Studies of Confidence in Software Cost Estimation Research Based on the Criterions MMRE and PRED

Dan Port (UHawaii) Vu Nguyen (USc) Tim Menzies (WVU)
dport@hawaii.edu, nguyenvu@usc.edu, tim@menzies.us

Jan 19, 2009

ABSTRACT

Confidence in cost estimation is different than model accuracy. It is related to the significance of results based on model accuracy measures such as MMRE and PRED. A lack of confidence places uncertainty in the accuracy for predicted values and the conclusions based on accuracy measure. This work is a critical empirical and analytic investigation of standard error in the accuracy criterions MMRE and PRED. Center in this is the conclusion instability problem which we show can be resolved and greater confidence can be found in the use of model accuracy criteria in practice. Five published studies are replicated and extended and their results are recast with consideration of standard error. We explore a number of practical applications such as estimating data size needed for confidence, reliability of estimators, and outlier detection. All results are based on publicly available data sets and models from the PROMISE repository. As a secondary contribution, we investigate the efficacy of the popular model accuracy criterion PRED, which unlike MMRE has received relatively little investigation.

Keywords

Cost Estimation; Cost Model; Standard Error; Confidence; MMRE; PRED; Model Selection; Parameters; Calibration; Bootstrapping; Confidence Interval; Convergence; Conclusion Instability

1. INTRODUCTION

This is not a typical cost estimation research paper. You will not find a proposal for a new cost estimation model, or a new means for improving accuracy, or fancy model calibration techniques. This is a critical and comprehensive study on the *confidence* in cost estimation and cost estimation research results. We will generally rely on the statistical description of confidence and its intimate relation, *significance*, but sometimes we will discuss confidence informally. We begin the discussion with a few words on the problems relating to confidence. (Kitchenham et al. 2001; Foss et al. 2003; Myrtveit et. al. 2005) and others have commented on *the conclusion instability problem* (also called the *convergence problem*) associated software cost estimation. They have doubt on the practicality of common methods and measures used for comparatively assessing different estimation model generators processing numerous data sets. The results of such a comparison, they argue, vary according to the sub-sample of the data being processed and the applied evaluation criteria. Foss et al. comment that it

> . . . is futile to search for the Holy Grail: a single, simple-to-use, universal goodness-of-fit kind of metric, which can be applied with ease to compare (different methods). (Foss et al. 2003, p993)

Shepperd, a co-author of (Myrtveit et al. 2005; Kitchenham et al. 2001), comments that "one of the reasons for the equivocal results in the previous example is that there is variation in the choice of response variable(s) (i.e. accuracy indicators) when comparing prediction performance. They claim that these tend to capture different properties that can lead to "rank reversal" problems i.e. accuracy indicator A prefers (generator1) over (generator2), whilst indicator B prefers (generator2) over (generator1)" (Shepperd 2007). He comments that one possible cause of the conclusion is the choice *validation strategy*, which falls into four general classes, namely, model fitting, jackknife, hold-out studies, and k-fold cross validation. Shepperd believes that

> "… it is a matter of some urgency that we as a research community define and agree reporting protocols and methods for comparison. Moreover this is a more important task than performing yet more empirical studies given our present difficulties."

Further, in (Myrtweit et al. 2005) it is stated that

> "… we are still not in a strong position to advise practitioners as to what prediction models they should select because the studies do no converge … Indeed, very contradictory results have been reported in studies … convergence of studies on software prediction models is poorly understood … we need to understand why we have obtained so wildly opposing conclusions on this matter."

Understandably, conclusion instability statements such as the above do not inspire high confidence in cost estimation research results. This lack of confidence ultimately leads to a lack of confidence in software cost models in general. In this paper we address this issue of "confidence" in software cost model accuracy results – both in accuracy measures and the research results predicated on them. We show that a detailed analytical and empirical analysis of the *standard error* (SE) of the accuracy measures resolves conclusion instability problems such as described above. We demonstrate that conclusion instability is not the result of data quirks, poorly understood evaluation methods or accuracy criteria, or even the complexity of effort estimation. Rather, it arises in a very well understood way from the selection of the particular evaluation basis, and the *failure to account for the inherent error* present when estimating a distribution's population parameter from sample data. In our re-investigation of previous studies we sometimes find confidence for the original study conclusions, at other times we find the results inconclusive and suggest further investigation is needed. For example, we find support for the result of (Foss et al. 2003) that MMRE will mis-select an overestimating model over a "true" model. We cast doubt on the conclusiveness of (Kitchenham et al. 2001) that MMRE and PRED inconsistent model rank selectors. Ancillary to our investigations, we also find and demonstrate, contrary to the pessimism of (Foss et al., 2001) that the simple, widely-used yet poorly studied performance measure PRED, the percentage relative error deviation (Conte et al. 1986), is adequate for the task of comparing different model generators. However, as many other studies have concluded, another widely-used measure, mean magnitude relative error (MMRE), is not adequate and we provide tangible, easily understood reasons to caution against its use. Outside of MMRE and PRED there are many other accuracy measures that have been considered and suggested. However we do not study these at present as we find that PRED and MMRE are, by a wide margin, the most commonly used model accuracy measures. Furthermore, our studies indicate that there is no basis to believe that other measures will prove superior to PRED with respect to their properties as statistical estimators.

This paper is organized as follows: background and objectives of the investigation; a discussion of related work; then a description of models and datasets used; discussion of criterions (MMRE, PRED) and some of their characteristics as statistical estimators; discussion on prevalence of MMRE and PRED; use of standard error for these accuracy estimators; elaboration on the conclusion instability problem; our research methodology; five replicated and extended studies of MMRE and PRED and their confidence implications; and lastly a discussion of empirical and analytic characteristics of MMRE and PRED; a brief discussion on increasing confidence. We conclude by providing a summary overview our studies and implications on confidence and the use of MMRE and PRED in general areas of software cost estimation research.

2. BACKGROUND AND OBJECTIVES

Over the last 20+ years there have been a large number of cost estimation research efforts. Improving estimation accuracy and methods to support better estimates has been a major focus of researchers and is also of great interest to practitioners. Generally the development, validation, and comparisons of this research is empirically based - applying model evaluation criterion to cost and effort *datasets* that have either been collected for this reason, or are publicly available (PROMISE repository Boetticher et al. 2007). Model criterions are also referred to as "accuracy measures" or "accuracy indicators" and the two most prevalent are MMRE and PRED (Conte et al. 1986). MMRE has been empirically studied and criticized in a number of works such as (Myrtweit et al. 2005; Foss et al. 2003; Kitchenham et al. 2001). Additionally, a dizzying array of alternative criteria has been suggested. Surprisingly though, few works have directly studied PRED or even included it as an alternative for comparison when studying cost model accuracy criteria.

Therefore, one objective of this work is to remedy this by providing an empirical and analytical comparative study of MMRE and PRED. In this, we replicate several key empirical studies of MMRE and expand them to include PRED. Replicating and extending studies is an essential research endeavor if we are to have confidence in the results of empirical studies. In our replication efforts we have verified a majority of the results we studied, but we have also uncovered numerous errors in public data sets, errors in methods and contradictory results. In a few cases we have not yet been able to replicate a published result. Another objective of this work is to advocate the appropriate use of standard error (SE) when discussing model criterion based on sample project data. This addresses the serious "conclusion instability problem" as noted in the introduction that show contradictory results when using model accuracy criterion to compare models. For example, in an ongoing discussion in the literature about local calibration, (Menzies et al. 2005) references eight different studies supporting local calibration, whereas (Wieczorek and Ruhe 2002) can't find any significant difference in the performance between local and global estimation models. (Briand et al. 2000; Kitchenham et al. 2007) also state inconclusive results in this matter. Such inconstancies promote a palpable lack of confidence in software cost estimation research results, and consequently, a lack of confidence in software cost estimation accuracy criteria (Moløkken and Jørgensen 2003; Kitchenham et al. 2001). By accounting for SE, we can explain and resolve these inconstancies facilitating greater confidence in our results.

All of our studies are based on publicly available cost estimation datasets and methods, and (Mair et al. 2005) gives an overview of a few of these. The study of confidence has some useful side benefits. For example, in cost estimation modeling the question of "how much and what quality of data is needed to obtain significant results?" is often raised (Menzies et al. 2005). This appears to be a difficult question to answer, and while there are some well-known rules of thumb (e.g. Boehm's heuristic of 10 data points for local calibration), few works have attempted to address the question in a meaningful and quantitative way. We demonstrate that analyzing a criterion's SE as a function of the size of a dataset is a natural and meaningful way to address this question. This approach will also prove to have other practical applications such as outlier detection and accuracy measure reliability.

The statements and analysis in this work are based on publicly available datasets and simulation of cost estimation data as specified in (Foss et al. 2003, Myrtweit et al. 2005). We do this because they are well-known, straightforward, and most importantly publicly available so that our results can be readily verified, replicated, criticized, and hopefully expanded. Our approach, however, is easily seen to not be tied to any particular dataset or model and the conclusions are assuredly general.

While the current work aims to provide, in particular, a better understanding and confidence in the use of MMRE and PRED as estimates of model accuracy, we would like to make it clear that we make no claims that either criterion is a superior accuracy measure to other measures discussed here or not discussed. These two criterions are the focus of this study simply because they are the most widely used in the literature. If we were to take a position, it would be to advocate the use of the more standard model statistics of mean squared error (MSE) or a maximum likelihood estimator (Larsen and Marx 1986) over any of the criterion generally used in the literature. It is a mystery as to why these appear not to have been considered, however this is not the focus of the current work and these will not be discussed further here.

With the above objectives in mind, this works aims to contribute:

- A theoretically justified practice for analyzing confidence in the accuracy of cost estimation models (this is not accuracy improvement)
- Simple resolution of the conclusion instability problem in cost estimation model research results

- Fill a gap in the lack of a detailed study of the popular accuracy criterion PRED
- An understanding of the consequences of standard error resulting from model accuracy results based on sample data
- Methods for increasing confidence in cost model accuracy (e.g. how many data points, outlier removal, data stratification, etc.)
- Verification of several key empirical cost estimation accuracy criteria studies (and indicating errors or contradictory results)
- Extend several key empirical cost estimation accuracy criteria studies (e.g. use of PRED, increased data size, etc.)
- Improve understanding of the relation of data sets and models to accuracy criteria (via standard error analysis)
- A detailed analysis of MMRE and PRED as statistical estimators (e.g. reliability, consistency, bias, etc.)
- Practical tools and techniques for analyzing the standard error model accuracy criteria (e.g. bootstrapping)
- Clarification (or corrections) and expansion of previous statements on the statistical nature of MMRE and PRED

3. RELATED WORK

The available resources and literature on cost estimation research can be overwhelming. Other than the related works already mentioned above, we mention a few of the basic areas that frequently contain work related to the subject of confidence of accuracy measures and their application. There exist a relatively large number of empirically based estimation methods. Non-model-based methods (e.g. "expert judgment") usually do not play an important role in the empirical literature. Generally such methods do not output point estimation data applicable to accuracy criterion (there are some research effort that are the exception). Still, they are widely practiced intuitive methods used frequently in organizations where a model-based approach would be too cumbersome or sufficient model-calibration data is unavailable. Model-based methods can be split into generic-model-based (e.g. COCOMO, SLIM, etc.) and domain specific-model-generation methods such as CART or Stepwise ANOVA. Jorgensen reviews 15 studies that compare model-based to expert-based estimation (Jorgensen 2004).

Besides the variety of cost estimation methods, there are a large diversity of studies on the topic - some on evaluation of cost estimation in different contexts, some assessing current practices in the software industry, others focusing on calibration of cost estimation models. See the *Encyclopedia of Software Engineering* (Briand and Wieczorek 2001) for an overview of cost estimation techniques as well as cost estimation studies. Also (Jørgensen and Shepperd 2007; BESTweb 2007) list current studies on software cost estimation.

As mentioned in the introduction, unlike with MMRE, no detailed study could be found on the nature and efficacy of PRED as a software cost estimation model criterion. In spite of this, PRED is a frequently used criterion as is evidenced by summations of model performances in (Menzies et al. 2006). There is a considerable amount of work that studies MMRE and other accuracy criterions. For example, (Foss et al. 2003) studies MMRE performance in comparison to numerous other criterions in selecting the best model (or as they call it, the "true" model) based on simulated data. In (Kitchenham et al. 2001) the authors investigate some the distributional qualities of MMRE and PRED and point out that they may be measures for two different things. The work in (Lum et al. 2006) suggests that the variation in MMRE under holdout studies and the variation in regression coefficients are important factors in cost model accuracy. Similarly there are a number of studies (Foss et al. 2001; Miyazaki et al. 1994) of the components of accuracy criterion, such as MRE.

4. COCOMO, DATASETS USED

We will make frequent use of COCOMO, the Constructive Cost Model (Boehm 1981) since, unlike other models such as PriceS or SLIM or and SEERSEM, it is a simple open model with substantial published project data sets. All details for COCOMO are published in the text "Software Engineering Economics" (Boehm 1981). There are several versions of the COCOMO model, the two most prevalent being COCOMO I and COCOMO II. The one we use here (COCOMO I) was chosen based on the publicly available COCOMO data within the PROMISE repository (Boetticher et al. 2007). Here we study variations of the classic COCOMO I model to exemplify our points and methods, and enable straightforward duplication and verification of our results and claims. However, our methods are not limited to such models, and it will be evident that COCOMO I is fully exchangeable with other, and perhaps better cost estimation models. The intent here is to define a set of experiments and examples that others may replicate in order to refute or improve on our results and methods. The particular datasets and cost models used here are simply a convenience.

Boehm's Post-Architecture version of COCOMO I:

$$effort = a * \left(\prod_{j=1}^{15} EM_j^{a_j} \right) * (size)^b * \omega \qquad (1)$$

Here, EM_j are "effort multiplier" parameters whose values are chosen based on a project's characteristics, and a_j, a, b are domain specific "calibration" coefficients, either given as specified by Boehm in (Boehm 1981) or determined statistically (generally via ordinary least squares regression) using historical project data. The dependent variable *size* is expressed either as KSLOC (thousand source lines of code) or in FP (function points) is estimated directly or computed from a function point analysis. The model error ω is a random variable with distribution D (not generally Normal). Model accuracy measures such as MMRE and PRED are estimating one or more parameters of D.

Table 1 shows the six COCOMO I model variations used in this work along with brief descriptions.

Table 1. COCOMO I model variations used in study

	Model	**a,b**	***EM$_j$***	***a$_j$***
(A)	ln_LSR_CAT	CLSR	categorical	CLSR
(B)	aSb	Given	none	None
(C)	given_EM	Given	given	None
(D)	ln_LSR_aSb	OLS	none	None
(E)	ln_LSR_EM	OLS	given	OLS
(F)	LSR_a+bS	OLS	none	None

The table entries are interpreted in the following way:

OLS: Ordinary Least Squares regression was used with the given project data set to determine the parameter values.

CLSR: Categorical Least Squares regression was used with the given project data set to determine the parameter values.

Given: The values of these parameters are given in (Menzies et al. 2006) and not derived statistically from the data set.

Categorical: The values of these parameters are considered non-numerical, non-ordinal categories (e.g. the implied order of "VL" "L" "N" "H" "XH" values for effort multipliers are ignored).

None: The parameters are not used in this model.

Models (A) - (E) use the functional form as the general COCOMO I model given equation (1) above, however model (F) uses a simple linear ***a+b*(size)*** form. When there is a "ln_" in the model name, the applicable project data was transformed by taking the natural logarithm (ln) for the analysis. All values were back transformed when used in the model and model calculations (e.g. calculating MMRE's).

An example reading of the table for model (B) states that it is the general COCOMO I model without using any effort multipliers (and hence no calibration coefficients) and the values of the parameters *a, b* are taken from the values given in (Menzies et al. 2006). No regression is performed on the data set.

