Ex Ante
Evaluation and Improvement of Forecasts

INFORMS Annual Meeting
Seattle, Washington
November 7, 2007

湯威頤

Víctor Tang PhD victang@alum.mit.edu
Kevin N. Otto PhD otto@robuststrategy.com
Warren P. Seering PhD seering@mit.edu
Abstract

Ex ante evaluation and Improvement of Forecasts
Victor Tang, Kevin N. Otto, Warren P. Seering

The dominant approach reported in the literature is to evaluate forecasts after the fact. We take a different approach, we present a way to evaluate and improve forecasts before the fact. We reconceptualize forecasts as thought experiments grounded on mental models. We show the results of our process which debiases and reduces the asymmetry of forecasters’ mental models. We also reconceptualize forecasting as measurements with errors. And to analyze and improve the entire forecasting process as a system, we use the methods of Design of Experiments (DOE) and Gage R&R from Measurement System Analysis (MSA). We show the results of our analyses using two new metrics, repeatability and reproducibility and discuss new opportunities for research.
Forecasting evaluation examples.

Mean absolute error

\[
MAE = \frac{1}{N} \sum_{i=1}^{N} |F_i - O_i| \quad 0 \leq MAE \leq \infty \\
0 = perfect \ score
\]

Brier score

\[
BS = \frac{1}{N} \sum_{k=1}^{K} n_k (p_k - \bar{o}_k)^2 - \frac{1}{N} \sum_{k=1}^{K} n_k (\bar{o}_k - \bar{o})^2 + \bar{o}(1-\bar{o}) \quad 0 \leq MBS \leq 1 \\
0 = perfect \ score
\]

Heidke skill score

\[
HSS = \left( \frac{\frac{1}{N} \sum_{i=1}^{K} n(F_i,O_i) - \frac{1}{N^2} \sum_{i=1}^{K} N(F_i)N(O_i)}{1 - \frac{1}{N^2} \sum_{i=1}^{K} N(F_i)N(O_i)} \right) \quad -\infty \leq HSS \leq 1 \\
0 = no \ skill \\
1 = perfect \ score
\]

... but typically, evaluation is after the fact.

- **ex ante**
  - improve
  - evaluate

- **input** → **forecasting** → **forecast**
- **after the fact**
  - evaluate
  - improve

- e.g. MAE, scoring methods, ...

---

We take a different approach.

We re-conceptualize **forecasting** as thought experiments, which are grounded on **mental models**.

We re-conceptualize **forecasts** as **measurements with errors**, which are grounded on a **measurement system** composed of forecasters, their databases, formal and informal procedures.

By addressing the mental models and analysis of the **measurement system**, we can **ex ante** evaluate and improve the entire forecasting system before the fact.

We re-conceptualize forecasting as thought experiments, which are grounded on mental models.

- Bias
- Group think,
- Herding

Problems

- Asymmetric mental models

Solution

- Debiasing
- Counter-argumentation
- Accountability

Debiasing $\Rightarrow$ stdev declines and confidence rises

forecasts = stdev ↓

forecasts ↑ stdev ↓

forecasts ↓ stdev ↓

confidence rises

stdev ↓
We develop an experimental process using Design of Experiments (DOE) where each treatment is a forecast.

1. Specify the desired output (dependent variable), $Y = \varphi (f_1, f_2, ..., f_m, n_1, n_2, ..., n_K)$
2. Specify the independent variables, $\{f_i\}$, $i=1, ..., m$
   Controllable and uncontrollable, $\{n_j\}$, $j=1, ..., k$
3. Specify the most frugal orthogonal array (OA) of treatments, $L_p(\alpha^m, \beta^k)$
4. Specify the most distinct treatments, $\{(f'_1, f'_2, ..., f'_m)\}_q, q=1, ..., s, s \sim q/4$
   relative to the orthogonal array using the Hat matrix. $H = L_p(L'_p L_p)^{-1} L'_p$
   Call this set the supplemental treatments, $S=\{(f'_1, f'_2, ..., f'_m)\}_q$
5. Forecast the output of all the treatments above.

