A Probabilistic Method for Predicting Software Code Growth: 2018 Update

ERIC M. SOMMER
USAF Space Command SMC/FMC, El Segundo, California

BOPHA SENG
Tecolote Research, Inc., El Segundo, California

DAVID L. LAPORTE
Tecolote Research, Inc., El Segundo, California

MICHAEL A. ROSS
r2Estimating, LLC, Lakeside, Montana

ABSTRACT

Software estimating is challenging. SMC's approach has evolved over time to tackle this challenge. Originally based on Mike Ross's 2011 DSLOC Estimate Growth Model, we've updated our model to include more recent SRDR data and an improved methodology (Orthogonal Distance Regression). Discussions will focus on non-linear relationships between size and growth, unique growth for new, modified, and unmodified DSLOC, as well as correlation between DSLOC types and future efforts to include space flight software data.

1. INTRODUCTION

The Delivered Source Lines Of Code (DSLOC) Estimate Growth Model version 8 (DEGM8) provides probabilistic growth adjustments to single-point Technical Baseline Estimates (TBEs) of Delivered Source Lines of Code (DSLOC), for New software, Modified software, and Unmodified software, that are sensitive to the estimate *maturity* of the DSLOC TBEs; i.e., when, in 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 and archived by the U.S. Department of Defense's Defense Cost and Resource Center (DCARC). This model represents a significant update and modernization of the Tecolote DSLOC Estimate Growth Model version 7 (DEGM7) (Ross, 2011) in that:

- It is based on recently-updated SRDR data.
- It is based on a better method of regressing the historical data.
- It recognizes non-linear relationships between size and growth.

- It accounts for error (uncertainty) in both the input DSLOC TBEs and the output DSLOC estimates.
- It decomposes the DEGM7 notion of Pre-existing reused software into Modified software and Unmodified software.
- It recognizes correlation between New, Modified, and Unmodified growth.

This new version will be released as DSLOC Estimate Growth Model version 8 (DEGM8).

This paper first summarizes the equations that comprise the model. It then provides a detailed description of the model's basis and components. Next, it describes an example of how to use the model to calculate growth-adjusted software size estimate distributions for the New, Modified, and Unmodified DSLOC that make up an example Computer Software Configuration Item (CSCI). The paper concludes with the authors' collective opinion of the value this model represents.

This paper includes:

- A section describing DEGM8SV (a special instance of the DEGM8) that estimates DSLOC associated with unmanned Space Vehicle (SV) flight software,
- An appendix containing Custom Cumulative Distribution Function (CDF) tables that can be copied into tools such as ACEIT or Crystal Ball in order to construct custom CDFs[1] that are needed to model the baseline New, Modified, and Unmodified DSLOC error factor parameter distributions,
- An appendix containing the regression results of three different data set filtering alternatives,
- An appendix containing a graphic comparison of DEGM7 and DEGM8 behavior,
- An appendix containing a detailed mathematical description of Orthogonal Distance Regression (ODR), a special case of Total Least Squares Regression, and how it is applied to the SRDR DSLOC data,
- An appendix containing a TRI ACEIT implementation of the paper's example growth-adjusted DSLOC estimate.

2. Model Summary

The DEGM8 equations for applying growth and uncertainty to TBE New, Modified, and Unmodified DSLOC are shown in Figure 1 below[2].

$$\boldsymbol{S_{DGANew}} \triangleq S_{DNew} + e^{-(Decay)(Maturity)}\left(\tilde{b}_{GN}\boldsymbol{\varepsilon_{GN}}\left(\frac{S_{DNew}}{K_N}\right)^{a_{GN}} K_N - S_{DNew}\right)$$

$$\boldsymbol{S_{DGAMod}} \triangleq S_{DMod} + e^{-(Decay)(Maturity)}\left(\tilde{b}_{GM}\boldsymbol{\varepsilon_{GM}}\left(\frac{S_{DMod}}{K_M}\right)^{a_{GM}} K_M - S_{DMod}\right)$$

$$\boldsymbol{S_{DGAUmod}} \triangleq S_{DUmod} + e^{-(Decay)(Maturity)}\left(\tilde{b}_{GU}\boldsymbol{\varepsilon_{GU}}\left(\frac{S_{DUmod}}{K_U}\right)^{a_{GU}} K_U - S_{DUmod}\right)$$

Figure 1 DEGM8 equations yield the sum of the appropriate TBE DSLOC value and its calculated DSLOC growth amount. The calculated DSLOC Growth amount is the product of the baseline DSLOC growth amount (zero maturity) and the calculated estimate maturity adjustment factor.

where

$\boldsymbol{S_{DGANew}}$	$\equiv$	***Output –*** growth-adjusted New DSLOC distribution of outcomes with associated attainment probability[3]
$\boldsymbol{S_{DGAMod}}$	$\equiv$	***Output –*** growth-adjusted Modified DSLOC estimate distribution of outcomes with associated attainment probability
$\boldsymbol{S_{DGAUmod}}$	$\equiv$	***Output –*** growth-adjusted Unmodified DSLOC estimate distribution of outcomes with associated attainment probability
$\triangleq$	$\equiv$	Estimator equality symbol; the right expression estimates the left expression
S_{DNew}	$\equiv$	***Input –*** Technical Baseline Estimate (TBE) of New DSLOC
S_{DMod}	$\equiv$	***Input –*** Technical Baseline Estimate (TBE) of Modified DSLOC

S_{DUnmod} ≡ ***Input*** **–** Technical Baseline Estimate (TBE) of Unmodified DSLOC

Decay ≡ ***Model*** **–** Decay constant; default is 3.466 based on Boehm's (1981 pp. 310-311) *Cone of Uncertainty*

Maturity ≡ ***Input*** **–** Estimate Maturity Parameter: (SDLCBegin (ATP, Contract Award) = 0%; SyRR = 10%; SwRR = 20%; SwPDR = 40%; SwCDR = 60%; SwTRR = 80%; SwAccept = 100%)[4],[5]

$\boldsymbol{\varepsilon_{GN}}$ ≡ ***Model*** **–** Baseline (SDLCBegin[6]) New DSLOC growth error factor distribution of outcomes with associated attainment probability; approximated by Custom CDF in Appendix A

$\boldsymbol{\varepsilon_{GM}}$ ≡ ***Model*** **–** Baseline (SDLCBegin) Modified DSLOC growth error factor distribution of outcomes with associated attainment probability; approximated by Custom CDF in Appendix A

$\boldsymbol{\varepsilon_{GU}}$ ≡ ***Model*** **–** Baseline (SDLCBegin) Unmodified DSLOC growth error factor distribution of outcomes with associated attainment probability; approximated by Custom CDF in Appendix A

a_{GN}, a_{GM}, a_{GU} ≡ ***Model*** **–** Exponent parameters for New, Modified, and Unmodified DSLOC growth estimating relationships that are calculated by the regression process

$\tilde{b}_{GN}, \tilde{b}_{GM}, \tilde{b}_{GU}$ ≡ ***Model*** **–** Geometric mean (arithmetic mean in log space) scale factor parameters for New, Modified, and Unmodified DSLOC growth estimating relationships that are calculated by the regression process

K_N, K_M, K_U ≡ ***Input*** **–** Software Item (SI) to Computer Software Configuration Item (CSCI) normalization factors for New, Modified, and Unmodified DSLOC

The following section of this paper describes the basis of these equations.

3. Components of the Model

The DEGM8 estimates, as shown in Figure 1 above, *growth-adjusted total amounts for New, Modified, and Unmodified DSLOC* that are each calculated as the sum of the particular *TBE DSLOC amount* and its associated *maturity-adjusted DSLOC growth amount*.

$$\begin{aligned}
\boldsymbol{S}_{\boldsymbol{DGANew}} &\triangleq S_{DNew} + \boldsymbol{f}\left(Maturity\right)\boldsymbol{S}_{\boldsymbol{DGAmountNewBL}} \\
\boldsymbol{S}_{\boldsymbol{DGAMod}} &\triangleq S_{DMod} + \boldsymbol{f}\left(Maturity\right)\boldsymbol{S}_{\boldsymbol{DGAmountModBL}} \\
\boldsymbol{S}_{\boldsymbol{DGAUmod}} &\triangleq S_{DUmod} + \boldsymbol{f}\left(Maturity\right)\boldsymbol{S}_{\boldsymbol{DGAmountUmodBL}}
\end{aligned} \tag{1}$$

where

$\boldsymbol{f}\left(Maturity\right)$	≡	Maturity adjustment factor function
$\boldsymbol{S}_{\boldsymbol{DGAmountNewBL}}$	≡	Baseline (SDLCBegin) growth amount of New DSLOC expressed as a distribution of outcomes with associated attainment probability
$\boldsymbol{S}_{\boldsymbol{DGAmountModBL}}$	≡	Baseline (SDLCBegin) growth amount of Modified DSLOC expressed as a distribution of outcomes with associated attainment probability
$\boldsymbol{S}_{\boldsymbol{DGAmountUmodBL}}$	≡	Baseline (SDLCBegin) growth amount of Unmodified DSLOC expressed as a distribution of outcomes with associated attainment probability

Technical Baseline Estimated DSLOC Amounts

The DEGM8 accepts, as input, Technical Baseline Estimate (TBE) amounts for New, Modified, and Unmodified DSLOC (S_{DNew}, S_{DMod} , and S_{DUmod}). These TBEs, often called point estimates, are rendered at various times during the program; typically by the program's technical team based on some combination of engineering analysis, relevant past program experience, and expert judgment. These estimates represent the technical team's best guess as to what the final outcome New, Modified, and Unmodified DSLOC values will be when the system is delivered and accepted. Note that these estimates are subject to error (size estimation uncertainty) caused by but not limited to (Ross, 2005):

- Technical team inability to accurately and precisely characterize the requirements as they are known in terms of New, Modified, and Unmodified DSLOC,
- DSLOC definition understanding, differences, and ambiguities,
- Automated code counting inconsistencies and discrepancies,
- Human error in completing the SRDR form,
- Human programmatic bias.

Maturity-Adjusted Growth Amounts

Each maturity-adjusted DSLOC growth amount is calculated as its particular *baseline DSLOC growth amount* (SDLCBegin to SwAccept) scaled by (multiplied by) its associated *maturity adjustment factor*. Note that the amount by which DSLOC grows (or shrinks) is subject to error caused by but not limited to (Ross, 2005):

- The customer doesn't know what he/she wants,
- The customer doesn't understand the problem,
- The mission has changed,
- The regulations that govern how this software should behave have changed,
- The vendor added a few extra features that he/she thought the customer would like,
- The project got behind schedule resulting in some requirements being dropped or postponed,
- The vendor finished early so the customer and/or the vendor thought up a few things to add.

Analysis of the preceding list suggests the following possible organization of issues that influence software size growth:

- Operational environment volatility
- Essence (requirements) volatility
- Essence understanding (requirements completeness and correctness)
- Essence versus implementation correspondence

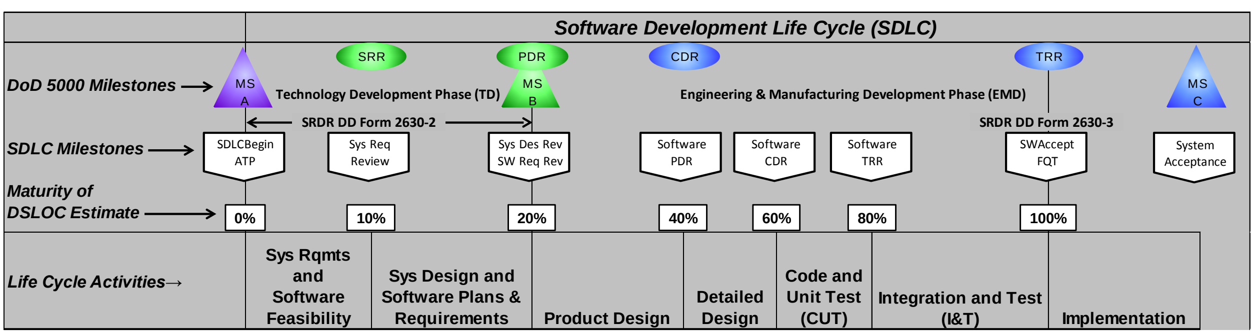
Maturity Adjustment Factor

The maturity adjustment factor component of the DEGM8 represents the portion (percentage) of the baseline (SDLCBegin) DSLOC growth amount that remains to be experienced by the program as a function of where in the SDLC the TBE amounts for New, Modified, and Unmodified DSLOC are rendered by the program's technical team. Obviously, if the estimates are rendered at SDLCBegin (i.e., are based on the knowledge that one would expect the program's technical team to have at SDLCBegin), then the maturity adjustment factor should be 100% (i.e., all of the baseline DSLOC growth amount remains to be experienced). The DEGM8 assumes that, as the program progresses through the SDLC, a sequence of progressively-more-informed DSLOC estimates will be performed by the program's technical team and that the corresponding maturity adjustment factors associated with each of these estimates will monotonically decrease and approach 0% at delivery and acceptance of the software (i.e., a point in the SDLC where virtually all of the growth has been realized).

Normalized Estimate Maturity

In addition to the TBE amounts for New, Modified, and Unmodified DSLOC, the DEGM8 also accepts, as input, normalized estimate *Maturity*. Normalized estimate *Maturity* is assumed to be the *earned* percentage of the program's *actual* (final) SDLCBegin to SwAccept duration that has elapsed at the time the estimate is rendered by the program's technical team. Since the actual SDLCBegin to SwAccept duration cannot be known until SwAccept has occurred, the DEGM8 provides a surrogate quantification of earned duration percentages associated with certain key milestones. This default quantification is contained in Table 1 below. For example, if the program's technical team performs an updated estimate of total New, Modified, and Unmodified DSLOC at Software Preliminary Design Review (SwPDR) then $Maturity = 40\%$. Note that if the DEGM8 were to be used at SwPDR using the TBE DSLOC values rendered at SDLCBegin then $Maturity = 0\%$; i.e., *Maturity* represents the maturity of the estimate. Estimate *Maturity* does *not* represent the maturity of the program or the maturity of the system under development. Note that there is some significant variability in when SRDR DD Form 2630-2's are submitted, which contributes to the variability in DEGM8 estimates.

Table 1 Default normalized estimate maturity scale



SDLCs come in all manner of activities, sequencing, and detail; the SDLC depicted in Table 1 is but one example. Customized versions of the milestone-maturity mapping can be developed based on alternative SDLCs given some knowledge of where in the SDLC's elapsed duration certain selected milestones occur; however, milestones associated with the endpoints ($Maturity = 0\%$ and $Maturity = 100\%$) must correspond to the two points in the SDLC where, respectively, the first SRDR DD Form 2630-2 and the final DD Form 2630-3 are submitted.

Note that it is best to use completed (earned) milestones to determine *Maturity* rather than just using the percentage of the currently-scheduled duration that has elapsed. This is because the currently-scheduled duration is an estimate (i.e., it is never necessarily the final actual duration until the program is complete) and subject to change as the program progresses.

Growth Decay

The DEGM8 assumes some normalized maturity adjustment factor function $\boldsymbol{f}$ of normalized $Maturity$ scaled such that $\boldsymbol{f}(Maturity) \in [0,1]$, where $\boldsymbol{f}(Maturity \mid Maturity = 0)\ \ 1$ represents the maximum (full scale) adjustment factor value (i.e., all the DSLOC growth has yet to be realized), and hypothesizes that $\boldsymbol{f}(Maturity)$ decreases (decays) at a rate proportional to its value (i.e., unrealized growth tends to decay (decrease) faster during the early stages of an SDLC when experience is low and tends to decay slower during the later stages of an SDLC when experience is high). We model this hypothetical behavior mathematically as

$$
\begin{aligned}
&\frac{d\,\boldsymbol{f}(Maturity)}{d(Maturity)} \propto -\boldsymbol{f}(Maturity) \\
&\therefore \frac{d\,\boldsymbol{f}(Maturity)}{d(Maturity)} = -(Decay)\,\boldsymbol{f}(Maturity)
\end{aligned}
\qquad \textbf{(2)}
$$

where

$\propto$	$\equiv$	Proportionality relation; the left expression is directly proportional to the right expression
$Decay$	$\equiv$	Constant of proportionality

Solving the ordinary differential Equation (2) yields

$$
\begin{aligned}
&\frac{d\,\boldsymbol{f}(Maturity)}{\boldsymbol{f}(Maturity)} = -(Decay)\,d(Maturity) \\
&\rightarrow \int \frac{d\,\boldsymbol{f}(Maturity)}{\boldsymbol{f}(Maturity)} = \int -(Decay)\,d(Maturity) \\
&\rightarrow \boldsymbol{ln}\big(\boldsymbol{f}(Maturity)\big) = -(Decay)(Maturity) + c \\
&\therefore \boldsymbol{f}(Maturity) = e^{-(Decay)(Maturity)}e^{c}
\end{aligned}
\qquad \textbf{(3)}
$$

Since we have already posited the constraint $\boldsymbol{f}(Maturity \mid Maturity = 0)\ \ 1$ we can solve Equation (3) for the constant of integration c as

$$f\left(0\right)=e^{-\left(Decay\right)\left(0\right)}e^{c}=1 \rightarrow e^{c}=1 \therefore c=0 \qquad \textbf{(4)}$$

Substituting the equivalent of c in Equation (4) for c in Equation (3) yields the maturity adjustment factor $\boldsymbol{f}\left(Maturity\right)$ portion of the DEGM8 equations in Figure 1 as

$$\begin{gathered}\boldsymbol{f}\left(Maturiy\right)=e^{-\left(Decay\right)\left(Maturity\right)}e^{\left(0\right)}\\ \therefore \underline{\underline{\boldsymbol{f}\left(Maturiy\right)=e^{-\left(Decay\right)\left(Maturity\right)}}}\end{gathered} \qquad \textbf{(5)}$$

In order to render the DEGM8 equations in Figure 1 useful in a particular estimating situation, we need to assume some value (or possibly some distribution) for the decay constant $Decay$. Two methods for accomplishing this are:

(1) to perform a regression analysis of relevant historical data to determine an expected value $Decay$ or distribution $\boldsymbol{Decay}$ and

(2) to assume a value for $Decay$ consistent with the slope of Boehm's (1981 pp. 310-311) *Cone of Uncertainty*. Given the dearth of granular, periodic, and relevant historical DSLOC estimate data available to the authors at the time of this study, the latter method is used as the DEGM8's default position. It is accomplished by scaling the top half of Boehm's *Cone of Uncertainty* to be consistent with the SDLC milestones and percentages in Table 1 and fitting an exponential curve over the scaled result. As shown in Figure 2 below, a near-perfect fit can be achieved with $Decay = 3.466$. Note that the DEGM8 assumes only the curvature (shape) of Boehm's *Cone of Uncertainty* and not its implied scaling (growth percentages).

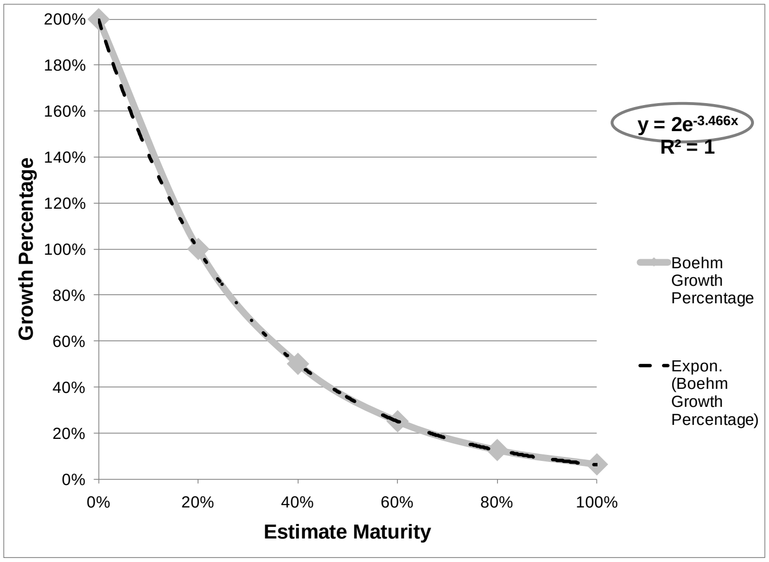


Figure 2 Curve fit of the top half of the Boehm *Cone of Uncertainty* – is near perfect when the fit function is assumed to be $y = 2e^{-3.466}$

Baseline DSLOC Growth Amounts

Without making any assumptions other than that the initial TBEs for New, Modified, and Unmodified DSLOC are related in some way to their respective final outcome values, we posit the following three definitions:

$$\begin{aligned}
\mathbf{S}_{\mathbf{DGAmountNewBL}} &\equiv \boldsymbol{f}_{\boldsymbol{GN}}\left(S_{DNew}\right) - S_{DNew} \\
\mathbf{S}_{\mathbf{DGAmountModBL}} &\equiv \boldsymbol{f}_{\boldsymbol{GM}}\left(S_{DMod}\right) - S_{DMod} \\
\mathbf{S}_{\mathbf{DGAmountUmodBL}} &\equiv \boldsymbol{f}_{\boldsymbol{GU}}\left(S_{DUmod}\right) - S_{DUmod}
\end{aligned} \tag{6}$$

where

$\mathbf{S}_{\mathbf{DGAmountNewBL}}$ ≡ Baseline (SDLCBegin) growth amount portion of New DSLOC expressed as a distribution of outcomes with associated attainment probability

$\mathbf{S}_{\mathbf{DGAmountModBL}}$ ≡ Baseline (SDLCBegin) growth amount portion of Modified DSLOC expressed as a distribution of outcomes with associated attainment probability

$\mathbf{S}_{\mathbf{DGAmountUmodBL}}$ ≡ Baseline (SDLCBegin) growth amount portion of Unmodified DSLOC expressed as a distribution of outcomes with associated attainment probability

S_{DNew} ≡ Technical Baseline Estimate (TBE) of New DSLOC

S_{DMod}	$\equiv$	Technical Baseline Estimate (TBE) of Modified DSLOC
S_{DUmod}	$\equiv$	Technical Baseline Estimate (TBE) of Unmodified DSLOC

Baseline DSLOC Estimate Growth Relationships

For the sake of economy, we will show only the mathematical derivation for specifying $\boldsymbol{f_{GN}}\left(S_{DNew}\right)$ above and assume that the same process can be similarly applied to the specification of $\boldsymbol{f_{GM}}\left(S_{DMod}\right)$ and $\boldsymbol{f_{GU}}\left(S_{DUmod}\right)$. We make certain assumptions about the nature of these relationships, the details of which are described as this section of the paper progresses. Making these assumptions and performing the linear algebra implied by these assumptions, we can elaborate $\boldsymbol{f_{GN}}\left(S_{DNew}\right)$ as follows:

Baseline New DSLOC Growth Relationship

$$\textbf{\textit{Baseline Growth Adjusted DSLOC}} \propto \boldsymbol{f}\left(\textbf{\textit{TBE DSLOC}}\right) \tag{7}$$

$$\boldsymbol{S_{DGANewBL}} \triangleq \tilde{b}_{GN}\boldsymbol{\varepsilon_{GN}}S_{DNew}{}^{a_{GN}}$$

where

$\propto$	$\equiv$	Proportionality operator; the left operand is directly proportional to the right operand.
$\triangleq$	$\equiv$	Estimator equality symbol; the right expression estimates the left expression
$\boldsymbol{\varepsilon_{GN}}$	$\equiv$	Baseline (SDLCBegin) New DSLOC growth error factor parameter distribution of outcomes with associated attainment probability
a_{GN}	$\equiv$	Exponent value parameter for the baseline New DSLOC growth estimating relationship that is calculated by the regression process; this exponent models the nonlinearity (economy or diseconomy of scale) present in the relationship between the initial DSLOC estimate and the final DSLOC actual; the baseline instance of the DEGM8 has a specific value for a_{GN}
$\tilde{b}_{GN}$	$\equiv$	Geometric mean (log space arithmetic mean) scale factor parameter for the New DSLOC growth estimating relationship that is calculated by the regression process

So how do we specify values for the exponent value parameter a_{GN}*, the scale factor parameter* $\tilde{b}_{GN}$*, and the error factor parameter distribution* $\boldsymbol{\varepsilon_{GN}}$ *in Equation (7)?* Obviously any data-driven methodology must start with some data. In this case we require lists of completed Software Item (SI) initially-estimated New, Modified, and Unmodified DSLOC values with lists of corresponding final-actual New, Modified, and Unmodified DSLOC values. These lists we collectively refer to as an historical data set.

DEGM8 Relevant Data

SRDR Data Filtering

A primary objective when stratifying (filtering) historical data sets is to maximize the similarity between included observations while at the same time maximizing the number of observations included in the set. This objective, while often difficult to achieve, is nonetheless intended to both reduce the amount of variability and to increase the statistical significance of the relationships that we derive from the data. The baseline (default) instance of the DEGM8 equation parameter values for New, Modified, and Unmodified DSLOC is based on a subset of Software Resources Data Report (SRDR) data collected and archived by the U.S. Department of Defense's Defense Cost and Resource Center (DCARC)[7]; this subset containing what we heretofore refer to as *relevant* data and satisfying the following filter criteria:

- SerialNo2015: >0 – observation must be included in the 2015 instance of the SRDR database (note that there are 8 observations in the 2011 SRDR database that do not appear in the 2015 SRDR database and have been assigned a SerialNo2015 value of 0)
- Report: 2630-3 – the source document for the observation must be a DD Form 2630-3, which documents the final actuals of a software development project (i.e., not estimated values); note that each DD Form 2630-3 observation in the SRDR database includes its initial-estimate DSLOC values from its associated DD Form 2630-2 form when such form exists
- SI: TRUE – the observation must represent a Computer Software Configuration Item (CSCI)-like Software Item (SI) (i.e., not a collection, summary, or roll-up of multiple CSCIs)
- Nonphysical: TRUE – the observation's DSLOC values must not be measured in units of straight physical lines of code (i.e., they must be measured in logical lines of code (language statements) or non-comment physical lines of code); note that a lack of consistent code counting standards and techniques is a significant source of error in DEGM8 estimates
- GFValid: TRUE – the observation must contain values for:
    - New_DSLOC_GF – New DSLOC implied growth factor; ratio of final actual New DSLOC (SRDR DD Form 2630-3) to initial estimated New DSLOC (SRDR DD Form 2630-2)

- Modified_DSLOC_GF – Modified DSLOC implied growth factor; ratio of final actual Modified DSLOC (SRDR DD Form 2630-3) to initial estimated Modified DSLOC (SRDR DD Form 2630-2)
- Unmodified_DSLOC_GF – Unmodified DSLOC implied growth factor; ratio of final actual Unmodified DSLOC (SRDR DD Form 2630-3) to initial estimated Unmodified DSLOC (SRDR DD Form 2630-2)

that are all inside three geometric standard deviations from their respective population (entire database) geometric mean (see Table 2 below).

Table 2 Statistical outlier filtering comparison; regression JCDER349 with 3 geometric standard deviation statistical outlier filtering was chosen as the basis for the DEGM8

	New DSLOC			Modified DSLOC			Unmodified DSLOC		
	JCDER345	JCDER349	JCDER346	JCDER345	JCDER349	JCDER346	JCDER345	JCDER349	JCDER346
Statistical Outlier Filtering:	None	**3 GeoSigma**	2 GeoSigma	None	**3 GeoSigma**	2 GeoSigma	None	**3 GeoSigma**	2 GeoSigma
Number of Data Points (observations):	302	**225**	213	169	**136**	125	190	**142**	132
Geometric (log space) mean of b:	0.7947	**1.2084**	1.1137	1.4203	**2.6508**	1.9364	0.3723	**0.6199**	0.5499
Arithmetic (unit space) mean of b:	3.4927	**1.7360**	1.4867	4.9077	**4.0600**	2.6865	0.9409	**0.7510**	0.6345
Standard deviation b:	19.1832	**1.8493**	1.3418	18.4790	**4.5566**	2.3590	3.1087	**0.6566**	0.4408
Coefficient of Variation (CV) b:	5.49	**1.07**	0.90	3.77	**1.12**	0.88	3.30	**0.87**	0.69
Arithmetic (unit space) mean of ε:	1.9368	**1.3665**	1.2819	2.2755	**1.5105**	1.4362	1.5683	**1.1911**	1.1280
Standard deviation of ε:	3.2795	**1.2238**	0.9590	4.0775	**1.6230**	1.4924	1.9782	**0.9296**	0.6398
Coefficient of Variation (CV) of ε:	1.69	**0.90**	0.75	1.79	**1.07**	1.04	1.26	**0.78**	0.57
Mean Magnitude of the Relative Error:	61%	**44%**	39%	67%	**50%**	44%	43%	**24%**	22%
Implied Growth Factor at data set arithmetic mean baseline DSLOC:	96% at 74,958 DSLOC	**53% at 59,443 DSLOC**	49% at 60,213 DSLOC	31% at 45,547 DSLOC	**11% at 22,934 DSLOC**	10% at 23,216 DSLOC	34% at 365,311 DSLOC	**7% at 251,323 DSLOC**	11% at 266,322 DSLOC
Implied Growth Factor at data set geometric mean baseline DSLOC:	80% at 25,635 DSLOC	**50% at 23,035 DSLOC**	45% at 23,672 DSLOC	33% at 9,161 DSLOC	**22% at 7,756 DSLOC**	17% at 7,808 DSLOC	14% at 72,523 DSLOC	**1% at 70,790 DSLOC**	3% at 75,292 DSLOC

See Appendix B for details and characteristics of the resulting relevant data set. Regarding the last filter criterion in the list above (GFValid), the geometric mean and the geometric standard deviation of a data set measure or metric are equivalent to the arithmetic mean and arithmetic standard deviation of that same measure or metric taken in log space. We choose the geometric mean and geometric standard deviation because they are more-suitable statistics for this historical data, the distributions of which have significant right skew. This is generally the case with software development essential measures historical data (Ross, 2008). The higher suitability comes from the fact that these geometric statistics provide an outlier determination that is more equitable to both high and low side outliers and thus tends to have less of an influence on the central tendency of the remaining observations.

The authors recognize that choosing to perform statistical filtering of outlier observations, as is done with GFValid above, is subject to some criticism; however, the statistics from the resulting data set show significant reduction in the Coefficients of Variation (CV) to values the authors consider somewhat more reasonable at the risk of possibly being unjustifiably more optimistic. The authors acknowledge the point of view that suggests a no-statistical-filtering strategy might have been more appropriate since it might have more-completely captured the inherent uncertainty. Whether or not this filtering is valid depends on the amount of uncertainty that is due to the ineffectiveness (or lack thereof) of the SRDR data collection, validation, and certification process versus the amount of uncertainty inherent in the SDLC process associated with that SRDR. Experience with the SRDR database has led the authors to be confident in the assumption that the former dominates the latter and are therefore using the statistical filtering as a surrogate for better validation and certification of the SRDR data. With regard to the specific statistical filtering that was used for the default instance of the DEGM8, we chose three geometric standard deviations as the filtering interval over two geometric standard deviations because the differences between the geometric means and CVs of the two alternatives does not, we believe, justify the increased degree of filtering implied by the two standard deviations alternative.

New, Modified, and Unmodified DSLOC Data Sets

Let's assume we have an historical data set containing ordered lists of the relevant measures that represent the initial-estimated values and the final-actual values for each of New, Modified, and Unmodified DSLOC for a population of N completed SIs.

