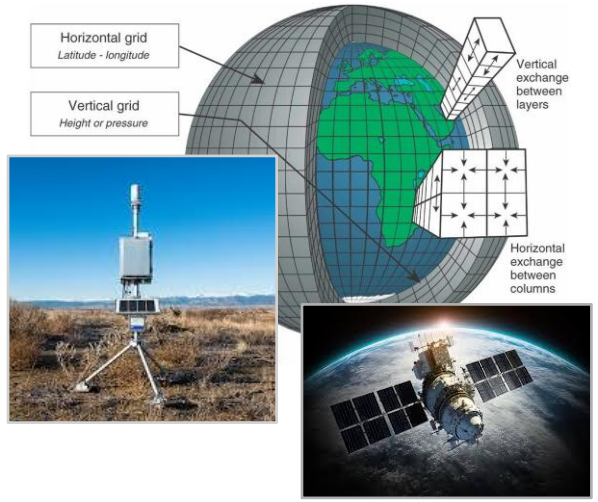
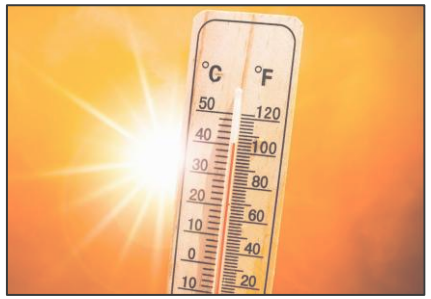
Measuring environmental effects on health

https://www.australiangeographic.com.au/news/2022/11/australian-mosquito-numbers-are-booming/

Recap of last week





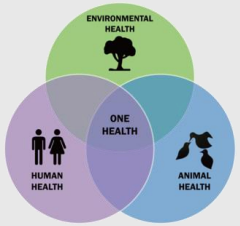
Rich **data on the climate and Earth system** are available from climate models and remote sensing – enable present and future impact studies.

Scenario frameworks allow the climate change and impacts research community to assess future risks, despite uncertainty.

Predicting risks to health and biodiversity requires **models that link environmental drivers to health or ecological outcomes.**

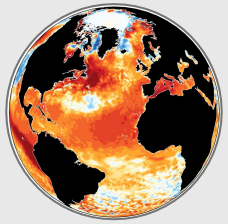
Today in context

Week 1



Future human and ecosystem health

Week 2

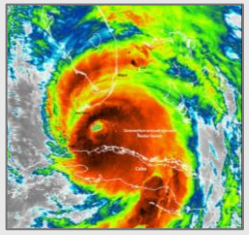


Scenarios of an uncertain future

Week 3

Measuring drivers of health risks

Week 4

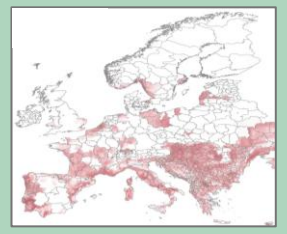


Prediction and forecasting

Week 5

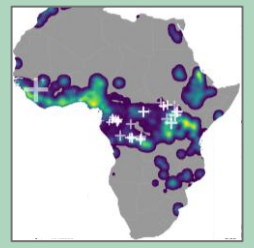
From models to solutions?

Week 7



Long-term futures for health

Week 8

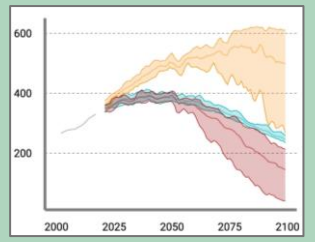


How well can we predict the future?

Week 9

Solutions for healthier cities

Week 10

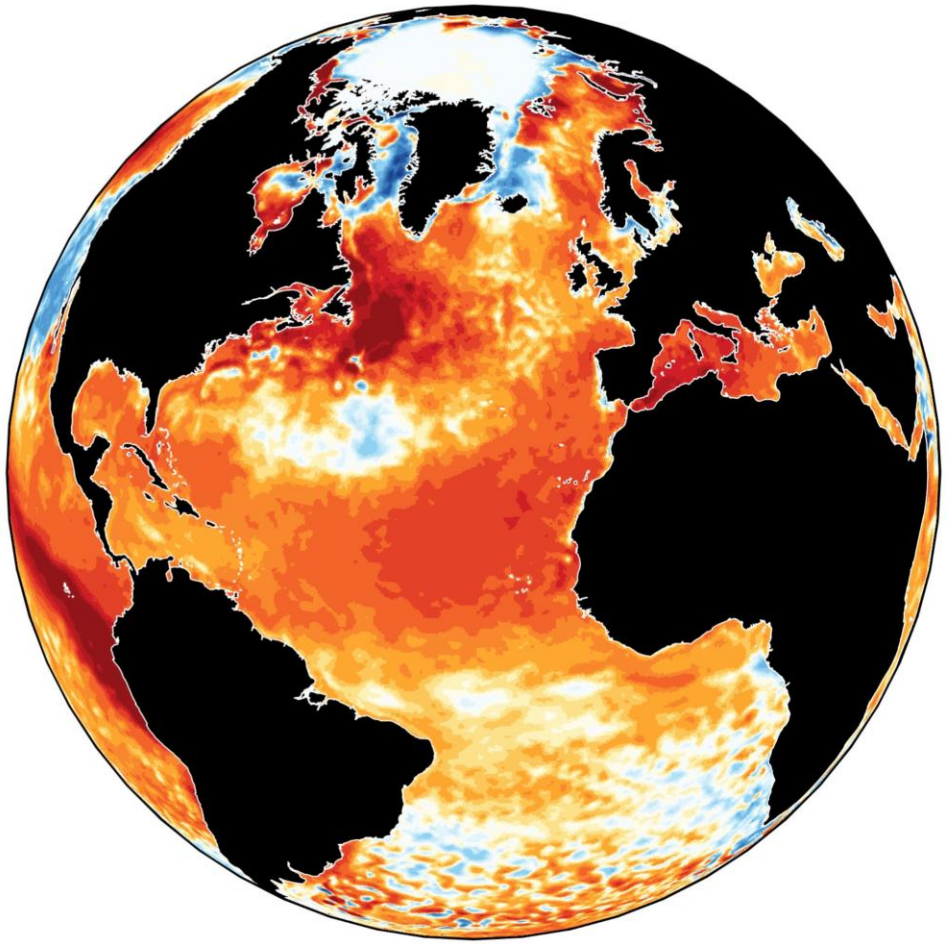


Communicating uncertainty

Week 11

Linking models to policy

Learning objectives



- Describe **major challenges in analysis** of observational ecological and health datasets (including confounding and autocorrelation)
- Discuss **statistical approaches** that can help to address these challenges
- Explain the role of climate and ecosystems in **vector-borne disease epidemiology**
- Develop statistical models relating environmental exposures to ecological outcomes (*practical*)

The structure of today

Lecture:
Measuring environmental drivers

09:00-11:00

Practical:
Climate and mosquito abundance in Italy

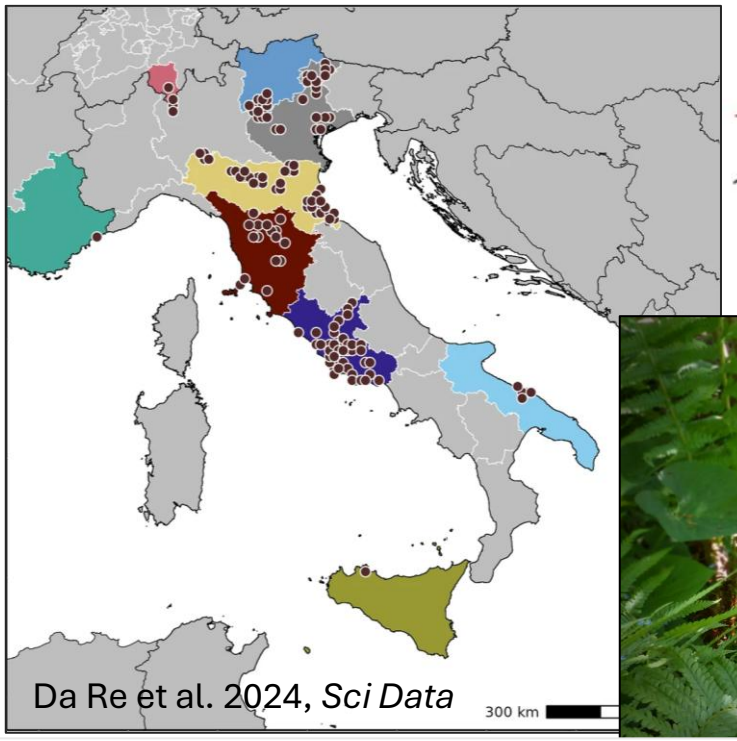
14:00-17:00

This afternoon's practical

R workshop: Measuring the influence of environment and climate on *Aedes albopictus* abundance in Italy, using vector survey data.

Workbook and data available at https://github.com/MSc-ECCH-UCL/BIOS0052-Human-And-Ecosystem-Health

14:00-17:00, Marshgate Room 636

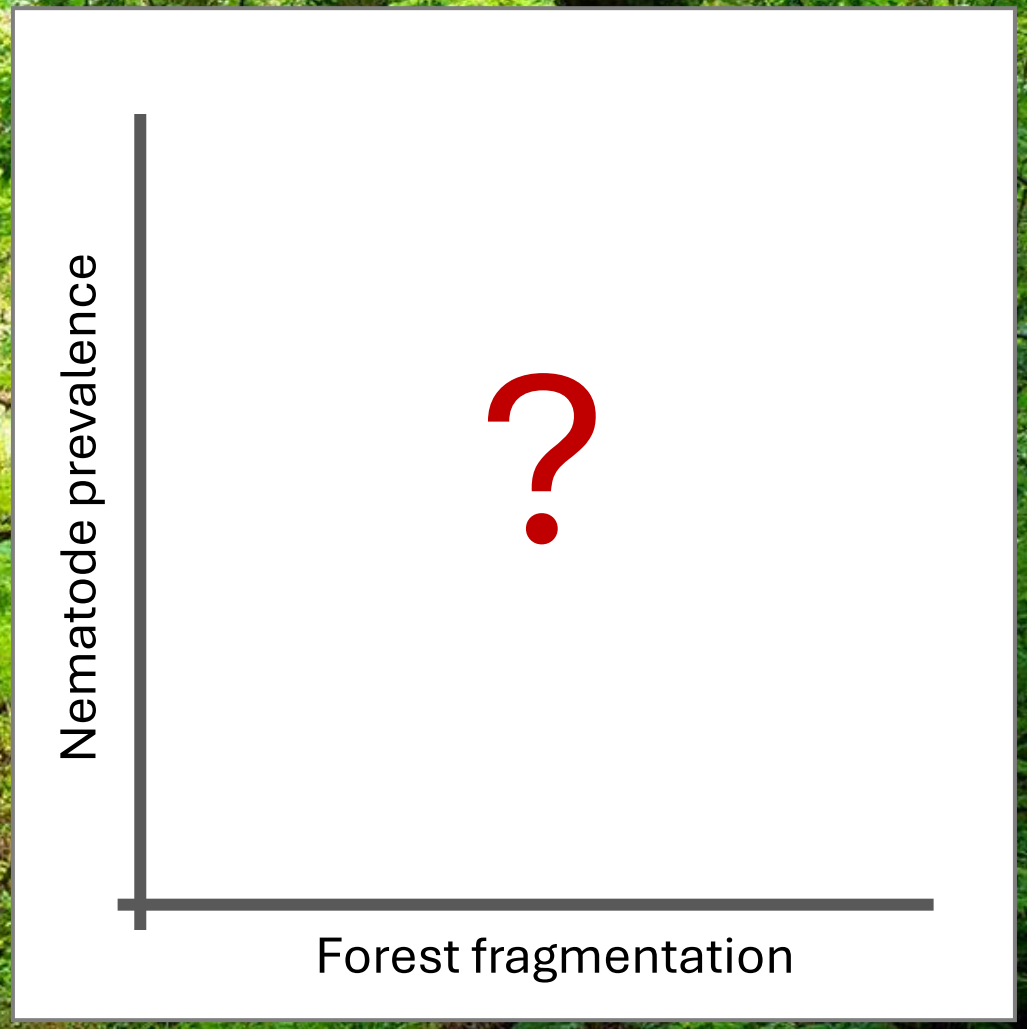


Should you worry about tiger mosquitos in Europe?

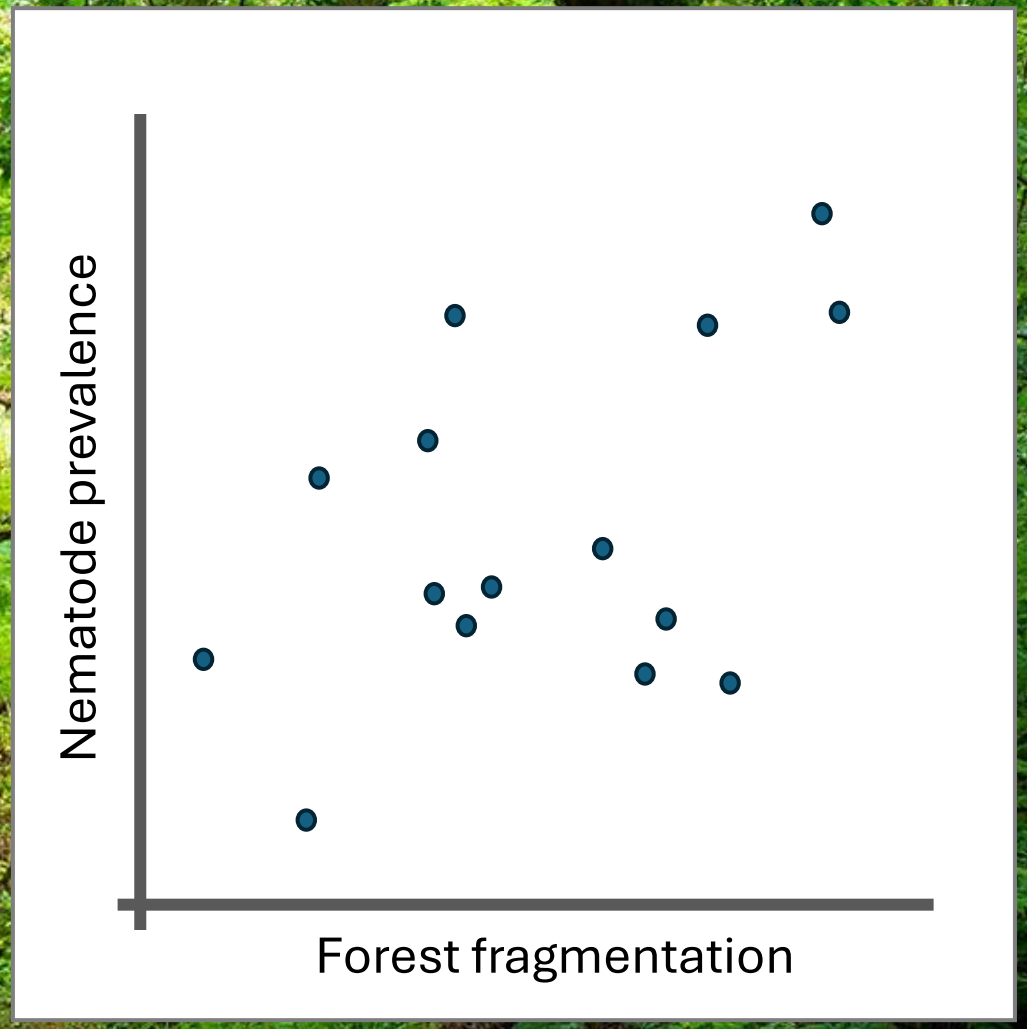
NATIONAL GEOGRAPHIC

Carrying dengue fever and other diseases, the insects have arrived in Europe from Asia thanks to rising temperatures—here's what travelers should know.

