Journal of Cost Analysis and Parametrics, 6:43–74, 2013
ISSN: 1941-658X print / 2160-4746 online
DOI: 10.1080/1941658X.2013.767073

ICEAA
www.iceaaonline.com

A Macro-Stochastic Model for Improving the Accuracy of Department of Defense Life Cycle Cost Estimates

ERIN T. RYAN, CHRISTINE M. SCHUBERT KABBAN, DAVID R. JACQUES, and JONATHAN D. RITSCHEL

Air Force Institute of Technology, WPAFB, Ohio

The authors present a prognostic cost model that is shown to provide significantly more accurate estimates of life cycle costs for Department of Defense programs. Unlike current cost estimation approaches, this model does not rely on the assumption of a fixed program baseline. Instead, the model presented here adopts a stochastic approach to program uncertainty, seeking to identify and incorporate top-level (i.e., "macro") drivers of estimating error to produce a cost estimate that is likely to be more accurate in the real world of shifting program baselines. The predicted improvement in estimating accuracy provided by this macro-stochastic cost model translates to hundreds of billions of dollars across the Department of Defense portfolio. Furthermore, improved cost estimate accuracy could reduce actual life cycle costs and/or allow defense acquisition officials the ability to make better decisions on the basis of more accurate assessments of value and affordability.

Introduction and Motivation

Many senior defense acquisition officials routinely make key decisions involving weapon systems that are projected to cost billions of—or perhaps even a trillion (Hebert, 2011)—dollars over their life cycle. These high-dollar decisions may involve how many units to procure, how to phase program funding, or even whether to fund a program at all. Typically, the decision will not only have major implications on the life of a given program, but it can also impact the Pentagon's overall budget and strategic direction. In light of the looming, significant reductions to the defense budget (GPO, 2011), these program decisions are bound to become both more difficult and more important, as questions of *value* and *affordability* increasingly take center stage.

For the senior decision-maker, a principal tool for assessing the value and/or affordability of a given defense program is via long-term program cost estimates, such as Life Cycle Cost (LCC) and per unit Operating and Support (O&S) cost. It is therefore essential that these estimates be reliable and accurate. But what if they are not? What if the forecasted ownership costs of a given program are far different from the actual costs? If there is a significant disconnect between estimated and actual costs, the concern naturally arises as to the utility of the estimates, and how sound are any decisions based upon them. These are not just hypothetical questions. The authors recently completed a study that shows

The authors welcome opportunities to share the data used to formulate this model. Please contact the primary author for any questions on this paper or to obtain source data.

Address correspondence to Erin T. Ryan, Air Force Institute of Technology, 2950 Hobson Way, WPAFB, Ohio 45433, USA. E-mail: erin.ryan@afit.edu

Department of Defense (DoD) estimates of long-term program cost are often highly inaccurate and—perhaps more surprisingly—improve very little, if at all, as programs mature (Ryan, Jacques, Ritschel, & Schubert, 2012).

This finding logically leads one to consider a more formidable challenge: How can the accuracy of DoD life cycle cost estimates be improved? In this article, the authors tackle the problem through a fundamentally different approach to cost estimating. We propose a technique that, in essence, models the error in the program estimate as a random variable whose value is determined by a salient group of top-level program summary indicators. This prediction of the estimate error is then used to adjust the official program estimate to a value that is, on average, significantly closer to the eventual, actual cost of the program. We refer to this technique as *macro-stochastic* cost estimation. The authors have borrowed the term "macro-stochastic" from the physical sciences where it is used to describe large-scale phenomenon that can only be analyzed effectively in a statistical manner, such as dynamic structural loads or earthquakes (Wijker, 2009).

This article is structured as follows. After providing some background on the nature of contemporary DoD cost estimating, we review the key elements of the characterization study that informed the creation of the two separate macro-stochastic cost models presented in this article. Next, we detail the mixed model methodology used to build each model as well as the list of independent variables to be evaluated for significance. In the *Results* section, we begin by showing a *theoretical* macro-stochastic cost model to illustrate the potential power of this technique. We then describe how we transform the theoretical model into a *prognostic* model and how its performance was validated. We conclude with a discussion of key findings, known model issues, and ideas for future improvements.

The authors found that the adjustments the macro-stochastic model makes to the program cost estimates achieve levels of accuracy significantly better than the original estimates. With these improved estimates of actual program cost, we contend that senior DoD decision-makers can expect to have far better insight into actual program value and affordability, and come closer to achieving an optimal allocation of diminishing DoD resources.

Background

Cost Estimating

Over the past couple of decades, DoD cost estimating has become increasingly sophisticated. This comes as a result of improved computing capabilities, revised policies (DoD, 1992; OSD CAIG, 2007), and the canonization of the best practices. Standard, contemporary cost estimation techniques include product-oriented WBS development, point estimating with associated confidence intervals, integration of probability distributions, stochastic parameterization through cumulative distribution functions (i.e., s-curves), Monte Carlo simulations, and uncertainty/sensitivity analyses (GAO, 2009; DAU, 2012). The resulting bottoms-up cost estimates are remarkably credible, highly detailed, fully traceable, and mathematically rigorous. But they are also resource intensive, and—as documented in the characterization study—often highly inaccurate (Ryan et al., 2012).

It is important to note that an *inaccurate* estimate is not necessarily the same thing as a *poor* estimate. It may be that the cost estimator's greatest ambition of a perfectly accurate cost estimate is simply unattainable in the highly uncertain realm of defense acquisition. As recently (and rather facetiously) articulated by the NASA advisory council in the context of space systems:

> [Cost estimating] involves using incomplete, inaccurate, and changing data for an outmoded & ineffective space system to derive the precise cost of purchasing an unknown quantity of an undefined new space system to satisfy an overly exaggerated & unvalidated requirement at some time in the future, under uncertain conditions, with a minimum of funds. (NASA, 2008)

Although defense cost estimators attempt to take into account many sources of uncertainty (e.g., inflation and discount rates, technical risks, commodity pricing, etc.), they are ultimately constrained in a fundamental and critical way: *They must assume a program baseline*. The DoD Acquisition Program Baseline (APB) reflects the key cost, schedule, and performance attributes of a program, and is the necessary anchor from which all statistical cost excursions are based. In fact, changes to the APB are one of the most frequent reasons for creating a new program cost estimate.

And while using the program baseline as a cost estimating baseline is a perfectly logical approach, it means that current cost estimating techniques are not only unable to account for unforeseen sources of uncertainty (i.e., the infamous "unknown-unknowns"), but they also preclude the possibility of capturing cost impacts that result from APB changes. If the aim is to construct a cost estimate that will be as accurate as possible in the long run, this link to the baseline represents a fundamental flaw in the estimate process because APB deviations are virtually inevitable (at least for major defense programs (Drezner & Krop, 1997)). A delay in the planned initial operating capability of the system; a reduction in the procurement quantity; an additional performance requirement: These are all common causes of an APB deviation, and each is likely to adversely affect the accuracy of the *original* estimate (no matter how good it may have been). Given the magnitude and frequency of baseline changes in major Pentagon programs, it really shouldn't be surprising that the original LCC and O&S estimates are so often inaccurate. The greater wonder, in fact, is that these estimates are ever relatively close.

The motivating principle of this article is that in order for a cost estimate to have a reasonable chance of being accurate in the real world of changing baselines, one must employ an estimating technique that does not assume the program baseline is fixed. This goal is at the core of the macro-stochastic cost estimating approach. Any change to the APB—or any other cost-impacting change, for that matter—is assumed to be part of a larger random process. In this way, we regard the top-level cost estimate as a stochastic value (vice the constituent cost elements, as is typically done in the traditional stochastic cost estimating approach). The premise of the macro-stochastic cost-estimating model, then, is that each official program estimate has some random error (vis-à-vis actual costs) that is related to a probability distribution. We further hypothesize that the nature of this distribution is unique to each program, and can be sufficiently characterized by a relatively small number of top-level program indicators easily gleaned from readily available program records.

Cost Measures

There are two distinct measures of cost that we assessed in the associated characterization study: The *Life Cycle Cost* (LCC) and the *Annual Unitized O&S Cost* (AUC[1]). Importantly, each measure serves as the basis of a separate macro-stochastic cost model presented in this article. The LCC measure is arguably the most comprehensive and intuitive cost indicator for system value assessments, and the first version of the model attempts to predict *the error in the program's official LCC estimate*. The AUC metric is also useful, however. The AUC data tends to be more broadly available (both in terms of estimates and actuals),

thereby enabling analysis of a greater number of programs over a longer span of time. Moreover, unitized O&S costs are a commonly employed metric for assessing sustainment costs, and can often provide a more valid comparative measure across similar contemporary or antecedent systems (DAU, 2012). The AUC constitutes the foundation of our second version of the model, which predicts *the error in the program's official AUC estimate*.

In the case of both dependent variables, values are reported as percentages, with negative values indicating that the estimate was lower than the actual cost, and positive values indicating the estimate was too high. Thus, a perfectly accurate estimate will have an error of 0%.

Data Structure

Only Major Defense Acquisition Programs (MDAPs) were evaluated in the previously completed characterization study. This is because only MDAPs provide the necessary level of cost insight for readily assessing the accuracy of the LCC and AUC estimates. By law, MDAPs are required to provide an annual report, known as the Selected Acquisition Report (SAR), which includes a full life cycle cost analysis. The SARs were the primary source used in the characterization study for official program cost estimates. They are nominally first provided upon program initiation (typically Milestone II/B), and continue every year until the program has been 90% acquired (DAU, 2012).

Each program SAR represents one observation in the MDAP data set. Further, each SAR for which a valid LCC estimate and valid LCC actuals are available also becomes an observation used in the development of the macro-stochastic LCC model. Similarly, each SAR for which a valid AUC estimate and valid AUC actuals can be obtained becomes an observation used to build the macro-stochastic AUC model. The specific count of LCC and AUC SARs by program is provided in Table 1. The LCC model is based on observations from 317 SARs across 31 MDAPs, and the AUC model is based on observations from 392 SARs spanning 35 MDAPs.

