Random Forests for the Social Sciences

Zachary Jones and Fridolin Linder[*]

Abstract

Machine learning techniques gain in popularity in many disciplines and increased computational power allows for easy implementation of such algorithms. However, they are still widely considered as "black box" models that are not suited for substantive research. We present one such method, random forests, with emphasis on practical application for exploratory analysis and substantive interpretation. Random forests detect interaction and nonlinearity without prespecification, have low generalization error, do not overfit, and can be used with many correlated predictors. Importantly, they can be interpreted in a substantively relevant way via measures of marginal variable importance and the partial dependence algorithm. We provide intuition as well as technical detail about how random forests work, in theory and in practice, as well as an empirical example from the literature on comparative politics. We also provide software to facilitate the substantive interpretation of random forests and guidance on when random forests may be useful.

Introduction

Political scientists have, in recent years, begun to utilize more flexible algorithmic methods for inferring substantive relationships from data (Beck and Jackman 1998; Beck, King, and Zeng 2000; Hainmueller and Hazlett 2013; Hill Jr. and Jones 2014). These methods can often outperform more commonly used regression methods at predicting data not used to fit the model, which is useful for policymakers and serves as a useful check of the explanatory power of our theories (Hill Jr. and Jones 2014). Many of these methods are commonly thought of as “black box,” that is, they predict well, but do not permit substantive interpretation (Breiman 2001b). We show that this is *not* the case with a broadly applicable, powerful, and underappreciated method (in political science): random forests (Breiman 2001a). Random forests are especially useful to political scientists because of their ability to approximate arbitrary functional forms, be used with continuous, discrete, and censored (survival) outcomes, and because they permit substantive interpretation via permutation importance measures and the partial dependence algorithm. We provide an introduction to the theory and use of random forests, a substantive example drawn from the literature on comparative politics, and provide software to make substantive interpretation of random forests easy.

We think that random forests would be useful in political science when relevant theory says little about the functional form of the relationship of interest, when the magnitude and degree of nonlinearity and interaction is unknown, when the number of possibly relevant predictors is large, and when prediction is important. Random forests can approximate many nonmonotone, nonlinear functional forms. Interactions and nonlinearities are identified by random forests without prespecification, which decreases prediction error and allows researchers to study the relationships discovered by the algorithm. Random forests allow the inclusion of more predictors than observations. Though

[*] Zachary M. Jones is a Ph.D. student in political science at Pennsylvania State University (zmj@zmjones.com). Fridolin Linder is a Ph.D. student in political science at Pennsylvania State University (fridolin.linder@gmail.com).

this situation is not common in political science (though see the literature on behavioral genetics, wherein this problem does occur), related issues such as a highly correlated predictors are not an issue for random forests. Prediction is important for theory evaluation and for policymakers, and random forests' predictive performance relative to common parametric regression methods and other nonparametric methods is strong.

Random forests are composed of classification and regression trees (CART). A CART is created by recursive partitioning, which is a method for finding homogeneous (in the outcome variable) subsets of the data using the predictors (Breiman et al. 1984). Different definitions of homogeneity allow CART to be used with different types of outcomes. The fitted values of CART have low bias but high variance. Random forests were developed to solve this problem via two mechanisms. First, growing multiple decision trees with bootstrap samples of the data and then averaging over the predictions made by each tree: a procedure known as "bagging," and second, randomly selecting a subset of the predictors at each possible split in each CART. The random selection of predictors at each node allows variables that have a weak or variable relationship with the outcome to have an influence on the fitted values. Together bagging and the random selection of predictors at each node substantially decreases the variance of predictions at the cost of a small increase in bias (of the fitted values).

Since CART is central to the understanding of random forests we explain it in detail in the first section. We then discuss random forests, variations on random forests, and methods for extracting substantively relevant information from them. We then apply these methods to an empirical example using data on state repression. Finally we conclude with directions for future research and areas where these methods might be fruitfully applied.

Classification and Regression Trees

Classification and regression trees (CART) are a regression method that relies on repeated partitioning of the data to estimate the conditional distribution of a response given a set of predictors. Let the outcome of interest be a vector of observations $\mathbf{y} = (y_1, \ldots, y_n)^T$ and the set of explanatory variables or predictors a matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$, where $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^T$ for $j \in \{1, \ldots, p\}$. The goal of the algorithm is, to partition $\mathbf{y}$ conditional on the values of $\mathbf{X}$ in such a way that the resulting subgroups of $\mathbf{y}$ are as homogeneous as possible.

The algorithm works by considering every unique value in each predictor as a candidate for a binary split, and calculating the homogeneity of the subgroups of the outcome variable that would result by grouping observations that fall on either side of this value. Consider the (artificial) example in Figure 1. $\mathbf{y}$ is the vote choice of $n = 40$ subjects (18 republicans and 22 democrats), $\mathbf{x}_1$ is the ideology of the voter and $\mathbf{x}_2$ is the age.

The goal of the algorithm is to find homogeneous partitions of $\mathbf{y}$ given the predictors. The algorithm starts at the upper right panel of Figure 1, the complete data is the first node of the tree. We could classify all cases as Democrats yielding a misclassification rate of $18/40 = 0.45$. But it is obvious that there is some relationship between ideology and vote choice (the D's are mostly on the right side and the R's mostly on the left side), so we could do better in terms of classification error using this information. Formally the algorithm searches through all unique values of both predictors and calculates the number of cases that would be misclassified if a split would be made at that value and all cases on the left and right of this split are classified according to the majority rule. The upper right panel displays this step for one value of ideology (which also turns out to be the best possible split). In the tree in the lower left panel of Figure 1 the split is indicated by the two branches growing out of the first node. The variable name in the node indicates that the split was made on ideology. To the left of an ideology value of 3.31 most of the subjects voted Republican and on the right most voted Democrat. Therefore we classify all cases on the left and right as Republican and Democrat respectively (indicated by the shaded areas in the scatterplots). Now only 8 cases are misclassified, yielding an error rate of $8/40 = 0.2$.

