An Empirical Evaluation of Explanations for State Repression[*]

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

Abstract

The empirical literature that examines cross-national patterns of state repression seeks to discover a set of political, economic, and social conditions that are consistently associated with government violations of human rights. Null hypothesis significance testing is the most common way of examining the relationship between repression and concepts of interest, but we argue that it is inadequate for this goal, and has produced potentially misleading results. To remedy this deficiency in the literature we use cross-validation and random forests to determine the *predictive* power of measures of concepts the literature identifies as important causes of repression. We find that few of these measures are able to substantially improve the predictive power of statistical models of repression. Further, the most studied concept in the literature, democratic political institutions, predicts certain kinds of repression much more accurately than others. We argue that this is due to conceptual and operational overlap between democracy and certain kinds of state repression. Finally, we argue that the impressive performance of certain features of domestic legal systems, as well as some economic and demographic factors, justifies a stronger focus on these concepts in future studies of repression.

Introduction

The past 20-30 years has witnessed the tremendous growth of an empirical, quantitative literature that examines cross-national patterns of state repression (See, e.g., **??????????????????????????**). The general purpose of this literature is to discover a set of political, economic, and social conditions that are consistently associated with government violations of the most basic human rights.[1] In other words, this literature aims to answer why some governments violate basic human rights more than others. This is an enormously important question since it relates directly to one of the fundamental problems of politics, which is how an entity given the exclusive authority to enforce rules through physical coercion (the state) can be prevented from abusing that authority (See, e.g., **?**). This literature deals specifically with violent, egregious abuses of such authority, but also addresses this broader problem which has clear implications for questions about democratization and the emergence of genuine constraints on government behavior (E.g., **??**).

[*] 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.

[1] "The most basic human rights" means freedom from political imprisonment, torture, kidnapping, and extrajudicial execution, generally referred to as personal, or physical, integrity rights (See, e.g. **??**).

Though the basic research question explored by this literature is of tremendous intrinsic importance, the standards currently used to assess claims about the causes of state repression are inadequate for the goals of this research. Specifically, scholars nearly always employ null hypothesis tests of statistical significance to determine if a particular covariate is a meaningful determinant of state repression. Under this approach, covariates whose coefficients achieve a *p*-value smaller than some arbitrary threshold (usually 0.05) are declared important determinants of state repression. Using this criteria, the literature has uncovered a number of empirical findings with respect to state repression. Some concepts have been so consistently related to repression that researchers are now effectively obligated to include measures of them in their models.[2] This list of "usual suspects" now includes, at minimum, measures of GDP per capita, population size, civil and international war, and democratic political institutions. Beyond these relationships, the list of concepts that influence repression has been steadily expanded to include international factors such as INGO presence and behavior (**???**), a country's position in the world economy (**?**), and participation in international financial institutions (**??**), to name just a few. Other recent additions to the list include a host of domestic legal institutions such as constitutional provisions for basic rights (**??**) and common law heritage (**?**).

While the current approach has value we argue that, by itself, the standard analysis used in the literature is incomplete at best, and is possibly misleading. First, since variables that are statistically significant may not meaningfully increase the ability of a model to predict the outcome of interest (**?**), the current approach effectively ignores the ability of a model to *predict* state repression. Second, since scholars routinely use all of the data to fit their models, they have no way of knowing if the patterns they uncover are the result of the peculiarities of a particular data-set or whether they are more general. That is, many of the results in the literature likely result from overfitting, meaning they reflect noise in the data rather than meaningful relationships. If indicators of theoretical concepts fail to produce relationships with state repression that generalize to other sets of data, or do not add predictive validity to a model of state repression, this calls into question the importance of these concepts in influencing repressive behavior. While tests for statistical significance have value, evaluating the ability of a model to predict state violence out-of-sample offers at least an additional, and perhaps a better, way of assessing the veracity of explanations for its occurrence (See, e.g. **?**). That is, significance tests for coefficient(s) are certainly not the only option available, and they may not be the best. This is a point which has implications for empirical research in any area of political science which still uses statistical significance as the primary (or perhaps only) criterion for evaluating results.

This study remedies this deficiency in the literature through the use of cross-validation and random forests. Cross-validation is a well-developed and widely-accepted method for assessing the relative predictive performance of statistical models (See, e.g., **????**), though its use is relatively rare in political science.[3] The cross-validation analysis below assesses the ability of covariates which the literature identifies as important to increase the predictive power of a model of government repression beyond models that include minimal, baseline sets of "usual suspect" covariates. Random forests, which are ensembles of decision trees, are another useful technique for determining how much predictive power is gained by adding a particular covariate to a statistical model (**?**). Random forests allow us to examine the predictive power that each covariate adds to models that include

[2] Researchers often justify their decisions about which covariates to include by appealing to past work that indicates that those covariates are important predictors of repression, and this suggests a misunderstanding about the purpose of control variables in regression models where the goal is to recover an unbiased estimate of some parameter. If the goal is causal inference then control variables are there to prevent spurious correlations, and so should only include variables that are correlated with both state repression *and* the variable of interest.

[3] See **?**, **?**, **?**, and **?** for exceptions.

various combinations of our other covariates, rather than what each covariate adds to the baseline model alone. Random forests are attractive for our purposes because they detect non-linear and interactive relationships that do not have to be pre-specified by the analyst. We find that some, but relatively few, of the concepts identified by the literature as important determinants of state repression are able to substantially improve the fit of statistical models predicting state repression. This means that researchers examining government violence have been drawing conclusions about the accuracy of theoretical explanations that are not necessarily supported by the data.

To foreshadow the results, out of all the covariates considered, civil conflict is the best predictor of most indicators of state repression. Indicators of democracy also perform well in the analysis, though they predict some types of repression much more accurately than others, which has gone unnoticed in this literature. These two results are strong and support the literature's principal findings (See **?**, pp. 7-14), but their importance is tempered by measurement issues: the most commonly employed operational definition of repression overlaps to some extent with the operational definitions of democracy and civil war typically adopted in this literature. Thus indicators of civil war and democracy partly measure repression, a point which we elaborate further below.

We also find that indicators of some concepts which have received relatively little attention in the literature, including domestic legal institutions, demographic youth bulges, and state reliance on natural resource rents, perform relatively well. The excellent performance of several aspects of domestic legal systems is anticipated by the comparative institutions literature, particularly arguments about the ability of constitutions and courts to constrain government behavior generally (E.g., **??????**). That literature has been largely ignored by scholars who study repression,[4] and we argue that it deserves more attention in the future.

Finally, indicators of some types of government violence are predicted well by few of the covariates examined, which indicates that disaggregating measures of repression will be useful in future studies. We conclude by offering suggestions about how researchers can incorporate the insights of this study into future theoretical and empirical work on state repression.

A Brief Tour of The Literature

Cross-national, quantitative research on government repression, which began in earnest in the mid-eighties, was facilitated by the publication of annual, national reports on human rights conditions by Freedom House, Amnesty International (AI), and the US State department (USSD).[5] Early work used cross-sections of these data to test hypotheses about the impact of various concepts on repression. The most seminal work in the field is due to **?**, who presented the first analysis using data covering a relatively large time span and a relatively large number of countries. These data were coded from the annual reports of AI and the USSD and measure the practices of political imprisonment, torture, disappearance, and summary execution. **?** found that the coefficients associated with measures of democracy and GDP per capita were negative and statistically significant, and those associated with population size, the occurrence of international and civil wars, and lagged repression, were significant and positive. A measure of "leftist" regimes, too, was found to be positive and significant, though only using the data coded from State Department reports.[6]

[4] Though see **?** who draw on this literature and argue that effective judicial institutions discourage torture.

[5] Previous data collection efforts such as the World Handbook of Political and Social Indicators (**?**) also facilitated early research on state repression, but data coded coded from AI and USSD reports have become the most commonly used in the literature.

[6] **?** later updated these results using data covering an even larger time period, and additionally found statistically significant relationships between repression and 1) military regimes (positive), 2) former British colonial status (negative), and 3) leftist governments, though this time the relationship was negative for the latter measure. For an

With the exception of democracy, which is the primary focus of much work on repression, most of the covariates listed above were simply adopted as standard "control" variables, particularly population size, GDP per capita, and international and civil war.

