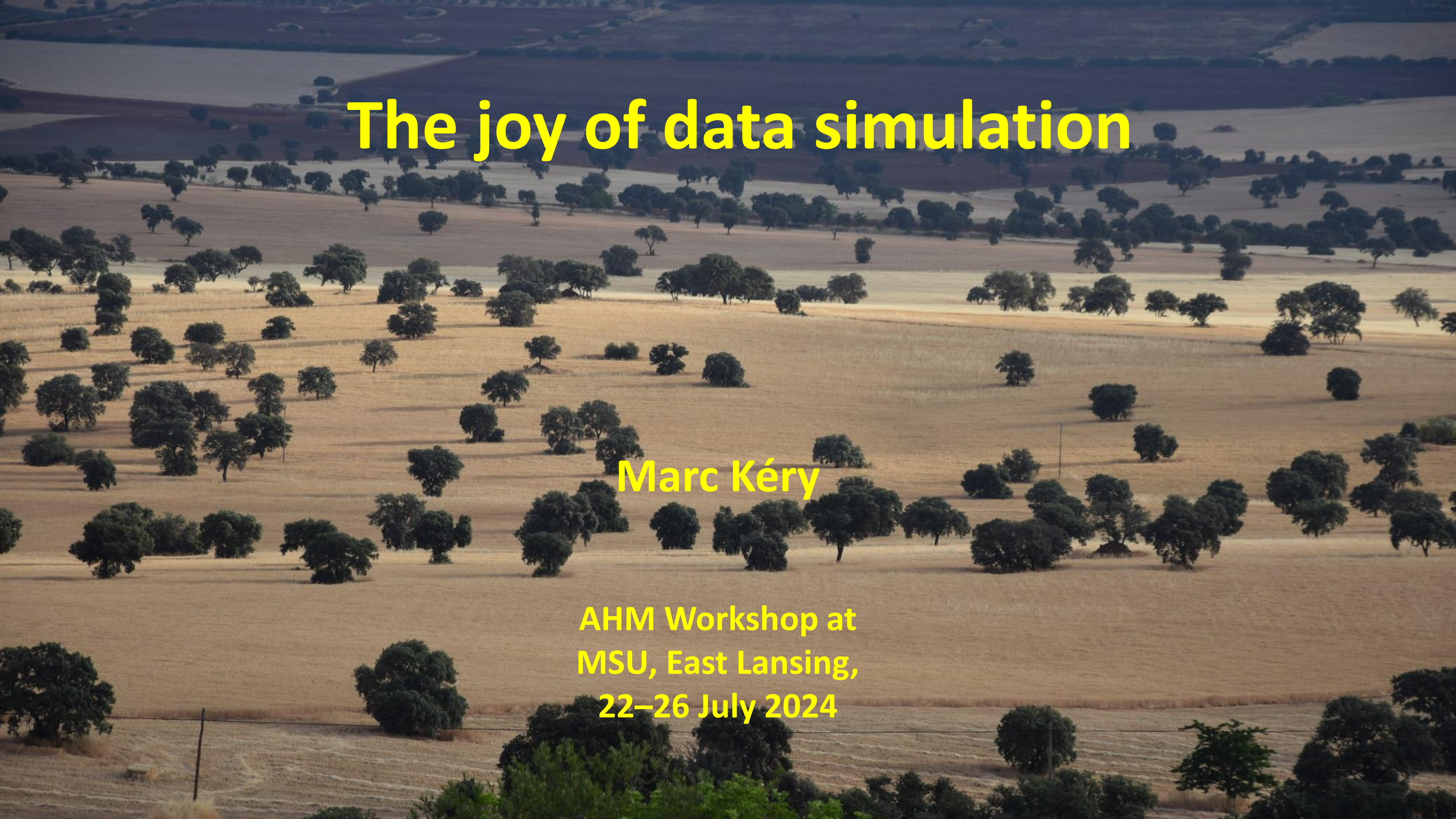


The joy of data simulation

Marc Kéry

AHM Workshop at
MSU, East Lansing,
22–26 July 2024

The joy of data simulation

a.k.a. "The experimental approach to statistics"

