Received: 4 August 2022 | Revised: 28 May 2024 | Accepted: 25 June 2024

DOI: 10.1111/ajps.12968

AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

ARTICLE

Measuring electoral democracy with observables

Daniel Weitzel[1] | John Gerring[2] | Daniel Pemstein[3] | Svend-Erik Skaaning[4]

[1]Department of Political Science, Colorado State University, Fort Collins, Colorado, USA

[2]Department of Government, University of Texas at Austin, Austin, Texas, USA

[3]Department of Political Science and Public Policy, North Dakota State University, Fargo, North Dakota, USA

[4]Department of Political Science, Aarhus University, Aarhus, Denmark

Correspondence
Daniel Weitzel, Department of Political Science, Colorado State University, Fort Collins, CO, USA.
Email: daniel.weitzel@colostate.edu

Abstract

Most cross-national indices of democracy rely centrally on coder judgments, which are susceptible to bias and error, and require expensive and time-consuming coding by experts. We present an approach to measurement based on observables that aim to preserve the nuanced quality of subjectively coded democracy indices. Our observable-to-subjective score mapping is free of idiosyncratic coder errors arising from misinformation, slack, or biases. It is less susceptible to systematic bias that may arise from coders' inferences about a country's regime, for example, from the ideology of the ruler. The data collection procedure and mode of analysis are fully transparent and replicable, and the procedure is based on random forests and 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 cross-national 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., 201; Munck, 2009). "If we were to renounce our judgmental faculties in the measurement of regime properties and regime dynamics," Schedler (2012, p. 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, that is, 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 cross-nationally (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

Verification Materials: The materials required to verify the computational reproducibility of the results, procedures, and analyses in this article are available on the *American Journal of Political Science* Dataverse within the Harvard Dataverse Network at: https://doi.org/10.7910/DVN/CWVQTZ.

The Cornell Center for Social Sciences verified that the data and replication code submitted to the AJPS Dataverse replicates the numerical results reported in the main text of this article.

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 across countries can be interpreted in specific terms, that is, 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 the 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.

EXTANT INDICES

We list the most widely used measures of democracy in Table 1, along with some key features. Appendix H (p. 15) in the Supporting Information 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.

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]

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, Chaps. 3 and 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, raise the sample of coders slightly—but

[1] We demonstrate these features in Appendix I (p. 17) in the Supporting Information 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.

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

TABLE 1 Extant democracy indices.

	Coder judgment	Scale	Raters	Polities	Years	Observations	Google Scholar
Freedom House (Freedom House, 2015)	High	Ordinal	1	202	1972–	7,598	1,780
Polity2 (Marshall et al., 2020)	High	Ordinal	1	182	1800–2018	15,772	2,360
Unified Democracy Scores (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
Boix-Miller-Rosato (Boix et al., 2013)	Low	Binary	1	208	1800–2015	15,620	688
Democracy-Dictatorship (Alvarez et al., 1996; Cheibub et al., 2010; Bjørnskov & Rode, 2020)	Low	Binary	1	208	1950–	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	1810–2018	14,984	331
Machine-learning democracy index (Gründler & Krieger, 2016, 2021)	Low	Interval, binary	N/A	186	1919–2019	12,588	151

Note: 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. Observations: country-year observations. Google Scholar citations (approximate) from 2015 to 2022. All measures of democracy are highly correlated (Appendix L, p. 27 in the Supporting Information).

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, for example, in favor of governments that pursue more liberal policies and against those who pursue more conservative policies.

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 United States 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 19th 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,

[3] The Chapel Hill survey enlists an average of 13 coders per country (Bakker et al., 2015), and the Electoral Integrity project enlists an average of 40 experts per country (Norris et al., 2013).

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 an 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, for example, 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 Boix-Miller-Rosato (BMR) dichotomous coding of democracy and the Democracy-Dictatorship 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 70 (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 19th 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 20th 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 cross-national 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 Evaluating Potential Biase).

Machine learning

A final index utilizes a method of aggregation that bears a 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 deciles 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, p. 27 in the Supporting Information).

