

Quantifying the Predictive Capability of Computational Models

MORS Workshop

October 15-17, 2002

Robert G. Easterling¹
consulting statistician
Cedar Crest, NM

rgeaste@comcast.net

Contents

Statistical Model

Case Study

Issues/Problems

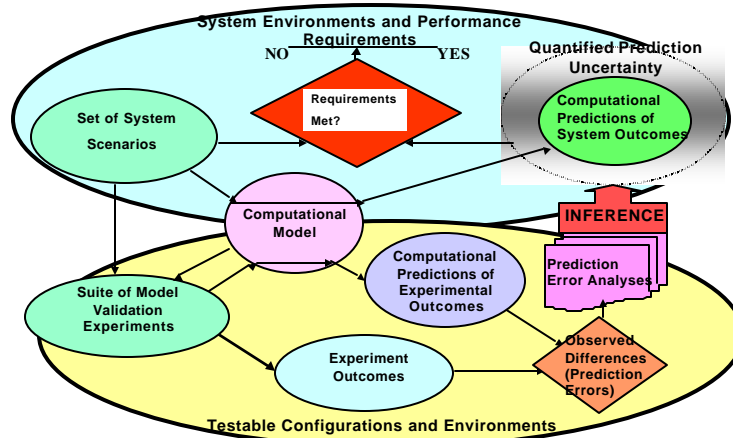
Conclusions

¹work supported by Sandia National Laboratories, Albuquerque, NM. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94AL85000

Introduction

- Computational Predictions` -- Inquiring minds want to know:
 - *How well does the computer model represent reality?*
 - *How well can the computer model predict reality under untried conditions?*
- Answers from: "MODEL VALIDATION"
 - comparisons of computations to data
 - » *(field or experimental)*

Measuring Predictive Capability: Purpose and Process



My View of Model Validation

- Essence of Model Validation
 - Comparison of computations to data
- This implies
 - design of experiments - to generate the right data
 - data analysis - to extract and communicate the information contained in the data
- *Thus, model-validation is fundamentally statistical*

Some Additional Motivation

- “Given the critical importance of model validation.. ., it is surprising that the constituent parts are not provided in the (DoD) directive concerning ... validation. A statistical perspective is almost entirely missing in these directives.”
 - *[National Academy of Sciences report on statistics, testing, and defense acquisition, Cohen et al. 1998]*

Goal:

- My goal is to be able to characterize a model's 'predictive capability' with statements like --
 - Our understanding of the underlying science, our ability to translate that understanding to a computational model, and an analysis of a robust set of experiments and corresponding calculations indicate that actual system performance is quite likely to be within P% of the computational prediction for the application of interest.
 - » *Then, e.g., if the computational prediction plus P% is less than the failure threshold, we can “confidently” declare that the system meets its requirement.*

Goal, cont.

- Use the results of a suite of model-validation experiments and computations to evaluate a computational model's "predictive capability"
(“the degree to which a model represents the real world”)
- **Constraint: The experimental region may not be the same as the application region, for which predictions are the objective.**
 - *Example. lab expts. on mock-ups vs. real device in field*
- **Evaluation should be:**
 - credible, defensible, communicable, ...
 - » (“Don't give me no statistics, Meathead. I want facts!” Archie Bunker)
- How do we (hope to) achieve the goal?
 - process: following slides

Mathematical Set-up:

Let x be a vector that defines an event of interest, often a system and the environment to which it is subjected; e.g.,

- **experiment**
 - » *hit an instrumented missile nose cone with a 500lb. hammer*
- **application**
 - » *subject a missile to hostile in-flight environment*
- Let y be event outcome
 - e.g., *stress on key missile parts*

Mathematical Set-up, cont.

Computational Model:

- $y^M(x) = M(x:j)$, where
- x = event-defining variables
- j = model parameters (constants in the equations within M):
 - » e.g., material properties of nose cone and hammer, damping coeffs., ...

Notes.

x , y^M , and j are all possibly vectors or fields.
Focus on deterministic M, but for stochastic M, y^M could be vector of realizations from a probability distribution
Computational parameters (e.g., grid size, convergence criteria) are included in the specification of M.
Assume that M has been 'verified.' It is deemed *validation-ready*

Statistical Model for Model-Validation

- Conduct experiment at x
- Experimental outcome: $y(x) = y(x,w)$,
 - where w = unmodeled variables that influence nature's outcome
 - statistical model: w varies randomly across expts.
 - » w has unknown probability distribution
 - $y(x)$ is a "realization" of the random variable, $y(x,w)$
- "Prediction Error" at x : $e_x = y(x) - y^M(x)$ (nature - model)
- Contributors to e_x
 - random effects, w , in nature, not in M
 - » Example: M is 2-D model; nature is 3-D
 - systematic differences between nature and M

Statistical Model, cont.

