Research Article

Research and Politics
January-March 2020: 1–7

DOI: 10.1177/2053168020905487
journals.sagepub.com/home/rap
SAGE

Predictive inference with random forests: A new perspective on classical analyses

Richard J. McAlexander[1] and Lucas Mentch[2]

Abstract

Despite the number of problems that can occur when core model assumptions are violated, nearly all quantitative political science research relies on inflexible regression models that require a linear relationship between dependent and independent variables for valid inference. We argue that nonparametric statistical learning methods like random forests are capable of combining the benefits of interpretability and flexibility. Recent work has shown that under suitable regularity conditions, averaging over predictions made by subsampled random forests produces asymptotically normal predictions. After estimating the variance, this property can be exploited to produce hypothesis tests and confidence intervals analogous to those produced within a parametric framework. We demonstrated the utility of this approach by replicating an important study on the determinants of civil war onset and show that subtle nonlinear relationships are uncovered, providing a new perspective on these ongoing research questions.

Keywords

Random forests, nonparametric, nonlinear

Introduction

Most quantitative political science relies on parametric statistical models that require strict model assumptions for valid inference, usually that the dependent variable is linearly related to the independent variables.[1] Obtaining unbiased coefficient estimates in a linear model requires specifying the exact correct model form, including any possible interaction or higher-order polynomial terms. Uncertainty about the correct functional form provides researchers wide latitude in testing interactions and higher-order polynomials, which can sometimes create multiple-testing issues. In addition, subtle nonlinear relationships – even if they are monotonic and not curvilinear – between the response and a predictor can severely bias coefficient estimates (including flipping the sign) for variables that do have a linear relationship, as was shown in Achen (2005) over a decade ago. Despite these potential issues and the existence of flexible statistical and machine learning alternatives, applied political science research continues to rely almost exclusively on classical approaches in order to perform inference. This reliance on more rigid tools is likely due to the perceived lack of interpretability and conventional variance estimates that available within a more flexible learning context.

In this paper, we argue that nonparametric ensemble methods – random forests (Breiman, 2001) in particular – can naturally accommodate nonlinearity in the data. We demonstrate that recent theoretical and methodological developments allow for some degree of interpretability including the use of null hypothesis significance testing via the production of conventional variance estimates. There has been a substantial amount of work in recent years that has demonstrated useful and desirable statistical properties of random forests when individual trees (or, more generally, base learners) are built with subsamples of the training data. Scornet et al. (2015) demonstrated that random forests are consistent whenever the underlying regression function is additive. Wager et al. (2014) applied the infinitesimal jackknife estimate of variance derived in Efron (2014) in order to estimate the variance of random forest predictions. Mentch and Hooker (2016) demonstrated that predictions from subsampled random forests can be viewed as infinite-order forms of classical U-statistics and as a result, are asymptotically normal under certain regularity conditions.

[1]Columbia University, USA
[2]University of Pittsburgh, USA

Corresponding author:
Richard J. McAlexander, Columbia University, 420 W 118th Street, New York, NY 10027, USA.
Email: rjm2187@columbia.edu

Here the authors provided a general set of conditions noting that any type of base learner (tree) that satisfies those technical conditions can be used. Peng et al. (2019) improved this result, weakening the necessary conditions somewhat and allowing larger subsamples to be used. Wager and Athey (2017) advocated for a particular class of base learners that conform to *honesty* and *regularity* conditions, meaning that trees are constructed on different data than that used to produce the estimates in the terminal nodes, and also that splits in the tree allow for a certain minimal percentage of observations to fall on each side. As most of these conditions are rather technical, in the following sections, we will use the phrase "suitably constructed base learners" to refer in general to any tree or other base learner that obeys the necessary conditions for asymptotic normality as put forth in the previous works just described.

Ultimately, these results provide a means by which the uncertainty in random forest predictions may be formally quantified. Variance estimates for the predictions made by subsampled random forests share the desirable properties of conventional variance estimates for linear models. These variance estimates can be used in a standard testing framework to evaluate hypotheses about the predictive significance of a variable. Though this more flexible approach comes with an increased computational cost, we stress that the perception that one must give up flexibility in the choice of model in order to interpret results and conduct inference is quickly becoming an outdated notion.