Marc Kéry

AHM Workshop at
MSU, East Lansing,
22–26 July 2024

Overview

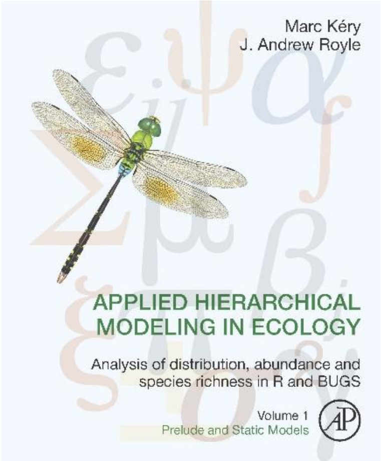
- What is data simulation
- Duality of fitting a model and simulating data under it
- Lots of benefits of data simulation
- Three examples:

  (1) Understanding a difficult concept in statistics

  (2) Assessing parameter identifiability

  (3) Power analysis

- See also chapter 4 in AHM1 book (handout for module)


Marc Kéry
J. Andrew Royle
APPLIED HIERARCHICAL MODELING IN ECOLOGY
Analysis of distribution, abundance and species richness in R and BUGS
Volume 1
Prelude and Static Models

What is data simulation

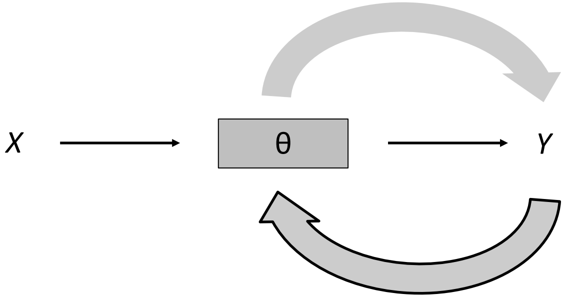
- Producing realizations from the stochastic process that is represented by a statistical model
- That is, generate random numbers from a model => use statistical model as a RNG
- [Can do this very formally in program NIMBLE: simulate under your model]
- Sometimes called "fake data simulation" ….

  Better avoid this term, since it has negative connotation …

  While data simulation is hugely important in applied statistics
- Can't do statistics without simulating data

Duality of model fitting and data simulation

Data simulation



X → θ → Y

Data analysis

- Same thing, but in opposite directions:
    - Data set + model given => can get parameter estimates
    - Model + parameter values given => can get data set(s)

Many benefits of data simulation

- Truth is known: can validate model and computer code
- Understand sampling error (=variability among different realizations of data set):

  e.g., WTF is a standard error ?
- Check frequentist operating characteristics of estimators:

  e.g. bias, precision
- Check parameter identifiability and estimability of paramaters in given data sets:

  i.e., do my data contain the necessary information to estimate my parameter ?
- Power analysis:

  e.g., how many sites do I have to survey to detect a trend ?
- Assess goodness-of-fit of your model:

  e.g., posterior predictive checks (Bayes) or parametric bootstrap (Freq.)
- Use for model selection via Cross-validation:

  i.e., simulate left-out data set and see which model predicts best

Many benefits of data simulation

- Check the robustness of model to violations of its assumptions:

  e.g., simulate data under a more complex model and then fit a simpler model
  [i.e., distinguish 'data generating model' from 'data analysis model']

- Calibrate derived parameters and data, also tune priors (prior predictive checks):

  see what data are implied by different values of the parameters

- Prove your understanding of a model:
  - i.e., if you cannot simulate data under your model, then you have probably not understood your model
  - Writing R code to simulate data under your model is one of the best ways to learn a new model and really understand what it means

