

New Ideas and Emerging Research: Evaluating Prediction System Accuracy

Martin Shepperd
School of Computing, IS & Maths
Brunel University
Uxbridge UB8 3PH, United Kingdom
martin.shepperd@brunel.ac.uk

Stephen G. MacDonell
SERL, Computing & Mathematical Sciences
AUT University, Private Bag 92006
Auckland 1142, New Zealand
stephen.macdonell@aut.ac.nz

ABSTRACT

BACKGROUND: Prediction e.g. of project cost is an important concern in software engineering.

PROBLEM: Although many empirical validations of software engineering prediction systems have been published, no one approach dominates and sense-making of conflicting empirical results is proving challenging.

METHOD: We propose a new approach to evaluating competing prediction systems based upon an unbiased statistic (Standardised Accuracy), analysis of results relative to the baseline technique of guessing and calculation of effect sizes.

RESULTS: Two empirical studies are revisited and the published results are shown to be misleading when re-analysed using our new approach.

CONCLUSION: Biased statistics such as MMRE are deprecated. By contrast our approach leads to valid results. Such steps will greatly assist in performing future meta-analyses.

Categories and Subject Descriptors

D.2.9 [Management]: Cost estimation

General Terms

Management, Measurement

Keywords

Software project management, empirical analysis, prediction system, prediction quality, forecasting, accuracy.

1. INTRODUCTION

An important feature of any engineering discipline is the ability to make timely and accurate predictions, and in order to do so we need prediction systems. Moreover we need to evaluate such models or prediction systems. Software engineering is no exception. However, judging predictive accuracy is a subtle task and our failure to do so appropriately

is a contributor to the present situation of a lack of conclusion stability across studies [13] and inconclusive results from systematic reviews [7].

So we find ourselves in a situation of increasing numbers of models and modelling techniques being offered, frequently with contradictory claims and results, for a wide range of data sets [9]. The picture is further complicated by the use of different accuracy statistics and validation schemes. The consequence is difficulties in sense making and increasing numbers of researchers commenting upon the challenge of conclusion instability.

In order to illustrate the challenges and show our proposed solution we revisit two published empirical studies of software project effort prediction systems and demonstrate how traditional evaluation techniques are misleading and how evaluation should be handled. The selected studies involved the first author (MS) and two contrasting data sets. Study 1 was published in an international conference with an acceptance rate of approximately 30% and Study 2 in the *IEEE Transactions on Software Engineering*.

2. METHOD

In order to bring some generality to our discussion we propose the following framework. We validate some prediction system P over a data set D using some accuracy statistic S . Empirical evaluation can be seen as an attempt to establish a set of preference relations such that $S(P_1, D) \preceq S(P_2, D), \dots, S(P_n, D)$. In this paper we restrict the discussion to predicting some continuous¹ output that is denoted Y , however, in principle the argument also applies to classifiers where the output is categorical.

When establishing these preference relations we need to be concerned with some other questions. For a given accuracy statistic S and data set D :

1. Does the prediction system P_1 outperform naïve guessing, a special case of a prediction system that we denote P_0 . In other words is $P_1 \prec P_0$ statistically significant for some pre-determined value of α ?
2. Is the difference $P_1 \prec P_2$ significant for some pre-determined value of α ?
3. Is the effect size large enough to justify $P_1 \prec P_2$ in practice?

¹Strictly speaking we also include the absolute scalar type i.e. counting.

Researchers have tended to focus on the second question typically by testing for the difference in means or medians. Exceptions are Jørgensen [3] who used sample mean productivity as fairly simple benchmark and Miller [10] who emphasises the need to estimate effect size in order to calculate the power of an experiment.

Typically statistics such as MMRE have been used as the accuracy statistic S for continuous prediction systems, where MMRE is given as:

$$\frac{\sum_1^n |(y_i - \hat{y}_i)|/y_i}{n} \quad (1)$$

and n is the number of cases in D . Obviously for classifiers different accuracy statistics e.g. AUC would be more appropriate.

Unfortunately it has been shown that this popular prediction accuracy statistic is flawed [8] in that it is a biased estimator of central tendency of the residuals of a prediction system because it is an asymmetric measure. Foss et al. [2] also highlighted problems associated with MMRE by means of simulation.

The fundamental variable of interest is the residual or prediction error, $y_i - \hat{y}_i$. As has been indicated there are potentially a number properties of this variable, however, for the present we assume the focus is upon central tendency rather than say bias or spread. As prediction system bias is not a concern, we use absolute residuals and for a set of predictions, mean absolute residual (MAR). This measure of centre is unbiased since it is not based on ratios unlike MMRE which must be bounded by zero in one direction and unbounded in the other.