Since random forests are a nonparametric method, there is no testing of the significance of a particular term-specific parameter as is the case in standard parametric models like linear regression. Instead, the predictions themselves take the place of these conventional parameters. While testing whether a variable makes a significant contribution to predicting an outcome is not the conventional approach utilized in political science research, we argue that it is in line with fundamental scientific principles. Indeed, though perhaps more subtle, even within a traditional linear model context, terms are only deemed "significant" when their inclusion in the model creates a more accurate fit than could be expected by random chance. Since real-world data generating processes may deviate substantially from the assumptions of linear models, the prime justification of using linear models has been their interpretability. Our goal here is to demonstrate that recent developments are beginning to allow for an analogous means by which the same inferential questions can be asked and answered in flexible machine learning contexts.

We note that variables with strong predictive power may not be statistically significant when included in a linear model if their relationship to the dependent variable is nonlinear and that nonlinearity is not taken into account in the model specification. To demonstrate the utility of our approach, we replicated Fearon and Laitin's (2003) (FL) central model, testing whether gross domestic product (GDP) or fractionalization is more important for explaining civil war onset. FL's work has led many scholars to debate the relevance of ethnolinguistic fractionalization (ELF) and religious fractionalization on civil war onset after they contradicted the conventional wisdom by presenting models where there was no statistically significant effect of these variables. Our tests showed that both ELF and religious fractionalization are statistically significant when tested individually.[2] We suggest that one reason for this difference is the nonlinear nature of the relationship between fractionalization and conflict onset.

Our main contribution in this article is to show that random forests can be used for inference, and to demonstrate that hypothesis testing using random forests can produce results analogous to the hypothesis tests typically conducted with simpler parametric models. The implication is that the perceived gap between the interpretability and inferential utility of linear parametric models and nonlinear nonparametric models is smaller than most presume.

A statistical context for random forests

We now begin to define the random forest procedure with more mathematical formality. Suppose we have data of the form $\mathrm{D} = \{Z_1 = (\mathrm{X}_1, Y_1), ..., Z_n = (\mathrm{X}_n, Y_n)\}$ that are independent and identically distributed (iid) where each observation Z_i is an ordered pair consisting of a vector of independent variables (features) $\mathrm{X}_i = (X_{1,i}, ..., X_{p,i})$ and a dependent variable (response) Y_i. It is convenient to assume our response is continuous and real valued, $Y_i \in \mathbb{R}$, though we note that in order for the theory to remain valid, we need only that the predictions are real valued. This allows us to work, for example, in the context where our response is binary ($Y_i \in \{0,1\}$), but where we might predict the probability of a positive outcome, $P[Y = 1 \mid X = x]$.

Now consider a new location (set of feature values) x_0 where we might like to utilize the original dataset D to make a prediction. The prediction at x_0 from a random forest consisting of m trees, each built with a subsample of size k, can be written as

$$\hat{y}_0^{RF} = \hat{\theta}_{RF}(\mathrm{x}_0) = \hat{\theta}_{RF} = \frac{1}{m}\sum_{i=1}^{m} T_{w_i}(\mathrm{x}_0; Z^*_{i,1}, ..., Z^*_{i,k})$$

where T_{w_i} denotes the tree function with randomization parameter w_i and $(Z^*_{i,1}, ..., Z^*_{i,k})$ denotes the subsample of k data points used to construct the i^{th} tree in the random forest.

In order to perform statistical inference on the predictions from a random forest, we must be able to say something about the distribution of $\hat{\theta}_{RF}$. Given the complexity of an estimation procedure like random forests, one might first consider resorting to a nonparametric approach like the bootstrap (Efron, 1979) in order to obtain an approximate

sampling distribution. However, this presents a substantial computational challenge: as random forests themselves already involve a resampling step, bootstrapping such a procedure adds another level of resampling. That is, drawing B bootstrap samples typically requires calculating B bootstrap estimates. However, in the case of random forests, since each estimate depends on constructing m base learners (tree estimates), a total of $B \times m$ estimates must be calculated. Even then, we would prefer our inference to concern all $B \times m$ outputs rather than only the original m.

