

Model: An Inference Machine With a Domain

IDE-015 · Fundamental Ideas

Concept essays for *The Universe and Its Observers*

Author: Juan Segurolo

Version: v0.8 EN · 19 September 2026

Status: Publication version

Article form: Concept essay

Scale / context: Analytical, statistical and computational models across the sciences

Boundary: Models as inference machines and how they are evaluated. The stand-in relation itself belongs to IDE-012; the measurement model inside a metrological chain belongs to IDE-014

Related works: IDE-016, IDE-012, IDE-014

Epistemic profile: Established knowledge with interpretive boundaries

Licence at publication: All Rights Reserved

Article Abstract

A model is a deliberately simplified structure built to support inference about something else. Its value is never absolute: it is conditional on a question, a domain and a set of assumptions, and a model that is excellent for one inference can be worthless for another built from the same system.[5,6] Box's aphorism that all models are wrong but some are useful is quoted constantly and read carelessly; what he argued was that the scientist should ask whether a model is illuminating rather than whether it is true, and should attend to badly wrong assumptions rather than to universal wrongness.[1] This article sets out what a model has to declare, why agreement with data is weaker evidence than it looks, how models are chosen when several fit, and what to do when good models disagree.

In 60 Seconds

Every model omits, and what it omits is chosen by the question it was built to answer, so the first thing a model has to declare is its purpose and its domain.[6] Fitting the data is the weakest kind of evidence it can offer. A numerical model of an open natural system cannot be verified in the logical sense, because agreement with observations is consistent with many model structures and with compensating errors; Oreskes and colleagues argue that such models can be confirmed to a degree and never established as true.[2] Parsimony has a precise version rather than an aesthetic one: information criteria estimate expected predictive loss and penalize parameters because unnecessary flexibility degrades prediction on new data, not because simplicity is elegant.[3,4] Multimodel inference evaluates conclusions across a set of plausible models and can identify conclusions that remain stable across that set.[4] And a model that fits for the wrong reasons, with parameters absorbing the effect of a mechanism left out, will fail exactly where it is extrapolated.

Why This Question Matters

Models are where scientific claims are actually made, and the most common failures are not errors inside the model but errors about it: using it outside its domain, treating agreement with data as proof, and forgetting which mechanisms were left out. Each of those is invisible unless the model's purpose and

assumptions were stated in the first place. The stakes are practical, because policy and engineering decisions rest on extrapolations, and an extrapolation is precisely the use for which the evidence of past fit is weakest.

1. A Model Is Built for a Question

A model keeps some features of its target and drops others. That is not a regrettable compromise but the source of its usefulness: a representation retaining everything would be as unmanageable as the system itself. What makes the selection defensible is that it was made with respect to a question, so the omitted features are ones the question does not turn on.[5,6]

Model-target representation and similarity are purpose-dependent.[5] Two people can build the same equations and be making different claims, because they have assigned different roles to the same terms.

The practical form of this is a declaration. A model should state what it is for, what it represents, what it deliberately excludes, and over what range of conditions it is expected to hold. A model without those is not more general; it is less checkable, because there is no stated condition under which it could be found to be misapplied.

Where That Leaves Us

Declaring purpose, exclusions and domain is what makes misapplication detectable. That declaration also changes what the familiar remark about all models being wrong actually means.

2. Wrong but Useful, Said Precisely

Box's 1976 paper is usually reduced to five words, and the surrounding argument is more useful than the slogan. Writing about statistical models of natural phenomena, he argued that since all such models are wrong the scientist cannot obtain a correct one by excessive elaboration, and should instead seek an economical description that captures the important features.[1] The recommendation is to look for the assumptions that are badly wrong rather than to note that all are wrong.

The scope is worth keeping. Box's subject is idealized empirical models, not every formal structure, and read as an unrestricted claim that nothing can ever be described correctly the aphorism becomes a philosophical thesis rather than the methodological one he argued for. Taken in its own scope the distinction does real work: universal wrongness would license nothing, since it would be equally true of the best and the worst model available, while whether a particular simplification is consequential for a particular inference is a question with an answer, and the answer is found by perturbing the assumption and seeing whether the conclusion moves.

Box also emphasized iteration. A model is criticized against data, its failures are diagnosed, and it is modified, which makes model building a process rather than a product.[1] Residuals are the productive part of that loop, because a systematic pattern in what the model failed to explain is information about what was left out.

Where That Leaves Us

Box's argument, about statistical models of natural phenomena, is that no amount of elaboration yields a correct one, so the work is to find which assumptions are consequentially wrong for the inference at hand, and to iterate against the residuals. Universal wrongness by itself excuses nothing. The next question is what agreement with data actually establishes once a model has been fitted to it.

3. Fit Is Not Confirmation

Oreskes, Shrader-Frechette and Belitz argued that numerical models of open natural systems cannot be verified or validated in the sense those words carry in logic and in software engineering. Natural systems are never closed, model results are never unique, and agreement with observation is consistent with more than one model structure.[2]

The mechanism is easy to state. Parameters are adjusted to fit; a parameter can absorb the effect of a mechanism that is missing; and two errors can cancel over the range of the data. The fitted model then reproduces the observations while encoding a wrong account of why they came out that way, which is undetectable from the fit alone and becomes visible only outside the fitted range.

