Democracy:
Global, Historical Measures Based on Observables

Daniel Weitzel
Department of Government
University of Vienna

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

Draft: 8 November 2021

Comments welcome!

A great deal of progress has been made in the measurement of democracy over the past several decades and there are now a wide range of options to choose from. However, most indices are either fully or partially dependent on judgmental coding by experts or individual scholars and their assistants. This situation is remarkable given the preference for directly observable data among many scholars and in the policy community. This perspective is based on the belief that data required by judgement-based measures "are hard, if not impossible, to obtain … [and therefore] entail coding created on the basis of inferences, extensions, and perhaps even guesses" (Cheibub, Gandhi & Vreeland 2010: 77).

Granted, some aspects of democracy, such as freedom of expression, require judgment on the part of knowledgeable coders versed in the history of a particular country (Bollen 1993; Bowman et al. 2005; Coppedge et al. 2019; Munck 2009). Schedler (2012: 33) even argues that: "If we were to renounce our judgmental faculties in the measurement of regime properties and regime dynamics, we would have to renounce the measurement of most of the most interesting regime properties and regime dynamics." However, 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. We show that most of the variability in widely used democracy indices can be captured by models constructed entirely on observable features of the world.

We are not the first to pursue this approach to measurement. Indeed, one of the very first attempts to measure democracy was based on observables (Cutright 1963). In this respect, our proposal harkens back to the inaugural era of crossnational research on democracy. However, extant objective measures of democracy 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, most indices reduce the conceptual space of democracy into binary or ordinal indices, with consequent loss of information. Third, they are limited in coverage.

Our approach seeks to combine the nuanced quality of subjectively coded democracy indices with a more objective approach to measurement. First, we gather data for a wide range of observable outcomes that seem to capture different aspects of the democratic process. Next, we employ a random forest model in which an existing democracy index (e.g., Polyarchy from the V-Dem project or Polity2 from the Polity IV project) is the outcome and the factual indicators are the predictors. The model that provides the best fit to the outcome is understood as an alternate index for that conceptualization of democracy, and can also be applied to out-of-sample cases.

The advantages of this alternate index are severalfold. First and foremost, it is free of idiosyncratic coder errors arising from

misinformation, slack, biases for or against a regime, or data entry mistakes. Second, it is free of systematic bias that may arise from coders' inferences about a country's regime status, e.g., from its recent economic performance, episodes of civil unrest, public policies (right- or left-wing), alliances (e.g., with the West or against the West), and the time-period under review (e.g., historical or contemporary). Note that insofar as these biases are widely shared they must be regarded as systematic rather than idiosyncratic. Third, it is easy to interpret an objective index insofar as changes in the index, or variation in scores across countries, are the product of a specific set of observable quantities. One can ascertain precisely which factors are driving variation. Fourth, the data collection procedure is fully transparent and replicable. Fifth, our index offers much broader coverage than any extant index (subjective or objective) because the observable features of politics in our index are fairly easier to gather and do not require in-depth knowledge of each polity. The resulting index can therefore be applied to micro-states, quasi-sovereign polities (e.g., colonies and dependencies), and now-defunct historical polities. Finally, the index is cheap to produce and easy to update.

We begin this article with a discussion of extant indices. Next, we present our methodology for measuring democracy with observables. The third section introduces a series of validity tests. We end with a brief discussion of the resulting index, offering annual coverage for most sovereign and semisovereign polities from 1789 to the present – the largest dataset of its kind.

I. Extant Indices

The most widely used measures of democracy are listed in Table 1 along with some of their key features. In this section we discuss problems of (a) coder judgment, (b) ambiguity, (c) coding observables, and (d) sample size and bias.

Table 1: **Democracy Indices**

Index	*Scale*	*Coder judgment*	*Polities*	*Years*	*Obs*
BMR (Boix, Miller, Rosato 2013)	Binary	Med	208	1800-2015	15,620
Democracy-Dictatorship ("DD") (Alvarez et al. 1996; Cheibub Gandhi, Vreeland 2010; Bjørnskov, Rode 2020)	Binary	Low	199	1946-2020	?
Political rights & Civil liberty (Freedom House 2015)	Ordinal	High	202	1972-	7,598
Democracy Barometer (Bühlmann Merkel, Müller, Weßels 2012)	Interval	Med	70	1990-	?
Polity2 (Marshall, Gurr, Jaggers 2013)	Ordinal	High	182	1800-	15,772
Unified Democracy Scores ("UDS") (Pemstein, Meserve, Melton 2010)	Interval	High	198	1946-	9,258
Lexical index of electoral democracy (Skaaning, Gerring, Bartusevičius 2015)	Ordinal	Low	224	1789-	17,020
Democracy (Vanhanen 1990, 1997, 2000, 2003, 2011)	Interval	Low	203	1810-2013	14,984
Polyarchy (Teorell, Coppedge, Lindberg, Skaaning 2019; Coppedge et al. 2020)	Interval	Med	177	1789-	22,734

Coder Judgment

Democracy is a latent concept so it is not surprising that most democracy indices – and all widely used democracy indices – rest to some degree on coder judgments (see Table 1). Coders might be outside experts, project directors, or research assistants under the direction of the principal investigators.