Note
1. Have parameterized the entire space of forecasts.
2. As well as, the entire uncertainty space.
3. The OA is sufficient to derive the output of any forecast for any uncertainty condition.
4. Comparing the supplemental forecasts versus derived outputs can give us a sense of the quality of the forecasts.
   Quality = repeatability and reproducibility

There is support for the choice of variables

**ANOVA**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Seq SS</th>
<th>Adj SS</th>
<th>Adj MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>SG&amp;A</td>
<td>1</td>
<td>30.366</td>
<td>54.974</td>
<td>54.974</td>
<td>164.66</td>
<td>0.000</td>
</tr>
<tr>
<td>COGS</td>
<td>1</td>
<td>676.518</td>
<td>92.328</td>
<td>92.328</td>
<td>276.55</td>
<td>0.000</td>
</tr>
<tr>
<td>capacity</td>
<td>1</td>
<td>27.992</td>
<td>14.933</td>
<td>14.933</td>
<td>44.73</td>
<td>0.000</td>
</tr>
<tr>
<td>portfolio</td>
<td>2</td>
<td>109.505</td>
<td>8.605</td>
<td>4.302</td>
<td>12.89</td>
<td>0.000</td>
</tr>
<tr>
<td>sales</td>
<td>1</td>
<td>44.122</td>
<td>33.583</td>
<td>33.583</td>
<td>100.59</td>
<td>0.000</td>
</tr>
<tr>
<td>financing</td>
<td>2</td>
<td>6.361</td>
<td>20.558</td>
<td>10.279</td>
<td>30.79</td>
<td>0.000</td>
</tr>
<tr>
<td>COGS*capacity</td>
<td>1</td>
<td>3.212</td>
<td>2.488</td>
<td>2.488</td>
<td>7.45</td>
<td>0.008</td>
</tr>
<tr>
<td>portfolio*sales</td>
<td>2</td>
<td>3.887</td>
<td>3.887</td>
<td>1.944</td>
<td>5.82</td>
<td>0.005</td>
</tr>
<tr>
<td>Error</td>
<td>67</td>
<td>22.368</td>
<td>22.368</td>
<td>0.334</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>78</td>
<td>934.330</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

S = 0.577800  R-Sq = 97.61%  R-Sq(adj) = 97.21%
We re-conceptualize forecasts as measurements with errors, which are can be analyzed using statistical methods.

Problems
- How do we know the extent of guessing?
- What are some *ex ante* metrics?

Solution
Consider the participants who are forecasting, their knowledge, data bases, formal and informal procedures, and their network of contacts as a *measurement system*.

Gage R&R from Measurement Systems Analysis (MSA) provides us with a method to determine *repeatability* and *reproducibility*.
Gage R&R

\[ \sigma^2_{rpt} = \sigma^2_{part} + \sigma^2_{rpd} + \sigma^2_{rpt} \]

Individual forecasts of 5 (test) treatments gives us an indication of *reproducibility* across “operators”.

![](image)
Forecasts vs. derived estimates give an indication of an operator’s *repeatability* across forecasts.

**Participant 1**

**Participant 2**

**Participant 3**

**Participant 4**

**Participant 5**

*Individual forecasts*

*Derived estimates using L18 operator data*

Source: gage r&r calculations.xls
We can improve low quality data

Overall variation in measurements (forecasts)

- Actual variation part-part
  - Variation over all treatments

- Measurement system variation
  - Gage R&R

Repeatability
- Variation in forecast by one operator for a given treatment

Reproducibility
- Variation in forecasts of different operators for a given treatment

<table>
<thead>
<tr>
<th></th>
<th>48 %</th>
<th>82 %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3 %</td>
<td>11 %</td>
</tr>
<tr>
<td></td>
<td>49 %</td>
<td>7 %</td>
</tr>
</tbody>
</table>

all w/o op 4
Summary

New way to think about forecasts, forecasting, and their evaluation.

- Forecasts are *thought experiments* based on mental models.
- Forecasts are *measurements* with errors.

*Ex ante* evaluation and improvement of the entire forecasting system before the fact.

- By debiasing and reducing the asymmetry of the mental models,
- By analyzing the measurements and their errors,
- By using the engineering methods of *Design of Experiments* (DOE) and *Gage R&R*.

Two new measures of forecasting quality:

- repeatability,
- reproducibility.