10-9-2015

Prediction: The Quintessential Model Validation Test

Wayne Wakeland
Portland State University, wakeland@pdx.edu

Follow this and additional works at: https://pdxscholar.library.pdx.edu/systems_science_seminar_series

Part of the Statistical Models Commons, and the Theory and Algorithms Commons

Let us know how access to this document benefits you.

Recommended Citation
https://pdxscholar.library.pdx.edu/systems_science_seminar_series/66

This Book is brought to you for free and open access. It has been accepted for inclusion in Systems Science Friday Noon Seminar Series by an authorized administrator of PDXScholar. Please contact us if we can make this document more accessible: pdxscholar@pdx.edu.
Prediction--the Quintessential Policy Model Validation Test

Wayne Wakeland
Systems Science Seminar Presentation
10/9/15
Assertion

• Models must, of course, be well suited to their intended application

• Thus, models for evaluating policies must be able to “predict” how the system is likely to respond to alternative policies
  – To a useful degree, and over a relevant time period

• One must, therefore, compare model predictions to what actually happens

• As recommended in the SD literature
  – But is rarely demonstrated
Must one Wait for Future to Unfold?

• It might be possible, for example, to blind oneself to the recent past and the use distant past to predict the more recent past
  – Concern is whether modeler is truly blind
    • Even a glance at a graph of recent outcomes could introduce subjective bias

• Another approach could be to use an algorithm for model calibration
  – Algorithms much less susceptible to subjective bias

• Predicting unknown future would be the most compelling test
Background

• Model testing has received considerable attention in SD literature
  – Predictive capability discussed some detail, but few examples are provided

• Model testing was often referred to as verification and validation
  – Authors have tended to avoid the word “validation” in order to avoid confusion with concept of statistical validity
    • Or the implication that SD models can be declared valid or invalid by running a set of tests
  – Emphasis is on rigorous and thorough testing processes, and establishing a model’s domain or boundary of applicability
Methods

• Revisit three SD policy / prescriptive models to determine accuracy of their predictions
  – In each case, model emphasized calibration of model against historical reference behavior

• Further, to examine underlying causes of prediction failures
Case 1: Fishery Regulation

• Stopping the decline of fish populations is very challenging
  – Rockfish landings were down nearly 80% and catch limits had been reduced by 78%-89%
  – West Coast ground fish fisheries were declared a federal disaster in 2000
• Likely due to ineffective natural resource management and short-term policies
  – Leading to a larger fishing fleet than could be supported
• Applications of SD to fisheries management are plentiful
Case 1: High-level CLD for Fisheries
Case 1: Model Calculations vs. Reference Data

Biomass

Acceptable Catch

Harvest
Case 2: Intracranial Pressure (ICP) Prediction

- Traumatic brain injury remains leading cause of death and disability in children
  - 30+% death rate for severe pediatric TBI
- Many sophisticated computer models have been created
- Parameters are typically estimated by calibrating models to fit patient-specific clinical data
  - Excellent results reported by Ursino and colleagues 2000
- Wakeland et al. 2009 was the first study to report actual prediction accuracy
  - Some studies refer to model calculations as predictions even though the study aim was to match (“predict”) reference data
    - Ursino, Minassian, Lodi et al. 2000
Case 2: Data Collection

- Patients given mild [IRB-approved] physiological challenges to estimate their state of autoregulation
  - Changing the head of bed between 0 and 30 degrees
  - Changing respiration rate to create mild hyper-ventilation and mild hypo-ventilation
- Patient ICP response carefully measured and recorded
- Goal: determine if patient-specific models could predict patient ICP response to interventions
  - And, ultimately, to use them to evaluate alternative treatments beforehand “in silico”
Case 2: Primary Stocks & Flows in ICP Dynamic Model


Subdural Bleeding → Hematoma

CSF Production → CSF Volume → Reabsorption of CSF

CSF Drainage
Case 2: ICP Model
(developed in STELLA and ported to Simulink for computation)
Case 2: Parameter Estimation Process to Create Patient-specific models

