Measuring Electoral Democracy with Observables

Daniel Weitzel
Department of Political Science
Colorado State Univesity

John Gerring
Department of Government
University of Texas at Austin

Daniel Pemstein
Department of Political Science
North Dakota State University

Svend-Erik Skaaning
Department of Political Science
Aarhus University

Wordcount (without abstract and appendices): 9,000
Appendices: 16 pages

ABSTRACT

Most crossnational indices of democracy rely centrally on coder judgments, which are susceptible to personal bias and error, and also require expensive and time-consuming coding by experts. The few measures based exclusively on observable indicators are either dichotomous or rely on a few rather crude proxies. This project lays out an approach to measurement based on observables that aims to preserve the nuanced quality of subjectively coded democracy indices.

First, we gather data for a wide range of observable indicators, $X´$, that capture different aspects of the democratic process. Next, we use supervised random forest machine learning to predict Z using factual indicators, $X´$, creating an **o**bservable-to-**s**ubjective score **m**apping (OSM). The mapping that provides the best cross-validated fit to the outcome serves as an alternate index, $Z´$, for that conceptualization of democracy.

Information loss from Z to $Z´$ is minimal for indices centered on an electoral conception of democracy and this loss may be advantageous for some purposes. It is free of idiosyncratic coder errors arising from misinformation, slack, or biases for or against a regime. It is also less susceptible to systematic bias that may arise from coders' inferences about a country's regime status, e.g., from the ideology of the current ruler. The data collection procedure and mode of analysis is fully transparent and replicable, and the procedure is cheap to produce, easy to update, and offers coverage for all polities with sovereign or semisovereign status, surpassing the sample of any existing index. We show that this expansive coverage makes a big difference to our understanding of some causal questions.

Most crossnational indices of democracy rely on coder judgments. This feature of measurement may be ineradicable, especially for aspects of democracy that are hard to observe and therefore require judgment by knowledgeable coders versed in the history of a particular country (Bollen 1993; Bowman et al. 2005; Coppedge et al. 2019; Munck 2009). “If we were to renounce our judgmental faculties in the measurement of regime properties and regime dynamics,” Schedler (2012: 33) argues, “we would have to renounce the measurement of most of the most interesting regime properties and regime dynamics.” At the same time, we must acknowledge that coder judgments are susceptible to bias and error and are also expensive to produce.

Fortuitously, many features of democracy leave an observable trace. For example, the freeness of an election may be inferred from the outcome of that contest, i.e., the share of votes won by the incumbent party, the margin of victory, and whether turnover occurred in control of the executive or parliament. These traces allow for measurements based on observables, an approach adopted by one of the very first attempts to measure democracy crossnationally (Cutright 1963).

Later projects following in Cutright’s footsteps (e.g., Alvarez et al. 1996; Vanhanen 2000) suffer from three common limitations. First, they are not always as objective as they seem, relying on subjective judgments or idiosyncratic coding instructions for key variables. Second, they reduce the conceptual space of democracy into binary or ordinal indices, with

consequent loss of information, or they rely on information from a small number of rather crude proxies. Finally, they are limited in coverage.

We seek to combine an objective approach to measurement with the nuance afforded by subjectively coded indices. We gather data for a wide range of observable outcomes that capture different aspects of the democratic electoral process. Next, we train a random forest to map factual indicators onto an existing index, Z, creating an observable-to-subjective score mapping (OSM). The mapping that provides the best cross-validated fit to the outcome serves as an alternate index, Z'.

Naturally, there is some information loss from Z to Z'. However, we show that the loss is minimal for a wide range of democracy indices. Accordingly, an index based on observables may be advantageous for some (though not all) purposes.

First, Z' is less prone to idiosyncratic coder errors arising from misinformation, slack, biases for or against a regime, or data-entry mistakes. It is also free of certain systematic biases that might be shared across coders such as ideological biases in favor of left- or right-wing governments.[1]

Second, the data collection procedure and mode of analysis used to construct Z' is transparent and replicable. Comparisons through time or

[1]We demonstrate these features in Appendix I through the introduction of large, simulated, biases. Across these extreme scenarios, our approach substantially reduces the introduced bias – by 83% in the easiest case of completely random bias and by 8% in a scenario where the bias is highly correlated with outcomes and predictors and forms a strong cluster at one end of the distribution.

across countries can be interpreted in specific terms, i.e., as the product of a specific set of observable quantities.

Third, the procedure is cheap to produce and easy to update. For any democracy index, Z, one can generate an OSM, $Z´$. Out-of-sample coverage for $Z´$ will include all polities with sovereign or semisovereign status, surpassing the sample of any extant index, Z. This is possible because the observable features of polities are fairly easy to gather and do not require in-depth knowledge of cases. $Z´$ can therefore be applied to micro-states, quasi-sovereign polities (e.g., colonies and dependencies), and defunct historical polities. We show that expansive coverage makes a difference to our understanding of some important causal questions.

We begin this article with a discussion of extant indices of democracy. Next, we present our methodology for measuring democracy with observables using the Polyarchy index from the Varieties of Democracy project as our test case. The third section assesses the fit between the original index and the OSM. The fourth section seeks to understand remaining deviations with a regression model focused on potential sources of disagreement. The fifth section generalizes our approach across other widely used democracy indices. The sixth section assesses potential ideological biases in extant indices using $Z´$ as a benchmark for Z. The seventh section examines what can be learned from extending our coverage from the usual country cases to a much larger set of unstudied cases.

A final section discusses the uses, and potential misuses, of this approach to measurement. It should be clear that we do not regard OSMs as wholesale replacements for subjective indices. Rather, we regard them as an important complement insofar as they provide estimates that are resistant to certain (not all) biases, are cheap to develop and replicate, and offer superior coverage.

I. Extant Indices

We list the most widely used measures of democracy in Table 1, along with some key features. Appendix H discusses coder judgments, which are summarized in the first column. In the sections that follow, we discuss problems of (a) subjective error, (b) ambiguity, and (c) coverage. We conclude with a brief discussion of a recent pioneering effort to produce an index of democracy using machine-learning.

Table 1: Extant Democracy Indices

	Coder judgment	*Scale*	*Raters (N)*	*Politi es*	*Years*	*Obs.*	*GS cites (2015 -)*
Freedom House (Freedom House 2015)	High	Ordinal	1	`202	1972-	7,598	1,780
Polity2 (Marshall et al. 2013)	High	Ordinal	1	182	1800-	15,772	2,360
Unified Democracy Scores ("UDS") (Pemstein et al. 2010)	High	Interval	N/A	198	1946-	9,258	457
Polyarchy (Teorell et al. 2019; Coppedge et al. 2020)	High	Interval	5	177	1789-	25,759	872
BMR (Boix, Miller, Rosato 2013)	Low	Binary	1	208	1800-2015	15,620	688
Democracy-Dictatorship ("DD") (Alvarez et al. 1996; Cheibub et al. 2010; Bjørnskov et al. 2020)	Low	Binary	1	208	1950-2018	13,728	459
Democracy Barometer (Bühlmann et al. 2012)	Low	Interval	N/A	70	1990-2017	1,431	481
Lexical index of electoral democracy (Skaaning et al. 2015)	Low	Ordinal	1	224	1789-	17,020	146
Democracy (Vanhanen 2000, 2011)	Low	Interval	1	203	1810-2013	14,984	331
Machine-learning democracy index ("MLI") (Gründler, Krieger 2016, 2021)	Low	Interval, Binary	N/A	186	1919-2019	12,588	151

The Freedom House index combines the Political rights and Civil liberties indices into a single index. *Raters:* average number of independent coders per country-year. *Obs:* country-year observations. *GS cites:* Google Scholar citations (approximate) from 2015 to 2022. All measures of democracy are highly correlated (Appendix L).

Subjective Error

All extant democracy indices involve some degree of coder judgment, which we have attempted to code (subjectively) in the first column of Table 1. This leads to a variety of potential sources of error.[2]

[2] Our discussion builds on Alvarez et al. (1996), Bollen (1990), Bollen and Paxton (2000), Cheibub et al (2010), Munck (2009), and Skaaning (2018).

Expert coders are not always strongly motivated and some may not be conscientious in undertaking a task that is time-consuming, onerous, and poorly remunerated. Some raters may not be fully qualified to assess the country they code. This is especially a problem with micro-states and historical states, neither of which are well-studied and by numerous qualified experts.

If the same coder assigns scores to all countries and all time-periods there is almost assuredly a problem of expertise, for who can master the history of every country? In this circumstance, coders are likely to rely on common perceptions rather than in-depth knowledge of the case at hand (Bowman et al. 2005). If, on the other hand, each expert covers a different country, region, or time-period it is difficult to achieve cross-coder comparability (Coppedge et al. 2020: chs 3-4).

Regardless of their expertise, coders may hold different views, which is likely to lead to varying judgments. Coders may also rely on different sources of information or assign different weights to the same sources. They may base their judgment on irrelevant issues and make inadvertent coding errors.

Stochastic error is problematic, as democracy measures do not employ a great number of coders per country. The modal number is one, as Table 1 shows. While Freedom House and Polity subject original scores to internal review processes, they do not report which cases are adjusted, how much scores change, or why revisions have been implemented. By contrast,

input from V-Dem experts is independent but there are only five coders per country-variable-year (on average), and just one or two coders for years before 1900. Pooling estimates from different projects, as UDS does, raises the sample of coders slightly – but not if the same people are working as coders for different projects. In any case, these are very small samples compared with other expert surveys,[3] not to mention surveys of the mass public.

More pernicious than random error is systematic error, of which several varieties deserve special mention. The first may be characterized as country-specific – where coders have an especially positive, or negative, view of the country they are coding, which then infects judgments on specific questions. From what we know about the V-Dem project (which publishes anonymized data about their experts) and what we can infer from other projects, democracy experts share a common set of characteristics. They usually have an advanced degree in political science (or related fields), are often associated with a university in the West (where they work or where they obtained their degree), and tend to hold liberal and cosmopolitan views. It is not hard to imagine they might also share certain biases, e.g., in favor of governments that pursue more liberal policies and against those who pursue more conservative policies.

[3] The Chapel Hill survey enlists an average of thirteen coders per country (Bakker et al. 2015) and the Electoral Integrity project enlists an average of forty experts per country (Norris et al. 2013).

Two prominent projects – Polity IV and Freedom House – are closely related to the US government, which provides ongoing funding. It is sometimes alleged these outfits, or at least Freedom House, project an American-centric measure of democracy and code countries close to the US more favorably than those outside the US orbit (Bush 2017; Giannone 2010; Steiner 2016).

Another sort of bias is historical. Because coders know a country's trajectory, they may unconsciously incorporate that knowledge into their judgments. For example, coders of Germany may assume that the Weimar period was not very democratic because of its subsequent collapse.

A third sort of bias is the assumption that good (bad) things go together, a "halo" effect. For example, suppose one is trying to judge the freeness and fairness of elections in Liberia during the nineteenth century. Coders may tacitly assume (without thinking consciously about it) that because the country is poor and located in a region where democracy was rare, elections were not very free and fair. In contemporary times, when Liberia was wracked with civil conflict, coders may assume that elections are not free and fair because of the existence of such conflict. In the post-conflict era, as Liberia recovered from economic crisis and things began to improve generally coders may assume that the quality of elections also improved.

All sorts of assumptions may be smuggled in when coders attempt to reach determinations on unobservable, hard-to-judge dimensions where

information is scarce. They could be true, or they could be false. In the latter case, they will induce spurious correlations between democracy and other phenomena, e.g., peace/conflict or economic development. Note that insofar as these biases are widely shared they must be regarded as systematic rather than idiosyncratic.

Ambiguity

An additional problem with subjective coding is that the resulting index of democracy is difficult to interpret. This problem is most obvious for indices that are broadly and vaguely defined like Freedom House and Polity2. It is true, a fortiori, for meta-indices such as UDS. We do not know what these indices mean because we do not know all the factors that may have contributed to coder judgments about each country's scores over time.

Binary indices are more precisely defined; however, they group together polities that are extremely heterogeneous. For example, both Singapore and North Korea receive a code of 0 (autocratic) in the BMR and DD datasets. This constitutes a considerable loss of information and leads to ambiguity of a different sort (Bollen 1990; Elkins 2000).

In principle, V-Dem's Polyarchy index is more interpretable as it can be disaggregated into specific indicators. However, these component indicators are not entirely independent. Codings related to the quality of elections may reflect impressions of human rights, media freedom, and other related matters. Consequently, we do not know precisely what causes

changes in a V-Dem index over time or what accounts for variation across cases.

Coverage

Whether resting on subjective coding or observable features of regimes, all democracy indices are limited in coverage, as noted in Table 1. The Democracy Barometer covers only seventy (largely democratic) countries from 1990 forward; it is, effectively, a “quality of democracy” index for countries that have surpassed a minimal threshold of democracy. Other indices treat only the contemporary era (e.g., DD, Freedom House, UDS). A small number extend back to the nineteenth century but include only sizeable sovereign countries (e.g., BMR, Polyarchy, Polity, Vanhanen). Many datasets are not regularly updated. No dataset includes a comprehensive set of sovereign and semisovereign units (e.g., colonies, dependencies) back to 1789.

The reason for this is presumably that expert coding is laborious and historical information required for coding is difficult to locate. Moreover, well-qualified country experts are rare, and not always willing to spend their scarce time on coding projects, especially if they require regular updates.

One might conclude that history is inessential to understanding the present, or that smaller countries, defunct countries, or entities that are not fully sovereign are inessential. For some questions this may be true.

However, the exclusion of polities that are older, smaller, non-sovereign, or for whatever reason less studied, constitutes an enormous loss of information. Moving back in time, colonies and other semisovereign units gain importance, constituting a large share of all polities and of the world's population prior to the turn of the twentieth century. Defunct states like Bavaria were just as important at the time, and just as sovereign, as many states that managed to survive. In comparative politics, as in international relations, we need to understand the losers as well as the winners. Survival bias is a problem.

Expanding the sample of available cases should also improve internal validity by reducing threats from stochastic error. This is a particular problem in crossnational analysis, where samples are small and extremely heterogeneous. Note that democracy is a sluggish variable, meaning that leverage is primarily latitudinal rather than longitudinal. Every case counts in a cross-sectionally dominated panel.

Finally, a more representative sample mitigates problems of external validity. We cannot be sure that commonly included and excluded countries are similar. Indeed, there are good reasons to think otherwise (see Section VII).

Machine Learning

A final index utilizes a method of aggregation that bears casual resemblance to our own and thus demands discussion. Gründler and Krieger (2016,

2021) use a support vector machine (SVM) trained on the Polyarchy and UDS indices (in their revised approach). The predictor variables are primarily observable but also include three factors measuring party pluralism and freedom of discussion that are classified as subjective. Models learn the relationship between democracy and these component variables from the upper and lower decile of the distribution for the Polyarchy and UDS indices. The SVM then predicts new democracy scores for all polities across the entire distribution, referred to as the machine-learning democracy index ("MLI").

In this fashion, Gründler and Krieger offer an innovative approach to the eternal aggregation problem. Naturally, it is not without assumptions. For present purposes, what bears emphasis is that our initiative is quite different. We do not seek to present a new index of democracy. Rather, we produce estimates of scores for existing indices using observable features of the world, a procedure which, if effective, reduces the scope for certain types of error, and also greatly expands the range of coverage. As expected, our OSM index is more strongly correlated with the original indices than the MLI, especially in the middle of the distribution (see Appendix L).

II. Methodology

Our protocol begins with the choice of an index and proceeds to the selection of observable indicators, the application of nonparametric

supervised machine learning techniques, followed by various model diagnostics.

Indices

The bane of composite indices is aggregation. Every democracy index struggles with this problem. Some rely on a set of necessary and sufficient conditions (Lexical, BMR, DD). Others establish categories, each with separate criteria (Freedom House). A third approach rests on formulas for aggregating component indicators (Polyarchy, Polity2). A fourth approach enlists principal components analysis (Coppedge et al. 2008) or latent variable models (Marquardt, Pemstein 2018).

All these approaches to aggregation are defensible and none clearly superior, accounting for the persistence of such radically different techniques. We offer no solutions to this eternal conundrum. Instead, we treat each existing composite index as an instantiation of a unique conception of democracy. For each conception (index), we propose an operationalization that relies entirely on observable features of the world.

Following common practice, we focus our attention on the electoral conception of democracy, understood as representative democracy achieved through competitive elections along with other supporting institutions. The Polyarchy index from the V-Dem project offers an illustration of this approach. (Section V discusses results for other widely used indices.)

Indicators

Having identified a conception of democracy and selected an index, we search for potential indicators. Criteria of inclusion include (a) relevance for the concept of electoral democracy, (b) observability, and (c) coverage.

Any feature that promises to facilitate the rule of the people through competitive elections is eligible for inclusion. We restrict our canvas to institutions, as the role of attitudes and values is uncertain. It is unclear, for example, whether a country in which people are strongly supportive of democracy is more democratic than another country – identical in all other respects – in which people are skeptical of democracy. Accordingly, we do not consider survey data or other measures of political culture. We also exclude indicators like per capita GDP that might predict democracy but are not constitutive or reflective of democracy.

Observability means that a feature can be collected and coded with little or no judgment on the part of the coder. It is factual in nature. Accordingly, replication of our dataset should be easy, following the guidelines in our codebook (see Appendix A). Granted, there are situations in which the historical record is unclear, e.g., where we do not know, or do not know for sure, what the vote or seat total was for the winning party. Here, data is missing or questionable, and reasonable people may disagree. Moreover, the discovery of new evidence may prompt revision of our data. However, we suspect that these cases are rare.