What such models can do is be confirmed to a degree by the number and variety of the observations they match, and be used heuristically to explore consequences of assumptions.[2] That is a real epistemic status and not a dismissal, but it is weaker than the language of validation suggests, and the difference matters most for extrapolation, which is the use the evidence supports least.

Where That Leaves Us

Agreement with data does not verify a model of an open system, because parameters can absorb missing mechanisms and errors can cancel across the fitted range. Confirmation admits of degree and comes from the variety of matched observations, not from goodness of fit alone. That leaves the practical question of how to choose when several models fit acceptably well.

4. Choosing Among Models

Adding parameters always improves fit to the data in hand, so fit cannot be the criterion. Akaike's insight was to define the target as expected predictive performance on new data from the same process, and to show that the maximized likelihood is a biased estimate of it, with the bias related to the number of estimated parameters.[3] The resulting criterion penalizes complexity for a stated reason rather than on principle.

This converts parsimony from an aesthetic preference into a prediction argument. A model with unnecessary flexibility fits the noise as well as the signal, and its predictions on new data are worse for it. Burnham and Anderson develop the practical apparatus, including the comparison of candidate models on a common data set and the weighting of models by their relative support.[4]

Two cautions belong with the tool. Criteria compare the candidates offered and say nothing about a better model absent from the set, so a well-supported model may still be the best of a poor field. And the comparison is only valid across models fitted to exactly the same data, which is a requirement that is easy to violate without noticing.[4]

Where That Leaves Us

Model choice cannot rest on fit, because fit always improves with parameters, so criteria estimate expected performance on new data and penalize complexity for that reason. Parsimony becomes a prediction argument rather than a taste. The criteria rank only the candidates offered, which raises the question of what to do when several plausible models survive together.

5. When Good Models Disagree

Multimodel inference evaluates conclusions across a set of plausible models and can identify conclusions that remain stable across that set. A molecular simulation, a continuum equation and a statistical surrogate can all be right in their own domains without being interchangeable, because they answer different questions at different costs.[5,6] Disagreement among them is informative rather than embarrassing.

This also gives the honest way to report a modelling result. State the question, the domain, the assumptions that were varied, the conclusions that were stable and the ones that were not. A single number from a single model, presented without that structure, is the form in which most modelling results are misused.

Where That Leaves Us

A result that flips when a structural choice changes is a fact about the model. IDE-016 takes up a property that models are often asked to quantify and that resists a single measure: complexity itself.

What We Know, and How

- Model-target representation and similarity are purpose-dependent.[5,6]
- Box argued, of statistical models of natural phenomena, that elaboration cannot produce a correct one and that the task is to find which assumptions are consequentially wrong, with model building proceeding iteratively against residuals.[1]
- Numerical models of open systems cannot be verified; agreement with data is consistent with multiple structures and with compensating errors, so confirmation is a matter of degree.[2]
- Information criteria penalize parameters because unnecessary flexibility degrades prediction on new data, and they rank only the candidate models actually offered.[3,4]

What We Still Don't Know

- How to characterize the predictive reliability of very high-dimensional models whose effective number of parameters is not well defined, where the classical penalty arguments do not straightforwardly apply.
- How to weight structurally different models against one another when they are not fitted to a common likelihood, which is the usual situation in multi-model climate and epidemiological ensembles.
- Whether robustness across a set of plausible models is evidence about the system or evidence about the shared assumptions of the modelling community that produced the set.

What Could Change Our Minds

- A demonstrated procedure that established a numerical model of an open natural system as true rather than confirmed to a degree, surviving the non-uniqueness and compensating-error objections, would overturn the standard epistemic status of such models.[2]
- Reproducible evidence that models selected by information criteria predict new data from the same process worse than models selected by in-sample fit, across varied settings with correctly matched data sets, would undermine the predictive justification for parsimony.[3,4]
- A worked case in which a conclusion robust across every plausible model structure was nonetheless wrong about the system, with the failure traced to a shared assumption rather than to chance, would weaken robustness analysis as evidence.[4,5]

If You Remember Only One Thing

A model is an inference machine with a stated question, domain and set of exclusions; fitting the data is the weakest evidence it can offer.

Related Articles

- IDE-016 — Complexity
- IDE-012 — Representation
- IDE-014 — Measurement

References

- [1] George E. P. Box. Science and Statistics. *Journal of the American Statistical Association* 71(356), 791-799 (1976). DOI 10.1080/01621459.1976.10480949.
- [2] Naomi Oreskes; Kristin Shrader-Frechette; Kenneth Belitz. Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences. *Science* 263, 641-646 (1994). DOI 10.1126/science.263.5147.641.
- [3] Hirotugu Akaike. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control* 19(6), 716-723 (1974). DOI 10.1109/TAC.1974.1100705.
- [4] Kenneth P. Burnham; David R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed. Springer (2002).
- [5] Michael Weisberg. *Simulation and Similarity: Using Models to Understand the World*. Oxford University Press (2013).
- [6] Ronald N. Giere. How Models Are Used to Represent Reality. *Philosophy of Science* 71(5), 742-752 (2004). DOI 10.1086/425063.