Three examples

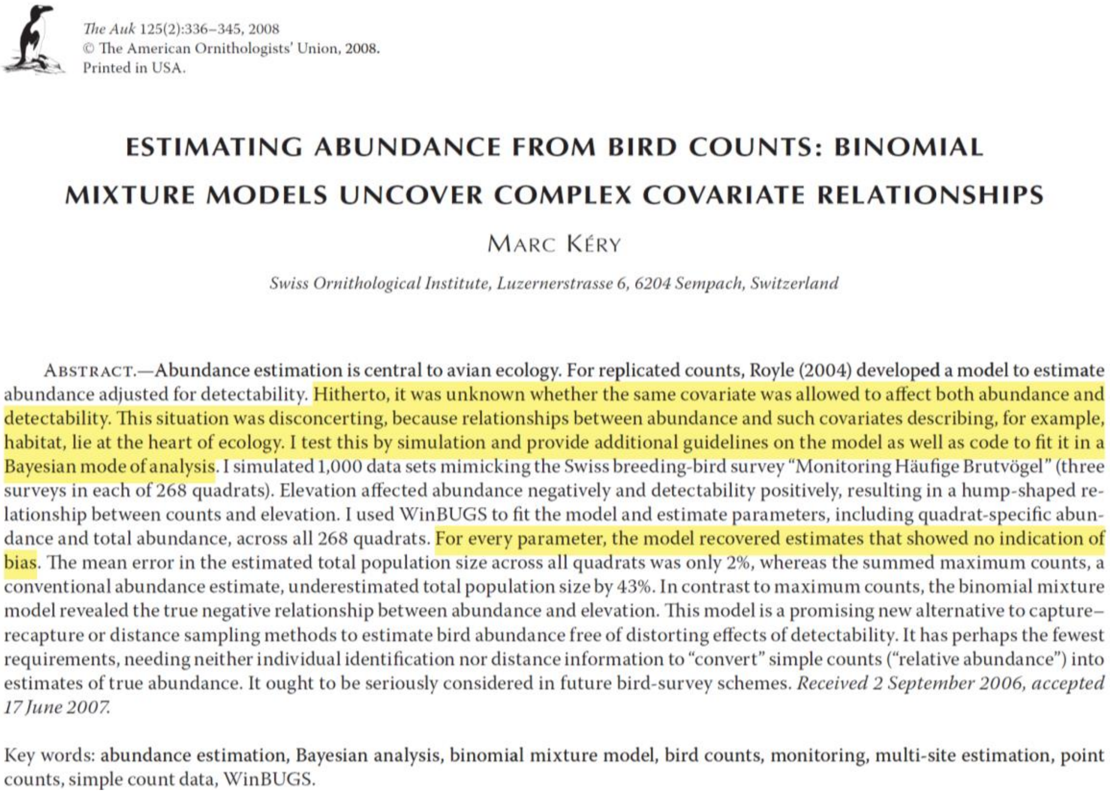
- (1) Understanding a difficult thing in statistics: wtf is a standard error ?
- **(2) Understanding parameter identifiability**
- (3) Power analysis

- Repeatedly simulate data in R and then analyse them
- Often package data simulation code into an R function => can more easily vary important settings of simulation, e.g., sample sizes, parameter values

Can we have same covariate in state and detection ?

Can we have same covariate in state and detection ?

- Remember that …



The Auk 125(2):336–345, 2008
© The American Ornithologists' Union, 2008.
Printed in USA.

ESTIMATING ABUNDANCE FROM BIRD COUNTS: BINOMIAL MIXTURE MODELS UNCOVER COMPLEX COVARIATE RELATIONSHIPS

MARC KÉRY

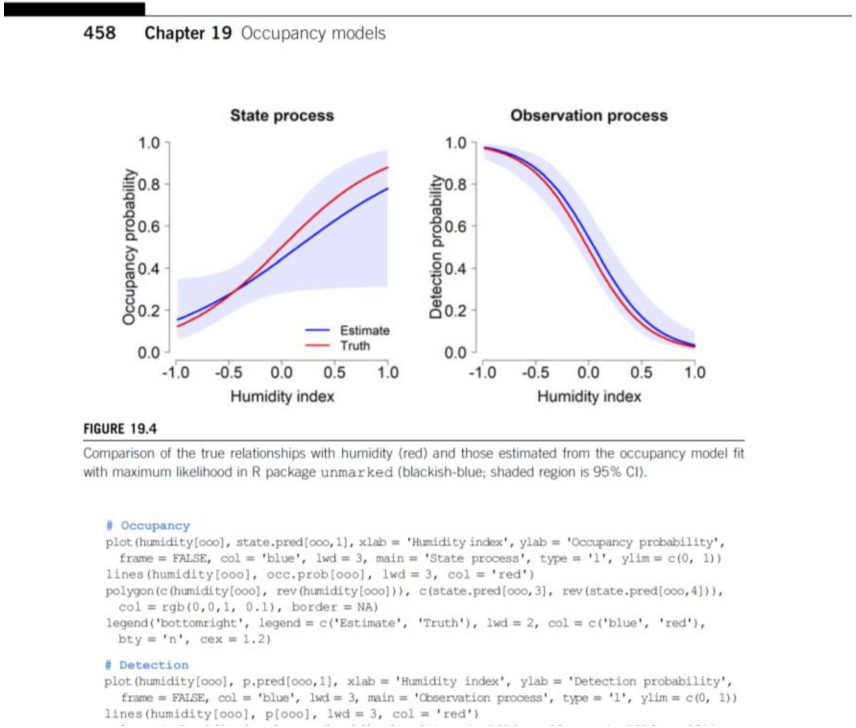
Swiss Ornithological Institute, Luzernerstrasse 6, 6204 Sempach, Switzerland

