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

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|F_i - O_i\right|$$

$$0 \le MAE \le \infty$$

$$0 = perfect \quad score$$

Brier score

$$BS = \frac{1}{N}\sum_{k=1}^{K} n_k\left(p_k - \bar{o}_k\right)^2 - \frac{1}{N}\sum_{k=1}^{K} n_k\left(\bar{o}_k - \bar{o}\right)^2 + \bar{o}\left(1 - \bar{o}\right)$$

$$0 \le MBS \le 1$$

$$0 = perfect \quad score$$

Heidke skill score

$$HSS = \frac{\frac{1}{N}\sum_{i=1}^{K} n(F_i, O_i) - \frac{1}{N^2}\sum_{i=1}^{K} N(F_i)N(O_i)}{1 - \frac{1}{N^2}\sum_{i=1}^{K} N(F_i)N(O_i)}$$

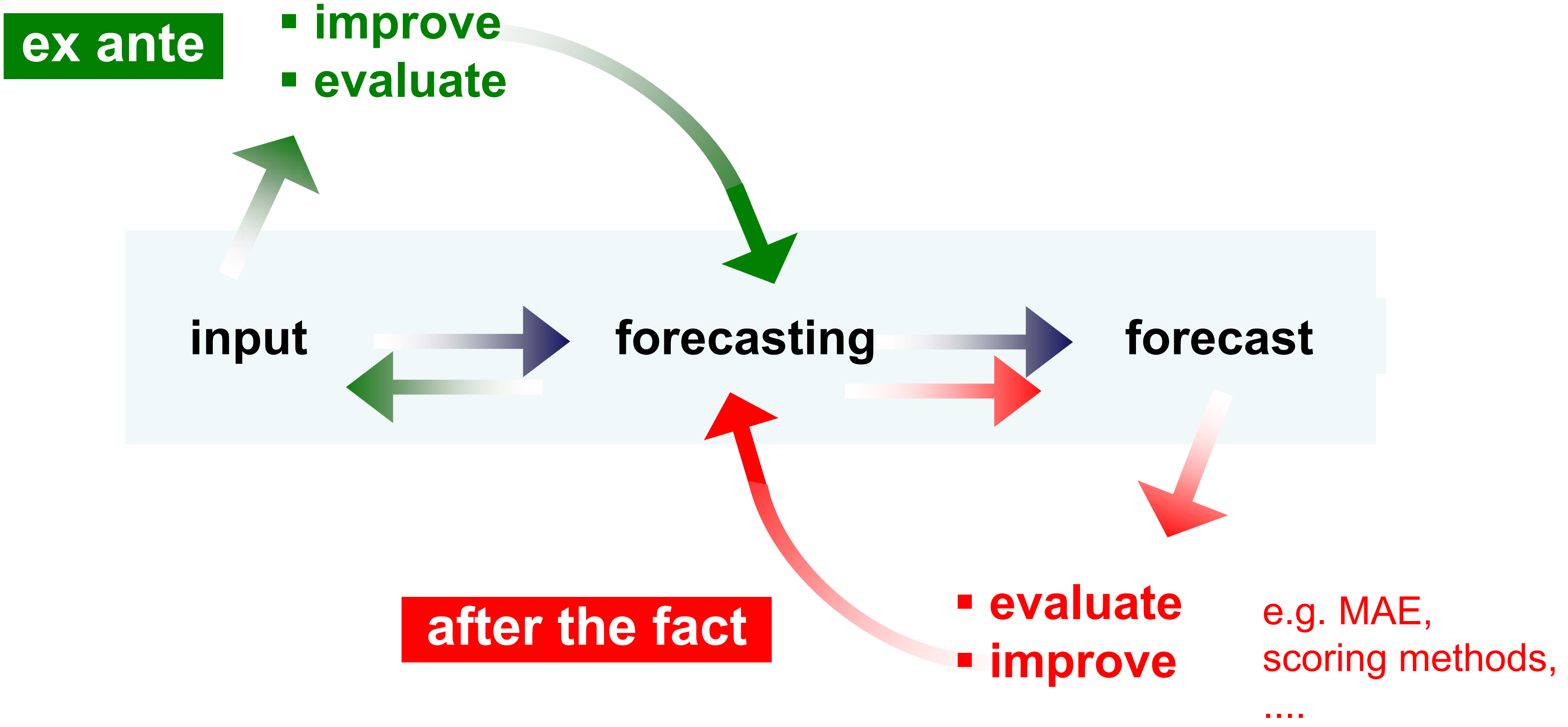
$$-\infty \le HSS \le 1$$

$$0 = no \quad skill$$

$$1 = perfect \quad score$$

1. Yates, J, Frank. 1982. External Correspondence: Decomposition of the Mean Probability Score. Org. Behavior and Human Performance 30: 132-156
2. Murphy, Allan H. 1973. A New Vector Partition of the Probability Score. Meteorology 12: 595-600.
3. Bryan, J.G., I. Enger. 1967. Use of probability Forecasts to Maximize Various Skill Scores. Journal of Applied Meteorology. 6:762-769

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



1. Singer, N C; W.P. Seering, 1990. Preshaping command inputs to reduce system vibration , ASME, Transactions, Journal of Dynamic Systems, Measurement, and Control. Vol. 112, pp. 76-82.