While some work has been done to investigate more computationally friendly approaches (see Sexton and Laake (2009) for one such popular example), recent results have provided a more direct approach. In particular, Mentch and Hooker (2016) showed that for suitable base learners, so long as the subsample size k grows slow relative to the training set size n, random forest predictions are asymptotically normal with mean $\mathbb{E}(\hat{\theta}_{RF})$ and variance

$$\frac{k^2}{m\alpha}\zeta_1 + \frac{1}{m}\zeta_k \tag{1}$$

where $\alpha = \lim_{n\to\infty} n/m$ and ζ_1 and ζ_k are variance parameters associated with classic U-statistics. Note that α here is simply the limiting ratio of the size of the dataset n to the size of the ensemble (number of base learners) m. In practice, we can use a simple plug-in estimate for α based on the size of the training set and the number of base learners actually built. Mentch and Hooker (2016) provided a brute force method for estimating this variance, though when very large random forests are built – the number of trees m is much larger than n – the infinitesimal jackknife procedure discussed in Wager et al. (2014) may also be used.

Confidence intervals for random forest predictions

Given this central limit theorem and feasible methods for estimating the variance, pointwise confidence intervals for random forest predictions may be readily constructed by simply estimating the parameters associated with the distribution and pulling out the desired quantiles. Importantly, we pause here for a moment to stress that such confidence intervals are valid for and centered at the *expected prediction* from a random forest and not necessarily the true value of the underlying regression function. In general, in order to guarantee probable coverage of the true value, stronger conditions on the base learners are needed; see Wager and Athey (2017) and Peng et al. (2019) for more details. For more general classes of base learners, if we consider the generic regression setup whereby $Y_i = f(\mathrm{X}_i) + \epsilon_i$ for some error ε_i, these intervals are centered at $\mathbb{E}(\hat{f}(\mathrm{x}_0))$ – where $\hat{f}$ denotes the random forest estimator – which may not necessarily be equal to $f(\mathrm{x}_0)$.

At first glance, this fact may seem disappointing and even appear to make such intervals prohibitive to practical usage, at least without strong conditions on the base learners. Note however that while perhaps more subtle, similar restrictions exist even with simple and more traditional statistical models. Consider, for example, the linear model whereby we assume a relationship of the form

$$Y_i = f(\mathrm{X}_i) + \epsilon_i = X_{1,i}\beta_1 + \cdots + X_{p,i}\beta_p + \epsilon_i. \tag{2}$$

For low dimensional problems ($p < n$), we can use ordinary least squares methods to calculate $\hat{\beta}$ and form predictions $\hat{f}(\mathrm{x}_i) = \sum_{j=1}^{p} \mathcal{X}_{j,i}\hat{\beta}_j$. Both the standard confidence intervals for $\mathbb{E}(Y_i \mid \mathrm{X}_i)$ and the standard prediction intervals for y_0 at a given location x_0 are constructed assuming that the relationship in equation (2) holds exactly. If the true relationship between the features and response is something different so that $\hat{f}(\mathrm{X}_i) \nrightarrow f(\mathrm{X}_i)$, then in an exactly analogous fashion, in general, the resulting intervals are valid only for the specified linear approximation rather than the ground truth. Thus, while demonstrating the consistency of random forests for large classes of underlying regression functions remains an important and worthwhile pursuit in the statistics and machine learning communities, the current scarcity of robustness proofs should not preclude the practical usage of these known distributional results for statistical inference purposes, especially when the predictive superiority of random forests can be well established empirically.

Hypothesis testing with random forests

In addition to providing confidence intervals for predictions, Mentch and Hooker (2016) also provided a means by which formal hypothesis tests for feature significance can be carried out. Following the previous notation, assume that we have a total of p features $X_1,...,X_p$ and now also suppose that we have a collection of test points (locations) $\mathrm{D}_{\text{test}} = \{\mathrm{x}_1,...,\mathrm{x}_N\}$ where we want to make predictions and evaluate feature importance. To assess the effect of the feature X_1, we can consider building two separate random forests; one *original* forest in which all features are used and another *reduced* forest, which differs only in that it is not allowed to use the feature of interest, X_1. As convenient shorthand, define $\widehat{RF}_{\text{orig}}(\mathrm{x})$ and $\widehat{RF}_{\text{red}}(\mathrm{x})$ to be the predictions at x made by the original and reduced random forest respectively. Then we can define

