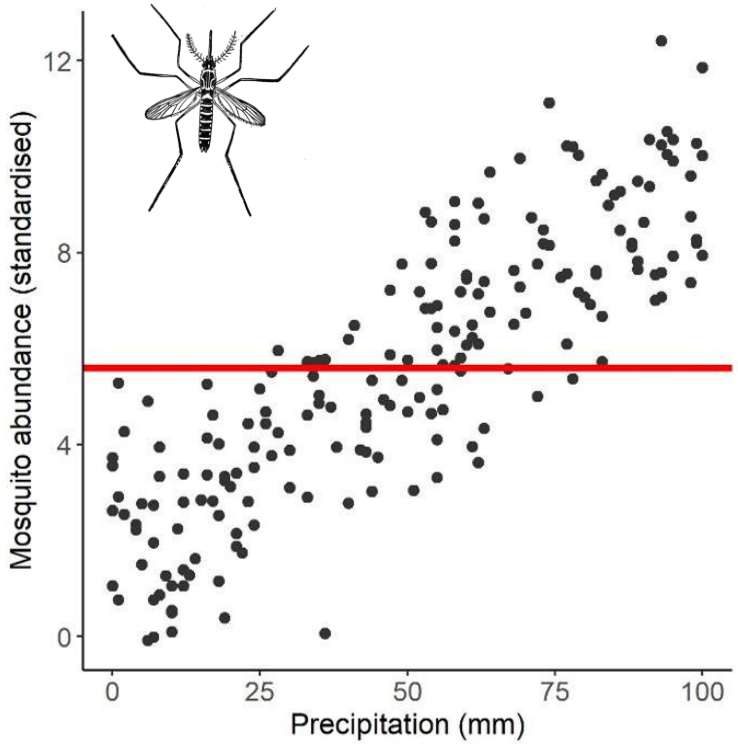


Simplest possible model

('null' model)

- 1 parameter (intercept)
- Maximum residual error
- Learning very little from the data so not generalisable

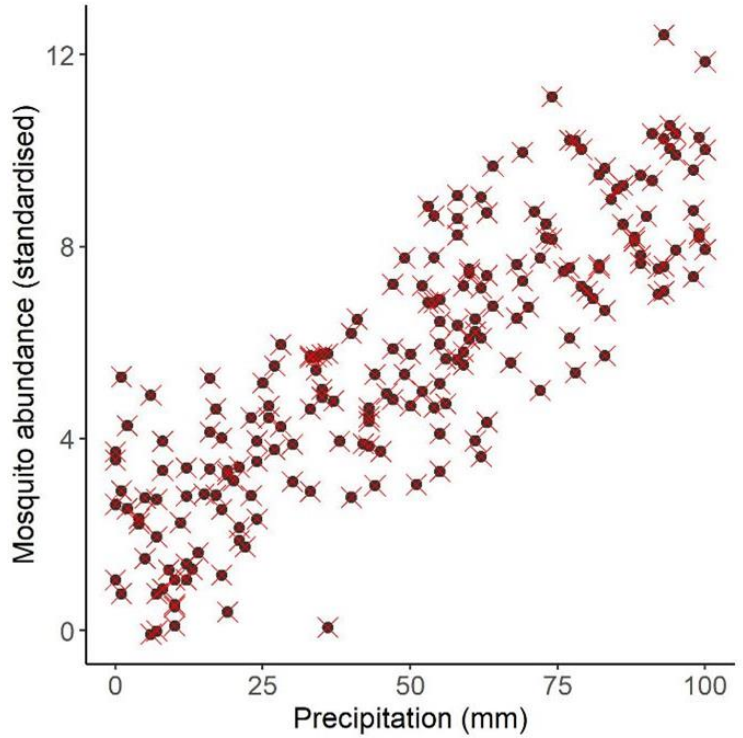
Model selection



Simplest possible model

('null' model)

- 1 parameter (intercept)
- Maximum residual error
- Learning very little from the data so not generalisable

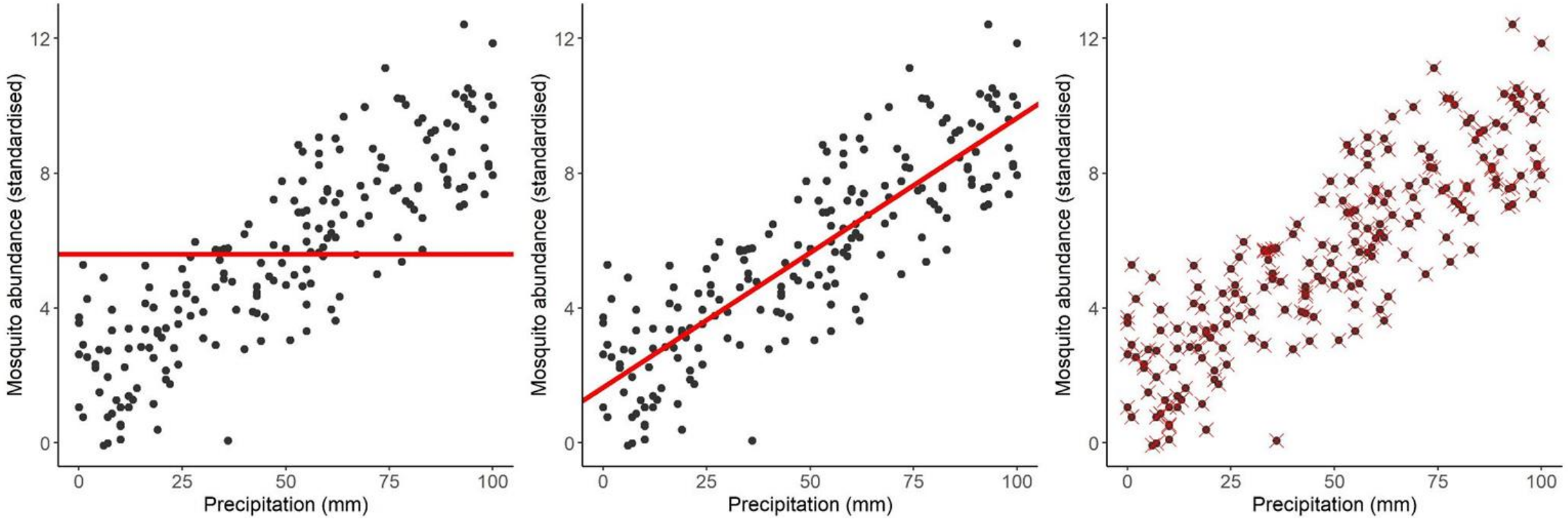


Most complex model

('saturated model')