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 Understanding Deviations 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, p. 2 in the Supporting Information). Granted, there are situations in which the historical record is unclear, for example, where we do not know, or do not know for sure, what the vote or seat total was for the winning party. Here, data are 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 (p. 2) in the Supporting Information, 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 is not much variance left to explain.

Later iterations of our model reduce the collection of variables from 40 to 13 to aid interpretability and to reduce the costs of extending or replicating this work. We selected the 13 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 40 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 & 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, and, most 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 and 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 validation set, a random sample of 15% of the total observations. The remaining 20% comprise the test set. This dataset 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 dataset, without stratification. This specification is used for an overall fit to 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-

[4] In Appendix F (p. 13) in the Supporting Information, we also report the results of a gradient boosting machine, XGBoost, and generalized linear models. We use a random forest to construct our OSM because it consistently outperforms other techniques in cross-validation and validation datasets. 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 (p. 22) in the Supporting Information.

[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 assessed the performance of our final OSM on the test set at publication time, after incorporating any changes suggested by reviewers. Results are reported in Appendix O in the Supporting Information.

[7] The dataset has numerous missing values. We apply various imputation approaches. Results, in Appendix D (p. 9) in the Supporting Information, are similar to those we report in the manuscript.

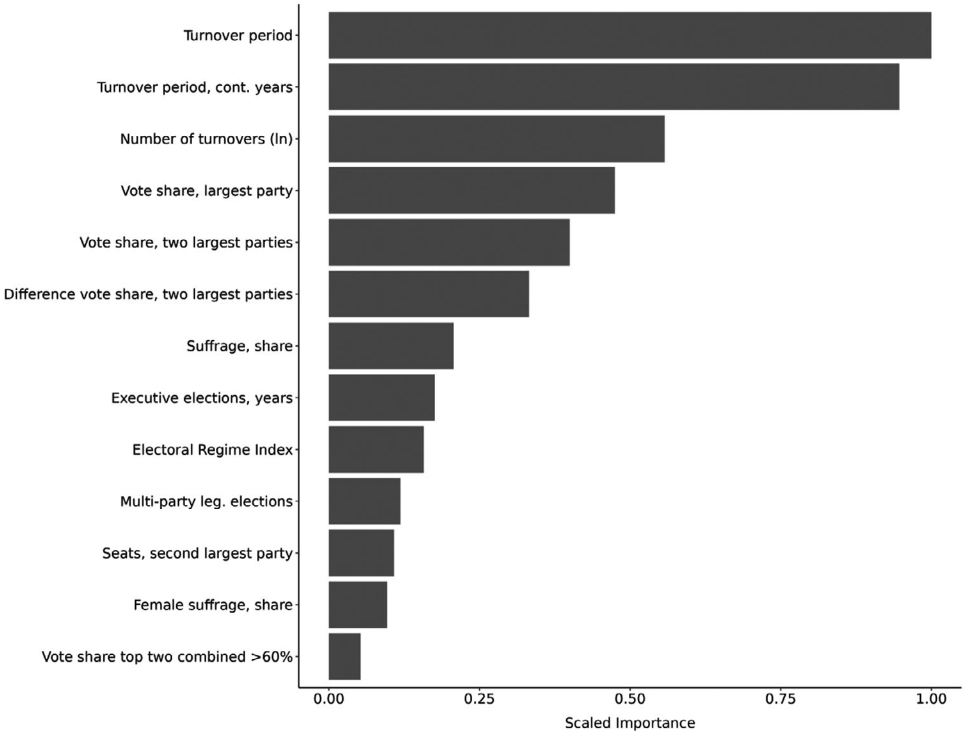

FIGURE 1 Variable importance. *Note*: The variables ($N = 13$) with the highest importance scores in the random forest model, with Polyarchy as the target.

stratified model specifications and explanations in Appendix K (p. 26) in the Supporting Information.

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 40 variables to 13, ranked according to their importance for the distributed random forest in Figure 1. Variable "importance" measures the extent to which the 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 13 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 are 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).