The historical data used for estimating the coefficients are taken from the COCOMO81 (COCOMO81 dataset 2007), COCOMONASA (COCOMONASA 2008), NASA93 (NASA93 2007) and Desharnais (Desharnais 2007) PROMISE repository data sets. We note that in the course of this work we contributed numerous corrections and clarifications for these data sets. Brief overviews of these data sets are:

COCOMO81: This dataset was first presented 1981. COCOMO81 was the foundation for Boehm's first COCOMO model (Boehm 1981). It holds 63 actual effort and actual size (in SLOC) data points each with fifteen numerical effort multipliers that further characterize each project. The data has been collected from software projects from 1964 to 1979 and comes from a variety of domains including engineering, science, financial, etc.

COCOMONASA: 60 NASA projects from the 1980s and 1990s. Since this data comes from NASA, it is stratified to aerospace applications. Each project data point includes actual effort and actual size (in SLOC) with 15 ordinal effort multipliers.

NASA93: Data from different centers for 93 NASA projects from 1980s and 1990s were collected by Jairus Hihn, JPL NASA Manager SQIP Measurement & Benchmarking. Projects contain both aerospace and IT applications. Besides lines of code and actual effort in person months, this dataset gives 15 standard COCOMO I discrete attributes and seven others describing each project.

Desharnais: Data for 81 software projects taken from (Desharnais 1989). Each project data point includes actual effort and actual size (in function points) with 9 additional project descriptors, some categorical, some ordinal, some numerical.

The simulated data used in our studies is based on (Foss et al. 2003) and (Myrtweit et al. 2005) and we direct the reader to these sources for details on their applicability and construction.

5. ACCURACY CRITERION AS STATISTICAL ESTIMATORS

The field of cost estimation research suffers a lack of clarity about the interpretation of model evaluation criterion. In particular, for model accuracy, various indicators of accuracy – both relative and absolute – have been introduced throughout the cost estimation literature. For example, mean squared error (MSE), absolute residuals (AR) or balanced residual error (BRE). Our literature review in the subsequent section indicates that the most commonly used, by far, are the "mean magnitude relative error" or MMRE, and "percentage relative error deviation within *x*" or PRED(*x*). Of these two, the MMRE is the most widely used. Both are based on the same basic unit-less value of magnitude relative error (MRE) which is defined as

$$MRE_i = \frac{|y_i - \hat{y}_i|}{y_i}, \qquad (2)$$

where y_i is the actual effort and $\hat{y}_i$ is the estimated effort for project *i*. It is argued that MRE is useful because it does not over penalize large projects and it is unit-less (i.e. scale independent). MMRE is defined as the sample average of the MRE's:

$$MMRE = \frac{1}{N}\sum_{i=1}^{N} MRE_i \qquad (3)$$

To be more precise we should label (3) as $MMRE_N$ to indicate that it is a sample statistic on N data points. To be consistent with customary usage we will drop the subscript when there is no confusion as to the number of data points used. Conte et al. (Conte et al. 1986) consider MMRE ≤ .25 as an acceptable level of performance for effort prediction models.

PRED(*x*) (Jørgensen 1995) defines the average fraction of the MRE's off by no more than *x* as defined by

$$PRED(x) = \frac{1}{N}\sum_{I=1}^{N}\begin{cases}1 & \text{if } MRE_i \le x\\ 0 & \text{otherwise}\end{cases} \qquad (4)$$

Typically PRED(.25) is used, but some studies also look at PRED(.3) with little difference in results. Generally PRED(.3) ≥ .75 is considered an acceptable model accuracy. There is some concern about what constitutes an appropriate value of *x* for PRED(*x*). Clearly the larger *x* is, the less information and confidence we have in an accuracy estimate. However interesting this question, we are interested in comparisons of PRED with MMRE and not the specific application of these measures. Note that inverse to MMRE, high PRED values are desirable. This is easily reversed to match MMRE by simply switching the 0-1 values in (4) if desired. This inverse relationship should be kept in mind when viewing our side by side comparisons. Another concern is that it is difficult to get an intuitive feel for comparing relative PRED values. For example, a generally accepted heuristic from (Conte et al. 1986) states that a model has acceptable accuracy if PRED(.25) ≥ .75, but how much "bad" relatively speaking is a model whose PRED(.25) = .65 as is the value for NASA93 with model C? What if PRED(.3) ≥ .75 and PRED(.25) < .75 as is the case with NASA93 with model C? Additionally, with PRED information is "lost" as to what degree estimates are off. A model with a PRED(.25) = .65 with its 35% of "outside the error range" estimates just slightly more then 25% off from the actuals (say 26%) is considered to have the same accuracy as a model with estimates perhaps ludicrously farther off (say 1000%). The same holds for the "inside" points, and this is troublesome for some research efforts.

Although MMRE and PRED are still today the de facto standards for cost model accuracy measurement (Myrtweit et al, 2005), they don't specifically measure accuracy. In fact, technically they are "estimators" of a function of the parameters related to the distribution of the MRE values. This in turn is presumably related to the error distribution of the model for the population data. As such, we will frequently refer to these as "accuracy indicators" rather than measures when it is more appropriate.

In practice it is difficult to obtain parametric knowledge of an error distribution, so we must resort to the use of estimators. In this regard, Wikipedia provides a concise description of estimator and the usual procedure in their use:

In statistics, an estimator is a function of the observable sample data that is used to estimate an unknown population parameter (which is called the estimand); an estimate is the result from the actual application of the function to a particular sample of data. Many different estimators are possible for any given parameter. Some criterion is used to choose between the estimators, although it is often the case that a criterion cannot be used to clearly pick one estimator over another. To estimate a parameter of interest (e.g., a population mean, a binomial proportion, a difference between two population means, or a ratio of two population standard deviations), the usual procedure is as follows:

1. *Select a random sample from the population of interest.*
2. *Calculate the point estimate of the parameter.*
3. *Calculate a measure of its variability, often a confidence interval.*
4. *Associate with this estimate a measure of variability.*

The standard error (SE) of an estimator is defined as the standard deviation (square-root of the variance) of the estimator's sample statistic over the distribution it estimates - which for MMRE and PRED would be the MRE distribution, not the project data distribution. The SE is the fundamental variability measure for which step 3 is based. In cost estimation research it has been commonplace to skip steps 3 and 4. This work aims to illustrate why this practice is ill-advised.

SE is the key to understanding confidence and so this will be our primary interest in this work. However, it is also important to have an idea as to what the estimators MMRE and PRED are estimating. Several studies have noted that MRE distributions are essentially related to the simpler distribution of the values:

$$z_i = \frac{\hat{y}_i}{y_i}, \qquad (5)$$

Which are clearly related to the distribution of the error residuals $\varepsilon_i = y_i - \hat{y}_i$ but in a non-trivial way (Kitchenham et al. 2001).

(Kitchenham et al., 2001) report that MMRE and PRED are directly related to measures of the *spread* and the *kurtosis* of the distribution of z_i values,.It is ,Both these parameters of the z distribution are legitimate and useful as model accuracy indicators so long as they are appropriately applied and interpreted. Because both MMRE and PRED are averages of a function of the MRE's, a useful fact that follows easily from the weak law of large numbers (Larsen and Marx 1986) is that both MMRE and PRED are *consistent estimators* (i.e. they converge in probability to some parameter of the distribution) (Larsen and Marx 1986). This provides a meaningful and precise interpretation of MMRE and PRED as "accuracy measure" when they are viewed as estimators for parameters of the error distribution *z* (presumably the spread and kurtosis as indicated earlier) for a given cost model and dataset. This however, does not say anything about how *good* they are as estimators (e.g. bias, uniform convergence, rate of convergence, MSE, SE, etc.). Indeed, there is substantial research that address bias and the so-called "reliability" of these estimators such as (Foss et al. 2003; Shepperd and Kadoda 2001), although not expressed or analyzed in the more standard statistical framework we take advantage of here. Numerous works such as (Kitchenham et al., 2001) discuss the bias of MMRE and PRED as model selection criterion (e.g. MMRE tends to favor under-estimating models). However bias is a characteristic of the estimator itself, not the sample data. It is certainly plausible that an estimator with high bias will also have high accuracy, and conversely, a low accuracy estimator may also have low bias. Standard error is a function of both the estimator and the particular sample data and so it has a more direct impact on accuracy. In particular, the SE provides insight into how well an MMRE or PRED estimate generalizes from the sample to the population, and therefore any result that is based on these estimates must account for SE if the analyst is to have confidence in the result. An understanding of SE forms the basis for understanding other estimator behaviors such as reliability and robustness. To our knowledge, this work is the first to comprehensively study the SE of MMRE and PRED and its applications to confidence. SE.

6. PREVALENCE OF MMRE AND PRED

The analysis of the rest this paper focuses on only the two model accuracy criterions MMRE and PRED. Before describing that analysis, we first digress to demonstrate that it is worthwhile discussing just these measures.

There is much evidence for the widespread use of PRED and MMRE in the literature. For example in (Foss et al. 2003) it is stated "… MMRE is still considered the de factor standard." Further evidence of this is shown in Table 2 which shows a survey of model accuracy criteria used in research results from 1986 through 2007. To be included, the work had to make use of a criterion to reach a conclusion, not simply cite a reference to a published result. Note that our survey was not exhaustive as we did not traverse all research outlets, but we believe it is representative and unbiased as we did not "pick and choose" the works cited. Rather, the survey is a comprehensive look at a few of the outlets that publish the majority of cost estimation studies as indicated in (Jørgensen and Shepperd, 2007).

Table 2. Survey of Model Accuracy Criteria Used in Previous Studies

Accuracy Indicator	Recommended by	Examples of studies using the accuracy indicator
$MMRE = \frac{1}{N}\sum_{i=1}^{N} MRE_i$ $MdMRE = \underset{i=1\ldots N}{median}(MRE_i)$	(Conte et al. 1986)	(Kemerer 1987; Walkerden and Jeffery 1999; Shepperd and Schofield 1997; Gray and MacDonell 1999; Boehm et al. 2000; Briand et al. 2000; Chulani et al. 1999; Menzies et al. 2006; Wittig and Finnie 1997; Srinivasan and Fisher 1995; Myrtveit and Stensrud 1999; Jørgensen, 1995; Kitchenham et al. 2002)
$PRED(x) = \frac{1}{N}\sum_{i=1}^{N}\begin{cases}1 \text{ if } \mathrm{MRE}_i \le \mathrm{x}\\ 0 \text{ otherwise}\end{cases}$	(Conte et al. 1986)	(Walkerden and Jeffery 1999; Shepperd and Schofield 1997; Gray and MacDonell 1999; Boehm et al. 2000; Briand et al. 2000; Chulani et al. 1999; Menzies et al. 2006; Wittig and Finnie 1997; Srinivasan and Fisher 1995; Jørgensen 1995; Kitchenham et al. 2002)
Mean of the balanced relative error (MBRE) $BRE_i = \frac{y_i - \hat{y}_i}{\min(y_i, \hat{y}_i)}$ $MBRE = \frac{1}{N}\sum_{i=1}^{N} BRE_i$	(Miyazaki et al. 1991)	(Miyazaki et al. 1994; Foss et al. 2003; Myrtweit et al. 2005; Gray and MacDonell 1999)
$AR(x) = \frac{1}{N}\sum_{i=1}^{N}\begin{cases}1 \text{ if } \mathrm{MRE}_i \le \mathrm{x}\\ 0 \text{ otherwise}\end{cases}$	(Miyazaki et al. 1991)	(Miyazaki et al. 1994; Gray and MacDonell 1999)

Mean of the magnitude of error relative to the estimate (MMER) and mean *z* $MMER = \frac{1}{N}\sum_{i=1}^{N}\frac{\lvert y_i - \hat{y}_i \rvert}{\hat{y}_i}$ $Mean\ z = \frac{1}{N}\sum_{i=1}^{N}\frac{\hat{y}_i}{y_i}$	(Kitchenham et al. 2001)	(Myrtweit et al. 2005; Lokan 2005; Foss et al. 2003)
Standard deviation (SD) $SD = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N-1}}$	(Foss et al. 03)	(Foss et al. 2003; Myrtweit et al. 2005; Lefley and Shepperd 2003)

One limitation of the above study is that the majority of dates of the publications range from 1986 to 2006. Perhaps, we conjectured, MMRE and PRED are antiquated measures and more recent publications use better approaches. To check this, we did a study where we listed senior venues (i.e. we now include conferences and workshops) in software engineering and examined publications at those venues, looking for what performance measures they applied to their work. We limited that study to the year 2007 (this year was chosen to be after the time period examined in the above study and not so recent so that very recent papers have not yet been included in indices). The forums we examined where

- The 2007 USC COCOMO forum
- ESEM'07: International Symposium on Empirical Software Engineering and Measurement
- ICSE'07 : International Conference on Software Engineering
- ASE '07: IEEE Automated Software Engineering
- 2007 IEEE Transactions on Software Engineering

The above list was not intended to be complete. We only sought to check if PRED and MMRE were still widely-used. In all we found 10 papers that presented performance results from automatically generated effort estimation models:

- One paper was a review of recent work (Kitchenham et al. 2007) that reviewed previous results in the period 1999 to 2005; i.e. a similar range to those papers discussed in Table 2. After ignoring the paper mentioned in both (Kitchenham et al. 2007) and Table 1, we found 9 studies in (Kitchenham et al. 2007), of which MMRE and PRED were used in 7 and 6 papers respectively.
- Of the remaining 2007 publications, apart from (Kitchenham et al. 2007), we found 1 with purely visual results (and no summary statistics); 1 that reported its results using correlation of expected to actual. The remaining 7 papers all used PRED, 6 of which also used MMRE.

In summary, since 1986 and persisting up until 2007, PRED and MMRE are widely used performance measures for evaluating automatically generated effort estimation models. The 2005 remarks of Foss et al. (mentioned in the introduction) suggest that the widespread use of these criteria is somewhat misguided. In the following, we will agree with Foss et al. in regards to MMRE, but will endorse the appropriate use of PRED.

7. STANDARD ERROR OF ACCURACY ESTIMATORS AND CONFIDENCE

One of Boehm's original motivations for creating COCOMO was to increase the confidence managers have when estimating software projects. Curiously, despite this original motivation for COCOMO, very little has been reported on the confidence in accuracy measures for COCOMO estimates. This has led to a surprising number of contradictory results in the theory and practice of cost estimation.

The primary concern is that COCOMO models (and more generally all cost estimation models) are "calibrated" with a relatively small amount of data (which is frequently biased or "sanitized"). Various measures such as PRED(.25)=.5 and MMRE=.35 are plainly presented stating just how "good" an analyst should feel about the model's accuracy and predictive capabilities. The reality of this is that these values only reflect the model accuracy for the data they were calibrated on. There is a serious question in our "confidence" in these measures for predicted values. Providing a standard error for these measures and a clearer understanding of what this implies (e.g. How much error is bad?) is key to addressing the confidence question. For example, if we understand the standard error from calibration data (i.e. the sample population), we can generate appropriate *confidence intervals* of these measures for the "true" population of values being predicted. For example, a COCOMO calibration that has a PRED(.25)=.5 for the sample data, one might state that with 95% confidence that the value of the unknown parameter for the population error distribution that is being estimated lies within the interval .38<PRED(.25)<.83.

It is critical that we distinguish the standard error in a model accuracy criterion from the standard error of a cost estimation model. The latter, while important, is not the subject of this work and they should not be confused for one another. Standard error of a model (such as the R^2 value) explains the variance in the degree a model accounts for the data is was calibrated on. While standard error of an accuracy criterion expresses the variance the criterion has as an estimate of the "population" or "true" accuracy parameter of an error (or accuracy) probability distribution based on a finite sample of the population data. One way to interpret this SE is as the "accuracy of the accuracy." In this sense we have a precise interpretation of model accuracy confidence that is distinct from model accuracy. For example we can have high confidence that a model has low accuracy, or have low confidence in a model with high accuracy, and so forth. Another distinguishing characteristic is that we expect that the standard error for an accuracy criterion will always decrease with increased data size, whereas we have no such expectation for the standard error of a model. Distinct too are the standard errors associated with model parameters (e.g. regression coefficients). While also very important and useful (e.g. calculating prediction intervals), they are not the focus of the current work.

