Comments on Blair and Sambanis, 2020, “Forecasting Civil Wars: Theory and Structure in an Age of ‘Big Data’ and Machine Learning”, *JCR*

Richard K. Morgan and Andreas Beger

12 May 2020

We highlight some issues with the analysis in the recent Blair and Sambanis (2020) paper that undermine their main empirical finding – that a theoretically-informed escalation model can forecast civil war onsets better than more inductive alternatives. Their paper has two key results:

1. A theoretically-informed model outperforms alternative specifications in predicting civil war onset.
2. Structural variables do not add, and sometimes detract, from the forecast accuracy of the escalation model they advance.

These results are highlighted in Table 1 and Table 2 (Blair and Sambanis 2020, pg. 13 and 19 respectively). They also evaluate forecasts for the first half of 2016 and conclude good accuracy in forecasting civil war (Table 4 on pg.23).

We find three main issues in their empirical analyses:

- Their primary finding – that the escalation model they advance consistently out-performs the alternation models – depends entirely on the unorthodox use of smoothed ROC curves when calculating the AUC-ROC values in Table 1 and Table 2. When using the standard (not smoothed) ROC curves, this finding is reversed. That is, the escalation performs no better and often worse than the other models.
- Regarding the forecast assessment in Table 4, the forecasts for 2016-H1 are erroneously assessed using civil war *incidence*, not *onset*, even though the models from which they are generated are clearly fitted to civil conflict onset. This vastly overstates the precision of the forecasts.
- The way the random forest models are fitted is very unorthodox.[1] The number of tree the model grow, the number of nodes in these trees, and the sample size used for each tree deviates drastically for traditional approaches (e.g., they set the number of trees to 100,000 where the default is 500).

We demonstrate the first point through model replication: we use the replication files that Blair and Sambanis provide to calculate the AUC-ROC scores using smoothed ROC curves (the published results) and then calculate the AUC-ROC scores using the standard approach – without the unorthodox smoother.[2] We demonstrate the second point using the replicated results from the original model Blair and Sambanis use to create Table 4 and then compare these results to the observed outcomes in the 2016 data.

For the final point, we evaluate two questions: (1) even if we allow for the use of smoothed ROC curves, does the key model’s superior test forecast performance generalize or is it a result of inadvertent fitting of the model hyperparameters to test forecast performance, and (2) does the inferior performance of the alternative models, and especially the CAMEO model with 1,159 covariates, reflect a lack of proper hyperparameter tuning for these models? (work still in progress).

[1] This is in fact what first led us to examine the analysis in more detail.

[2] In this draft, we focus our replication efforts on what Blair and Sambanis call their “Base specification”. This set of models are what Blair and Sambanis use when discussing their main findings and the implications of their results.

Review of the Blair and Sambanis paper

Blair and Sambanis (2020) argue that theory contributes to forecasting accuracy even when using non-parametric machine learning models (in this case, the random forest model) and alternative specifications that are not informed by theory. They arrive at this conclusion by examining the problem of predicting civil war onset, and find that a parsimonious model using a small number of covariates derived from escalation theories of conflict can forecast civil war onset better than alternative specifications based on generic covariates not specifically informed by theory and a kitchen sink model with more than 1,000 covariates.

Their data consist of country-month observations covering 2001 to 2015. The main outcome is civil war onset, measured using Sambanis' civil war dataset. The independent variables for the first set of analysis reported in Table 1 in the paper are all derived from the ICEWS event data, using domestic events between actors within a country. The key results are based on a comparison of the forecast accuracy of a random forest (RF) model with 4 covariate specifications as well as an ensemble model. These models are:

- Escalation: a set of 10 theoretically-informed indicators.
- Quad: ICEWS quad counts, i.e. material conflict, material cooperation, verbal conflict, verbal cooperation.
- Goldstein: -10 (conflictual) to 10 (cooperative) scores derived from the ICEWS data for interactions between the government one one side and opposition or rebel actors on the other. These are directed, thus making for 4 total covariates.
- CAMEO: counts for all CAMEO event codes, thus a total of 1,159 covariates.
- Average: unweighted average of the predictions from the 4 models above.

