Improving the Accuracy of Cost Estimating Relationships (CERs) for NSS Software Systems

Dr. David L. Wang
Dr. Austin Lee
Affiliated with The Aerospace Corporation

2014 ICEAA Professional Development & Training Workshop
10-13 June 2014

Agenda

Estimating the Uncertainty in Cost Estimating Relationships (CERs)

- Statement of the Problem
- Introduction
- Basic Concepts and Terminology Used in Parametric Modeling
- Empirical Analysis Results
- Statistical Characterization of Normalized ESLOC
- Proposed Solution
- Empirical Results from Applying Modified PI Equation
- S-Curve Generation
- Application to Cost Prediction
- Summary

Statement of the Problem

- Current Department of Defense acquisition policy guidance mandates funding at a set percentile of confidence level
  - The confidence level percentile estimate is typically derived from Cost Estimating Relationship (CERs), the CER prediction interval (PI), and associated S-curve
- Numerous studies by GAO and others have shown there is significant cost growth in many National Security Space (NSS) acquisition programs
  - The results from these studies suggest that the CERs and associated S-curves may be underestimating the true cost
- A more accurate and robust CER would allow decision-makers to be better informed on how much money is needed to fund a particular NSS acquisition program
- Our analysis results suggest the conventional Prediction Interval equation may be too optimistic
- We show in this presentation a practical method for improving the accuracy of the prediction interval estimate, thereby improving the accuracy of the resulting S-curve

Introduction

Software Permeates All Elements of National Security Space (NSS) Systems [Eslinger, 2010]

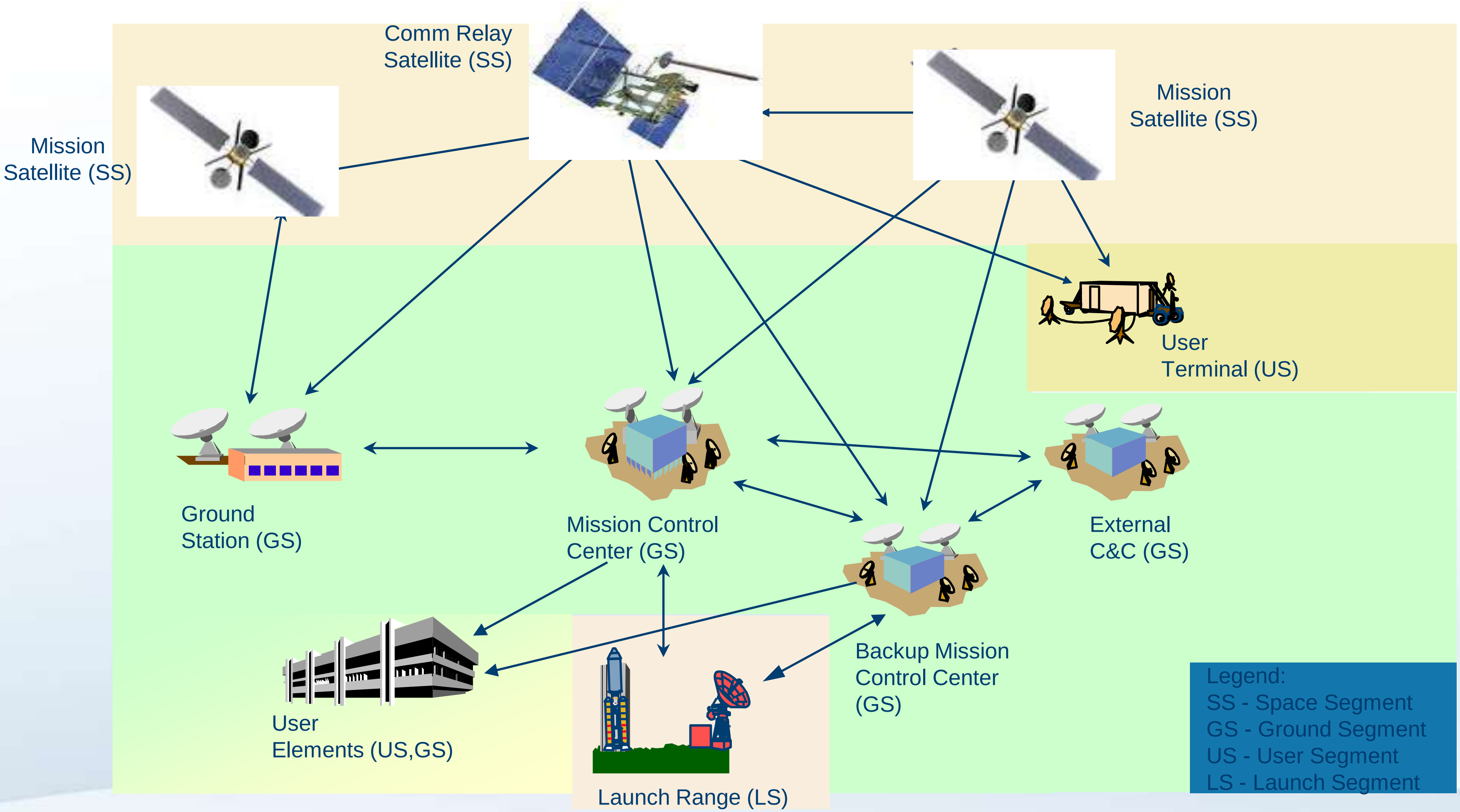


Figure 1

In order to develop useful and predictive CER for NSS systems, it is necessary to develop a predictive CER for NSS software systems

Basic Concepts and Terminology Used in Parametric Modeling

Introduction to CERs

- CERs express cost as a function of one or more independent cost drivers

  $Y = f(x, \beta)$;

  where

  x is a vector representing the cost driver variables

  β is a vector of coefficients to be estimated by the regression analysis of the sample cost data points

- Below are examples of the common parametric cost model equations for hardware or software systems:

  – *Linear:* $Y = \sum_{i}^{N} \beta_i \cdot x_i$

  – *Non-Linear:* $Y = A \cdot X^B$

  where A and B are constants derived from the regression

Linear CER Model

- Linear CER Model: $Y = A + B \cdot X$
- CER residual error, ε_i, is represented as additive errors: $Y = A + B \cdot x_i + \varepsilon_i$
- Problem: Find A and B such that the Sum of the Squared Error ($SSE = \sum_i^N \varepsilon_i^2$) is minimized

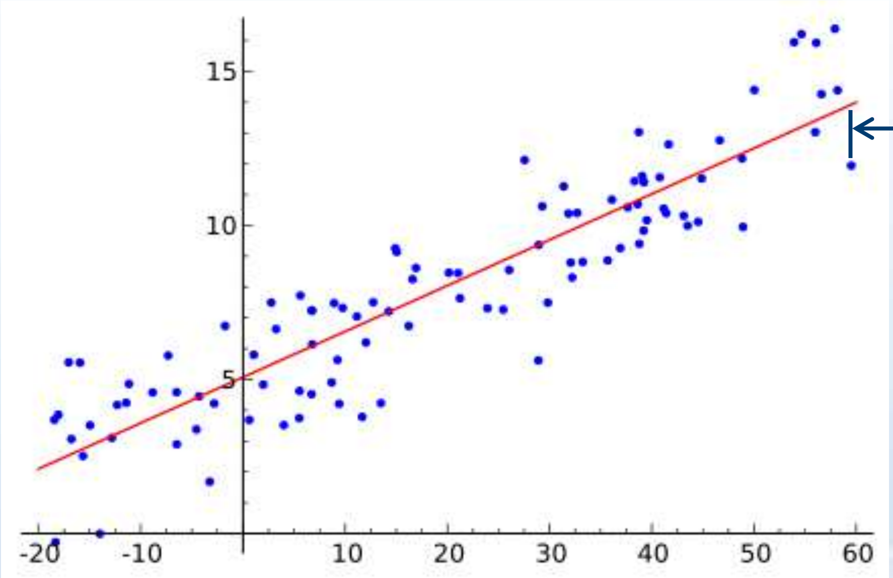


Figure 2 Linear CER Model

ε_i : delta between the i^{th} data point and the linear regression line

Linear CER assumption:
(1) An error around the regression line, ε, is distributed normally, and is symmetric; or
(2) The number of an observation, N, is sufficiently large so that Central Limit Theorem is applicable