Generally the standard error of an estimator is difficult to compute analytically. However, bootstrapping (Mooney and Duval 1993; Efron 1979) is a well-known, well-accepted, and straightforward approach to approximating the standard error of an estimator. Briefly, bootstrapping is a "computer intensive" technique similar to Monte-Carlo simulation that re-samples the data with replacement to "reconstruct" the general population distribution. The bootstrap is distinguished from the *jackknife*, used to detect outliers, and *cross-validation* or "holdouts" used to ensure that results have predictive validity.

Table 3. Overview of datasets with model (C)

Data Set	Size	MMRE	SE	PRED(.25)	SE	95% Confidence
NASA93	93	.6	.14	.48	.05	.37 ≤ MMRE ≤ .94 .38 ≤ PRED ≤ .6
COCOMO81	63	.37	.04	.37	.06	.31 ≤ MMRE ≤ .45 .25 ≤ PRED ≤ .49
COCOMO NASA	60	.25	.03	.65	.06	.2 ≤ MMRE ≤ .32 .55 ≤ PRED ≤ .78

We use bootstrapping in various capacities to understand the standard error of PRED and MMRE for the COCOMO model using the three PROMISE data sets indicated previously. Are we making a fuss over nothing here? As a preliminary investigation Table 3 shows manifestly that the standard error for MMRE and PRED(.25) for various COCOMO models are significant, and clearly worthy of more detailed study.

8. CONCLUSION INSTABLITY

(Kitchenham et al. 2001; Foss et al. 2003; Myrtveit et al. 2005) caution that, historically, ranking estimation methods has been done quite poorly and unreliably. Based on an analysis of two data sets, as well as simulations over artificially generated data sets, (Foss et al. 2003) and (Myrtveit et al. 2005) concluded that numerous commonly used criterion such as the MMRE are unreliable measures of estimation effectiveness. Also, the conclusions reached from these standard measures can vary wildly depending on which subset of the data is being used for testing. Figure 1 demonstrates conclusion instability for results based on MMRE. It shows two experimental runs. In each run, 30 times, effort estimate models were built from 19 effort estimation data sets ranging in size from 20 to 83 examples (this data came from subsets of the *NASA93* and *COCOMO81* data sets, discussed above). Models were built using two methods: method1 and method2. Each time, an effort model was built from a randomly selected subset of 90% of the data. Results are expressed in terms of the difference in MMRE between the two subsets; e.g. in Run #1, method1 had a much larger MMRE than method2.

After Run #1, the results endorse method2 since that method either (a) did better (i.e. lower MMRE) as method1 or (b) had similar performance to method1 (i.e. near 0). However, this conclusion is *not stable*. We have ordered the data sets so the Run #1 graph in Figure 1 goes from minimum to maximum difference to help better visualize the instability. Observe in Run #2 that:

- The improvements of method2 over method1 disappeared in subsets 1,2,3,7, and 11.
- Worse, in subsets 1, 2, and 11 method1 performed dramatically better than method2.

If the results were stable, we would expect similar points of differences, perhaps in different locations due to the random selection of data subsets. For example, in Run #1 we see a maximum difference of more than -100 while in Run #2 there is no corresponding large difference. Some of this can be explained as normal deviation resulting from random selections of the subset data. But we note that the deviations seen in 30 repeats of the above procedure were quite large: within each data set, the standard deviation on the MRE's were {median, max} = {150%, 649%}

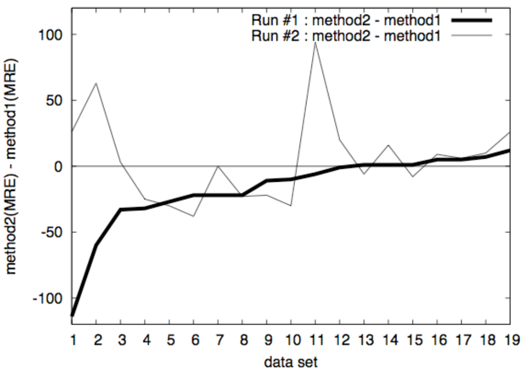


Figure 1. Results of 2 different runs of COSEEKMO (Menzies et al. 2005) comparing two effort estimation generation methods using mean MRE values. Points on the Y-axis show the difference in mean relative error (MRE) between method1 and method2. Lower values endorse method2 since, when such values occur, method2 has a lower error than method1. Data sets sorted in terms of the results of Run #1.

One troubling result from the Figure 1 study is that the number of training examples was not connected to the size of standard deviation. A pre-experimental intuition was that the smaller the training set, the worse the prediction instability. On the contrary, we found small and large instability (i.e. MRE standard deviations) for both small and large data sets (Menzies et al. 2006). That is, instability cannot be tamed by further data collection or by a more reliable estimator than MMRE. Rather, the data must be processed and analyzed in some better fashion (e.g. the methods discussed below).

These large instabilities explain the contradictory results already observed in the effort estimation literature. Jorgensen reviews fifteen studies that compare model-based to expert-based estimation. Five of those studies found in favor of expert-based methods; five found no difference; and five found in favor of model-based estimation (Jorgensen 2004). Such diverse conclusions are to be expected if model evaluation results routinely exhibit large instabilities in their performance. This puts into question the confidence we can have in the conclusions when there is instability leading to the new possibility of an "inconclusive" result. We will show that instability often emerges from the inappropriate use of estimators for sample data to represent whole population data parameters without considering their standard error. As seen in our example above, this is particularly severe in hold-out studies where the hold-out sets are relatively small. In many cases the hold-out sets must be small as the sample data set is relatively small itself and increasing the hold-out set size would leave an insufficient amount of date for model training. In this case, hold-out studies are inappropriate for obtaining a conclusive result and other methods must be explored. We will investigate the stability of hold-out studies in greater detail in a subsequent section.

9. RESEARCH METHOD

As mentioned previously, we use four different PROMISE datasets to give a good overview of the effects of SE on research results based on accuracy criterion. We calculated MMRE, PRED(.25) and PRED(.30) for the models (A)-(E) on the COCOMO81, COCOMONASA and NASA93 datasets. The same accuracy indicators were calculated using the models (E) and (F) for the Desharnais dataset using adjusted and also raw function points as in (Kitchenham et al. 2001). The reason there are fewer models for this dataset is that the Desharnais data does not provide COCOMO I effort multipliers. Where appropriate, we also consider simulated data sets that will be described subsequently.

Many of the studies we replicate and extend here rely on comparing MMRE or PRED values calculated on the data sets above. Our general research approach is to re-consider the confidence we have in such comparisons when accounting for SE. This provides a third possible outcome to a comparison –"inconclusive," in addition to "equal" and "not-equal."

We aim to obtain the standard error for MMRE and PRED for the various models (A) - (F) and four PROMISE data sets. The parameters of the *z*-distribution (related to the error distribution, see Section 5) are unknown, and we cannot assume this to be normally distributed. However the SE, which is the variance of estimators MMRE and PRED as estimators of this distribution, are difficult to obtain analytically. A well-established method for obtaining approximations for the standard error of estimators is to use *bootstrapping* (Mooney and Duval 1993; Efron 1979).

Standard confidence intervals are also difficult to compute analytically, so here too we resort to bootstrapping. A notable concern in bootstrapping confidence intervals is the effect of non-normally distributed data. In particular, confidence intervals for highly skewed distributions are poorly approximated with the basic bootstrapping method. We discuss the distributional characteristics in a later section, we found that the BC-percentile, or "bias corrected" method has been shown effective in approximating confidence intervals for the type of distributions we are concerned with. For most calculations we chose 15,000 bootstrap iterations (well beyond the suggested number) using the Excel Add-In poptools (Hood 2008) or in some cases, the R statistics environment (R 2008). Our bootstrapping results have also been replicated using other bootstrapping Excel Add-In's, manual calculations, and some custom developed bootstrapping software that performs bootstrap iterations to a desired precision rather than using the arbitrarily chosen 15,000 iterations to ensure the integrity of our results. All results were seen to be consistent, some of which we now describe. An example bootstrap run is provided in the appendix.

In practice there are two simple naïve approaches to testing significance of a comparison:

1) Do the values SE intervals (that is, the value plus or minus its SE) overlap?
2) Do the values 95% confidence intervals overlap (the standard of a 95% confidence level is considered reasonable for a conclusive result)

It has been shown in (Payton et al. 2000) that (1) tends to be too forgiving by indicating false significance above 5%, while (2) tends to be overly conservative falsely indicating inconclusive outcomes more than 5% of the time. Unfortunately, due to the atypical nature of MMRE and PRED as statistics, the non-normal distribution of MRE (we will elaborate on this later), and the particular experimental studies considered here, standard hypothesis tests statistical comparison such as t-tests, and Wilcox generally are not easily applicable and cannot be relied on to test for significance (e.g. difficulty in calculating SE, non-normality, unequal variances, different populations, etc.). This poses a problem in trying to obtain more confident and consistent results. One option is to use bootstrapping to estimate SE and apply various transformations to massage the data into application to a standard hypothesis test. Not only does this approach get computationally intense, it is also is difficult to grasp the precision one can expect in the results due to compounding inaccuracies. Fortunately in our studies a high degree of precision in the interpretation of significance is not required. In (Payton et al. 2003) it is suggested that considering the overlap of 83% to 84% confidence intervals will provide an approximately 5% significance level when the SE's are close in value (typical in our comparison efforts). When making comparisons at the 95% confidence level, this is the method we will use.

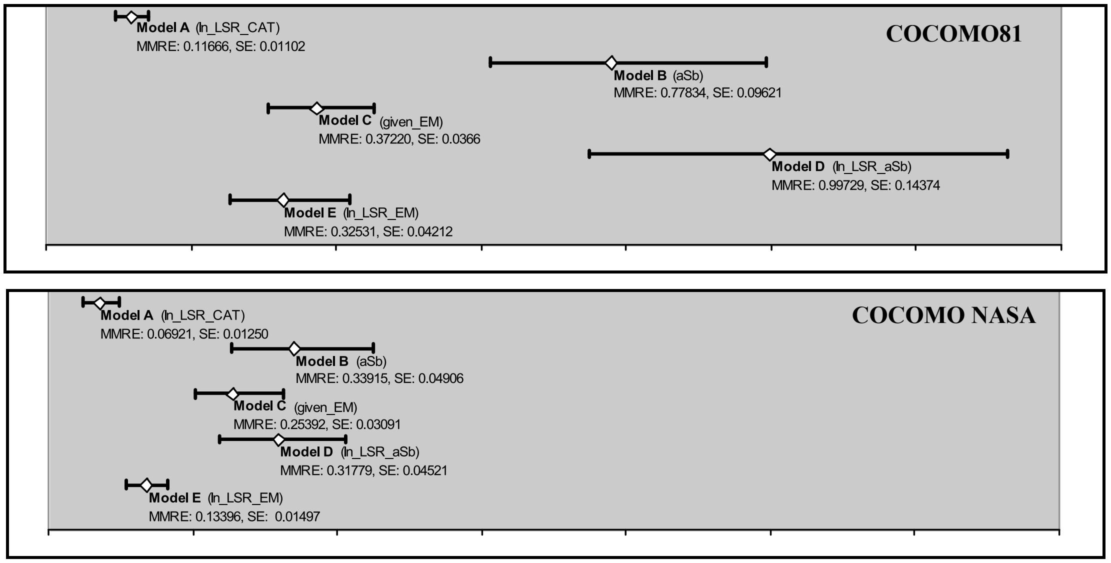
10. STUDIES OF CONFIDENCE IN RESEARCH RESULTS

In this section we provide empirical studies that replicate and extend existing studies. We then analyze the results of each study accounting for standard error. We have performed a dozen such studies and these are just a few representative examples.

Study 1: confidence in single criterion model ranking

We have already introduced the popular cost estimation research area of *model selection*. Model selection commonly involves advocating methods and criteria for choosing a particular estimation model format, calibration method, or use of calibration data (e.g. "pruning", stratification, etc.) or a combination thereof (Jørgensen and Shepperd 2007). Model selection research results often appear to be contradictory across different data sets. Validation methods such as "holdout" experiments, while intuitively may seem reasonable, are difficult to justify formally and are inherently unstable when applied to small data sets. Many model selection research results compare COCOMO models and calibration approaches to (presumably better) alternatives. To illustrate how standard error can be used to obtain more confident results, we choose to study a number of variations of COCOMO I itself. This study replicates results from several other studies (at least in part), or are analogous enough to indicate how the approach could be used with alternative models and calibration methods, including analogy models (Shepperd and Schofield 1997), COSEEKMO (Menzies et al. 2006), and simulated project data approaches (Foss et al. 2003).

Figure 2 visualizes the performance of MMRE with respect to the PROMISE datasets. Each graph on a diagram shows the estimator location (MMRE) within its 95%-confidence interval and is labeled with the model name and MMRE value. Remarkably, the standard error for the MMRE's for the same model vary greatly over different datasets. This perhaps in part explains some of the inconstant results in the literature when different data sets were used.



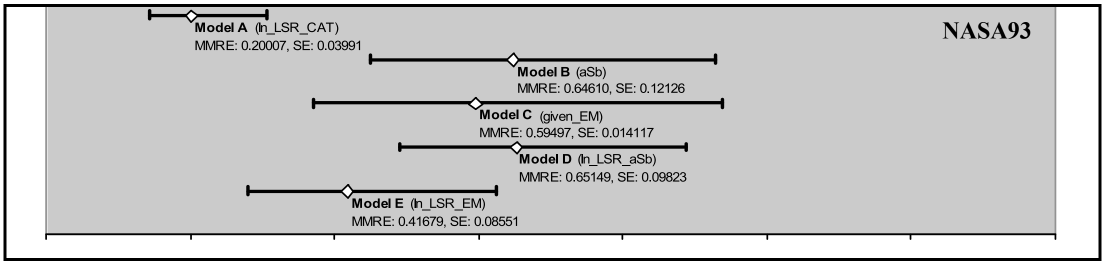


Figure 2. MMRE bootstrapped 95% confidence intervals

While the precise probabilities for the likelihood of one value being greater than another were not calculated, it is clear enough to see that the more overlap two confidence intervals have, the less one is able to say "in confidence" that one value is greater (or less than) another. This informal statement can be made more precise by considering Vysochanskiï-Petunin inequality (Wikipedia 2008) which is a refinement of Chebyshëv's inequality that places an upper bound on the probability of how far a random variable can be from its mean with respect to its variance. Such tools are essential for reliably understanding dispersion measures such as the coefficient of variation (such as used in (Menzies et al. 2006) and presumably MMRE and PRED. For our purposes here, the visual amount of overlap of the confidence intervals will suffice to illustrate the effects of standard error in our confidence of MMRE and PRED comparisons.

Performance ranking (where lower MMRE's have higher rank) is not consistent over all four models as indicated in Table 4 (i.e. ranks 4 and 5 are switched for COCOMONASA). Therefore we have "rank reversal" conclusion instability. Furthermore, the standard error intervals vary greatly in size: on the COCOMO NASA dataset the biggest interval ranges 19 points whereas the 95%-confidence interval for model (D) on the COCOMO81 dataset has a range of more than 66 points. This indicates that MMRE is sensitive to both the data and the particular model used. That is, some models may provide more confident accuracy results for a given data set than others. As a result, a more confident performance ranking accounting for this might be something like that listed in Table 5. Model ranking based on MMRE at 95% confidence where two models will be given the same rank if their 84% confidence intervals overlap (recall that in the introduction it was discussed why overlapping 95% confidence intervals do not provide 95% confidence). That is, the models cannot be distinguished from one another in terms of the model accuracy criterion performance.