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 13 variables), recalibrating the algorithm each time. The variations that result from these serial omissions are very slight as shown in Appendix B (p. 6) in the Supporting Information. 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

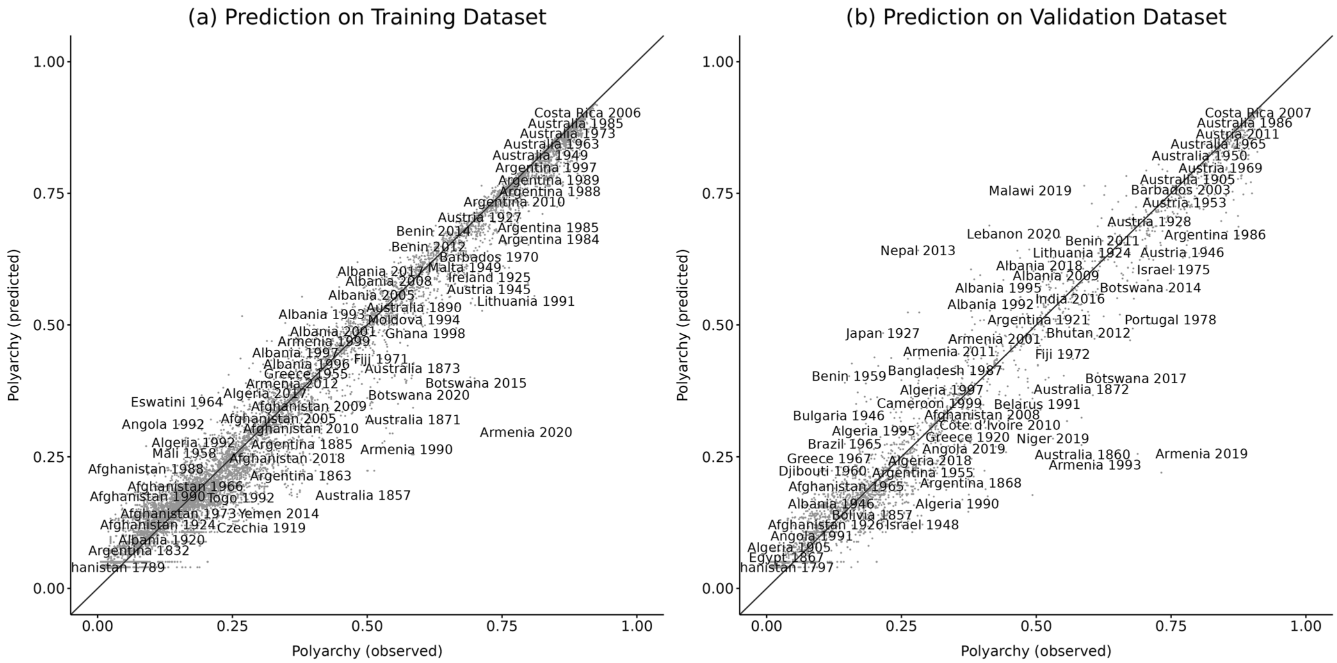

FIGURE 2 Actual versus predicted polyarchy scores. *Note*: Predicted versus actual polyarchy scores for all observations in our training dataset. The line indicates a perfect match between scores. The further points are from the line the more are they under- or overpredicted. Labels are shown for selected country-years. Predictions in the validation and test (see Appendix O, p. 36 in the Supporting Information) dataset yield similar performance. Training set with $N = 16{,}016$ (65% of V-Dem Data) and validation set with $N = 4{,}004$ (15% of V-Dem Data).

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, the 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 and not to propose an entirely novel set of observable measures.

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 .95 in the training, validation, and cross-validation sets with a mean squared error of .003 in the validation data.[8] Since the outcome, Polyarchy, ranges from 0 to 1, this is 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 new incumbent party's enormous gain in 2018, achieving 88 seats (70% of the seats in the National Assembly), is driving our model's conservatism.

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 (code available in the replication files). 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.

