

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 and Michael D. Ward

03 June 2020

Issues in the B&S replication:

- The AUC-ROC statistics are calculated using smoothed ROC curves
- The “Weighted by PITF” model is incorrectly implemented
- The “PITF Split Population” model is incorrectly implemented
- The forecasts are scored using incidence, not onset

Pre-registration should have had:

- The escalation model specification, prior to looking at data!
- The hyper-parameter tuning strategy.

With rare events, also a lot of variance over RNG state.

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 that:

- the first key finding depends on using AUC values calculated for *smoothed*, not the original, ROC curves
- the 2016 forecasts are incorrectly assessed using the occurrence (incidence), not onset, of civil war

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 on 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.

Smoothed ROC curves

Comparison of the escalation specification to alternatives (Table 1)

The results for the core analysis are summarized in Table 1 in Blair and Sambanis (2020), which we copied in Figure 1 below.

The first row, each, for the one-month and 6-month forecasts are the core base specification. The additional rows reflect various robustness tests.

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.

A small detail, which is mentioned in the text and visually obvious in the ROC curves they show in B&S Figure 1, is that these AUC-ROC values are computed using smoothed ROC curves. We used their replication code to recreate the ROC curves in their figure, which are shown in the left two panels in the figure below. The curves indeed appear to be very smooth. But in the underlying test data for which these curves are, there are only 11 positive events, out of 2,366 rows for the 6-month data and 13,748 rows for the 1-month data. The ROC curves in the two panels on the right are the original, non-smoothed versions, which correctly reflect the small number of positive events.¹

¹The diagonal lines in the ROC curves are there because there are groups of cases which have identical predictions, causing jumps in both the true and false positive rates at the same time.