2. V. Tang, "Corporate Decision Analysis: An Engineering Approach." Ph.D. dissertation, Engineering Systems Division, MIT, Cambridge, MA, 2006. available: http://esd.mit.edu/people/dissertations/tang_victor.pdf

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.

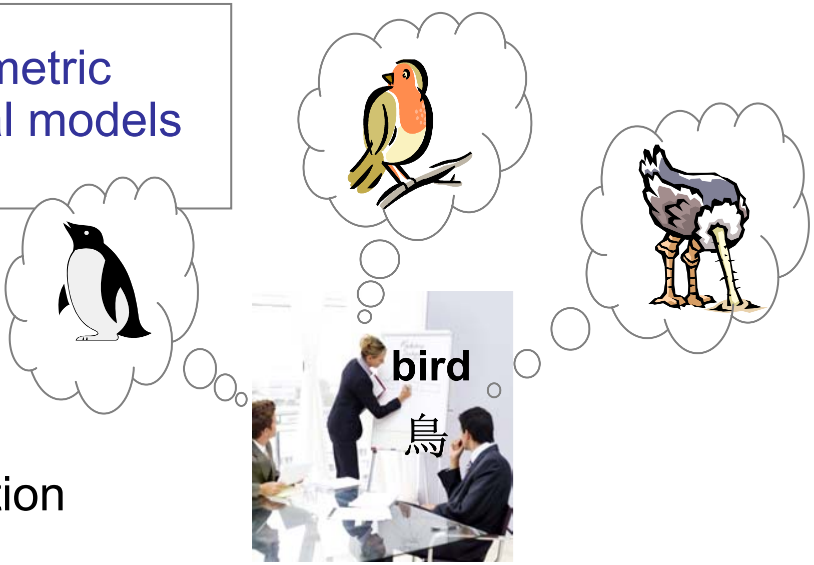
1. Mathieu, J.E., T.S. Heffner, G.F. Goodwin, E. Salas, J.A. Cannon-Bowers. 2000. The influence of shared mental models on team process and performance. Journal of Applied Psychology. Apr. 85(2): 273-283.
2. Gentner, D., A.L. Stevens. 1983. Mental Models. Lawrence Erlbaum Associates, Inc. Hillside, NJ.
3. Poses, R.M., C. Bekes, R.L. Winkler, W.E. Scott, F.J. Copre. 1990. Are two (inexperienced) heads better than one (experienced) head? Averaging house officers' prognostic judgments for critically ill patients. Arch. Of Internal Medicine. 150(9).
4. Kee, F. T. Owen, R. Leathem. 2007. Offering a prognosis in lung cancer: when is a team of experts an expert team? Journal of Epidemiology and Community Health. 61: 308-313.
5. Yaniv, I. 2004. Receiving other people's : Influence and benefit. Organizational Behavior and Human Decision Processes 93(1) 1-13.
6. Hubbard, A., R.H. Ashton.1985. Aggregating subjective forecasts: Some empirical results. Management Science 12(December): 1499-1508.