The data used to construct and validate the model was obtained from all Air Force and Navy aviation, maritime, and munitions MDAPs. Therefore, the model is deemed to be valid for use only against these services and types of programs. Additional data would be required in order to assess model utility against space and information MDAPs. More details on data set sources and compilation (as well as why Army programs could not be evaluated) is available in the characterization paper.

Methodology

Mixed Models

The preceding characterization study, as well as the resulting model presented in this article, are based on longitudinal data (Ryan et al., 2012), which is to say that the source data consists of repeated measurements on different subjects over time. Importantly, the nature of longitudinal data precludes the possibility of assuming an identical and independent distribution (i.i.d.) of the random variables. Because the data is clustered into programs, with repeated measurements of each program over time, there necessarily exists a correlation between the repeated measurements for a given program—and therefore the statistical errors of the observations—that must be accounted for in the statistical analysis. Further, one expects these correlations to be greater for data points close in time, such as for successive SARs from the same program. This means that the statistical errors will be correlated as well.

TABLE 1 MDAP data used for model development

#	Program name	SubProgram name	Lead component	System type	SAR years	# of SARs	LCC SARs	AUC SARs
1	AIM-9X	AIM-9X (Navy)	Navy	Munition	1996–2001	6	6	6
2	AMRAAM (AF)	AMRAAM (AF)	Air Force	Munition	1988–1992	5	3	3
3	AMRAAM (Joint)	AMRAAM (Joint)	Air Force	Munition	1992–2010	18	18	18
4	AOE 6	AOE 6	Navy	Maritime	1988–1997	11	7	7
5	AV-8B	AV-8B REMAN.	Navy	Aviation	1994–2002	10	NA	10
6	C-130J	C-130J	Air Force	Aviation	1996–2010	13	12	12
7A	C-17A	C-17A (BY1981)	Air Force	Aviation	1987–1994	10	8	8
7B	C-17A	C-17A (BY1996)	Air Force	Aviation	1995–2010	14	14	14
8	C/MH-53E	C/MH-53E	Navy	Aviation	1987–1994	9	1	5
9	CVN 68 (74/75)	CVN 68 (74/75)	Navy	Maritime	1987–1998	13	2	2
10	CVN 68 (76)	CVN 68 (76)	Navy	Maritime	1994–2002	9	4	4
11	DDG 51	DDG 51	Navy	Maritime	1987–2010	25	20	20
12	E-2C	E-2C	Navy	Aviation	1994–2006	14	4	4
13	F-14D	F-14D	Navy	Aviation	1987–1993	9	5	9
14	F-16C/D	F-16C/D	Air Force	Aviation	1987–1994	8	4	4
15A	F-22	F-22 (BY1990)	Air Force	Aviation	1991–2004	16	4	16
15B	F-22	F-22 (BY2005)	Air Force	Aviation	2005–2010	6	6	6
16	F/A-18C	F/A-18C	Navy	Aviation	1987–1994	10	NA	7
17A	F/A-18E/F	F/A-18E/F (BY1990)	Navy	Aviation	1992–1999	9	9	9
17B	F/A-18E/F	F/A-18E/F (BY2000)	Navy	Aviation	2000–2010	10	10	10
18	GLOBAL HAWK	GLOBAL HAWK	Air Force	Aviation	2001–2010	11	NA	11
19	JASSM	JASSM	Air Force	Munition	1999–2009	12	11	12
20A	JPATS	JPATS (BY1995)	Air Force	Aviation	1995–1999	5	NA	5
20B	JPATS	JPATS (BY2002)	Air Force	Aviation	2001–2010	9	8	9

(*Continued*)

TABLE 1 (Continued)

#	Program name	SubProgram name	Lead component	System type	SAR years	# of SARs	LCC SARs	AUC SARs
21	JSOW	JSOW	Navy	Munition	1997–2010	14	14	14
22A	JSTARS	JSTARS (BY1983)	Air Force	Aviation	1989–1996	10	8	10
22B	JSTARS	JSTARS (BY1998)	Air Force	Aviation	1997–2003	6	3	6
23	KC-135R	KC-135R	Air Force	Aviation	1987–1994	8	NA	5
24	LHD 1	LHD 1	Navy	Maritime	1987–2005	18	15	15
25	LPD 17	LPD 17	Navy	Maritime	1996–2010	16	16	16
26A	MH-60R	MH-60R (BY1993)	Navy	Aviation	1994–2005	14	12	12
26B	MH-60R	MH-60R (BY2006)	Navy	Aviation	2006–2010	5	2	2
27	MH-60S	MH-60S	Navy	Aviation	1998–2010	17	17	17
28	MHC 51	MHC 51	Navy	Maritime	1991–1998	8	8	8
29	PREDATOR	PREDATOR	Air Force	Aviation	2009–2010	2	2	NA
30	SSGN	SSGN	Navy	Maritime	2002–2007	6	6	6
31	SSN 21	SSN 21	Navy	Maritime	1987–1999	15	11	11
32	SSN 774	SSN 774	Navy	Maritime	1995–2010	16	16	16
33	STRAT. SEALIFT	STRAT. SEALIFT	Navy	Maritime	1993–2001	11	NA	11
34A	T-45TS	T-45TS (BY1984)	Navy	Aviation	1987–1993	10	5	5
34B	T-45TS	T-45TS (BY1995)	Navy	Aviation	1994–2007	14	12	13
35	T-AKE	T-AKE	Navy	Maritime	2001–2010	10	10	10
36	T-AO 187	T-AO 187	Navy	Maritime	1987–1994	8	4	4
TOTAL						**470**	**317**	**392**

Importantly, the fact that we expect correlated errors for the programs in this study invalidates the underlying assumptions of simple analysis of variance and regression models, namely i.i.d. observations. To compensate for this, we instead employ *mixed model* techniques for the data in this study. Mixed models use both fixed (i.e., entire population) effects and random (i.e., subject-specific) effects within the same analysis. The key distinction between mixed models and simple regression models is that the former can produce valid models even if the subject observations are not independent. In essence, mixed models allow the data to exhibit inherent correlations and non-constant variability that arise from the program-specific effects. This allows one to effectively model not only the measures of central tendency for the data, but also the covariance structure attributable to the repeated measurements (Diggle, Liang, & Zeger, 1994; Verbeke & Molenberghs, 2000).

Relative to the standard General Linear Model (GLM), the use of a mixed model for this analysis provides several advantages, primarily relating to flexibility. A mixed model allows the use of input variables even if data is missing for one or more of the subjects (i.e., programs). Mixed models can also automatically accommodate for unequal spacing of the repeated measurements (i.e., ensure minimum variance), which is a characteristic of this data set. In addition, the mixed model allows more efficient and direct modeling of the within-subject covariance structure for the entire dataset, as opposed to unique covariances for every data pair. Finally, the results from the mixed model can be readily extended to outcomes that do not conform to a normal distribution. In this study, we have assumed the cost estimate errors are normally distributed (i.e., the solution to the mixed model equations is a maximum likelihood estimate where the distribution of the errors is normal), but the mixed model can accommodate nonlinear approaches, should they be considered more appropriate (Patetta, 2002).

To put this in mathematical terms, the GLM in matrix form is given as:

$$y = X\beta + \varepsilon, \tag{1}$$

where

y = the observed data vector, where $E(y) = X\beta$ and $var(y) = \sigma^2 I$;
x = the fixed effect design (i.e., model) matrix;
β = the vector of fixed effect parameter estimates (same for all subjects);
ε = the vector of residual errors, where $E(\varepsilon) = 0$ and $var(\varepsilon) = \sigma^2 I$.

For the mixed model version, a random-effects term is added:

$$y = X\beta + Zy + \varepsilon, \tag{2}$$

where

Z = the random effect design (i.e., model) matrix;
γ = the vector of random effect parameter estimates (varies by subject).

In addition,

$$E\begin{bmatrix}\gamma \\ \varepsilon\end{bmatrix} = 0 \text{ and } var\begin{bmatrix}\gamma \\ \varepsilon\end{bmatrix} = \begin{bmatrix}G & 0 \\ 0 & R\end{bmatrix} \Rightarrow var(y) = ZGZ + R, \tag{3}$$

where

G = the random effects covariance matrix;
R = the fixed effects covariance matrix.

One of the key inputs for a mixed model analysis is what structure should be used for the random covariance matrix, *G*. For this data set, since we tend to observe high correlations in the response variables reported in successive SARs, but increasingly less correlation as the time between SARs grows larger, a covariance structure that captures diminishing levels of correlation is desired. Therefore, a sensible choice for model development is the autoregressive (AR) structure, which has homogeneous variances and correlations that will decline exponentially with temporal distance (Wolfinger, 1993). Multiple other covariance matrix structures were also examined, but overall model performance was best using first-order autoregressive, i.e., AR(1).

To obtain the estimates of *G* and *R*, we solve for the values that optimize an objective function, in this case the Restricted Maximum Likelihood (REML) criterion. The method for computing the denominator degrees of freedom for the tests of fixed effects was Kenward-Roger. Thousands of model iterations were executed to find the best set of variables from Table 1 to use in each model: The Bayesian Information Criterion (BIC) was used as the primary method of discrimination between potential models. All model analysis was accomplished using SAS version 9.3 (http://www.sas.com/software/sas9/).

Independent Variables and the Unit of Analysis

As noted earlier, a central assumption of the macro-stochastic cost estimating approach is that there exists a relatively small set of high-level program parameters that, in aggregate, significantly relate to the LCC and AUC estimate errors observed for a given program. Table 2 lists and defines all of the independent variables we evaluated as potential fixed or random effect parameters for both the LCC and AUC cost models. All variables in this table are based on information available in the program SARs. Some of the variables are taken directly from the SAR, some are calculated based on information available in different sections of a single SAR, and some are calculated from information available across successive SARs. All cost figures are in native (i.e., SAR-specific) base year (BY) dollars, with the exception of variable #14. Although there are only 50 variables listed in Table 2, the inclusion of "trending versions" of several variables (see Table 2, footnote #a) brings the total count of independent variables to 252.

Table 2 is also interesting for what is not included. Defense acquisition professionals and cost estimators alike are keenly interested in the cost impacts of a number of strategic policies related to procurement. Three of the most intriguing—and controversial—relate to acquisition strategy (e.g., traditional vs. evolutionary), contracting strategy (fixed-price vs. cost-reimbursement), and sustainment strategies (organic vs. contractor). Although each of these policy topics could potentially serve as an excellent macro-level predictor of cost estimating accuracy, we were unable to incorporate variables related to any of these topics.