- **Parameters Estimated**
  - Autoregulation factor (smooth muscle compliance effect)
  - Basal cranial volume -- CSF drainage rate -- Hematoma increase rate
  - $\Delta$ pressure time constant (a smoothing parameter associated with HOB elevation change)
  - ETCO2 time constant (a smoothing parameter associated with RR changes)
  - Smooth muscle gain (a multiplicative factor related to the impact of smooth muscle tension)
  - Systemic venous pressure -- “Baseline” ICP -- Pressure volume index (PVI)
Case 2: Model Calibration Results (1)
Case 2: Model Calibration Results (2)
Case 2: Model Fitness (MAE/MAD) by patient, type of challenge, challenges/session, length of session, mean ICP

<table>
<thead>
<tr>
<th></th>
<th>P004</th>
<th>P006</th>
<th>P007</th>
<th>P201</th>
<th>P202</th>
<th>P204</th>
<th>P205</th>
<th>P206</th>
<th>P207</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MAE/MAD</strong></td>
<td>.53</td>
<td>.43</td>
<td>.99</td>
<td>1.06</td>
<td>.45</td>
<td>.52</td>
<td>.30</td>
<td>.53</td>
<td>.96</td>
<td>.72</td>
</tr>
<tr>
<td><strong>N</strong></td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>24</td>
</tr>
<tr>
<td><strong>Only HOB</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Challenges</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>MAE/MAD</strong></td>
<td>.89</td>
<td>.50</td>
<td>.61</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>N</strong></td>
<td>14</td>
<td>3</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>14</td>
</tr>
<tr>
<td><strong>Length of Session (minutes)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>&lt;=40</strong></td>
<td>.54</td>
<td>.62</td>
<td>.69</td>
<td>.93</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>41-60</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.47</td>
<td>.77</td>
<td>.91</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>61-80</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>&gt;80</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Mean ICP for Session (mmHg)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Low (&lt;12)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.47</td>
<td>.77</td>
<td>.91</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Medium (12-18)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>High (&gt;18)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>MAE/MAD</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.47</td>
<td>.77</td>
<td>.91</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>N</strong></td>
<td>5</td>
<td>9</td>
<td>6</td>
<td>4</td>
<td>8</td>
<td>10</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Case 3: Opioid Diversion & Abuse

• Motivation: dramatic rise in the nonmedical use of pharmaceutical opioid pain medicine and fatal overdoses; ineffective government policies and regulations

• SD models often used to study health policy

• Modeled medical use of pharmaceutical opioids to treat pain, drug diversion, and nonmedical use/outcomes

• 7 state variables, 90 support variables, 40 parameters

• Data from literature and other public sources
  – Direct empirical support for 12 params. indirect for 17 more

• Expert panel judgment for model structure and parameters lacking empirical support

• All but two highly influential parameters had some degree of empirical support
Case 3: High-level CLD

Popularity of Opioids for Nonmedical Use

Initiation + Nonmedical Users

Supply of Opioids

Dr. Shopping & Forgery + Chronic Pain Patients

Profit Motive

Overdose Deaths

Perceived Risk

B

MEDICAL USE

NONMEDICAL USE

Diversion +

Demand -

R

B

Clipped Text: Popularity of Opioids for Nonmedical Use

Nonmedical Users

Initiation +

Supply of Opioids

Dr. Shopping & Forgery +

Chronic Pain Patients

Profit Motive

Overdose Deaths

Perceived Risk

B

MEDICAL USE

NONMEDICAL USE

Diversion +

Demand -

R

B

Case 3: High-level CLD
Case 3: SFD for Medical Use Sector

- Low Frequency Nonmedical Opioid Users
- High Frequency Nonmedical Opioid Users
- Total Number of Individuals Using Opioids Nonmedically
- Rate of Initiation of Nonmedical Use
- Initiating Nonmedical Use
- Increasing Frequency
- Total Demand for Opioids
- US Population Aged Twelve Plus
- Opioid Popularity
- Number of Individuals Using Drugs Nonmedically (Excluding Marijuana and Pharmaceutical Opioids)
- Supply of Opioids Diverted by Patients
- Accessibility of Pharmaceutical Opioids
- Fraction of Demand Met from Chronic Pain Trafficking
- Rate of Initiation During Unlimited Accessibility
- B
- R
- <US Population Aged Twelve Plus>
Case 3: Model vs. Reference Behavior

Number of Initiates - RBP vs. Model Behavior

- Reference Behavior for the Number of Initiating Nonmedical Users: baseline
- Initiating Nonmedical Use of Opioids: baseline

