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

04 May 2020

We highlight some issues with the analysis in the recent Blair and Sambanis paper that undermine their main empirical finding, that a theoretically-informed model can forecast civil war onsets better than more inductive alternatives.

- The key findings depend on the decision to use smoothed rather than original ROC curves when calculating area under the ROC curve (AUC-ROC) fit statistics. When using regular, not smoothed ROC curves, the findings reverse.
- The 2016 forecasts for the risk of civil war onset appear to be evaluated using civil war *incidence* not *onset*. This distorts the assessment of how accurate they were. (TBD)
- The way the random forest models are fitted by B&S is very unorthodox.[1] We evaluate two questions: (1) does the key model's superior test forecast performance generalize or is it a result of inadvertant 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? (TBD)

Review of the Blair and Sambanis paper

Blair and Sambanis (2020) (B&S) argue that theory contributes to forecasting accuracy even when using non-parametric machine learning models 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 1,100 covariates.

Their data consist of country-month observations covering 2001 to 2015. The main outcome is civil war onset, measured using Sambanis' data. 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. They 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:

- 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.

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

To assess forecast accuracy, B&S:

1. Split the training data into training (2001 - 2007) and test (2008 - 2015) sets.
2. Estimate 5 random forest models with the specifications from above.
3. Create out-of-sample (OOS) predictions from each model for 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. B&S also conduct a series of robustness tests that (1) vary hyperparameters of the RF model, (2) change the split year for the trainin/test data split, and (3) alternative codings of the dependent variable.

The results for this analysis are summarized in Table 1 in the B&S paper, which we reproduce 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 outperforms the alternative models/specifications, and on that basis B&S conclude that theory does indeed provide better forecast accuracy when compared to models not informed by theory.

Before listing the primary issues we identified, we should note that we are partial to their approach. Theory is important. It is helpful for forecasting problems, especially with regards to identifying which variables to

include in a model. This is particularly true when the goal is to assess how ... (RM: I want to add something here about identifying potential policy interventions that can help advert onset...)

Issues

Smoothed ROC curves

```
## Parsed with column specification:
## cols(
##   model = col_character(),
##   horizon = col_character(),
##   specification = col_character(),
##   smoothed = col_double(),
##   original = col_double()
## )
```

model	horizon	specification	smoothed	original
base specification	1 month	escalation	0.85	0.79
base specification	1 month	quad	0.80	0.79
base specification	1 month	goldstein	0.79	0.79
base specification	1 month	CAMEO	0.84	0.81
base specification	1 month	avg	0.82	0.82
base specification	6 month	escalation	0.82	0.77
base specification	6 month	quad	0.77	0.79
base specification	6 month	goldstein	0.82	0.83
base specification	6 month	CAMEO	0.77	0.78
base specification	6 month	avg	0.79	0.81

Using AUC-ROC values from un-smoothed, original ROC curves reverses the original conclusions from B&S Table 1. In both the 1 month-ahead and 6-month ahead forecasts, the escalation model is outperformed by alternative models.

Assessment of the 2016 forecasts

The main data set covering 2001 - 2015 contains 20 civil war onsets. But there are 15 "Persistence" cases (top of table 4) and 16 "Change" cases (bottom of table 4)

Random forest model tuning

3. Unorthodox RF hyperparameter choices that may be driving the OOS test prediction results. There are two issues here. One is that the RF models are so unorthodox. I'm not sure that this in itself is a problem. But the second potential problem is that the hyperparameter values may not have been tuned in a way that works the best for the non-escalation model, and especially the CAMEO model with 1,100+ features.

- Improperly tuning of models
  - RM: In a twitter DM, I asked Blair to clarify – "[I] was wondering how you all came to your tuning procedure for the RF model hyper-parameters?" He responded "mostly trial and error, honestly. ... Trial and error was on early data. Forecasts were for much later data"
- Using a regression framework in a classification problem
- Rounding of AUC/ROC scores
  - AB: smoothing or rounding? I think rounding is ok.
- Treating the output of a RF regression model as Pr() – (RM: I guess this is akin to a Linear Probability Model, but I'm not sure if this is possible in a RF framework
  - It's hacky but it works. The resulting forecasts are within the 0-1 interval.

The table below shows the randomForest package default hyperparameter values for a binary classification problem like the one at hand and compares them to the B&S base settings. Some hyperparameters have heuristics that determine the default value based on characteristics of the input data; we note these in the second column and what the realized default setting would be for the basic Escalation model (first row in Table 1 in the paper) in the 3rd column. The last column has the settings B&S use.

Hyperparameter	Default heuristic	Default values (Escalation)	B&S value
type		classification	regression
ntree		500	100,000
mtry	`floor(sqrt(ncol(x)))`	3	3
replace		true	false
sampsize	`nrow(x)` if replace, else `ceiling(.632*nrow(x))`	11,869	100
nodesize	1 for classification	1	1
maxnodes		null	5

It is worth noting that commonly the ntree, mtry, and nodesize parameters are the main parameters tuned; of these three in the B&S specification only 1, the ntree parameter, deviates from the default settings.