In the case of New DSLOC, we map each of the New initial-estimated DSLOC values list $\mathbf{S_{DNEst}}$ and the New final-actual DSLOC values list $\mathbf{S_{DNAct}}$ to a different dimension of 2-dimension ($\mathbb{R}^2$) space: $\mathbf{S_{DNEst}}$ is represented by displacement along the $\hat{\mathbf{e}}_1$-axis (x-axis) and $\mathbf{S_{DNAct}}$ is represented by displacement along the $\hat{\mathbf{e}}_2$-axis (y-axis).[8] We organize this historical data set of points as the matrix $\mathbf{P}$.

$$\mathbf{P} \equiv \begin{bmatrix} S_{DNEst_1} & S_{DNAct_1} \\ S_{DNEst_2} & S_{DNAct_2} \\ \vdots & \vdots \\ S_{DNEst_N} & S_{DNAct_N} \end{bmatrix} \tag{8}$$

Choosing the Functional Form of the Model

The notions of lines and planes in analytic geometry and linear algebra are *linear*; however, we cannot claim that estimated and actual size can be modeled as a linear combination; in other words we cannot claim that the relationship between $\mathbf{S_{DNEst}}$ and $\mathbf{S_{DNAct}}$ is additive.

$$\boldsymbol{f}\left(S_{DNEst}, S_{DNAct}\right) \cancel{\approx} c_1 S_{DNEst} + c_2 S_{DNAct} + c_3 \tag{9}$$

where c_1 and c_2 are scale factor parameters and c_3 is an offset parameter . Experience with software development project historical data has shown that the relationships between essential measures tend to be *nonlinear* (in this case multiplicative; i.e., the *amount* of change in the output is *not* proportional to the *amount* change in the input which implies the existence of an economy or diseconomy of scale) (Ross, 2008); i.e.,

$$\boldsymbol{f}\left(S_{DNEst}, S_{DNAct}\right) \rightarrow S_{DNEst}{}^{c_1} S_{DNAct}{}^{c_2} c_3 \tag{10}$$

where c_1 and c_2 are exponent parameters and c_3 is a scale factor parameter.

Log Transformation Makes the Problem Linear

One method for applying linear algebra to nonlinear problems is called *log transformation*. If we apply the log transformation function $\boldsymbol{g}$ to the multiplicative function $\boldsymbol{f}$ above; i.e., if we take the natural logarithm of both sides of Equation (10) we get

$$\begin{gathered} \boldsymbol{g}(\boldsymbol{f}) \equiv \boldsymbol{ln}\left(\boldsymbol{f}\left(S_{DNEst}, S_{DNAct}\right)\right) = \boldsymbol{ln}\left(S_{DNEst}{}^{c_1} S_{DNAct}{}^{c_2} c_3\right) \\ \therefore \boldsymbol{g}(\boldsymbol{f}) = c_1\,\boldsymbol{ln}\left(S_{DNEst}\right) \quad c_2\,\boldsymbol{ln}\left(S_{DNAct}\right) \quad \boldsymbol{ln}\left(c_3\right) \end{gathered} \tag{11}$$

which *is* a linear combination (additive). We refer to the function $\boldsymbol{f}$ above as the function in *unit space* and the log-transformed function $\boldsymbol{g}$ as the function in *log space* or *fit space*. The idea is to transform the given problem to log space (i.e., to transform the given problem into something that looks linear), apply appropriate linear algebra to the log-transformed problem, and then transform the results back to unit space. Transforming the log-space result of the linear algebra back to unit space involves exponentiating the result; i.e., raising the Euler number e to the power of the result.

Log transformation provides the added benefit of making the scaling of each dimension somewhat (but not perfectly) *commensurable* and *scale invariant*.[9] In the case we are describing here, scale commensurability and invariance are guaranteed by virtue of the fact that both dimensions are scaled in the same units of measure (DSLOC). We mention this because commensurable and invariant scaling are a prerequisite for Orthogonal Distance Regression, the regression method we will be using later in this section.

If we log transform our historical data set $\mathbf{P}$ in Equation (8) we get $\boldsymbol{ln}(\mathbf{P})$.

$$\boldsymbol{ln}(\mathbf{P}) \equiv \begin{bmatrix} \boldsymbol{ln}\left(S_{DNEst_1}\right) & \boldsymbol{ln}\left(S_{DNAct_1}\right) \\ \boldsymbol{ln}\left(S_{DNEst_2}\right) & \boldsymbol{ln}\left(S_{DNAct_2}\right) \\ \vdots & \vdots \\ \boldsymbol{ln}\left(S_{DNEst_N}\right) & \boldsymbol{ln}\left(S_{DNAct_N}\right) \end{bmatrix} \tag{12}$$

Best Fit Line Through a Log-transformed Data Set

Now that we have the historical data organized, the next step in instantiating the exponent parameter value, the scale factor parameter value, and the error factor distribution is to find the *best fit line* through our N -element log-transformed data set of points $\boldsymbol{ln}(\mathbf{P})$.

Using the linear algebra definition of the parametric system of equations form of a line we can describe our desired best fit line as

$$L_{Best\,Fit} \equiv \begin{cases} P_{Best\,Fit_{S_{DNEst}}} = P'_{Best\,Fit_{S_{DNEst}}} + ta_{S_{DNEst}} \\ P_{Best\,Fit_{S_{DNAct}}} = P'_{Best\,Fit_{S_{DNAct}}} + ta_{S_{DNAct}} \end{cases} \tag{13}$$

where

$\begin{pmatrix} P_{Best\,Fit_{S_{DNEst}}}, \\ P_{Best\,Fit_{S_{DNAct}}} \end{pmatrix}$	$\equiv$	any point on $L_{Best\,Fit}$
$\begin{pmatrix} P'_{Best\,Fit_{S_{DNEst}}}, \\ P'_{Best\,Fit_{S_{DNAct}}} \end{pmatrix}$	$\equiv$	some known point on $L_{Best\,Fit}$
$\begin{bmatrix} a_{S_{DNEst}} \\ a_{S_{DNAct}} \end{bmatrix}$	$\equiv$	a direction vector of $L_{Best\,Fit}$
$t \in \mathbb{R}$	$\equiv$	the scaling parameter of $L_{Best\,Fit}$

Orthogonal Distance Regression

At this point we choose to define *best fit line* as the *Orthogonal Distance Regression* (ODR) *line*; i.e., the line that results from minimizing the sum of the squared orthogonal (shortest) distances from the log-transformed data set points $\boldsymbol{ln}(\mathbf{P})$ to an ODR best fit line L_{ODR} (see Appendix D); i.e., $L_{Best\,Fit} \equiv L_{ODR}$. Note that the application of ODR is a special case of *Total Least Squares Regression*.[10]

Specifying the ODR best fit line L_{ODR} involves finding some point on L_{ODR} and finding a direction vector $\mathbf{a}$ that is parallel to L_{ODR}. We prove, in Appendix D, that the data set centroid of a data set always lies on its ODR best fit line; therefore, L_{ODR} can be specified in terms of the direction vector $\mathbf{a}$ and the centroid $C_{\boldsymbol{ln}(\mathbf{P})}$ of our log-transformed data set as the system of parametric equations shown below in Equations (14) where t is a scaling parameter.

$$L_{ODR} \equiv \begin{cases} P_{ODR_{S_{DNEst}}} = C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}} + ta_{S_{DNEst}} \\ P_{ODR_{S_{DNAct}}} = C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNAct}}} + ta_{S_{DNAct}} \end{cases} \qquad \textbf{(14)}$$

Since we can compute the centroid $\left(C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}}, C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNAct}}}\right)$ from our log-transformed data set $\boldsymbol{ln}(\mathbf{P})$ using the equations

$$\begin{aligned} C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}} &\equiv average\left(\mathbf{S_{DNEst}}\right) = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{ln}\left(S_{DNEst_i}\right) \\ C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNAct}}} &\equiv average\left(\mathbf{S_{DNAct}}\right) = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{ln}\left(S_{DNAct_i}\right) \end{aligned} \qquad \textbf{(15)}$$

we need only find values for the coordinates of the direction vector $\mathbf{a}$ in order to fully specify L_{ODR}.

Singular Value Decomposition

We now use the Singular Value Decomposition (SVD) to solve for the components of the L_{ODR} direction vector $\mathbf{a}$. See Appendix D for more detail about the SVD. In order to use the SVD for our purpose, we must first center our matrix about the origin of $\mathbb{R}^2$. We do this by subtracting, from each data set coordinate, its corresponding centroid $C_{\boldsymbol{ln}(\mathbf{P})}$ coordinate as is shown in Equation (16). By doing this we create matrix $\mathbf{M}$, a $2 \times N$ matrix, the centroid of which lies at the origin of $\mathbb{R}^2$; i.e., at the point $(0,0)$.

$$\mathbf{M} \equiv \begin{bmatrix} \boldsymbol{ln}\left(S_{DNEst_1}\right) - C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}} & \boldsymbol{ln}\left(S_{DNAct_1}\right) - C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNAct}}} \\ \boldsymbol{ln}\left(S_{DNEst_2}\right) - C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}} & \boldsymbol{ln}\left(S_{DNAct_2}\right) - C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNAct}}} \\ \vdots & \vdots \\ \boldsymbol{ln}\left(S_{DNEst_N}\right) - C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}} & \boldsymbol{ln}\left(S_{DNAct_N}\right) - C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNAct}}} \end{bmatrix} \qquad \textbf{(16)}$$

We next perform an SVD of matrix $\mathbf{M}$. $\boldsymbol{SVD}(\mathbf{M})$ is a special factorization of $\mathbf{M}$ such that

$$\boldsymbol{SVD}(\mathbf{M}) \equiv \{\mathbf{U}, \boldsymbol{\Sigma}, \mathbf{V}^{\mathsf{T}}\} \mid \mathbf{M} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathsf{T}} \qquad \textbf{(17)}$$

where

$\mathbf{M}$	$\equiv$	Given origin-centered data set matrix $(N \times 2)$
$\mathbf{U}$	$\equiv$	Orthogonal matrix $(N \times 2)$***; not used as part of the ODR process***
$\boldsymbol{\Sigma}$	$\equiv$	(Greek sigma, not to be confused with summation) Square diagonal matrix (2×2); contains the singular values of $\mathbf{M}$
$\mathbf{V}$	$\equiv$	Orthogonal matrix (2×2); the transpose of this matrix $\mathbf{V}^{\mathsf{T}}$ (as shown below) contains the singular vectors of $\mathbf{M}$ organized as column vectors

$$\begin{aligned}\mathbf{M} &= \mathbf{U}\begin{bmatrix}\Sigma_{1,1} & 0 \\ 0 & \Sigma_{2,2}\end{bmatrix}\begin{bmatrix}V_{1,1} & V_{1,2} \\ V_{2,1} & V_{2,2}\end{bmatrix}^{\mathsf{T}} \\ &= \mathbf{U}\begin{bmatrix}\Sigma_{max} & 0 \\ 0 & \Sigma_{2,2}\end{bmatrix}\begin{bmatrix}a_{S_{DNEst}} & V_{1,2} \\ a_{S_{DNAct}} & V_{2,2}\end{bmatrix}^{\mathsf{T}}\end{aligned} \qquad \textbf{(18)}$$

The ODR best fit line's direction vector $\mathbf{a}$ is the singular vector of $\mathbf{M}$ that corresponds to the largest singular value of $\mathbf{M}$; in this case either $\Sigma_{1,1}$ or $\Sigma_{2,2}$. A feature of the algorithm we are using to implement the SVD is that it returns $\boldsymbol{\Sigma}$ and $\mathbf{V}$ such that their contained values / vectors are sorted in descending singular value order from left to right. This implies the vector we are looking for is always in the leftmost column of $\mathbf{V}$ as shown in Equation (18).

Instantiating the ODR Best-Fit Line

We now have values for the components of direction vector $\mathbf{a}$ from Equation (18) and the components of the log-transformed data set centroid $C_{\boldsymbol{ln}(\mathbf{P})}$, which can be used to specify the L_{ODR} system of Equations (14). Since the ODR best fit line in log space represents our desired estimating relationship, it follows that any log-transformed single-point estimate of growth-adjusted New DSLOC $S_{DGANewBL}$, as represented by the point $P = (\boldsymbol{ln}(S_{DNew}), \boldsymbol{ln}(S_{DGANewBL}))$, must necessarily lie on L_{ODR}. Substituting the coordinates of this single-point estimate P into Equations (14) and then solving each of the equations for the scalar t gives us

$$L_{ODR} \equiv \begin{cases} t = \dfrac{\boldsymbol{ln}\left(S_{DNew}\right) - C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}}}{a_{S_{DNEst}}} \\ t = \dfrac{\boldsymbol{ln}\left(S_{DGANewBL}\right) - C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNAct}}}}{a_{S_{DNAct}}} \end{cases} \tag{19}$$

The right side expression of each equation in the L_{ODR} system of Equations (19) is equal to t; therefore, we can set these expressions equal to each other.

$$\frac{\boldsymbol{ln}\left(S_{DNew}\right) - C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}}}{a_{S_{DNEst}}} = \frac{\boldsymbol{ln}\left(S_{DGANewBL}\right) - C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNAct}}}}{a_{S_{DNAct}}} \tag{20}$$

Solving Equation (20) for $\boldsymbol{ln}\left(S_{DGANewBL}\right)$, the estimated value we are looking for, yields

$$\boldsymbol{ln}\left(S_{DGANewBL}\right) = \frac{a_{S_{DNAct}}}{a_{S_{DNEst}}}\left(\boldsymbol{ln}\left(S_{DNew}\right) \quad C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}}\right) + C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNAct}}} \tag{21}$$

Transforming the ODR Best-Fit Line to Unit Space

We now transform our log space Equation (21) to unit space by exponentiating both sides of the equation; i.e., taking each of the two equivalent expressions as a power of e such that $\boldsymbol{exp}(x) \leftrightarrow e^x$. This leads to the following algebraic manipulation and simplification:

$$
\begin{aligned}
\boldsymbol{exp}\left(\boldsymbol{ln}\left(S_{DGANewBL}\right)\right) &= \boldsymbol{exp}\left(\frac{a_{S_{DNAct}}}{a_{S_{DNEst}}}\left(\boldsymbol{ln}\left(S_{DNew}\right) - C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}}\right) + C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNAct}}}\right) \\
\rightarrow S_{DGANewBL} &= \boldsymbol{exp}\left(\frac{a_{S_{DNAct}}}{a_{S_{DNEst}}}\left(\boldsymbol{ln}\left(S_{DNew}\right) - C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}}\right)\right)\boldsymbol{exp}\left(C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNAct}}}\right) \\
\rightarrow S_{DGANewBL} &= \frac{\boldsymbol{exp}\left(\frac{a_{S_{DNAct}}}{a_{S_{DNEst}}}\boldsymbol{ln}\left(S_{DNew}\right)\right)}{\boldsymbol{exp}\left(\frac{a_{S_{DNAct}}}{a_{S_{DNEst}}}C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}}\right)}\boldsymbol{exp}\left(C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNAct}}}\right) \qquad \textbf{(22)} \\
\therefore S_{DGANewBL} &= \frac{\boldsymbol{exp}\left(C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNAct}}}\right)}{\boldsymbol{exp}\left(\frac{a_{S_{DNAct}}}{a_{S_{DNEst}}}C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}}\right)} S_{DNew}^{\frac{a_{S_{DNAct}}}{a_{S_{DNEst}}}}
\end{aligned}
$$

Letting $a_{GN} \equiv \frac{a_{S_{DNAct}}}{a_{S_{DNEst}}}$ and $b_{GN} \equiv \frac{\boldsymbol{exp}\left(C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNAct}}}\right)}{\boldsymbol{exp}\left(\frac{a_{S_{DNAct}}}{a_{S_{DNEst}}}C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}}\right)}$, and substituting a_{GN} and b_{GN} in Equation (22) with their equivalent expressions above gives us

$$S_{DGANewBL} = b_{GN} S_{DNew}{}^{a_{GN}} \qquad \textbf{(23)}$$

The application of the SVD gives us specific values for $a_{S_{DNAct}}$ and $a_{S_{DNEst}}$, and, therefore, a specific value for a_{GN}. However we, as yet, have no specific value for b_{GN}. We can remedy this by first instantiating Equation (23) with the appropriate data set lists $\mathbf{S_{DNEst}}$ and $\mathbf{S_{DNAct}}$ of observations as defined in Equation (8) and then solve for the list of corresponding scale factor parameter values b_{GN} to get the list $\mathbf{b_{GN}}$.

$$\left.\left(b_{GN_i} = \frac{S_{DNAct_i}}{S_{DNEst_i}{}^{a_{GN}}}\right)\right|_{i=1}^{N} \rightarrow \mathbf{b_{GN}} = \frac{\mathbf{S_{DNAct}}}{\mathbf{S_{DNEst}}{}^{a_{GN}}} \qquad \textbf{(24)}$$

We then use an appropriate central tendency value of the list $\mathbf{b_{GN}}$ to estimate b_{GN}. Since b_{GN} exists in unit space but is based on ODR performed in log space, we choose to use the arithmetic mean of $\mathbf{b_{GN}}$ in log space (geometric mean of $\mathbf{b_{GN}}$ in unit space) as the estimator of b_{GN}. Since we can now compute a value $\tilde{b}_{GN}$ for the geometric mean of the list $\mathbf{b_{GN}}$,

$$\tilde{b}_{GN} \equiv \boldsymbol{GeoMean}\left(\mathbf{b_{GN}}\right) = \boldsymbol{exp}\left(\frac{1}{n}\sum_{i=1}^{N}\boldsymbol{ln}\left(b_{GN_i}\right)\right) \tag{25}$$

we instantiate Equation (23) with the value $\tilde{b}_{GN}$ to get

$$S_{DGANewBL} \triangleq \tilde{b}_{GN} S_{DNew}{}^{a_{GN}} \tag{26}$$

Error and Uncertainty

Orthogonal Distance Represents the Error

Notice that Equation (26) is a single-point (non-probabilistic) version of the baseline New DSLOC estimate growth Equation (7). It estimates specific baseline growth-adjusted New DSLOC as a function of specific single values of a_{GN}, $\tilde{b}_{GN}$, and TBE New DSLOC S_{DNew}. We refer to Equation (26) as being expressed in *estimator form*, recognized by the fact that baseline growth-adjusted DSLOC $S_{DGANewBL}$ is being estimated by the expression to the right of the "$\triangleq$" symbol.

A reality of estimation is the existence of *error* in both the independent and dependent variables of an estimating relationship. There exist numerous technological, programmatic, personnel, and modelling factors that together determine the magnitude and direction of this error (Ross, 2003). Earlier in the paper we listed several causes for this error specific to software growth. When using ODR, the resultant orthogonal error vector $\mathbf{n}_{\boldsymbol{\varepsilon}_{\mathbf{GN}_i}}$ for a given observation data point $\boldsymbol{ln}(P_i) = \left[\boldsymbol{ln}\left(S_{DNEst_i}\right) \quad \boldsymbol{ln}\left(S_{DNAct_i}\right)\right]$ is the vector $\overline{Q_i\,\boldsymbol{ln}(P_i)}$ where point Q_i is on the best fit line L_{ODR} and where $\overline{Q_i\,\boldsymbol{ln}(P_i)}$ is orthogonal (normal) to L_{ODR}. The point Q_i necessarily represents the point on L_{ODR} that is closest to the point $\boldsymbol{ln}(P_i)$. In our case, each error vector $\mathbf{n}_{\boldsymbol{\varepsilon}_{\mathbf{GN}_i}}$ has one component vector in each of the two dimensions of our New DSLOC data subset. One of these component vectors $\boldsymbol{\varepsilon}_{\mathbf{GNEst}_i}$ is associated with the independent variable $\boldsymbol{ln}\left(S_{DNEst_i}\right)$ and is parallel t0 the $\hat{\mathbf{e}}_1$-axis (x-axis) and the other component vector $\boldsymbol{\varepsilon}_{\mathbf{GNAct}_i}$ is associated with the dependent variable $\boldsymbol{ln}\left(S_{DNAct_i}\right)$ and is parallel to the $\hat{\mathbf{e}}_2$-axis (y-axis). The geometry of the resultant error and its components is illustrated in Figure 3 below.

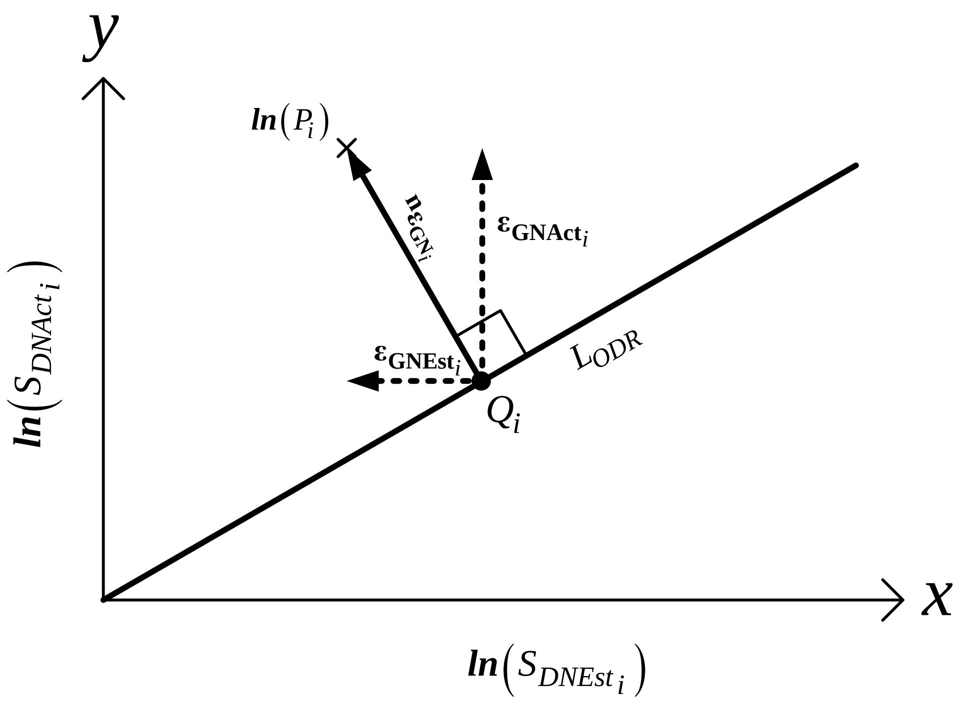


Figure 3 Geometry of an example point $\boldsymbol{ln}(P_i)$, its resultant orthogonal error vector $\mathbf{n}_{\boldsymbol{\varepsilon}_{\mathbf{GN}_i}}$, and its component vectors $\boldsymbol{\varepsilon}_{\mathbf{GNEst}_i}$ and $\boldsymbol{\varepsilon}_{\mathbf{GNAct}_i}$

Analytic geometry[11] provides that, for a line given by the equation $ax+by+c=0$ where a, b, and c are real constants with a and b not both zero, the point $Q=(x,y)$ on this line that is closest to some point (x_0,y_0) has the coordinates

$$x=\frac{b\left(bx_0-ay_0\right)-ac}{a^2+b^2} \text{ and } y=\frac{a\left(-bx_0+ay_0\right)-bc}{a^2+b^2} \tag{27}$$

We can rewrite Equation (21) to match the $ax+by+c=0$ form as

$$\begin{gathered}\boldsymbol{ln}\left(S_{DGANewBL}\right)=\frac{a_{S_{DNAct}}}{a_{S_{DNEst}}}\left(\boldsymbol{ln}\left(S_{DNew}\right) \quad C_{\boldsymbol{ln}(\mathbf{P})S_{DNEst}}^{+}\right) \quad C_{\boldsymbol{ln}(\mathbf{P})S_{DNAct}} \\ \therefore\frac{a_{S_{DNAct}}}{a_{S_{DNEst}}}\boldsymbol{ln}\left(S_{DNew}\right)-\boldsymbol{ln}\left(S_{DGANewBL}\right)+C_{\boldsymbol{ln}(\mathbf{P})S_{DNAct}}-\frac{a_{S_{DNAct}}}{a_{S_{DNEst}}}C_{\boldsymbol{ln}(\mathbf{P})S_{DNEst}}=0\end{gathered} \tag{28}$$

which implies

$$a \equiv \frac{a_{S_{DNAct}}}{a_{S_{DNEst}}}=a_{GN} \text{ (previously defined)}$$

$$b \equiv -1$$

$$c \equiv C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNAct}}} - \frac{a_{S_{DNAct}}}{a_{S_{DNEst}}} C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}}$$

We let the point (x_0, y_0) represent a log-transformed observation data point $\boldsymbol{ln}(P_i) = \left(\boldsymbol{ln}\left(S_{DNEst_i}\right), \boldsymbol{ln}\left(S_{DNAct_i}\right)\right)$; therefore,

$$x_0 \equiv \boldsymbol{ln}\left(S_{DNEst_i}\right)$$

$$y_0 \equiv \boldsymbol{ln}\left(S_{DNAct_i}\right)$$

Making the appropriate substitutions into Equations (27) and simplifying the result yields

$$Q_i = \left(Q_{i_x}, Q_{i_y}\right)$$

$$\left(\frac{\left(\boldsymbol{ln}\left(S_{DNEst_i}\right) + a_{GN}\, \boldsymbol{ln}\left(S_{DNAct_i}\right)\right) - a_{GN}\left(C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNAct}}} - a_{GN} C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}}\right)}{a_{GN}{}^2 + 1}, \frac{a_{GN}\left(\boldsymbol{ln}\left(S_{DNEst_i}\right) + a_{GN}\, \boldsymbol{ln}\left(S_{DNAct_i}\right)\right) + \left(C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNAct}}} - a_{GN} C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}}\right)}{a_{GN}{}^2 + 1} \right) \quad \mathbf{(29)}$$

We can now define each of the signed magnitudes of each error component as

$$\varepsilon_{SDNEst_i} \equiv \boldsymbol{ln}\left(S_{DNEst_i}\right) - Q_{i_x} \quad \mathbf{(30)}$$

and

$$\varepsilon_{SDNAct_i} \equiv \boldsymbol{ln}\left(S_{DNAct_i}\right) - Q_{i_y} \quad \mathbf{(31)}$$

using Equation (29) to instantiate each of Q_{i_x} and Q_{i_y}.

We can instantiate Equations (30) and (31) with the appropriate lists $\mathbf{S_{DNEst}}$, $\mathbf{S_{DNAct}}$, $\mathbf{Q_x}$, and $\mathbf{Q_y}$, to get corresponding error lists $\boldsymbol{\varepsilon}_{\mathbf{SDNEst}}$ and $\boldsymbol{\varepsilon}_{\mathbf{SDNAct}}$ as

$$\left(\varepsilon_{SDNEst_i} \equiv \boldsymbol{ln}\left(S_{DNEst_i}\right) - Q_{i_x}\right)\Big|_{i=1}^{N} \to \boldsymbol{\varepsilon}_{\mathbf{SDNEst}} = \boldsymbol{ln}\left(\mathbf{S_{DNEst}}\right) \quad \mathbf{Q_x} \quad \mathbf{(32)}$$

and

$$\left(\varepsilon_{SDNAct_i} \equiv \boldsymbol{ln}\left(S_{DNAct_i}\right) - Q_{i_y}\right)\Big|_{i=1}^{N} \to \boldsymbol{\varepsilon}_{\mathbf{SDNAct}} = \boldsymbol{ln}\left(\mathbf{S_{DNAct}}\right) \quad \mathbf{Q_y} \quad \mathbf{(33)}$$

One advantage of using ODR instead of an Ordinary Least Squares (OLS)-based regression method is this ability to isolate the error in each of the estimating relationship's variables rather than using just the dependent variable's residual[12] value to represent the error.

Because we have developed our estimating relationship using ODR in log space; the error, while additive in log space; must be transformed to unit space and treated as multiplicative. We can rewrite Equation (26) to account for this multiplicative error in each variable (Equations (32) and (33)) as

$$S_{DGANewBL}\,\boldsymbol{exp}\left(\boldsymbol{\varepsilon}_{\mathbf{S_{DNAct}}}\right) \triangleq \tilde{b}_{GN}\left(S_{DNew}\,\boldsymbol{exp}\left(\boldsymbol{\varepsilon}_{\mathbf{S_{DNEst}}}\right)\right)^{a_{GN}}$$
$$\therefore \mathbf{S_{DGANewBL}} \triangleq \tilde{b}_{GN}\frac{\boldsymbol{exp}\left(\boldsymbol{\varepsilon}_{\mathbf{S_{DNEst}}}\right)^{a_{GN}}}{\boldsymbol{exp}\left(\boldsymbol{\varepsilon}_{\mathbf{S_{DNAct}}}\right)}S_{DNew}{}^{a_{GN}} \quad \mathbf{(34)}$$

where

$\boldsymbol{\varepsilon}_{\mathbf{S_{DNEst}}}$ ≡ Log space error component list associated with the list $\boldsymbol{ln}\left(\mathbf{S_{DNEst}}\right)$

$\boldsymbol{\varepsilon}_{\mathbf{S_{DNAct}}}$ ≡ Log space error component list associated with the list $\boldsymbol{ln}\left(\mathbf{S_{DNAct}}\right)$

Letting

$$\boldsymbol{\varepsilon}_{\mathbf{GN}} \equiv \frac{\boldsymbol{exp}\left(\boldsymbol{\varepsilon}_{\mathbf{S_{DNEst}}}\right)^{a_{GN}}}{\boldsymbol{exp}\left(\boldsymbol{\varepsilon}_{\mathbf{S_{DNAct}}}\right)} \quad \mathbf{(35)}$$

and then substituting $\boldsymbol{\varepsilon}_{\mathbf{GN}}$ for its equivalent into Equation (34) gives us

$$\mathbf{S_{DGANewBL}} \triangleq \tilde{b}_{GN}\boldsymbol{\varepsilon}_{\mathbf{GN}}S_{DNew}{}^{a_{GN}} \quad \mathbf{(36)}$$

If we were to assume that the historical values in the two corresponding $\boldsymbol{ln}\left(\mathbf{S_{DNEst}}\right)$ and $\boldsymbol{ln}\left(\mathbf{S_{DNAct}}\right)$ lists of log-transformed relevant measures are perfectly correlated; i.e., each and every historical observation in the data set $\boldsymbol{ln}\left(P_i\right)=\left[\boldsymbol{ln}\left(S_{DNEst_i}\right) \quad \boldsymbol{ln}\left(S_{DNAct_i}\right)\right]$ lies on the ODR best-fit line L_{ODR}, then every element of the error factor parameter list $\boldsymbol{\varepsilon}_{\mathbf{GN}}$ in Equation (36) would have the same value (specifically a value of 1); therefore, the list $\boldsymbol{\varepsilon}_{\mathbf{GN}}$ would be unnecessary and the list $\mathbf{S_{DGANewBL}}$ would simply be the single value $S_{DGANewBL}$. The estimator form of the baseline New DSLOC estimate growth equation sans $\boldsymbol{\varepsilon}_{\mathbf{GN}}$ would be sufficient to do perfectly-certain (i.e., *deterministic*) estimating with respect to some given value of S_{DNew}. This scenario, however, is not realistic since, for each observation, error is almost always present in each of the relevant measures. We will henceforth refer to the net effect of this error as *uncertainty* since its presence is what makes estimation uncertain; i.e., *stochastic* or *probabilistic* and not *deterministic* or a sure thing. Since Equation (36) contains only three parameters, a_{GN}, $\tilde{b}_{GN}$, and $\boldsymbol{\varepsilon}_{\mathbf{GN}}$; and since a_{GN} and $\tilde{b}_{GN}$ are constants of the regression, it follows that the range and distribution of the values in the $\boldsymbol{\varepsilon}_{\mathbf{GN}}$ list can be used to model the uncertainty in Equation (36). Because the list $\boldsymbol{\varepsilon}_{\mathbf{GN}}$ is being used to model a range of possible outcomes, we must indicate Equation (36) as a probabilistic estimating relationship by replacing the calibration form list variable $\boldsymbol{\varepsilon}_{\mathbf{GN}}$ with its corresponding estimator form random (distribution) variable $\boldsymbol{\varepsilon_{GN}}$. Since the variable $S_{DGANewBL}$, represented in Equation (36) as the list of estimates $\mathbf{S_{DGANewBL}}$, is now being estimated as a function of $\boldsymbol{\varepsilon_{GN}}$, it follows that $S_{DGANewBL}$ is now the random variable $\boldsymbol{S_{DGANewBL}}$. Note that S_{DNew} remains a single-value variable since we have specified that this be given as a single-value TBE. Therefore,

$$\underline{\underline{\boldsymbol{S_{DGANewBL}} \triangleq \tilde{b}_{GN}\boldsymbol{\varepsilon_{GN}}S_{DNew}{}^{a_{GN}}}} \qquad \textbf{(37)}$$

which is identical to Equation (7).

