The power of simulating data: a tool to design experiments, understand data limitations and improve scientific reasoning

Frederik Baumgarten[1,2], EM Wolkovich[1], invite Andrew Gelman

June 7, 2024

[1] Department of Forest and Conservation, Faculty of Forestry, University of British Columbia, 2424 Main Mall Vancouver, BC Canada V6T 1Z4.

[2] Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Zürcherstr. 111, Birmensdorf 8903, Switzerland

Corresponding Author: Frederik Baumgarten; frederik.baumgarten@ubc.ca
Journal: statistical report in Ecology?

Abstract

Science continuously tries to explain patterns in nature that appear in data, hidden by the noise of variation caused by a myriad of influencing factors. To unravel pattern from noise, link correlation with causality and build upon the existing house of knowledge, researcher follow a scientific workflow. However, concerns about several aspects of this workflow has been raised: formulating meaningless (null) hypothesis, poor experimental designs, the replication crisis and the problematic usage of p-values and "significance" has led to biased reporting of findings, over-interpretation or wrong conclusion with patterns emerging from mathematical artefacts or by sheer chance.
Here we propose the integration of data simulation to the scientific workflow to facilitate overcoming these challenges by 1) developing meaningful and testable hypotheses through the formulation of mathematical/mechanistic models, 2) playing with effect size, variance and replication to generate fake-datasets, 3) exploring the potential and limitations of both the statistical method and the data set obtained, and 4) drawing conclusions that can build on the existing 'body of knowledge'.
Furthermore, we believe that in a future world with powerful AI and machine learning algorithms that can be applied to ever larger data sets, meaningful, theoretically grounded hypotheses will be more important than ever to understand the underlying causalities. Without them, we may be able to make excellent predictions (within known limits) but without knowing the driving variables.
Keywords: bayesian statistics, statistcal methods, quantitative ecology, hypothesis testing, scientific reasoning, experimental design, power analysis

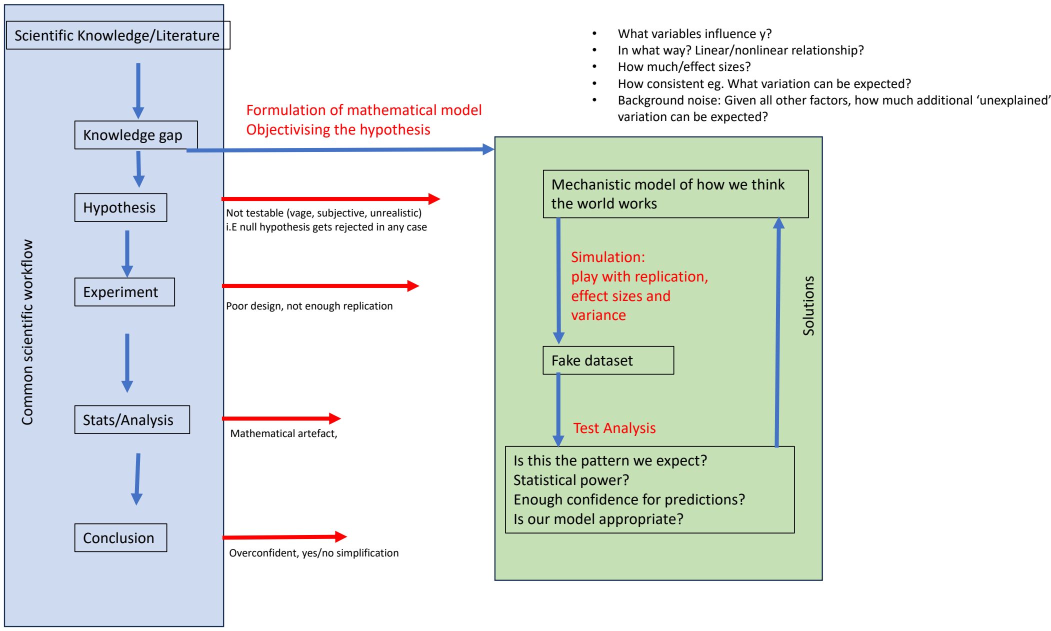

Figure 1: Schematic overview of the current scientific workflow with its related problems and how they could be overcome with the help of data simulation. We believe that this 'extraround' not just helps to detect and solves some of these problems but that it leads to better science by attaching new findings to the established theoretical framework

1 Introduction

Opening example

A renowned researcher once explained to me the progression of a curve deriving from a dendrometer — a device attached to the bark of a tree that stood beside us. As we observed the graph on the screen extending in real-time, he provided a passionate and insightful explanation on water relations and plant physiology. However, moments later, the responsible technician entrusted me with the fact, that the device hasn't yet recorded anything meaningful and that all data displayed couldn't possibly depict anything biologically relevant. This example - and I think most of us can share similar stories - shows the incredible ability of our minds to make sense of patterns that underpin our current believes or argument. But it also underpins our blindness of how randomness can look like. However, separating noise from a pattern is perhaps the most important skill of any researcher.
Human desire for patterns even in pure noise
Confirmation bias
But noisy data
how can we ensure a standard?

Current solutions

scientific workflow (fig)
include experiments, null hypothesis testing and their limits(not meaningull, not testable, not interesting) , analysis -> conclusions

Growing evidence that this is not enough

p-values/replication crisis/overconfidence/overinterpretation/mathematical artefacts
Show famous examples. -> could expand into box

Some people hope to address this through machine learning

machines search for patterns without or less bias (or at least in a systematic/objective way)
machine learning is usually amechanistic
problem with hypothesis remain. searching for a model so it includes much of the assumptions/hypotheses expected from a useful model

Aim

We propose an updated approach focussed on simulation + show how to do it
Helps to address current gaps/limitations:
build hypothesis, then formulation of mathematical model
better design experiments
avoid overconfidence/overinterpretation and mathematical artefacts

2 Bus example with simulation workflow

We all manage to go through this world with the help of models in our minds. Most of the time we are not aware of this fact but their implications are all around us. For instance when we wait for a bus we expect the waiting time to be lets say around 7±3 min or when we readjust our expected travel time after an incidence and add the variable 'traffic jam' to our internal model. In short, whenever we have an intuitive understanding of how the world works, our brain suggests basic models - mostly oversimplified and with increasing bias as soon as we enter non-linear relationships. - In many practical cases we get the chance to recalibrate our internal model (or lets call it intuition). For example when the bus after 10 min is still not there and we start realizing that we need to incorporate the variable 'traffic jam'. But in many cases we don't get this kind of feedback or at least not in an informative way because it is too complex to grasp for our minds. While daily life is arguably not a big drawback to this issue, in science it can be.

Situation

model - show poisson distr. simulate

no what?

still waiting for the bus...outlayer?
update the model - assumptions
add variable traffic jam

3 Biological example - how to do it

question based + we walk through each one.

what influences y?

nitrogen

what form? linear/nonlinear, near Gaussian, Poisson

linear

What assumptions are reasonable?

for y, alpha, beta
for effect sizes (parameters)
for x data
-> pick some for this example

Simulate!

How to use this - Play!

so many ways, we highlight just a few

Power analysis

to better design experiments

Avoiding overconfidence

play with replication while holding variance and effect size constant. p-value figure

Increasing importance in the future

Avoiding overconfidence

evergrowing Lit with AI we must learn to ask better questions
the right questions and hypothesis that are testable with current + new methods. Build up house of knowledge instead of using new pattern finding algorithms
how to integrate with AI?

References