DUNEDIN NEW ZEALAND

Assessing Prediction Systems

Barbara Kitchenham
Stephen MacDonell
Lesley Pickard
Martin Shepperd

The Information Science Discussion Paper Series

Number 99/14
June 1999
ISSN 1172-6024

University of Otago

Department of Information Science

The Department of Information Science is one of six departments that make up the Division of Commerce at the University of Otago. The department offers courses of study leading to a major in Information Science within the BCom, BA and BSc degrees. In addition to undergraduate teaching, the department is also strongly involved in postgraduate research programmes leading to MCom, MA, MSc and PhD degrees. Research projects in spatial information processing, connectionist-based information systems, software engineering and software development, information engineering and database, software metrics, distributed information systems, multimedia information systems and information systems security are particularly well supported.

The views expressed in this paper are not necessarily those of the department as a whole. The accuracy of the information presented in this paper is the sole responsibility of the authors.

Correspondence

This paper represents work to date and may not necessarily form the basis for the authors' final conclusions relating to this topic. It is likely, however, that the paper will appear in some form in a journal or in conference proceedings in the near future. The authors would be pleased to receive correspondence in connection with any of the issues raised in this paper, or for subsequent publication details. Please write directly to the authors at the address provided below. (Details of final journal/conference publication venues for these papers are also provided on the Department's publications web pages: http://divcom.otago.ac.nz:800/COM/INFOSCI/Publctns/home.htm). Any other correspondence concerning the Series should be sent to the DPS Coordinator.

Department of Information Science
University of Otago
P O Box 56
Dunedin
NEW ZEALAND

Fax: +64 3 479 8311
email: dps@infoscience.otago.ac.nz
www: http://divcom.otago.ac.nz:800/COM/INFOSCI/

Assessing Prediction Systems[1]

Barbara A. Kitchenham
Department of Computer Science
University of Keele
Keele, ST5 5BG UK
+44 1782 583413
barbara.kitchenham@cs.keele.ac.uk

Stephen G. MacDonell
Department of Information Science
University of Otago
P.O. Box 56, Dunedin, New Zealand
+64 3 479 8142
stevemac@infoscience.otago.ac.nz

Lesley M. Pickard
University of Keele
lesley@cs.keele.ac.uk

Martin J. Shepperd
Empirical Software Engineering Research Group
Department of Computing
Bournemouth University, Talbot Campus
Poole, BH12 5BB, UK
+44 1202 595503
mshepper@bmth.ac.uk

ABSTRACT

For some years software engineers have been attempting to develop useful prediction systems to estimate such attributes as the effort to develop a piece of software and the likely number of defects. Typically, prediction systems are proposed and then subjected to empirical evaluation. Claims are then made with regard to the quality of the prediction systems. A wide variety of prediction quality indicators have been suggested in the literature. Unfortunately, we believe that a somewhat confusing state of affairs prevails and that this impedes research progress. This paper aims to provide the research community with a better understanding of the meaning of, and relationship between, these indicators. We critically review twelve different approaches by considering them as descriptors of the residual variable. We demonstrate that the two most popular indicators MMRE and pred(25) are in fact indicators of the spread and shape respectively of prediction accuracy where prediction accuracy is the ratio of estimate to

[1] Version 11 28/5/99

actual (or actual to estimate). Next we highlight the impact of the choice of indicator by comparing three prediction systems derived using four different simulated datasets. We demonstrate that the results of such a comparison depend upon the choice of indicator, the analysis technique, and the nature of the dataset used to derive the predictive model. We conclude that prediction systems cannot be characterised by a single summary statistic. We suggest that we need indicators of the central tendency and spread of accuracy as well as indicators of shape and bias. For this reason, boxplots of relative error or residuals are useful alternatives to simple summary metrics.

Keywords
Prediction systems, estimation, empirical analysis, metrics, goodness-of-fit statististics.

1. The Problem

A major challenge for managers of software projects is to be able to make accurate predictions. For example, how long will a project take, how much effort will it require and how many defects will a particular component contain? To answer this type of question has been a major goal of workers in the field of software metrics over the past 25 years. In general, the approach adopted has been to collect various measures that can then be used to construct a prediction system. For example, one might count the number of function points or perhaps count the number of reports that are to be generated.

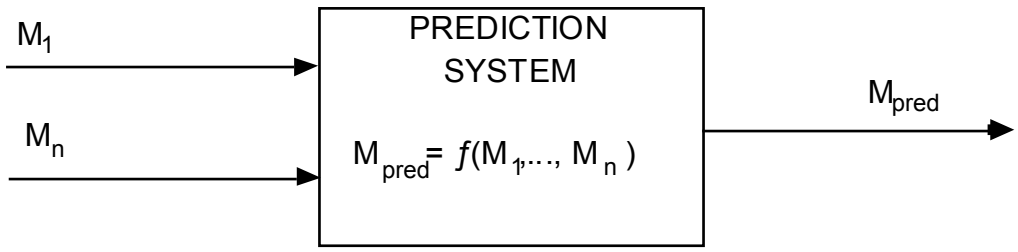


Figure 1: The Structure of Prediction Systems

Figure 1 illustrates the basic structure of a prediction system. Prediction systems must have at least three components. These are (i) a vector of one or more input measures ($M_{1...n}$); (ii) the output measure (M_{pred}) which is the quantity being predicted; and (iii) the prediction system itself which will be a system of equations that enable us to derive a value for M_{pred} from $M_{1...n}$.

The critical issue for software metrics researchers is how to formulate the prediction system. In addition, if there are alternative methods of formulating a prediction system, we would like to have some criterion for selecting the most appropriate system.

An example of a prediction system would be to use function points to predict development effort for software projects in a given environment. Here we might use a technique such as linear regression analysis in order to determine an equation to link the input, that is function points (FP), to the predicted variable, that is effort. A simple linear regression equation will require two further inputs or coefficients β_0 and β_1 such that:

$$\text{effort} = \beta_0 + \beta_1 FP$$

By collecting local data it may be possible to find values for β_0 and β_1 to give useful predictions of effort. Note that more complex models may require a greater number of coefficients. Examples of work following this type of approach include the MERMAID projects (Kok *et al.* 1990) which have reported promising results. Another approach would be to use a predefined generic model such as COCOMO II (Chulani, Clark and Boehm 1998) and use data from previous projects to calibrate the model to our particular circumstances. Prediction systems do not necessarily involve mathematical relationships between inputs and outputs. Machine learning algorithms and expert systems can be used to develop rule-based prediction systems. Pattern-matching and analogy-based approaches can lead to prediction systems that choose the most similar past project and base estimates on the actual values of that project. These methods of generating a prediction system are all data intensive. This can be contrasted with human-intensive estimating systems which do not depend on the availability of a systematic dataset of past projects or an explicit prediction equation.