[8] Linear regression achieves an adjusted R^2 of. 71 for the reduced and .80 for the full set of predictors. See Appendix N (p. 33) in the Supporting Information.

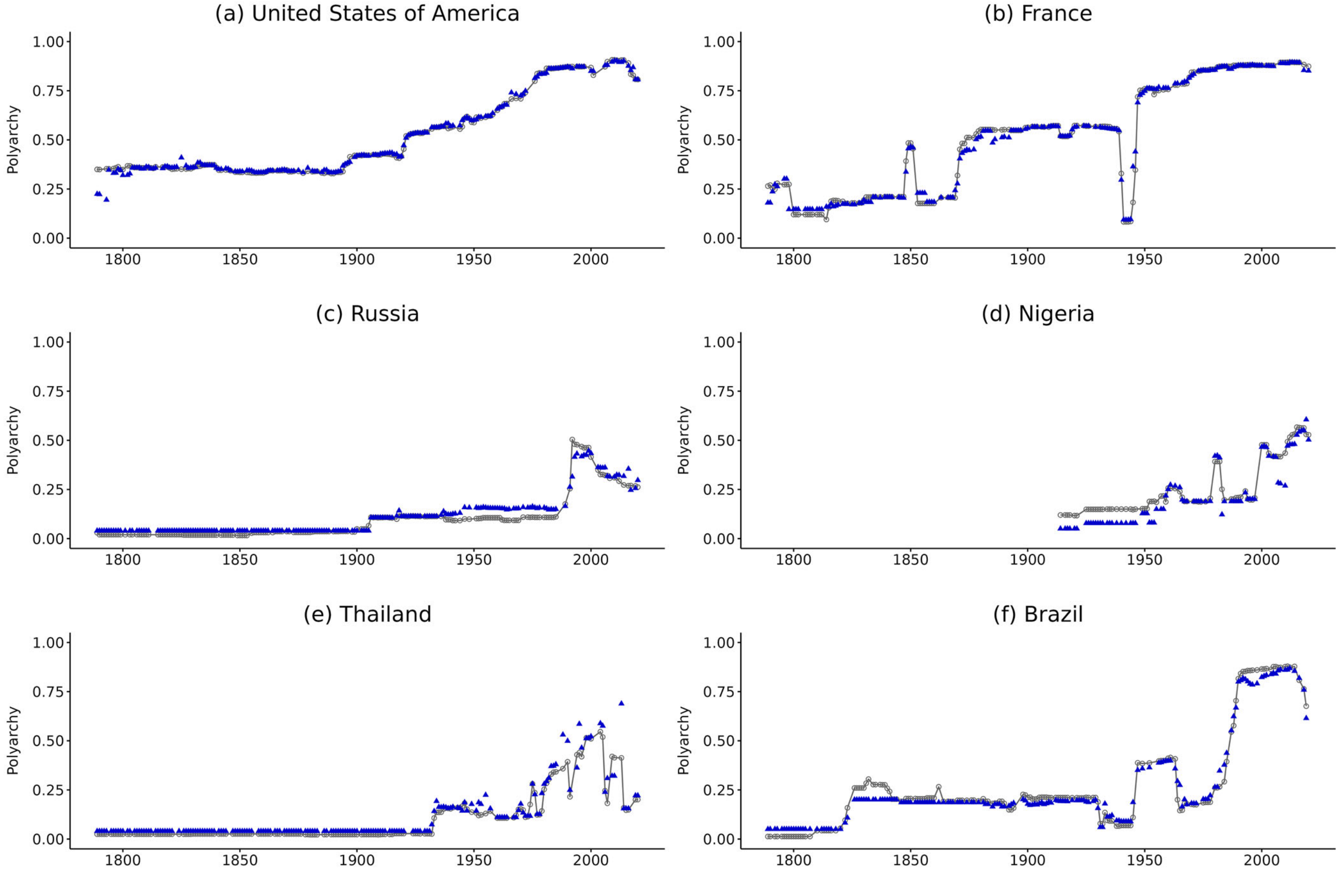

FIGURE 3 Actual and predicted polyarchy scores for selected countries. *Note*: Polyarchy scores (gray circles) and predicted scores based on the random forest (blue triangles). A complete set of country graphs is available in the Online Appendix.

