An Improved Method for Predicting Software Code Growth:

Tecolote DSLOC Estimate Growth Model

Michael A. Ross
Technical Expert
Tecolote Research, Inc.
3601 Aviation Blvd., Suite 1600
Manhattan Beach, CA 90266-3756
310-536-0011
mross@tecolote.com
http://www.tecolote.com

[1,2]***Abstract***— This paper describes a model and methodology developed for Tecolote Research, Inc. by this paper's author and referred to as the Tecolote DSLOC Estimate Growth Model v06. The model provides probabilistic growth adjustment to single-point Technical Baseline Estimates (TBEs) of Delivered Source Lines of Code (DSLOC), for both New software and Pre-Existing Reused (PER) software, that is sensitive to the *maturity* of the estimate; i.e., when, in the Software Development Life Cycle (SDLC), the DSLOC TBE is performed. The model is based on Software Resources Data Report (SRDR) data collected by Dr. Wilson Rosa of the U.S. Air Force Cost Analysis Agency (AFCAA). This model provides an alternative to other software code growth methodologies such as Mr. Barry Holchin's (2003) code growth matrix.

TABLE OF CONTENTS

1. INTRODUCTION

The Tecolote DSLOC Estimate Growth Model v06, developed for Tecolote Research, Inc. by the author of this paper, provides probabilistic growth adjustment to single-point Technical Baseline Estimates (TBEs) of Delivered Source Lines of Code (DSLOC), for both New software and Pre-Existing Reused (PER) software, that are sensitive to the *maturity* of the estimate; i.e., when, in

[2] 1st Edition, Version 05; April 18, 2011

the Software Development Life Cycle (SDLC), the DSLOC TBE is performed. It is a *data driven* model and methodology that is based on Software Resources Data Report (SRDR) data collected by Dr. Wilson Rosa of the U.S. Air Force Cost Analysis Agency (AFCAA). This model provides an alternative to other software code growth methodologies such as Mr. Barry Holchin's (2003) code growth matrix.

The paper includes custom Cumulative Distribution Function (CDF) tables that can be copied into tools such as Tecolote's ACEIT or Oracle's Crystal Ball in order to create the Custom CDFs that are needed to model the baseline New DSLOC growth factor distribution and to model the baseline Pre-Existing Reused DSLOC growth factor distribution. The paper also includes a set of DSLOC growth factor multipliers as a function of Estimate Maturity for each of New DSLOC and Pre-Existing DSLOC such that appropriate application of these factors to a DSLOC TBE yields corresponding Least, Likely, and Most DSLOC values that, if input to Galorath's SEER-SEM, will reasonably model growth and uncertainty consistent with the SRDR historical data.

2. MODEL SUMMARY

The Tecolote DSLOC Estimate Growth Model v06 equations for applying growth and uncertainty to TBE New and PER DSLOC are[3]

$$\boldsymbol{S_{D\_Adj\_New}} \equiv S_{D\_New}\left(e^{-bt}\left(\boldsymbol{K_{GF\_New}}-1\right)+1\right) \tag{1}$$

and

$$\boldsymbol{S_{D\_Adj\_PER}} \equiv S_{D\_PER}\left(e^{-bt}\left(\boldsymbol{K_{GF\_PER}}-1\right)+1\right) \tag{2}$$

where

$\boldsymbol{S_{D\_Adj\_New}} \equiv$ growth-adjusted New DSLOC estimate distribution

$\boldsymbol{S_{D\_Adj\_PER}} \equiv$ growth-adjusted PER DSLOC estimate distribution

$S_{D\_New} \equiv$ Technical Baseline Estimate (TBE) of New DSLOC

$S_{D\_PER} \equiv$ Technical Baseline Estimate (TBE) of PER DSLOC

$\boldsymbol{K_{GF\_New}} \equiv$ baseline (Estimate Maturity = 0%) New DSLOC growth factor distribution (provided custom CDF in Table 3)

$\boldsymbol{K_{GF\_PER}} \equiv$ baseline (Estimate Maturity = 0%) PER DSLOC growth factor distribution (provided custom CDF in Table 3)

$b \equiv$ decay constant; default is 3.466 based on Boehm's "Cone of Uncertainty" (Boehm, 1981, p. 311)

$t \equiv$ Estimate Maturity Parameter: (SDLCBegin = 0%; SyRR = 20%; SwRR = 40%; SwPDR = 60%; SwCDR = 80%; SwAccept = 100%)

[3] We use the ***Arial bold italic*** font to denote random variables; i.e., variables that can take on values according to some probability distribution.

The equations for providing the appropriate New and PER $\langle Least, Likely, Most \rangle$ DSLOC inputs to Galorath's SEER-SEM tool are

Growth-Adjusted New DSLOC	Growth-Adjusted PER DSLOC
$S_{D\_Adj\_New\_Least} = S_{D\_New}\left(-0.828071e^{-3.466t}+1\right)$	$S_{D\_Adj\_PER\_Least} = S_{D\_PER}\left(-0.687191e^{-3.466t}+1\right)$
$S_{D\_Adj\_New\_Likely} = S_{D\_New}\left(-0.828071e^{-3.466t}+1\right)$	$S_{D\_Adj\_PER\_Likely} = S_{D\_PER}\left(-0.687192e^{-3.466t}+1\right)$
$S_{D\_Adj\_New\_Most} = S_{D\_New}\left(5.366128e^{-3.466t}+1\right)$	$S_{D\_Adj\_PER\_Most} = S_{D\_PER}\left(3.658219e^{-3.466t}+1\right)$

(3)

The remainder of this paper describes the basis of these equations.

3. COMPONENTS OF THE MODEL

Normalized Estimate Maturity

The single parameter input to the Tecolote DSLOC Estimate Growth Model is normalized Estimate Maturity t. By default, Estimate Maturity is quantified by the scale contained in Table 1 below. This scale is consistent with the model defaults for the baseline New and Pre-Existing DSLOC growth factor distributions (which is based on the SRDR data) and with the uncertainty decay factor (which is based on Boehm's (1981) *Cone of Uncertainty*. Tailored instances of the model can be created for different SDLCs as long as historical data exists where the projects followed that particular SDLC and where this data has been used to determine corresponding baseline growth factor distributions and uncertainty decay factor values or distributions.

Table 1 *Default Normalized Estimate Maturity Scale*

Estimate Maturity Scale	
t = 0%	Begin SDLC
t = 20%	System Requirements Review
t = 40%	System Design Review / Software Requirements Review
t = 60%	Software Preliminary Design Review
t = 80%	Software Critical Design Review
t = 100%	Software Acceptance

DSLOC Baseline Growth Factor Distributions

DSLOC estimate growth is modeled at the computer program (CSCI) level and is applied by multiplying the TBEs of New and Pre-Existing DSLOC by the appropriate decay-adjusted growth factor distribution. The baseline (zero Estimate Maturity) growth factor distributions for New DSLOC and for Pre-Existing DSLOC have the following characteristics (Table 2) and custom CDFs (Table 3).