$$\widehat{D}(\mathbf{x}) = \widehat{RF}_{\text{orig}}(\mathrm{x}) - \widehat{RF}_{\text{red}}(\mathrm{x})$$

as the difference in predictions between the two forests in order to form $\widehat{D} = (\widehat{D}(\mathrm{x}_1),...,\widehat{D}(\mathrm{x}_N))^T$, the vector of prediction differences at each point in the test set. The theory

provided in Mentch and Hooker (2016) provides that, asymptotically, $\widehat{D}^T \hat{\Sigma}_D \widehat{D}$ follows a χ^2 distribution with N degrees of freedom where the empirical covariance $\hat{\Sigma}_D$ is simply the multivariate analogue of the variance defined in equation (1). Thus, $\widehat{D}^T \hat{\Sigma}_D \widehat{D}$ may be used as a test statistic in assessing whether X_1 is predictively significant: whenever the statistic falls above the $1-\alpha$ quantile of the χ^2_N distribution, we reject the null hypothesis that X_1 is not important and conclude that the feature makes a significant contribution to the predictions at some (at least one) point in D_{test}. Note that in theory, at the extreme, this thus implies that the test may reject the null hypothesis in the case where predictions are identical at all test locations but one. In order for this to occur in practice however, the magnitude of the difference at that location must be quite substantial, especially when several test points are used. Nonetheless, this serves as an important reminder that test locations should ideally be chosen in dense areas of the feature space rather than in regions where there are few to no observations, so as to prevent deciding importance based on extrapolation by the random forests.

Due to the inherent volatility of tree-based estimators, the authors note that this procedure may sometimes produce a high number of false positives depending on the structure of the data. That is, because trees in random forests are typically grown to near maximum depth, even noise features (with no true relationship to the response) can alter predictions enough to be considered significant. As a robust practical workaround to remove this susceptibility, Mentch and Hooker (2016) recommend using a permuted version of X_1 in the reduced random forest instead of simply removing it. A significant result from this amended test allows for a more robust and decisive conclusion: not only is the feature predictively significant, but it is more significant than could be expected by simply adding random noise to the model. The issue can also be remedied replacing the original variable with a *knockoff* variable (Barber et al., 2015) rather than a permuted version.

These procedures naturally extend to more general hypothesis tests. To test whether a group of features is significant, each feature in that group can simply be removed or permuted in the reduced random forest. Furthermore, in follow-up work, Mentch and Hooker (2017) showed that when the set of test points D_{test} is arranged in a grid structure, this testing procedure can also assess the additivity of the underlying regression function to determine whether interactions exist between specific features. Finally, we note that because the described testing procedure requires the estimation of an $N \times N$ covariance matrix, large test sets may make the procedure computationally prohibitive. In very recent work, Coleman et al. (2019) developed an alternative permutation-style test that scales in a computationally feasible manner even when very large numbers of test points are used.

Replication

In the previous section, we discussed the fact that for suitably constructed base learners, predictions from random forests follow an asymptotically normal distribution. This allows for confidence intervals to accompany predictions and also allows us to conduct hypothesis tests for whether an independent variable contributes significantly to the prediction of the dependent variable. This has important implications for political science research. If the linear or otherwise model-specific assumptions in standard parametric models are violated, then both the coefficient estimates and confidence intervals are biased. Random forests make no such model-specific assumptions about the functional form. This section uses random forests to make inferences and conduct significance tests by replicating FL's seminal paper on civil conflict.

Finally, we stress the importance and role of the test and training sets in conducting the aforementioned hypothesis tests. Unlike traditional linear models, the significance tests are conducted on held-out test data – a subsample of the original dataset not used to train the model. This reduces concerns about overfitting in the trained model but also means that results can sometimes depend on the particular test locations chosen. To mitigate this effect, large test sets can be used in conjunction with more efficient procedures as in Coleman et al. (2019). Alternatively, tests can be performed across multiple train–test splits and results combined to form a more robust procedure. Meinshausen et al. (2009) provided one such means of combining results, and more efficient and general procedures are currently being developed in the statistics and machine learning community.