The important question is of course, how accurate is a prediction system? For a data intensive prediction system, we might want to know if it could be improved if we changed the values of β_0 and β_1? Should we consider using a non-linear model? How do machine learning approaches compare with algorithmic models? All these questions require some indication of predictive performance in order to be able to obtain meaningful answers. Although this question would seem a straightforward one, there are in fact a large number of different prediction quality indicators that that have been used in the literature. Moreover, it would appear that they are not all assessing the same thing. We believe that the lack of understanding of what different prediction quality indicators are in fact assessing is hindering progress in this important branch of software engineering. We suggest that there are at least three dimensions to the quality of a prediction system. First, there is the central tendency of the errors or residuals, in other words, what would be a typical error? Second, there is the spread or variance. Here we are more interested in the range of values, perhaps what is the worst case? In each case indicators will be more or less robust to the presence of skewed distributions of error values. Third, there will be the shape of the distribution of the errors. For example, a skewed distribution might be important if one were risk averse.

The next section examines a range of indicators of prediction quality that have been discussed in the literature. We compare these indicators by considering their underlying meaning defined in terms of describing the observed residual values of the prediction system. This is followed by a simulation study in which we illustrate how

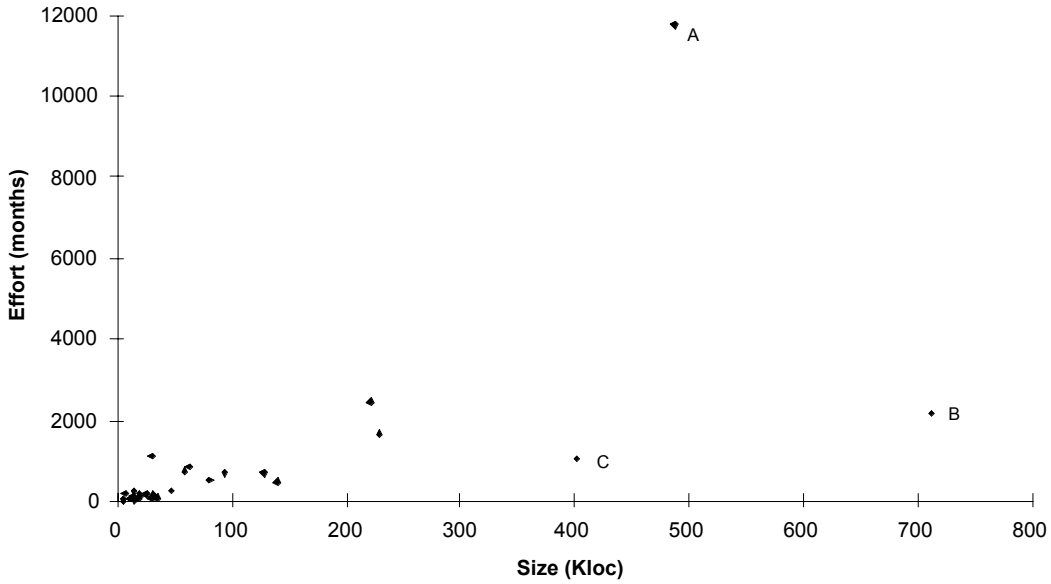
the different indicators give conflicting results and how the choice of indicator depends upon one's objectives in using the prediction system as well as the choice of analysis technique and the nature of the dataset used to derive the prediction system. We conclude by offering some guidelines as to the choice of appropriate indicators.

2. Prediction Quality Indicators

Many different prediction quality indicators have been proposed over the years by researchers when attempting to evaluate or compare prediction systems. This section describes twelve different indicators that have either been widely used or embody significantly different notions of prediction quality.

In this section, we illustrate the behaviour of the prediction quality indicators using the dataset collected by Belady and Lehman in 1979 and reported by Conte *et al.*(1986). The relationship between effort and size for this dataset is shown in Figure 2. Although it is an old dataset, it illustrates a number of points about software datasets. They are usually positively skewed i.e. have a large number of small datapoints and relatively few large data points. Furthermore, they usually include atypical data points, for example the points labelled, A, B, and C. Atypical points may be an effect of skewness or the effect of a non-stable variance for the relationship between effort and size. For whatever reason the points labelled A, B and C are likely to be high influence points with respect to any model fitted to the dataset.

Figure 2. The relationship between size and effort reported by Belady and Lehman



In order to assess the quality of predictions, it is necessary to compare the predictions produced by a prediction system with actual values. Ideally, the evaluation of prediction quality should be independent of the construction of the prediction system. This means we should take a random sample of past projects and use that data to

construct our prediction system. Then we should take a second random sample, and apply the prediction system to the second sample. In practice, we seldom have a sufficiently large population of past projects to use this approach. There are alternative strategies that can be used:

1. Leave one out: Leave one data point out of the dataset, construct the prediction system using the remaining projects, and predict the value of the omitted project. Repeat this process for each project in the dataset. This approach, sometimes known as jackknifing, is useful if you have a small dataset (e.g. 10-30 projects). Shepperd and Schofield 1997 is an example of this type of validation.
2. Training and testing subsamples. Divide the dataset into two random subsamples. Use one of the subsamples to construct the prediction system and use the prediction system to obtain predictions for the other subsample. It is usual for the training subsample to be based on 2/3 of the dataset. This is useful when you have a large dataset (e.g. >100 projects). MacDonell, *et al.* 1997 is an example of this type of validation.
3. *x*-fold cross-validation: This is a mixture of the two approaches. Create a number of different training and testing subsets by dividing the dataset into *x* (e.g. 6) mutually exclusive subsamples of approximately the same size. Each of these subsamples is regarded as a testing subsample for the training subsample made up of the other projects. This give *x* different testing-learning subsets. Generate the prediction system on each learning subsample and apply it to the corresponding testing subsample. This is useful for a moderate dataset (e.g. 50-100 projects).

It is also common to use the full dataset to generate the prediction system and to use the prediction system to predict the value of each of point in the dataset. If the predictions are based on the full dataset, we are assessing goodness of fit rather than prediction quality although the quality indicators are constructed in exactly the same way. Kemerer's validation of four project cost models is such an example (Kemerer 1987).