Table 2: *SRDR Data Set Distribution Statistics*

ACE DSLOC Baseline Growth Factor Distribution Statistics				
New DSLOC Growth Factor			**Pre-Existing DSLOC Growth Factor**	
Number of Data Points (N)	56		**Number of Data Points (N)**	45
Data Set Mean (m)	1.75		**Data Set Mean (m)**	1.43
CDF Mean (m')	1.75		**CDF Mean (m')**	1.42
%ile @ Data Set Mean (P(m))	69%		**%ile @ Data Set Mean (P(m))**	71%
%ile @ CDF Mean (P(m'))	69%		**%ile @ CDF Mean (P(m'))**	71%
%ile @ Point (P(pt))	29%		**%ile @ Point (P(pt))**	29%
Data Set Median m[~]	1.20		**Data Set Median m[~]**	1.04
CDF Median m'[~] *Define a baseline growth factor distribution in ACE by using this value as the "Equation / Throughput" field entry with a custom CDF containing corresponding median-normalized growth factor values.*	**1.204296**		**CDF Median m'[~]** *Define a baseline growth factor distribution in ACE by using this value as the "Equation / Throughput" field entry with a custom CDF containing corresponding median-normalized growth factor values.*	**1.037044**
Data Set Std Dev s	1.33		**Data Set Std Dev s**	0.91
CDF Std Dev s'	1.32		**CDF Std Dev s'**	0.90
Data Set CV (C[V])	0.76		**Data Set CV c[V]**	0.64
CDF CV (C'[V])	0.75		**CDF CV (C'[V])**	0.63

Table 3: *DSLOC Estimate Growth Factor Distribution CDFs*[4]

ACE DSLOC Baseline Growth Factor Distribution CDFs						
Copy red columns into ACE Custom CDF Dialog Box				*Copy red columns into ACE Custom CDF Dialog Box*		
New DSLOC Growth Factor CDF				**Pre-Existing DSLOC Growth Factor CDF**		
%ile	**Raw Growth Factor**	**Median-Normalized Growth Factor**		**%ile**	**Raw Growth Factor**	**Median-Normalized Growth Factor**
0.0	0.547902	**0.4549560272208**		**0.0**	0.655131	**0.6317293787416**
10.0	0.676993	**0.5621483902387**		**10.0**	0.725186	**0.6992822451771**
20.0	0.968243	**0.8039911843758**		**20.0**	0.947745	**0.9138907707378**
30.0	1.001516	**0.8316194196262**		**30.0**	1.000010	**0.9642887324668**
40.0	1.061531	**0.8814541263447**		**40.0**	1.000096	**0.9643717324103**
50.0	1.204296	**1.0000000000000**		**50.0**	1.037044	**1.0000000000000**
60.0	1.403391	**1.1653207912851**		**60.0**	1.118300	**1.0783540487449**
70.0	1.791218	**1.4873573359220**		**70.0**	1.394266	**1.3444623081028**
80.0	2.516756	**2.0898160858878**		**80.0**	1.775599	**1.7121742117209**
90.0	3.710696	**3.0812166786418**		**90.0**	2.571689	**2.4798271957032**
100.0	6.253957	**5.1930414674842**		**100.0**	5.265691	**5.0775979934902**

The default DSLOC baseline growth factor distribution statistics and CDF tables shown above are developed from historical data reported in SRDRs and collected by Dr. Wilson Rosa at AFCAA. This data is filtered first by eliminating all data points where the New or PER growth factor is zero or undefined (i.e., the estimated value cannot be zero and the final actual value cannot be zero):

$$Candidate_i = Est\_New_i \neq 0 \;\wedge\; Est\_PER_i \neq 0 \;\wedge\; Act\_New_i \neq 0 \;\wedge\; Act\_PER_i \neq 0 \qquad \textbf{(4)}$$

The resulting filtered data is then filtered again to eliminate all data points that are outside above and below two multiplicative standard deviations of the filtered data set mean:

[4] A 1001-element version of each CDF can be obtained by e-mail request to mross@tecolote.com.

$$Candidate\_New_i =$$

$$Act\_New_i / Est\_New_i = K_{GF\_New_i} \in \left(\left(\%SEE_{GF\_New} + 1 \right)^{-2} \bar{K}_{GF\_New}, \left(\%SEE_{GF\_New} + 1 \right)^{2} \bar{K}_{GF\_New} \right)$$

where

$$\%SEE_{GF\_New} \equiv \sqrt{ \frac{1}{(N-1)} \sum_{i=1}^{N} \left(\frac{\left(K_{GF\_New\_i} - \bar{K}_{GF\_New} \right)}{\bar{K}_{GF\_New}} \right)^2 }$$

Within two multiplicative standard deviations of the data set mean

$$Candidate\_PER_i =$$

$$Act\_PER_i / Est\_PER_i = K_{GF\_PER_i} \in \left(\left(\%SEE_{GF\_PER} + 1 \right)^{-2} \bar{K}_{GF\_PER}, \left(\%SEE_{GF\_PER} + 1 \right)^{2} \bar{K}_{GF\_PER} \right)$$

where

$$\%SEE_{GF\_PER} \equiv \sqrt{ \frac{1}{(N-1)} \sum_{i=1}^{N} \left(\frac{\left(K_{GF\_PER\_i} - \bar{K}_{GF\_PER} \right)}{\bar{K}_{GF\_PER}} \right)^2 }$$

DSLOC Estimate Uncertainty Decay

Decrease (decay) of the uncertainty implied by DSLOC estimate growth factor distributions as a project progresses from start to finish is modeled by the general form

$$\boldsymbol{K_{GF\_Adj}} = e^{-bt} \left(\boldsymbol{K_{GF}} - 1 \right) + 1 \qquad \textbf{(5)}$$

where

- t ≡ normalized Estimate Maturity (percentage of the development process ***duration*** at which the estimate is performed); $t_{start} \equiv t_0 \equiv 0\%$ and $t_{finish} \equiv 100\%$
- b ≡ decay parameter; by default is set to a value of 3.466 which emulates the decay behavior of Boehm's "Cone of Uncertainty".[5]
- $\boldsymbol{K_{GF}}$ ≡ growth factor distribution at time t_0
- $\boldsymbol{K_{GF\_Adj}}$ ≡ decay-adjusted growth factor distribution at Estimate Maturity t

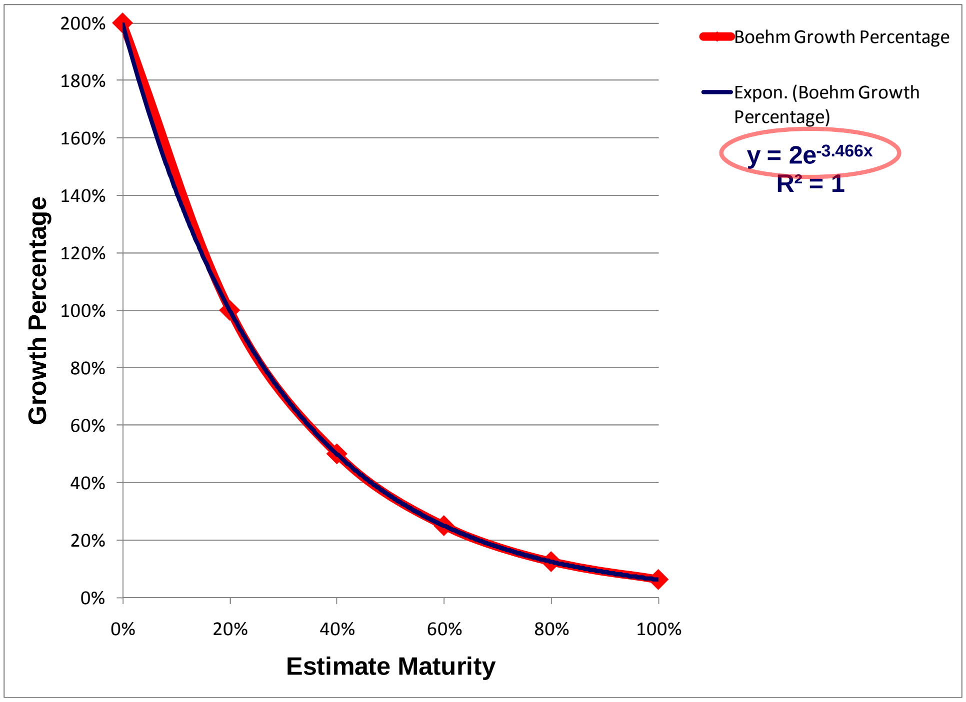
The practical effect of applying this model is time-progressive compression of the DSLOC estimate distribution about the TBE position approaching no uncertainty at process completion.