The fundamental obstacle in all three cases was being able to effectively quantify these variables in the context of fluctuating and disparate acquisition efforts. Consider, for instance, an evolutionary acquisition strategy, which may not be implemented until late in the program when technical maturity is sufficient or may only be applied to a particular element of the system in development. It may also be that an evolutionary strategy is abandoned midway through development or blended with more traditional practices into a hybrid approach. This is just one example, but these types of subtleties tend to dominate these three important procurement policy topics, thus regrettably precluding definitive categorization and analysis.

The last item involving methodology that the reader should be aware of pertains to the unit of analysis, which is equivalent to the *model subject*. This is a subtle, but critical, analytical element that changes throughout model development, characterization, and

TABLE 2 Listing of independent variables evaluated in error-correction models

#	Variable name	Msmnt. level	Description (values)
1	Program Year	Interval	Number of years since Milestone B (II) or program initiation
2	DoD Component	Nominal	Lead acquisition service component ("AF" or "Navy")
3	Joint	Nominal	Are units being procured for more than one service? ("Yes" or "No")
4	System Type	Nominal	Type of system ("Aviation," "Maritime" or "Munition")
5	Acq Phase	Nominal	Acquisition phase ("Development" or "Production")
6	Acq Type	Nominal	Type of acquisition ("New," "Modification," or "Variant")
7	Maturity	Ordinal	Program maturity level; categories based on *Expended* (#18).
8	Total Dev APBs	Interval	Cumulative number of development APBs to date
9	Avg Dev APBs	Interval	Average number of development APBs per year
10	Total Prod APBs	Interval	Cumulative number of production APBs to date
11	Avg Prod APBs	Interval	Average number of production APBs per year
12	Prime Contractor	Nominal	Contractor for 3 largest active contracts ("Boeing," "GD," "Lockheed-Martin," "Northrup-Grumman," "Raytheon," "Other," or "Multiple")
13	Acq Cost Est	Interval	Current estimate of total acquisition cost
14	Acq Cost Est, BY10[a,b]	Interval	Current estimate of total acq cost standardized to BY10 dollars
15	AUC Est[a,b]	Interval	Current estimate of annual unit O&S cost
16	O&S Cost Est[a,b]	Interval	Current estimate of total O&S cost
17	LCC Est[a,b]	Interval	Current estimate of total LCC cost
18	Expended	Interval	Percentage of *Acq Cost Est* (#13) expended to date

(*Continued*)

TABLE 2 (Continued)

#	Variable name	Msmnt. level	Description (values)
19	Funding Years	Interval	Current total planned funding years of program
20	PAUC Change, Dev	Interval	Percentage change in Program Acquisition Unit Cost (PAUC) from Development baseline
21	PAUC Change, Prod	Interval	Percentage change in Program Acquisition Unit Cost (PAUC) from Production baseline
22	APUC Change, Dev	Interval	Percentage change in Average Procurement Unit Cost (APUC) from Development baseline
23	APUC Change, Prod	Interval	Percentage change in Average Procurement Unit Cost (PAUC) from Production baseline
24	CV, Engr[a]	Interval	Total cost variance (CV) to date in engineering category as % of *Acq Cost Est* (#13)
25	CV, Est[a]	Interval	Total CV to date in estimating category as % of *Acq Cost Est* (#13)
26	CV, Quan[a]	Interval	Total CV to date in quantity category as % of *Acq Cost Est* (#13)
27	CV, Total[a]	Interval	Total CV to date in all CV categories as % of *Acq Cost Est* (#13)
28	CV, Total-Quan[a]	Interval	Total CV to date in all CV categories (except Quantity) as % of *Acq Cost Est* (#13)
29	Breaches, Sched	Interval	Cumulative number of schedule breaches to date
30	Breaches, Perf	Interval	Cumulative number of performance breaches to date
31	Breaches, Cost	Interval	Cumulative number of cost breaches to date
32	Breaches, UC	Interval	Cumulative number of unit cost breaches to date
33	Breaches, Total	Interval	Cumulative total of all breaches to date
34	Breach, N-M	Nominal	Has program incurred a Nunn-McCurdy breach? (“Yes” or “No”)
35	CDR-MSII	Interval	Time between Critical Design Review (CDR) and Milestone II
36	CDR-PDR	Interval	Time between CDR and Preliminary Design Review (PDR)
37	LRIP-MSII	Interval	Time between Low Rate Initial Production (LRIP) and Milestone II

38	MSIII-MSII	Interval	Time between Milestone III and Milestone II
39	IOC-MSIII	Interval	Time between Initial Operating Capability (IOC) and Milestone III
40	IOC-MSII	Interval	Time between Initial Operating Capability (IOC) and Milestone II
41	Reqmnts, New	Interval	Cumulative # of new requirements added to performance baseline
42	Reqmnts, Deleted	Interval	Cumulative # of existing requirements removed from performance baseline
43	Reqmnts, Total	Interval	Total number of requirements in current performance baseline
44	Reqmnts, Obj	Interval	Percentage of total requirements to date in which objective value was made more stringent
45	Reqmnts, Thresh	Interval	Percentage of total requirements to date in which threshold value was made more stringent
46	Reqmnts, Change	Interval	Percentage of total requirements to date in which threshold or objective value was modified
47	Procure, Plan[a,b]	Interval	Current total planned procurement quantity
48	Procure, Change[a,b]	Interval	Percentage change in *Procure, Plan* (#47) relative to baseline
49	Procured	Interval	Percentage of *Procure, Plan* (#47) currently procured
50	Unit Acq Ratio	Interval	Ratio of *AUC Est* (#15) to *Acq Cost Est* (#13)

[a]Includes trend versions of variable to date, i.e., minimum, maximum, range, mean, weighted mean (by Program Year), standard deviation, and the slope of the regression line.

[b]One or more transformations applied (i.e., unitary normalization, scalar reduction, square root, and natural log) to better achieve model stability, interpretability, and/or to capture nonlinear relationships.

validation. We begin with the *SubProgram* as the unit of analysis, but then switch to a broader subject defined as the *Program Category*. This transformation is crucial to infusing predictive capability into the macro-stochastic cost model. During validation, however, the unit of analysis reverts to the full *Program* in order to present model performance in a context most likely to resonate with target users. This nonstandard progression regarding the unit of analysis (i.e., model subject) is explained in greater detail at each step of model characterization.

Results

A Theoretical Macro-Stochastic Model

The first task is to assess the theoretical premise of a macro-stochastic cost model. A reasonable suspicion would be that the nature of cost estimating errors for defense programs—along with the underlying uncertainty which drives them—is inherently chaotic, such that attempting to characterize these errors via a stochastic process is misguided at best. Thus, the fundamental question that must be answered at the outset is whether there is any meaningful correlation between the variables in Table 2 and the level of accuracy in a given SAR's cost estimate. We believe that the data shown in Figures 1 through 4 offers a compelling response to this question.

Figure 1 is a plot of the percentage error in the empirical LCC estimates for all of the MDAPs listed in Table 1. Overall, the data exhibits a high level of dispersion. Although the mean error across all programs is only –4.7%, the mean *magnitude* of the errors (i.e., the mean absolute value of the errors) is over 22%. The magnitude error does appear to reduce slightly as time increases, suggesting that the accuracy of LCC estimates may be improving slightly as program acquisition matures. However, as noted in the characterization paper,

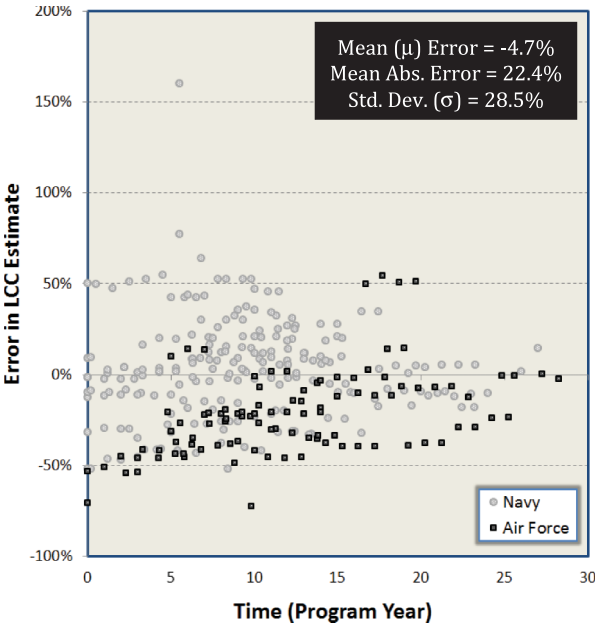

FIGURE 1 Error in LCC estimate as a function of time (empirical data) (color figure available online).

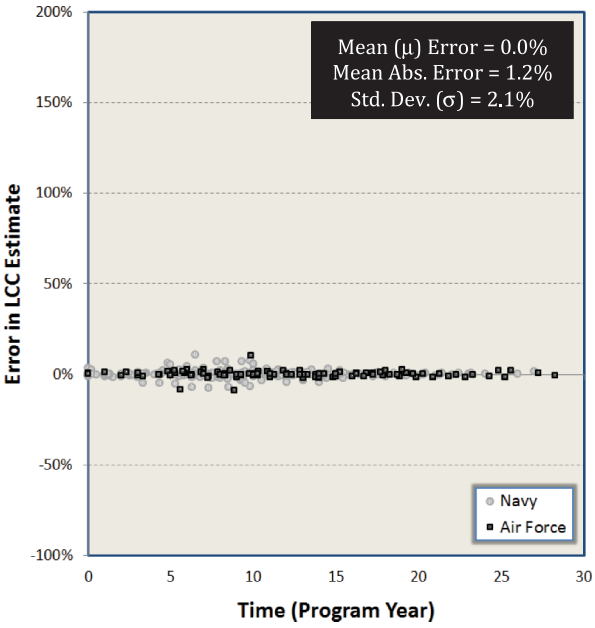

FIGURE 2 Error in LCC estimate as a function of time (theoretical macro-stochastic model) (color figure available online).