Common Measures of CER Uncertainty

- The Standard Error of the Estimate (SEE) is the standard deviation of the cost estimates from a CER
  - SEE is <u>not</u> the CER regression error
- The Confidence Interval (CI) is expressed as $(1 - \alpha) \cdot 100\%$ confident that the true mean value is contained within the calculated range; where α is the probability that the population mean for a parameter lies outside of the CI; $(0 \leq \alpha \leq 1)$
  - e.g., An α of 0.20 represents a confidence level of 80% (i.e., there is 80% certainty that the true value of the mean lies within the CI)
- The Prediction Interval (PI) measures the range of uncertainty around the cost estimates from a CER

Prediction Interval Equation

For single variate linear CER

$$\hat{Y} \pm t_{\alpha/2,df} \times \text{SEE}\sqrt{\frac{n+1}{n} + \frac{(X-\bar{X})^2}{\sum X^2 - n\bar{X}^2}} \quad \text{(Eqn 1)}$$

Where

- $\hat{Y}$ is the CER prediction
- $t_{\alpha/2,df}$ is the upper α/2 cut-off point of the student's t distribution (for the simple linear regression, df = n-2)
- n is the number of observations
- SEE is the Standard Error of the Estimate
- X is the value of the independent variable used in calculating the estimate

Basic Parametric Software Cost Model Equation

- Basic Parametric Software Cost Model: $Cost = A \cdot ESLOC^B$ (Eqn 2)

  where

  Cost is the Development Effort in Person-months

  A is the proportionality constant calculated from the cost driver parameters

  ESLOC is the Equivalent Software Lines of Code which normalizes the amount of new, modified, and re-used code applied to calculate the effort to produce the total software product

  B is an exponent (depends on the specific software cost model used, but always > 1)

- Translate into linear CER by transforming into the natural log domain
  - Ln (Cost) = ln (A) + B * ln (ESLOC)

Empirical Analysis Results

Linear Regression Model

(Data Samples from NSS Software Systems)