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

Problems

- Bias
- Group think,
- Herding

- Asymmetric mental models

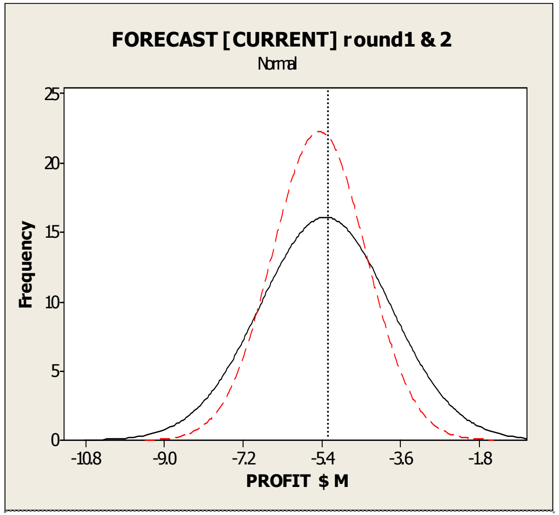


Solution

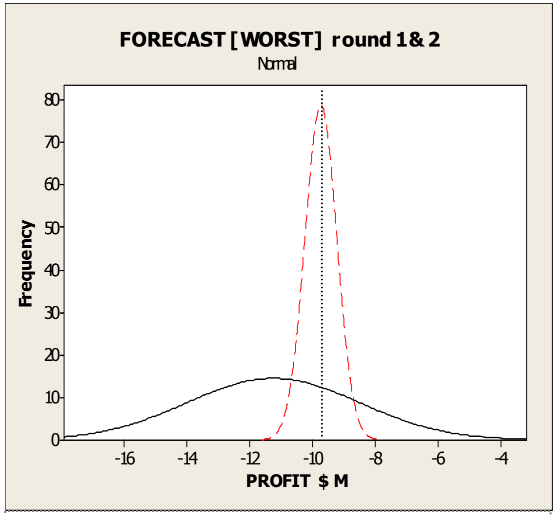
- Debiasing
- Counter-argumentation
- Accountability

1. D. Kahneman, P. Slovic, and A. Tversky. (1982). “Judgment under uncertainty: heuristics and biases.” Ch. 1, *Judgment under Uncertainty: Heuristics and Biases*. Cambridge, UK: Cambridge Univ. Press, pp. 3-20.
2. R. Hastie and T. Kameda.. (2005, April). “The robust beauty of majority rules in group decisions.” *Psychology Review* 112(2), pp. 494-508.
3. H. R. Arkes. (2001). “Overconfidence in judgmental forecasting.” In *Principles of Forecasting: A Handbook for Researchers and Practitioners,*
4. J. Scott Armstrong, Ed. Dordrecht, Netherlands: Kluwer Academic Publishers, pp. 495-515.
5. Janis. (1992). “Causes and consequences of defective policy making: A New Theoretical Analysis.” Ch. 2, *Decision-Making and Leadership*,
6. F. Heller, Ed. Cambridge, UK: Cambridge Univ. Press, pp. 11-45.
7. L. J. Kray and A. D. Galinsky. (2003). “The debiasing effect of counterfactual mind-sets: Increasing the search for disconfirmatory information in group decisions.” *Organizational Behavior and Human Decision Processes* (91), pp. 69-81.
8. H. R. Arkes. (2001). “Overconfidence in judgmental forecasting.” In *Principles of Forecasting: A Handbook for Researchers and Practitioners,*
9. J. E. Russo and P. J. Schoemaker. (1992, Winter). “Managing overconfidence.” *Sloan Management Review* (33), pp. 7-17.
10. B. Fishoff. (1999). “Debasing.” Chapter 31 in *Judgment under Uncertainty: Heuristic and Biases,* D. Kahneman, P. Slovic, A. Tversky, Eds. Cambridge, UK: Cambridge Univ. Press, pp. 422-444.
11. J. F. Yates, E. S. Veinott, and A. L. Patalano. (2003). “Hard decisions, bad decisions: On decision quality and decision aiding.” Ch. 1, *Emerging Perspectives on Judgment and Decision Research,* S. L. Schneider and J. Shanteau, Eds. Cambridge, UK: Cambridge Univ. Press, pp. 34.
12. W. Edwards and D. von Winterfeldt. (1986). “On cognitive illusions and their implications.” Chapter 40 in *Judgment and Decision Making: An Interdisciplinary Reader,* H. R. Arkes and K.R. Hammond, Eds. Cambridge, UK: Cambridge Univ. Press, pp 642-679.