- 1 parameter per observation
- Maximises likelihood (no residual error)
- Learning too much from this specific dataset - not generalisable

Model selection



Underfitting ←

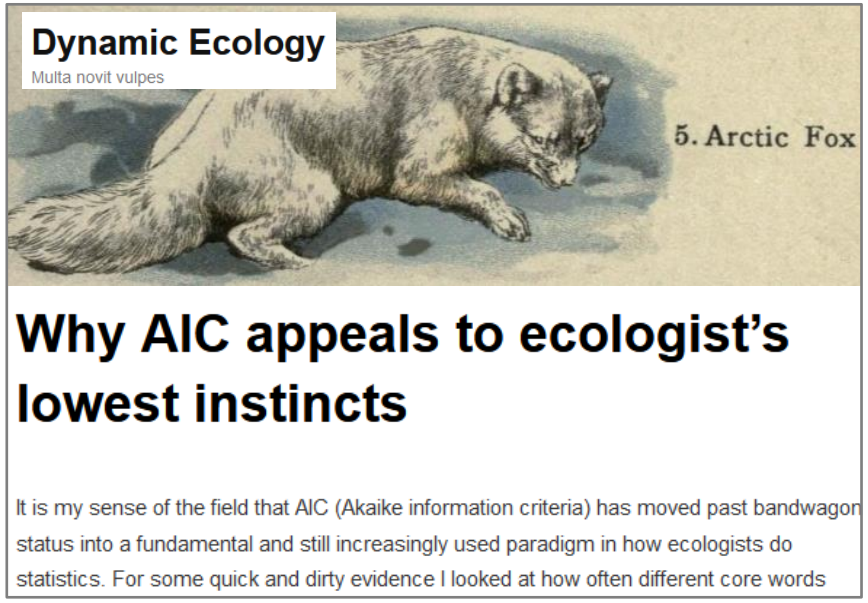
‘Goldilocks’ model
(not too simple, not too complex)

- Learned enough from the data for (potentially) general and transferable inference

→ **Overfitting**

Information criteria

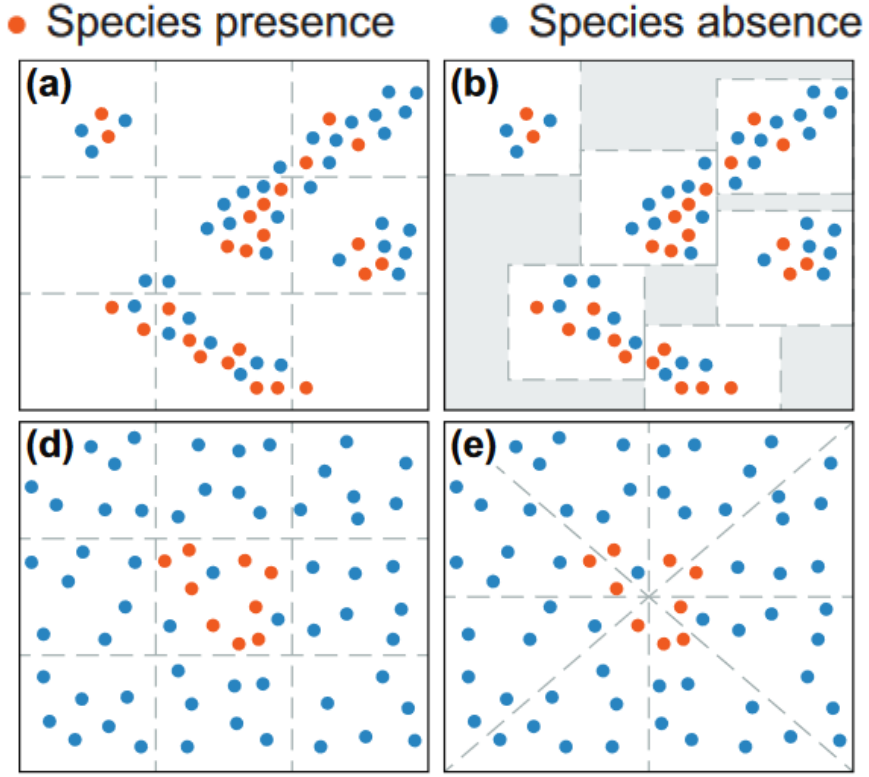
- A suite of criteria designed to balance this trade-off by penalising more complex models.
- Most well-known is Akaike Information Criterion.
- **AIC = 2k – 2ln(L)**
  where *k* is the number of parameters in the model, and *L* is the model likelihood – lower values indicate a better model.
- Numerous others: DIC, BIC, WAIC…
- Not a silver bullet!



https://dynamicecology.wordpress.com/2015/05/21/why-aic-appeals-to-ecologists-lowest-instincts/

Out-of-sample evaluation approaches

- Ecologists seek generalisable explanations for ecological phenomena
- A good model is one that accurately predicts **out-of-sample** (unobserved) data
  (doesn't just shrinkwrap to training data)
- Gold standard: challenge the model by predicting to an independent test dataset.
- Cross-validation: divide the study dataset into *k* folds and estimate predictive accuracy across all train-test splits
  (train on 80%, predict on 20%)