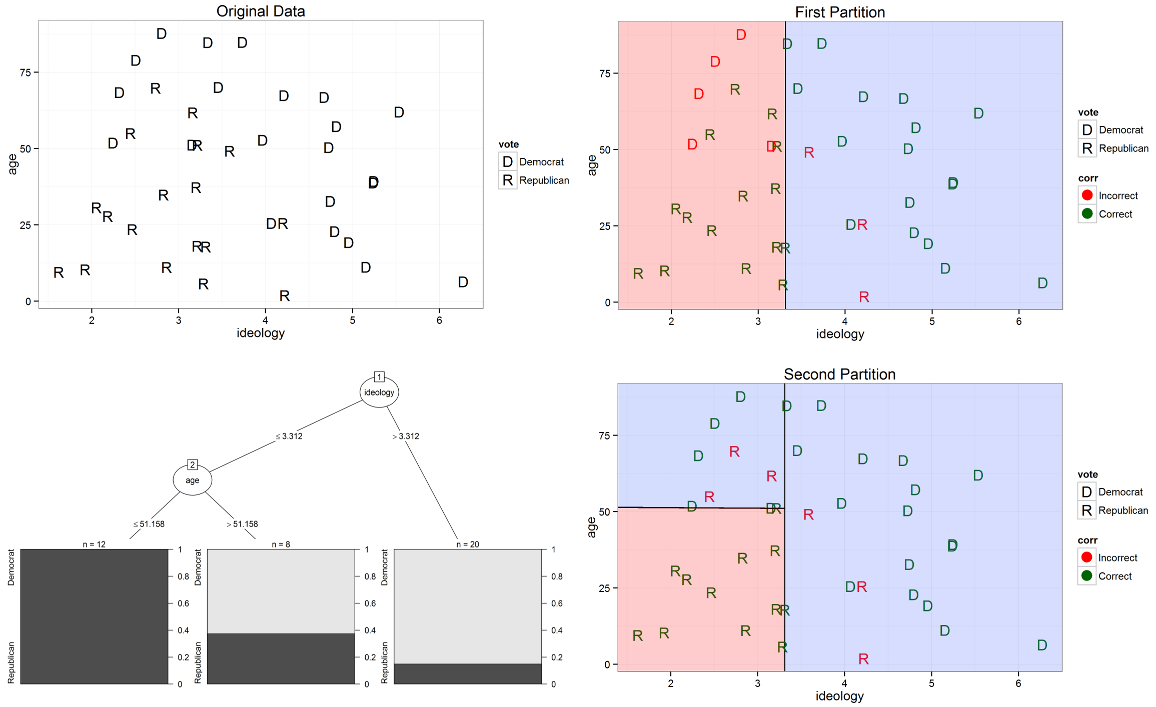

Figure 1: Visualization of a recursive partitioning algorithm for classification. The upper left panel displays the original data. The two panels on the right display the partitions of the original data after the first and the second split respectively. The lower left panel displays the corresponding decision tree. The blue and red shaded areas in the right panels indicate the value for the fitted value of the terminal node. The shading of the area visualizes classification as Republican (red) or Democrat (blue) by majority rule. The red colored letters indicate incorrect classifications under this rule.

The algorithm then looks for further splits within the two new partitions (left and right of $c_{x_1} = 3.21$. It turns out that for the right side there is no split that decreases the misclassification rate sufficiently (we talk about the criteria for making stopping decisions later). This is shown in the tree as a so called terminal node on the right branch of the ideology split. The plot in the terminal node displays the distribution of cases in this partition of the data.

However, age still contains information to improve the partitioning. At the second node (i.e. all data that falls left of the first split), when splitting the data into subjects older or younger then 51 years, we can obtain a completely homogeneous partition where all subjects voted Republican. Additionally those subjects older then 51 and with an ideology value lower than 3.21 are now classified as democrats. Note that the four democratic cases in this region of the data, which were misclassified before, are now correctly classified. The three republicans in the upper right partition are now misclassified. The classification error has therefore been reduced from 8/40 to 6/40.

We now extend the logic of CART from this very simple example of binary classification with two continuous predictors to other types of outcome variables. When extending the algorithm to other types of outcome variables we have to think about loss functions explicitly. In fact, we used a loss function in the illustration above. We calculated the classification error when just using the modal category of the outcome variable and argued that further splits of the data are justified because they decrease this error. More formally let $\mathbf{y}^{(m)} = (y_1^{(m)}, \dots, y_{n^{(m)}}^{(m)})$ and $\mathbf{X}^{(m)} = (\mathbf{x}_1^{(m)}, \dots, \mathbf{x}_p^{(m)})$ be the data at the current node m, $\mathbf{x}_s^{(m)}$ the predictor that is to be used for a split, with unique values $\mathcal{C}^{(m)} = \{x_i^{(m)}\}_{i \in \{1,\dots,n^{(m)}\}}$ and $c \in \mathcal{C}^{(m)}$ the value considered for a split. Then the data in the daughter nodes resulting from a split in c are $\mathbf{y}^{(m_l)}$ and $\mathbf{y}^{(m_r)}$. Where $\mathbf{y}^{(m_l)}$ contains all elements of $\mathbf{y}^{(m)}$ whose corresponding values of $\mathbf{x}_s^{(m)} \leq c$ and $\mathbf{y}^{(m_r)}$ all elements where $\mathbf{x}_s^{(m)} > c$. The gain (or reduction in error) from a split at node m in predictor $\mathbf{x}_s$ at value c is defined as:

$$\Delta(\mathbf{y}^{(m)}) = L(\mathbf{y}^{(m)}) - \left[\frac{n^{(m_l)}}{n^{(m)}} L(\mathbf{y}^{(m_l)}) + \frac{n^{(m_r)}}{n^{(m)}} L(\mathbf{y}^{(m_r)})\right]$$

.

Where $n^{(m_l)}$ and $n^{(m_r)}$ are the number of cases that fall to the right and to the left of the split, and $L(\cdot)$ is the loss function.

In the example above we made the intuitive choice to use the number of cases incorrectly classified when assigning the mode as the fitted value, divided by the number of cases in the node, as the loss function. This proportion can also be interpreted as the impurity of the data in the node, to return to our goal stated at the beginning: to partition the data in a way that produces homogeneous subgroups. Therefore it is intuitive to use the amount of impurity as a measure of loss. This is how the algorithm can be used for outcomes with more than two unique values (i.e. for nominal or ordinal outcomes with more than two categories, or continuous outcomes). By choosing a loss function that is appropriate to measure the impurity of a variable at a certain level of measurement, the algorithm can be extended to those outcomes.

For categorical outcomes, denote the set of unique categories of $\mathbf{y}^{(m)}$ as $\mathcal{D}^{(m)} = \{y_i^{(m)}\}_{i \in \{1,\dots,n^{(m)}\}}$. In order to asses the impurity of the node we first calculate the proportion of cases pertaining to each class $d \in \mathcal{D}^{(m)}$ and denote it as $p^{(m)}(d)$. Denote further the class that occurs most frequent as:

$$\hat{y}^{(m)} = \underset{d}{\operatorname{argmax}}\, p^{(m)}(d)$$

Then the loss function can be applied to obtain the impurity of the node. The familiar misclassification loss is obtained from:

$$L_d^{(m)}(\mathbf{y}^{(m)}) = \frac{1}{n^{(m)}} \sum_{i=1}^{n^{(m)}} \mathbb{I}(y_i^{(m)} \neq \hat{y}^{(m)}) = 1 - p^{(m)}(\hat{y}^{(m)})$$

.

Where $\mathbb{I}(\cdot)$ is the indicator function that is equal to one when its argument is true. This formalizes the intuition used above: the impurity of the node is the proportion of cases that would be misclassified under "majority rule."[1]

In the continuous case, the fitted value in a node is not calculated by majority vote. Typically the mean of the observations in that node is used as the predicted value for the observations in that node. To measure the impurity of the node usually the mean squared error (MSE) is used: $\hat{y}^{(m)} = \bar{y}^{(m)}$, where $\bar{y}^{(m)}$ is the mean of the observations in $\mathbf{y}^{(m)}$.[2]

[1] The other two loss functions that are most often used are the Gini loss $L_{\text{gini}}(\mathbf{y}^{(m)}) = \sum_{d \in \mathcal{D}^{(m)}} p^{(m)}(d)[1 - p^{(m)}(d)]$, and the entropy of the node $L_{\text{ent}}(\mathbf{y}^{(m)}) = -\sum_{d \in \mathcal{D}^{(m)}} p^{(m)}(d) \log[p^{(m)}(d)]$. Extensive theoretical (e.g. Raileanu and Stoffel 2004) and empirical (e.g. Mingers 1989) work in the machine learning literature concluded that the choice between those measures does not have a significant impact on the results of the algorithm.

[2] This algorithm can also be applied to censored data. See Ishwaran et al. (2008) and Hothorn et al. (2006) for details.

$$L_{\text{mse}}(\mathbf{y}^{(m)}) = \sum_{i=1}^{n^{(m)}} (y_i^{(m)} - \hat{y}^{(m)})^2$$

.