However, MAR does have the disadvantage that it is hard to interpret and comparisons cannot be made across data sets since the residuals are not standardised. Therefore we propose to measure accuracy as the MAR relative to naïve guessing P_0 hence we offer a standardised accuracy measure SA for prediction technique P_i :

$$SA_{P_i} = 1 - \frac{MAR_{P_i}}{\overline{MAR}_{P_0}} \quad (2)$$

where \overline{MAR}_{P_0} is the mean value of a large number, typically 1000, runs of naïve guessing. This is defined as – to predict a \hat{y} for the target case randomly sample over all the remaining cases and take $\hat{y} = y_{RAND}$. This is the most naïve approach possible without being perverse. It also provides a relevant baseline irrespective of the exact form of P_1 . Over many runs the \overline{MAR}_{P_0} will converge on simply using the sample mean. The advantage of not using the sample mean is one can estimate the distribution of MARs for determining likelihood of any observed MAR value along with the variance of MAR. Note that whilst SA, like MMRE is a ratio, this is not problematic since we are only interested in one direction i.e. better than random.

The interpretation of SA is that the ratio represents how much better P_i is than naïve guessing. Clearly a negative value would be worrisome and close to zero discouraging!

To judge the effect size we use a standardised measure due to Glass [11] which is:

$$\Delta = \frac{MAR_{P_i} - \overline{MAR}_{P_0}}{s_{P_0}} \quad (3)$$

where s_{P_0} is the sample standard deviation of the naïve

guessing strategy. Note we do not use a pooled measure as in Cohen’s d since (i) we cannot assume the variances of P_i and P_0 are homogenous and (ii) the comparison is with respect to the control i.e. naïve guessing². Even if comparing between two prediction systems the rationale still tends to be P_1 represents the *status quo* with which P_2 is to contrasted and hence P_1 is effectively a control.

Having defined a standardised accuracy measure SA and an effect size measure Δ we are now in a position to revisit some typical empirical validation studies of project effort prediction systems and pose our three questions.

3. STUDY 1: THE ATKINSON DATA SET

This is a small data set of telecoms projects that uses real-time function points as a size estimator. It was one of two data sets employed by the replication study (Study 1) [12] of a proposed regression to the mean (R2M) prediction method [4]. The details are not important, suffice to say that the aim of Study 1 was to empirically compare the accuracy of R2M with an estimation-by-analogy (EBA) prediction system as a baseline. The reported accuracy statistics were MAR and MMRE (for interpretation purposes).

Table 1: Atkinson Data Set Results

Prediction Method	MAR	MMRE	SA
EBA-prod	331.6	99%	-17%
R2M-prod	291.6	84%	-3%
NG	283.0	86.2%	0%
NG _{0.05}	210.8	56.8%	26%
Expert judgment	117.5	24.1%	58%

Table 1 gives the results for SA sorted and expressed as a percentage including EBA-prod and R2M-prod which was the main theme for study Study 1. However, we also add the results for naïve guessing (P_0) and the 5% quantile for the cumulative distribution of MAR values from 1000 runs of P_0 (the histogram is shown in Fig. 1). An encouraging, and to be expected, observation is the near Gaussian distribution which contrasts with a more skewed MMRE distribution from the same 1000 runs (see Fig. 2). Note also that MMRE does not preserve the same ranking of prediction systems.

We can see that the two methods under review (EBA-prod and R2M-prod) both perform *worse* than naïve guessing. It is therefore pointless to even consider whether the differences between the two methods are significant because one would be better off guessing. In a sense one might argue we have non-prediction systems. Although not included in Study 1, we add results using predictions made by experts at the time which formed part of the data set. We see that expert judgment is best and a 58% improvement over guessing.

4. STUDY 2: THE DESHARNAIS DATA SET

Study 2 considers nine data sets in an empirical comparison of stepwise regression (SWR) prediction, intended as a benchmark, and estimation by analogy (EBA). We choose the Desharnais data set out of the nine as it is substantially larger than Atkinson (77 cf. 16 cases) and because it

²One note of caution is that Glass’s Δ is known to be a biased estimator for small sample sizes or if there are large discrepancies in sample sizes, in which case Hedges’s g might be preferred.

is widely used [9]. The main accuracy statistic employed in Study 2 [14] was MMRE.