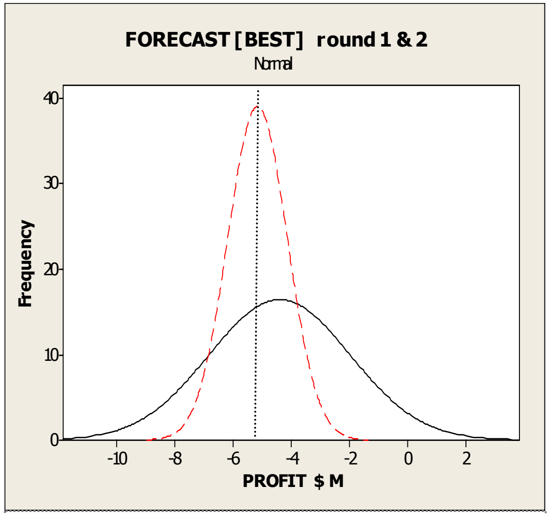
Debiasing → stdev declines and confidence rises



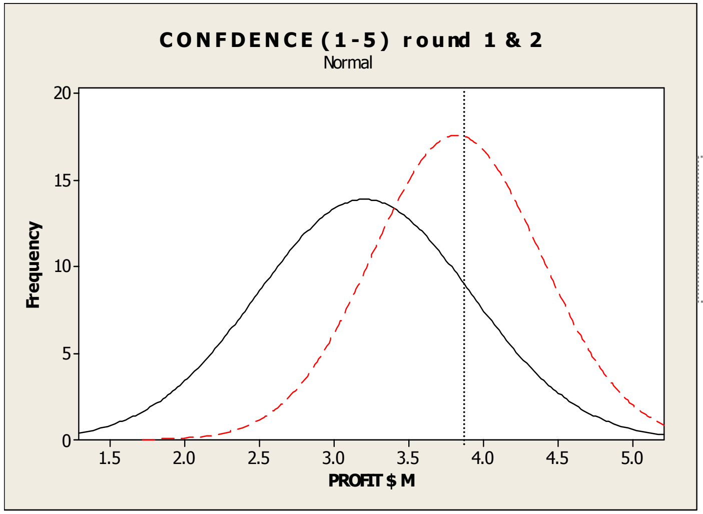
forecasts =
stdev ↓



forecasts ↑
stdev ↓



forecasts ↓
stdev ↓



confidence rises
stdev ↓

— round 1
- - - round 2

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

1. Specify the desired output (dependent variable), $Y = \varphi\ (f_1,f_2,\ldots,f_m,n_1,n_2,\ldots,n_K)$
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, $\mathbf{L_p(\alpha^m,\beta^k)}$
4. Specify the most distinct treatments, $\{(f'_1,f'_2,\ldots,f'_m)\}_q$, q=1,...,s, s~q/4
   relative to the orthogonal array using the Hat matrix. $\mathbf{H = L_p(L'_pL_p)^{-1}L'_p}$
   Call this set the supplemental treatments, $S=\{(f'_1,f'_2,\ldots,f'_m)\}_q$
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

I. N. Vucjkov and L. N. Boyadjieva. (2001). *Quality Improvement with Design of Experiments.* Dordrecht, Netherlands: Kluwer Academic Publishers.