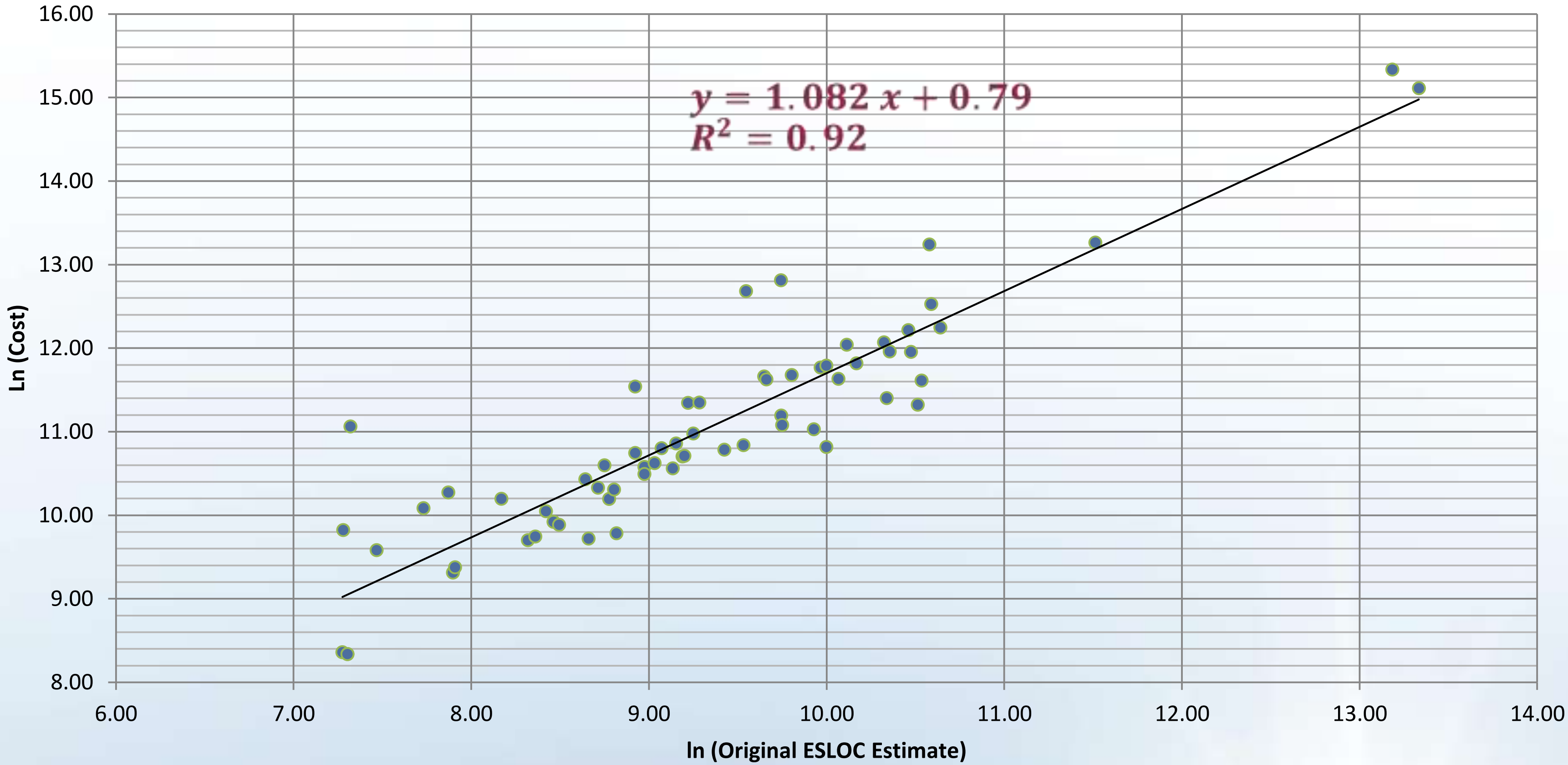


Figure 3

Applying 95% Confidence Bands from PI Equation to Historical Data

Data Collection Samples from NSS Software Systems

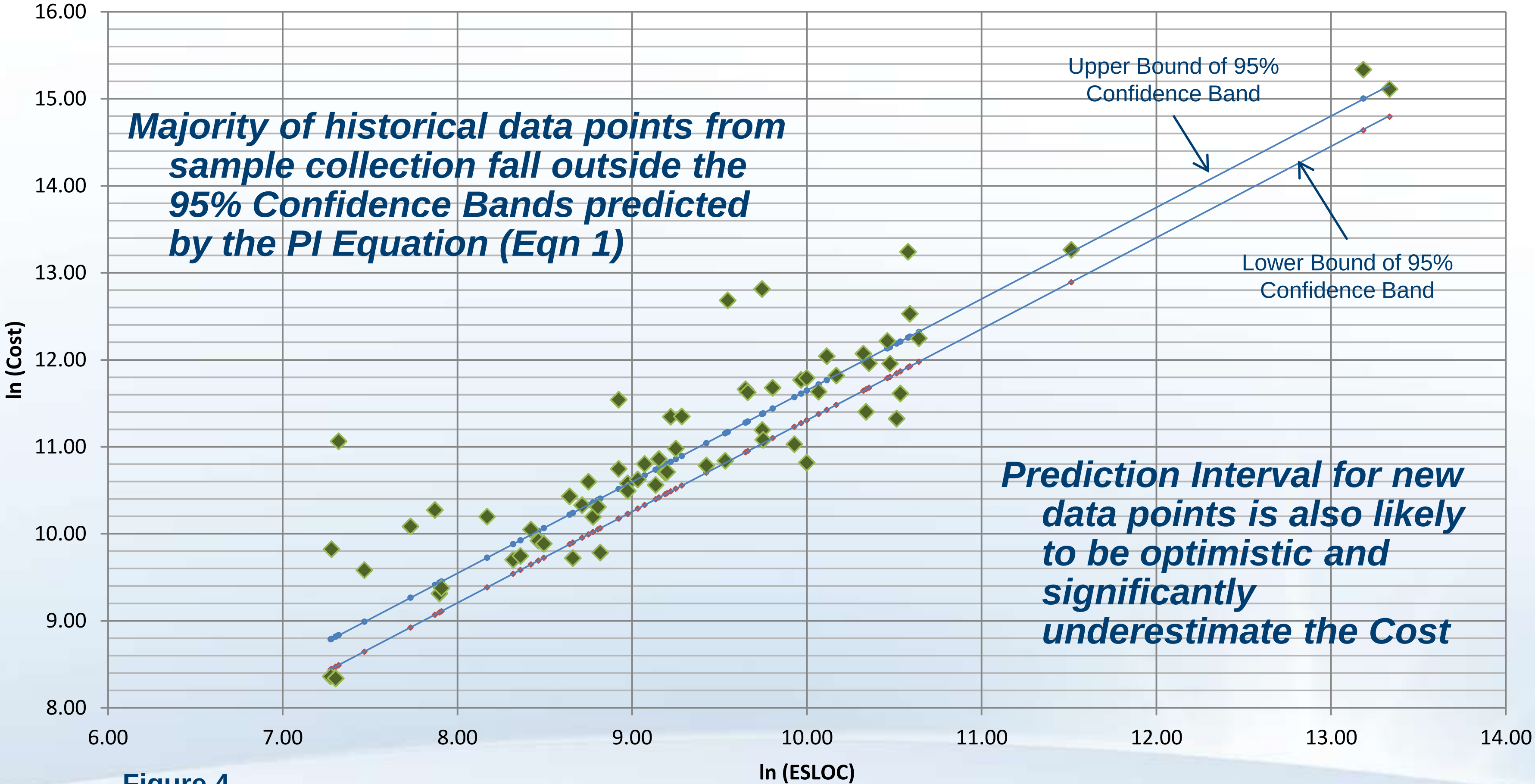


Figure 4

There are significant statistical variation in addition to the regression errors that the PI Equation is not accounting for

Observation

- There are significant statistical variations in addition to the regression errors that the PI Equation is modeling
  - *PI equation estimates the prediction interval based on the second-order statistics of the CER cost estimates,*
    - $Cost = A * ESLOC^{B} * \varepsilon$
      - *Where ε is the CER regression error*
      - *A and B are constants*
    - SEE (the standard error of the cost estimate) is a function of ε and ESLOC
  - *The independent driver variable (ESLOC) is typically assumed to have insignificant variations relative to the regression errors*
  - *If ESLOC varies significantly, then the SEE term in the PI Equation will significantly underestimate the true prediction interval*
- Question:
  - *How much does ESLOC vary?*

Empirical and historical data for ESLOC growth provides a definitive answer!!

Empirical Data on Uncertainties of ESLOC Estimates

High degree of uncertainty and variation in ESLOC estimates!

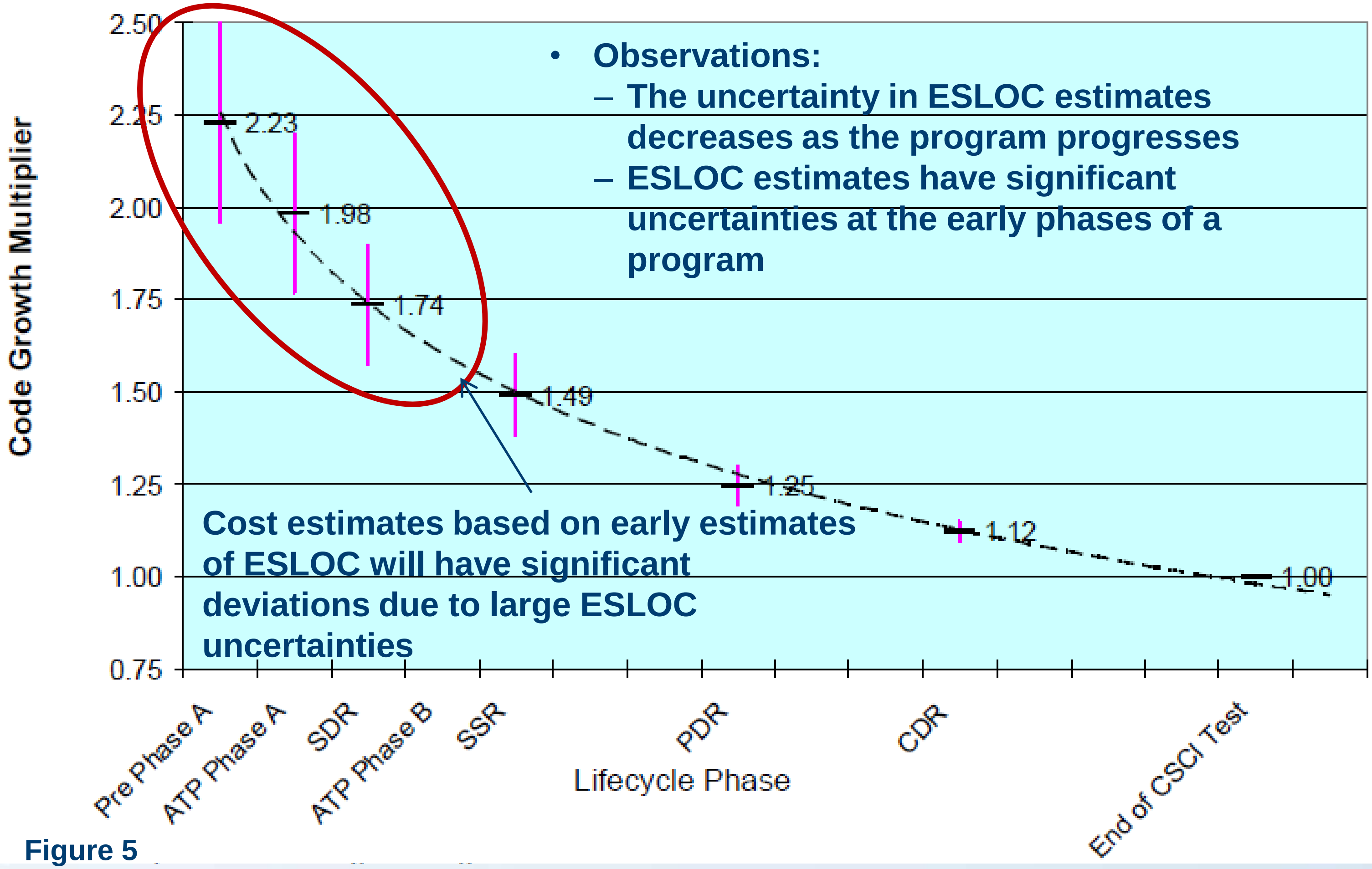


Figure 5

Based on Barry Holchin's code growth algorithm for medium-to-high complexity software from: [Holchin, 2003]