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 the complexities is offered in the Polity handbook (Marshall, Gurr, and Jaggers 2013: 73; quoted in Marzagão 2017: 32), 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 are asked to 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. 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 is partly reflecting whether government turnover has taken place but not exclusively so (as it is neither understood as a necessary or sufficient condition), meaning that judgment influences the score.

Given the pervasiveness of coder judgments in extant democracy indices, we must consider possible sources of error. For example, coders may use different sources of information. They may assign different weights to the selected information. They may base their judgment on irrelevant issues. Their scores can be affected by differences in coding procedures. If the same experts code all countries and all time-periods there is a problem of expertise, meaning that they are likely to rely on common perceptions rather than in-depth knowledge of the case at hand (Bowman, Lehoucq, Mahoney 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, Gerring, Glynn et al. 2020: chs 3-4). Across indices produced by different projects – e.g., when comparing V-Dem with Freedom House or Polity – one must contend with the varying perspectives of the coders recruited for these projects (Elff, Ziaja 2018). If they have different views of politics, history, or democracy this might account for varying judgments. Moreover, these differences would persist even if the instructions and coding frames for these projects were identical.

Accordingly, measures of democracy based on coder judgments are susceptible to systematic and unsystematic error, an oft-discussed issue in the literature on measuring democracy (e.g., Alvarez, Cheibub, Limongi, Przeworski 1996; Bollen 1990, 1993; Bollen, Paxton 1998, 2000; Cheibub et al 2010; Giebler 2012).

Stochastic error is to be expected, as projects to measure democracy (at-large) do not employ a great number of coders. Typically, the number is one. Although the V-Dem project enlists thousands of experts across the world, each expert typically codes only one country. This means that there is an average of five coders per country-variable-year, and just one or two coders for years prior to 1900. These are very small samples compared with

other expert surveys,[1] not to mention surveys of the mass public. 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).[2]

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 infects their 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.

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 1016).

Another sort of bias consists of *assumptions* that coders may use to reach determinations on unobservable, hard-to-judge dimensions where information is scarce. 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 in a region where democracy was not widespread, 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 of these assumptions could be true. But they could also be false. In the latter case, they will induce spurious correlations between democracy and other phenomena, e.g., peace/conflict, economic development.

---

[1] 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).

[2] Unfortunately, this is probably not a solvable problem. Because few people are intimate with the details of political life in a country across the course of a century or two – and these people tend to be very busy and uninterested in the time-consuming process of filling out a detailed questionnaire – increasing the number of coders usually entails a loss of expertise.

There is no easy way to escape this sort of bias because coders are asked to make lots of judgments and do not always have the full set of facts that would enable them to make those judgments in an unbiased way. Even if the facts are available, coders may not take the time required to marshal those facts if they assume they know what they will find. Bear in mind that coding is time-consuming and poorly remunerated (if it is remunerated at all). We suspect that most coders rely on what they already know, as it is onerous to consult secondary and primary accounts.

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 Political rights and Polity2. It is true, a fortiori, for meta-indices such as UDS. We do not know what these indices mean because we don't know all the factors that may have contributed to coder judgments about each country's scores over time.

Binary indices are 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 imprecision of a different sort.

In principle, the Polyarchy index is more interpretable, as it can be disaggregated into specific indicators. However, those 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.

Coding Observables

Mindful of these difficulties, a number of indices attempt to measure the ambient concept of democracy with minimal coder judgment. Most efforts of this nature result in binary or ordinal scales, as noted in Table 1. This means that the concept of democracy is reduced to a very small number of variables, each of which must be coded in a binary fashion – a considerable loss of information, as noted (Bollen 1990; Elkins 2000).

Equally important, indices based on observables are not quite as objective as they purport to be. Indeed, they often involve subjective judgments.

Democracy-Dictatorship. 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 autocratic until an alternation occurs. If an alternation occurs, it is recoded as democratic back to the date when the ruling party first gained power. This 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.

Vanhanen. In a series of articles and books stretching back over several decades Vanhanen (1990, 1997, 2000, 2003, 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 the largest party is automatically assigned a score of 30%. If the vote (or seat) share garnered by the largest party falls below 30% it is nonetheless assigned a score of 30%, under the assumption that any further diminution is a product of electoral system laws and is irrelevant to the quality of democracy. If candidates are not aligned with a political party, but parties are allowed, the share of the "largest party" is again assigned a score of 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 when these "reflect power relations more realistically" (2003: 57). It is not possible to tell how many observations these (and other) ad hoc coding decisions affect.

Text analysis. Another innovative approach employs text analysis, using the Wordscores algorithm (Laver et al. 2003), to derive a democracy score for countries around the world. To do so, Marzagão (2017) draws on over 6,000 English language news sources contained in the LexisNexis Academic, which provide roughly 42 million articles judged relevant to the question of regime type. The Wordscores algorithm is trained on the UDS scores for 1992 and then applied to other years up to the present, generating regime scores, and standard errors, for all countries across a two-decade period. The resulting indices are correlated with leading democracy indices at about 0.70. Of course, one might wonder what the resulting scores mean. They reflect any words in the chosen articles that help the algorithm predict variation in UDS scores across countries in 1992. Some of these words are surely central to democracy, but others may not be, or may be ancillary to that concept. There is no possibility of defining democracy and little possibility of identifying which factors are driving variation across cases or through time. It is a bit of a penumbra. One might also wonder what biases an English-language database of news sources might contain, and a correlation of . A final limitation is the coverage, which is constrained by LexisNexis, a database that currently extends back to 1980.