The general theoretical framework for most of this research could be described as an informal, decision-theoretic approach that focuses on conditions which make repressive tactics costlier/more beneficial to political leaders.[7] For example, the positive relationship between violent (civil and international) conflicts and repression is usually interpreted to mean that leaders perceive repression to be more useful as real or perceived threats to their position in power increase, which is consistent with the idea that repression is a response to internal or external political challenges (See, e.g. **?????**). Indeed, empirical studies have so consistently found a relationship between dissent and repression[8] that this constitutes one of the literature's principal findings, and the reciprocal relationship between the two has become incorproated into more recent, formal, strategic models as an assumption (**?**).[9]

In line with this general theoretical framework, many take the negative relationship between democracy and repression to indicate that institutional constraints in democracies create a higher expected cost for using repression.[10] As noted above, the relationship between repression and democracy uncovered by early work (**??**) has been explored in-depth by a number of scholars, who have examined various topics such as how transitions to/from democracy affect repression (**?**), the functional form of the relationship between democracy and repression (**??**), and which aspects of democracy are most strongly related to repression (**???**). The negative relationship between democracy and repression represents the literature's other principal finding, but much research on democracy and repression is plagued by measurement problems. This is because governments that target political opposition with violence are less democratic by definition, given the way democracy is usually defined and operationalized in this literature.[11] The most commonly employed measure of democracy in studies of repression is the Polity index (**?**), which primarily measures the competition (or "opposition") dimension of democracy discussed by **?**, i.e. the extent to which the government tolerates competing policy preferences.[12] Since the definition of repression is the use of coercion against potential and actual opponents of the government, measures of repression will be related by construction to measures of democracy that include information about violence used to suppress political competition.[13] We discuss the implications of this problem in more detail below.

A recent and promising development is a body of work that examines the effects of various domestic *legal* institutions on state repression (**?????**). We view these studies as a promising development because a large amount of theoretical work in comparative politics suggests there should be a meaningful relationship between legal institutions and repression. In particular, the comparative institutions literature views constitutions and courts as instrumental in helping citizens overcome the coordination problem they face when attempting to resist government encroachment on basic

analysis of the differences between the Amnesty and State Department reports see **?**.

[7] See **?**. For an excellent example of this kind of approach see **?**.

[8] One line of research more closely examines this so-called "dissent-repression nexus." This research rarely uses the data based on Amnesty/State Department annual reports, but rather employs sub-national data collected at low levels of temporal aggregation, since it is interested in conflict dynamics that are not easily captured at the level of the country-year. See, e.g. **????????**.

[9] See **?** for another formal, strategic model of dissent and repression.

[10] For a review of arguments linking democracy to state repression see **?**.

[11] See **?**.

[12] See **?** for a discussion of the connection between Polity (and other commonly used measures of democracy) and Dahl's definition.

[13] Some recent research circumvents this problem by disaggregating democracy into its constituent parts, separating political competition and participation from constraints on policy change, for example. See, e.g., **???**. But most researchers adopt the "off-the-shelf" Poe and Tate model, which includes the Polity scale.

rights (??????). This suggests that constitutions and courts are useful for generating credible commitments on the part of the government to observe limits on its authority and refrain from encroaching on rights generally, including civil/political liberties as well as property rights. Arguments from the institutional literature on constitutions and courts have implications for empirical work on repression, though their connection to the repression literature is not widely appreciated.[14] Specifically, these arguments anticipate negative relationships between repression and 1) constitutional provisions which set explicit limits on government authority, and 2) judicial independence. Concerning empirical findings, ?, ?, and ? find negative relationships between repression and certain kinds of constitutional provisions, while ? find that common law legal systems are associated with less repression. ? and ? report a negative relationship between *de facto* judicial independence and state violence. The claim that domestic legal institutions are good predictors of repressive behavior is strongly supported by the results we present below, and we discuss the implications of these findings in the conclusion.

Two more recent studies have examined how other macro-level domestic factors influence the use of repression. One evaluates how state reliance on natural resource rents, rather than tax revenue, affects incentives for governments to use repression (?),[15] building on theoretical insights from the literatures on natural resource revenue and civil war, and natural resource revenue and democratization. The other study analyzes the relationship between so-called youth bulges and cross-national levels of state violence, arguing that governments in countries with large youth populations use repression in anticipation of high levels of dissent and conflict (?).[16] We think this is also a promising development since it suggests a focus on macro-economic and demographic factors beyond per capita wealth and population size.

Alongside research that examines how domestic conditions (primarily dissent and democracy) affect a government's use of repression, there has developed a large body of work examining the relationships between repression and a variety of international factors such as international human rights law and a state's position in the global economy. In general, the findings in this literature are much less consistent than those in research on domestic political behavior/institutions and repression (??), which indicates that these influences may be more tenuous. Much of this work also adopts an essentially decision-theoretic approach, arguing that various international influences affect the costs/benefits to political leaders for using repression. For example, one branch of this research focuses on the impact of international economic factors such as exposure to trade and foreign investment, pitting classic Marxist arguments about the role of international capital in degrading human rights practices against arguments that expect trade and investment to improve human rights practices by virtue of their beneficial effects on the domestic economy. Arguments in favor of a positive relationship between foreign investment and repression typically claim that influxes of foreign capital harm the domestic economy as a whole (though they benefit political elites), which creates dissent, thus repression becomes net beneficial because it maintains regime stability and encourages further investment (See, e.g. ?). In terms of empirical results, recent work on this topic has generally found a negative relationship between repression and openness to trade and investment (???).[17]

[14] Though see ?, who draw on arguments from this literature to argue for the relevance of judicial effectiveness for protection from torture. ? also argue that judicial independence helps reduce repression, and that common law systems help reduce repression, in part, because they promote judicial independence.

[15] The measure of resource rents comes from ?.

[16] They employ a measure from ?.

[17] ? provides an extensive discussion of this literature, and performs an extreme bounds analysis (?) to address this literature's inconsistent empirical findings. This is a valuable effort, but is motivated by different concerns than those motivating the analysis below. Hafner-Burton's study examines the sensitivity of statistical relationships to the inclusion of different groups of covariates. Her inferences are based on models fitted using all the available data and

Another international economic factor examined in this literature is participation in IMF and World Bank structural adjustment programs. Employing an argument similar to the one discussed above with respect to foreign investment and repression, **??** find a positive relationship between repression and participation in such programs. **?** focuses on human rights clauses in preferential trade agreements, arguing that explicitly tying human rights practices to trade policy makes repression costlier, and finds a negative relationship between such agreements and state repression.[18]

Another line of research examines the impact of global civil society broadly, and human rights NGO/INGO and Western media activity specifically, on human rights practices (**????**). The impact of NGO/INGO presence on repression has been found to be negative, while results concerning the effects of "naming and shaming" are more mixed.[19]

Beyond the international economy and global civil society, there is also a large literature on the effect of international legal agreements on human rights practices (**?????????**). Much research examining the impact of international law focuses on UN treaties and has found no relationship, or even a *positive* relationship, between treaty ratification and repression (**???**),[20] while other studies have found negative relationships conditional on domestic factors such as democratic political institutions (**??**),[21] strong domestic courts/rule of law (**?**), a large NGO presence (**?**), the expected tenure of political leaders (**?**), and legal standards of proof for particular rights violations (**?**).

For our purposes we do not need to exhaustively review all of the theoretical arguments presented in the studies cited above. Our goal is to evaluate the empirical implications of existing theoretical arguments using predictive validity as a criterion for inference. We are not interested in prediction for the sake of prediction, but rather for empirically sorting through the many hypotheses advanced in this literature. Our goal is essentially the same as **?**: to determine which of the many posited causes of repression receive the strongest support in the available data. To make this determination we examine whether 1) the statistical relationships discovered by this broad literature are generalizable beyond the particular data-sets which produced these relationships, and 2) indicators of the concepts identified as important determinants of state repression improve the predictive power of statistical models of state repression. In the next section we discuss the methods and data used to accomplish these goals.

Evaluating Models of State Repression

As discussed above, the standard criterion for assessing the veracity of a potential explanation for state repression is a null hypothesis significance test for one or more covariates which measure theoretically relevant concepts. The shortcomings of this criterion for social science research are well documented (See, e.g. **?**), and we do not discuss all of them here. Our concern is that the use of this criterion alone has hindered the development of generalizable and accurate explanations for

are drawn on the basis of (a large number of) null hypothesis significance tests. Thus her analysis does not guard against over-fitting or provide any information about the predictive power of the included covariates.

[18] Though see **?**, who use matching techniques prior to their regression analysis and find no relationship between PTAs and repression.

[19] **?** finds positive/null relationships between repression and NGO shaming measures while **?** and **?** find a negative relationship. The differences are due to different measures, different samples, and the fact that both **?** and **?** interact NGO activity with other covariates.

[20] See **?** and **?** for explanations for this finding. See also **?**, who argues that the positive relationship is an artifact of changes over time in the way information about state repression has been evaluated and finds that there is a small, negative correlation between signing the Convention Against Torture (CAT) and violations of personal integrity rights.

[21] **?** finds a negative relationship between ratification and repression among democracies, while **?** finds a negative relationship in transitioning/weakly democratic regimes.

repression. For one, strict adherence to null hypothesis significance tests alone ignores the ability of a model to predict instances of state repression. This means that a variable which is a "statistically significant" predictor of repression may not actually improve our ability to correctly classify governments as more or less repressive. Recent work on civil war has shown that statistical significance and predictive validity can actually be at odds with one another, i.e. covariates with statistically significant coefficients can actually impair a model's predictive performance (**?**). This means that attention to statistical significance alone is misleading researchers about what are, and what are not, important determinants of state repression.[22] Rather than evaluating statistical significance alone, researchers should evaluate the *fit* of their model to the data. If a theoretically informed statistical model of state repression is offered as evidence that one has discovered an important cause of repression, then the model should be able to produce reasonably accurate predictions, i.e. predicted values that closely match observed values. Until now this literature has given little attention to predictive validity,[23] so it is not even obvious what "reasonably accurate" means. We provide some indication of which theoretically motivated variables add the most accuracy to models using the most commonly analyzed data on state repression. By doing this we hope to establish a baseline for future work, and provide a better way to adjudicate between existing theoretical explanations for repression.