To assess forecast accuracy, Blair and Sambanis conduct the following procedures:

1. Split the training data into training (2001 - 2007) and test (2008 - 2015) sets.
2. Estimate 4 random forest models with the specifications from above as well as a unweighted average of the models.
3. Create out-of-sample (OOS) predictions from each model using the test set.
4. Calculate AUC-ROC measures for each set of OOS predictions.

This is done for both 1-month ahead and 6-month ahead forecasts. Blair and Sambanis also conduct a series of robustness tests that (1) vary hyperparameters of the RF model, (2) change the split year for the training/test data split, and (3) alternative codings of the dependent variable.

The results for this analysis are summarized in Table 1 in Blair and Sambanis (2020), which we copied in here:

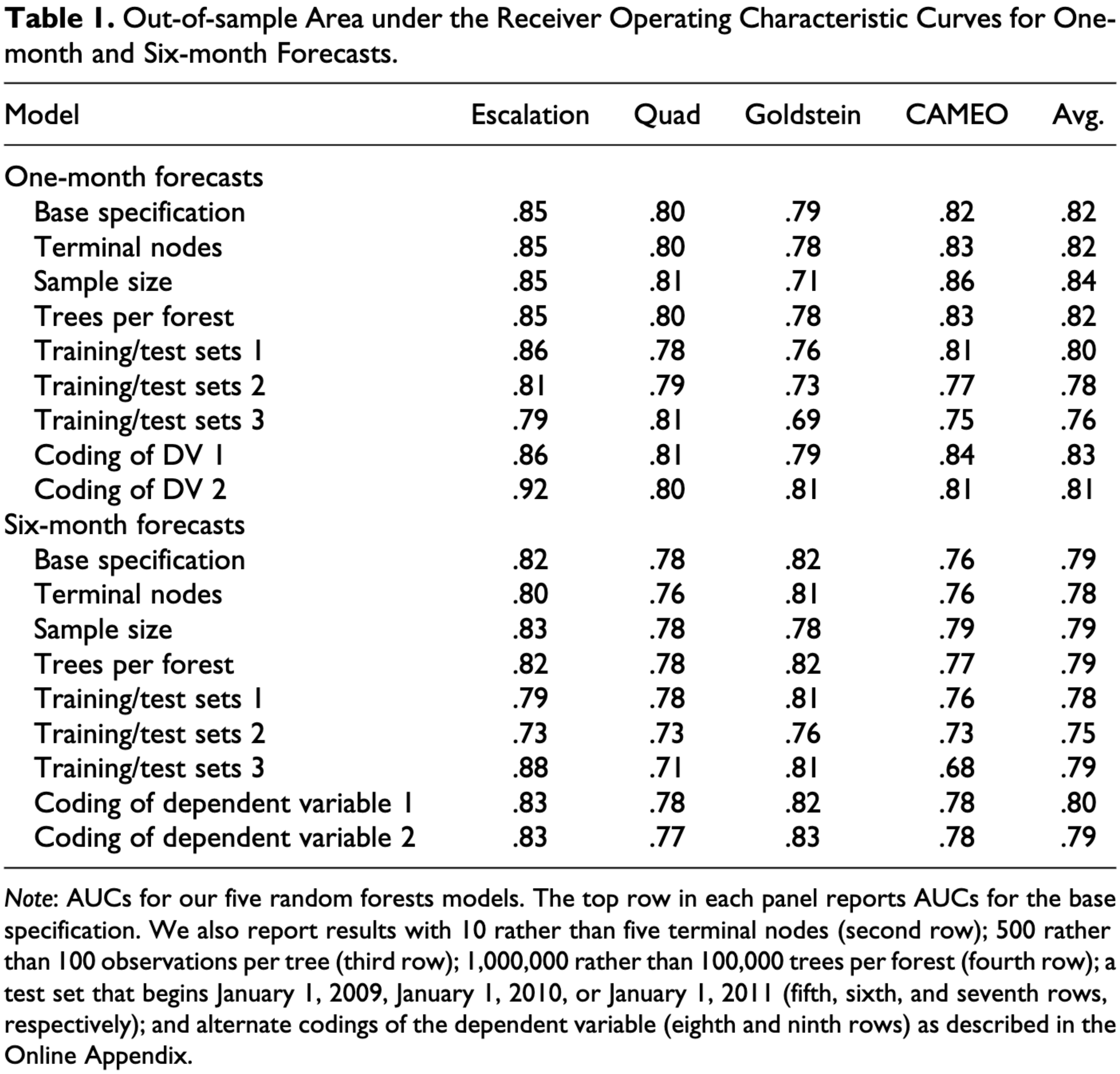

Table 1. Out-of-sample Area under the Receiver Operating Characteristic Curves for One-month and Six-month Forecasts.