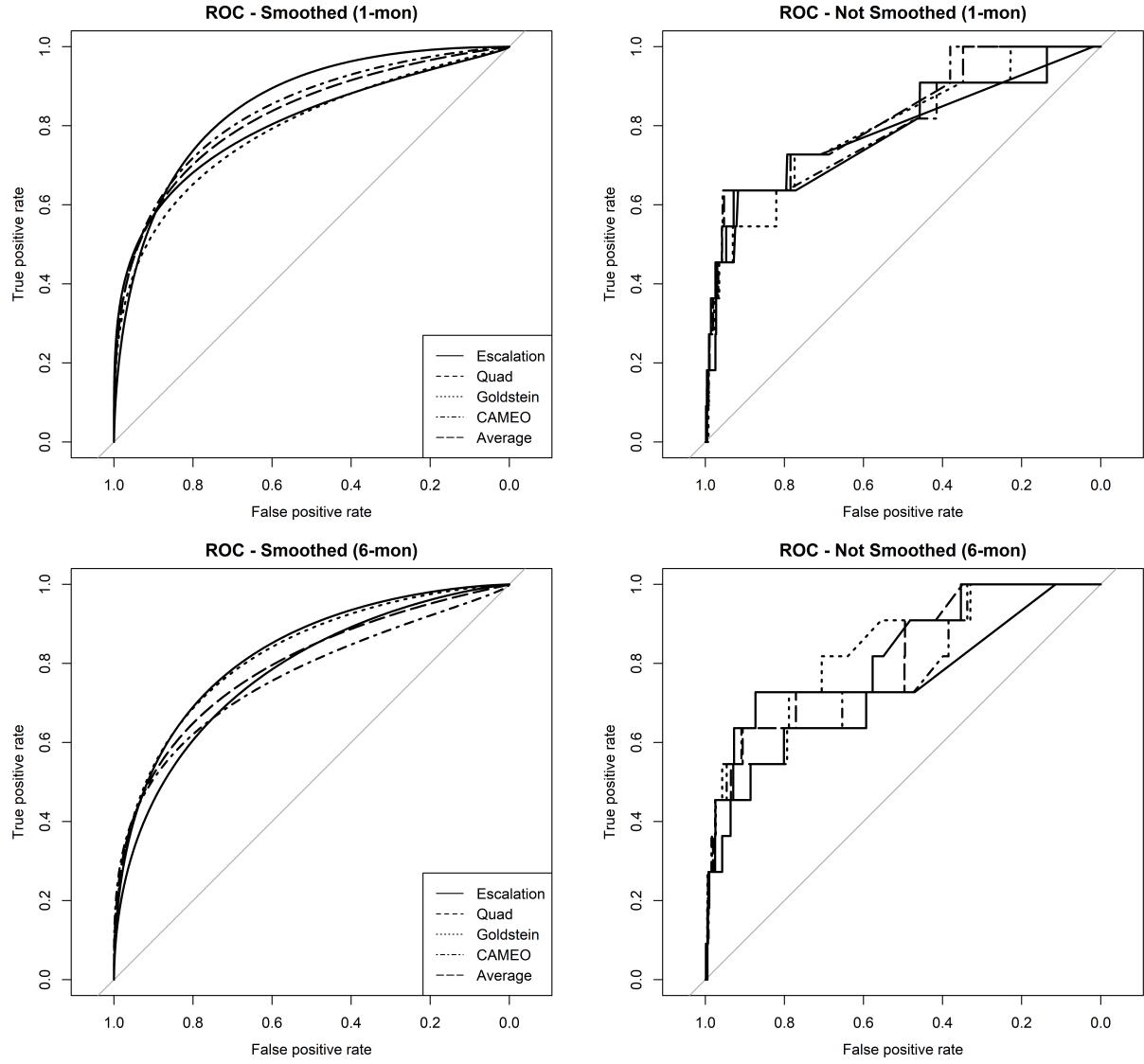


Figure 1: Replication of B&S Figure 1 with both smooth and non-smoothed ROC curves.

Using these smoothed and non-smoothed curves, we then replicate B&S 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. We focus on the core “Base specification” models only, not the additional robustness checks shown in the other rows in the table. The results of this are shown in our Table 1 above.

First, note that we were unable to exactly reproduce the AUC-ROC values for all models. For example, in B&S the 1-month CAMEO has a AUC-ROC of 0.82, while our reproduction has a AUC-ROC value (using the smoothed version) of 0.84. 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. Despite these small differences, the way we would interpret the smoothed-version results does not change. In the 1-month version of the Base Specification, the Escalation model outperforms all alternatives; in the 6-month version it outperforms all but one alternative and has equal performance with the Goldstein model.

When we calculate the AUC-ROC values using the original, non-smoothed curves, the findings change substantially. Rows 2 and 4 in Table 1 show the non-smoothed AUC-ROC values for the 1-month and 6-month versions, respectively. In the 1-month version, the escalation model is outperformed by 2 alternatives and even with the other 2. In the 6-month version, all 4 alternatives outperform the escalation model. When using the original, non-smoothed ROC curves, the escalation model has the worst overall performance out of the 5 models considered. The first key finding in the paper, that the escalation model is superior to the alternative models, thus depends on using smoothed ROC curves when calculating the AUC-ROC statistic.

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

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. A copy of B&S Table 2 is included in Figure 3.

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.

Table 2 is our recreation of B&S Table 2, copied in Figure 3, with AUC values for both smoothed and the original ROC curves.

Note that we were again 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 differ 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 B&S 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%. Since the positive event rate in the confusion matrices far exceeds the rate of observed civil war onsets in both the training and test data, we suspected that the forecasts were erraneously 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

Table 1: 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

for 2016-H1 actually indicates the risk of civil war onset in 2016-H2. In the Table 4 script referenced above, 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 also 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 hyperparameters 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.²

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 analysis.

²On page 10, B&S write that they use classification trees; this is strictly speaking incorrect and the random forest is built using regression trees on the 0 or 1 outcome.

Table 2: Test set AUC-ROC; replication of the original Table 1 with smoothed ROC curves.