Second, since scholars typically use all of the data available to estimate their models, there is a significant danger of over-fitting. This means that researchers may be discovering a relationship that is the result of the unusual features of a particular data-set rather than a meaningful, generalizable relationship between repression and a concept of interest. It has been demonstrated elsewhere that selecting sets of covariates based on p-values can result in models with significant (at the 0.05 level) coefficients for variables whose relationship with some response variable is truly random (**?**). This is potentially a serious problem for cross-national research on state repression since the purpose of this literature is to uncover *general* empirical regularities between repression and concepts of interest. Examining the fit of a model does not necessarily circumvent this problem, because any model will almost certainly provide a better fit to the data used for its estimation than any other set of data (See, e.g. **?**). This is why some have proposed the use of *out-of-sample* fit as a heuristic for evaluating model performance in conflict studies (**??**). Such an analysis avoids drawing conclusions based on idiosyncratic results: if a model has not produced a generalizable result, then it will produce poor predictions in a set of data which was not used for its estimation.

As a final point, models of state repression with predictive validity will be of much more interest to policymakers than models with statistically significant coefficients. If covariates with significant coefficients do not provide any leverage in predicting when and where government violence will occur, then they will not be of much value for making policy decisions. These are important points that have been largely ignored in the quantitative literature on state repression. Cross-validation techniques and random forests, which we discuss below, can address these omissions.

Cross-Validation

The purpose of cross-validation is to examine the out-of-sample predictive power of a statistical model. The cross-validation procedure we use below proceeds as follows: the analyst divides the data into k subsets, estimates a model using $k-1$ of the subsets (the "training" set), uses these estimates to generate predictions for the remaining subset (the "test" set), and calculates some

[22] This problem is likely exacerbated by the common practic of treating dependent observations as if they were independent, which increases statistical power and thus the model's ability to detect *small* effects for variables which may not be important causes or correlates of state repression.

[23] See **?** for a notable exception.

measure of prediction error in the test set. The data are "rotated" k times so that each of the k folds is eventually used as the test set, and the prediction error is summarized based on all test sets. This is often called k-fold cross-validation. Typically the data are divided up a number of times in this fashion to ensure that results are not dependent on dividing the data in a particular way. For the analysis below we perform 10-fold cross-validation.[24] We randomly divide the data into 10 folds, estimate the model, and calculate the prediction error across all folds 1,000 times for each model. Resampling this many times allows us to approximate the uncertainty around the median prediction error for each model, which is useful for comparing performance across models. We describe the statistics used to evaluate predictive performance below.

Random Forests

We also estimate a set of random forests to assess each covariate's predictive power. Random forests, and their constituent decision trees, are a class of supervised machine learning algorithms that are commonly used for prediction as well as assessing which variables are the most important predictors of the outcome of interest (**?**).[25] There are several advantages to this nonparametric approach. For one, it allows us to consider the predictive power of all the covariates, rather than comparing the fit of a model with a particular covariate in addition to a base model, as in the cross-validation analysis. Random forests also allow for non-linear functional forms and complex interactions among the covariates, without the analyst having to pre-specify a particular functional form or interaction term.[26] Decision trees, or base learners, the constituent parts of a random forest, find an optimal partition of the covariate space (the space spanned by all of the predictor variables) through recursive partitioning, or "growing" the tree. In brief, the recursive partitioning algorithm we use (a single decision tree in an ensemble) works by:[27]

1. selecting a set of observations (by subsampling from the full set of data)

2. selecting a subset of covariates

3. finding the variable in the selected subset that is most strongly related to the dependent variable

4. finding the point in the selected variable that optimally classifies the dependent variable

5. repeating steps two through five on the resulting partitions (daughter nodes) until a stopping criteria is met

For a random forest, this process is repeated a large number of times, resulting in a forest of decision trees. Each tree is grown with a randomly sampled set of data taken from the full set

[24] In practice the choice of k does not seem to be very consequential, and $k = 10$ is fairly standard in the machine learning literature (See, e.g. **?**).

[25] Random forests are necessary because decision trees are high variance estimators. Using an ensemble of decision trees decreases the variance of the fitted values (**??**). Typically bagging, or bootstrapped aggregating, is used to decorrelate the predictions made by each tree in the forest by sampling observations with replacement. We instead use subsample aggregating, which has been shown to work better under weaker conditions (**??**). Random forests add, in addition to the resampling of cases, a random selection of predictors at each splitting node.

[26] Random forests are also equipped to accommodate missing data via surrogate splits (**?**). Surrogate splits proceed by ignoring the missing values, finding the variable most strongly related to the dependent variable within the node, finding an optimal split in the selected variable, and then searching for a non-missing variable that results in a similar split. Thus, using random forests is also a check on our imputation model.

[27] The second step is specific to decision trees in a random forest. If the algorithm were not used as a part of an ensemble there would be no random selection of predictors.

of data and each node may have different predictors randomly selected. The predicted value for an observation is the most commonly predicted value for that observation across all the terminal nodes (the node at which the stopping criteria is met) in each decision tree in the forest. A non-linear relationship between a particular covariate and the outcome can be detected because the partitioning algorithm can make multiple splits on the same variable within a single decision tree in addition to making different splits (i.e. at different points in the variable) across trees in the forest (See **?**)). The detection of interactions between covariates works similarly. A more in depth explanation of decision trees and random forests can be found in **?**.

As an example consider a model of political imprisonment as measured by the political imprisonment component of the CIRI scale. The political imprisonment variable is an ordered categorical variable that ranges from 0 to 2, with higher values indicating less political imprisonment. Suppose we wished to predict what level of political imprisonment would be observed in a particular set of country-years using the measure of civil war from the UCDP/PRIO data (**?**) and the measure of youth bulges from **?** used by **?**. These are thought to capture incentives to repress based on current, and prospective, levels of dissident activity. Figure **??** shows the result from a decision tree using a set of 500 randomly selected observations (the first step in the algorithm described above) along with the two aforementioned variables.[28] At the first node the youth bulges variable is selected because, at least in the 500 observations that were randomly selected, it was most strongly related to the CIRI measure of political imprisonment. After the youth bulges variable is selected, an optimal partition of the variable is found, whereby the dependent variable is best classified. This results in two more partitions, on each side of the split. Each of these daughter nodes is then partitioned further using the civil war variable, which is most strongly related to the dependent variable at nodes 2 and 7. Node 2 undergoes one more split using the youth bulges variable, resulting in a set of terminal nodes (the row along the bottom of Figure **??**). At these nodes the stopping criteria is reached: the increase in classification performance from further partitioning is low at this point. The terminal nodes are used to classify observations by using the most common class in each node. If this were a random forest instead of a decision tree (i.e. an ensemble of decision trees), the variables included in each node in the tree (and thus available for selection) would be randomly selected, and the predicted class of each observation would be the most commonly predicted class across the predictions made by each tree.[29]

There are a variety of implementations of random forests, some of which have more or less desirable statistical properties (**??**). We utilize the unbiased decision tree algorithm (referred to as a conditional inference tree) developed by **?**. These trees first test the global null hypothesis of no relation between the covariates in the partition (a particular node) P, X_j^P, (the variables in that node) where j indexes each covariate, and the dependent variable. If this global null hypothesis can be rejected at a pre-specified level of confidence α, then the covariate with the smallest p-value is selected, and an optimal split in the selected covariate is found. If the null hypothesis cannot be rejected, then partitioning stops. This stopping criteria avoids the established bias towards variables with many possible partitions that occurs in many other random forest implementations, allowing for unbiased variable selection (**??**). It also prevents overfitting since it does not optimize global information, as is common in other decision tree implementations.

[28] In this case the variables were not randomly selected at each node. We selected them because they are both strongly related to political imprisonment and because one is binary and the other numeric.

[29] In the case where the response variable is continuous the prediction for observations in a given terminal node would be the mean of the response for all observations in said node and the forest prediction would be the mean of the tree predictions for each observation.

tree.png

Figure 1: The results of using a decision tree to predict the level of political imprisonment using 500 randomly sampled observations and two covariates, civil war and youth bulges. The number of observations at each node (or partition) is indicated next to the node's number, and the bar plots indicate the distribution of the values of the dependent variable at the node. At node 1 (the parent node) the youth bulges variable is most strongly related to political imprisonment and is selected. The optimal split in youth bulges is 29.6, resulting in two daughter nodes, wherein the process repeats. At each daughter node (2 and 7) civil war is the most strongly related to political imprisonment and is selected, resulting in node 3, where youth bulges is again selected (a new split is found at 17.6), and then the terminal nodes (nodes 4, 5, 6, 8, and 9), which are used to predict the dependent variable (the most common category in each terminal node is the predicted value of all observations in that node). Note how the variance of the distribution of dependent variable decreases at each node.