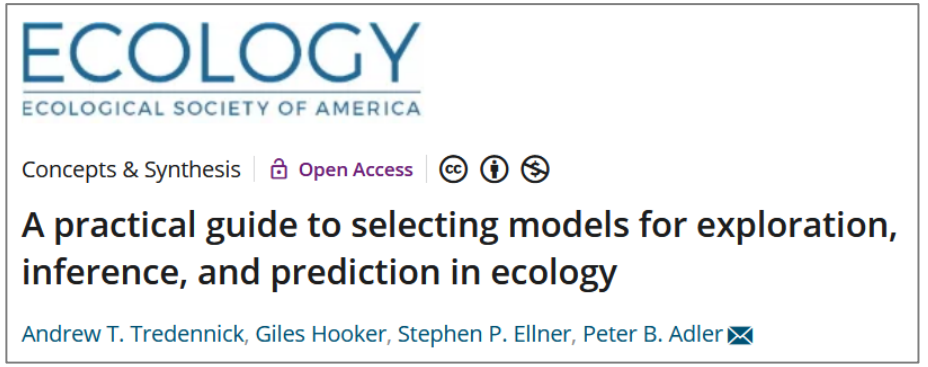


Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure

David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta G

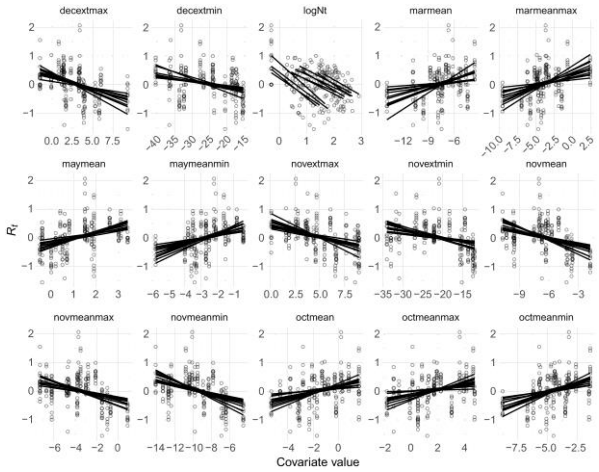
Model selection depends on the study objective

“Confusion about how to do model selection is confusion about how to do ecology” (Tredennick et al. 2021)



Model selection depends on the study objective

Description

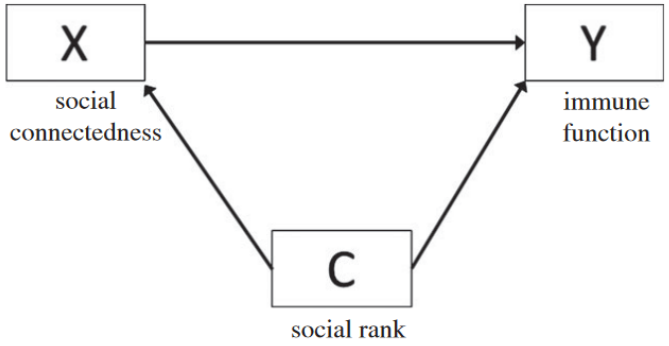


Objective: describing patterns and relationships to generate hypotheses.

Best model(s) consider all variables of interest that might influence the outcome.

Examples: *many (most?) ecological studies!*

Explanation (causality)



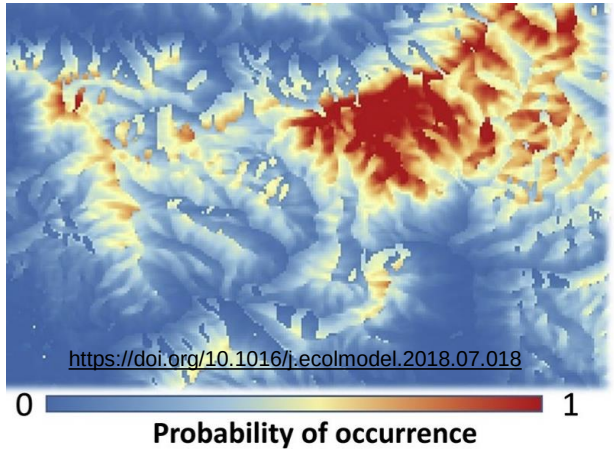
https://doi.org/10.1098/rspb.2020.2815

Objective: answering a causal question (*"does X cause Y?"*)

Best model includes the covariates needed to achieve an unbiased estimate of X -> Y

Examples: *hypothesis testing; causal inference*

Prediction



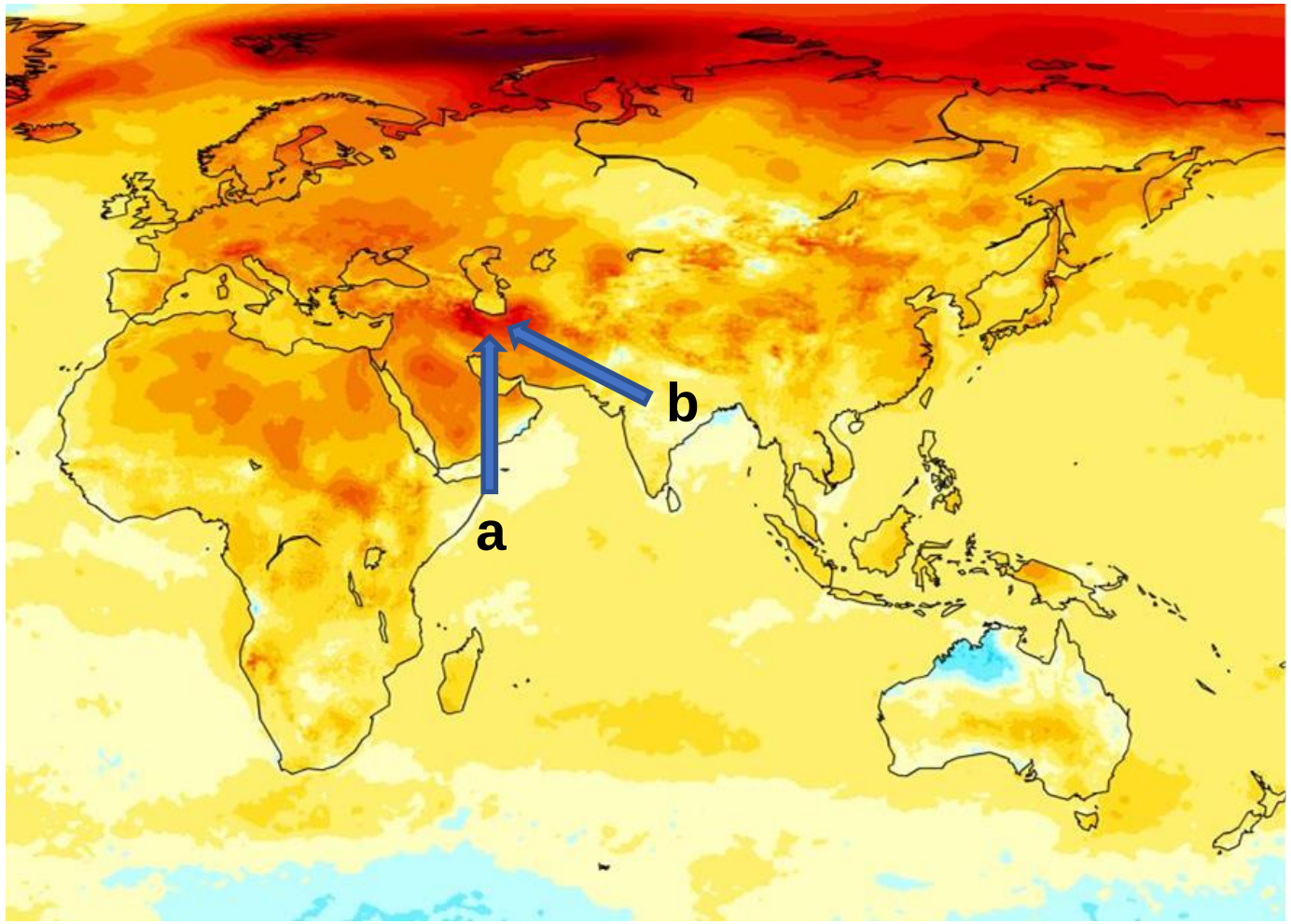
Objective: predicting phenomena across space, time or phylogeny