Table 4. Model ranking based on MMRE

	COCOMO81	COCOMONASA	NASA93
1.	A	A	A
2.	E	E	E
3.	C	C	C
4.	B	D	B
5.	D	B	D

Table 5. Model ranking based on MMRE at 95% confidence

	COCOMO81	COCOMONASA	NASA93
1.	A	A	A
2.	C, E	E	B, C, D, E
3.	B, D	B, C, D	-

Looking at Figure 2 and Table 5. Model ranking based on MMRE at 95% confidence one might get excited about model (A) as it appears to perform consistently better than other models and with high confidence (i.e. even the 95% confidence intervals do not overlap). However this illustrates a fallacy of using purely statistical results as a basis for model selection. By allowing our model parameters to vary unconstrained, we are indeed able to calibrate a model that fits the data very well. However, a quick look at the parameter coefficients generated for this model reveal a number of absurd effort parameter relationships. For example model (A) applied to NASA93 CLSR estimated the parameter values for the "required complexity" CPLX effort multiplier (see Menzies et al. (2006) for details for this) to be L=-0.483, N=0.989, H=0.677, and VH=-0.745. Generally it is believed that higher required complexity requires higher effort. This obviously runs contrary to such beliefs (which other studies have empirically validated). Data enthusiasts might counter this objection with "perhaps this actually describes the true nature of required complexity" in that it is not ordinal. But then looking at the values estimated for model (A) applied to COCOMONASA where L=-0.66, N=0.659, H=0.653, and VH=7.24 contradicts this. The enthusiast may counter again by stating that perhaps it is the true nature of required complexity varies with respect to the kind of projects each data set represents.

Fair enough, but closer inspection of NASA93 and COCOMO NASA would reveal that the kinds of projects in both are very similar. In fact, a large number of the projects in NASA93 are from the COCOMONASA dataset. Surely similar data sets should have similarly behaving required complexity within the same category value. The reasonable conclusion here is that model (A) is not a realistic effort model for the data sets despite its statistical performance. There is no confidence that the model is accurate for predictions (i.e. data outside the calibration set).

So what can be concluded with confidence based on the MMRE results? The intervals for COCOMO81 in Figure 2 indicate that model E is significantly better than model D. Hence for COCOMO81 data, we can be confident that adding effort multipliers improves MMRE. The same result holds for the other data sets except NASA93. This provides reasonable confidence that in general adding effort multipliers will improve indeed MMRE.

Figure 3 compares PRED results for the same models and datasets as Figure 2. Unlike MMRE, the PRED rankings are consistent over *all datasets* (i.e. A > E > C > B > D where ">" means "higher PRED rank than"). Also, note that the confidence intervals vary less in size. This, from a perspective that is easily accessed and justified, supports a variety of assertions made in the literature that claim PRED is more consistent and "robust" than MMRE. In fact, as we will illustrate in a later section, that unlike MMRE, PRED is not dependent on the variance of the MRE's. Thus PRED is immune to large variances from outliers in the data (i.e. it is more robust). This property explains why the MMRE confidence intervals for models (E) and (D) overlap for NASA93, but not for the other datasets. At the 95% confidence level we cannot claim that the PRED rankings are significant except for model (A) which can be thrown out for the same reasons discussed previously. However, none of the PRED confidence intervals for (E) and (D) overlap, so we can be confident that effort multipliers improve accuracy in OLS calibrated COCOMO I models. This result is consistent with the MMRE results above and further strengthens our confidence in it.

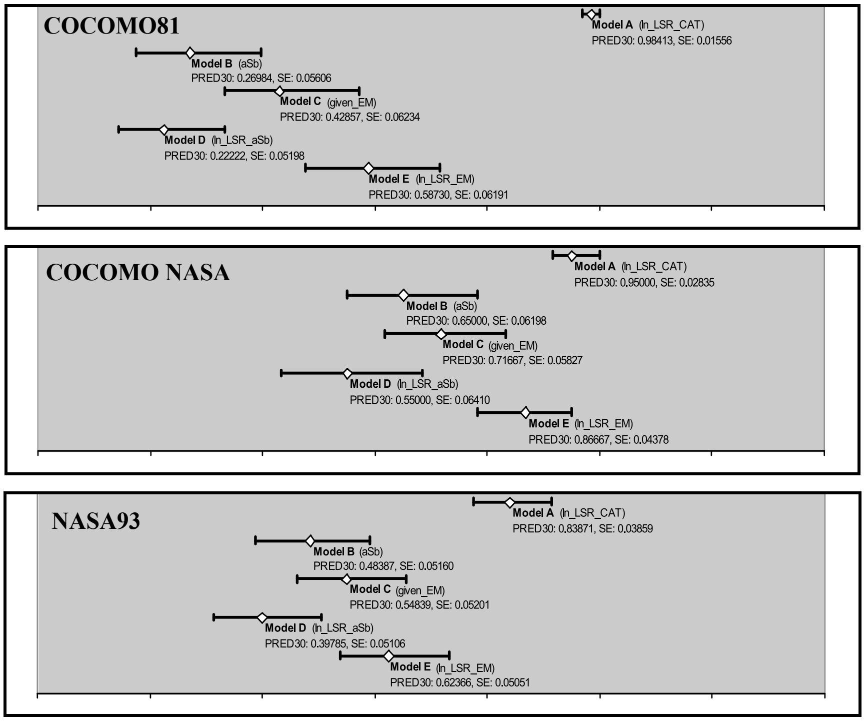


Figure 3. PRED(.3) bootstrapped 95% confidence intervals

In summary we confirm that SE (and hence confidence intervals) is affected by the type of model, particular data, and the estimator. From our confidence studies we find there is not much support for distinguishing one model of type (B) through (E) taken together, however, using COCOMO effort multipliers does improve accuracy.

Study 2: Confidence in evaluation with multiple criterions

Figure 4 presents MMRE and PRED 95% confidence intervals for two models using function point (FP) size values in the Desharnais data.

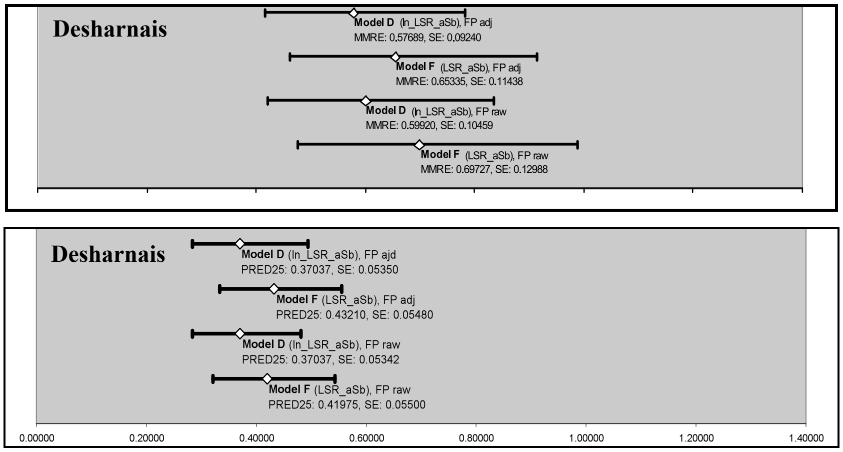


Figure 4. MMRE and PRED 95% confidence intervals

Table 6. Model ranking (Desharnais, FP adj)

	MMRE ranking	PRED(.25) ranking
1.	D	F
2.	F	D

Table 7. Model ranking at 95% confidence level (Desharnais, FP adj)

	MMRE ranking	PRED(.25) ranking
1.	F, G	F, G

Table 6 and Table 7 indicate the MMRE and PRED(.25) performance rankings for linear (F) and non-linear (D) models based. It has been suggested that multiple criteria be used to obtain more confident results, however it is evident in Table 6 and Table 7 that using multiple criteria can lead to another kind of rank reversal problem. In (Kitchenham et al, 2001) it is suggested that the fact that MMRE and PRED(.25) present inconsistent rankings for selecting models (D) and (F) is evidence that they are measuring two different aspects of the model's error distribution. This may indeed be true, however we suggest if SE is taken into account as it is in Table 7 the ranking results are actually inconclusive. In Figure 4 one can see that there is a substantial overlap of the 95% confidence intervals. Therefore we cannot determine if the difference in values is significant or the result of standard error. Hence, to have confidence that a rank reversal is present, one would need a great deal more data to reduce the SE, or use an alternative approach that is not subject to SE. The question of how much more data would be needed is taken up in Section 11 with a simplified, yet analogous example. However, from the methods presented there, we estimate that at least 4808 project data points would be needed to achieve 95% confidence that Model (F) has greater PRED(.25) than Model (D), and the number for MMRE would be much higher.

Table 6. Model ranking (Desharnais, FP adj)

	MMRE ranking	PRED(.25) ranking
1.	D	F
2.	F	D

Table 7. Model ranking at 95% confidence level (Desharnais, FP adj)

	MMRE ranking	PRED(.25) ranking
1.	F, G	F, G

To summarize, it is inconclusive that MMRE and PRED are subject to the rank reversal problem.

Study 3: Simulation study of validity of MMRE and PRED as selection criterion

This study replicates the cost estimation simulation results in (Foss et al. 2003) which suggest that MMRE is an invalid[1] criterion for selecting among competing prediction models. The evidence presented for this was in observing the frequency from 1000 trials for which MMRE would select a "true" model over four other models deliberately constructed to either over estimate or underestimate 30 simulated Desharnais-like effort and size data points. See (Foss et al. 2003) for further details of this investigation and construction and justification of simulated data.[2] While numerous alternative criterions to MMRE were also investigated, curiously PRED was left out. This notwithstanding, here we extend the results in (Foss et al. 2003) by including PRED and also account for SE in the results.

The appeal of using simulated data is that, in addition to being able to create a large number of sample data sets, we also know precisely the distributional properties of the population. Using this information we should be able to determine a premise for when a "true" model is deemed "best" over competing models. An OLS regression was performed on the Desharnais data set to determine parameters for a model of type (D) (see Table 1). These parameters were then assumed to represent the parameters for the whole population rather than just the Desharnais sample data itself. The model with these parameters, now of type (B), is called the "true" model as it is, in theory, the "best" fit to the population data. As per (Foss et al. 2003), the simulated data set is generated from the true model $effort = e^{3.03}(size)^{0.943}e^{u}$ by creating 30 normally distributed values u with mean of -.18 and standard deviation .6 and then calculating effort from the true model assuming these are the residuals for this model. Size values are multiples of 50 (that is, $size = 50 \bullet i$ for $i = 1, 2, \ldots, 30$). The competing models are of type (B) with differing values for the a and b parameters to selectively over- or under-estimate the predictions. Model(28) has parameters selected to severely underestimate the simulated data. Model(29)'s parameters were selected to also underestimate, but only moderately. Model(30) severely overestimates, while Model(31) only moderately overestimates.

1000 simulated data sets were generated and MMRE and PRED(x) were computed for each set. One model is "selected" over another if it has a "better" value – e.g. lower MMRE or higher PRED(x). When comparing at the 95% confidence level, if the result is inconclusive, it is not counted as selected for either model. Our replicated and extended results are given in Table 8.

[1] The authors use the term "unreliable" but we use a different term to avoid confusion with later studies

[2] There are numerous errors in (Foss et al. 2003) and care should be taken if replicating the study described there.

Table 8. Number of times competing model selected over true model using MMRE and PRED(.25) criterion

	Model(28)/true/inc	Model(29)/true/inc	Model(30)/true/inc	Model(31)/true/inc
MMRE	974/26/*	1000/0/*	0/1000/*	0/1000/*
MMRE @ 95%	360/0/640	0/0/1000	0/886/114	0/1/999
PRED(.25)	369/631/*	488/512/*	110/890/*	231/769/*
PRED(.25) @ 95%	40/52/908	4/0/996	0/210/790	1/8/991

From Table 8 we find that our MMRE results are consistent with those listed in (Foss et al. 2003). In our extended results, we find that PRED(.25) performs about as well as the standard deviation of the MRE's, the best criterion as described in (Foss et al. 2003). When applying the 95% confidence intervals we find good evidence that MMRE indeed seems to "incorrectly" select Model(28) over the true model in over 50% of the trials which is well outside the approximately 5% expected significance level. There is insufficient evidence to support MMRE improperly selecting Model(29). Remarkably we see no evidence that PRED(.25) is an unreliable model selection criterion. In fact, it appears to perform reliably in this capacity given it does not generally select the incorrect model. But we cannot be confident in this because the majority of the 95% confidence comparisons are inconclusive. That is, it did not generally select the incorrect model, but it also did not confidently (i.e. at 95% confidence) select the true model in general.

As an additional validation, we repeated the trials increasing to 500 simulated data points. In this case we expect smaller SE's and hence stronger evidence at the 95% confidence level. We see in Table 9 a dramatic strengthening of evidence that MMRE improperly selects Model(28) and Model(29), at least with respect to the discussion in (Foss et al. 2003) since at the 95% confidence level it did so for nearly 100% of the 1000 trials and never selected the true model. We see that PRED(.25) is inconclusive at the 95% confidence level for all but Model(30) where we can be confident that PRED(.25) selects the true model. We cannot conclude that PRED(.25) is unreliable. Curiously PRED(.25) seems to select Model(29) frequently when not comparing at 95% confidence. We have no explanation for this currently.

Table 9. Number of times competing model selected over true model using MMRE and PRED(.25) criterion 500 pts

	Model(28)/true/inc	Model(29)/true/inc	Model(30)/true/inc	Model(31)/true/inc
MMRE	1000/0/*	1000/0/*	0/1000/*	0/1000/*
MMRE @ 95%	1000/0/0	953/0/47	0/1000/0	0/1000/0
PRED(.25)	268/732/*	766/234/*	0/1000/*	3/997/*
PRED(.25) @ 95%	2/7/950	2/0/998	0/971/29	0/180/820

There is an additional consideration regarding the assumption that the "true model" is also the "best model" that also may play a role in the results we have discussed. What basis do we have that the OLS model used to generate the data is also the most accurate? Recall that the errors are generated by transforming a normal random variable into a log-normal variable whose mean is equal to 1, and the OLS model is itself transformed to an exponential model. This is no longer a simple linear least squares problem. Other optimizations (i.e. "best") are possible and would lead to different model selection results. This however, is well beyond the scope of the current study and we only make a note of this for future investigation.

Since PRED(x) has the tolerance level parameter x, we were curious what its effect it might have on model selection. In Figure 5 the left plot shows how the frequency of selections for the four models as x ranges from 0 to 1.5 (in increments of .1) whereas the plot on the right is when the 95% confidence interval was used. As the tolerance level is increased, we see that the underestimating models are selected more frequently and overestimating models less frequently perhaps indicating that PRED(x), like MMRE, favors models that underestimate. However there is little confidence in this assertion given that at the 95% level we see that PRED(x) rapidly begins to select Model(28) – the most severe underestimate – only after x is above a ludicrous level of .7, and even then, the maximum is about 40% of the trails at PRED(1) (note the different scales).