Data and Model Evaluation

Most of the empirical research on repression uses either the indicator used by **?**, known as the "Political Terror Scale" (PTS) (**?**), or an indicator known as the "Physical Integrity Index" from the Cingranelli-Richards (CIRI) human rights project (**?**). While there are some differences between the two,[30] both of these are ordinal indicators coded from annual AI and USSD reprorts, and both measure instances of political imprisonment, torture, kidnapping, and summary executions. The most important difference between these two indicators is that the CIRI physical integrity rights index can be disaggregated into components that measure each of these abusive practices separately. Though disaggregation of the CIRI index is possible, it is not common practice; few studies theorize about the use of any of the four specific practices measured by CIRI, and those that do typically focus on torture (See, e.g. **?????**).[31] The analysis below employs each of the CIRI components in addition to the PTS and the aggregated CIRI index.[32] This allows us to evaluate whether theoretically informed covariates are better at predicting some repressive practices than others. Our results reveal important differences between the individual components, a point to which we return below.[33] In addition to CIRI and PTS we employ a new measure from **?**, which is created using a Bayesian measurement model[34] and incorporates the indicators mentioned above in addition to data from many other sources.[35]

For models using the PTS, the aggregate CIRI index, and the variable created by **?** we estimate linear models, fit using ordinary least squares, which is common practice in the literature. We estimate ordinal logit models for the CIRI component scales, and an additional ordinal model for the PTS.[36] For the linear models, root mean squared error provides a straightforward way of assessing predictive performance.[37] For ordinal variables such as the CIRI components the choice of a fit statistic is not as obvious. We use Somer's D, a rank correlation coefficient (**?**), as our discrepancy statistic for the ordinal logit models. Somer's D is closely related to Goodman and Kruskal's γ and Kendall's τ, differing only in the denominator.[38] Somer's D makes a distinction between the independent and dependent variable in a bivariate distribution, correcting for ties

[30] See **?** and **?**.

[31] There is a sociological literature on state repression, informed mainly by resource mobilization theory (See, e.g. **?**), that does theorize about/examine variation in the repressive tactics used by governments in response to dissent (See, e.g. **???**), but the typologies presented by these authors are distinct from the categorization of repressive tactics used by PTS/CIRI.

[32] Note that PTS uses higher values to indicate *more abuse* of personal integrity, while CIRI uses higher values to indicate *more respect* for personal integrity.

[33] See **?**, who perform a Mokken Scale analysis using an early version of the CIRI data. Their analysis suggests that the CIRI components measure a unidimensional construct, and that summing the components does not introduce too much measurement error, i.e. the sum of the components is nondecreasing in the latent construct measured by the scale. We do not challenge the conclusions of their analysis, but rather suggest that the components themselves may not be identically related to indicators of various determinants of repression.

[34] The measurement model used by Fariss accounts for changing standards of accountability in human rights reports, and produces a measure of repression which indicates that state practices have, on average, improved over time. Since this contrasts with CIRI and PTS it will be useful to compare results using the three different scales. The model by **?** builds on a similar latent variable model by **?**.

[35] The other measures used in this model include two indicators of torture from **?** and **?**, a binary measure of measure of government one-sided killings adapted from **?**, measures of genocide/politicide from **???**, and a binary measure of political executions adapted from **?**.

[36] Cross-validation results for the ordinal models using the PTS can be found in the appendix.

[37] Root mean squared error is $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{Y}_i - Y_i)^2}$

[38] Somer's D is similar to the commonly used τ_b, which is equal to $\frac{P-Q}{(P+Q+X_0)(P+Q+Y_0)}$, where Y_0 is the number of ties in Y, and γ, which is equal to $\frac{P-Q}{P+Q}$.

within the independent variable. With Y being treated as the independent variable it is denoted D_{xy}. Specifically:

$$D_{xy} = \frac{P - Q}{P + Q + X_0}$$

where P is the number of concordant pairs, Q is the number discordant pairs, and X_0 is the number of ties in X. This is simply a measure of association for ordinal variables, so our approach is essentially to calculate the correlation between predicted and observed values. Like all correlation coefficients, the D statistic lies in the interval $[-1, 1]$, with values closer to 1 indicating more rank agreement and values closer to 1 indicated less rank agreement, so values closer to 0 indicate more prediction error. In the results section below we discuss how we use these performance measures to judge whether covariates add substantially to a model's predictive ability.

For the random forests, variable importance is assessed using an unscaled permutation test which measures the mean decrease in classification performance (% of cases classified correctly) after permuting each element of the set of predictors X_j, where j indexes each covariate, over all trees in the forest. Permuting important variables will result in a systematic decrease in classification accuracy, whereas permuting unimportant variables will result in a random decrease, or no decrease, in classification accuracy. The variable importance scores do not measure the importance of the variable conditional on the importance of the other predictors (they measure marginal importance), thus scores can be confounded by correlations between predictors. Although it is possible in principle to conduct a conditional permutation test, such a test is computationally infeasible given the large number of predictors in this study. A correlation matrix of all the predictors used in this study is available in the online appendix. Although there are some highly correlated pairs, the covariates are not so highly correlated as to make this comparison uninformative. Notable are the correlations between Polity and its components, as well as those between the media coverage covariates from **?**. Youth bulges are negatively correlated with Polity and its components, and the INGO measure is positively correlated with Polity. To deal with the possible inflation of the importance scores of possibly unimportant covariates we set the number of variables selected at each node to 10 (the default is 5) and increase the total number of trees in the forest. Additionally, we bootstrap the permutation importance scores by taking samples from the full set of data (with replacement and of the same size as the full data), re-fitting the random forest, and re-calculating the permutation importance scores 100 times. In our discussion of the results we present summary statistics of the bootstrapped sampling distribution for the permutation importance scores. Additionally, we estimate the concordance of ranked permutation importance across different values of the aforementioned tuning parameters in the online appendix.

Our explanatory variables are drawn from the literature. We use indicators of concepts that are "usual suspect" covariates (i.e., standard control variables) as well as indicators for concepts whose relationships with repression are less well-established. Table **??** lists the measures used below and the sources from which they were obtained. Full descriptions of these data can be found in the appendix. In our cross-validation analysis we assess the increase in predictive validity which results from adding each variable to three different baseline models: one that includes only (the natural logs of) GDP per capita and population size, another that includes both of these variables and an indicator of civil war from the UCDP/PRIO armed conflict data-set (**?**),[39] and another that includes GDP per capita, population, and a lagged dependent variable. We employ the last of these specifications because, partly as a result of a significant coefficient for a lagged dependent variable

[39] We employ the measure of "high-intensity" conflict, i.e. conflict producing ≥ 1000 annual battle-related deaths, as this measure performs much better in cross-validation than the "low-intensity" measure, which uses a death threshold of 25.

in **?**, it has become standard to include a lagged dependent variable in models of repression. There is also a theoretical argument which suggests that bureaucratic inertia and elite habituation to the use of violence creates strong patterns of temporal dependence in state repression (E.g., **??**). To save space we present the cross-validation results from our third baseline specification in the appendix, but these largely confirm our findings from the first two baseline models.

Table 1: Measures and Sources

Measure	Source
Demographics	
Population Size	Gleditsch (2002)
Youth Population	Urdal (2006)
Macroeconomic Factors	
GDP per capita	Gleditsch (2002)
Oil Revenue	Ross (2006)
Violent Conflict	
Civil War	UCDP/PRIO armed conflict
Interstate War	UCDP/PRIO armed conflict
Political Institutions	
Democracy	Polity IV
Military Regime	Database of Political Institutions
Left/Right Regime	Database of Political Institutions
Domestic Legal Institutions	
de facto Judicial Independence	CIRI
Constitutional Provisions	Keith, Tate, and Poe (2009)
Common Law System	Mitchell, Ring, and Spellman (2013)
International Economic Factors	
Trade Openness	World Bank
Foreign Direct Investment	World Bank
Structural Adjustment (WB and IMF)	Abouharb and Cingranelli (2007)
PTA Agreement w/ Human Rights Clause	Spilker and Bohmelt (2012)
Civil Society/INGOs	
INGO Presence	Hafner-Burton and Tsutsui (2005)
INGO Shaming	Ron, Ramos, and Rodgers (2005)
Western Media Shaming	Ron, Ramos, and Rodgers (2005)
HRO Shaming	Murdie and Davis (2012)
International Law	
ICCPR Ratification	UN website via `untreaties`
CAT Ratification	UN website via `untreaties`

Several points about the variables used in the analysis are worth mentioning. First, for several of the concepts listed in Table **??** we use multiple measures. These include our measure of democracy, the Polity IV scale (**?**), for which we employ both the commonly used "democracy minus autocracy" scale, as well as each of the democracy scale components. One study analyzing the Polity data found that the aggregated scale primarily reflects the executive constraints subcomponent (?),[40] and studies of repression have found that the competition subcomponent of Polity is more strongly related to measures of repression than the other subcomponents (??). We obtain a similar result, which we discuss in more detail below.