In order to render Equation (5) useful in a particular estimating situation, we need to assume some value (or distribution) for the uncertainty decay function proportionality constant b. Two methods for accomplishing this are 1) perform a regression analysis of relevant historical data to determine an expected value or distribution for b and 2) assume uncertainty decay consistent with Dr. Barry Boehm's (1981 pp. 310-311) *Cone of Uncertainty*. The latter is assumed to be the model default and can be accomplished by assuming $b = 3.466$ (see Figure 1 below) and by as-

[5] Note that the model only uses the rate of uncertainty decay implied by Boehm's "Cone of Uncertainty". The model does not use Boehm's growth factors but instead uses growth factors derived from the SRDR data.

suming time t to be normalized according to the SDLC Estimate Maturity scale in Table 1 above.

Figure 1: *Boehm "Cone of Uncertainty" – Top Half*



Decay-Adjusted DSLOC Growth Factor Distributions

Uncertainty Decay

We assume some normalized uncertainty scale factor function $\boldsymbol{K}_U$ of time t where $\boldsymbol{K}_U(t) \in [0,1]$, where $\boldsymbol{K}_U(t \mid t=0) = 1$ represents maximum (full scale) uncertainty, and hypothesize $\boldsymbol{K}_U(t)$ decreases (decays) at a rate proportional to its value (i.e., uncertainty tends to decay faster during the early stages of a process when experience is low and tends to decay slower during the later stages of a process when experience is high). We model this hypothetical behavior mathematically as

$$\frac{d\,\boldsymbol{K}_U(t)}{dt} \propto -\boldsymbol{K}_U(t) \quad \therefore \frac{d\,\boldsymbol{K}_U(t)}{dt} = -b\,\boldsymbol{K}_U(t) \tag{6}$$

where b is the constant of proportionality. Solving the ordinary differential Equation (6) yields

$$\frac{d\,\boldsymbol{K}_U(t)}{\boldsymbol{K}_U(t)} = -bdt \quad \rightarrow \quad \int \frac{d\,\boldsymbol{K}_U(t)}{\boldsymbol{K}_U(t)} = \int -bdt \quad \rightarrow \quad \boldsymbol{ln}\left(\boldsymbol{K}_U(t)\right) = -bt + c$$
$$\therefore \boldsymbol{K}_U(t) = e^{-bt} e^{c} \tag{7}$$

Since we have already posited the constraint $\boldsymbol{K}_U\left(t \mid t=0\right)=1$ we can solve Equation (7) for the constant of integration c

$$\boldsymbol{K}_U\left(0\right)=e^{-b(0)}e^c=1 \quad \rightarrow \quad e^c=1 \quad \therefore c=0 \tag{8}$$

Substituting the equivalent of c in Equation (8) for c in Equation (7) yields

$$\boldsymbol{K}_U\left(t\right)=e^{-bt}e^{(0)} \quad \therefore \boldsymbol{K}_U\left(t\right)=e^{-bt} \tag{9}$$

Applying Uncertainty Decay to Growth Factor Distributions

Suppose we have a baseline DSLOC estimate growth factor distribution $\boldsymbol{K}_{GF}$, which has been developed from historical data and which models the amount of uncertainty that exists about the TBE of DSLOC assuming that this estimate is done at the beginning of a software development process; i.e., Estimate Maturity is zero, consistent with the processes from which the historical data was collected. Suppose this baseline distribution is represented as a CDF; i.e., a mapping of growth factor values to percentiles. We would like to model what happens to the uncertainty modeled by this baseline distribution as activities in the process progress to completion. We have already hypothesized that uncertainty decays over time and have developed a model for this decay in Equation (9). Since the function $\boldsymbol{K}_U\left(t\right)$ in Equation (9) is normalized (i.e., yields uncertainty factors that are percentages of full scale), we can scale our baseline DSLOC estimate growth factor distribution by the transformation

$$\boldsymbol{K}_{GF\_Adj}=\boldsymbol{K}_U\left(t\right)\left(\boldsymbol{K}_{GF}-1\right)+1 \tag{10}$$

where

$\boldsymbol{K}_{GF}$ ≡ baseline growth factor distribution at $t=0$ (0% Estimate Maturity) which is given as a custom CDF (see Table 3)

$\boldsymbol{K}_{GF\_Adj}$ ≡ decay-adjusted growth factor distribution at some Estimate Maturity t

This transformation effectively scales the percentage differences between the growth factors in the baseline growth factor distribution and no growth (a growth factor of 1).

Substituting the value of $\boldsymbol{K}_U\left(t\right)$ in Equation (9) for $\boldsymbol{K}_U\left(t\right)$ in Equation (10) yields

$$\boldsymbol{K}_{GF\_Adj}=e^{-bt}\left(\boldsymbol{K}_{GF}-1\right)+1 \tag{11}$$

As stated earlier, in order to render Equation (11) useful in a particular estimating situation, we need to assume some value (or distribution) for the uncertainty decay function proportionality constant b; either by assuming $b=3.466$ (Boehm's "Cone of Uncertainty") or by analyzing relevant historical data to model decay as a single value b or as a distribution $\boldsymbol{B}$.

Figure 2 and Figure 3 below illustrate the behavior of Equation (11) with decay constant $b=3.466$ over the range of possible Estimate Maturity values $t\in\left[0,1\right]$.