Historical Data on ESLOC Growth

Significant growth between initial ESLOC estimate and actual ESLOC

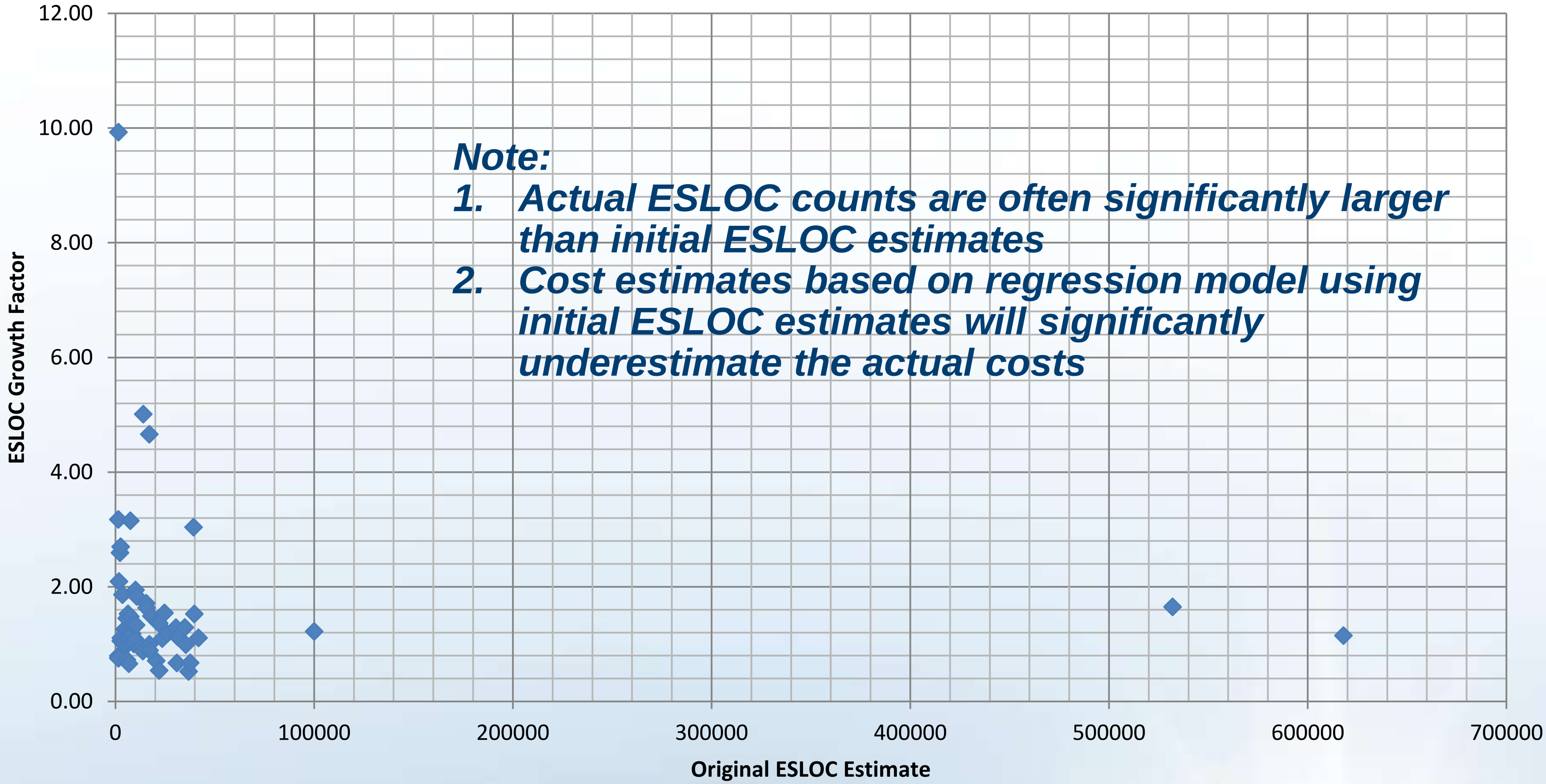


Figure 6

Improving the prediction interval estimate requires a better statistical characterization of ESLOC growth

Statistical Characterization of Normalized ESLOC

Basic Software Schedule and Software Cost Models

Cost is the development effort (Person-months)

T is the development Time or Duration (Months)

C is a proportionality Constant

D is an exponent

(where C and D depend on the specific software cost model used)

BASIC PARAMETRIC SOFTWARE SCHEDULE MODEL EQUATION

$$Cost = A \bullet ESLOC^{B}$$

Cost is the development effort (Person-months)

A is a proportionality constant calculated from cost driver parameters

$ESLOC$ is the Effective number of SLOC

B is an exponent (depends on specific software cost model used, but always >1)

BASIC PARAMETRIC SOFTWARE COST MODEL EQUATION

$$T = C_1 \bullet ESLOC^{B_1}$$

T is the development Time or Duration (Months)

C_1 is the proportionality Constant (i.e., $C \bullet A^{D}$)

B_1 is an exponent (i.e., $B \bullet D$)

Alternate Formulation of Parametric Software Schedule Model Equation

- ESLOC is the independent variable
- ESLOC is assumed to be a known value

Known Results from Prior Studies

- Schedule delays exhibit fat-tail behaviors [Wang, 2013], [Smart, 2013], [Wang, 2012]
  - Schedule delays extreme statistics can be approximated by Extreme Value distribution or Log Normal distribution
- Cost growths exhibit fat-tail behaviors [Smart, 2013], [Smart, 2011]
  - Cost growth extreme statistics can be approximated by Log Normal distribution

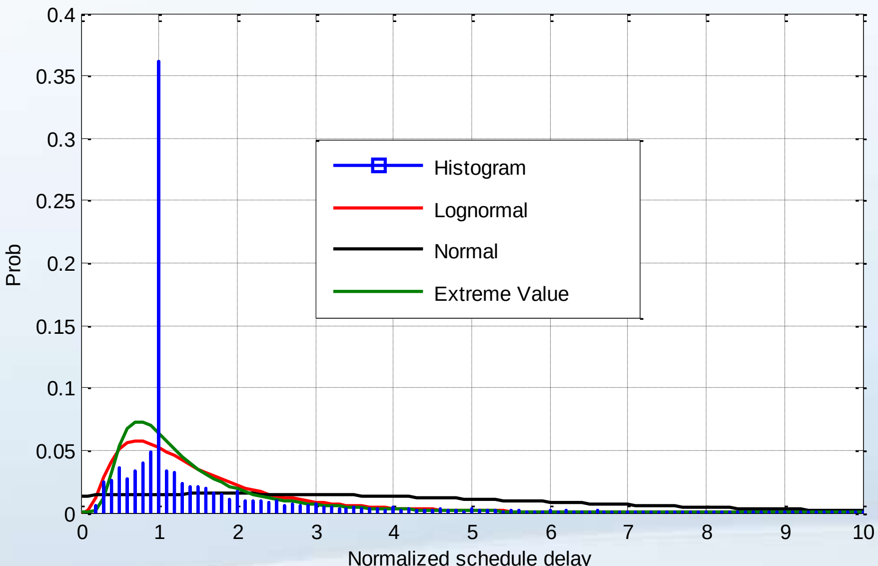
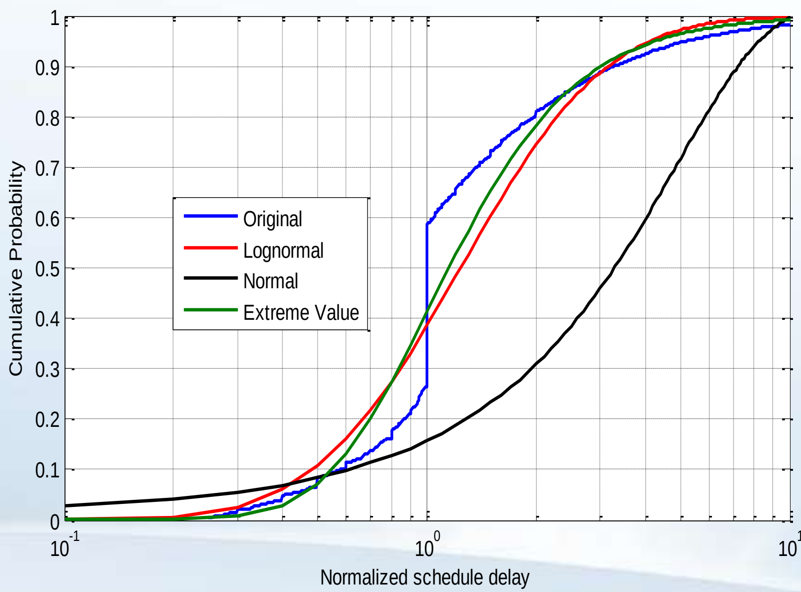