- Problem: Measurement Error
 - $y(x)$, the “true” experimental outcome is, in general, not observable.
 - Measured experimental result: $y^E(x) = y(x) + d_x$, where d_x is a random variable with an unknown probability distribution
 - » “gage studies” provide estimate of distribution of d_x
- **THUS ...**

Statistical Framework, Bottom Line

- The resulting statistical relationship between $y^E(x)$ and $y^M(x)$ is:

$$y^E(x) = y^M(x) + e_x + d_x, \quad \boxed{\text{Data} = \text{Signal} + \text{Noise}}$$

where e_x and d_x are random variables with unknown dist'ns. that, in general, depend on x

~~$\boxed{\text{OT\&E analog: } y = \text{miss distance; } x = \text{range}}$~~

- **The Task (should you choose to accept it) is to conduct suite of experiments and computations that provide for a credible, defensible, communicable, ... characterization of the probability distribution of e_x**

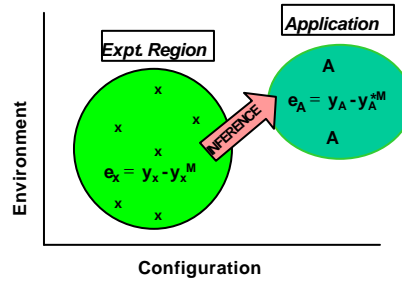
Graphically, ...

1. Experimental Design

Design and conduct a set of experiments and corresponding calculations (the x -points in Test Region, which is defined by two meta-variables: configuration and environment)

2. Data Analysis

Evaluate predictive capability $\{e_x\}$ for the experiments in the Test Region



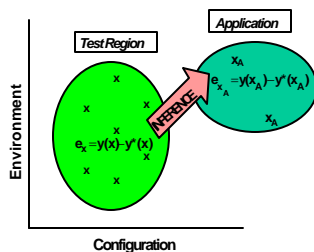
3. Inference

Estimate predictive capability at untested situations (A-points in the Application region)

Technical Issues

1. Experimental Design

- building blocks for inference
- realistic est. of prediction error
- manageable number
- efficient exploration of test region



2. Data Analysis

- estimated characteristics of e_x will be (statistically) uncertain
- ~~statistical methods appropriate~~

3. Inference

- interpolation is desirable
- extrapolation, if required, generally difficult to justify
- requires extending:
 - a. modeled physics, via y^M
 - b. unmodeled physics - the prediction errors
 - empirical, judgment-based
- may be a hard plank to walk, but it's been done

Foam Degradation Case Study

Requirements, Polyurethane foam -

1. structural support, normal environments
2. role (not requirement): insulates weapon components in accident-induced thermal environments

Applications:

system: systems in fires --
 x = fuel source, temperature profile, orientation, duration, weather, system-damage state, ...
 component: x = system-fire induced thermal environment

Models :

CPUF (foam decomposition)
 • Newly developed comp. model,
 • better accounting for foam effects

Validation test program -- to date

simulated-component experiments - decomp., diffusion, radiation

Analysis:

Evaluate predictive capability

Measuring Predictive Capability - Case Study: Foam Vaporization Experiments

Nature: Eight experiments in Sandia's Radiant Heat Facility:

y^E = decomposition-front position vs. time, measured via x-ray imagery (unconnected dots, plot)

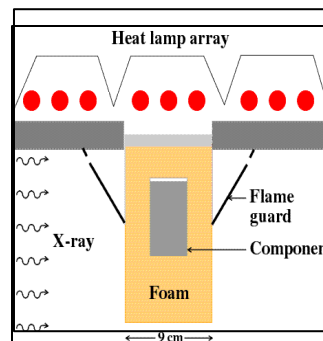
Model: $M(x;j) = \text{CPUF}$, where

x = experimental factors:

base plate temp. (600, 750, 900, 1000C -- after 1.5 min. ramp)

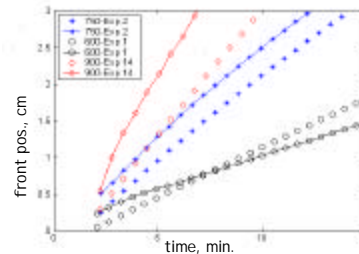
j = activation energies for foam decomposition, emissivity, ... (obtained from other sources)

y^M = calculated decomposition-front position vs. time



Subset of Results

Plot shows computational predictions (connected points) and experimental results (unconnected) for experiments at 600, 750, and 900C.