Figure 2: *New DSLOC Growth Factor Decay*

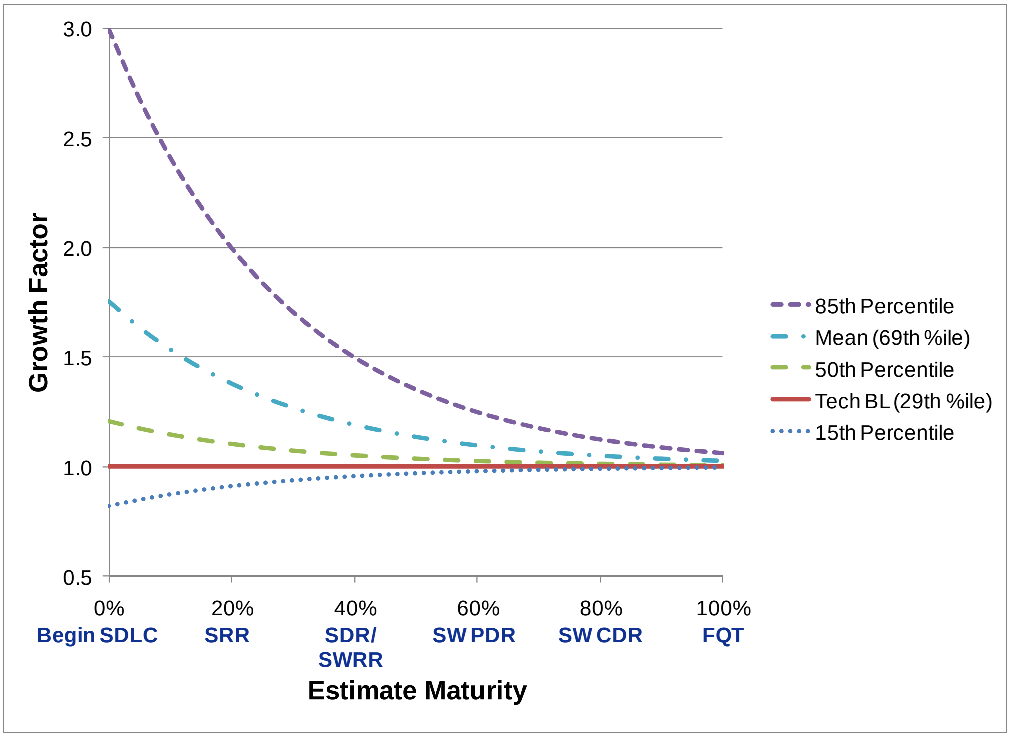
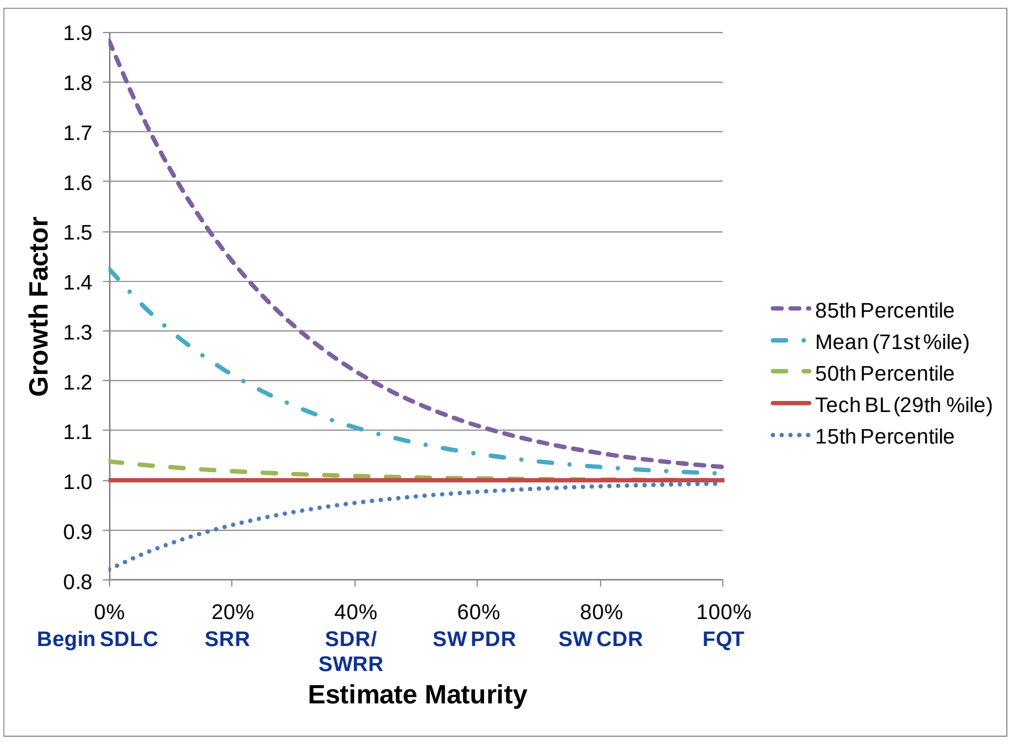


Figure 3: *Pre-Existing Reuse DSLOC Growth Factor Decay*



Applying Growth Factor Distributions to TBEs of New and PER DSLOC

We can now transform single-point TBEs of New S_{D_New} and PER S_{D_PER} DSLOC into growth-adjusted distributions of New $\boldsymbol{S_{D\_Adj\_New}}$ and PER $\boldsymbol{S_{D\_Adj\_PER}}$ DSLOC by simply scaling the appropriate instantiation of Equation (11) above (a distribution) by the corresponding single-point TBE:

$$\boldsymbol{S_{D\_Adj\_New}} \equiv S_{D\_New}\left(e^{-bt}\left(\boldsymbol{K_{GF\_New}}-1\right)+1\right) \quad \textbf{(12)}$$

and

$$\boldsymbol{S_{D\_Adj\_PER}} \equiv S_{D\_PER}\left(e^{-bt}\left(\boldsymbol{K_{GF\_PER}}-1\right)+1\right) \quad \textbf{(13)}$$

Figure 4 and Figure 5 below illustrate the behaviors of the growth-adjusted New DSLOC estimate distribution (Equation (12)) and the growth-adjusted PER DSLOC estimate distribution (Equation (13)) for given New and PER TBEs and a given Estimate Maturity.

Figure 4: *Example Growth-Adjusted New DSLOC Distribution vs. Estimate Maturity*

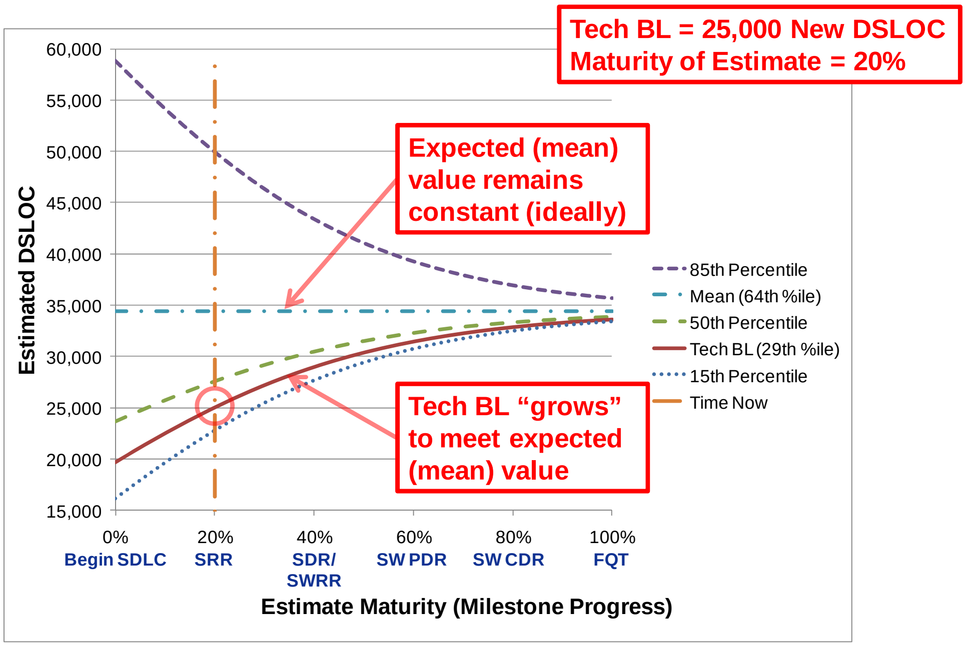
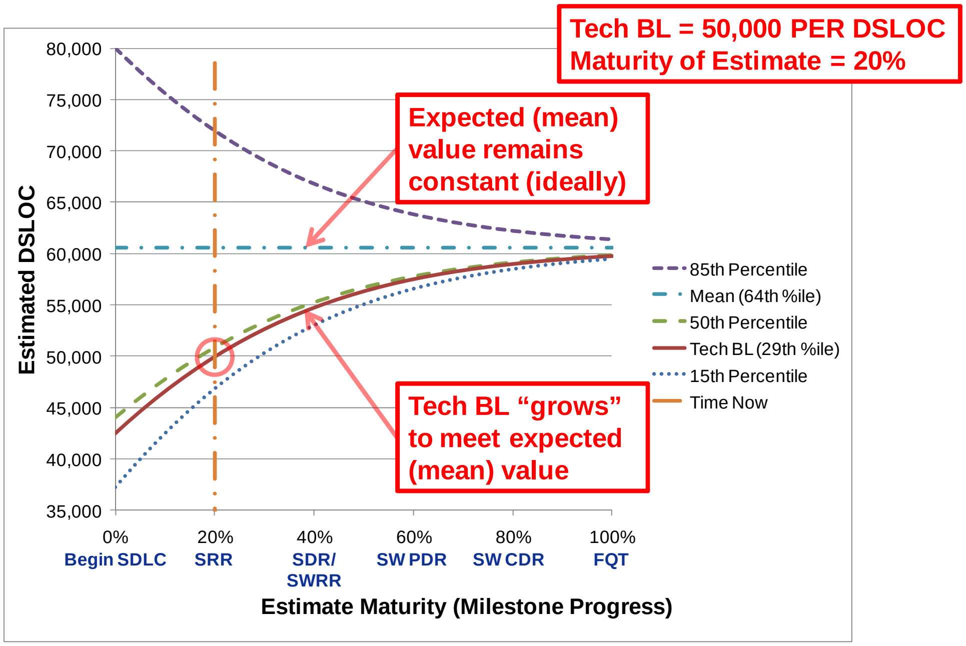