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, with results posted in Table 2.

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 are scarcer or more error-prone.

Other predictors—per capita Gross Domestic Product (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 of Polyarchy's posterior distribution (for a given country-year) is also associated with greater error.[9] Evidently, we have a harder time replicating scores for Polyarchy where the V-Dem experts are themselves in disagreement. This does not necessarily mean that Z' offers a better estimate than Z. What it shows is that when the signal (Z) is noisy, the

[9] Polyarchy is aggregated from multiple sub-indicators. The posterior standard deviations therefore reflect multiple sources of measurement uncertainty, among which expert disagreement in the constituent indicators is the most important (Coppedge et al., 2022).

TABLE 2 Modeling the difference between predicted and observed polyarchy.

	(1) OLS	(2) OLS	(3) OLS, FE
Population (log)	−.077**	−.054**	−.052**
	(.013)	(.013)	(.013)
GDP per capita (log)	−.115**	−.045†	−.020
	(.026)	(.026)	(.024)
GDP per capita (log), first difference	−.277	−.145	−.274
	(.265)	(.257)	(.234)
Protestant	−.002†	−.001	
	(.001)	(.001)	
Muslim	.000	.000	
	(.001)	(.001)	
English legal origin	−.079	−.018	
	(.053)	(.057)	
Year	.000	−.000	.000
	(.000)	(.000)	(.000)
Missing observations (%)		.491**	.508**
		(.068)	(.060)
Polyarchy		.072	−.026
		(.221)	(.170)
Polyarchy, first-difference, absolute value		2.151**	2.191**
		(.222)	(.218)
Polyarchy, first-difference, absolute value, lagged		1.025**	1.069**
		(.205)	(.190)
Polyarchy, standard deviation		5.201*	5.049**
		(2.231)	(1.761)
Turnover period		−.007**	−.007**
		(.002)	(.001)
Countries	184	171	175
Years	229	227	230
Observations	14,985	13,003	14,494
R-squared	.076	.164	.150

Note: 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. Country fixed-effects (FE) included in Model 3. Intercept not shown.
$^{\dagger}p < .10$; $^{*}p < .05$; $^{**}p < .01$.

random forest estimate (*Z*′) is not as precise. Having said that, we should add that in circumstances where unreliability is associated with bias, one may interpret *Z*′ as a less biased estimate of *Z*.

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.

Model 3 focuses on variability through time, dropping predictors that register little or no change through time and adding country fixed-effects. Results are striking similar to those registered in the benchmark OLS models.

Overall, patterns evident 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 Evaluating Poten-

TABLE 3 Model-fit across a set of democracy indices.

			Normalized root mean square error	
	Range	**Scale**	**Full OSM**	**Reduced OSM**
Polyarchy	0 to 1	Interval	.05	.06
UDS	−2 to 2	Interval	.05	.05
Polity2	−10 to 10	Ordinal	.10	.12
Freedom House	1 to 15	Ordinal	.08	.09
BMR	0/1	Dichotomous	.11	.13

Notes: Goodness-of-fit statistics for five democracy indices as predicted by each observable-to-subjective score mapping (OSM). Measures are calculated on out-of-bag training samples. Full OSM: all 40 variables (see Appendix A, p. 2 in the Supporting Information). Reduced OSM: the 12–13 most important variables for that particular outcome. UDS refers to the Unified Democracy Scores (Pemstein et al., 2010), Polity2 is based on Marshall et al. 2020, and the BMR is based on Boix et al. (2013).

tial Biases, this may also account for some of the deviations between Z and Z'.

GENERALIZING THE APPROACH

The protocol described in the methodology section may be applied to any democracy index—or more broadly, to any subjective measure for which sufficient observable proxy data exists. In Appendix C (p. 7) in the Supporting Information, 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 12 or 13 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 the methodology section —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.