Sample Size and Bias

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). Still others go back to the nineteenth century but include only sizeable sovereign countries (e.g., BMR, Polyarchy, Polity, Vanhanen). No dataset includes a comprehensive set of sovereign and semisovereign units (e.g., colonies, dependencies) back to 1789, though Lexical comes close. The reason for this shortcoming, we surmise is that coding is laborious and historical information required for coding – whether subject or objective – is often difficult to find.

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 our understanding of large nation-states. For some questions this may be true. However, the exclusion of polities that are older, smaller, or non-sovereign constitutes an enormous loss of information, which is surely useful for some questions of theoretical interest (Gerring, Veenendaal 2020: ch 1).

Note that countries with a population of less than a million (the threshold Polity uses to justify inclusion in their dataset) constitute roughly ten percent of countries at the current time. Moving back in time, colonies

and other semisovereign units gain importance, constituting somewhere about half(?) of all polities – and a majority of the world's population – at the turn of the twentieth century. And defunct states like Bavaria were just as important at the time, and just as sovereign, as many states that managed to endure. In comparative politics, as in international relations, we need to understand the losers as well as the winners.

A survey of individuals based on a non-random sample of 150 people would not be regarded as a reasonable basis for making inferences about a larger population. And yet, this is standard practice in the world of crossnational analysis. It is not clear whether, or to what extent, a comprehensive sample (a "census") of polities might alter our view of democracy's causes and effects. We will not know until we look at these cases, and we cannot do so unless they are measured systematically.

Adding cases will also assist in the search for greater *internal* validity. Insofar as the sample of polities can be expanded it provides longer time-series and more extensive cross-case leverage. These will not overcome all the difficulties encountered in reaching causal inference with observational data. But it will surely help.

II. Methodology

Three criteria inform our search for indicators: (a) observability, (b) coverage, and (c) relevance for the concept of democracy (variously understood). We limit ourselves to indicators that reflect or embody democracy, understood as a set of institutions rather than an assemblage of attitudes and values.

Observability means that a feature that 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 (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 current codings. However, these cases are rare and fall primarily in the early part of our period of observation or with respect to very small states whose histories are not well-preserved or well-researched.

We recognize that there may be some disagreement over *which* observable features of polities to focus on. Our goal is to include all observable features that are measurable across the vast set of extant polities and potentially indicative of the state of democracy (according to a variety of conceptualizations). We exclude others that, while they might enhance prediction, are not constitutive or reflective of democracy. For example, per capita GDP is excluded because it is not generally regarded as a component of democracy.

Chosen variables focus on political institutions rather than on economics, sociology, or culture. We do not attempt to measure attitudes like trust, belief in democracy, and so forth, as it is difficult to do so even where surveys exist and impossible where they do not. Likewise, attitudes occupy an uncertain position in the measurement of democracy, which is usually conceptualized as a set of institutions. (If trust is lower in one polity than another, but they are institutionally speaking identical, it is not obvious that the low-trust polity is less democratic.)

Chosen indicators are listed in Appendix A along with detailed definitions and sources. Many are drawn from the V-Dem project (which includes many factual variables such as suffrage rights), with additional coding by the authors that extends these variables to cover states outside the V-Dem universe. A total of forty-four variables are included, though many of these are variations on a smaller number of core variables (e.g., count variables, logarithmic transformations, and so forth).

Since the selection of indicators is an important part of this exercise we conduct a series of robustness tests in which individual variables are removed and the model is recalibrated. As it happens, the variations that result from these serial omissions are very slight (see Section III). Accordingly, we can state fairly confidently that the results reported in this study are not contingent upon the inclusion of any single variable.

The impact of possible *omissions* is more difficult to address. It is possible that important observable indicators are inadvertently omitted from our dataset. However, the extremely tight fit obtained from the set of chosen variables assures that any additional variables are unlikely to change index scores by very much. There isn't much unexplained variance left to explain.

A Random Forest Model

The bane of composite indices is aggregation. Every democracy index struggles with this problem. Some offer formulas for aggregating component indicators (Polyarchy, Polity2). Others establish categories, each with separate criteria (Lexical, Freedom House, BMR, DD). A third approach enlists principal components analysis (Coppedge et al. 2008) or IRT models (Marquardt, Pemstein 2018). A fourth approach starts with the identification of polities that are assumed to be highly democratic or highly autocratic, from which machine-learning algorithms build a model of attributes that explain variation along this scale (Gründler, Krieger 2016).

All of these approaches are defensible and none clearly superior, which accounts for the survival of such radically different approaches to aggregation. We offer no solutions to this eternal problem. 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, as discussed.