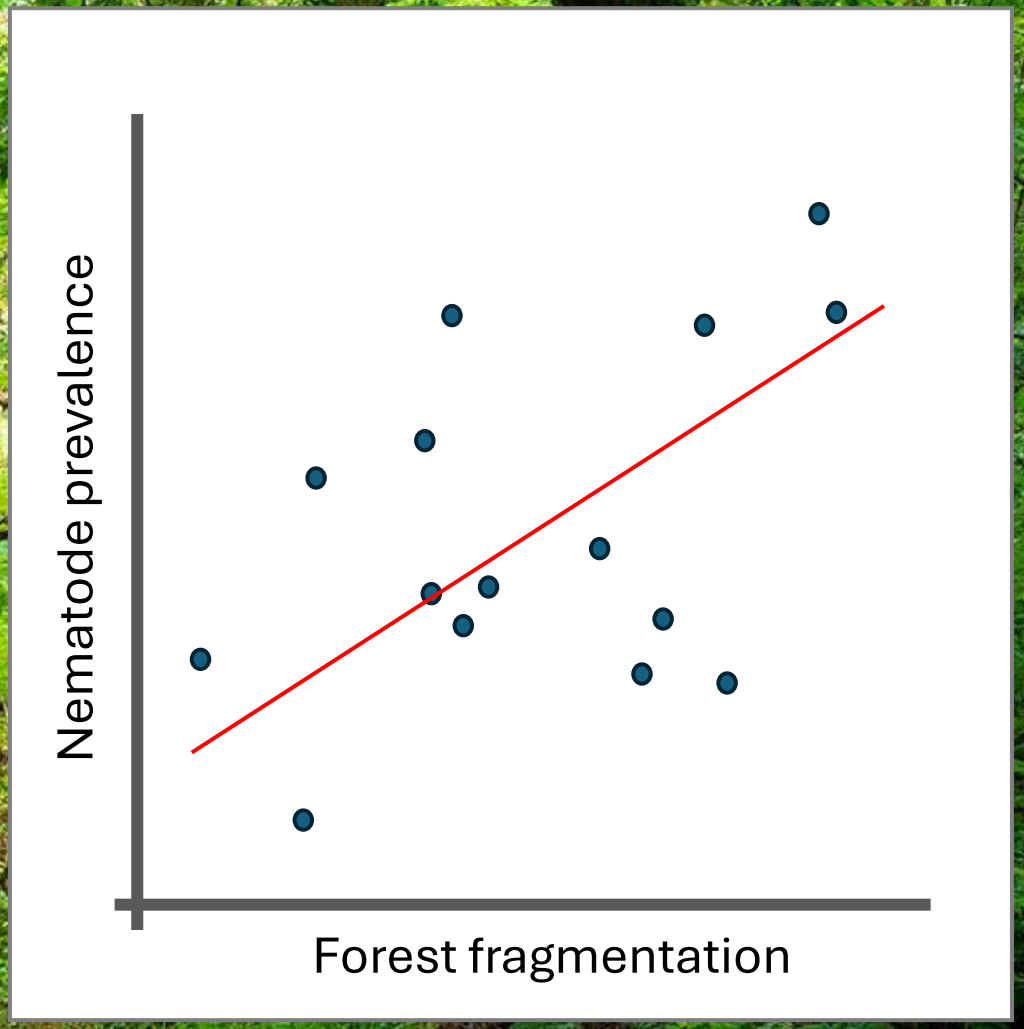
Learning about nature



Learning about nature

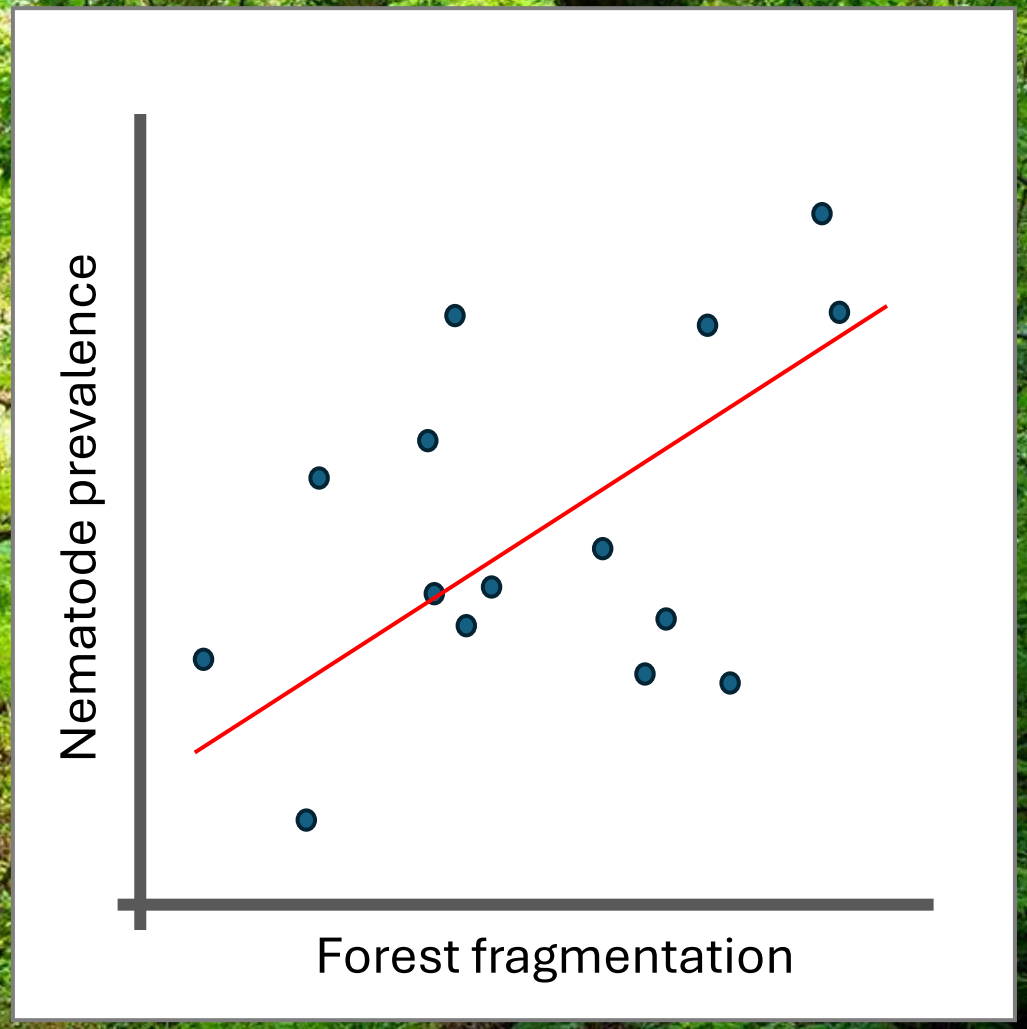


Learning about nature



Learning about nature

- What is a model?
- Challenges of ecological and health data
- Modelling tools for complex data
- Innovations in inference



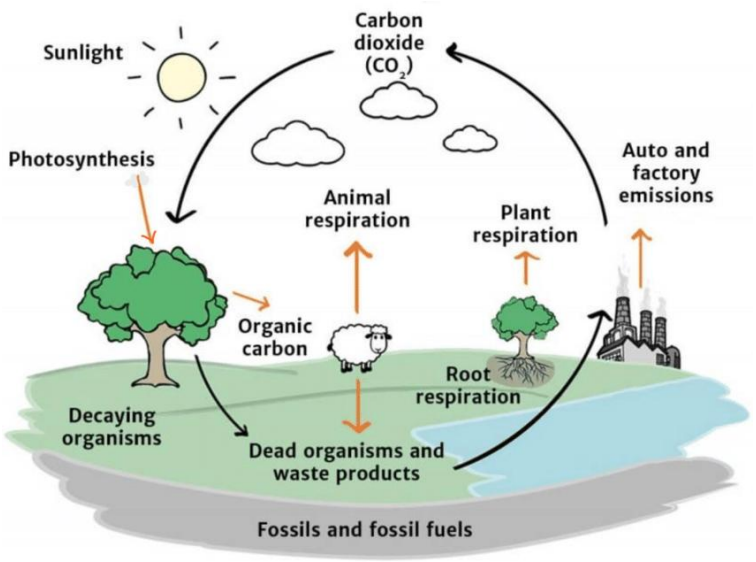
What is a model?

What is a model?

https://en.wikipedia.org/wiki/Eise_Eisinga_Planetarium

What is a model?

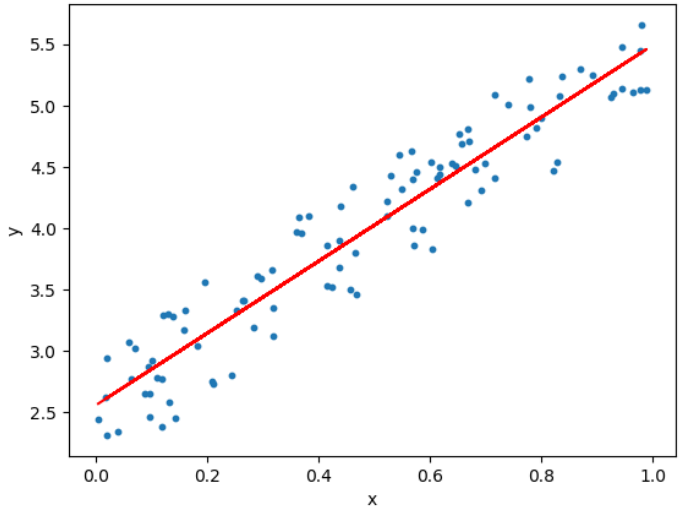
Conceptual



Conceptual representation of known/ hypothesised relationships in a system

Can make qualitative predictions about a system's expected behaviour

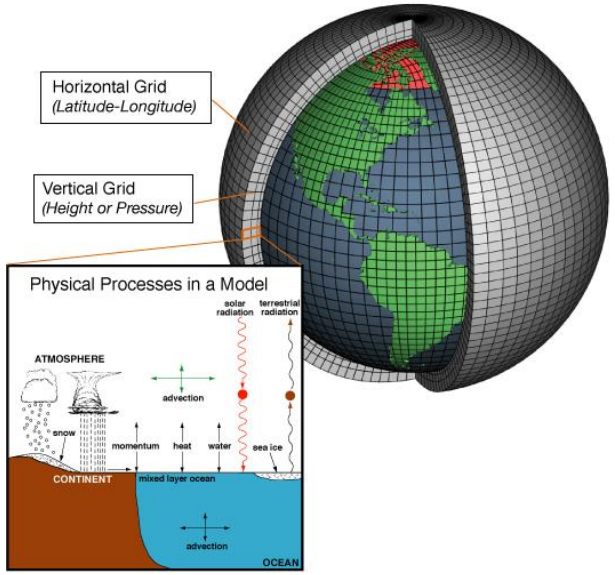
Statistical



Estimation of parameters of interest by fitting to data (observations)

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

Mathematical (mechanistic)



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

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

What is a model?

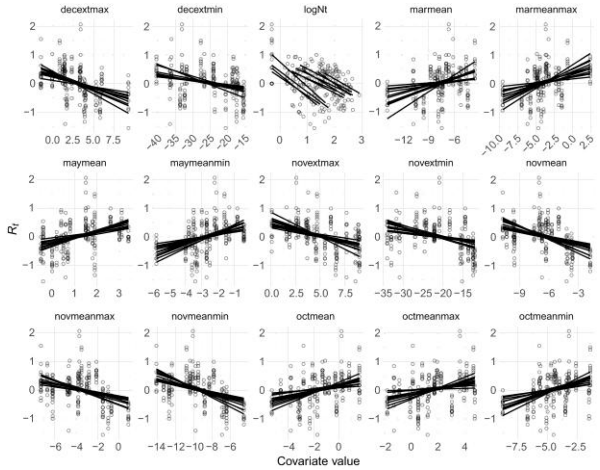
What is a model?

Conceptual

Statistical

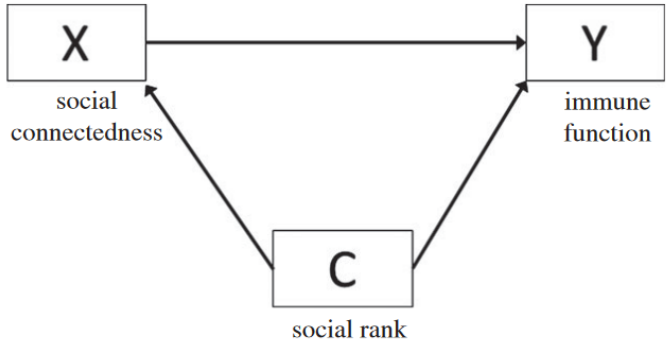
What can statistical models help us to do?

Describe



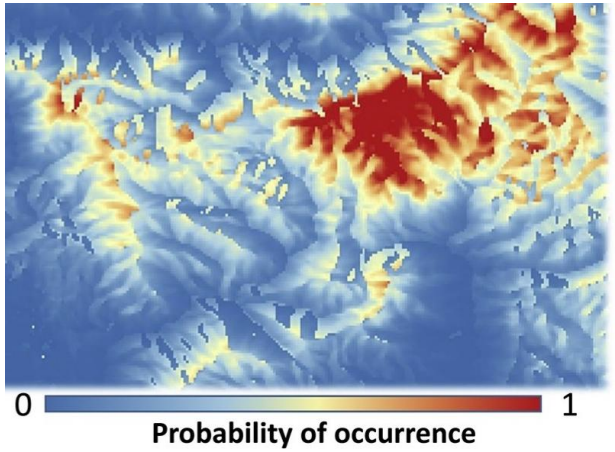
Describe associations among variables, to help generate hypotheses

Explain



Answer questions about probable **causal** relationships, i.e. *"does X cause Y?"*

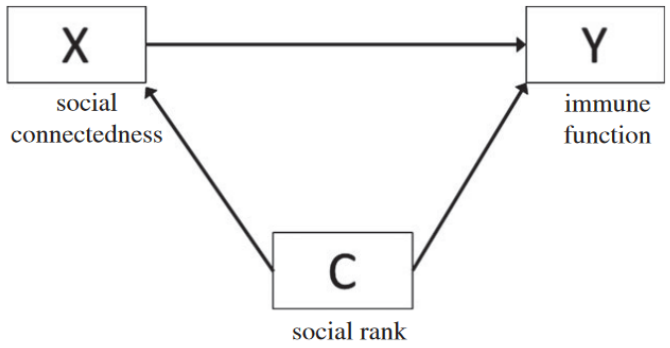
Predict



Predict unobserved outcomes in different places, times or species, *e.g. mapping/forecasting*

What can statistical models help us to do?

Explain

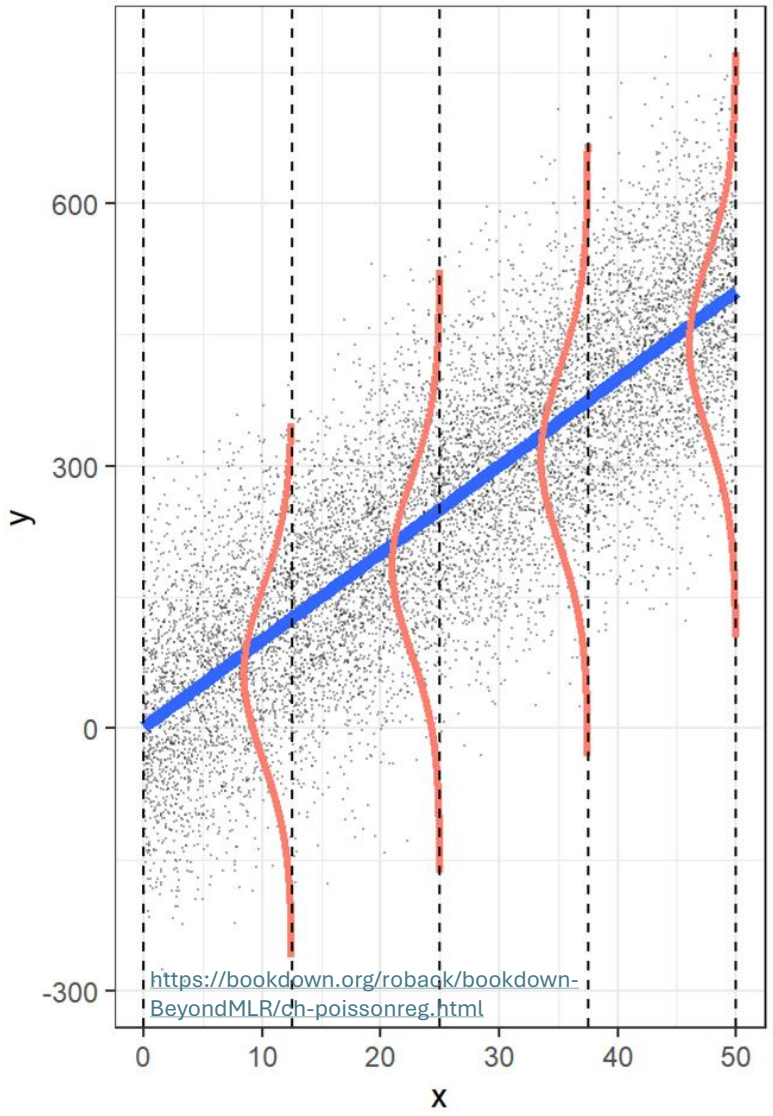


Answer questions about probable **causal** relationships, i.e. *"does X cause Y?"*

Ecological and health sciences are heavily skewed towards **observational** studies

How well are we able to attribute causes using observational data?