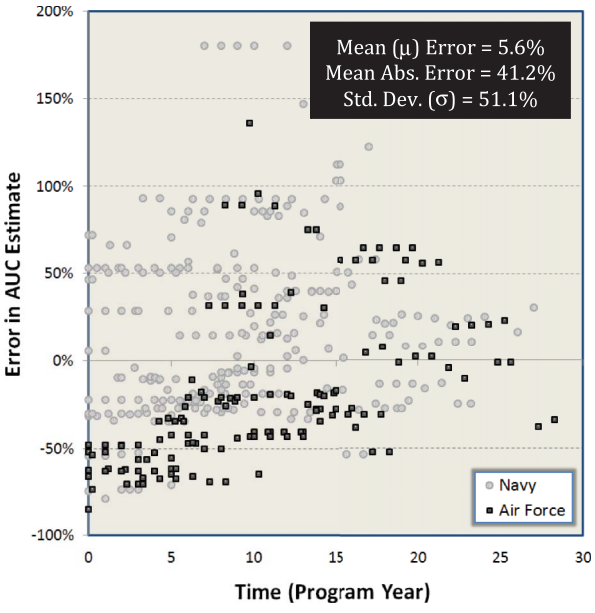

FIGURE 3 Error in AUC estimate as a function of time (empirical data) (color figure available online).

this is likely an artifact of the acquisition cost component of the LCC converging to a known value by the end of the acquisition phase (Ryan et al., 2012). When examining total O&S cost, per se, there is no significant improvement in LCC estimating accuracy as time goes on.

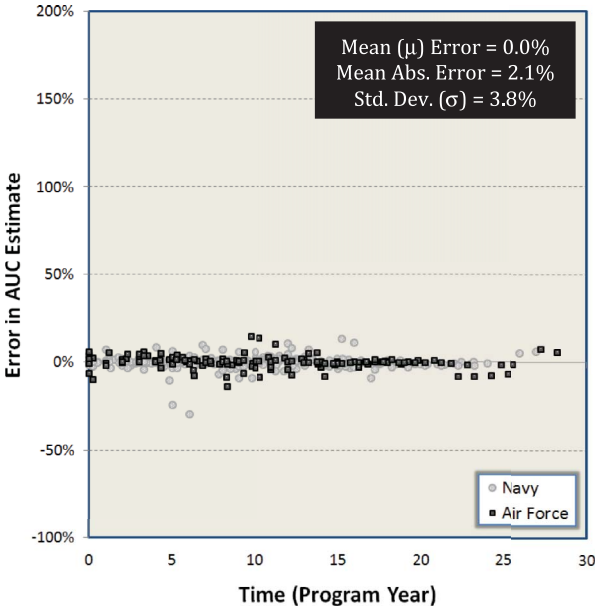

FIGURE 4 Error in AUC estimate as a function of time (theoretical macro-stochastic model) (color figure available online).

Figure 2 plots the results from a macro-stochastic mixed model that attempts to predict the error in each SAR and then compensate for it. The subject of this model is the SubProgram (for reasons explained in the characterization paper, the SubProgram—vice the Program—is the more appropriate unit of analysis). Designating the SubProgram (or the Program, for that matter) as the model subject is a logical choice, but it has important implications to model utility to be discussed shortly.

This so-called "theoretical macro-stochastic cost model" depicted in Figure 2 consists of just three variables: *Procure, Change* (#48 in Table 2), the standard deviation of the natural logarithm of *Acq Cost Est, BY10* (#14), and the natural logarithm of *LCC Est* (#17). All three variables are modeled as fixed effects, while the first two—along with an intercept term—are also modeled as random effects. The way to interpret this result is that the broad pattern (i.e., the fixed effects) of life cycle cost estimating errors in all Navy and Air Force MDAPs can be captured by examining the extent of procurement quantity changes to date, the variability of the acquisition cost estimates to date, and the current LCC estimate. Further, each program has its own pattern of errors (i.e., the random effects) driven by the procurement quantity changes and the variability of the acquisition cost estimate to date, as well as a unique starting point as defined by the intercept term.

Figures 3 and 4 are the AUC versions of Figures 1 and 2. Figure 3 shows that empirical AUC estimating accuracy for MDAPs is considerably worse than LCC estimating accuracy, with the magnitude of the errors and accompanying standard deviation almost twice as high. Figure 4 depicts the same data using a macro-stochastic model, and again, only three variables are used. This time the variables are the *Unit Acq Ratio* (#50), the standard deviation of the natural logarithm of *Acq Cost Est, BY10* (#14), and the natural logarithm of *AUC Est* (#15). As before, the first two variables are modeled as both fixed and random effects, and the model includes a random intercept term. The model subject remains the SubProgram.

In both cases, the theoretical macro-stochastic model performs impressively, driving down the magnitude of the mean error in the original prediction to a little over 1% in the case of LCC estimates, and just over 2% in the case of AUC estimates. Since the result is represented in percentage terms, it is easy to lose context of the amount of money involved. But these potential improvements in estimating accuracy typically represent billions of dollars. Since the mean magnitude error in the original LCC estimate is over 20%, a program estimated to cost $30.0 billion over its life cycle could be expected to actually cost somewhere in the range of $24.0 to $36 billion, a $12 billion range. On the other hand, the macro-stochastic model might predict a life cycle cost of $34.0 billion, but its equivalent expected error range would only be $800 million. Clearly, such a massive reduction in cost uncertainty would be of tremendous benefit to defense acquisition officials.

In one respect, this significant estimating improvement is an extremely important result. Figure 2 and Figure 4 are remarkable because they show the tremendous potential utility of the macro-stochastic cost modeling approach. With a highly parsimonious model, the model is able to predict the actual LCC and AUC estimating errors for all of the programs in this study with exceptional accuracy. Moreover, the random (subject-specific) effects are very powerful, strongly suggesting there is a unique pattern for each unit of analysis. This result is especially impressive given that there are over 35 SubPrograms in both models, over half of which consist of at least 10 data points (i.e., SARs) that must be "fitted."

However, in another—arguably more relevant—respect, this finding is of little utility. The problem with the preceding approach is that it is inherently a *post-hoc* analysis. This is why we refer to this model as "theoretical." One cannot expect that the exact cost estimating error patterns of these programs will occur again. So although using the SubProgram as the model subject may reveal powerfully descriptive random effects, the theoretical macro-stochastic model has no meaningful *predictive* capability.

The fact remains, however, that we now have some measure of confidence in the principle of macro-stochastic cost estimating of DoD programs. The challenge becomes how to translate this technique into a useful prognostic model.

Program Categories

In order to construct a predictive macro-stochastic model, the authors have devised a template-based solution involving the creation of Program Categories. This approach aims to achieve a better balance between model accuracy and utility by structuring the data into broader categories comprising multiple programs and using criteria that apply to both current and foreseeable programs. In this way, the Program Category supplants the SubProgram as the model subject and the unit of analysis.

To use a stock market analogy, the Program Category notion is the equivalent of forecasting an individual company's performance based on the business sector to which it is assigned. In the absence of company-specific performance indicators (which would be preferred, but may not be available until too late), we assume that the company's future performance will roughly conform to the average pattern of all the other companies in the same sector. A key to making this approach work, of course, is ensuring that companies (i.e., programs) are assigned to representative sectors (i.e., categories).

Indeed, establishing the exact Program Categories and ontological criteria was one of the most challenging aspects of model development. Our first goal was to be able to employ the model as early as possible, so the criteria used to assign a program to a particular Program Category had to be clearly discernible at the outset of a program. Second, we wanted the Program Category criteria to be simple and logical, easily derived from the list

of independent variables in Table 2. Third, we sought to have each category consist of programs similar enough to one another that the new model subject (i.e. the Program Category) would continue to exhibit statistically significant subject-specific patterns that could be captured by the random design matrix of the mixed model. (Given the complex interactions between various fixed and random effect model terms and the constituent covariance matrices, identifying meaningfully similar programs is often far from clear.)

In addition, the total number of program categories needed to be carefully considered as it represented another source of tension between accuracy and utility. If we create too few categories (i.e., many programs in a single category), the power of the mixed model is bound to be diminished as there will likely be little in the way of subject-specific effects to model. If we create too many categories, then we run the risk of building a model that is still too program-specific. In other words, if we have a large number of categories with a few programs in each, then we cannot—without additional data—have confidence that we have identified a valid Program Category that will effectively subsume a future program of interest.

We evaluated many different categorization structures defined via various variables and attribute thresholds, as well as varying numbers of categories. In the end, we empirically determined that the best balance of performance and utility was achieved through seven Program Categories defined by the following three variables: DoD component (#2), System type (#4), and Program size based on Acq cost est, BY10 (#14). Although the Program Category criteria were the same for both the AUC and LCC model, the specific programs and SAR counts are slightly different due to differences in data availability (see Table 1). Tables 3 and 4 show the Program Category structure and program assignments for each model.

Note that while the acquiring service component and the system type would not be expected to change during a program's life, the size of the program does change as acquisition cost estimates vary—sometimes significantly—over time. The dependence of the Program Category assignment on acquisition cost estimates introduces the possibility that a program's category assignment might change at some point in development. For the programs in our data set, this did not happen, but it could for some future program. If this were to occur, it's not clear whether that means the differently-sized program is in fact behaving more like the programs in its newly assigned category, or whether the size thresholds we have established here would need to be modified.

In addition, the fact that a surface maritime system (i.e., DDG 51) and a submarine system (i.e., SSN 774) are grouped together into a single category is likely to aggrieve the traditional cost estimator (as presumably would the grouping of fixed and rotary-wing aviation systems). Although both the surface vessel and the submarine are maritime systems, the Navy cost estimator knows that there are key cost-impacting differences between how each type of program is acquired and operated. With respect to the modeling approach pursued here, the point to keep in mind is that the pattern of program costs for similar systems is a fundamentally different phenomenon than the pattern of program cost *errors*. It is the latter that is relevant to our approach, and using this metric, the groupings in Tables 3 and 4 proved to be the most effective.