Coverage, the third criterion, is a matter of degrees. The greater the spatial and temporal coverage, the more useful an indicator is (ceteris paribus), especially if coverage for a prospective indicator complements coverage for other indicators.

In summary, our goal is to identify factual indicators of all institutions that are potentially indicative of the state of electoral democracy and are measurable globally and historically. Forty variables, described in Appendix A, meet these criteria.

The impact of possible omissions from this list of variables is difficult to address. Conceivably, important observable indicators are missing from our collection. However, the extremely tight fit obtained from the set of chosen variables suggests that any additional variables are unlikely to change index scores by very much. There simply isn't much variance left to explain.

Later iterations of our model reduce the collection of variables from forty to thirteen to aid interpretability and to reduce the costs of extending or replicating this work. We selected the thirteen indicators in the revised model based on their importance scores, as described below.

A Random Forest Model

To compose an objective index, Z', based on an existing democracy index, Z, we train a random forest algorithm on Z using the set of forty variables

introduced above, producing an OSM through prediction.[4] Random forests are meta estimators, averaging over a large collection of individual decision trees. The main idea behind this ensemble learning method is to combine multiple decision trees to offer more accurate and robust predictions. Decision trees partition the covariate space through recursive binary splitting. These smaller subsets are based on a certain feature, and the tree continues to grow until the split results in pure subsets (i.e., subsets that only contain data belonging to one class of the dependent variable). Each decision tree is therefore restricted to a random sample of observations and predictors, never the full set (Hill and Jones 2014).[5]

The process of sampling from the predictors at each node allows the algorithm to learn the optimal split decisions to partition the data. However, it also makes each individual decision tree noisier. Random forests extend decision trees by creating multiple trees, as mentioned above, and combining their predictions to make a more accurate final prediction. Growing many decision trees and averaging improves prediction accuracy, makes the random forests robust to highly correlated variables, most

[4] In Appendix F we also report the results of a gradient boosting machine (GBM), XGBoost, and generalized linear models (GLM). We use a random forest to construct our OSM because it consistently out-performs other techniques in cross-validation and validation data sets. Throughout we use models from the H2O package in R, allowing researchers to implement this approach with their preferred programming language.

[5] We further explain the use of random forests in Appendix J.

importantly, reduces the danger of overfitting idiosyncrasies in the training data.

Whether applied to continuous (regression) or discrete (classification) response variables, this non-parametric, supervised machine learning algorithm is ideal for evaluating the value of multiple predictors and interactions among them, differentiating those with strong predictive power from those that are redundant or predict idiosyncratic variation in the training data. In addition to model-fit statistics, as in more conventional models like OLS, random forests also provide metrics on the predictive power of individual predictors included in the model. These “importance” scores allow the reader to assess which predictors are central to the performance. The ability of random forests to accommodate response and predictor variables of different types, missing data, as well as variation in the balance of classes (dependent variable values) accounts for their popularity across the sciences and social sciences (Breiman 2001; Hastie et al. 2013).

To ensure that our model produces generalizable predictions, avoiding overfitting, we divide our data into different groups. The Polyarchy index provides 25,759 country-year observations for 195 countries from 1789 to 2021. We split this dataset into three parts: a training set, a validation set, and a test set. The training set consists of a random subset of 65% of the total observations, which we use to train our random forest. In this dataset the algorithm learns about the relationship between our target

variables, the democracy measures, and the objective predictors. We iteratively test the performance of the trained OSM on the out-of-sample validation set, a random sample of 15% of the total observations. The remaining 20% comprise the test set. This data set will be used to assess the final model performance at the end of the project on data that the model has never seen or been calibrated against.

We use cross-validation to train the model, splitting the training set into 15 folds for k-fold cross-validation. This approach involves dividing the data into k subsets (called "folds"), training the model on k-1 folds, and evaluating the performance on the remaining fold. We repeat the process k times, with each fold serving as a separate validation set. We estimate the model's performance on unseen data with the average performance across all k folds. Cross-validation greatly reduces the risk of overfitting.[6]

In the reduced model, we estimate 130 (number of variables*10) trees, allowing for a maximum tree depth of 20.[7] We estimate this random forest in two specifications. First, we randomly select country-year observations for training, validation, and test data based on the entire data set, without stratification. This specification is used for an overall fit to

[6] Cross-validation on the training set makes over-fitting of the validation dataset unlikely. Thus, the test set serves largely as a fail-safe to ensure that we have not accidentally contaminated our results. We will assess the performance of our final OSM on the test set at publication time, after incorporating any changes suggested by reviewers.

[7] The dataset has numerous missing values. We apply various imputation approaches. Results, in Appendix D, are similar to those we report in the manuscript.

assess biases. Second, we stratify the dataset by country, assigning all country-year observations from each country to either the training, validation, cross-validation, or test set. This is a somewhat more realistic test of out-of-sample performance. Researchers primarily interested in out-of-sample prediction can find country-stratified model specifications and explanations in Appendix K.

Variable Importance

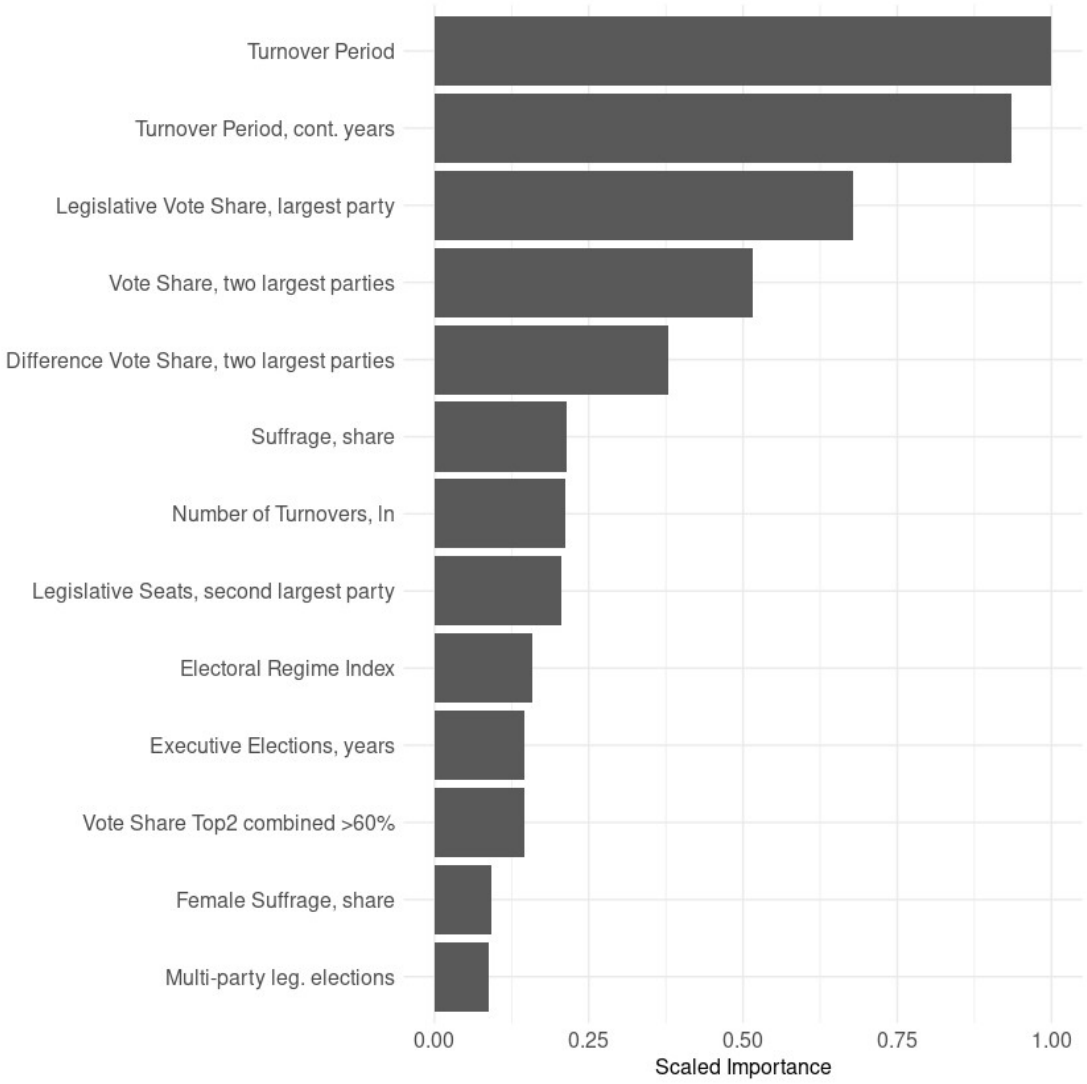
Some observable indicators are more useful than others in predicting a particular democracy index. To simplify the procedure, we produce a second OSM that eliminates indicators that contribute very little to overall fit. In the case of Polyarchy, we reduce the initial set of forty variables to thirteen, ranked according to their importance for the distributed random forest in Figure 1. Variable "importance" measures the extent to which inclusion of a variable decreases the entire forest's squared prediction error and how valuable the variable is for splitting the data within individual trees. Important variables produce highly informative splits and thus show up near tree "roots."

The thirteen variables of special importance to Polyarchy may be understood conceptually along four dimensions. Five variables reflect the vote or seat shares of the top parties. Three variables reflect turnover in control of the executive. Three variables measure the existence of elections,

whether key offices are elective, and whether multiple parties were allowed to compete in those elections. Two variables capture the extent of suffrage.

For each of these dimensions there are several variables, attesting to the varying ways in which these concepts can be operationalized. Consider the key concept, *turnover period*, which is scored zero until an election-instigated turnover of control over the executive, and one thereafter – unless multi-party elections are suspended, at which point the scoring reverts to zero until another election-instigated turnover takes place. One variable ("Turnover period") measures whether a given year falls within a turnover period, another ("Turnover period, cont. years") measures how many years a country has been within a turnover period, and a third ("Number of Turnovers, ln") measures the number of turnovers in a country's history (logged).

Figure 1: **Variable Importance**



The variables (N=13) with the highest importance scores in the random forest model, with Polyarchy as the target.

Since the selection of indicators is a crucial part of this exercise, we conduct a series of robustness tests in which individual variables are removed from the benchmark model (composed of thirteen variables), recalibrating the algorithm each time. The variations that result from these serial omissions are very slight, as shown in Appendix B. Accordingly, the

results reported in this study are not contingent upon the inclusion of any single variable.

Before concluding we must call attention to an important feature of our protocol. Any democracy index that incorporates observable features of the world is likely to see those same features included in an OSM developed to predict the index. In the case of Polyarchy, the overlap involves two variables – suffrage and electoral regime. In the case of democracy indices resting largely on observables such as the Lexical index the overlap would be even greater. By contrast, for democracy indices resting entirely on coder judgments, such as Freedom House, there is no overlap.

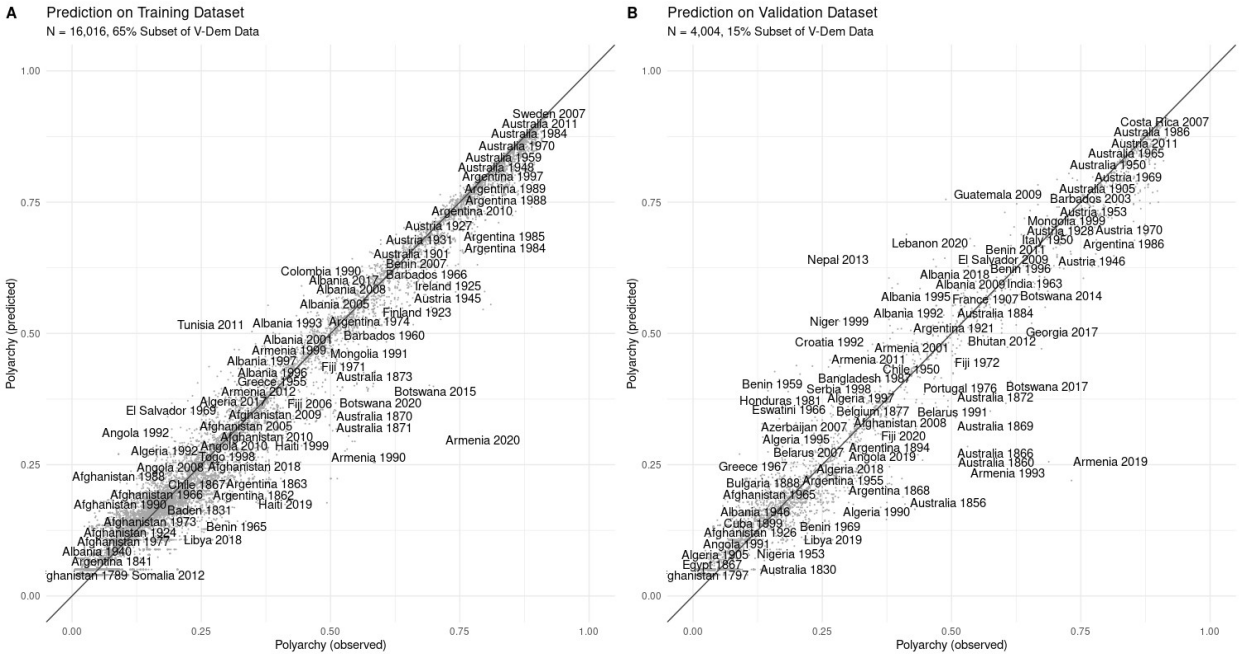
Although there is some circularity to our approach (with respect to indices that incorporate observables) it should be clear that the set of observables composing Z' is much larger than the set of observables in Z. Note also that attempting to predict Polyarchy with only suffrage and electoral regime would not get you very far. Moreover, *excluding* these variables from our OSM scarcely attenuates fit, as neither is of high importance (see Figure 1). In any case, our goal is predictive, not causal. Accordingly, overlap between Z and Z' is regarded as a feature rather than a bug. The purpose of our venture is to purge existing indices of subjectively coded components, not to propose an entirely novel set of observable measures.

III. Assessing the Fit

Because Polyarchy is continuous, we use a regression estimator within the random forest. The resulting model performs well, producing R squared values of 0.95 in the training, validation and cross-validation sets with a mean squared error (MSE) of 0.003 in the validation data. Since the outcome, Polyarchy, ranges from 0 to 1 this a very low average squared difference between the predicted and the observed values.

In Figure 2 we plot the original Polyarchy index against predictions from the random forest, labeling a random subset of those observations. The distribution of points lies in a symmetrical fashion near the 45-degree line. Some instances such as Armenia in 2020 are underpredicted. In this case, we suspect that the enormous gain of 83 seats of the new incumbent party in 2018 (70% of the seats in the National Assembly) is driving our model's conservatism.

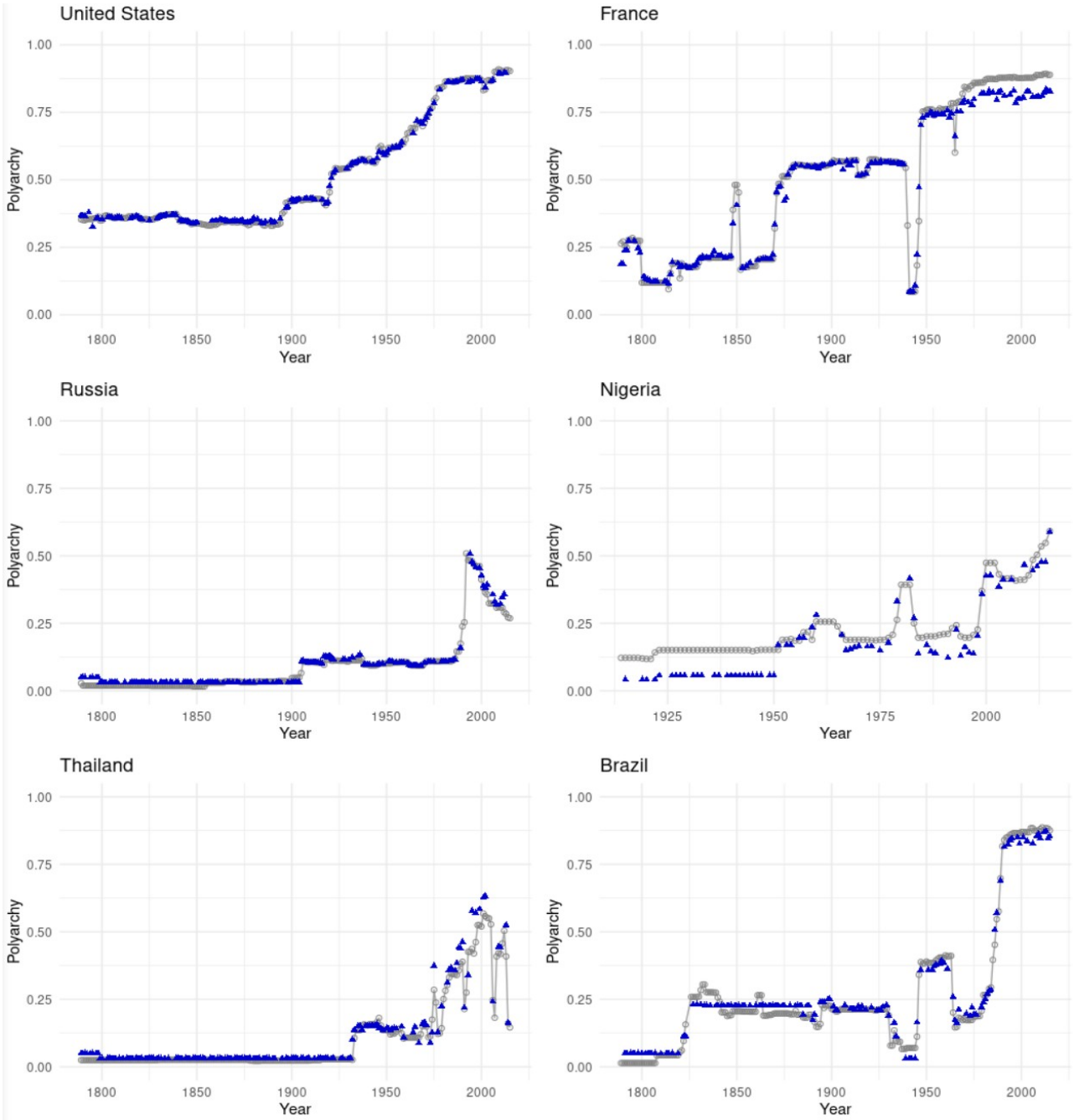
Figure 2: **Actual vs. Predicted Polyarchy Scores**



Predicted vs actual Polyarchy scores for all observations in our training dataset. The line indicates a perfect match between scores. The further points are from the line the more are they under- or overpredicted. Labels shown for selected country-years. Predictions in the validation data set yield similar performance. (The test set has not yet been deployed for model performance evaluation.)