Best model(s) minimizes out-of-sample prediction error

Examples: *species distribution models, forecasting, climate change projections, imputation, mapping*.

Modelling ecological phenomena in space and time

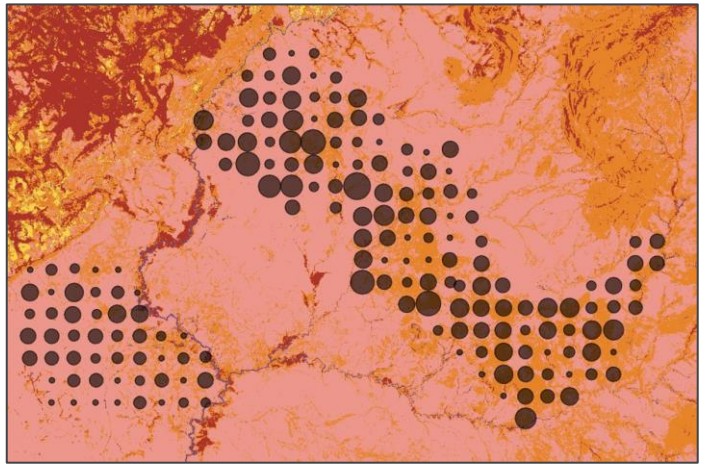
What is autocorrelation?



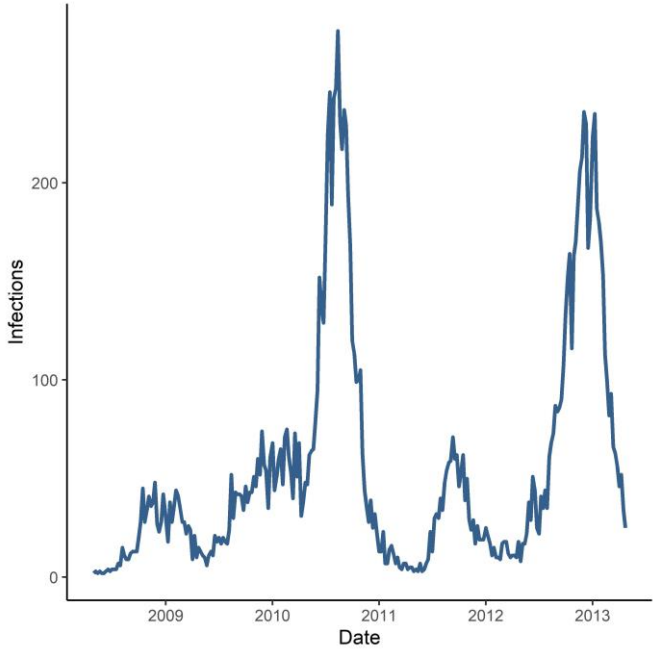
Observations nearby to one another (in space or time) are **non-independent**

-> they are realisations of the same underlying (unobserved) processes

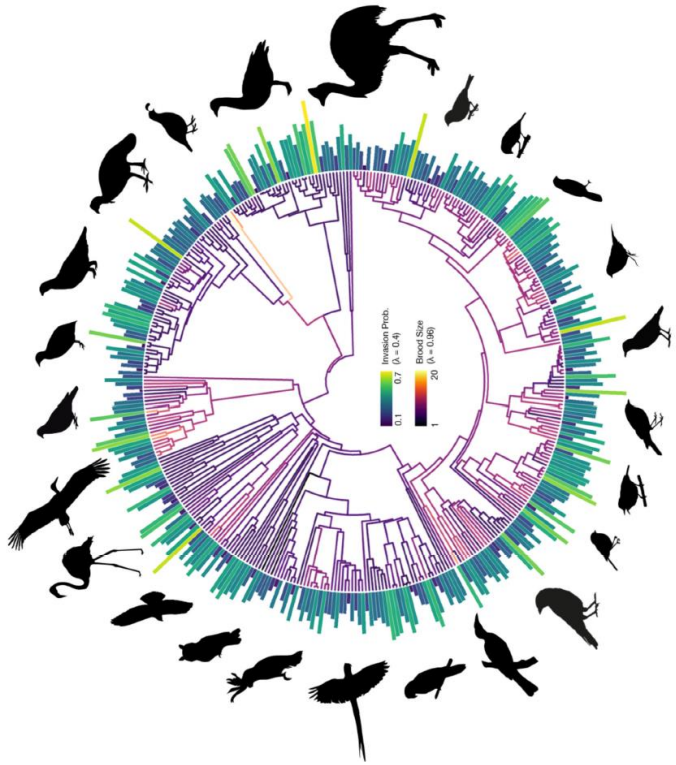
Autocorrelation over space, time and phylogeny