We also employ multiple measures of INGO shaming. Three of these come from ? and were employed by ?. These are counts of the number of AI press releases and background reports issued about a particular country during a given year. These variables are all lagged by one year. Our other measure of INGO shaming comes from ?. This measure is based on events data and is a count of the annual number of conflictual actions sent by human rights organizations (beyond AI alone) towards a particular government. The constitutional protection data from ? also includes multiple measures, all found to be statistically significant in regressions using the PTS: provisions for a fair trial, provisions for a public trial, provisions stating that the decisions of high/constitutional courts are final, and provisions which require legislative approval to suspend constitutional liberties.

Second, as mentioned above, past results for some of the indicators in Table ?? are slightly mixed. These are the measures of shaming by INGO/HROs and Western media, and measures of ratification for two core UN human rights conventions: the International Covenant on Civil and Political Rights (ICCPR), and the Convention Against Torture (CAT). ? finds that shaming by Amnesty International (AI) is actually *positively* associated with repression, but shaming by Western media bears no relationship to repression. ? find a negative relationship between NGO shaming and repression conditional on NGO presence and shaming by other actors (such as governments), and ? finds a similar relationship conditional on dependency on foreign aid and investment. Results using human rights treaties data have also been mixed, and we employ these indicators because of unexpected, statistically significant findings in the literature, and because we believe the volume of recent work on this topic justifies the inclusion of human rights treaty ratification.

Finally, many of the variables we use have substantial missingness. First, we restrict our analysis to the period 1981-1999, which is well covered by most of the variables we consider. Since the assumption that these data are missing at random is implausible, we use model-based imputation of the missing values prior to cross-validation.[41] We perform five imputations of the missing values, cross-validate our models on each imputed data-set, combine our discrepancy statistics across them, and then compute summary statistics. We now turn to the results from our analysis.

Results

For our cross-validation analysis, we adopted the following rule to determine whether a covariate is an important predictor of state repression: if the lower bound (the .025 quantile) of the prediction error for the model including that covariate is above the upper bound (the .975 quantile) of the prediction error for the baseline model, then the covariate is marginally important.[42] This is a rather strict rule, but it is justified since we are evaluating the performance of models which include the covariate in question against models that are stripped-down relative to those common in the literature. In the interest of space we limit most of our discussion to the handful of variables

[40] Though see ?, which suggests that more recent version of the democracy scale is driven by the competition component as much as it is the executive constraints component.

[41] The technical details of the imputation model can be found in the online appendix.

[42] Since higher values of Somer's D indicate more predictive power, this is the rule for the ordered logit models. For RMSE lower values indicate better predictions, so the rule is reversed, i.e. the upper bound of the model which includes the covariate in question should be below the lower bound for the baseline model. The importance is marginal because the increase in predictive power is only conditional on the covariates in the base specification.

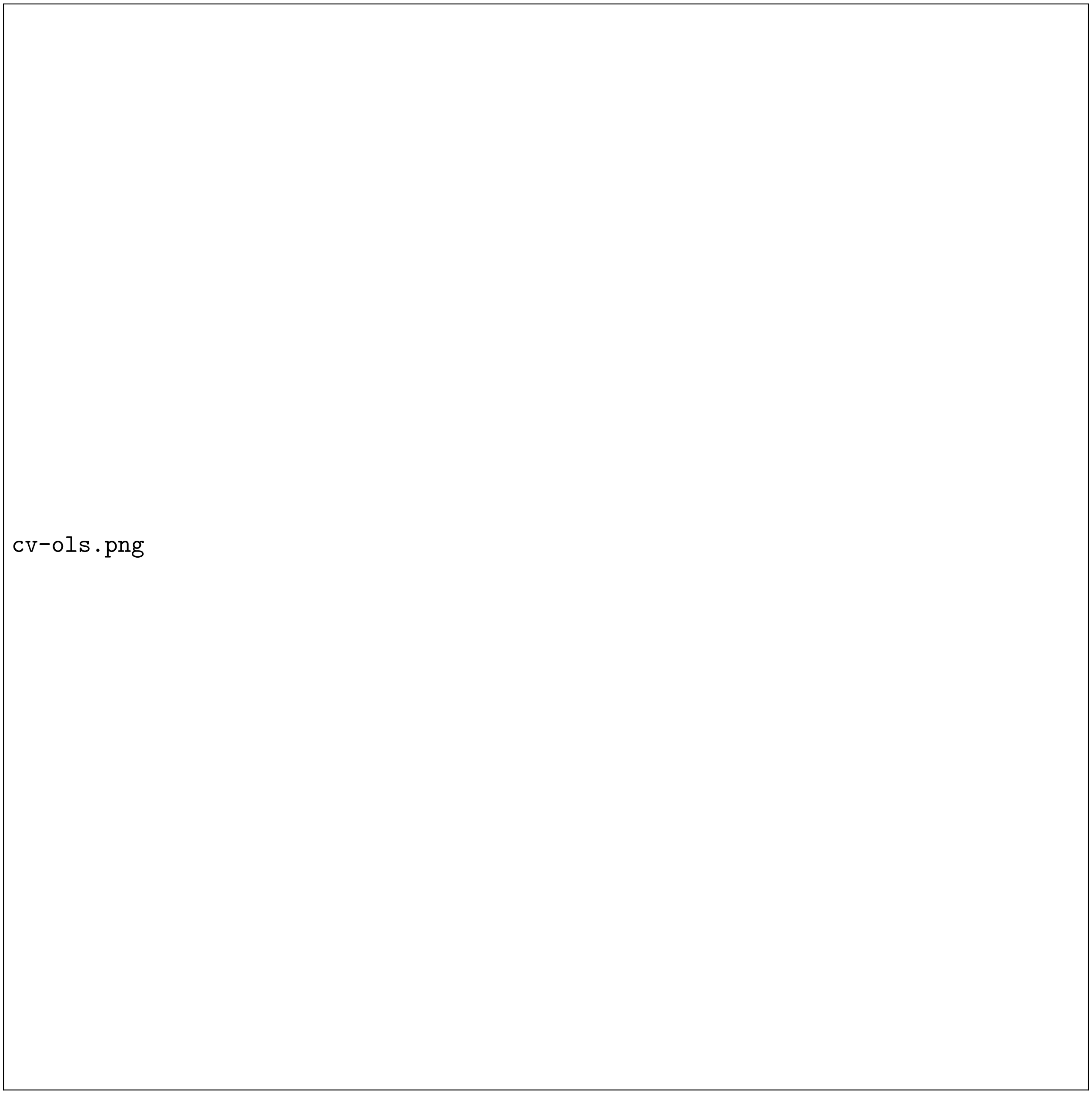

Figure 2: Results from 10-fold cross-validation with 1000 resampling iterations for linear (OLS) models of repression using (the natural logs of) GDP per capita and population. The x-axis shows root mean squared error (RMSE). The y-axis represents model specifications which are composed of a base model, which is indicated by the gray band, and the variable indicated on the y-axis. The dots show the median of the sampling distribution of the RMSE statistic, along with the .025 and .975 quantiles. The dotted line shows the .025 quantile of the sampling distribution of RMSE for the base model. Model specifications whose intervals overlap with this line do not add significantly to the fit of the model compared to the base specification.

cv-lrm.png

Figure 3: Results from 10-fold cross-validation with 1000 resampling iterations for ordinal logistic regression models of state repression using (the natural logs of) GDP per capita and population. The x-axis shows Somer's D_{xy}, a rank correlation coefficient that ranges from -1 to 1. The y-axis represents model specifications which are composed of a base model, which is indicated by the gray band, and the variable indicated on the y-axis. The dots show the median of the sampling distribution of the Somer's D_{xy} statistic, along with the .025 and .975 quantiles. The dotted line shows the .975 quantile of the sampling distribution of the D_{xy} statistic for the base model. Model specifications whose intervals overlap with this line do not add significantly to the fit of the model compared to the base specification.

that add the most predictive power *and* perform well across most of the models.[43] Figures **??**-**??** display the median prediction error (shown as dots) as well as the .975 and .025 quantiles of the sampling distribution of the error statistic (shown as lines) for each model we estimate.[44] In each figure a dashed line is placed at the .975 quantile (or .025 quantile, depending on the discrepancy statistic) of the error for the baseline model. The gray, horizontal bands in each figure highlight the baseline models.

The first two figures (**??** and **??**) show the consequences of adding different covariates to a model that includes only the natural logs of GDP per capita and population. A passing glance at these figures immediately conveys that civil conflict, for most measures of repression, adds much more predictive power to this baseline model than any other covariate examined here. This is consistent with one of the literature's two principal findings, a phenomenon that has been labeled the "law of coercive response" (**?**). The fact that this relationship is labeled a "law" gives some indication of its regularity. Though this result is strong, we would point out that indicators of civil war overlap empirically with indicators of state repression: indicators of repression include information about non-combatant casualties during violent conflicts, and these casualties also contribute to a conflict reaching the death threshold necessary to classify it as a civil war. We return to this result in the discussion section below.