To further assess the model's performance, we generate country-year plots of predicted and actual values for all countries in the V-Dem sample (available upon request). For illustrative purposes, Figure 3 presents graphs for a subset of six countries that reflect a variety of political systems and histories. OSM performance is impressive, judging by the overlap between circles (representing Polyarchy scores) and triangles (representing OSM predictions). For most country-years these symbols are virtually indistinguishable. In Nigeria, the random forest frequently underpredicts Polyarchy even though the trend-lines are highly correlated.

Figure 3: **Actual and Predicted Polyarchy Scores for Selected Countries**



Polyarchy scores (gray circles) and predicted scores based on the random forest (blue triangles). A complete set of country graphs is available upon request.

IV. Understanding Deviations

Although the fit between random forest predictions and actual Polyarchy scores is remarkably strong, it is important to understand the remaining deviations. To assess this issue, we calculate the difference between the original Polyarchy scores and our OSM estimate of those values, operationalized as the natural logarithm of the absolute difference. We then regress this outcome against factors that plausibly influence deviations. Table 2 adopts a cross-sectional format. (A fixed-effect format, appropriate for right-side variables that are not static, shows similar results.)

Model 1 includes characteristics of countries that may be regarded as exogenous (or nearly so) relative to democracy. We find that larger and richer countries are associated with smaller deviations. This could be because smaller and poorer countries are less well-understood by expert coders and/or because observable data is scarcer or more error-prone.

Other predictors – per capita GDP growth, Protestantism, Islam, English legal origin, and year – are not associated (or are only very weakly associated) with deviation. Importantly, the estimated coefficient for year is almost exactly zero, suggesting that there is no attenuation in the OSM's ability to predict Polyarchy as one moves back in time.

Model 2 adds variables that measure elements of democracy or features that are likely to be endogenous to democracy. We find that the degree of missingness among our chosen set of observable indicators (the inputs to the OSM) is associated with greater error, as one might expect.

The Polyarchy score itself is not associated with error, which is reassuring. However, year-to-year variability in Polyarchy is associated with greater deviations. This may be related to the fact that most of the observable features of democracy that inform the OSM occur during elections; in between elections we have much less information about the status of regimes.

The standard deviation across expert codings of Polyarchy (for a given country-year) is also associated with greater error. This demonstrates that we have a harder time replicating scores for Polyarchy where the V-Dem experts are themselves in disagreement.

Finally, we find that there is less deviation between Polyarchy scores and the random forest model during turnover periods, presumably because the model has more information about the status of democracy during those periods.

Overall, the patterns in Table 2 are consistent with our priors. An OSM will have greater difficulty replicating an index where there is greater uncertainty or less (observable) information about the outcome.

Importantly, neither of these models explains very much of the variance in predictive errors, judging by the low R squares. The remaining deviations may be largely stochastic. If the OSM is less subject to coder biases, an issue taken up in Section VI, this may also account for some of the deviation between Z and $Z´$.

Table 2: **Modeling the Deviations**

	1	**2**
Population (log)	-0.077***	-0.054***
	(-6.108)	(-4.232)
GDP per capita (log)	-0.115***	-0.045*
	(-4.439)	(-1.756)
GDP per capita (log), first difference	-0.277	-0.145
	(-1.046)	(-0.562)
Protestant	-0.002*	-0.001
	(-1.942)	(-0.839)
Muslim	0.000	0.000
	(0.294)	(0.339)
English legal origin	-0.079	-0.018
	(-1.485)	(-0.321)
Year	0.000	-0.000
	(1.118)	(-0.338)
Missing obs (%)		0.491***
		(7.238)
Polyarchy		0.072
		(0.326)
Polyarchy, first-difference, abs value		2.151***
		(9.673)
Polyarchy, first-difference, abs value, lagged		1.025***
		(5.005)
Polyarchy, standard deviation		5.201**
		(2.332)
Turnover period		-0.007***
		(-3.965)
Countries	184	171
Years	229	227
Observations	14,985	13,003
R-squared	0.0758	0.164

Outcome: absolute value of the Polyarchy score (Z) minus the random forest estimate ($Z´$), transformed by the natural log. *Estimator:* ordinary least squares, standard errors clustered by country, t statistics in parentheses. Intercept not shown. *** p<0.01, ** p<0.05, * p<0.1

V. Generalizing the Approach

The protocol described in Section II may be applied to any democracy index – or more broadly, to any subjective measure for which sufficient observable proxy data exists. In Appendix C we produce OSMs for four of the most

widely employed indices: UDS, Polity2, Freedom House, and BMR. For each index, we construct a random forest using our entire set of observable indicators of democracy. We whittle this set down to twelve or thirteen variables that explain most of the variability, based on their estimated importance. We then generate predictions for each index, in- and out-of-sample.

Table 3 reports the accuracy for each of these (in-sample) exercises – along with the Polyarchy index from Section II – assessed through the normalized root mean square error. We find that OSM models are more successful in replicating indices based on interval scales or ordinal scales with many levels (mimicking interval scales). They are somewhat less successful with the binary scale adopted by BMR.

Even so, random forest models based on observables explain most of the variability across all of these indices, suggesting that our approach is generalizable across the broad – and perpetually growing – field of democracy indicators.

Table 3: **Model-fit Across a Set of Democracy Indices**

			Normalized root mean square error	
Index	*Range*	*Scale*	*Full OSM*	*Reduced OSM*
Polyarchy	0 to 1	Interval	0.05	0.06
UDS	-2 to 2	Interval	0.05	0.05
Polity2	-10 to 10	Ordinal	0.06	0.06
Freedom House	1 to 15	Ordinal	0.07	0.08
BMR	0/1	Dichotomous	0.11	0.13

Goodness of fit statistics for five democracy indices as predicted by each OSM. Measures are calculated on out-of-bag training samples. *Full OSM:* all 40 variables (see Appendix A). *Reduced OSM:* the 12-13 most important variables for that particular outcome.

VI. Evaluating Potential Biases

In Section I, we reviewed ways in which the subjective coding of democracy might be biased. This is not an easy matter to assess empirically. However, an approach to measurement based on observables offers a benchmark against which potential coder biases may be identified. Specifically, if scores from an index (Z) are consistently higher (lower) than predictions from the OSM for that index (Z'), and if the difference is associated with background factors of no apparent relevance to the quality of democracy, a *prima facie* case for bias exists.

As an example, we interrogate potential biases cued by the political ideology of governments. Political scientists, like most academics, lean to the left (Cardiff, Klein 2005). Since political scientists are primarily responsible for producing measures of democracy it would not be too surprising if their ideological predilections affected their views of democracy and hence the indices that they generate (working as project

directors or as coders). Thus, we hypothesize that most subjective indices lean to the left. By contrast, one coding project – conducted by Freedom House – is alleged to hold more conservative views closely aligned with US interests, at least during the Cold War period (Bush 2017; Giannone 2010; Steiner 2016). Accordingly, we hypothesize that the Freedom House index leans to the right.

To examine this question, we enlist a new dataset (Herre 2022) that measures the ideology of heads of government from 1945 to 2020. Heads of government are classified as leftist, centrist, rightist, or non-ideological depending upon their attitudes towards redistributive state interventions into the economy. We employ a dummy for those classified on the right.

In Table 4, we regress the four indices of democracy (re-scaled from 0-1) against this dummy variable along with the OSM prediction for each index (intended to capture the observable features of that index). Each model also includes background covariates measuring per capita GDP (log), the first-difference of per capita GDP (log), and country and year dummies (two-way fixed effects).

Results from these analyses confirm our hunches. Right-wing governments are associated with lower scores for Polyarchy, UDS, and Polity2, suggesting that coders informing these projects might be influenced by the ideological complexion of the country they are coding. Freedom House appears to register a slight right-wing bias prior to 1990, though it does not surpass standard thresholds of statistical significance (perhaps

because the sample is considerably smaller). The fact that UDS registers a somewhat weaker left-wing bias than Polyarchy and Polity2 may reflect its composite nature; specifically, components of the UDS with a left-wing bias such as Polity2 are balanced by the right-wing bias of Freedom House.

Two caveats must be added to this set of findings. First, we do not find similar patterns when testing right- and left-wing heads of state ("leaders" in the Herre dataset), perhaps because their role is often centered on foreign policy or is largely symbolic.

Second, we must entertain the possibility that right-wing heads of government are bad for democracy in ways that are not reflected in observable measures, and thus deserve lower scores. For example, it is possible that right-wing leaders are especially hostile to the press and to free speech more generally, in which case the patterns apparent in Table 4 may be the product of an unmeasured confounder – civil liberties – rather than coder bias.

Despite these qualifications, we have demonstrated the utility of our approach for identifying *potential* biases, an approach that might be adapted to test other biases such as those discussed in Section I.

Importantly, when this methodology is reversed – when the estimate drawn from the OSM model is on the left side of the model – the head of government's ideology is no longer a statistically significant predictor. (This remains true whether the original index of democracy is included or excluded from the right side of the model.) This test offers some assurance

that OSM estimates are resistant to systematic coder biases. Appendix I describes the application of our method to datasets into which we have injected large, simulated, biases. We show that our method is largely robust even in the presence of improbably high systematic bias in the target measure, even when such bias is correlated with both predictors and outcomes.

Table 4: **Potential Ideological Bias**

Democracy index	Polyarchy	UDS	Polity2	Freedom House	Freedom House (-1990)
	1	**2**	**3**	**4**	**5**
Right-wing head of government	-0.006**	-0.004*	-0.012**	0.002	0.008
	(-2.216)	(-1.800)	(-2.469)	(0.542)	(1.309)
Countries	176	176	169	175	149
Years	75	67	74	47	17
Observations	8,317	7,295	7,804	6,015	1,985
R-squared	0.931	0.894	0.904	0.906	0.826

Covariates: OSM prediction, per capita GDP, per capita GDP first-difference, country dummies, year dummies, intercept. *Estimator:* ordinary least squares, standard errors clustered by country, t statistics in parentheses. *** p<0.01, ** p<0.05, * p<0.1

VII. Expanding the Universe of Cases

A key advantage of an OSM approach is that the usual sample of cases can be expanded, providing something close to a census of all sovereign and semi-sovereign polities in the world. Recall that determining the level of democracy in a polity through the usual procedures requires in-depth knowledge and expertise. This is plentiful for well-studied countries but often absent for less-studied cases. It is easy, for example, to find experts to

judge the quality of democracy in India but much harder to find qualified experts for São Tomé (today) or Bavaria (in the nineteenth century).

By contrast, collecting observable features of democracy is fairly straightforward and requires a low resource investment. (The notable exception is a handful of cases where elections occurred but there is no record of their results.) Accordingly, we can generate in-sample and out-of-sample democracy scores for 348 sovereign and semi-sovereign states. These states are observed over any period(s) of time during which they enjoyed a minimal degree of sovereignty, beginning in 1789 and ending in 2021 (the last year in our sample). Our full dataset provides estimates of democracy for 48,448 country-years. This may be contrasted with 25,759 observations covered by the Polyarchy index and considerably fewer observations for all other extant indices (see Table 1).

To be sure, we do not know how reliable the out-of-sample estimates are. Recall that although we test our random forest predictions with a validation set, the validation set is drawn from the population of the original index. It is possible that once one moves outside that population, the OSM is less successful in producing (synthetic) Polyarchy scores. In particular, one might worry that smaller countries, poorer countries, semisovereign entities, and historical cases are different in some unmeasurable fashion from the cases that predominate among extant indices. Indeed, Table 2

shows that smaller population and lower per capita GDP are associated with larger errors.

Although we do not have a foolproof method for testing the validity of estimates falling outside the population of an original index, we believe that out-of-sample estimates from our random forest model offer a substantial improvement over a status quo in which all of this potentially valuable information is simply ignored. Indeed, the more "different" the out-of-sample cases are from the observed cases, the more we ought to be concerned about sample bias. Analogies to the problem of missing data, and the potential solution provided by missing-data algorithms, are apt (Little, Rubin 2019).

Coverage: An illustrative analysis

Assuming that out-of-sample predictions are reliable (even if not precise), what can be learned from them? How much might this extension of coverage affect our understanding of the causes and effects of democracy?

For an illustrative example, we focus on the time-honored question of geography's impact on regime type. Since Montesquieu, geography has been considered a factor in conditioning a polity's democratic prospects. Among the many factors that have been proposed, we focus on two that are easy to measure and well-established in the literature: *islands* and *equatorial distance*.

Many writers regard island status as a force in favor of democratic outcomes in the modern era (Anckar 2008; Srebrnik 2004). First, island states are exposed to oceans and this may influence the propensity of a state to democratize. Second, islands offered appealing ports of call and colonies of settlement for Europeans, including Britishers and Protestants, and they were often subjected to an extensive tutelary relationship with a European power, culminating in many years' experience with electoral politics and semi-autonomous governance prior to independence. For a variety of reasons, one may suppose that the colonial experience was more transformative for island-states than for other states.

Third, most islands depend upon international trade or tourism for a large share of their national income. This may encourage a more open attitude toward democracy. Fourth, islands tend to be small, limiting the population. And with natural borders provided by the sea island living may foster a greater sense of national community than one finds in land-based states. These features are often regarded as conducive to democracy. Finally, being geographically isolated, island-states may be less militarist because their sovereignty is more secure than land-based states and because expansionist policies are more difficult to pursue.

Distance from the equator is also commonly regarded as a factor conducive to democracy. First, equatorial distance is correlated with economic development (Easterly, Levine 2003); insofar as the latter is a cause of democracy (or democratic consolidation), geography is an

antecedent cause. Second, tropical climates affect the epidemiological environment, fostering malaria and many other communicable diseases, which limit human capital and economic productivity at large. Third, for this reason, Europeans were less likely to settle in large numbers, which may, in turn, have had important repercussions for the sort of regimes that developed in the modern world (Gerring et al. 2022: part III). Fourth, tropical climates are also conducive to plantation agriculture, which served as a spur to slavery and other coercive labor systems. This, in turn, fostered vast inequality in landholding and wealth, and extractive institutions in subsequent centuries (Engerman, Sokoloff 2012).

In summary, there are plenty of reasons to regard islands and equatorial distance as important influences on regime type, and empirical results seem to support this view. However, work on these subjects relies on extant indices of democracy with limited coverage. What happens when we expand the usual scope of cases?

In Table 5, the outcome of interest is the Polyarchy index, which we interrogate in three tests. The first incorporates the original index. The second employs the OSM in-sample estimate. The third test employs the OSM estimate for all available cases, in-sample and out-of-sample (though limited by the availability of coverage for right-side covariates).

A linear trend variable (Year) and a panel of region dummies are included in all analyses in order to mitigate potential confounders associated with time and spatial location. (Results are robust when these

background factors are removed.) Models 4-9 introduce a sovereignty variable, measuring whether a state is fully sovereign or a colony/dependency. (Since sovereignty may be downstream from geography we do not include this factor in Models 1-3.) Models 7-9 are limited to the contemporary era.

Across the original index and the OSM estimate there are minimal differences, as one might expect given how highly correlated they are. Island and Equator distance matter quite a lot, corroborating the conventional finding. When the sample is expanded, however, there are appreciable differences. Specifically, island and equator distance are much stronger predictors of democracy in the restricted (V-Dem) sample than in the full sample. Indeed, full sample estimates for island and equator distance are less than half the size of estimates based on the restricted V-Dem sample.

This does not mean that these geographic factors can be discarded; after all, most of the estimates are statistically significant in the predicted direction. However, they may play less of a role than we had thought.

In any case, our purpose is not to make strong causal claims. It is, rather, to show that sample size – and potential bias – matters. Presumably, this holds for other variables of theoretical interest. In this respect, OSMs promise to expand our leverage on important research questions.