In this paper we are interested in the behaviour of the prediction quality indicators rather than obtaining the actual predictive quality of any particular prediction system. Consequently, we have simply used the full dataset to generate the predictions, and hence calculate the values of the quality indicators.

2.1 Coefficient of Determination - R^2

The coefficient of determination is defined as:

$$R^2 = 1 - \left(\frac{\sum_{i=1}^{i=n} (x_i - \hat{x}_i)^2}{\sum_{i=1}^{i=n} (x_i - \bar{x})^2} \right)$$

where there are n predictions, x_i is the ith observed value, $\hat{x}_i$ (x-hat) is the ith predicted value and $\bar{x}$ (x-bar) is the mean of the n observed values. Note that if $\hat{x}_i$ is constructed using least squares regression with one independent variable (i.e. one input variable), R^2 is the square of the Pearson correlation coefficient between the independent variable and x_i. In addition, it is equal to the square of the Pearson correlation between $\hat{x}_i$ and x_i. The coefficient of determination is only defined for ordinary least squares multivariate regression.

The R^2 value attained for a predictive model provides an indication of the degree to which the model (i.e. the regression equation) accounts for the variation in the value of the estimated commodity known as the dependent variable. It is therefore often interpreted as the explanatory capability of the estimation model - the higher the R^2 value, the more effectively the model accounts for the change in estimate value. Interpreted in another way, a low R^2 value indicates that the model may be unsatisfactory, in that there are factors contributing to the value being estimated that the model has failed to capture. This may assist the model builder in reformulating the predictive hypothesis to include other characteristics. In terms of the *accuracy* of prediction, the R^2 indicator is inadequate, as it provides no information concerning the extent of correspondence between actual and predicted values.

The R^2 value should not be used unless each independent variable included in the regression model contributes significantly to the equation i.e. the value of every multiplicative parameter β_i included in the model is significantly different from zero. An R^2 value that is significantly different from zero should be regarded as a general minimum criterion for any predictive system. If it is not significantly different from zero, *there is no prediction system at all*, irrespective of the value of other criteria!

The R^2 statistic is particularly vulnerable to high influence points (i.e. data points radically different from the majority of points in a dataset). The effect of high influence points on the value of R^2 can be demonstrated by performing an ordinary least squares regression on five different variants of the Belady-Lehman dataset as shown in Table 1.

Table 1 The value of R^2 for different variants of the Belady-Lehman dataset

Dataset variant	R^2
All data points	0.4056
All data points except A	0.6056
All data points except B	0.6149
All data points except C	0.4528
All data points except A, B, and C	0.7512

If least squares regression has not been used to generate a predictive model, it is possible to calculate the square of the correlation between $\hat{x}_i$ and x_i, but the value obtained is not the same as the coefficient of determination.

2.2 Adjusted Coefficient of Determination - R^2 (Adjusted)

In general, the adjusted R^2 value is used in preference to the raw value, as this takes into account the number of parameters included in the model so it is possible to compare competing models with different numbers of independent variables. When using multiple regression the introduction of new independent variables can only increase the R^2 value, so even the addition of a random variable might increase the R^2 – the adjusted R^2 value compensates for this tendency. Clearly, if the inclusion of variables that do not contribute significantly to a model is avoided, the use of the adjusted R^2 is not so important. As with the R^2 indicator, the adjusted R^2 provides no information value concerning the extent of correspondence between actual and predicted values.

2.3 Total Error

Total error is defined as:

$$\sum_{i=1}^{i=n}(x_i - \hat{x}_i)$$

where there are n predictions, x is the true value and x-hat is the predicted value. In statistical terminology, $x_i - \hat{x}_i$ is referred to as the residual error (usually referred to as e_i) and the total error is the sum of the residuals.

If $\hat{x}_i$ is an unbiased estimator of x_i, then $x_i - \hat{x}_i$ has an expected value of 0. Furthermore if $\hat{x}_i$ and x_i are both Normally distributed, with variances $\sigma_{\hat{x}}^2$ and σ_x^2 respectively, then the variance of $x_i - \hat{x}_i$ is obtained from the standard formula for the variance of the difference between two Normal variables as:

$$\sigma_e^2 = \sigma_{\hat{x}}^2 + \sigma_x^2 + \mathrm{cov}(\sigma_{\hat{x}}, \sigma_x)$$
$$\sigma_e^2 = \sigma_{\hat{x}}^2 + \sigma_x^2 + r_{\hat{x},x}\sigma_{\hat{x}}, \sigma_x$$

Where $\mathrm{cov}(x_i, \hat{x}_i)$ is the covariance between x and $\hat{x}_i$, and $r_{\hat{x},x}$ is the Pearson correlation coefficient between x_i and $\hat{x}_i$.

Whilst particular estimates may be too high or too low, overall predictive performance may achieve a level close to zero in terms of total error. With a long-term view of predictive performance, individual variation may be tolerated as long as global error falls within a threshold value. This global performance can be measured using the total error indicator, and an optimal model chosen based upon its value. In terms of organisational liquidity there may also be a desire that, whatever the individual estimate error, the global prediction must not be less than the actual value, whilst still attempting to reach a zero balance.

2.4 Total Relative Error

Total Relative Error is defined as:

$$\frac{\sum_{i=1}^{i=n}(x_i - \hat{x}_i)}{\sum_{i=1}^{i=n}\hat{x}_i}$$

The total relative error is used to overcome one major drawback of total error as an indicator, which is its lack of scope of the values being estimated. For example, it may be acceptable to have a total error of 100 person-days over a set of projects totalling 10 person-years, but this may be quite unacceptable for a set of projects totally just one person-year.

Of course, relative error also conceals problems. For example, an underestimate of 10% on a set of projects totalling 1 years effort may be manageable, whereas an underestimate of 10% on a set of projects totalling 10 person-years may not.

2.5 Average Error

In addition to the total error for a set of projects, it is also useful to consider the average error per project (i.e. the average residual). If the residuals are not Normally distributed, it is usually preferable to use the median residual rather than the mean residual.

Total error, mean error and median error are affected by the technique used to construct the prediction system. Table 2 shows the total error, the mean error and median error for the Belady-Lehman dataset based on four different analysis techniques:

1. ordinary least squares regression;
2. robust regression (which is based on ignoring severe outliers and giving low weights to moderate outliers);
3. median regression (which is based on minimising the deviation from the median);
4. ordinary least squares applied to the dataset with points A, B and C removed (which is equivalent to deriving a prediction system restricted projects less than 300 KLOC in size).