Model	Escalation	Quad	Goldstein	CAMEO	Avg.
One-month forecasts					
Base specification	.85	.80	.79	.82	.82
Terminal nodes	.85	.80	.78	.83	.82
Sample size	.85	.81	.71	.86	.84
Trees per forest	.85	.80	.78	.83	.82
Training/test sets 1	.86	.78	.76	.81	.80
Training/test sets 2	.81	.79	.73	.77	.78
Training/test sets 3	.79	.81	.69	.75	.76
Coding of DV 1	.86	.81	.79	.84	.83
Coding of DV 2	.92	.80	.81	.81	.81
Six-month forecasts					
Base specification	.82	.78	.82	.76	.79
Terminal nodes	.80	.76	.81	.76	.78
Sample size	.83	.78	.78	.79	.79
Trees per forest	.82	.78	.82	.77	.79
Training/test sets 1	.79	.78	.81	.76	.78
Training/test sets 2	.73	.73	.76	.73	.75
Training/test sets 3	.88	.71	.81	.68	.79
Coding of dependent variable 1	.83	.78	.82	.78	.80
Coding of dependent variable 2	.83	.77	.83	.78	.79

Note: AUCs for our five random forests models. The top row in each panel reports AUCs for the base specification. We also report results with 10 rather than five terminal nodes (second row); 500 rather than 100 observations per tree (third row); 1,000,000 rather than 100,000 trees per forest (fourth row); a test set that begins January 1, 2009, January 1, 2010, or January 1, 2011 (fifth, sixth, and seventh rows, respectively); and alternate codings of the dependent variable (eighth and ninth rows) as described in the Online Appendix.

The escalation model generally out-performs the alternative models/specifications, and on that basis Blair and Sambanis conclude that theory does indeed provide better forecast accuracy when compared to models not informed by theory.

After establishing that the escalation model they advance out-performs all other covariate specifications, Blair and Sambanis move to show how the inclusion of structural risk variables – the set of structural variables identified by the Political Instability Task Force (PITF) that can contribute to an increased risk of civil conflict. The figure below is Table 2 from Blair and Sambanis (2020), which presents these findings.

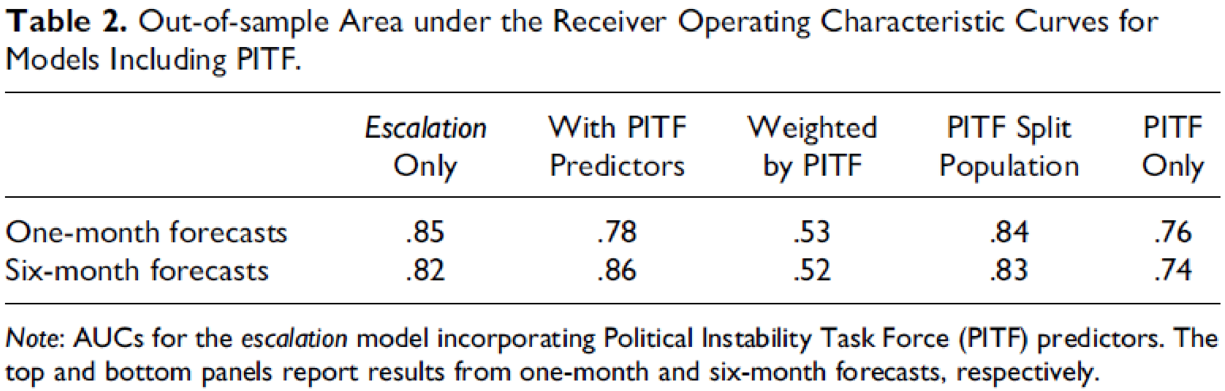

Table 2. Out-of-sample Area under the Receiver Operating Characteristic Curves for Models Including PITF.