Model	Escalation	Quad	Goldstein	CAMEO	Average
One-month forecasts					
Base specification	0.85	0.80	0.79	0.83	0.82
Terminal nodes	0.86	0.79	0.78	0.83	0.82
Sample size	0.85	0.81	0.70	0.86	0.84
Trees per forest	0.84	0.80	0.78	0.83	0.82
Training/test sets 1	0.86	0.78	0.75	0.81	0.80
Training/test sets 2	0.81	0.79	0.72	0.78	0.78
Training/test sets 3	0.79	0.80	0.69	0.75	0.75
Coding of DV 1	0.86	0.81	0.80	0.84	0.83
Coding of DV 2	0.92	0.81	0.81	0.82	0.81
Six-month forecasts					
Base specification	0.82	0.78	0.82	0.77	0.79
Terminal nodes	0.80	0.75	0.81	0.75	0.77
Sample size	0.83	0.78	0.78	0.78	0.79
Trees per forest	0.82	0.78	0.82	0.77	0.79
Training/test sets 1	0.79	0.77	0.81	0.75	0.78
Training/test sets 2	0.72	0.73	0.77	0.74	0.75
Training/test sets 3	0.88	0.70	0.81	0.68	0.79
Coding of DV 1	0.83	0.77	0.82	0.79	0.80
Coding of DV 2	0.83	0.77	0.82	0.79	0.79

Table 3: Test set AUC-ROC; replication of the original Table 1 *without* smoothed ROC curves.

Model	Escalation	Quad	Goldstein	CAMEO	Average
One-month forecasts					
Base specification	0.78	0.78	0.79	0.80	0.82
Terminal nodes	0.79	0.78	0.78	0.81	0.82
Sample size	0.79	0.80	0.74	0.82	0.84
Trees per forest	0.78	0.78	0.79	0.81	0.82
Training/test sets 1	0.78	0.76	0.76	0.79	0.80
Training/test sets 2	0.75	0.77	0.73	0.76	0.78
Training/test sets 3	0.70	0.79	0.69	0.73	0.74
Coding of DV 1	0.80	0.79	0.80	0.82	0.83
Coding of DV 2	0.80	0.82	0.78	0.83	0.81
Six-month forecasts					
Base specification	0.77	0.78	0.83	0.79	0.81
Terminal nodes	0.77	0.78	0.82	0.78	0.79
Sample size	0.78	0.77	0.80	0.80	0.82
Trees per forest	0.77	0.79	0.83	0.78	0.81
Training/test sets 1	0.75	0.79	0.82	0.78	0.80
Training/test sets 2	0.70	0.75	0.78	0.75	0.77
Training/test sets 3	0.85	0.72	0.84	0.71	0.81
Coding of DV 1	0.77	0.80	0.83	0.79	0.82
Coding of DV 2	0.80	0.78	0.84	0.80	0.81

Table 4: Replication of BS Table 2: Test set AUC-ROC for Escalation model with and without structural PITF contribution

Escalation Only	With PITF Predictors	Weighted by PITF	PITF Split Population	PITF Only
One-month forecasts				
0.69	0.71	0.69	0.81	0.67
Six-month forecasts				
0.79	0.87	0.80	0.69	0.71

Note: Differences from the original BS Table 2: (1) AUC-ROC values are computed on the common subset of cases, meaning that N is equal in each row; (2) AUC-ROC values are computed using original, non-smoothed ROC curves.

Table 5: Number of valid test predictions for each cell in BS Table 2

Horizon	Escalation Only	With PITF Predictors	Weighted by PITF	PITF Split Population	PITF Only
Original model-specific cases					
1 month	13748	13155	13461	13748	13510
6 months	2366	2264	2317	2366	2265
Cases adjusted to common subset					
1 month	9811	9811	9811	9811	9811
6 months	1915	1915	1915	1915	1915

Table 6: Replication of BS Table 2 with smoothed/original ROC and with original varying N cases or adjusting for common case set with constant N

	Smoothed ROC	Escalation Only	With PITF Predictors	Weighted by PITF	PITF Split Population	PITF Only
Original model-specific cases						
<i>1 month</i>						
Yes		0.85	0.76	0.80	0.81	0.77
No		0.78	0.77	0.80	0.80	0.75
<i>6 months</i>						
Yes		0.82	0.86	0.81	0.79	0.74
No		0.77	0.86	0.81	0.82	0.74
Cases adjusted to common subset						
<i>1 month</i>						
Yes		0.74	0.68	0.69	0.80	0.69
No		0.69	0.71	0.69	0.81	0.67
<i>6 months</i>						
Yes		0.85	0.86	0.79	0.70	0.71
No		0.79	0.87	0.80	0.69	0.71

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*.