Custom CDFs Approximate Random Variable Distributions

One of the features of TRI's ACE tool (as is the case with other similar statistical tools) is that it supports the notion of defining random variables in terms of what ACE refers to as a *custom Cumulative Distribution Function (CDF)* (see Endnote 1). We wish to approximate $\boldsymbol{\varepsilon_{GN}}$ with a custom CDF that is based on the list $\boldsymbol{\varepsilon}_{\mathbf{GN}}$; therefore, $\boldsymbol{\varepsilon_{GN}} \cong \boldsymbol{customCDF}\left(\boldsymbol{\varepsilon}_{\mathbf{GN}}\right)$ where $\boldsymbol{customCDF}\left(\boldsymbol{\varepsilon}_{\mathbf{GN}}\right)$ represents a process that returns an $N \times 2$ matrix $\mathbf{C}$, where N is the number of elements in $\boldsymbol{\varepsilon}_{\mathbf{GN}}$, such that

$$\begin{aligned} C_{1,2}, C_{2,2} \ldots C_{N,2} &\equiv \boldsymbol{ascending\_sort}\left(\boldsymbol{\varepsilon}_{\mathbf{GN}}\right) \\ C_{1,1}, C_{2,1} \ldots C_{N,1} &\equiv \boldsymbol{percentile\_rank}\left(C_{1,2}, C_{2,2} \ldots C_{N,2}\right) \end{aligned} \qquad \textbf{(38)}$$

The result is a matrix $\mathbf{C}$ where the second column contains an ascending-sorted list of the values from $\boldsymbol{\varepsilon}_{\mathbf{GN}}$ and where the first column contains the percentile rank of its associated value in the second column. This matrix $\mathbf{C}$ can be copied into ACE as the definition of a custom CDF that can be used to approximate the probability distribution associated with $\boldsymbol{\varepsilon_{GN}}$.

Equations for Baseline New, Modified, and Unmodified DSLOC

If we apply the same overall baseline DSLOC estimate growth relationship derivation, that we used for New DSLOC above, to each of Modified and Unmodified DSLOC we get all three equations we are looking for:

$$\begin{aligned} \boldsymbol{S_{DGANewBL}} &\triangleq \tilde{b}_{GN} \boldsymbol{\varepsilon_{GN}} S_{DNew}{}^{a_{GN}} \\ \boldsymbol{S_{DGAModBL}} &\triangleq \tilde{b}_{GM} \boldsymbol{\varepsilon_{GM}} S_{DMod}{}^{a_{GM}} \\ \boldsymbol{S_{DGAUmodBL}} &\triangleq \tilde{b}_{GU} \boldsymbol{\varepsilon_{GU}} S_{DUmod}{}^{a_{GU}} \end{aligned} \tag{39}$$

Custom CDFs for each of $\boldsymbol{\varepsilon_{GN}}$, $\boldsymbol{\varepsilon_{GM}}$, and $\boldsymbol{\varepsilon_{GU}}$ that are specific to the default instance of the DEGM8 can be found in Appendix A.

Software Item Normalization Factor

The observations in the relevant data set being used as the basis for the default instance of the DEGM8 are the result of overall SRDR database filtering that includes only those observations that represent Computer Software Configuration Item (CSCI)-like Software Items (SIs). In cases where the DEGM8 is being used to estimate growth-adjusted DSLOC for SIs other than CSCI-like SIs, a normalization of S_{DNew} , S_{DMod} , and S_{DUmod} is necessary to make the scaling of these three values consistent with the historical data in order to prevent disproportionate application of the economy or diseconomy of scale effects resulting from the exponents a_{GN} , a_{GM} , and a_{GU} . We therefore enhance Equations (39) to incorporate this scaling normalization as

$$\begin{aligned} \boldsymbol{S_{DGANewBL}} &\triangleq \tilde{b}_{GN} \boldsymbol{\varepsilon_{GN}} \left(\frac{S_{DNew}}{K_N}\right)^{a_{GN}} K_N \\ \boldsymbol{S_{DGAModBL}} &\triangleq \tilde{b}_{GM} \boldsymbol{\varepsilon_{GM}} \left(\frac{S_{DMod}}{K_M}\right)^{a_{GM}} K_M \\ \boldsymbol{S_{DGAUmodBL}} &\triangleq \tilde{b}_{GU} \boldsymbol{\varepsilon_{GU}} \left(\frac{S_{DUmod}}{K_U}\right)^{a_{GU}} K_U \end{aligned} \tag{40}$$

where

K_N, K_M, K_U ≡ Software Item (SI) to Computer Software Configuration Item (CSCI) normalization factors for New, Modified, and Unmodified DSLOC such that each equals the number of CSCIs represented by the SI; for example, if the SI being estimated is one of four equal-size components of a CSCI, then each of K_N, K_M, and K_U would equal 0.25. Likewise, if the SI being estimated is a collection of four CSCIs, then each of K_N, K_M, and K_U would equal 4.

Converting Growth-Adjusted Totals to Growth Amounts

Notice that, as written, Equations (40) estimate baseline growth-adjusted *total amounts* of DSLOC; however, we will next be applying a maturity adjustment but only wish to apply that adjustment to the *growth amount* portion of the total and not to the whole growth-adjusted total. Therefore, we rewrite Equations (40) to include subtracting the given TBE DSLOC amounts from their associated equations in order to isolate the portions of growth-adjusted DSLOC that represent the amount of DSLOC growth.

$$\begin{aligned}
\boldsymbol{S}_{\boldsymbol{DGAmountNewBL}} &\triangleq \tilde{b}_{GN}\boldsymbol{\varepsilon}_{\boldsymbol{GN}} \left(\frac{S_{DNew}}{K_N}\right)^{a_{GN}} K_N - S_{DNew} \\
\boldsymbol{S}_{\boldsymbol{DGAmountModBL}} &\triangleq \tilde{b}_{GM}\boldsymbol{\varepsilon}_{\boldsymbol{GM}} \left(\frac{S_{DMod}}{K_M}\right)^{a_{GM}} K_M - S_{DMod} \qquad \textbf{(41)} \\
\boldsymbol{S}_{\boldsymbol{DGAmountUmodBL}} &\triangleq \tilde{b}_{GU}\boldsymbol{\varepsilon}_{\boldsymbol{GU}} \left(\frac{S_{DUmod}}{K_U}\right)^{a_{GU}} K_U - S_{DUmod}
\end{aligned}$$

Assembling the Model Components

Maturity-Adjusted Growth Amount

We begin with the baseline (SDLCBegin) DSLOC estimate growth amount distributions of Equations (41), which have been developed from historical data and which model the amount of uncertainty that exists about the TBEs of New, Modified, and Unmodified DSLOC. We assume that these estimates are done at the beginning of the SDLC; i.e., $Maturity = 0$, consistent with the SDLCs from which the historical data were collected.

Suppose these baseline distributions are represented as CDFs; i.e., mappings of growth amount values to probability of attainment percentages. We would like to model what happens to the uncertainty modeled by these distributions as activities in the SDLC progress to completion and TBEs are updated to reflect the evolving knowledge. We have already hypothesized that uncertainty decays over time and have developed a model for this decay in Equation (5). Since the maturity adjustment factor function of Equation (5) is normalized (i.e., yields factors that are percentages of full scale), we can use this factor to scale our baseline estimated growth amounts of New, Modified, and Unmodified DSLOC to reflect the estimate maturity of subsequent TBEs.

Equations (41) represent the baseline estimated growth amounts of New, Modified, and Unmodified DSLOC. We wish to adjust these amounts as a function of estimate Maturity (i.e., when in the SDLC the estimate is rendered). We accomplish this by multiplying these amounts by the maturity adjustment factor function Equation (5) to yield the expressions

$$\begin{aligned} &\boldsymbol{f}\left(Maturity\right)\boldsymbol{S}_{\boldsymbol{DGAmountNewBL}} \\ &\boldsymbol{f}\left(Maturity\right)\boldsymbol{S}_{\boldsymbol{DGAmountModBL}} \\ &\boldsymbol{f}\left(Maturity\right)\boldsymbol{S}_{\boldsymbol{DGAmountUmodBL}} \end{aligned} \tag{42}$$

We can expand the three Expressions (42) by substituting each factor with its equivalent in Equations (5) and (41) respectively, which gives us the maturity-adjusted growth amounts

$$\begin{aligned} &\underline{\underline{e^{-(Decay)(Maturity)}\left(\tilde{b}_{GN}\boldsymbol{\varepsilon}_{\boldsymbol{GN}}\left(\frac{S_{DNew}}{K_N}\right)^{a_{GN}}K_N - S_{DNew}\right)}} \\ &\underline{\underline{e^{-(Decay)(Maturity)}\left(\tilde{b}_{GM}\boldsymbol{\varepsilon}_{\boldsymbol{GM}}\left(\frac{S_{DMod}}{K_M}\right)^{a_{GM}}K_M - S_{DMod}\right)}} \\ &\underline{\underline{e^{-(Decay)(Maturity)}\left(\tilde{b}_{GU}\boldsymbol{\varepsilon}_{\boldsymbol{GU}}\left(\frac{S_{DUmod}}{K_U}\right)^{a_{GU}}K_U - S_{DUmod}\right)}} \end{aligned} \tag{43}$$

Notice that the estimate maturity factor portion of Expressions (43) serves to decay (exponentially decrease) the estimated DSLOC growth amount distributions. The practical effect of applying this decay is time-progressive compression of the effective DSLOC growth factor distributions about the TBE position approaching no uncertainty at SDLC completion (SWAccept, $Maturity = 100\%$).

As stated earlier, in order to render Expressions (43) useful in a particular estimating situation, we need to assume some value (or distribution) for the decay constant $Decay$ in Expressions (43); either by assuming $Decay = 3.466$ (Boehm's *Cone of Uncertainty*) or by analyzing relevant historical data to model decay as a single value $Decay$ or as a distribution $\boldsymbol{Decay}$.

Total Growth-Adjusted DSLOC Estimate Amounts

We now simply add the TBEs for New, Modified, and Unmodified DSLOC to each of the Expressions (43) to produce the final form DEGM8 equations:

$$\begin{aligned}
&\underline{\underline{\boldsymbol{S_{DGANew}} \triangleq S_{DNew} + e^{-(Decay)(Maturity)}\left(\tilde{b}_{GN}\boldsymbol{\varepsilon_{GN}}\left(\frac{S_{DNew}}{K_N}\right)^{a_{GN}} K_N - S_{DNew}\right)}}\\
&\underline{\underline{\boldsymbol{S_{DGAMod}} \triangleq S_{DMod} + e^{-(Decay)(Maturity)}\left(\tilde{b}_{GM}\boldsymbol{\varepsilon_{GM}}\left(\frac{S_{DMod}}{K_M}\right)^{a_{GM}} K_M - S_{DMod}\right)}}\\
&\underline{\underline{\boldsymbol{S_{DGAUmod}} \triangleq S_{DUmod} + e^{-(Decay)(Maturity)}\left(\tilde{b}_{GU}\boldsymbol{\varepsilon_{GU}}\left(\frac{S_{DUmod}}{K_U}\right)^{a_{GU}} K_U - S_{DUmod}\right)}}
\end{aligned} \qquad \textbf{(44)}$$

Note that these equations match those of Figure 1.

Since each of the right-side expressions of Equations (44) represent total growth-adjusted DSLOC amounts, we can divide each expression by its associated TBE DSLOC amount to create associated expressions that represent implied growth factors which we can use as a metric to describe the effective growth being applied by the model to the TBEs of New, Modified, and Unmodified DSLOC:

$$\begin{aligned}
&\left(S_{DNew} + e^{-(Decay)(Maturity)}\left(\tilde{b}_{GN}\boldsymbol{\varepsilon_{GN}}\left(\frac{S_{DNew}}{K_N}\right)^{a_{GN}} K_N - S_{DNew}\right)\right)\Big/ S_{DNew}\\
&\left(S_{DMod} + e^{-(Decay)(Maturity)}\left(\tilde{b}_{GM}\boldsymbol{\varepsilon_{GM}}\left(\frac{S_{DMod}}{K_M}\right)^{a_{GM}} K_M - S_{DMod}\right)\right)\Big/ S_{DMod}\\
&\left(S_{DUmod} + e^{-(Decay)(Maturity)}\left(\tilde{b}_{GU}\boldsymbol{\varepsilon_{GU}}\left(\frac{S_{DUmod}}{K_U}\right)^{a_{GU}} K_U - S_{DUmod}\right)\right)\Big/ S_{DUmod}
\end{aligned} \qquad \textbf{(45)}$$

Correlation is Necessary to Properly Add Random Variables

Equations (44) yield random variables for each of growth-adjusted New, Modified, and Unmodified DSLOC. It is likely, as part of an overall estimation process of which the DEGM8 is a part, that there will be a need to define some random variable as the sum of $\boldsymbol{S_{DGANew}}$, $\boldsymbol{S_{DGAModified}}$, and $\boldsymbol{S_{DGAUnmodified}}$ (i.e., total DSLOC) or to define some random variable as the sum of random variables that are functions of $\boldsymbol{S_{DGANew}}$,

$S_{DGAModified}$, and $S_{DGAUnmodified}$. The process of summing random variables is called *convolution*[13], the mathematics of which are beyond the scope of this paper. However, several commercially available software tools provide the capability to sum both independent (uncorrelated) and dependent (correlated) random variables. TRI's ACEIT is one such software tool; convolution is performed using Monte Carlo simulation techniques and correlation is applied using an internal algorithm that requires providing pairwise correlation between the summands. The DEGM8 provides this pairwise correlation in the form of a correlation matrix that is developed as part of the overall historical data regression process. Table 3 shows the correlation matrix for the default instance of the DEGM8. The three values from this matrix necessary as inputs to ACEIT in order to properly convolve growth-adjusted New, Modified, and Unmodified DSLOC are:

$$\begin{aligned} \boldsymbol{correl}\left(b_{GN}, b_{GM}\right) &\equiv 2.56966251E-03 \\ \boldsymbol{correl}\left(b_{GN}, b_{GU}\right) &\equiv 3.02466226E-01 \\ \boldsymbol{correl}\left(b_{GM}, b_{GU}\right) &\equiv 7.46696118E-02 \end{aligned} \tag{46}$$

Table 3 Correlation matrix from JCDER349 regression (DEGM8 default instance)

Correlation Matrix			
	New DSLOC	Modified DSLOC	Unmodified DSLOC
New DSLOC	1		
Modified DSLOC	2.569663E-03	1	
Unmodified DSLOC	3.024662E-01	7.466961E-02	1

4. DEGM8 Example Application

In this section of the paper we describe an example of how we can use the default version of the DEGM8 to adjust TBEs of New, Modified, and Unmodified DSLOC to develop growth-adjusted probabilistic estimates of what we expect these values to be when the software is done and accepted. Appendix E contains a description and tables that show how this example can be implemented in TRI's ACEIT software tool.

Ground Rules and Assumptions for the Example

For this example, we are given the task of providing a growth-adjusted probabilistic size estimate for the Navigation (NAV) Function Computer Software Configuration Item (CSCI) that is part of an unmanned satellite ground control system. The following list represents the Ground Rules and Assumptions (GRAs) that describe the scenario we will be using:

- The given TBEs for New, Modified, and Unmodified software size are 25,000 DSLOC, 50,000 DSLOC, and 100,000 DSLOC respectively.
- We assume no redelivered code from a previous increment, iteration, or release of the NAV function.
- Since the NAV function is a CSCI; i.e., it is of a size and scope that is normally associated with a single observation in the SRDR database, we assume $K[N] \equiv 1$, $K[M] \equiv 1$, and $K[U] \equiv 1$; i.e., normalization of the TBEs to the historical data is unnecessary.
- We assume that the TBEs for New, Modified, and Unmodified DSLOC are rendered at the completion of Software Requirements Review (SwRR); therefore, $Maturity \equiv 20\%$ per Table 1.
- We assume $Decay \equiv 3.466$ based on Boehm's (1981 pp. 310-311) *Cone of Uncertainty* as is specified for the default instance of the DEGM8.
- We assume the DEGM8 default values for the exponent parameters associated with each of New, Modified, and Unmodified DSLOC as $a[GN] \equiv 1.021$, $a[GM] \equiv 0.913$, and $a[GU] \equiv 1.044$ per the second page of Appendix B.
- We assume the DEGM8 default values for the geometric means of the scale factor parameters associated with each of New, Modified, and Unmodified DSLOC as $\tilde{b}[GN] \equiv 1.208$, $\tilde{b}[GM] \equiv 2.651$, and $\tilde{b}[GU] \equiv 0.6199$ per the second page of Appendix B.
- We assume the DEGM8 default correlation values shown in Equations (46).

Instantiating the DEGM Equations

Given the GRAs listed above, we can instantiate the DEGM equations in Figure 1 as

$$\begin{Bmatrix} \boldsymbol{S_{DGANew}} \triangleq S_{DNew} + e^{-(Decay)(Maturity)}\left(\tilde{b}_{GN}\boldsymbol{\varepsilon_{GN}}\left(\frac{S_{DNew}}{K_N}\right)^{a_{GN}} K_N - S_{DNew}\right) \\ \boldsymbol{S_{DGAMod}} \triangleq S_{DMod} + e^{-(Decay)(Maturity)}\left(\tilde{b}_{GM}\boldsymbol{\varepsilon_{GM}}\left(\frac{S_{DMod}}{K_M}\right)^{a_{GM}} K_M - S_{DMod}\right) \\ \boldsymbol{S_{DGAUmod}} \triangleq S_{DUmod} + e^{-(Decay)(Maturity)}\left(\tilde{b}_{GU}\boldsymbol{\varepsilon_{GU}}\left(\frac{S_{DUmod}}{1}\right)^{a_{GU}} K_U - S_{DUmod}\right) \end{Bmatrix}$$

$$\rightarrow \begin{Bmatrix} \boldsymbol{S_{DGANew}} \triangleq 25,000 + e^{-(3.466)(20\%)}\left((1.208)\boldsymbol{\varepsilon_{GN}}\left(\frac{25,000}{1}\right)^{1.021}(1) - 25,000\right) \\ \boldsymbol{S_{DGAMod}} \triangleq 50,000 + e^{-(3.466)(20\%)}\left((2.651)\boldsymbol{\varepsilon_{GM}}\left(\frac{50,000}{1}\right)^{0.913}(1) - 50,000\right) \\ \boldsymbol{S_{DGAUmod}} \triangleq 100,000 + e^{-(3.466)(20\%)}\left((0.6199)\boldsymbol{\varepsilon_{GU}}\left(\frac{100,000}{1}\right)^{1.044}(1) - 100,000\right) \end{Bmatrix} \quad \textbf{(47)}$$

Figure 4, Figure 5, and Figure 6 below illustrate the behaviors of the resulting growth-adjusted New DSLOC, Modified DSLOC, and Unmodified DSLOC estimate distributions as described in Equations (47) for the given TBEs and estimate *Maturity* .

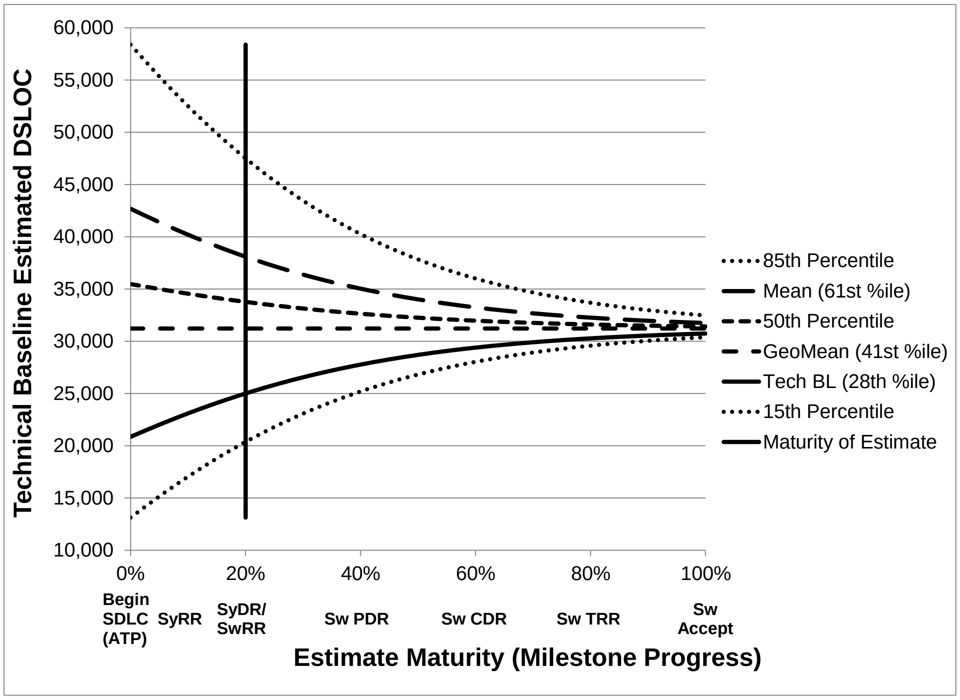


Figure 4: Example growth-adjusted New DSLOC distribution vs. estimate maturity

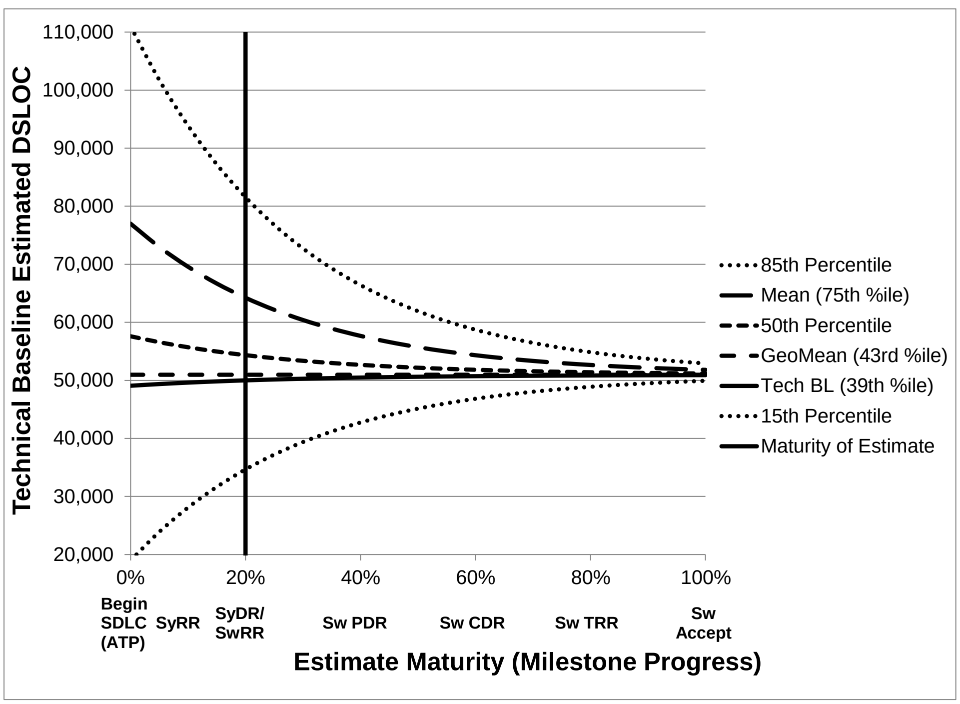


Figure 5: Example growth-adjusted Modified DSLOC distribution vs. estimate maturity

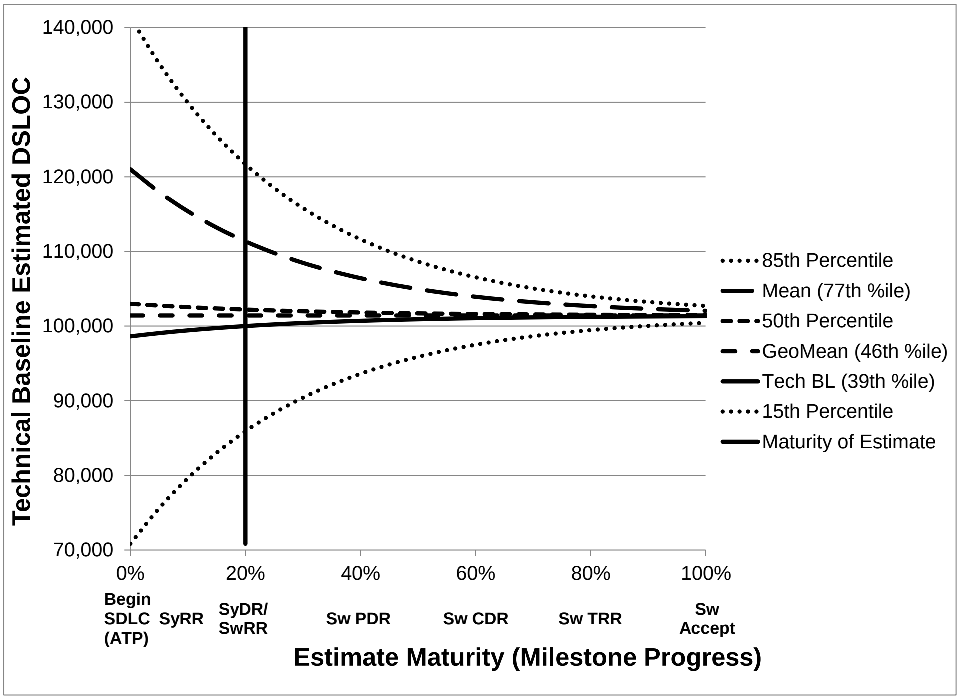


Figure 6: Example growth-adjusted Unmodified DSLOC distribution vs. estimate maturity

Implied Growth Factors as the Estimates Mature

We can instantiate the implied growth factor Expressions (45) with values from our example's GRAs to get

$$
\left\{
\begin{array}{l}
\left(S_{DNew} + e^{-(Decay)(Maturity)}\left(\tilde{b}_{GN}\boldsymbol{\varepsilon}_{\mathbf{GN}}\left(\dfrac{S_{DNew}}{K_N}\right)^{a_{GN}} K_N - S_{DNew}\right)\right) \Big/ S_{DNew} \\
\left(S_{DMod} + e^{-(Decay)(Maturity)}\left(\tilde{b}_{GM}\boldsymbol{\varepsilon}_{\mathbf{GM}}\left(\dfrac{S_{DMod}}{K_M}\right)^{a_{GM}} K_M - S_{DMod}\right)\right) \Big/ S_{DMod} \\
\left(S_{DUmod} + e^{-(Decay)(Maturity)}\left(\tilde{b}_{GU}\boldsymbol{\varepsilon}_{\mathbf{GU}}\left(\dfrac{S_{DUmod}}{1}\right)^{a_{GU}} K_U - S_{DUmod}\right)\right) \Big/ S_{DUmod}
\end{array}
\right\}
$$

$$
\rightarrow
\left\{
\begin{array}{l}
\left(25,000 + e^{-(3.466)(Maturity)}\left((1.208)\,\boldsymbol{\varepsilon}_{\mathbf{GN}}\left(\dfrac{25,000}{1}\right)^{1.021}(1) - 25,000\right)\right) \Big/ 25,000 \\
\left(50,000 + e^{-(3.466)(Maturity)}\left((2.651)\,\boldsymbol{\varepsilon}_{\mathbf{GM}}\left(\dfrac{50,000}{1}\right)^{0.913}(1) - 50,000\right)\right) \Big/ 50,000 \\
\left(100,000 + e^{-(3.466)(Maturity)}\left((0.6199)\,\boldsymbol{\varepsilon}_{\mathbf{GU}}\left(\dfrac{100,000}{1}\right)^{1.044}(1) - 100,000\right)\right) \Big/ 100,000
\end{array}
\right\}
\quad \mathbf{(48)}
$$

We can use Expressions (48) to solve for the New, Modified, and Unmodified DSLOC growth factors implied by the DEGM8 as a function of increasing estimate $Maturity$. Figure 7, Figure 8, and Figure 9 below illustrate this behavior of over the range of possible estimate $Maturity$ values $Maturity \in [0\%, 100\%]$.