Figure 7



Fundamental Theorem

[Papoulis] "Probability, Random Variables, and Stochastic Processes"

- For $Y = A * X^B$, (where A and B are real number, and B > 1)

If Y is (Extreme Value , or Log Normal, or Normal) distributed, then X is also (Extreme Value , or Log Normal, or Normal) distributed

If X is (Extreme Value , or Log Normal, or Normal) distributed, then Y is also (Extreme Value , or Log Normal, or Normal) distributed

Numerical Simulation Results

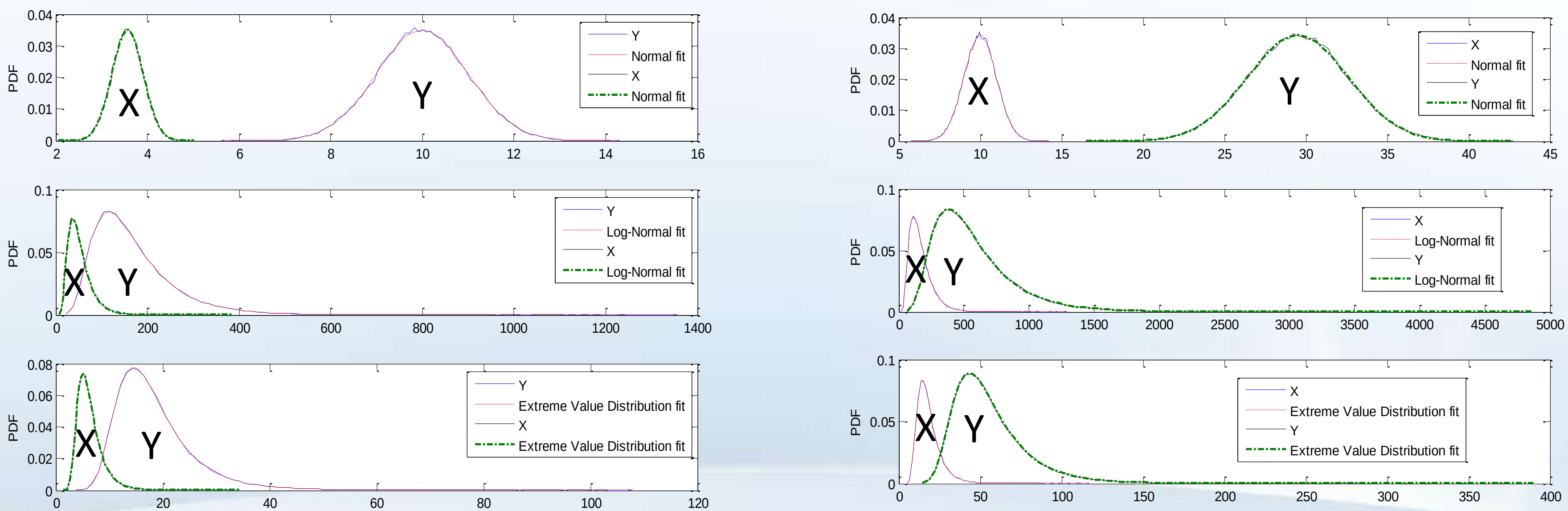
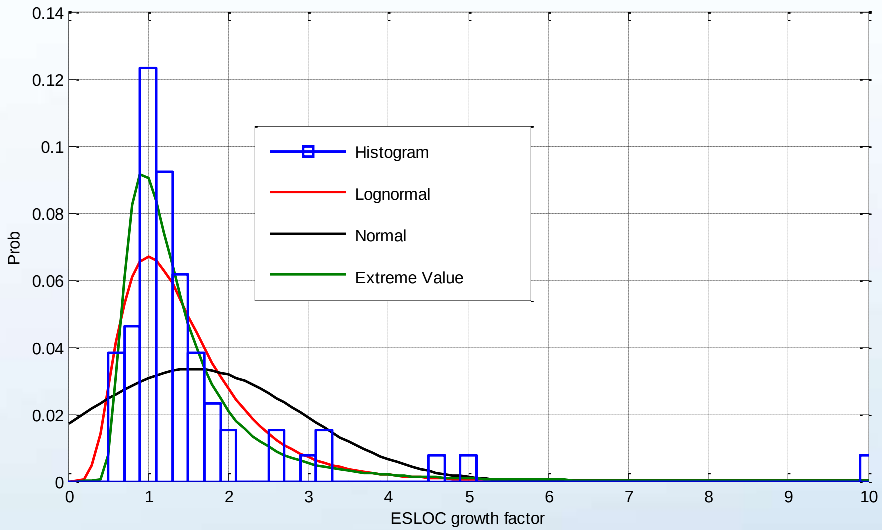


Figure 8

Conclusion: Normalized ESLOC is characterized by Extreme Value distribution or Log Normal distribution

Empirical Data on Normalized ESLOC Statistics

Empirical Data confirmed that Normalized ESLOC statistics are approximated by Extreme Value distribution or Log Normal distribution



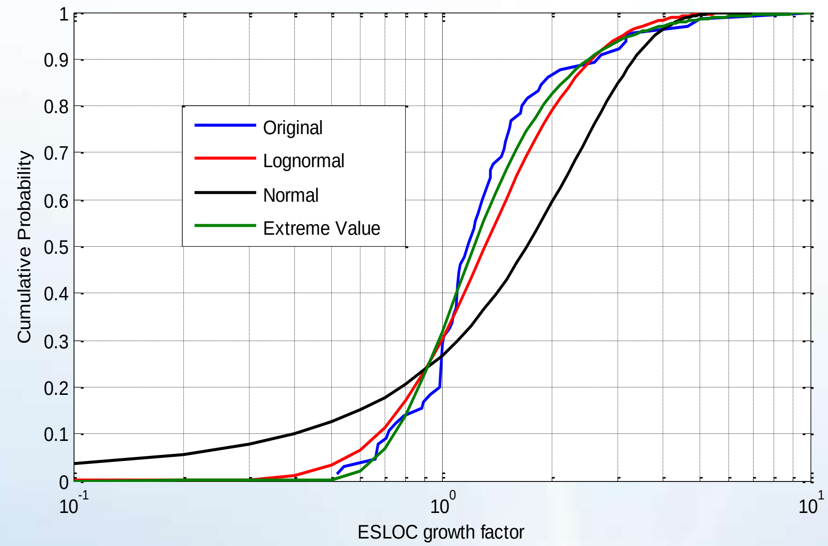


Figure 9

Large Variability of ESLOC violates key assumption in PI Equation

Effect of Normalized ESLOC Statistical Analysis Results

- The PI equation (Eqn 1) will significantly underestimate the prediction interval range for a given α, and thus overestimate the confidence level of a cost estimate or schedule estimate, because:
  - Empirical and historical data show clearly that the key assumption of a regression model's SEE is not applicable for NSS software systems
    - Normalized ESLOC (i.e., ESLOC Growth) can be approximated by fat-tail distributions (e.g., Extreme Value distribution or Log Normal distribution)
    - the variation of Normalized ESLOC is significantly larger relative to the regression error ε
- Adjustment to the Prediction Interval equation is needed to account for the large variability of Normalized ESLOC