Total Nbr of People Using Nonmedically vs. Reference Behavior

- Total Number of Individuals Using Opioids Nonmedically: baseline
- Reference Behavior for the Number of Nonmedical Users of Pharm Opioids: baseline

Total Overdose Deaths vs. Reference Behavior

- Total Number of Opioid Overdose Deaths per Year: baseline
- Reference Behavior for the Number of Overdose Deaths: baseline

Mean Absolute Percentage Error (MAPE): 10%, 9%, 22%
### Case 1: New Data

<table>
<thead>
<tr>
<th>Year</th>
<th>Harvest</th>
<th>Spawning Biomass</th>
<th>ABC</th>
<th>MSY (OY)</th>
<th>Likely Sp. Biomass</th>
</tr>
</thead>
<tbody>
<tr>
<td>1992</td>
<td></td>
<td>18,000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1995</td>
<td></td>
<td>15,822</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1998</td>
<td></td>
<td>15,735</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1999</td>
<td></td>
<td>16,955</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td>3735</td>
<td>17,909</td>
<td>3539</td>
<td></td>
</tr>
<tr>
<td>2001</td>
<td></td>
<td>2142</td>
<td>18,467</td>
<td>3146</td>
<td></td>
</tr>
<tr>
<td>2002</td>
<td></td>
<td>1260</td>
<td>18,783</td>
<td>3146</td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td></td>
<td>551</td>
<td>16,324</td>
<td>3146</td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td></td>
<td>618</td>
<td>17686</td>
<td>4320</td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td></td>
<td>892</td>
<td>16915</td>
<td>4320</td>
<td>4940</td>
</tr>
<tr>
<td>2006</td>
<td></td>
<td></td>
<td></td>
<td>4680 (4548)</td>
<td>4743</td>
</tr>
<tr>
<td>2007</td>
<td></td>
<td></td>
<td></td>
<td>4634</td>
<td>15,717</td>
</tr>
</tbody>
</table>

**Note:** The values for MSY (OY) and Likely Sp. Biomass are indicative of the decision table’s context.
Case 1: Prediction Accuracy

<table>
<thead>
<tr>
<th></th>
<th>Spawning Biomass</th>
<th>ABC</th>
<th>Harvest</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model Fit Error</td>
<td>19%</td>
<td>24%</td>
<td>27%</td>
<td>20</td>
</tr>
<tr>
<td>Model Prediction Error</td>
<td>14%</td>
<td>51%</td>
<td>601%</td>
<td>6 for Harvest, 8 for SB and ABC</td>
</tr>
</tbody>
</table>
Case 1: Prediction Discussion

- Model did not capture regulatory agencies behavior
- Small changes $\rightarrow$ significant effect
  - *Spawning biomass* levels indicate “normal” fishing: $ABC = 18\%$ of *mature fish*
  - But, regulators chose to leave the fishery as “precautionary” w/$ABC = 12\%$
  - This accounts for much of the model prediction error for $ABC$
- Results question whether endogenously modeling fishery regulation is possible
  - Regulators use judgment and do not set rules based only on the numbers
    - Big challenge for modelers striving to model fishery regulatory processes
  - E.g., closing a fishery because a co-mingled fishery is in danger
    - Model boundary issue
  - Supports Pilkey and Pilkey-Jarvis (2007) assertion that environmental scientists “cannot predict the future” even with (or perhaps because of) their reliance on quantitative models
Case 2: Example Prediction Results (1)
Case 2: Example Prediction Results (2)
## Case 2: Prediction Error w/in Segment (MAE/MAD)

<table>
<thead>
<tr>
<th>Patient</th>
<th>Best Fit</th>
<th>Predicted</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>P004</td>
<td>.43</td>
<td>1.88</td>
<td>3</td>
</tr>
<tr>
<td>P006</td>
<td>.48</td>
<td>.59</td>
<td>5</td>
</tr>
<tr>
<td>P007</td>
<td>.83</td>
<td>3.49</td>
<td>3</td>
</tr>
<tr>
<td>P201</td>
<td>1.81</td>
<td>1.79</td>
<td>4</td>
</tr>
<tr>
<td>P202</td>
<td>.38</td>
<td>3.50</td>
<td>2</td>
</tr>
<tr>
<td>P204</td>
<td>.81</td>
<td>2.57</td>
<td>2</td>
</tr>
<tr>
<td>P205</td>
<td>.76</td>
<td>1.43</td>
<td>1</td>
</tr>
<tr>
<td>P206</td>
<td>.62</td>
<td>1.61</td>
<td>1</td>
</tr>
<tr>
<td>P207</td>
<td>.94</td>
<td>1.03</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>.82</td>
<td>1.90</td>
<td>22</td>
</tr>
</tbody>
</table>
## Case 2: Prediction Error between Sessions