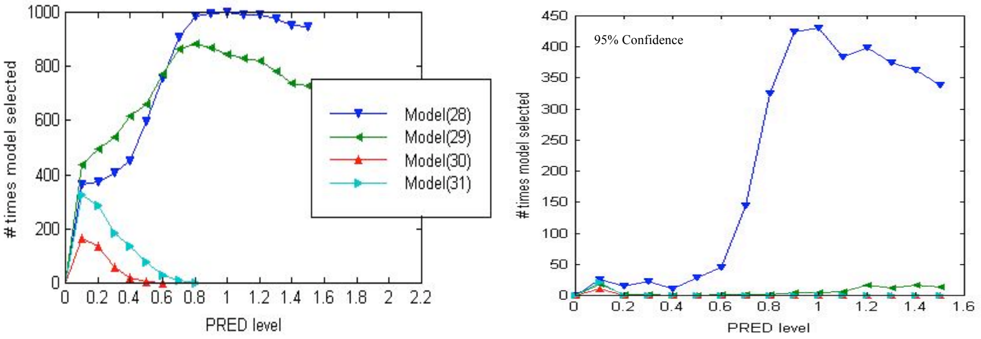


Figure 5. PRED(*x*) selection frequencies vs. PRED level *x*

In summary, we are confident in the conclusion that MMRE is an invalid model selector, but it is inconclusive if PRED is too. s

Study 4: Simulation study of reliability of cross validated accuracy criterion

In this study we study MMRE and PRED in combination with validation methods for selecting models. We replicate and extend the reliability and validity results described in (Myrtweit et al. 2005). It is similar to study 3 above in that simulated data is generated from a model with known error parameters and we consider the number of times competing models are selected over a 1000 trails. The data is simulated in nearly the same way as in (Foss et al. 2003) except the parameters are based on the characteristics of the Finnish data set (Briand and Wieczorek 2001). The models being compared are simple OLS (model B in Table 1) and a "nearest neighbor" estimation by analogy example of an arbitrary function approximator (AFA). Other than the use of the AFA model, another notable difference to the above study is the use of n-fold cross-validation, a variant of k-fold cross validation with *k* equal to *n*, to compute the MMRE and PRED values (the average of all n cross-validations is used as the representative value). It is unclear that this would ever produce a substantially different value over computing the MMRE or PRED from the entire data set, however, here we repeat this n-fold cross-validation to maintain veracity of the study. In theory this process could also produce an estimate for the SE, however we continue to use bootstrapping for the reasons mentioned previously and for its proven superiority over n-fold cross-validation as discussed in the statistics literature (Effron 1979).

As with study 3, an OLS regression was performed on the Finnish data set to determine parameters for a model of type (D) (see Table 1). These parameters were then assumed to represent the parameters for the whole population rather than just the Finnish sample data itself. As per (Myrtweit et al. 2005), the simulated data set is generated from the true model $effort = e^{1.70}(size)e^{u}$ by creating 40 normally distributed values *u* with mean of -.31 and standard deviation .79 and then calculating effort from the true model assuming these are the residuals for this model. The *size* value is uniformly distributed random variable between 0 and 2000. For each trial, these 40 data points are used to fit both an OLS and an AFA model along with n-fold cross-validation as specified in (Myrtweit et al. 2005). A summary of the MMRE's, PRED's, and their respective SE's and 95% confidence intervals is given in Figure 6.

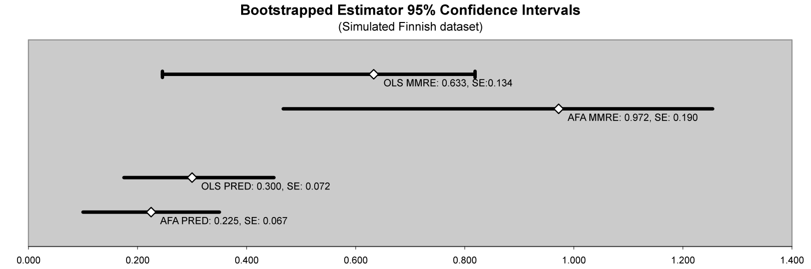


Figure 6: MMRE and PRED SE's and 95% Confidence Intervals for Simulated Finnish Data

Table 10 displays the percentage a given estimator selects AFA or OLS (i.e. that model had a "better" value for that estimator over the other model). When comparing at the 95% confidence level, we again used our standard criteria that if the 84% confidence intervals overlap, the result is inconclusive. Note that unlike Study 3, where we were trying to ascertain if the criterions consistently select the "correct" model, here we are investigating how consistent they are in selecting a model. The assumption here is that one of AFA or OLS must always be "best" (i.e. more accurate) on any set of the type of data being generated because there must only be one best model for the population. This assumption is somewhat questionable, but we will not discuss this further here.

Table 10: Results of Comparison between AFA and OLS Models, Highest Accuracy (40 data points)

Estimator	N	AFA(%)	OLS(%)	inconclusive

MMRE	1000	0.8%	99.2%	NA
MMRE @ 95%	1000	0%	5.7%	94.3%
PRED(.25)	1000	24.2%	75.8%	NA
PRED(.25) @ 95%	1000	0%	.6%	99.4%

We see from Table 10 that MMRE and PRED are consistent in selecting which model is best. Curiously, our results for MMRE are opposite to those stated in (Myrtweit et al. 2005). Since our MMRE results are very close to the MMER results they stated, we suspect that the authors of this study may have mixed the results for these. We are currently working with the authors to resolve the inconsistency.

We may be tempted to conclude that MMRE and PRED are consistent model selection criterions. However, we can not be confident in this because at the 95% confidence level the vast majority of the comparisons are inconclusive. In fact, the small percentages OLS was selected at the 95% level are reasonably within the expected 5% error range. Hence contrary to the conclusions stated in (Myrtweit et al. 2005), we cannot conclude from this study that MMRE or PRED are unreliable or invalid as model selection criterions. Perhaps using a larger data set will help drive more conclusive results. To investigate this, we increased the simulated data set size from 40 to 80 and repeated the experiment above with results listed in Table 10.

Table 11: Results of 95% Confidence Comparison between AFA and OLS Models, Highest Accuracy (80 data points)

Estimator	N	AFA(%)	OLS(%)	inconclusive
MMRE @ 95%	100	0%	60%	40%
PRED(.25) @ 95%	100	0%	7%	93%

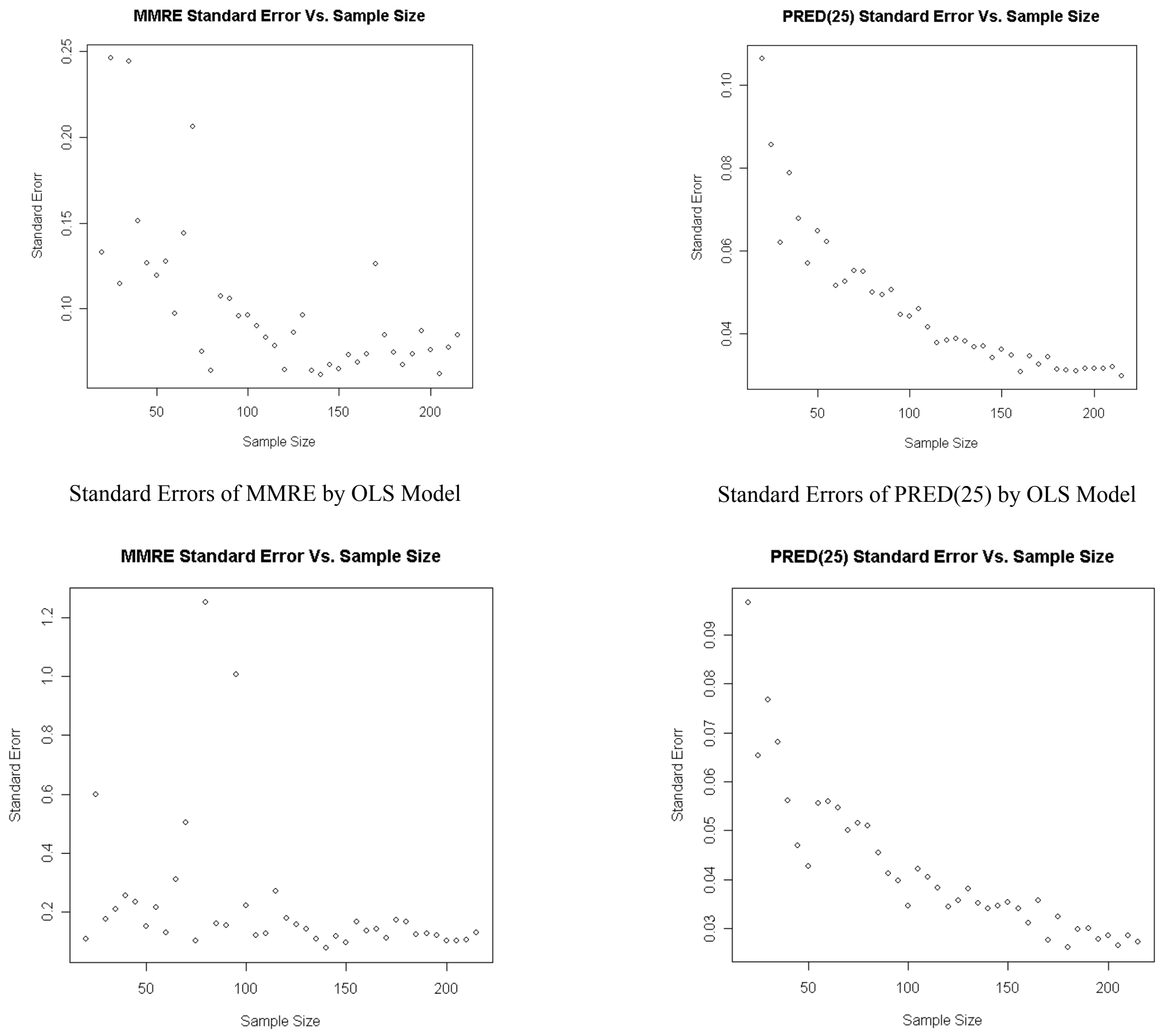
Interestingly we appear to get more conclusive results for MMRE but not PRED. However we would need more conclusive results in both if we are to have confidence in the result that MMRE and PRED reliably (or unreliably) select AFA and OLS models using cross-validation. A more refined approach is needed here that accounts for standard error due to sample size.

We believe that the reliability of estimators is better understood by considering the behavior of the SE with respect to sample size. To reliably select a model from two competing based on sample data, we must be able to reliably estimate the SE. Figure 7 illustrates the behavior of the SE for MMRE and PRED(.25) for both OLS and AFA as the sample size increases.

Here we see directly why the results are inconclusive a majority of the time. Despite a fully controlled data set, at a sample size of 40 the SE is large and the MMRE and PRED values are close enough to be within the expected variations. Even though there are a few significant cases where OLS beats AFA, we should not conclude that in general this is the case. This small number of cases can easily result from the expected standard error variations in the comparison values. Furthermore, the fact that there are zero significant cases where AFA beats OLS should not lead us to conclude that this is a general phenomenon. This could just as likely be the result of *bias* in the estimators (and indeed, MMRE is known to be biased).

We note a further lack of confidence in the MMRE as their SE's are themselves unpredictable. In Figure 7 the MMRE SE's values generally decrease as the data size increases, but not in a discernable way. They are clearly *unstable* in that there are seemingly arbitrarily large "jumps" in the SE, especially for data sets of less than 100 in size. If we repeated the experiment, the jumps would be different in size and occur at different locations. As such we particularly have no confidence in what the expected SE for relatively small data sets (although larger data sets may also be questionable). It is notable that the PRED SE's are predictable and decrease at the rate of $1/\sqrt{n}$, a useful fact we will explore and exploit later. This in part explains why in Table 11 we see that there are more significant cases where MMRE selects OLS then for PRED than might be predicted for comparisons at 95% confidence. From the instability perspective, this corroborates a few other studies such as (Kirsopp and Shepperd, 2002) that recommend not forming comparison conclusions based on MMRE with cross-validation when the data set is relatively small.

In summary, our replicated study indicates that we cannot be confident that MMRE and PRED are unreliable as model selectors. Increasing the number of sample points appears to help improve confidence in MMRE as a consistent selector where the inconclusive rate dropped from 94.3% to 40%, but this did not do much for PRED which dropped from 99.4% to 93%. Why this is so is clear from Figure 6 and the behavior of SE in Figure 7. The AFA and OLS PRED values in Figure 6 are much closer to each other than the MMRE values, and in Figure 7 at 80 data points the PRED SE's are still relatively large compared the MMRE SE's (but recall that these are unstable so we can't be sure). At 200 points, both the PRED and MMRE SE's are relatively low, and we might get confident results. Unfortunately, given the computation time for 80 points and 1000 trails was about 7 hours on a 2Ghz dual-core Intel processor, we estimate 4 days of computation time would be required for 200 points.



Standard Errors of MMRE by OLS Model

Standard Errors of PRED(25) by OLS Model

Standard Errors of MMRE by AFA Model

Standard Errors of PRED(.25) by AFA Model

Figure 7: Behavior of Standard Error of Estimators as Data Size Increases for AFA and OLS Models Predicting Simulated Data

Study 5: Stability of hold-out validations

Earlier we suggested that conclusion instability may be tamed by deprecating hold-out experiments and changing the evaluation methods for effort estimation models. Hold-out experiments are widely used in data mining (Witten and Frank 2005; Menzies et al. 2005) and proponents of those methods might argue for retaining these methods, and apply other methods to reduce conclusion stability such as row pruning of irrelevant examples or column pruning of noisy attributes. In this section we show empirically that the combination of hold-outs and pruning is fundamentally unstable and inadequate for taming conclusion instability.

Hold-out experiments divide the available data into a training set and a smaller test set. A model is generated from the training data and applied to the test data. A standard hold out study divides the data into a 66% training set and a 33% test set. For small data sets, there may be some concerns over the tiny sets so, in the following experiment, we used a fixed test set size of 10 instances. Each row of Table 12 shows the results of hold-out experiments over various subsets of the NASA93 and COCOMO81 data sets studied in this paper. Each row shows some subset of the data. For example:

- Rows 9 and 16 show data from different NASA centers, located in different parts of the USA.
- Rows 12 and 15 show just the data related to separate developments*: Project X* and *Project Y.*
- Rows 13 and 19 show just the data related *to semi-detached* and *embedded system* (*embedded systems* are those strongly coupled

with hardware and *semi-detached* systems are less embedded).

Further details on these stratifications and others can be found in the NASA93 PROMISE repository (NASA93 dataset, 2007). A key point to keep in mind here is that the subsets are not machine generated. They are groupings based on project organizational characteristics. In our experiments, for each row, 10 examples were selected at random. Models were generated using Boehms's local calibration procedure (Boehm 1981) which is a regression methods specialized for learning COCOMO models. This procedure was repeated 30 times.

The performance of the models learned over the training set was assessed using the test set via MMRE. The rows in Table 12 are sorted first by their source (COCOMO81 or NASA93) and then by MMRE.

Table 12: Hold-out Experiments

	subset	#training examples	#test examples	MMRE	sd MRE
1	coc81:mode.org	13	10	32	27
2	coc81:lang.mol	10	10	34	29
3	coc81:all	53	10	42	45
4	coc81:mode.e	18	10	42	47
5	coc81:kind.max	21	10	47	51
6	coc81:kind/min	11	10	47	66
7	coc81:lang.ftn	14	10	50	48
8	nasa93:cat.avionicsmonitor	20	10	43	47
9	nasa93:center.2	27	10	43	148
10	nasa93:cat.missionplan	10	10	46	45
11	nasa93:fg.g	70	10	53	126
12	nasa93:project.Y	13	10	56	168
13	nasa93:mode.sd	59	10	58	149
14	nasa93:all	83	10	60	157
15	nasa93:project.X	28	10	68	142
16	nasa93:center.5	29	10	80	169
17	nasa93:year.980	28	10	81	211
18	nasa93:year.1975	27	10	82	192
19	nasa93:mode.e	11	10	188	649

A common claim is that software effort models work better when calibrated with local data (Boehm et al. 2000; Ferens and Christensen 1998; Lum et al. 2002). In that view better models can be built from *row pruning* the data; i.e. train from subsets of the data rather than all the available data. According to this argument, we should not mix up training data from (Center 1, Center 2) or (Project 1, Project 2) or (embedded, semi-detached) systems or indeed any two subset rows of Table 12.