	Escalation Only	With PITF Predictors	Weighted by PITF	PITF Split Population	PITF Only
One-month forecasts	.85	.78	.53	.84	.76
Six-month forecasts	.82	.86	.52	.83	.74

Note: AUCs for the *escalation* model incorporating Political Instability Task Force (PITF) predictors. The top and bottom panels report results from one-month and six-month forecasts, respectively.

Issues

Smoothed ROC curves

One issue that stood out as we read Blair and Sambanis (2020) was how smooth their Receiver Operating Characteristic curves where in Figure 1 on page 14. A Receiver Operating Characteristic curve displays the balance between the true positive rate and the false positive rate, across a range of discrimination thresholds – the value in which a zero to one probability is assumed to be an "onset". Given that the DV is civil conflict onset (1/0) and that this is a rare event, these curves should resemble a step-like function. Looking through the Blair/Sambanis replication code, we noticed that they use a smoothing option when producing their Receiver Operating Characteristic curves.

This is important because Blair and Sambanis use a common performance metric – Area Under the Receiver Operating Characteristic Curves (AUC-ROC) – as evidence that their escalation model out preforms all others. However, by smoothing the Receiver Operating Characteristic Curves, they are misrepresenting the true AUC-ROC scores of their various models. To help make this point clear, we replicate Figure 1 below. Rather than a two panel figure (one ROC plot their one-month-ahead forecasts and one for their six-months-ahead forecasts, as is presented in the article), we now include four panels: one panel for the two forecast windows with the smoothed curve, as they produce, and one panel for the two forecast windows that are not smoothed.

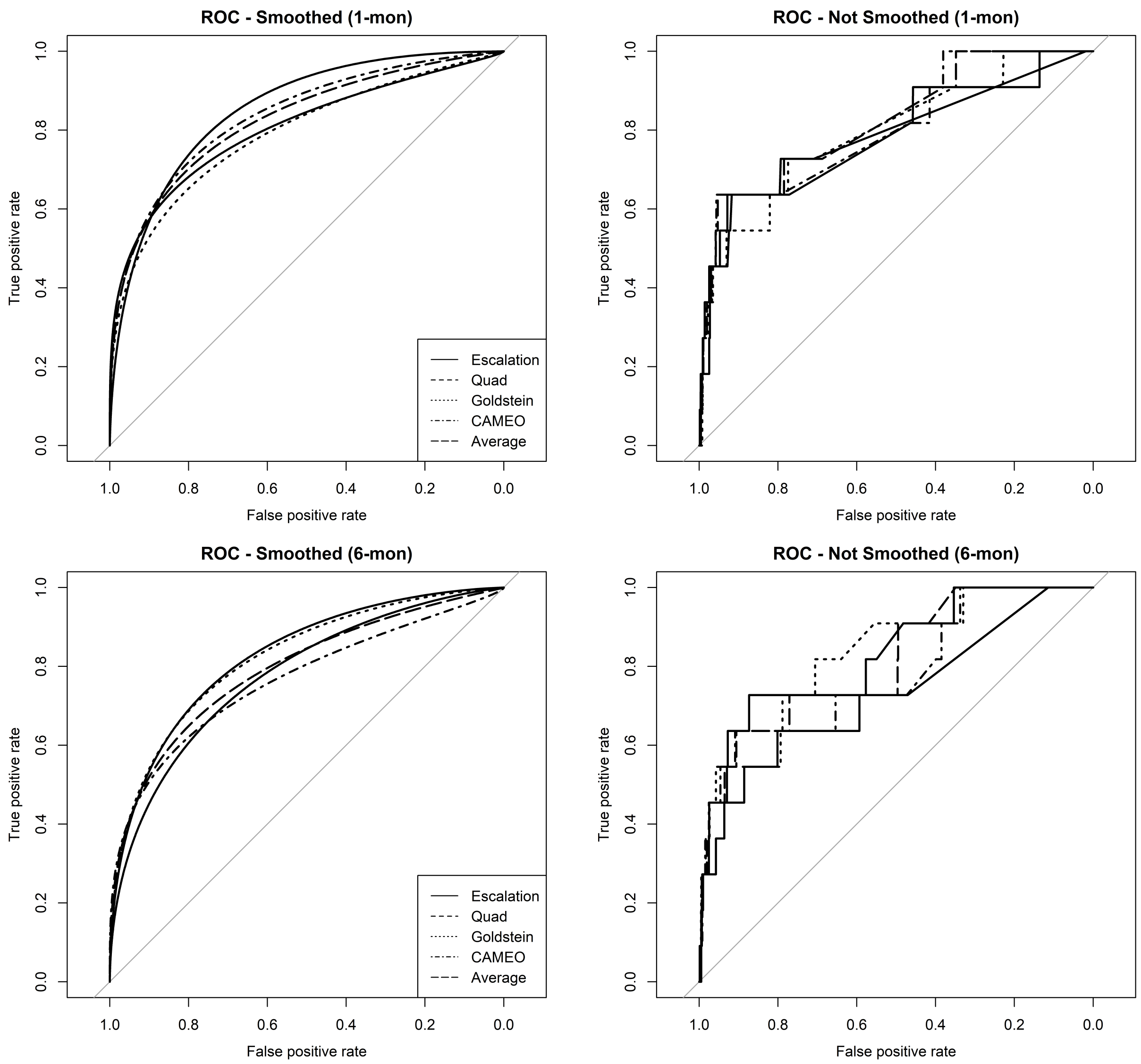
ROC - Smoothed (1-mon)
ROC - Not Smoothed (1-mon)
ROC - Smoothed (6-mon)
ROC - Not Smoothed (6-mon)
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
Escalation
Quad
Goldstein
CAMEO
Average