Nearby locations in **space** share common environmental features and are linked by organism movement



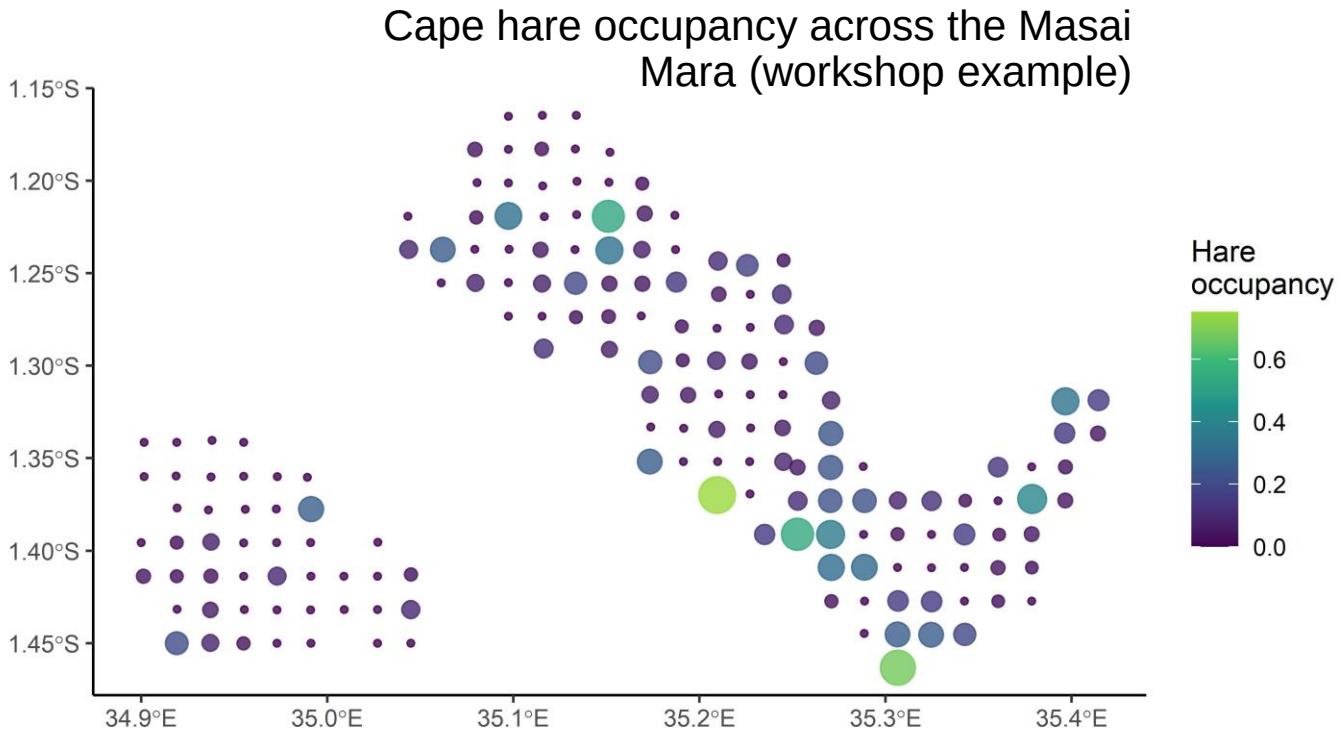
The value of a phenomenon at **time** *t* depends on its value at time *t-1*



Species traits cluster in **phylogenetic** space due to shared evolutionary history

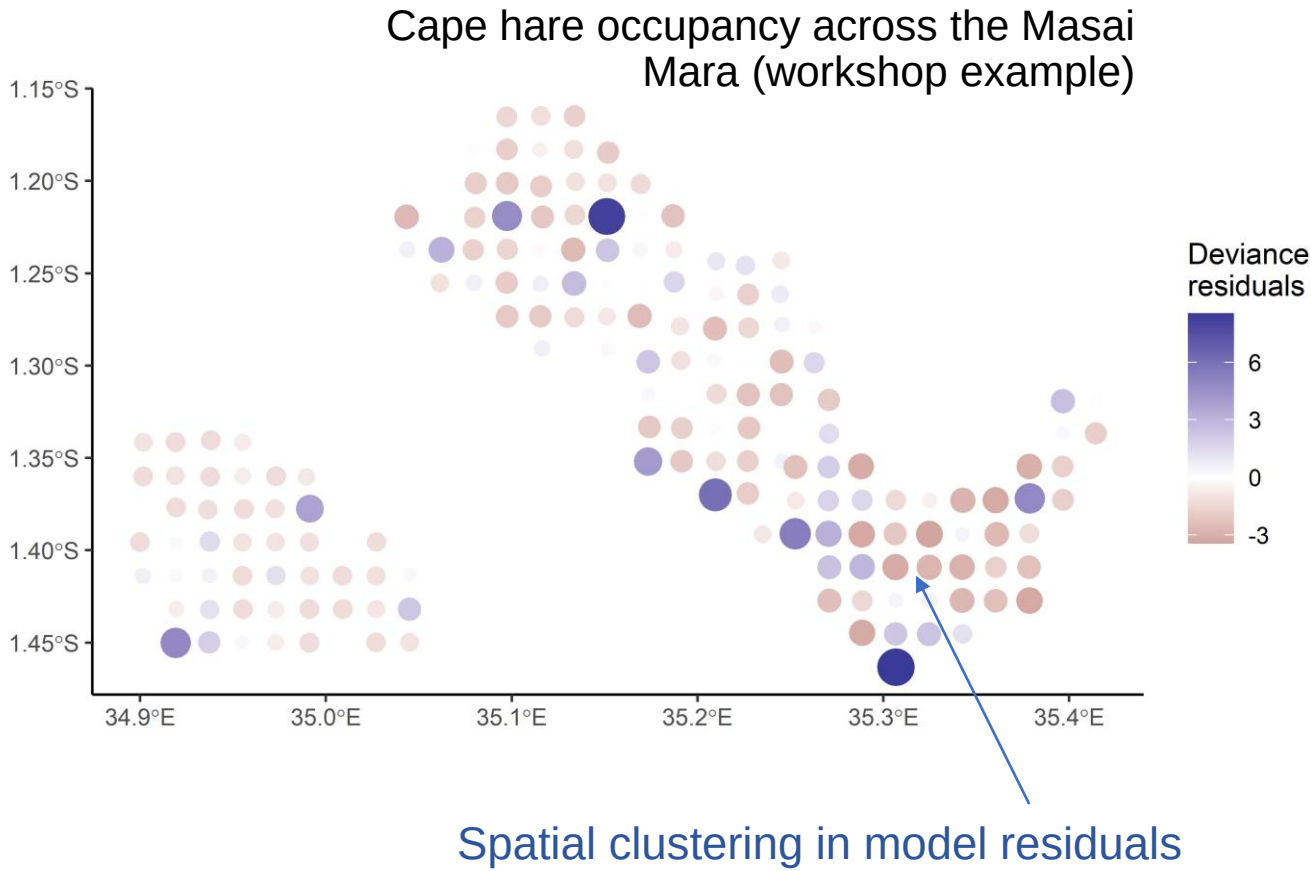
Autocorrelation and ecological inference and prediction