Civil conflict onset

One of FL's more notable arguments is that the share of mountainous terrain is an important predictor of civil conflict onset, while markers of ethnic and religious fractionalization are not significant predictors. Using random forests with the accompanying variance estimates on their dataset suggests one reason why these results were produced: the marginal relationship between both ELF and religious fractionalization on civil conflict onset appears to be highly nonlinear.

We examined the data using the methods already outlined. We present the marginal relationship between the dependent variable and an independent variable using predicted probability plots – the change in the predicted probability of the dependent variable taking on a particular value while holding all other variables at their mean – in Figure 1 for all continuous covariates in FL's main models.[3] For all predicted probability plots, the variance estimates of the predictions from our random forests model are included.

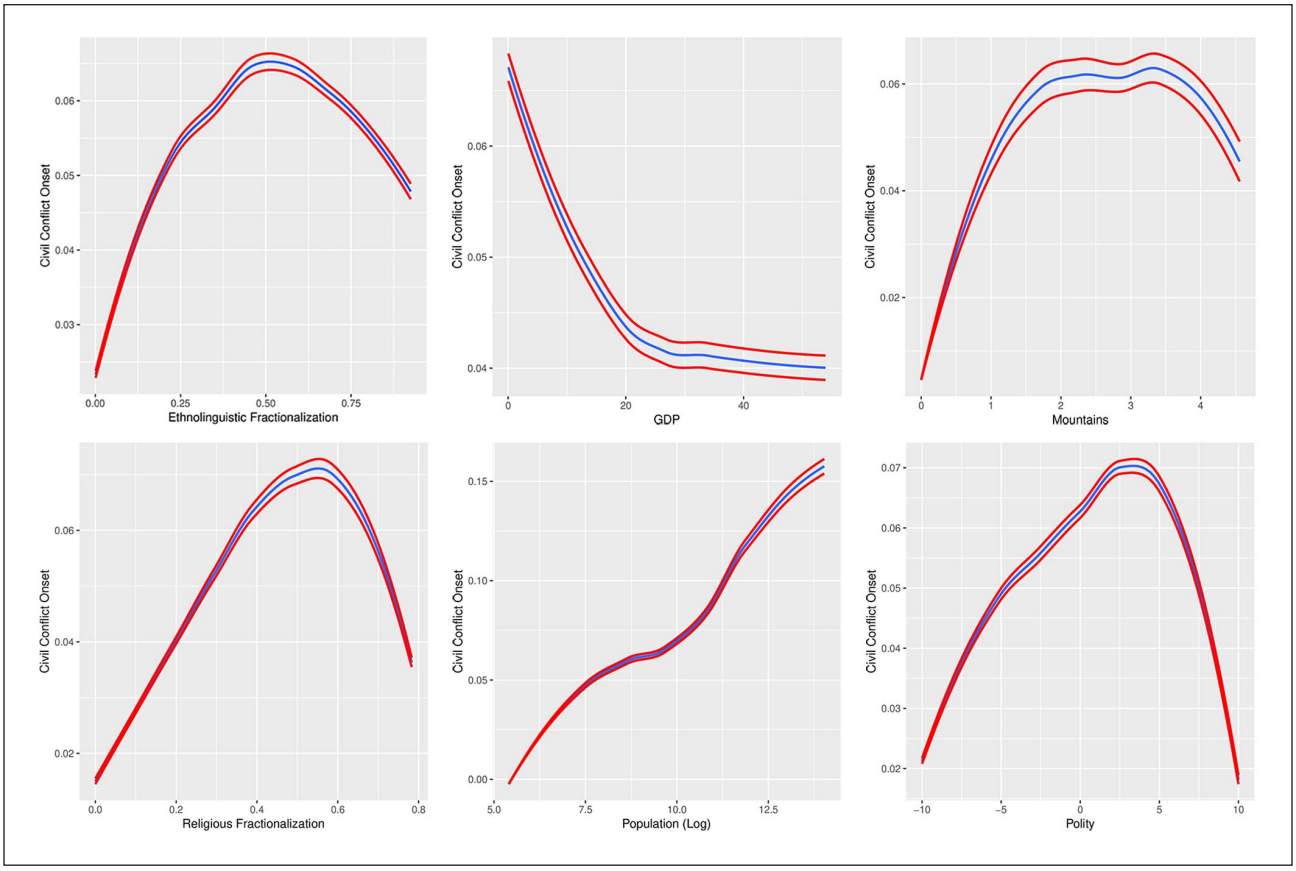

Figure 1. Marginal effects of a change in an independent variable on the likelihood of civil conflict onset. The blue lines represent the LOESS smoothing estimate of those predictions, while the red lines represent LOESS smoothed 95% confidence intervals for those predictions.