To aggregate across these indicators we rely on a random forest model (Breiman 2001, Hastie et al. 2013). This sort of estimator is ideal for investigating multiple predictors and interactions among them, differentiating those with strong predictive power from those that are redundant. 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 makes them an excellent and widely used model (Hastie et al. 2013). A random forest, both for continuous (regression) and discrete (classification) response variables, generates a large amount of individual decision trees that partition the data (Breiman 2001). Each of these decision trees only includes a random sample of predictors, never the full set. This makes the random forest model, among other things, robust to highly correlated variables. However, it also makes each individual decision tree more likely to be noisy. The process of sampling from the predictors at each node allows the model to learn about the optimal split decisions to partition the data. In the end, the random forest model averages over all decision trees that were generated to produce a weighted result that can be used for prediction (Hastie et al. 2013). In addition to the prediction we also receive, for classification tasks, error rates in the classification task, as well as, variable importance measures for all models.

We build a dataset of 44 objective measures of democracy in a country. These include the vote shares of the top three parties in legislative election or top two candidates in presidential elections, the age of political parties, sovereignty of a territory, or the extension of the suffrage.[3] In total we have 24,128 country-year observations for which we can predict Polyarchy values. We split this dataset into three parts. A training, a validation, and a test set. The training set consists of a random subset of 65% of the total observations and is used to train our random forest model. In this data set the model learns about the relationship between our target variables, the democracy measures, and all the objective measures we collected. We iteratively test the performance of the trained model on the validation set, which is a random sample of 15% of the total observations. Finally, the remaining 20% are the test set. This data set will in the future be used to assess the final model performance and has not yet been used. The dataset has a considerable share of missing values and in order to mitigate this problem we apply a K-nearest neighbor imputation that omits zero variance variables. The test set is separated from the training and the validation set before the imputation to prevent leaking of information from the test set into the training and validation set. 26 variables are fully imputed and included in the model. We estimate a random forest model with 15 folds for K-fold cross-validation, in order to improve the predictive performance of our model. In total we estimate 260 (number of variables*10) trees, allowing for a maximum tree depth of 20. Due to the continuous nature of our dependent variable Polyarchy (it can take any value between 0 and 1) we use random forest regression. We also apply

[3]A complete list of variables and their description is given in Appendix A.

regression for the UDS democracy measure examined later on (ranging from approx. -2 to 2). The Freedom House Political Rights, BMR, and Polity2 measures are treated as ordinal and we use random forest classification. Our random forest performs rather well. The model produces R squared values of 0.94 in the training, validation and cross-validation sets with a mean squared error (MSE) of 0.004 in the validation data. Since the outcome, Polyarchy, ranges from 0 to 1 we consider this a very low average squared difference between the predicted and the observed values. The model has not yet been validated on the test data set. In Figure 1 we plot the predicted against the actual Polyarchy values as coded by V-Dem and label a select set of observations. Overall, the prediction seems to work very well. The distribution of the points is near the 45 degree line. Some instances, such as Armenia in 2020, are being 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 models conservatism.

In order to assess our models predictive performance further we generated country-year plots of the predicted vs the actual values for all countries in our sample. The complete set of plots is in Appendix C. In Figure 2 we present six cases that reflect a variety of political and electoral systems, heterogeneity to the stability and length of democracy, and also include countries with transitions into and out of democracy. Overall the model performance is highly satisfying. The fit is evident as illustrated by the overlap between dots and triangles. For the majority of country-years, they are virtually identical. Only in Nigeria before 1951 and a couple of years in the late 1980s, there are notable differences. The predictions for the U.S., a long-lasting and very stable democracy are very much matching the actual values coded by V-Dem. In France our model is able to follow the strong variations in Polyarchy scores over time but appears to underpredict more recent years. Russia, on the other hand, is predicted with little error. In Nigeria we face the issue of consistent underprediction up until the more recent years. Although, it should be remarked that the model does well in following the general trajectory of the country. Thailand and Brazil, the last two countries we included are also predicted very well. This is impressive since in both cases the Polyarchy scores as coded by V-Dem vary substantially over short periods of time.

