Leveraging the 'Black Box': Covariate Adjustment via Random Forests

Miles D. Williams*

02 March, 2022

Abstract

Multiple regression is a workhorse method for adjusting for the confounding influence of covariates in observational studies. However, its reliability is sensitive to violations of its parametric functional form assumptions. This note describes an alternative approach. By using a nonparametric ensemble machine learner (random forests) to first partial out the influence of covariates on a response and causal variable of interest, reliable estimates of the effect of the causal variable can be recovered. A simulation demonstrates the efficiency and reduced bias provided by random forest adjustment (RFA) relative to multiple regression, especially when dealing with nonlinear confounding. RFA is then applied to a replication of Nielsen et al. (2011) who assess the effect of foreign aid shocks on civil war onset. An additional strength of RFA is illustrated through the use of random forest diagnostics. Variable importance metrics can be used to identify which among 20 covariates pose the greatest source of confounding in testing the effect of aid shocks, and regression weights generated from random forests can used to characterize the effective relative to the nominal data sample used to identify the effect of aid shocks. These diagnostics highlight how RFA estimates, despite being facilitated via a "black box" method like random forests, can be made transparent and interpretable.

*University of Illinois at Urbana-Champaign, milesdw2@illinois.edu

1 Introduction

Covariate adjustment remains a perennial concern in observational studies. Most existing approaches impose necessary but potentially fragile parametric assumptions, while many robust non-parametric alternatives can be difficult to interpret. I propose a simple to implement method called random forest adjustment (RFA) that blends the interpretability of parametric estimators with the power and flexibility of non-parametric approaches.

As its name suggests, RFA applies the powerful and popular random forests machine learner to the problem of covariate adjustment. Random forests is a technique that is well known for its predictive power and ability to handle many forms of nonlinearity and covariate interactions in data. These are ideal qualities in any setting but especially so when adjusting for covariates when neither theory nor past research provide adequate guidance about an appropriate functional form.

RFA consists of a two-step process. In the first stage, random forest regressions are fit to a response variable and a causal variable of interest using relevant covariates in the data. Predictions from these models are then used to re-center the outcome and causal variable so that only residual variation in each not explained by covariates remains. In the second stage, the residualized response is regressed on the residualized causal variable, and estimation is done via OLS.

Simulations show that this approach is unbiased and efficient in the face of both linear and nonlinear forms of confounding. While a correctly specified parametric model performs slightly better, an incorrectly specified parametric model performs poorly while RFA remains insensitive to different functional forms.

Following simulation, I apply RFA to replication data from Nielsen et al. (2011). The authors examined the effect of sudden and sizable negative shocks in the foreign aid received by developing countries as a share of GDP on the likelihood of civil war onset. Using RFA, in contrast to the findings in the original study, I find that *both* negative *and* positive shocks affect the likelihood of civil war rather than only negative shocks. RFA estimates further suggest that a continuous version of the measure for changes in aid per GDP also has a significant negative relationship with civil war onset.

Various diagnostics for random forests are also demonstrated to highlight the ease of demystifying RFA estimates. A common complaint about non-parametric machine learners like random forests is that they are "black boxes" that may make good predictions but lack interpretability. This raises natural questions about how to interpret RFA estimates. What comparisons in the data do estimates reflect? What does the effective sample used to generate these estimates look like after adjustment?

Well established processes exist for answering these questions. Variable importance metrics can be used to identify which variables were most prognostic of the outcome and causal variable in the random forests. Further, using the method outlined by Aronow and Samii (2016) it is possible to draw comparisons between the nominal data sample and the effective data sample upon which estimates are based. Both methods lift the veil on RFA estimates, showing that random forests are more "white" than "black" boxes than often realized.

Though RFA is not a one-size-fits-all solution for covariate adjustment, it has many desirable qualities that make it a powerful addition to any methodological toolbox. This novel application of random forests also provides a foundation for exploring other adjustment procedures that substitute random forests with gradient boosting methods, support vector machines, or even super learner approaches that use an ensemble of machine learners to generate predictions. There are many untried but promising extensions, and it is hoped this introduction to RFA spurs further innovations.

2 Random Forest Regression

The past two decades witnessed the development and evolution of numerous machine learning techniques for causal inference. Among these methods, random forest regression has gained special popularity. Since its development by Brieman (2001), random forest regression has been widely used across a range of industries and scientific fields, due in no small part to its high level of predictive performance and simultaneous resistance to overfitting (Wager 2016).

It comes as little surprise, then, that random forests have found increasing use among political scientists. Hill and Jones (2014), for instance, use random forests to test the predictive power of various political, economic, and social conditions on state repression. Bonica (2018) uses the method to assess the predictive power of campaign contributions on roll call votes in the U.S. Congress. Meanwhile, Carroll and Kenkel (2019) use the approach, in addition to other machine learning algorithms, to generate a proxy for relative military power among states.

Beyond political science, social scientists across fields and disciplines have applied RF in their research—e.g., Athey and Imbens (2015) and Foster, Taylor, and Ruberg (2011). These studies are joined by a broader literature that has developed creative applications of other machine learning techniques for causal inference (for examples see Wager 2016).

Random forest regression is a powerful technique for a number of reasons. Random forests are an ensemble machine learner—so called because random forests are an ensemble of many individual regression trees. These individual trees, or base learners, are "grown" on either random subsamples, or bootstrapped samples, of the full dataset. For each sample, a random set of predictors is then selected. Among this set of predictors, a partition is made along one of these predictors such that the residual sum of squares between the predicted and observed value of the response is minimized (this is the first set of "branches" in the tree). Within these partitions, a new subset of variables is sampled, the best predictor of the response within this subset is identified, and a new optimal partition is made. This process continues until some criterion is reached.

By repeating this process multiple times, a "forest" of these base tree learners is produced. The random forest prediction for an individual observation is then generated by a “voting” process, whereby the average of the regression tree predictions for a given observation is used to generate an overall forest prediction.[1]

The combination of using regression trees as base learners, bootstrapping samples and subsampling predictors, and bagging predictions makes random forest regression robust to various forms of nonlinearity and variable interactions. At the same time, these features confer resistance to overfitting and the influence of outlier observations. These features, thus, make random forests a powerful ally in adjusting for the confounding influence of covariates in observational data.