K. S. Phadke. (1989). *Quality Engineering Using Robust Design*. Englewood Cliffs, NJ: Prentice Hall.

C. F. J. Wu and M. Hamada. (2000). *Planning, Analysis, and Parameter Design Optimization.* Hoboken, NJ: Wiley Series in Probability and Statistics.

There is support for the choice of variables

ANOVA

Supplemental treatments

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

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

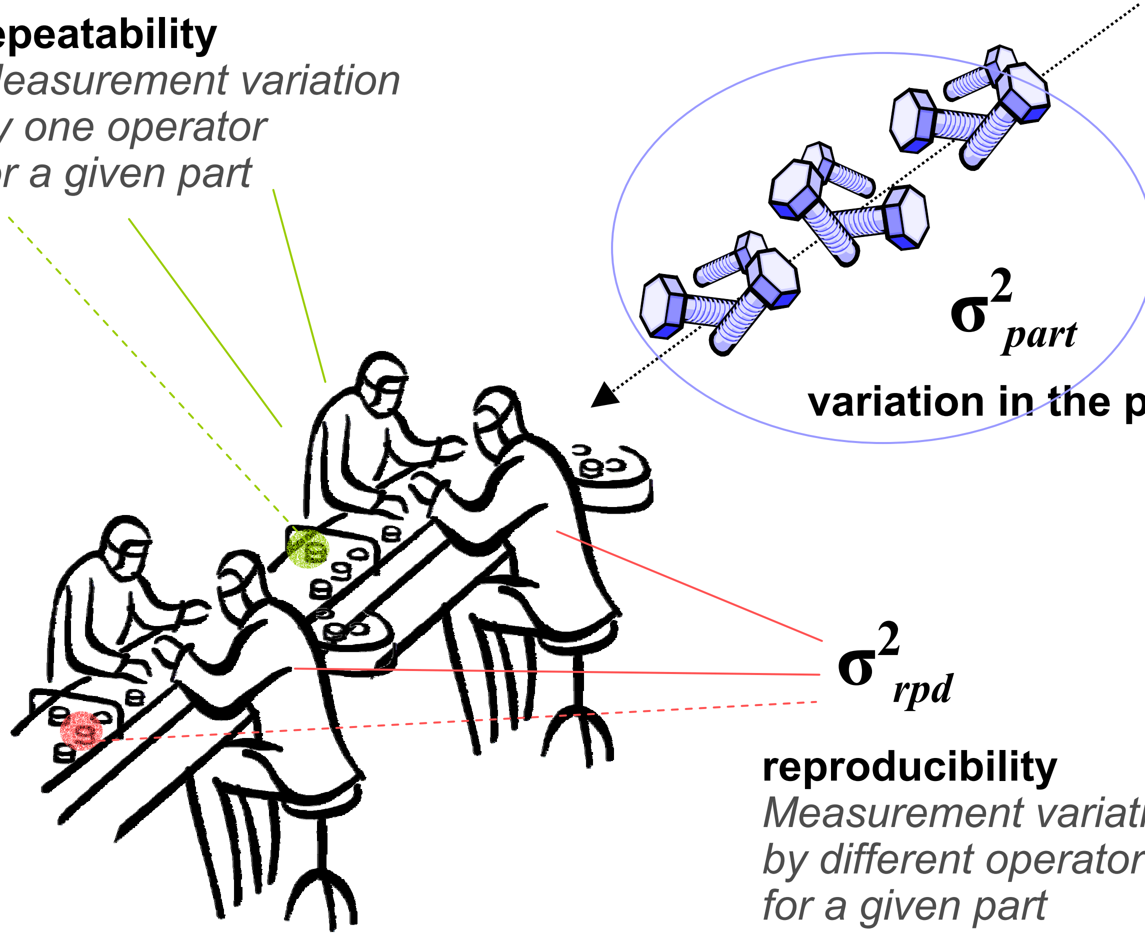
Consider the participants who are forecasting, their knowledge, data bases, formal and informal procedures, and their network of contacts as a ***measurement system***.