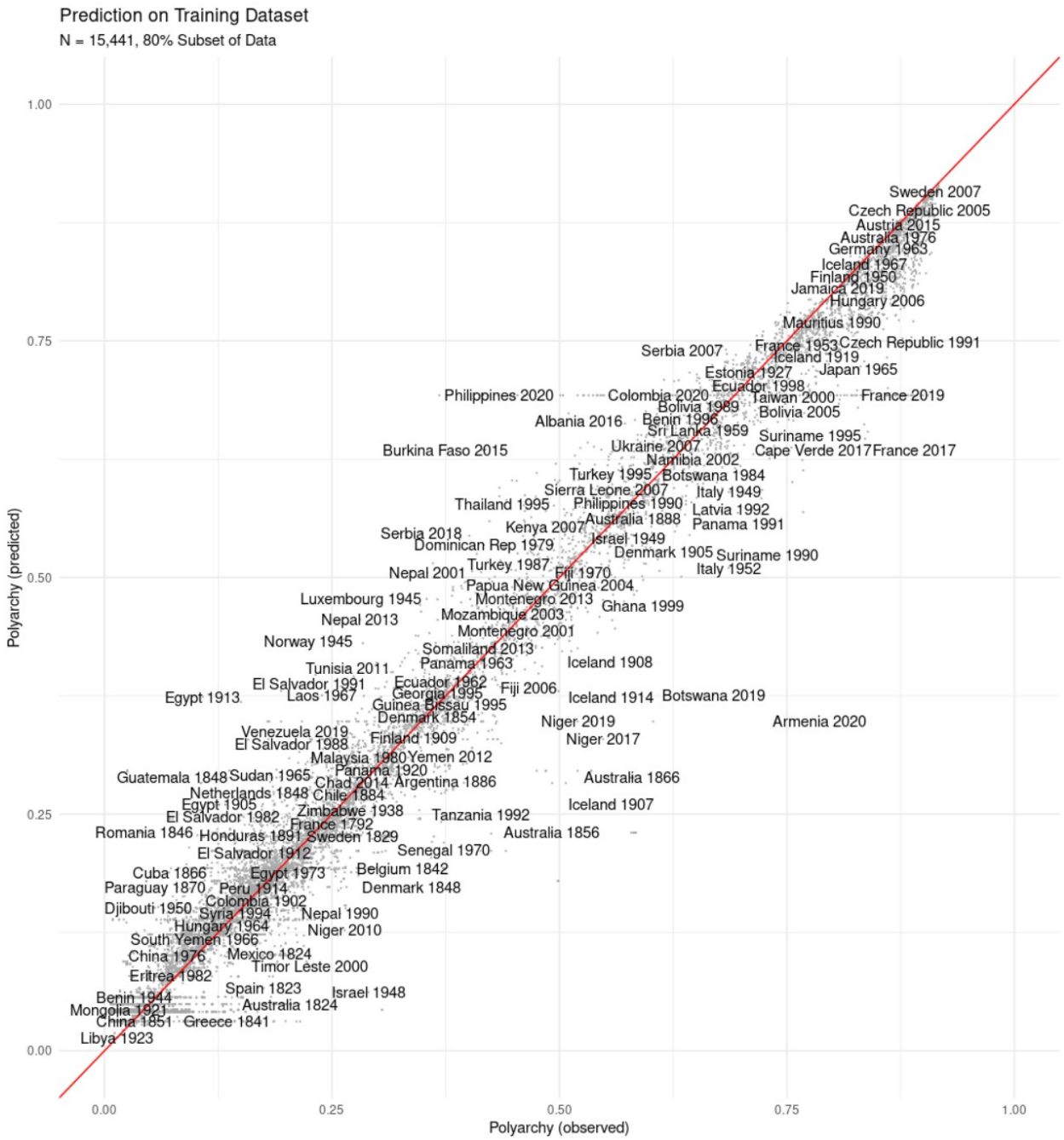

Figure 1: Predicted vs. V-Dem Polyarchy Values

Note: Shown are predicted vs actual Polyarchy scores as they were coded by V-Dem for all observations in our training dataset. The red 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 used for model performance evaluation.