A Prognostic Macro-Stochastic Model

By restructuring the data from individual programs into Program Categories, we can now use the model to make predictions. Given the assumption that future programs are essentially like the programs in this data set, then as long as a future program can be assigned to one of the existing categories, the macro-stochastic model can be reasonably applied at any

TABLE 3 Summary of LCC macro-stochastic cost model program categories (PCats)

PCat	DoD comp	System type	Size (mean acq cost est, BY10)	SARs	# of programs	Assigned programs
1	AF	Aviation	Small (≤ $18.0B)	33	4	C-130J, JPATS, JSTARS, PREDATOR
2	Navy	Aviation	Small (≤ $18.0B)	53	5	C/MH-53E, E-2C, MH-60R, MH-60S, T-45TS
3	Both	Aviation	Large (> $18.0B)	60	5	C-17A, F-16C/D, F-22, F-14D, F/A-18E/F
4	Navy	Maritime	Small (≤ $8.5B)	41	7	AOE 6, CVN68 (74/75), CVN68 (76),MHC 51, SSGN, T-AKE, T-AO 187
5	Navy	Maritime	Medium ($8.5B–$30.0B)	42	3	LHD 1, LPD 17, SSN 21
6	Navy	Maritime	Large (> $30.0B)	36	2	DDG 51, SSN 774
7	Both	Munition	All	52	5	AIM-9X, AMRAAM-AF, AMRAAM-JT, JASSM, JSOW
TOTAL				**317**	**31**	

TABLE 4 Summary of AUC macro-stochastic cost model program categories (PCats)

PCat	DoD comp	System type	Size (mean acq cost est, BY10)	SARs	# of programs	Assigned programs
1	AF	Aviation	Small ($\leq$ \$18.0B)	58	5	C-130J, GLOBAL HAWK, KC-135A, JPATS, JSTARS
2	Navy	Aviation	Small ($\leq$ \$18.0B)	68	6	AV-8B, C/MH-53E, E-2C, MH-60R, MH-60S, T-45TS
3	Both	Aviation	Large ($>$ \$18.0B)	83	6	C-17A, F-16C/D, F-22, F-14D, F/A-18C,F/A-18E/F
4	Navy	Maritime	Small ($\leq$ \$8.5B)	52	8	AOE 6, CVN68 (74/75), CVN68 (76), MHC 51, SSGN, STRAT. SEALIFT, T-AKE, T-AO 187
5	Navy	Maritime	Medium (\$8.5B–\$30.0B)	42	3	LHD 1, LPD 17, SSN 21
6	Navy	Maritime	Large ($>$ \$30.0B)	36	2	DDG 51, SSN 774
7	Both	Munition	All	53	5	AIM-9X, AMRAAM-AF, AMRAAM-JT, JASSM, JSOW
TOTAL				**392**	**35**	

time after program initiation to predict the expected error in the program's cost estimate and, by extension, predict the actual LCC or AUC.

This improved utility has come at a cost, however. The powerful program-specific trends depicted in Figure 2 and Figure 4, which consisted of only three independent variables, are diluted by the amalgamation with other—albeit similarly behaving—programs. In essence, the new model subject of Program Category requires that the random effects design matrix (**Z**) compromise between multiple, different program trends, resulting in reduced model performance. Or, to continue the market analogy, a particular company's performance is not likely to *exactly* follow the average of its assigned sector: There will be important company-specific deviations. Fortunately, we can restore a large degree of expected model performance though the inclusion of additional variables.

The final LCC macro-stochastic model incorporated 12 variables (to include 5 random variables) from Table 2 and the final AUC macro-stochastic model incorporated 14 (to include 6 random variables). The selected fixed and random variables, along with their estimated parameter values, are listed in Tables 5 through 8. Since the random variables vary by Program Category, they are specified in their own tables. The reader should be cautious in making inferences based on relative parameter estimate values as not all variables are normalized, and the relationship between parameters is complicated by the inclusion of both fixed and random effects.

Note that over half of the final variables from both models are capturing, in some manner, program trends to date regarding the estimated cost and/or production quantity. These variables may capture trends either directly by what is being measured (e.g., Nunn McCurdy Breach, Cost Variance, etc.) or indirectly via changes in a given variable to date (e.g., standard deviation, mean, etc.). Regardless, a consequence of this predominance of trending variables is that a program should have at least one previous SAR on which to construct a trend value; without a previous SAR, we find that model performance diminishes considerably. In practice, this results in a small impact on the utility of the model in that

TABLE 5 LCC macro-stochastic model variables and fixed effects parameter estimates

#	LCC model variable	Random effect?	Fixed effect estimate
1a	*DoD Component* (#2)—Navy	No	0.4157
1b	*DoD Component* (#2)—Air Force	No	0.0000
2a	*Acq Type* (#6)—New	No	0.2132
2b	*Acq Type* (#6)—Modification	No	0.2183
2c	*Acq Type* (#6)—Variant	No	0.0000
3	Weighted Mean of Normalized *Acq Cost Est,* BY10 (#14)	Yes	3.2555
4	Std. Dev. of Natural Log of *Acq Cost Est,* BY10 (#14)	No	9.3392
5	Natural Log of *LCC Est* (#17)	No	7.1928
6	Mean of Natural Log of *LCC Est* (#17)	No	−4.8595
7	Maximum *CV, Est* (#25)	Yes	−0.7387
8	Slope of Regression Line of *CV, Quan* (#26)	Yes	0.2188
9	Standard Deviation of *CV, Total* (#27)	No	−1.6512
10	Range of *CV, Total-Quan* (#28)	Yes	0.7593
11a	*Breach, N-M* (#34)—No	Yes	−3.1063
11b	*Breach, N-M* (#34)—Yes	Yes	−3.1440
12	Std. Dev. of Square Root of *Procure, Change* (#48)	No	−0.3264

TABLE 6 LCC macro-stochastic model random effects parameter estimates by program category (PCat)

#	LCC model variable	Random effect estimate						
		PCat1	PCat2	PCat3	PCat4	PCat5	PCat6	PCat7
1	Wtd. Mean of Normalized *Acq Cost Est, BY10*	−3.5641	4.4451	−2.0476	−3.4837	3.2109	−3.4782	4.9177
2	Maximum *CV, Est*	0.2384	−2.0194	−1.1120	0.5359	1.0697	0.7222	0.5652
3	Slope of Regression Line of *CV, Quan*	0.3989	0.1223	−1.1514	−0.4866	2.3672	−0.6491	−0.6013
4	Range of *CV*, *Total-Quan*	−0.1111	1.0039	0.3164	1.8285	−2.2551	−0.2410	−0.5417
5a	*Breach, N-M*—No	0.4924	−0.0670	−0.4322	0.2778	−0.5395	−0.0921	0.3606
5b	*Breach, N-M*—Yes	0.5163	−0.0541	−0.4083	0.5018	0.0801	0.0442	−0.7883

TABLE 7 AUC macro-stochastic model variables and fixed effects parameter estimates

#	AUC model variable	Random effect?	Fixed effect estimate
1a	*DoD Component* (#2)—Navy	No	0.4687
1b	*DoD Component* (#2)—Air Force	No	0.0000
2a	*Acq Phase* (#5)—Development	Yes	1.6995
2b	*Acq Phase* (#5)—Production	Yes	1.6816
3a	*Acq Type* (#6)—New	No	0.3993
3b	*Acq Type* (#6)—Modification	No	−0.1132
3c	*Acq Type* (#6)—Variant	No	0.0000
4	Mean of Scaled *Acq Cost Est, BY10* (#14)	Yes	0.4536
5	Natural Log of *AUC Est* (#15)	No	0.6391
6	Mean of Natural Log of *AUC Est* (#15)	No	−0.5730
7	Maximum *CV, Engr* (#24)	Yes	1.4515
8	Weighted Mean of *CV, Est* (#25)	No	1.5208
9	*CV, Quan* (#26)	No	0.8438
10	Mean *CV, Total* (#27)	No	−1.2817
11	Wtd. Mean of Natural Log of *Procure, Plan* (#47)	Yes	0.1570
12	Mean of Square Root of *Procure, Plan* (#47)	No	0.2402
13	Wtd. Mean of Square Root of *Procure, Change* (#48)	Yes	−0.1111
14	*Unit Acq Ratio* (#50)	Yes	5.8501

it is not suitable for use until the second SAR, which is nominally one year after program initiation.

Figure 5 and 6 show, respectively, the performance of the LCC and AUC *prognostic* macro-stochastic models where the subject equals the Program Category. Each model is capable of predicting the accuracy of a current LCC or AUC point estimate at any point in a program's life where at least two SARs are available, and then compensating for that error to provide a statistically more accurate estimate. Although model performance is not as impressive as it was for the theoretical model (where the model subject was SubProgram), it is still far better than current estimate performance. The mean magnitude error in the prognostic LCC macro-stochastic model is more than a fourfold improvement of the empirical estimate; for the AUC model, the improvement is over fivefold.