Fundamental tools: linear regression



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*

Fundamental tools: generalised linear models

Likelihood and link function relax assumptions around linearity and error distribution

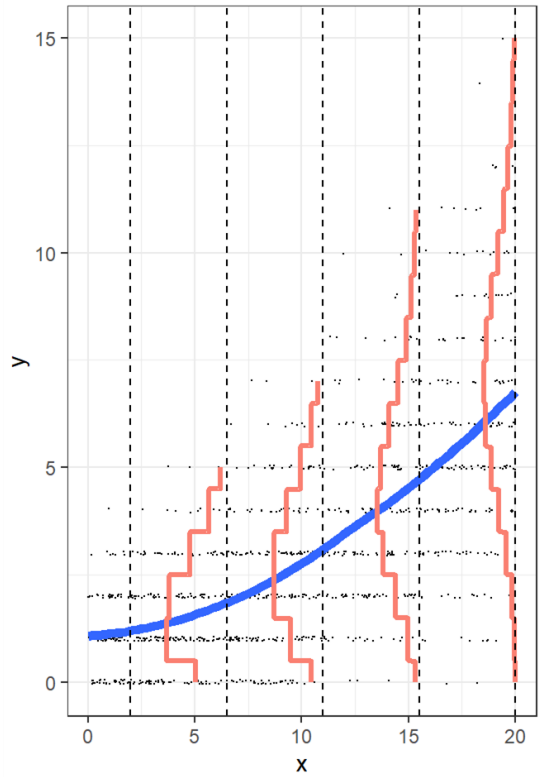
<u>Still assume residual errors are independent!</u>

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)



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

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

Fundamental tools: generalised linear models

Likelihood and link function relax assumptions around linearity and error distribution

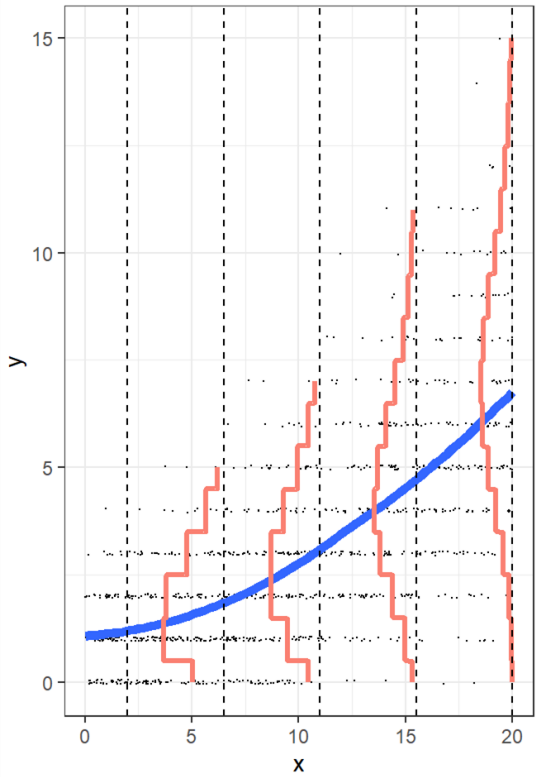
<u>Still assume residual errors are independent!</u>

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)*

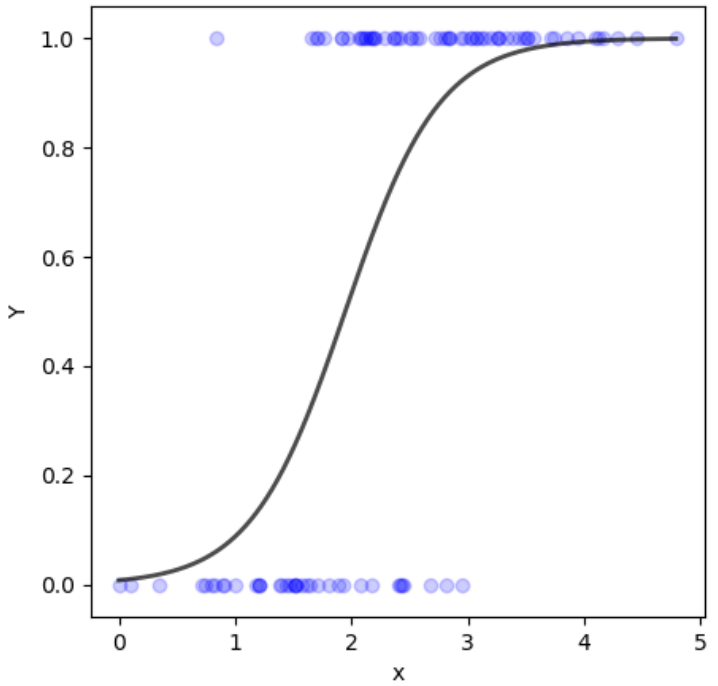
Poisson

(e.g. butterfly counts)



Logistic (binomial)

(e.g. hatching success)



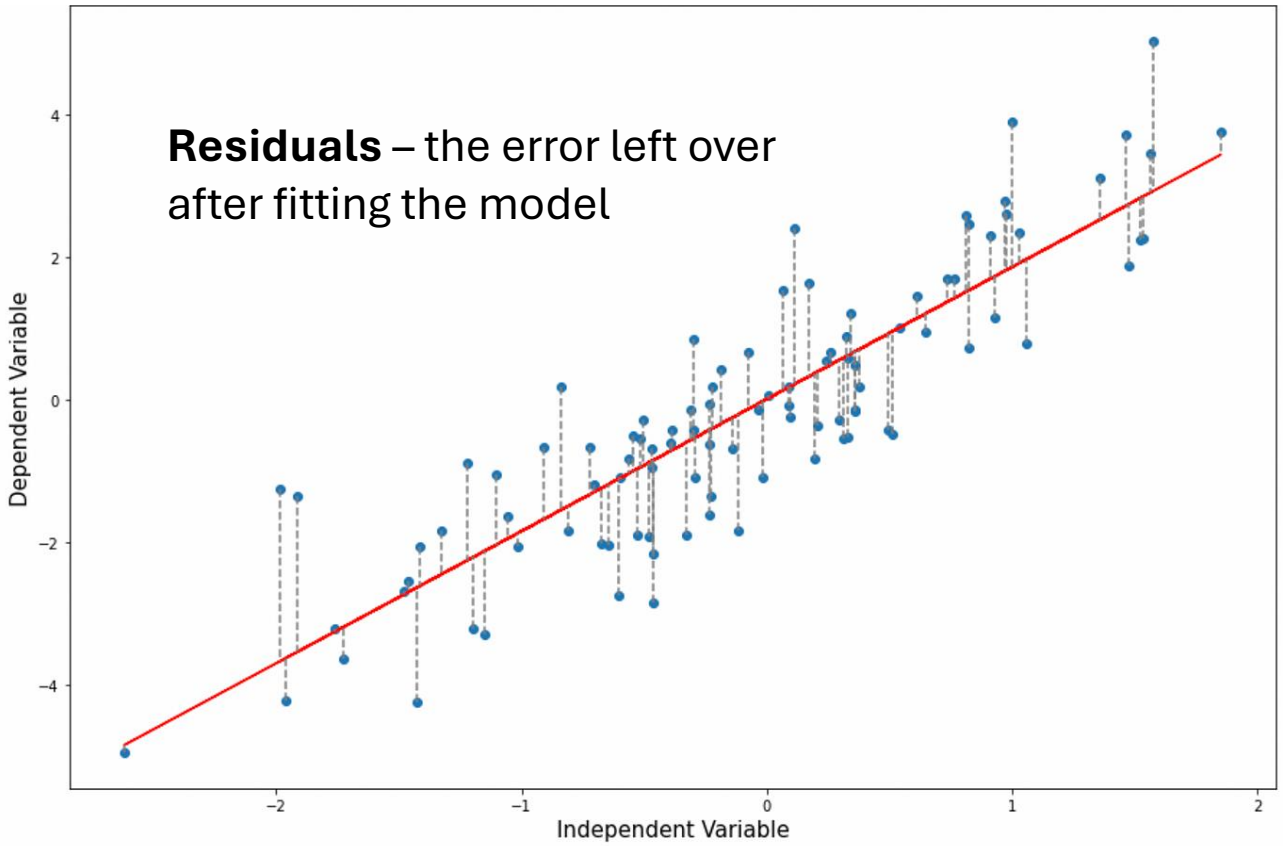
Likelihood function
(describes error distribution) → $Y_i \sim Pois(\lambda_i)$

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

$Y_i \sim Binom(p_i)$

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

Why is independence of residual errors so important?



We assume that the **errors (residuals) are statistically independent** *(i.e. uncorrelated)*

If violated, this can severely impact the reliability of inference

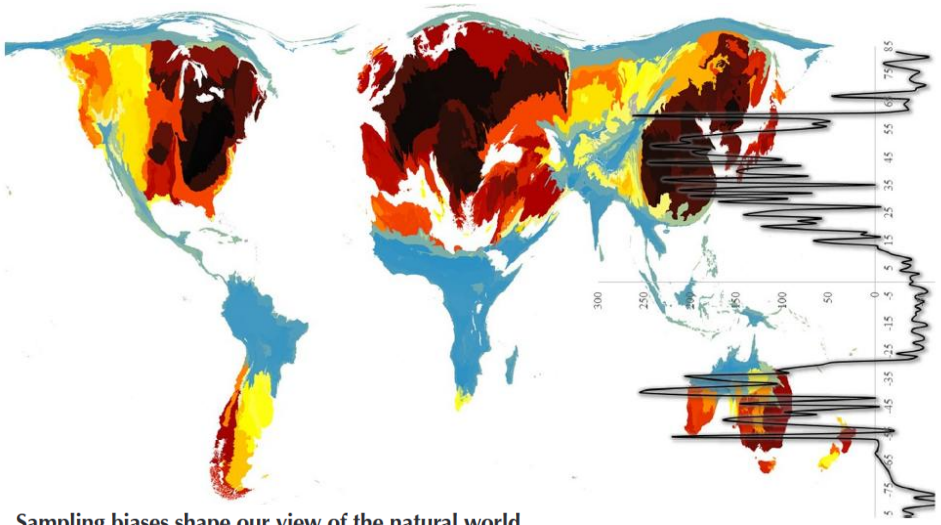
Conventionally address this via careful sampling design, e.g. randomization

Difficult with ecological and health data – inherently spatial and temporal in nature, true randomization not always (or often) possible

What are the challenges of analysing observational data on ecology and health?

Real worlds are noisy, and so is the observation process

Real worlds are noisy, and so is the observation process



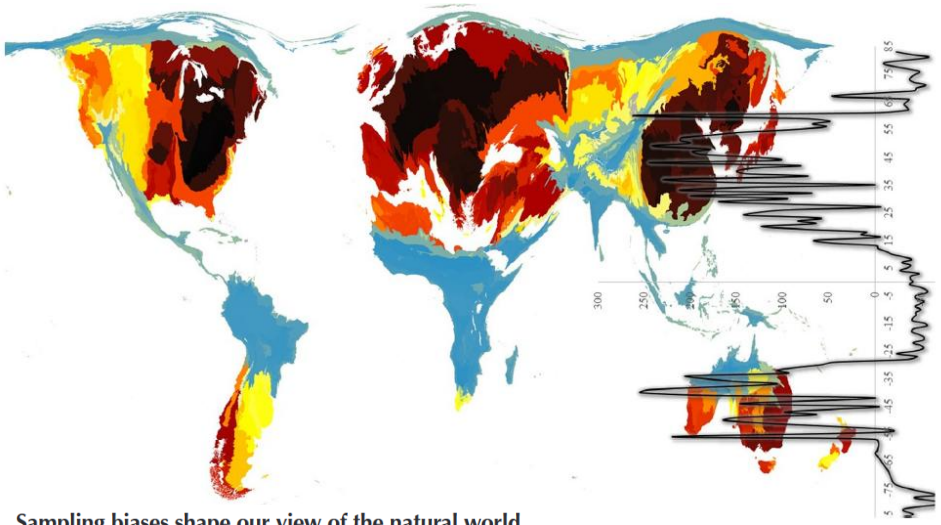
Sampling biases shape our view of the natural world

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)*

Real worlds are noisy, and so is the observation process



Sampling biases shape our view of the natural world

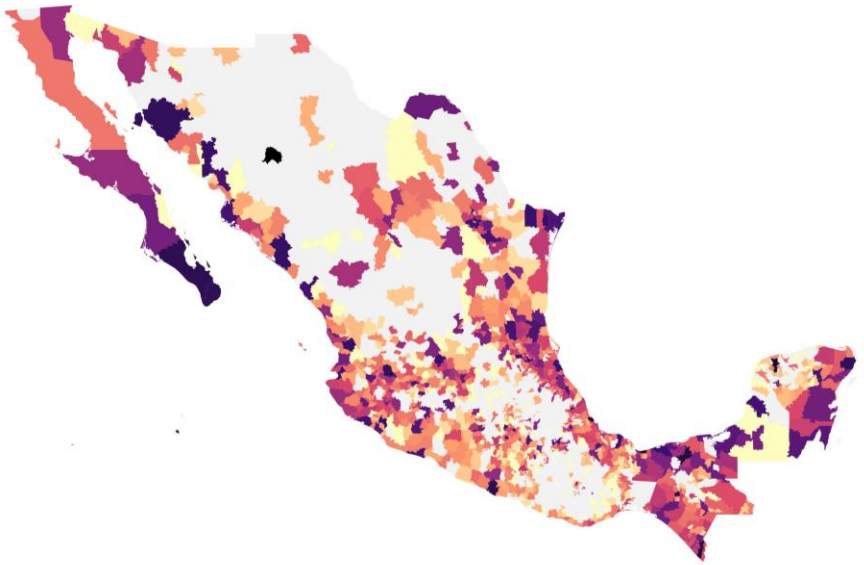
Alice C. Hughes, Michael C. Orr, Keping Ma, Mark J. Costello, John Waller, Pieter Provoost, Qinmin Yang, Chaodong Zhu and Huijie Qiao

Sampling biases shape our understanding

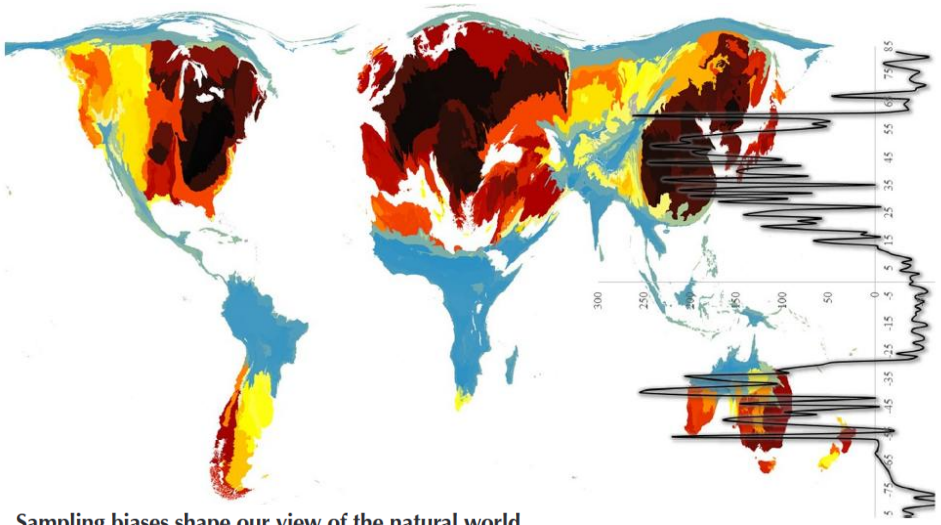
*(historical and evolving **biases in survey effort** confound our knowledge of pattern and process)*

Dependency in space and time

*(data points **closer together are more closely related**)*



Real worlds are noisy, and so is the observation process



Sampling biases shape our view of the natural world

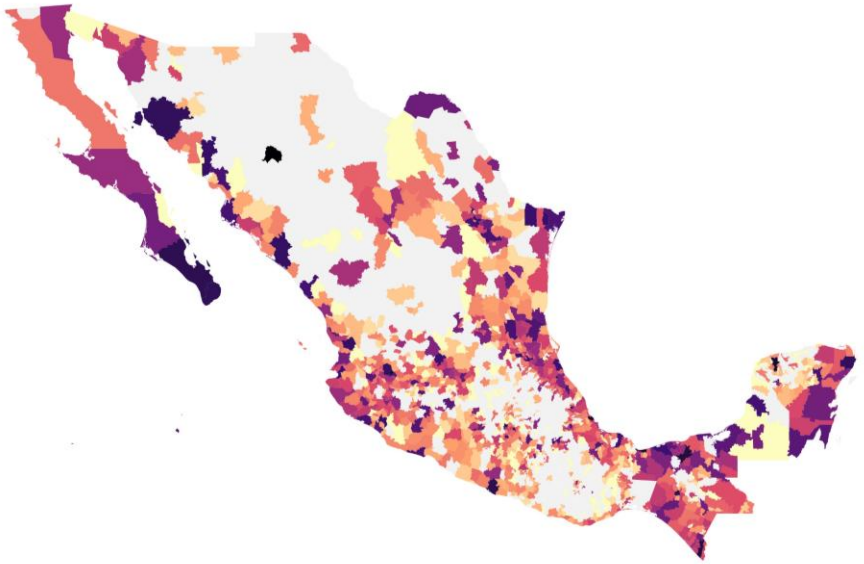
Alice C. Hughes, Michael C. Orr, Keping Ma, Mark J. Costello, John Waller, Pieter Provoost, Qinmin Yang, Chaodong Zhu and Huijie Qiao