Clearly civil war predicts repression better than nearly all of the other covariates, but there are exceptions to this pattern. For the political imprisonment component of the CIRI physical integrity index, civil war is outperformed by the aggregated Polity scale, the CIRI judicial independence measure, and three of the components of the Polity democracy scale, most notably the political competition component. This latter result is consistent with previous studies (**??**), though no study we are aware of has noted that the Polity measure of democracy predicts political imprisonment more accurately than it does other kinds of government violence. While the performance of Polity and its democracy components is impressive, the ability of the aggregated scale, and the political competition component, to predict political imprisonment is driven by the problem noted above: the way Polity defines and operationalizes democracy makes any relationship between Polity and political imprisonment tautological, i.e. governments who engage in political imprisonment must be considered less democratic given the operational definition. Political imprisonment is the only component of CIRI that considers *only* violence directed at political opposition,[45] so it is necessarily related to the component of Polity that measures restrictions on political competition, and the aggregate Polity scale (which includes this component). And of course, since both the aggregate CIRI index and PTS include information about political imprisonment, Polity and the competition component are necessarily related to these measures as well. This is less of a problem with the other Polity components, especially the executive constraints component. We return to this point below after discussing our other results.

The other measure of government violence for which civil war adds less predictive power than other covariates is the torture component of the CIRI scale. For this indicator the CIRI measure of judicial independence and a measure of youth population from **?** employed by **?**[46] both add more predictive power to the baseline model than civil conflict. This is notable because the concepts

[43] Since we combine our plots we cannot preserve a best-to-worst ordering of the covariates, which makes it harder to see which add more predictive power than others for a single dependent variable. However, it makes it easier to compare variable performance across dependent variables.

[44] Recall that the data were randomly divided into 10 folds 1000 times for each model.

[45] While the CIRI components measuring torture, disappearance, and summary execution also measure violence against political opposition, they also much more likely than the political imprisonment scale to include non-political violence against criminals and marginalized members of society, such as migrant workers and the homeless. For a discussion/analysis of this kind of government violence and how it relates to democracy see **?**.

[46] This indicator measures the proportion of the adult population (older than 15) that is younger than 25.

measured by the these covariates have received relatively little attention in the literature. The results for judicial independence lends plausibility to the theoretical connection between strong courts/legal systems and violations of basic rights (E.g., **?**), while the result for youth bulges lends credence to the theory advanced in **?** that leaders apply repression in anticipation of dissent/conflict.

cv-ols-cwar.png

Figure 4: Cross-validation results from linear models (OLS) of the same form as Figure **??**, but with a base specification which consists of (the natural logs of) GDP per capita and population as well as civil war.

Figures **??** and **??** display results from cross-validation analyses in which the baseline model now includes civil war in addition to the natural logs of GDP per capita and population. Using the latent variable constructed by **?**, the most predictive power is added by the measure of political competition from Polity, followed by youth bulges and the CIRI judicial independence indicator.

cv-lrm-cwar.png

Figure 5: Cross-validation results from ordinal logistic regressions of the same form as Figure **??**, but with a base specification which consists of (the natural logs of) GDP per capita and population as well as civil war.

Common law, oil rents, constitutional provisions for fair trials, and PTAs with human rights clauses also do well. For the aggregated CIRI scale, the CIRI measure of *de facto* judicial independence adds the most predictive power to the baseline model, followed by the measure of youth bulges and Polity's executive constraints and competition scales. Fair trial provisions, oil rents, and common law legal systems, concepts which have only recently received attention in the literature, also perform very well in this model. For the linear model using the PTS,[47] youth bulges, political competition, judicial independence, common law legal systems, executive constraints, and fair trial provisions perform best. The impressive performances of constitutional provisions for fair trials and common law legal heritage, in addition to judicial independence, justify a stronger focus on legal institutions in future studies of repression, and lend further support to theoretical insights from the comparative institutions literature. This is another point to which we return below.

Turning to results from the individual CIRI components, the political imprisonment models are largely consistent with those above: the Polity scale and its democracy subcomponents perform best, particularly the competition scale. Judicial independence, fair trial provisions, and oil rents also add a substantial amount of predictive power. The results pertaining to torture are also consistent with the analysis above, with judicial independence and youth bulges adding the most predictive power to the baseline model.

The other two CIRI components, which measure disappearance and extrajudicial killing, behave much differently than the other indicators used in this study. Precious few of the covariates included in the analysis add much predictive power to the baseline model. Strict adherence to the decision rule mentioned above leads to the conclusion that only judicial independence, oil rents, and youth bulges increase the predictive power of the second baseline model for disappearance. For extrajudicial killing, only youth bulges, judicial independence, and one of the (lagged) NGO shaming indicators (Western media reports) improve the performance of the baseline model. The relatively poor performance of the covariates in these models is partly due to the fact that killings and disappearances simply occur with less frequency than political imprisonment and torture.[48] But the analysis does suggest that there are important differences between the separate components of the physical integrity scale that are ignored when one uses the aggregate scale. One potentially important difference between these two components and the political imprisonment/torture scales is that disappearances and extrajudicial killings occur, by definition, outside of the legal system. This is perhaps another reason why, with the exception of judicial independence, the features of legal systems that do well in predicting political imprisonment and torture do not add much predictive power to models of disappearances and killings. More fundamentally, this suggests that theories of repression need greater refinement. We return to this point also in our conclusion.

Figures displaying results for our third baseline specification, which includes a lagged dependent variable in addition to GDP per capita and population, can be found in the appendix. These results are largely consistent with the results presented from the first two specifications. Not surprisingly, the third baseline model provides a better fit than the first two: inclusion of a lagged dependent variable markedly improves predictive validity, which lends support to theories which suggest that governments can become habituated to the use of violence to resolve political conflict (**?**). This improvement dampens the predictive power that other covariates add to the model, with the result being that even fewer covariates perform well in the cross-validation analysis. Civil war still improves the fit of all the models. For linear models using the aggregate scales judicial independence, youth bulges, and Polity's political competition scale still do well. For the disappearance and killing components of the CIRI scale only civil war improves out-of-sample fit. The results for the political

[47] Results for the ordinal logit models using PTS can be found in the appendix.

[48] See **?**.

imprisonment scale are consistent with the other two baseline models: Polity and its components do very well, as do the measures of judicial independence, oil rents, and fair trial provisions. For the CIRI torture scale judicial independence and youth bulges also improve model fit.

10 of the 31 covariates included in the analysis failed to add predictive power to *any* of the baseline models. These are: military regime, British colonial status, two of the three variables measuring participation in IMF and World Bank structural adjustment programs (World Bank structural adjustment program participation measured alone marginally improves the fit of some models), constitutional provisions stating that high court decisions are final, constitutional provisions giving the legislature authority over declaration of states of emergency, the measure of HRO shaming used by **?**, foreign direct investment, ratification of the ICCPR, and international war. That British colonial status and international war fail to improve the fit of our baseline models is surprising given that these are often included as standard control variables. This underscores that statistical significance is not the best criterion to use for variable selection. A number of variables only marginally improve model fit, despite being technically important according to our decision rule.

For the most part features of domestic politics, rather than international politics, are adding the most explanatory power to these models. This is consistent with previous empirical findings in the literature: analyses of international determinants of human rights practices such as international economic standing (**?**), international law (**????????**), and NGO shaming (**???**) tend to produce inconsistent results. The contrast between results for domestic/international factors suggests that the institutional (political and legal) constraints that exist at the domestic level are more important for the decision to repress than are any international constraints arising from treaties, NGO activity, or a state's situation in the global economy. However, as we discuss below, the relationships between international political factors and repression may be more complex than the cross-validation analysis allows for. But this analysis suggests that international political factors are, in general, not as useful for predicting instances of repression as domestic factors.

We next turn to the permutation importance measures from the random forests, which are depicted in Figures **??** and **??**. These figures show each covariate's importance score from the permutation test described above. For the most part the results of this analysis echo those from the cross-validation: across most dependent variables, civil war, youth bulges, and judicial independence remain among the most important predictors. As above, Polity and its various components, particularly the competition component, do extremely well in the political imprisonment model. Constitutional provisions for fair trials and common law legal systems do not do as well in this analysis, though fair trial provisions is an important predictor of political imprisonment. The most notable contrasts with the cross-validation results discussed above are the performances of trade openness and INGO presence. While neither of these performed especially well in the cross-validation analysis,[49] they score relatively high on our variable importance measure; trade openness is judged to be among the most important predictors of the CIRI killing scale, the CIRI torture scale, the dynamic latent variable from **?**, the aggregated CIRI scale, and the PTS, which gives further credence to theories which posit some relationship between general economic openness and repression. INGOs does well relative to other variables for both the CIRI torture scale and the PTS. This suggests that these variables may have an interactive or non-linear relationship with measures of repression, which suggests that international factors broadly may be related to state violence in complex ways. The impact of trade openness and INGO presence may be conditional on other variables, for instance.