The extension to ordered discrete predictors is straightforward. Since the observed values of a continuous random variable are discrete, the partitioning algorithm described above works in the same way for ordered discrete random variables. Unorderd categorical variables are handled differently. If a split in category c of an unordered discrete variable is considered, the categorization in values to the left and to the right of c has no meaning since there is no ordering to make sense of "left" and "right." Therefore all possible combinations of the elements of $\mathcal{D}^{(m)}$ that could be chosen for a split are considered. This can lead to problems for variables with many categories. For an ordered discrete variable the number of splits that the algorithm has to consider is $|\mathcal{D}^{(m)}| - 2$, however, for an unordered variable it is $2^{|\mathcal{D}^{(m)}|-1} - 1$. This number gets large very quickly. For example the inclusion of a country indicator might be prohibitive if there are more than a handful of countries (e.g. if there are 21 countries in the sample the number of splits that have to be considered for that variable at each node is more than a million). Solutions to that problem are to include a binary variable for each category or to randomly draw a subset of categories at each node (See Louppe 2014 for details of the latter method).

After a loss function is chosen, the algorithm proceeds as described in our example. At each node m, $\Delta(\mathbf{y}^{(m)})$ is calculated for all variables and all possible splits in the variables. The variable-split combination that produces the highest Δ is selected and the process is repeated for the data in the resulting daughter nodes $\mathbf{y}^{(m_l)}$ and $\mathbf{y}^{(m_r)}$ until a stopping criterion is met. The stopping criterion may be that the tree has reached sufficient depth, that the number of observations that fall into the daughter nodes is too small, or the distribution of $\mathbf{y}^{(m)}$ is sufficiently homogeneous. These stopping criteria are arbitrary and should be understood as tuning parameters, that is, they should be chosen to minimize the expected generalization error. The longer partitioning continues, the smaller the resulting terminal nodes. That is, the resulting model is more tailored to the data. The result is the possibility of overfitting, resulting in higher generalization error. A predicted value for each observation is obtained, as in our example, by assigning a summary statistic for the terminal node the observation ended up in. For continuous data usually the mean of the distribution in the terminal node is used. For categorical data, either the majority category, or a vector of predicted probabilities for each category is assigned. Figure 2 illustrates how the predicted values from CART can approximate the function connecting the outcome and the predictor.

After a tree has been "grown" on the data, predicted values for new data can be obtained in a straightforward manner. Starting at the first node of the tree, a new observation i is "dropped down the tree", according to its values of the predictors $(x_{i1}, ..., x_{ip})$. That is, at each node, the observation is either dropped to the right or the left daughter node depending on its value on the predictor that was used to make a split at that node. This way, each new observation ends up in one terminal node. Then the predicted value of this terminal node is assigned as the prediction of the tree for observation i.

As previously mentioned CART has two main problems: fitted values have high variance and there is a substantial risk of overfitting. Fitted values can be unstable, producing different classifications when changes to the data used to fit the model are made (i.e., the estimator has high variance). There are several related reasons why this occurs. The first is that CART is locally optimal, that is, each split is optimal only at the node at which it occurs. Globally optimal partitioning is generally computationally intractable. Instead heuristic algorithms that are locally optimal (greedy) are used.[3] Given this locally optimal optimizaton, order effects result, that is, the order in which the variables

[3] Though see Grubinger, Zeileis, and Pfeiffer for an example of a stochastic search algorithm for this problem.

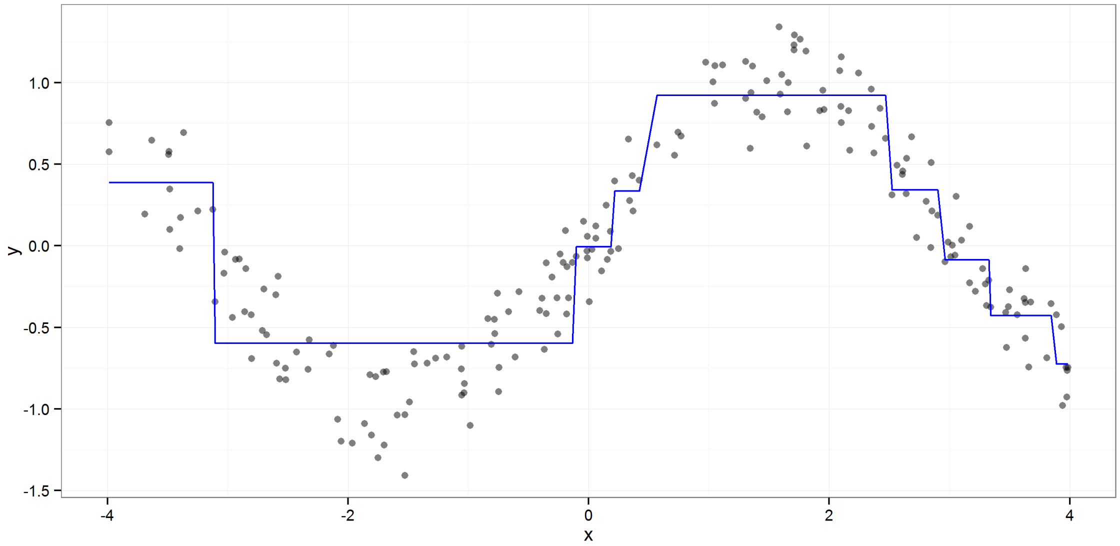

Figure 2: Approximating $\mathbf{y} = \sin(\mathbf{x}) + \epsilon$ with a regression tree which is a piecewise constant function.

are split can result in different resulting tree structures, and thus, different predictions. Random forests, which we discuss in the next section, have much lower variance and remove the effects of ordering.

Random Forests

In this section we demonstrate the utility of random forest and the methods to interpret it substantively described above, to explore relationships in complex data sets.

Breiman (1996) proposed bootstrap aggregating, commonly called "bagging," to decrease the variance of fitted values from CART. This innovation also can be used to reduce the risk of overfitting. A set of bootstrap samples are drawn from the data: samples drawn with replacement and of the same size as the original data. A CART is fit to each of these samples. Each bootstrap sample excludes some portion of the data, which is commonly referred to as the out-of-bag (OOB) data. Each tree makes predictions for the OOB data by dropping it down the tree that was grown without that data. Thus each observation will have a prediction made by each tree where it was not in the bootstrap sample drawn for that tree. The predicted values for each observation are combined to produce an ensemble estimate which has a lower variance than would a prediction made by a single CART grown on the original data. For continuous outcomes the predictions made by each tree are averaged. For discrete outcomes the majority class is used. Relying on the OOB data for predictions also eliminates the risk of overfitting since the each tree's prediction is made with data not used for fitting.

Breiman (2001a) extended the logic of bagging to predictors, resulting in random forests. Instead of choosing from all predictors for the split at each node, only a random subset of the predictors are used: increasing the diversity of splits across trees, which allows weaker predictors to have an opportunity to influence the models' predictions. This results in a further decrease in the variance of the fitted values (beyond bagging observations) and allows the use of large numbers of potentially relevant predictors (many more predictors than observations in some cases). A particular observation can fall in the terminal nodes of many trees in the forest, each of which, potentially, can give a different prediction. Again the OOB data, that is, data that was *not* drawn in the bootstrap sample used to fit a particular tree, is used to make each tree's prediction. For continuous outcomes, the prediction of the forest is then the average of the predictions of each tree:

$$\hat{f}(\mathbf{X}) = \frac{1}{T}\sum_{t=1}^{T} f^{(t)}(\mathbf{X}_{i \in \bar{\mathcal{B}}^{(t)}})$$

where T is the total number of trees in the forest, and $f^{(t)}(\cdot)$ is the t'th tree, $\bar{\mathcal{B}}^{(t)}$ is the out-of-bag data for the t'th tree, that is, observations in $\mathbf{X}^{(t)}$ and not in $\mathcal{B}^{(t)}$, the bootstrap sample for the t'th tree. For discrete outcomes, the prediction is the majority prediction from all trees that have been grown without the respective observation. Figure 3 displays the approximation to a function relating a continuous outcome to a single predictor obtained from a random selection of 25 trees from a random forest. It can be observed that the approximation is much smoother compared to the approximation by a single tree (see 2).