In any case, there is a stark contrast in the default RF settings and the way B&S use the RF models. The default approach is to train a relatively small number (ntree; 500) of classification trees, but where each tree is fairly big in that it is trained on data that has the same number of rows as the training data, albeit sampled with replacement (replace is true; sampsize is 11,869), and allowed to grow fairly deep (this is governed by nodesize (1), which is the minimum size a terminal node must have). In contrast, B&S grow very extensive forests with a large number of trees (100,000 compared to 500), but each tree is very small and shallow; only a 100 rows are sampled from the training data for each tree, and the trees are constrained to at most 5 terminal nodes (maxnodes).

This approach only works due to the other unorthodox choice, which is to use regression, not classificaton, trees. Trying to use classification trees with the other parameter settings in fact does not work at all because it is almost guaranteed that a sample of 100 from the 11,869 training data rows with 9 positive cases will only include 0 (negative) outcomes in the sample. As it is, using regression with a 0 or 1 outcome produces warnings when estimating the models:

```
Warning message:
In randomForest.default(y = as.integer(train_df$incidence_civil_ns_plus1 ==  :
  The response has five or fewer unique values.  Are you sure you want to do regression?
```

While unorthodox, the approach does work. It produces predictions that are within the 0 to 1 interval. The concern is that the settings work well only for the escalation model, and only for the particular test set at hand. Specifically, we wonder (1) if the hyperparameter settings only work well for the particular test set chosen, and (2) if the CAMEO model with 1,100+ features, compared to 10 or less for the other specifications, would perform better with a more explicit tuning procedure.

Are the RF hyperparameters overfit to the test set?

Does the base specification's good performance in the test set generalize?

One initial piece of evidence is already available from Table 1. In addition to the test set starting 2008, B&S also evaluate test sets starting in 2009, 2010, and 2011 (sets 1, 2, and 3 in Table 1). For the 1 month models, when we compare the base specification performance of each model to the performance in the alternate test sets, the AUC-ROC increases in 2 (of 15) cases, is the same in 0 cases, and decreases in 13 cases. For the 6 month models, it increases in 1 case, is the same in 3 cases, and decreases in 11 cases. Altering the test set thus generally shows reduced performance. The table below shows cross-validated training and OOS test AUC-ROC for the original 1 month escalation model, a RF model with default hyperparameter settings, and a tuned RF model with 10,000 trees and otherwise the default settings. The first set of results are

out of sampel results from repeated cross-validations performed on the training data. We show the average AUC-ROC, it's lower and upper 95% CI, obtained via bootstrapping, and the standard deviation of the distribution of resampled AUC-ROC values.

Because there are only 9 positive cases in the whole training data, and thus to ensure that any given split will include at least 1 positive case in each data partition, we use 2-fold CV, i.e. splitting the original training data into equally-sized new training and validation sets. This is repeated 21 times for a total of 42 OOS performance samples for each model. The last column shows the OOS performance on the original test set. The first value in this column corresponds to the base escalation model results reported Table 1 in the B&S paper when *not* using smoothed ROC curves.

Model	Avg_CV_ROC_AUC	ci_lower	ci_upper	SD_CV_ROC_AUC	Test_ROC_AUC
Escalation, 1mo	0.68	0.63	0.72	0.15	0.78
Modified Escalation, 1mo	0.62	0.58	0.66	0.12	0.59
Tuned Escalation, 1mo	0.67	0.63	0.71	0.12	0.58

The results show that:

- Both of the alternative RF models are able to achieve roughly similar OOS performance in the training data split. T-tests comparing model 1 to the other models fail to reject the null hypothesis at a 95% confidence level. (A t-test comparing model 1 and 2 average AUC-ROC just barely fails to reject the null hypothesis (p slightly above 0.05).)
- Only the base escalation model is able to achive good test performance; the other two models, despite achieving similar training data performance, have significantly lower test performance.

This suggests that the base RF specifications are (over-)fit to the test data.

4. Design choices

Rick, I think the points below are more subjective choices. The first two points above are IMO potentially objectively incorrect technical errors that could undermine the original B&S findings. The 3rd point about the RF models is somewhere between objective error and subjective choice. I'm not sure. But the points below I would say are subjective choices that could alter their findings and that I think are justifiable criticisms, but I think it'll be easier for someone to push back on this if they wanted to.

- Their train/test approach
  - 5-year forecasts for sep plots and AUC scores)
- Smoothing of AUC/ROC scores
- Lack of yearly test forecasts in favor of a single 5-year test forecast
- The lack of procedures to account for rare events in an RF model

To-do

1. Full replication
   - Can we get their results
   - RM: maybe I can go through their code and bring it "up-to-date" i.e. make it tidy...
     - AB: that sounds like a lot of work
   - Figure out how they built table 4 (It looks like this is done in STATA)
   - Check to see if it matter that they are using regression for a classification problem
     - AB: i tried switching the models as they have them just from regression to classification and it breaks because the data samples end up being all 0's.
2. Address tuning
   - Does properly tuning the CAMEO model improve its performance?
3. Address AUC/ROC smoothing...
   - How drastic is the change; does it change their conclusions?
4. Run a yearly test forecast – as is the current standard in TSCS forecasts (Is it the current standard?)
5.

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*.