Table 12 finds no evidence that row pruning improves effort estimation. Measured in terms of MMRE, learning from data subsets yielded worse models nearly half the time. In the COCOMO81 data, four of the six subsets have an error equal to or larger than the error found after applying all the COCOMO81 data. In the NASA93 data, five of the 12 subsets have a larger error than the error found after applying all the NASA93 data.

(One objection to the above result is that the learner used in that study is too simplistic since it uses a method defined by Boehm in 1981. We reject that argument as follows. Previously (Menzies et al. 2006) we have tried to out-perform Boehm's methods using a variety of data mining methods. In extensive experiments with a large number of data mining methods, we found that 15/19 of rows of Table 12, Boehm's 1981 methods performed as well as anything else.)

In terms of conclusion instability, the important feature of Table 12 is the large standard deviations: up to 649%, in the worst case. These can grow disturbingly large: in 14 rows the standard deviation is larger than the means. These large deviations are not tamed by row pruning. In COCOMO81 and NASA93, four and five data sets (respectively) have a larger variance that that seen after training from all the data. Such large deviations, make it difficult, to say the least, to distinguish the performance of different effort estimation models.

If row pruning cannot tame conclusion instability then perhaps *column pruning* might be useful. Miller offers an extensive survey of attribute column pruning for linear models and regression (Miller 2002). That survey includes a very strong argument for attribute pruning: the deviation of a linear model learned by minimizing least squares error decreases as the number of columns in the model is reduced. That is, the fewer the columns, the more restrained are the model predictions.

Elsewhere (Menzies et al. 2006), we have applied column pruning to the Table 12 data using a data mining technique called the WRAPPER (Kohavi and John 1997). The WRAPPER uses some subset of the N columns of data in a training set. That subset is grown if it the mean performance of N+1 attributes is better N. An AI heuristic search technique (best-first search) is used to avoid a 2^N exploration of all combinations of the attributes.

In terms of reducing conclusion instability, the WRAPPER experiments were not useful. Use of t-tests (95% confidence) could not distinguish between models in nearly half our experiments. This was due to the large standard deviations remaining in the results, even after applying the WRAPPER: in 14/19 of the rows of Table 12, the MRE standard deviation remained larger than the MMRE. Furthermore, it is easy to observe that the MRE's are not normally distributed, sometimes grossly so, thereby making standard statistical tests difficult to apply.

In summary, after extensive experimentation we could find no conclusive evidence that row or column pruning improves our ability to distinguish between models. A common feature to all the above experimentation is the use of hold-out experiments (Table 12 and our prior work (Menzies et al. 2006) used hold-outs of 10 test cases). This paper was motivated by the following speculation: perhaps the real cause of conclusion instability is not quirks in the data or problems with the evaluation metric. Rather, the real cause might be the hold-out experiments themselves. In the sequel we show that conclusion instability disappears if we track the standard error of the generated models rather than use hold-out experiments. That is, we strongly recommend *against* hold-out experiments. In the sequel, we will revisit this study from the perspective of standard error. We will find that the cause of instability in Table 12 was the hold-out analysis used and that a different analysis can yield stable results.

11. EMPIRICAL AND ANALYTICAL CHARACTERISTICS OF MMRE AND PRED

Here we consider two important characteristics of PRED and MMRE – their sample distributions and the behavior of SE with respect to data set size. Example applications of these characteristics are also presented. The results are generally not replications or extensions of previous studies, but were chosen for their interest as well as ease of validation.

Distributional characteristics of PRED and MMRE

The left side of Figure 8 is an example "reconstructed" distribution of the bootstrapped MMRE's from use of model (A) on the NASA93 data set. It is, in theory, asymptotically normally distributed (Mooney and Duval 1993), yet clearly it is not normally distributed for the relatively small data set used as is evident by looking at the closest fitting normal curve that is superimposed on the histogram in Figure 8. Other non-normality tests such as skewness and kurtosis, and normal p-plot are consistent with this finding (note that we would not expect a Mode given that MMRE is not discrete). Our empirical investigations on the other models and data sets, including the simulated data set, reveal that in general the MRE's are log-normally distributed, and while unimodal, they are skewed to the left. In addition to some analytic results that support this belief, it is evidenced empirically by considering the log-transformed distribution displayed on the right of Figure 5 which looks decidedly normally distributed.

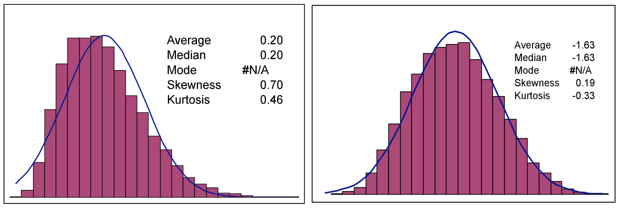


Figure 8. Histogram of bootstrapped MMRE and log-transformed MMRE for model (A), NASA93 dataset

A complete analytical characterization of the distribution of the MRE's and consequently of the sample MMRE for COCOMO models is complex, but empirically it is clear that assuming normality is unjustified and will likely lead to inconsistent results.

In contrast, PRED as seen in equation (4) is based on the sum of indicator values for the MREs which will have a more tractable binomial distribution (Larsen and Marx 1986). Even though resulting distribution for sample PRED estimators is discrete[3], it is approximated very

[3] for a given data set of size of n there are only 2^n possible sums of 0-1 indicator values and therefore only that many possible PRED values

well as a normal distribution as can be seen in Figure 9. This is due to the classic normal approximation to the binomial (Larsen and Marx 1986) which makes PRED easier analytically to work with over MMRE as we now show in the next consideration.

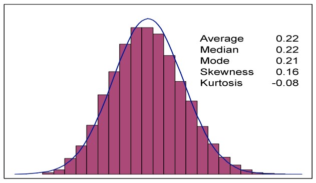


Figure 9. Histogram of bootstrapped PRED(.3) for model (D), COCOMO81 dataset

Application: how much data for data significance?

Thus far we have only looked at 95%-confidence intervals. However, especially from a model selection point of view it is interesting to ask: "How confident can I be that my chosen model will be significantly more accurate?" Therefore we chose a typical model selection example, COCOMO NASA in Figure 2, and decreased the significance level until the intervals no longer overlap. In Figure 10, we found that this occurs at a confidence level of 32% or lower. Meaning that 68% of the time the actual value of the parameter for the error distribution may lie outside of the interval where we cannot be certain how it compares to other values. A problem with this analysis is that it should only be applied to pairs of models, not all at the same time. When two models are essentially the same, the confidence will be reduced greatly to avoid interval overlap, yet there may be another model outside of these that would not overlap at a much higher level of confidence. None the less, it illustrates an approach to obtaining a confidence level in choosing between two models.

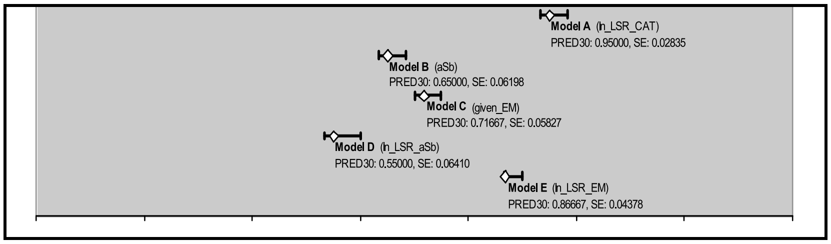


Figure 10. Bootstrapped PRED(.3) non-overlapping 32% confidence intervals COCOMO NASA

Alternatively we might ask the question: "How much data is needed to get significant results?" Because MMRE and PRED are consistent estimators, we know that the standard error must decrease as the number of data points increase. MMRE has a complicated relationship between standard error and data size and is not easy to work with. However PRED is more tractable and it can be shown that

$$SE_{PRED(x)} \approx \frac{SD_{1(MRE\le x)}}{\sqrt{N}} \qquad (6)$$

(≈ means approximately equal) where *SE* stands for standard error and *SD* is the sample standard deviation of the indicator values of the MRE values less than *x* (i.e. 1 for the MRE's less than *x*, 0 otherwise) for *N* data points. Since the bootstrap distributions for PRED are approximately normal (or student t-distributed for small *N*), we can estimate *N* such that the 95% confidence intervals for models C and E are unlikely to overlap by considering where the PRED's are not within $z_{0.05} \approx 1.645\ SE$ for each of the model's respective standard errors where $z_{0.05}$ is the 95% percentile for standard normal.

For our example, we assume PRED for C is less than PRED for E and so the intervals will not overlap when

$$PRED_C + 1.645 SE_{PRED_C} < PRED_E - 1.645 SE_{PRED_E} \qquad (7)$$

Substituting (6) and solving for *N* gives:

$$N > \left(\frac{1.645(SD_{1(MRE_C \le x)} + SD_{1(MRE_E \le x)})}{PRED_E(x) - PRED_C(x)} \right)^2 \qquad (8)$$

Applying equation (8) to our previous example, the COCOMO NASA data set with PRED(.3) we find *N*>76 if 95% confidence is desired that $PRED_E$>$PRED_C$ i.e. to have confidence that local calibration is PRED-superior over using a general model. To match Figure 10 where all intervals do not overlap we have at least as much data as the maximum for any pair (again, keep in mind that there may be pairs that require less data). For COCOMO NASA this turns out to be Models (B) and (C) and (8) suggests *N*>756. Given that the data set has only 60, we do not have a sufficient amount of data to conclude in confidence the rankings in Table 4. However, we do have a reasonable estimate of how much more data we might need collect to get to such a conclusion.

The method presented is somewhat crude and conservative, but simple and informative because we can easily visualize its meaning - the PRED(.3) intervals do not overlap. Strictly speaking we cannot say "the intervals do not overlap 95% of the time" without resorting to more refined calculations, however these are more complex and not without their own interpretation challenges. As discussed earlier, we could also follow the advice in (Payton et al. 2003) and use $z_{0.16} \approx 1.645\ SE$ as a reasonable approximation for 95% confidence of non-overlap. There are also more refined methods for approximating two-sample or multiple-sample confidence intervals that may provide tighter size estimates. However the example in (8) provides a reasonable feel for the relationship of SE to data size needed for significance and so we are able to address "how much is enough data" questions.

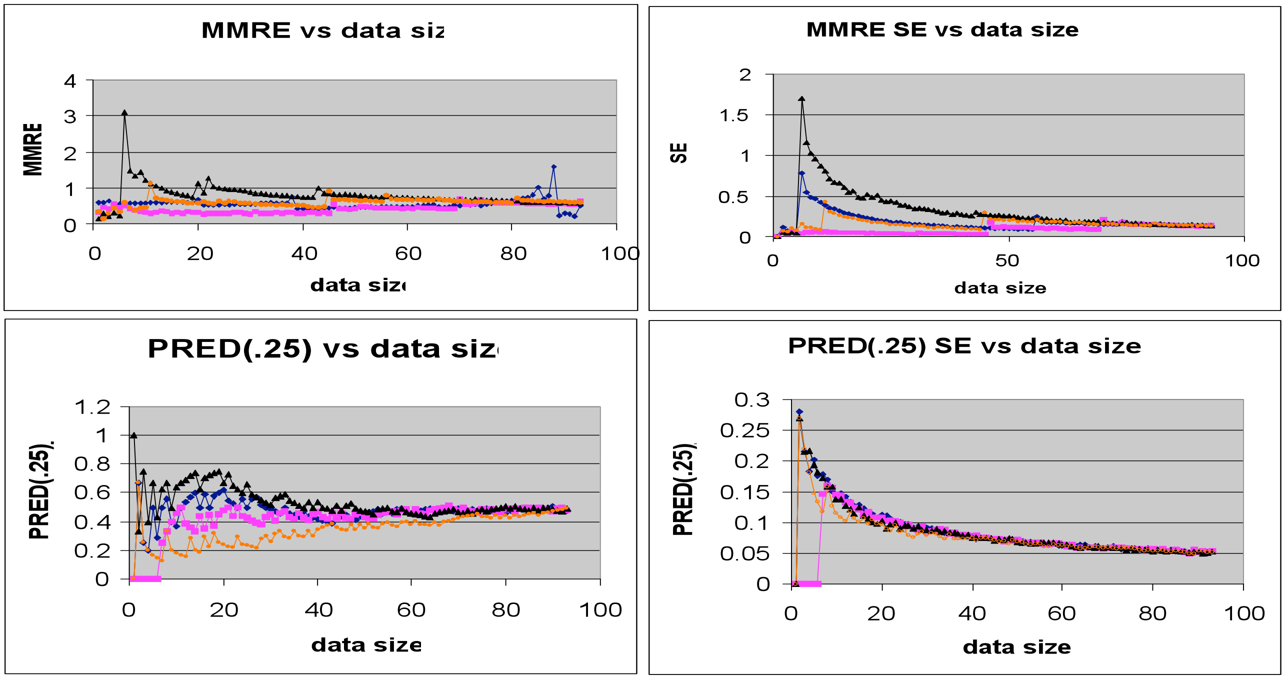


Figure 11. MMRE and PRED(.25) performance with respect to increasing data set size for model (C), NASA93

PRED and MMRE Estimator Performance

We now focus our empirical microscope on the question of how well, as estimators, MMRE and PRED perform with respect to data set size. The primary purpose of the experiments here is to compare expected versus observed behaviours of MMRE and PRED when increasing sample size (i.e. more project data points). In doing so, we illuminate a number of striking differences between the two criterions and clarify the notion of estimator reliability and it's consequences to model accuracy confidence.

The experiments were performed using Model (C) with the NASA93 data set. This was selected as both model and data are reasonably representative. Neither are highly specialized or "tuned and pruned" for specific results. We note in advance that we have performed the same experiments repeatedly with all the data sets and simulation data and have observed no significant differences with respect to the analysis for the example presented. Figure 11 shows a sample of four different runs of randomizing the NASA93 data set and creating subsets of size n=1,2,…,93 by progressively adding in points from the set. The MMRE and PRED(.25) is calculated for each subset and bootstrapping is used to estimate their respective SE's. From the distributional results discussed above we could approximate the SE without bootstrapping, but we find it more confidence inspiring to use bootstrapping then to verify the results with a non-bootstrap approximation. Figure 11 shows the results plotted as functions of data set size.

We now analyze the experimental results both empirically and analytically as a series of "expected" and "observed" behaviors as data size increases in Figure 11.

<u>Case *n*=1</u>

Expect: MMRE will be the MRE for the single data point, SE = 0 as there is no variance in the sample data. PRED will be 0 if MRE > .25 or 1 otherwise, SE = 0 for similar reasons.

Observed: Exactly as expected. This serves as a basic verification that the experiment is correctly configured.

<u>Case *n*=93</u>

Expect: All runs should have same MMRE and PRED values and respective SE's as all data is being used and there is no random variation from selecting subsets.

Observed: Exactly as expected. This serves as another basic verification that the experiment is correctly configured. We note that the bootstrapping introduces less than a .01 error in the approximating SE for MMRE and .004 for approximating SE for PRED(.25).

Case *n*→93

Expect: MMRE and PRED are both consistent estimators for non-negative random variables hence we expect that their respective values will uniformly converge (i.e. does not depend on the particular value converged to) in probability to the distribution parameters they estimate. This also implies that the SE's should uniformly converge to 0. MMRE is an average of continuous random variables (approximately log-normal) and we would not expect a large deviation in the rate of convergence, especially as the data set gets large. In contrast with this, we anticipate a degree of bounded variation for PRED(.25) since each indicator $1(MRE \le .25)$ has expected value $P(MRE \le .25)$ and so this is essentially a series of Bernoulli trials (i.e. flipping a weighted coin) where at each *n* the PRED may increase by $(1 - PRED_{n-1})/n$ or decrease by $PRED_{n-1}/n$ where $PRED_{n-1}$ is the PRED(.25) value at *n*-1 data points, hence the variations are tightly bounded by the previous values. Since each increase or decrease is divided by *n*, the magnitude of these variations decrease as *n* increases. Hence we may see a few rare long runs up or down, but generally a repeated series of short increases followed by short decreases that are smaller and smaller as *n* approaches 93.