ABSTRACT.—Abundance estimation is central to avian ecology. For replicated counts, Royle (2004) developed a model to estimate abundance adjusted for detectability. Hitherto, it was unknown whether the same covariate was allowed to affect both abundance and detectability. This situation was disconcerting, because relationships between abundance and such covariates describing, for example, habitat, lie at the heart of ecology. I test this by simulation and provide additional guidelines on the model as well as code to fit it in a Bayesian mode of analysis. I simulated 1,000 data sets mimicking the Swiss breeding-bird survey "Monitoring Häufige Brutvögel" (three surveys in each of 268 quadrats). Elevation affected abundance negatively and detectability positively, resulting in a hump-shaped relationship between counts and elevation. I used WinBUGS to fit the model and estimate parameters, including quadrat-specific abundance and total abundance, across all 268 quadrats. For every parameter, the model recovered estimates that showed no indication of bias. The mean error in the estimated total population size across all quadrats was only 2%, whereas the summed maximum counts, a conventional abundance estimate, underestimated total population size by 43%. In contrast to maximum counts, the binomial mixture model revealed the true negative relationship between abundance and elevation. This model is a promising new alternative to capture–recapture or distance sampling methods to estimate bird abundance free of distorting effects of detectability. It has perhaps the fewest requirements, needing neither individual identification nor distance information to "convert" simple counts ("relative abundance") into estimates of true abundance. It ought to be seriously considered in future bird-survey schemes. *Received 2 September 2006, accepted 17 June 2007.*

Key words: abundance estimation, Bayesian analysis, binomial mixture model, bird counts, monitoring, multi-site estimation, point counts, simple count data, WinBUGS.

Can we have same covariate in state and detection ?

- And also that… (from Module 1 on static occupancy models)



458 **Chapter 19** Occupancy models

FIGURE 19.4

Comparison of the true relationships with humidity (red) and those estimated from the occupancy model fit with maximum likelihood in R package `unmarked` (blackish-blue; shaded region is 95% CI).

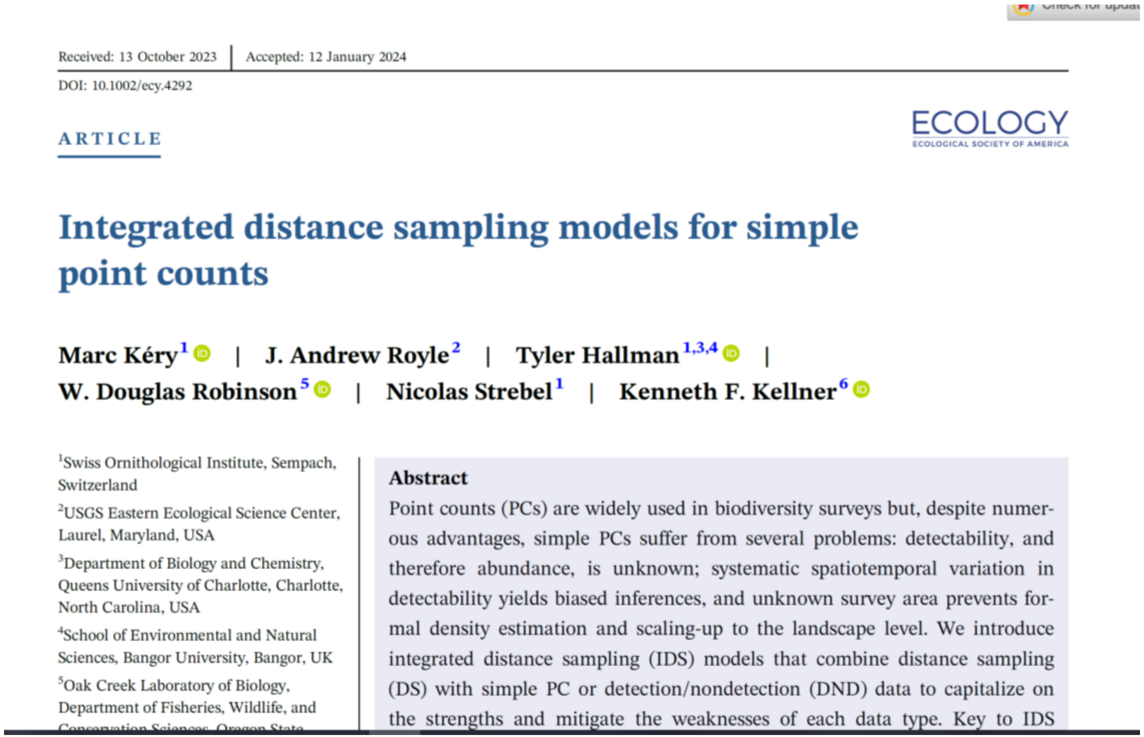
```
# Occupancy
plot(humidity[ooo], state.pred[ooo,1], xlab = 'Humidity index', ylab = 'Occupancy probability',
  frame = FALSE, col = 'blue', lwd = 3, main = 'State process', type = 'l', ylim = c(0, 1))
lines(humidity[ooo], occ.prob[ooo], lwd = 3, col = 'red')
polygon(c(humidity[ooo], rev(humidity[ooo])), c(state.pred[ooo,3], rev(state.pred[ooo,4])),
  col = rgb(0,0,1, 0.1), border = NA)
legend('bottomright', legend = c('Estimate', 'Truth'), lwd = 2, col = c('blue', 'red'),
  bty = 'n', cex = 1.2)

# Detection
plot(humidity[ooo], p.pred[ooo,1], xlab = 'Humidity index', ylab = 'Detection probability',
  frame = FALSE, col = 'blue', lwd = 3, main = 'Observation process', type = 'l', ylim = c(0, 1))
lines(humidity[ooo], p[ooo], lwd = 3, col = 'red')
```