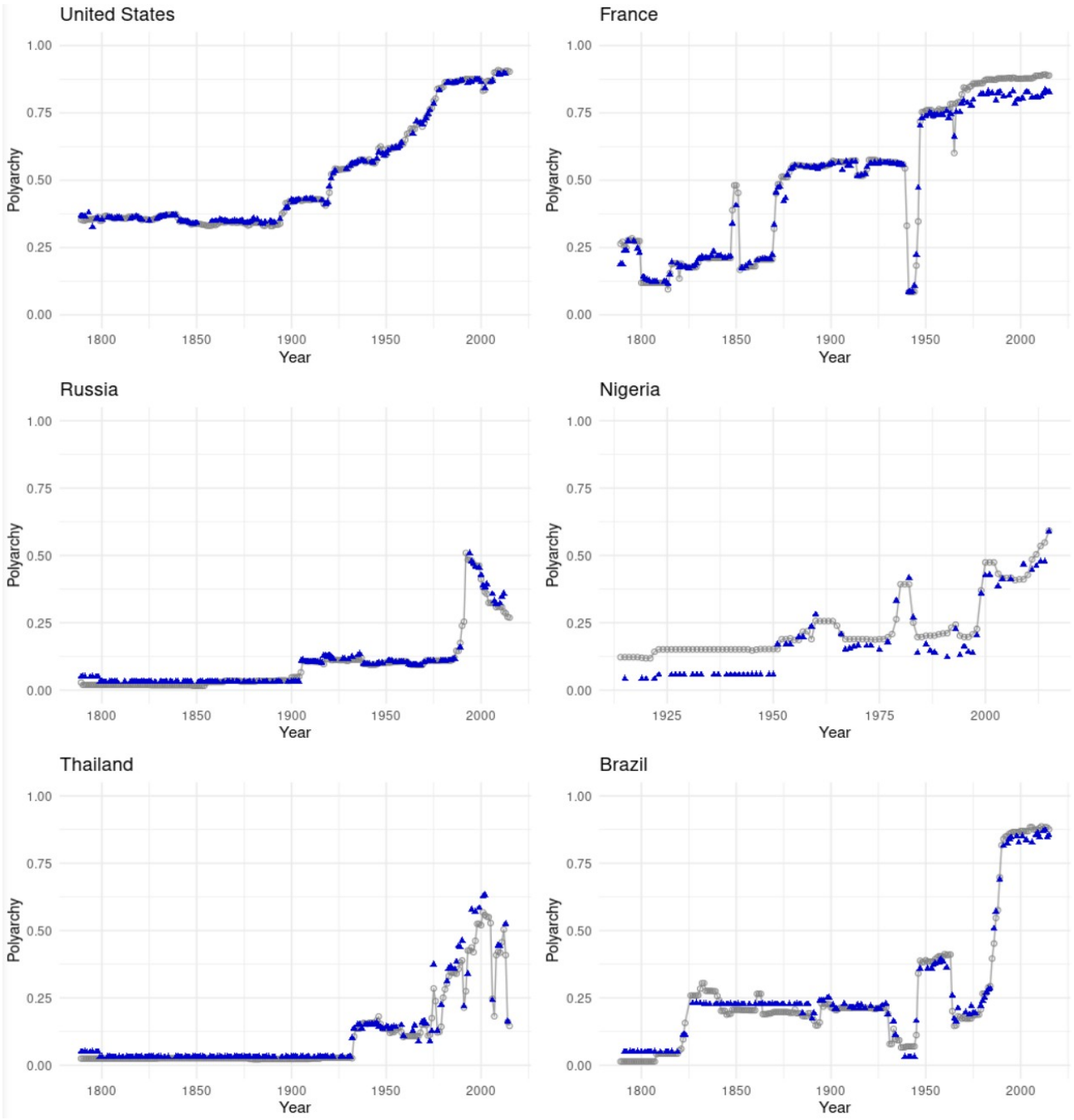

Figure 2: Predicted vs V-Dem Polyarchy Scores, Selected Countries

Note: Predicted and actual Polyarchy scores as coded by V-Dem for six countries. The gray circles are the V-Dem Polyarchy values connected by a line. Blue triangles indicate predicted Polyarchy values based on objective measures of democracy in the training dataset. The countries have been selected to represent the algorithms ability to deal with a variety of electoral situations.

But which indicators among the many included in the model tend to do the heavy lifting? Figure 3 provides an overview of the fifteen most important indicators ranked according to their importance for the distributed random forest. The measure calculates the relative importance of a variable by incorporating information on whether it was selected to split on during tree building and how much the squared error over all trees decreased as a result of the inclusion of this variable. Interestingly, four of the five best predictors are not from the V-Dem database and have thus not been used to calculate the original Polyarchy scores.

The indicator showing the highest importance value is the dichotomous Turnover Period variable *(turnover_period*) from LIED. It indicates whether a particular country-year is part of a period between an initial electoral government alternation (as indicated by a turnover event) and an interruption of electoral practices (multi-party elections interrupted/no longer on track). The prominence of this indicator suggests that Przeworski and collaborators were right in arguing the value of this distinction for measuring democracy, even though government turnover is not a defining feature of democracy. The second most important indicator, the variable indicating the existence of multi-party legislative elections (*multi-party_legislative_elections*) from LIED, reflects whether the lower house (or unicameral chamber) of the legislature is (at least in part) elected by voters facing more than one choice. Only on the third place we find a V-Dem indicator. It reflects the extension of suffrage (*v2asuffrage*). The importance of this indicator is not surprising since suffrage is one of the constituent subcomponent that the original index scores are based on). Next in the ranking is the executive elections indicator from LIED. It captures whether the chief executive is either directly or indirectly elected (i.e., chosen by people who have been elected).

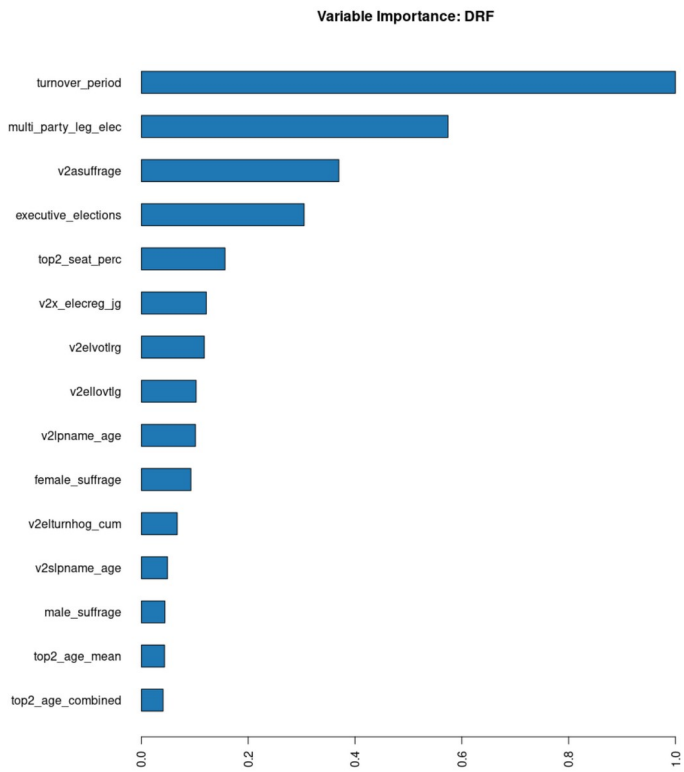

Figure 3: What matters? The Importance of Variables

Note: Shown are the top 15 variables with the highest importance scores for the distributed random forest regression with Polyarchy scores as coded by V-Dem as the target.