Figure 5: *Example Growth-Adjusted PER DSLOC Distribution vs. Estimate Maturity*



4. Modeling DSLOC Growth in Tecolote's ACEIT Tool

The process for using the Tecolote DSLOC Estimate Growth Model within a Tecolote ACEIT model for each of a particular set of computer programs (CSCIs) is (see Figure 6 below):

- Define a variable for each CSCI for each of New and PER that will represent the particular CSCI's New and PER DSLOC baseline growth factor distributions; e.g., SI010101_New_BL_GF and SI010101_PER_BL_GF. These will represent the random variables (distributions) K_{GF_New} and K_{GF_PER} in Equations (1) and (2) respectively.

  These variables must be described as distributions using ACEIT's custom CDF feature. The model default position CDFs are shown in Table 3 above. Note that when using ACEIT's custom CDF feature, it is best to normalize the growth factor values about the median growth factor value (right-most (red) columns of Table 3 above) and set the point estimate to the median (50th percentile) growth factor value in order to see reasonable point estimate values and percentages that are calculated from the CDF.

- Define a variable for each CSCI for each of New and PER; e.g., SI010101_New_Sd_Est_Mat and SI010101_PER_Sd_Est_Mat for each of New and PER; that will represent the Estimate Maturity variable t in Equations (1) and (2). For example, if the current TBE of New DSLOC for SI010101 was performed at successful completion of a System Requirements Review (System Requirements Analysis is complete) then the variable SI010101_New_Sd_Est_Mat would, from Table 1, be entered as 0.2 (20%).

- Define a new variable for each CSCI for each of New and PER; e.g., SI010101_New_GF_Decay and SI010101_PER_GF_Decay; that will represent the decay parameter variable b in Equations (1) and (2). Note that these variables could alternatively be described as a random variables (distributions) ***B*** using ACE's custom CDF feature based on some program-specific historical data. The model default is a constant value for b of 3.466.

- Define a variable for each CSCI for each of New and PER; e.g., SI010101_New_Adj_GUF and SI010101_PER_Adj_GUF; that will represent the uncertainty-decay-adjusted version of the New DSLOC and PER DSLOC growth factor distributions for that CSCI. The equation field for each of these variables implements Equation (11); e.g., exp(–SI010101_New_GF_Decay * SI010101_New_Sd_Est_Mat) * (SI010101_New_BL_GF – 1) + 1.
- If the decay constant is being described as a random variable ***B*** (distribution) then, because each decay constant random variable is inversely related to its corresponding growth factor random variable, as can be seen in Equation (11), we would need to negatively correlate each growth factor / decay constant pair in order for the convolution of these two variables to work properly in ACEIT[6]. For example, we would group SI010101_New_GF_Decay and SI010101_New_BL_GF and call the group SI010101_Growth_Decay_Group. We would then set the Group Strength of SI010101_New_GF_Decay to "–1" and set the Group Strength of SI010101_New_BL_GF to "D". Note that none of this step is necessary if using the model defaults based on SRDR data and assuming the decay to be constant with a value of 3.466.

Figure 6: *Example ACEIT Model Application of New DSLOC Estimate Growth*

-	New Growth-Adjusted DSLOC	SI010101_New_Adj_Sd	SI010101_New_Adj_GUF * SI010101_New_Sd
-	Technical Baseline DSLOC Point Estimate	SI010101_New_Sd	25000 [Given]
-	Maturity at DSLOC Estimate	SI010101_New_Sd_Est_Mat	0.20 [Sys Req Rev Complete = 20% Estimate Maturity]
-	Baseline Growth Factor	SI010101_New_BL_GF	1.204296 [Tecolote DSLOC Estimate Growth Model v06 Median of SRDR New DSLOC Data Set]
-	Decay Constant	SI010101_New_GF_Decay	3.466 [Tecolote DSLOC Estimate Growth Model v06 Default]
-	Adjusted Growth Factor	SI010101_New_Adj_GUF	exp(-SI010101_New_GF_Decay * SI010101_New_Sd_Est_Mat) * (SI010101_New_BL_GF - 1) + 1 [Tecolote DSLOC Estimate Growth Model v06]

5. MODELING DSLOC GROWTH IN GALORATH'S SEER-SEM TOOL

SEER PERT Distribution

The Galorath, Inc. SEER family of estimating tools incorporate a rather unique probability distribution to model input uncertainty. I refer to this distribution here as a SEER PERT distribution; Program Evaluation and Review Technique (PERT) because it borrows from the mathemat-

[6] Note that e^{-bt} is equivalent to $1/e^{bt}$.

ics that the PERT methodology uses to relate elicited expert opinion about the input parameters Least, Likely, and Most. The SEER PERT distribution combines the left half of one Normal (Gaussian) distribution with the right half of another Normal distribution. Because Normal distribution Probability Density Functions (PDFs) are range symmetrical, the mean of the left (low side) half-distribution μ_L is always equal to the mean of the right (high side) half-distribution μ_H. However, the low side half-distribution standard deviation σ_L need not equal the high-side half-distribution standard deviation σ_H. When $\sigma_H > \sigma_L$ the overall SEER PERT distribution is skewed to the right, when $\sigma_L > \sigma_H$ the distribution is skewed to the left, and when $\sigma_L = \sigma_H$ the distribution is symmetrical and classically Normal. Because the two halves of the distribution are Normal, because each of their means is always equal, and because the mean of a Normal distribution is always its median (50th percentile) value, it follows that each half-distribution contains half of the probability density of the overall SEER PERT distribution. Therefore, μ_L and μ_H are always equal to the overall SEER PERT distribution's median value $\tilde{\mu}$. Note, however, that $\tilde{\mu}$ is not necessarily equal to the overall SEER PERT distribution's mean μ; this is only true when $\sigma_L = \sigma_H$.

