What can artificial life offer ecology?

Jason Noble, Donna Clarke
and Rob Mills

University of Southampton

Relating simulation results to real data

- How can we show which of two simulations is a better match to real data?
- Choosing between models is central to statistics.
- In stats, a *likelihood function* allows us to quantify how well a model accounts for data:
  - Example: toss a coin 100 times. Is it a fair coin or a 25/75 biased coin? Binomial probability allows us to calculate the likelihood of each model.
  - Important point: the likelihood function is something we can calculate.

ALife's ambivalent relationship with data: "fact free science"?

- For strong ALife, simulation output *is* the data.
  - A controversial position.
- For most of us, the simulation is not intended to be "matched" against real data, but is used as a theoretical playground, proof of concept, intuition pump, "opaque thought experiment", etc.
  - Legitimate and valuable tool.
- But can we do more? Can we fit simulations to data just as statistical models are fitted?
- Problem: complicated simulations have no analytically tractable likelihood function. We might be stuck with qualitative comparisons.

Introducing ecology

- The study of the distribution of organisms and their interactions with each other and their environment.
- Complex ecosystems seen as arising from the interactions of many component organisms.
- Obvious affinity with the theoretical perspective of ALife.

Models in ecology

- Ecologists know they are studying a complex dynamical system:
  - Animal A eats plant B, influencing its distribution, which in turn influences the distribution of other animals that may compete with A, etc.
- But the standard model-building tool in ecology is multiple regression -- a bit like a multi-way correlation analysis.
- Regression models do not make it easy to capture complex causal relationships acting over time and space.
- Ecologists are potentially interested in the *process* models of ALife but they need to know that these models can relate to real data.

A typical problem from ecology

A typical problem from ecology

- Effects of power-line corridors on the distributions of native and introduced mammals in Australia.
- Clarke, Pearce and White, 2006. *Wildlife Research, 33* .
- Goals of the research:
  - Reporting on species prevalence.
  - Could the corridors provide a benefit for some species that favour a transitional forest environment?
  - Suggestions for more subtle management policy: manage fire risk while maximizing species diversity.

Modelling power-line corridors

- Clarke et al. used regression models of species prevalence on environmental variables such as grass, shrubs, fallen logs, tree cover, etc.
- Simplified view?

- A simulation of processes could shed light on the system.
- ALife has no problem representing animals in an environment.
- The difficulty is in judging whether one simulation variant is a better fit to the real data than another.

Indirect inference

- The trick we need is called indirect inference (Gourieroux et al., 1993. *Journal of Applied Econometrics, 8* .)
- Can't calculate likelihood functions for simulations.
- But we can fit an *auxiliary model* to both the real data and to the output of candidate simulations.
- Auxiliary model need not be perfect, but it does need a likelihood function, e.g., regression.
- We fit the auxiliary to both the real data and the simulated data, giving us a location in the auxiliary's parameter space for each case.
- The simulation we prefer is the one whose location in aux-p-space is closest to that of the real data!