A Prognostic (and Validated!) Macro-Stochastic Model

The performance shown in Figures 5 and 6 was achieved under conditions in which the training data set and the test (i.e., validation) data set were equivalent. Thus, it is reasonable to suspect that actual model performance against future programs will be reduced (Larson, 1931; Hart & Wehrly, 1986). In order to validate the model, we need to test its performance against data that is not available to the model. We certainly don't want to wait several years for new program data to become available, but the current size of the data set is an impediment to a standard data partitioning techniques (i.e., dedicated training and test data sets). With respect to validation, the most logical unit of analysis is the program, as that is the fundamental entity for cost estimation and cost accrual accounting in the DoD. For both the LCC and AUC model, however, we have fewer than three dozen programs available for

TABLE 8 AUC macro-stochastic model random effects parameter estimates by program category (PCat)

#	AUC model variable	Random effect estimate						
		PCat1	PCat2	PCat3	PCat4	PCat5	PCat6	PCat7
1a	*Acq Phase* (Development)	−1.2914	0.4545	0.3163	0.0417	−0.2775	−0.0037	0.7601
1b	*Acq Phase* (Production)	−1.2125	0.5695	0.2867	−0.4582	−0.2023	−0.0122	1.0289
2	Mean of Scaled *Acq Cost Est, BY10*	−0.1246	0.3191	−0.4065	−0.6093	0.4997	−0.2703	0.5919
3	Maximum *CV, Engr*	−1.2665	−0.5047	1.8679	−0.0688	−0.2330	0.2027	0.0024
4	Wtd. Mean of Natural Log of *Procure, Plan*	0.0549	−0.0547	0.2851	0.2652	−0.2904	0.4902	−0.7503
5	Wtd. Mean of Square Root of *Procure, Change*	0.4584	0.1701	0.3614	0.0257	−0.6929	−0.4415	0.1188
6	*Unit Acq Ratio*	−2.0932	−0.2902	1.0105	−0.0981	−0.4456	1.4998	0.4168

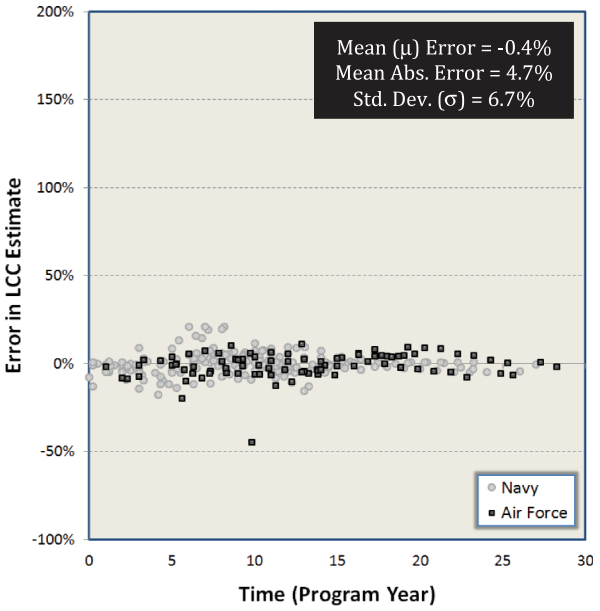

FIGURE 5 Error in LCC estimate as a function of time (prognostic macro-stochastic model) (color figure available online).

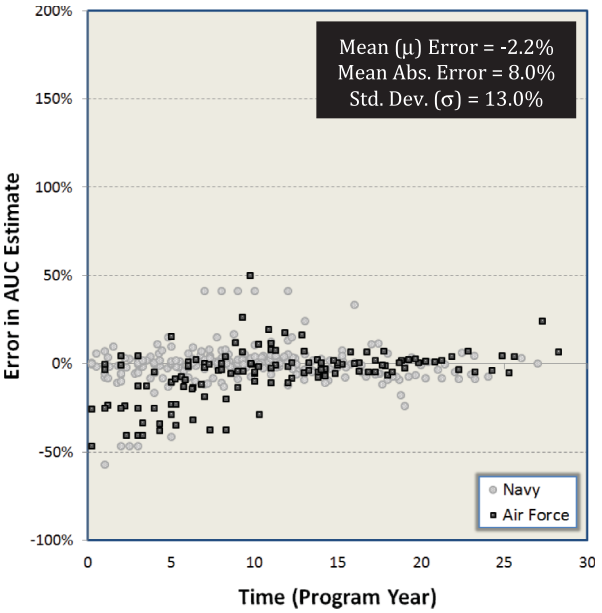

FIGURE 6 Error in AUC estimate as a function of time (prognostic macro-stochastic model) (color figure available online).

analysis, hardly sufficient to execute a robust validation involving separate training and test data sets.

This leads us to cross-validation. However, the specific method of cross-validation for the macro-stochastic model is more complicated than it might at first seem. The

non-i.i.d. nature of the data also invalidates standard cross-validation techniques: omitting an observation (i.e., one SAR) does not remove the associated information due to correlations with other observations from that subject (Opsomer, Wang, & Yang, 2001; Arlot, 2010). Suggested techniques to work around this problem include modified cross-validation (Chu & Marron, 1991), *h*-block cross-validation (Burman, 1994), and sequential validation (Bengio & Chapados, 2003).

Unfortunately, none of these techniques are well suited for the structure of the MDAP data. Not only is the correlation distance (i.e., the strength of the correlation) highly dependent on the program, but several programs have an insufficient number of SARs to faithfully implement the given technique. For instance, in the case of *h*-block cross-validation, determining the theoretically appropriate size of *h* in our data set is not clear, but it must be relatively large, and any value of *h* greater than two could eliminate as many as six programs from the validation.

As a result, we have implemented a tailored version of Leave One Out Cross Validation (LOOCV). Ordinarily, the "one" in LOOCV refers to a single observation, which is held out from the data set and used for validation after the model is trained on the remaining data. This process is then repeated for every data observation. Given the correlations within a program, we have redefined the "one" to denote an *entire Program*. This is an appealing strategy for two reasons. First, this is the level at which the correlations exist, so omitting an entire Program is the only assured method for fully eliminating the correlations. Second, despite restructuring the data into Program Categories, principal cost estimating interest remains with the Program, so that is the appropriate level for assessing model performance. Thus, for validation purposes, the entire Program (not just the SubProgram) becomes the unit of analysis and the observation left out.

After removing a given Program from the data set, we train the model using the remaining data and use the omitted Program as the test set. Then we record how the model performed against that Program. We repeat this process for every Program in the data set. This results in 30 (the C/MH-53E program cannot be validated because it only has one valid LCC SAR) separate validations for the LCC model and 35 for the AUC model, which are then amalgamated into a single summary of overall validated model performance. This is a particularly rigorous validation as no information regarding the program to be tested remains embedded in the model. Also note that the Program Category structure still applies. This means that when validating certain programs (particularly the large and medium maritime categories) very few programs remain in the category to form the basis of the (i.e., train) Program Categorization parameters (refer to Tables 3 and 4). Nevertheless, the validated version of each model performs well.

Figure 7 shows the resultant validated performance of the macro-stochastic prognostic LCC model based on our tailored LOOCV technique. This performance reflects model-corrected LCC estimates for every program with at least two valid SAR-derived LCC estimates. The analogous results for the AUC version of the model can be seen in Figure 8. As one would expect, model performance has diminished relative to the non-validated version of the model, but it still remains significantly better than empirical performance. The mean magnitude error in the validated LCC macro-stochastic model is 2.1 times better than the empirical estimate; for the AUC model, the model is 2.6 times better.

Figure 9 compares the mean magnitude error per SAR in the empirical data to that of both the AUC and LCC validated models across all programs. For reference, performance of the non-validated version of each model is also shown. To ensure a fair comparison, all SARs omitted from the macro-stochastic models (i.e., initial SARs) were also omitted from the empirical data, which is why the mean magnitude errors for the empirical data are slightly different from those shown in Figures 1 and 3.

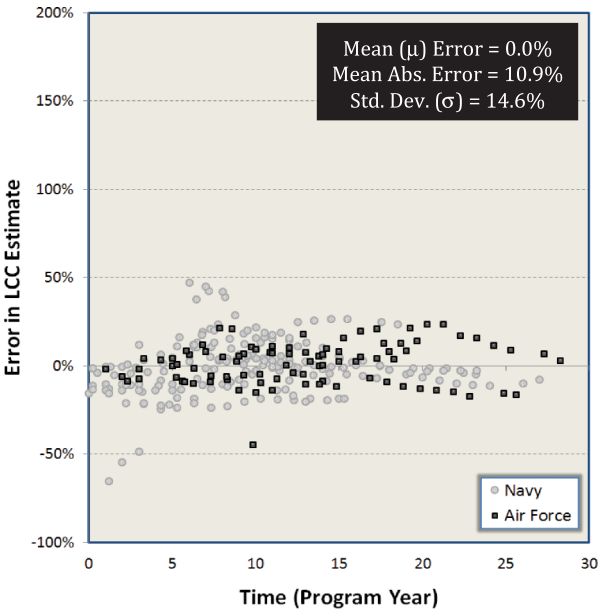

FIGURE 7 Error in LCC estimate as a function of time (validated prognostic macro-stochastic model) (color figure available online).

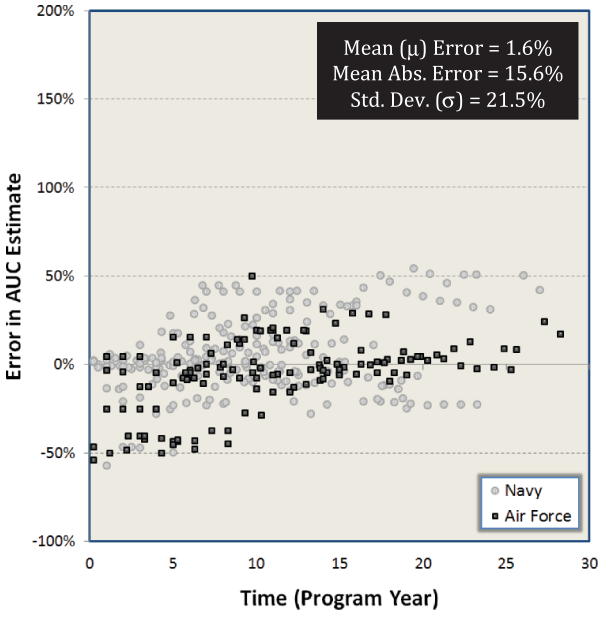

FIGURE 8 Error in AUC estimate as a function of time (validated prognostic macro-stochastic model) (color figure available online).

Figure 10 shows another measure of model effectiveness, which is essentially "head-to-head" performance of each macro-stochastic model to the empirical estimates. This program-by-program comparison shows that the validated LCC model performs better (i.e., has an overall lower error across all the SARs of a given program) in 23 of 30 cases. The validated AUC model performs better for 31 of the 35 programs.

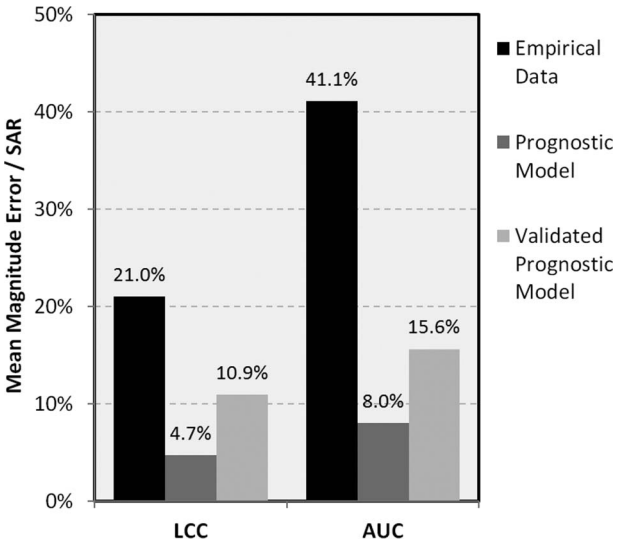

FIGURE 9 Model performance as measured by mean magnitude error per SAR.