EVALUATING POTENTIAL BIASES

In the Section Extant Indices, we reviewed ways in which the subjective coding of democracy might be biased. Our expectation is that an approach to measurement based on observables is resistant to some, if not all, of these biases. This is not an easy matter to assess, as one must have a hypothesis about the direction of bias and a way of measuring it. In this section, we assess possible ideological biases.

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 and hence the indices that they generate (working as project directors or as coders). Thus, one might 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 no such bias appears in the Freedom House measure.

To examine this question, we enlist a new dataset (Herre, 2023) 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 toward redistributive state interventions into the economy. We employ a dummy for those classified on the right. This is the variable of theoretical interest in a series of tests presented in Table 4.

We begin by examining potential biases at the coder level, as revealed by coder-level responses to individual indicators on the V-Dem questionnaire. The most general question about electoral democracy on that lengthy survey is posed as follows: “Taking all aspects of the pre-election period, election day, and the post-election process into account, would you consider this national election to be free and fair?” Responses from 1,734 country experts (who coded this question for some portion of the contemporary era) are registered on a 5-point Likert scale. This outcome is regressed against the right-wing head of state dummy along with fixed effects for each coder and each year. Results show that these country experts coded elections as less free and fair when governed by a right-wing head of state, offering prima facie evidence of coder bias (caveats to follow).

TABLE 4 Potential ideological bias.

	Free and fair elections (V-dem)	Polyarchy minus OSM	UDS minus OSM	Polity2 minus OSM	Freedom House minus OSM
	Coder-years	Country-years	Country-years	Country-years	Country-years
	(1)	(2)	(3)	(4)	(5)
Right-wing head of government	−.020**	−.006†	−.004†	−.011*	.002
	(.008)	(.003)	(.002)	(.004)	(.003)
Unit fixed effect	✓	✓	✓	✓	✓
Year dummies	✓	✓	✓	✓	✓
Countries	173	177	177	170	176
Years	76	76	67	74	48
Observations	24,108	8,577	7,397	7,909	6,214
R-squared	.001	.002	.002	.004	.000

Note: Estimator: ordinary least squares, standard errors clustered by coder. UDS refers to the Unified Democracy Scores (Pemstein et al., 2010), Polity2 is based on Marshall et al. (2020), and the BMR is based on Boix et al. (2013).
$^{\dagger}p < .10$; $^{*}p < .05$; $^{**}p < .01$.

Subsequent tests in Table 4 are focused on democracy indices, where coder-level decisions are aggregated up to a single point estimate for each country-year and where the objective is to capture a more comprehensive measure of democracy. To identify potential bias, we subtract the OSM estimate from the original index and regress this outcome against the right-wing head-of-state dummy along with country and year fixed effects. The procedure is repeated for Polyarchy (Model 2), UDS (Model 3), Polity2 (Model 4), and Freedom House (Model 5). All variables (original indices and OSM estimates) are re-scaled from 0 to 1, so coefficients for right-wing head of state are comparable across Models 2–5.

It will be seen that right-wing governments are associated with lower scores for Polyarchy, UDS, and Polity2—relative to the OSM estimates for those indices—suggesting that coders working on these projects might be influenced by the ideological complexion of the country they are coding. No such relationship appears for Freedom House as expected. The fact that UDS registers a weaker left-wing bias than Polyarchy and Polity2 may reflect its composite nature; components of the UDS with a left-wing bias such as Polyarchy and Polity2 are presumably balanced by 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, and more importantly, we must consider 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 the Section Extant Indices.

In Appendix I (p. 17) in the Supporting Information, we describe 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.

In another publication, we demonstrate that OSMs are helpful in identifying biases in an index through time (Weitzel et al., 2024). For example, Little and Meng (2024) allege that democracy indices showing a global downward trend in recent decades are in fact registering a widespread bias: Expert coders are applying different standards, or are seeing the world differently, than they did before. While the OSM cannot detect which of several coding protocols is correct (in the sense of concept validity), it can detect breakpoints in a time series that indicate a different data-generating process is in play. Indeed, we find several breakpoints in the Freedom House scores, suggesting that changes through time may be influenced by changes to the coding protocol, rather than (or in addition to) changes in the world.

