Online Appendix for An Empirical Evaluation of Explanations for State Repression*

Daniel W. Hill, Jr.† and Zachary M. Jones‡

Stability of Random Forest Results

The random forest algorithm used has two main tuning parameters, the number of variables selected at each node and the number of trees grown in the forest. If the results are unstable across these tuning parameters this casts doubt on the reliability of the results. To verify the stability of the results we estimate permutation importance using random forests with different combinations of tuning parameters. We rank the importance scores for each model and dependent variable combination and compare the concordance in rankings across different tuning parameter combinations for each dependent variable. We present summary statistics below, however, the replication archive enables the curious reader to examine the ranks or raw importance scores directly, for any tuning parameter and dependent variable combination considered.

	W	α
Disappearances	0.95	0.95
Killings	0.94	0.93
Political Imprisonment	0.98	0.98
Torture	0.96	0.96
Political Terror Scale	0.95	0.95
Physical Integrity Index	0.97	0.97
Dynamic Latent Score	0.98	0.98

Table 1: Kendall's coefficient of concordance W and Krippendorff's α, computed for the ranks of the permutation importance estimates for each dependent variable (indicated by the rows) across different values of the tuning parameters. The number of variables considered at each split is set to be 3, 5, 10, 15 (the value used in the main results is 10), and the number of trees grown in each forest is set to be 500, 1000, or 3000. Every possible combination of these tuning parameters is used.

*Complete history of the code and manuscript are available at http://github.com/zmjones/eeesr/, along with the data and further information about how to reproduce these analyses. Thanks to Christopher Fariss, Luke Keele, and Will Moore for helpful comments.

†Assistant Professor, Department of International Affairs, University of Georgia. email: dwhill@uga.edu. Responsible for the research question, design of the cross-validation analysis, selection of the data, and the majority of the writing.

‡Ph.D. student, Department of Political Science, Pennsylvania State University. email: zmj@zmjones.com. Responsible for design of the random forest analysis and multiple imputation, all data analysis and visualization, and description of the methods.

For the main results we set 10 variables to be randomly selected for each node and 1000 trees to be grown for each forest. Additionally, we use the bootstrap to display the stability of the permutation importance estimates subject to variation in the input data. While it is possible to bootstrap each of the possible tuning parameter and dependent variable combinations above and then compare the overlap of the simulated sampling distribution of the permutation importance scores, the necessary computation time was extreme. For the interested reader it simple to make this comparison given our replication code.