Table 1: Comparison of smoothed and not smoothed AUC-ROC values

horizon	ROC Smoothed	Escalation	Quad	Goldstein	CAMEO	Avg
Base Specification						
1 month	Yes	0.85	0.80	0.79	0.84	0.82
	No	0.79	0.79	0.79	0.81	0.82
6 month	Yes	0.82	0.77	0.82	0.77	0.79
	No	0.77	0.79	0.83	0.78	0.81

Comparison of the escalation specification to alternatives (Table 1)

Using these smoothed and non-smoothed curves, we then replicate Table 1 (page 13), but now the table reports the AUC-ROC scores for each model specification with and without the use of a smoother. If the non-smoothed AUC-ROC scores are drastically different from those of the smoothed AUC-ROC score (a difference of ≥ 0.05) or if the non-smoothed AUC-ROC scores are < 0.8, would suggest the findings presented in Blair and Sambanis (2020) are fragile to the use of a non-standard procedure.

As the table above shows, the results presented in Blair and Sambanis (2020) are fragile to the use of a smoothing option when calculating the AUC-ROC performance metric. Further, even when we rely solely on the replication files provided by Blair and Sambanis, we cannot replicate their findings exactly. For example, in our replication, the AUC-ROC score for the 1-month CAMEO model with the use of the smoother is 0.84; however, in the paper, the reported score is 0.82. This is most likely a function of the version of `R` Blair and Sambanis used to conduct their analysis. The current replication project was done using `R version 4.0.0`, which updated their random number seeding procedure, so small differences should be expected. (RM: I mention this last part because later when we replicate the PITF table, the results are very far off, beyond what we would expect from random fluctuations)

Comparison of the escalation model to structural extensions (Table 2)

Table 2 in the B&S paper (p. 19) compares the escalation model, based only on ICEWS-derived indicators, to alternatives that incorporate aspects of a structural model in different ways.