Table 5: **Estimated Impact of Geography on Democracy in Varying Samples**

Years	1789-2021			1789-2021			1946-2021		
Outcome	*Polyarchy*	*OSM (in-sample)*	*OSM (full-sample)*	*Polyarchy*	*OSM (in-sample)*	*OSM (full-sample)*	*Polyarchy*	*OSM (in-sample)*	*OSM (full-sample)*
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Island	0.092***	0.081***	-0.022	0.112***	0.101***	0.051***	0.151***	0.141***	0.071***
	(2.863)	(2.744)	(-1.460)	(3.332)	(3.260)	(3.223)	(3.884)	(3.789)	(2.629)
Equator	0.005***	0.005***	0.002**	0.005***	0.004***	0.002***	0.006***	0.006***	0.003**
	(4.943)	(4.898)	(2.427)	(4.688)	(4.633)	(2.851)	(5.290)	(5.311)	(2.485)
distance									
Sovereign				✓	✓	✓	✓	✓	✓
Countries	192	192	347	192	192	347	180	180	244
Years	232	232	232	232	232	232	75	75	75
Observations	20,020	20,020	43,398	20,020	20,020	43,398	9,764	9,764	13,939
R-squared	0.512	0.521	0.429	0.531	0.542	0.505	0.466	0.476	0.411

Additional covariates: year, region dummies (Europe, Americas, MENA, sub-Saharan Africa, Asia), intercept. *Estimator:* Ordinary least squares, standard errors clustered by country, t statistics in parentheses. *** p<0.01, ** p<0.05, * p<0.1. (The OSM samples will be 15% larger when the test-set is included, in the final version.)

VIII. Discussion

Many decisions are required when one composes an index for a complex, latent concept such as democracy. At the very least, one must define the concept, measure its components, and aggregate the resulting indicators (if more than one). The approach introduced in this study offers an objective strategy for measurement while side-stepping questions of conceptualization and aggregation.

There is, to be sure, a cost, which can be represented formally in a simple model:

$$Z = Z' + \varepsilon$$

where Z = a subjective index, Z' = an OSM estimate, and ε = error. The tricky aspect of this equation is that the error term encapsulates both coder

error *and* information loss, i.e., elements of the chosen concept of democracy that we have not found a way to measure with observables. Unfortunately, we have no easy way of distinguishing between error and information loss.

In the context of electoral democracy, we expect greater information loss *in between* elections, as elections provide most of the observable features of democracy. Sometimes, information loss affects countries unequally. For example, if civil liberty is missing in our index of observables, countries that offer greater protection for civil liberty than the global mean will receive a score on our index that is too low; and vice-versa for countries offering a level of civil liberty that is lower than the global mean.

Our commonsense conclusion is that an OSM arrived at in the fashion outlined in this paper is correctly regarded as superior to the original (subjective) version in some respects (mitigated coder error and overall coverage) and inferior in others (loss of information). It could be that researchers conducting crossnational analyses with democracy will want to use both versions – one as a benchmark and the other as a robustness check. OSMs are also useful tools for interrogating potential biases in subjectively coded measures, as we demonstrate in Section VI. Clearly, an OSM will never suffice as a wholesale replacement of the original index; indeed, an OSM cannot exist without a subjective index to mimic.

Before concluding we also want to guard against a potential misunderstanding. In describing our index as free from coder bias we do not mean to suggest it is free of all bias. After all, bias can take many forms. There may be biases associated with the observable indicators we have chosen to represent the concept of democracy. Here, "bias" is understood by reference to some (unbiased) concept of democracy. One must therefore consider whether observable aspects of democracy bias the measurement of democracy in a particular way, and in what direction the bias might run. An OSM also has the potential to "learn" biases in the target measure when those biases are correlated with objective predictors. Nonetheless, as Appendix *I* shows, the method that we demonstrate here is quite robust, even in the face of large biases that are correlated with both predictors and target outcomes. Given relevant measures, one can also investigate potential biases, as we demonstrate in section VI.

There may also be biases associated with an OSM when it is used as a predictor or an outcome in a causal model. For example, if one is using an OSM to test the relationship between democracy and growth and the OSM features a measure of turnover, one must be cognizant that poor growth performance may enhance turnover, introducing endogeneity between the left and right sides of a causal model (Knutsen, Wig 2015). This may be handled by introducing lags of the dependent variable, by lagging the predictor several periods prior to the outcome, or by reconstructing the OSM without turnover. If one is particularly concerned about the issue, all

three approaches may be employed, providing an extensive set of robustness tests. The general point is that lack of bias in data collection does not mean that the resulting index is free of bias in the context of a causal analysis. It may, or it may not be.

Evidently, the more ingredients are included in an OSM the greater the prospect for circularity between a predictor and an outcome. That is why we opt for a parsimonious selection of variables, excluding those that add little predictive power (but might confound causal analyses). In some circumstances, researchers will want to construct even more parsimonious OSMs, shorn of any variable that might be subject to endogeneity.

In any case, the same problem besets subjectively coded indices. The difference is that it is difficult to tell when a problem of endogeneity exists and when it can be ignored. Consider the situation of a country undergoing a civil conflict. Experts enlisted to code the quality of elections may assume that it is lower during times of conflict – a reasonable assumption, especially if the conduct of elections is not directly observable for a particular year. There is no way to purge the index of these sorts of assumptions, which are innumerable and sometimes not even apparent, even to those doing the coding. By contrast, with an index based on observables we know precisely which factors contribute to a country's score in each year. Moreover, we can purge the index of any indicator that poses a potential problem of interpretation (with some informational costs, depending upon the indicator).

IX. References

Alvarez, Mike, Jose A. Cheibub, Fernando Limongi, Adam Przeworski. 1996. “Classifying Political Regimes.” *Studies in Comparative International Development* 31(2): 3–36.

Anckar, Carsten. 2008. “Size, Islandness, and Democracy: A Global Comparison.” *International Political Science Review* 29(4): 440-441.

Bakker, Ryan, Catherine De Vries, Erica Edwards, Liesbet Hooghe, Seth Jolly, Gary Marks, Jonathan Polk, Jan Rovny, Marco Steenbergen, Milada Anna Vachudova. 2015. "Measuring party positions in Europe: The Chapel Hill expert survey trend file, 1999–2010." *Party Politics* 211): 143-152.

Bjørnskov, Christian, Martin Rode. 2020. "Regime types and regime change: A new dataset on democracy, coups, and political institutions." *Review of International Organizations* 15, 531-51.

Boix, Carles, Michael Miller, Sebastian Rosato. 2013. “A Complete Dataset of Political Regimes, 1800-2007.” *Comparative Political Studies* 46(12): 1523-1554.

Bollen, Kenneth. 1990. “Political democracy: Conceptual and measurement traps.” *Studies in Comparative International Development* 25(1): 7–24.

Bollen, Kenneth, Pamela Paxton. 2000. “Subjective measures of political democracy.” *Comparative Political Studies* 33(1): 58–86.

Bowman, Kirk, Fabrice Lehoucq, James Mahoney. 2005. "Measuring political democracy: Case expertise, data adequacy, and Central America." *Comparative Political Studies* 38(8): 939-970.

Breiman, Leo. 2001. "Random forests." *Machine Learning* 45(1): 5-32.

Bühlmann, Marc, Wolfgang Merkel, Lisa Müller, Bernhard Weßels. 2012. "The Democracy Barometer: A New Instrument to Measure the Quality of Democracy and Its Potential for Comparative Research." *European Political Science* 11(4): 519-536.

Bush, Sarah Sunn. 2017. "The politics of rating freedom: Ideological affinity, private authority, and the freedom in the world ratings." *Perspectives on Politics* 15(3): 711–731.

Cardiff, Christopher F., Daniel B. Klein. 2005. "Faculty partisan affiliations in all disciplines: A voter-registration study." *Critical Review* 17(3-4): 237-255.

Cheibub, Jose Antonio, Jennifer Gandhi, James Raymond Vreeland. 2010. "Democracy and Dictatorship Revisited." *Public Choice* 143(1–2): 67–101.

Coppedge, Michael, Angel Alvarez, Claudia Maldonado. 2008. "Two persistent dimensions of democracy: Contestation and inclusiveness." *The Journal of Politics* 70(3): 632-647.

Coppedge, Michael, John Gerring, Adam Glynn, Carl Henrik Knutsen, Staffan I. Lindberg, Daniel Pemstein, Brigitte Seim, Svend-Erik Skaaning, Jan Teorell. 2020. *Varieties of Democracy: Measuring a Century of Political Change*. Cambridge: Cambridge University Press.

Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Jan Teorell, David Altman, Michael Bernhard, Agnes Cornell, M. Steven Fish, Lisa Gastaldi, Haakon Gjerløw, Adam Glynn, Allen Hicken, Anna Lührmann, Seraphine F. Maerz, Kyle L. Marquardt, Kelly McMann, Valeriya

Mechkova, Pamela Paxton, Daniel Pemstein, Johannes von Römer, Brigitte Seim, Rachel Sigman, Svend-Erik Skaaning, Jeffrey Staton, Aksel Sundtröm, Eitan Tzelgov, Luca Uberti, Yi-ting Wang, Tore Wig, and Daniel Ziblatt. 2021. *V-Dem Codebook v11.1* Varieties of Democracy (V-Dem) Project.

Cutright, Phillips. 1963. “National political development. Measurement and analysis.” *American Sociological Review* 28(2): 253-264.

Easterly, William, Ross Levine. 2003. "Tropics, germs, and crops: how endowments influence economic development." *Journal of Monetary Economics* 50(1): 3-39.

Elkins, Zachary. 2000. “Gradations of democracy? Empirical tests of alternative conceptualizations.” *American Journal of Political Science* 44(2): 293–300.

Engerman, Stanley L., Kenneth L. Sokoloff. 2012. *Economic Development in the Americas since 1500: Endowments and Institutions*. Cambridge: Cambridge University Press.

Freedom House. 2015. *Methodology: Freedom in the World 2015*. New York. (https://freedomhouse.org/sites/default/files/Methodology_FIW_2015.pdf), accessed December 2, 2015.

Gerring, John, Brendan Apfeld, Tore Wig, Andreas Tollefsen. 2022. *The Deep Roots of Modern Democracy: Geography and the Diffusion of Political Institutions*. Cambridge: Cambridge University Press.

Giannone, Diego. 2010. “Political and ideological aspects in the measurement of democracy: The Freedom House case.” *Democratization* 17(1): 68–97.

Gründler, Klaus, Tommy Krieger. 2016. "Democracy and growth: Evidence from a machine learning indicator." *European Journal of Political Economy* 45: 85-107.

Gründler, Klaus, Tommy Krieger. 2021. "Using Machine Learning for measuring democracy: A practitioners guide and a new updated dataset for 186 countries from 1919 to 2019." *European Journal of Political Economy* 70: 102047.

Hastie, Trevor, Robert Tibshirani, Jerome Friedman. 2013. *The elements of statistical learning*. New York: Springer.

Herre, Bastian. 2022. “Identifying Ideologues: A Global Dataset on Political Leaders, 1945-2020.” *British Journal of Political Science* (forthcoming).

Hill, Daniel W., and Zachary M. Jones. 2014. "An Empirical Evaluation of Explanations for State Repression." American Political Science Review 108 (3): 661-687.

Knutsen, Carl Henrik, Tore Wig. 2015. "Government turnover and the effects of regime type: How requiring alternation in power biases against the estimated economic benefits of democracy." *Comparative Political Studies* 48(7): 882-914.

Laver, Michael, Kenneth Benoit, J. Garry. 2003. “Extracting policy positions from political texts using words as data.” *American Political Science Review* 97(2):311-331.

Little, Roderick JA, Donald B. Rubin. 2019. *Statistical analysis with missing data*. New York: John Wiley & Sons.

Marquardt, Kyle L., Daniel Pemstein. 2018. “IRT models for expert-coded panel data.” *Political Analysis* 26(4): 431–456.

Marquardt, Kyle L., Daniel Pemstein, Brigitte Seim, Yi-ting Wang. 2019. “What makes experts reliable? Expert reliability and the estimation of latent traits.” *Research & Politics*: 10.1177/2053168019879561

Marshall, Monty G. 2020. *POLITY5 Political Regime Characteristics and Transitions, 1800-2018 Dataset Users’ Manual*. Center for Systemic Peace and Societal-Systems Research. (www.systemicpeace.org/inscr/p5manualv2018.pdf)

Marshall, Monty G., Ted Gurr, Keith Jaggers. 2013. *Polity IV Project: Political Regime Characteristics and Transitions, 1800–2012, Dataset Users' Manual*. Center for Systemic Peace, Vienna, VA.

Norris, Pippa, Richard W. Frank, Ferran Martinez i Coma. 2013. "Assessing the quality of elections." *Journal of Democracy* 24(4): 124-135.

Pemstein, Daniel, Stephen Meserve, James Melton. 2010. “Democratic Compromise: A Latent Variable Analysis of Ten Measures of Regime Type.” *Political Analysis* 18(4): 426–449.

Schedler, Andreas. 2012. “Judgment and measurement in political science.” *Perspectives on Politics* 10(1), 21–36.

Skaaning, Svend-Erik. 2018. "Different Types of Data and the Validity of Democracy Measures." *Politics and Governance* 6(1): 105-116.

Skaaning, Svend-Erik, John Gerring, Henrikas Bartusevičius. 2015. “A Lexical Index of Electoral Democracy.” *Comparative Political Studies* 48(12): 1491-1525.

Srebrnik, Henry. 2004. “Small Island Nations and Democratic Values.” *World Development* 32(2), 329-41.

Steiner, Nils D. 2016. “Comparing Freedom House Democracy Scores to Alternative Indices and Testing for Political Bias: Are U.S. Allies Rated as More Democratic by Freedom House?” *Journal of Comparative Policy Analysis* 18(4): 329–49.

Teorell, Jan, Michael Coppedge, Staffan Lindberg, Svend-Erik Skaaning. 2019. "Measuring polyarchy across the globe, 1900–2017." *Studies in Comparative International Development* 54(1): 71-95.

Vanhanen, Tatu. 2000. "A new dataset for measuring democracy, 1810-1998." *Journal of Peace Research* 37(2): 251-265.

Vanhanen, Tatu. 2011. "Measures of democracy 1810–2010." *FSD1289, version* 5. www.fsd.tuni.fi/fi/aineistot/taustatietoa/FSD1289/Introduction_2010.pdf

Appendix A: Codebook

Democracy Indices

Polyarchy. Electoral democracy index. *Source:* V-Dem (Coppedge et al. 2018; Teorell et al. 2016). *Scale:* interval. *v2x_polyarchy*

Freedom House. Combines the Polity rights and Civil liberties indices into an additive single index. Political rights enable people to participate freely in the political process, including the right to vote freely for distinct alternatives in legitimate elections, compete for public office, join political parties and organizations, and elect representatives who have a decisive impact on public policies and are accountable to the electorate. The specific list of rights considered varies over the years. Civil liberties include freedoms of expression, assembly, association, education, and religion; an established and generally fair legal system that ensures the rule of law (including an independent judiciary), allows free economic activity, and tends to strive for equality of opportunity for everyone, including women and minority groups. *Source:* Freedom House (2018). *Scale:* ordinal. *e_fh_combined*

Polity2. Computed by subtracting the autocracy score from the democracy score. The resulting unified POLITY scale ranges from +10 (strongly democratic) to -10 (strongly autocratic). *Source:* Polity V (Marshall 2020). *Scale:* ordinal. *e_polity2*

BMR. Dichotomous democracy measure based on contestation and participation. Countries coded democratic have (1) political leaders that are chosen through free and fair elections and (2) a minimal level of suffrage. *Source:* Boix, Miller, Rosato (2013). *Scale:* Dichotomous. *e_boix_regime*

UDS. Democracy score posterior (mean). *Source:* Pemstein et al. (2010). *Scale:* Interval. *e_uds_mean.*

Variables in the Full OSM

Principal data sources: Nohlen (2005), Nohlen, Grotz, Harmann (2002), Nohlen, Krennerich, Thibaut (1999), Nohlen, Stover (2010), Chronicle of Parliamentary Elections (IPU), Wikipedia entries focused on particular elections, PIPE (Przeworski 2013), Skaaning et al. (2015).

Difference Vote Share, two largest parties. Difference in the share of votes received by the largest and the second largest party in the first (or only) round of the election to the lower (or unicameral) chamber of the legislature. *Scale:* Interval. *top2_difference*

Electoral Regime Index. Coded 1 if regularly scheduled national elections are on course, as stipulated by election law or well-established precedent. *Source:* V-Dem 11 (Coppedge et al. 2021), with additional coding by authors. *Scale:* binary. *v2x_elecreg_JG*

Executive Elections. Are executive elections taking place? *Scale:* Dichotomous. *executive_elections*

Executive Elections, years. Years the executive has been elected. *Scale:* Interval. *years_exec_elec_continuous*

Female Suffrage, share. Share of enfranchised women of voting age. *Scale:* Interval. *female_suffrage*

Independent states. A state is considered to be an independent polity if it (a) has a relatively autonomous administration over some territory, (b) is considered a distinct entity by local actors or the state it is dependent on. *Scale:* dichotomous. *v2svindep*

Independents, legislature, share. Independents as share (%) of seats in lower or unicameral chamber of the national legislature. Independents defined as members who are not declared members of a political party. *Scale:* interval. *v2elindss*

Independents, votes, share. Votes won by independents as share (%) of total votes for lower or unicameral chamber of the national legislature. Independents defined as members who are not declared members of a political party. *Scale:* interval. *v2elindsv*

Largest party votes, presidential. Share (%) of votes received by the winning candidate in the first (or only) round of a presidential election. *Scale:* interval. *v2elvotlrg*

Legislative Elections. Are legislative elections taking place? Scale: Dichotomous. *legislative_elections.*

Legislative Seats, second largest party. In the last election: How many lower chamber election seats did the second largest party win? *Scale:* interval. *v2ellostsm*

Legislative Vote Share, largest party. Share of votes received by the largest party in the first (or only) round of the election to the lower (or unicameral) chamber of the legislature. *Scale:* Interval. *v2ellovtlg*

Length of HOS/HOG tenure, ln. Length of HOS or HOG tenure in office, transformed by the natural logarithm. *Scale:* interval. *hos_hog_tenure_ln*

Lower chamber election seat share, largest party. Share (%) of seats in the lower (or unicameral) chamber of the legislature obtained by the largest party. *Scale:* interval. *v2ellostsl*

Lower chamber election seat share, second largest party. Share (%) of seats in the lower (or unicameral) chamber of the legislature obtained by the second-largest party. *Scale:* interval.

