Calibration Statistics
### Selecting a Statistic and Setting a Standard

1996 System Dynamics Conference

Kimberly Sklar Reichelt,
Dr. James M. Lyneis & Carl G. Bespolka
PUGH-ROBERTS ASSOCIATES

Calibration Statistics

- Why We Care About Statistics
- Alternative Statistics to Consider
- Selecting a Best Statistic
- Setting Standards

Why We Care about Statistics

- To enhance the confidence of others
- For comparison against other calibrations
- To assure that minimum standards are met
- For use in "automatic" calibration and sensitivity analysis software
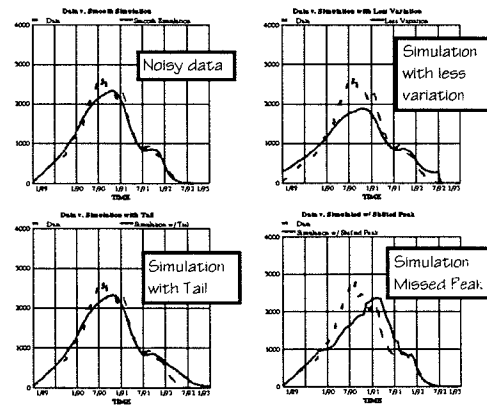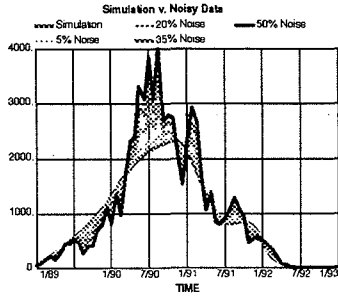

The Ideal Statistic

- Provides meaningful results in a variety of circumstances
- Easily interpreted
- Easy to explain ("intuitive")
- Provides a consistent measure
- Is sensitive to important differences
- Is relatively insensitive to small differences when either the simulation or the data is close to zero
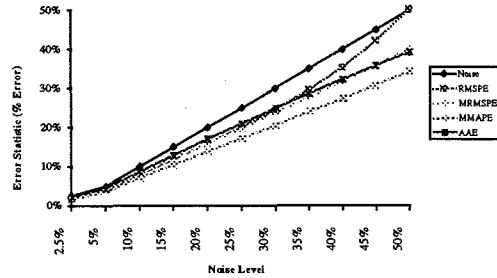
Alternative Statistics

- R^2
- Mean Absolute Percent Error (MAPE)
- Root Mean Square Percent Error (RMSPE)
- Modified Mean Absolute Percent Error (MMAPE)
- Modified Root Mean Square Percent Error (MRMSPE)
- Theil Statistics (U)
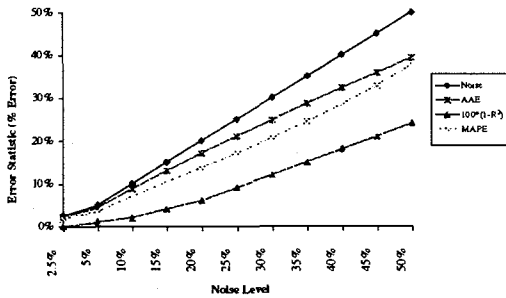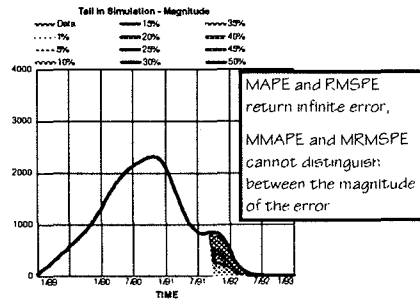- Average Absolute Error (AAE)

Handling Noisy Data

Simulation v. Noisy Data

Legend: Simulation | 20% Noise | 50% Noise | 5% Noise | 35% Noise

Y-axis: 4000, 3000, 2000, 1000, 0

X-axis (TIME): 1/89, 1/90, 7/90, 1/91, 7/91, 1/92, 7/92, 1/93

Response of RMSPE, MRMSPE, MMAPE and AAE to Variations in "Noise"