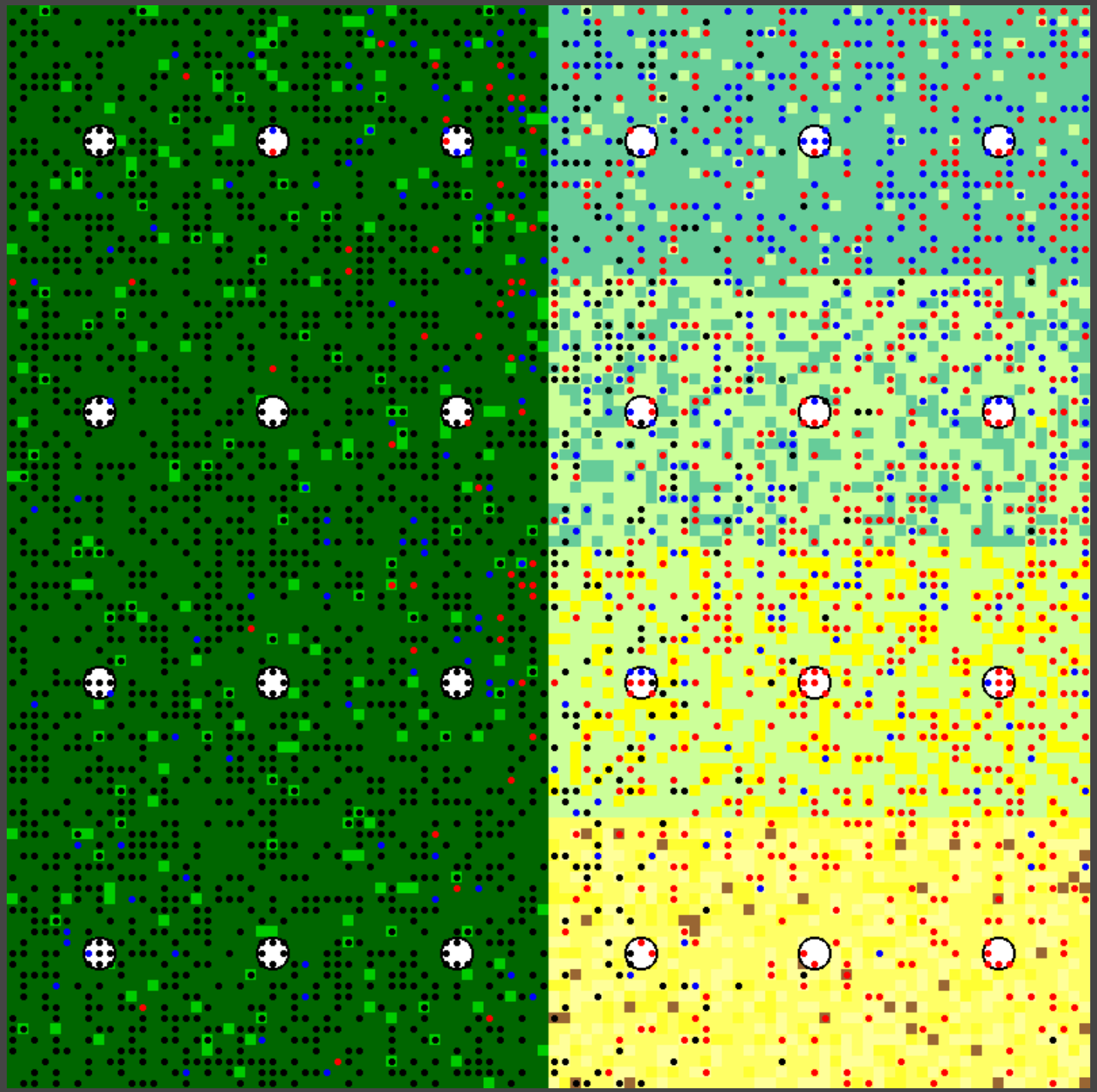
Ways of thinking about indirect inference

- Indirect inference looks a bit magical at first.
- Think of the auxiliary model as a window into the larger space of all possible simulations.
- Alternatively, view it as compressing the many degrees of freedom of the simulation output down into only a few dimensions, those of the auxiliary's parameter space.
- Distance from the real data in this parameter space stands in for the tractable likelihood function we've been missing all along.

A fictional data set: why?

- Can we test this idea using Clarke et al.'s real data?
- Problem: we don't know the real causal story behind this data. It's an open empirical question.
- So we need a fictional data set where the truth is known from the outset.
- Devised a fictional system based on Clarke's real system:
  - Three small-mammal species coexisting in forest and corridor environments, sections of the power-line corridor periodically cleared, etc.
- We test the indirect inference method by seeing whether it identifies the right model as being closest to the "real data".

- Forest on the left, cleared ground on the right.
- Vegetation changes as time since clearance increases (most recent at bottom)
- Three small mammal species in red, blue and black
- Trap sites shown as white circles