<table>
<thead>
<tr>
<th>Patient</th>
<th>Prediction Error (MAE/MAD)</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>P004</td>
<td>1.93</td>
<td>6</td>
</tr>
<tr>
<td>P006</td>
<td>1.99</td>
<td>10</td>
</tr>
<tr>
<td>P007</td>
<td>2.34</td>
<td>3</td>
</tr>
<tr>
<td>P201</td>
<td>2.99</td>
<td>6</td>
</tr>
<tr>
<td>P202</td>
<td>2.88</td>
<td>6</td>
</tr>
<tr>
<td>Overall</td>
<td>2.41</td>
<td>31</td>
</tr>
</tbody>
</table>
Case 2: Discussion

• Model prediction error for ICP is far too large to be clinically useful
  – Disappointing, as model fitness to RBP was much better
  – Fitness to RBP may not indicate model’s utility for prescriptive analysis

• Prediction is hard, especially for human physiology
  – Due, in part, to high degree of non-stationarity

• Ultimately, the patient-specific model research was abandoned
  – Due to high intra-patient non-stationarity / variability
  – Though well-known to clinicians and easily seen in the data, it was the attempt to make predictions that forced researchers to revise their expectations...
Case 3: Prediction Errors (2009-2013)

- 5-year MAPE
  - 7%, 14%, 3%
Case 3: Discussion

• Five-year prediction errors of 7%, 14%, and 3% seem respectable
• But, these predictions did not capture the reduction in initiation and number of nonmedical users
• Might not be a bad thing altogether, because the baseline model assumed no policy change
  – Whereas, in 2011, the most abused medicine, OxyContin©, was re-issued in a truly tamper-resistant formulation, and since then, it has been less diverted and abused
  – Also, prescription drug monitoring programs are now operating in 49 states
    • Prescribers can check to see if their patients are getting medicines from other docs; and, some prescribers are being more cautious
• Making predictions and checking their accuracy added value beyond the replication of reference behavior
Study Limitations

• Was based on three projects led by a single researcher
  – Findings could be highly biased and non-representative
  – Future work should involve models created by multiple researchers to avoid potential biases and idiosyncrasies

• Method was retrospective, subjective, and did not employ a refutable hypothesis coupled with earnest efforts to refute that hypothesis
  – Such an approach could strengthen support for the assertion that prediction tests are the quintessential model tests for SD-based policy/prescriptive models
Conclusion

• When model objectives include forward-looking policy evaluation, testing prediction accuracy can be important.
• When automated calibration algorithms are used, it may be sufficient to hold back part of the data, calibrate model using a training subset, and measure prediction performance using the holdout sample.
• If manual calibration is used, modeler must be blind to recent outcomes, make predictions of recent outcomes, get the actual data, and measure prediction performance.
A Nagging Worry

• Do complex models that more fully reflect system interconnectivity and dynamics actually predict system behavior better?
  – Conventional wisdom, and likely empirical evidence, may suggest otherwise
  – When forecasting, simple models often outperform complex models

• These modeling cases are thought-provoking, and seem to indicate that complex models should be used with considerable caution...
Further Reflections

• More complex SD models can lead to deep insights into structure and behavior that are likely not possible with simple non-parametric models
  – The point is not that SD models should be used for making predictions, but rather that prediction testing is useful to test whether a policy-oriented model is ready to be deployed

• Hmmm. Does “policy analysis” actually require prediction?
  – Certainly prescriptive models (such as the ICP dynamics model) must be able to predict
  – But do policy analysis models need to make accurate predictions?
  – Could a model with poor numerical predictive ability still make useful qualitative predictions that lead to deep and useful insights?
    • If so, then how might a modeler assess qualitative predictive utility?