3 Applying Random Forests within a Covariate Adjustment Framework

Consider the simple linear regression model, with a real valued response Y_i and a real valued predictor (or causal variable) X_i, where $i = 1, ..., n$:

$$Y_i = \alpha + \beta X_i + \epsilon_i. \tag{1}$$

The parameter β denotes the unit change in Y_i per a unit change in X_i, α is the expected value of Y_i when $X_i = 0$, and ϵ_i is unexplained noise. Since α is not of primary importance, we may de-mean both Y_i and X_i and fix $\alpha = 0$ without doing violence to the estimation of β. This gives the modified specification:

$$Y_i - \overline{Y} = \beta\left(X_i - \overline{X}\right) + \epsilon_i. \tag{2}$$

Many estimators exist for recovering β, but the most common solution is ordinary

[1] This process is slightly modified when the response is a discrete category. Random forests can be used both for categorization and regression—in the latter case, the base learners are optimized to predict the conditional mean of the response.

least squares (OLS). This entails finding $\hat{\beta} \in \mathbb{R}$ that minimizes the residual sum of squares:

$$\text{rss} = \sum_{i=1}^{n} \left[\left(Y_i - \overline{Y} \right) - \hat{\beta} \left(X_i - \overline{X} \right) \right]^2 . \tag{3}$$

This has the closed-form solution:

$$\hat{\beta} = \frac{\text{cov}\left(X_i, Y_i \right)}{\text{var}\left(X_i \right)} = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X} \right) \left(Y_i - \overline{Y} \right)}{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2}. \tag{4}$$

The residual values $\hat{\epsilon}_i$ are of course calculated as $\hat{\epsilon}_i = \left(Y_i - \overline{Y} \right) - \hat{\beta} \left(X_i - \overline{X} \right)$.

A prominent feature in equations 2-4 is that the sample means of Y_i and X_i act as pivot points that, once fixed, provide the appropriate variation in Y_i and X_i necessary for estimating $\hat{\beta}$. In a simple regression model with only one predictor, these pivot points are held constant. However, when additional covariates are controlled for in a regression model, these means are conditional. Suppose we extend the regression model by including a second predictor Z_i:

$$Y_i = \alpha + \beta X_i + \delta Z_i + \epsilon_i. \tag{5}$$

The least squares solution for β now takes into account the *conditional* means of Y_i and X_i given Z_i in fixing the pivot points for each variable. Let $\hat{Y}_i^z$ equal the conditional mean of Y_i given Z_i, and let $\hat{X}_i^z$ equal the conditional mean of X_i given Z_i. The solution for $\hat{\beta}$ now is:

$$\hat{\beta} = \frac{\text{cov}\left(X_i - \hat{X}_i^z, Y_i - \hat{Y}_i^z \right)}{\text{var}\left(X_i - \hat{X}_i^z \right)} = \frac{\sum_{i=1}^{n} \left(X_i - \hat{X}_i^z \right) \left(Y_i - \hat{Y}_i^z \right)}{\sum_{i=1}^{n} \left(X_i - \hat{X}_i^z \right)^2}. \tag{6}$$

Within the multiple regression framework, the pivot values for Y_i and X_i are linear functions of Z_i:

$$Y_i = \eta + \eta_z Z_i + v_i; \quad X_i = \gamma + \gamma_z Z_i + \mu_i. \tag{7}$$

Essentially, this means that the solution for $\hat{\beta}$ is equivalently identified by regressing the residual variation in Y_i on the residual variation in X_i from the above linear models:

$$\hat{v}_i = \beta \hat{\mu}_i + \epsilon_i \quad \text{s.t.} \quad \hat{\beta} = \frac{\sum_{i=1}^{n} \hat{\mu}_i \hat{v}_i}{\sum_{i=1}^{n} \hat{\mu}_i^2}. \tag{8}$$

A limitation of this approach, and the one that motivates this note, is that it is valid only to the extent that Y_i and X_i can be adequately expressed as linear functions of Z_i. Any violations in functional form impend bias in the solution for $\hat{\beta}$. Enter random forest adjustment.

Regression adjustment via random forests innovates on the above multiple regression framework by generalizing the process of identifying the appropriate conditional means (pivot points) of the response and the variable of interest given other observed covariates. For the above example where the relationship between Y_i and X_i accounts for Z_i, this entails replacing the linear model predictions for these variables given Z_i with their random forest generated predictions given Z_i.

To be more explicit, suppose we generalize the functional forms shown in 7:

$$Y_i = f_y(Z_i) + v_i; \quad X_i = f_x(Z_i) + \mu_i. \tag{9}$$

Let $\hat{f}_x(Z_i)$ denote the expected values of X_i given Z_i, and let $\hat{f}_y(Z_i)$ denote the expected values of Y_i given Z_i. Using these values as the new pivot points the solution for $\hat{\beta}$ now is:

$$\hat{\beta} = \frac{\text{cov}\left[X_i - \hat{f}_x(Z_i), Y_i - \hat{f}_y(Z_i)\right]}{\text{var}\left[X_i - \hat{f}_x(Z_i)\right]}. \tag{10}$$

In a multiple regression framework the $f(\cdot)$ functions can only ever be linear. In random forest adjustment, these functions may take any number of forms. In the event that a linear form best approximates the relationships in the data, multiple regression will do well. However, to the extent that the confounding influence of Z_i deviates from the linear assumptions of a linear multiple regression model, the pivot points used to center the response and causal variable will be erroneous and $\hat{\beta}$ will reflect inappropriate comparisons in the data sample. But, as far as the data-driven approach of random forests is robust to such deviations, random forest adjustment will yield better pivot points and thus $\hat{\beta}$ will reflect more appropriate comparisons.

4 Limitations

The proposed strength of random forest adjustment (RFA), then, is its robustness to functional form violations in the confounding influence of one or several covariates in a data sample. This is a real virtue, but before demonstrating its advantages some limitations bear noting.

While its atheoretical approach to regression adjustment confers a certain degree of robustness, this same atheoreticism is also a point of concern. Random forests, while powerful, are criticized for their limited interpretability. By extension, this raises natural questions about how RFA estimates should be interpreted. If the method of controlling for covariates is a "black box," this implies obscurity about the comparisons in the data sample that RFA supports as well. However, there are numerous methods for lifting the lid on this black box that should mitigate concerns about interpretability. Such methods will be discussed in greater detail in a subsequent section.

Another limitation of RFA is also simply a limitation of all observational studies: unknown or unmeasured sources of confounding. RFA estimates are only as good as the data used by a researcher. If an important confounding variable is missing, the quality of

RFA estimates cannot be guaranteed.

Finally, the RFA framework as described here is constructed to address situations where a research has an interest in estimating the association between *one* explanatory variable and an outcome. There of course may be situations where a scholar is interested in an interactive effect or more complex causal relationships. Modifying RFA to accommodate such studies is an important goal for future development.