Histogram of MAR values from Naive Guessing for the Atkinson Data S

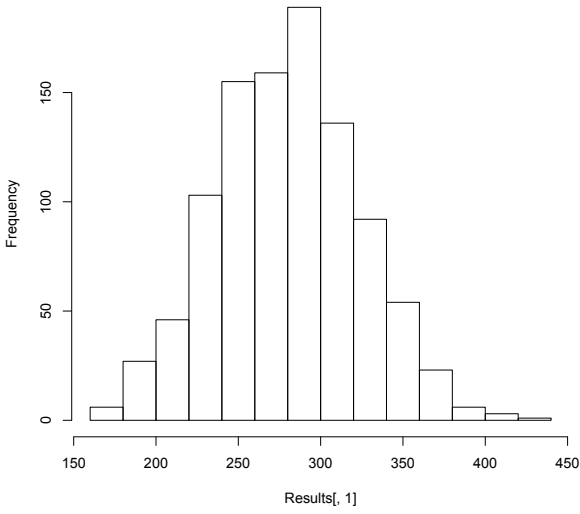


Figure 1:

MMRE values from Naive Guessing for the Atkinson Data Set

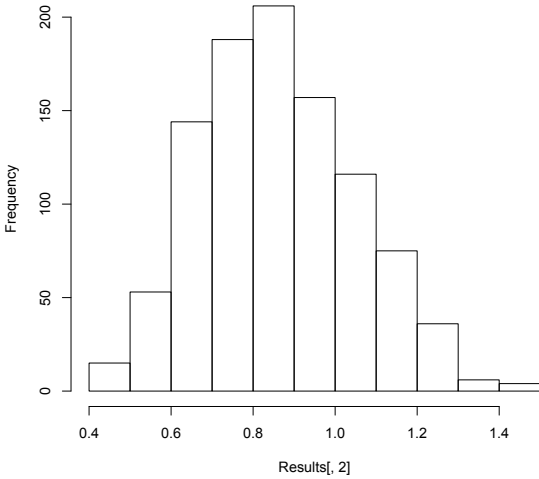


Figure 2:

Table 2: Desharnais Data Set Results

Prediction Method	MAR	MMRE	SA
NG	4149	142%	0%
NG _{0.05}	3556	110%	13.9%
EBA, $k = 3$	2289	63%	44.9%
SWR	2022	71%	51.3%
EBA-FSS, $k = 3$	1682	41%	55%

Table 2 sets out the prediction results from SWR, EBA and, in addition, we add EBA that uses a better (more recent) feature subset selection method (EBA-FSS) [6]. To these prediction systems are added the baseline technique of naïve guessing and the 5% quantile ($NG_{0.05}$).

In order to obtain the residuals we re-ran the regression analysis and EBA keeping the procedure as similar as possible to [14]. Due to changes in the EBA tool over the intervening 12 years our results differ slightly from Study 2 although the rankings are preserved. Running a paired one-tailed t-test yields $p \approx 0.3$ which of course is not significant. However, in fairness to the original study, the argument was made with respect to the overall analysis of nine data sets.

We also see the bias of MMRE since it indicates EBA is to be preferred to the SWR prediction system yet by examining the residuals and MAR we see this is fallacious. Furthermore, SWR seeks to minimise the sum of the squares of the residuals so using MMRE is inappropriate. Examining the SA measure we see that both prediction systems improve upon guessing and fall beyond the 0.05 confidence level. SWR is the more accurate technique and offers a 51.3% improvement over guessing. We observe that the difference in SA between SWR and EBA is small (6.4%) and in the opposite direction to that indicated by MMRE i.e. $P_{SWR} < P_{EBA}$.

Table 3: Desharnais Effect Sizes

Prediction Method	MAR SD	MMRE SD	Δ wrt P_0	Δ wrt SWR
NG	4220	258%	n.a.	n.a.
EBA, $k=3$	2684	82%	0.436	-0.123
SWR	2171	118%	0.499	n.a.

Next we consider effect size, Δ (see Table 3) using Eqn. 3. Cohen [1] suggested that one might interpret a d or Δ of 0.2 as indicating a small effect, 0.5 as medium and 0.8 as large. Note that Δ may exceed 1. First, we see that the improvement in accuracy over guessing due to using SWR or EBA tends towards a medium effect size. Clearly each approach is doing substantially better than naïve guessing but then this is not saying a lot! Study 2 set out to compare EBA with SWR as a benchmark so we recalculate the Δ 's with respect to SWR and use the SWR MAR_{SD} value. Here we see a small, statistically insignificant, *negative* effect size between SWR and EBA.

The above analysis uses MAR as the accuracy statistic for the reasons discussed previously, however, to indicate yet another disadvantage with MMRE we provide the standardised standard deviations (i.e. mean/SD) in Table 4. We

see that in all cases the standardised variance is greater for MMRE than for MAR. This means that any effects due to differences in prediction system are harder to detect since they will be masked by variance inherent in the accuracy statistic.

Table 4: Variance due to MAR cf. MMRE

P	MAR/SD MAR	MMRE / SD MMRE
NG	1.16	2.35
EBA, $k=3$	1.80	4.10
SWR	2.04	3.64

5. DISCUSSION AND CONCLUSIONS

First we need to stress that our purpose is not the evaluation of specific prediction systems but rather the question of *how* one evaluates prediction systems. From the foregoing it is clear that there are a number of problems with how we presently view and analyse results from empirical validation studies of competing prediction systems. We have re-visited two such studies and show that the published results — both studies underwent peer reviewing for prestigious outlets — were misleading.

