PA

# Seeing the Forest through the Trees

**David Alan Muchlinski[1], David Siroky[2], Jingrui He[3] and Matthew Adam Kocher[4]**

[1] *Georgia Institute of Technology, Sam Nunn School of International Affairs, Atlanta, GA 30332, USA. Email: d.muchlinski@unsw.edu.au*
[2] *Arizona State University—School of Politics and Global Studies, Tempe, AZ 85287, USA. Email: david.siroky@asu.edu*
[3] *Arizona State University—Computer Science and Engineering, Tempe, AZ 85281, USA. Email: jingrui.he@asu.edu*
[4] *Johns Hopkins University—Political Science and SAIS, Baltimore, MD 21218, USA. Email: matthew.kocher@jhu.edu*

*Keywords:* binary logistic regression, data analysis algorithms, model selection

We thank the editors of *Political Analysis* (PA) for the opportunity to respond to these critiques of our article (Muchlinski, Siroky, He, and Kocher 2016). We also thank Yu Wang, Marcel Neunhoeffer, and Sebastian Sternberg for their constructive commentary and careful attention to our work.

In this reply, we do three things. First, we acknowledge two significant errors. Second, we point out that the substantive conclusions of our article hold, in spite of these errors. Third, having just been put through the replication process, we offer a few comments on how such a process should ideally work.

At the outset, we also want to indicate one thing we do *not* do. Wang, as well as Neunhoeffer and Sternberg, identify alternative algorithmic prediction procedures that they argue outperform our implementation of Random Forests. This is no great surprise. We chose to highlight Random Forests not because it is optimal, but rather because it is simple, widely available, easy to implement, and relatively straightforward to interpret compared to some other machine learning approaches. Our goal was modest: to point out that the canonical method for analyzing civil war data, logistic regression, does not predict civil war well, and that we can do much better with readily available techniques, including but not limited to Random Forests. Given widely divergent statistical models and a profound lack of consensus about the underlying causal processes, we hoped to convince readers that successful prediction is an appropriate scientific standard to aim for, building on seminal work in political science by Schrodt (1991), Beck, King, and Zeng (2000), Ward, Greenhill, and Bakke (2010), Greenhill, Ward, and Sacks (2011), and others. We congratulate Wang, Neunhoeffer, and Sternberg for further advancing this research agenda!

## 1 Acknowledgment of Errors

Wang points out that the AUC value of 0.91 we report for our implementation of Random Forests is inconsistent with the ROC curve depicted in Figure 2, which implies an AUC of 0.97. Further, he notes that the separation plot reported in Figure 1 implies no false negatives, when in fact our procedure generated three false negatives according to Wang's replication. Wang's observations are correct. In sum, Wang confirms that the AUC value reported in the text (0.91) is correct, and he replicated this number in his analysis. Figures 1 and 2 are not correct, however, as Wang points out. Corrected figures and code have been uploaded to the Harvard Dataverse and may be found here: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/KRKWK8.

Wang also points out that our implementation of Random Forests produces a much larger number of false positives than the logistic regression models to which we compare it. While this is true, it is also misleading. Logistic regression achieves a high rate of successful prediction for

***Political Analysis (2019)***
**vol. 27:111–113**
**DOI:** 10.1017/pan.2018.45

**Corresponding author**
David Alan Muchlinski

**Edited by**
Jeff Gill

civil war by predicting *no* onsets. Since approximately 1.6% of country/years have civil war onsets, logistic regression predicts 98.4% of the cases correctly and has a false positive rate of zero (!).[1] The entire prediction problem for class-imbalanced data is to find a procedure that strikes a balance between accuracy and sensitivity. The ROC curves and the AUC statistic capture this tradeoff and show, unequivocally, that Random Forests strikes a better balance than the common methods to which we compare it.

The problem identified by Neunhoeffer and Sternberg is ultimately less consequential but much more embarrassing. During PA's second manuscript review of our article, a reviewer suggested carrying out an out-of-sample test using completely different years (post 2000) for a subset of countries, which we did by collecting new data to produce Table 1. When our replication data and code were prepared, the member of our team who was responsible for this step posted the wrong code and data for this portion of the analysis (Table 1), and subsequently lost portions of the data and code used in our original analyses.

Working from our posted materials, Neunhoeffer and Sternberg concluded that we did not perform the out-of-sample analysis reported in the article (Wang makes the same observation in his Fn. 2). While this was not an unreasonable conclusion to draw based on the replication materials, it is not correct. When PA notified us of Neunhoeffer and Sternberg's findings, we immediately reconstructed the data and code from scratch and generated a new replication file that produced results very similar to those reported in our article. We could not replicate the original results *exactly* because we imputed the values of some variables; randomness in the imputation procedure implies that no two versions of the data will be exactly alike. Random Forests also utilizes randomness both in selecting cases and in the feature set (predictors) to use for prediction (Breiman 2001; Siroky 2009). For these reasons, the predictions will vary to some extent each time. A revised Table 1 is included with the corrected replication file we have uploaded to the Harvard Dataverse.