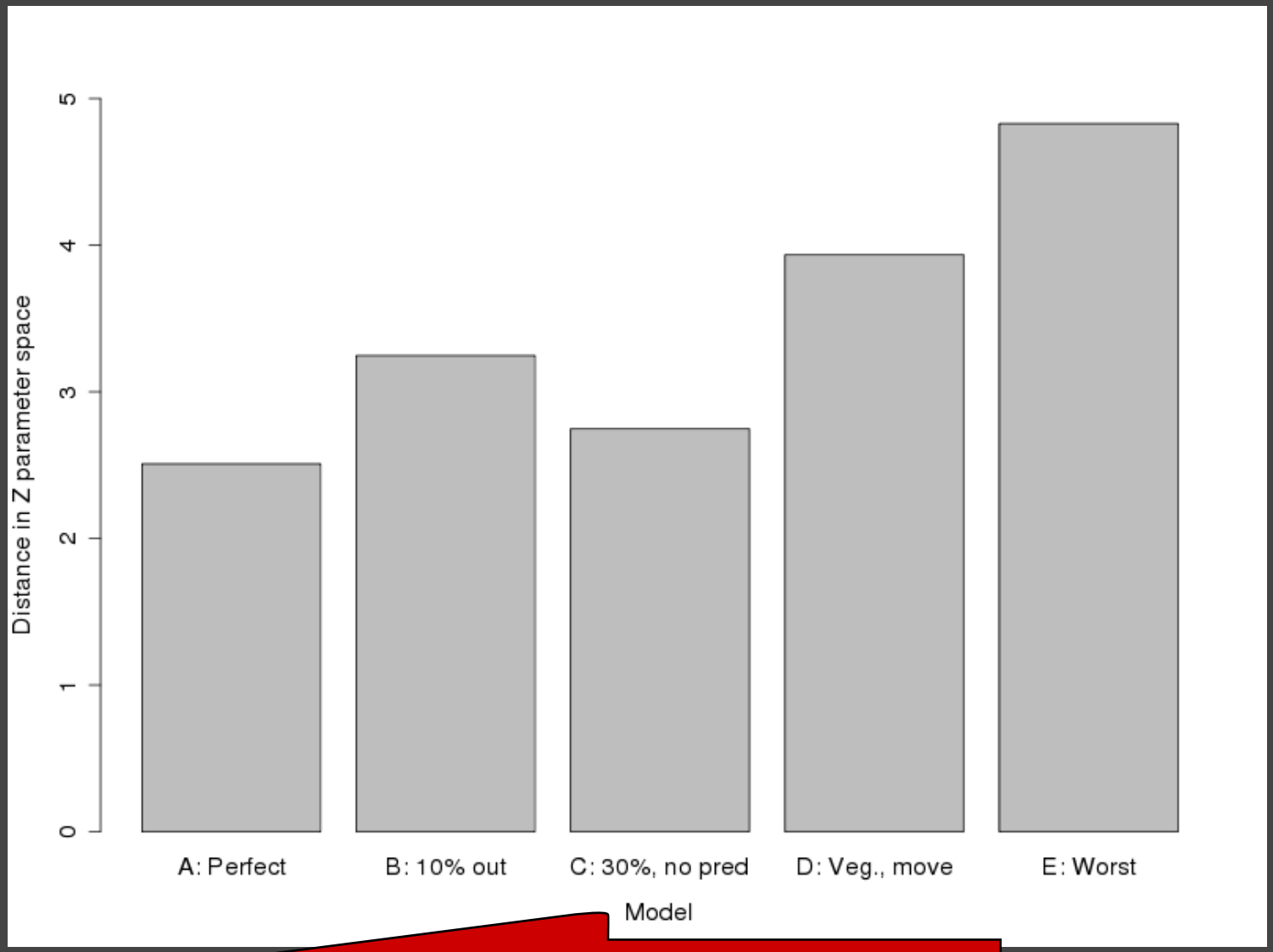
Statistical analysis of "real" data

The auxiliary model: regression of species A density, reduced using AIC.

```
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -2.5122404  0.9387635  -2.676  0.00836 **
ShortShrubs      0.1486472  0.0232270   6.400 2.29e-09 ***
ForestDist       0.0805033  0.0137331   5.862 3.25e-08 ***
TSM              0.0248839  0.0112968   2.203  0.02928 *
SpB              0.3292179  0.1563186   2.106  0.03702 *
ShortShrubs:TSM -0.0005703  0.0002517  -2.266  0.02502 *
TSM:SpB         -0.0006755  0.0003728  -1.812  0.07214 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1

Residual standard error: 2.14 on 137 degrees of freedom
Multiple R-squared: 0.7669,     Adjusted R-squared: 0.7567
F-statistic: 75.12 on 6 and 137 DF,  p-value: < 2.2e-16
```

Results: distance from the "real" data



Closer to the parameter values of the real data

Objectively better models

Conclusions and future work

- The right model was identified as the preferred simulation.
- Indirect inference is a promising method for getting the complex simulations of ALife to connect to real data in ecology and other sciences.
- Future work:
  - How to specify p-space: we used Z-scores across all observed parameter values but other methods possible.
  - How rich does the auxiliary need to be to capture enough complexity?
  - Comparison with a machine learning approach: search for good simulation parameters by splitting the real data, training on one half and testing on the other.