From Figure 1, we see that GDP and population appear to have relationships to conflict onset that are at least monotonic and close enough to linear that we would expect such a signal to be captured in a standard linear model. The figure also suggests that a reasonable functional form for the effect of a country's polity score is also specified in FL's models, since the random forest output shows a roughly quadratic relationship.

When we turn to the marginal effects of ELF and religious fractionalization, a more complicated picture emerges. The plots for these variables show a curvilinear and nonmonotonic relationship. For both fractionalization measures, there is a positive and largely linear relationship when both variables are at the lower half of their range of values. When the variables take on values above 0.50, that is, when two randomly selected individuals are more rather than less likely to belong to different ethnolinguistic or religious groups, the marginal relationship is negative.

We now use the method described in "Hypothesis testing with random forests" to test whether ELF or religious fractionalization are statistically significant predictors of civil conflict onset. As already described, this is performed by constructing two large random forests built on proper subsamples of the training data – one where the variables we are interested in testing are randomly permuted, and one where the data are undisturbed – and comparing predicted values at various locations throughout the feature space. Extracting the variance estimates allows us to compute a p-value to determine whether the models constructed with these different training sets produce significantly different predictions.[4]

The results of these hypothesis tests for the predictive significance of individual variables are presented in Table 1. The results show that several of the variables achieve marginal predictive significance at conventional ($\alpha = 0.05$) levels. Lagged values for civil war, oil, and the instability measure are not statistically significant. Both fractionalization measures are significant predictors of conflict onset. This contrasts with one of the main findings of FL, that ELF is unrelated to whether a civil war occurs. One possibility for the difference in results between our random forest model and their logit model is the nonlinear relationship between the fractionalization measures and the probability of onset.

Table 1. *P*-values for marginal predictive significance for a number of predictors of civil conflict onset.

Variable	*p*-value
ELF	0.0023
Mountains	0.0024
Religious fractionalization	0.0009
New state	0.0174
War (lagged)	0.5104
GDP	0.0001
Population	0.0003
Contiguity	0.0004
Oil	0.6664
Instability	0.7351
Polity	0.0537

This random forest testing procedure is also capable of conducting joint significance tests for predictive importance. This is conducted in the same way as testing whether one variable is significant: we simply permute the values of the entire set of predictor variables we want to jointly test, and then compare random forest models trained on the original data with models trained on data where the values of those variables have been randomly permuted. This is an important follow-up step for variables that appear to have qualitatively similar marginal relationships to the response variable. Indeed, in an analogous fashion to linear models, when two highly correlated predictors are available, it is entirely possible that neither looks marginally important since the other can be used to uncover the same signal. We carried out this test for both religious fractionalization and ELF, since these variables have similar effects according to the predicted probability plots. The resulting *p*-value was 0.00003. We note that, quite importantly, a joint significance test for these same variables in a logit model produces a *p*-value of 0.723.

Conclusion

Random forests are often considered the best off-the-shelf black box algorithm for making accurate predictions. They can readily accommodate missing values, nonlinear relationships, interactions, and a large numbers of covariates. A number of studies have used random forests to predict the onset of civil or interstate wars (Muchlinski et al., 2015). We have argued here that the superior predictive performance of random forests can be harnessed to examine the same kinds of relationships in the data that political scientists typically seek to uncover with conventional parametric models, including making inferences about the marginal effect of independent variables.