The PDF of the SEER PERT distribution can thus be described as

$$\boldsymbol{f}_{\boldsymbol{SEER}}\left(x \mid \tilde{\mu}, \sigma_L{}^2, \sigma_H{}^2\right) = \begin{cases} \dfrac{1}{\sqrt{2\pi\sigma_L{}^2}} e^{\frac{-(x-\tilde{\mu})^2}{2\sigma_L{}^2}} & x \le \tilde{\mu} \\ \dfrac{1}{\sqrt{2\pi\sigma_H{}^2}} e^{\frac{-(x-\tilde{\mu})^2}{2\sigma_H{}^2}} & x \ge \tilde{\mu} \end{cases} \qquad \textbf{(14)}$$

the CDF can be described as

$$\boldsymbol{F}_{\boldsymbol{SEER}}\left(x \mid \tilde{\mu}, \sigma_L{}^2, \sigma_H{}^2\right) = \begin{cases} \dfrac{1}{2}\left(1 + \boldsymbol{erf}\left(\dfrac{x-\tilde{\mu}}{\sqrt{2\sigma_L{}^2}}\right)\right) & x \le \tilde{\mu} \\ \dfrac{1}{2}\left(1 + \boldsymbol{erf}\left(\dfrac{x-\tilde{\mu}}{\sqrt{2\sigma_H{}^2}}\right)\right) & x \ge \tilde{\mu} \end{cases} \qquad \textbf{(15)}$$

and the inverse CDF (quantile function) can be described as

$$\boldsymbol{F}^{-1}_{\boldsymbol{SEER}}\left(p \mid \tilde{\mu}, \sigma_L{}^2, \sigma_H{}^2\right) = \begin{cases} \tilde{\mu} + \sqrt{2\sigma_L{}^2}\,\boldsymbol{erf}^{-1}\left(2p-1\right) & p <= 0.5 \\ \tilde{\mu} + \sqrt{2\sigma_H{}^2}\,\boldsymbol{erf}^{-1}\left(2p-1\right) & p >= 0.5 \end{cases} \qquad \textbf{(16)}$$

$$p \in (0,1)$$

The SEER PERT distribution borrows from PERT methodology in how it relates the distribution parameters $\tilde{\mu}$, σ_L, and σ_H to expert opinion elicitations of estimated values that describe a CSCI's DSLOC range; these values being referred to as Least L, Likely M, and Most H. The resulting relationships are

$$\tilde{\mu} = \frac{(L + 4M + H)}{6} \tag{17}$$

and

$$\sigma_L = \frac{\tilde{\mu} - L}{3} \tag{18}$$

and

$$\sigma_H = \frac{H - \tilde{\mu}}{3} \tag{19}$$

It is important to note here that the PERT relationship in Equation (17) constrains the amount of skew that can be modeled by the SEER PERT distribution. Maximum right (high side) skew occurs when $M = L$ and maximum left (low-side) skew occurs when $M = H$. Examples of the SEER PERT PDF can be seen in Figure 7 and Figure 8 below.

Least, Likely, Most Multipliers

SEER-SEM requires that the uncertainty about a DSLOC estimate be characterized as a Least, Likely, Most triple. Since SEER-SEM provides no facility for specifying DSLOC growth and growth uncertainty decay, DSLOC inputs to SEER-SEM must already be growth and uncertainty adjusted. Therefore, in order to model DSLOC growth in SEER-SEM according to the Tecolote DSLOC Estimate Growth Model, DSLOC Least, Likely, and Most values must be chosen that cause SEER-SEM's SEER PERT distribution to match, as closely as possible, the distributions described in Table 2 and Table 3 and adjusted for uncertainty decay as a function of Estimate Maturity.

Growth-adjusted Least L_{Adj}, Likely M_{Adj}, and Most H_{Adj} DSLOC inputs to SEER-SEM can be calculated for each of New and PER as functions of the given New and PER DSLOC TBEs S_{D_New} and S_{D_PER} with given Estimate Maturity t. For each of New and PER we define a set of three DSLOC estimate growth multipliers K_{L_Adj}, K_{M_Adj}, and K_{H_Adj} using Equation (11):

$$K_{L\_Adj} = e^{-3.466t}(K_L - 1) + 1 \tag{20}$$

and

$$K_{M\_Adj} = e^{-3.466t}(K_M - 1) + 1 \tag{21}$$

and

$$K_{H\_Adj} = e^{-3.466t}(K_H - 1) + 1 \tag{22}$$

such that

$$L_{Adj} = K_{L\_Adj} S_D \quad \text{and} \quad M_{Adj} = K_{M\_Adj} S_D \quad \text{and} \quad H_{Adj} = K_{H\_Adj} S_D \tag{23}$$

We first instantiate Equations (18), (19) and, (17) with K_L, K_H, and K_M respectively to yield:

$$\sigma_L = \frac{\tilde{\mu} - K_L}{3} \quad \therefore K_L = \tilde{\mu} - 3\sigma_L \tag{24}$$

and

$$\sigma_H = \frac{K_H - \tilde{\mu}}{3} \quad \therefore K_H = \tilde{\mu} + 3\sigma_H \tag{25}$$

and

$$\tilde{\mu} = \frac{\left(K_L + 4K_M + K_H\right)}{6} \rightarrow K_M = \frac{6\tilde{\mu} - K_L - K_H}{4} \rightarrow K_M = \frac{6\tilde{\mu} - \left(\tilde{\mu} - 3\sigma_L\right) - \left(\tilde{\mu} + 3\sigma_H\right)}{4}$$
$$\therefore K_M = \frac{4\tilde{\mu} + 3\sigma_L - 3\sigma_H}{4} \tag{26}$$

Recall that $\tilde{\mu}$ is always equal to the overall SEER PERT distribution median. We wish to force this value to be equal to the SRDR data set median $\tilde{m}$; therefore,

$$K_L = \tilde{m} - 3\sigma_L \quad \text{and} \quad K_H = \tilde{m} + 3\sigma_H \quad \text{and} \quad K_M = \frac{4\tilde{m} + 3\sigma_L - 3\sigma_H}{4} \tag{27}$$

Substituting the equivalents of K_L, K_H, and K_M in Equations (27) for K_L, K_H, and K_M in Equations (20), (22), and (21) respectively yields

$$K_{L\_Adj} = e^{-3.466t}\left(\tilde{m} - 3\sigma_L - 1\right) + 1 \tag{28}$$

and

$$K_{M\_Adj} = e^{-3.466t}\left(\left(\frac{4\tilde{m} + 3\sigma_L - 3\sigma_H}{4}\right) - 1\right) + 1 \tag{29}$$

and

$$K_{H\_Adj} = e^{-3.466t}\left(\tilde{m} + 3\sigma_H - 1\right) + 1 \tag{30}$$

Appropriate values for $\tilde{m}$ can be found in Table 2 above. Appropriate values for σ_L and σ_H have been determined by using the Microsoft Excel Solver add-in to minimize the percentage standard error of estimate $\%SEE$ between each SRDR data set CDF and its corresponding SEER PERT CDF varying σ_L and σ_H. The results from running Solver and then calculating K_L, K_M, and K_H are shown in Table 4 below.