Table 2 was derived by applying a particular analysis technique to the dataset deriving a model relating effort to size, and then using the model and the actual size value to "predict" the value of effort.

Table 2 Total, average and median error for prediction systems derived using different analysis methods

Regression Analysis Technique	Total error	Mean error	Median error
Least squares	3.663×10^{-3}	-1.11×10^{-5}	-80.11
Robust	8893.52	269.5	-6.63
Median	10705.91	324.4	0
Least Squares on restricted dataset	-2.03×10^{-6}	-6.1×10^{-5}	-36.46

Least squares regression minimises the sum of squares of the residuals which forces the total error and average error to zero. However, if the residuals are biased or some of the residuals are outliers, the median error will not be close to zero. Median regression minimises the sum of the absolute residuals. This forces the median residual to zero, but if the residuals are not Normally distributed, the total error and average error of the residuals may not be close to zero. Robust regression is based on reducing the influence of atypical data points. It will therefore behave similarly to a median regression.

2.6 Average Relative Error

Just as there is an argument that total error conceals the context of the estimate, there is a similar problem with average error. Relative error can be accounted for using the average relative error indicator:

$$\frac{1}{n}\sum_{i=1}^{i=n}\left(\frac{x_i-\hat{x}_i}{\hat{x}_i}\right)$$

If the relative error per project is not Normally distributed it may be preferable to use the median relative error rather than the mean relative error.

2.7 Mean Magnitude of Relative Error - MMRE

The Mean Magnitude Relative Error (MMRE) prediction quality indicator is probably the most widely used indicator in recent years, particularly when assessing the performance of software effort estimation models. The MMRE is defined by Conte *et al*. (1986) as:

$$\frac{1}{n}\sum_{i=1}^{i=n}\left(\frac{\left|x_i-\hat{x}_i\right|}{x_i}\right)$$

It is, however, not particularly meaningful for assessing predictions (as opposed to providing a goodness of fit statistic). If the aim is to generate an estimate of the effort for a new project, upper and lower bounds about the estimate are normally required, in order to present a range of values likely to contain the actual value. In other words

interest is in the deviation relative to the *estimate* not relative to the *actual*. This formulation of the MMRE can be referred to as the EMMRE (Estimation MMRE).

The MMRE differs from relative error since the absolute value of the difference between actual and predicted is used. This has the effect of preventing under- and overestimates from canceling each other out. On the other hand, it obscures the fact as to whether a prediction system has any bias, the knowledge of which could be used to make corrections to any prediction generated.

In order to understand what the MMRE measures, consider a random variable x distributed Normally with mean μ and variance σ^2. It has been demonstrated by Iglewicz (1983) that for a sample of size n where x-bar is the average of the n observations:

$$d_n = \frac{1}{n}\sum |x_i - \bar{x}| \rightarrow \sigma\sqrt{\frac{\pi}{2}} \text{ as } n \rightarrow \infty$$

If we rewrite the MMRE as follows:

$$\frac{1}{n}\sum_{i=1}^{i=n}\left|\frac{\hat{x}_i}{x_i} - 1\right| = \frac{1}{n}\sum_{i=1}^{i=n}|z_i - 1|$$

it is clear that if $\hat{x}_i$ is an unbiased estimator of x_i, the expected value of $z_i = \frac{\hat{x}_i}{x_i}$ is 1 and if z_i is distributed Normally with mean 1 and variance σ_z, the MMRE tends to the value $\sigma_z\sqrt{\frac{\pi}{2}}$. This demonstrates that the MMRE is an estimate of the *spread* of the variable *z* that will not be so vulnerable to large outliers as the root mean square estimate. Since MMRE is a measure of spread it is incorrect to refer to it as a measure of prediction accuracy. The variable *z* is a better indicator of prediction accuracy since it has a defined optimum value (i.e. 1) which indicates clearly whether or not the prediction system under- or overestimates.

Using the above argument, the EMMRE will be an estimate of spread of the variable $q = \frac{1}{z}$.

This discussion indicates that the quality of a prediction system can be reported in terms of the average or median value of the prediction accuracy variables *z* or *q*, and the MMRE or EMMRE should be used to assess the variability of *z* and *q* respectively.

Table 3 shows these statistics for the Belady and Lehman dataset for each of the four analysis techniques.

Table 3. Relative error and spread statistics for the Belady-Lehman dataset

Regression Analysis Technique	MMRE	Mean z	Median z	EMMRE	Mean q	Median q
Least squares	1.4179	2.2227	1.8578	0.6499	0.8268	0.5383
Robust	0.5815	1.1806	1.0906	0.8508	1.467	0.9170
Median	0.8618	1.0954	1.0	0.9205	1.5896	1.0
Least Squares on restricted dataset	0.9172	1.6603	1.5070	0.6400	1.025	0.6636

As would be expected, if we use the median z or median q statistics to assess which prediction system is best we would assume that the prediction system derived from the median regression was best with the system obtained using robust regression a close second.

Similarly it is not surprising that if we were to use the mean q statistic, the least squares based models appear best. However, it is surprising that the least squares based prediction systems appear worst when judged on the mean z statistic. All the relative error statistics make it clear that least squares models overestimate. Thus, the large mean z value is probably due to one or more severe overestimates. This can be confirmed by inspecting boxplots of the z and q variables.

Pickard et al. (1999) recommend inspecting boxplots of the residuals to compare models. This gives a good indication of the distribution of the residuals and can help explain the behaviour of the summary statistics. Similarly a boxplot of the accuracy values can clarify the various prediction quality statistics. Figure 3 shows a boxplot of the q values and Figure 4 shows a boxplot of the z values for each prediction system. Note, the figures do not include a boxplot of accuracy from the prediction system based on the restricted dataset because, being based on a different number of data points, it would not be comparable with the boxplots from the other prediction systems. Figure 4 confirms that there is one very large z value that will increase the mean value.

Table 3 makes it clear that all the prediction systems overestimate (with the exception of the system based on median regression). This is why the EMMRE is less than the MMRE. This bias is also clear from the boxplots in Figures 3 and 4.