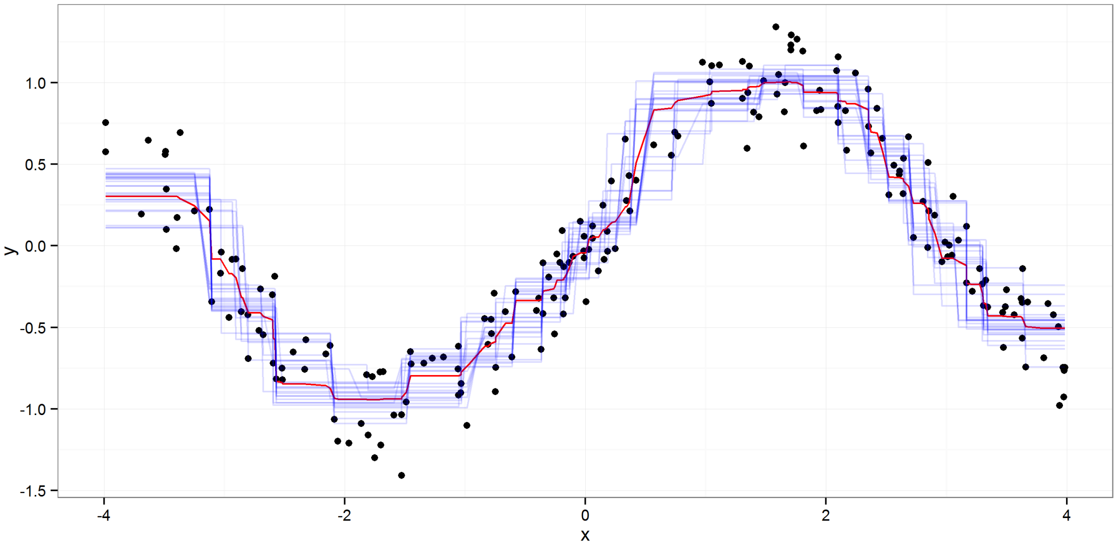

Figure 3: 25 randomly selected trees (shown in blue) in a random forest (prediction shown in red) each grown with a subsample (.632) of the training data.

The number of candidate predictors available at each node is a tuning parameter and should be chosen to minimize expected generalization error. Random forests compare favorably with other popular nonparametric methods in prediction tasks and can be interpreted substantively as well (See e.g., Breiman 2001a; Breiman 2001b; Cutler et al. 2007; Murphy 2012; Hastie et al. 2009).

Random forests are not without issue however. The CART which they are composed of often rely on biased splitting criteria: some types of variables, specifically variables with many unique values, are artificially preferred to variables with fewer categories, which has effects on measures of variable importance which are one of the ways random forests are interpreted substantively (Hothorn, Hornik, and Zeileis 2006; Strobl et al. 2007). Recent developments have resulted in unbiased recursive partitioning algorithms that separate the variable selection and split selection parts of the CART algorithm, and utilize subsampling rather than bootstrapping (Hothorn, Hornik, and Zeileis 2006).

Substantive Interpretation

Since, with random forests both predictors and observations are being sampled, no particular tree will give great insight into the model's overall prediction for each observation. There are, however, several ways to extract substantive insight from random forests. The most simple is partial dependence (Hastie et al. 2009). The partial dependence algorithm works as follows:

1. Let $\mathbf{x}_j$ be the predictor of interest, $\mathbf{X}_{-j}$ be the other predictors, $\mathbf{y}$ be the outcome, and $\hat{f}(\mathbf{X})$ the fitted forest.
2. For $\mathbf{x}_j$ sort the unique values $\mathcal{V} = \{\mathbf{x}_j\}_{i \in \{1,\ldots,n\}}$ resulting in $\mathcal{V}^*$, where $|\mathcal{V}^*| = K$. Create K new matrices $\mathbf{X}^{(k)} = (\mathbf{x}_j = \mathcal{V}_k^*, \mathbf{X}_{-j})$, $\forall\, k = (1, \ldots, K)$.
3. Drop each of the K new datasets, $\mathbf{X}^{(k)}$ down the fitted forest resulting in a predicted value for each observation in all k datasets: $\hat{\mathbf{y}}^{(k)} = \hat{f}(\mathbf{X}^{(k)})$, $\forall\, k = (1, \ldots, K)$.
4. Average the predictions in each of the K datasets, $\hat{y}_k^* = \frac{1}{n}\sum_{i=1}^{N} \hat{y}_i^{(k)}$, $\forall\, k = (1, \ldots, K)$.
5. Visualize the relationship by plotting $\mathbf{V}^*$ against $\hat{\mathbf{y}}^*$.

With slight modification, this method can also be used to visualize any joint relationships (i.e. interactions) the algorithm may have found. To do this create a dataset for each of the possible combinations of unique values of the explanatory variables of interest, predict the outcome in each of these observations, and then find the mean or modal prediction for each of these unique value combinations. For computational reasons we do not always use every unique value when an explanatory variable takes on more an arbitrary number of unique values. In this paper we use a random sample of 24 unique values that $\mathbf{x}_j$ takes on.[4] This logic can be generalized to joint relationships of an arbitrary dimension, but we limit ourselves here to pairwise partial dependence. The interpretation of partial dependence: the average predicted value for a particular value of an explanatory variable averaged within the joint values of the other predictors, is intuitive.

Another approach to extracting information from random forests relies on permutation tests for variable importance. Rather than attempting to characterize the partial dependence of one or more predictors on the response, the goal is instead to describe how the model's ability to predict y depends on a particular predictor. If a particular column of $\mathbf{X}$, say $\mathbf{x}_j$, is unrelated to $\mathbf{y}$, then randomly permuting $\mathbf{x}_j$ within $\mathbf{X}$ should not meaningfully decrease the model's ability to predict $\mathbf{y}$. However, if $\mathbf{x}_j$ is strongly related to $\mathbf{y}$, then permuting its values will produce a systematic decrease in the model's ability to predict $\mathbf{y}$, and the stronger the relationship between $\mathbf{x}_j$ and $\mathbf{y}$, the larger this decrease. Averaging the amount of change in the fitted values from permuting $\mathbf{x}_j$ across all the trees in the forest gives the marginal permutation importance of a predictor.[5] Formally, for classification, the importance of explanatory variable $\mathbf{x}_j$ in tree $t \in T$ is:

$$\text{VI}^{(t)}(\mathbf{x}_j) = \frac{\sum_{i \in \bar{\mathcal{B}}^{(t)}} \mathbb{I}(y_i = \hat{y}_i^{(t)})}{|\bar{\mathcal{B}}^{(t)}|} - \frac{\sum_{i \in \bar{\mathcal{B}}^{(t)}} \mathbb{I}(y_i = \hat{y}_{i\pi j}^{(t)})}{|\bar{\mathcal{B}}^{(t)}|}$$

$$\text{VI}(\mathbf{x}_j) = \frac{1}{T}\sum_{t=1}^{T} \text{VI}^{(t)}(\mathbf{x}_j)$$

[4] It is also possible to use an evenly spaced grid, however, this may result in extrapolation. Both of these options are implemented in our R package edarf.