For Study 1 the absence of some fundamental baseline (P_0) meant that the accuracy statistics were hard to interpret (MAR=292 and MMRE=84%) and so the problem that both techniques were substantially worse than guessing was overlooked. In such a situation the relative improvement of R2M-prod is irrelevant and in any case is most likely the consequence of the technique converging on using the sample mean as the correlation between predicted and actual productivity tends to zero.

In Study 2, unlike Study 1, the prediction techniques perform significantly better than P_0 . Here however, the problem is reliance on a biased accuracy statistic MMRE which reverses the preference relation between EBA and SWR. Further the effect with respect to P_0 would barely be considered medium and the difference between EBA and SWR is negligible.

However before both papers are consigned to the scrapheap of history it should be added that both papers use multiple data sets and so some form of meta-analysis is required before commenting upon the overall conclusions of each study.

So where do we go from here? First, biased accuracy statistics should be absolutely deprecated. We see no good purpose in using MMRE. We have proposed a new standardised measure SA that is unbiased, enables meaningful interpretation and can be used across different data sets. Second, validation studies should address the three questions we pose in this paper, namely is P_1 better than guessing (P_0), is P_1 significantly better than P_2 and is the effect size large enough to be of practical import? Biased statistics and noisy data sets will lead to high levels of variance resulting in the between-prediction system variance being difficult to detect. Finally, this approach will better support meta-analysis via effect sizes even when different response variables have been used [5].

Acknowledgments

Martin Shepperd was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant

EP/H050329/1. The authors wish to thank Magne Jørgensen, Barbara Kitchenham, Carolyn Mair and Audris Mockus for helpful comments and discussions.

6. REFERENCES

- [1] J. Cohen. *Statistical power analysis for the behavioral sciences*. Lawrence Earlbaum Associates., Hillsdale, NJ, 2nd edition, 1988.
- [2] T. Foss, E. Stensrud, B. Kitchenham, and I. Myrvtveit. A simulation study of the model evaluation criterion mmre. *IEEE Transactions on Software Engineering*, 29(11):985–995, 2003.
- [3] M. Jørgensen. Experience with the accuracy of software maintenance task effort prediction models. *IEEE Transactions on Software Engineering*, 21(8):674–681, 1995.
- [4] M. Jørgensen, U. Indahl, and D. I. K. Sjøberg. Software effort estimation by analogy and “regression toward the mean”. *J. of Systems & Software*, 68(3):253–262, 2003.
- [5] V. Kampenes, T. Dybå, J.E. Hannay, and D.I.K. Sjøberg. A systematic review of effect size in software engineering experiments. *Information & Software Technology*, 49:1073–1086, 2007.
- [6] C. Kirsopp, M.J. Shepperd, and J. Hart. Search heuristics, case-based reasoning and software project effort prediction. In *Genetic and Evolutionary Computation Conf.*, New York, 2002. AAAI.
- [7] B. Kitchenham, P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman. Systematic literature reviews in software engineering – a systematic literature review. *Information & Software Technology*, 51(1):7–15, 2009.
- [8] B.A. Kitchenham, S.G. MacDonell, L. Pickard, and M.J. Shepperd. What accuracy statistics really measure. *IEE Proceedings - Software Engineering*, 148(3):81–85, 2001.
- [9] C. Mair, M. Shepperd, and M. Jørgensen. An analysis of data sets used to train and validate cost prediction systems. In *PROMISE 2005*, St Louis, MI, 2005. ACM Computer Press.
- [10] J. Miller, J. Daly, M. Wood, M. Roper, and A. Brooks. Statistical power and its subcomponents - missing and misunderstood concepts in empirical software engineering research. *Information & Software Technology*, 39(4):285–295, 1997.
- [11] R. Rosenthal. Parametric measures of effect size. In H. Cooper and L. Hedges, editors, *The Handbook of Research Synthesis*. Sage, New York, 1994.
- [12] M. Shepperd and M. Cartwright. A replication of the use of regression towards the mean (r2m) as an adjustment to effort estimation models. In *11th IEEE Intl. Softw. Metrics Symp. (Metrics05)*, Como, Italy, 2005. Computer Society Press.
- [13] M.J. Shepperd. Software project economics: a roadmap. In *Future of Software Engineering (ICSE 2007)*, Minneapolis, 2007. ACM.
- [14] M.J. Shepperd and C. Schofield. Estimating software project effort using analogies. *IEEE Transactions on Software Engineering*, 23(11):736–743, 1997.