2.8 Pred (n)

Another widely used prediction quality indicator is Pred (n), which is simply the percentage of estimates that are within n% of the actual value. Typically n is set to 25 so the indicator reveals what proportion of estimates are within a tolerance of 25%. Clearly, Pred (n) is insensitive to the degree of inaccuracy of estimates outside the specified tolerance level. For example, a Pred (25) indicator will not distinguish between a prediction system for which predictions deviate by 26% and one for which predictions deviate by 260%.

As with MMRE, it is preferable to formulate Pred (n) for estimating by considering the percentage of actuals within n% of the estimate.

Based on the discussion of EMMRE above, it is clear that when the prediction accuracy (i.e. actual/estimate) is approximately Normal, Pred (n) has (asymptotically) a functional relationship with EMMRE. If $q_i = \frac{x_i}{\hat{x}_i}$ is distributed normally with mean $\mu = 1$ and variance σ_q^2, then the proportion of actuals within n% of the estimate depends on the size of the variance compared with a Standard Normal variate which has a variance equal to 1. The EMMRE provides an estimate of the variance of *q*. Recalling that the mean of *q* is 1, the proportion of actuals within n% of the estimate can be calculated using the tables of the standard normal variate and the ratio:

$$\frac{n}{100} \Big/ \sigma_q$$

For example, if n=25% and EMMRE=0.5, an estimate of σ_q is $0.5 / \sqrt{\frac{\pi}{2}}$ which is approximately 0.5/1.2533=0.3989. The proportion of actuals within 25% of the estimate corresponds to the number of actuals in the range 0.75 to 1.25. This depends on how the variance of q compares with the proportion n/100. In this case an upper and lower bound of 0.25 around the mean, a standard deviation of 0.3989 corresponds to plus or minus .25/.3989=0.627 standard deviations about the mean. From tables of the standard normal deviate, this range corresponds to a probability of 0.46. Thus if a sample comprises 100 estimate-actual pairs, 46 of the actuals should be within 25% of the estimate.

Alternatively working backwards, in order to achieve a Pred (25)≥75%, the probability of a value of *q* between the values .75 and 1.25 corresponds to 0.75 would be needed. Thus, *q* would need to have a standard deviation of 0.216, and EMMRE would need to have an asymptotic value of 0.27.

However, Pred(25) is *not* a measure of the spread of q. To understand what it measures, consider what happens it a distribution is more peaked than a Normal distribution. A sample from a more peaked distribution would have more values within 25% of the mean than normal. Similarly a sample from a flatter distribution would have less values within 25% of the distribution. Thus, Pred(25) is related to the *shape* of the distribution q. Shape has two dimensions: *skewness* which describes whether or not the distribution is symmetrical about a central value and *kurtosis* which describes the extent to which the distribution peaks around its central value. Pred(25) is therefore a measure of kurtosis.

2.9 Balanced MMRE

$$\frac{1}{n}\sum_{i=1}^{i=n}\frac{\left|x_i-\hat{x}_i\right|}{\min(x_i,\hat{x}_i)}$$

A Balanced MMRE was suggested by Miyazaki *et al*. (1991, 1994) to 'equally' weight over- and underestimates. A weakness of MMRE is that it is asymmetric in that underestimates cannot exceed 100% whereas overestimates are unbounded. This is particularly perverse inasmuch as this will tend to lead the estimator to choose a prediction system that under-estimates. The Balanced MMRE indicator overcomes this problem by dividing the absolute difference of actual and predicted by the *lesser* of the actual and predicted. For this reason the Balanced MMRE will tend to give a higher indication of error than the straightforward MMRE. Thus can be seen by comparing the balanced MMRE shown in Table 4 with the MMRE and EMMRE shown in Table 3.

Table 4. Other spread statistics for the Belady-Lehman dataset

Regression Analysis Technique	Pred (25)	Balanced MMRE	Median MMRE	Median EMMRE	Root Mean Square Error	Relative RMS
Least squares	0.12	1.5586	0.858	0.563	1591.07	0.0501
Robust	0.33	1.0419	0.484	0.504	1641.44	0.0509
Median	0.30	1.0720	0.537	0.464	1669.48	0.0513
Least Squares on restricted dataset	0.20	1.0755	0.562	0.429	274.05	0.0208

The balanced MMRE has not been generally used, but Miyazaki *et al*. (1994) have gone on to define other variants such as the inverted balanced MMRE and logarithmic relative error. We are unaware of other researchers adopting these indicators.

2.10 Median MRE

A Median MRE indicator was used by Jorgensen (1995) instead of MMRE in order to avoid the influence of outlier MRE values. It is intended that a Median MRE is more representative of a typical estimate error than a mean. In a situation where there are a few very large estimation errors and many smaller errors, that is, the error distribution is positively skewed, the Median MRE will be less than the MMRE value. For a negatively skewed distribution the reverse will be true. The Median MRE is useful when the focus of concern is upon typical estimates rather than extreme cases. In the light of the discussion of MMRE, it will be clear, that the median MRE is going to be a more robust estimate of spread than the MMRE since it will be less susceptible to outliers.

From Jorgensen's analysis it appears that positive skewing is more commonplace and in all cases the MMRE exceeds the Median MRE. This is the case for the Belay-Lehman data as can be seen by comparing the MMRE and median MMRE in Tables 3 and 4.

An implication of Jorgensen's result is that if the estimator is concerned with putting sensible bounds on an estimate, then the variability of underestimates should be considered separately from the variability of overestimates, allowing non-symmetric bounds about an estimate.

2.11 Mean Square Error

Conte et al. define the mean square error as:

$$MSE = \frac{1}{n}\sum_{i=1}^{i=n}(x_i - \hat{x}_i)^2$$

The argument for using the mean square error is that if an estimator is risk averse it penalises large deviations more than small deviations since it is based upon the mean of the sum of the squares of the residuals. It should be noted that if you have used least squares regression, the mean square error should be adjusted according to the number of independent variables (i.e. you would divide the sum of squares of the residuals by n-p-1 rather than n, where p is the number of independent variables). This method of calculation would give an unbiased estimate of the variance of the residuals. The formula given above is useful for comparing prediction systems derived using different analysis techniques.

An alternative to the Mean Square Error is the Root Mean Square Error which is the square root of the MSE.

The mean square error suffers from the problem that it can be difficult to compare values derived from different data sets although it is clearly possible to compare the use of alternative models on the same data set.