Lower chamber election seat share, third largest party. Share (%) of seats in the lower (or unicameral) chamber of the legislature obtained by the third-largest party. *Scale:* interval.

Lower chamber election seats. Number of seats in the lower chamber. *Scale:* interval. *v2elloseat*

Lower chamber election seats, largest party. In this election to the lower (or unicameral) chamber of the legislature, how many seats were obtained by the largest party? *Scale:* interval. *v2ellostlg*

Lower chamber election seats, second largest party. *I*n this election, how many seats in the lower (or unicameral) chamber of the legislature were obtained by the next-largest party? *Scale:* interval. *v2ellostsm*

Lower chamber election seats, third largest party. In this election, how many seats in the lower (or unicameral) chamber of the legislature were obtained by the next-largest party? *Scale:* interval. *v2ellosttm*

Lower chamber election vote share, second-largest party. In this election to the lower (or unicameral) chamber of the legislature, what percentage (%) of the vote was received by the second largest party in the first/only round? *Scale:* interval. *v2ellovtsm*

Lower chamber election vote share, third-largest party. In this election, what percentage (%) of the total seats in the lower (or unicameral)? *Scale:* interval.

chamber of the legislature was obtained by the next-largest party? *Scale:* interval. *v2ellostts*

Male Suffrage, share. Share of enfranchised men of voting age. Scale: Interval. *male_suffrage*

Multi-party leg. Elections. Dummy variable indicating whether there were multi-party elections. *Scale:* dichotomous. *multi_party_leg_elec*

Number of Turnovers, ln. Number of electoral turnovers, logged**.** *Scale:* interval. *turnover_total_ln*

Presidential election vote share, second-largest party. In this presidential election, what percentage (%) of the vote was received by the second most successful candidate in the first round? *Scale:* interval. *v2elvotsml*

Seat Share, two largest parties. Share (%) of seats in the lower or unicameral house held by the top two parties in the last election. *Scale:* interval. *top2_seat_perc*

Sovereignty. A state is considered to be sovereign if it (a) has a relatively autonomous administration over some territory, (b) is considered a distinct entity by local actors or the state it is dependent on. This excludes colonies, states that have some form of limited autonomy (e.g. Scotland), are alleged to be independent but are contiguous to the dominant entity (Ukraine and Belarus prior to 1991), de facto independent polities but recognized by at most one other state (Turkish Republic of Northern Cyprus). Occupations or foreign rule are considered to be an actual loss of statehood when they extend beyond a decade. This means that cases such as the Baltic Republic during Soviet occupation are not considered independent states, but independent statehood is retained for European countries occupied during World War II. *Scale:* dichotomous. *Sources:* Gleditsch and Ward (1999), v2svindep variable from V-Dem 11 (Coppedge et al. 2021), with additional coding by authors. *sovereign_erik*

Suffrage, share. The share (%) of enfranchised adults older than the minimal voting age who are legally allowed to vote. *Sources:* Bilinski (2015) along with sources listed above. *Scale:* interval. *v2asuffrage*

Turnover Event. Indicator event for turnover in government. *Scale:* dichotomous. *turnover_event*

Turnover HOG, cumulative. Was there turnover in the office of the head of government (HOG) as a result of this national election? This variable counts the number of turnovers. Source(s): Henisz (2000; 2002); Lentz (1994; 1999);

worldstatesmen.org; V-Dem Country Coordinators. *Scale:* interval. *v2elturnhog_cum*

Turnover HOS, cumulative. Was there turnover in the office of the head of state (HOS) as a result of this national election? This variable counts the number of turnovers. *Sources:* Henisz (2000; 2002); Lentz (1994; 1999); worldstatesmen.org; V-Dem Country Coordinators. *Scale:* interval. *v2elturnhos_cum*

Turnover Period. Dummy variable indicating whether there was a turnover in an election. After the first turnover the variable takes the value 1 and remains 1 until multi-party elections for the executive and/or legislature are interrupted. *Scale:* dichotomous.

Turnover Period, cont. Count of years since first turnover. Resets at electoral interruptions. *Scale:* interval. *years_turnover_period_cont*

Two Turnover Period. Indicator variable for instances where at least two electoral turnovers happened. *Scale:* dichotomous. *two_turnover_period*

Vote Share Top2 combined >60%. Dummy variable indicating whether the top two parties in the lower house gain more than 2/3 of the votes. *Scale:* dichotomous. *top2_monopoly*

Vote Share, two largest parties. Combined sum of the share of votes received by the largest and the second largest party in the first (or only) round of the election to the lower (or unicameral) chamber of the legislature. *Scale:* interval. *top2_combined*

Years since turnover event. Count variable counting the years since a turnover event. *Scale:* interval. *years_turnover_event_yes*

Total number of independents. Total number of independents in the legislature. *Scale:* interval. *v2elinds*

Appendix B: Serial Omission

Table B-1: Serial Omission

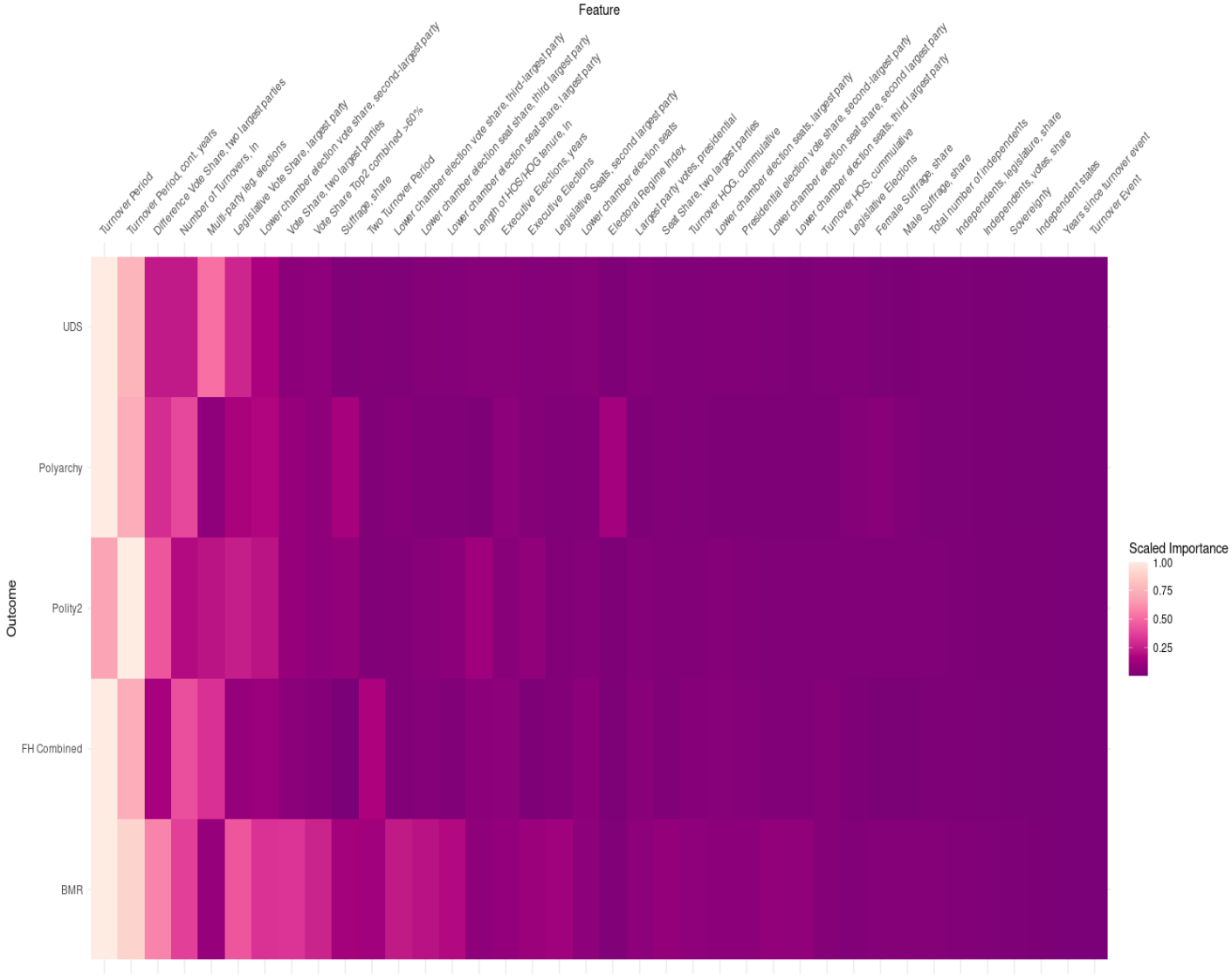
	Training			Validation			Cross-validation		
Excluded	*MSE*	*RMSE*	*R2*	*MSE*	*RMSE*	*R2*	*MSE*	*RMSE*	*R2*
turnover_period	0.0033	0.0574	0.9524	0.0032	0.0567	0.9544	0.0033	0.0577	0.9519
years_turnover_period_cont	0.0032	0.0567	0.9542	0.0029	0.0536	0.9572	0.0033	0.0576	0.9528
top2_difference	0.0034	0.0587	0.9500	0.0031	0.0561	0.9559	0.0035	0.0590	0.9496
top2_combined	0.0033	0.0577	0.9517	0.0031	0.0560	0.9537	0.0034	0.0584	0.9505
v2ellovtlg	0.0033	0.0570	0.9527	0.0035	0.0588	0.9494	0.0033	0.0578	0.9515
v2ellostsm	0.0036	0.0600	0.9479	0.0034	0.0586	0.9495	0.0037	0.0606	0.9468
v2x_suffr	0.0038	0.0616	0.9456	0.0039	0.0621	0.9431	0.0039	0.0621	0.9448
v2x_elecreg_jg	0.0035	0.0591	0.9495	0.0032	0.0568	0.9518	0.0035	0.0596	0.9487
top2_monopoly	0.0031	0.0558	0.9551	0.0031	0.0557	0.9536	0.0031	0.0559	0.9550
turnover_total_ln	0.0036	0.0601	0.9480	0.0036	0.0597	0.9471	0.0037	0.0607	0.9470
multi_party_leg_elec	0.0037	0.0609	0.9462	0.0039	0.0622	0.9435	0.0037	0.0610	0.9461
years_exec_elec_continuous	0.0035	0.0592	0.9484	0.0037	0.0612	0.9470	0.0036	0.0601	0.9469
female_suffrage	0.0034	0.0585	0.9510	0.0030	0.0551	0.9568	0.0034	0.0585	0.9504

Appendix C: Goodness of Fit of Full and Reduced Models

Table C-1: **Goodness of Fit**

		Training			Validation			Cross-validation		
Measure	*OSM*	*MSE*	*RMSE*	*R2*	*MSE*	*RMSE*	*R2*	*MSE*	*RMSE*	*R2*
Polyarchy	Full	0.002	0.05	0.97	0.002	0.05	0.97	0.002	0.05	0.97
	Reduced	0.003	0.06	0.95	0.003	0.06	0.95	0.003	0.06	0.95
Freedom House	Full	0.94	0.97	0.94	0.87	0.93	0.95	0.97	0.99	0.94
	Reduced	1.2	1.1	0.93	1.1	1.1	0.93	1.2	1.1	0.93
Polity2	Full	4.33	2.11	0.91	4.39	2.09	0.91	4.56	2.13	0.91
	Reduced	5.91	2.43	0.88	6.03	2.46	0.88	6.10	2.47	0.88
UDS	Full	0.04	0.20	0.96	0.04	0.19	0.96	0.04	0.21	0.96
	Reduced	0.05	0.23	0.94	0.05	0.23	0.94	0.06	0.24	0.94
BMR	Full	0.01	0.11	0.94	0.01	0.12	0.95	0.01	0.12	0.94
	Reduced	0.02	0.13	0.92	0.02	0.13	0.93	0.02	0.13	0.92

Figure C-1: VIP Heatmap



Appendix D: Imputation

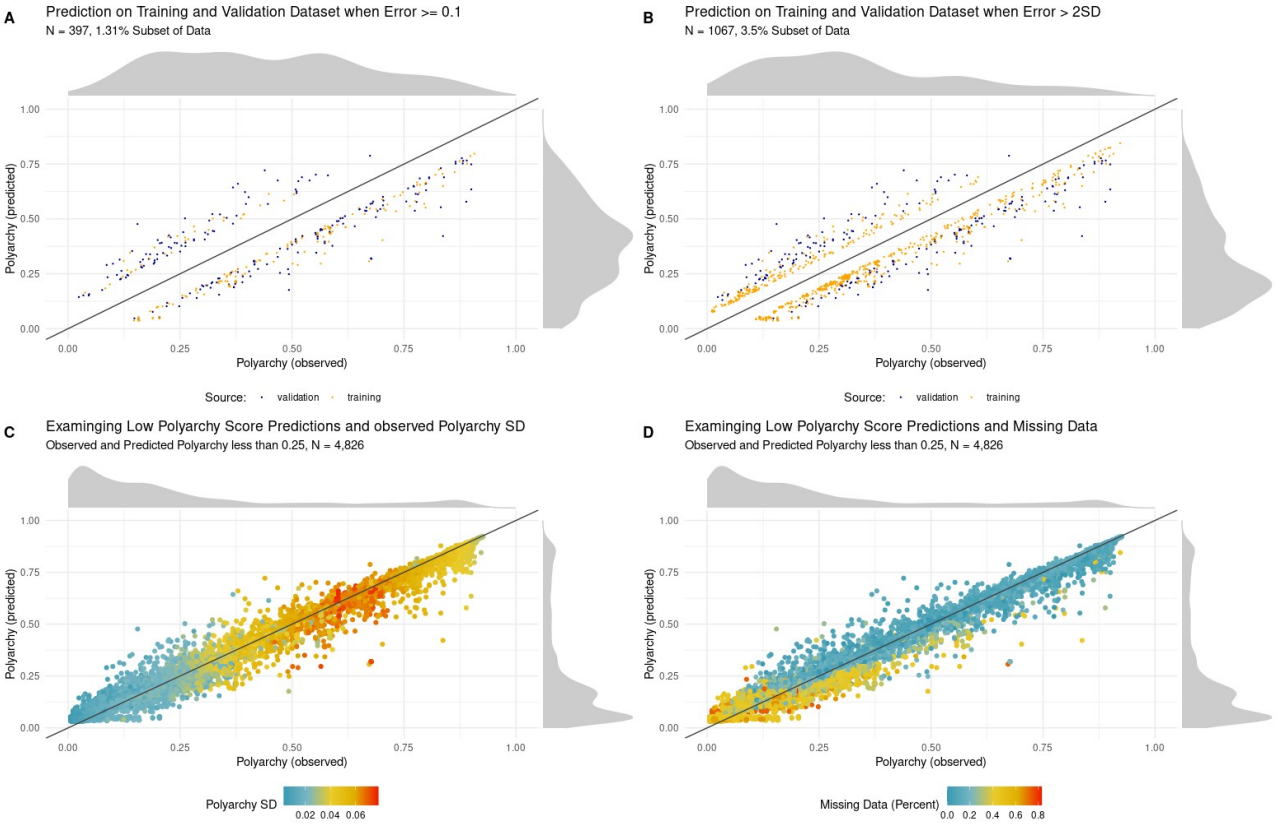
Table D-1: **Different Imputation Strategies**

	Training			Validation			Cross-Validation		
Method	*MSE*	*RMSE*	*R2*	*MSE*	*RMSE*	*R2*	*MSE*	*RMSE*	*R2*
KNN	0.003	0.058	0.950	0.004	0.059	0.950	0.004	0.059	0.950
Bagged Trees	0.003	0.055	0.957	0.003	0.055	0.956	0.003	0.055	0.956
MICE	0.005	0.067	0.934	0.005	0.067	0.935	0.005	0.067	0.935

Appendix E: Further Examining Predictive Performance

Panels A and B in Figure E-1 show that large differences between the observed and the predicted Polyarchy scores are rather rare for the training and the validation data sets. Only 397 observations (1.31% of the sample) have an error that is equal to or larger than ten percentage points of the Polyarchy scale and only 1,067 observations have an error that is larger than two times the standard deviation. The prediction on cross-validated training data and the validation dataset are performing remarkably well. Panel C and D also show that the coder disagreement (operationalized as the standard deviation of the original Polyarchy variable) as well as the missingness in the objective features we use for the prediction do not systematically relate to prediction error of our model. Larger Polyarchy scores have a larger standard deviation and they also appear to be associated with an overprediction of Polyarchy. (Panel C). Panel D shows that lower Polyarchy scores tend to have significantly more missingness in the objective features that we use to predict and that a higher level of missingness also tends to come with a tendency to overpredict Polyarchy values.