Kéry & Kellner,
brand-new ASM book

Can we have same covariate in state and detection ?

- So, there is absolutely no reason that this should be different for DS …. Also note this:



Received: 13 October 2023 | Accepted: 12 January 2024

DOI: 10.1002/ecy.4292

ECOLOGY
ECOLOGICAL SOCIETY OF AMERICA

ARTICLE

Integrated distance sampling models for simple point counts

Marc Kéry[1] | J. Andrew Royle[2] | Tyler Hallman[1,3,4] |
W. Douglas Robinson[5] | Nicolas Strebel[1] | Kenneth F. Kellner[6]

[1]Swiss Ornithological Institute, Sempach, Switzerland

[2]USGS Eastern Ecological Science Center, Laurel, Maryland, USA

[3]Department of Biology and Chemistry, Queens University of Charlotte, Charlotte, North Carolina, USA

[4]School of Environmental and Natural Sciences, Bangor University, Bangor, UK

[5]Oak Creek Laboratory of Biology, Department of Fisheries, Wildlife, and Conservation Sciences, Oregon State

Abstract

Point counts (PCs) are widely used in biodiversity surveys but, despite numerous advantages, simple PCs suffer from several problems: detectability, and therefore abundance, is unknown; systematic spatiotemporal variation in detectability yields biased inferences, and unknown survey area prevents formal density estimation and scaling-up to the landscape level. We introduce integrated distance sampling (IDS) models that combine distance sampling (DS) with simple PC or detection/nondetection (DND) data to capitalize on the strengths and mitigate the weaknesses of each data type. Key to IDS

Kéry et al., *Ecology*, 2024

Can we have same covariate in state and detection ?

- So, there is absolutely no reason that this should be different for DS …. Also note this:

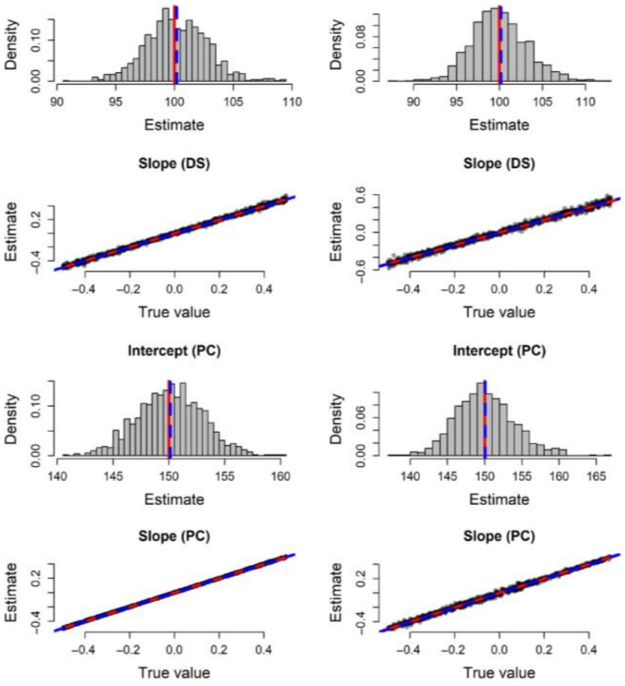
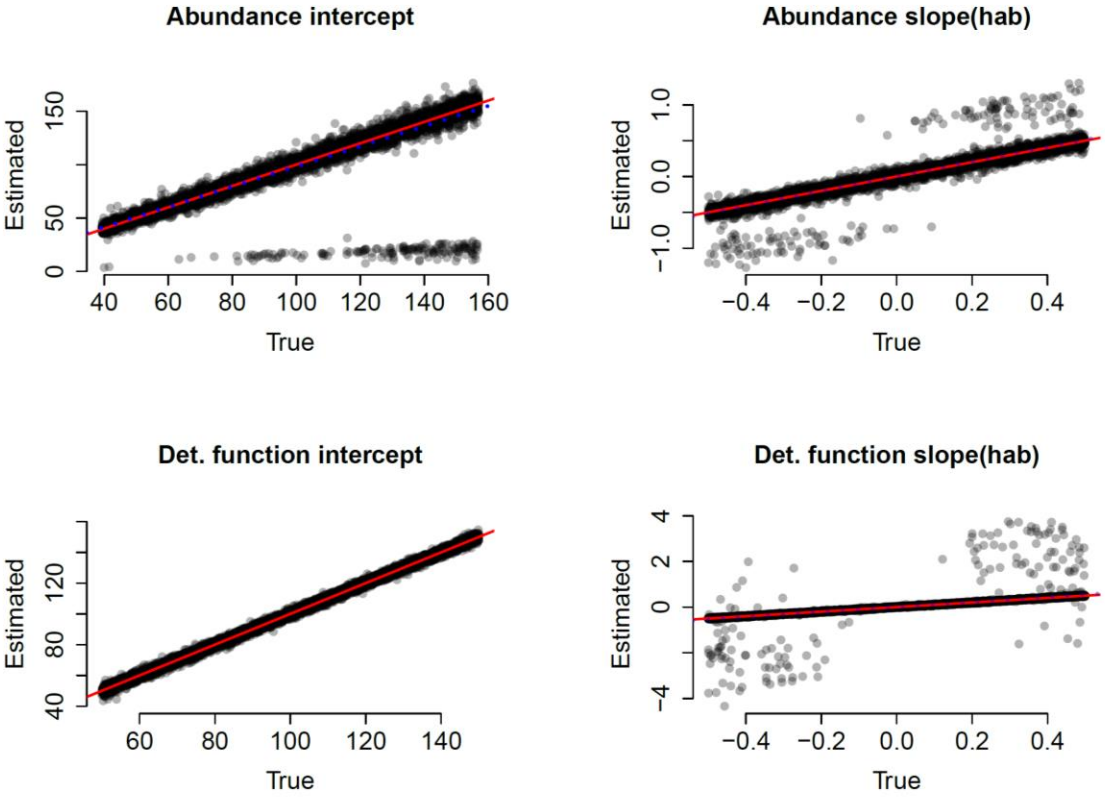


FIGURE 2 Another simulation-based validation of IDS1 combining DS and PC data (Simulation 2). Left, Simulation 2a: Sampling distributions of intercept and slope estimates for detection function parameters with independent effects in the distance sampling (top) and the point count (bottom) parts of the data. Right, Simulation 2b: Intercept and slope estimates for detection function parameters with independent effects in the distance sampling (top) and the point count (bottom) parts of the data, when the same covariate also has an effect on density. Red denotes truth, dashed blue shows mean of estimates. Sample size in both simulations is 1000 data sets. See also Appendix S2: Table S3.

Kéry et al., *Ecology*, 2024

Can we have same covariate in state and detection ?

- Do a custom simulation (see Demo). Result --- Yes, we can !


Abundance intercept
Estimated
True
Abundance slope(hab)
Estimated
True
Det. function intercept
Estimated
True
Det. function slope(hab)
Estimated
True