From that fact that MRE is an absolute value, the SE for MMRE is dependent on $P(z > 1)P(z \le 1)$ where the probability is from the distribution of z (see Section 4). For large enough *n* this value should be fairly constant when approximating this with an empirically derived distribution (as we are doing with bootstrapping). More importantly however, the SE is also dependent on the *variance* of z. This implies that for large point variations in the data, perhaps due to outliers, we would expect large increases in an otherwise decreasing SE. Note that MMRE is the average of *n* MRE's, so its sample variance will be divided by n^2 and hence the SE will be divided by *n.* The PRED SE should be quite well behaved as it only depends on $P(MRE \le .25)P(MRE > .25)$ and not the variance of the MRE's as with MMRE. As stated previously, we expect the approximate values for these probabilities as derived from the empirical distribution used by the bootstrap to rapidly stabilize and therefore we would expect the PRED SE to generally decrease on the order of *1/n*.

Observed: The experimental runs have the characteristics as described above with a few additional notable items. First is that contrary to popular belief in the cost estimation folklore, we see that in general, more data does not imply better accuracy values. Even in our analysis of expected behavior above, we anticipated the possibility of both decreasing and increasing MMRE and PRED values as data size increases. All we can say with confidence is that the criterions will eventually converge, and a bit on how much variability to expect along the way. That we cannot predict for any given set of data that more data will improve the MMRE or PRED is an *absolutely crucial fact* about the use of these as accuracy criteria. This indicates that it is imperative to always consider the SE of these accuracy measures when using them to avoid erroneous results due to sampling phenomenon. We believe that this is a major source of inconstant cost estimation research results, a few of which we have exemplified and resolved in section 10.

Small *n*

Expect: We expect that both MMRE and PRED will be unstable due to high sensitivity. A single new data point may have a large effect on the average value of a small data set. For similar reasons the SE's should also be unstable.

Observed: For each *n* < 11 we observe that there is a large variation in that data set's MMRE, PRED, and respective SE values.

Sensitivity to outliers and estimator reliability

Another notable observation is that we clearly see from the large variations in the MMRE runs in Figure 11 that there are at least 3 significant outliers in the NASA93 data. Hence as has been frequently claimed in the literature (and expected as above), MMRE is indeed very sensitive to outliers. We see that even very near the full 93 data points that an outlier can cause a radical variation in the MMRE. Although this introduces uncertainty in the estimated value of MMRE, this by itself does not necessarily make MMRE unreliable as an accuracy criterion. What does make it unreliable is how outliers affect the MMRE SE. A reliable estimator is one that is likely to give the same estimate for repeated trials on the same data. The interpretation of "likely" here is defined by "within a known range of the SE." In our case the SE must be "strictly decreasing with probability 1", which allows for the occasional increase, but overall is decreasing when sample size increases. The reason for this is subtle, but simple. As the sample size increases towards the entire population the SE must converge to zero (i.e. there is no SE for an estimator calculated on the entire population data). Hence for a reliable accuracy estimate, the SE must be predicable. Given that an outlier may radically increase the MMRE SE at any time, and continue increasing, we cannot reliably estimate it for a given data size, and thus we are uncertain about the true value of the error parameter estimated by MMRE.

In contrast, PRED is a reliable estimator. As expected, and observed, the PRED SE is "generally" decreasing and we never see and long running increase (as we do see for MMRE SE). Note that the small increases seen in the PRED SE runs in Figure 11 are the result of approximation errors and random variation from bootstrapping. Because the sample MMRE's are non-discrete, the SE is less effected by these kinds of errors. As seen in Figure 11, the PRED estimate varies quite but between different trials, but always within a very well predicted range as defined by the SE. Some meditation on this by comparing the magnitude of the PRED SE for a given data size (right side of Figure 11) and the corresponding range of PRED values (left side of the figure) will make this property apparent. A similar meditation on the MMRE SE and MMRE reveals that MMRE is clearly not reliable, even though paradoxically, the MMRE values themselves appear to be more consistent than PRED over the different trials. We also observe that unlike MMRE SE, PRED SE tends to stabilize quickly, and that after 10 data points all runs more or less converge on the same trajectory. Thus we can reliably predict the SE and have greater confidence in results based on PRED so long as we account for this SE. We also see that PRED is much less affected by the outliers we know are present in the NASA93 data set used in this example. Indeed, very unlike MMRE, there seems to be no discernable perturbations in the PRED and PRED SE trajectories. In this sense, PRED is "robust" as an estimator.

But reliability is more useful than robustness. Take for example a statement made by Barry Boehm in talking about how much data is needed to reliably locally calibrate a COCOMO model. He suggested that 10 data points is generally sufficient. In Figure 11 and Figure 14 we observe that the PRED SE tends to "stabilize" rapidly after 10 points, and given that PRED is a reliable estimator, we are confident that the PRED SE can be used to bound the error in the PRED and hence we know what confidence we have in the estimates given by the model (although this does not say that they are accurate). The point here is that we can determine if the calibration results in a significant improvement. To illustrate, in our studies of the three COCOMO data sets in the PROMISE repository in Figure 3 we can be confident that that local calibration using model(E) outperforms the non-locally calibrated model(C) in the COCOMONASA data set at the 95% confidence level (recall that we consider the overlap of 84% confidence intervals, not the 95% intervals shown in the figure). To make this conclusion we must know that the PRED SE has stabilized and that PRED is a reliable estimator. Hence there is some support for Boehm's "10 is enough" heuristic for locally calibrating models.

Behavior with respect to randomized subsets

In the above analysis each "run" was started by randomizing the original data set, then adding in one point at a time from this set, calculated the MMRE, PRED, and their respective SE's, and graphing the results. We believe this is a practical perspective because it represents how the estimators and their standard errors change as more data is *collected*. The downside of this is that it does not show directly how population samples behave as a function of the size of a sample. An alternative perspective that captures this is to select a random subset of increasing size from the original sample data. This differs from the previous procedure in that the previously selected points are not reused. That is, the sample of size *n*-1 will *not* in general be a subset of the sample of size *n*. This perspective is shown in Figure 12.

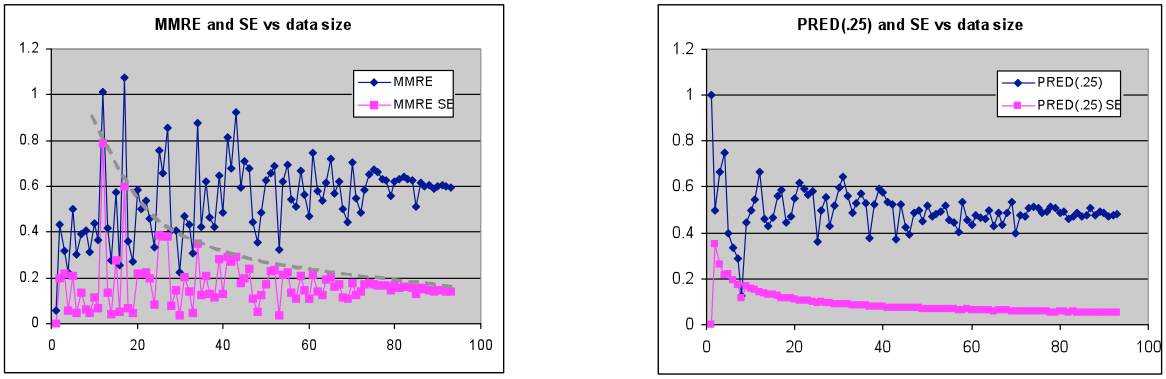


Figure 12: MMRE and PRED(.25) performance with respect to randomized subsets of increasing size for model (C), NASA93

While the perspective in Figure 12 does not provide different results from the previous discussion, it more clearly and quickly summarizes the estimator reliability issue. The randomized subsets are equivalent to samplings from many runs of the collection approach illustrated in Figure 11 and so there is no need to have multiple runs. It will always represent the average behavior. Consistency is also clear in that as expected the SE, on average, decreases. Here it is clear that PRED is reliable and MMRE is not. We see that the trajectory for PRED SE is dependent only on the data size and not the data itself (recall that each subset is a random selection of data). As discussed earlier, this is expected because the MMRE SE is dependent on the variance of the data while PRED is not. One thing this perspective offers over the collected data perspective is that it shows that the maximum MMRE SE may serve as a reliable predictor. To make use of this, we need only find, as exemplified as the dashed line in Figure 12, the σ that $\sigma/\sqrt{n}$ fits the maximum MMRE SE points best. This function can subsequently be used as a reliable estimate for the MMRE SE for a data set of size *n*.

This perspective is not useful in detecting outliers, but the behavior of the SE is more easily analyzed. The two perspectives are consistent, however, as an external sanity check, we note that our results in this section are consistent with similar studies of properties of cost estimation accuracy with respect increasing data set sizes as found in (Mair et al. 2005; Kirsopp and Shepperd 2002).

12. INCREASING CONFIDENCE

Given a consistent, stable and reliable accuracy criterion, in most cases more conclusive results can be obtained by increasing the sample data size. We can estimate how many data points are needed to obtain a given level of confidence. This may imply a large number of new data points. In practice it may not be possible or practical to obtain this new data. In this section we explore two alternative means that may help improve confidence by reducing data – stratification (also known as row pruning as discussed previously) and outlier removal.

Can stratification increase confidence?

Study 5 in Section 10 introduced stratification (row pruning), or calibrating a model with strategically chosen subsets of data. We found little evidence that this practice significantly improves estimation accuracy. But this does not mean stratification is valueless. Here we investigate the possibility that stratification can increase confidence by reducing SE. Intuitively this may not seem viable given that stratification reduces the number of data points which we would generally expect to increases SE. The idea here is that highly related data will have less variation than is incurred from the increased sensitivity (i.e. variations have larger effect) from having fewer data points. Table 13 displays three stratifications of the NASA93 data set based on system type (recall that these groupings were supplied by the organization, in this case NASA).

Table 13: SE of NASA93 Stratification Subsets

Stratification	Data Size (n)	MMRE SE	PRED(.25) SE	PRED25 sigma
nasa93:all	93	0.14	0.05	0.47
nasa93:cat.avionicsmonitor	30	0.1	0.09	0.49
nasa93:cat.missionplan	20	0.07	0.11	0.49
nasa93:mode.sd	69	0.09	0.06	0.5
nasa93:mode.e	21	0.48	0.08	0.37
nasa93:cat.avionics	11	0.55	0.14	0.46

As is seen in Table 13, the SE's for the MMRE are about the same or decreased. Stratification may indeed increase confidence. However, recall from Section 11 that MMRE is an unreliable estimator for this data using COCOMO models in the sense that the relation of SE to data size is unstable. We do not know what the range of values this SE has, so these reductions may be due to random variation. We cannot conclude with confidence that stratifications improve SE based on MMRE SE reductions.

However we also found in Section 11 that PRED is reliable. In Table 13 we see that the SE's for PRED(.25) tend to increase. This is expected due to the reduced data sizes. To adjust for this and see if the variance had any reduction we calculate PRED25 sigma or PRED(.25) SE multiplied by $\sqrt{n}$. We do this because PRED is a reliable estimator, so if its population variance is sigma, then its average SE will be $\sigma/\sqrt{n}$ .We see that the PRED25 sigma's are about equal for the nasa93:cat indicating no variance reduction for these stratifications. We spoke with the originator of the NASA93 data set about this, he said that the stratifications listed it NASA93 are "really bad, almost meaningless" and he's not surprised with this result and he suggested we may have better luck with the nasa:mode stratifications: Here we do see a reduction for the nasa93:mode.e stratification. Since the variance in the PRED(.25) SE estimates generated from bootstrapping are well understood, we should perform some sort test such as double bootstrapping or a t-test to ensure the reduction in SE from 0.47 on 93 points to 0.37 with 21 point is significant. Since this is an involved calculation, we skip this step for now since the point here is that by study of SE with a reliable estimator we have a practical tool for investigating the efficacy of a stratification. It should be noted that if a stratification is found to reduce SE, then care should be taken not to generalize the subsequent model learned from that data subset to the entire population. The expected reduced SE only applies to the population data in the same stratification - in this case, "embedded mode" NASA93 type projects.

Does removing outliers increase confidence?

In our analysis of Figure 11 we noted that the NASA93 data set must contain a few points that cause extreme variations in the MMRE SE. An alternative to grouping like data together to reduce variance is to remove (or prune) these points from the data set. This has the advantage that the resulting subset is still generally representative of the population, albeit not for the extreme cases which are typically not captured faithfully in the estimation models anyway. As discussed previously we could calculate the MMRE SE on collectively increasing data sets and whenever there is an unexpectedly large jump in the MMRE SE (i.e. outside of what might be expected as variation from a reliable estimator), the point that was just added can be classified as an *outlier* and subsequently removed from the data set. A more aggressive pruning approach would be to prune points that cause large jumps in the MMRE rather than the MMRE SE. Since our goal here is to reduce SE, we shall explore the more conservative approach. We made several runs and pruned out points where there was a positive jump of more than 10% of the previous SE. The choice of 10% is somewhat arbitrary, but we believe it is conservative. We pruned only positive jumps since the SE should decrease with increased data size. Furthermore, we only considered jumps that occurred after 50% of the data set as we know that MMRE is unstable for small data sizes. We use 50% rather than a fixed value because the data set size is decreasing as we prune the outliers. Note that a single run will not generally identify all outliers due to successive dominance. This occurs when the jump in SE in a subsequent outlier is smaller and nearby the preceding outlier which "dominates" it so its effect on SE is negligible. The result of four runs after pruning is shown in Figure 13.

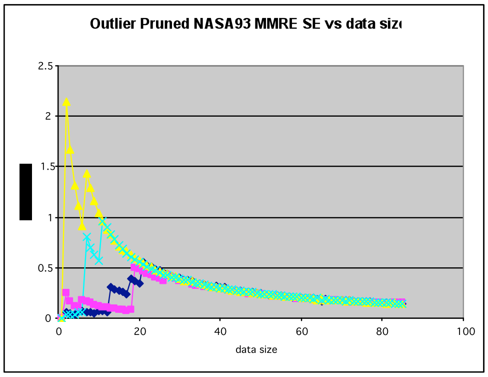


Figure 13. MMRE performance with respect to increasing data set size for model (C), NASA93

We could continue pruning until the MMRE SE curve is generally smoothed, but the trend is clear – the MMRE SE is more stable. What is remarkable is how quickly the MMRE SE stabilizes after *n* = 20 data points. After this point, we no longer see the erratic behavior exhibited in the MMRE SE's runs of Figure 11. For the pruned data set now with *n* = 85 data points the bootstrapped MMRE SE increased to 0.147. Adjusting for the 8 fewer data points indicates the data set has the same SE as the original and so it appears that pruning outliers does not decrease SE. However pruning outliers improves reliability and therefore may be useful for improving confidence in the use of MMRE SE to when forming conclusions based on MMRE comparisons.

As a final check we look at the COCOMO81 data set which is known to have pruned outliers, though a different pruning process was used. In Figure 14 the MMRE and MMRE SE are much better behaved than in the NASA93 data set in Figure 11. The properties discussed previously still hold. Here again, after 20 points the MMRE SE is reasonably stable and therefore may be useful for determining the confidence we can have in MMRE as an accuracy criterion.