On the fifth place is an indicator representing the share of seats in the lower or unicameral house held by the top two parties in the last election (*top2_seat_perc*). Only thereafter, we find the V-Dem electoral regime index (*v2x_elecreg_jg*), which plays an important role in the construction of the Polyarchy index. It is interesting to note that among these fifteen indicators ranked below, about half directly tap into constitutive aspect of democracy (the presence of multi-party elections with inclusive suffrage), while the rest referring to matters such as turnover, seat share, and party age only indirectly reflect the level of democracy.

We also test our ability to use observable indicators to predict democracy classification on other indices. In Figure 4 below we replicate our initial random forest model with the democracy measures from UDS (mean), BMR, Freedom House Political Rights, and Polity2 as outcomes. Plotting the predicted against the actual values as they were coded by the different projects we can see that our approach does fairly well in all instances. Our predictions for the UDS (mean) scores, another continuous measure of democracy, follow the 45 degree line fairly smoothly and, as we report in Table 4, the model has an out-of-bag training sample MSE of 0.04 and an R squared of 0.95. Considering that the UDS measure is based in measures that have a high reliance on subjective coder judgments on the

classification of countries the high performance of our model, which only relies on objective measures, is interesting.

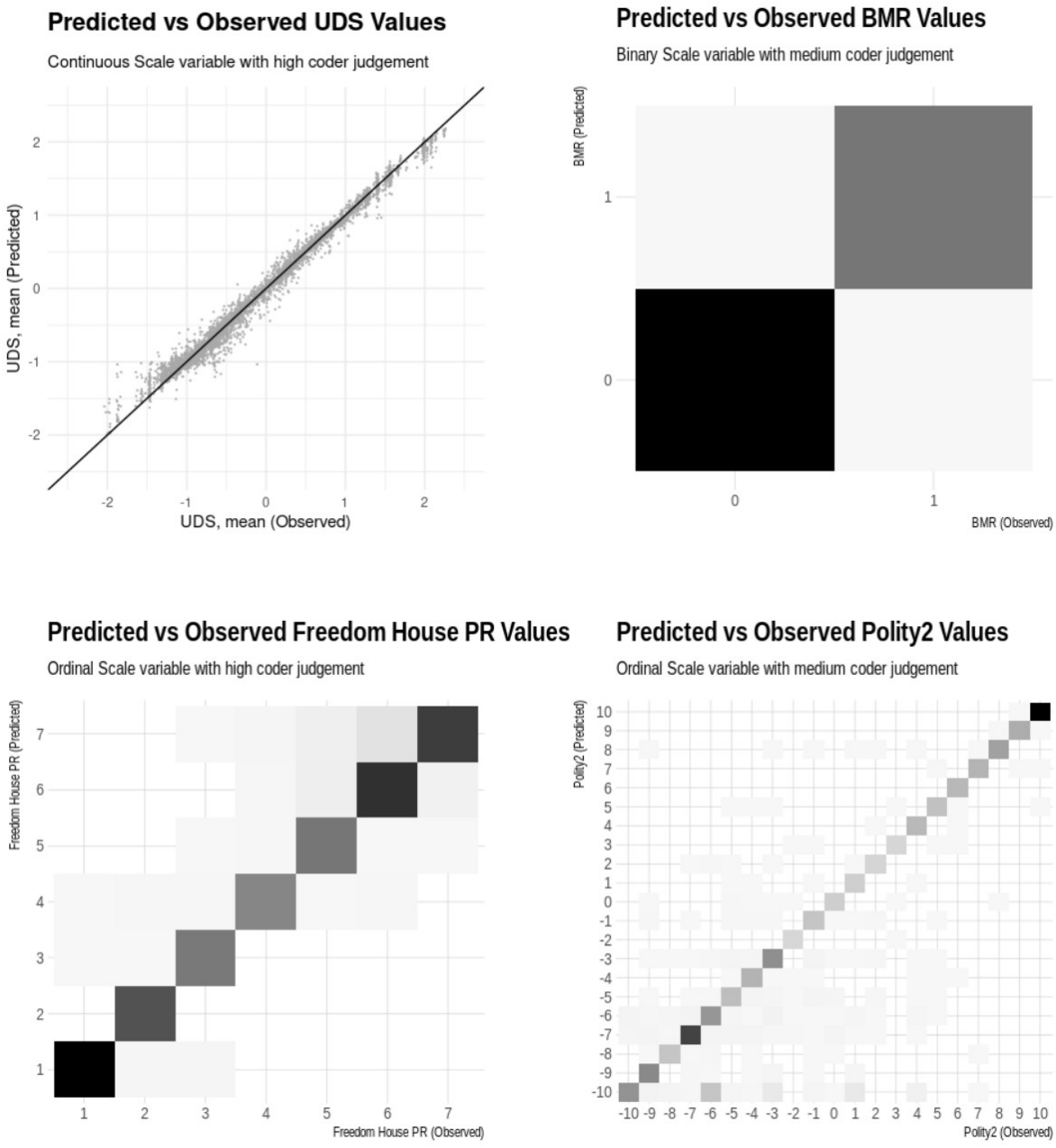

Figure 4: Predictions for Other Indices

Note: Predicted vs actual values as coded by four other indices. These indices vary in their nuance from binary to continuous measures, the required coder judgment from medium to high (see Table 1), and the concept that they are measuring.