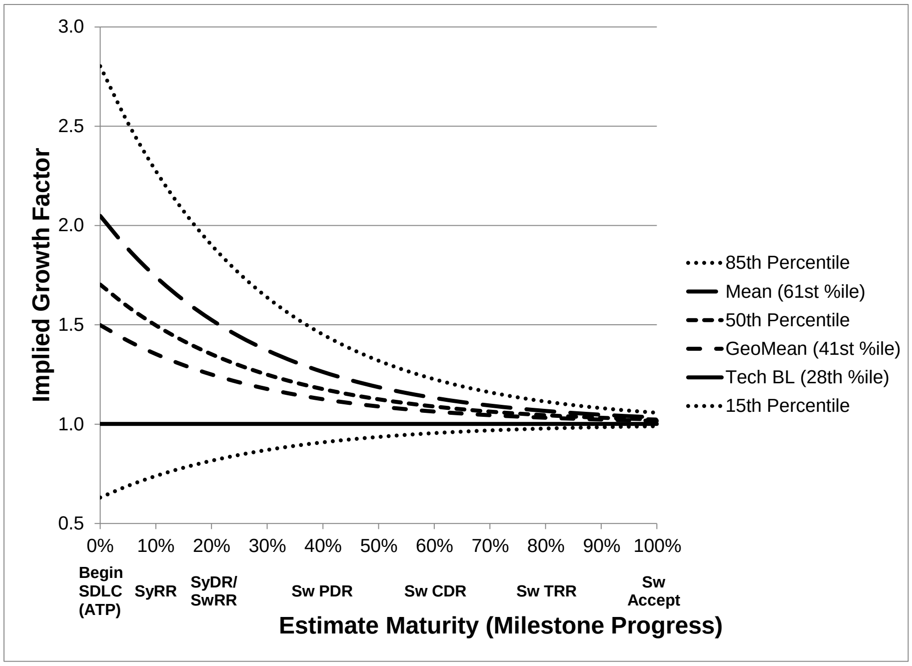


Figure 7 New DSLOC implied growth-factor decay; the implied growth factor for a TBE of 25,000 New DSLOC at SwRR ($Maturity = 20\%$) is 1.52 (52%) at the arithmetic mean (61st percentile) and 1.25 (25%) at the geometric mean (41st percentile)

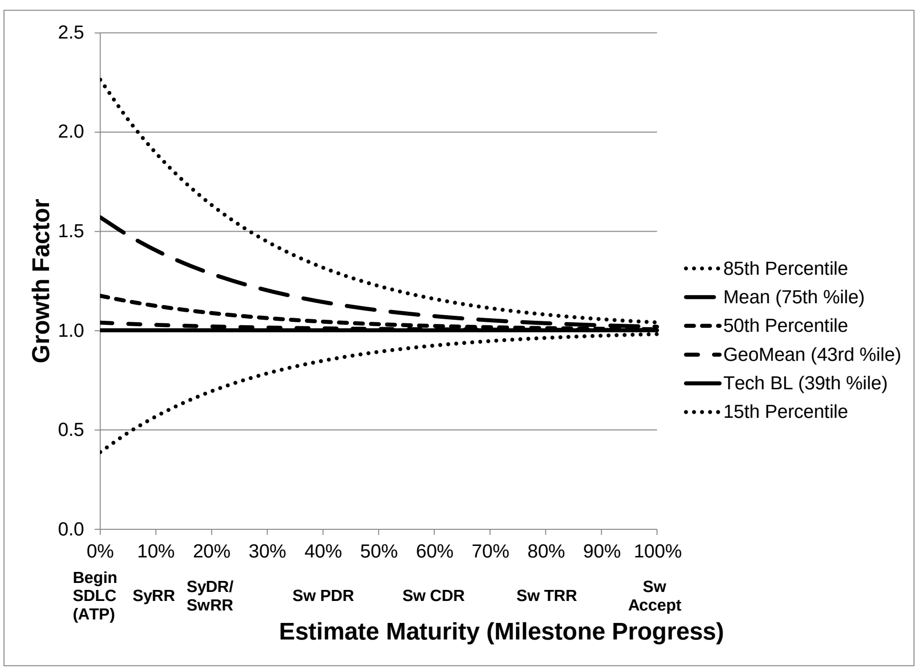


Figure 8 Modified DSLOC implied growth-factor decay; the implied growth factor for a TBE of 50,000 Modified DSLOC at SwRR ($Maturity = 20\%$) is 1.28 (28%) at the arithmetic mean (75th percentile) and 1.02 (2%) at the geometric mean (43rd percentile)

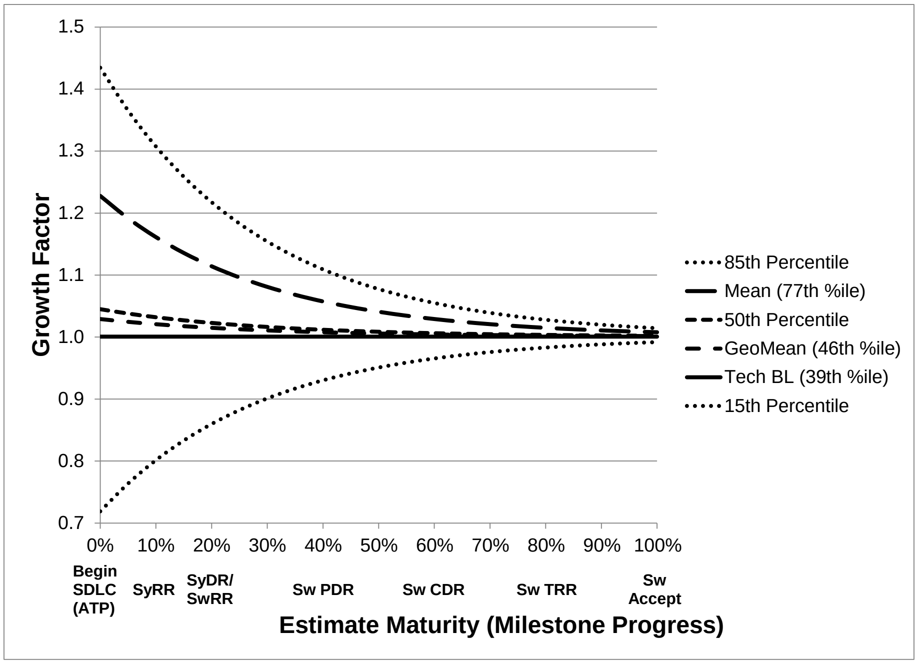


Figure 9 Unmodified DSLOC implied growth-factor decay; the implied growth factor for a TBE of 100,000 Unmodified DSLOC at SwRR ($Maturity = 20\%$) is 1.11 (11%) at the arithmetic mean (77th percentile) and 1.01 (1%) at the geometric mean (46th percentile)

5. A Special Instance of DEGM8 for Estimating Unmanned Spacecraft Flight Software

A particular challenge for USAF Space and Missile Systems Center (SMC) is estimating the size of unmanned Space Vehicle (SV) embedded software (bus software and payload software). One thing that makes this a challenge is the lack of readily-available relevant historical data. The SRDR database, as of this writing, contains very few observations of this kind of software. Apparently, Government Contractors have typically not been required to furnish SV software measurement and metric data to the Government in the form of submitted SRDRs.

SMC with assistance from Tecolote Research, Inc. has collected some relevant historical DSLOC estimates and final-actual DSLOC results for several space programs in order to improve on existing growth, cost, and duration estimates for unmanned SV software; however, the estimates were performed at various points in the SDLC and most of the programs they represent did not perform DSLOC estimates at SDLCBegin. As such, this data is insufficient to develop Growth Estimating Relationships (GERs) using the approach that was taken for the default instance of the DEGM8. The remainder of this section describes a new and unique variant approach and resulting GERs that can be used for SV programs to perform growth-adjusted New, Modified, and Unmodified DSLOC estimates as a function of TBE New, Modified, and Unmodified

DSLOC size; and estimate maturity. This model variant is henceforth referred to as DEGM8SV and is based on relevant data collected from programs where there exists a final-actual DSLOC result and at least one DSLOC estimate along with its corresponding estimate maturity.

Once again, for the sake of economy, we focus on New DSLOC and assume that the same process will hold for Modified and Unmodified DSLOC.

Identify the Desired Functional Form

We start by recognizing the fact that we have three unique data fields for each historical observation: S_{DNEst}, S_{DNAct}, and $Maturity_{GN}$. Since we wish to maturity-adjust only the growth portion of a TBE and not the entire estimate, we have created a new implied growth factor field $S_{DNGF} \equiv \frac{S_{DNAct}}{S_{DNEst}}$ that represents the amount of growth yet to be realized. It is to this field that we can multiply the decaying estimate $Maturity$ adjustment factor $e^{-(Decay_{GN})(Maturity_{GN})}$ that we developed earlier in Equation (5).

Using the above observations and assumptions, we propose the following multiplicative combination with which we will derive and specify our GERs for New, Modified, and Unmodified DSLOC.

$$S_{DNEst}{}^{c_1} S_{DNGF}{}^{c_2} e^{c_3(Maturity_{GN})} c_4 \equiv 1 \;\Big|\; S_{DNGF} \quad \frac{S_{DNAct}}{S_{DNEst}} \qquad \textbf{(49)}$$

Normalize the Measures to Ensure Scale Invariance and Commensurability

In order to ensure scale invariance and measurement commensurability we sigma-normalize each of the historical data N-element lists $\mathbf{S_{DNEst}}$, $\mathbf{S_{DNGF}}$, and $\mathbf{Maturity_{GN}}$.

$$\mathbf{S}'_{\mathbf{DNEst}} \equiv \frac{\mathbf{S}_{\mathbf{DNEst}}}{\sigma_{\mathbf{S}_{\mathbf{DNEst}}}} = \frac{\mathbf{S}_{\mathbf{DNEst}}}{\boldsymbol{stdev}\left(\mathbf{S}_{\mathbf{DNEst}}\right)}$$

$$\left.\frac{S_{DNEst_j}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(S_{DNEst_i} - \mu_{\mathbf{S}_{\mathbf{DNEst}}}\right)^2}}\right|_{j=1}^{N} \quad \left|\mu_{\mathbf{S}_{\mathbf{DNEst}}} \equiv \frac{1}{N}\sum_{k=1}^{N} S_{DNEst_k}\right.$$

$$\mathbf{S}'_{\mathbf{DNGF}} \equiv \frac{\boldsymbol{ln}\left(\mathbf{S}_{\mathbf{DNGF}}\right)}{\sigma_{\boldsymbol{ln}\left(\mathbf{S}_{\mathbf{DNGF}}\right)}} = \frac{\boldsymbol{ln}\left(\mathbf{S}_{\mathbf{DNGF}}\right)}{\boldsymbol{stdev}\left(\boldsymbol{ln}\left(\mathbf{S}_{\mathbf{DNGF}}\right)\right)}$$

$$\left.\frac{S_{DNGF_j}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(S_{DNGF_i} - \mu_{\mathbf{S}_{\mathbf{DNGF}}}\right)^2}}\right|_{j=1}^{N} \quad \left|\mu_{\mathbf{S}_{\mathbf{DNGF}}} \equiv \frac{1}{N}\sum_{k=1}^{N} S_{DNGF_k}\right.$$

$$\mathbf{Maturity}'_{\mathbf{GN}} \equiv \frac{\mathbf{Maturity}_{\mathbf{GN}}}{\sigma_{\mathbf{Maturity}_{\mathbf{GN}}}} = \frac{\mathbf{Maturity}_{\mathbf{GN}}}{\boldsymbol{stdev}\left(\mathbf{Maturity}_{\mathbf{GN}}\right)}$$

$$\left.\frac{Maturity_{GN_j}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{ln}\left(Maturity_{GN_i}\right) - \mu_{\mathbf{Maturity}_{\mathbf{GN}}}\right)^2}}\right|_{j=1}^{N} \quad \left|\mu_{\mathbf{Maturity}_{\mathbf{GN}}} \equiv \frac{1}{N}\sum_{k=1}^{N} Maturity_{GN_k}\right. \quad \mathbf{(50)}$$

Log Transform the Data to Make the Relationship Linear

$$\begin{gathered}\boldsymbol{g}(\boldsymbol{f}) \equiv \boldsymbol{ln}\left(\boldsymbol{f}\left(S'_{DNEst}, S'_{DNAct}, e^{\left(Maturity'_{GN}\right)}\right)\right) = \boldsymbol{ln}\left(S'^{\,c_1}_{DNEst} S'^{\,c_2}_{DNGF} e^{c_3\left(Maturity'_{GN}\right)} c_4\right) \\ \therefore \boldsymbol{g}(\boldsymbol{f}) = c_1\,\boldsymbol{ln}\left(S'_{DNEst}\right) \quad c_2\,\boldsymbol{ln}\left(S'_{DNGF}\right) \quad c_3\left(Maturity'_{GN}\right) \quad \boldsymbol{ln}\left(c_4\right)\end{gathered} \quad \mathbf{(51)}$$

Organize the Historical Data as a Matrix

$$\boldsymbol{ln}'(\mathbf{P}) \equiv \begin{bmatrix} \boldsymbol{ln}\left(S'_{DNEst_1}\right) & \boldsymbol{ln}\left(S'_{DNGF_1}\right) & Maturity'_{GN_1} \\ \boldsymbol{ln}\left(S'_{DNEst_2}\right) & \boldsymbol{ln}\left(S'_{DNGF_2}\right) & Maturity'_{GN_1} \\ \vdots & \vdots & \vdots \\ \boldsymbol{ln}\left(S'_{DNEst_N}\right) & \boldsymbol{ln}\left(S'_{DNGF_N}\right) & Maturity'_{GN_N} \end{bmatrix} \quad \mathbf{(52)}$$

Define the ODR Best Fit Line

$$L_{ODR} \equiv \begin{cases} P_{ODR_{S_{DNEst}}} = P'_{ODR_{S_{DNEst}}} + ta_{S_{DNEst}} \\ P_{ODR_{S_{DNGF}}} = P'_{ODR_{S_{DNGF}}} + ta_{S_{DNGF}} \\ P_{ODR_{Maturity_{GN}}} = P'_{ODR_{Maturity_{GN}}} + ta_{Maturity_{GN}} \end{cases} \tag{53}$$

Find the Data Set Centroid Point Coordinates

$$\begin{pmatrix} C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}} \equiv \boldsymbol{average}\left(\boldsymbol{ln}\left(\mathbf{S}'_{\mathbf{DNEst}}\right)\right) = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{ln}\left(S'_{DNEst_i}\right), \\ C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNGF}}} \equiv \boldsymbol{average}\left(\boldsymbol{ln}\left(\mathbf{S}'_{\mathbf{DNGF}}\right)\right) = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{ln}'\left(S'_{DNGF_i}\right), \\ C_{\boldsymbol{ln}(\mathbf{P})_{Maturity_{GN}}} \equiv \boldsymbol{average}\left(\mathbf{Maturity}'_{\mathbf{GN}}\right) = \frac{1}{N}\sum_{i=1}^{N} Maturity'_{GN_i} \end{pmatrix} \tag{54}$$

Instantiate the ODR Best Fit Line with the Data Set Centroid

$$L_{ODR} \equiv \begin{cases} P_{ODR_{S_{DNEst}}} = C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}} + ta_{S_{DNEst}} \\ P_{ODR_{S_{DNGF}}} = C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNGF}}} + ta_{S_{DNGF}} \\ P_{ODR_{Maturity_{GN}}} = C_{\mathbf{P}_{Maturity_{GN}}} + ta_{Maturity_{GN}} \end{cases} \tag{55}$$

Center the Data Set Matrix about the Coordinate System Origin

$$\mathbf{M} \equiv \begin{bmatrix} \boldsymbol{ln}\left(S'_{DNEst_1}\right) - C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}} & \boldsymbol{ln}\left(S'_{DNGF_1}\right) - C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNGF}}} & Maturity'_{GN_1} - C_{\mathbf{P}_{Maturity_{GN}}} \\ \boldsymbol{ln}\left(S'_{DNEst_2}\right) - C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}} & \boldsymbol{ln}\left(S'_{DNGF_2}\right) - C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNGF}}} & Maturity'_{GN_2} - C_{\mathbf{P}_{Maturity_{GN}}} \\ \vdots & \vdots & \vdots \\ \boldsymbol{ln}\left(S'_{DNEst_N}\right) - C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}} & \boldsymbol{ln}\left(S'_{DNGF_N}\right) - C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNGF}}} & Maturity'_{GN_N} - C_{\mathbf{P}_{Maturity_{GN}}} \end{bmatrix} \tag{56}$$

Apply the SVD to the Centered Data Set Matrix

$$\boldsymbol{SVD}(\mathbf{M}) \equiv \left\{\mathbf{U}, \boldsymbol{\Sigma}, \mathbf{V}^{\mathsf{T}}\right\} \;\middle|\; \mathbf{M} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathsf{T}} \tag{57}$$

$$\boldsymbol{SVD}(\mathbf{M}) \equiv \{\mathbf{U}, \boldsymbol{\Sigma}, \mathbf{V}^{\mathsf{T}}\} \;\middle|\; \mathbf{M} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathsf{T}}$$

$$\mathbf{M} = \mathbf{U}\begin{bmatrix} \Sigma_{1,1} & 0 & 0 \\ 0 & \Sigma_{2,2} & 0 \\ 0 & 0 & \Sigma_{3,3} \end{bmatrix}\begin{bmatrix} V_{1,1} & V_{1,2} & V_{1,3} \\ V_{2,1} & V_{2,2} & V_{2,3} \\ V_{3,1} & V_{3,2} & V_{3,3} \end{bmatrix}^{\mathsf{T}} \tag{58}$$

$$\mathbf{M} = \mathbf{U}\begin{bmatrix} \Sigma_{max} & 0 & 0 \\ 0 & \Sigma_{2,2} & 0 \\ 0 & 0 & \Sigma_{3,3} \end{bmatrix}\begin{bmatrix} a_{S_{DNEst}} & V_{1,2} & V_{1,3} \\ a_{S_{DNGF}} & V_{2,2} & V_{2,3} \\ a_{Maturity_{GN}} & V_{3,2} & V_{3,3} \end{bmatrix}^{\mathsf{T}}$$

Define the Growth Estimating Relationship Hyperplane

We now introduce and define the notion of a growth estimating relationship (GER) *hyperplane*. We define a GER hyperplane H_{GER} as a 2-dimension flat subspace of $\mathbb{R}^3$ (a plane in this case) that is orthogonal to the ODR best fit line L_{ODR} and that passes through a point on L_{ODR} regarded as a fundamental solution P_{GER} on L_{ODR}. The importance of a GER hyperplane is that it contains all the points in $\mathbb{R}^3$ that orthogonally project onto L_{ODR} at P_{GER}; i.e., P_{GER} is the *fundamental solution* for every point on H_{GER}. *This implies that the GER hyperplane H_{GER} contains all the potential outcome coordinate combinations that are best estimated by the coordinate combination represented by point P_{GER}.*

The nature of our desired GER is such that we wish to structure the GER hyperplane equation so that the final-actual estimate of New DSLOC growth amount S_{DNGAmt} is *dependent* on the TBE of New DSLOC S_{DNew}. Note that in estimating situations we are not trying to solve for S_{DNew} but rather are given S_{DNew} as a single value based on when in the SDLC the estimate is being rendered; i.e., its estimate $Maturity$. We instantiate the variables in the general scalar equation form of a hyperplane

Theorem: Scalar or Implicit Form Equation of a Hyperplane in m-Dimension Space

$$\begin{aligned} &\mathbf{n}\bullet\mathbf{r} + d = 0 \\ &n_1 r_1 + n_2 r_2 + \ldots + n_m r_m + d = 0 \end{aligned} \tag{59}$$

where

$$\begin{aligned} d &\equiv -\mathbf{n}\bullet\mathbf{r}' \\ &= -n_1 r_1' - n_2 r_2' - \ldots - n_m r_m' \end{aligned}$$

using the following observations and assumptions, the result being an equation describing H_{GER}.

- Since the best fit GER hyperplane is orthogonal to our ODR best fit line and since a hyperplane is partially described by a vector that is orthogonal (normal) to that hyperplane, we let the direction vector $\mathbf{a}$ from the definition of our ODR best fit line L_{ODR} also be the normal vector $\mathbf{n}$ of our JCDER hyperplane H_{JCDER}. Therefore, $\mathbf{n} \equiv \mathbf{a}$.
- The GER hyperplane H_{GER} must include the fundamental solution point P_{GER}. We therefore use P_{GER} as our known point on H_{GER}, the position vector of which is represented in Equation (59) as $\mathbf{r}'$. Since P_{GER} must lie on our ODR best fit line L_{ODR} we can instantiate Equations (55) as

$$L_{ODR} \equiv \begin{cases} r_1' = C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}} + t a_{S_{DNEst}} \\ r_2' = C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNGF}}} + t a_{S_{DNGF}} \\ r_3' = C_{\mathbf{P}_{Maturity_{GN}}} + t a_{Maturity_{GN}} \end{cases} \tag{60}$$

Since P_{GER} must also be dependent on the given value for TBE New DSLOC it follows that the P_{GER} position vector $\boldsymbol{ln}\left(S_{DNEst}'\right)$ component must be

$$r_1' \equiv \boldsymbol{ln}\left(S_{DNew}'\right) \tag{61}$$

Substituting r_1' in the first of Equations (60) with its equivalent in Equation (61) yields

$$\boldsymbol{ln}\left(S_{DNew}'\right) = C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}} + t a_{S_{DNEst}} \tag{62}$$

Solving for the parameter t in Equation (62) we get

$$t = \frac{\boldsymbol{ln}\left(S_{DNew}'\right) - C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}}}{a_{S_{DNEst}}} \tag{63}$$

Substituting each instance of t in Equations (60) with its equivalent in Equation (63) gives us a New DSLOC TBE-constrained specification of the fundamental solution point P_{GER} and therefore the components of the GER hyperplane known point's position vector $\mathbf{r}'$.

$$r_1' = \boldsymbol{ln}\left(S'_{DNew}\right)$$
$$r_2' = C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNGF}}} + \frac{\boldsymbol{ln}\left(S'_{DNew}\right) - C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}}}{a_{S_{DNEst}}} a_{S_{DNGF}} \qquad \textbf{(64)}$$
$$r_3' = C_{\mathbf{P}_{Maturity_{GN}}} + \frac{\boldsymbol{ln}\left(S'_{DNew}\right) - C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}}}{a_{S_{DNEst}}} a_{Maturity_{GN}}$$

- Any potential sigma-normalized log-transformed estimate $\left(\boldsymbol{ln}\left(S'_{DNew}\right), \boldsymbol{ln}\left(S'_{DNGF}\right), Maturity'_{GN}\right)$ must lie somewhere on our GER hyperplane. Therefore, $r_1 \equiv \boldsymbol{ln}\left(S'_{DNew}\right)$, $r_2 \equiv \boldsymbol{ln}\left(S'_{DNGF}\right)$, and $r_3 \equiv Maturity'_{GN}$.

This gives us the GER hyperplane

$$a_{S_{DNEst}} \boldsymbol{ln}\left(S'_{DNew}\right) + a_{S_{DNGAmt}} \boldsymbol{ln}\left(S'_{DNGAmt}\right) + a_{Maturity_{GN}} Maturity'_{GN} + d = 0 \qquad \textbf{(65)}$$

where

$$d = \left(\begin{array}{l} -a_{S_{DNEst}} \boldsymbol{ln}\left(S'_{DNew}\right) - \\ a_{S_{DNGF}} \left(C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNGF}}} + \frac{\boldsymbol{ln}\left(S'_{DNew}\right) - C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}}}{a_{S_{DNEst}}} a_{S_{DNGF}} \right) \\ a_{Maturity_{GN}} \left(C_{\mathbf{P}_{Maturity_{GN}}} + \frac{\boldsymbol{ln}\left(S'_{DNew}\right) - C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}}}{a_{S_{DNEst}}} a_{Maturity_{GN}} \right) \end{array} \right) \qquad \textbf{(66)}$$

Next, we substitute d in Equation (65) with its equivalent in Equation (66) to get

$$a_{S_{DNEst}} \boldsymbol{ln}\left(S'_{DNew}\right) + a_{S_{DNGF}} \boldsymbol{ln}\left(S'_{DNGF}\right) + a_{Maturity_{GN}} Maturity'_{GN} +$$
$$\left(\begin{array}{l} -a_{S_{DNEst}} \boldsymbol{ln}\left(S'_{DNew}\right) - \\ a_{S_{DNGF}} \left(C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNGF}}} + \frac{\boldsymbol{ln}\left(S'_{DNew}\right) - C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}}}{a_{S_{DNEst}}} a_{S_{DNGF}} \right) - \\ a_{Maturity_{GN}} \left(C_{\mathbf{P}_{Maturity_{GN}}} + \frac{\boldsymbol{ln}\left(S'_{DNew}\right) - C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}}}{a_{S_{DNEst}}} a_{Maturity_{GN}} \right) \end{array} \right) = 0 \qquad \textbf{(67)}$$

Algebraically rearranging the factors and terms of Equation (67) gives us

$$
\begin{aligned}
& a_{S_{DNGF}}\, \boldsymbol{ln}\left(S'_{DNGF}\right) = \\
& \left(\frac{a_{S_{DNGF}}{}^{2} + a_{Maturity_{GN}}{}^{2}}{a_{S_{DNEst}}}\right) \boldsymbol{ln}\left(S'_{DNew}\right) - \\
& \left(\begin{array}{l} \left(\dfrac{a_{S_{DNGF}}{}^{2} + a_{Maturity_{GN}}{}^{2}}{a_{S_{DNEst}}}\right) C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}} + \\ a_{S_{DNGF}}\, C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNGF}}} + a_{Maturity_{GN}}\, C_{\mathbf{P}_{Maturity_{GN}}} \end{array} \right) - a_{Maturity_{GN}}\, Maturity'_{GN}
\end{aligned}
\tag{68}
$$

Transforming Equation (68) to unit space yields

$$
\begin{aligned}
& \boldsymbol{exp}\left(a_{S_{DNGF}}\, \boldsymbol{ln}\left(S'_{DNGF}\right)\right) = \\
& \frac{\boldsymbol{exp}\left(\left(\dfrac{a_{S_{DNGF}}{}^{2} + a_{Maturity_{GN}}{}^{2}}{a_{S_{DNEst}}}\right) \boldsymbol{ln}\left(S'_{DNew}\right)\right) \boldsymbol{exp}\left(-a_{Maturity_{GN}}\, Maturity'_{GN}\right)}{\boldsymbol{exp}\left(\begin{array}{l} \left(\dfrac{a_{S_{DNGF}}{}^{2} + a_{Maturity_{GN}}{}^{2}}{a_{S_{DNEst}}}\right) C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}} + \\ a_{S_{DNGF}}\, C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNGF}}} + a_{Maturity_{GN}}\, C_{\mathbf{P}_{Maturity_{GN}}} \end{array} \right)}
\end{aligned}
$$

$$
\therefore S'_{DNGF} = \left(\frac{ S'^{\left(\frac{a_{S_{DNGF}} + a_{Maturity_{GN}}{}^{2}}{a_{S_{DNEst}}}\right)}_{DNew}\, \boldsymbol{exp}\left(-\left(\dfrac{a_{Maturity_{GN}}}{a_{S_{DNGF}}}\right) Maturity'_{GN}\right) }{ \boldsymbol{exp}\left(\begin{array}{l} \left(\dfrac{a_{S_{DNGF}}{}^{2} + a_{Maturity_{GN}}{}^{2}}{a_{S_{DNEst}}}\right) C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}} + \\ a_{S_{DNGF}}\, C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNGF}}} + a_{Maturity_{GN}}\, C_{\mathbf{P}_{Maturity_{GN}}} \end{array} \right) } \right)^{1/a_{S_{DNGF}}}
\tag{69}
$$

We next convert the normalized variables S'_{DNew}, S'_{DNGF}, and $Maturity'_{GN}$ back to their un-normalized state by multiplying each by their corresponding standard deviations from Definitions (50) to get

$$
S_{DNGF} = \left(\frac{ S_{DNew}^{\left(\frac{a_{S_{DNGF}} + a_{Maturity_{GN}}{}^{2}}{a_{S_{DNEst}}}\right)}\, \boldsymbol{exp}\left(-\left(\dfrac{a_{Maturity_{GN}}}{a_{S_{DNGF}}}\right) Maturity_{GN}\right) }{ \boldsymbol{exp}\left(\begin{array}{l} \left(\dfrac{a_{S_{DNGF}}{}^{2} + a_{Maturity_{GN}}{}^{2}}{a_{S_{DNEst}}}\right) C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}} + \\ a_{S_{DNGF}}\, C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNGF}}} + a_{Maturity_{GN}}\, C_{\mathbf{P}_{Maturity_{GN}}} \end{array} \right) } \right)^{1/a_{S_{DNGF}}}
\tag{70}
$$

Letting $a_{GN} - 1 \equiv \frac{a_{S_{DNGF}}{}^2 + a_{Maturity_{GN}}{}^2}{a_{S_{DNEst}}}$, $Decay_{GN} \equiv \frac{a_{Maturity_{GN}}}{a_{S_{DNGF}}}$, and

$$b_{GN} \equiv \boldsymbol{exp}\left(\begin{array}{l} \left(\frac{a_{S_{DNGF}}{}^2 + a_{Maturity_{GN}}{}^2}{a_{S_{DNEst}}} \right) C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNEst}}} + \\ a_{S_{DNGF}} C_{\boldsymbol{ln}(\mathbf{P})_{S_{DNGF}}} + a_{Maturity_{GN}} C_{\mathbf{P}_{Maturity_{GN}}} \end{array} \right)^{-1/a_{S_{DNGF}}} \quad \text{gives us}$$

$$S_{DNGF} = \frac{S_{DNEst}{}^{a_{GN}-1} \boldsymbol{exp}\left(-Decay_{GN} Maturity_{GN}\right)}{b_{GN}{}^{-1}} \qquad \textbf{(71)}$$

$$\therefore S_{DNGF} = e^{-\left(Decay_{GN}\right)\left(Maturity_{GN}\right)} b_{GN} S_{DNew}{}^{a_{GN}-1}$$

Acknowledging b_{GN} as a list $\mathbf{b_{GN}}$ with geometric mean central tendency $\tilde{b}_{GN}$

$$\left(b_{GN_i} = \frac{S_{DNGF_i}}{e^{-\left(Decay_{GN}\right)\left(Maturity_{GN}\right)} S_{DNEst_i}{}^{a_{GN}-1}} \right) \Bigg|_{i=1}^{N} \qquad \textbf{(72)}$$

$$\rightarrow \mathbf{b_{GN}} = \frac{\mathbf{S_{DNGF}}}{e^{-\left(Decay_{GN}\right)\left(\mathbf{Maturity_{GN}}\right)} \mathbf{S_{DNEst}}{}^{a_{GN}-1}}$$

$$\tilde{b}_{GN} \equiv \boldsymbol{GeoMean}\left(\mathbf{b_{GN}}\right) = \boldsymbol{exp}\left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{ln}\left(b_{GN_i}\right) \right) \qquad \textbf{(73)}$$

we instantiate Equation (71) with the value $\tilde{b}_{GN}$ to get

$$S_{DNGF} \triangleq e^{-\left(Decay_{GN}\right)\left(Maturity_{GN}\right)} \tilde{b}_{GN} S_{DNew}{}^{a_{GN}-1} \qquad \textbf{(74)}$$

Apply Error and Uncertainty Attributed to the GER

Linear algebra provides that in $\mathbb{R}^m$ space any line can be described by the vector equation

$$\mathbf{r} = \mathbf{r}' + t\mathbf{a} \;\; \left| \begin{array}{l} \mathbf{r}, \mathbf{r}', \mathbf{a} \in \mathbb{R}^m \\ t \in \mathbb{R} \end{array} \right. \qquad \textbf{(75)}$$

where $\mathbf{r}$ is a position vector representing any point on that line, $\mathbf{r}'$ is a position vector representing a known point on that line, $\mathbf{a}$ is a direction vector of that line, and t is a scaling parameter such that for every specific unique value of t there is a specific unique value of $\mathbf{r}$. We can instantiate Equation (75) with what we already know about L_{ODR} and the data upon which it is based:

- We have already formatted and translated (offset) all of the sigma-normalized and log-transformed data points as the matrix $\mathbf{M}$ such that the data set centroid $C'_{\boldsymbol{ln}(\mathbf{P})_{S_{DN}}}$ of the translated data set has the position vector $[0 \quad 0 \quad 0]$; aka, the zero vector $\mathbf{0}$.
- We have already proven that a best fit ODR line passes through (contains) its data set centroid; therefore, since our data set is now centroid-translated, we can let $\mathbf{r}' \equiv [0 \quad 0 \quad 0]$.
- We have already determined a direction vector that is appropriate for this best fit ODR line through the sigma-normalized log-transformed centroid-translated data set when we earlier applied the SVD process; therefore, $\mathbf{a} \equiv \mathbf{a}_{\mathbf{S_{DN}}} = \left[a_{S_{DNEst}} \quad a_{S_{DNGF}} \quad a_{Maturity_{GN}} \right]$.