[5] This measure is not truly marginal since the importance of a variable within a particular tree is conditional on all previous splits in the tree. It is possible to conduct a conditional permutation test which permutes $\mathbf{x}_j$ with variables related to $\mathbf{x}_j$ "held constant," reducing the possibility that a variable is deemed important when it is actually spurious (Strobl et al. 2008). However, this procedure is prohibitively costly in terms of computational resources.

where $\bar{\mathcal{B}}$ is the out-of-bag data for tree t, $\mathbf{x}_j$ is a particular predictor, $\hat{y}_i^{(t)}$ is the fitted value for observation i in tree t, $\hat{y}_{i\pi j}^{(t)}$ is the predicted class for the i'th observation after permuting $\mathbf{x}_j$, and $|\bar{\mathcal{B}}^{(t)}|$ is the number of observations *not* in the bootstrap sample used for fitting tree t. The importance of variable $\mathbf{x}_j$ in tree t is averaged across all trees to give the permutation importance for the forest (Breiman 2001a; Strobl et al. 2008). This can be thought of as testing the null hypothesis of independence between $\mathbf{x}_i$ and $\mathbf{y}$ as well as all other explanatory variables $\mathbf{X}_{-j}$. That is, $\mathbb{P}(\mathbf{y}, \mathbf{x}_j, \mathbf{X}_{-j}) = \mathbb{P}(\mathbf{x}_j)\mathbb{P}(\mathbf{y}, \mathbf{X}_{-j})$. For regression, the permutation is defined similarly, by the average increase in the mean squared error across trees that results from permuting $\mathbf{x}_j$.

Dependent Data

As previously mentioned, these methods are not designed for use with dependent data, such as is common in political science. Not modeling the dependence structure may decrease predictive performance on new data and mislead us about the importance of variables strongly related to different features of the unmodeled dependence structure. In its basic implementation the estimated regression function is the result of complete pooling of the data. There are several ways this dependence can be modeled. One way is to include a categorical variable with unit indicators as an explanatory variable. Then, this explanatory variable has a chance of being included in the set of variables that the algorithm may select for splitting at a particular node. This is computationally intensive and not always possible. Alternatively, a random effects approach could be used: the outcome of interest is treated as a function of an unknown regression function, which is estimated using random forests and completely pools the data, and set of unit random effects for which we estimate the variance, and idiosyncratic error which is assumed uncorrelated with the random effects (Hajjem, Bellavance, and Larocque 2014; Hajjem, Bellavance, and Larocque 2011). Alternative approaches include sampling units (Adler et al. 2011) or sampling units and then an observation from within that unit (Adler et al. 2011). Alternatively, the analyst can transform the dependent variable, such as by subtracting the within-unit mean. This, however, invalidates the use of the fitted random forest for the prediction of new data. In the following applications we discuss what approach we use and leave the development of other approaches to future work.

Missingness

Missing values are a perennial problem in real world data. Often missingness is ignored, which, at best, decreases the precision of estimates of model parameters, and at worst biases them. Multiple imputation is frequently used to provide estimates of the values of missing values: this relies on the assumption that, conditional on the observed covariates, missingness is random (Rubin 2004; Honaker and King 2010). When analysis is exploratory or predictive, missingness has a different epistemological status and can be *useful* (Shmueli 2010). When the goal is prediction alone, informative missingness improves our ability to predict the outcome, and when doing exploration, understanding how missingness is informative can be useful for future work. When predictors with missingness are categorical, missingness can simply be recoded as an additional category. However, with continuous variables this is not possible nor is it always desireable in the categorical case. When using permutation importance to measure the degree to which a model's fit is degraded by permuting said variable recoding in this way conflates missingness that is informative with how informative the observed values are. An alternative would be to create a set of variables that indicate whether missingness in a particular predictor (or set of predictors) is present. This does not fix the problem of partially observed predictors, but this can be handled using multiple imputation or other ways to handle missing values such as surrogate splitting, which, since unbiased estimation of causal effects is not the goal, do not have to satisfy the stringent conditionally missing at random assumption previously mentioned. Frequently data are imputed once, and the data are treated as completely observed. This does however, result in a tendency to underestimate generalization error.

State Repression

There is a large literature in political science on the determinants of state repression, especially violations of physical integrity such as killings, disappearances, torture, and political imprisonment. Quantitative analysis of cross-national patterns of state repression relies on annual country reports from Amnesty International and the United States Department of State, which are used to code ordinal measures of state repression such as the Cingranelli and Richards Physical Integrity Index and the Political Terror Scale (Cingranelli and Richards 2010; Wood and Gibney 2010). We use the measure from Fariss (2014) which is based on a dynamic measurement model, which aggregates information from multiple sources on state repression in each country-year into a continuous measure.

The literature on state repression has, like most areas of political science, been dominated by data modeling. A particular stochastic model for the data is assumed, and the parameters of the model are estimated using the data on hand. Then statistical inference about the model's parameters is used to support claims about the generality of the results (implicitly about causality or population quantities). It seems unlikely that the assumed model of the data generating process is correctly specified or even a reasonable approximation given the complexity of the dynamics of dissent and repression well known from micro-level quantitative and qualitative research, leaving open the question of what can be learned from estimates of parameters that govern such models absent strong predictive validity. We think that researchers often sell the predictive power of their theories short by making strict assumptions about functional form that are not a part of their theory.

We instead aim to explore the available data, summarizing the relationships between the an expanded set of variables identified in Hill Jr. and Jones (2014) as important explanations for state repression in the extant literature and the measure of respect for physical integrity rights from Fariss (2014).[6] We do this using random forests. We present the marginal permutation importance of each variable and the partial dependence of the estimated function on the values of a subset of the explanatory variables as well as a subset of the large set of possible two-dimensional joint dependencies discovered by the model. This builds on the analysis of Hill Jr. and Jones (2014) by examining the substantive relationships between the set of explanatory variables and Fariss' measure of state repression; only the marginal permutation importance is considered in the original paper. Additionally, we check the predictive performance of random forests on this data as well as comparing it to least angle regression (LARS), ordinary least squares (OLS), and support vector regression machines (SVM) (Efron et al. 2004; Drucker et al. 1997).

To reiterate, the marginal permutation importance shows the mean decrease in prediction error that results from randomly permuting an explanatory variable. If the variable is an important predictor, permuting its values will result in a systematic increase in prediction error, while an unimportant predictor will result in no decrease or a random decrease in prediction error. Figure 4 shows the permutation importance of predictors of state repression taken from Hill Jr. and Jones (2014). The rank ordering of the predictor importance is somewhat different from that of the original study, though the set of data used here is somewhat different.

Participation competitiveness, a component of Polity that measures "the extent to which alternative preferences for policy and leadership can be pursued in the political arena," is the most important predictor according to this measure (Marshall and Jaggers 2009). That is, the prediction error

[6] As discussed in the last paragraph of the sub-section title "Classification and Regression Trees" it is not possible to include a categorical variable of country labels in this example. There are 189 countries in the data for this example, resulting in $2^{189-1} - 1$ possible splits.