5 Monte Carlo Analysis

5.1 The Simulation

To assess the performance of RFA in comparison to multiple regression adjustment (MRA), I perform a Monte Carlo analysis using two different data-generating processes. For the first, I simulate a simple linear d.g.p. For n observations where $i = 1, ..., n$, I specify:

$$\begin{aligned}
&Y_i \sim \mathcal{N}\left(\mu_i = 1 + \beta X_i + 0.5 Z_i, \sigma = 10\right), \\
&X_i \sim \mathcal{B}(n = 1, p = 1/(1 + \exp\{2 - 0.05 Z_i\})), \\
&Z_i \sim \mathcal{N}\left(\mu_i = 50, \sigma = 10\right).
\end{aligned}$$

For three different possible ns ($n = 500; 1,000; 2,000$), I iteratively simulate the d.g.p. 10,000 times. For each iteration, I recover an estimate of β using both MRA and RFA. For the sake of this exercise, the true value of $\beta = 5$.

I further simulate a nonlinear d.g.p., specified as

$$\begin{aligned}
&Y_i \sim \mathcal{N}\left(\mu_i = 1 + \delta X_i + 0.5 Z_i + Z_i^2, \sigma = 10\right), \\
&X_i \sim \mathcal{B}(n = 1, p = 1/(1 + \exp\{2 - 0.05[1.15 Z_i - \bar{Z}]^2\})), \\
&Z_i \sim \mathcal{N}\left(\mu_i = 50, \sigma = 10\right).
\end{aligned}$$

As before, for n of 500, 1,000, and 2,000 I iteratively simulate the d.g.p. 10,000 times and for each iteration I recover estimates of β via MRA and RFA.

The use of a single confounder, Z_i, is certainly not representative of most real-world cases. However, by focusing on one confounder it is all the easier to bring into stark relief the very real implications of confounding and how even simple model mispecification (like failure to include a quadratic term) can produce inconsistent estimates on the causal variable of interest. It further aids in clear exposition of the robustness of RFA to such errors.

After running the simulations, I summarize the performance of each approach according to the average percent bias in parameter estimates, the percent coverage of 95% confidence intervals[2] the average size of the standard errors, and the root mean squared error (RMSE) of the parameter estimates.

5.2 Results

A summary of the performance of RFA and MRA is shown in Figure 1. For each metric used to assess the approaches, results are shown for the different samples sizes ($N = 500; 1,000; 2,000$), and for the linear (left) and nonlinear (right) cases.

Upon examination of the results, it is immediately apparent that RFA confers a substantial advantage with respect to the nonlinear d.g.p. The average percent bias of RFA estimates are -2.48, -5.8, and -0.36 for N of 500, 1,000, and 2,000 respectively. For MRA, the average percent bias is orders of magnitude worse: 1,859.27, 1,887.39, and 1,864.28 for each successive N. With respect to coverage, for each sample size in ascending order, RFA 95% confidence intervals contain the true estimate 91, 84, and 96 percent of the time. For MRA, none contain the true effect.

RFA estimates are also more efficient, and less prone to error in general. The average standard error for RFA estimates for each sample size is 1.6, 1.08, and 0.76 in ascending order. Meanwhile, on average the standard errors of MRA estimates are several times

[2] Based on HC1 robust estimates of parameter variance.

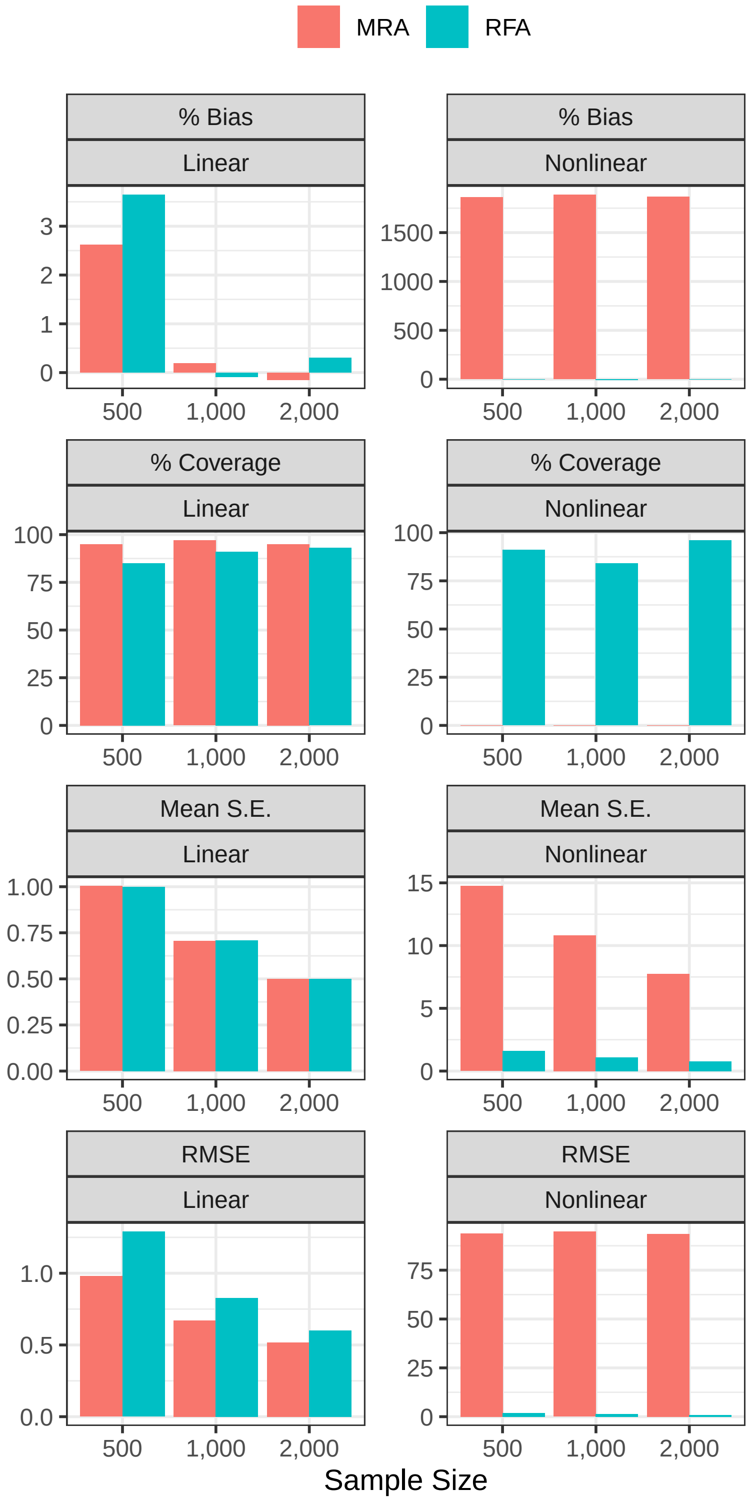

Figure 1: Performance of RFA vs. MRA

wider: 14.75, 10.81, and 7.74 respectively. Further, RMSE for the RFA estimates for each sample size are 1.92, 1.37, and 0.79; while for MRA, RMSE is 93.86, 94.72, and 93.49.

