Researcher Bias: The Use of Machine Learning in Software Defect Prediction: Supplementary Materials

Martin Shepperd, David Bowes and Tracy Hall

Statistical Procedures

In these Supplementary Materials we describe some threats to our procedures and demonstrate how the alternative analyses that we performed do not materially alter our findings and therefore our conclusions.

The Response Variable is non-Gaussian

As mentioned in the main paper, one slightly awkward property of the Matthews Correlation Coefficient (MCC) is that depending upon the marginal distributions of the confusion matrix, plus or minus unity may not be attainable and so the theoretical maxima and minima are constrained. Some statisticians propose a ϕ/ϕ_{max} rescaling [1]. We choose not to follow this procedure since it results in an over-sensitive measures of association in that a very small change in the count of correctly classified instances (TP or TN) lead to unintuitively large changes in the correlation. This is a particular problem in the setting of imbalanced data sets which are the norm for software defects. The deleterious effects of this can be seen very plainly in Fig. 1 in particular for the tails, when compared with the original distribution shown by Fig. 2.

Correlation coefficients tend not to normally distributed due to their tails being constrained to unity and minus unity (as opposed to plus or minus infinity). This is usually corrected using the Fisher r to Z transformation which is based on an inverse hyperbolic function [2]. In principle this also has the beneficial impact of stabilising the variance so that it doesn't diminish as it approaches either limit. (NB Our data set contains values of MCC=1, therefore 0.0001 was subtracted before applying the r to Z transformation to prevent mapping to infinity). Unfortunately our data reveals highly unstable behaviour as illustrated by the qqplot in Figure 3.

- *Martin Shepperd and Tracy Hall are with Brunel University, Uxbridge, Middlesex, UB8 3PH,UK.*
  E-mail: {martin.shepperd,tracy.hall}@brunel.ac.uk
- *David Bowes is with Science and Technology Research Institute, University of Hertfordshire, Hatfield, Hertfordshire, AL10 9AB, UK.*
  E-mail: d.h.bowes@herts.ac.uk

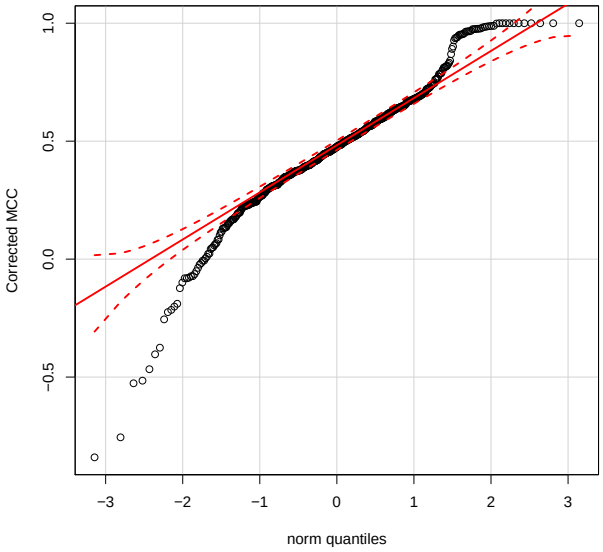

Figure 1. qqplot of MCC After the Phi/Phimax Correction

The explanation seems to be that the tails and in particular the upper tail are rather over-represented prior to the transformation (see Figure 2) so the 'correction' exacerbates this deviation. This is probably explained by the tendency of researchers to select their most 'interesting' results thus our response variable is not a random sample of all experimental results [3]. Thus we do not use this transformation.

Use of Robust Alternatives to ANOVA

Given these possible problems we also explore a robust analysis based upon the following. The square of the MCC (or phi) is approximately a chi-squared distribution with one degree of freedom [4]. A chi-squared distribution is a special case of a gamma distribution so we check the linear model using MCC-squared and a generalised linear model (GLM) where the choice of exponential family of distributions is gamma and the

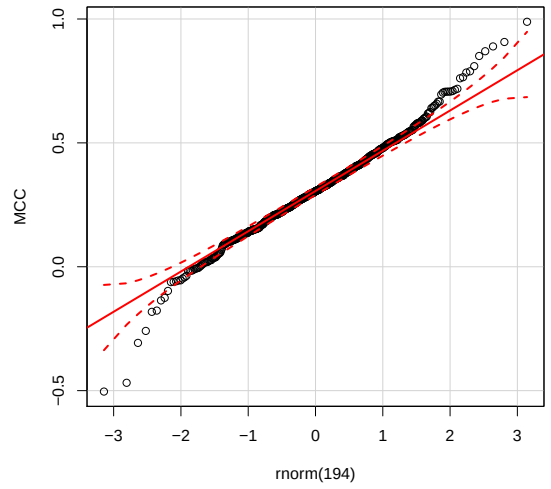

Figure 2. qqplot of the Matthews Correlation Coefficients without transformation

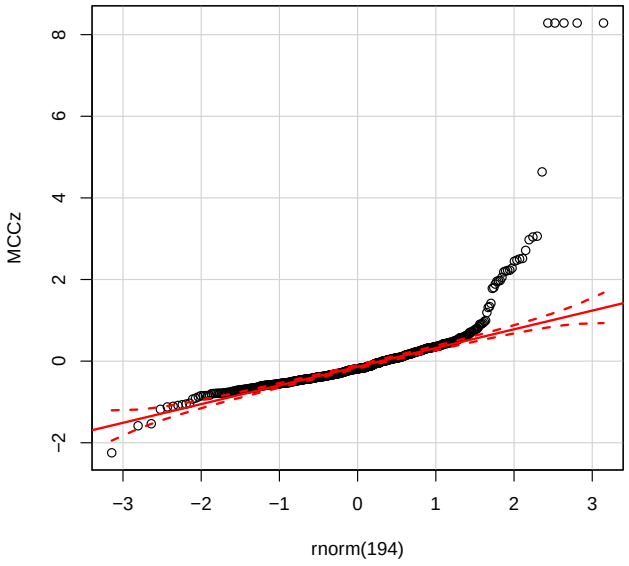

Figure 3. qqplot of MCC after the Application of the Fisher r to Z Transformation

link function is inverse (in R this is Family=gamma and link=inverse). The results are shown in Table 1 which again indicate that Researcher Group is the dominant factor in terms of deviance (rather than variance which we use for ANOVA).