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

Gage R&R

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

σ^2_{rpt} **repeatability**
Measurement variation
by one operator
for a given part



σ^2_{part}

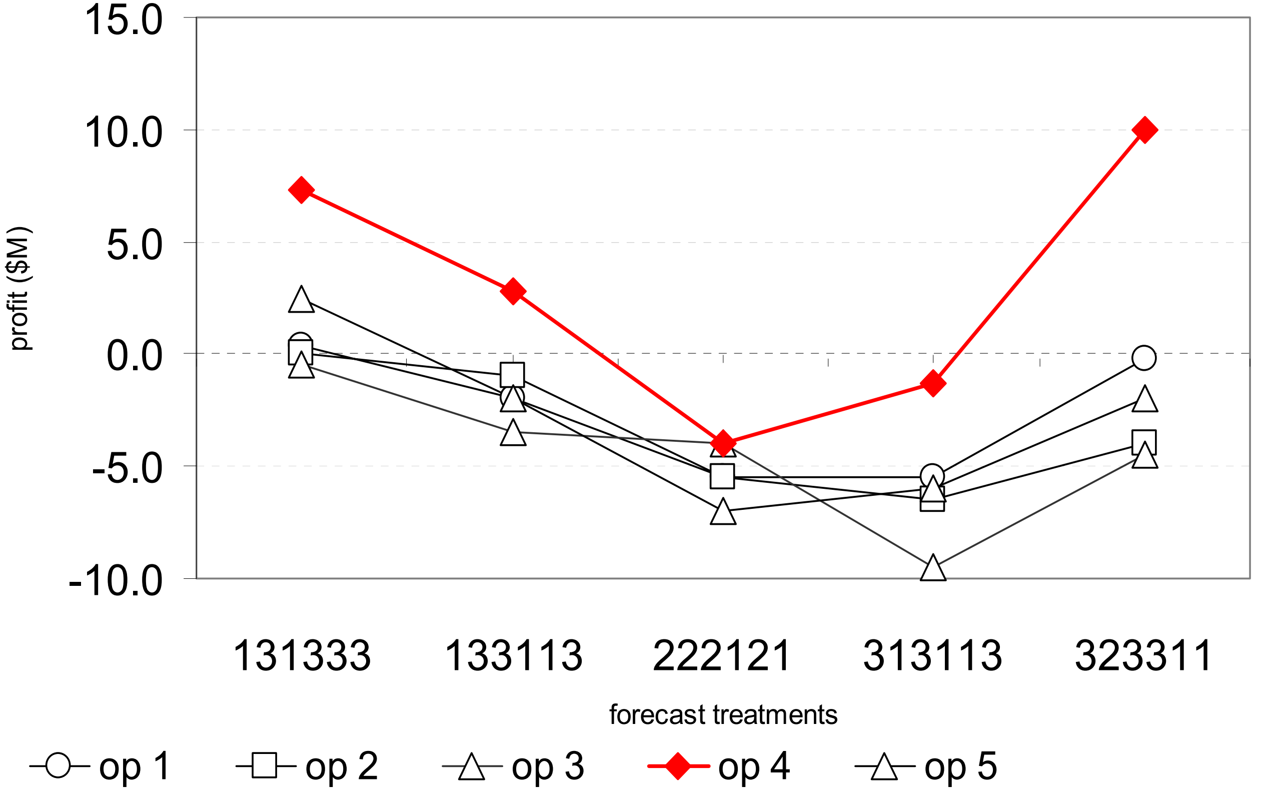
variation in the parts

σ^2_{rpd}

reproducibility