While RFA performs better in terms of both bias and efficiency when dealing with a nonlinear d.g.p., as the results shown in the left panel show, MRA is moderately less biased and prone to slightly less error if the d.g.p. is linear. The average percent bias of RFA estimates are 3.65, -0.09, and 0.31 for N of 500, 1,000, and 2,000 respectively. For MRA, the average percent bias is moderately improved: 2.61, 0.19, and -0.15 for each N. With respect to coverage, for each sample size in ascending order, RFA 95% confidence intervals contain the true estimate 85, 91, and 93 percent of the time. For MRA, they contain the true estimate 95, 97, and 95 of the time.

Further, RFA estimates, while no more or less efficient than MRA estimates, tend to be slightly more prone to error. The average standard error for RFA estimates for each sample size is 1, 0.71, and 0.5 in ascending order. Meanwhile, on average the standard errors of MRA estimates nearly identical: 1, 0.71, and 0.5 respectively. However, RMSE for the RFA estimates for each sample size are 1.29, 0.83, and 0.6; while for MRA, RMSE is 0.98, 0.67, and 0.52.

The finding that RFA is slightly more prone to error and bias when the true d.g.p. is linear makes a good deal of sense. The nonparametric approach of random forests, though powerful and capable of detecting a wide array of relationships, is just not as efficient as a parametric linear regression model when dealing with a linear d.g.p. It is nonetheless encouraging that RFA does not perform *much* worse than parametric regression adjustment when dealing with linear confounding. Further, as suggested by the variation in each estimator's performance with increasing sample sizes, RFA's error converges to that of MRA given a sufficient number of observations. However, while RFA performs only moderately worse than MRA in the face of a linear d.g.p., it is far and away superior when dealing with nonlinear confounding in the data.

The reason for this result is simple. A correctly specified parametric regression model is generally more efficient than the nonparametric approach taken by random forests, especially when dealing with finite samples. However, random forests is generally a robust method of approximating the form of confounding in the data absent an appropriately specified parametric model, and its performance in this task improves with the sample size.

6 Random Forest Adjustment, Aid Shocks, and Civil War

To demonstrate how RFA might be used with real-world data, I apply it to replication data made publicly available by Nielsen et al. (2011). In their study, the authors assess the impact of negative aid shocks on the likelihood of civil war onset in developing countries that regularly receive foreign aid from industrialized donor countries. Theirs is an ideal dataset to use due to the potential for myriad violations of linearity and additivity in the confounding influence of covariates. In their study, Nielsen et al. (2011) control for some 20 variables in estimating the effect of aid shocks—providing ample opportunity for complex forms of confounding.

This dataset is also useful for demonstrating how various diagnostics can be applied to demystify random forest adjusted estimates. Using variable importance metrics for random forests, it is possible to uncover which of the numerous covariates is the greatest source of confounding in the data. Further, using the method outlined by Aronow and Samii (2016), it is possible to assess how the effective sample produced via random forest adjustment compares to the nominal sample used at the outset.

6.1 Background: Aid Shocks and Civil War

Before diving into the analysis, it first will be helpful to summarize Nielsen et al.'s (2011) motivation for assessing the link between aid shocks and civil war. The authors note the underexamined role of aid instability in the onset of intra-state conflict. They posit that sudden changes in aid received by developing country governments can upset the balance of power between rebels and the state. Many countries rely on aid flows for a sizable share of their total outlays, and thus their budgets are subject to substantial windfalls or shortfalls as a consequence of fluctuations in total aid received. In their own words, the authors contend "that rapid changes in aid flows—aid shocks—can grow large enough to materially affect the balance of power between a government—the sovereign recipient of aid flows—and potential rebels, which we define as individuals or groups that might use violence to oppose the government" (Nielsen et al. 2011, 221).

Their contention is that sudden and sizable declines in aid flows can diminish a government's capacity to deter would-be rebels and to continue side-payments to rebels to keep them appeased. The sudden shift in power that ensues generates a commitment problem between rebels and the state. Reduced deterrence increases rebel's assessment of their likelihood of victory, thus leading insurgents to demand greater side-payments in exchange for peace. At the same time, the government faces a shortfall in funds due to the decline in aid flows, making it harder to afford the necessary side-payments to keep rebels at bay.

The converse of negative aid shocks—positive aid shocks—according to the authors has more ambiguous implications. There are reasons to anticipate a sudden windfall in revenue would strengthen the government's position relative to insurgents, thus enhancing deterrence. However, the increase in resources might also invoke its own kind of commitment problem, namely, that rebels will demand larger side-payments. Given these competing logics, Nielsen et al. (2011) leave the effect of positive aid shocks an open question.

To test whether aid shocks influence the probability of conflict the authors compiled a time-series, cross-sectional dataset of up to 139 aid recipient countries from 1981 to 2005. The outcome variable in the dataset is a binary indicator of armed conflict onset drawn from the UCDP/PRIO database (Gleditsch et al. 2002). The dataset codes a conflict as "1"

Table 1: List of Confounding Variables

(1) Ethnic Fractionalization; (2) Religious Fractionalization; (3) Oil Exportation;
(4) Political Instability; (5) Population; (6) Territorial Contiguity;
(7) Mountainous Terrain; (8) Human Rights Violations; (9) Assassinations;
(10) General Strikes; (11) Riots; (12) Antigovernment Demonstrations;
(13) Infant Mortality; (14) Neighboring Countries Experiencing Civil or Ethnic Conflict;
(15) The Cold War; (16) Democracy (Goldstone et al. 2010); (17) GDP per capita.

("0" otherwise) if at least 25 battle deaths occurred in a given year in a conflict between two or more actors, where at least one of the combatants was government forces. In their primary specification, subsequent years of civil war are dropped from the analysis, but any subsequent resumptions of conflict are kept.

The primary explanatory variable of interest is the incidence of an "aid shock," coded as "1" for the bottom 15% of observations in changes in aid flows per recipient GDP. Aid flows are measured as the total of Official Development Assistance (ODA) as grants and loans received by a developing country in a given year, as recorded in the AidData database (Tierney et al. 2011). The authors consider additional thresholds for aid shocks beyond being below the 15th percentile. A positive aid shock, conversely, is coded as "1" ("0" otherwise) for any changes in aid per GDP at or above the 85th percentile. To adjust for possible time delays in the effect of aid shocks, values are lagged by one year.

The authors adjust for a wide range of possible confounding covariates. These are listed in Table 3. Each of these has been used elsewhere in the civil war literature to predict armed conflict.