- Studying autocorrelation can provide valuable insight into the processes shaping biological phenomena
  (e.g. phylogenetic signal in species traits)
- If not accounted for in models, can violate the assumption of independence of errors
  (and lead to confounding by unobserved drivers)
- This can be particularly problematic when sampling is clustered or biased
  (as in many observational ecology studies)



Autocorrelation and ecological inference and prediction

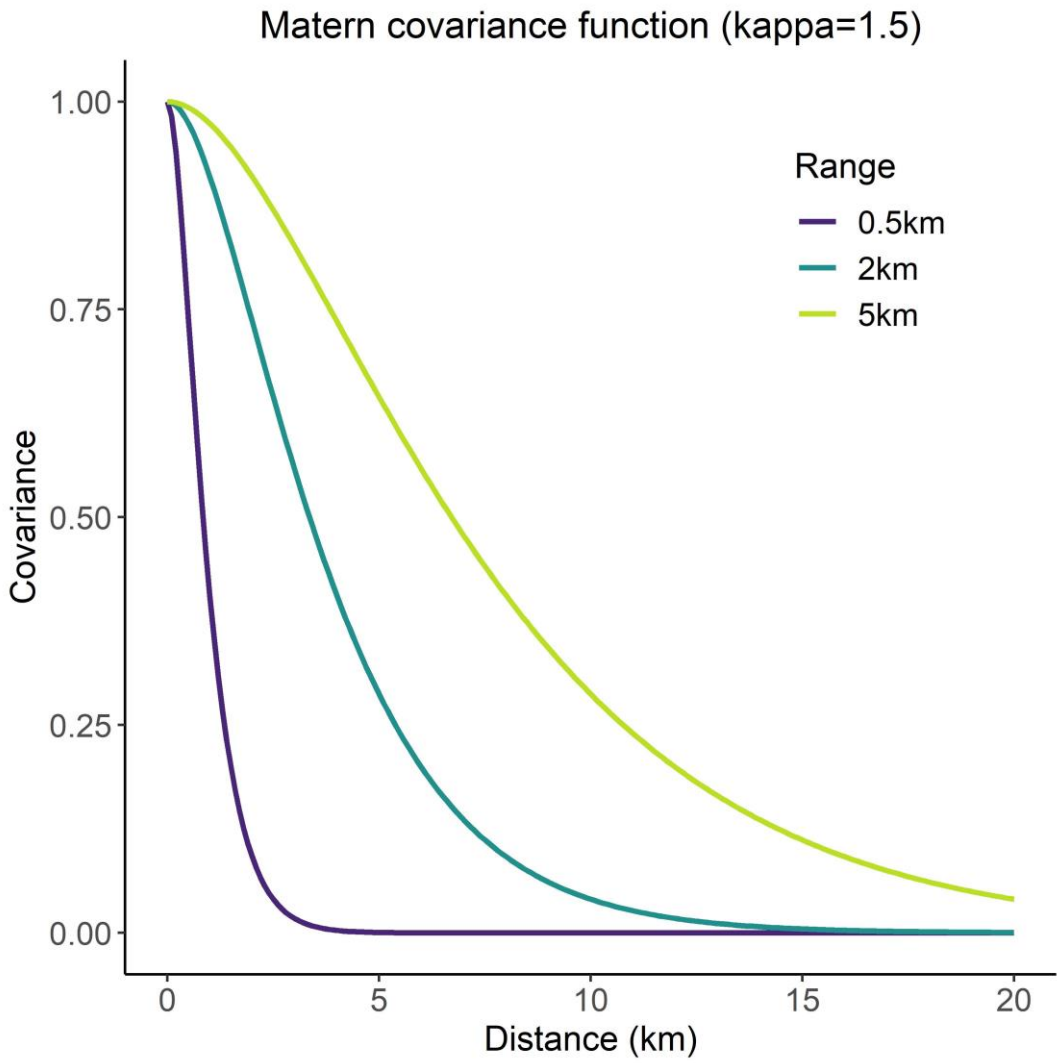
- Studying autocorrelation can provide valuable insight into the processes shaping biological phenomena
  (e.g. phylogenetic signal in species traits)
- If not accounted for in models, can violate the assumption of independence of errors
  (and lead to confounding by unobserved drivers)
- This can be particularly problematic when sampling is clustered or biased
  (as in many observational ecology studies)



Tobler’s first law of geography

“Everything is related to everything else, but near things are more related than distant things” (Tobler 1970)