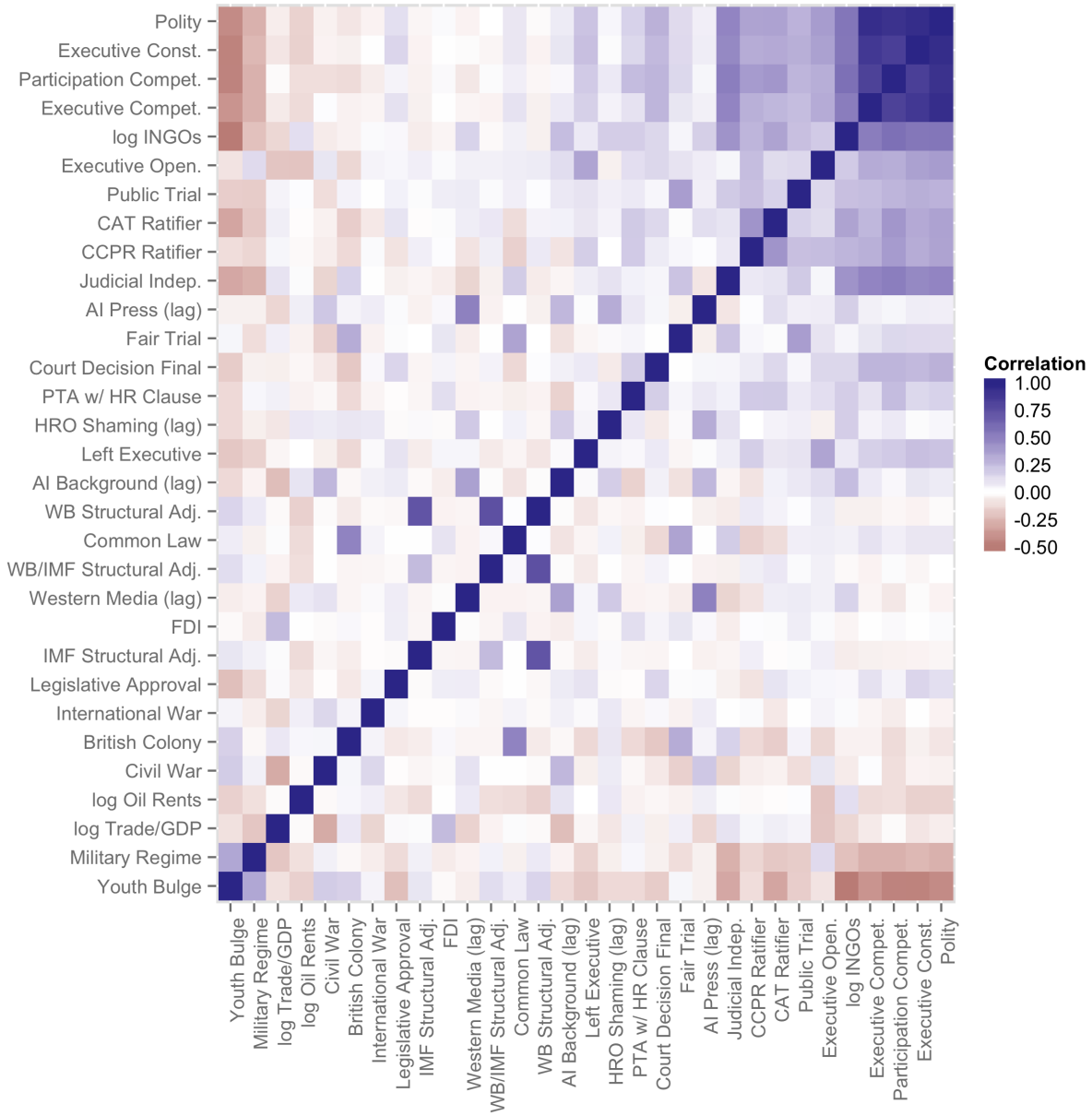

Figure 1: Correlations Between All Covariates. Correlations between numeric variables are Pearson product-moment correlations, correlations between numeric and ordinal variables are polyserial correlations, and correlations between ordinal variables are polychoric correlations.

Imputation

To impute missing values in our data we use a combination of random forests and the random indicator method (Buuren and Groothuis-Oudshoorn 2011; Jolani 2012). The random indicator method is designed to deal with data that is not missing at random, that is, when the missing values are related to the probability of missingness. Imputing data that is not missing at random (NMAR) as if it is missing at random (MAR) may skew the distribution of the variable, biasing results. Formally, for NMAR data, $\mathbb{P}(Y|X, R = 1) \neq \mathbb{P}(Y|X, R = 0)$ if we let Y be a random variable with some missingness, R be the indicator function which equals 1 when the value of Y is observed and 0 otherwise, and X be a fully-observed covariate. The random indicator method works by first estimating a model of the probability of response $\hat{R}$ and comparing the conditional distribution of Y given R and $\hat{R}$. An offset $\delta_R = \mathbb{E}(Y|R = 1, \hat{R} = 1) - \mathbb{E}(Y|R = 1, \hat{R} = 0)$ is estimated, which, under the assumption that the variance of the missing values and the observed values of Y are equal, is equivalent to $\delta_{NR} = \mathbb{E}(Y|R = 0, \hat{R} = 1) - \mathbb{E}(Y|R = 0, \hat{R} = 0)$ (Jolani 2012). We use the random indicator method for all numeric covariates and random forests for all binary, ordinal, or categorical covariates (excluding Polity and its components, which are imputed using the random indicator method). The distribution of the imputed and observed values for all variables with missingness are shown in Figure 2. Note that when data are NMAR we should expect the location of the distribution of imputed values to be different than the observed values.