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 semisovereign 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 19th 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 semisovereign 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.

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 cov-

TABLE 5 Estimated impact of geography on democracy in varying samples.

	1789–2021			1789–2021			1946–2021		
		OSM	OSM		OSM	OSM		OSM	OSM
	Polyarchy	in-sample	full-sample	Polyarchy	in-sample	full-sample	Polyarchy	in-sample	full-sample
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Island	.127**	.114**	.003	.143**	.130**	.074**	.180**	.173**	.118**
	(.034)	(.031)	(.014)	(.034)	(.031)	(.015)	(.038)	(.037)	(.025)
Equator distance	.004**	.004**	.002**	.004**	.004**	.002**	.004**	.004**	.002**
	(.001)	(.001)	(.001)	(.001)	(.001)	(.001)	(.001)	(.001)	(.001)
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	.455	.466	.406	.484	.497	.483	.395	.399	.357

Note: Additional covariates: year, region dummies (Europe, Americas, MENA, sub-Saharan Africa, Asia), intercept. Estimator: Ordinary least squares, standard errors clustered by country, standard errors in parentheses.
$^{\dagger}p < .10$; $^{*}p < .05$; $^{**}p < .01$. (The OSM samples will be 15% larger when the test set is included.)

erage. 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.

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 is a subjective index, Z' is an OSM estimate, and ε is an error. The tricky aspect of this equation is that the error term encapsulates both coder error *and* information loss, that is, 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 cross-national analyses with democracy will want to use both versions—one as a benchmark and the other as a robustness check. 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.

OSMs are also useful tools for interrogating potential biases in subjectively coded measures as we demonstrate in Section VI with respect to ideological biases and elsewhere with respect to biases through time (Weitzel et al., 2024).

Before concluding, we 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 (p. 17) in the Supporting Information 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).

ACKNOWLEDGMENTS

The article has benefited greatly from comments at workshops and events at Aarhus University, Modeling @Alphabet Talk Series, the University of Vienna, and Colorado State University. We thank Lee Cojocaru for his valuable research assistance.

ORCID

Daniel Weitzel https://orcid.org/0000-0003-1431-2542
John Gerring https://orcid.org/0000-0001-9858-2050
Daniel Pemstein https://orcid.org/0000-0002-1144-6337
Svend-Erik Skaaning https://orcid.org/0000-0002-8587-8183

REFERENCES

Alvarez, Mike, Jose A. Cheibub, Fernando Limongi, and 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): 440–41.

Bakker, Ryan, Catherine De Vries, Erica Edwards, Liesbet Hooghe, Seth Jolly, Gary Marks, Jonathan Polk, Jan Rovny, Marco Steenbergen, and Milada Anna Vachudova. 2015. "Measuring

Party Positions in Europe: The Chapel Hill Expert Survey Trend File, 1999–2010." *Party Politics* 21(1): 143–52.

Bjørnskov, Christian, and Martin Rode. 2020. "Regime Types and Regime Change: A New Dataset on Democracy, Coups, and Political Institutions." *Review of International Organizations* 15(2): 531–51.

Boix, Carles, Michael Miller, and Sebastian Rosato. 2013. "A Complete Dataset of Political Regimes, 1800–2007." *Comparative Political Studies* 46(12): 1523–54.

Bollen, Kenneth. 1990. "Political Democracy: Conceptual and Measurement Traps." *Studies in Comparative International Development* 25(1): 7–24.

Bollen, Kenneth. 1993. "Liberal Democracy: Validity and Method Factors in Cross-National Measures." *American Journal of Political Science*, 37(4): 1207–30.

Bollen, Kenneth, and Pamela Paxton. 2000. "Subjective Measures of Political Democracy." *Comparative Political Studies* 33(1): 58–86.