Y-axis (Error Statistic (% Error)): 50%, 40%, 30%, 20%, 10%, 0%

X-axis (Noise Level): 2.5%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%

Legend: Noise | RMSPE | MRMSPE | MMAPE | AAE

Response of AAE, R² and MAPE to Variations in "Noise"

Y-axis (Error Statistic (% Error)): 50%, 40%, 30%, 20%, 10%, 0%

X-axis (Noise Level): 2.5%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%

Legend: Noise | AAE | 100*(1-R²) | MAPE

Measuring Sensitivity to Tail Height

Tail in Simulation - Magnitude

Legend: Data | 15% | 35% | 1% | 20% | 40% | 5% | 25% | 45% | 10% | 30% | 50%

Y-axis: 4000, 3000, 2000, 1000, 0

X-axis (TIME): 1/89, 1/90, 7/90, 1/91, 7/91, 1/92, 1/93

MAPE and RMSPE return infinite error,

MMAPE and MRMSPE cannot distinguish between the magnitude of the error

---

Measuring Sensitivity to Tail Length

Tail in Simulation - Length

Legend: Simulation | .75 | 1.75 | .0 | 2 | .25 | 1.25 | 2.25 | .50 | 1.5 | 2.5

Y-axis: 4000, 3000, 2000, 1000, 0

X-axis (TIME): 1/89, 1/90, 7/90, 1/91, 7/91, 1/92, 1/93

MAPE and RMSPE return infinite error.

The modified statistics cannot distinguish between the magnitude of the error at any one point, however, they do return greater errors for longer tails

---
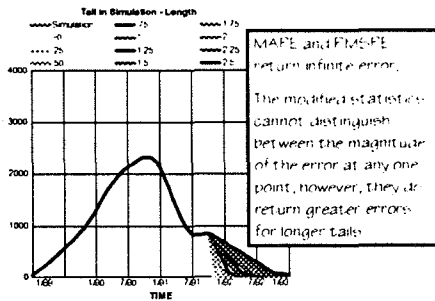
R² Ignores *Magnitude* of the Data

Multiple of Data

Legend: Data | 8 | 1 | 5 | 6 | 4 | 7 | 2

Y-axis: 4000, 3000, 2000, 1000, 0

X-axis (TIME): 1/89, 1/90, 1/90, 1/91, 7/91, 1/92, 7/92, 1/93

All of these simulations generate an R² of 1 measured against the data
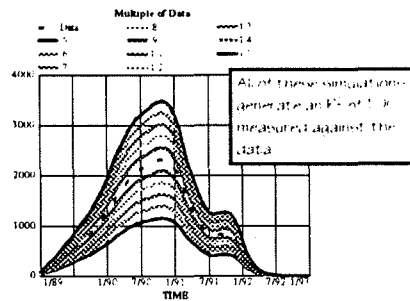
Most Statistics Do Not Produce Symmetrical Results for Multiples
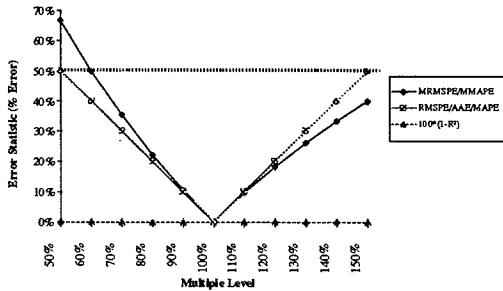
AAE Provides the Best Measure for Life Cycle Models

- Returns symmetric errors for bias
- Gives reasonable errors for increasing magnitude and duration of any "tails"
- Provides predictable response to increasing levels of noise
- Yields similar values to other statistics "the rest of the time"
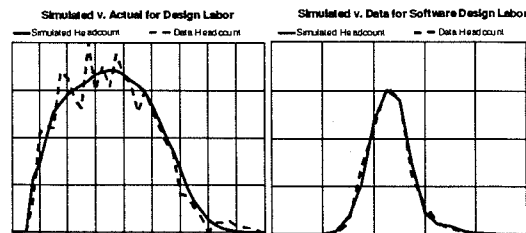- Is intuitive and easy to explain

All other measures suffer some fatal flaw
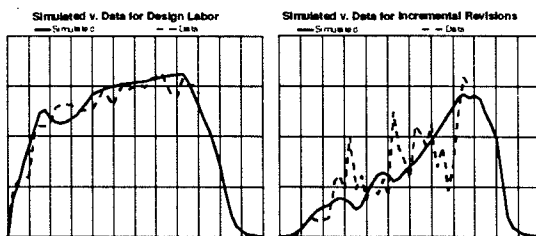
How Good a Fit is Necessary?