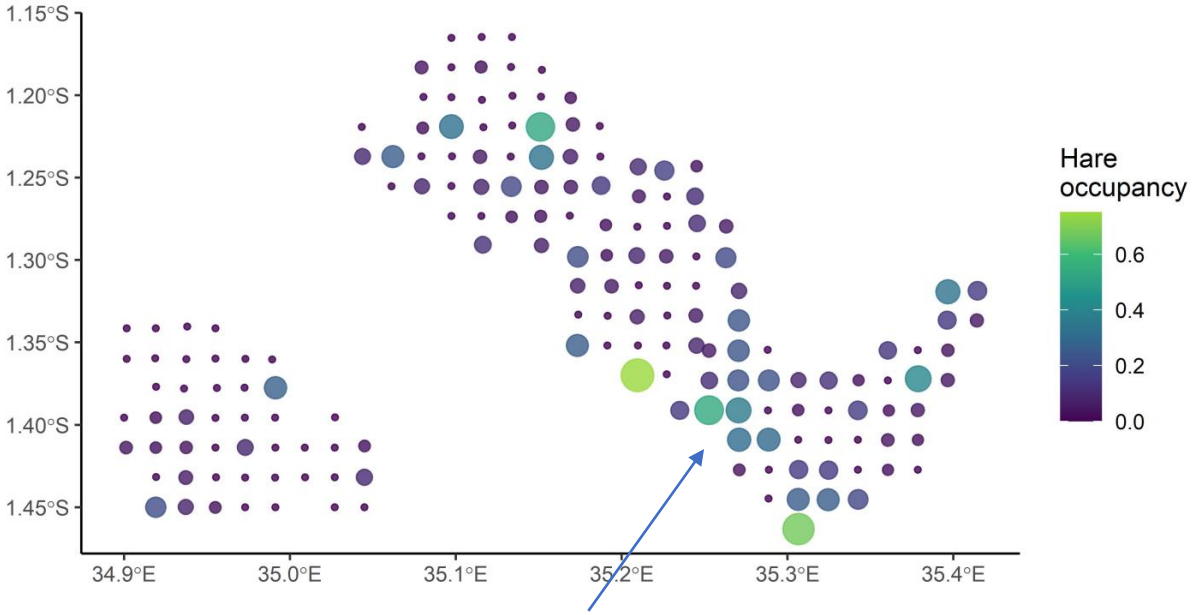
Covariance functions



- Covariance functions describe how the correlation between any pair of observations declines as a function of distance (in space, time, temperature…)
- e.g. Matern covariance function – 2 parameters together determine the shape and distance decay.
- Used in spatial statistics for interpolation (kriging) and in Gaussian process models.

Geospatial model: Gaussian process

Geospatial data = phenomena in continuous space sampled at discrete (point) locations



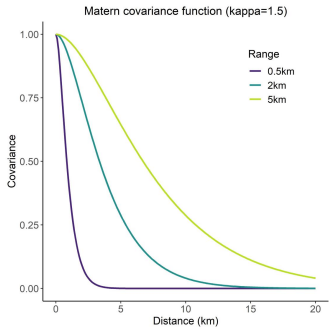
Spatial covariance in hare occupancy: nearby camera trap locations share unmeasured environmental factors and linked by hare movement

- Model hare occupancy per camera trap as a spatially-structured random intercept using a Gaussian process
  (a generalization of the multivariate normal distribution)

$$Y_i \sim Binom(p_i)$$

$$logit(p_i) = \beta_0 + \alpha_c$$

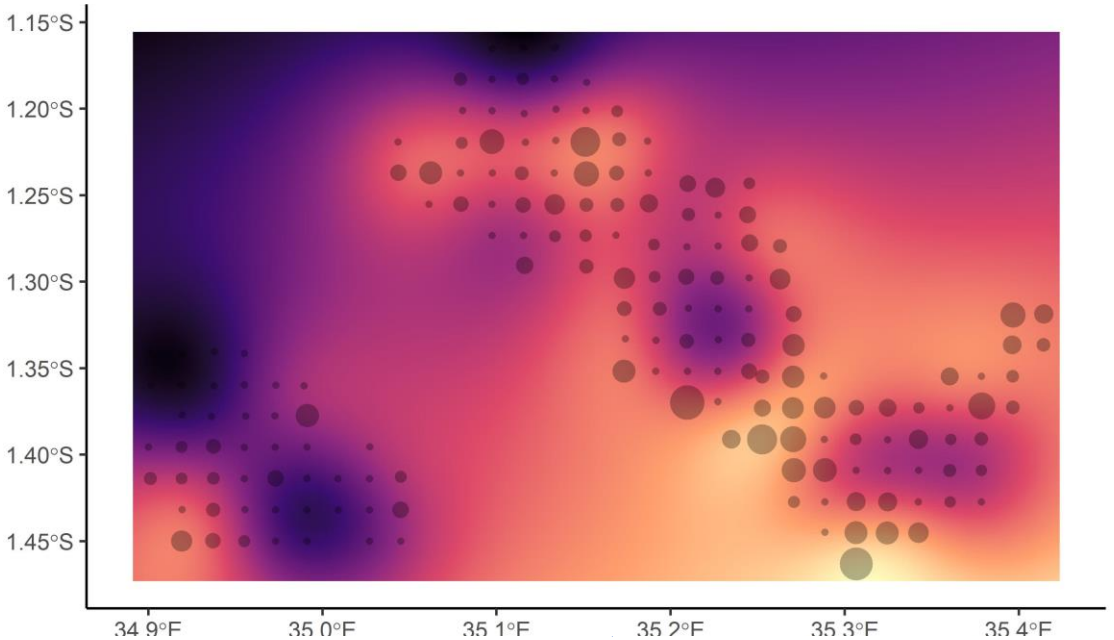
$$\alpha_c \sim MVNormal(0, K)$$



K is a matrix of covariances between all pairs of camera traps *c*, **determined by their distance as described by a covariance function**

Geospatial model: Gaussian process

Geospatial data = phenomena in continuous space sampled at discrete (point) locations