Table 4: *SEER PERT Distribution Parameters and Resulting Multiplier Values*

	Solver Change Values (Results)		Solver Target (Objective)	SEER-SEM Multiplier Expression Scale Factors		
DSLOC Type	σ[A]	σ[B]	|μ[ACE]-μ[SEER]|	K[L]	K[M]	K[H]
New	0.344122	1.720611	0.017010	-0.828071	-0.828071	5.366128
Pre-Existing	0.241411	1.207058	0.014898	-0.687191	-0.687192	3.658219

Substituting the computed values of K_L, K_M, and K_H in Table 4 for K_L, K_M, and K_H in Equations (20), (21), and (22) for each of New and PER DSLOC yields:

- For New DSLOC:

$$K_{L\_Adj} = -0.828071e^{-3.466t} + 1$$
$$K_{M\_Adj} = -0.828071e^{-3.466t} + 1 \quad \textbf{(31)}$$
$$K_{H\_Adj} = 5.366128e^{-3.466t} + 1$$

- For Pre-Existing DSLOC:

$$K_{L\_Adj} = -0.687191e^{-3.466t} + 1$$
$$K_{M\_Adj} = -0.687192e^{-3.466t} + 1 \quad \textbf{(32)}$$
$$K_{H\_Adj} = 3.658219e^{-3.466t} + 1$$

Substituting the multiplier expressions in the sets of Equations (31) and (32) for the multiplier variables in Equations (23), yields the sets of equations for determining the appropriate Least, Likely, and Most DSLOC values to input into SEER-SEM such that growth, growth uncertainty, and growth uncertainty decay are modeled consistent with the Tecolote DSLOC Estimate Growth Model and with the SRDR data upon which it is based.

- For New DSLOC:

$$Least = S_{D\_New}\left(-0.828071e^{-3.466t} + 1\right)$$
$$Likely = S_{D\_New}\left(-0.828071e^{-3.466t} + 1\right) \quad \textbf{(33)}$$
$$Most = S_{D\_New}\left(5.366128e^{-3.466t} + 1\right)$$

- For Pre-Existing DSLOC:

$$Least = S_{D\_PER}\left(-0.687191e^{-3.466t} + 1\right)$$
$$Likely = S_{D\_PER}\left(-0.687192e^{-3.466t} + 1\right) \quad \textbf{(34)}$$
$$Most = S_{D\_PER}\left(3.658219e^{-3.466t} + 1\right)$$

Figure 7 and Figure 8 below illustrate maximally-right-skewed SEER PERT distribution PDFs that approximate the New DSLOC and PER DSLOC growth factor distributions implied by the SRDR data set CDFs contained in Table 3 above. Figure 9 and Figure 10 below show comparisons between the resulting SEER PERT CDFs and the corresponding SRDR data set CDFs.