Sampling biases shape our understanding

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

Dependency in space and time

*(data points **closer together are more closely related**)*

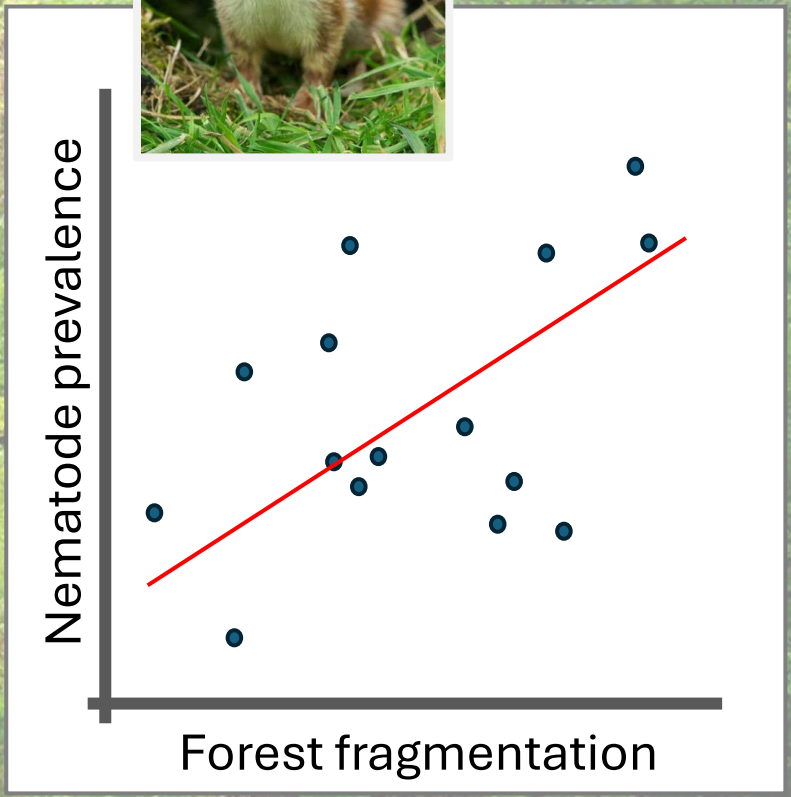


Observation is imperfect

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

Confounding

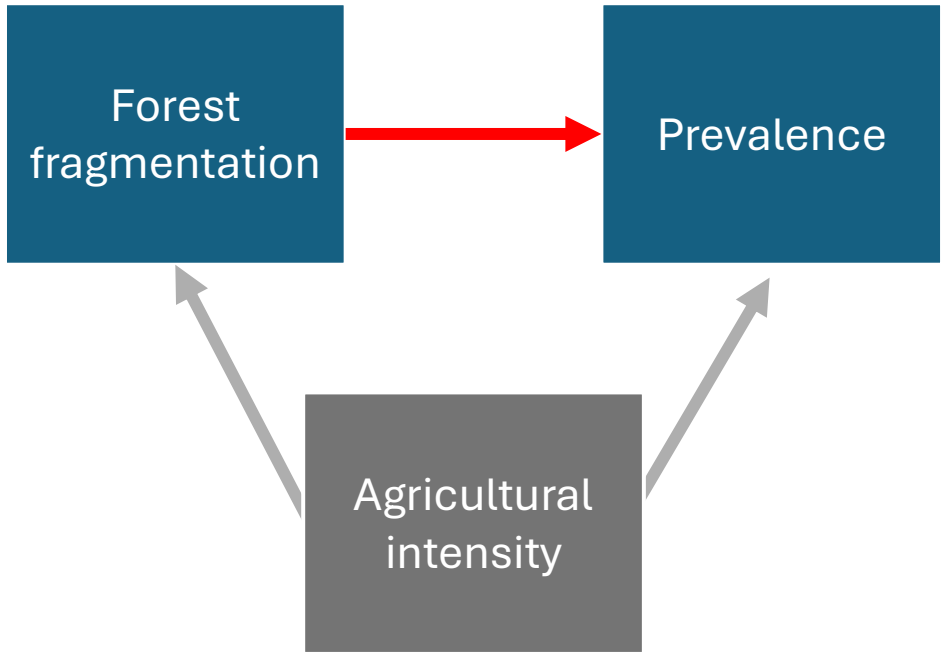
Forest fragmentation → Prevalence

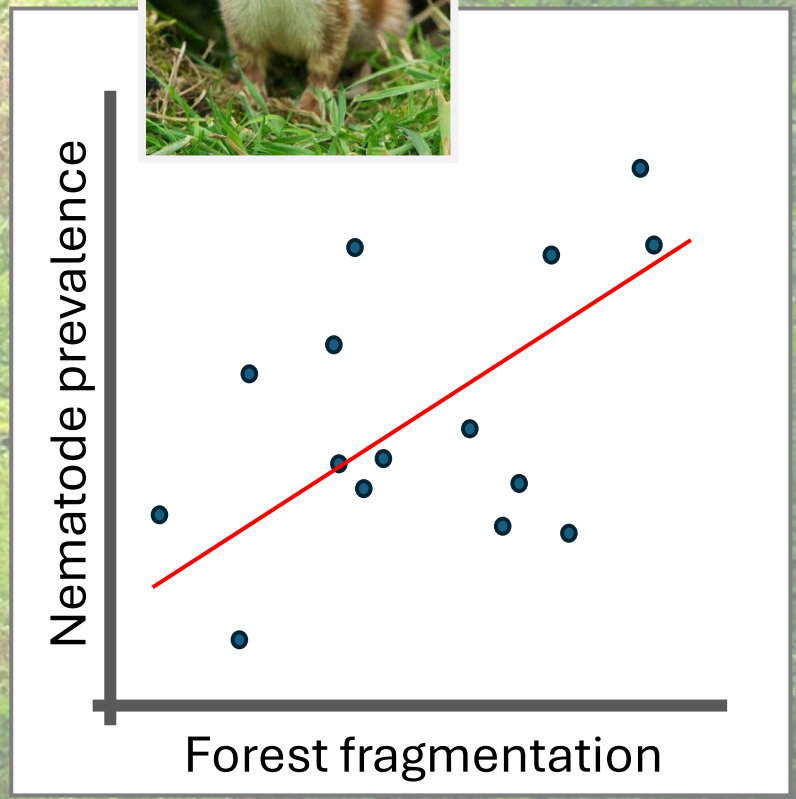


Confounding

A confounder is a **variable that influences *both* the exposure and the outcome** variable

(and we don't necessarily observe it!)

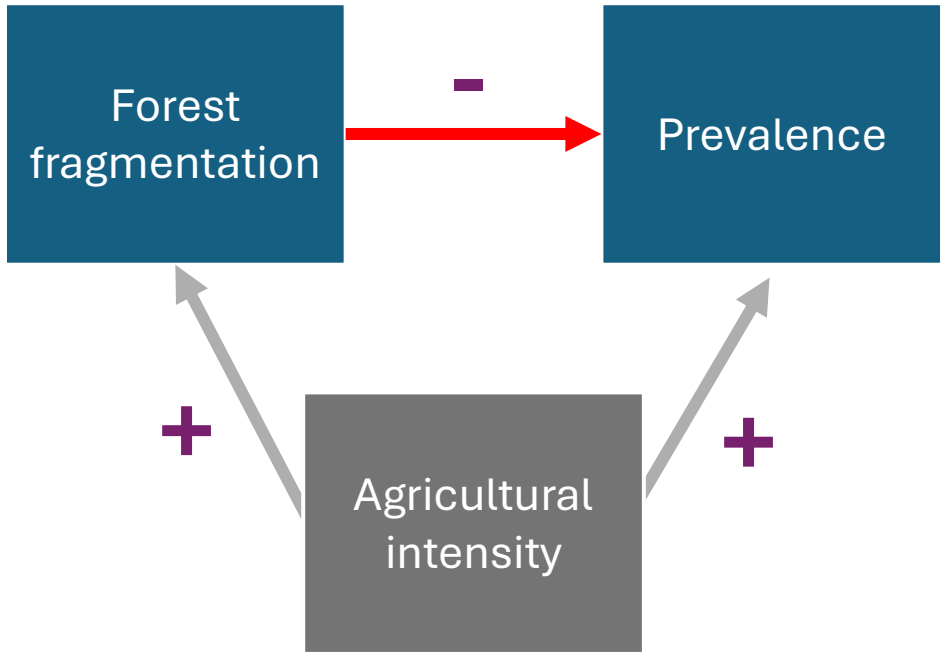


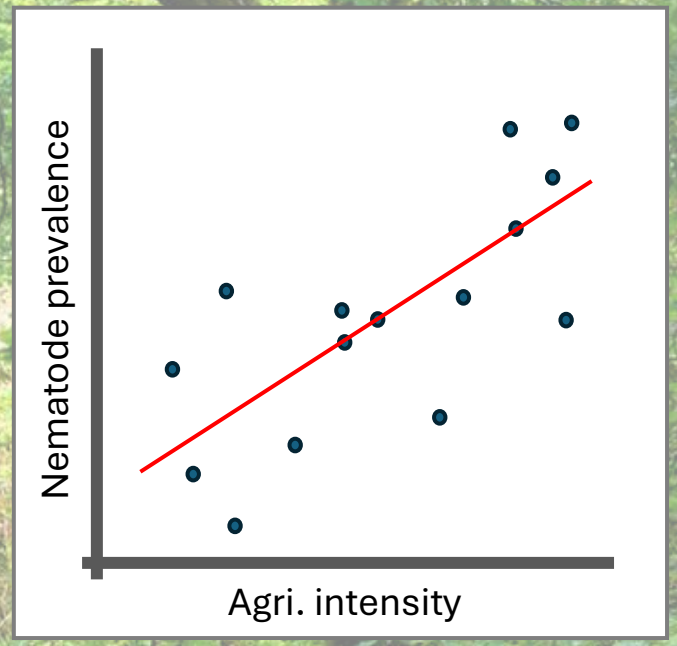


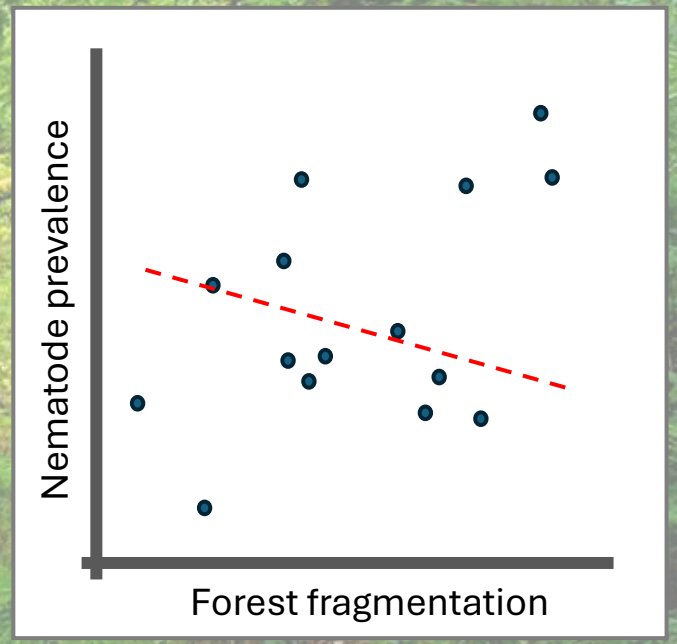
Confounding

A confounder is a **variable that influences *both* the exposure and the outcome** variable

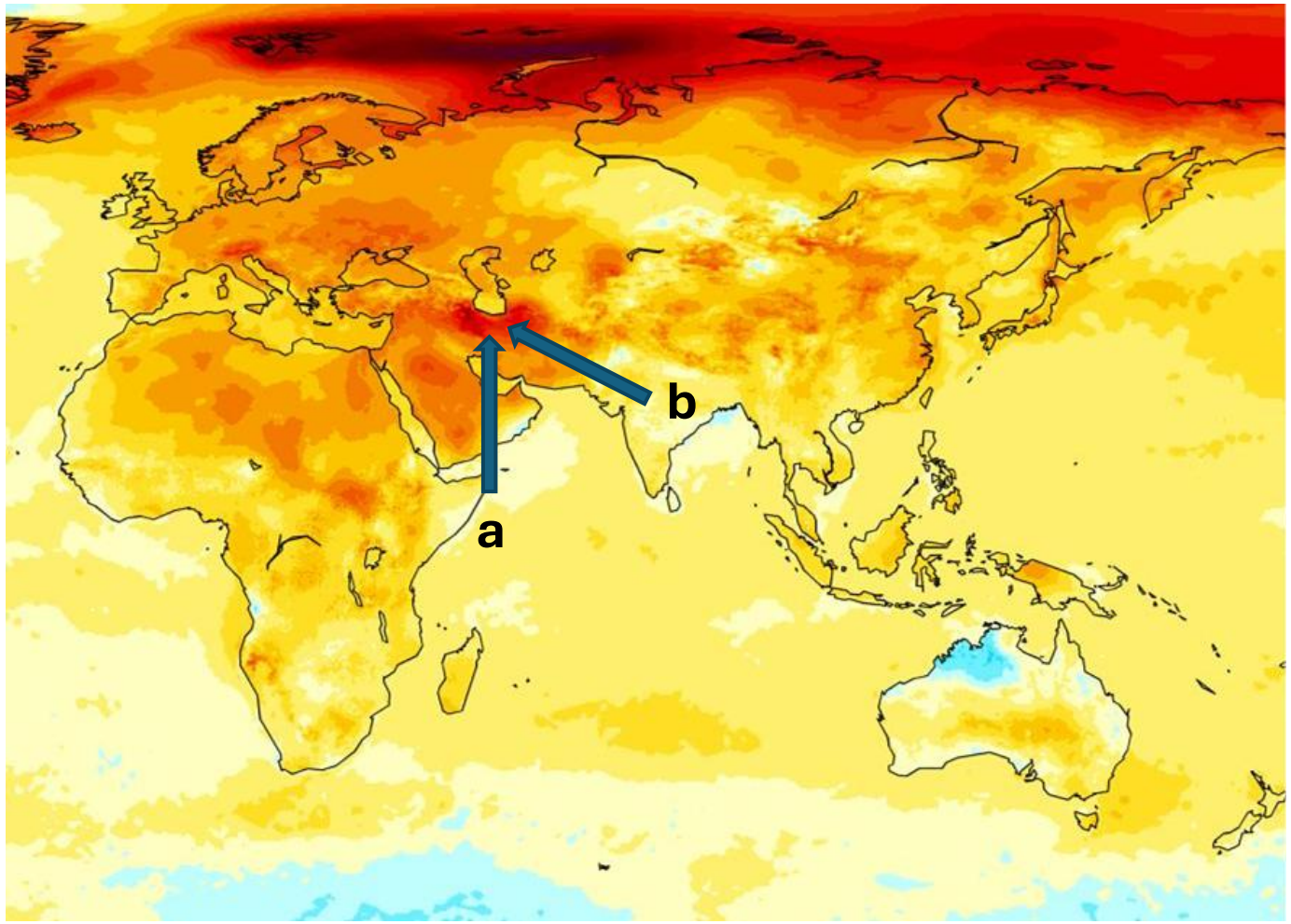
(and we don't necessarily observe it!)







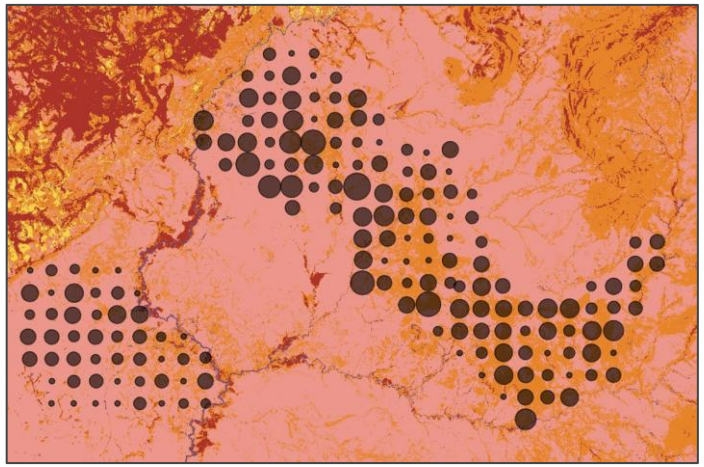
Autocorrelation over space, time and phylogeny



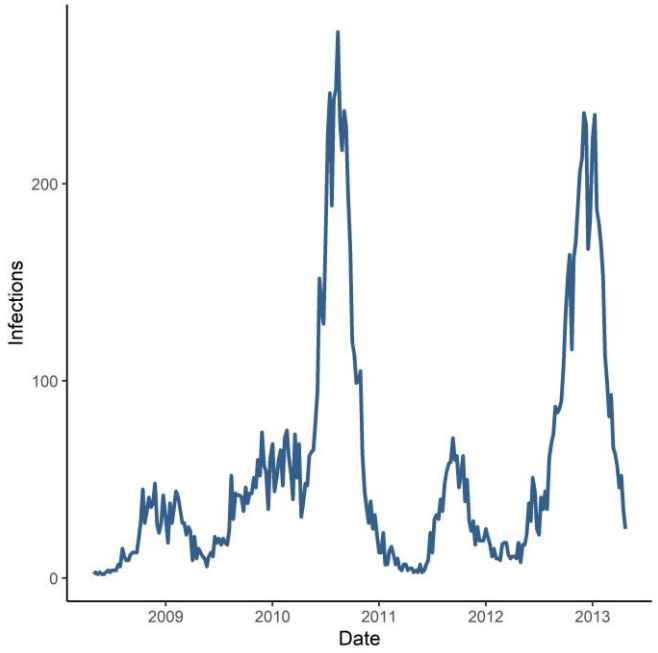
Observations nearby to one another (in space or time) are **non-independent**

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

Autocorrelation over space, time and phylogeny

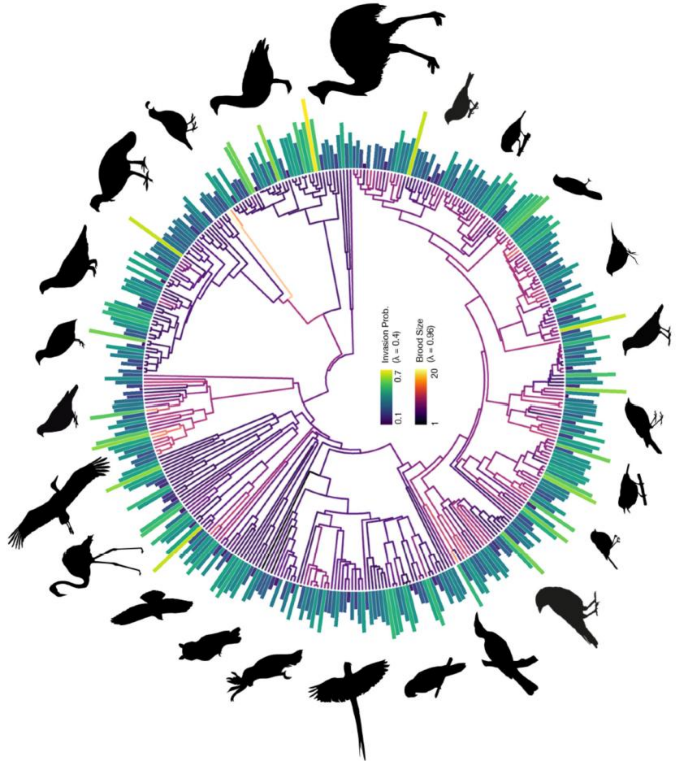


Nearby locations in **space** share common environmental features and are linked by organism movement