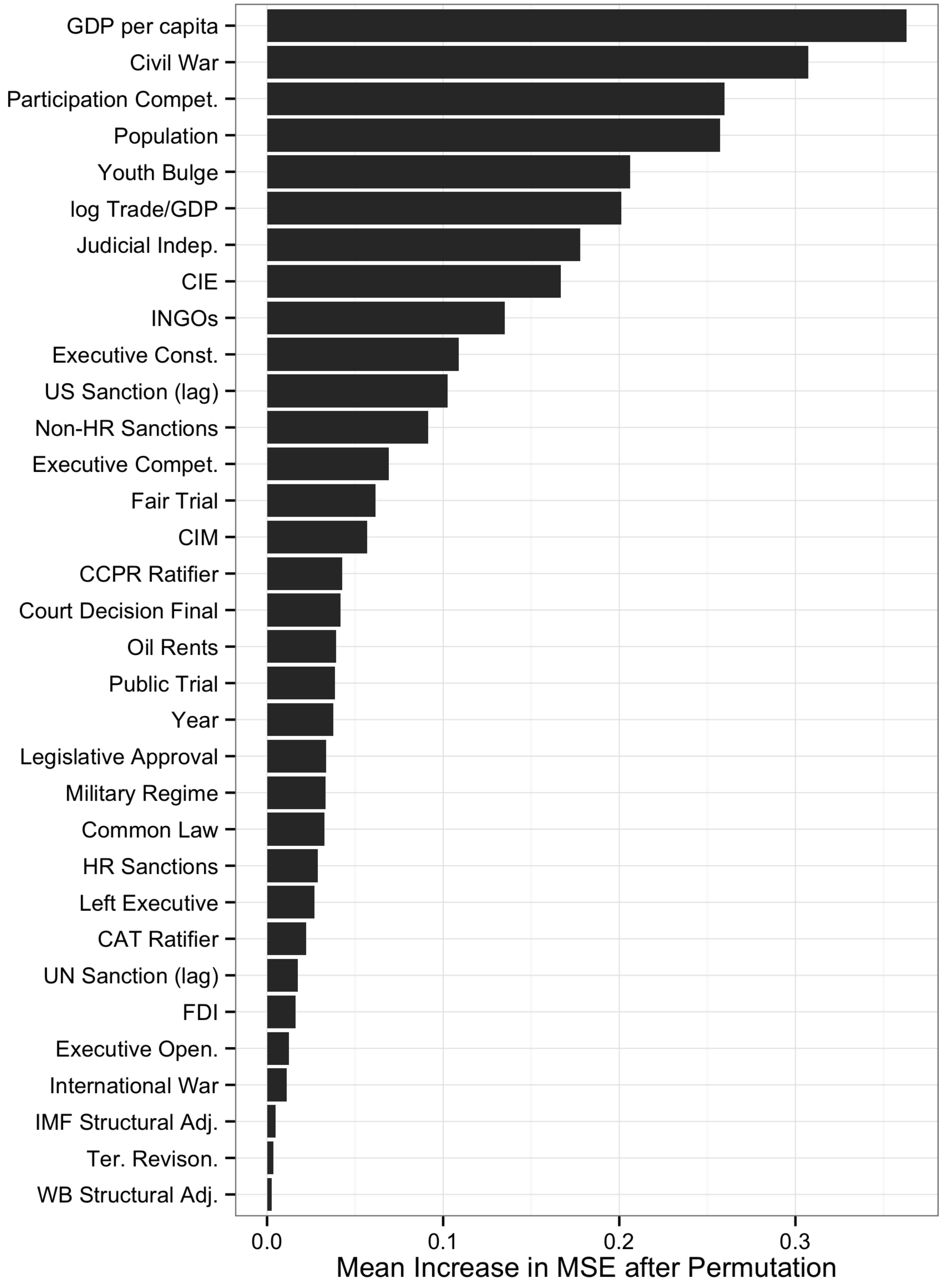

Figure 4: The marginal permutation importance of explanatory variables estimated using random forests, using the dynamic latent measure of state repression developed by Fariss (2014). The bar shows the mean decrease in mean squared error that results from randomly permuting the independent variable indicated. If the variable is important, permuting its values should systematically *decrease* predictive performance, whereas an unimportant variable should produce no decrease, or a random decrease in mean squared error. The error bars result the .025 and .975 quantiles of the distribution of this statistic from the repeated draws from the posterior of the latent variable estimated by Fariss (2014).

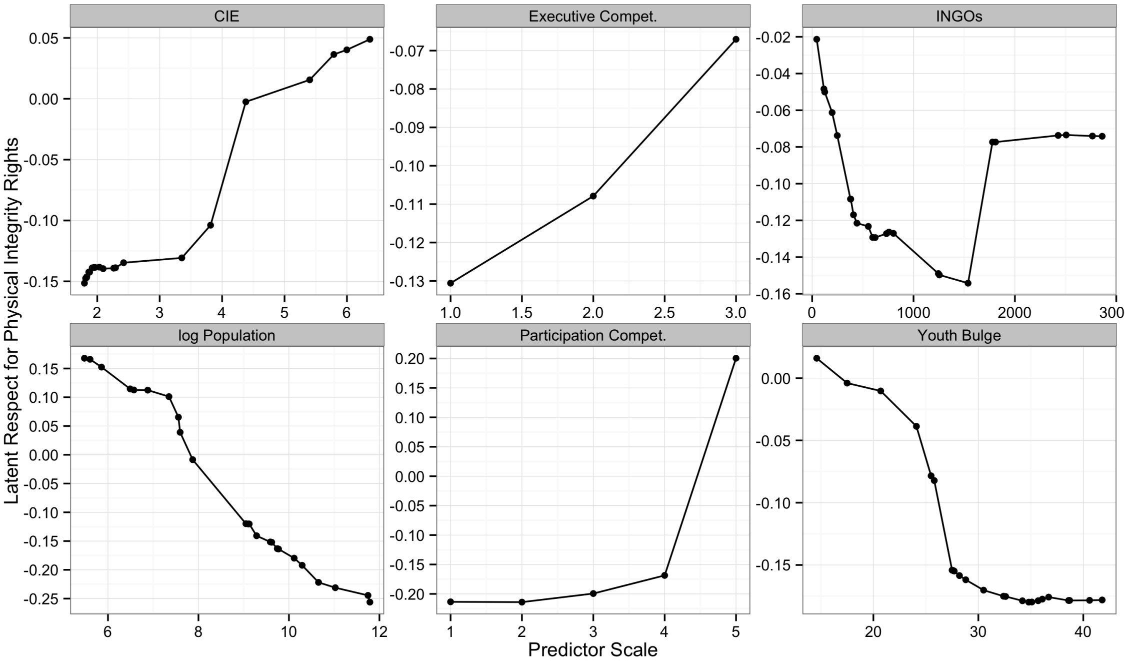

Figure 5: The partial dependence of the response of year, percentage of the population that is male and aged 15-25, executive constraint and participation competitiveness (from Polity), the natural logarithm of population, a count of the number of INGOs of which a government's citizens are members measured by Hafner-Burton and Tsutsui (2005) (labeled INGOs), and contract intensive economy (CIE), a measure of life insurance contracting used by Mousseau and Mousseau (2008) (from Beck and Webb (2003)). On the y-axis is the dynamic latent measure of repression developed by Fariss (2014) and on the x-axis the predictor indicated by the above label. Note that the scale of y differs for each sub-plot. The relationship between each predictor and the response is averaged within the joint values of the other predictors conditional on the relationships discovered by the random forest. All values of the explanatory variable for which there are predictions are observed or imputed values.

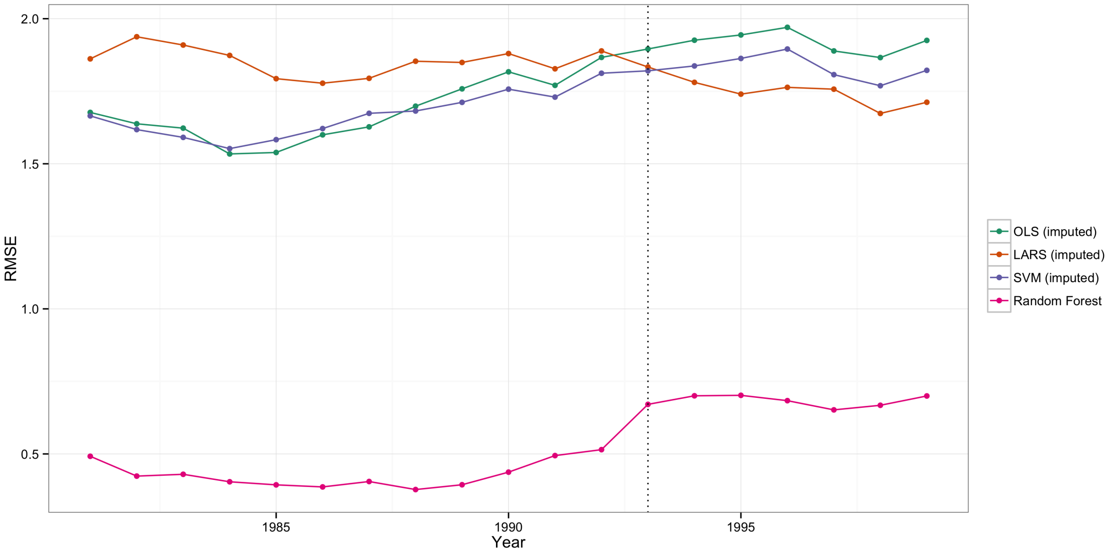

Figure 6: The root mean squared error (RMSE) computed using for the training set (1981-1992) and the test set (1993-1999). The RMSE is calculated for each model fit to a random draw (100 in total) from the posterior of the measure of respect for physical integrity due to Fariss (2014), and the error bars show the lower .025 and upper .975 quantiles. The point shows the median. The models shown are Least Angle Regression (LARS) (Efron et al. 2004), a basis-expanded Generalized Additive Model regularized via the adaptive least absolute shrinkage and selection operator (LASSO) (Kenkel and Signorino 2013; Tibshirani 1996), and a random forest (Breiman 2001a).