Making the appropriate substitution for the known point position vector $\mathbf{r}' \equiv [0 \quad 0 \quad 0]$ yields

$$\mathbf{r} = [0 \quad 0 \quad 0] \quad t\mathbf{a}_{\mathbf{S_{DN}}} \left| \begin{array}{l} \mathbf{r}, \mathbf{a}_{\mathbf{S_{DN}}} \in \mathbb{R}^m \\ t \in \mathbb{R} \end{array} \right. \qquad \textbf{(76)}$$

$$\rightarrow \mathbf{r} = t\mathbf{a}_{\mathbf{S_{DN}}}$$

Linear algebra also gives us a theorem for projecting one vector $\mathbf{v}$ in $\mathbb{R}^m$ space onto another vector $\mathbf{u}$ in the same vector space

$$\boldsymbol{proj}_{\mathbf{u}}(\mathbf{v}) = \left(\frac{\mathbf{u} \bullet \mathbf{v}}{\mathbf{u} \bullet \mathbf{u}} \right) \mathbf{u} \qquad \textbf{(77)}$$

We can find the position vector $\mathbf{r}_i''$ of the point P_i'' on L'_{ODR} that is closest to a sigma-normalized log-transformed centroid-translated historical observation $\boldsymbol{ln}''(P_i)$ by projecting the $\boldsymbol{ln}''(P_i)$ position vector $\mathbf{r}_i'''$ onto the position vector $\mathbf{r}$ of any point on L'_{ODR}. We set this up by letting $\boldsymbol{proj}_{\mathbf{u}}(\mathbf{v}) \equiv \mathbf{r}''$, $\mathbf{u} \equiv \mathbf{r}$, and $\mathbf{v} \equiv \mathbf{r}'''$; and substituting these equivalencies into Theorem (77) to get

$$\mathbf{r}_i'' = \left(\frac{\mathbf{r} \bullet \mathbf{r}_i'''}{\mathbf{r} \bullet \mathbf{r}} \right) \mathbf{r} \qquad \textbf{(78)}$$

Substituting $\mathbf{r}$ in Equation (78) with its equivalent $t\mathbf{a}_{\mathbf{S_{DN}}}$ in Equation (76) gives us

$$\mathbf{r}_i'' = \left(\frac{t\mathbf{a}_{S_{DN}} \bullet \mathbf{r}_i'''}{t\mathbf{a} \bullet t\mathbf{a}} \right) t\mathbf{a}$$
$$\rightarrow \mathbf{r}_i'' = \left(\frac{\mathbf{a}_{S_{DN}} \bullet \mathbf{r}_i'''}{\mathbf{a}_{S_{DN}} \bullet \mathbf{a}_{S_{DN}}} \right) t\mathbf{a}_{S_{DN}} \tag{79}$$

The error vector in sigma-normalized log-transformed centroid-translated space $\boldsymbol{\varepsilon}_i'$ associated with a data point $\boldsymbol{ln}''(P_i)$ is expressed as the difference $\mathbf{r}_i''' - \mathbf{r}_i''$ where each coordinate of $\boldsymbol{\varepsilon}_i'$ represents its associated dimensional component of the data point's error vector. Therefore

$$\boldsymbol{\varepsilon}_i' \equiv \mathbf{r}_i''' - \mathbf{r}''$$
$$\rightarrow \mathbf{r}'' \equiv \mathbf{r}_i''' - \boldsymbol{\varepsilon}_i' \tag{80}$$

Substituting $\mathbf{r}_i''$ in Equation (79) with its equivalent in Equation (80) gives us

$$\mathbf{r}_i''' - \boldsymbol{\varepsilon}_i' = \left(\frac{\mathbf{a}_{S_{DN}} \bullet \mathbf{r}_i'''}{\mathbf{a}_{S_{DN}} \bullet \mathbf{a}_{S_{DN}}} \right) \mathbf{a}$$
$$\rightarrow \boldsymbol{\varepsilon}_i' = \mathbf{r}_i''' - \left(\frac{\mathbf{a}_{S_{DN}} \bullet \mathbf{r}_i'''}{\mathbf{a}_{S_{DN}} \bullet \mathbf{a}_{S_{DN}}} \right) \mathbf{a}_{S_{DN}} \tag{81}$$

We can now define the signed magnitudes of each New DSLOC estimate error vector component as

$$\begin{bmatrix} \varepsilon'_{S_{DNEst_i}} & \varepsilon'_{S_{DNEst_i}} & \varepsilon'_{Maturity_{GN_i}} \end{bmatrix} \equiv \boldsymbol{\varepsilon}_i' = \mathbf{r}_i''' - \left(\frac{\mathbf{a}_{S_{DN}} \bullet \mathbf{r}_i'''}{\mathbf{a}_{S_{DN}} \bullet \mathbf{a}_{S_{DN}}} \right) \mathbf{a}_{S_{DN}} \tag{82}$$

We next instantiate and evaluate Equation (82) with each of our three sigma-normalized, log-transformed, centroid-translated lists $\boldsymbol{ln}(\mathbf{S}''_{\mathbf{DNEst}})$, $\boldsymbol{ln}(\mathbf{S}''_{\mathbf{DNGF}})$, and $\mathbf{Maturity}''_{\mathbf{GN}}$ to produce the sigma-normalized, log-transformed error lists $\boldsymbol{\varepsilon}'_{\mathbf{S_{DNEst}}}$, $\boldsymbol{\varepsilon}'_{\mathbf{S_{DNGF}}}$, and $\boldsymbol{\varepsilon}'_{\mathbf{Maturity_{GN}}}$ as

$$\begin{aligned}
&\left[\boldsymbol{\varepsilon}'_{\mathbf{S_{DNEst}}} \quad \boldsymbol{\varepsilon}'_{\mathbf{S_{DNEst}}} \quad \boldsymbol{\varepsilon}'_{\mathbf{Maturity_{GN}}}\right] \equiv \\
&\qquad \left.\begin{pmatrix}
\left[S''_{DNEst_i} \quad S''_{DNGF_i} \quad Maturity''_{GN_i}\right] - \\
\left(\dfrac{\left[S''_{DNEst_i} \quad S''_{DNGF_i} \quad Maturity''_{GN_i}\right]}{\left[a_{S_{DNEst}} \quad a_{S_{DNGF}} \quad a_{Maturity_{GN}}\right] \bullet \left[a_{S_{DNEst}} \quad a_{S_{DNGF}} \quad a_{Maturity_{GN}}\right]}\right) \\
\left[a_{S_{DNEst}} \quad a_{S_{DNGF}} \quad a_{Maturity_{GN}}\right]
\end{pmatrix}\right|_{i=1}^{N}
\end{aligned} \tag{83}$$

Note that we do not refer to the error component lists as being centroid-translated since error is a relative quantity with respect to a data set and its best fit ODR line; therefore, if the data set and hence its best fit ODR line are re-positioned within the space, the error component lists are not affected.

We next convert the sigma-normalized log-transformed error component lists $\boldsymbol{\varepsilon}'_{\mathbf{S_{DNEst}}}$, $\boldsymbol{\varepsilon}'_{\mathbf{S_{DNGF}}}$, and $\boldsymbol{\varepsilon}'_{\mathbf{Maturity_{GN}}}$ to un-sigma-normalized unit space by first exponentiating each of $\boldsymbol{\varepsilon}'_{\mathbf{S_{DNEst}}}$ and $\boldsymbol{\varepsilon}'_{\mathbf{S_{DNGF}}}$ (NOT $\boldsymbol{\varepsilon}_{\mathbf{Maturity_{GN}}}$ since it remains exponentiated in our unit space equation) and then by multiplying all three of $\boldsymbol{\varepsilon}'_{\mathbf{S_{DNEst}}}$, $\boldsymbol{\varepsilon}'_{\mathbf{S_{DNGF}}}$, and $\boldsymbol{\varepsilon}'_{\mathbf{Maturity_{GN}}}$ by the standard deviation value that corresponds to its respective dimension (see Definitions (50)) to get $\boldsymbol{\varepsilon}_{\mathbf{S_{DNEst}}}$, $\boldsymbol{\varepsilon}_{\mathbf{S_{DNGF}}}$, and $\boldsymbol{\varepsilon}_{\mathbf{Maturity_{GN}}}$.

Because we have developed our estimating relationship using ODR in log space; the error represented by $\boldsymbol{\varepsilon}_{\mathbf{S_{DNEst}}}$ and $\boldsymbol{\varepsilon}_{\mathbf{S_{DNGF}}}$, while additive in log space; must be transformed to unit space and treated as multiplicative. We can rewrite Equation (74) to account for this multiplicative error in each variable as

$$\begin{aligned}
&S_{DNGF}\left(\boldsymbol{\varepsilon}_{\mathbf{S_{DNGF}}}\right) \triangleq \\
&\qquad e^{-\left(Decay_{GN}\right)\left(Maturity_{GN} + \boldsymbol{\varepsilon}_{\mathbf{Maturity}}\right)} \tilde{b}_{GN} \left(S_{DNew}\left(\boldsymbol{\varepsilon}_{\mathbf{S_{DNEst}}}\right)\right)^{a_{GN}-1} \\
&\qquad \therefore \mathbf{S_{DNGF}} \triangleq e^{-\left(Decay_{GN}\right)\left(Maturity_{GN}\right)} \tilde{b}_{GN} \\
&\qquad\qquad \frac{\left(\boldsymbol{\varepsilon}_{\mathbf{S_{DNEst}}}\right)^{a_{GN}-1}}{\left(e^{Decay_{GN} \boldsymbol{\varepsilon}_{\mathbf{Maturity_{GN}}}}\right)\left(\boldsymbol{\varepsilon}_{\mathbf{S_{DNGF}}}\right)} S_{DNEst}{}^{a_{GN}-1}
\end{aligned} \tag{84}$$

Letting

$$\boldsymbol{\varepsilon}_{\mathbf{GN}} \equiv \frac{\boldsymbol{\varepsilon}_{\mathbf{S_{DNEst}}}{}^{a_{GN}-1}}{\left(e^{Decay_{GN}\boldsymbol{\varepsilon}_{\mathbf{Maturity_{GN}}}}\right)\boldsymbol{\varepsilon}_{\mathbf{S_{DNGF}}}} \tag{85}$$

and then substituting $\boldsymbol{\varepsilon}_{\mathbf{GN}}$ for its equivalent into Equation (84) gives us

$$\mathbf{S_{DNGF}} \triangleq e^{-(Decay_{GN})(Maturity_{GN})}\tilde{b}_{GN}\boldsymbol{\varepsilon}_{\mathbf{GN}}S_{DNew}{}^{a_{GN}-1} \tag{86}$$

Substituting $\mathbf{S_{DNGF}}$ in Equation (86) with its equivalent defined in the constraint portion of Equation (49) yields

$$\begin{gathered}\frac{\mathbf{S_{DGANew}}}{S_{DNew}} \triangleq e^{-(Decay_{GN})(Maturity_{GN})}\tilde{b}_{GN}\boldsymbol{\varepsilon}_{\mathbf{GN}}S_{DNew}{}^{a_{GN}-1} \\ \therefore \mathbf{S_{DGANew}} = e^{-(Decay_{GN})(Maturity_{GN})}b_{GN}S_{DNew}{}^{a_{GN}}\end{gathered} \tag{87}$$

Express the GER as a Random Variable Estimating Relationship

$$\boldsymbol{S}_{\boldsymbol{DGANew}} \triangleq e^{-(Decay_{GN})(Maturity_{GN})}\tilde{b}_{GN}\boldsymbol{\varepsilon}_{\boldsymbol{GN}}S_{DNew}{}^{a_{GN}} \tag{88}$$

Add CSCI Normalization and Extend to All 3 Types of DSLOC

$$\begin{aligned}\boldsymbol{S}_{\boldsymbol{DGANew}} &\triangleq e^{-(Decay_{GN})(Maturity_{GN})}\tilde{b}_{GN}\boldsymbol{\varepsilon}_{\boldsymbol{GN}}\left(\frac{S_{DNew}}{K_N}\right)^{a_{GN}}K_N \\ \boldsymbol{S}_{\boldsymbol{DGAMod}} &\triangleq e^{-(Decay_{GM})(Maturity_{GM})}\tilde{b}_{GN}\boldsymbol{\varepsilon}_{\boldsymbol{GM}}\left(\frac{S_{DMod}}{K_M}\right)^{a_{GM}}K_M \\ \boldsymbol{S}_{\boldsymbol{DGAUmod}} &\triangleq e^{-(Decay_{GU})(Maturity_{GU})}\tilde{b}_{GU}\boldsymbol{\varepsilon}_{\boldsymbol{GU}}\left(\frac{S_{DUmod}}{K_U}\right)^{a_{GU}}K_U\end{aligned} \tag{89}$$

Results

Custom CDFs for each of $\boldsymbol{\varepsilon}_{\boldsymbol{GN}}$, $\boldsymbol{\varepsilon}_{\boldsymbol{GM}}$, and $\boldsymbol{\varepsilon}_{\boldsymbol{GU}}$ that are specific to DEGM8SV can be found in Appendix A. Measurements, parameters, and statistics specific to DEGM8SV can be found in the tables contained in Appendix F.

6. Summary and Conclusions

A significant challenge that many cost analysts and project managers face is predicting by how much their estimates of software development cost and schedule will change over the life of the project. Examination of currently-accepted software cost, schedule, and defect estimation algorithms reveals a common acknowledgment that estimated software size is the single most influential independent variable (Ross, 2003) (Ross, 2005). Unfortunately, the most important business decisions about a software project are made at its beginning; the time when most estimating is done; and coincidently the time of minimum knowledge, maximum uncertainty, and hysterical optimism (Ross, 2005). This paper describes a model and methodology, DEGM8, that provides probabilistic growth adjustment to single-point TBEs of DSLOC, for New, Modified, and Unmodified software that is sensitive to the maturity of the estimates. The model is based on more and more-recent SRDR data collected by the DoD combined with a state-of-the-art data regression technique (Orthogonal Distance Regression using Singular Value Decomposition) that more-accurately models the data and its error (uncertainty).

It continues to be the authors' collective opinion that the DEGM8, as described in this paper, represents a quantum improvement over the field of available software code growth methodologies. Specifically, among the advantages of this model over the Holchin (2003) and Jensen (2008) code growth matrices are the following:

- DEGM8 is based on DoD-collected SRDR data versus Holchin's Delphi survey of *experts* approach and Jensen's data from multiple proprietary sources.
- DEGM8 requires only one parameter, estimate *Maturity*, which is reasonably objective versus Holchin's and Jensen's rather subjective and vaguely-defined Complexity and Maturity parameters.
- DEGM8 equations are non-linear allowing for the existence of economies/diseconomies of scale. Holchin and Jensen model growth as a constant factor (as does DEGM7).
- DEGM8 equations include a composite error factor distribution (embodies uncertainty derived from the data) versus Holchin's single-point growth-factor result. (Jensen uses the lognormal distribution to model uncertainty.)
- DEGM8 provides error factor distribution decay based on updated estimate *Maturity* versus Holchin's single-point growth factor reduction based on updated Complexity and Maturity parameters. (Jensen defines Maturity in terms of defined program phases.)
- DEGM8 differentiates between New, Modified, and Unmodified DSLOC growth versus Holchin's and Jensen's one-growth-factor-fits-all approach.

- DEGM8 provides correlation (a correlation matrix) between each of growth-adjusted New, Modified, and Unmodified DSLOC to support proper summation of their distributions. Neither Holchin nor Jensen provide this capability.

The DEGM8 represents a significant update and improvement to the DEGM7, which has been used as part of the basis for numerous USAF Program Objective Memoranda (POM), Program Office Estimates (POEs), Independent Cost Estimates (ICEs), and Service Cost Positions (SCPs). Planned enhancements to this model include:

- Updating the model to incorporate new data from the 2017 update of the SRDR database;
- Rerunning the data analysis using specific stratifications of the SRDR database in order to create growth models that are unique to specific software operating environments, application domains, and other characteristics of interest;
- Improving the DEGM8SV by searching for and collecting additional data (including SDLCBegin / SwAccept pairs) and rerunning the regressions;
- Creating specific growth model instances regressed from the same data subset that is regressed to create specific Joint Cost and Duration Estimating Relationships, thus creating unified software estimating methodologies that are specific to certain types of software.

Appendix A Custom CDFs for DEGM8 Default and DEGM8SV

DEGM8 Default

Percentile	*Value*	*Percentile*	*Value*	*Percentile*	*Value*
JCDER349_e_GN_CDF		*JCDER349_e_GM_CDF*		*JCDER349_e_GU_CDF*	
0.21645022	0.09821762	0.35211268	0.09492565	0.33783784	0.10673836
0.64935065	0.10232146	1.05633803	0.10684445	1.01351351	0.18960627
1.08225108	0.11109953	1.76056338	0.12369618	1.68918919	0.19945773
1.51515152	0.13489884	2.81690141	0.15941357	2.36486486	0.20979475
1.94805195	0.16370668	2.81690142	0.15941358	3.04054054	0.24345644
2.38095238	0.18335752	3.87323944	0.17850355	3.71621622	0.27187318
2.81385281	0.19330899	4.57746479	0.19543874	4.39189189	0.30710571
3.24675325	0.21836560	5.28169014	0.20389943	5.06756757	0.31986433
3.67965368	0.22039978	5.98591549	0.23714712	5.74324324	0.39364068
4.11255411	0.22058416	6.69014085	0.24288171	6.41891892	0.47735093
4.54545455	0.22563979	7.39436620	0.24705582	7.09459459	0.50690072
4.97835498	0.22852010	8.09859155	0.25971336	7.77027027	0.51225232
5.41125541	0.23557328	8.80281690	0.26497418	8.44594595	0.51762484
5.84415584	0.23867731	9.50704225	0.26548364	9.12162162	0.52028920
6.27705628	0.23970764	10.21126761	0.30119541	9.79729730	0.52204448
6.70995671	0.24467688	10.91549296	0.31100004	10.47297297	0.54042888
7.14285714	0.25017776	11.61971831	0.33708118	11.14864865	0.54975280
7.57575758	0.25514710	12.32394366	0.35378674	12.16216216	0.56340323
8.00865801	0.26989293	13.02816901	0.36074439	12.16216217	0.56340324
8.44155844	0.27366389	13.73239437	0.36277815	13.17567568	0.62790738
8.87445887	0.28133024	14.43661972	0.37305545	13.85135135	0.63718801
9.30735931	0.28768647	15.14084507	0.37885928	14.52702703	0.71219519
9.74025974	0.29379456	15.84507042	0.38878340	15.20270270	0.73881429
10.38961039	0.31633894	16.54929577	0.39663624	15.87837838	0.76143309
10.38961040	0.31633895	17.25352113	0.41137016	16.55405405	0.76266655
11.03896104	0.31650047	18.30985915	0.42150043	17.22972973	0.77549967
11.47186147	0.32006359	18.30985916	0.42150044	17.90540541	0.78325797
11.90476190	0.32967199	19.36619718	0.45023719	18.58108108	0.79477746
12.33766234	0.34455362	20.07042254	0.45038705	19.25675676	0.80329462
12.77056277	0.37395471	20.77464789	0.49132637	19.93243243	0.82562109
13.20346320	0.39148263	21.47887324	0.53930522	20.60810811	0.83501158
13.63636364	0.39451336	22.18309859	0.55709646	21.28378378	0.83821397
14.06926407	0.40104391	22.88732394	0.55733744	21.95945946	0.84197220
14.50216450	0.41907499	23.59154930	0.57964684	22.63513514	0.84281395
14.93506494	0.42299016	24.29577465	0.58134245	23.31081081	0.86196544
15.36796537	0.42958110	25.00000000	0.59308040	23.98648649	0.87663594
15.80086580	0.45026453	25.70422535	0.61536825	24.66216216	0.88095989
16.23376623	0.45854228	26.40845070	0.63787364	25.33783784	0.89568406
16.66666667	0.46373383	27.11267606	0.66070194	26.01351351	0.90678107
17.09956710	0.47131963	27.81690141	0.72561716	26.68918919	0.92262526

Percentile	Value	Percentile	Value	Percentile	Value
JCDER349_e_GN_CDF		JCDER349_e_GM_CDF		JCDER349_e_GU_CDF	
17.53246753	0.49222882	28.52112676	0.73552024	27.36486486	0.92435012
17.96536797	0.49240754	29.22535211	0.77890545	28.04054054	0.92902123
18.39826840	0.50098235	29.92957746	0.78419067	28.71621622	0.93234456
18.83116883	0.50347304	30.63380282	0.79861919	29.39189189	0.94481865
19.26406926	0.50505288	31.33802817	0.81071575	30.06756757	0.95134657
19.69696970	0.51629065	32.04225352	0.84507406	30.74324324	0.95232862
20.34632035	0.53523730	32.74647887	0.86474552	31.41891892	0.95443446
20.34632036	0.53523731	33.45070423	0.88659754	32.09459459	0.95471250
20.99567100	0.53893907	34.15492958	0.89616397	32.77027027	0.95515820
21.42857143	0.55008859	34.85915493	0.90125044	33.44594595	0.95787124
21.86147186	0.55124128	35.56338028	0.90723316	34.12162162	0.96128995
22.29437229	0.55278169	36.26760563	0.94435818	35.13513514	0.96274656
22.72727273	0.56331881	36.97183099	0.96496430	35.13513515	0.96274657
23.16017316	0.57413751	38.02816901	0.96607188	36.14864865	0.96521768
23.59307359	0.58451797	38.02816902	0.96607189	36.82432432	0.96527064
24.02597403	0.58949251	39.08450704	0.96898028	37.50000000	0.96962100
24.45887446	0.58963747	39.78873239	0.98164758	38.17567568	0.97447003
24.89177489	0.59406530	40.49295775	0.99460500	38.85135135	0.97682458
25.32467532	0.64763349	41.19718310	1.00062964	39.52702703	0.97726240
25.75757576	0.64955919	41.90140845	1.00363883	40.20270270	0.98569819
26.19047619	0.65082979	42.60563380	1.02232975	40.87837838	0.99271735
26.62337662	0.65498366	43.30985915	1.04921698	41.55405405	0.99281532
27.05627706	0.66313574	44.01408451	1.07738733	42.22972973	0.99291536
27.48917749	0.68437058	44.71830986	1.08732480	42.90540541	0.99605774
27.92207792	0.70634203	45.42253521	1.08844706	43.58108108	0.99781065
28.35497835	0.71032622	46.12676056	1.09278189	44.25675676	0.99832844
28.78787879	0.71211396	46.83098592	1.10297469	44.93243243	1.00065760
29.22077922	0.72495221	47.53521127	1.12043120	45.94594595	1.00480454
29.65367965	0.76954993	48.23943662	1.13932301	45.94594596	1.00480455
30.08658009	0.78069905	48.94366197	1.15290836	46.95945946	1.01054113
30.51948052	0.78866247	49.64788732	1.15422336	47.63513514	1.01313500
30.95238095	0.80004616	50.35211268	1.15632827	48.31081081	1.01431860
31.60173160	0.83899489	51.05633803	1.15862717	48.98648649	1.01598499
31.60173161	0.83899490	51.76056338	1.16817627	49.66216216	1.01679682
32.25108225	0.84355358	52.46478873	1.17823225	50.33783784	1.01993995
32.68398268	0.84804201	53.16901408	1.18058853	51.01351351	1.02045178
33.11688312	0.84836838	53.87323944	1.18243449	51.68918919	1.02471509
33.54978355	0.85737260	54.57746479	1.18407902	52.36486486	1.02517320
33.98268398	0.86288518	55.28169014	1.19389059	53.04054054	1.02643147
34.41558442	0.86565230	57.04225352	1.19844250	53.71621622	1.03068399
34.84848485	0.88588056	57.04225353	1.19844251	54.39189189	1.03345167

Percentile	*Value*	*Percentile*	*Value*	*Percentile*	*Value*
JCDER349_e_GN_CDF		*JCDER349_e_GM_CDF*		*JCDER349_e_GU_CDF*	
35.28138528	0.88962728	57.04225354	1.19844252	55.06756757	1.03518492
35.71428571	0.89028080	57.04225355	1.19844253	55.74324324	1.03941485
36.14718615	0.89959510	58.80281690	1.21153873	56.41891892	1.04126235
36.58008658	0.90328560	59.50704225	1.22300063	57.09459459	1.04138225
37.01298701	0.90634077	60.21126761	1.23364639	57.77027027	1.04692846
37.44588745	0.91754061	60.91549296	1.25008616	58.44594595	1.04792167
37.87878788	0.92323309	61.61971831	1.25448392	59.12162162	1.05557466
38.31168831	0.93631473	62.32394366	1.29524374	59.79729730	1.06808282
38.74458874	0.94482377	63.02816901	1.31605101	60.47297297	1.07376488
39.17748918	0.95128544	63.73239437	1.32506191	61.14864865	1.08935287
39.61038961	0.96145076	64.43661972	1.34117075	61.82432432	1.08981529
40.04329004	0.98448707	65.49295775	1.35439267	62.83783784	1.09961564
40.47619048	1.01076477	65.49295776	1.35439268	62.83783785	1.09961565
40.90909091	1.02892104	66.54929577	1.37658361	63.85135135	1.10651320
41.34199134	1.03768307	67.25352113	1.37915579	64.52702703	1.10891397
41.77489177	1.04725298	67.95774648	1.40701851	65.20270270	1.11295598
42.20779221	1.05118515	68.66197183	1.41422111	65.87837838	1.11882794
42.64069264	1.05132735	69.36619718	1.42637124	66.55405405	1.12844118
43.07359307	1.06718546	70.07042254	1.44531403	67.22972973	1.13443558
43.50649351	1.07451964	70.77464789	1.47315839	67.90540541	1.13542148
43.93939394	1.07659579	71.47887324	1.49081150	68.58108108	1.15473245
44.37229437	1.08163376	72.18309859	1.54708991	69.59459459	1.15539873
44.80519481	1.08474770	72.88732394	1.55583947	69.59459460	1.15539874
45.23809524	1.08922249	73.59154930	1.56382122	70.60810811	1.16009834
45.67099567	1.09511497	74.29577465	1.56491152	71.28378378	1.16157482
46.10389610	1.09565189	75.00000000	1.56797051	72.29729730	1.16754135
46.53679654	1.09800178	75.70422535	1.58043573	72.29729731	1.16754136
46.96969697	1.10938441	76.40845070	1.62531725	73.31081081	1.16759715
47.40259740	1.11135363	77.11267606	1.78452498	73.98648649	1.18450417
47.83549784	1.12682812	77.81690141	1.80415794	74.66216216	1.20403030
48.26839827	1.12748891	78.52112676	1.83920114	75.33783784	1.21281155
48.70129870	1.13572756	79.22535211	1.88882286	76.01351351	1.21689269
49.13419913	1.13838012	79.92957746	1.96622901	76.68918919	1.25925016
49.56709957	1.15478281	80.63380282	1.97495992	77.36486486	1.28065136
50.00000000	1.16895040	81.33802817	2.17124647	78.04054054	1.28508203
50.43290043	1.17980294	82.04225352	2.23511451	78.71621622	1.28712532
50.86580087	1.18763087	82.74647887	2.31668772	79.39189189	1.30733194
51.29870130	1.19559110	83.45070423	2.33666407	80.06756757	1.34235065
51.73160173	1.20847670	84.15492958	2.42965630	80.74324324	1.35490194
52.16450216	1.21009716	84.85915493	2.52126116	81.41891892	1.38417187

Percentile	Value	Percentile	Value	Percentile	Value
JCDER349_e_GN_CDF		JCDER349_e_GM_CDF		JCDER349_e_GU_CDF	
52.59740260	1.21718712	85.56338028	3.94768260	82.09459459	1.42239106
53.67965368	1.21856639	86.26760563	4.03681511	82.77027027	1.47491346
53.67965369	1.21856640	86.97183099	4.08535697	83.44594595	1.54087554
53.67965370	1.21856641	87.67605634	4.31916018	84.12162162	1.56165460
53.67965371	1.21856642	88.38028169	4.38631400	84.79729730	1.63760872
54.76190476	1.22070022	89.08450704	4.46549480	85.47297297	1.64581226
55.19480519	1.24898802	89.78873239	4.69566894	86.14864865	1.74727755
55.62770563	1.26316574	90.49295775	5.36466737	86.82432432	1.85380525
56.06060606	1.30071886	91.19718310	5.64451232	87.50000000	1.96463749
56.49350649	1.30311379	91.90140845	6.15092302	88.17567568	2.19879720
56.92640693	1.30328666	92.60563380	6.60824135	88.85135135	2.25059828
57.35930736	1.30469479	93.30985915	6.88934213	89.52702703	2.30862992
57.79220779	1.31474442	94.01408451	6.93822543	90.20270270	2.45242958
58.22510823	1.32808421	94.71830986	8.08527304	90.87837838	2.49114222
58.65800866	1.33893139	95.42253521	8.68058303	91.55405405	2.56578076
59.09090909	1.34017032			92.22972973	2.84820035
59.52380952	1.36144435			92.90540541	3.42449546
59.95670996	1.36740683			93.58108108	4.36850399
60.38961039	1.40539438			94.25675676	4.58082833
60.82251082	1.40738383			94.93243243	6.70576472
61.25541126	1.40801947			95.60810811	6.98289033
61.68831169	1.40953550				
62.12121212	1.41229135				
62.55411255	1.42075900				
62.98701299	1.43790840				
63.41991342	1.44077958				
63.85281385	1.44130955				
64.28571429	1.44795403				
64.71861472	1.45670678				
65.15151515	1.45923646				
65.58441558	1.46438521				
66.01731602	1.46447619				
66.45021645	1.46648988				
66.88311688	1.47076292				
67.31601732	1.47296703				
67.74891775	1.47756225				
68.18181818	1.48139681				
68.61471861	1.48462273				
69.04761905	1.48575596				
69.48051948	1.49090125				

Percentile	*Value*	*Percentile*	*Value*	*Percentile*	*Value*
JCDER349_e_GN_CDF		*JCDER349_e_GM_CDF*		*JCDER349_e_GU_CDF*	
69.91341991	1.49141098				
70.34632035	1.49472454				
70.77922078	1.49737276				
71.21212121	1.50360791				
71.64502165	1.50593145				
72.07792208	1.51197987				
72.51082251	1.51238818				
72.94372294	1.51418200				
73.37662338	1.51956967				
73.80952381	1.52604731				
74.24242424	1.52724671				
74.67532468	1.53910715				
75.10822511	1.54909324				
75.54112554	1.55196396				
75.97402597	1.55407957				
76.40692641	1.60155165				
76.83982684	1.61218716				
77.27272727	1.62130637				
77.70562771	1.63103661				
78.13852814	1.63662640				
78.57142857	1.63891713				
79.00432900	1.63919405				
79.43722944	1.65367135				
79.87012987	1.67889383				
80.30303030	1.71500546				
80.73593074	1.72094563				
81.16883117	1.75016409				
81.60173160	1.75624422				
82.03463203	1.79557587				
82.46753247	1.86888860				
82.90043290	1.86969853				
83.33333333	2.00941077				
83.76623377	2.01070395				
84.19913420	2.03365666				
84.63203463	2.10800121				
85.06493506	2.24289924				
85.49783550	2.26979724				
85.93073593	2.42286591				
86.36363636	2.42915557				
86.79653680	2.43417811				