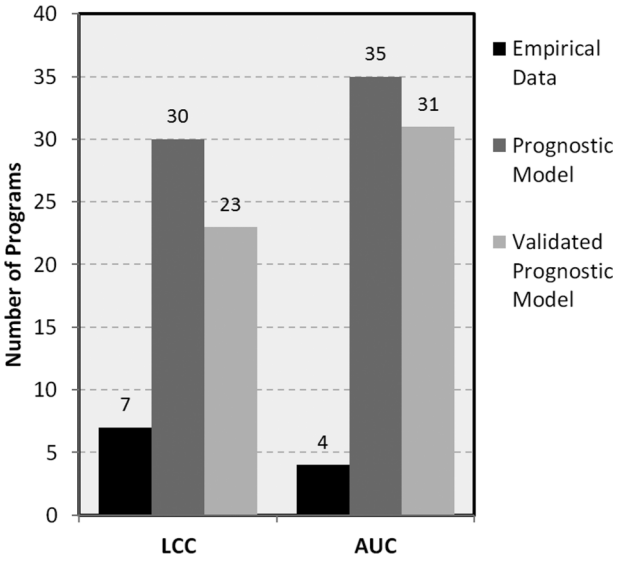

FIGURE 10 Model performance as measured by number of programs with lower overall error.

Discussion

Key Findings

Although the validation results of the LCC and AUC macro-stochastic prognostic models yield sizeable errors, we find that overall accuracy for both models is significantly better than what was achieved in the original SAR estimates. The predicted LCC value from the validated macro-stochastic model had a mean absolute error of just under 11% compared to a 21% error in the historical program estimates. Given that the total LCC across all

of the programs we evaluated was approximately $800 billion, the model-predicted LCC estimates represent an improvement in estimating performance of about $80 billion, or an average of $2.6 billion per program. For the AUC estimates, the improvement was even greater. The mean magnitude error of the historical cost estimates was over 40%, while the model estimates had a mean magnitude error of less than 16%. Again, this translates to cost fidelity improvements measured in billions of dollars.

Improvements in the mean errors tell part of the story, but program-by-program performance is also important, and the macro-stochastic models performed well by this metric as well. If the macro-stochastic prognostic model presented here had been used to estimate LCC costs for every SAR of the programs in the LCC data set (aside from the first), the model-based estimate would have had a lower overall error than the original estimate for 23 of 30 programs (77%). In the case of the AUC estimates, the model would have performed better for 31 of 35 programs (89%).

Not only can the original program estimate be improved dramatically using a macro-stochastic derived correction factor, but it can also be accomplished with minimal effort. The specific variables that feed each model are easily derived from data routinely available in the program's SARs. Program values observed for these variables can be transcribed into the model formula at any point after the program's second SAR, and a macro-stochastic estimate derived in just a few hours.

The fact that trending variables were found to be statistically significant predictors of LCC and AUC estimate errors is an intriguing result, but difficult to fully explain. Recall that the original estimates developed by the program had access to all of the same information (and far more) available to us in the SAR. Thus, any cost-impacting changes to the program *should* have been incorporated into the latest SAR estimate. It may be that the full cost implications of certain types of baseline changes are not fully understood until later in the program. Or it may be that certain types of historical program instability are likely to persist and/or permeate other elements of the program in ways that distort expected costs. In any case, the prominence of the trending variables make it tempting to conclude that change and cost instability tends to beget further change and cost instability. But this interpretation is too simplistic, and frankly not warranted based on the data. Instead, our conclusion is more nuanced: *Certain types and degrees of change in certain types of programs do tend to affect the ultimate accuracy of the current cost estimates in relatively predictable ways*.

Perhaps of equal interest to the parameters included in the model are those that were omitted, i.e., those that never significantly contributed to model performance. Notable non-contributors were *Joint* (#3), APB-related variables (#8–11), *Prime Contractor* (#12), PAUC/APUC-related variables (#20–23), Requirement Changes (#41–46), *Program Year* (#1), *Maturity* (#7), and *Expended* (#18). The last three are perhaps the most surprising, as one would expect that variables that capture program age would be a good indicator of cost estimate accuracy (with the presumption that estimate accuracy improves as programs mature). Since they were not, this serves as additional evidence of the finding presented in the characterization paper, i.e., that LCC and O&S cost estimates for MDAPs are improving very little, if at all, over time (Ryan et al., 2012).

We believe that the LCC and AUC macro-stochastic cost models presented here are ready for trial use. However, it is important to understand a fundamental constraint on their intended implementation. Note that both the LCC and AUC models require as an input all of the subject program's respective cost estimates to date (see Tables 5 and 7). This means, for one, that the output of the macro-stochastic model would generally not be suitable for internal program use. Unless perhaps implemented as a final validation check, awareness of the macro-stochastic output could influence the official SAR cost estimates, which in turn,

would likely bias the output of the macro-stochastic model. This is because the macro-stochastic model implicitly relies on the continuation of current cost estimating practices; any deviation from these practices, to include modifying the estimate based on the results of the macro-stochastic model, could fundamentally change the stochastic nature of this key input variable.

The dependence of the macro-stochastic cost model on the program's cost estimate also means that it is *not meant to be used in lieu of existing program estimates*. The traditional cost estimate may be perfectly accurate given the current baseline, which is an important input, per se, for senior decision-makers. The macro-stochastic model, on the other hand, is intended to be a complementary data point—it provides leadership the equivalent of a stochastic cost vector, i.e., a *probabilistic indication of where program costs are likely to end up*.

As a consequence of these implementation constraints, the authors envision that these models could be most effectively employed by cost validation entities outside the acquisition chain of command. An independent cost estimate is required for all MDAPs, which is provided by either the service cost agency or the Office of the Secretary of Defense, Cost Assessment and Program Evaluation (OSD/CAPE). Either of these entities may find the output of the macro-stochastic model highly useful when conducting their independent analyses. The Defense Acquisition Executive (DAE) and/or the Defense Acquisition Board (DAB) are also potential consumers, as they each require independent cost estimates as part of their review process, and the macro-stochastic model estimate could serve as an alternate source of realistic cost validation (GAO, 2009; DAU, 2012).

Another potential user of this type of model would be the DoD service component acquisition portfolio manager, who is often required to manage the execution of several similar defense systems. The macro-stochastic model may be especially suitable in this case, as the portfolio manager is likely to be responsible for multiple systems from the same Program Category, and more accurate insights into overall portfolio cost commitments could be invaluable. Moreover, using the model for several contemporaneous programs would reduce the susceptibility of the predicted values being skewed by statistical outliers. Although the macro-stochastic model may certainly be applied to—and has been validated against—individual programs, one would expect it to perform more consistently when multiple programs are being simultaneously evaluated. This suspicion can be partially confirmed by examining aggregated program performance at the Program Category level. Although the results are not presented here due to space considerations, we did find that both models provided significantly improved estimates across every Program Category.

Issues and Concerns

Not surprisingly, the macro-stochastic model will sometimes predict an error estimate that overcorrects the program estimate, such that an underestimate becomes an overestimate, and vice versa. This is a natural consequence of that fact that the model is attempting to minimize variance around a "perfect" estimate (i.e., zero error), which means that it implicitly regards an overestimate as equally undesirable as an underestimate. This can (and does) create the following type of situation: The original estimate is 20% too high (or too low), but the model-corrected estimate is 10% too low (or too high). The question arises of whether we would be better off budgeting 20% too much or 10% too little. Although both underestimates and overestimates are undesirable from a budget planning perspective, there are situations where one type of error may be preferred to the other. The macro-stochastic model could certainly be tailored to reflect such preferences through a zero error offset.

Somewhat related to the issue of overcorrections are the occasional instances where the model predicts an extremely large estimate error. While these predictions of *massive* errors—once applied to the original estimated cost—sometimes produce a more accurate estimate, they can also lead to unrealistic results, such as when the model predicts that the actual LCC or AUC has been underestimated by more than 100% (unless one wishes to advocate the possibility that Pentagon programs could turn a profit!). To avoid these types of nonsensical outcomes, we have embedded a threshold mechanism into the prognostic model such that the original estimates—regardless of what error the model predicts—are not corrected by more than a factor of two. In other words, the prediction of actual cost after correction for the model-predicted error will never be more than double the empirical estimate, nor less than half. In principle, the threshold could be much higher, but this level seemed appropriate from a practical standpoint. Although the program LCC and AUC estimates are sometimes inaccurate by a factor greater than two, corrections that require more than doubling or halving of the program estimate would—even if valid—likely be regarded with justifiable skepticism. Note that while thresholding did provide an improvement to overall model performance, the effect was marginal, and it was not implemented often. The threshold constraint affected the output in 26 of 709 cases (3.7%), and nearly half of these instances occurred on a single program (C-130J).

Another potential concern is long-term model reliability. As discussed in the previous section, the current iteration of both macro-stochastic models relies on official program estimates to produce its own estimate. This fact introduces an inherently recursive—and potentially unstable—element to longer-term model use. We know that senior defense leaders make key decisions based on the traditional program cost estimates, and that these estimates are often highly inaccurate. The nature of those decisions—and thus the ultimate trajectory of certain types of programs—may be substantively different if the decision-maker has access to more accurate cost estimates. For instance, programs that would otherwise be cancelled might instead be funded, and vice versa. This in turn, could create a negative feedback loop where cost estimate trajectories of certain program categories no longer conform to the patterns that characterize the programs that we have seen to date, thereby reducing the predictive capacity of the macro-stochastic model. Though highly speculative, this argument points to the need for continued refinement of the model as more data becomes available.