In their original analysis, the authors estimated a rare-events logit model (King and Zang 2001) using the pooled country-year data, and used robust standard errors clustered by aid recipient country. After imputing missing values, and cutting subsequent years of conflict from the data, the final dataset consisted of 2,627 observations. The choice of rare-events logit was made given the relatively small number of occurrences of civil war onset in the data. As Table 4 summarizes, the sample contained only 93 instances of civil war onset—only 4% of observations. Since the focus here is on demonstrating an

Table 2: Instances of Civil War Onset

Civil War	N	Proportion
No	2,535	0.96
Yes	93	0.04

Table 3: Random Forest Adjusted Estimates

	Negative	Positive	Negative (25th)	Positive (75th)
Shock	0.009	-0.020	0.027^{*}	-0.025^{*}
	(0.014)	(0.014)	(0.012)	(0.011)
Num. obs.	2,627	2,627	2,627	2,627
N Clusters	139	139	139	139

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

application of random forest adjustment, for the sake of space a replication of the authors' method will not be considered here.

6.2 Results

RFA estimates are shown in Table 5. Entries denote estimated coefficients, with robust standard errors, clustered by aid recipient, shown in parentheses. The first column denotes the effect of negative aid shocks, with the same 15th percentile cutoff used by Nielsen et al. (2011). Unlike the effect identified by the authors, the RFA estimate falls short of statistical significance. The same is true for the effect of a positive shock (defined according to the 85th percentile).

To assess whether alternative percentile cutoffs yielded different results, the third and fourth columns of Table 5 show the effects of negative and positive aid shocks, defined in terms of the 25th and 75th percentiles respectively. Recovered RFA estimates for these versions of the variable are statistically significant. The first suggests that a negative aid shock increases the likelihood of civil war onset by 2.7 percentage points ($p < 0.05$), while the second suggests that a positive aid shock decreases the likelihood of civil war onset by 2.5 percentage points ($p < 0.05$).

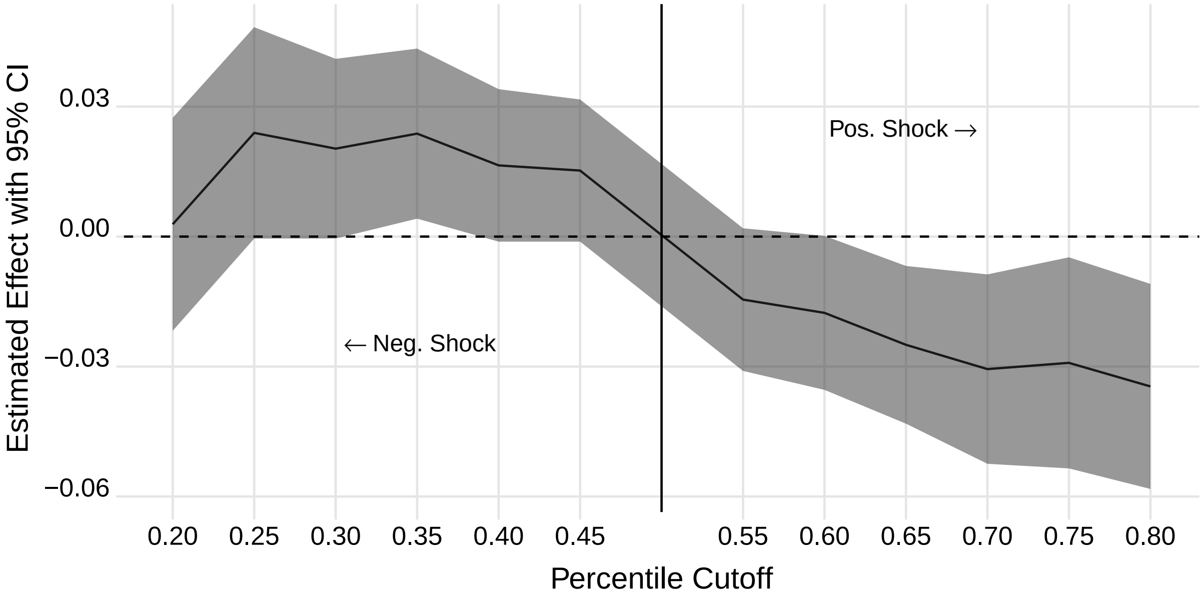

Figure 2: The estimated effect of aid shocks on civil war onset.

Figure 1 shows the range of alternative RFA estimates recovered by applying alternative percentile cutoffs for aid shocks. Unlike Nielsen et al. (2011) who find that only negative aid shocks, and only within a limited range of cutoffs, effect the likelihood of civil war, the results shown in Figure 1 suggest that aid shocks defined over a range of cut-points have a significant relationship with civil war onset after adjusting for covariates.

These results suggest a monotonic relationship between changes in aid per GDP and civil war. To test the monotonicity of this relationship, I further applied RFA to assess the relationship between three continuous versions of the explanatory variable of interest and the response. Table 6 shows the results. In the first column, the raw values of change in aid per GDP are used. In the second, changes in aid per GDP are rank transformed—and further scaled to standard deviation units. In the third column, the variable is trichotomized to values of -1, 0, and 1 based on whether changes in aid per GDP fall within the bottom, middle, or top percentiles.[3]

The raw variable has no significant relationship with the likelihood of civil war. However, this null finding may result from noisiness in the data. As the results for the two

[3] That is, the bottom, middle, and top thirds of the data.

Table 4: Random Forest Adjusted Estimates (Continuous Predictor)

	Aid/GDP	Rank	Percentile
Change	-0.199	-0.010^{*}	-0.016^{**}
	(0.178)	(0.004)	(0.005)
Num. obs.	2,627	2,627	2,627
N Clusters	139	139	139

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

transformed versions of the variable suggest, when rank transformed, or trichotomized by percentiles, changes in aid per GDP do appear to have a significant monotonic relationship with the probability of civil war. After random forest adjustment, recovered estimates suggest that a standard deviation unit increase in the rank-based measure of change in aid per GDP is associated with a decrease the probability of civil war by 1 percentage point ($p < 0.05$). Alternatively, when trichotimized by percentiles, a change from the bottom to the middle, or from the middle to the top, third of the data is associated with a 1.6 percentage point decrease ($p < 0.01$) in the probability of civil war onset.

In terms of real-world significance, both of these estimates suggest quite sizable changes in aid flows are required to generate even a modest change in the likelihood of conflict between governments and insurgents. For a sense of scale, one standard deviation of the rank-transformed measure of change in aid per GDP is 757.28. For reference, a standard deviation unit decline in the rank-based measure relative to its mean corresponds to a decline of roughly 60 cents per 100 dollars of GDP. Since the average country in the sample receives 2.27 dollars in aid per 100 dollars in GDP, such a decline in total aid inflows constitutes, on average, a 26.4% decline in total aid. This is a considerable change in total aid received for such a marginal change (a little more than 1 percentage point) in the likelihood of conflict onset. In short, the random forest adjusted estimate is statistically distinguishable from zero, but quite modest.