The Proposed Solution

Proposed Adjustment to the PI Equation

- Recall from PI Equation (Eqn. 1)

$$\hat{Y} \pm t_{\alpha/2,df} \times \text{SEE} \sqrt{\frac{n+1}{n} + \frac{(X-\bar{X})^2}{\sum X^2 - n\bar{X}^2}}$$

  - *Where*
    - X is the independent cost driver, i.e., ESLOC
    - SEE is a function of X_α and α: SEE(X_α, α)
    - X_α is the value of X that corresponds to a confidence of $(1 - \alpha/2)$
      - *As the $(1 - \alpha/2)$ percentile increases, the corresponding value for SEE(X, α) will also increase*
  - *As the value for SEE(X_α, α) increases, the prediction interval will increase accordingly*

Empirical Results from Applying Modified PI Equation

Modified PI Equation Confidence Level Bands

(in natural log space)

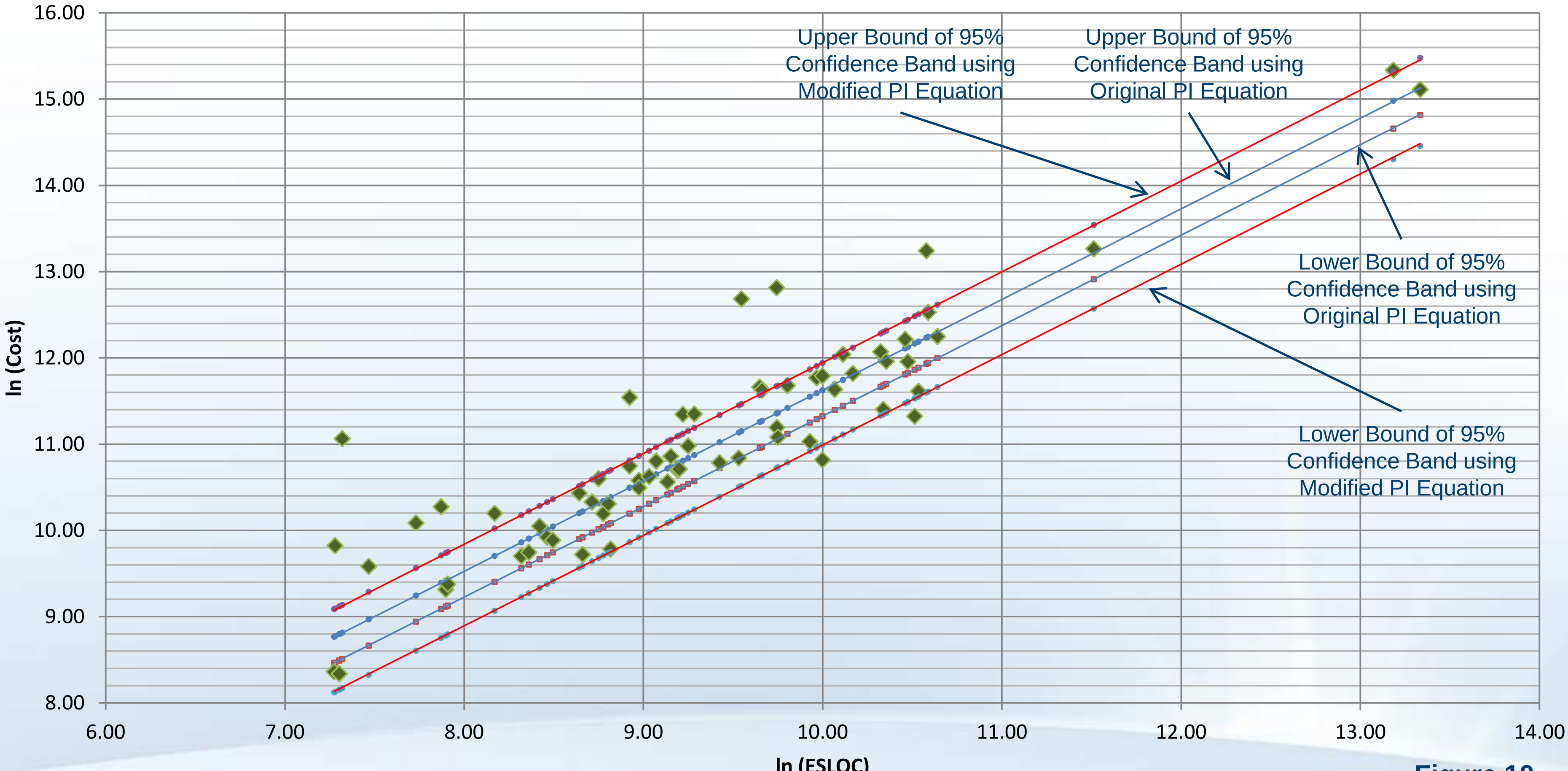


Figure 10

Empirical results show clearly that the modified PI Equation Confidence Level Bands are more accurate than the original PI Equation's

Modified PI Equation Confidence Level Bands

(in ESLOC, Cost space)