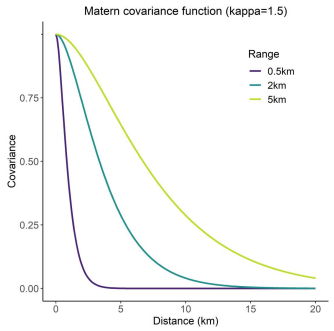


- The fitted effect accounts for spatial dependency among nearby points
  (whereas fitting a random intercept for each camera trap would treat each as independent)

$$Y_i \sim Binom(p_i)$$

$$logit(p_i) = \beta_0 + \alpha_c$$

$$\alpha_c \sim MVNormal(0, K)$$



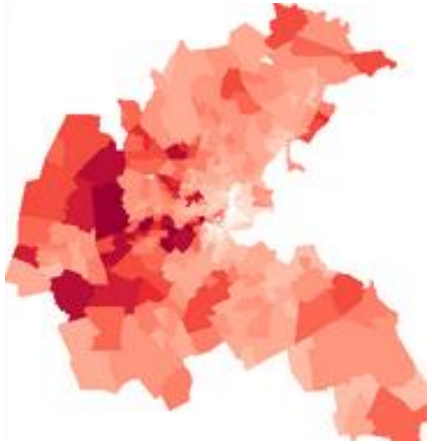
K is a matrix of covariances between all pairs of camera traps *c*, **determined by their distance as described by a covariance function**

https://distill.pub/2019/visual-exploration-gaussian-processes/

Neighbourhood matrices and areal data

- Spatial proximity can instead be defined in terms of adjacency for areal or grid data
  (instead of distance, ask, which areas are neighbours?)
- Markov property: assume that a value in any location depends only on the values of its immediate neighbours
- e.g. conditional autoregressive model

Neighbourhood matrix (Boston admin areas)

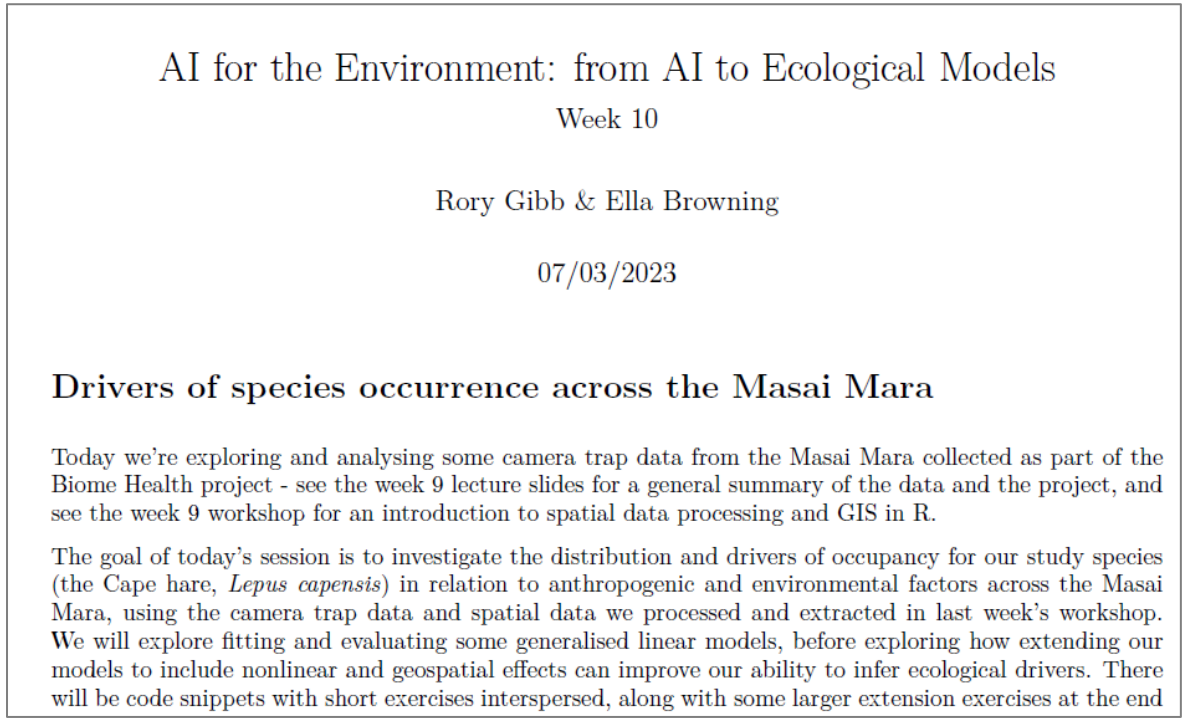


Fitted spatial effect

https://becarioprecario.bitbucket.io/inla-gitbook/ch-spatial.html

Spatial and temporal modelling in R

- Rich ecosystem of resources for spatial, temporal (and phylogenetic) modelling in R
- In today's workshop we're using the 'mgcv' package which is fast and flexible.
- Many others including R-INLA, inlabru, brms, glmmTMB…



AI for the Environment: from AI to Ecological Models

Week 10

Rory Gibb & Ella Browning

07/03/2023

Drivers of species occurrence across the Masai Mara

Today we're exploring and analysing some camera trap data from the Masai Mara collected as part of the Biome Health project - see the week 9 lecture slides for a general summary of the data and the project, and see the week 9 workshop for an introduction to spatial data processing and GIS in R.

The goal of today's session is to investigate the distribution and drivers of occupancy for our study species (the Cape hare, *Lepus capensis*) in relation to anthropogenic and environmental factors across the Masai Mara, using the camera trap data and spatial data we processed and extracted in last week's workshop. We will explore fitting and evaluating some generalised linear models, before exploring how extending our models to include nonlinear and geospatial effects can improve our ability to infer ecological drivers. There will be code snippets with short exercises interspersed, along with some larger extension exercises at the end

Summary

- When developing statistical models, start with the question and the conceptual (scientific) understanding.
- Ecosystems and the measurement process are noisy – think carefully about *how* the data were generated.
- Statistical models assume errors are independent – a suite of approaches, from multi-level to spatial, temporal and phylogenetic models, help to account for underlying non-independence.
- What is a “good” model depends on the objective of the analysis – are you seeking to describe, explain, or predict?