2.12 Relative RMS

Conte *et al.* (1986) extend the mean square indicator to give the relative root mean square error (Relative RMS) to represent the relative mean value of the error minimised by the regression. This is defined as:

$$RMS = \sqrt{\frac{1}{n}\sum_{i=1}^{i=n}\left(x_i - \hat{x}_i\right)^2}$$

$$\overline{RMS} = \frac{RMS}{\frac{1}{n}\sum_{i=1}^{i=n} x_i}$$

Since the relative root mean square error divides the square root of the mean square error by the average value of the dependent variable (which is obviously the same value for all models fitted to the same dataset), it is directly correlated with the mean

square error. That is, the rank order of the root mean square for different prediction systems must be exactly the same as the rank order of the means square error. This can be confirmed by inspection of Table 4.

Conte *et al*. state that the relative RMS is only appropriate when using regression analysis to derive a prediction system. However, it is not clear why this is so, nor what the asymptotic properties of the statistic are.

2.13 Other Indicators

Other indicators have been proposed such as weighted mean of quartiles of mean relative errors. Prediction quality indictors related to quartiles can be related to shape or spread. For example, consider the variable $q = \frac{x}{\bar{x}}$. If q is distributed Normally with mean 1 and variance σ_q^2 , and a boxplot is constructed from a set of n variables q_i, the box length (i.e the difference between the 75 percentile value and the 25 percentile value) will tend to $1.347 \times \sigma_q$. However, if the distribution is more peaked than a Normal distribution the interquartile range will be smaller, if the distribution is flatter the interquartile range will be wider. Thus, the interquartile range would be a measure of shape. However, none of these indicators have been used to any significant degree. For this reason they have been excluded from this analysis. For further details the reader is referred to Lo and Gao (1997).

From this section we see that many different accuracy indicators have been proposed and that a naïve interpretation could lead to the conclusion that many indicators conflict with one another. In other words indictor A favours prediction system 1 over prediction system 2 whilst indictor B favours prediction system 2 over 1. We have argued that it is more helpful to regard the accuracy indicators as statistics describing different properties of the residuals, or error terms. There are four such properties, namely central tendency, variance, kurtosis and skew. Differences in accuracy indicators can therefore easily be accommodated on the grounds that they are measuring different properties of the residual variable. Which properties are most important depend upon one's objectives. This will be further explored in Section 4.

3. A Simulation Study

The discussion in Section 2 suggests that the behaviour of prediction quality indicators is strongly influenced both by the analysis technique and by the characteristics of the dataset used to derive the prediction system. In order to confirm this finding, we undertook a simulation study.

We simulated four different datasets. In each case, the dataset comprised 31 data points each of which was a vector of two variables: one independent variable and one dependent variable. We chose a dataset of 31 points to represent a moderate size

dataset and an odd number of points to simplify calculation of medians. The dependent variable was related to the independent variable by a functional model including an error term. The four datasets each had different characteristics:

1. A dataset with Normally distributed independent variables and a Normal error term (i.e. a Normal dataset).
2. A dataset with a Gamma distributed independent variables and a Normal error term, leading to a skewed dependent variable (i.e. a Gamma dataset).
3. A Normal dataset with 5% severe outliers (which were created by selecting 2 data points at random and multiplying the dependent value by 10).
4. A Gamma dataset with 5% severe outliers.

Note. This procedure does not guarantee non-negative predictions, so any simulated errors leading to negative predictions were truncated. This makes the "Normal" errors actually truncated Normal error. Each simulated dataset was analysed using three different analysis techniques: least squares multivariate regression, robust regression and median regression.

Table 5 Dataset A

Prediction Quality Indicator	Technique		
	T1	T2	T3
Total Error	166166.5	-0.00128	174496
Total Relative Error	0.327	-1.90e-09	0.350
Average Error	5360.211	-0.0000413	5328.902
Median Error	-0.0000529	-5360.211	268.691
Average Relative Error	0.607	0.0071	0.790
Median Relative Error	-3.23.e-09	-0.273	0.0167
Average Accuracy (actual/estimate)	1.607	1.008	1.790
Median Accuracy (actual/estimate)	1.000	0.727	1.017
MMRE	34.987	67.557	29.257
EMMRE	1.020	0.735	1.171
Pred (25) -% estimates	52%	29%	42%
Pred (25) - % actuals	52%	39%	45%
Median MRE	0.250	0.563	0.316
Median MMRE	0.250	0.360	0.341
Balanced MMRE	35.651	67.860	30.062
MSE	6.40e08	2.87e07	6.48e08
RRMS	1.164	1.136	1.172

We were interested in confirming the extent to which the prediction quality indicators were indicators of the analysis technique and the dataset type. In order to avoid post-hoc rationalisation we organised our simulation as a small blind experiment. One of the authors (Pickard) simulated the datasets and then analysed them using the different analysis techniques. She produced the prediction quality statistics shown in Tables 5 to 8. These tables concealed the identity of the analysis technique and the dataset. One of the other authors (Kitchenham) was then asked to identify the datasets and the techniques from the tables. Readers might like to attempt the task for themselves before reading on.

Table 6 Dataset B

Prediction Quality Indicator	Technique		
	T1	T2	T3
Total Error	52541.29	0.00111	69519.56
Total Relative Error	0.143	2.65e-09	0.199
Average Error	1694.88	0.0000359	2242.567
Median Error	0.0000157	-1533.828	578.100
Average Relative Error	0.178	-0.0210	0.197
Median Relative Error	1.22e-09	-0.144	0.056
Average Accuracy	1.178	0.979	1.197
Median Accuracy	1.000	0.856	1.056
MMRE	1.474	1.967	1.475
EMMRE	0.521	0.481	0.556
Pred (25) -% estimates	52	42	55
Pred (25)- % actuals	55	52	52
Median MRE	0.205	0.267	0.262
Median MMRE	0.215	0.349	0.243
Balanced MMRE	1.704	2.129	1.717
MSE	2.05e08	2.01e08	2.06e08
RRMS	1.057	1.048	1.062

Table 7 Dataset C

Prediction Quality Indicator	Technique		
	T1	T2	T3
Total Error	-0.0135	0.000211	-6579.908
Total Relative Error	-4.04e-08	6.30e-10	-0.0193
Average Error	-0.0000435	6.79e-6	0.114
Median Error	-0.0000539	104.121	-0.00467
Average Relative Error	-0.0174	0.113	-0.179
Median Relative Error	-1.21e-07	0.0170	-0.0193
Average Accuracy	0.983	1.049	1.050
Median Accuracy	1.000	1.017	0.995
MMRE	1.410	1.436	1.463
EMMRE	0.331	0.459	0.473
Pred (25) -% estimates	45%	52%	52%
Pred (25)- % actuals	45%	45%	52%
Median MRE	0.220	0.222	0.241
Median EMRE	0.254	0.264	0.250
Balanced MMRE	1.459	1.569	1.613
MSE	2.03e07	2.03e07	2.03e07
RRMS	0.418	0.418	0.418