Percentile	Value	Percentile	Value	Percentile	Value
JCDER349_e_GN_CDF		JCDER349_e_GM_CDF		JCDER349_e_GU_CDF	
87.22943723	2.53313064				
87.66233766	2.68118931				
88.09523810	2.81748607				
88.52813853	2.83066175				
88.96103896	2.88697964				
89.39393939	3.02215573				
89.82683983	3.07194389				
90.25974026	3.14832918				
90.69264069	3.17650058				
91.12554113	3.33077404				
91.55844156	3.61764732				
91.99134199	3.62766747				
92.42424242	3.85337172				
92.85714286	3.87914435				
93.29004329	4.20076391				
93.72294372	4.31303466				
94.15584416	4.53838753				
94.58874459	4.99452737				
95.02164502	5.16629480				
95.45454545	5.48454971				
95.88744589	5.68416997				
96.32034632	6.35782957				
96.75324675	7.35600544				
97.18614719	8.80095370				

DEGM8SV

Percentile	Value	Percentile	Value	Percentile	Value
JCDER101_e_GN_CDF		JCDER101_e_GM_CDF		JCDER101_e_GU_CDF	
1.42857143	0.39797761	2.17391304	0.12594765	1.61290323	0.13570110
4.28571429	0.44850623	6.52173913	0.15264906	4.83870968	0.13926041
7.14285714	0.45502048	10.86956522	0.29956749	8.06451613	0.15562935
10.00000000	0.49553145	15.21739130	0.30107109	11.29032258	0.15998283
12.85714286	0.53769705	19.56521739	0.51047319	14.51612903	0.17382815
15.71428571	0.59597535	23.91304348	0.52432572	17.74193548	0.45944600
18.57142857	0.63250633	28.26086957	0.56496728	20.96774194	0.46209723
21.42857143	0.64519672	32.60869565	0.56835342	24.19354839	0.46922325
24.28571429	0.77058335	36.95652174	0.58355543	27.41935484	0.50365112
27.14285714	0.85185971	41.30434783	0.61270300	30.64516129	0.51921489
30.00000000	0.90596191	45.65217391	1.10082647	33.87096774	0.61009536
32.85714286	0.96374280	50.00000000	1.12965659	37.09677419	0.65629299
35.71428571	1.02491388	54.34782609	1.26664643	40.32258065	0.68347624
38.57142857	1.04556007	60.86956522	1.49520849	43.54838710	0.70730690
41.42857143	1.07067023	60.86956523	1.49520850	46.77419355	0.72038296
44.28571429	1.07451465	67.39130435	1.62277675	50.00000000	0.77768603
47.14285714	1.11223696	71.73913043	1.64009976	53.22580645	0.79931977
50.00000000	1.12029311	78.26086957	1.74582811	56.45161290	1.25171660
52.85714286	1.12912742	78.26086958	1.74582812	59.67741935	1.30113261
55.71428571	1.14227580	84.78260870	1.85588735	62.90322581	1.33957377
58.57142857	1.15615324	89.13043478	6.12074288	66.12903226	1.47613209
61.42857143	1.15739641	93.47826087	7.37744324	69.35483871	1.53213683
64.28571429	1.16677484	97.82608696	7.83946109	72.58064516	2.05349824
67.14285714	1.20762787			75.80645161	2.13644682
70.00000000	1.22701029			79.03225806	2.35487073
72.85714286	1.23870624			82.25806452	3.77161536
75.71428571	1.25187892			85.48387097	4.52797607
78.57142857	1.29971870			88.70967742	5.73543246
81.42857143	1.35686365			91.93548387	8.82957076
84.28571429	1.37357149			95.16129032	10.07136883
87.14285714	1.41537438			98.38709677	12.19462228
90.00000000	1.51039009				
92.85714286	1.53989881				
95.71428571	1.63834787				
98.57142857	3.56984543				

Appendix B DEGM8 Regressions Summary

r2 Software Estimating Framework (r2SEF)								
Joint Cost and Duration Estimating Relationship (JCDER) Data Sheet								
JCDER349: Version 8 DSLOC Growth Model Baseline w/ GF3 Valid Filtering Only								
Source Database:	SRDR 2015							
Filter Criteria:	SerialNo2015: >0; Report: 2630-3; SI: TRUE; Nonphysical: TRUE; GF3Va TRUE							
ESLOC Calculation Method:	Popp 2015							
Custom Select:	All							
DSLOC Measurement Point:	Final							
Effort SDLC Scope:	Core							
Duration SDLC Scope:	Core							
Data Set Statistics								
Number of Data Points (observations):			578					
	Smallest		*Largest*		*Mean*		*Geomean*	
Size (ESLOC):	12		913,602		70,172		24,859	
Core Effort (pm):	0.1		9,526.9		421.1		129.8	
Core Duration (cm):	1.15		115.02		35.59		28.71	
All-All Productivity (ESLOC / pm):	0.06		20,657.48		327.96		161.64	

DSLOC Estimate Growth Model Version:	Version 8
Version 8 DSLOC Estimate Growth Model Regression Method:	ODR

DSLOC Estimate Growth Model Equations and Variables

New DSLOC Growth Equation:	**S[DGANew]** ≙ **exp**(−(Decay*Maturity))*($\tilde{b}$[GN]***ε[GN]***(S[DNew]/K[N])^a[GN]*K[N]−S[DNew])+S[DNew]
Modified DSLOC Growth Equation:	**S[DGAMod]** ≙ **exp**(−(Decay*Maturity))*($\tilde{b}$[GM]***ε[GM]***(S[DMod]/K[M])^a[GM]*K[M]−S[DMod])+S[DMod]
Unmodified DSLOC Growth Equation:	**S[DGAUmod]** ≙ **exp**(−(Decay*Maturity))*($\tilde{b}$[GU]***ε[GU]***(S[DUmod]/K[U])^a[GU]*K[U]−S[DUmod])+S[DUMod]

where:

a[GN] =	1.021	a[GM] =	0.913	a[GU] =	1.044
e ≡	2.7183	Decay ≡	3.466		

List Statistics	[GN]	[GM]	[GU]
Number of Data Points (observations):	225	136	142
Geometric (log space) mean of b:	1.208E+00	2.651E+00	6.199E-01
Arithmetic (unit space) mean of b:	1.736E+00	4.060E+00	7.510E-01
Standard deviation of b:	1.849E+00	4.557E+00	6.566E-01
Coefficient of Variation (CV) b:	1.07	1.12	0.87
Arithmetic (unit space) mean of ε:	1.366E+00	1.510E+00	1.191E+00
Standard deviation of ε:	1.224E+00	1.623E+00	9.296E-01
Coefficient of Variation (CV) of ε:	0.90	1.07	0.78
Mean Magnitude of the Relative Error:	44%	50%	24%

New to Modified DSLOC Correlation:	2.570E-03
New to Unmodified DSLOC Growth Correlation:	3.025E-01

Growth Factor Estimating Relationships Behavior	*New DSLOC Growth*	*Modified DSLOC Growth*	*Unmodified DSLOC Growth*
Implied Growth Factor at data set mean baseline DSLOC:	53% at 59,443 DSLOC	11% at 22,934 DSLOC	7% at 251,323 DSLOC
Implied Growth Factor at data set geometric mean baseline DSLOC:	50% at 23,035 DSLOC	22% at 7,756 DSLOC	1% at 70,790 DSLOC

	Variables and Notation
Variables Used in JCDER and DSLOC Estimate Growth Equations	
S ≡	Size (ESLOC) distribution
E ≡	Core Effort (person-months) or (pm) distribution (data set SDLC)
T ≡	Core Duration (calendar months) or (cm) distribution (data set SDLC)
E[A-A] ≡	All-All Effort (person-months) or (pm) distribution (All-All SDLC)
T[A-A] ≡	All-All Duration (calendar months) or (cm) distribution (All-All SDLC)
alpha[E] ≡	Effort dimension exponent (economy/diseconomy of scale)
alpha[T] ≡	Duration dimension exponent (economy/diseconomy of scale)
alpha[S] ≡	Size dimension exponent (economy/diseconomy of scale)
D ≡	Difficulty (nonlinear person-month months per ESLOC) distribution -- time-sensitive inverse productivity
I ≡	Intensity (nonlinear person-months per calendar month) distribution -- staffing or burn
Eaf ≡	Effor Adjustment Factor (data set SDLC to All-All SDLC)
Daf ≡	Duration Adjustment Factor (data set SDLC to All-All SDLC)
K[S] ≡	Software Item (SI) being estimated to CSCI normalization factor for ESLOC
a[GN] ≡	New DSLOC estimating relationship exponent (economy/diseconomy of scale)
a[GM] ≡	Modified DSLOC estimating relationship exponent (economy/diseconomy of scale)
a[GU] ≡	Unmodified DSLOC estimating relationship exponent (economy/diseconomy of scale
ε[GN] ≡	Baseline New DSLOC estimate growth error factor distribution
ε[GM] ≡	Baseline Modified DSLOC estimate growth error factor distribution
ε[GU] ≡	Baseline Unmodified DSLOC estimate growth error factor distribution
Decay ≡	Decay Constant Parameter; default is 3.466 based on Boehm's "Cone of Uncertainty" (Boehm, 1981, p.311)
$\tilde{b}[GN], \tilde{b}[GM], \tilde{b}[GU]$ ≡	Geometric mean scale factor parameters for New, Modified, and Unmodified DSLOC growth estimating relationships
Maturity ≡	Estimate Maturity Parameter (e.g., SDLCBegin=0%; SyRR=10%; SWRR=20%; SwPDR=40%; SWCDR=60%; TRR=80%; SWAccept=100%)
K[N], K[M], K[U] ≡	Software Item (SI) to CSCI normalization factor for New, Mod, and Umod DSLOC
S[DNew] ≡	Technical Baseline Estimate (TBE) of New DSLOC
S[DMod] ≡	Technical Baseline Estimate (TBE) of Modified DSLOC
S[DUmod] ≡	Technical Baseline Estimate (TBE) of Unmodified DSLOC
S[DGANew] ≡	Growth-adjusted New DSLOC estimate distribution
S[DGAMod] ≡	Growth-adjusted Modified DSLOC estimate distribution
S[DGAUmod] ≡	Growth-adjusted Unmodified DSLOC estimate distribution
Notation Conventions	
'≡' is used to indicate identity; left expression is defined by right expression	
'≜' is used to indicate estimation; the right expression estimates the left expression	
Arial Bold Italic font indicates a random variable; range and distribution of possible outcomes	
Times New Roman Italic font indicates a specific-value (single-value) variable	
Bracks '[]' are used to indicate the subscripted characters of a variable name	

Appendix C DEGM8 vs. DEGM7 Implied Growth Factor Comparison

This appendix contains three scatter charts that show the implied growth percentage for each of New, Modified, and Unmodified DSLOC as a function of the given TBE DSLOC. We can summarize the information on these charts as follows:

- New DSLOC Implied Growth Percentage at SDLCBegin ($Maturity = 0\%$) for TBE DSLOC in the range 137 DSLOC to 858,984 DSLOC (225 observations)
    - DEGM8 Default: 34% to 62%
    - DEGM7: 75%
- Modified DSLOC Implied Growth Percentage at SDLCBegin ($Maturity = 0\%$) for TBE DSLOC in the range 10 DSLOC to 208,082 DSLOC (136 observations)
    - DEGM8 Default: 117% to -8%
    - DEGM7: 43%
- Unmodified DSLOC Implied Growth Percentage at SDLCBegin ($Maturity = 0\%$) for TBE in the range 1,100 DSLOC to 6,564,104 DSLOC (142 observations)
    - DEGM8 Default: -16% to 24%
    - DEGM7: 43%

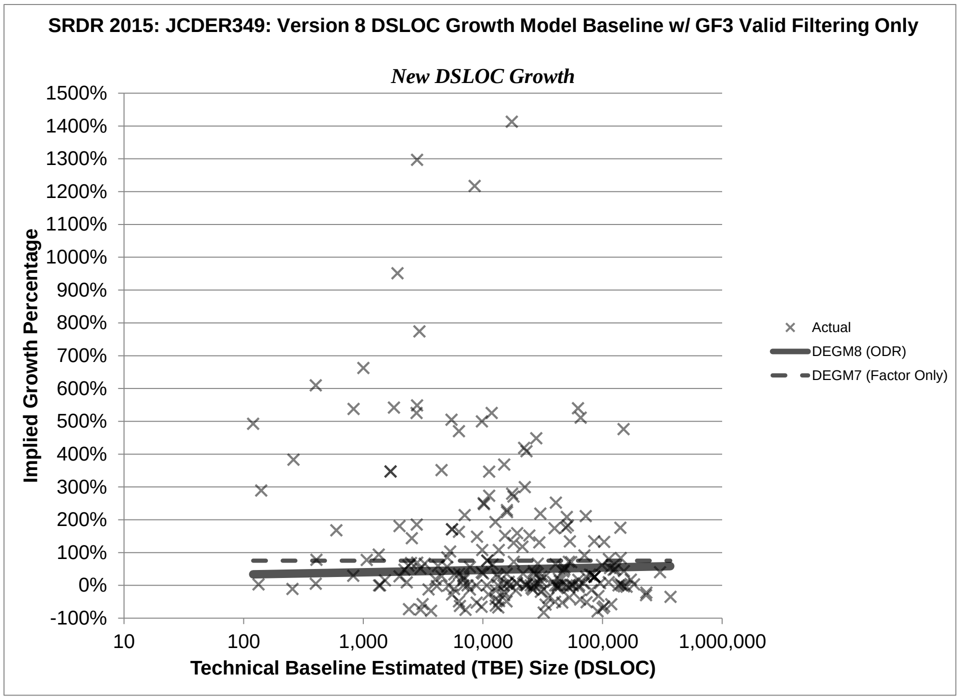


Figure 10 Filtered 2015 SRDR New DSLOC data regressed with ODR in log space (DEGM8 default) yields a better fit than does simply using the DEGM7 relationship.

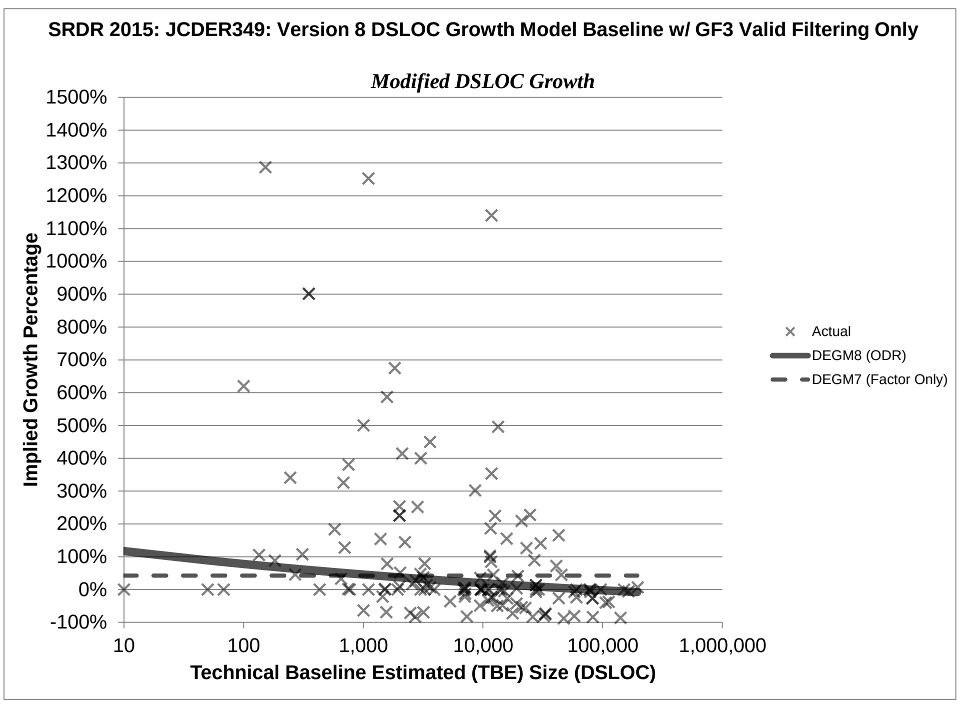


Figure 11 Filtered 2015 SRDR Modified DSLOC data regressed with ODR in log space (DEGM8 default) yields a better fit than does simply using the DEGM7 relationship.

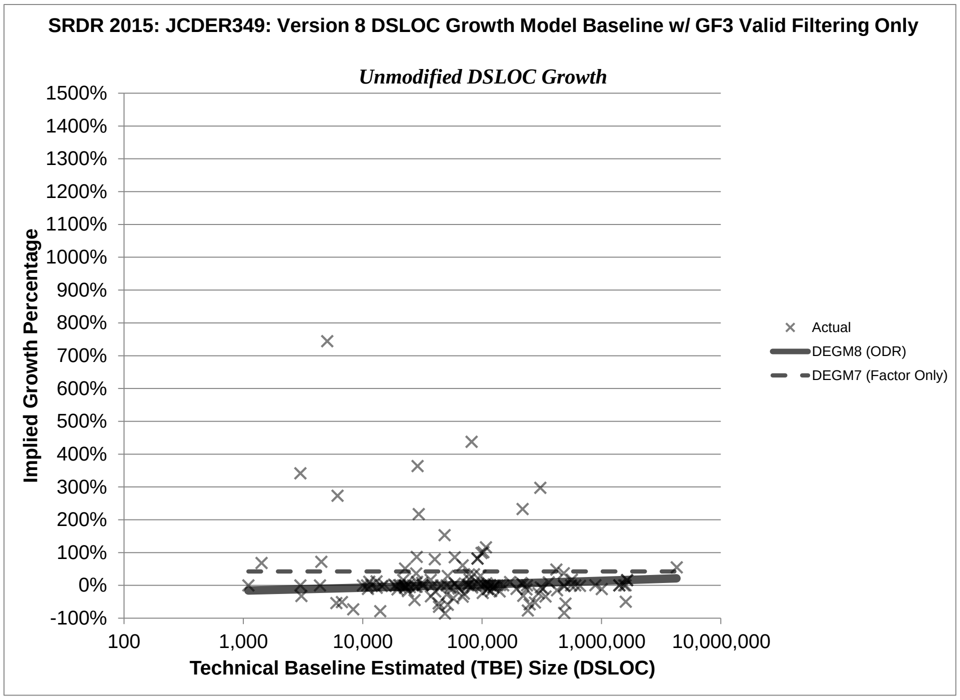


Figure 12 Filtered 2015 SRDR Unmodified DSLOC data regressed with ODR in log space (DEGM8 default) yields a better fit than does simply using the DEGM7 relationship.

Appendix D Orthogonal Distance Regression

Orthogonal Distance Regression (ODR) is the name given to the computational problem associated with finding the maximum likelihood estimators of parameters in measurement error models in the case of normally distributed errors.[14] For some data set $\mathbf{P}$ that contains n observations, each containing unique values for the same m observable parameters, ODR seeks to find a line in $\mathbb{R}^m$ space that minimizes the sum of the squared orthogonal (shortest) distances between each data point and that line. In other words, ODR is a process for finding a *best fit* line (an estimator) through a multi-dimension set of data points (observations).

Why is ODR better than Ordinary Least Squares (OLS) regression and its variants?

- ODR works in situations where there are more than two dimensions (measures) without making assumptions about which measures are dependent and which are independent.
- ODR acknowledges the existence of measurement error in all dimensions; not just in the dimension associated with a single variable deemed to be "dependent".

Data Sets

Let $\mathbf{P}$ be a set of data points (relevant measure observations) in $\mathbb{R}^m$ space. We can describe $\mathbf{P}$ as an $n \times m$ matrix of data point coordinate values

$$P_{i,j} \in \mathbf{P} \equiv \begin{bmatrix} P_{1,1} & P_{1,2} & \cdots & P_{1,m} \\ P_{2,1} & P_{2,2} & \cdots & P_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ P_{n,1} & P_{n,2} & \cdots & P_{n,m} \end{bmatrix} \tag{90}$$

where m is the number of dimensions (relevant measures) and where n is the number of observations (data points).

Data Set Centroid

Given the data set $\mathbf{P}$ of m relevant measures and n observations, we define the *centroid* point $C_{\mathbf{P}}$ of $\mathbf{P}$ with position vectors $\mathbf{r}_i$ as

Definition: Data Set Centroid

$$
\begin{aligned}
C_{\mathbf{P}} &\equiv \frac{1}{n}\sum_{i=1}^{n}\mathbf{r}_i \quad \Big|\, P_{i,j} \in \mathbf{P} \subseteq \mathbb{R}^m \\
&= \left(\frac{1}{n}\sum_{i=1}^{n}P_{i,1}, \frac{1}{n}\sum_{i=1}^{n}P_{i,2}, \ldots, \frac{1}{n}\sum_{i=1}^{n}P_{i,m}\right)
\end{aligned} \qquad \textbf{(91)}
$$

Orthogonal Distance Regression to Find the Best Fit Line

Orthogonal Distance Regression (ODR) (a special case of *Total Least Squares regression*) in m-dimension space ($\mathbb{R}^m$) can be used to determine the equation of a *best fit line* L_{ODR} according to the following definition:

Definition: ODR Best Fit Line

An ODR best fit line L_{ODR} in $\mathbb{R}^m$ space is one that minimizes the sum of the squared orthogonal distances $\sum_{i=1}^{n}\delta_i^2$ from L_{ODR} to each point P_i of a given set of data points $\mathbf{P}$ in $\mathbb{R}^m$.

One way to specify the best fit line L_{ODR} is to specify a particular point P' on L_{ODR} and to specify a direction vector $\mathbf{a}$; i.e., a vector that is collinear with or parallel to L_{ODR}.

Specifying a Point on the ODR Best Fit Line

We first specify an arbitrary reference point on L_{ODR} which we call P' with position vector $\mathbf{r}'$. We next identify each of points $\tilde{P}_i$ on L_{ODR} with associated position vectors $\tilde{\mathbf{r}}_i$ from which we measure the orthogonal distance to each corresponding data point P_i with position vector $\mathbf{r}_i$. In this appendix we use the *overstrike tilde* ($\sim$) notation on a point or a position vector to indicate an *estimated* value[15]. In this case $\tilde{P}_i$ is an estimate of its corresponding actual data point P_i since, practically speaking, if L_{ODR} represents the best fit line and $\tilde{P}_i$ is, by definition, on L_{ODR}, then $\tilde{P}_i$ represents the best estimate of actual outcome P_i. This implies each vector $\overrightarrow{\tilde{P}_iP_i}$ is orthogonal to L_{ODR} and, hence, any one of these vectors qualifies as a normal vector of L_{ODR}. Since each vector $\overrightarrow{P'\tilde{P}_i}$ is obviously on L_{ODR}, it follows that each $\overrightarrow{P'\tilde{P}_i}$ is orthogonal to each $\overrightarrow{\tilde{P}_iP_i}$.

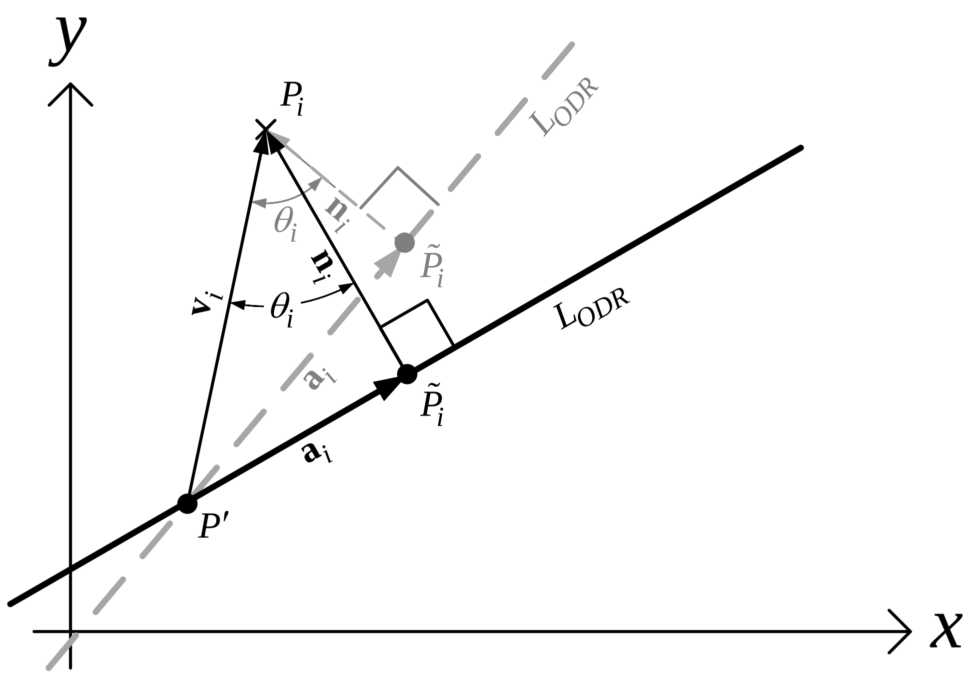


Figure 13 Geometry of the orthogonal distance problem; maximizing the sum of the squared angles θ_i serves to minimize the sum of the squared orthogonal distances $\|\mathbf{n}_i\|$

The previous paragraph and Figure 13 describe a right triangle in $\mathbb{R}^m$ space with base leg $\mathbf{a}_i \equiv \overrightarrow{P'\tilde{P}_i}$ (on L_{ODR}), normal leg $\mathbf{n}_i \equiv \overrightarrow{\tilde{P}_i P_i}$ (orthogonal to L_{ODR}), and hypotenuse $\mathbf{v}_i \equiv \overrightarrow{P'P_i}$. By the angle between vectors definition

Definition: Angle Between Vectors

$$\begin{aligned} \cos\theta &= \frac{\mathbf{u}\bullet\mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|} \\ &= \frac{\sum_{i=1}^{m} u_i v_i}{\sqrt{\sum_{i=1}^{m} u_1^{\ 2}}\sqrt{\sum_{i=1}^{m} v_1^{\ 2}}} \end{aligned} \tag{92}$$

and the dot product associativity theorem,

Theorem: Dot Product Associativity

$$(c\mathbf{u})\bullet\mathbf{v} = c(\mathbf{u}\bullet\mathbf{v}) \tag{93}$$

we define the angle θ_i between the unit vector of any normal leg $\hat{\mathbf{n}}$ [16] and each hypotenuse vector $\mathbf{v}_i$ to be

$$\begin{gathered}\theta_i \;\middle|\; \cos\theta_i = \frac{\hat{\mathbf{n}}\bullet\mathbf{v}_i}{\|\hat{\mathbf{n}}\|\|\mathbf{v}_i\|} = \frac{\hat{\mathbf{n}}\bullet\mathbf{v}_i}{\|\mathbf{v}_i\|} = \frac{1}{\|\mathbf{v}_i\|}\left(\hat{\mathbf{n}}\bullet\mathbf{v}_i\right) \\ \therefore \hat{\mathbf{n}}\bullet\mathbf{v}_i = \|\mathbf{v}_i\|\cos\theta_i\end{gathered} \tag{94}$$

Substituting vector $\mathbf{v}_i$ in Equation (94) with its equivalent position vector difference $\mathbf{r}_i - \mathbf{r}'$ per the definition of a direction vector

Definition: Direction Vector

$$\mathbf{v} = k\left(\mathbf{r}'' - \mathbf{r}'\right) \;\middle|\; \mathbf{v},\mathbf{r}',\mathbf{r}'' \in \mathbb{R}^m;\, k \in \mathbb{R} \tag{95}$$

gives us

$$\hat{\mathbf{n}}\bullet\left(\mathbf{r}_i - \mathbf{r}'\right) = \|\mathbf{r}_i - \mathbf{r}'\|\cos\theta_i \tag{96}$$

By the dot product distributivity theorem

Theorem: Dot Product Distributivity

$$\mathbf{u}\bullet\left(\mathbf{v}+\mathbf{w}\right) = \mathbf{u}\bullet\mathbf{v} + \mathbf{u}\bullet\mathbf{w} \tag{97}$$

we can rewrite Equation (96) as

$$\begin{gathered}\hat{\mathbf{n}}\bullet\mathbf{r}_i - \hat{\mathbf{n}}\bullet\mathbf{r}' = \|\mathbf{r}_i - \mathbf{r}'\|\cos\theta_i \\ \hat{\mathbf{n}}\bullet\mathbf{r}' = \hat{\mathbf{n}}\bullet\mathbf{r}_i - \|\mathbf{r}_i - \mathbf{r}'\|\cos\theta_i\end{gathered} \tag{98}$$

If we let $d \equiv \hat{\mathbf{n}}\bullet\mathbf{r}_i - \|\mathbf{r}_i - \mathbf{r}'\|\cos\theta_i$ then we can say

$$d = \hat{\mathbf{n}}\bullet\mathbf{r}' \tag{99}$$

From the geometry of Figure 13 above, it should be obvious that minimizing the sum of the squared magnitudes of each vector $\mathbf{n}_i$ can be accomplished by maximizing the sum of each squared θ_i which, by inspection of Equations (94) and (99), implies minimizing the sum of each squared d_i. We have already specified the locations of points P_i (the given data points) and point P' (the arbitrary reference point on L_{ODR}) which necessarily specifies the magnitudes of vectors $\mathbf{v}_i$. Because each vector $\mathbf{v}_i$ is the hypotenuse of a right triangle, the magnitude of each normal leg $\|\mathbf{n}_i\|$ can be determined as the magnitude of the vector projection of each $\mathbf{v}_i$ onto the unit vector $\hat{\mathbf{n}}$. The unit vector $\hat{\mathbf{n}}$, per the vector unitizing theorem,

Theorem: Unitizing a Vector

$$\hat{\mathbf{u}} = \frac{1}{\|\mathbf{v}\|}\mathbf{v} \tag{100}$$

is equal to $\mathbf{n}_i / \|\mathbf{n}_i\|$. The magnitude $\|\mathbf{n}_i\|$ of each vector $\mathbf{n}_i$ can therefore be specified as

Definition: Normal Vector Magnitude

$$\|\mathbf{n}_i\| = \left\| \boldsymbol{proj}_{\mathbf{n}_i}(\mathbf{v}_i) \right\| \quad | \ \mathbf{n}_i \equiv \frac{\mathbf{n}_i}{\|\mathbf{n}_i\|} \tag{101}$$

By the vector projection definition

Definition: Vector Projection

$$\boldsymbol{proj}_{\mathbf{u}}(\mathbf{v}) = \left(\frac{\mathbf{u} \bullet \mathbf{v}}{\mathbf{u} \bullet \mathbf{u}} \right) \mathbf{u} \tag{102}$$

we can rewrite Definition (101) as

$$\|\mathbf{n}_i\| = \left\| \frac{\hat{\mathbf{n}} \bullet \mathbf{v}_i}{\hat{\mathbf{n}} \bullet \hat{\mathbf{n}}} \hat{\mathbf{n}} \right\| \tag{103}$$

Substituting vector $\mathbf{v}_i$ in Equation (103) with its equivalent position vector difference $\mathbf{r}_i - \mathbf{r}'$ per Definition (95) gives us

$$\|\mathbf{n}_i\| = \left\| \frac{\hat{\mathbf{n}} \bullet (\mathbf{r}_i - \mathbf{r}')}{\hat{\mathbf{n}} \bullet \hat{\mathbf{n}}} \hat{\mathbf{n}} \right\| \tag{104}$$

By the dot product associativity Theorem (97), Equation (104) becomes

$$\|\mathbf{n}_i\| = \left\| \frac{\hat{\mathbf{n}} \bullet \mathbf{r}_i - \hat{\mathbf{n}} \bullet \mathbf{r}'}{\hat{\mathbf{n}} \bullet \hat{\mathbf{n}}} \hat{\mathbf{n}} \right\| \tag{105}$$