Figure 7: *SEER PERT PDF of New DSLOC Growth Factor at 0% Estimate Maturity*

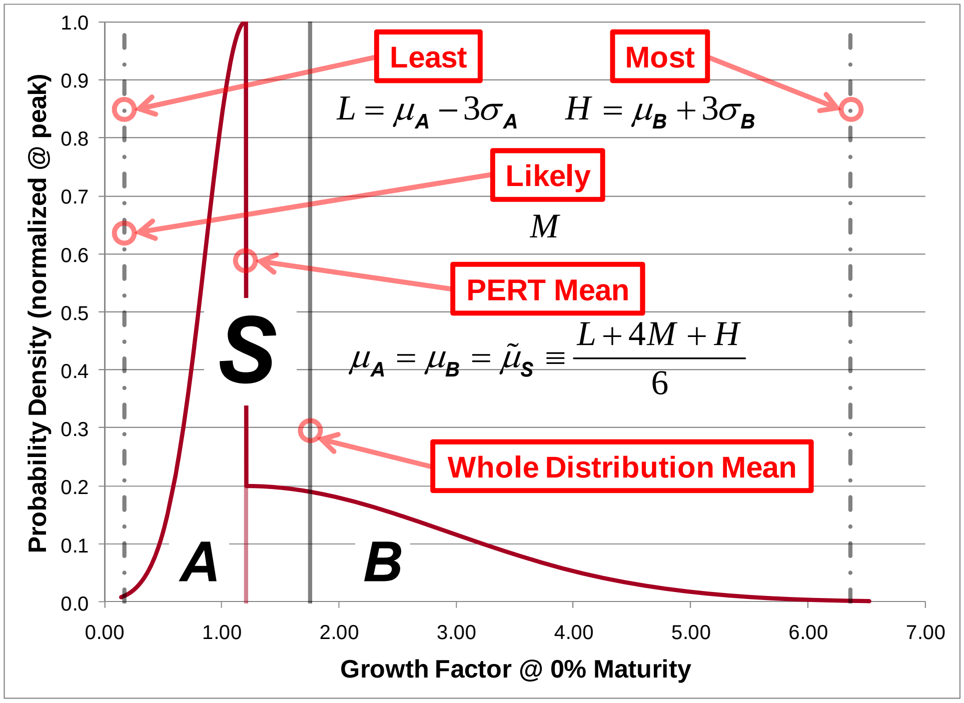


Figure 8: *SEER PERT PDF of PER DSLOC Growth Factor at 0% Estimate Maturity*

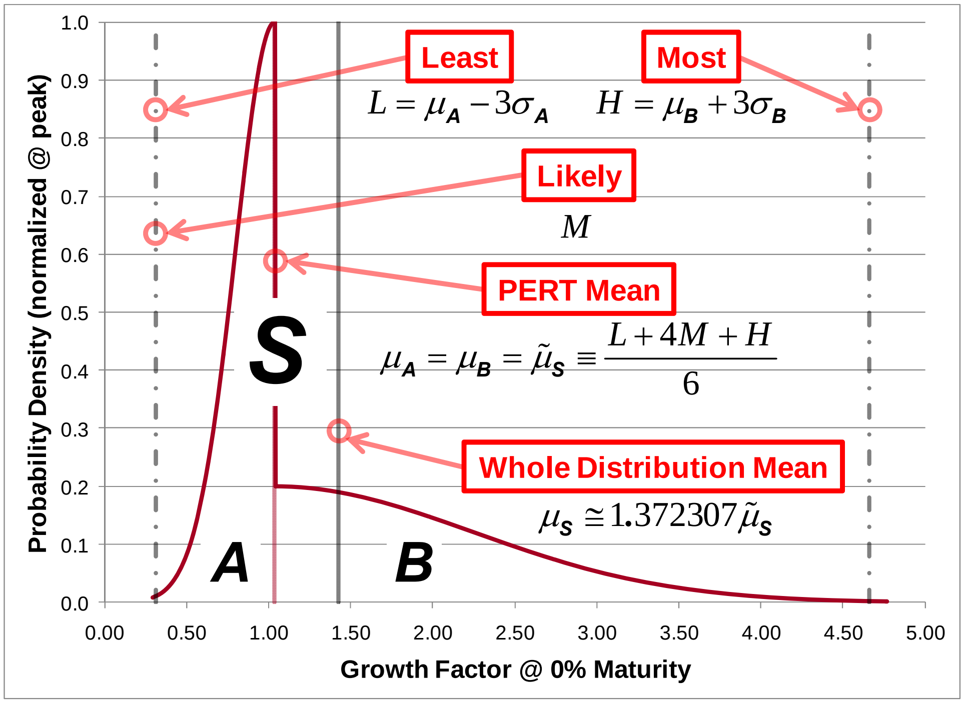


Figure 9: *Comparison of New DSLOC Growth Factor CDFs – SEER PERT vs. SRDR Data*

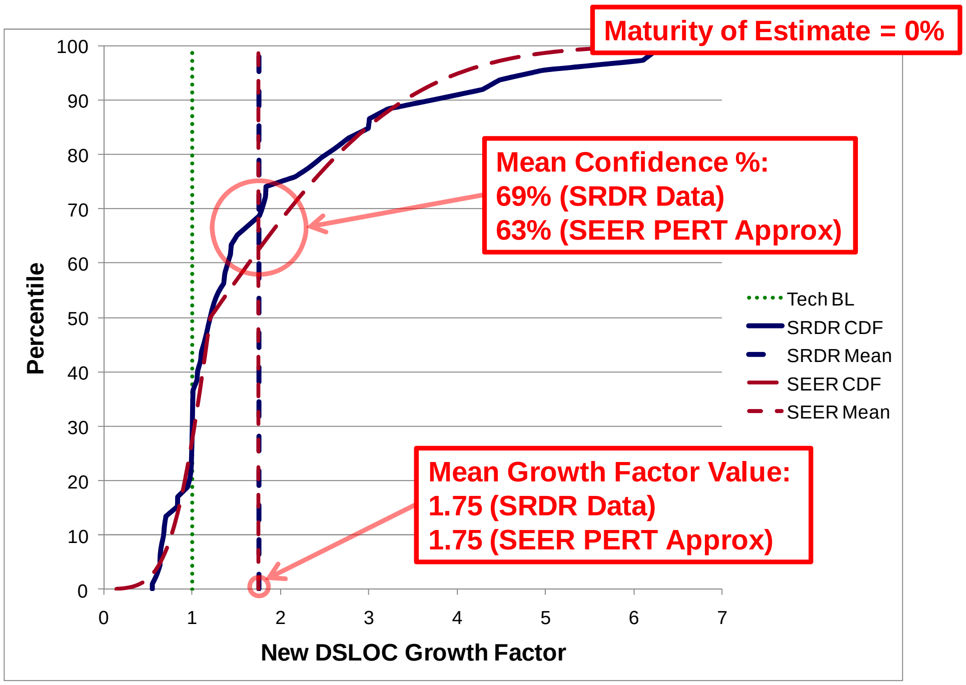
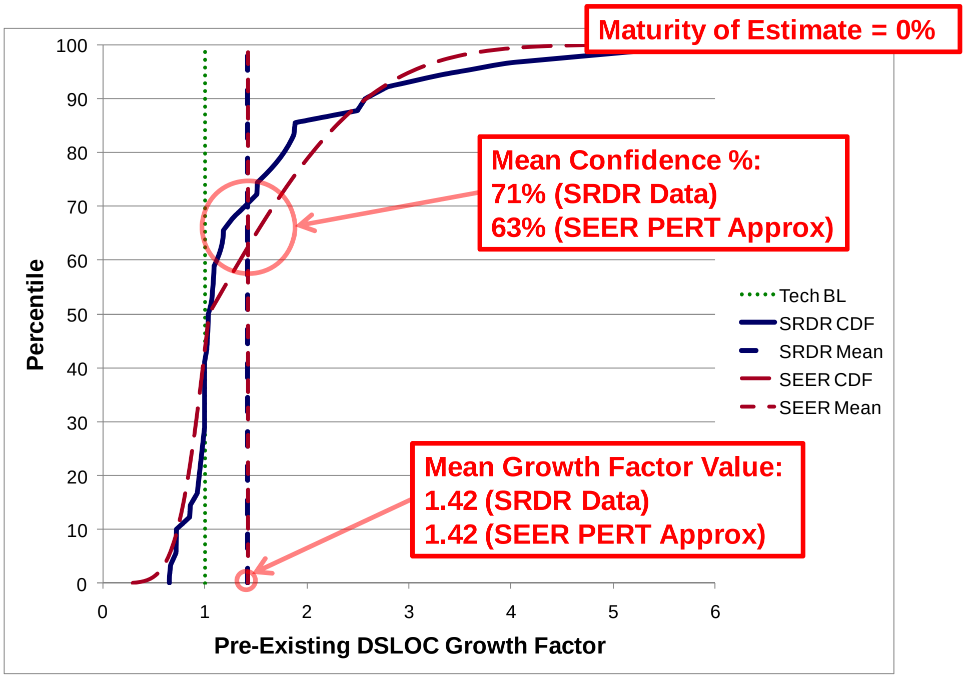


Figure 10: *Comparison of New DSLOC Growth Factor CDFs – SEER PERT vs. SRDR Data*



6. CONCLUSION

It is this author's opinion that the Tecolote DSLOC Estimate Growth Model as described in this paper represents a quantum improvement to the field of available software code growth methodologies. Specifically, advantages of this model over the Holchin (2003) code growth matrix are:

- The Tecolote model is based on AFCAA-collected SRDR data versus Holchin's Delphi survey of *experts* approach.
- The Tecolote model requires only one parameter, Estimate Maturity, which is reasonably objective versus Holchin's rather subjective and vaguely-defined Complexity and Maturity parameters.
- The Tecolote model produces a growth factor distribution result (embodies uncertainty) versus Holchin's single-point growth factor result.
- The Tecolote model provides growth factor distribution decay based on updated Estimate Maturity parameter versus Holchin's single-point growth factor reduction based on updated Complexity and Maturity parameters.
- This model differentiates between New and Pre-Existing DSLOC growth versus Holchin's one-growth-factor-fits-all approach.

This model is currently used as part of the basis for several USAF program office estimates and independent cost estimates. Planned enhancements to this model include rerunning the data analysis using a recently-updated version of the AFCAA SRDR data set. The number of programs and possible stratifications in this new data set may lead to unique baseline growth factor distributions for a particular program type and/or characteristic.

REFERENCES

Boehm, Barry W. 1981. *Software Engineering Economics.* Englewood Cliffs : Prentice-Hall, Inc., 1981. ISBN 0-13-822122-7.

Holchin, Barry. 2003. Code Growth Study. September 17, 2003.

BIOGRAPHY

Michael A. Ross *has over 35 years of experience in software engineering as a developer, manager, process expert, consultant, instructor, and award-winning international speaker. Mr. Ross is currently a Technical Expert for Tecolote Research, Inc. Mr. Ross's previous experience includes three years as President and CEO of r2Estimating, LLC (makers of the r2Estimator software estimation tool), three years as Chief Scientist of Galorath Inc. (makers of the SEER suite of estimation tools), seven years with Quantitative Software Management, Inc. (makers of the SLIM suite of software estimating tools) where he was a senior consultant and Vice President of Education Services, and 17 years with Honeywell Air Transport Systems (formerly Sperry Flight Systems) and 2 years with Tracor Aerospace where he developed and/or managed the development of real-time embedded software for various military and commercial avionics systems. Mr. Ross did his undergraduate work at the United States Air Force Academy and Arizona State University, receiving a Bachelor of Science in Computer Engineering.*