[49] INGO presence improves the fit of models using the dynamic latent score, the aggregate CIRI scale, and political imprisonment. Trade openness only does well in the linear PTS model that does not include civil war.

imp-ciri.png

Figure 6: The marginal permutation importance of independent variables estimated using random forests, with the CIRI components as dependent variables. Each panel pertains to a random forest model of the dependent variable indicated by the gray bar located at the top of the panel. Each dot represents the median of the bootstrapped sampling distribution of the mean decrease in classification performance that results from randomly permuting the variable indicated in the y-axis across all decision trees in the random forest. If the variable is truly important, permuting its values should systematically *decrease* performance, whereas a truly unimportant variable should produce no decrease, or a random decrease in classification performance. The error bars show a bootstrapped 95% credible interval from 100 bootstrap iterations.

imp-aggregate.png

Figure 7: The marginal permutation importance, as described in Figure ??, with the CIRI physical integrity index, the PTS, and the dynamic latent score estimated by ? as the dependent variables.

all-ols.png

Figure 8: Coefficient estimates for linear (OLS) models of the dependent variable indicated by the gray bar at the top of each panel. The dot represents the point estimate and the error bar a 95% confidence interval. The y-axis shows the model specification which is a base specification (the natural logs of GDP per capita and population in this case) in addition to the variable indicated. Only the coefficient of the variable of interest is shown. If the variable's 95% confidence interval includes zero (indicated by the dashed line), then the variable is *not* significant at $p < .05$.

As a more general point, it is worth noting that statistical significance does not perfectly correlate with variable importance. There is certainly a positive correlation between the two, but statistical significance is neither necessary nor sufficient for predictive validity. To illustrate that statistical significance is not sufficient for predictive validity, consider Figure **??**, which displays coefficient estimates and 95% confidence intervals for all variables from linear models using the aggregate repression scales as dependent variables. These models are fit using all of the available data and include GDP per capital and population size as control variables.[50] Notable in each panel of Figure **??** is the coefficient estimate for international war, which is larger than any other coefficient save that for civil war. Contrast this with Figure **??**, which indicates that international war adds no predictive power to either model. International war is a very rare event and thus in this instance, where the distribution of values in the dependent variable is much more balanced, it will be unable to improve predictive validity markedly. Additionally, the precision of the coefficient estimate is likely misleading due to the inflated statistical power that results from assuming the observations are independent. The larger point about predictive power versus significance is in line with the findings of at least one previous study (**?**), but bears mentioning as it is not widely appreciated.

Discussion/Conclusion

What do these results tell us about the state of the empirical literature on repression? The two concepts most widely recognized in the literature as important have predictive power and exhibit relationships with repressive violence that generalize beyond particular sets of data: civil war and democracy. However, the importance of these results is tempered by the problems we note above. The first of these results strongly suggests, in line with previous studies, that when government/dissident military violence produces a large number of deaths, governments often target non-combatants with violence. But again, this result is tainted by measurement problems. Indicators of civil war record only battle-related deaths, but will still pick up a subset of the violations recorded by PTS and CIRI. Many government-inflicted casualties in civil conflicts will be non-combatants and thus are likely to be picked up in measures of state repression, at least those that measure deadly uses of force (i.e. not the political imprisonment scale). Of obvious interest is the reciprocal relationship between dissident and government violence which results in conflicts escalating to the point where they are classified as civil wars,[51] but it is not reasonable to expect to capture these dynamics in annual, country-level data. Given that indicators of civil war tend to dampen the predictive power of many of the other covariates considered here, and will overlap empirically with the most widely available measures of repression, we would urge researchers to consider excluding them from their models.[52]

The relationship between democracy and repression is also general, but suffers from similar problems. To measure democracy we employed the Polity data, which contains the most commonly used measures in the literature. Results were particularly strong with respect to the "competitiveness of participation " component, and this component along with the other components and the aggregated scale added a tremendous amount of predictive power to the political imprisonment models. Part of this result is due to the fact that a government that imprisons their political

[50] Similar plots for the other models can be found in the appendix.

[51] Theoretical work on the dynamics of political violence is thin, but see **????**.

[52] We do not mean to impugn studies of the dissent/repression nexus (see Footnote 11), as research in this tradition typically employs (sub-national, sub-annual) data on *dissident violence itself* rather than civil war, which is a combination of government/dissident violence.

competitors cannot be considered "fully" democratic, given the way democracy is usually operationalized in this literature. This point is further underscored by the Polity codebook (**?**, p. 26), which makes it clear that the competition subcomponent in particular partly *measures repression.* The lowest category, called "repressed," is defined as follows:

> No significant oppositional activity is permitted outside the ranks of the regime and ruling party. Totalitarian party systems, authoritarian military dictatorships, and despotic monarchies are typically coded here. *However, the mere existence of these structures is not sufficient for a "Repressed" coding. The regime's institutional structure must also be matched by its demonstrated ability to repress oppositional competition* (emphasis added).

Examples of activities that may justify coding a state in the bottom three categories of this scale are:

> Systematic harassment of political opposition (*leaders killed, jailed,* or sent into exile; candidates regularly ruled off ballots; opposition media banned, etc.) (emphasis added).

Thus the fact that Polity predicts very well the imprisonment of political opponents and the aggregated repression scales, which both include information about political imprisonment, should not be surprising. This means that one of the strongest results in the literature is partially the result of estimating what are essentially tautological statistical models.

This is not to say that the predictive power of indicators of democracy is entirely meaningless. Notably, the executive constraints component of Polity does well in most of the models,[53] and this component does not suffer from the measurement issues that plague the competition component. But researchers who wish to employ the PTS or CIRI data should avoid using the aggregated Polity scale as well as the competition component. For the PTS this problem is especially bad since the scale cannot be disaggregated to exclude political imprisonment. Unfortunately this is also a problem for the CIRI scale to the extent that governments are repressing political opposition through the use of torture, kidnapping, and summary execution. This is because targeting political opponents with these tactics also reduces a government's level of democracy as defined/measured by the Polity scale. For future work, correcting this problem entails removing government violence that explicitly targets political opponents from the study's dependent variable, if one wishes to employ the aggregate Polity scale or the political competition scale as a covariate. Or, one could model violence against political opposition but remove Polity/political competition from the list of covariates included in the model.

Future theoretical work on repression should also take this point seriously, which entails developing arguments about why governments would have incentives to repress political opponents that do not use political competition as an explanatory concept. One possibility is suggested by the results for oil rents, which **?** theorize as related to repression because of the lack of incentives to protect human rights that results from increasing non-reliance on citizen-generated revenue. They note that this is consistent with arguments from the literature on democratization (E.g. **??**). This is a nice theoretical insight, since explanations for why governments would stop violently suppressing political competition are, in some sense, explanations for the emergence of democracy. Future work would benefit from incorporating more insights from the democratization literature about how other economic conditions, for example asset mobility and inequality, affect leaders' incentives to repress political opposition (See, e.g. **???**). This might help to further understand the relationship between per capita wealth and repression, which is empirically well-established but very rarely discussed in theoretical terms.

[53] Interestingly, this component of Polity is sometime used as an indicator of judicial independence. See, e.g. **???**.

Other promising results here are those for certain features of domestic legal systems, including judicial independence, constitutional guarantees for fair trials, and common law heritage. Judicial independence in particular performed well in both analyses, outperforming even civil war in predicting political imprisonment and torture. As discussed above, these results are consistent with the comparative literature on institutional constraints on government behavior, which views an independent judiciary as crucial in limiting government encroachment on basic rights (See esp. **???**). This also suggests that insights from theories of judicial behavior and the construction of judicial power may be useful for future work on repression (See **????**).

There is a nascent literature on domestic courts and human rights violations, but most of this literature examines the interplay between domestic courts and international legal obligations (**?????**). The results here strongly suggest that the relationship between domestic courts and repression is general and does not depend on a government's ratification status for various international human rights treaties, so studying the impact of domestic courts themselves on repression would be useful.[54]

Regarding constitutional rights protections, the results for fair trial provisions suggest that formal legal protection of basic rights may be more than a "parchment barrier," (See **?**) and justify more attention to the law itself, including criminal trial procedures, in future studies.[55] The performance of the common law heritage measure used by **?** also suggests that trial procedures and judicial behavior are relevant to research on state repression; the theoretical argument connecting common law systems to respect for rights focuses on the trial procedures typical of common law systems (adversarial trials, oral argumentation) and the principle of *stare decisis*, i.e. that legal precedent constrains subsequent interpretation. Overall, legal institutions have received far less attention in the literature than democratic political institutions, and the performance of these three legal institutions merits further research. Further, since these measures of legal institutions are not *tautologically* related to commonly used measures of repression they represent a more fruitful path for future research than additional studies examining the relationship between the Polity scale and PTS/CIRI.

Another promising result was the excellent performance of the youth bulges measure used by **?**. The theoretical reason for the relationship between a large per capita youth population and repression is preemptive action, on the part of the government, to prevent large-scale rebellion. The strength of this result suggests that demographic factors beyond mere population size should be more closely examined in the future.