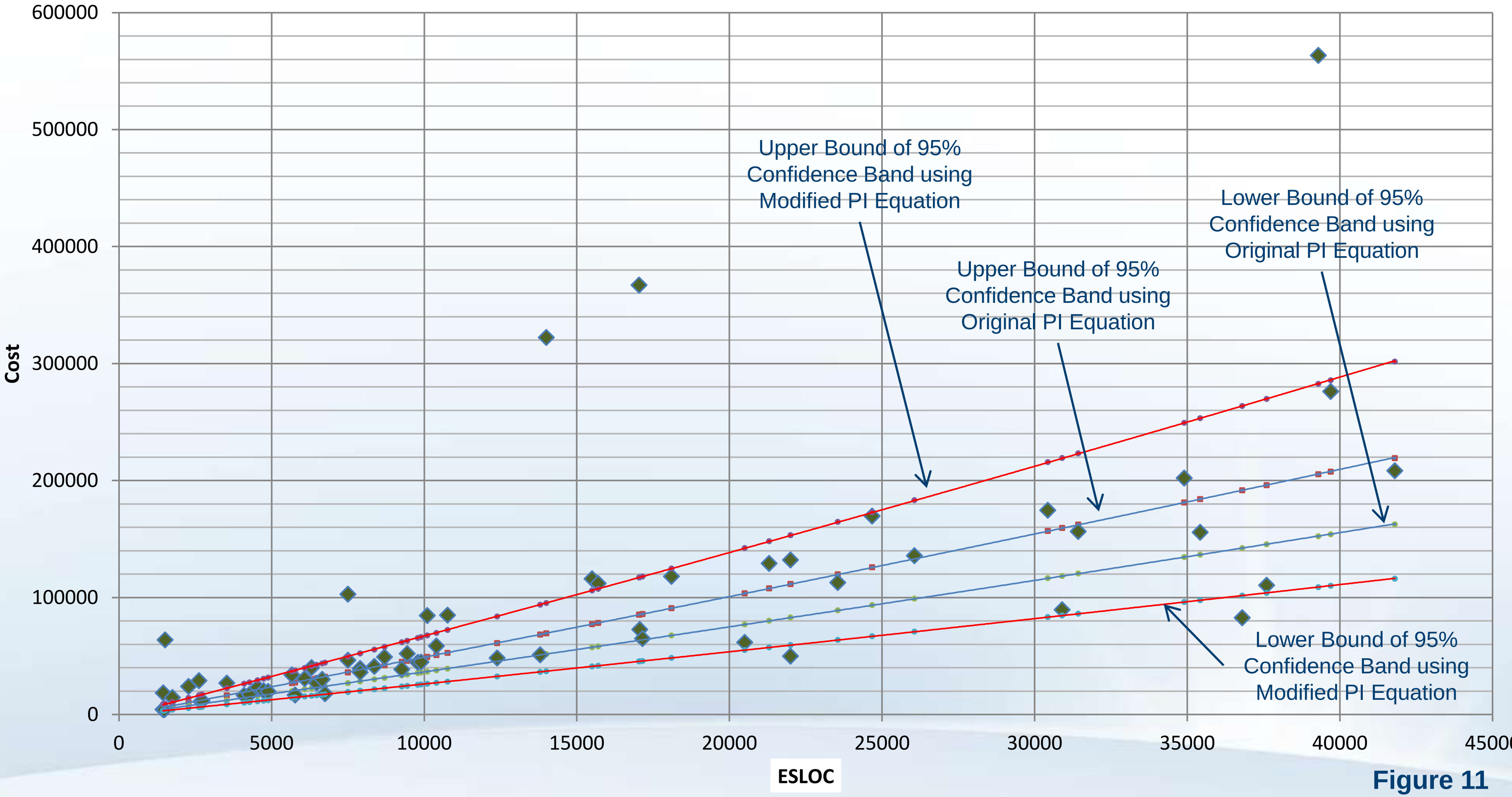


Figure 11

Empirical results show clearly that the modified PI Equation Confidence Level Bands are more accurate than the original PI Equation's

S-Curve Generation

Generating an S-Curve from a Set of PI Curves

Notional Example

The upper PI bound corresponds to (1 - α/2) percentile on the cumulative distribution

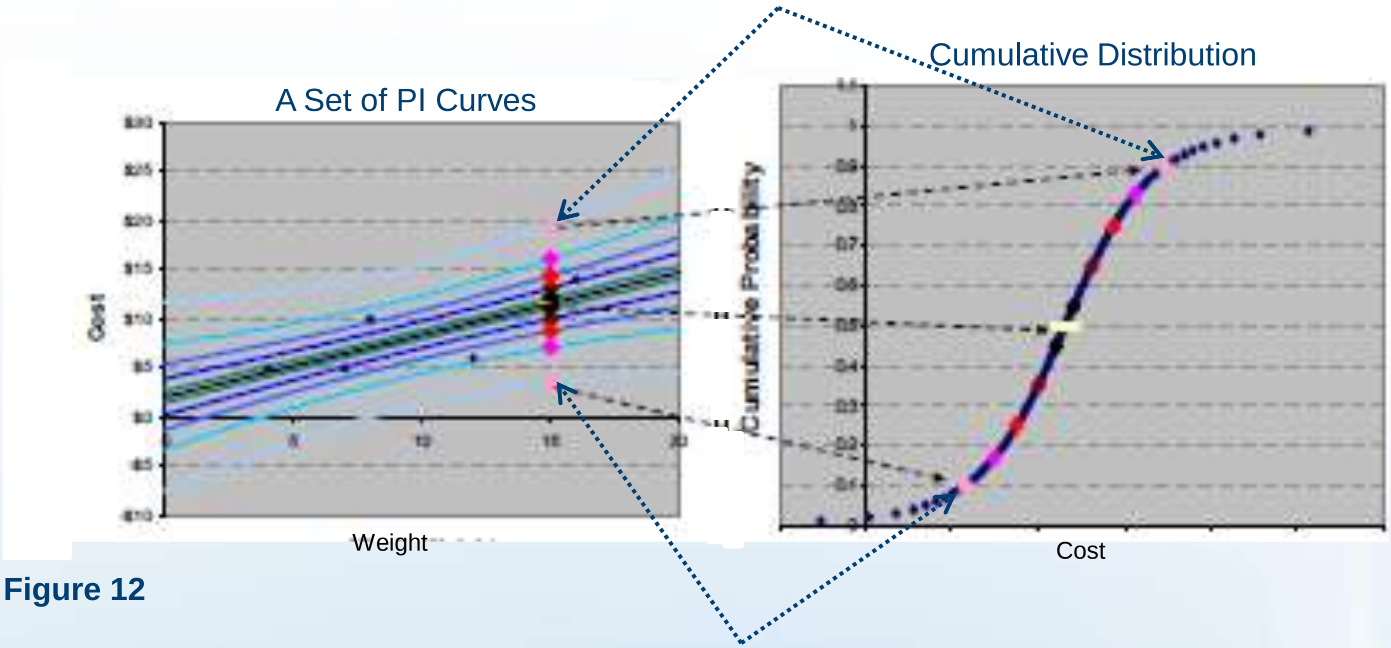


Figure 12

The lower PI bound corresponds to α/2 percentile on the cumulative distribution

- Generating an S-Curve by varying α from 0 to 1

Note: the S-Curve will overestimate the cumulative probability, if the Prediction Interval is underestimating the true variation of CER prediction

Empirical S-Curves vs S-Curves from CER Predictions

Notional Example

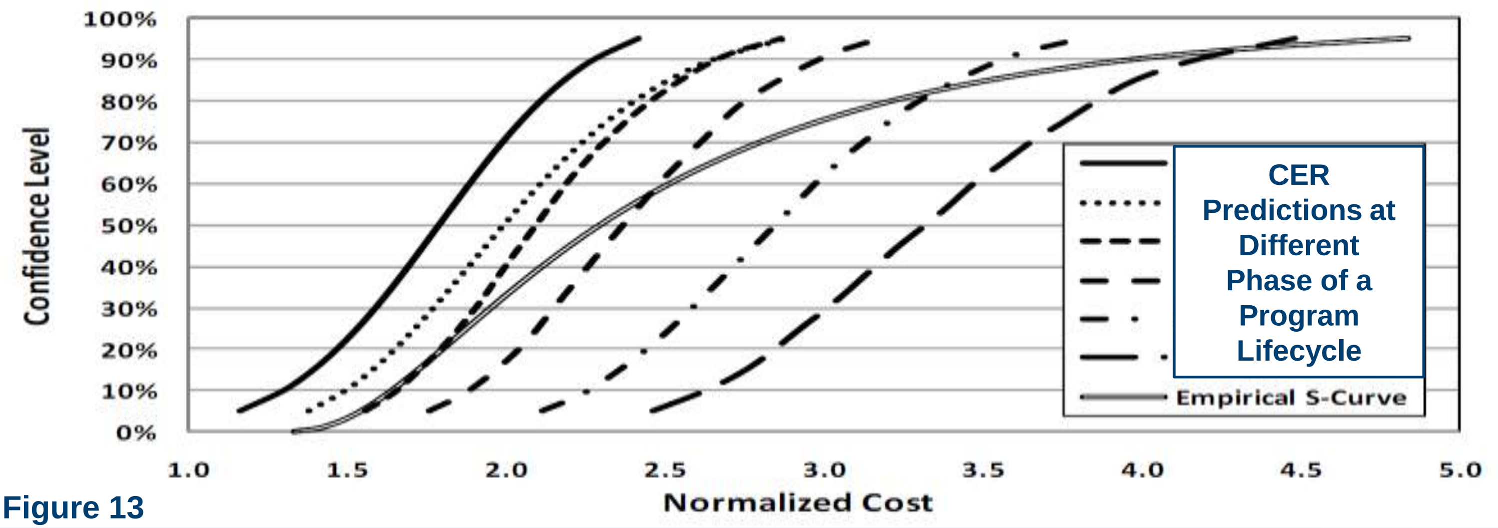


Figure 13

- S-curves from the CER predictions shift to the right as the program progresses
- Empirical S-curves show a more accurate description of the actual cost behaviors