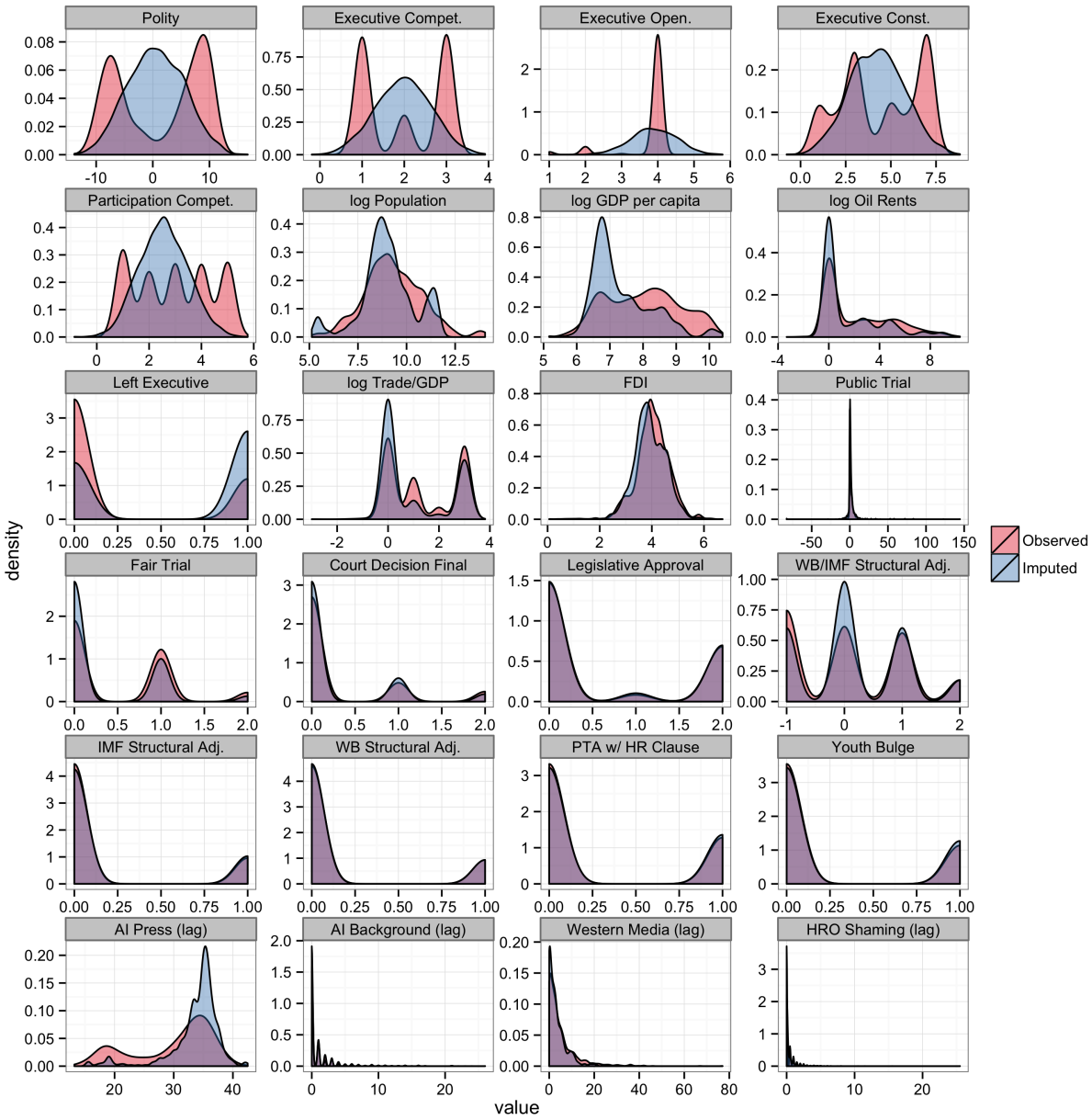

Figure 2: The distribution of observed (shown in red) and imputed (shown in blue) values for all covariates with missingness. Categorical variables are imputed using random forests and numeric variables are imputed using the random indicator method.

References

Buuren, Stef and Karin Groothuis-Oudshoorn. 2011. "MICE: Multivariate imputation by chained equations in R." *Journal of statistical software* 45(3).

Jolani, S. 2012. Dual Imputation Strategies for Analyzing Incomplete Data PhD thesis University of Utrecht.