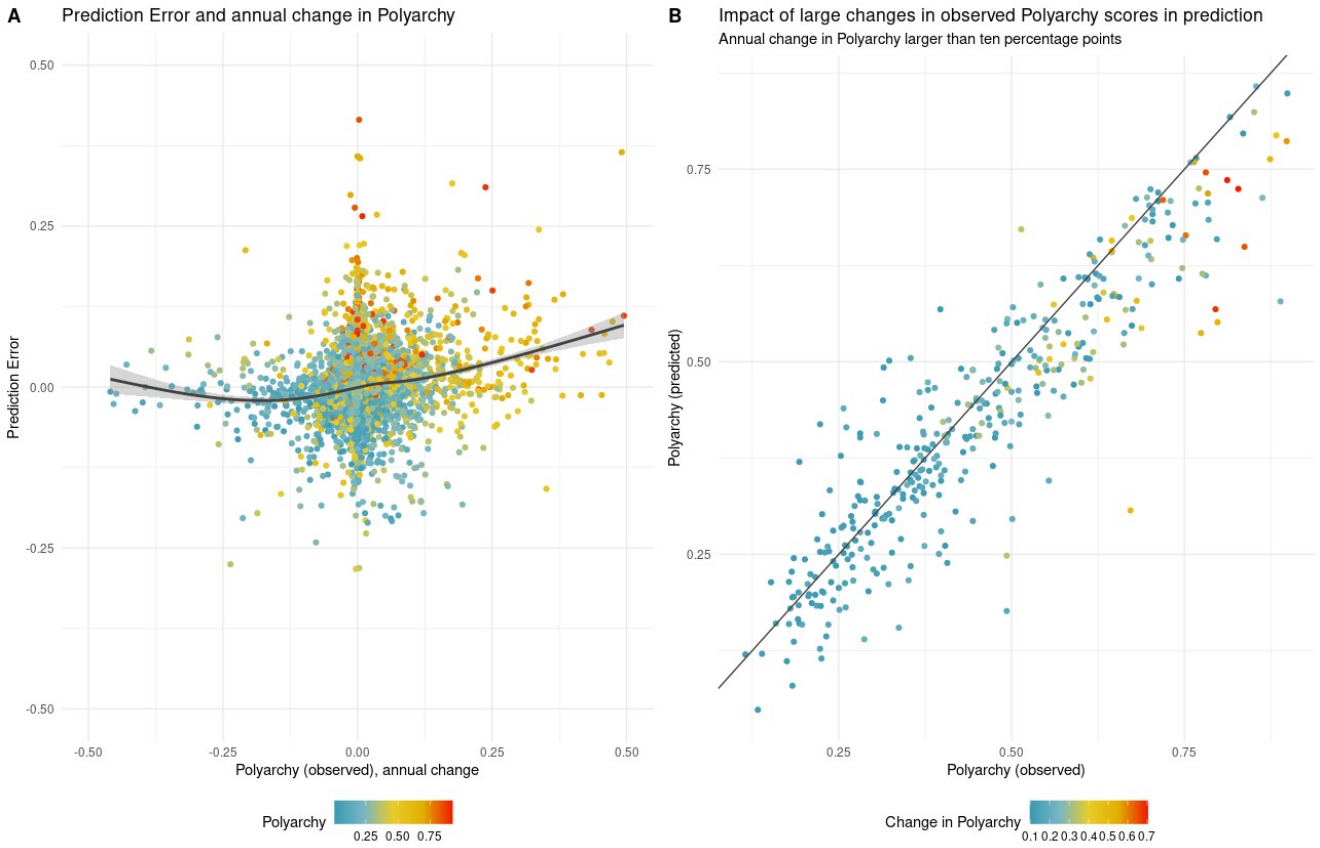
Figure E-1: **Predictive quality**



Panel A plots observed vs. predicted Polyarchy scores. Points are colored based on the standard deviation of the observed Polyarchy measure. Warmer colors indicate a higher standard deviation in the observed Polyarchy measure. Panel B also plots observed vs predicted Polyarchy scores. The color indicates missingness of the data used in the random forest model (such as election data). Warmer colors indicate more missingness in the underlying data. The remaining two panels plot observed vs predicted Polyarchy score based on the absolute error being larger than 0.1 (Panel C) or larger than 2SD (Panel D).

A further source of potential problems could be in rapid changes in democracy scores. Countries that experience a sudden in- or decrease in the assigned democracy scores might be experiencing different dynamics and the learned relationship between our observed features and the democracy scores might not hold in these particular circumstances. Figure 5 below plots the annual changes in a country's Polyarchy score. Panel A shows the relationship between the prediction error (observed Polyarchy score -predicted Polyarchy score) against the annual change in Polyarchy as coded by the V-Dem Project. Panel B shows, for instances where these changes was larger than 10 percentage points (at least a change of 0.1 on the 0-1 Polyarchy scale), the relationship between the predicted and the observed values. Large changes are rather rare. Panel A indicates that increases in Polyarchy scores are associated with larger prediction errors, the first quadrant of the cartesian plane has warmer colors than the third.

Figure E-2: **Relationship between prediction error and annual changes in Polyarchy**



These results indicate that predictions of our model are weakest in cases where there is a lot of missigness in the data, there are large changes in the democracy coding from year to year, and there also is larger coder disagreement. We argue that these are good news, our model performs worst in scenarios where we also expect human coders to struggle with coming up with a reliable estimate. Polities in transition, experiencing a rapid decline or increase in democracy or polities with no information (early polities or autocratic polities) are difficult to assess for humans and algorithms.

Appendix F: Other ML Models

Table F-1 shows the performance metrics for three additional models. We trained an XGBoost, a Gradient Boosting Machine (GBM), and a Generalized Linear Model (GLM) on the Polyarchy outcome variable with the full set of predictors. The XGBoost model performs comparatively well in the training data. However, its performance in the validation and cross-validation set drop significantly. The random forest model remains preferable. The Gradient Boosting Machine and the Generalized Linear Model never perform as well as the random forest or XGBoost in the training data and experience significant performance losses in the validation and cross-validation data set. The model of choice is therefore the random forest.

Table F-1: **Different Algorithms**

	Training			**Validation**			**Cross-Validation**		
Method	*MSE*	*RMSE*	*R2*	*MSE*	*RMSE*	*R2*	*MSE*	*RMSE*	*R2*
XGBoost	0.0341	0.1846	0.9979	0.9183	0.9583	0.9455	0.8493	0.9216	0.9475
GBM	0.3210	0.5666	0.9801	1.0145	1.0072	0.9398	0.9482	0.9738	0.9413
GLM	3.7938	1.9478	0.7655	3.7031	1.9243	0.7720	3.8572	1.9640	0.7616

Appendix G: Country-Year Plots for Polyarchy

Available upon request (hundreds of pages long).

Appendix H: Coder Judgment

Democracy is a latent concept so it is not surprising that all widely used democracy indices rest to some degree on coder judgments, as indicated in Table 1. Coders might be outside experts, project directors, or research assistants under their direction.

The role of judgment is most apparent in indices like Polity2 and Freedom House, where the coding categories are extremely broad and therefore open to interpretation. A glimpse of these complexities is offered in the Polity codebook (Marshall, Gurr, Jaggers 2013: 73), which instructs:

> If the regime bans all major rival parties but allows minor political parties to operate, it is coded here. However, these parties must have some degree of autonomy from the ruling party/faction and must represent a moderate ideological/philosophical, although not political, challenge to the incumbent regime.

It is not hard to see why different coders might have different interpretations of this coding rule.

The V-Dem expert survey disaggregates the concept of democracy into highly specific questions, which in principle should be more determinate. However, they still require interpretation. Questions incorporated into the Polyarchy index focus, among other things, on government censorship, harassment of journalists, media self-censorship, media bias, freedom of discussion, and freedom of academic and cultural expression – which expert coders rate on a Likert scale. Because they are not directly observable, and because they depend upon anticipated actions (How would the government respond if a citizen did *X*?), reasonable people with extensive knowledge of a country may disagree on the answers. And they do, as shown by coder-level responses in the V-Dem dataset (Marquardt et al. 2019). The measurement model developed by the project is designed to minimize random error and to correct for some coder biases. However, not all biases are amenable to algorithmic adjustment.

Even the more objective indices listed in Table 1 involve some sort of coder judgment. For example, in the Lexical index and BMR, the assessment of whether elections are genuinely competitive rests not only on whether government turnover has taken place but also on coder judgments.

The DD Index regards a polity as democratic if four conditions hold: (1) the chief executive is chosen (directly or indirectly) by popular election, (2) the legislature is popularly elected, (3) more than one party competes in elections, (4) an alternation in power occurs under electoral rules identical to the ones that brought the incumbent to office (Cheibub et al. 2010: 69). These rules are fairly clear in most instances but encounter ambiguity in others. Condition (1) is unclear where unelected and elected officials share power, as in many constitutional monarchies or polities where the military

exercises power sotto voce behind the throne. Condition (2) is complicated if there are multiple chambers or legislatures, some of which are elective and others appointive. Condition (3) is ambiguous in cases where the independence of "opposition" parties is in doubt.

Condition (4) has elicited the most controversy. The authors stipulate that because turnover is not known, ex ante, polities are coded as autocratic until an alternation occurs. If an alternation occurs, the country is recoded as democratic back to the date when the ruling party first gained power. This approach is potentially problematic, as the authors acknowledge, since codes are uncertain until an alternation has occurred. Another feature of the coding requires (in our opinion) some judgment on the part of the coder: when did electoral rules change? The authors state that the electoral rules in Mexico changed under Zedillo, when the PRI relinquished control of the Federal Electoral Institute, which means that 2000 – the first peaceful, election-based alternation of power – in Mexico's history also corresponds to its first year of democracy. Others might see things differently. And one faces the same problem in every regime in which the first three conditions (above) are met. Currently, Botswana poses a problem for DD, as one party has held power since independence under conditions that look (in other respects) quite democratic.

In a series of articles and books stretching back over several decades Vanhanen (2000, 2011) proposes a democracy index formed by the multiplication of two indices. One is focused on competition (100 minus the size of the largest party as a share of all votes or seats in an election) and the other on participation (the share of the eligible population who vote). Of all the extant indices, this is perhaps closest to our own approach.

However, Vanhanen's influential work is marred by several difficulties. First, it is unclear how he obtains turnout data for historical elections. Second, there are some seemingly arbitrary decision rules used to adjust scores for the Competition index. For example, if competitors in legislative elections are independent candidates rather than organized parties, Vanhanen automatically assigns the largest party a score of 30%. If the vote (or seat) share garnered by the largest party falls below 30% he nonetheless assigns a score of 30%, under the assumption that any further diminution is a product of electoral laws and is irrelevant to the quality of democracy. If candidates are not aligned with a political party, but parties are allowed, he again sets the share of the "largest party" to 30%. The size of the largest party cannot fall below 30% on the assumption that further attenuation must be the product of electoral system oddities. Where elections involve several rounds, Vanhanen usually uses second round results but occasionally shifts to first round results. It is not possible to tell how many observations these (and other) ad hoc coding decisions affect.

Appendix I: Bias Reduction

In this appendix we demonstrate with a series of examples how our approach reduces random and systematic sources of biases in the data. As in any other regression type model researchers do not face problems with their estimation strategy if the bias is uncorrelated with the covariates in the model. However, in the social world that social scientists study that is frequently not the case. We test our claim that the implemented approach can reduce biases in the data with a series of particularly hard and challenging cases. We vary the degree to which the bias introduced is correlated with both our *outcome* as well as our *predictors*. We find that across different types of random and systematic biases our model significantly reduces the introduced bias at a minimal cost of increased random error, or noise, in the predictions.

Across all tests we always introduce bias generated with a draw from a normal distribution with mean .1 and variance .1. As a reminder, the Polyarchy index ranges from 0 to 1 and has a variance of 0.069, a N(.1,.1) bias introduction is hence always substantial in size. The number of biased observations in our data set varies based on condition and ranges between 862 and 1,997. This means that in different scenarios between 6% (highly liberal democracies) and 26% (left-wing governments) of the observations in the dataset are biased. Even for the completely random assignment of bias to observations we are hence in a worst-case scenario. Biased data points are introduced before the data set are split into training and validation set, replicating the real world scenario with which we would encounter such biases and setting the hurdle for the test higher.

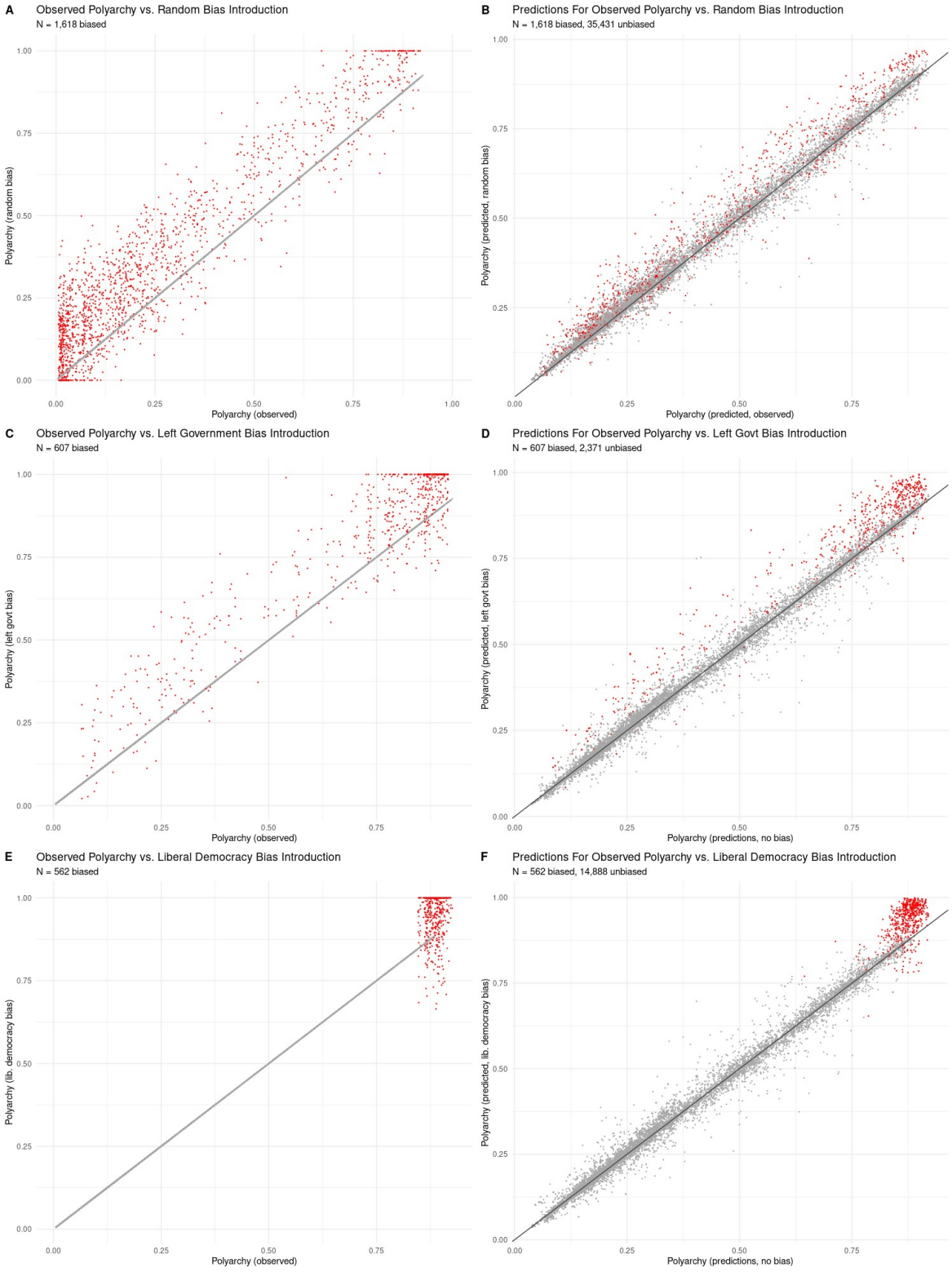
Table I-1: **Bias introduced**

Bias	**Obs.**	**Clustered Error**	**Reduction (%)**
Random	10%	No	84
Left-government	26%	Moderately	28
Highly democratic	6%	Highly	8

We first introduce random error. This error is not correlated to our outcome or our predictors and is not clustered in the data distribution. We introduce this bias for 1,997 out of 20,020 observations. This bias introduction is 1.445 times the variance of the Polyarchy measure and hence a substantial introduction of upward bias. Panel A in Figure I-1 shows the relationship between the original Polyarchy measure and the Polyarchy measure with bias. The gray observations on the 45-degree line are the unchanged Polyarchy variables, the red observations are the biased Polyarchy scores. In panel B we plot the predictions of a model that was

trained on the unbiased Polyarchy data against the predictions of a model that was trained on the biased Polyarchy data. It is immediately visible that the biased observations in red are sitting much tighter to the 45-degree line in panel B than in panel A. Calculating the average distance between biased and unbiased observations we can see that there was a drop from 0.0987 in the original data to 0.0162 in the prediction. This came at an increase of the mean error in the prediction from 0.0096 in the unbiased model to 0.0103 in the biased model. We can hence conclude that our random forest approach is able reduce significant biases that are uncorrelated with any of our measures.