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

(e.g. population size, infectious disease prevalence)

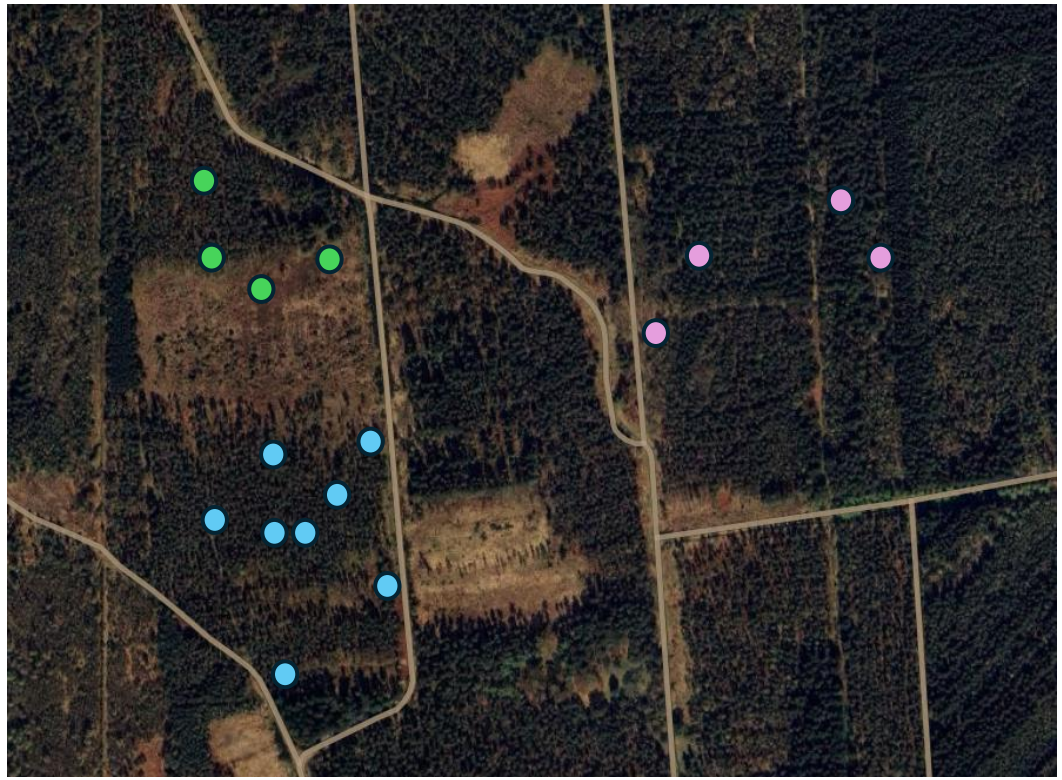


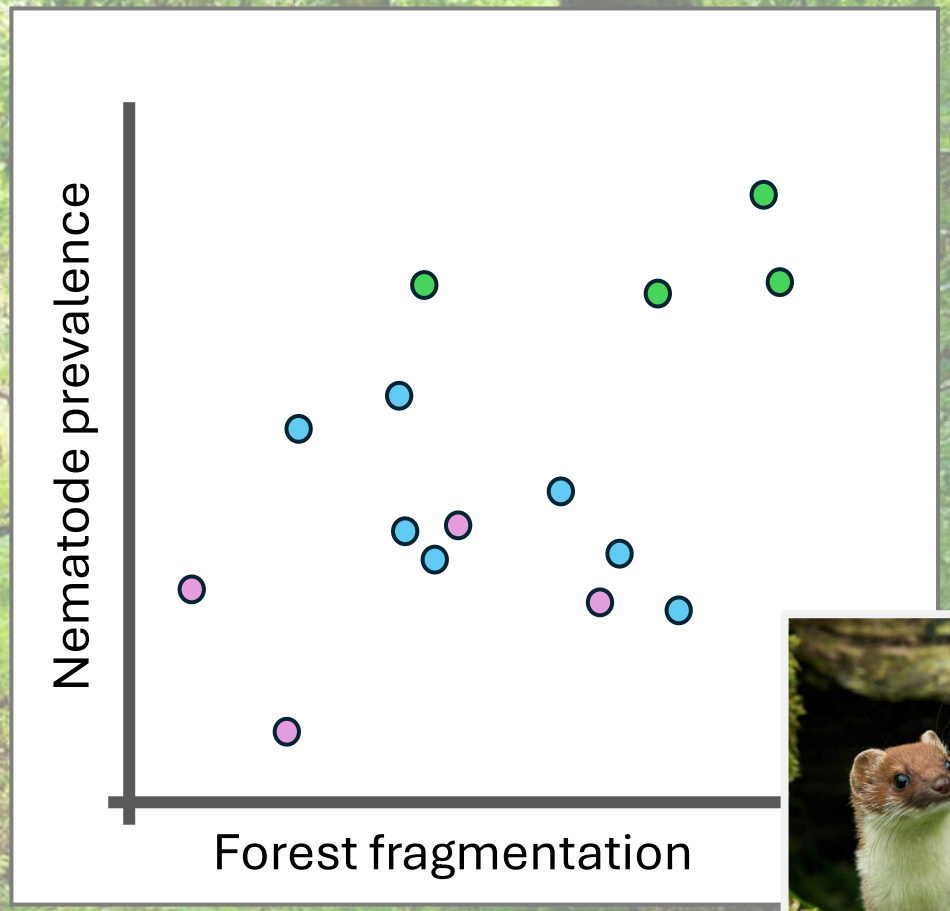
Species traits cluster in **phylogenetic** space due to shared evolutionary history

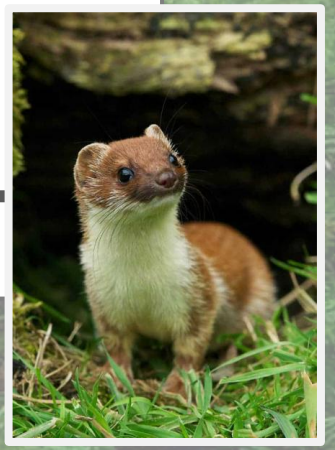
Autocorrelation

Sampling was **spatially clustered**: other unobserved factors could influence results

(e.g. host movement)







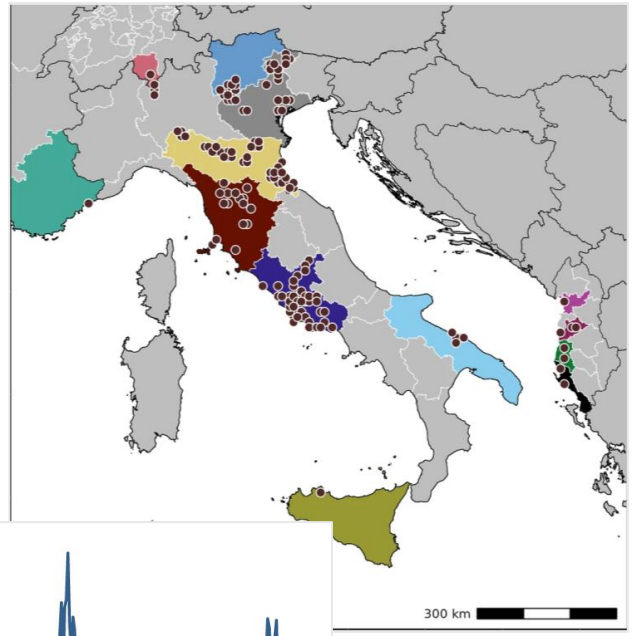
How does autocorrelation affect ecological and health data analysis?

In ecology/epidemiology we often study phenomena with **spatial or temporal structure**, from populations to epidemics
(and we almost never measure *everything* that could influence this structure)

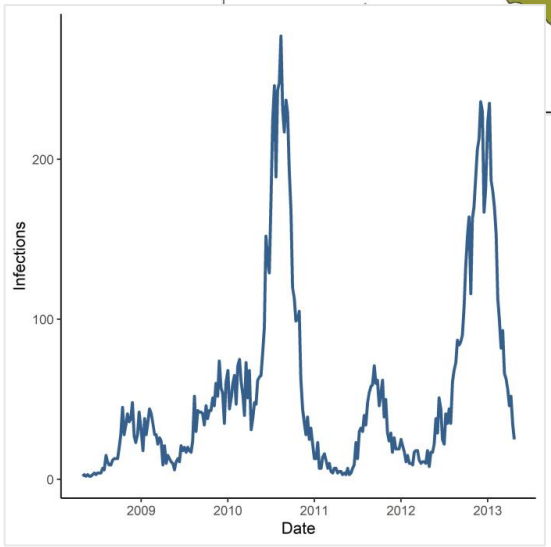
If not accounted for, can violate the assumption of independence of errors
(increased likelihood of false positive results, or confounding by unobserved variables)

This can be particularly problematic when sampling is clustered or biased
(as in many observational studies)

Health reports in admin. areas

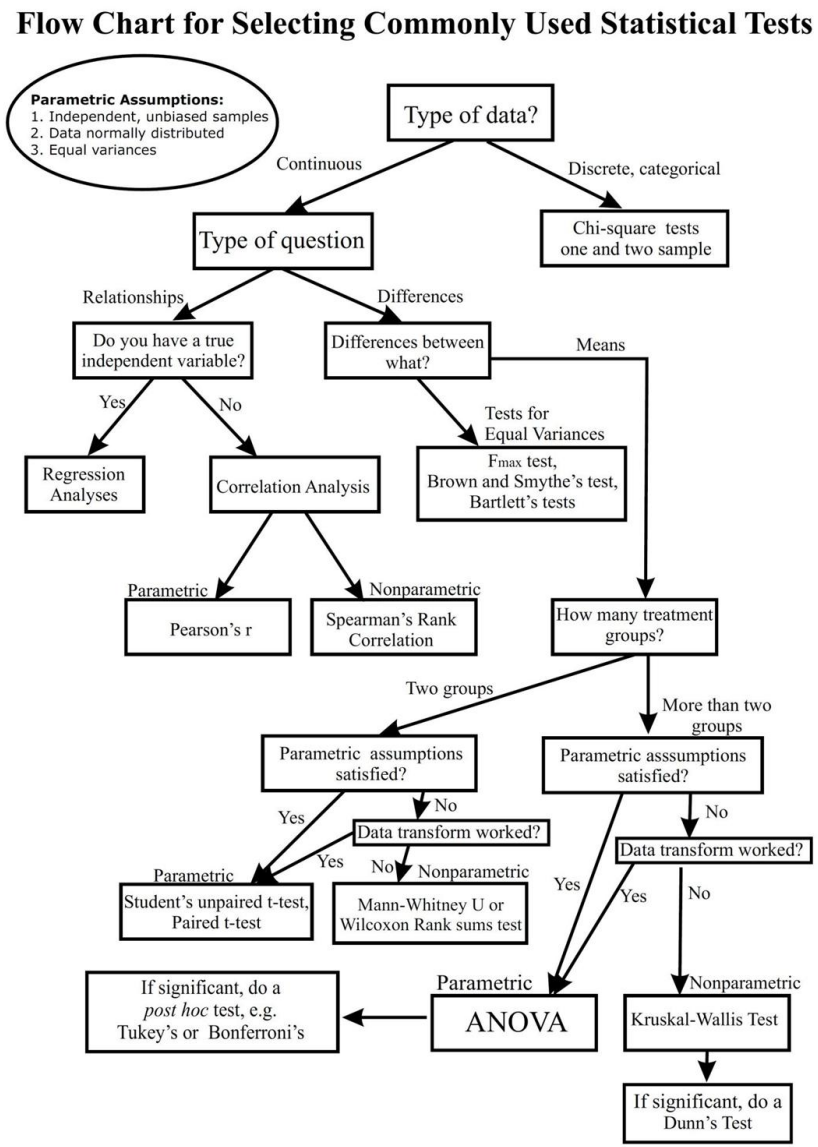


Ecological survey locations



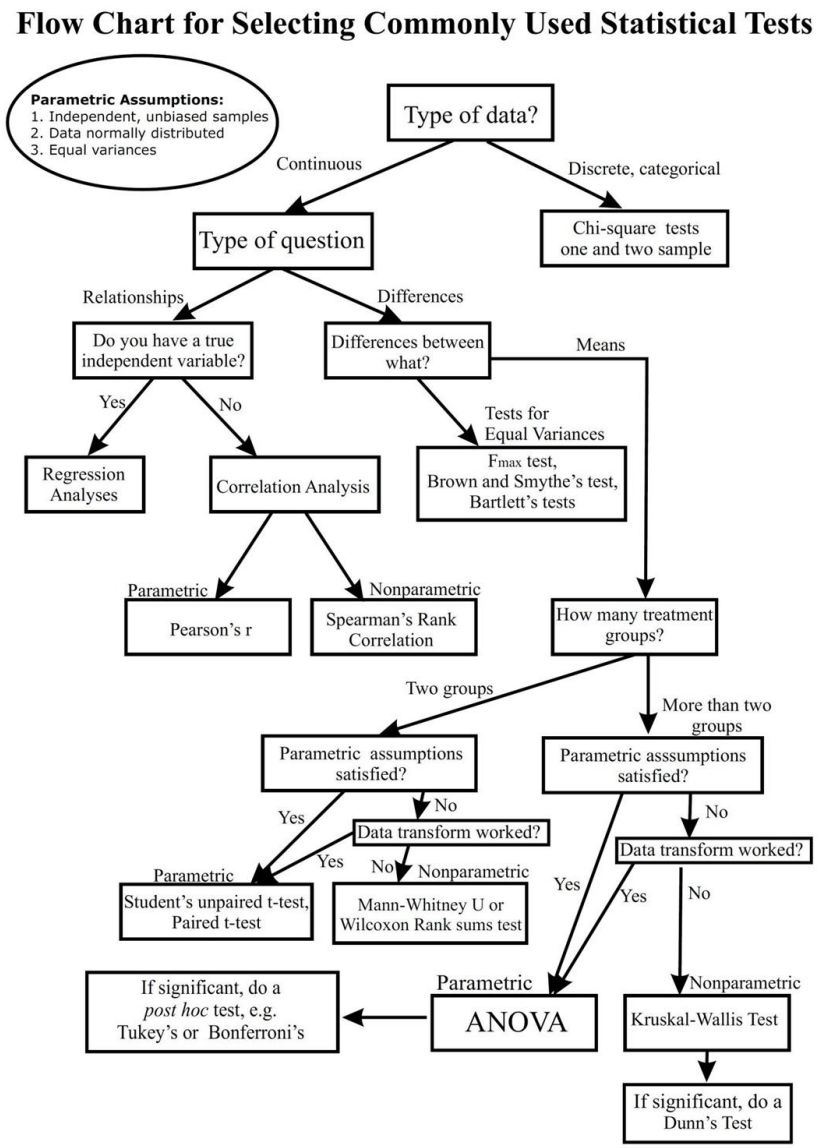
Weekly disease surveillance reports

The problem with “cookbook” statistics for complex systems



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

The problem with “cookbook” statistics for complex systems



This approach ignores the most important things that help us design an appropriate analysis!