While indicators of potential international influences on repression did not perform as well as features of domestic politics, trade openness in INGO presence performed well in our random forest analysis, which allows for more complex relationships (non-linear/interactive) than the cross-validation analysis. This suggests that some of the inconsistent results in this literature may be due to complicated relationships between international political factors and repression that commonly used models will fail to detect. Thus rather than downplay the importance of these factors for future research, we would suggest that the complex nature of these relationships is something that deserves more attention.

Another notable finding was that the performance for most covariates was uneven across indicators of repression. Most of the covariates that perform well do so for the aggregate scales and the CIRI political imprisonment and torture scales, but not the disappearance or killing scales. This means that theoretically motivated models of repression are often explaining only part of what they are supposed to explain. As noted above, this is partly due to the relative rarity of disappearances

[54] See **?** for a review of existing measures of judicial independence, and **?** for a promising approach to measuring that concept.

[55] See **?**, who laments a lack of attention to the law in research on human rights violations.

and summary executions. Still, most analyses examine all of these practices together, treating them as homogenous, and the results above suggest that the different repressive acts measured by PTS/CIRI may be driven by different processes. Thus researchers would do well to not simply assume *a priori* that their covariates are related to each practice identically, i.e. they should disaggregate these indicators if possible.[56]

In short, our analysis suggests that there are many potential, fruitful paths for future research on state repression. It is not our intention to treat these results as the final, definitive statement about what are the "most important" causes of repression. Rather, we have shown that some of the hypotheses advanced in the literature receive much stronger support than others, offered an appraisal of existing explanations for repression in light of these results, and suggested how the patterns our analysis reveals can usefully inform future theoretical and empirical research on this topic.

The broader problem we outline with respect to the repression literature, i.e. attention to statistical significance alone, is one that is common in many areas of political science research. Researchers in all subfields should supplement their usual analysis with some examination of model fit (as measured by various statistics such as RMSE, area under the ROC curve (for binary response models), likelihood-based statistics like AIC, etc.) and in particular an examination of whether including covariates their theory suggests are important *improves* the fit of the model. Just as important, some effort should be made to assess whether the inferences drawn from a model generalize to other sets of data. Cross-validation addresses both of these problems, is easy to understand, and fits comfortably with the regression analyses that political scientists often conduct, so we would urge researchers to familiarize themselves with these techniques.[57]

A final point is that prediction *per se* was not our goal. If it was, there are several ways we could improve the accuracy of the regression models used in the cross-validation analysis. We mention above the improved accuracy that would likely result from the inclusion of a lagged dependent variable. More complicated strategies for modeling temporal dynamics may also be helpful,[58] though data on repression are usually collected at a high level of temporal aggregation. Mixture models, such as "zero-inflated" models, offer another promising approach (See, e.g. **??**). Also, the estimators we use treat these observations as independent, which is standard practice in the literature, but using a model that more realistically accounts for the structure of the data would undoubtedly result in better predictions.[59] Most importantly, we wish to stress that predictive validity should be used more often to evaluate the accuracy of theoretical explanations for repressive government violence.

[56] There is a nascent body of research that examines the relationships among the repressive practices themselves to determine whether they are generally complements (i.e. states typically employ these practices in combination) or substitutes (i.e. greater use of one reduces use of the others). See, e.g. **???**.

[57] For a useful introduction to these methods for political science research see **?** and **?**.

[58] See, e.g., (**?**)

[59] E.g., mixed effects models would markedly improve our predictions. See the modeling strategy presented in **?**.

Appendices

A Data Descriptions

Data on population and trade come from **?**. These are data from the Penn World Tables (**?**) with missing values imputed using information from the CIA World Factbook and procedures described fully in **?**.

The measure of youth bulges used above comes from **?**, who uses demographic data from the UN to construct a measure of the proportion of the adult population (older than 15) that is younger than 25.

The indicator of oil rents is due to **?**, and measures the total value of oil and natural gas production, accounting for extraction costs. This figure is divided by mid-year population.

Data on civil and interstate war come from the UCDP/PRIO armed conflict data-set (**?**). The civil war variable is equal to one for all years in which a country experienced conflict between the government and rebel groups resulting in at least 1000 battle-related deaths. The interstate war variable is equal to one for years in which a country's government was involved in a militarized conflict with another government resulting in at least 1000 battle deaths.

All of our measures of democracy come from the Polity IV regime characteristics data (**?**). The democracy component of Polity is comprised of four subcomponents which measure competitiveness of executive recruitment, openness of executive recruitment, executive constraints, and the competitiveness of participation. We use each of these individual components in the analysis. The most commonly used indicator of democracy results from subtracting the aggregated autocracy scale (which measures the four characteristics above in addition to the regulation of participation) from the aggregated democracy scale. We also use this measure in the analysis.

Information on military regimes and leftist regimes comes from the Database of Political Institutions (**?**). The military regime variable is coded one if the chief executive is a military officer or an officer who has not formally retired from the military before assuming office. The leftist regime variable is coded one for chief executives identified as communist, socialist, social democratic, or left-wing based on their economic policies.

Data on constitutional provisions come from **?**. These are all binary and are created by coding the text of national constitutions. The variables we use indicate the presence of provisions for a fair trial, provisions for a public trial, provisions stating that the decisions of high/constitutional courts are final, and provisions which require legislative approval to suspend constitutional liberties.

The measure of common law legal systems comes from **?**. This is a binary variable coded one if a country's legal system has primarily features of a common law system. Other possible categories are civil law, Islamic law, and mixed legal system.

Measures of trade openness and foreign direct investment both come from the World Bank's World Development Indicators (**?**). These measure trade as a percentage of GDP and FDI net inflows as a percentage of GDP.

Indicators for participation in IMF and World Bank structural adjustment programs come from **?**. We use three binary indicators, one which is coded one if a government is currently participating in an IMF structural adjustment program, another which is equal to one if a government is participating in a World Bank structural adjustment program, and another which is coded one if a government is participating in structural adjustment programs with both the World Bank and the IMF.

Data on preferential trade agreements (PTAs) with human rights clauses comes from **?**. This variable is coded one for all years a government is a member of at least one PTA with a "hard" human rights clause. A hard clause is defined as one that explicitly mentions human rights principles and also declares that the benefits of the agreement are conditional on observing those principles.

Our measure of INGO presence comes from **?**, and is a count of the number of INGOs of which a government's citizens are members. Two of our three INGO shaming measures come from **?**. These are counts of the annual number of press releases and background reports issued by Amnesty International about a particular country. From **?** we use an events data-based measure which is a count of the annual number of conflictual actions human rights organizations send to a particular government. As a fourth shaming measure we use another variable from **?** which measures the annual, average number of stories about a particular country published in Western media outlets (Newsweek and The Economist) which mention human rights practices.

Finally, our measures of UN treaty ratification status are taken from the UN website via the `untreaties` utility.[60] We use two indicators, one coded one for every year a country has ratified the Convention Against Torture, and another indicating ratification status for the International Covenant on Civil and Political Rights.

B Additional Cross-Validation Results

[60] Available at http://github.com/zmjones/untreaties

cv-ols-ldv.png

Figure 9: Cross-validation results from linear models (OLS) of the same form as Figure **??**, but with a base specification which consists of (the natural logs of) GDP per capita and population as well as a lagged dependent variable. Note that the number of observations when including a lagged dependent variable is somewhat lower since we do not impute missing values of the lagged dependent variable (the first year for each country) since we have no contemporaneous information on its likely value.

cv-lrm-ldv.png

Figure 10: Cross-validation results from ordinal logistic regressions of the same form as Figure **??**, but with a base specification which consists of (the natural logs of) GDP per capita and population as well as a lagged dependent variable.

cv-lrm-pts.png

Figure 11: Cross-validation results from ordinal logistic regressions of the same form as Figure ?? with the addition of the Political Terror Scale (omitted in Figure ??).

cv-lrm-cwar-pts.png

Figure 12: Cross-validation results from ordinal logistic regressions of the same form as Figure **??** (the base specification includes the natural logs of GDP per capita and population as well as civil war) with the addition of the Political Terror Scale (omitted in Figure **??**).

C Coefficient Estimates

all-ols-cwar.png

Figure 13: Coefficient estimates of the same form as Figure **??**, with the base specification being the natural logs of GDP per capita and population in addition to civil war.

all-ols-ldv.png

Figure 14: Coefficient estimates of the same form as Figure **??**, with the base specification being the natural logs of GDP per capita and population in addition to a lagged dependent variable.

all-lrm.png

Figure 15: Coefficient estimates for ordinal logistic regression models of the dependent variable indicated by the gray bar at the top of each panel. The dot represents the point estimate and the error bar a 95% confidence interval. The y-axis shows the model specification which is a base specification (the natural logs of GDP per capita and population in this case) in addition to the variable indicated. Only the coefficient of the variable of interest is shown. If the variable's 95% confidence interval includes zero, then the variable is *not* significant at $p < .05$.

all-lrm-cwar.png

Figure 16: Coefficient estimates of the same form as Figure ??, with the base specification being the natural logs of GDP per capita and population in addition to civil war.

all-lrm-ldv.png

Figure 17: Coefficient estimates of the same form as Figure **??**, with the base specification being the natural logs of GDP per capita and population in addition to a lagged dependent variable.