Figure I-1: Introduction of bias with N (.1,.1)



In a next step we introduce bias for country-year observations in which a country is ruled by a left government. The coding of left governments is based on Brambor, Lindvall and Stjernquist (2017). In this scenario, the introduced bias is somewhat correlated to our outcome variable, as Panel C in Figure I-1 shows. Country-year observations with left governments have a larger representation among higher Polyarchy scores. The average Polyarchy score for left-government observations is 0.69, the one for non-left-government observation is 0.47. The measure is also correlated with our predictors. Left leaning governments have higher suffrage rates and female suffrage rates. Nevertheless, as panel C shows we can find country-year observations with left governments across the entire distribution of Polyarchy scores. As before the second panel, panel D, shows the predictions of a model that was trained on the unbiased Polyarchy data against the predictions of a model that was trained on the biased Polyarchy data. Comparing the two panels in Figure 2 we can visually see a reduction in the bias. The average distance of biased observations to their unbiased, original value decreased by 0.025 from 0.088 to 0.0628. This came at an increase in random error of 0.016 from -0.0085 to 0.0079 for all observations. We hence have a 28 percentage point reduction in bias at a minimal cost.

Finally, we introduce bias that by design is strongly correlated with the outcome variable, as shown in Figure I-1. Specifically, we add bias drawn from N(.1,.1) to all countries that the V-Dem Liberal Democracy Index classifies as highly liberal democratic (coded as 1 on the scale of the *e_v2x_libdem_5C* variable). The Liberal Democracy Index is a combination of the *v2x_liberal* variable and our outcome variable, *v2x_polyarchy* (Coppedge et al., 2021). The specific aggregation function is:

$$v2x\_libdem = .25*v2x\_polyarchy^{1.585} + .25*v2x\_liberal + .5*v2x\_polyarchy^{1.585} *v2x\_liberal$$

Furthermore, this bias is also correlated with several of our predictors. First, the bias is correlated with the availability of observations for our model. For example, highly liberal democracies have frequent elections and make the results of these elections public. Missing data on election results for these country-year observations is lower than for other countries. Secondly, the bias is directly correlated to several of our key predictors. Highly liberal democracies, by design, have high suffrage shares as well as high female suffrage shares. In addition, this bias also clusters in a specific area of the democracy score distribution. While the random error as well as the left government error introduced bias across the entire spectrum of democracy score values (as can be seen in Figure I-1 Panel A and Figure I-1 Panel C) countries high on the Liberal Democracy Index are per definition all scoring relatively high on the Polyarchy measure. The introduction of this bias is hence the hardest possible case to test our claim: a very large

bias introduced with strong correlation with the predictors that clusters in a specific area of the distribution of our outcome.

Panel E in Figure I-1 plots the observed, unchanged Polyarchy values against the biased observations with bias based on their e_v2x_libdem_5C score. In panel F we once more show the predictions of a model that was trained on the unbiased Polyarchy data against the predictions of a model that was trained on the biased Polyarchy data. Comparing Panel E to F it can be seen that the red observations are now closer to the 45 degree line. The red, biased observations still form a cluster above the line in the top corner of the Polyarchy distribution but even in this worst case scenario the bias has been reduced. The average distance of biased observations to their unbiased, original value decreased by 0.005 from 0.065 to 0.060. We hence have a reduction of incredibly strong bias of almost eight percentage points. While this reduction is somewhat modest, it nonetheless demonstrates that OSMs can attenuate bias even in the absolute worst of worlds.

We would like to highlight that this approach offers researchers their own way of assessing the role and influence of all possible sources of bias. It is straightforward to introduce specific types of biases that researchers might suspect influence or drive democracy scores. Researchers might have a theoretical reason to suspect that specific observations in existing democracy indices are subject to particular biases. As we demonstrate in this appendix, it is possible to specify the type of bias that one is concerned about, vary the intensity of that bias, and assess how well the random forest is dealing with these types of biases, across various levels of bias intensity. Researchers applying our approach can modify bias type and intensity to suit their theoretical expectations, research needs or curiosity. The results of such analyses can be used to put upper and lower bounds on inherited bias in OSMs, subject to reasonable assumptions.

Appendix J: Decision Trees and Random Forest

In this section we present a detailed explanation of decision trees and random forests with reference to application in political science. We start by explaining the logic of decision trees, how trees become a forest, and point to Guyon (1997) and research following her work on common rules in random forest analysis.

The starting point: Decision Trees

Random forests are a machine learning algorithm that is based on decision trees. Decision trees are a type of supervised learning algorithm used for classification and regression tasks. They work by recursively splitting the input data into subsets based on the value of a chosen feature to create a tree-like model of decisions and their possible consequences. Each internal node of the tree represents a feature, and each leaf node represents a class or a regression value. To make a prediction on a new data point, the decision tree starts at the root node and follows the appropriate branch based on the value of the corresponding feature, until it reaches a leaf node and outputs the class or regression value associated with that node (Hastie, Tibshirani Friedman 2013 but also McAlexander and Mentch 2020 or Hill and Jones 2014 for an application in political science).

The idea behind a decision tree is hence to build a series of binary decisions based on the input features (independent variables), that lead to a prediction of the output class (dependent variable). Each decision is a split and creates a branch in the tree (the name is very literal), which leads to a different set of decisions or a prediction. In this way, the decision tree can be seen as a series of if-then statements that make a prediction at the end.

Decision trees are based on similar data set structures as more common statistical methods in the social sciences. We have an outcome variable and a set of predictor or explanatory variables. The decision tree will analyze this data in the search for the best possible split of the data (Step 1). The decision tree algorithm selects the input feature that best separates the data based on some criterion, such as information gain, gain ratio, or Gini index (Step 2).[8] Once the best split has been selected, the decision tree algorithm creates a new node in the tree (Step 3). This node represents the decision based on the selected input feature. The data is then split into two branches, one for each possible outcome of the decision.

The algorithm then recursively repeats split selection and node creation for each of the two branches created in step 3. This continues until

[8]Information gain measures the reduction in entropy or uncertainty in the data, while gain ratio normalizes the information gain by the intrinsic information of the feature. Gini index measures the impurity of the data, or the probability of misclassification of a randomly chosen data point.

a previously determined stopping criterion is met, such as a maximum depth of the tree, a minimum number of data points in a leaf node, or a minimum information gain. Each branch in the tree represents a sequence of decisions that lead to a prediction of the output variable.

Once the tree has been built, making a prediction for a new data point involves traversing the tree from the root node to a leaf node, based on the values of the input features. At each node, the decision based on the input feature is made, and the traversal continues down the corresponding branch until a leaf node is reached. The value in the leaf node represents the predicted value of the output variable (Hastie, Tibshirani Friedman 2013, Greenwell 2022)

From a tree to a forest

Decision trees are prone to overfitting, especially when the tree becomes very deep or complex. This means that they can capture the idiosyncrasies of the training data too closely, and fail to generalize well to new, unseen data. To address this problem, random forests use an ensemble of decision trees to make more robust and accurate (out-of-sample) predictions.

In a random forest, each tree is trained on a randomly selected subset of the training data, and only considers a random subset of the features at each split. This helps to reduce the correlation between the trees and decorrelate their predictions, while preserving their individual strengths. The final prediction of the random forest is then based on the majority vote of the predictions of all the trees, for classification tasks, or the mean of the predictions, for regression tasks (Greenwell 2022).

In summary, decision trees form the building blocks of random forests and provide the framework for making individual predictions, while random forests aggregate the predictions of multiple decision trees to make a more reliable and generalizable prediction (see for example Muchlinski et al. 2016 as well as the responses by Wang 2019 and Muchlinski et al 2019).

Random Forest model application

In random forest analysis, the data is typically divided into several subsets, each of which serves a different purpose. These subsets include a training set, a validation set, cross-validation, and a test set (Guyon 1997, Dubbs 2021, and Aria 2023. In political science see, for example, Hill and Jones 2014 or McAlexander and Mentch 2020).

1. Training set: This is the portion of the data used to train the random forest model. The model uses the training set to learn the relationships between the input features and the target variable(s). The size of the training set should be large enough to capture the variability in the data, but not so large that it slows down the training process or overfits the model.

2. Validation set: This is a subset of the data that is used to evaluate the performance of the model during training. The validation set is used to tune the hyperparameters of the model, such as the number of trees, the maximum depth of the trees, and the minimum number of samples required to split a node. The validation set should be large enough to provide a reliable estimate of the model's performance, but not so large that it overfits the hyperparameters.
3. Cross-validation: Cross-validation is a technique for estimating the performance of a model by splitting the data into multiple folds and training the model on each fold while evaluating its performance on the remaining folds. Cross-validation is used to estimate the generalization error of the model and to select the best model from a set of candidate models. The number of folds used in cross-validation depends on the size of the data and the computational resources available. One cross-validates within the training set.
4. Test set: This is a subset of the data that is used to evaluate the final performance of the model after training and hyperparameter tuning. The test set is used to estimate the model's generalization error and to compare its performance to other models. The test set should be large enough to provide a reliable estimate of the model's performance, but not so large that it's computationally prohibitive, or reduces the size of the training set too substantially.

The size of the data splits in random forest analysis depends on several factors, including the size of the data, the complexity of the model, and the computational resources available. In the following, we present several “common rules” on data splits. Highlighting, however, that decisions on data splits are also dependent on the distribution of the data and that split selection needs to make sure that key features of the data are represented in the training, validation, and test data.

1. Training set: The training set should be large enough to capture the variability in the data and to prevent overfitting, but not so large that it slows down the training process. A common rule of thumb is to use 60-80% of the data for training.
2. Validation set: The validation set should be large enough to provide a reliable estimate of the model's performance during training, but not so large that it overfits the hyperparameters. A common rule of thumb is to use 10-20% of the data for validation.
3. Cross-validation: The number of folds used in cross-validation depends on the size of the data and the computational resources available. A common rule of thumb is to use 5-10 folds for small to medium-sized data sets, and 3-5 folds for large data sets.
4. Test set: The test set should be large enough to provide a reliable estimate of the model's generalization error, but not so large that it's

computationally prohibitive. A common rule of thumb is to use 20-30% of the data for testing.

Appendix K: Out of Sample Application

Researchers interested in applying the random forest model to specific out-of-sample prediction of completely new polities that have never been coded before should stratify the sampling into training, (cross-)validation, and test set. In order to make sure that the random forest model does as well as possible in the out-of-sample prediction it is necessary to simulate out-of-sample prediction during the model training. This can be achieved through stratified sampling that assigns all country-year observations of specific countries to either the training, (cross-) validation, or test set.

As an example, in a stratified sampling approach all country-year observations of the United States of America might end up in the training set, all country-year observations of Mexico might end up in the validation set, and all country-year observations of Canada in the test set. The model is then trained and validated on the USA and Mexico and the out-of-sample performance is assessed with Canada.

As a demonstration, we implemented this approach by assigning all country-years of countries to either the training or the test set and by generating six blocks for the cross-validation data set that randomly assign entire countries of the training set into one of six blocks.

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Total
3,890	3,755	3,482	2,781	2,546	3,278	19,732

The random forest then iteratively trains the model on five of these six folds and predicts on the sixth. Trying to maximize the predictive performance in the fold that was left out of the training. The performance for out-of-bag samples using the training data is listed in Table K-1 below.

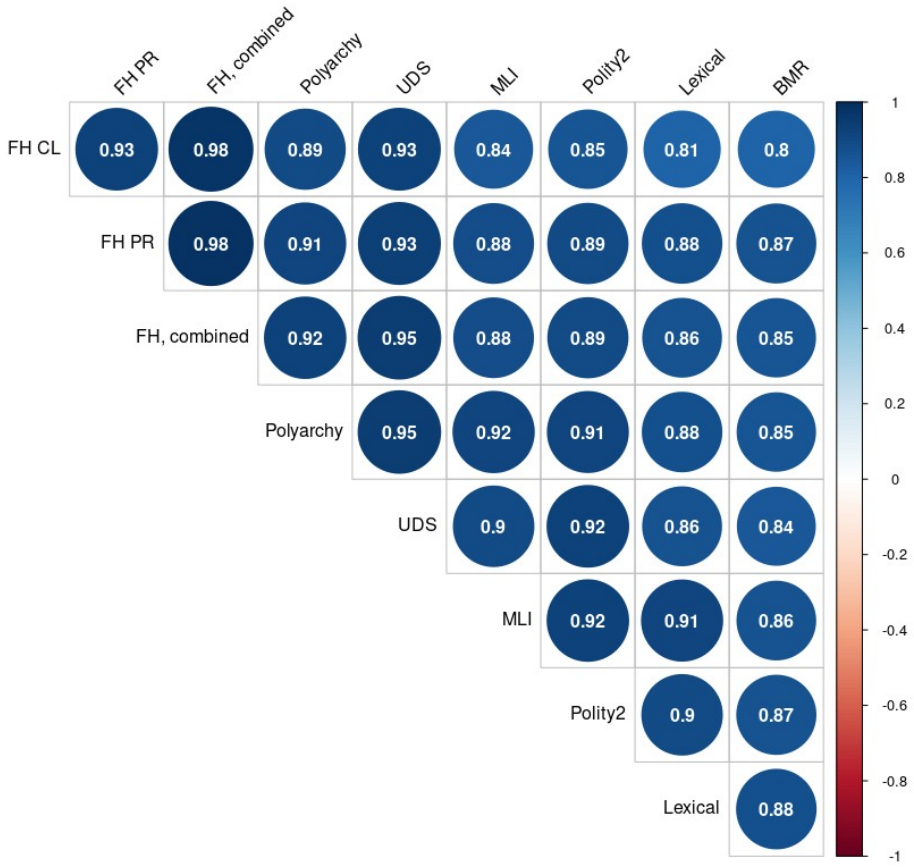
Table K-1: **Cross-Validation Results**

	Data	
Metric	**Training**	**Cross-validation**
Mean Squared Error	0.002	0.009
Root MSE	0.039	0.097
Mean Absolute Error	0.024	0.069
Root Mean Squared Log Error	0.030	0.070
Mean Residual Deviance	0.002	0.009
R2	0.977	0.857

Appendix L: Correlations among Democracy Indicators

Democracy indices differ in how they define and measure the latent concept democracy. Yet, since the core concept is shared it is reasonable to expect that these indices might be correlated. Figure L-1 demonstrates this simple point, showing correlations across a series of influential democracy indices. All correlations are positive and all are quite high (Pearson's r ≥ 0.8).

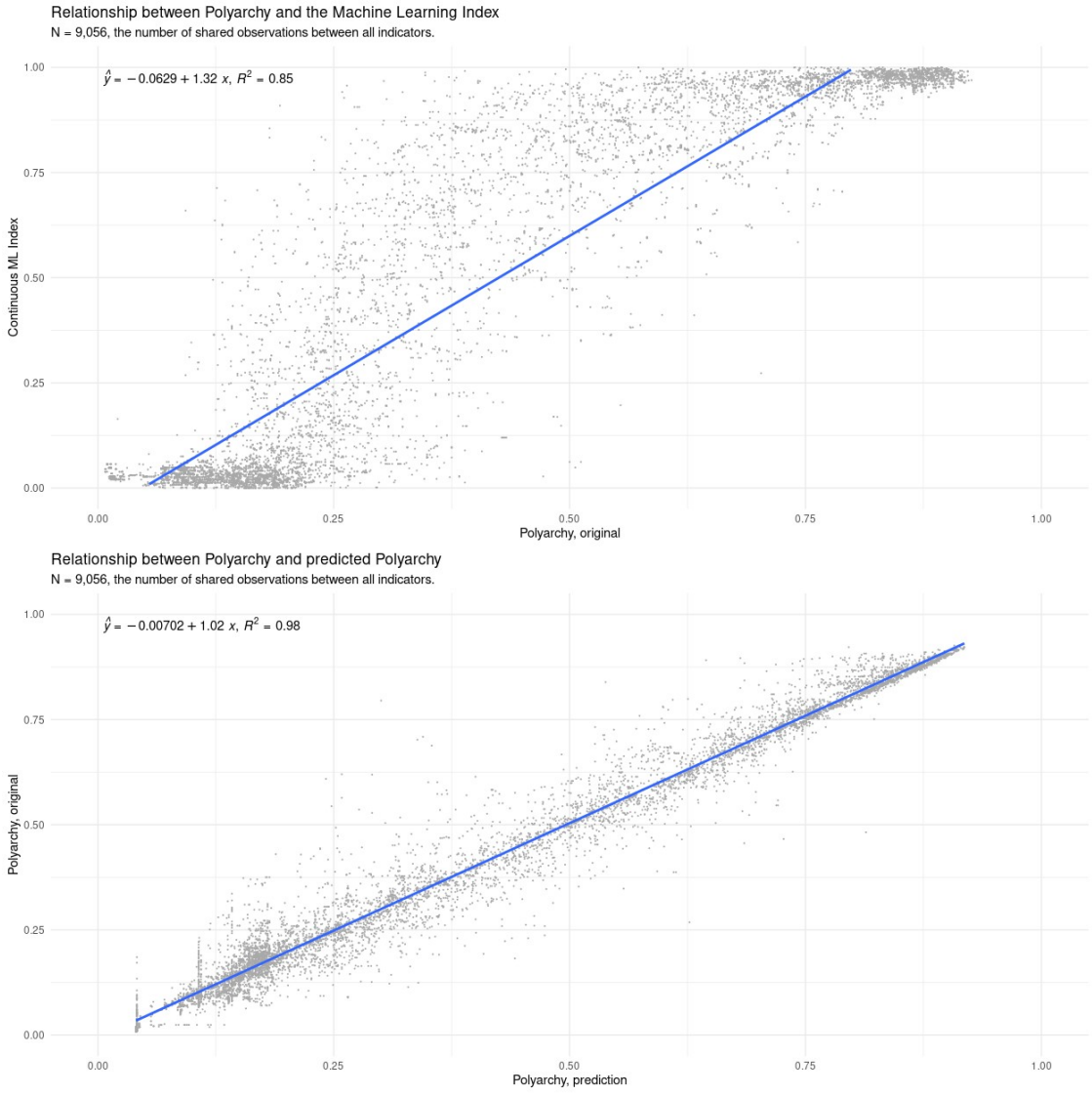
Figure L-1: **Correlation of Democracy Indices**



Note: Shown are the correlations between the democracy indices that we are using in our analysis. Blue indicates positive correlations, red negative correlations, and darker colors signify stronger correlations.