Table 8 Dataset D

Prediction Quality Indicator	Technique		
	T1	T2	T3
Total Error	35718.43	0.0000932	5013.392
Total Relative Error	0.0757	1.84e-10	0.0098
Average Error	1152.2	3.01e-06	161.722
Median Error	0.000148	-621.445	-415.347
Average Relative Error	0.110	0.0134	0.0367
Median Relative Error	2.70e-08	-0.056	-0.057
Average Accuracy	1.110	1.013	1.037
Median Accuracy	1.000	0.947	0.943
MMRE	27.804	31.001	29.923
EMMRE	0.473	0.431	0.443
Pred (25) -% estimates	42%	42	42
Pred (25)- % actuals	48%	42	42
Median MRE	0.333	0.275	0.300
Median EMRE	0.251	0.326	0.303
Balanced MMRE	27.945	31.102	30.035
MSE	2.25e07	2.04e07	2.04e07
RRMS	0.290	0.276	0.276

Kitchenham was able to identify correctly the techniques and correctly identified the Gamma with outliers dataset and Normal without outliers dataset. She misidentified the Normal with outliers dataset and the Gamma without outliers dataset. The key to Tables 5 to 8 is as follows:

- Technique 1 is median regression.
- Technique 2 is ordinary least squares regression.
- Technique 3 is robust regression.
- Dataset A is Gamma with outliers.
- Dataset B is Normal with outliers.
- Dataset C is Normal without outliers
- Dataset D is Gamma without outliers.

The techniques are easy to distinguish because ordinary least squares always results in a total error very close to zero, and median regression always results in a median average accuracy of 1. It is more difficult to distinguish the type of dataset from the prediction indicators. However, the gamma distribution can be detected by order of magnitude differences between the MMRE and the EMMRE. The effect of outliers can be detected by a relatively poor median accuracy and Pred(25) values for prediction systems derived using ordinary least squares.

Through this simple experiment we have shown that the by use of the *entire* set of indicators it is possible to differentiate between prediction systems even when done on a blind basis. The relationship between the underlying nature of the dataset and the indicators is, however, more complex.

4. Choosing Prediction Quality Indicators

Conte *et al*. (1986) provide a useful but less comprehensive comparison of predictive model evaluation criteria, examining the R^2 value along with variations of the MMRE and RRMS measures over a set of four predictive models. While certainly of some general value, the responsibility for making use of the comparison is then effectively passed to the reader. Conte *et al.* make following comments: "It is unfortunate that these criteria... are often not in agreement... in the sense that we cannot say which model is best without making a subjective judgement on the relative importance of the evaluation criteria." (p. 175); "It is still the researcher's responsibility to decide which model is best by weighing the objective scores *subjectively* when they do not provide consistent results." (p. 166). They go on to suggest that the pair of measures MMRE and Pred (25) seems the most suitable, in the absence of any generally accepted standard.

In this paper we have shown that the MMRE and Pred(25) statistics measure different attributes of the distribution of the prediction accuracy variable z=estimate/actual, so it is not surprising that they are not always in agreement. In addition, since the are respectively measures of spread and kurtosis, it is also necessary to consider a measure of the central tendency, and a measure of the skewness, of the distribution of z. We have also suggested that it is preferable to use the variable q=actual/estimate rather than z with appropriate adjustments to the mean magnitude relative error and the Pred(25) statistics.

Theoretically, we could base a comparison of different models on the statistic that best fitted our prediction objectives. If we wanted unbiased estimates we might prefer the prediction system for which the mean accuracy was closest to 1. If we wanted a prediction system that produced the most stable estimates, we would select the prediction system with the smallest MMRE. If we wanted a prediction system that produced that largest number of very close estimates, we would choose the prediction system with the largest Pred(25) value. However, we have shown that the analysis technique and the characteristics used to generate a prediction system will affect the value of prediction quality statistics. Thus, if we have a defined prediction objective we may need to select the method of generating the prediction system rather than the quality statistic.

Other issues affect our choice of central tendency and spread statistics. If we wanted a prediction system that was unbiased or stable for the most typical estimating situations we might prefer the median accuracy and the median MRE respectively. Basing quality indicators on means corresponds to a risk averse strategy, since it would select the prediction system that did best under worst conditions. Basing quality indicators on medians corresponds to a risk seeking strategy, since it selects prediction system that copes best under normal circumstances.

In addition, our prediction objectives are influenced by our role. For example, an estimator might be happy to characterise the "best estimate" in terms of accuracy

(actual/estimate). A project manager might be more concerned about estimate error (actual-estimate). Error indicates the extent to which a project is profitable or not and the extent to which deviations from plans can be accommodated. Thus a project manager might be more interested in statistics related to the distribution of error rather than prediction accuracy.

Project managers and more senior managers have different concerns. Senior managers are concerned about a set of projects i.e. a portfolio. The portfolio approach, as suggested in Kitchenham and Linkman (1997), requires a prediction system that optimises characteristics of the portfolio such as minimising overall loss or maximising overall gain. It would be concerned about statistics relating to total error and also the extent to which individual project estimates were unbiased.

Table 9 indicates that different prediction quality indicators are geared towards different estimation objectives associated with the roles of the people involved in using estimates. It is interesting to note that none of the usual summary statistics report measures of skewness, although the ratio of (i) the mean and median error or (ii) ratio of mean to median accuracy are simple indicators of skewness.

Where the objectives are unknown, researchers should choose an indicator from each category in order to identify for what purposes a prediction system is best suited. In addition, the R^2 and adjusted R^2 may be generally useful in indicating whether there is any empirical basis for the prediction system. Any model where the R^2 value is not significant can automatically be discarded.