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

(2) How the data were generated

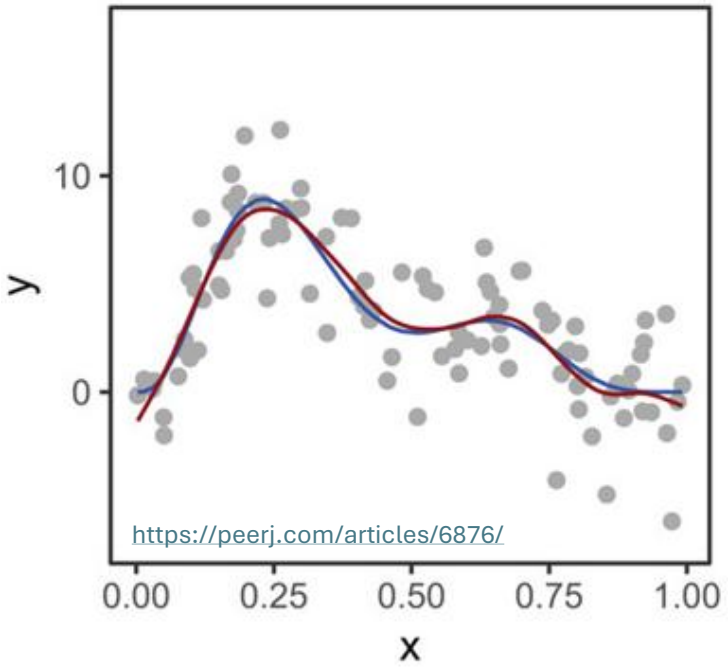
https://twitter.com/TrevorABranch/status/1082135735757504512

Statistical modelling tools for complex real-world data

Statistical modelling tools for complex real-world data

Nonlinear

(e.g. generalised additive models)



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

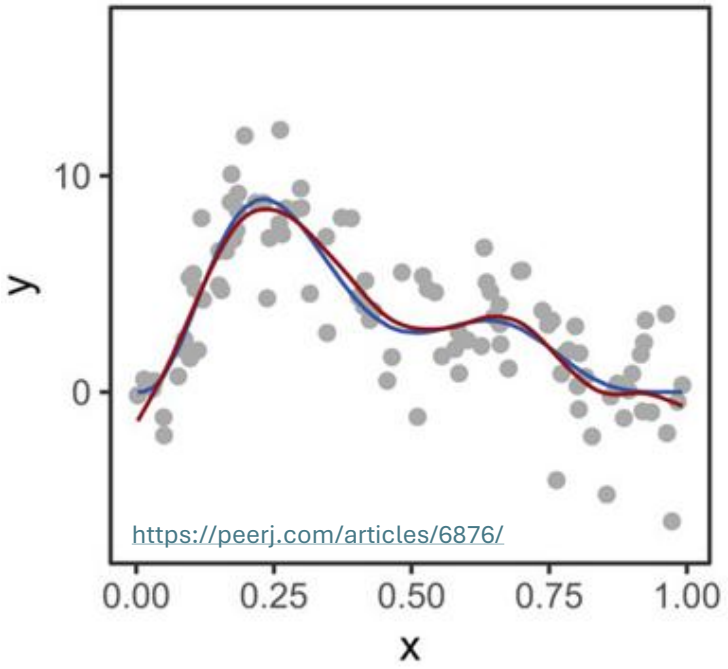
Nonlinear function of covariates X

(e.g. penalised splines)

Statistical modelling tools for complex real-world data

Nonlinear

(e.g. generalised additive models)

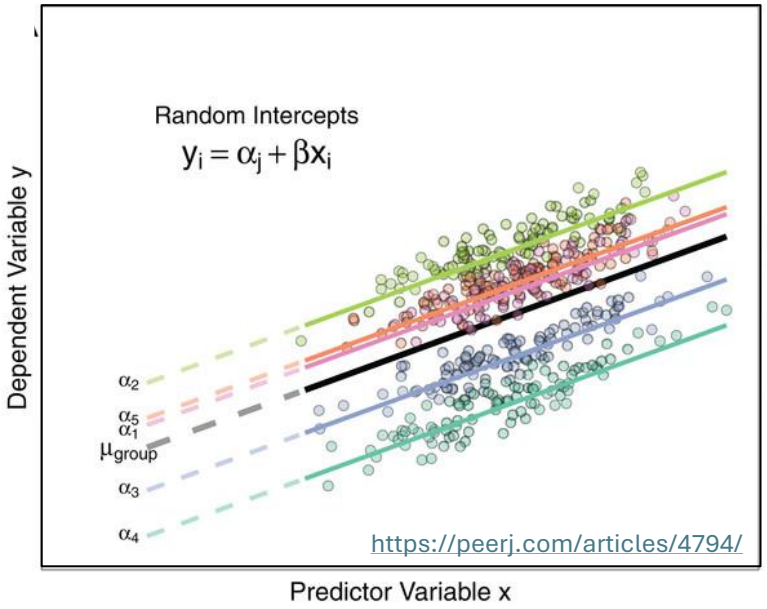


$$Y_i = f(\eta_i) = \beta_0 + f(X_{i1})$$

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

Multilevel (mixed effects)

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



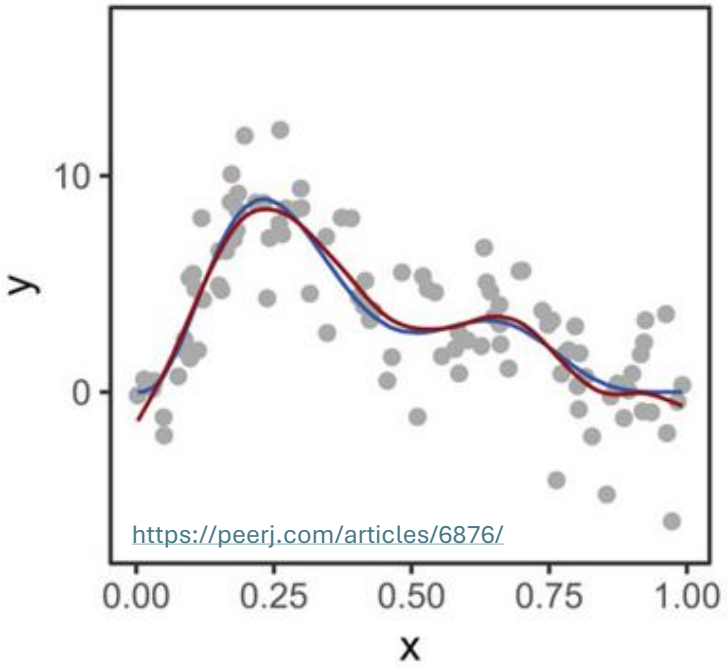
$$Y_i = f(\eta_i) = \beta_0 + \beta_1 X_{i1} + \alpha_{s(i)}$$

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

Statistical modelling tools for complex real-world data

Nonlinear

(e.g. generalised additive models)

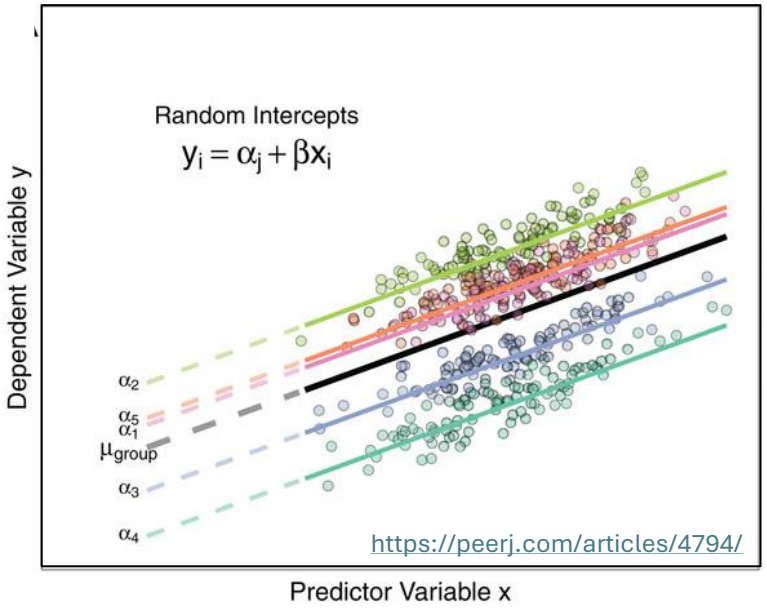


$$Y_i = f(\eta_i) = \beta_0 + f(X_{i1})$$

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

Multilevel (mixed effects)

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

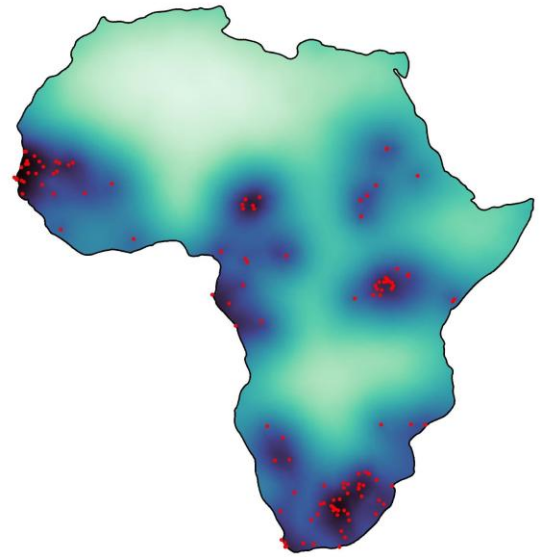


$$Y_i = f(\eta_i) = \beta_0 + \beta_1 X_{i1} + \alpha_{s(i)}$$

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)

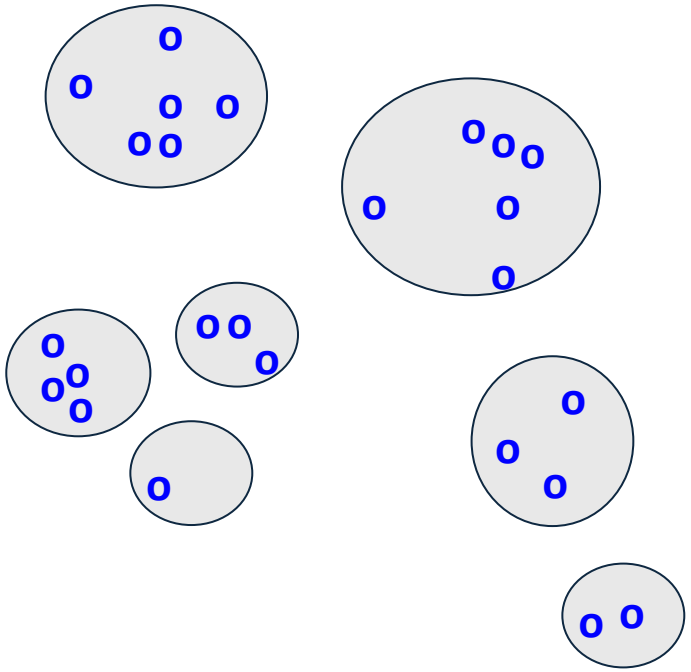


$$Y_i = f(\eta_i) = \beta_0 + \beta_1 X_{i1} + s_i$$

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

Mixed-effects (multilevel/hierarchical) models

Ecological study of bird abundance across an island archipelago

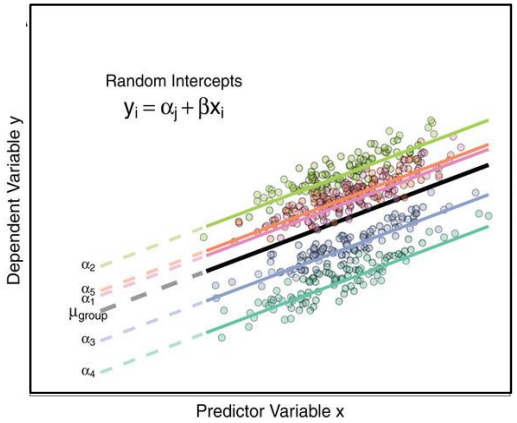


O = Abundance record of bird species X
Sampled across 7 islands *s* (s = 1…7)

Data often contain some clustered/nested structure (e.g. space, time, phylogeny, individual, population)

Observations <u>within</u> each cluster may be non-independent - account for this by allowing intercepts/slopes to vary between clusters (*"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?)

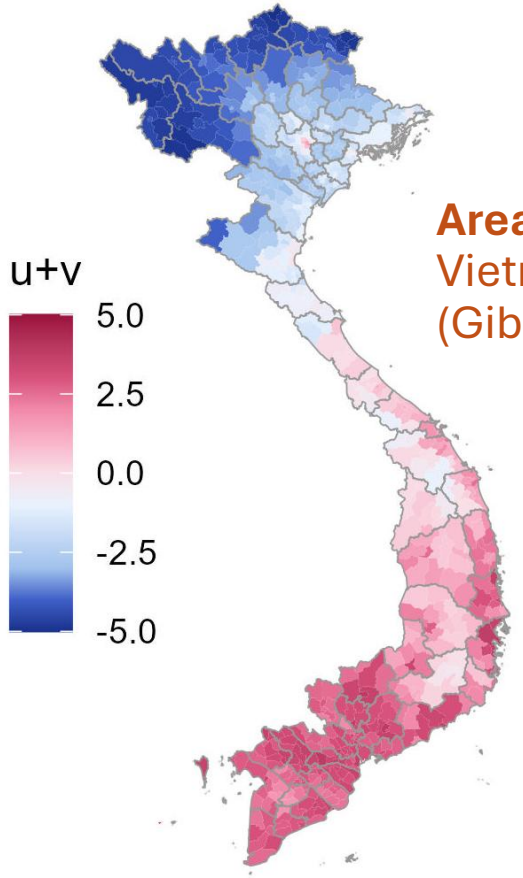
A model within a model - hence "hierarchical"!
(σ is a '*hyperparameter*')

Spatial statistics

Develop models that **explicitly account for the spatial dependency** among nearby observations.

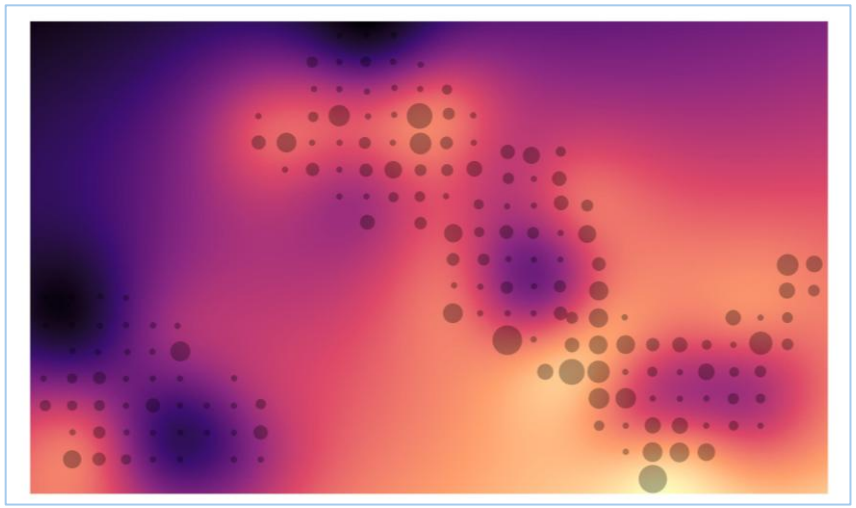
Key assumption: closer proximity = more similar.

Huge branches of statistics focused on modelling both **areal** and **geospatial** data - critical tools for both ecology and health research.



Areal data: dengue incidence in Vietnamese districts (Gibb et al 2023, *Nat. Comms*.)

Geospatial data: species detections at camera trap locations (Maasai Mara)

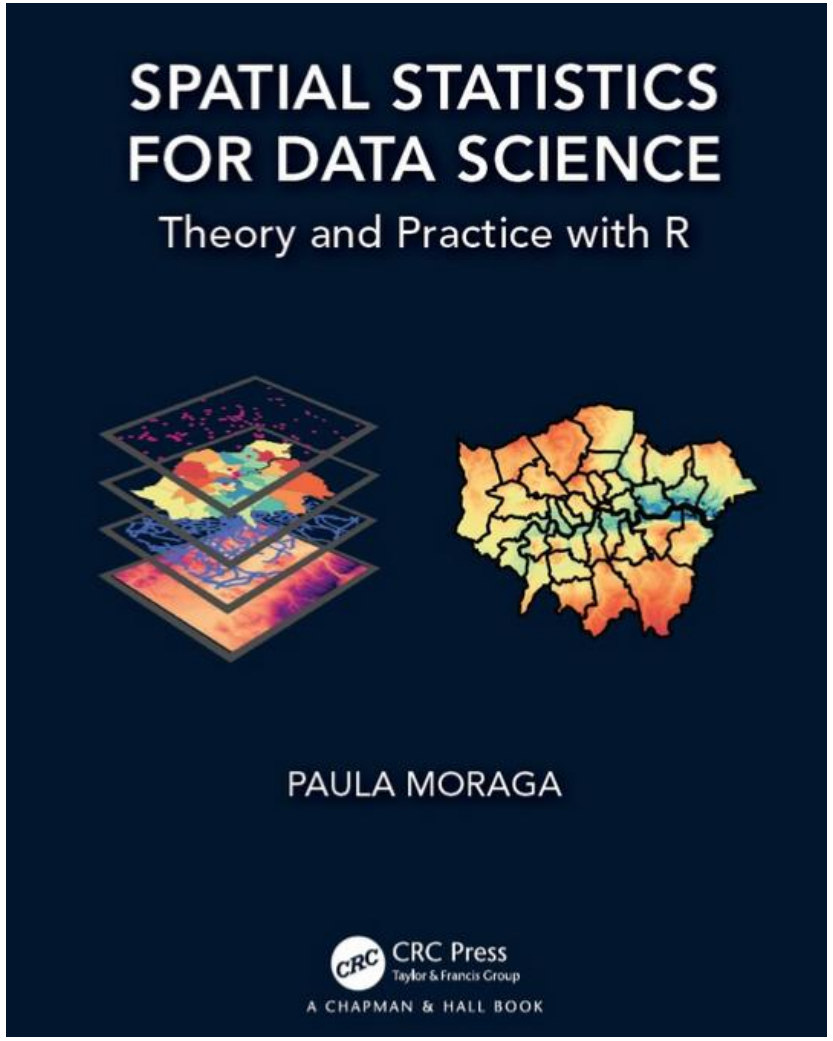


Spatial statistics

Develop models that **explicitly account for the spatial dependency** among nearby observations.

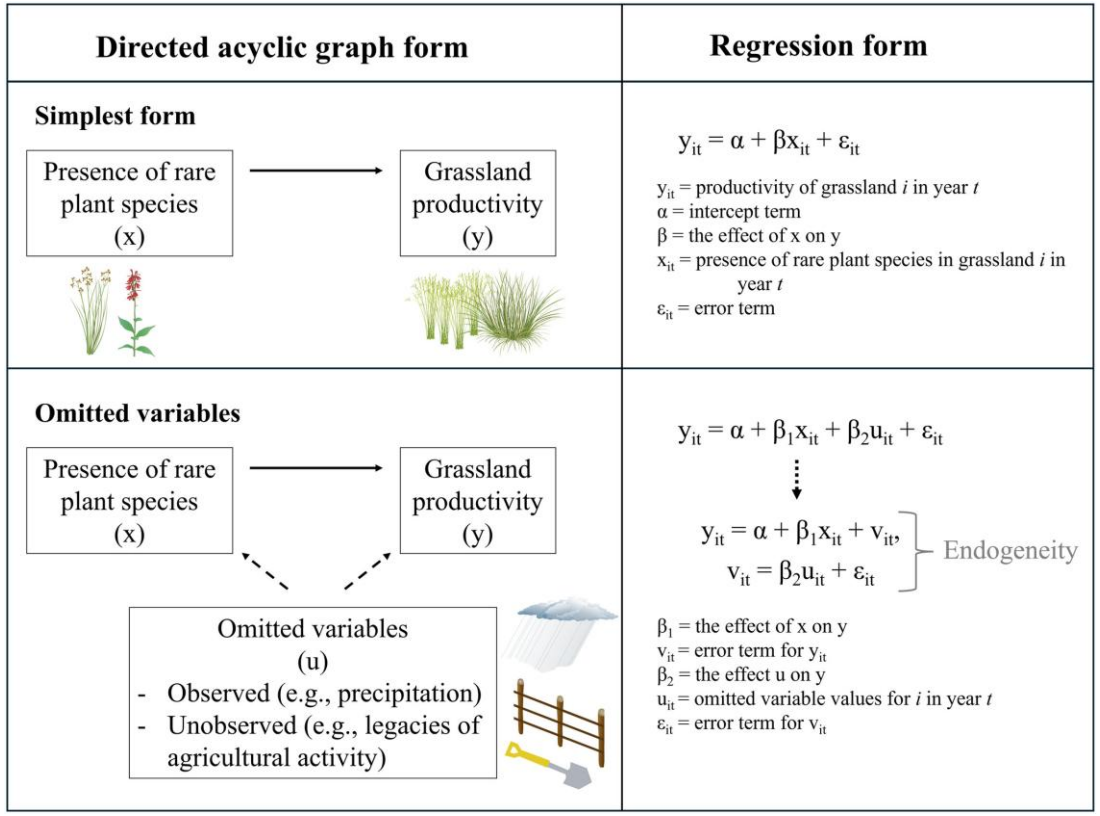
Key assumption: closer proximity = more similar.