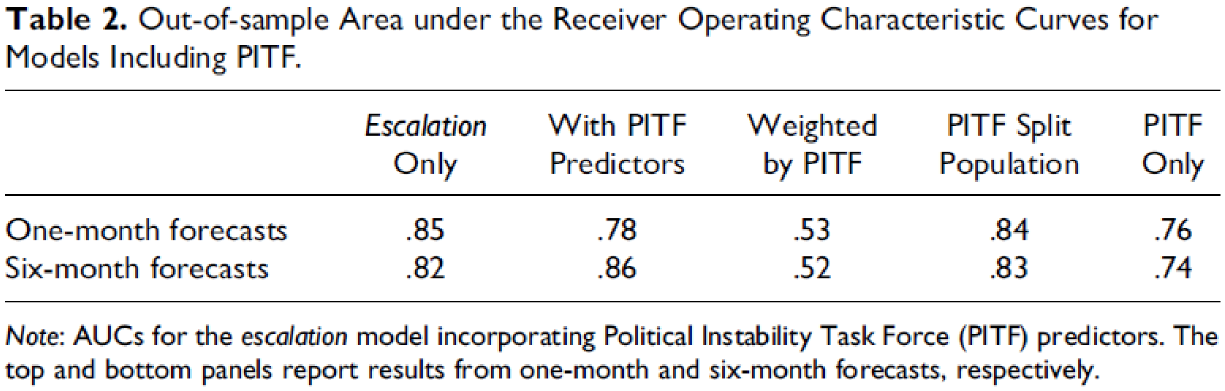

Table 2. Out-of-sample Area under the Receiver Operating Characteristic Curves for Models Including PITF.

	Escalation Only	With PITF Predictors	Weighted by PITF	PITF Split Population	PITF Only
One-month forecasts	.85	.78	.53	.84	.76
Six-month forecasts	.82	.86	.52	.83	.74

Note: AUCs for the *escalation* model incorporating Political Instability Task Force (PITF) predictors. The top and bottom panels report results from one-month and six-month forecasts, respectively.

On page 20:

> Overall, our results suggest that while measures of structural risk may improve predictive performance, the value they add is marginal and inconsistent. Of the approaches we test, the split-population analog is most promising, with AUCs of .84 and .83 over one-month and six-month windows, respectively.

Below is a recreation of this table with AUC values for both smoothed and the original ROC curves.

Table 2: Comparison of smoothed and not smoothed AUC-ROC values

Horizon	ROC Smoothed	Escalation Only	With PITF Predictors	Weighted by PITF	PITF Split Population	PITF Only
1 month	Yes	0.85	0.78	0.72	0.87	0.76
	No	0.79	0.79	0.71	0.79	0.76
6 months	Yes	0.82	0.85	0.67	0.83	0.75
	No	0.77	0.86	0.67	0.67	0.67

Note that we were unable to reproduce some of the AUC values in Table 2. Values that are notably different are 1-month "Weighted by PITF" and "PITF Split Population", and 6-month "PITF Split Population". For all three of these, the AUC-ROC values we obtained are much better than those in the original Table 2. The values for 6-month "PITF Only" and "With PITF Predictors" are slightly off but could be rounding errors.

Despite these reproduction differences, the original interpretation mostly still stands. The various efforts to incorporate structural variables into the forecasts largely do not work or do not add much predictive value. In the 6-month forecasts, the value added by simply adding the annual structural predictors to the existing random forest model is more pronounced than before, while the PITF split population approach fares much worse.

Although the substantive interpretation does not change much as a result of using ROC curve smoothing, the absolute AUC-ROC values between the smoothed and original versions do differe by quite a bit in several instances. The AUC-ROC values differ by 0.05 or more in 5 of the 10 cells in the original Table 2, and the direction of change is in all cases such that using smoothed ROC curves present more optimistic performance values.

Assessment of the 2016 forecasts

The predictions for the first half of 2016 are assessed in Table 4:

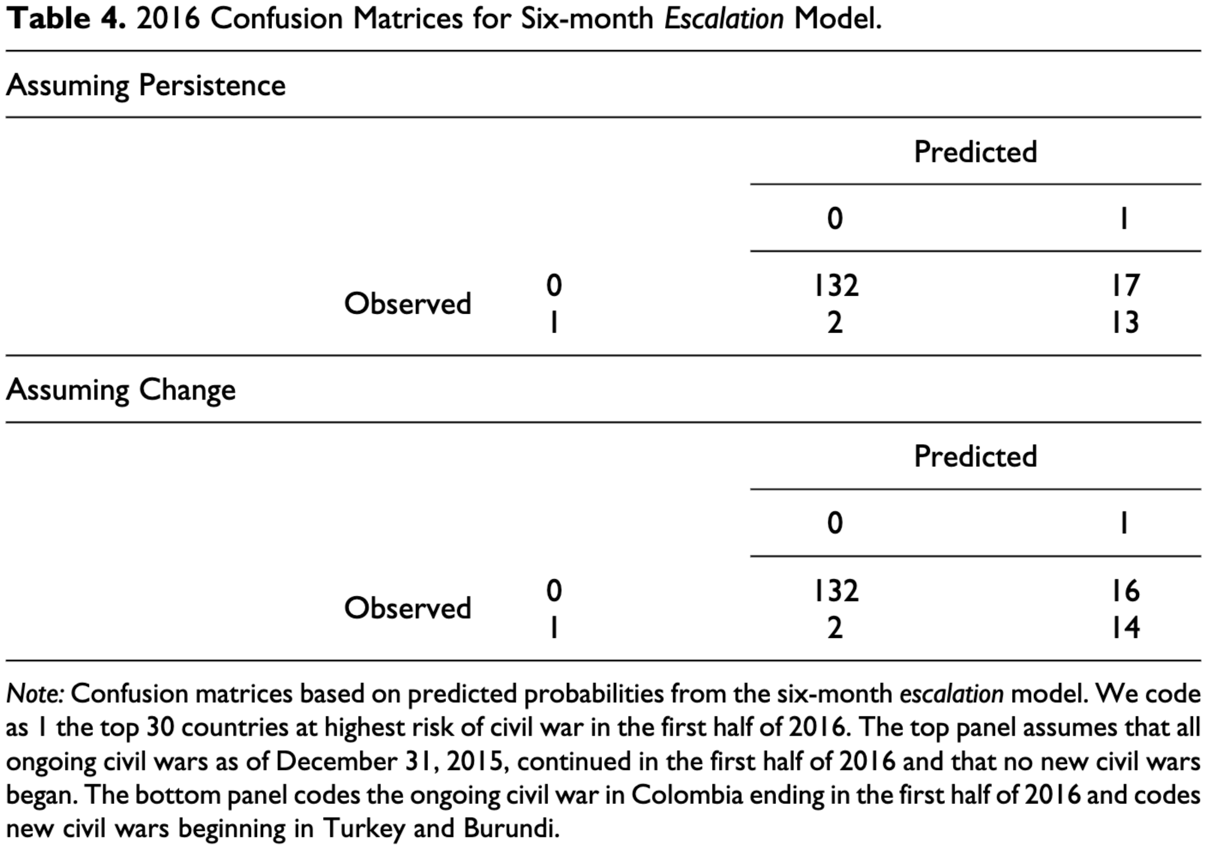

Table 4. 2016 Confusion Matrices for Six-month *Escalation* Model.

Assuming Persistence

		Predicted	
		0	1
Observed	0	132	17
	1	2	13

Assuming Change

		Predicted	
		0	1
Observed	0	132	16
	1	2	14

Note: Confusion matrices based on predicted probabilities from the six-month *escalation* model. We code as 1 the top 30 countries at highest risk of civil war in the first half of 2016. The top panel assumes that all ongoing civil wars as of December 31, 2015, continued in the first half of 2016 and that no new civil wars began. The bottom panel codes the ongoing civil war in Colombia ending in the first half of 2016 and codes new civil wars beginning in Turkey and Burundi.

The table presents 15 or 16 positive cases for 2016-H1, depending on a minor coding decision. This is the difference between the top and bottom confusion matrices. From the confusion matrices we can infer that the test data had a reported positive rate of around 9.5% for the first half of 2016. In contrast, the corresponding 6-month version of the data from 2001 to 2015, with 30 half-years, has in total 20 civil war onset events, for a positive rate of around 0.5%. This suggests that the civil war onset forecasts are erranously assessed using civil war incidence, not onset.

The relevant variables in the data are "incidence_civil_ns" and "incidence_civil_ns_plus1", which appears to be a 1-period lead version of the DV that is used in the actual prediction models. The incidence DV contains both 0/1 and missing values. By examining the pattern of missing values, it seems clear that this was originally an incidence variable indicating whether a country was at civil war in a given year or not, and which was converted to an onset version so that onsets retain the value of 1 but continuing civil war years are coded as missing.