increases a substantial amount (30%) when it is randomly shuffled. This should not be surprising, since, as noted in Hill Jr. (2013) and Hill Jr. and Jones (2014), there is conceptual overlap between this measure and measures of state repression. The natural logarithm of population and GDP per capita, and civil war are unsurprisingly important as well. More interesting are the importance of measures of contract intensive economy (CIE, used by Mousseau and Mousseau (2008)), rule of law (from the International Country Risk Guide), judicial independence (from Cingranelli and Richards (2010)), and youth bulges (measured by Urdal (2006) and used by Nordås and Davenport (2013)), which represent relatively understudied concepts that may be worth further investigation.

The partial dependence plots shown in Figure 5 indicates that the model discovers a variety of nonlinearity which may be missed by analysts relying on more restrictive data models. These plots show the predicted values of repeated draws of the latent measure of state repression averaged within the values of the other variables included. However it is possible that this "nonlinearity" is due to interactions with other included or excluded explanatory variables, or reflect a pattern due to unmodeled dependence in the data. Since this is EDA this can be investigated, for example by looking at the average fitted value within values of multiple predictors. A plot showing partial dependence for all of the predictors is available in the Appendix. The random forest has clearly discovered patterns in the data that do not comport with the assumptions commonly made about the function mapping the predictors to the response in commonly used data models. We suggest that a great deal more value could be extracted from observational data already collected but analyzed with methods that impose structure on the data that is inappropriate. Additionally, the quantities extracted from the model pertain directly to prediction (in the case of permutation importance), or are the result of a fitting process that is primarily concerned with prediction (in the case of partial dependence), which we think makes their application to observational data in political science quite natural.

Lastly we examine how well the model does at predicting latent respect for physical integrity rights. We split the data into two sets: a training set that covers 1981-1992 and a test set that covers 1993-1999. We repeatedly draw from the latent outcome variable's distribution, fit a model to the training set, predict the outcome in the test set, calculate the root mean square error,[7] and then summarize the distribution of these error statistics across a large number of draws from the latent outcome variable using the lower .025 and upper .975 quantiles and the median. Absent intuitive scale for the outcome variable it is hard to judge the model's performance in absolute terms, thus we provide comparisons to other models. One thing to note is that there is an increase in the prediction error over time within the training and test data suggesting that there there is over-time variation in respect for physical integrity rights that we are not able to explain with the included measures. We hope that this provides a useful predictive benchmark for this measure of physical integrity rights.

Conclusion

In situations where relevant theory says little about the functional form of the relationship of interest, the magnitude and degree of nonlinearity and interaction is unknown, the number of possibly relevant predictors is large, or when prediction is important, random forests may be more useful than a generalized linear model. We believe that this situation is common, especially in comparative politics and international relations.

We have provided a technical introduction to CART and random forests, with the goal of providing researchers with enough background knowledge to use these methods in an informed manner, as well as companion software to reduce the technical burden of their use. Random forests are but one member of a broad class of supervised machine learning methods that could be useful to political scientists. We hope that in the future these methods and the issues that they are designed to solve

[7] $\text{RMSE}(\mathbf{y}, \hat{\mathbf{y}}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$.

are incorporated into political science. Particularly, we see the development of machine learning methods for dependent data as a fruitful area for future research.

References

Appendix

Adler, Werner, Alexander Brenning, Sergej Potapov, Matthias Schmid, and Berthold Lausen. 2011. "Ensemble Classification of Paired Data." *Computational Statistics & Data Analysis* 55 (5). Elsevier: 1933–41.

Beck, Nathaniel, and Simon Jackman. 1998. "Beyond Linearity by Default: Generalized Additive Models." *American Journal of Political Science* 42. University of Texas Press: 596–627.

Beck, Nathaniel, Gary King, and Langche Zeng. 2000. "Improving Quantitative Studies of International Conflict: A Conjecture." *American Political Science Review*. JSTOR, 21–35.

Beck, Thorsten, and Ian Webb. 2003. "Economic, Demographic, and Institutional Determinants of Life Insurance Consumption Across Countries." *The World Bank Economic Review* 17 (1). World Bank: 51–88.

Breiman, Leo. 1996. "Bagging Predictors." *Machine Learning* 24 (2). Springer: 123–40.

———. 2001a. "Random Forests." *Machine Learning* 45 (1). Springer: 5–32.

———. 2001b. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." *Statistical Science* 16 (3). The Institute of Mathematical Statistics: 199–231. doi:10.1214/ss/1009213726.

Breiman, Leo, Jerome Friedman, Charles J Stone, and Richard A Olshen. 1984. *Classification and Regression Trees*. CRC press.

Cingranelli, David L., and David L. Richards. 2010. "The Cingranelli-Richards (CIRI) Human Rights Dataset." http://www.humanrightsdata.org.

Cutler, D Richard, Thomas C Edwards Jr, Karen H Beard, Adele Cutler, Kyle T Hess, Jacob Gibson, and Joshua J Lawler. 2007. "Random Forests for Classification in Ecology." *Ecology* 88 (11). Eco Soc America: 2783–92.

Drucker, Harris, Chris JC Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. 1997. "Support Vector Regression Machines." *Advances in Neural Information Processing Systems* 9. Morgan Kaufmann Publishers: 155–61.

Efron, Bradley, Trevor Hastie, Iain Johnstone, Robert Tibshirani, and others. 2004. "Least Angle Regression." *The Annals of Statistics* 32 (2). Institute of Mathematical Statistics: 407–99.

Fariss, Christopher J. 2014. "Respect for Human Rights Has Improved over Time: Modeling the Changing Standard of Accountability." *American Political Science Review*. Cambridge Univ Press, 1–22.

Grubinger, Thomas, Achim Zeileis, and Karl-Peter Pfeiffer. "Evtree: Evolutionary Learning of Globally Optimal Classification and Regression Trees in R."

Hafner-Burton, Emilie M, and Kiyoteru Tsutsui. 2005. "Human Rights in a Globalizing World: The Paradox of Empty Promises1." *American Journal of Sociology* 110 (5). JSTOR: 1373–1411.