A 4-way ANOVA model is over complex

The complexity of the 4-way model means that from a factorial point of view many of the cells are empty since if we count the levels this yields ($23 \times 24 \times 7 \times 8 = 30912$). For this reason we looked at "dense" subsets of our data. We performed the following analysis for both ECLIPSE and NASA datasets. We only used the most popular class of Metric and for the single data set thereby reducing the model to two-way.

The results using just Eclipse are given in Table 2 and for NASA in Table 3. The factors are listed in order of importance. In both analyses both factors and the interaction term is significant and Research Group remains dominant (accounting for between 5 and 60 times more variance than the choice of Classifier technique). Note, however that for the Eclipse data set the model does not show a very good fit with the residual term accounting for more than 80% of the total variance. This suggest that adding extra factors improves the explanatory value of our model and provides the ability to fit to a wider range of experimental results.

Table 2
Eclipse Only Analysis: 2-way ANOVA (MCC = ResearcherGroup*Classifier)

	Df	Sum Sq	% of total variance	F value	Pr(>F)
ResearcherGroup	4	0.304	18.95	3.595	0.011
Classifier	4	0.006	0.35	0.067	0.992
ResearcherGroup: Classifier	2	0.005	0.31	0.118	0.889
Residuals	61	1.290	80.39		

Table 3
NASA Only Analysis: 2-way ANOVA (MCC = ResearcherGroup*Classifier)

	Df	Sum Sq	% of total variance	F value	Pr(>F)
ResearcherGroup	13	1.546	20.12	7.168	0.000
ResearcherGroup: Classifier	19	0.600	7.81	1.904	0.013
Classifier	7	0.379	4.92	3.259	0.002
Residuals	311	5.161	67.15		

Lastly we repeat the 4-way analysis on the three most widely used data sets (Eclipse, NASA and Mozilla) to explore whether our main findings are the result of a few infrequently used data sets.

We see from Table 4, which is again arranged in decreasing importance of factors, that the results remain broadly similar to our 4-way model for all data (given in the main paper). Researcher Group still dominates, although in this case the second order interaction term of Researcher Group with Classifier now becomes the second term. Given Data Set is now limited to three levels this may explain it's reduced importance. Overall the model accounts for about 40% of the overall variance in MCC which represents a small reduction in fit from 43.6% in our model for all experimental results contained within the main paper.

Table 1
4-way ANOVA (MCC^2 = Response Variable)

	Df	Deviance	Resid. Df	Resid. Dev
NULL			599	879.41
ResearcherGroup	22	150.03	577	729.38
Dataset	20	87.47	557	641.91
Metric	5	15.12	552	626.80
Classifier	7	18.25	545	608.55
ResearcherGroup:Dataset	3	5.76	542	602.79
ResearcherGroup:Metric	6	15.30	536	587.49
Dataset:Metric	1	0.11	535	587.38
ResearcherGroup:Classifier	32	95.63	503	491.75
Dataset:Classifier	7	2.14	496	489.61
Metric:Classifier	4	0.27	492	489.34
ResearcherGroup:Dataset:Metric	0	0.00	492	489.34
ResearcherGroup:Dataset:Classifier	0	0.00	492	489.34
ResearcherGroup:Metric:Classifier	0	0.00	492	489.34
Dataset:Metric:Classifier	0	0.00	492	489.34
ResearcherGroup:Dataset:Metric:Classifier	0	0.00	492	489.34

Table 4
4-way ANOVA Using the 3 Most Frequently Used Datasets (MCC = Response Variable)

	Df	Sum Sq	% of total variance	F value	Pr(>F)
ResearcherGroup	17	2.380	22.16	8.449	0.000
ResearcherGroup:Classifier	23	0.636	5.93	1.670	0.028
Dataset	2	0.578	5.38	17.452	0.000
Classifier	7	0.378	3.52	3.263	0.002
ResearcherGroup:Dataset	2	0.236	2.19	7.113	0.001
Residuals	394	6.528	60.78		

RAW DATA

Our raw data and the R scripts are available from https:// codefeedback.cs.herts.ac.uk/mlbias.

REFERENCES

[1] E. Davenport and N. El-Sanhurry, "Phi/phimax: Review and synthesis," *Educational and Psychological Measurement*, vol. 51, pp. 821–828, 1991.

[2] M. Warner, *Applied statistics: from bivariate through multivariate techniques*. London: SAGE Publications, 2013.

[3] K. Dickersin, "The existence of publication bias and risk factors for its occurrence," *JAMA: the Journal of the American Medical Association*, vol. 263, no. 10, pp. 1385–1389, 1990.

[4] S. Siegel and N. Castellan, "Nonparametric statistics for the behavioral sciences (Mcgraw-hill, New York)," 1988.