By examining the code used to generated Table 4, we were able to confirm that the onset forecasts are assessed using incidence, not onset. In the file `6mo_make_confusion_matrix.do` on line 52, missing values in "incidence_civil_ns" are recoded to 1, thus reverting the onset coding of this variable back to incidence.

The correct confusion matrices when using observed onset (or the lack of it) are shown below.

There are no civil war onsets in the data for 2016-H1. Thus the recall values is undefined, while the precision is 0/30 = 0, compared to recall and precision values of 13/15 = 0.87 and 13/30 = 0.43 before. The the alternative coding ("Assuming Change") at the bottom of Table 4, there are 2 civil war onsets. Recall is 1 compared to 14/16 = 0.88 before, and precision is 2/30 = 0.07 instead of 14/30 = 0.47.

Another, minor issue or rather coding error, is related to using a lead version of the DV. With the lead version of the DV, "incidence_civil_ns_plus1", which is what the models are predicting, the predicted value for 2016-H1 actually indicates the risk of civil war onset in 2016-H2. In the Table 4 script referenced above,

Table 3: 2016 Confusion Matrices for Six-month Escalation Model.

header	Observed	0	1
Assuming Persistence	0	134	30
	1	0	0
Assuming Change	0	134	28
	1	0	2

the 2016-H1 predictions (for 2016-H2) are assessed using the raw DV, “incidence_civil_ns”, not the lead version. Essentially, the forecasts for 2016-H2 are assessed using observed outcomes for 2016-H1. In this case it probably doesn’t make a difference since both the raw DV and lead version for 2016-H1 do not have any positive values.

Discussion

When using the standard method of calculating AUC-ROC values, i.e. without smoothing the ROC curve beforehand, one of the principal B&S results reverses. The theory-driven escalation model does not in fact outperform the models that use alternative covariate specifications. In both the 1-month and 6-month versions of the base specification, all four other alternatives achieve equal or better AUC-ROC values compared to the escalation model. This includes the 1,159 covariate kitchen-sink CAMEO model.

We find no general change in the results regarding whether adding structural factors to the escalation model’s forecasts improves performance, although the gain from incorporating the PITF structural variables into the 6-month escalation model is much more pronounced. There is no reason to rule out the possibility that the escalation model alternatives would not also similarly gain from incorporating structural variables in their 6-month versions.

Furthermore, it appears that the forecasts for the risk of civil war onset in the first half of 2016 were erroneously evaluated using *incidence*, not *onset*. When correcting this mistake, the forecast evaluation changes significantly. There were no onsets in 2016, resulting in a precision of 0 and undefined recall; while with the manual re-coding reflected in the bottom portion of Table 4, the forecasts capture both civil war onsets for a recall of 1 and precision of 0.07.

Both of these issues are relatively objective problems. It seems relatively clear to us that using incidence to assess onset forecasts is simply incorrect. Using smoothed ROC curves to calculate AUC is non-standard to the extent of our knowledge, and the decision to use them is not discussed or justified in the paper. Maybe it is possible to defend this choice, but given that one of the principal findings rests on it, this should be done explicitly.

There are also other, more subjective concerns that we did not raise. For example, the paper does not explain how the random forest model hyperparemeters were selected. Although they do evaluate several changes in the hyperparameters as part of the robustness tests, the overall choice of hyperparameter values is unusual. Of note is also that this unusual approach only works (with warnings issued by the software) because B&S use random forests based on regression, not decision, trees. The concern to rule out would be that the hyperparameters are (over-) fit on the test data and that the reported performance values do not generalize.[ˆOn page 10, B&S write that they use classification trees; this is strictly speaking incorrect and the random forest is built using regression tress on the 0 or 1 outcome.]

We do not disagree with the overall point that theory is important for recognizing which variables or indicators might be important for forecasting. But it does not appear that B&S analysis demonstrates this, and that the paper’s conclusions to this effect rest on faulty results.

–>

References

Blair, Robert A., and Nicholas Sambanis. 2020. “Forecasting Civil Wars: Theory and Structure in an Age of ‘Big Data’ and Machine Learning.” *Journal of Conflict Resolution.*