Measurement variation
by different operators
for a given part

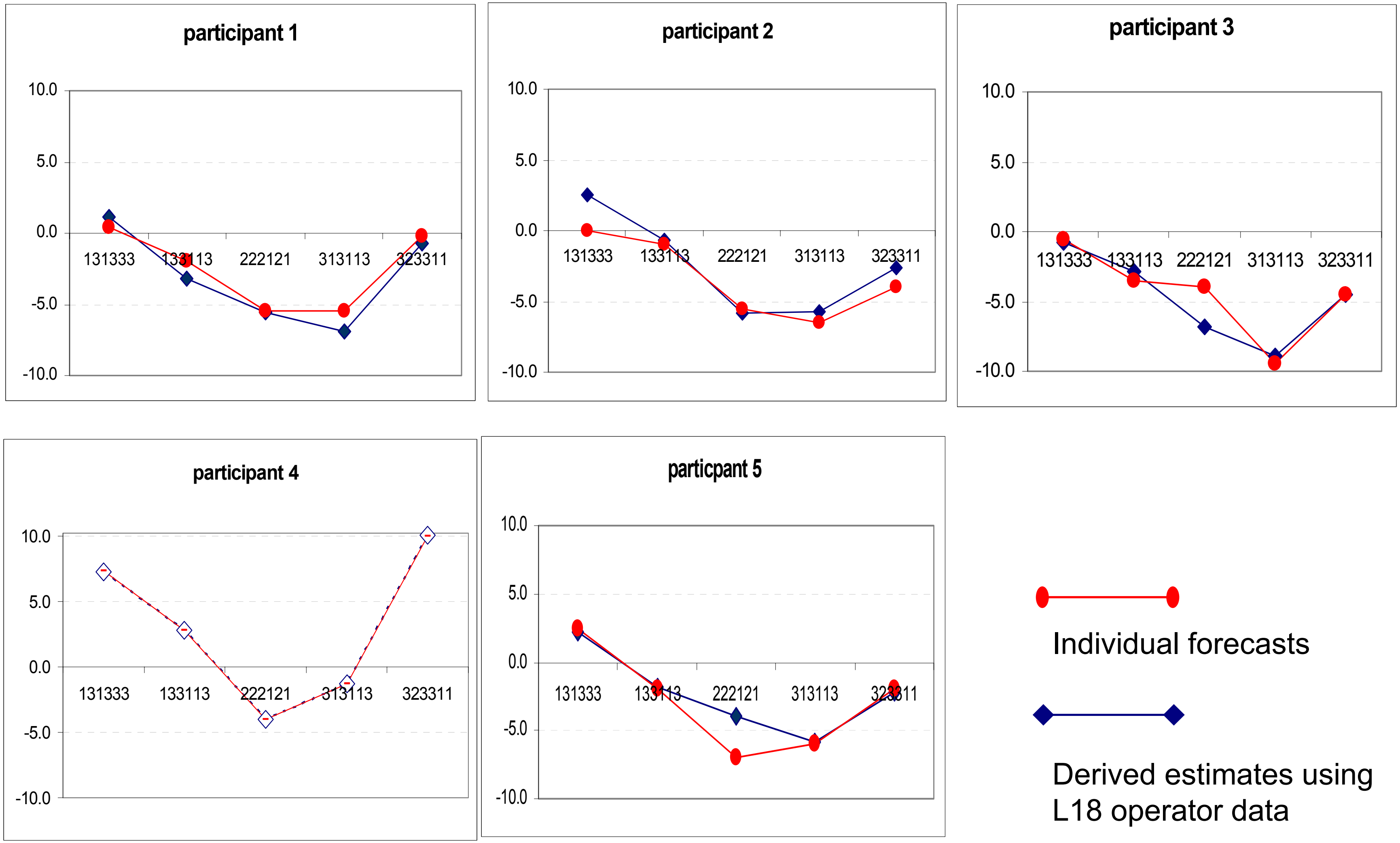
1. D. C. Montgomery. (2001). *Design and Analysis of Experiments*. Hoboken, NJ: John Wiley & Sons.
2. *Measurement Systems Analysis*. Reference Manual (3rd Edition). (2002). Copyright DaimlerChrysler, Ford Motor Co., GM..