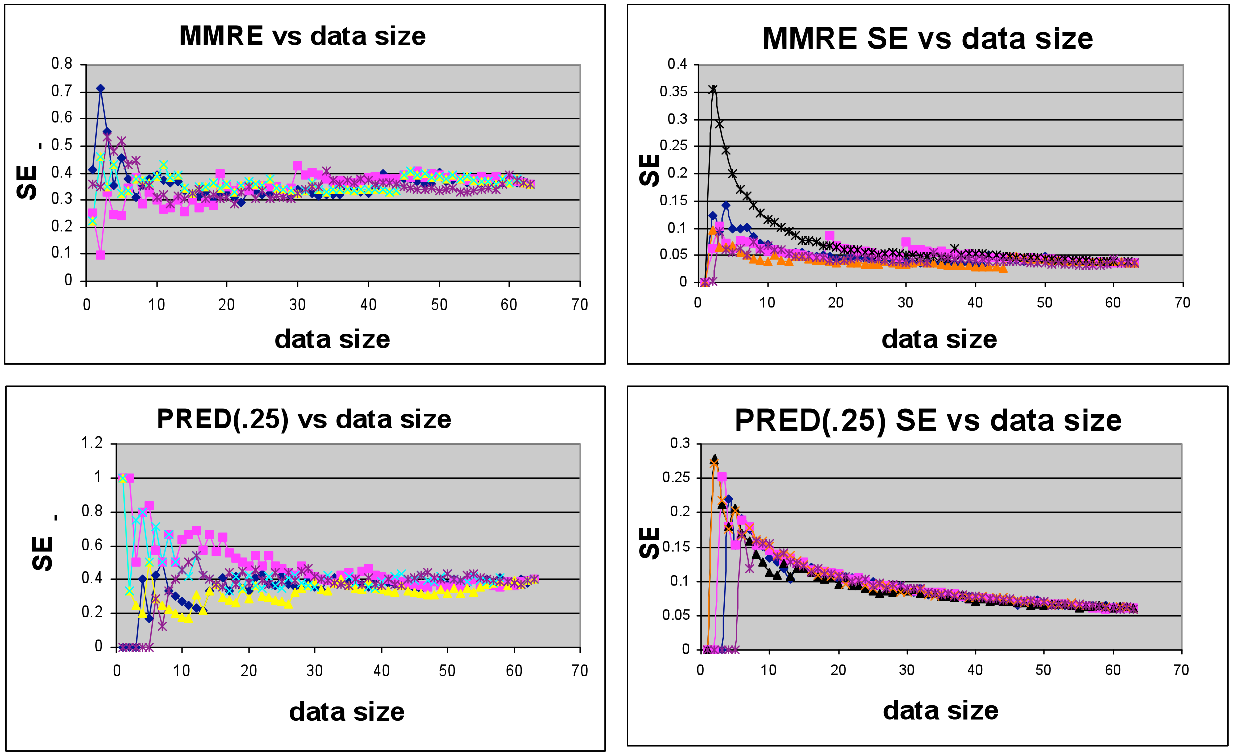


Figure 14. MMRE and PRED(.25) performance with respect to increasing data set size for model (C), COCOMO81

13. SUMMARY AND CONCLUSIONS

The question on whether to use PRED or MMRE is probably beside the point. In the above, we show that either are useful (but MMRE must be used carefully, see the discussion around Figure 13). However, what is vital is that whatever accuracy measure is used, the standard error must also be considered. We advise understanding clearly the properties of these criterions and ensure that they are appropriate to the problem at hand. When using these criterion, or with any accuracy estimator, if one wants to have confident results, care should be taken to understand the SE and the consequences of data quality and data size. We have demonstrated that in the light of standard error, many research results and methods are erroneous or inconclusive. This can constructively explain the conclusion instability seen in preciously published studies. By offering an "inconclusive" alternative to a studies outcome, we have shown several examples that resolve the conclusion instability problem both within a given study and across multiple studies. When a result is inconclusive, it does not imply that it is incorrect. Rather, we need to increase confidence the result by reducing the standard error, or by searching for more powerful methods to provide significant results.

In this study we have applied standard, easily analyzed and easily replicated statistical methods from the theory of estimation – SE and bootstrapping - to a number of exemplar cost estimation model research results. We would like to reiterate that our primary purpose is to investigate the confidence we may have on results based on PRED and MMRE and not to advocate the use of these over other accuracy criterions. To summarize some of the specific confidence based findings from our replicated and extended studies:

- Study 1: Confident that use effort multipliers for COCOMO type models improve accuracy [supports (Menzies et al, 2005)]
- Study 2: Inconclusive that MMRE and PRED are inconsistent model ranking criteria [unconfident for results in (Kitchenham et al, 2001)]
- Study 3: MMRE is an invalid model selector and favors models that underestimate [supports (Foss et al. 2003)], inconclusive if PRED is a valid or invalid model selector or favors models that underestimate [extension to (Foss et al. 2003)]
- Study 4: Inconclusive that MMRE and PRED are unreliable model selectors [unconfident for results and extension for (Myrtweit et al. 2005)]
- Study 5: Confident that hold-out tests are unstable for small data sets [contrary to (Menzies et al. 2006)]

However recall that a secondary objective of this work is to establish a better understanding of the second most utilized cost estimation model criterion, PRED. So in addition to the replicated results in the 5 studies, we also found that

- PRED is a reliable estimator, MMRE is not (reliability here is respect to the estimator itself, not as a selection criterion as in Study 4). Reliability is (in our opinion) more important than robustness for cost estimation accuracy research. We have shown several examples in Section 10 that, to draw conclusions with confidence, the SE must be estimated reliably. Our 95% confidence interval studies (Figure 2 - Figure 4) show that PRED is more consistent than MMRE when used to rank model accuracy performance both across models and across different data sets.
- As a model selection criterion PRED outperforms MMRE in selecting the "true" model with simulated data (Table 4 - Table 7). PRED performs remarkably well with respect to other proposed model selection criteria indicated in the original study (Foss et al. 2003) such as RSD, LSD, and SD.
- When using PRED(x) the particular choice of x results in an additional degree of uncertainty in that results may be sensitive to the particular level used. However studies such as Figure 5 indicate that PRED may be relatively robust with respect to the choice of x and that results will only vary at unrealistically low and high values.
- The distributional properties of PRED are more easily analyzed and worked with than MMRE. For example, in deriving equation (6) used later in (8) to estimate the amount of data needed for significance in PRED between two models, we made use of the fact that PRED is binomial distributed and can approximated very well as a normal distribution.
- Judicious stratification may improve SE
- In comparing Figure 11 whose data set has several outliers with Figure 13 whose data set has been "pruned," we observed that MMRE is highly sensitive to outliers. Different subsets of the same data may give significantly different MMRE values, even for very large subsets. This observation has been noted in several other works such as (Kirsopp and Shepperd 2002), but here we have provided what we hope to be a straightforward perspective on this based on real project data.
- MMRE may be useful as an "outlier detector" and we have demonstrated a procedure for using it as such, but pruning outliers does not improve SE.
- Often suggested in previous works, but actually observed here, is that PRED appears to be robust and nearly immune to outliers. That is, PRED values will be nearly the same for any large enough random subset.

We hope that our methods have the potential for increasing confidence in cost estimation research results and cost estimation practice, In particular, resolving contradictory research results where PRED and MMRE are used.

14. REFERENCES

BESTweb (2007) BESTweb – Better Estimation of Software Tasks, http://www.simula.no/~simula/se/bestweb/

Boehm BW (1981) *Software Engineering* Economics. Prentice Hall

Boehm BW, Horowitz E, Madachy R, Reifer D, Clark BK, Steece B, Brown AW, Chulani S, and Abts C (2000) Software Cost Estimation with COCOMO II. Prentice Hall.

Boetticher G, Menzies T, and Ostrand T (2007) The PROMISE Repository of Empirical Software Engineering Data. http://promisedata.org/repository.

Briand LC, Langley T, and Wieczorek I (2000) A replicated assessment and comparison of common software cost modeling techniques. Proceedings of the 22nd International Conference on Software Engineering, Limerick, Ireland, pp. 377–386.

Briand LC, Wieczorek I (2001) Resource Estimation in Software Engineering. Encyclopedia of Software Engineering, Pp. 1160 – 1196, Wiley-Interscience Publishing

Chulani S, Boehm BW and Steece B (1999) Bayesian analysis of empirical software engineering cost models. IEEE Transactions on Software Engineering, vol. 25 n.4, pp. 573-583.

COCOMO81 dataset (2007) http://promisedata.org/repository/#coc81, 12/29/2007

COCOMONASA dataset (2008) http://promisedata.org/repository/#cocomonasa_v1, 01/19/2008

Conte SD, Dunsmore HE, and Shen VY (1986) Software engineering metrics and models. Benjamin-Cummings Publishing.

Desharnais JM (1989) Analyse statistique de la productivitie des projets informatique a partie de la technique des point des function. Masters thesis, Univ. of Montreal.

Desharnais dataset (2007) http://promisedata.org/repository/#desharnais, 12/29/2007

Efron B (1979) Bootstrap methods: Another look at the jackknife. The Annals of Statistics, 7, 1-26

Ferens D and Christensen D (1998) Calibrating software cost models to Department of Defense Database: A review of ten studies. Journal of Parametrics, 18(1):55–74.

Foss T, Myrtveit I, and Stensrud E (2001) MRE and heteroscedasticity. Proc. 12th European Software Control and Metrics Conference (ESCOM 2001), Shaker Publishing BV, The Netherlands, pp. 157-164.

Foss T, Stensrud E, Kitchenham B, Myrtveit I (2003) A simulation study of the Model Evaluation Criterion MMRE. IEEE Transactions on Software Engineering, Vol. 20, No. 11

Gray A, MacDonell S (1999) Software Metrics Data Analysis-Exploring the Relative Performance of Some Commonly Used Modeling Techniques. Empirical Software Engineering.

Hood G (2008) http://www.cse.csiro.au/poptools, 01/19/2008

Jørgensen M (1995) Experience with the accuracy of software maintenance task effort prediction models. IEEE Transactions on Software Engineering, Vol. 21, No. 8

Jørgensen M (2003) How Much Does a Vacation Cost? or What is a Software Cost Estimate?. ACM SIGSOFT Software Engineering Notes, P. 5, Vol. 28, No. 6

Jørgensen M, Shepperd M (2007) A Systematic Review of Software Development Cost Estimation Studies. IEEE Trans. Software Eng., Vol. 33, No. 1

Jorgensen M, (2004) A Review of Studies on Expert Estimation of Software Development Effort. J. Systems and Software, vol. 70, nos. 1-2, pp. 37-60

Kemerer CF (1987) An Empirical Validation of Software Cost Estimation Models. Communications of the ACM.

Kirsopp C, Shepperd M (2002) Making Inferences with Small Numbers of Training Sets. IEE Proceedings - Software, Volume: 149, Issues 5, Pages 123-130

Kitchenham B, Pickard L, MacDonell S, Shepperd M (2001) What accuracy statistics really measure. Proceedings of the IEEE, Vol. 148, No. 3

Kitchenham B, Pfleeger SL, McColl B and Eagan S (2002) An Empirical Study of Maintenance and Development Estimation Accuracy. Journal of Systems and Software.

Kitchenham B, Mendes E, Travassoss G (2007) Cross- vs. Within-Company Cost Estimation Studies: A Systematic Review. *IEEE Transactions on Software Engineering, Vol. 33, No. 5*

R. Kohavi and John GH (1997) Wrappers for feature subset selection. Artificial Intelligence, vol. 97, no. 1-2, pp. 273–324, 1997.

Land CE (1971) Confidence intervals for linear functions of the normal mean and variance. Annals of Mathematical Statistics, 42, 1187-1205

Larsen R, Marx M (1986) An Introduction to Mathematical Statistics and its Applications. Second Edition, Prentice Hall.

Lefley M and Shepperd M (2003) Using genetic programming to improve software effort estimation based on general data sets. In Proceedings of GECCO 2003, volume 2724 of LNCS, pages 2477–2487. Springer-Verlag.

Lokan C (2005) What Should You Optimize When Building an Estimation Model? Proc. 11th IEEE Int'l Software Metrics Symp.

Lum K, Powell J, and Hihn J (2002) Validation of spacecraft cost estimation models for flight and ground systems. In ISPA Conference Proceedings, Software Modeling Track.

Lum K, Hihn J, Menzies T, (2006) Studies in Software Cost Model Behavior: Do We Really Understand Cost Model Performance? Proceedings of the ISPA International Conference, Seattle, WA

C. Mair, Shepperd M, Jørgensen M (2005) An Analysis of Data Sets Used to Train and Validate Cost Prediction Systems. International Conference on Software Engineering, St. Louis, Missouri, USA

Menzies T, Port D, Chen Z, Hihn J, Stukes S (2005) Validation Methods for Calibrating Software Effort Models. Proceedings of the 27th international conference on Software engineering

Menzies T, Chen Z, Hihn J, Lum K (2006) Selecting Best Practices for Effort Estimation. IEEE Transactions on Software Engineering, Vol. 32, No. 11

Miller A (2002) Subset Selection in Regression (second edition). Chapman & Hall.

Miyazaki Y, Takanou A, Nozaki H, Nakagawa N, Okada K (1991) Method to Estimate Parameter Values in Software Prediction Models. Information and Software Technology, 1991

Miyazaki Y, Terakado M, Ozaki K, and Nozaki H (1994) Robust Regression for Developing Software Estimation Models. Journal of Systems and Software, Vol. 27, pp. 3-16.

Moløkken K, Jørgensen M (2003) A Review of Surveys on Software Effort Estimation. International Symposium on Empirical Software Engineering, Rome, Italy

Mooney C, Duval R (1993) Bootstrapping: A Nonparametric Approach to Statistical Inference. Sage Publications; 1. edition.

Myrtveit I and Stensrud E (1999) A Controlled Experiment to Assess the Benefits of Estimating with Analogy and Regression Models. IEEE Trans. Software Eng., vol. 25, no. 4, pp. 510-525.

Myrtweit I, Stensrud E, Shepperd M (2005) Reliability and Validity in Comparative Studies of Software Prediction Models. IEEE Transactions on Software Engineering, Vol. 31, No. 5

NASA93 dataset (2007) http://promisedata.org/repository/#nasa93, 12/29/2007

Payton M, Miller A, Raun W (2000) Testing statistical hypotheses using standard error bars and confidence intervals. Communications in Soil Science and Plant Analysis. 31:547–552.

Payton M, Greenstone M, Schenker N (2003) Overlapping confidence intervals or standard error intervals: What do they mean in terms of statistical significance? Insect Sci. 3: 34. , Published online 2003 October 30. PMCID: PMC524673

R Project (2008) http://www.r-project.org/. Version 2.8.0.

Shepperd M, Schofield C (1997) Estimating software project effort using analogies. IEEE Transactions on Software Engineering, 23(12).

Shepperd M, Kadoda G (2001) Using Simulation to Evaluate Prediction Techniques. Proc. Fifth Int'l Software Metrics Symp.

Shepperd M (2005) Evaluating Software Project Prediction Systems. 11th IEEE International Software Metrics Symposium, Como, Italy

Shepperd M (2007) Software project economics: a roadmap. Future of Software Engineering (FOSE'07)

Srinivasan K and Fisher D (1995) Machine learning approaches to estimating software development effort. IEEE Trans. Soft. Eng., pages 126–137.

Walkerden F, Jeffery R (1999) An Empirical Study of Analogy-Based Software Effort Estimation. Empirical Software Engineering.

Wieczorek I, Ruhe M (2002) How valuable is company-specific data compared to multi-company data for software cost estimation? Proceeding for the Eights IEEE Symposium on Software Metrics (METRICS 02)

Wikipedia (2008) http://en.wikipedia.org/wiki/Vysochanskii-Petunin_inequality, 01/19/2008

Witten IH and Frank E (2005) Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann.

Wittig G and Finnie G (1997) Estimating software development effort with connectionist models. Information and Software Technology, 39(7):469–476.