In other words, our analysis followed Neunhoeffer and Sternberg's procedure #6, not procedure #5 as they suggest.[2] The responsibility for this confusion is entirely our own, and we thank Neunhoeffer and Sternberg for identifying this error in the replication materials.

## 2 Fidelity of the Replicated Results to Our Original Claims

When a published article fails to replicate—i.e. the authors' code run on the authors' data does not reproduce the *exact* published results—it should be the beginning rather than the end of the discussion. While this standard of replication is important, it should not be fetishized. It is important to see the forest for the trees. A substantively erroneous or even fraudulent paper can replicate perfectly, while a paper that fails to replicate fully can be methodologically sound and arrive at accurate conclusions.

The central claims advanced in our article are that logistic regression is a relatively poor classifier for class-imbalanced outcomes such as civil war onset, and that statistical learning methods such as Random Forests give significantly better predictive results. Wang's replication, reported in Figures 2 and 3, fully supports these conclusions. Colaresi and Mahmood (2017) have also shown that Random Forests predicts civil war more accurately than logistic regression, and significantly improved on the earlier logit models through an adaptive iterative process of model criticism and building, which combines Random Forests and Logistic Regression, that we

1 Hegre and Sambanis's (2006) data include 115 civil war onsets out of 7140 country/years (1.61%).

2 In their replication exercise, Neunhoeffer and Sternberg divided the civil war data into training and testing sets at the year 1989, while we divided the data at 2000. It is not surprising that these analyses give rather different results. On substantive grounds, 1989 may be a particularly bad year at which to divide a civil war dataset, given prior evidence that the end of the Cold War created an unusually sharp temporal break in the global pattern of civil wars. However, Colaresi and Mahmood (2017) also split the data in 1989, and rely on the data from Hegre and Sambanis (2006), yet arrive at conclusions and numbers similar to ours, though again not exactly the same.

think should be more widely utilized in political science for the reasons described in the article. Our corrected replication materials demonstrate that our arguments are fully supported by the available evidence.

## 3 Institutionalizing Replication

Nobody likes to be criticized, especially when the critics have a point. Nevertheless, when Wang, Neunhoeffer, and Sternberg found significant errors in our work, they did exactly what they should have done: write up the details and forward them to the journal editors. The editors at PA notified us promptly of these criticisms and asked us for an explanation, which we readily provided, together with updated data and code. Most importantly, in our view, PA decided to publish the criticisms together with our reply.

As a discipline, political science has made great progress in creating a norm among quantitative analysts of making replication materials available. At the moment, however, we lack well-established norms about what to do when postpublication replications identify errors. Unless journals are prepared to publish replication results, scholars attempting to correct mistakes in published work run the risk of angering colleagues while deriving no professional benefit from their efforts. Even worse, very significant errors may stand uncorrected, either because critics bury their findings, or because unpublished critiques fail to achieve visibility comparable to that of the published articles they rectify. By publicizing the replication of our article and providing us the opportunity to explain our errors, the editors of *Political Analysis* are helping to establish a constructive normative framework for critical engagement in the discipline.

To conclude, we reiterate our appreciation for the effort Wang, Neunhoeffer, and Sternberg, as well as the editors of PA, put into identifying and correcting the errors in our article. We regret these mistakes, and we stand by our findings.

## References

Beck, Nathaniel, Gray King, and Langche Zeng. 2000. Improving quantitative studies of international conflict. *American Political Science Review* 94(March):21–36.

Breiman, Leo. 2001. Random forests. *Machine Learning* 45(1):5–32.

Colaresi, M., and Z. Mahmood. 2017. Do the robot: Lesson from machine learning to improve conflict forecasting. *Journal of Conflict Resolution* 54(2):193–214.

Greenhill, B. D., M. D. Ward, and A. Sacks. 2011. The separation plot: A new visual method for evaluating the fit of binary models. *American Journal of Political Science* 55(4):990–1002.

Hegre, H., and N. Sambanis. 2006. Sensitivity analysis of empirical results on civil war onset. *Journal of Conflict Resolution* 50(4):508–535.

Muchlinski, D., D. Siroky, J. He, and M. Kocher. 2016. Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. *Political Analysis* 24(1):87–103.

Schrodt, Philip A. 1991. Prediction of interstate conflict outcomes using a neural network. *Social Science Computer Review* 9(3):359–380.

Siroky, David S. 2009. Navigating random forest and related advances in algorithmic modeling. *Statistics Survey* 3:147–163.

Ward, M. D., B. D. Greenhill, and K. M. Bakke. 2010. The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research* 47(4):363–375.