Analysis: focus on front velocity (slope of curves) between heat source and insulated component (1-2 cm)

Eyeball Analysis: Model is OK at 750C, over-predicts velocity at 900C, under-predicts at 600C. Issue:

"real" model error or "in the noise?"

The following analysis will substantiate the eyeball analysis

The Data

Exp.	Temp.	Heat Orient.	Int'l. Comp.	v^M	v^E	e	$ln e$
2	750	bottom	none	0.246	0.232	-0.013	-0.056
10	750	overhead	none	0.234	0.211	-0.023	-0.105
11	750	side	none	0.262	0.258	-0.004	-0.014
13	750	side	none	0.228	0.215	-0.012	-0.056
15	750	bottom	AL cyl.	0.284	0.275	-0.009	-0.030
1	600	bottom	none	0.091	0.131	0.039	0.358
14	900	bottom	none	0.450	0.349	-0.100	-0.253
16	1000	bottom	AL cyl.	0.770	0.558	-0.212	-0.322

Table shows logarithmic error ($\ln[v^E/v^M]$, denoted $ln e$) because preliminary analysis led to this transformation based on theoretical and potential variance-stabilizing properties

Note: First five experiments are all essentially the same; the v^M values vary because the measured boundary conditions, used as model input, varied

Analysis 1 -- Characterize Predictive Capability at 750C

Working assumption: the variability of the observed prediction log-errors among the 5 experiments at 750C is random extra-model variability (primarily specimen-to-specimen; measurement error not addressed here, appears to be negligible):

Analysis:

summary statistics, $\ln e$: ave = -.05 stdev = .034

evidence of bias?

$t = \text{ave}/(\text{stdev}/\sqrt{5}) = -3.40$, on 4df

$P(2\text{-tail}) = .03^*$

Fairly strong statistical evidence of bias; may not be practically significant, and bias is in conservative direction

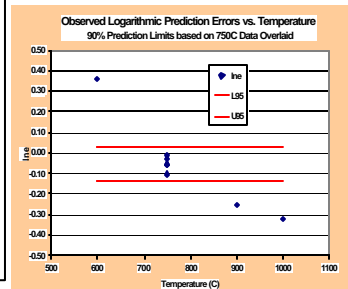
*P = $\text{Prob}(|t_4| > 3.40)$, based on Normal distribution assumption for $\ln e$

Analysis 2 -- Inference to Temp. Extremes

Emulate the inference process by extending 750C findings to 600C, 900C, 1000C

90% prediction interval for future log-error:

$$\begin{aligned} &\text{ave} \pm t_{.05}(4) \cdot \text{stdev} \cdot \sqrt{1+1/n} \\ &= -.05 \pm .080 \\ &= (-.13, .03) \end{aligned}$$



Inference, based on (leap-of-faith, judgment-based) assumption that the $\ln e$ distribution is independent of temperature over the experimental range: to be consistent (@90% level) with 750C data, $\ln e$ at temp. extremes should be in $(-.13, .03)$

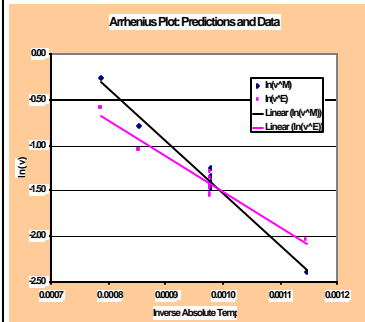
Finding (see plot): Inference grossly in error!

Analysis 3 -- Look at the Data Again

- Use subject-matter insight -
- Arrhenius model:
 $v \propto \exp(E/\text{abs. Temp})$

Chart: $\ln(v)$ vs. $1/(\text{abs. Temp})$

Both the model predictions and the experimental data exhibit fairly good linearity on these scales, BUT with different slopes.