Standard will depend on ...
- The purpose for which the model is constructed
- The amount of noise in the data
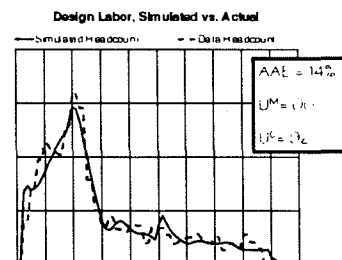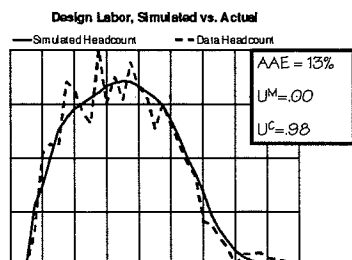- Time and budget available

Noise Levels Vary Between Projects
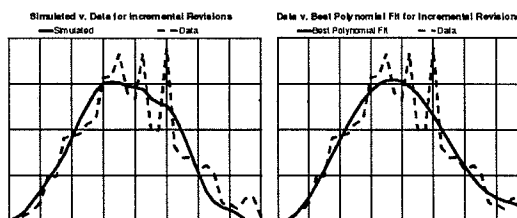

Noise Levels Vary Within a Project

Examples of Good Fit

Examples of Good Fit

Design Labor, Simulated vs. Actual

—Simulated Headcount   – – Data Headcount

AAE = 13%
U^M=.00
U^C=.98

Smoothed Data May Provide a More Fair Basis for Comparison

Simulated v. Data for Incremental Revisions

—Simulated   – –Data

Data v. Best Polynomial Fit for Incremental Revisions

—Best Polynomial Fit   – –Data

Summary

- Statistics cannot substitute for a careful visual inspection and analysis of the fit to historical data
- We should continue to compute and report all statistics to allow comparison to other results
- The "Average Absolute Error" statistic should be used as the standard
- What fit is "good enough" is inherently subjective.
- The Theil components should be computed and monitored to ensure that "bias" is small (<=10%) and most error is due to covariance (>=70%).

References

○ Barlas, Yaman. "Multiple Tests for Validation of System Dynamics Type of Simulation Models," European Journal of Operational Research 42 (1989).

● Forrester, Jay W. Industrial Dynamics. The MIT Press, Cambridge, Massachusetts, 1961.

● Mathews, Brian P. and Adamantios Diamantopoulos. Towards a Taxonomy of Forecast Error Measures: A Factor-comparative Investigation of Forecast Error Dimensions. Journal of Forecasting, Vol. 13, 409-416, 1994.

● Sterman, John, Nelson Repenning and Fred Kofman. Unanticipated Side Effects of Successful Quality Programs: Exploring a Paradox of Organizational Improvement. Management Science (forthcoming).

● Lyneis, James M. and Alexander L. Pugh III. "Hand vs. Automated Tuning: An Experimental Analysis," Proceedings of the 1996 International System Dynamics Conference, Cambridge, Massachusetts, July 1996.

● Sterman, John D. Appropriate Summary Statistics for Evaluating the Historical Fit of System Dynamics Models. Dynamica. 10(2) pp. 51-66, 1984.

● Sterman, John D., George P. Richardson and Pal Davidsen. Modeling the Estimation of Petroleum Resources in the United States Technological Forecasting and Social Change 33, 219-249, 1988

For a copy of the complete paper, contact the authors at:

Pugh-Roberts Associates
41 William Linskey Way
Cambridge, MA 02142

telephone: (617) 864-8880
facsimile: (617) 864-8884

428