Huge branches of statistics focused on modelling both **areal** and **geospatial** data - critical tools for both ecology and health research.



https://www.paulamoraga.com/book-spatial/

Causal inference methods



Siegel & Dee, 2025, *Ecology Letters*

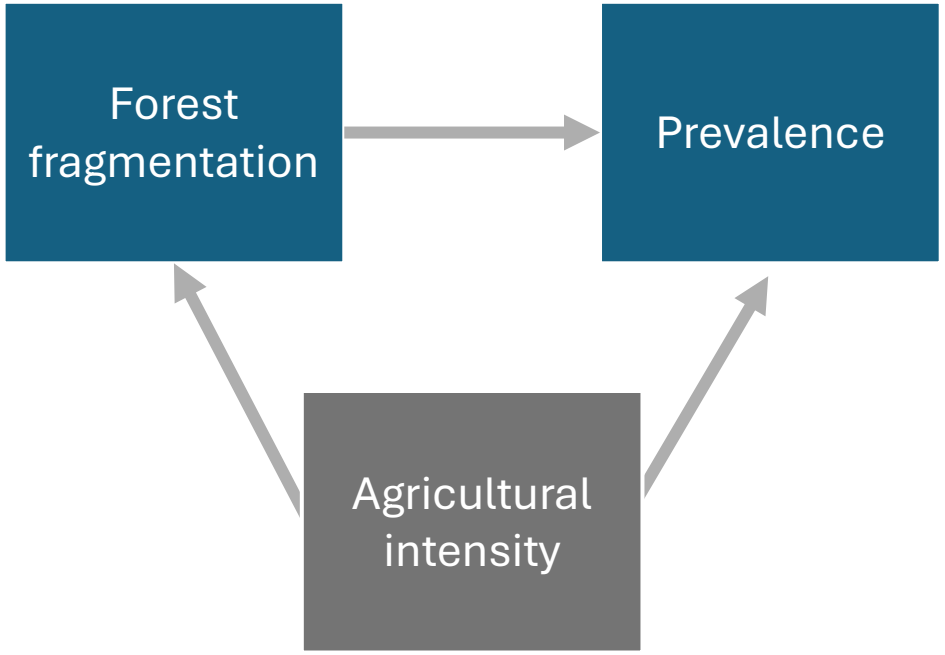
Growing body of methods, largely from epidemiology and econometrics, for **inferring causes from observational data**

Represent relationships among variables (measured and unmeasured) as **causal diagrams** (DAGs), and use these to design models

Huge topic and growing importance for ecology and biodiversity research.

Causal inference methods

A very simple **directed acyclic graph (DAG)**
(causal diagram)



Growing body of methods, largely from epidemiology and econometrics, for **inferring causes from observational data**

Represent relationships among variables (measured and unmeasured) as **causal diagrams** (DAGs), and use these to design models

Huge topic and growing importance for ecology and biodiversity research.

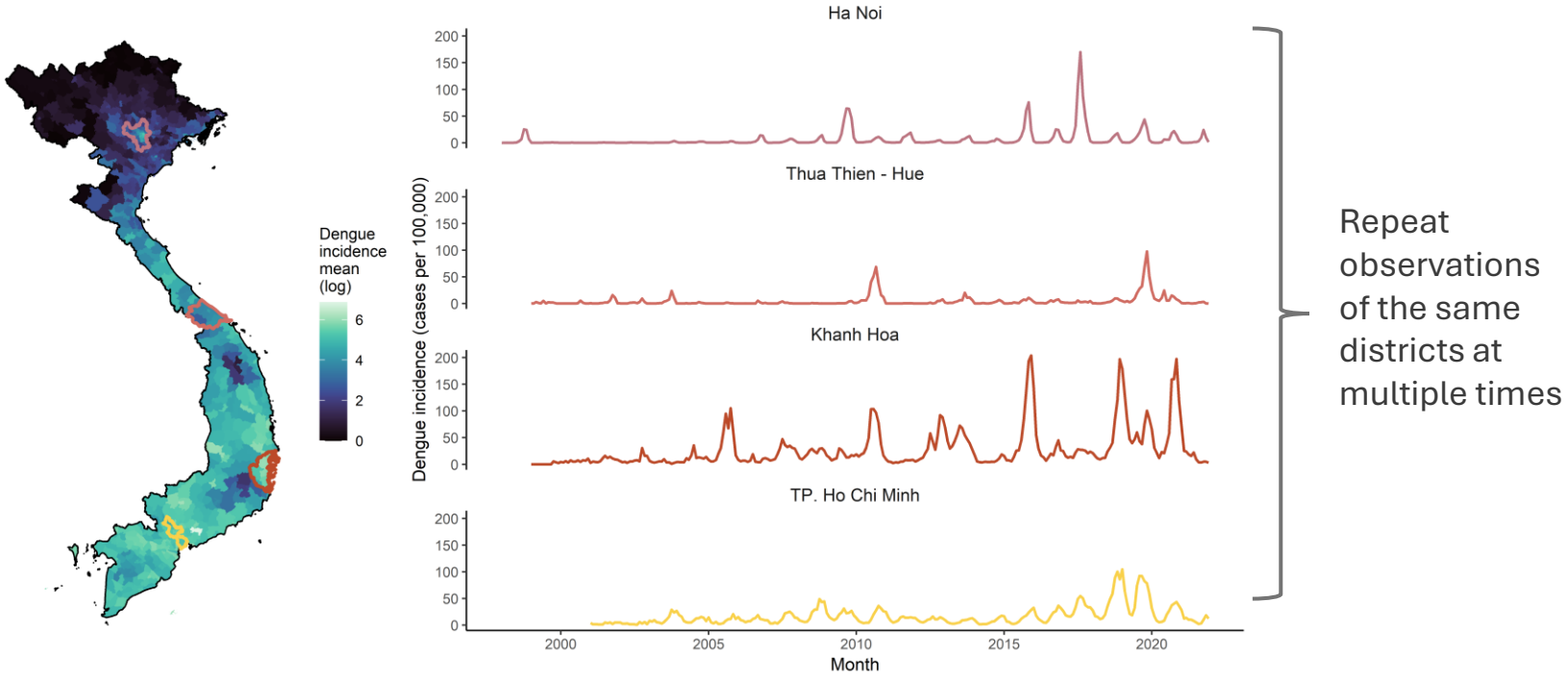
Panel data – repeat observations of same units over time

The same units **observed repeatedly over time** *(e.g. individuals, locations)* – provides information on **within- and between-unit variation**

Very useful for addressing the issue of unmeasured confounders, and for studying the influence of dynamic drivers (e.g. climate)

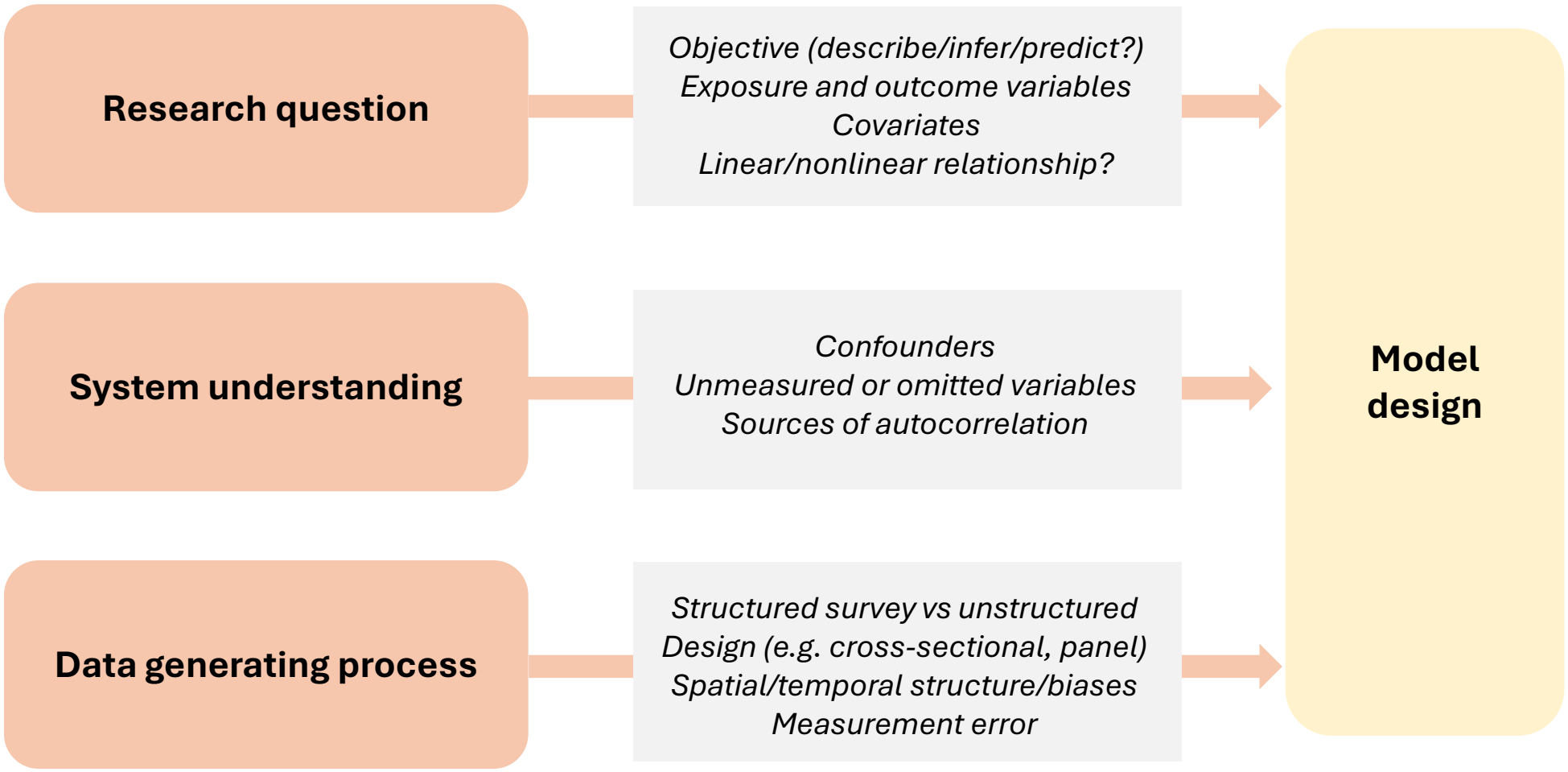
Common in health datasets – rarer, but useful, in ecology

Longitudinal health surveillance data: dengue in Vietnam



Gibb et al. 2023, *Nature Communications*

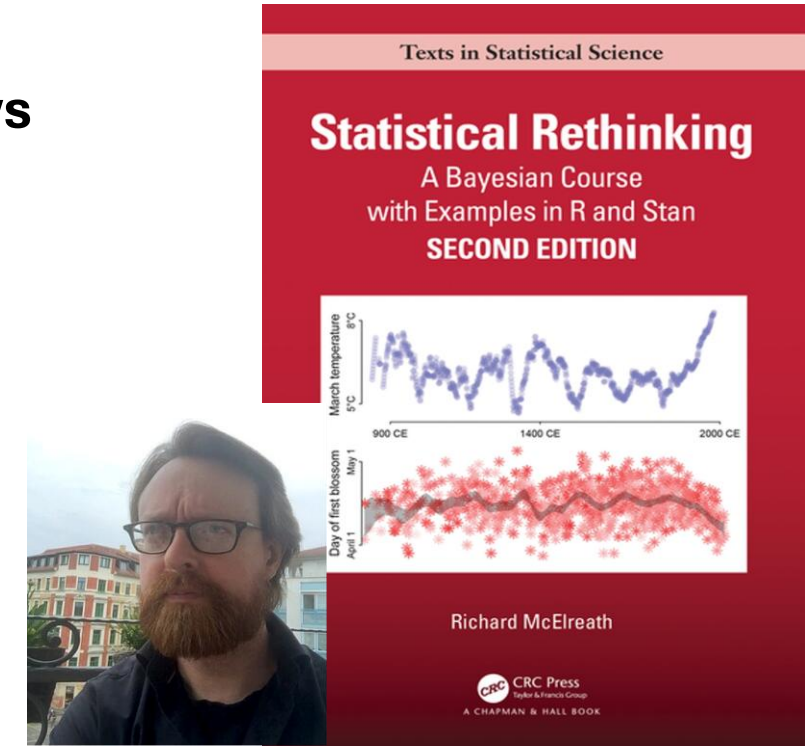
From question, to data, to model design



From question, to data, to model design

Statistical models are powerful machines for helping us to understand and predict the world, but lack insight.

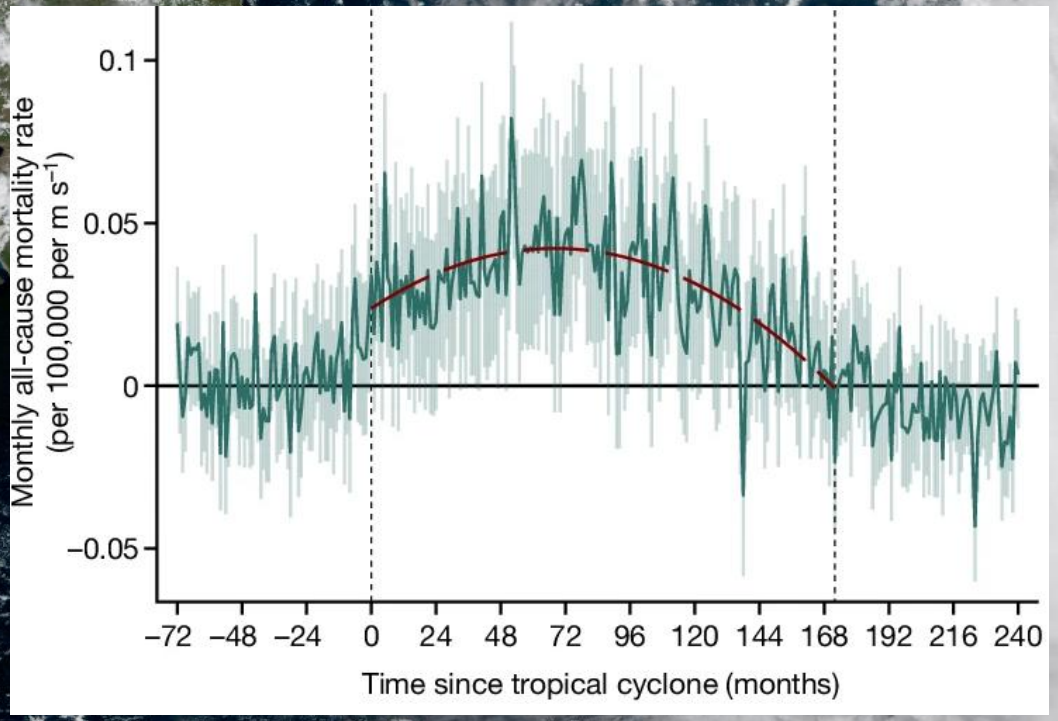
A model only knows what you tell it!