Table 9: Estimation Objectives and Prediction Indicators

Role	Error Indicators	Error Variance Indicators	Error Skewness	Error Kurtosis
Project Manager (risk averse)	Average Error	Mean Square Error		
Project manager (risk seeking)	Median Error	Median Absolute Error[2]		
Senior Manager (portfolio)	Total Error			
	Total Relative Error			
Estimator (risk averse)	Mean Accuracy	MMRE, EMMRE, Balanced MMRE		Pred(25)
Estimator (risk seeking)	Median Accuracy	Median MRE		Interquartile accuracy range

Table 9 also indicates that portfolio based measures are somewhat different to error and accuracy measures. For portfolio analysis you are likely to be interested in

[2] This measure is not discussed in section 2.

whether or not you have a homogeneous portfolio. So instead of measures of spread or shape, you might be interested in total error or total relative error for different segments of the portfolio (e.g. very large projects, average projects and very small projects).

Problems arise because estimating objectives are seldom one-dimensional. Although we might have a preference for an unbiased estimate, it is unlikely we would really want a prediction system that ensured lack of bias at the expense of large estimate error. Thus, except for portfolio assessment, we recommend selecting at least one measure of accuracy or error and one measure of error or accuracy variance. In addition it may be necessary to provide appropriate shape measures. Furthermore, measures of accuracy and accuracy variance and accuracy shape should be consistent with one another, for example, if you are using actual/estimate as your accuracy measure, you should use EMMRE not MMRE as a measure of accuracy variance.

However, a more complete understanding of the distribution can be obtained by constructing boxplots of the residuals or the accuracy. Boxplots can be regarded as an improvement upon simple summary statistics because they allow a visual display of central value, spread and shape which also highlight the extent to which the fitted model is vulnerable to outliers.

5. Conclusions

Our analysis and results suggest that the two statistics most frequently used to assess the quality of prediction systems, MMRE and pred(25), are respectively measures of the spread (variance) and shape (kurtosis) of the accuracy (estimate/actual). We believe that it is necessary to report measures of accuracy (i.e. actual/estimate) as well as measures of the spread and shape of the accuracy distribution. Furthermore, we suggest that boxplots of the residuals (actual error) and accuracy give a better assessment of prediction quality than one or two summary statistics.

We have presented evidence that prediction quality indicators are affected both by the analysis technique and the characteristics of the dataset from which the prediction system was derived. This relationship suggests another good reason for presenting results in the form of boxplots. Boxplots make it relatively easy to compare estimates derived from different prediction systems while at the same time making clear the nature of the dataset in terms of estimate bias and the impact of outliers.

In addition, the concept of errors and accuracy values being represented as a distribution of values supports improved methods of comparing prediction systems. For example, Stensrud and Myrveit (1998) suggest using a paired *t* test to test whether the absolute relative error obtained using one prediction system is significantly different from the absolute relative error obtained using another system. Pickard *et al.* (1999) used a non-parametric sign test on the residual values obtained from different prediction systems. Both these approaches allow two prediction systems to be compared using a

formal statistical test of significance rather than compared subjectively by means of simple descriptive statistics.

Thus to summarise, whilst many of the arguments in this paper may appear arcane to the non-statistician, it is essential that we understand how to make comparisons between competing prediction systems. Researchers have employed a wide range of different accuracy indicators, some of which appear to give contradictory results. Without understanding what the various indicators are describing, meaningful comparison is not possible. And if we cannot make meaningful comparisons we cannot make progress. We have argued that the indicators are statistics describing residual values and that a number of different properties of the residuals need to be described. Moreover, different properties will be of interest in different circumstances. For this reason we urge researchers to provide a range of indicators such as offered by descriptive techniques such as boxplots.

References

Belady, L.A. and Lehman, M.M. (1979) The characteristics of large systems. In Research Directions in Software Technology. P.Wegner (ed..), Cambridge , MA, MIT Press, p106-138.

Chulani, S, Clark, B. Boehm, B.W. (1998), "Calibrating the COCOMO II Post Architecture Model," 20th International Conference on Software Engineering,

Conte, S.D., Dunsmore, H.E., and Shen, V.Y.(1986). *Software Engineering Metrics and Models*, Benjamin/Cummings, Menlo Park CA.

Iglewicz, B. (1983) Robust scale estimators and confidence intervals for location. In Understanding Robust and exploratory data analysis. Hoaglin, D.C., Mosteller, F. and Tuket, J.W. (eds), John Wiley & sons Inc.

Jorgensen, M., (1995). Experience with the accuracy of software maintenance task effort prediction models, *IEEE Trans. Soft. Eng.* 21, 674-681.

Kemerer, C.F., (1987). An empirical validation of software cost estimation models, *CACM*, 30(5), pp416-429.

Kitchenham, B.A., (1992). Empirical studies of assumptions that underlie software cost estimation models, *Information & Softw. Technol.*, 34(4), pp211-218.

Kitchenham, B. A., and Linkman, S. G. (1997). Estimates, uncertainty and risk, *IEEE Software*, 14(3), 69-74.

Kok, P., B.A. Kitchenham, and J. Kirakowski. (1990). The MERMAID approach to software cost estimation, in *Proc. Esprit Technical Week*, 1990.

Lo, B.W.N., and Gao, X. (1997). Assessing software cost estimation models: criteria for accuracy, consistency and regression. *Australian J. of Information Systems*, 5(1), 30-44.

MacDonell, S.G., Shepperd, M.J., and Sallis, P.J. (1997). Metrics for Database Systems: An Empirical Study. In *4th IEEE Intl. Metrics Symp.,* Alberqueque, NM

Miyazaki, Y., Takanou, A., Nozaki, H., Nakagawa, N., and Okada, K, (1991). Method to estimate parameter values in software prediction models, *Information & Softw. Technol.*, 33(3), 239-243

Miyazaki, Y., *et al.*, (1994). Robust Regression for Developing Software Estimation Models, *J. of Systems & Software*, pp3-16,.

Pickard, L.M., Kitchenham, B.A. and Linkman, S.J. (1999) Investigation of Analysis Techniques for Software datasets. Dept. Computer Science. Keele University, TR99-05..

Rousseeuw, P.J., and Leroy, A.M., (1987). *Robust Regression and Outlier Detection*, John Wiley & Sons, New York.

Stensrud, E. and I. Myrveit. (1998) "Human performance estimating with analogy and regression models: An empirical validation, Proceedings of the Fifth International Software Metrics Symposium IEEE Computer Society Press.

Shepperd, M.J. and C. Schofield, (1997). Estimating software project effort using analogies, *IEEE Trans. on Softw. Eng.*, 23(11), pp736-743, 1997.

Acknowledgements
Dr Lesley Pickard's work is supported by the EPRSC Project GR/M 33709.