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



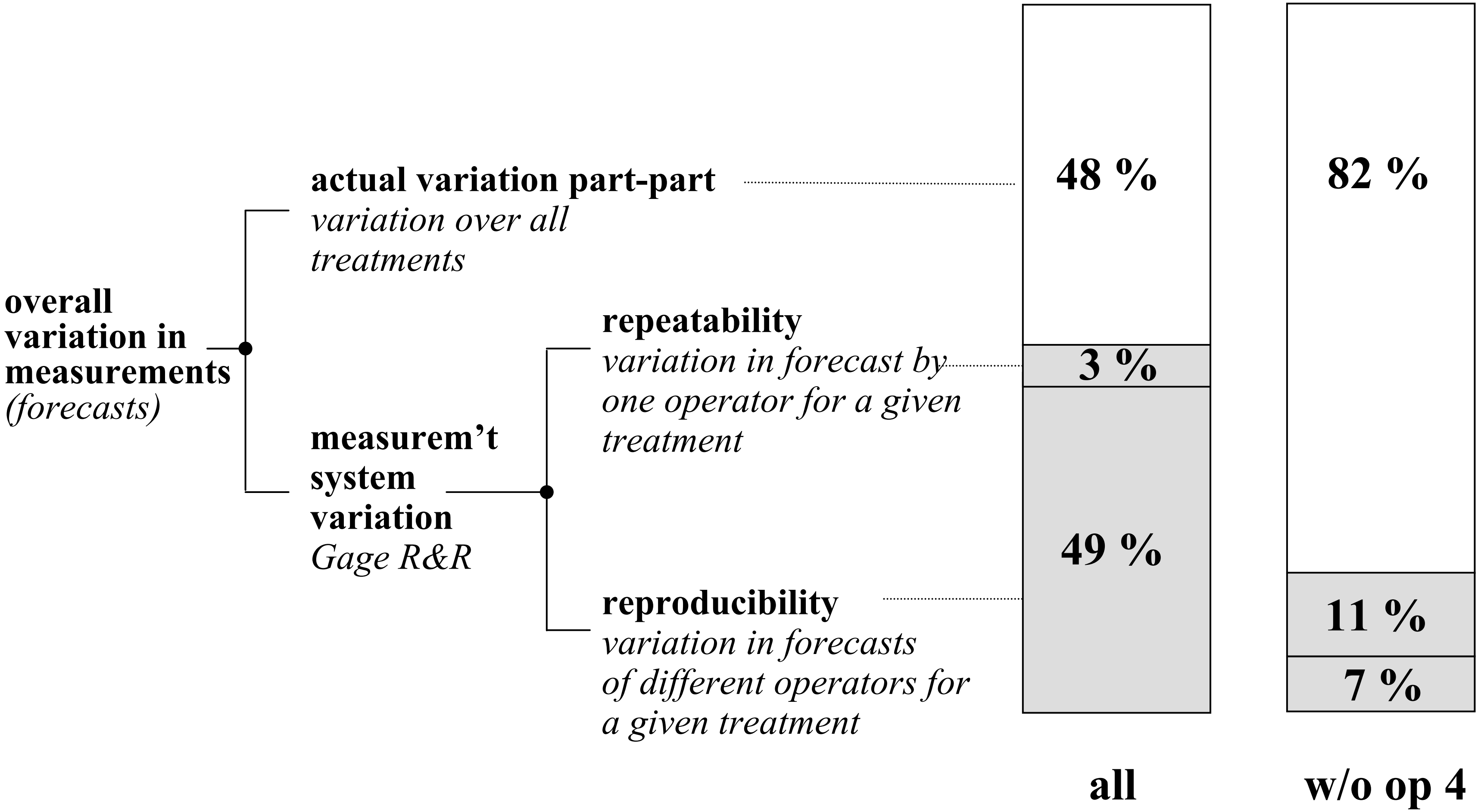
©vtang

Forecasts vs. derived estimates give an indication of an operator’s *repeatability* across forecasts.



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*

measurem’t 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*

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.