We next substitute $\hat{\mathbf{n}} \bullet \hat{\mathbf{n}}$ in Equation (105) with its equivalent $\|\hat{\mathbf{n}}\|^2$ implied by the definition of vector magnitude

Definition: Vector Magnitude

$$\begin{aligned} \|\mathbf{v}\| &= \sqrt{\mathbf{v} \bullet \mathbf{v}} \\ &= \sqrt{v_1^2 + v_2^2 + \dots + v_m^2} \\ &= \sqrt{\sum_{j=1}^{m} v_j^2} \end{aligned} \tag{106}$$

which yields

$$\|\mathbf{n}_i\| = \left\| \left(\frac{\hat{\mathbf{n}} \bullet \mathbf{r}_i - \hat{\mathbf{n}} \bullet \mathbf{r}'}{\|\hat{\mathbf{n}}\|^2} \right) (\hat{\mathbf{n}}) \right\| \tag{107}$$

Observing that the quantities $\hat{\mathbf{n}} \bullet \mathbf{r}_i$, $\hat{\mathbf{n}} \bullet \mathbf{r}'$, and $\|\hat{\mathbf{n}}\|^2$ all evaluate to scalar values, we can use the vector magnitude factoring theorem

Theorem: Vector Magnitude Factoring

$$\|c\mathbf{v}\| = c\|\mathbf{v}\| \tag{108}$$

to rewrite Equation (107) as

$$\begin{aligned} \|\mathbf{n}_i\| &= \frac{\hat{\mathbf{n}} \bullet \mathbf{r}_i - \hat{\mathbf{n}} \bullet \mathbf{r}'}{\|\hat{\mathbf{n}}\|^2} \|\hat{\mathbf{n}}\| \\ &= \frac{\hat{\mathbf{n}} \bullet \mathbf{r}_i - \hat{\mathbf{n}} \bullet \mathbf{r}'}{\|\hat{\mathbf{n}}\|} \\ &= \hat{\mathbf{n}} \bullet \mathbf{r}_i - \hat{\mathbf{n}} \bullet \mathbf{r}' \end{aligned} \tag{109}$$

Substituting $\hat{\mathbf{n}} \bullet \mathbf{r}'$ in Equation (109) with its equivalent d in Equation (99) yields

$$\|\mathbf{n}_i\| = \hat{\mathbf{n}} \bullet \mathbf{r}_i - d \tag{110}$$

The sum of the squared normal vector magnitudes can now be written as

$$\sum_{i=1}^{n} \left\| \mathbf{n}_i \right\|^2 = \sum_{i=1}^{n} \left(\hat{\mathbf{n}} \bullet \mathbf{r}_i - d \right)^2 \tag{111}$$

We define a function $\boldsymbol{f}$ to represent our sum of squared vector magnitudes

$$\begin{aligned} \boldsymbol{f}\left(\hat{\mathbf{n}}, d \mid \mathbf{r} \right) &\equiv \sum_{i=1}^{n} \left\| \mathbf{n}_i \right\|^2 \\ &= \sum_{i=1}^{n} \left(\hat{\mathbf{n}} \bullet \mathbf{r}_i - d \right)^2 \end{aligned} \tag{112}$$

We can solve for the value of each d that minimizes $\boldsymbol{f}$ by setting the partial derivative of $\boldsymbol{f}$ with respect to d in Equation (112) to zero. This yields

$$\frac{\partial}{\partial d} \boldsymbol{f} = \frac{\partial}{\partial d} \sum_{i=1}^{n} \left(\hat{\mathbf{n}} \bullet \mathbf{r}_i - d \right)^2 = 0 \tag{113}$$

To solve for d we first expand the square quantity within the summation.

$$\frac{\partial}{\partial d} \sum_{i=1}^{n} \left(\left(\hat{\mathbf{n}} \bullet \mathbf{r}_i \right)^2 - 2\left(\hat{\mathbf{n}} \bullet \mathbf{r}_i \right) d + d^2 \right) = 0 \tag{114}$$

Next we apply the summation to each term within the summation and factor out the scalars.

$$\frac{\partial}{\partial d} \left(\sum_{i=1}^{n} \left(\hat{\mathbf{n}} \bullet \mathbf{r}_i \right)^2 - 2d \sum_{i=1}^{n} \left(\hat{\mathbf{n}} \bullet \mathbf{r}_i \right) + d^2 n \right) = 0 \tag{115}$$

Taking the partial derivative of each term with respect to d gives us

$$\frac{\partial}{\partial d} \left(\sum_{i=1}^{n} \left(\hat{\mathbf{n}} \bullet \mathbf{r}_i \right)^2 \right) - \frac{\partial}{\partial d} \left(2d \sum_{i=1}^{n} \hat{\mathbf{n}} \bullet \mathbf{r}_i \right) + \frac{\partial}{\partial d} \left(d^2 n \right) = 0 \tag{116}$$

Resolving the partial derivatives yields

$$\begin{gathered} \frac{\partial}{\partial d} \left(\sum_{i=1}^{n} \left(\hat{\mathbf{n}} \bullet \mathbf{r}_i \right)^2 \right) - \frac{\partial}{\partial d} \left(2d \sum_{i=1}^{n} \hat{\mathbf{n}} \bullet \mathbf{r}_i \right) + \frac{\partial}{\partial d} \left(d^2 n \right) = 0 \\ \therefore -2 \sum_{i=1}^{n} \hat{\mathbf{n}} \bullet \mathbf{r}_i + 2dn = 0 \end{gathered} \tag{117}$$

Since $\hat{\mathbf{n}}$ is a constant vector (i.e., is the same for each $\mathbf{r}_i$), we can use the dot product distributivity Theorem (97) to remove $\hat{\mathbf{n}}$ from the summation to get

$$-2\hat{\mathbf{n}}\bullet\sum_{i=1}^{n}\mathbf{r}_i + 2dn = 0 \tag{118}$$

Multiplying the summation by $\frac{n}{n}$ (the equivalent of 1) gives us

$$\begin{gathered}-2\hat{\mathbf{n}}\bullet\frac{n}{n}\sum_{i=1}^{n}\mathbf{r}_i + 2dn = 0\\ \therefore -2\hat{\mathbf{n}}\bullet n\frac{1}{n}\sum_{i=1}^{n}\mathbf{r}_i + 2dn = 0\end{gathered} \tag{119}$$

We observe that Equation (119) contains an expression $(1/n)\sum_{i=1}^{n}\mathbf{r}_i$ that is equivalent to the data set centroid $C_{\mathbf{P}}$ in Definition(91). Making the substitution yields

$$-2\hat{\mathbf{n}}\bullet nC_{\mathbf{P}} + 2dn = 0 \tag{120}$$

We now apply dot product commutativity

Theorem: Dot Product Commutativity

$$\mathbf{u}\bullet\mathbf{v} = \mathbf{v}\bullet\mathbf{u} \tag{121}$$

and dot product associativity Theorem (93) to Equation(120), the result being

$$-2n\left(\hat{\mathbf{n}}\bullet C_{\mathbf{P}}\right) + 2dn = 0 \tag{122}$$

Finally, solving for d gives us

$$\begin{gathered}-2n\left(\hat{\mathbf{n}}\bullet C_{\mathbf{P}}\right) + 2dn = 0\\ \therefore d = \hat{\mathbf{n}}\bullet C_{\mathbf{P}}\end{gathered} \tag{123}$$

Comparing Equation (123) to Equation (99) leads to the conclusion that the arbitrary reference point P' on L_{ODR} and with position vector $\mathbf{r}'$ is equal to the data set centroid $C_{\mathbf{P}}$ which proves

Theorem: The ODR Best Fit Line Contains the Data Set Centroid

$$C_{\mathbf{P}} \in L_{ODR} \tag{124}$$

Once again, a way to specify L_{ODR} is to specify a particular point P' on the line and to specify a direction vector $\mathbf{a}$; i.e., a vector that is collinear with or parallel to the line. Substituting the appropriate vector components for $\mathbf{a}$, $\tilde{\mathbf{r}}_i$, and $C_{\mathbf{P}}$ into the generalized vector equation of a line in multi-dimension space

Theorem: Vector Equation of a Line in m-Dimension Space

$$\mathbf{r} = \mathbf{r}' + t\mathbf{a} \quad \left| \begin{array}{l} \mathbf{r}, \mathbf{r}', \mathbf{a} \in \mathbb{R}^m \\ t \in \mathbb{R} \end{array} \right. \tag{125}$$

gives us

Theorem: The ODR Best Fit Line, Vector Form

$$\tilde{\mathbf{r}} = C_{\mathbf{P}} + t\mathbf{a} \Leftrightarrow \tilde{\mathbf{r}} \in L_{ODR} \quad \left| \begin{array}{l} L_{ODR} \subset \mathbb{R}^m \\ \tilde{\mathbf{r}}, C_{\mathbf{P}}, \mathbf{a} \in \mathbb{R}^m \\ t \in \mathbb{R} \end{array} \right. \tag{126}$$

Substituting the vector component form

Definition: Vector Components and Notations

$$\begin{aligned} \mathbf{v} &= [v_1, v_2, \ldots, v_m] \\ &= [v_1 \quad v_2 \quad \cdots \quad v_m] \\ &= \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{bmatrix} \end{aligned} \tag{127}$$

of vectors $\tilde{\mathbf{r}}$, $C_{\mathbf{P}}$, and $\mathbf{a}$ into the equation of Theorem (126) gives us

$$\begin{bmatrix} \tilde{r}_1 \\ \tilde{r}_2 \\ \vdots \\ \tilde{r}_m \end{bmatrix} = \begin{bmatrix} C_{\mathbf{P}1} \\ C_{\mathbf{P}2} \\ \vdots \\ C_{\mathbf{P}m} \end{bmatrix} + t \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} \quad \textbf{(128)}$$

The parametric form equation of L_{ODR} is therefore

Theorem: ODR Best Fit Line, Parametric Form

$$\begin{cases} \tilde{r}_1 = C_{\mathbf{P}1} + ta_1 \\ \tilde{r}_2 = C_{\mathbf{P}2} + ta_2 \\ \quad \vdots \\ \tilde{r}_m = C_{\mathbf{P}m} + ta_m \end{cases} \quad \textbf{(129)}$$

The scalar or implicit form equation of L_{ODR} is therefore

Theorem: ODR Best Fit Line, Symmetric Form

$$\left(\frac{\tilde{r}_1 - C_{\mathbf{P}1}}{a_1} \right) = \left(\frac{\tilde{r}_2 - C_{\mathbf{P}2}}{a_2} \right) = \cdots = \left(\frac{\tilde{r}_m - C_{\mathbf{P}m}}{a_m} \right) \quad \textbf{(130)}$$

Using Singular Value Decomposition to Specify a Direction Vector for the ODR Best Fit Line

Since we can easily specify a particular point on the ODR best fit line L_{ODR} by computing the centroid $C_{\mathbf{P}}$ from the values in our data set $\mathbf{P}$ using Definition (91)

$$\begin{aligned} C_{\mathbf{P}} &\equiv \frac{1}{n} \sum_{i=1}^{n} \mathbf{r}_i \\ &= \left(\frac{1}{n} \sum_{i=1}^{n} P_{i,1}, \frac{1}{n} \sum_{i=1}^{n} P_{i,2}, \ldots, \frac{1}{n} \sum_{i=1}^{n} P_{i,m} \right) \end{aligned} \quad \textbf{(131)}$$

we need only find values for the direction vector $\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}$ in order to fully specify L_{ODR}.

We can find these values by using Principal Component Analysis (PCA) to perform a principal component transformation on our data set matrix $\mathbf{P}$. PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by some projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. [17] It turns out that the first principal component axis has the same direction as does L_{ODR}. One way to perform a principal component transformation is to use the *Singular Value Decomposition* (SVD). The SVD is a special matrix factorization that, when applied to a data-set-centroid-centered version our data set matrix $\mathbf{P}$, will yield, among other things, the components of the L_{ODR} direction vector $\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}$. To apply SVD, we first center our data set matrix $\mathbf{P}$ about the centroid point $C_{\mathbf{P}}$ to create an $n \times m$ matrix $\mathbf{M}$; i.e., we create $\mathbf{M}$ as the differences in each dimension between each data point P_i and the data set centroid $C_{\mathbf{P}}$ such that

$$\mathbf{M} \equiv \begin{bmatrix} P_{1,1} - C_{\mathbf{P}_1} & P_{1,2} - C_{\mathbf{P}_2} & \cdots & P_{1,m} - C_{\mathbf{P}_m} \\ P_{2,1} - C_{\mathbf{P}_1} & P_{2,2} - C_{\mathbf{P}_2} & \cdots & P_{2,m} - C_{\mathbf{P}_m} \\ \vdots & \vdots & \ddots & \vdots \\ P_{n,1} - C_{\mathbf{P}_1} & P_{n,2} - C_{\mathbf{P}_2} & \cdots & P_{n,m} - C_{\mathbf{P}_m} \end{bmatrix} \tag{132}$$

The SVD of matrix $\mathbf{M}$ is a factorization of $\mathbf{M}$ such that

$$\boldsymbol{SVD}(\mathbf{M}) \equiv \{\mathbf{U}, \boldsymbol{\Sigma}, \mathbf{V}^{\mathsf{T}}\} \quad \Big| \, \mathbf{M} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathsf{T}} \tag{133}$$

where

$\mathbf{M}$	$\equiv$	Given centroid-centered data set matrix $(n \times m)$
$\mathbf{U}$	$\equiv$	Orthogonal matrix $(n \times p)$; $p = \boldsymbol{min}(n, m)$; not used as part of the ODR process
$\boldsymbol{\Sigma}$	$\equiv$	Square diagonal matrix $(p \times p)$; singular values
$\mathbf{V}$	$\equiv$	Orthogonal matrix $(m \times p)$; singular vectors organized by row; $\mathbf{V}^{\mathsf{T}}$ contains singular vectors organized by column

$$\mathbf{M} = \mathbf{U}\begin{bmatrix} \Sigma_{1,1} & 0 & \cdots & 0 \\ 0 & \Sigma_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma_{p,p} \end{bmatrix} \begin{bmatrix} V_{1,1} & V_{1,2} & \cdots & V_{1,3} \\ V_{2,1} & V_{2,2} & \cdots & V_{2,3} \\ \vdots & \vdots & \ddots & \vdots \\ V_{3,1} & V_{3,2} & \cdots & V_{m,p} \end{bmatrix}^{\mathrm{T}}$$

$$\mathbf{M} = \mathbf{U}\begin{bmatrix} \Sigma_{max} & 0 & \cdots & 0 \\ 0 & \Sigma_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma_{p,p} \end{bmatrix} \begin{bmatrix} a_1 & V_{1,2} & \cdots & V_{1,3} \\ a_2 & V_{2,2} & \cdots & V_{2,3} \\ \vdots & \vdots & \ddots & \vdots \\ a_m & V_{m,2} & \cdots & V_{m,p} \end{bmatrix}^{\mathrm{T}} \quad \textbf{(134)}$$

Note that a feature of the algorithm being used to implement SVD is that it returns $\mathbf{V}$ such that its contained vectors (columns) are sorted in descending singular value order from left to right. PCA asserts that the ODR best fit line's direction vector $\mathbf{a}$ is equal to the direction of PCA's first principal component axis which is the singular vector of $\mathbf{M}$ that corresponds to its largest singular value; in this implementation, the leftmost column vector in $\mathbf{V}$. Therefore

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} V_{1,1} \\ V_{2,1} \\ \vdots \\ V_{3,1} \end{bmatrix} \quad \textbf{(135)}$$

In summary, we can find the system of equations that define an ODR best fit line through any multi-dimension commensurable and scale-invariant data set by first finding the data set centroid, then centering the data set matrix about the data space origin, and finally applying the SVD to the centered matrix. The resulting ODR best fit line is specified by a known point on the line (the data set centroid) and a direction vector (the column of the singular vector matrix that is associated with the largest singular value in the singular value matrix).

Appendix E TRI ACEIT Implementation of the Example Estimate

The following two tables show results from the DEGM8 example estimate described in the body of this paper. They compare DEGM8 idealized values produced by an MS Excel workbook against the output of an implementation of the DEGM8 example estimate in the TRI ACEIT tool.

Table 4 Example estimate DSLOC values comparison; MS Excel idealized from data set versus ACEIT risk output; differences in Total DSLOC because MS Excel assumes no correlation between summands (independence) while ACEIT uses given DEGM8 correlation values (group strength) between each summand in a Monte Carlo convolution (partial dependence). Notice that the Total DSLOC mean values are very close to each other which is expected since probability theory proves that the sum of the means is equal to the mean of the sum.

	TBE DSLOC	MS Excel Geomean DSLOC	ACEIT Point Estimate Position DSLOC	MS Excel Median DSLOC	ACEIT Median DSLOC	MS Excel Mean DSLOC	ACEIT Mean DSLOC
Total DSLOC	175,000	183,600	183,703	190,325	192,755	213,666	213,672
Attainment Probability	#N/A	#N/A	30%	#N/A	50%	#N/A	68%
Implied Growth	0%	5%	5%	9%	10%	22%	22%
New DSLOC	25,000	31,217	31,227	33,769	33,768	38,091	38,090
Attainment Probability	28%	41%	41%	50%	50%	62%	62%
Implied Growth	0%	25%	25%	35%	35%	52%	52%
Total DSLOC	50,000	50,965	50,970	54,342	54,344	64,226	64,227
Attainment Probability	39%	43%	43%	50%	50%	75%	75%
Implied Growth	0%	2%	2%	9%	9%	28%	28%
Unmodified DSLOC	100,000	101,418	101,506	102,213	102,213	111,350	111,355
Attainment Probability	39%	46%	47%	50%	50%	77%	76%
Implied Growth	0%	1%	2%	2%	2%	11%	11%

Table 5 Example estimate CDFs comparison; MS Excel idealized from data set versus ACEIT risk output; differences in Total DSLOC table because MS Excel assumes no correlation between summands (independence) while ACEIT uses given DEGM8 correlation values (group strength) between each summand in a Monte Carlo convolution (partial dependence).

Total DSLOC			New DSLOC			Modified DSLOC			Unmodified DSLOC		
%ile	MS Excel CDF (DSLOC) (no correl)	ACE CDF (DSLOC) (w/ correl)	%ile	MS Excel CDF (DSLOC)	ACE CDF (DSLOC)	%ile	MS Excel CDF (DSLOC)	ACE CDF (DSLOC)	%ile	MS Excel CDF (DSLOC)	ACE CDF (DSLOC)
5	113,123	144,611	5	16,767	16,767	5	30,142	30,143	5	66,214	66,213
10	126,853	154,073	10	18,002	18,002	10	31,988	31,988	10	76,863	76,862
15	140,940	160,765	15	20,367	20,367	15	34,662	34,661	15	85,911	85,918
20	149,939	166,549	20	22,064	22,063	20	36,544	36,545	20	91,331	91,329
25	158,770	170,839	25	23,542	23,542	25	40,076	40,076	25	95,153	95,151
30	168,602	175,190	30	26,077	26,085	30	44,440	44,444	30	98,086	98,088
35	176,094	179,519	35	28,673	28,673	35	48,050	48,049	35	99,372	99,371
40	180,591	184,028	40	30,255	30,255	40	50,109	50,109	40	100,228	100,227
45	186,030	188,374	45	32,652	32,652	45	52,038	52,038	45	101,339	101,339
50	190,325	192,755	50	33,769	33,768	50	54,342	54,344	50	102,213	102,213
55	193,776	197,488	55	35,318	35,318	55	55,616	55,616	55	102,842	102,842
60	197,515	202,922	60	37,473	37,474	60	56,205	56,204	60	103,838	103,838
65	204,400	208,970	65	39,468	39,468	65	58,531	58,529	65	106,401	106,401
70	209,437	218,239	70	40,242	40,243	70	60,796	60,796	70	108,399	108,399
75	215,360	230,280	75	40,882	40,882	75	64,446	64,449	75	110,033	110,033
80	226,514	248,547	80	43,097	43,097	80	68,448	68,435	80	114,969	114,974
85	250,747	272,034	85	47,510	47,510	85	81,551	81,546	85	121,686	121,690
90	333,950	307,845	90	62,711	62,723	90	129,599	129,588	90	141,640	141,656
95	434,026	364,916	95	84,782	84,782	95	169,400	169,403	95	179,843	179,837

The remaining tables in this appendix illustrate an implementation of the DEGM8 in the TRI ACEIT software tool that models the example estimate. These tables are listed in the order in which they must appear in the AECIT file. The order of computation is from last appearing to first appearing; each table referencing necessary predecessor tables through relative (indexed) addressing supported by the index values in the ***Ri*** and ***i*** columns of each referencing table. The ***Equation / Throughput*** column holds ACEIT expressions that represent various parts of the DEGM8 equations.

Table 6 Growth Adjusted Accumulated DSLOC Distributions Table

WBS/CES Description	Unique ID	Equation / Throughput	WBS Indent Level	PE Position in Distribution	Distribution Form	Grouping	Group Strength	Random Seed	CDF Keyword	Ri (+) Rev Row Index	i (+) Row Index
*EXAMPLE - DSLOC Estimate Growth Model v8 (DEGM8)											
* Growth Adjusted Accumulated DSLOC Distributions (includes Redelivery) (combined)											
*THIS TABLE MUST BE IN REVERSE ORDER											
Table Header Row -- DO NOT DELETE (table rows must be at WBS Level 2)	GADSLOCTbl		1								
Example Software Item with Unmodified DSLOC Only		FYTot(@GMSdNTbl+Ri)+FYTot(@GMSdMTbl+Ri)+FYTot(@GMSdUTbl+Ri)+FYIVal(@STbl+Ri,2033)*(if(FYIVal(@STbl+Ri,2027)=0,0,FYTot(@GADSLOCTbl+FYIVal(@STbl+Ri,2027)))-FYIVal(@STbl+Ri,2028))	2							4	1
Example Software Item with Modified DSLOC Only		FYTot(@GMSdNTbl+Ri)+FYTot(@GMSdMTbl+Ri)+FYTot(@GMSdUTbl+Ri)+FYIVal(@STbl+Ri,2033)*(if(FYIVal(@STbl+Ri,2027)=0,0,FYTot(@GADSLOCTbl+FYIVal(@STbl+Ri,2027)))-FYIVal(@STbl+Ri,2028))	2							3	2
Example Software Item with New DSLOC Only		FYTot(@GMSdNTbl+Ri)+FYTot(@GMSdMTbl+Ri)+FYTot(@GMSdUTbl+Ri)+FYIVal(@STbl+Ri,2033)*(if(FYIVal(@STbl+Ri,2027)=0,0,FYTot(@GADSLOCTbl+FYIVal(@STbl+Ri,2027)))-FYIVal(@STbl+Ri,2028))	2							2	3
Example Software Item with New, Modified, and Unmodified DSLOC		FYTot(@GMSdNTbl+Ri)+FYTot(@GMSdMTbl+Ri)+FYTot(@GMSdUTbl+Ri)+FYIVal(@STbl+Ri,2033)*(if(FYIVal(@STbl+Ri,2027)=0,0,FYTot(@GADSLOCTbl+FYIVal(@STbl+Ri,2027)))-FYIVal(@STbl+Ri,2028))	2							1	4

Table 7 Growth/Maturity-Adjusted New DSLOC Distributions Table

WBS/CES Description	Unique ID	Equation / Throughput	WBS Indent Level	PE Position in Distribution	Distribution Form	Grouping	Group Strength	Random Seed	CDF Keyword	Ri (+) Rev Row Index	i (+) Row Index
* Growth/Maturity-Adjusted New DSLOC Distributions											
Table Header Row -- DO NOT DELETE (table rows must be at WBS Level 2)	GMSdNTbl		1								
Example Software Item with New, Modified, and Unmodified DSLOC		exp(-3.466*FYIVal(@STbl+i,2012))*(FYTot(@GSdNTbl+i)-FYIVal(@STbl+i,2011))+FYIVal(@STbl+i,2011) [DEGM8]	2							4	1
Example Software Item with New DSLOC Only		exp(-3.466*FYIVal(@STbl+i,2012))*(FYTot(@GSdNTbl+i)-FYIVal(@STbl+i,2011))+FYIVal(@STbl+i,2011) [DEGM8]	2							3	2
Example Software Item with Modified DSLOC Only		exp(-3.466*FYIVal(@STbl+i,2012))*(FYTot(@GSdNTbl+i)-FYIVal(@STbl+i,2011))+FYIVal(@STbl+i,2011) [DEGM8]	2							2	3
Example Software Item with Unmodified DSLOC Only		exp(-3.466*FYIVal(@STbl+i,2012))*(FYTot(@GSdNTbl+i)-FYIVal(@STbl+i,2011))+FYIVal(@STbl+i,2011) [DEGM8]	2							1	4

Table 8 Growth/Maturity-Adjusted Modified DSLOC Distributions Table

WBS/CES Description	*Unique ID*	*Equation / Throughput*	*WBS Indent Level*	*PE Position in Distribution*	*Distribution Form*	*Grouping*	*Group Strength*	*Random Seed*	*CDF Keyword*	*Ri (+) Rev Row Index*	*i (+) Row Index*
* Growth/Maturity-Adjusted Modified DSLOC Distributions											
Table Header Row -- DO NOT DELETE (table rows must be at WBS Level 2)	GMSdMTbl		1								
Example Software Item with New, Modified, and Unmodified DSLOC		exp(-3.466*FYIVal(@STbl+i,2015))*(FYTot(@GSdMTbl+i)-(FYIVal(@STbl+i,2013)-FYIVal(@STbl+i,2014)))+(FYIVal(@STbl+i,2013)-FYIVal(@STbl+i,2014)) [DEGM8]	2							4	1
Example Software Item with New DSLOC Only		exp(-3.466*FYIVal(@STbl+i,2015))*(FYTot(@GSdMTbl+i)-(FYIVal(@STbl+i,2013)-FYIVal(@STbl+i,2014)))+(FYIVal(@STbl+i,2013)-FYIVal(@STbl+i,2014)) [DEGM8]	2							3	2
Example Software Item with Modified DSLOC Only		exp(-3.466*FYIVal(@STbl+i,2015))*(FYTot(@GSdMTbl+i)-(FYIVal(@STbl+i,2013)-FYIVal(@STbl+i,2014)))+(FYIVal(@STbl+i,2013)-FYIVal(@STbl+i,2014)) [DEGM8]	2							2	3
Example Software Item with Unmodified DSLOC Only		exp(-3.466*FYIVal(@STbl+i,2015))*(FYTot(@GSdMTbl+i)-(FYIVal(@STbl+i,2013)-FYIVal(@STbl+i,2014)))+(FYIVal(@STbl+i,2013)-FYIVal(@STbl+i,2014)) [DEGM8]	2							1	4

Table 9 Growth/Maturity-Adjusted Unmodified DSLOC Distributions Table

WBS/CES Description	*Unique ID*	*Equation / Throughput*	*WBS Indent Level*	*PE Position in Distribution*	*Distribution Form*	*Grouping*	*Group Strength*	*Random Seed*	*CDF Keyword*	*Ri (+) Rev Row Index*	*i (+) Row Index*
* Growth/Maturity-Adjusted Unmodified DSLOC Distributions											
Table Header Row -- DO NOT DELETE (table rows must be at WBS Level 2)	GMSdUTbl		1								
Example Software Item with New, Modified, and Unmodified DSLOC		exp(-3.466*FYIVal(@STbl+i,2022))*(FYTot(@GSdUTbl+i)-(FYIVal(@STbl+i,2020)-FYIVal(@STbl+i,2021)))+(FYIVal(@STbl+i,2020)-FYIVal(@STbl+i,2021)) [DEGM8]	2							4	1
Example Software Item with New DSLOC Only		exp(-3.466*FYIVal(@STbl+i,2022))*(FYTot(@GSdUTbl+i)-(FYIVal(@STbl+i,2020)-FYIVal(@STbl+i,2021)))+(FYIVal(@STbl+i,2020)-FYIVal(@STbl+i,2021)) [DEGM8]	2							3	2
Example Software Item with Modified DSLOC Only		exp(-3.466*FYIVal(@STbl+i,2022))*(FYTot(@GSdUTbl+i)-(FYIVal(@STbl+i,2020)-FYIVal(@STbl+i,2021)))+(FYIVal(@STbl+i,2020)-FYIVal(@STbl+i,2021)) [DEGM8]	2							2	3
Example Software Item with Unmodified DSLOC Only		exp(-3.466*FYIVal(@STbl+i,2022))*(FYTot(@GSdUTbl+i)-(FYIVal(@STbl+i,2020)-FYIVal(@STbl+i,2021)))+(FYIVal(@STbl+i,2020)-FYIVal(@STbl+i,2021)) [DEGM8]	2							1	4

Table 10 Growth Adjusted New DSLOC Distributions Table

WBS/CES Description	Unique ID	Equation / Throughput	WBS Indent Level	PE Position in Distribution	Distribution Form	Grouping	Group Strength	Random Seed	CDF Keyword	Ri (+) Rev Row Index	i (+) Row Index
* Baseline Growth-Adjusted New DSLOC Distributions											
Table Header Row -- DO NOT DELETE (table rows must be at WBS Level 2)	GSdNTbl		1								
Example Software Item with New, Modified, and Unmodified DSLOC		FYIVal(@CDERTbl+FYIVal(@STbl+i,2001),2018)*(FYIVal(@STbl+i,2011)/if(FYIVal(@STbl+i,2034)=0,0.00000001,FYIVal(@STbl+i,2034)))^FYIVal(@CDERTbl+FYIVal(@STbl+i,2001),2017)*FYIVal(@STbl+i,2034) [DEGM8]	2	Undefined	CDF	KG_1_1_1	D	3806309	DGEM8_New_Default_CDF	4	1
Example Software Item with New DSLOC Only		FYIVal(@CDERTbl+FYIVal(@STbl+i,2001),2018)*(FYIVal(@STbl+i,2011)/if(FYIVal(@STbl+i,2034)=0,0.00000001,FYIVal(@STbl+i,2034)))^FYIVal(@CDERTbl+FYIVal(@STbl+i,2001),2017)*FYIVal(@STbl+i,2034) [DEGM8]	2	Undefined	CDF			1237242	DGEM8_New_Default_CDF	3	2
Example Software Item with Modified DSLOC Only		FYIVal(@CDERTbl+FYIVal(@STbl+i,2001),2018)*(FYIVal(@STbl+i,2011)/if(FYIVal(@STbl+i,2034)=0,0.00000001,FYIVal(@STbl+i,2034)))^FYIVal(@CDERTbl+FYIVal(@STbl+i,2001),2017)*FYIVal(@STbl+i,2034) [DEGM8]	2	Undefined	CDF			1754079	DGEM8_New_Default_CDF	2	3
Example Software Item with Unmodified DSLOC Only		FYIVal(@CDERTbl+FYIVal(@STbl+i,2001),2018)*(FYIVal(@STbl+i,2011)/if(FYIVal(@STbl+i,2034)=0,0.00000001,FYIVal(@STbl+i,2034)))^FYIVal(@CDERTbl+FYIVal(@STbl+i,2001),2017)*FYIVal(@STbl+i,2034) [DEGM8]	2	Undefined	CDF			2544617	DGEM8_New_Default_CDF	1	4