Perhaps the most significant barrier to macro-stochastic model implementation relates to the fact that it represents a fundamentally different approach to DoD cost estimating. In particular, it could be viewed in many respects as inherently non-transparent. In contrast to a traditional bottoms-up cost estimate, the specific drivers of the macro-stochastic cost estimate are not directly traceable, nor fully explainable. Users could be inclined to view this type of model as too opaque, in that the output may in fact be probabilistically more accurate, but the internal workings are inscrutable. Nevertheless, the results presented here are compelling: Independent cross-validation verifies the improvements in long-term DoD cost estimates that may be achieved by adjusting the cost estimates using the model-predicted error.

In practice, the most important caveat to using this model pertains to the Program Category structure. This construct was a strategy employed to transform the theoretical macro-stochastic model into a useful prediction tool. However, it is only a valid construct to the extent that current programs are representative of future programs, and those future programs really do "fit" into one of these established categories. Expanding on this point, the number of programs in Program Categories five (medium maritime) and six (large maritime) are fewer than we would prefer. By only having two to three constituent programs, we run the risk identified early on, i.e., that the defined Program Category may not be

sufficiently representative of the next program to be assigned. Therefore, users of the current iteration of the macro-stochastic model may wish to be more wary when employing the model against these two Program Categories. This concern can be significantly mitigated only with the passage of time and the inclusion of more data.

Finally, a methodological note of caution: The specific model variables selected as well as the parameter estimates are based on the results of the previously completed characterization study (Ryan et al., 2012). Therefore, we recommend that potential users familiarize themselves with that study in order to understand the potential issues and biases documented there before employing the macro-stochastic model. If the specific findings of the characterization study are not valid, then the specific variables and parameter values of this model are not likely to be valid either. Note, however, that concerns about the methodology of the characterization study would not be expected to weaken the underlying premise of macro-stochastic cost estimation; it would only affect its specific formulation.

Future Work

There are a number of ways in which the reliability, accuracy, or utility of the macro-stochastic model could be improved. One beneficial task would involve using more current data to reproduce the characterization study and rebuild the model. The current data set is based on information available as of mid-2011. By expanding the data set to incorporate more recent SAR data and cost actuals, one could conceivably expand the data set by approximately ten percent in terms of the SAR count, and five percent with respect to program count. This additional data could help identify flaws in the model or increase confidence in the current implementation, especially if conducted by an independent source. Alternatively, model reproduction on a larger data set could allow the model parameters and Program Categories to be further optimized.

The utility of the macro-stochastic model might be significantly improved by extending its applicability to earlier MDAP cost estimates. The availability of a more accurate cost estimate prior to Milestone B could be especially valuable, as this milestone requires independent certification of program cost reasonableness and affordability (DAU, 2012). But, as previously noted, the model is currently constrained by the need for certain trending variables, which are ostensibly not available until the second SAR. Ordinarily, SARs are not required until after Milestone II/B, but some MDAPs do submit what are known as RDT&E (Research, Development, Test, and Evaluation) SARs prior to Milestone II/B. These SARs nominally exclude certain key cost categories (e.g., O&S costs), but if enough of these types of SARs exist, and they are otherwise sufficiently extensive, they could be evaluated for model inclusion. Alternatively, it may be possible to obtain values for the model parameters from non-SAR sources.

Lastly, a couple of specific model concerns articulated in the previous section relate to the fact that the model is dependent on the program's cost estimate. Therefore, a key improvement would be to build the model without using any of the program's estimates as independent variables. This approach might reduce model performance, but it would also improve long-term model reliability and ensure the output is truly a functionally *independent* cost estimate. This modification could also be instrumental in beginning to characterize the cost-effectiveness of the aforementioned procurement strategy polices related to acquisition, contracting, and sustainment. Although we previously commented on the difficulty of quantifying these elements via the SARs, some degree of quantification could be very beneficial, and alternate data sources (besides the SAR) could be considered.

Conclusion

Despite the fact that DoD cost estimating practices have become increasingly sophisticated, the actual program cost estimates that are produced remain poor, at least when compared to the final, actual costs of the program. Our hypothesis is that this deficiency is largely due to the fact that current cost estimating techniques must assume a fixed program baseline. As a way around this unrealistic assumption, we have proposed a fundamentally different approach to cost estimating that attempts to capture this uncertainty by modeling the error in the program estimate as a random variable. We found that the value of this variable is largely unique to a given program—and even a group of programs, to some extent—and could be predicted reasonably well through a relatively small number of top-level program summary indicators gleaned from the annual SARs.

The macro-stochastic model represents an intriguing option for vetting program estimates of Life Cycle Cost and Annual Unit O&S Cost. It not only appears to provide cost estimates that are significantly more accurate than those reported in the original SAR estimates, but the amount of effort needed to construct the estimates is minimal. Although the current version of the macro-stochastic model is not suited for replacing existing program cost estimates, the authors believe it could be extremely useful to independent costing entities outside the acquisition chain of command who are seeking a more realistic assessment of system value or program affordability.

Note

1. "AUC" is not a standard DoD acronym; the authors have coined it for convenience in the context of this application. Further, AUC should not be confused with the APUC (Average Procurement Unit Cost).

References

Arlot, S. (2010). A Survey of Cross-Validation Procedures for Model Selection. *Statistics Surveys*, 4, 40–79.

Bengio, Y., & Chapados, N. (2003). Extensions to Metric-Based Model Selection. *Journal of Machine Learning Research*, 3, 1209–1227.

Burman, P. (1994). A Cross-Validatory Method for Dependent Data. *Biometrika*, 81(2), 351–358.

Chu, C., & Marron, J. (1991). Comparison of Two Bandwidth Selectors with Dependent Errors. *Annals of Statistics*, 19(4), 1906–1918.

DAU. (2012). *Defense Acquisition Guidebook*. Retrieved from https://dag.dau.mil/Pages/Default.aspx

Diggle, P., Liang, K., & Zeger, S. (1994). *Analysis of Longitudinal Data*. Oxford: Clarendon Press.

DoD. (1992). *Cost Analysis Guidance and Procedures* (DoD 5000.04-M). Washington, D.C.

Drezner, J., & Krop, R. (1997). *The Use of Baselining in Acquisition Program Management*. Santa Monica, CA: RAND.

GAO. (2009). Cost Estimating and Assessment Guide. Washington D.C.: GAO.

GPO. (2011). Budget Control Act of 2011. U.S. Government Printing Office. Retrieved from http://www.gpo.gov/fdsys/pkg/PLAW-112publ25/pdf/PLAW-112publ25.pdf

Hart, J., & Wehrly, T. (1986). Kernel Regression Estimation Using Repeated Measurements Data. *Journal of the American Statistical Association*, 81(396), 1080–1088.

Hebert, A. (2011, July). Lies, Damn Lies, and the Trillion-Dollar F-35. *Air Force Magazine*, 94(7), 4.

Larson, S. (1931). The Shrinkage of the Coefficient of Multiple Correlation. *Journal of Educational Psychology*, 22, 45–55.

NASA. (2008). Report of Audit and Finance Committee. NASA Advisory Council. Retrieved from http://www.nasa.gov/pdf/314880main_AFC_KSC_NAC_Feb-5-2009.pdf

Opsomer, J., Wang, Y., & Yang, Y. (2001). Nonparametric Regression with Correlated Errors. *Statistical Science*, 16(2), 134–153.

OSD CAIG. (2007). Operating & Support Cost Estimating Guide. Acquisition Community Connection. Office of the Secretary of Defense Cost Analysis Improvement Group. Retrieved from https://acc.dau.mil/adl/en-US/142233/file/27619/O_S_Cost_Estimating_Guide_Oct_2007.pdf

Patetta, M. (2002). *Longitudinal Data Analysis with Discrete and Continuous Responses Course Notes*. Cary, NC: SAS Institute.

Ryan, E., Jacques, D., Ritschel, J., & Schubert, C. (2013). Characterizing the Accuracy of DoD Operating and Support Cost Estimates. *Journal of Public Procurement*, 13(1), 71–101.

Verbeke, G., & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.

Wijker, J. (2009). *Random Vibrations in Spacecraft Structures Design: Theory and Applications*. New York: Springer.

Wolfinger, R. (1993). Covariance Structure Selection in General Mixed Models. *Communications in Statistics, Simulation and Computation*, 22(4), 1079–1106.

About the Authors

Major Erin Ryan, Ph.D., is an Assistant Professor in the Department of Systems Engineering and Management at the Air Force Institute of Technology (AFIT). In addition to his Ph.D. in Systems Engineering from AFIT, Maj Ryan also holds degrees in Electrical Engineering from the University of Washington and National Security Studies from New Mexico State University. Major Ryan's principal experience to date has been in the intelligence and space communities, serving as the Contracting Officer's Technical Representative or Program Manager for multiple space-related programs. His principal research interests are stochastic modeling, decision analysis, life cycle cost estimating, and space architectures.

Dr. Christine Schubert Kabban is an Assistant Professor of Statistics in the Department of Mathematics and Statistics at AFIT. She received her Ph.D. in Applied Mathematics from AFIT and returned to AFIT after five years in the Department of Biostatistics at VCU. Her research interests include classification and detection methods, ROC surfaces and diagnostic testing, information fusion, and multi-level modeling.

Dr. David Jacques (Lt Col, USAF-Ret), is an Associate Professor of Systems Engineering on the faculty at AFIT. He holds a Ph.D. and M.S. in Aeronautical Engineering from AFIT, and a B.S. in Mechanical Engineering from Lehigh University. Dr. Jacques is currently curriculum chair for Graduate Systems Engineering at AFIT. His research interests are in the areas of concept definition and evaluation, architecture modeling, and optimal system design. Varied applications of his research involve experimental test of multi-UAV cooperative control and autonomous munition concepts, and networked sensor approaches to chem/bio detection within a building.

Lieutenant Colonel Jonathan D. Ritschel, Ph.D., is an Assistant Professor and Director Cost Analysis Program in the Department of Systems Engineering and Management at AFIT. He received his BBA in Accountancy from the University of Notre Dame, his M.S. in Cost Analysis from AFIT, and his Ph.D. in Economics from George Mason University. Lt Col Ritschel's research interests include public choice, the effects of acquisition reforms on cost growth in DOD weapon systems, research and development cost estimation, and economic institutional analysis.