Possible Further Actions

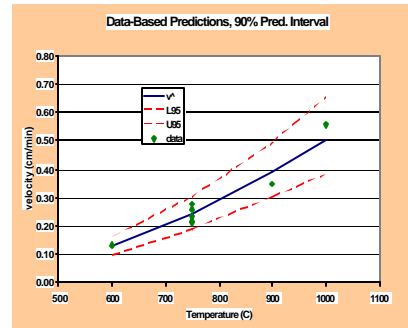
- Do bias-corrected prediction:
 - $y^M(x) + b(x) \pm$ prediction-error limits,
where $b(x)$ is estimated bias-correction function
 - » *weak science - unappealing*
 - » *not feasible in predictions for applications*
- Fix/improve the model
 - modify parameter estimates (activation energies)
 - put more physics/chemistry into model
 - » *incorporate specimen-specific variables (w 's) into model*
 - *requires measuring those variables on each specimen*
 - » *expensive, time-consuming*
- Abandon theoretical model; use semi-empirical model (for limited purpose of predicting front velocity in experimental region) - next slide

Statistical Prediction Intervals

Plot shows regression fit and 90% statistical prediction intervals*, on original velocity and temperature axes.

Prediction interval width is temperature-dependent: there are wider logarithmic error limits the greater the distance from the center of the data.

*assump.: log-errors Normally distributed, homogeneous variances



Interpretation: At a given temperature, with 90% confidence the measured velocity in a future experiment like these would fall within the indicated limits

Inference to 1500C? 2000C? Other geometries? ... ?

Requires foam - expert judgment

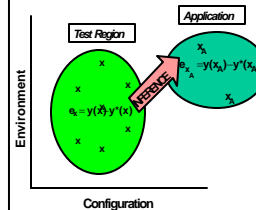
Case Study - Interim Findings

- Predictive capability is not too good
 - Doesn't get the temperature effect right
 - Model's sensitivity to temperature is about 2x that of experimental data
- The data used to evaluate predictive capability can be used directly to construct a semi-empirical model directly

Issue: What if we fail?

What if we cannot bridge the inference gap from experiments to application?
Possible solutions:

- test in more application-like configurations and environments (science)
 - » extreme example: resumption of underground nuclear testing
- redesign system (engineering)
 - » design out features that are most difficult to model
- rework the requirements (program mgt.)
- rework the model (comp. science)
 - » put more science into model
- "softer" methods -- expert opinion:
 - E.g. We never saw more than a 25% prediction error in the experiments we could do, but differences between those conditions and the application lead us to think that an additional factor of two would be prudent -- i.e., 50% prediction error limit. Trust us.



Even if we fail on one loop, knowing why and what the obstacles are is useful in deciding what next to do.

Issue:

~~Conflicting Objectives: Model-Builder vs. Experimenter~~

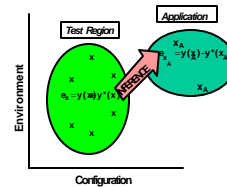
- **Model-Builder:**
 - If I put more science (more x's) into the model, I can drive error to negligible.
 - that's science, progress, ...
- **Experimenter:**
 - If you put stuff in your model I can't measure or control, I can't do the experiments.
 - If you put too much stuff in, I can't do enough experiments
 - We'll never know if error is negligible
- **Sponsor/User/Decision-Maker:**
What about us?

The framework proposed here for evaluating predictive capability provides a means of expressing these arguments and for resolving trade-offs between the model and the ability to measure predictive capability.

Summary

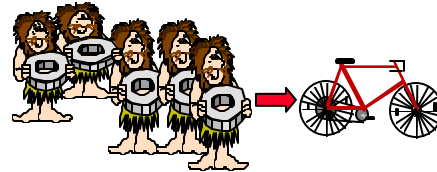
Measuring predictive-capability poses numerous difficult problems:

- scientific, experimental, statistical, organizational, management



Statistical ideas and methods can contribute to successful resolution of these problems, or clear understanding of why they cannot be solved

We all gotta work together



Model-Validation in the News

DoD comparison of computer simulations versus live-fire tests of the effect of gunfire on helicopter blades:

- On a scale of 1 to 10, the models scored:
 - 7 in predicting how the shell would penetrate the blade,
 - 3 in predicting the destruction of the helicopter blade,
 - 2 in predicting the loss of a helicopter,
 - » [defense news item, 10/17/96]
- modeling hierarchy: phenomenon - component - system
- predictive capability decreases as complexity increases
- scoring rule? interpretation? how good is good enough? ...