Bowman, Kirk, Fabrice Lehoucq, and James Mahoney. 2005. "Measuring Political Democracy: Case Expertise, Data Adequacy, and Central America." *Comparative Political Studies* 38(8): 939–70.

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

Bühlmann, Marc, Wolfgang Merkel, Lisa Müller, and 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): 711–31.

Cardiff, Christopher F., and Daniel B. Klein. 2005. "Faculty Partisan Affiliations in all Disciplines: A Voter-registration Study." *Critical Review* 17(3-4): 237–55.

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

Coppedge, Michael, Angel Alvarez, and Claudia Maldonado. 2008. "Two Persistent Dimensions of Democracy: Contestation and Inclusiveness." *Journal of Politics* 70(3): 632–47.

Coppedge, Michael, John Gerring, Adam Glynn, Carl Henrik Knutsen, Staffan I. Lindberg, Daniel Pemstein, Brigitte Seim, Svend-Erik Skaaning, and 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, Kyle L. Marquardt, Juraj Medzihorsky, et al. 2022. "V-Dem Methodology v. 12." *Varieties of Democracy (V-Dem) Project*. https://v-dem.net/documents/55/codebook.pdf

Cutright, Phillips. 1963. "National Political Development. Measurement and Analysis." *American Sociological Review* 28(2): 253–64.

Easterly, William, and 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., and 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. https://freedomhouse.org/sites/default/files/Methodology_FIW_2015.pdf.

Gerring, John, Brendan Apfeld, Tore Wig, and 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, and Tommy Krieger. 2016. "Democracy and Growth: Evidence From a Machine Learning Indicator." *European Journal of Political Economy* 45 (Supplement): 85–107.

Gründler, Klaus, and 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, and Jerome Friedman. 2013. *The Elements of Statistical Learning*. New York: Springer.

Herre, Bastian. 2023. "Identifying Ideologues: A Global Dataset on Political Leaders, 1945–2020." *British Journal of Political Science* 53(2): 740–48.

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

Knutsen, Carl Henrik, and 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.

Little, Andrew, and Anne Meng. 2024. "Subjective and Objective Measurement of Democratic Backsliding." *PS: Political Science & Politics* 57(2): 149–61.

Little, Roderick J. A., and Donald B. Rubin. 2019. *Statistical analysis With missing data*. New York: John Wiley & Sons.

Marquardt, Kyle L., and Daniel Pemstein. 2018. "IRT Models for Expert-coded Panel Data." *Political Analysis* 26(4): 431–56.

Marquardt, Kyle L., Daniel Pemstein, Brigitte Seim, and Yi-ting Wang. 2019. "What Makes Experts Reliable? Expert Reliability and the Estimation of Latent Traits." *Research & Politics* 6(4).

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.

Munck, Gerardo L. 2009. *Measuring Democracy: A Bridge Between Scholarship and Politics*. Baltimore: Johns Hopkins University Press.

Norris, Pippa, Richard W. Frank, and Ferran Martinez i Coma. 2013. "Assessing the Quality of Elections." *Journal of Democracy* 24(4): 124–35.

Pemstein, Daniel, Stephen Meserve, and 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): 105–16.

Skaaning, Svend-Erik, John Gerring, and Henrikas Bartusevičius. 2015. "A Lexical Index of Electoral Democracy." *Comparative Political Studies* 48(12): 1491–525.

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, and 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): 251–65.

Vanhanen, Tatu. 2011. “Measures of Democracy 1810–2010.” FSD1289, version 5. www.fsd.tuni.fi/fi/aineistot/taustatietoa/FSD1289/Introduction_2010.pdf.

Weitzel, Daniel, John Gerring, Daniel Pemstein, and Svend-Erik Skaaning. 2024. “Measuring Backsliding With Observables: Observable-to-Subjective Score Mapping.” *PS: Political Science & Politics* 57(2): 216–23.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Weitzel, Daniel, John Gerring, Daniel Pemstein, and Svend-Erik Skaaning. 2025. “Measuring electoral democracy with observables.” *American Journal of Political Science* 1–17. https://doi.org/10.1111/ajps.12968