Table 11 Growth Adjusted Modified DSLOC Distributions Table

WBS/CES Description	Unique ID	Equation / Throughput	WBS Indent Level	PE Position in Distribution	Distribution Form	Grouping	Group Strength	Random Seed	CDF Keyword	Ri (+) Rev Row Index	i (+) Row Index
* Baseline Growth-Adjusted Modified DSLOC Distributions											
Table Header Row -- DO NOT DELETE (table rows must be at WBS Level 2)	GSdMTbl		1								
Example Software Item with New, Modified, and Unmodified DSLOC		FYIVal(@CDERTbl+FYIVal(@STbl+i,2001),2020)*((FYIVal(@STbl+i,2013)-FYIVal(@STbl+i,2014))/if(FYIVal(@STbl+i,2035)=0,0.00000001,FYIVal(@STbl+i,2035)))^FYIVal(@CDERTbl+FYIVal(@STbl+i,2001),2019)*FYIVal(@STbl+i,2035) [DEGM8]	2	Undefined	CDF	KG_1_1_1	3.09E-01	3263314	DGEM8_Mod_Default_CDF	4	1
Example Software Item with New DSLOC Only		FYIVal(@CDERTbl+FYIVal(@STbl+i,2001),2020)*((FYIVal(@STbl+i,2013)-FYIVal(@STbl+i,2014))/if(FYIVal(@STbl+i,2035)=0,0.00000001,FYIVal(@STbl+i,2035)))^FYIVal(@CDERTbl+FYIVal(@STbl+i,2001),2019)*FYIVal(@STbl+i,2035) [DEGM8]	2	Undefined	CDF			3857033	DGEM8_Mod_Default_CDF	3	2
Example Software Item with Modified DSLOC Only		FYIVal(@CDERTbl+FYIVal(@STbl+i,2001),2020)*((FYIVal(@STbl+i,2013)-FYIVal(@STbl+i,2014))/if(FYIVal(@STbl+i,2035)=0,0.00000001,FYIVal(@STbl+i,2035)))^FYIVal(@CDERTbl+FYIVal(@STbl+i,2001),2019)*FYIVal(@STbl+i,2035) [DEGM8]	2	Undefined	CDF			3901410	DGEM8_Mod_Default_CDF	2	3
Example Software Item with Unmodified DSLOC Only		FYIVal(@CDERTbl+FYIVal(@STbl+i,2001),2020)*((FYIVal(@STbl+i,2013)-FYIVal(@STbl+i,2014))/if(FYIVal(@STbl+i,2035)=0,0.00000001,FYIVal(@STbl+i,2035)))^FYIVal(@CDERTbl+FYIVal(@STbl+i,2001),2019)*FYIVal(@STbl+i,2035) [DEGM8]	2	Undefined	CDF			1618071	DGEM8_Mod_Default_CDF	1	4

Table 12 Growth Adjusted Unmodified DSLOC Distributions Table

WBS/CES Description	*Unique ID*	*Equation / Throughput*	*WBS Indent Level*	*PE Position in Distribution*	*Distribution Form*	*Grouping*	*Group Strength*	*Random Seed*	*CDF Keyword*	*Ri (+) Rev Row Index*	*i (+) Row Index*
* Baseline Growth-Adjusted Unmodified DSLOC Distributions											
Table Header Row -- DO NOT DELETE (table rows must be at WBS Level 2)	GSdUTbl		1								
Example Software Item with New, Modified, and Unmodified DSLOC		FYIVal(@CDERTbl+FYIVal(@STbl+i,2001),2022)*((FYIVal(@STbl+i,2020)-FYIVal(@STbl+i,2021))/if(FYIVal(@STbl+i,2036)=0,0.00000001,FYIVal(@STbl+i,2036)))^FYIVal(@CDERTbl+FYIVal(@STbl+i,2001),2021)*FYIVal(@STbl+i,2036) [DEGM8]	2	Undefined	CDF	KG_1_1_1	-1.06E-02	1501525	DGEM8_Umod_Default_CDF	4	1
Example Software Item with New DSLOC Only		FYIVal(@CDERTbl+FYIVal(@STbl+i,2001),2022)*((FYIVal(@STbl+i,2020)-FYIVal(@STbl+i,2021))/if(FYIVal(@STbl+i,2036)=0,0.00000001,FYIVal(@STbl+i,2036)))^FYIVal(@CDERTbl+FYIVal(@STbl+i,2001),2021)*FYIVal(@STbl+i,2036) [DEGM8]	2	Undefined	CDF			3527083	DGEM8_Umod_Default_CDF	3	2
Example Software Item with Modified DSLOC Only		FYIVal(@CDERTbl+FYIVal(@STbl+i,2001),2022)*((FYIVal(@STbl+i,2020)-FYIVal(@STbl+i,2021))/if(FYIVal(@STbl+i,2036)=0,0.00000001,FYIVal(@STbl+i,2036)))^FYIVal(@CDERTbl+FYIVal(@STbl+i,2001),2021)*FYIVal(@STbl+i,2036) [DEGM8]	2	Undefined	CDF			2864677	DGEM8_Umod_Default_CDF	2	3
Example Software Item with Unmodified DSLOC Only		FYIVal(@CDERTbl+FYIVal(@STbl+i,2001),2022)*((FYIVal(@STbl+i,2020)-FYIVal(@STbl+i,2021))/if(FYIVal(@STbl+i,2036)=0,0.00000001,FYIVal(@STbl+i,2036)))^FYIVal(@CDERTbl+FYIVal(@STbl+i,2001),2021)*FYIVal(@STbl+i,2036) [DEGM8]	2	Undefined	CDF			2778686	DGEM8_Umod_Default_CDF	1	4

Table 13 Software Item-Specific Inputs, Parameters, and Controls Table

WBS/CES Description	*Unique ID*	*Equation / Throughput*	*Phasing Method*	*FY 2001*	*FY 2011*	*FY 2012*	*FY 2013*	*FY 2014*	*FY 2015*	*FY 2020*	*FY 2021*	*FY 2022*	*FY 2027*	*FY 2028*	*FY 2033*	*FY 2034*	*FY 2035*	*FY 2036*
* Software Item-Specific Inputs, Parameters, and Controls (GR&As)				CDER Select	TBE New DSLOC	New DSLOC Est Maturity	TBE Modified DSLOC	TBE Modified Deleted DSLOC	Modified DSLOC Est Maturity	TBE Unmodified DSLOC	TBE Unmodified Deleted DSLOC	Unmodified DSLOC Est Maturity	Redelivery Source Reverse Index	TBE Redelivered Deleted DSLOC	Redelivery Multiplier	GN CSCI Norm Factor	GM CSCI Norm Factor	GU CSCI Norm Factor
Table Header Row -- DO NOT DELETE (value entries contained in FY columns)	STbl	0																
Example Software Item with New, Modified, and Unmodified DSLOC		[Input Throughput]	I	1	25000	0.2	50000	0	0.2	100000	0	0.2	0	0	1	1	1	1
Example Software Item with New DSLOC Only		[Input Throughput]	I	1	25000	0.2	0	0	0.2	0	0	0.2	0	0	1	1	1	1
Example Software Item with Modified DSLOC Only		[Input Throughput]	I	1	0	0.2	50000	0	0.2	0	0	0.2	0	0	1	1	1	1
Example Software Item with Unmodified DSLOC Only		[Input Throughput]	I	1	0	0.2	0	0	0.2	100000	0	0.2	0	0	1	1	1	1

Table 14 JCDER Library Table

WBS/CES Description	*Unique ID*	*Equation / Throughput*	*Phasing Method*	*FY 2017*	*FY 2018*	*FY 2019*	*FY 2020*	*FY 2021*	*FY 2022*	*FY 2023*	*FY 2024*	*FY 2025*
*JCDER Library				a[GN]	b[GN] geomean	a[GM]	b[GM] geomean	a[GU]	b[GU] geomean	e[GN]e[GM] correl	e[GN]e[GU] correl	e[GM]e[GU] correl
Table Header Row -- DO NOT DELETE (value entries contained in FY columns)	JCDERTbl	0										
JCDER349: SRDR 2015: Version 8 DSLOC Growth Model Baseline w/ GF3 Valid Filtering Only		[Input Throughput]	I	1.02E+00	1.21E+00	9.13E-01	2.65E+00	1.04E+00	6.20E-01	2.57E-03	3.02E-01	7.47E-02

Appendix F DEGM8SV Measures, Parameters, and Statistics

JCDER501: All Growth Eligible

DSLOC Estimate Growth Model Version:	Version 8
Version 8 DSLOC Estimate Growth Model Regression Method:	ODR

DSLOC Estimate Growth Model Equations and Variables

New DSLOC Growth Equation:	$S[DGANew] \triangleq \exp(-(Decay*Maturity))*(\tilde{b}[GN]*\varepsilon[GN]*S[DNew]/K[N])^{\wedge}a[GN]*K[N]$
Modified DSLOC Growth Equation:	$S[DGAMod] \triangleq \exp(-(Decay*Maturity))*(\tilde{b}[GM]*\varepsilon[GM]*S[DMod]/K[M])^{\wedge}a[GM]*K[M]$
Unmodified DSLOC Growth Equation:	$S[DGAUmod] \triangleq \exp(-(Decay*Maturity))*(\tilde{b}[GU]*\varepsilon[GU]*S[DUmod]/K[U])^{\wedge}a[GU]*K[U]$

where:

$a[GN]$ = 1.074	$a[GM]$ = 0.709	$a[GU]$ = 1.265
$Decay[GN]$ = 0.088	$Decay[GM]$ = 0.333	$Decay[GU]$ = 2.336

List Statistics	[GN]	[GM]	[GU]
Number of Data Points (observations):	35	26	33
Geometric (log space) mean of b:	5.187E-01	1.421E+01	5.016E-02
Arithmetic (unit space) mean of b:	5.656E-01	2.296E+01	6.070E-02
Standard deviation of b:	2.510E-01	2.959E+01	3.430E-02
Coefficient of Variation (CV) b:	0.44	1.29	0.57
Arithmetic (unit space) mean of ε:	1.101E+00	1.769E+00	2.152E+00
Standard deviation of ε:	5.436E-01	2.205E+00	3.060E+00
Coefficient of Variation (CV) of ε:	0.49	1.25	1.42
Mean Magnitude of the Relative Error:	110%	177%	215%

New to Modified DSLOC Correlation:	5.868E-01
New to Unmodified DSLOC Growth Correlation:	-8.293E-02

Growth Factor Estimating Relationships Behavior	*New DSLOC Growth*	*Modified DSLOC Growth*	*Unmodified DSLOC Growth*
Implied Growth Factor at data set mean baseline DSLOC:	23% at 119,750 DSLOC	-37% at 45,778 DSLOC	-3% at 70,527 DSLOC
Implied Growth Factor at data set geometric mean baseline DSLOC:	16% at 55,026 DSLOC	-10% at 13,140 DSLOC	-10% at 52,272 DSLOC

	Variables and Notation
Variables Used in JCDER and DSLOC Estimate Growth Equations	
S ≡	Size (ESLOC) distribution
E ≡	Core Effort (person-months) or (pm) distribution (data set SDLC)
T ≡	Core Duration (calendar months) or (cm) distribution (data set SDLC)
E[A-A] ≡	All-All Effort (person-months) or (pm) distribution (All-All SDLC)
T[A-A] ≡	All-All Duration (calendar months) or (cm) distribution (All-All SDLC)
alpha[E] ≡	Effort dimension exponent (economy/diseconomy of scale)
alpha[T] ≡	Duration dimension exponent (economy/diseconomy of scale)
alpha[S] ≡	Size dimension exponent (economy/diseconomy of scale)
D ≡	Difficulty (nonlinear person-month months per ESLOC) distribution -- time-sensitive inverse productivity
I ≡	Intensity (nonlinear person-months per calendar month) distribution -- staffing or burn
Eaf ≡	Effor Adjustment Factor (data set SDLC to All-All SDLC)
Daf ≡	Duration Adjustment Factor (data set SDLC to All-All SDLC)
K[S] ≡	Software Item (SI) being estimated to CSCI normalization factor for ESLOC
a[GN] ≡	New DSLOC estimating relationship exponent (economy/diseconomy of scale)
a[GM] ≡	Modified DSLOC estimating relationship exponent (economy/diseconomy of scale)
a[GU] ≡	Unmodified DSLOC estimating relationship exponent (economy/diseconomy of scale
ε[GN] ≡	Baseline New DSLOC estimate growth error factor distribution
ε[GM] ≡	Baseline Modified DSLOC estimate growth error factor distribution
ε[GU] ≡	Baseline Unmodified DSLOC estimate growth error factor distribution
Decay[GN],[GM],[GU] ≡	Decay parameters for New, Modified, and Unmodified DSLOC growth estimating relationships
$\tilde{b}[GN]$, $\tilde{b}[GM]$, $\tilde{b}[GU]$ ≡	Geometric mean scale factor parameters for New, Modified, and Unmodified DSLOC growth estimating relationships
Maturity ≡	Estimate Maturity Parameter (e.g., SDLCBegin=0%; SyRR=10%; SWRR=20%; SwPDR=40%; SWCDR=60%; TRR=80%; SWAccept=100%)
K[N], K[M], K[U] ≡	Software Item (SI) to CSCI normalization factor for New, Mod, and Umod DSLOC
S[DNew] ≡	Technical Baseline Estimate (TBE) of New DSLOC
S[DMod] ≡	Technical Baseline Estimate (TBE) of Modified DSLOC
S[DUmod] ≡	Technical Baseline Estimate (TBE) of Unmodified DSLOC
S[DGANew] ≡	Growth-adjusted New DSLOC estimate distribution
S[DGAMod] ≡	Growth-adjusted Modified DSLOC estimate distribution
S[DGAUmod] ≡	Growth-adjusted Unmodified DSLOC estimate distribution
Notation Conventions	
'≡' is used to indicate identity; the left expression is defined by right expression	
'≙' is used to indicate estimation; the right expression estimates the left expression	
Arial Bold Italic font indicates a random variable; range and distribution of possible outcomes	
Times New Roman Italic font indicates a specific-value (single-value) variable	
Bracks '[]' are used to indicate the subscripted characters of a variable name	

ODR -- SVDD(M[Est,Act])			ODR -- SVDV(M[Est,Act])		
6.517E+00	0.000E+00	0.000E+00	**9.973E-01**	1.109E-02	-7.323E-02
0.000E+00	**5.559E+00**	0.000E+00	**7.377E-02**	-6.154E-02	9.954E-01
0.000E+00	0.000E+00	**2.456E+00**	**-6.528E-03**	9.980E-01	6.218E-02
a[GN]	1.074		*Generalized R^2*	0.90	
b[GN] (geometric mean)	5.187E-01		*%SOEE Unit Space*	**49.84%**	
b[GN] data set arithmetic mean	5.656E-01		*min(SDNAct)*	5,998	
b[GN] data set standard deviation	2.510E-01		*max(SDNAct)*	480,086	
b[GN] data set CV	0.44		*Growth @ min(SDNAct)*	-1%	
e[GN] data set mean	1.101E+00		*Growth @ max(SDNAct)*	37%	
e[GN] data set standard deviation	5.436E-01		*Relationship MMROE*	110%	
e[GN] data set CV	0.49		*Decay*	0.088	
Implied growth factor at data set mean size	23% at 119,750 DSLOC				
Implied growth factor at data set geomean size	16% at 55,026 DSLOC				

ODR -- SVDD(M[Est,Act])			ODR -- SVDV(M[Est,Act])		
8.336E+00	0.000E+00	0.000E+00	**9.535E-01**	-2.241E-01	-2.017E-01
0.000E+00	**5.184E+00**	0.000E+00	**-2.861E-01**	-4.615E-01	-8.397E-01
0.000E+00	0.000E+00	**3.977E+00**	**-9.515E-02**	-8.584E-01	5.042E-01
a[GM]	0.709		*Generalized R^2*	0.91	
b[GM] (geometric mean)	1.421E+01		*%SOEE Unit Space*	**64.67%**	
b[GM] data set arithmetic mean	2.296E+01		*min(SDMAct)*	2,000	
b[GM] data set standard deviation	2.959E+01		*max(SDMAct)*	240,100	
b[GM] data set CV	1.29		*Growth @ min(SDMAct)*	56%	
e[GM] data set mean	1.769E+00		*Growth @ max(SDMAct)*	-61%	
e[GM] data set standard deviation	2.205E+00		*Relationship MMROE*	177%	
e[GM] data set CV	1.25		*Decay*	0.333	
Implied growth factor at data set mean size	-37% at 45,778 DSLOC				
Implied growth factor at data set geomean size	-10% at 13,140 DSLOC				

ODR -- SVDD(M[Est,Act])			ODR -- SVDV(M[Est,Act])		
6.305E+00	0.000E+00	0.000E+00	**6.660E-01**	-6.875E-01	2.895E-01
0.000E+00	**4.892E+00**	0.000E+00	**-2.935E-01**	1.152E-01	9.490E-01
0.000E+00	0.000E+00	**2.268E+00**	**-6.857E-01**	-7.170E-01	-1.251E-01
a[GU]	1.265		*Generalized R^2*	0.20	
b[GU] (geometric mean)	5.016E-02		*%SOEE Unit Space*	**49.11%**	
b[GU] data set arithmetic mean	6.070E-02		*min(SDUAct)*	8,000	
b[GU] data set standard deviation	3.430E-02		*max(SDUAct)*	213,010	
b[GU] data set CV	0.57		*Growth @ min(SDUAct)*	-46%	
e[GU] data set mean	2.152E+00		*Growth @ max(SDUAct)*	30%	
e[GU] data set standard deviation	3.060E+00		*Relationship MMROE*	215%	
e[GU] data set CV	1.42		*Decay*	2.336	
Implied growth factor at data set mean size	-3% at 70,527 DSLOC				
Implied growth factor at data set geomean size	-10% at 52,272 DSLOC				

Correlation Matrix			
	New DSLOC	**Modified DSLOC**	**Unmodified DSLOC**
New DSLOC	1		
Modified DSLOC	5.867595E-01	1	
Unmodified DSLOC	-8.293332E-02	-8.104507E-02	1

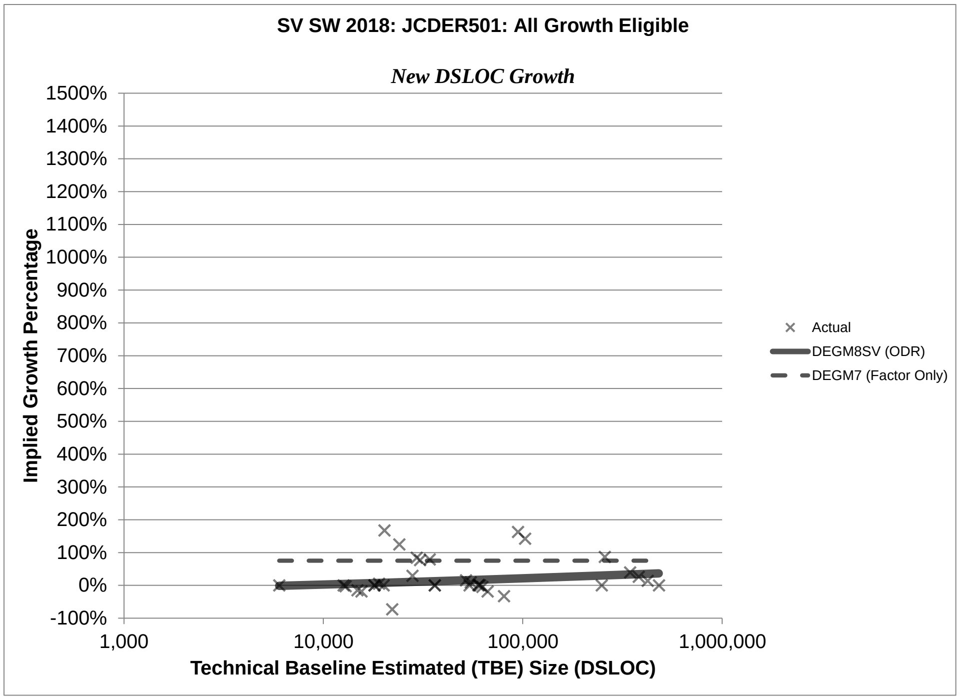


Figure 14 SV New DSLOC data regressed with ODR in log space (DEGM8SV) yields a better fit than does simply using the DEGM7 relationship.

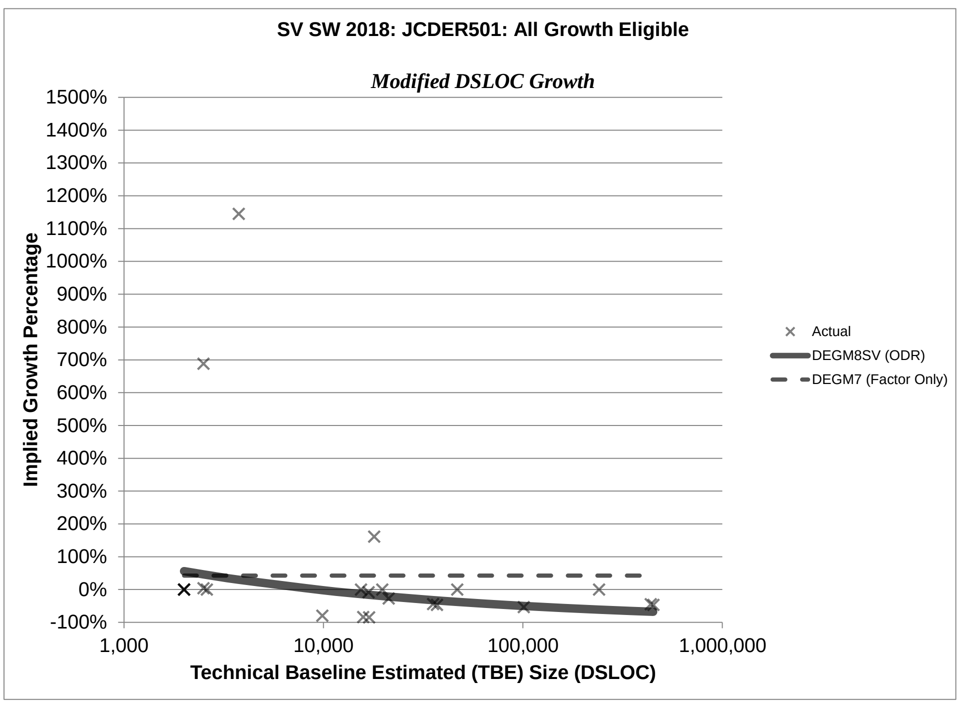


Figure 15 SV Modified DSLOC data regressed with ODR in log space (DEGM8SV) yields a better fit than does simply using the DEGM7 relationship.

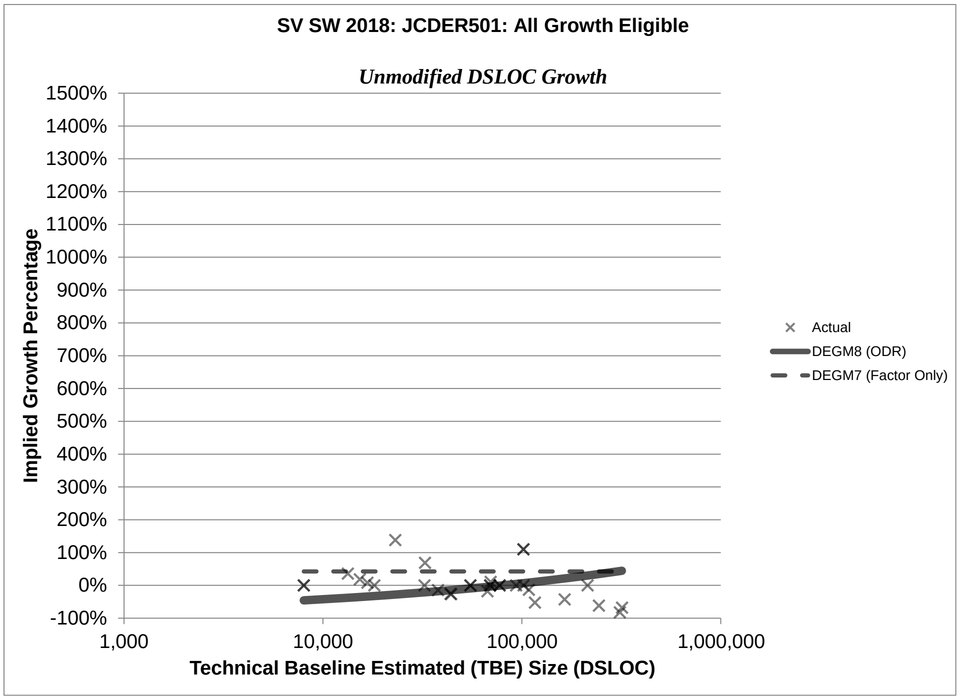


Figure 16 SV Unmodified DSLOC data regressed with ODR in log space (DEGM8SV) yields a better fit than does simply using the DEGM7 relationship.

NOTES

1 The term *custom CDF* refers to a feature in Tecolote Research, Inc.'s ACE software tool that allows distributions to be specified as a discrete range-value-to-percentile mapping as opposed to a mapping described by some mathematical distribution function such as "lognormal".

2 We use the ***Arial bold italic*** font to denote a random variable; i.e., a variable that can take on values according to some probability distribution, the ***Times New Roman bold italic*** font to denote a function, the **Times New Roman bold font** to denote a vector or matrix or list or array, the *Times New Roman italic font* to denote a simple variable, and the Times New Roman normal font to denote a number. We use the overstrike caret (^) character to indicate an object that represents an estimated value. We use the overstrike tilde (~) character to indicate an object that represents the geometric mean value (arithmetic mean value in log space) of an associated list or random variable (acts as a descriptor), or the geometric mean value of a list or random variable (acts as a function); depending on context.

3 We use the term *attainment probability* to describe the probability that the actual outcome will be less than or equal to a particular value in the distribution of outcomes; i.e., a particular outcome's percentile rank.

4 These percentages are default values representing rough approximations of the percentages used by various software estimating models.

5 Note that it is possible to expand the estimate *Maturity* input to include specific (and possibly different) *Maturity* values for New, Modified, and Unmodified DSLOC. This can be done as part of the model implementation since each of New, Modified, and Unmodified DSLOC has its own unique growth equation.

6 SDLCBegin assumes zero maturity.

7 This data set, circa 2015, contains 2,861 observations (submitted DD Form 2630's).

8 Linear algebra defines the *standard basis* in $\mathbb{R}^m$ as an ordered set of m distinct standard *unit vectors* $\{\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \ldots, \hat{\mathbf{e}}_n\}$ where each $\mathbf{e}_i$ denotes a vector with the value of 1 as its i-th coordinate and the value of 0 as every other coordinate. Locating the initial points of each standard unit vector $\hat{\mathbf{e}}_i$ at the $\mathbb{R}^m$ reference origin P_0 serves to locate all the positive coordinate axes in $\mathbb{R}^m$.

9 Physical quantities that are *commensurable* have the same dimension and can be directly compared to each other, even if they are originally expressed in differing units of measure (such as inches and meters, or pounds and newtons). If physical quantities have different dimensions (such as length vs. mass), they cannot be expressed in terms of similar units and cannot be compared in quantity (also called incommensurable). For example, asking whether a kilogram is greater than, equal to, or less than an hour is meaningless.
[https://en.wikipedia.org/wiki/Dimensional_analysis]

In physics, mathematics, statistics, and economics, *scale invariance* is a feature of objects or laws that do not change if scales of length, energy, or other variables, are multiplied by a common factor, thus represent a universality. [https://en.wikipedia.org/wiki/Scale_invariance]

10 [https://en.wikipedia.org/wiki/Total_least_squares]

11 [https://en.wikipedia.org/wiki/Distance_from_a_point_to_a_line]

12 In regression analysis, the difference between the observed value of the dependent variable and its predicted value is called the residual. [stattrek.com/statistics/dictionary.aspx?definition=residual]

13 The convolution of probability distributions arises in probability theory and statistics as the operation in terms of probability distributions that corresponds to the addition of independent random variables and, by extension, to forming linear combinations of random variables. [https://en.wikipedia.org/wiki/Convolution_of_probability_distributions]

14 [http://www.mechanicalkern.com/static/odr_ams.pdf]

15 Not to be confused with use of the overstrike tilde (~) notation to indicate that a value is a geometric mean as is done in the body of the paper.

16 We use the overstrike caret or *hat* (^) notation on a vector to indicate that it is a unit vector.

17 [https://en.wikipedia.org/wiki/Principal_component_analysis]

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.

Jensen, Randall. 2008. Estimating Software Growth. Reston, VA : Space Systems Cost Analysis Group (SSCAG), October 15-16, 2008.

Ross, Michael A. 2011. A Probabilistic Method for Predicting Software Code Growth. [ed.] Stephen A. Book and Edward White III. *Journal of Cost Analysis and Parametrics.* Jul-Dec 2011. Vol. 4 No. 2, pp. 127-147. ISSN 1941-658X.

—. 2003. Continuous Software Size Estimating and Tracking: Size Does Matter. *Proceedings, Joint ISPA / SCEA 2003 Conference.* Orlando, Florida, USA : The International Society of Parametric Analysts and The Society of Cost Estimating and Analysis, June 2003.

—. 2008. Next Generation Software Estimating Framework: 25 Years and Thousands of Projects Later. [ed.] Stephen A. Book and Edward White III. *Journal of Cost Analysis and Parametrics.* s.l. : Society of Cost Estimating and Analysis - International Society of Parametric Analysts, Fall 2008. Vol. 1, 2, pp. 7-30. ISSN 1941-658X.

—. 2005. Software Size Uncertainty: The Effects of Growth and Estimation Variability. *Proceedings, Joint ISPA / SCEA 2005 Conference.* Denver, Colorado, USA : The International Society of Parametric Analysts and The Society of Cost Estimating and Analysis, 2005.

About the Authors

Eric Sommer has a Bachelor of Science degree in Mathematics and a Finance focused MBA. He began his career as a high school math teacher, teaching algebra to AP Calculus, before joining the cost estimating community at the Space and Missile Systems Center (SMC). He currently is an Acting Cost Chief and an Operations Research Analyst at SMC focusing on software cost estimating and office automation.

Bopha Seng is a senior analyst with Tecolote Research supporting the Cost Estimating and Research Divisions at the Space and Missile Systems Center (SMC). Her primary responsibilities include space and ground cost estimating/modelling and methodology development. Prior to Tecolote, she was a civilian at SMC where she was the MILSATCOM advanced concepts cost lead and POM lead. Bopha is a CCE/A and has a BA in Statistics from UC Berkeley.

David LaPorte is an analyst with Tecolote Research supporting the Cost Research Division at the Space and Missile Systems Center. His chief research efforts include data collections, space vehicle mass growth analysis, and software growth analysis. He has been working with Tecolote for 6 months after graduating top of his class from California State University Dominguez Hills with a degree in Finance.

Michael Ross has 40 years of experience in software engineering as a developer, manager, consultant, instructor, and award-winning international speaker. Mr. Ross is currently President and CEO of r2Estimating, LLC (developers of the r2-v2 Software Estimating Framework). Mr. Ross is a Life Member of ICEAA and regularly presents papers at its annual conferences (four of which have been recognized with Best Paper Awards). Mr. Ross has a BS in Computer Engineering from Arizona State University.