6.3 Contextualizing Random Forest Adjusted Estimates

A strength of RFA is that it provides a way to adjust for control variables that is robust to violations of linearity and additivity in their confounding relationships with the response and causal variable of interest. However, some may worry that this robustness comes at the cost of interpretability. RFA may recover a residual relationship between some outcome and explanatory variable, but what can we say about the effective sample used to generate this estimate? Can we say something meaningful about the confounding influence of covariates?

Concern about the transparency of RFA estimates is understandable; though lifting the lid on the random forests trained on the confounding variables is quite easy. On this front, I demonstrate two sets of diagnostics that can improve the interpretability of recovered RFA estimates: (1) variable importance metrics and (2) evaluation of sample distortions incurred from covariate adjustment.

The first set of diagnostics quantifies the predictive import of specific factors included in a random forest regression. Two general measures of variable importance exist: Gini Importance, otherwise known as Mean Decrease in Impurity (MDI), and Permutation Importance, otherwise known as Mean Decrease in Accuracy (MDA). Since my main goal is demonstration, I constrain my choice of importance measure to only one, specifically, MDI. The approach entails simply summing the number of splits, across all trees in the random forest, that are based on a partition in a particular variable, proportional to the total of samples split. The greater the number of times a factor is used in a split, the more important it is in predicting variation in the response (see Louppe (2014) for a helpful summary).

The second diagnostic is not unique to random forests, but applies generally to various forms of regression adjustment. The approach, outlined by Aronow and Samii (2016), quantifies the reweighting of sample observations in generating some causal estimate of interest incurred by regression-based adjustment. The authors distinguish between what they call *nominal* and *effective* samples. The former comprises the complete, unweighted data sample used to estimate some regression model. The latter constitutes the reweighted data sample generated as a product of multiple regression weights. As Aronow and Samii (2016) note, the process of adjusting for the confounding influence of covariates included in the nominal sample leads to distortions in the effective sample upon which causal estimates are identified. Causal estimates produced for a seemingly representative sample in fact often reflect local effects for observations in the data most poorly predicted by the confounding variables controlled for in a regression. As they demonstrate, the effective sample can differ substantially from an otherwise representative nominal sample.

The effective sample regression weights are recovered by simply taking the square of the residual variation in the causal variable of interest left over after regressing it on confounding covariates. For random forest adjustment, this is simply:

$$w_i \equiv \left(z_i - \hat{f}_z(x_i)\right)^2$$

where $\hat{f}_z(x_i)$ generates predicted values of z_i as a function of a vector of confounding factors x_i.

With the computed weights, it is then possible to summarize the characteristics of the effective sample in comparison to the nominal sample. Though not highlighted by Aronow and Samii (2016), one useful approach to making this comparison is, for each of the covariates contained in the vector x_i, standardize values in standard deviation units and mean center, and then compute the weighted average for the variable. Distortions produced by regression adjustment will be apparent by the degree to which the weighted average of the standardized variable diverges from zero.[4] Meanwhile, since values are standardized, it is possible to make comparisons across covariates to assess the relative

[4] Since the standardized value for the nominal sample is mean-centered at zero, any divergence from the zero for the weighted mean reflects a distortion between the effective and nominal samples.

magnitude of distortion. For the k^{th} covariate, relative distortion is calculated as

$$\text{Distortion} = \frac{\sum_{i=1}^{n} w_i \frac{(x_{ik} - \bar{x}_k)}{\text{sd}(x_{ik})}}{\sum_{i=1}^{n} w_i}.$$

6.4 Which Covariates Are Most Confounding?

I begin by summarizing the variable importance metrics for the response, civil war onset, and the causal variable of interest, aid shocks. For example's sake, I use negative aid shocks defined as such for all observations in the bottom 25% of changes in aid per GDP. MDI estimates of importance are shown in Figure 2.

Upon inspection, it is immediately apparent that many of the variables that are most prognostic of civil war onset also predict negative aid shocks. The correlation between MDI measures between responses is shown in the lower right caption of the figure: $\rho = 0.91$. This is a strong correlation, to say the least, and it thus speaks to the wisdom of adjusting for covariates in assessing the relationship between aid shocks and civil war—the variables that most strongly predict one, also most strongly predict the other.

The identities of the top-three confounders should come as little surprise: GDP per capita, population, and infant mortality. Individually, and together, these variables are generally known to be related to civil war onset and receipt of foreign aid. Lower GDP per capita, which often is a proxy for average income, and infant mortality are both markers of poverty, which as Braithwaite, Dasandi, and Hudson (2016) summarize is strongly linked to greater likelihood of conflict. Poverty, likewise, is related to receipt of more total foreign aid. Additionally, population has a positive relationship with the probability of civil war onset (Bleaney and Dimico 2011) and with receipt of foreign aid (Bermeo 2017).