A nonparametric approach that retains the ability to conduct valid inference opens up a number of possibilities for political scientists. In addition, advances in plotting the results of random forests can enable researchers to present results about how variables interact without specifying the functional form (Goldstein et al., 2013). To ensure comparability, we chose to replicate a study that did not use fixed effects or clustered standard errors. Future research should explore how random forest models can incorporate dependence between observations. Predictions from random forest models have also been shown to be effective when used in multiple imputation for missing variables (Tang and Ishwaran, 2017). As our exploration of the relationship between ELF and religious fractionalization measures and civil conflict onset show, the full benefits of nonparametric methods in general and random forests in particular have yet to be utilized by political scientists.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Richard J. McAlexander https://orcid.org/0000-0002-2165-2652

Supplemental materials

The supplemental files are available at http://journals.sagepub.com/doi/suppl/10.1177/2053168020905487.
The replication files are available at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/ZLEFIB

Notes

1. We use the term "linear" to refer to a model that is linear in the parameters.
2. We also compared the predictive performance of our random forest model and FL's logit models on this dataset, and present confusion matrices for both models in Section 2 of the Online Appendix.
3. In Section 3 of the Online Appendix, we show predicted probability plots when the continuous covariates are held at different quantiles. For ELF, the U shape appears across most specifications. For religious fractionalization, the U shape only appears when the other covariates are at their 50th through 70th percentiles, suggesting the U-shaped relationship between religious fractionalization and conflict onset is conditional on other covariate values.
4. We selected the following tuning parameters: $k = 75$, nx1 $= 50$, nmc $= 1000$, minsplit $= 3$, with 20 observations in the test set. These parameter values ensured that a standard uniform distribution of *p*-values was produced when testing the significance of random draws from a standard normal distribution. Results when the test set was increased to 40 are presented in Section 1 in the Online Appendix.

Carnegie Corporation of New York Grant

This publication was made possible (in part) by a grant from the Carnegie Corporation of New York. The statements made and views expressed are solely the responsibility of the author.

References

Achen CH (2005) Let's put garbage-can regressions and garbage-can probits where they belong. *Conflict Management and Peace Science* 22(4): 327–339.

Barber RF and Candes EJ (2015) Controlling the false discovery rate via knockoffs. *The Annals of Statistics* 43(5): 2055–2085.

Breiman L (2001) Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statististical Science* 16(3): 199–231.

Coleman T, Peng W and Mentch L (2019) Scalable and efficient hypothesis testing with random forests. Available at: https://arxiv.org/abs/1904.07830 (accessed 9 October 2019).

Efron B (1979) Bootstrap methods: Another look at the jackknife. *The Annals of Statistics* 7(1): 1–26.

Efron B (2014) Estimation and accuracy after model selection. *Journal of the American Statistical Association* 109(507): 991–1007.

Fearon JD and Laitin DD (2003) Ethnicity, insurgency, and civil war. *American Political Science Review* 97(1): 75–90.

Goldstein A, Kapelner A, Bleich J, et al. (2013) Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. Available at: https://arxiv.org/abs/1309.6392 (accessed 9 October 2019).

Meinshausen N, Meier L and Buhlmann P (2009) P-values for high-dimensional regression. *Journal of the American Statistical Association* 104(488): 1671–1681.

Mentch L and Hooker G (2016) Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *The Journal of Machine Learning Research* 17(1): 841–881.

Mentch L and Hooker G (2017) Formal hypothesis tests for additive structure in random forests. *Journal of Computational and Graphical* Statistics 26(3): 589–597.

Muchlinski D, Siroky D, He J, et al. (2015) Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. *Political Analysis* 24(1): 87–103.

Peng W, Coleman T and Mentch L (2019) Asymptotic distributions and rates of convergence for random forests and other resampled ensemble learners. Available at: https://arxiv.org/abs/1905.10651 (accessed 9 October 2019).

Scornet E, Biau G, Vert J-P, et al. (2015) Consistency of random forests. *The Annals of Statistics* 43(4): 1716–1741.

Sexton J and Laake P (2009) Standard errors for bagged and random forest estimators. *Computational Statistics & Data Analysis* 53(3): 801–811.

Tang F and Ishwaran H (2017) Random forest missing data algorithms. *Statistical Analysis and Data Mining* 10(6): 363–377.

Wager S and Athey S (2018) Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* 113(523): 1228–1242.

Wager S, Hastie T and Efron B (2014) Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research* 15(1): 1625–1651.