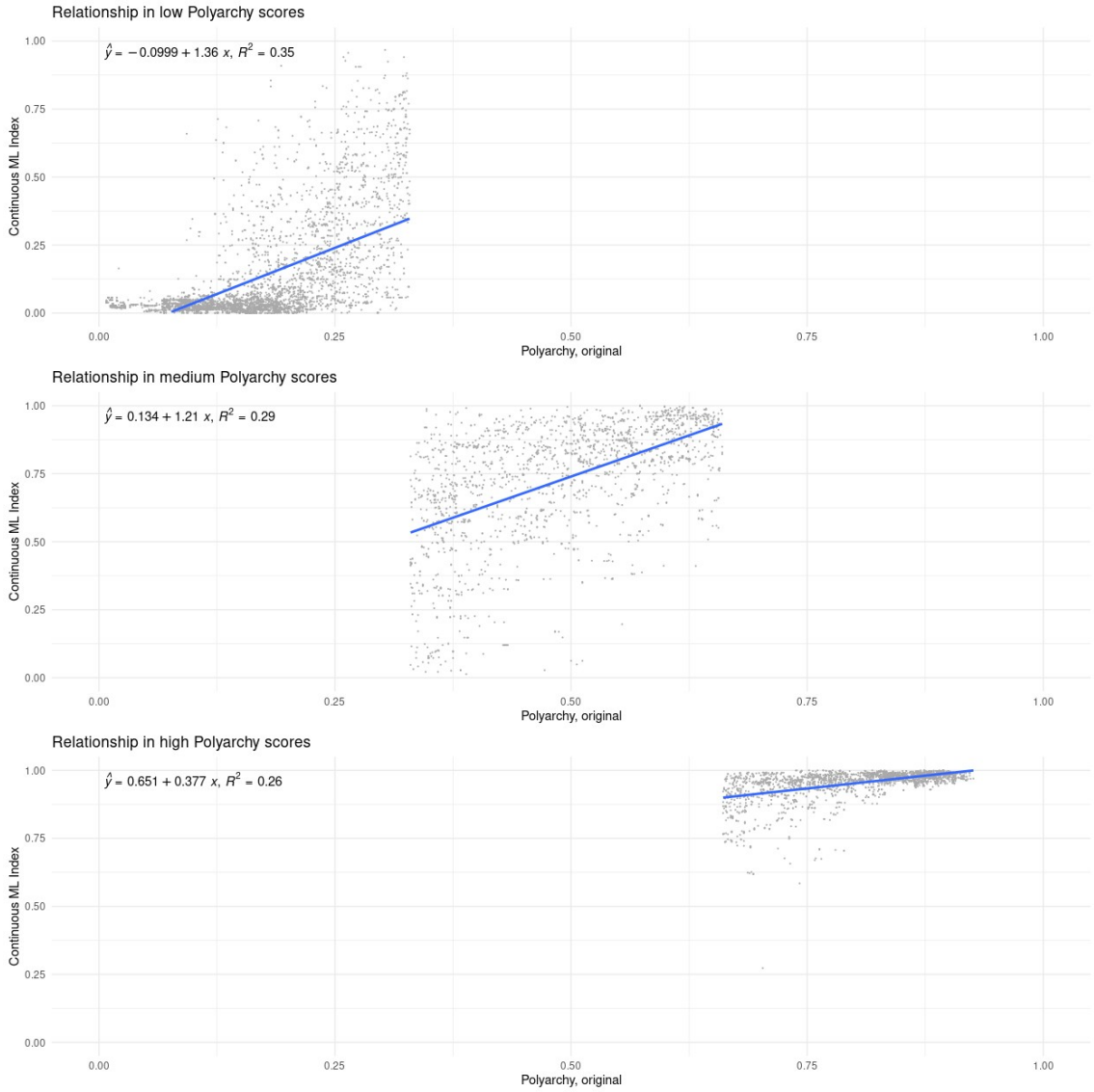
In further analyses, shown in Figure L-2, we focus on the machine-learning democracy index, MLI (Gründler and Krieger 2021). The upper panel shows the relationship between the MLI and Polyarchy. The scatterplot presents a good deal of scatter around the middle, which is not surprising given that the model is trained on the extremes (the top and bottom deciles). This is not a problem, per se, and Gründler and Krieger (2021) highlight scenarios under which this is even desirable. By contrast, the association between the our OSM and Polyarchy, shown in the lower panel, is much closer and does scarcely varies across the distribution. This reinforces our main conclusion, namely, that Gründler and Krieger present a new democracy index while we present a way to extend existing indices.

Figure L-2: **Relationship between MLI and Polyarchy**



In Figure L-3 we take a closer look at the fit between the observed Polyarchy scores and the MLI. As can be seen, the relationship between the two variables varies considerably across the three slices of data. We split the Polyarchy score in the lower (0-0.33), middle (above 0.33-0.66), and the upper third (above 0.66-1). The overall fit and slope of bivariate regression lines varies across all subsets of the data. Although the overall correlation between the MLI and Polyarchy is 0.92, in the lower quarter it is 0.35, in the middle it is 0.29, and in the top quarter it is 0.25.[9] (Analogous plots for Polyarchy and OSM are displayed in Appendix M.)

Figure L-3: **Relationship between Polyarchy and the MLI in subsets of the data**



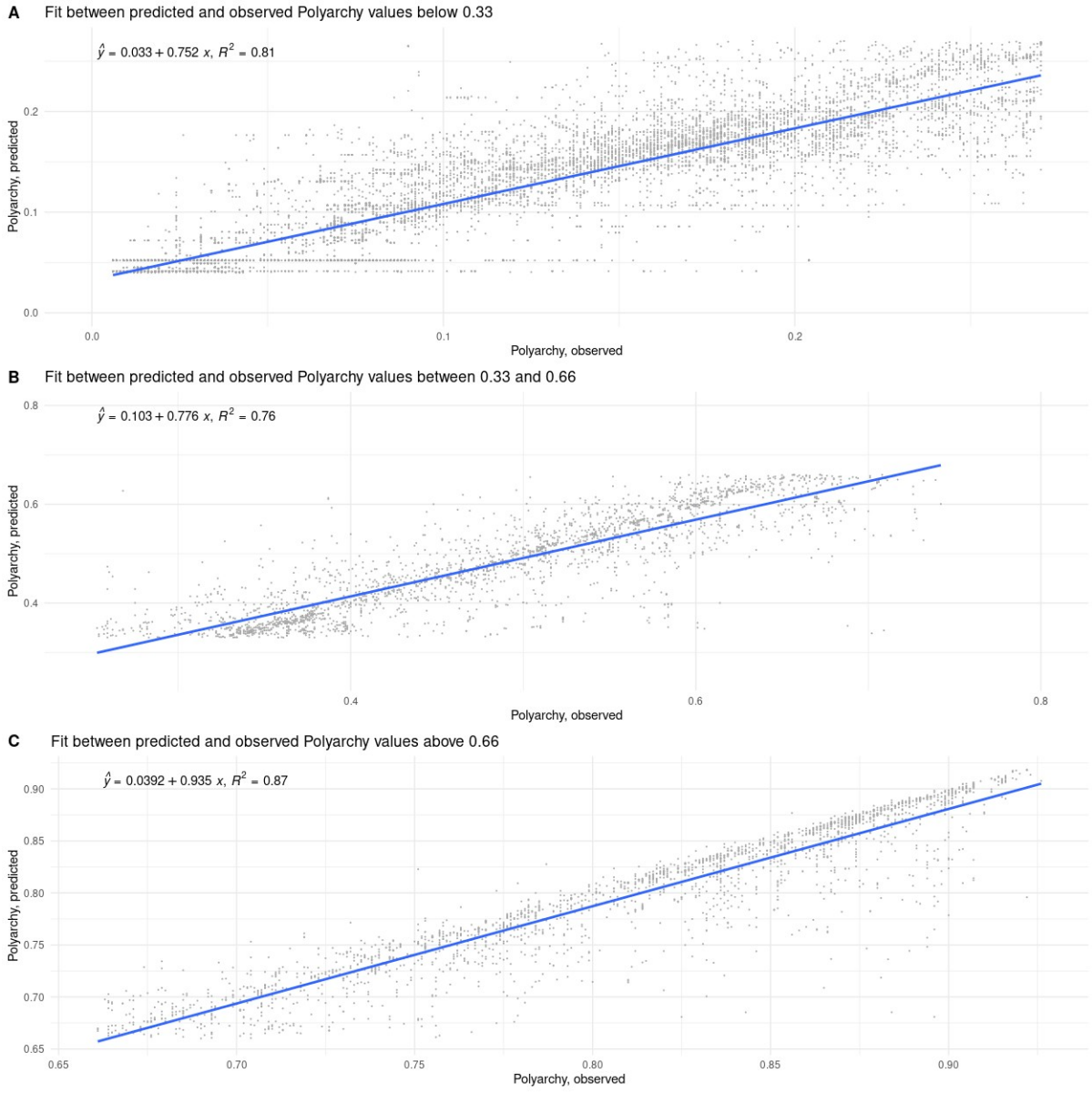
[9]Note that the overall correlation between two variables does not necessarily equal the sum of correlation of different subsets of the same data.

Note: The relationship between the MLI (Gründler and Krieger 2021) and Polyarchy across three equal-sized sub-sections of the Polyarchy index. The blue line is a bivariate linear regression line.

Appendix M: Conditionality of fit

In order to validate the robustness of the predictions we demonstrate that the fit of the model does not depend on specific Polyarchy scores. We split the data into three groups with values ranging from 0 to 0.33, above 0.33 to 0.66, and above 0.66. By doing so we can highlight that potentially easy classification cases (0-0.33 and 0.67-1) are not fundamentally driving the performance of the model. Extremely well fitting predictions for obvious classifications at the upper and lower end of the Polyarchy distribution are not causing the fit. As Figure M-1 shows there is some variation across the groups with the key difference being between the low + medium democracy score cases (which we would argue are the harder cases) and the high democracy score cases.

Figure M-1: Fit across different values



Appendix References

Alvarez, Mike, Jose A. Cheibub, Fernando Limongi, Adam Przeworski. 1996. “Classifying Political Regimes.” *Studies in Comparative International Development* 31(2): 3–36.

Aria, Massimo, Agostino Gnasso, Carmela Iorio, and Giuseppe Pandolfo. "Explainable Ensemble Trees." Computational Statistics (2023): 1-17.

Boix, Carles, Michael Miller, Sebastian Rosato. 2013. “A Complete Dataset of Political Regimes, 1800-2007.” *Comparative Political Studies* 46(12), 1523-1554.

Brambor, Thomas, Johannes Lindvall, and Annika Stjernquist. 2017. “The Ideology of Heads of Government, 1870–2012.” Version 1.5. Department of Political Science, Lund University.

Cheibub, Jose Antonio, Jennifer Gandhi, James Raymond Vreeland. 2010. “Democracy and Dictatorship Revisited.” *Public Choice* 143(1–2): 67–101.

Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Jan Teorell, David Altman, Michael Bernhard, Agnes Cornell, M. Steven Fish, Lisa Gastaldi, Haakon Gjerløw, Adam Glynn, Allen Hicken, Anna Lührmann, Seraphine F. Maerz, Kyle L. Marquardt, Kelly McMann, Valeriya Mechkova, Pamela Paxton, Daniel Pemstein, Johannes von Römer, Brigitte Seim, Rachel Sigman, Svend-Erik Skaaning, Jeffrey Staton, Aksel Sundtröm, Eitan Tzelgov, Luca Uberti, Yi-ting Wang, Tore Wig, and Daniel Ziblatt. 2021. "V-Dem Codebook v11.1" Varieties of Democracy (V-Dem) Project.

Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Jan Teorell, David Altman, Michael Bernhard, Agnes Cornell, M. Steven Fish, Lisa Gastaldi, Haakon Gjerløw, Adam Glynn, Allen Hicken, Anna Lührmann, Seraphine F. Maerz, Kyle L. Marquardt, Kelly McMann, Valeriya Mechkova, Pamela Paxton, Daniel Pemstein, Johannes von Römer, Brigitte Seim, Rachel Sigman, Svend-Erik Skaaning, Jeffrey Staton, Aksel Sundtröm, Eitan Tzelgov, Luca Uberti, Yi-ting Wang, Tore Wig, and Daniel Ziblatt. 2021. "V-Dem Codebook v11.1" Varieties of Democracy (V-Dem) Project.

Dubbs, Alexander. "Test set sizing via random matrix theory." *arXiv preprint arXiv:2112.05977* (2021).

Freedom House. 2015. “Methodology: Freedom in the World 2015.” New York. (https://freedomhouse.org/sites/default/files/Methodology_FIW_2015.pdf), accessed December 2, 2015.

Gleditsch, Kristian S., Michael D. Ward. 1999. "A revised list of independent states since the Congress of Vienna." *International Interactions* 25.4: 393-413.

Greenwell, Brandon M. Tree-based Methods for Statistical Learning in R. CRC Press, 2022.

Guyon, Isabelle. "A scaling law for the validation-set training-set size ratio." *AT&T Bell Laboratories* 1, no. 11 (1997).

Hastie, Trevor, Robert Tibshirani, Jerome Friedman. 2013. The elements of statistical learning. New York: Springer.

Hill, Daniel W., and Zachary M. Jones. 2014. "An Empirical Evaluation of Explanations for State Repression." American Political Science Review 108 (3): 661-687.

Marquardt, Kyle L., Daniel Pemstein. 2018. “IRT models for expert-coded panel data.” *Political Analysis* 26(4): 431–456.

Marshall, Monty G. 2020. “POLITY5 Political Regime Characteristics and Transitions, 1800-2018 Dataset Users’ Manual.” Center for Systemic Peace and Societal-Systems Research. (www.systemicpeace.org/inscr/p5manualv2018.pdf)

Marshall, Monty G., Ted Gurr, Keith Jaggers. 2013. “Polity IV Project: Political Regime Characteristics and Transitions, 1800–2012, Dataset Users' Manual.” Center for Systemic Peace, Viena, VA.

McAlexander, Richard J., and Lucas Mentch. "Predictive inference with random forests: A new perspective on classical analyses." Research & Politics 7.1 (2020).

Muchlinski, David, David Siroky, Jingrui He, and Matthew Kocher. 2016. "Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data." Political Analysis 24, no. 1: 87-103.

Muchlinski, David Alan, David Siroky, Jingrui He, and Matthew Adam Kocher. 2019. "Seeing the forest through the trees." Political Analysis 27, no. 1: 111-113.

Nohlen, Dieter (ed). 2005. *Elections in the Americas: A Data Handbook, vols 1-2*. New York: Oxford University Press.

Nohlen, Dieter, Florian Grotz; Christof Harmann (eds). 2002. *Elections in Asia and the Pacific: A Data Handbook, vols 1-2*. New York: Oxford University Press.

Nohlen, Dieter, Michael Krennerich, Bernhard Thibaut (eds). 1999. *Elections in Africa: A Data Handbook*. Oxford: Oxford University Press.

Nohlen, Dieter; Philip Stover (eds). 2010. *Elections in Europe: A Data Handbook*. Nomos Verlagsgesellschaft.

Pemstein, Daniel, Stephen Meserve, James Melton. 2010. “Democratic Compromise: A Latent Variable Analysis of Ten Measures of Regime Type.” *Political Analysis* 18(4): 426–449.

Przeworski, Adam. 2013. “Political Institutions and Political Events (PIPE) Data Set.” Available at https:// sites.google.com/a/nyu.edu/adam-przeworski/home/data

Skaaning, Svend-Erik, John Gerring, Henrikas Bartusevičius. 2015. “A Lexical Index of Electoral Democracy.” *Comparative Political Studies* 48(12): 1491-1525.

Teorell, Jan, Michael Coppedge, Staffan Lindberg, Svend-Erik Skaaning. 2019. "Measuring polyarchy across the globe, 1900–2017." *Studies in Comparative International Development* 54, no. 1: 71-95.

Vanhanen, Tatu. 2000. "A new dataset for measuring democracy, 1810-1998." *Journal of peace research* 37.2: 251-265.

Vanhanen, Tatu. 2011. "Measures of democracy 1810–2010." *FSD1289, version* 5. www.fsd.tuni.fi/fi/aineistot/taustatietoa/FSD1289/Introduction_2010.pdf

Wang, Yu. 2019."Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data: A comment." Political Analysis 27, no. 1: 107-110.