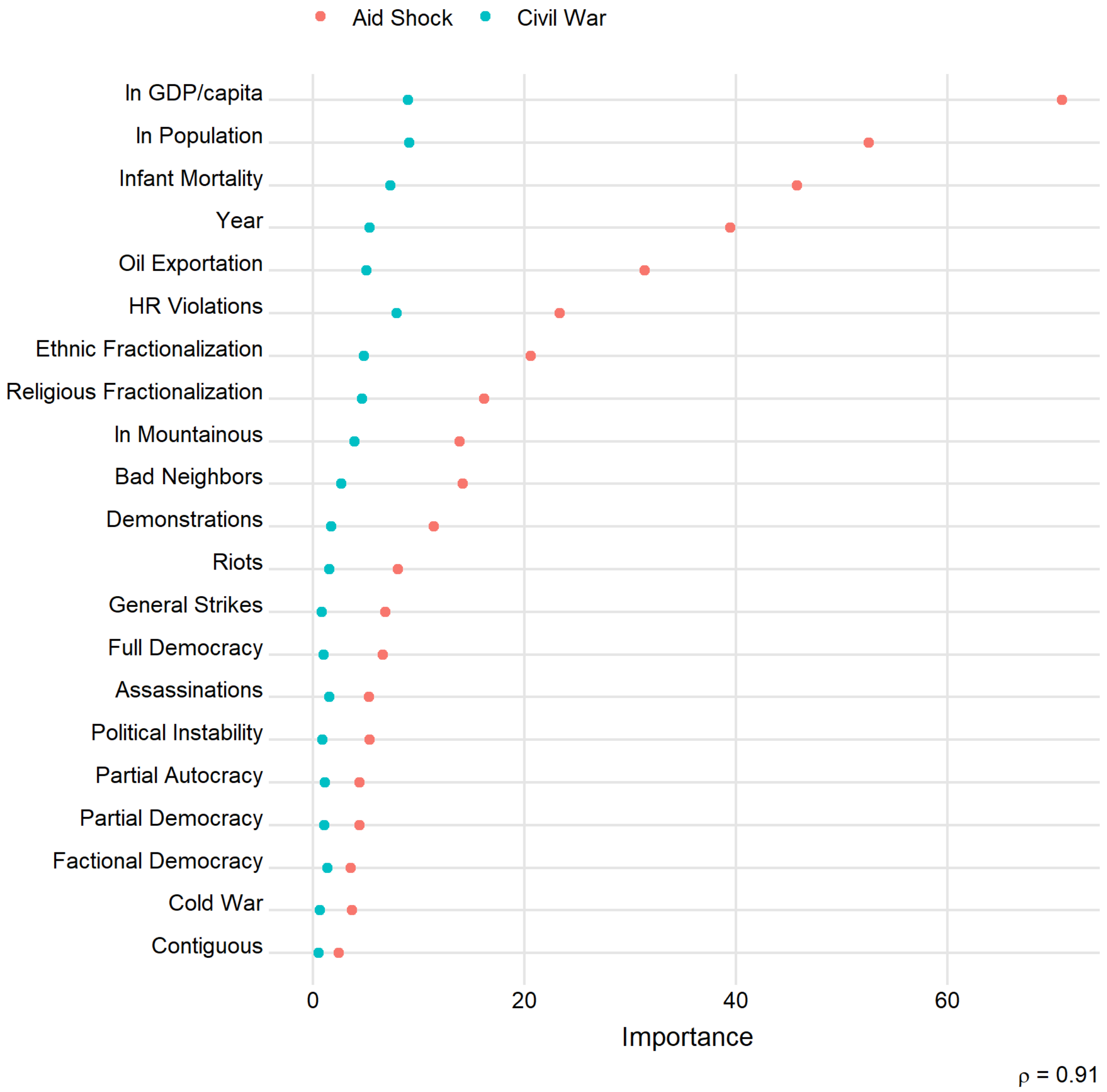

Figure 3: MDI of confounders for civil war onset and aid shocks (25^{th} percentile cutoff).

6.5 How Local is the Effect of Aid Shocks?

Having gained some sense of which variables are the greatest source of confounding in the data, an equally important concern centers on how accounting for confounding via random forests distorts the effective sample used to identify the effect of aid shocks on civil war onset. Figure 3 shows the relative distortion in the effective sample for each of the confounding variables used in the analysis. 95 percent confidence intervals are included, which reflect the weighted standard error of the weighted effective sample means. If these

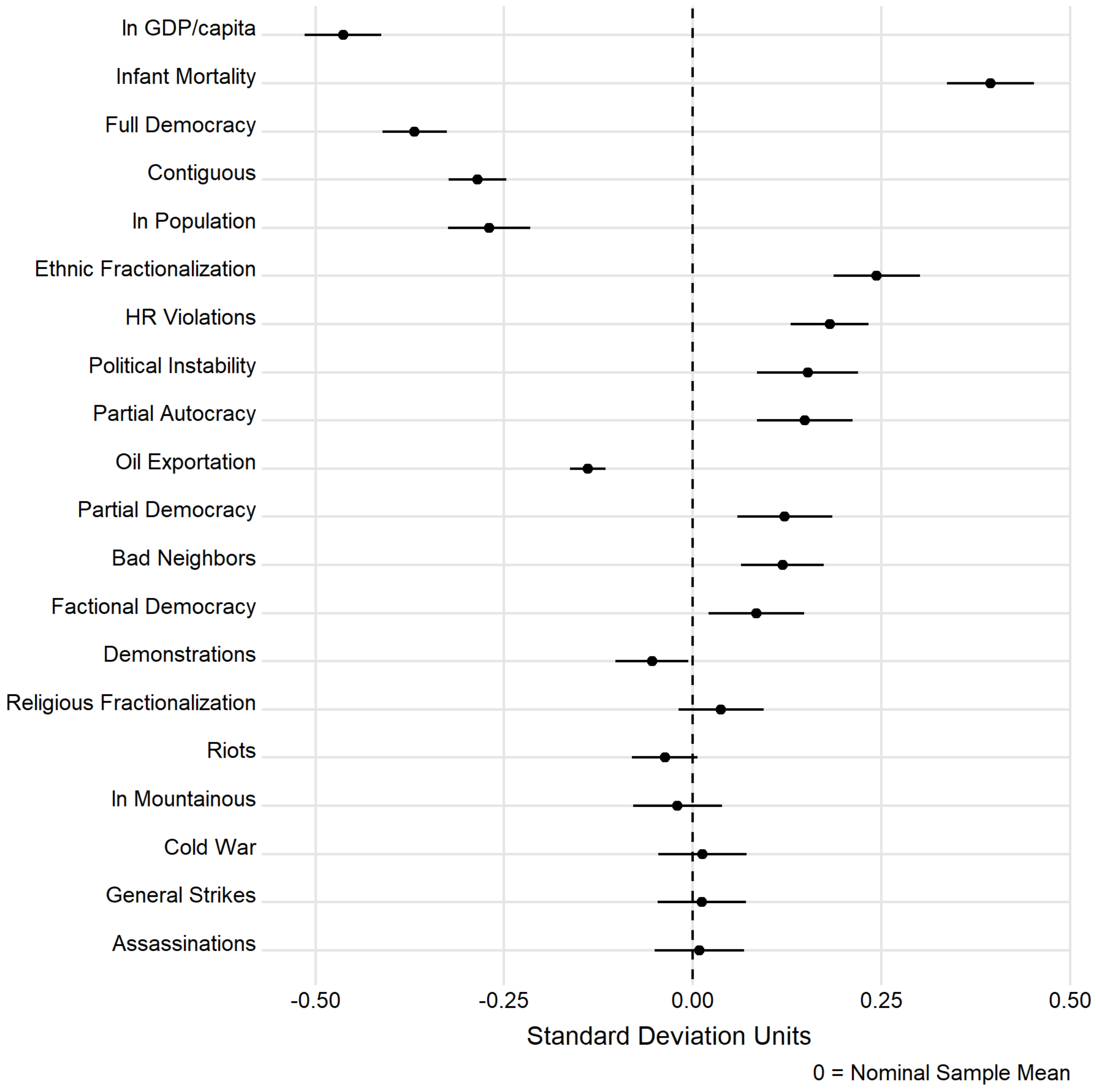

Figure 4: Relative distortions between effective and nominal samples for identifying the effect of aid shocks (25th percentile cutoff).

intervals do not contain zero, this suggests that the effective sample mean for a covariate differs significantly from the nominal sample.