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

Innovations in inference for ecology, climate change and health

Innovations in econometrics – attributing climate impacts on health



Econometric methods – attributing all-cause mortality (direct and indirect) to tropical cyclones using data since 1930

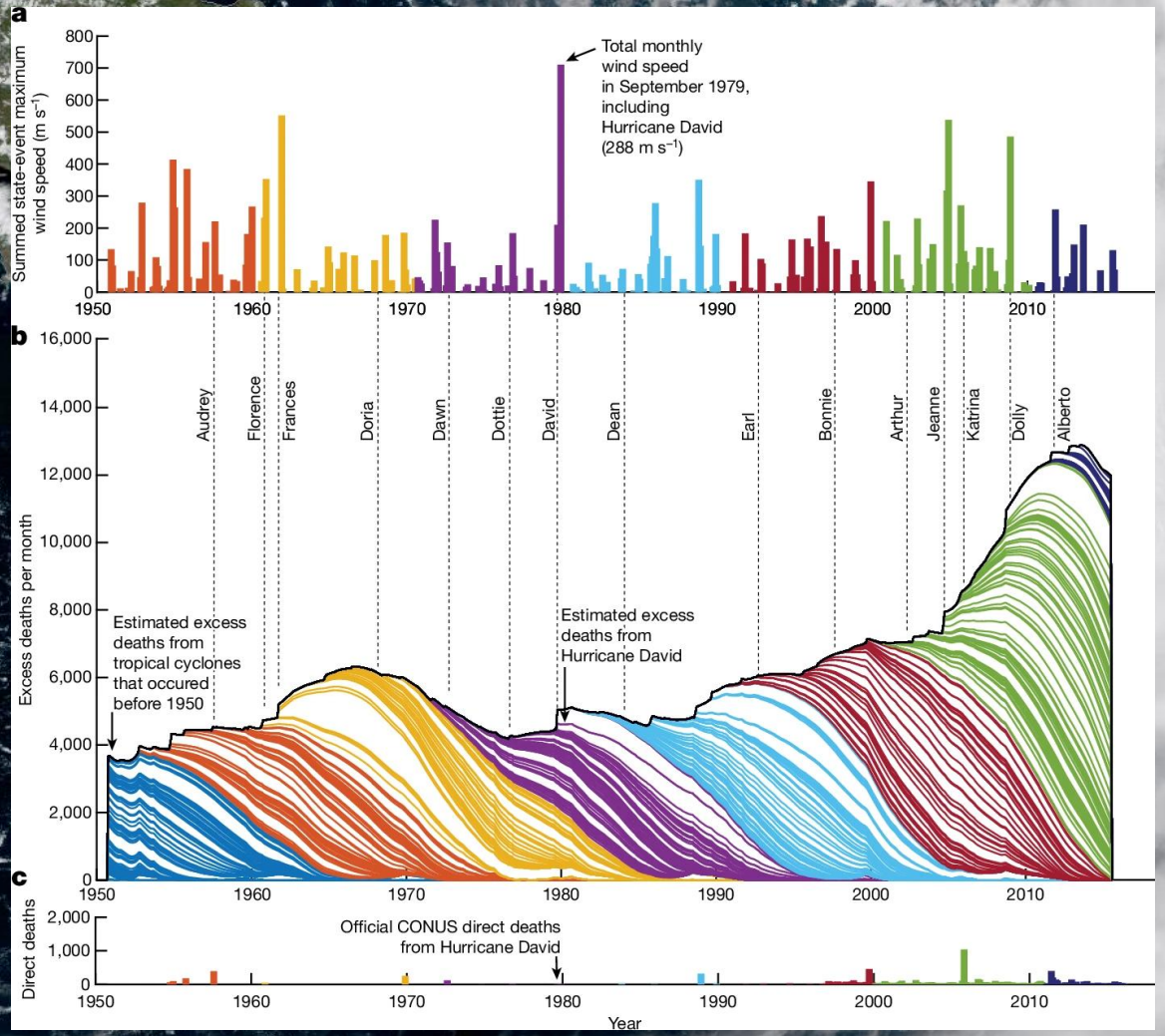
Article

Mortality caused by tropical cyclones in the United States

Young & Hsiang, *Nature*, 2024

NOAA

Innovations in econometrics – attributing climate impacts on health



Econometric methods – attributing all-cause mortality (direct and indirect) to tropical cyclones using data since 1930

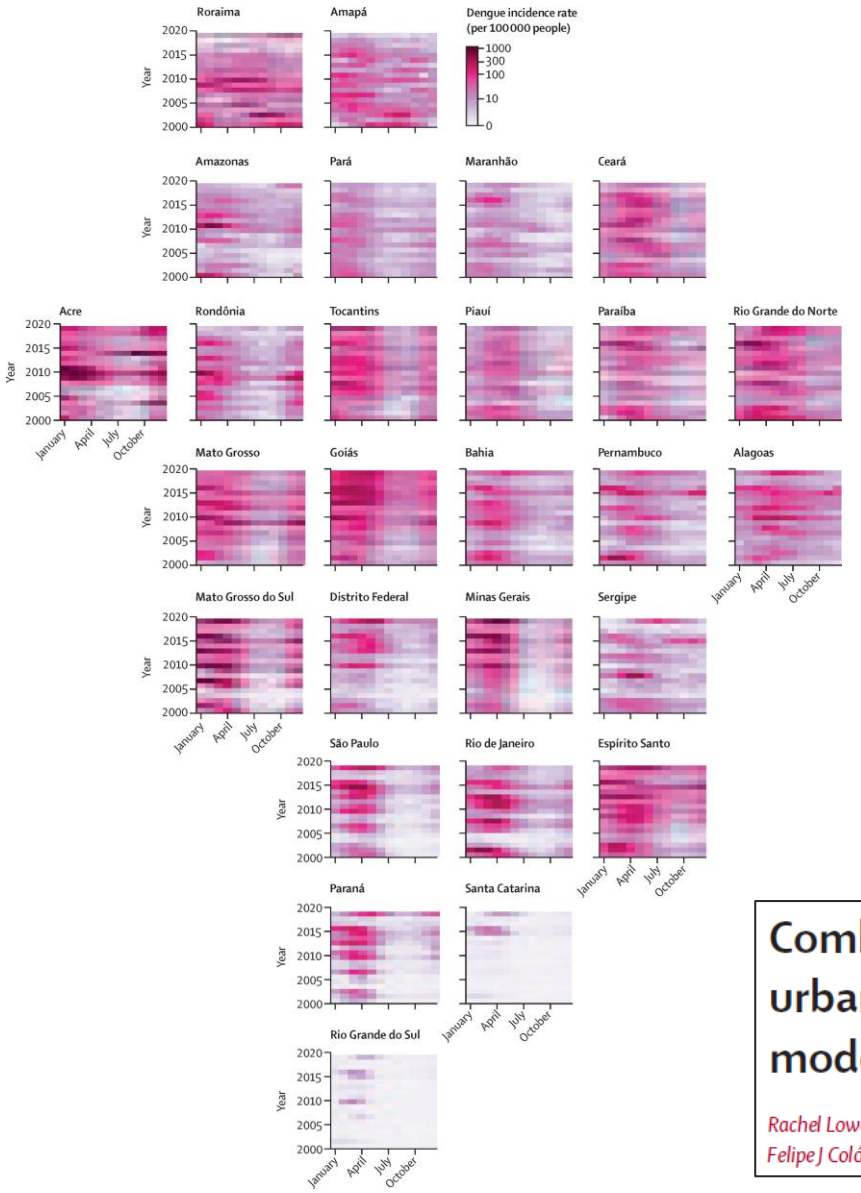
Massive **indirect mortality** estimate – up to 5% of deaths in US Gulf of Mexico coastal states

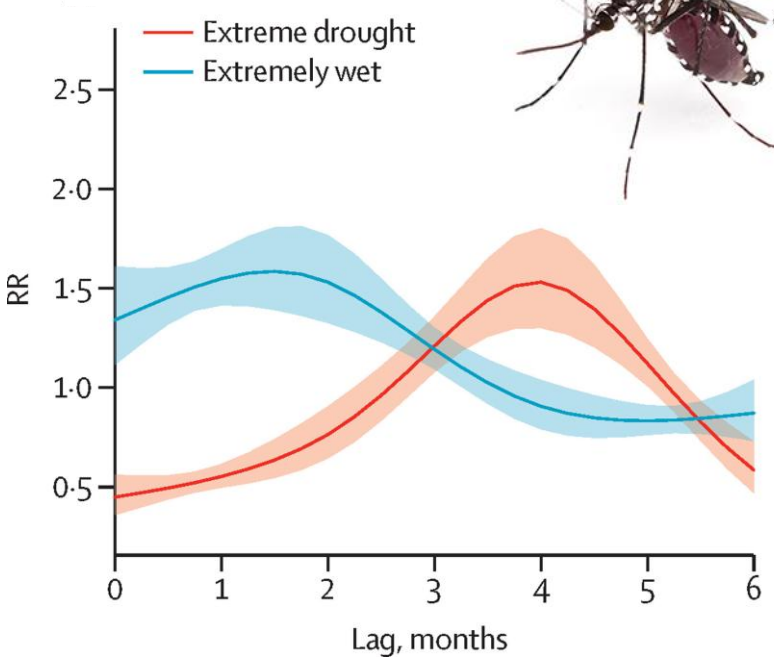
Article

Mortality caused by tropical cyclones in the United States

Young & Hsiang, *Nature*, 2024

Innovations in epidemiology – delayed impacts of drought on dengue





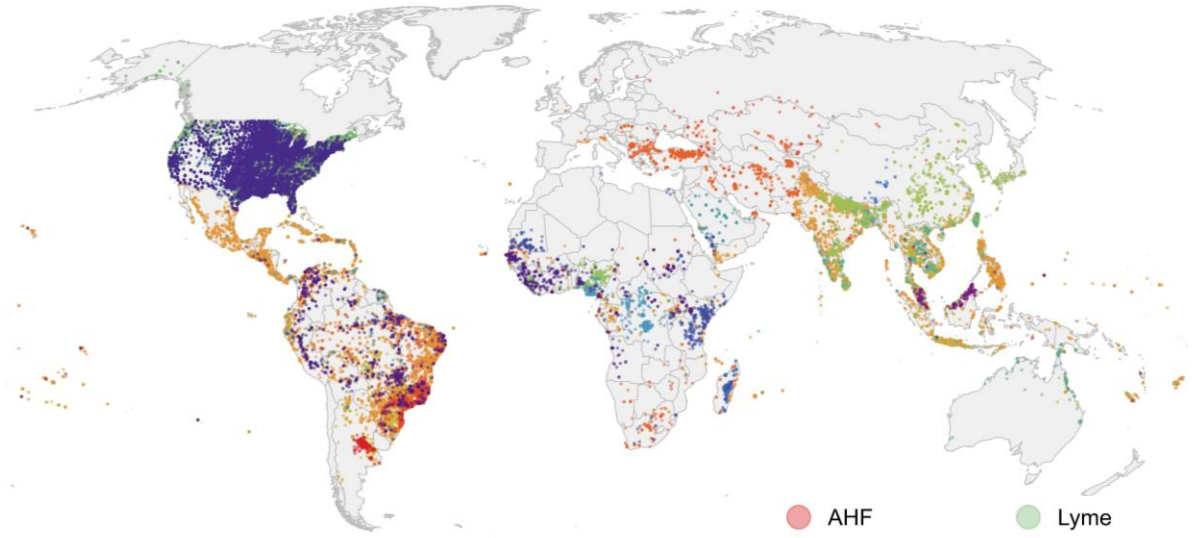
Distributed-lag nonlinear models – measure covariate effects at multiple time delays

Dengue incidence surges **immediately after extreme rain**, and **several months after extreme drought**

Combined effects of hydrometeorological hazards and urbanisation on dengue risk in Brazil: a spatiotemporal modelling study

Rachel Lowe, Sophie A Lee, Kathleen M O'Reilly, Oliver J Brady, Leonardo Bastos, Gabriel Carrasco-Escobar, Rafael de Castro Catão, Felipe J Colón-González, Christovam Barcellos, Marilia Sá Carvalho, Marta Blangiardo, Håvard Rue, Antonio Gasparrini

Innovations in ecology – measuring drivers of zoonotic spillover



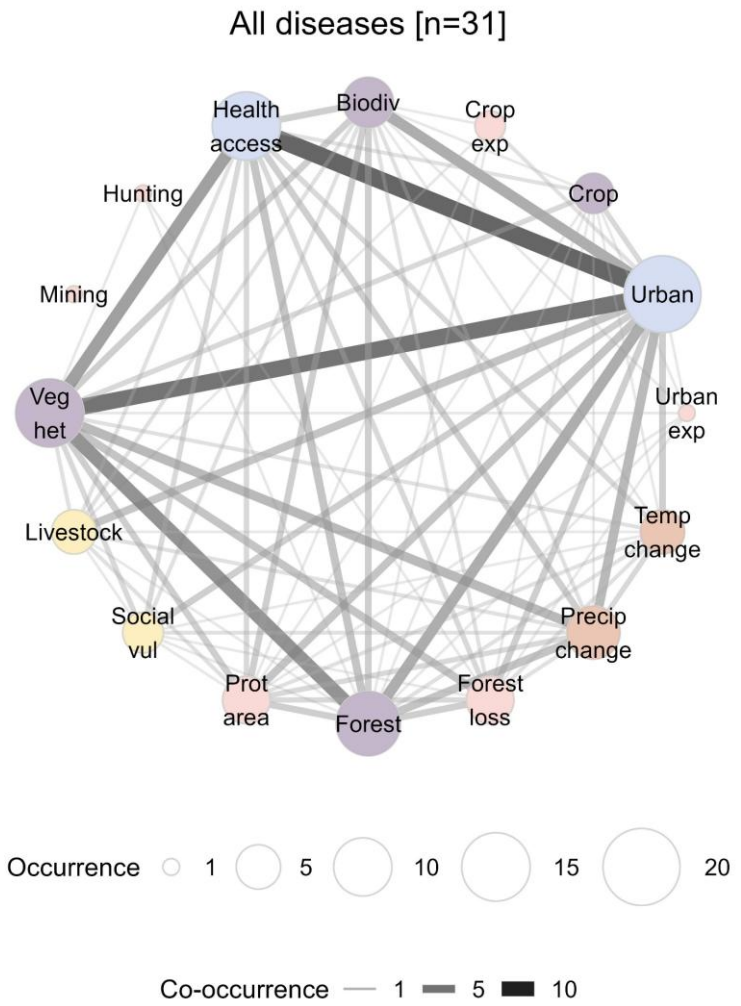
How general are **drivers of zoonotic outbreaks** across diverse diseases?

Huge dataset of outbreak data (32 diseases) + **geospatial driver inference** framework

The anthropogenic fingerprint on emerging infectious diseases

Rory Gibb, Sadie J. Ryan, David Pigott, Maria del Pilar Fernandez, Renata L. Muylaert, Gregory F. Albery, Daniel J. Becker, Jason K. Blackburn Hernan Caceres-Escobar, Michael Celone, Evan A. Eskew, Hannah K. Frank,

Innovations in ecology – measuring drivers of zoonotic spillover



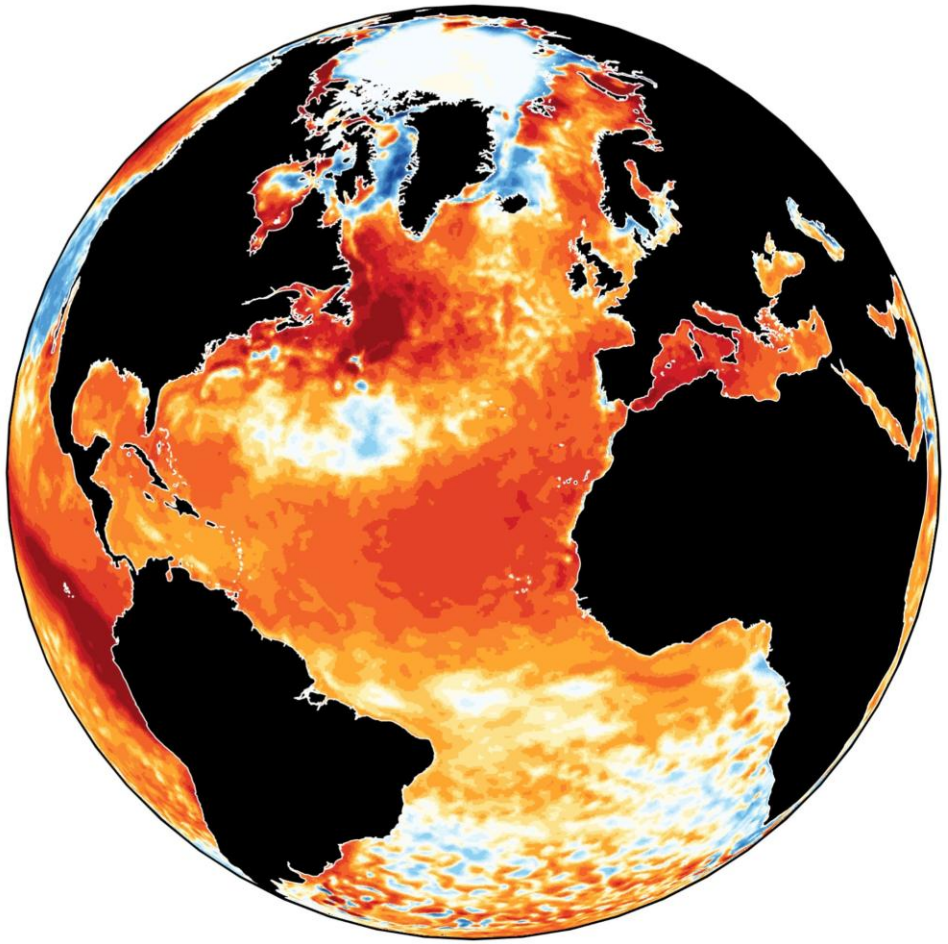
General ecological drivers are rare – most drivers are disease-specific.

Most prevalent "drivers" relate to health systems and detection biases – **we observe outbreaks in cities and close to clinics!**

The anthropogenic fingerprint on emerging infectious diseases

Rory Gibb, Sadie J. Ryan, David Pigott, Maria del Pilar Fernandez, Renata L. Muylaert, Gregory F. Albery, Daniel J. Becker, Jason K. Blackburn Hernan Caceres-Escobar, Michael Celone, Evan A. Eskew, Hannah K. Frank,

In summary…



- Often in ecology and health we are limited to working with **observational data** – cannot always easily (or ethically) experimentally manipulate systems!
- **Measuring the influence of environmental change using these data poses significant challenges**, particularly for causal inference – for example, confounding, observation biases, and autocorrelation.
- **Nonlinear, spatial and hierarchical models** can provide powerful frameworks to account for these issues in analysis design.
- Modelling analysis design needs to consider the **research question**, the **wider system context**, and the **data-generating process**!

Questions?