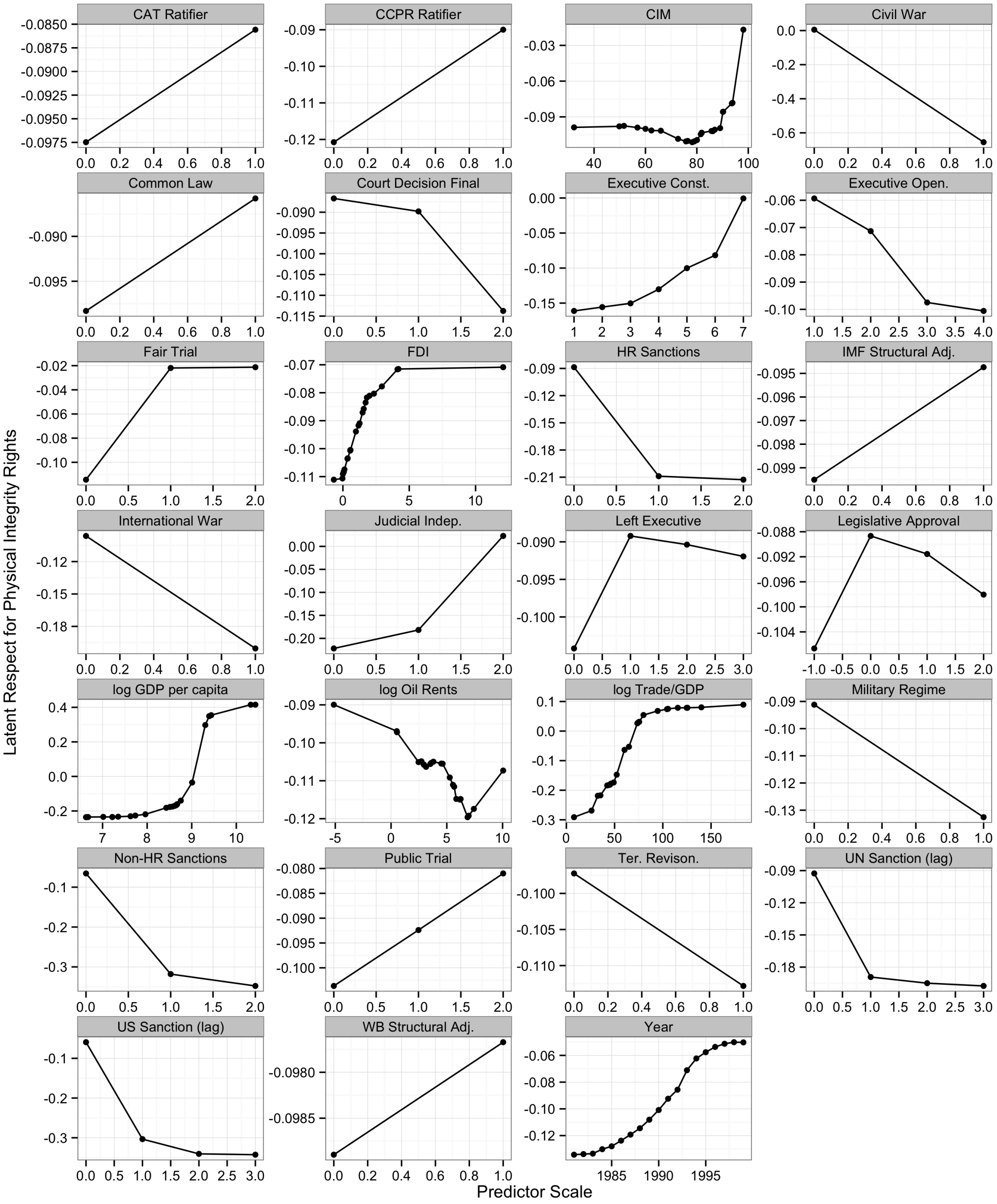

Figure 7: Two-way partial dependence for all predictors included in the model of latent respect for physical integrity rights. The interpretation is the same as Figure 5. Note that imputation of missing values results in some categorical/binary variables taking on non-integer values.

Hainmueller, Jens, and Chad Hazlett. 2013. “Kernel Regularized Least Squares: Reducing Misspecification Bias with a Flexible and Interpretable Machine Learning Approach.” *Political Analysis.* SPM-PMSAPSA, mpt019.

Hajjem, Ahlem, François Bellavance, and Denis Larocque. 2011. “Mixed Effects Regression Trees for Clustered Data.” *Statistics & Probability Letters* 81 (4). Elsevier: 451–59.

———. 2014. “Mixed-Effects Random Forest for Clustered Data.” *Journal of Statistical Computation and Simulation* 84 (6). Taylor & Francis: 1313–28.

Hastie, Trevor, Robert Tibshirani, Jerome Friedman, T Hastie, J Friedman, and R Tibshirani. 2009. *The Elements of Statistical Learning.* Vol. 2. 1. Springer.

Hill Jr., Daniel W. 2013. “The Concept of Personal Integrity Rights in Empirical Research.”

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

Honaker, James, and Gary King. 2010. “What to Do About Missing Values in Time-Series Cross-Section Data.” *American Journal of Political Science* 54 (2). Wiley Online Library: 561–81.

Hothorn, Torsten, Peter Bühlmann, Sandrine Dudoit, Annette Molinaro, and Mark J Van Der Laan. 2006. “Survival Ensembles.” *Biostatistics* 7 (3). Biometrika Trust: 355–73.

Hothorn, Torsten, Kurt Hornik, and Achim Zeileis. 2006. “Unbiased Recursive Partitioning: A Conditional Inference Framework.” *Journal of Computational and Graphical Statistics* 15 (3).

Ishwaran, Hemant, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. 2008. “Random Survival Forests.” *The Annals of Applied Statistics* 2 (3). JSTOR: 841–60.

Kenkel, Brenton, and Curtis S Signorino. 2013. “Bootstrapped Basis Regression with Variable Selection: A New Method for Flexible Functional Form Estimation.” *Manuscript, University of Rochester.*

Louppe, Gilles. 2014. “Understanding Random Forests: From Theory to Practice.” *ArXiv Preprint ArXiv:1407.7502.*

Marshall, Monty, and Keith Jaggers. 2009. *Polity IV Project: Political Regime Characteristics and Transitions, 1800-2007. Data Users’ Manual.* Center for Systemic Peace.

Mingers, John. 1989. “An Empirical Comparison of Selection Measures for Decision-Tree Induction.” *Machine Learning* 3 (4). Springer: 319–42.

Mousseau, Michael, and Demet Yalcin Mousseau. 2008. “The Contracting Roots of Human Rights.” *Journal of Peace Research* 45 (3). Sage Publications: 327–44.

Murphy, Kevin P. 2012. *Machine Learning: A Probabilistic Perspective.* The MIT Press.

Nordås, Ragnhild, and Christian Davenport. 2013. “Fight the Youth: Youth Bulges and State Repression.” *American Journal of Political Science.* Wiley Online Library.

Raileanu, Laura Elena, and Kilian Stoffel. 2004. “Theoretical Comparison Between the Gini Index and Information Gain Criteria.” *Annals of Mathematics and Artificial Intelligence* 41 (1). Springer: 77–93.

Rubin, Donald B. 2004. *Multiple Imputation for Nonresponse in Surveys.* Vol. 81. John Wiley & Sons.

Shmueli, Galit. 2010. “To Explain or to Predict?” *Statistical Science* 25 (3). Institute of Mathematical Statistics: 289–310.

Strobl, Carolin, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. 2008. “Conditional Variable Importance for Random Forests.” *BMC Bioinformatics* 9 (1). BioMed Central Ltd: 307.

Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. 2007. “Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution.” *BMC Bioinformatics* 8 (1). BioMed Central Ltd: 25.

Tibshirani, Robert. 1996. “Regression Shrinkage and Selection via the Lasso.” *Journal of the Royal Statistical Society. Series B (Methodological)*. JSTOR, 267–88.

Urdal, Henrik. 2006. “A Clash of Generations? Youth Bulges and Political Violence.” *International Studies Quarterly* 50 (3). Wiley Online Library: 607–29.

Wood, Reed M., and Mark Gibney. 2010. “The Political Terror Scale (PTS): A Re-Introduction and a Comparison to CIRI.” *Human Rights Quarterly* 32 (2). The Johns Hopkins University Press: 367–400.