Conjecture: If we improve the accuracy of the Prediction Interval, then the resulting S-curve should better approximate the actual cost behavior

S-Curves for NSS Software Systems

The modified PI equation results in S-Curves that approximate the actual cost behavior

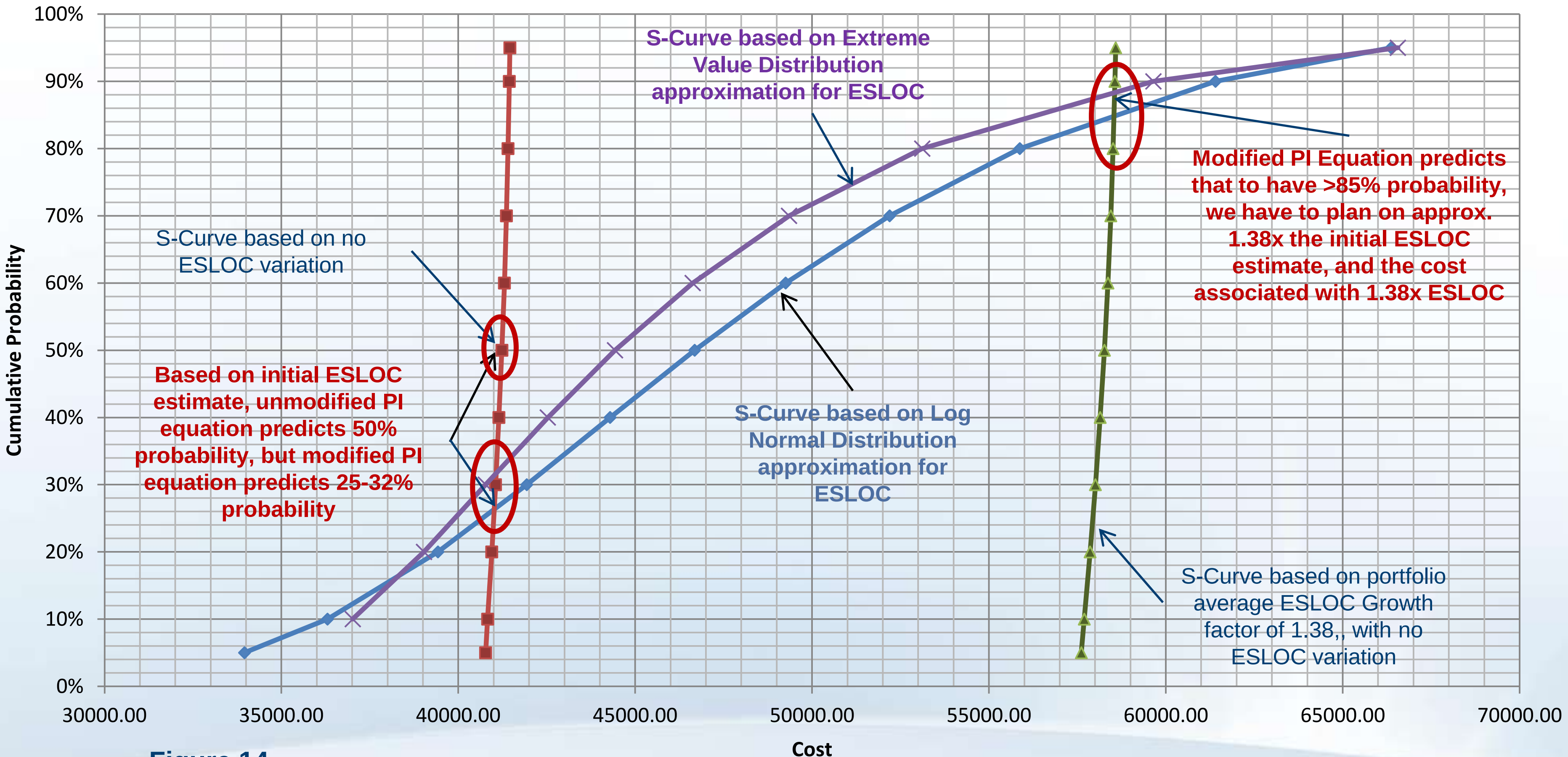


Figure 14

Most recent actual program experience confirms the prediction from the modified PI equation

Application to Cost Prediction

Cost Prediction Example

- S-Curves based on modified PI Equation predicts there will be 38% ESLOC Growth on average for a Cumulative Probability of 85-90%
  - Actual program data: 39% ESLOC Growth
- Apply regression model derived from historical data, CER (with modified PI eqn) predicts a 43% Cost Growth
  - Detailed SEER-SEM model with 39% ESLOC Growth predicts a 47% Cost Growth

The Modified PI Equation produces a more realistic forecast of ESLOC Growth and Cost Growth than unmodified PI Equation

Summary

- In this presentation, we presented analytical analysis as well as empirical data that the existing well-known PI equation consistently underestimates the prediction interval.
  - This underestimation of the prediction interval results in an inflated S-curve confidence level.
- We presented results that show the cause of the PI equation underestimating the prediction interval.
- We presented a proposed modification to the PI equation to account for the variability of the independent cost driver, ESLOC.
- We applied the proposed modification to the PI equation, and showed that the prediction of the modified PI equation is more accurate.
- We generated S-Curves based on the modified PI equation.
  - Our S-Curves better approximate the S-Curve derived from empirical data.
  - Our S-Curve prediction was confirmed by actual program experience.
- Cost Prediction based on our modified PI equation is a close approximation of the Cost Prediction using a detailed SEER-SEM model.

References

- Wang, D., “Estimating the Uncertainty in Cost Estimating Relationship (CER),” 25th Annual International Integrated Program Management Conference November 2013
- Smart, C., “Advanced Cost Risks,” ICEAA 2013 Annual Professional Development & Training Workshop
- Wang, D., “Improving Realism of Cost and Schedule Risk Analysis,” SCEA/ISPA Joint Annual Conference & Training Workshop, June 2012
- Papoulis, A., “Probability, Random Variables, and Stochastic Processes,” p126
- Boehm, et al., “US DoD Application Domain Empirical Software Cost Analysis,” IEEE Computer Society, 2011
- Smart, C., “Covered With Oil: Incorporating Realism In Cost Risk Analysis,” June 2011
- Eslinger, S. “The Dynamics of Software Project Management,” 2010 PMAG Symposium
- Book, S.A., “Cost S-Curves Through Project Phases,” NASA Independent Project Assessment Office technical report, September 2007
- Gayek, J., et al., “Software Cost and Productivity Model,” Aerospace Report No. ATR-2004(8311)-1, 2004
- Holchin, Barry, et al. “Code Growth Algorithm,” Tecolote Inc, 2003

Backup