The results suggest that the effect of aid shocks is identified using an effective sample that differs significantly from the original. The relationship between aid shocks and civil war onset holds for an effective sample that is generally much poorer (lower GDP per capita), worse off in terms of infant mortality, less democratic, less likely to have territorial contiguity, smaller in terms of population, experiences greater levels of ethnic fractionalization, suffers more human rights violations, and has greater political instability—to summarize only some of the results.

These findings highlight how the process of random forest adjustment (though this is generally true of any form of covariate adjustment) constrains the generalizability of the recovered causal estimate of interest. This trade-off, however, comes with its own reward. Because it is possible to transparently quantify the distortions incurred by covariate adjustment, one can speak directly to the cases in the data where a causal effect was identifiable. From a policy perspective, it might be imperative, for example, to note that the most impoverished, least democratic, and smallest developing countries are also the ones where an effect of aid shocks on conflict can be observed. This information, easily recoverable for random forest adjusted estimates, has the potential to be quite helpful indeed.

7 Conclusion

Multiple regression adjustment will likely remain the method of choice for political scientists for some time. This is for good reason. The method outlined here, random forest adjustment (RFA), is not meant to displace the standard approach—though there may certainly be cases where it is preferred. Rather, RFA should be treated as yet another tool in the arsenal of researchers; a tool that has its place among the various methods at our disposal.

This note demonstrates the strengths of RFA—namely, its robustness to nonlinear, nonadditive forms of confounding in observational data. Absent well-supported or clear theoretical guidance for adjusting for confounding, the nonparametric approach taken by RFA has clear advantages. RFA is able to recover unbiased, efficient estimates of the relationship between a response and explanatory variable of interest, without requiring *ex ante* specification of the form of confounding. Even in instances where confounding truly is linear and additive, the approach does not perform poorly; albeit, with slightly less efficiency than a parametric regression model.

Further, despite its reputation as a "black box," it is quite easy to lift the lid on random forests. By using variable importance metrics, and using residual regression weights (Aronow and Samii 2016), RFA estimates can be straightforwardly demystified. These diagnostics help to reveal which covariates in a data sample are the greatest source of confounding, and how the process of covariate adjustment via random forests distorts the effective sample used to identify the relationship between the response and the explanatory variable of interest.

Using RFA, and associated diagnostics, I replicate the results from Nielsen et al. (2011), who assessed the effect of aid shocks—sudden changes in aggregate aid flows to developing countries—on the likelihood of civil war onset. The results suggest not only that negative aid shocks increase the likelihood of civil war, but that positive aid shocks decrease the likelihood. In fact, unlike Nielsen et al. (2011), RFA estimates suggest a monotonic, negative relationship between changes in aid received by developing countries and the likelihood of civil war. This relationship is modest, to be sure, but statistically distinguishable after using random forests to adjust for some 20 covariates.

Diagnostics help to contextualize these results. Variable importance metrics for both the response (civil war onset) and the key explanatory variable (aid shocks) are highly and positively correlated, suggesting that the strongest predictors of one also are the strongest predictors of the other. The greatest sources of confounding include average income, population size, and infant mortality.

Comparison of the effective and nominal samples further helps to summarize the cases in the data for which an effect of aid shocks can be identified. This effective sample skews toward especially low income countries, with higher infant mortality, lower levels of democracy, higher rates of political instability and human rights violations, among other differences.

These results should not be viewed as definitive, or even as a refutation of Nielsen et al. (2011). The analysis is meant merely to illustrate how RFA might be applied in practice, and how the results can be dissected to tell a much richer story than that conveyed by focusing only on the single point estimated and test-statistic returned from the procedure. RFA can be leveraged not only as a flexible means to partial out the confounding influence of covariates in observational data, it can also be leveraged to shed light on confoundingness itself.

References

Aronow, Peter M., and Cyrus Samii. 2016. "Does Regression Produce Representative Estimates of Causal Effects?" *American Journal of Political Science* 60(1): 250–67.

Athey, Susan, and Guido Imbens. 2015. "Machine Learning for Estimating Heterogeneous Causal Effects." Research Paper. Stanford University, Graduate School of Business.

Bermeo, Sarah B. 2017. "Aid Allocation and Targeted Development in an Increasingly Connected World." *International Organization* 74(1): 735–66.

Bleaney, Michael, and Arcangelo Dimico. 2011. "How Different Are the Correlates of Onset and Continuation of Civil Wars?" *Journal of Peace Research* 48(2): 145–55.

Bonica, Adam. 2018. "Inferring Roll-Call Scores from Campaign Contributions Using Supervised Machine Learning." *American Journal of Political Science* 62(4): 830–48.

Braithwaite, Alex, Niheer Dasandi, and David Hudson. 2016. "Does Poverty Cause Conflict? Isolating the Causal Origins of the Conflict Trap." *Conflict Management and Peace Science* 33(1): 45–66.

Brieman, Leo. 2001. "Random Forests." *Machine Learning* 45: 5–32.

Carroll, Robert J., and Brenton Kenkel. 2019. "Prediction, Proxies, and Power." *American Journal of Political Science* 63(3): 577–93.

Foster, Jared C., Jeremy M.G. Taylor, and Stephen Ruberg. 2011. "Subgroup Identification from Randomized Clinical Trial Data." *Statistics in Medicine* 30(24).

Gleditsch, Nils Petter, Peter Wallensteen, Mikael Eriksson, Margareta Sollenberg, and Håvard Strand. 2002. "Armed Conflict 1946-2001: A New Dataset." *Journal of Peace Research* 39(5): 615–37.

Hill, Daniel W., and Zachary M. Jones. 2014. "An Empirical Evaluation of Explanations for State Repression." *American Political Science Review* 108(3): 661–87.

King, Gary, and Langche Zang. 2001. "Logistic Regression in Rare Events Data." *Political Analysis* 9(2): 137–63.

Louppe, Gilles. 2014. "Understanding Random Forests: From Theory to Practice." PhD thesis, Université de Liège.

Nielsen, Richard A., Michael G. Findley, Zachary S. Davis, Tara Candland, and Daniel L. Nielson. 2011. "Foreign Aid Shocks as a Cause of Violent Armed Conflict." *American Journal of Political Science* 55(2): 219–32.

Tierney, Michael J., Daniel L. Nielson, Darren G. Hawkins, J. Timmons Roberts, Michael G. Findley, Ryan M. Powers, Bradley Parks, Sven E. Wilson, and Robert L. Hicks. 2011. "More Dollars Than Sense: Refining Our Knowledge of Development Finance Using AidData." *World Development* 39(11): 1891–1906.

Wager, Stefan. 2016. "Comments on: A Random Forest Guided Tour." *TEST* 25: 261–63.