The other three measures are treated as ordinal and we switch from regression to classification. BMR is a binary measure of countries into democratic or authoritarian. In Figure 4 we plot the predicted against the actual values as coded by BMR and in Table 2 we list the cell totals and the error rate for our prediction with objective measures. Overall, we classify 10,822 country years and have an error rate of 0.017. 91 democratic country years are classified by us as autocratic country years and 85 autocratic country years as democratic. This total misclassification of 177 units make up only a small share compared to the 7,041 correctly classified autocratic and 3,404 democratic country years.

		Predicted		**Error**	**Rate**
		0	**1**		
Observed	**1**	92	3,404	0.03	92/3,496
	0	7,041	85	0.01	85/7,126
Total		7,133	3,489	0.017	177/10,822

Table 2: Predictive Performance BMR

The third democracy measure we use to train our model on is the Freedom House Political Rights index. This is a categorical variable ranging from 1 to 7. In Figure 4 we can see that the model overall does well in predicting the actual values as coded by Freedom House. However, we do have a non-trivial amount of predictions that fall under or above the main diagonal. Table 3 reports the cell totals of the confusion matrix as well as the row and overall error rate. Here we can see that the row as well as the overall error rate is considerable higher than for the BMR measure. The goodness-of-fit statistics in Table 4 support this further. For the out-of-bag training samples our mean squared error is 0.22 and the mean per-class classification error is 0.24. Treating the Freedom House index as continuous and using random forest regression does not yield substantively different results (see Appendix C). The predictions of the measure with objective variables is not as successful as with other measures of democracy.

		Predicted						**Error**	**Rate**
		1	**2**	**3**	**4**	**5**	**6**	**7**		
Observed	**7**	0	2	1	10	20	119	637	0.19	152/789
	6	1	1	4	13	60	671	160	0.27	239/910
	5	4	3	5	43	356	222	53	0.38	218/574
	4	6	13	42	299	54	29	17	0.35	161/460
	3	27	47	374	47	8	6	1	0.27	136/510
	2	60	568	34	7	4	6	1	0.16	112/680
	1	1,004	28	6	4	0	0	0	0.04	38/1,042
Total		1,102	661	466	423	502	942	869	0.21	1,056/4,965

Table 3: Predictive Performance Freedom House PR

Lastly, we predict the Polity2 measure, which is a categorical measure with twenty distinct categories. Figure 4, once more, shows the predicted vs. observed values and Table 4 reports the goodness-of-fit statistics. For the

out-of-bag training samples our mean squared error is at a small 0.31 compared to the 20 point dependent variable. Similarly the mean per-class classification error is at 0.29. With an R squared of 0.99 the model performs rather well in the prediction. As a robustness check we also treated Polity2 as a continuous measure and we arrive at similar results (see Appendix C).

Measure	Scale	MSE	MCE	Rsq.
Polyarchy	0 to 1	0.004		0.94
UDS	-2 to 2	0.04		0.95
BMR	0/1	0.02	0.02	0.92
FH PR	1/7	0.22	0.24	0.96
Polity2	-10/10	0.31	0.29	0.99

Table 4: Goodness of Fit Statistics

Note: Shown are goodness of fit statistics for the five democracy measures we predict with objective measures. Measures are calculated on out-of-bag training samples. Continuous variables are indicated by “to”, ordinal variables by “/”. We report the scale the measure is on, the mean squared error, the mean per-class classification error where applicable, and the R squared.

For each of these models we also produced variable importance plots similar to the one presented in Figure 3 (see Appendix B). Here the variation in which variables matter the most and how the importance is distributed across all variables in the model is interesting. For BMR (Figure A2-2) the turnover period variable dominates the importance scores. While the existence of multi-party elections and the seat share of the largest party in the last election to the lower house also matter they (and the other predictors) are not as central as the turnover measure. The Freedom House Political Rights measure (Figure A2-3) has a rather equally distributed variable importance plot. While turnover and years since sovereignty matter the most all other measures are following very close in their importance. For Polity2 (Figure A2-3) the sovereignty variable has the highest importance score. All other variables have an almost equal, and considerable smaller effects. Lastly, UDS (Figure A2-4) is dominated by turnover period, multi-party elections, and the seat shares in the lower house. All other variables have minuscule importance scores.

Overall, we can say that predicting existing democracy scores, which combine objective and subjective coding, with a random forest model trained only using objective measures of the features of democratic countries does extraordinarily well. The model is able to predict the indices coded by Polyarchy, BMR, and Polity2 with a high reliability. However, it does struggle with the Political Rights index coded by Freedom House. Here we suspect that, while all indices are highly correlated with each other (see Figure 5), the measure of political rights is nevertheless distinct from all other measures used. For example, the BMR index measure whether a country is a democracy or an autocracy. The Freedom House Political Rights Index consists of components that likely distinguish democracies

from autocracies but that the variables we include in our models might not be able to predict as well.

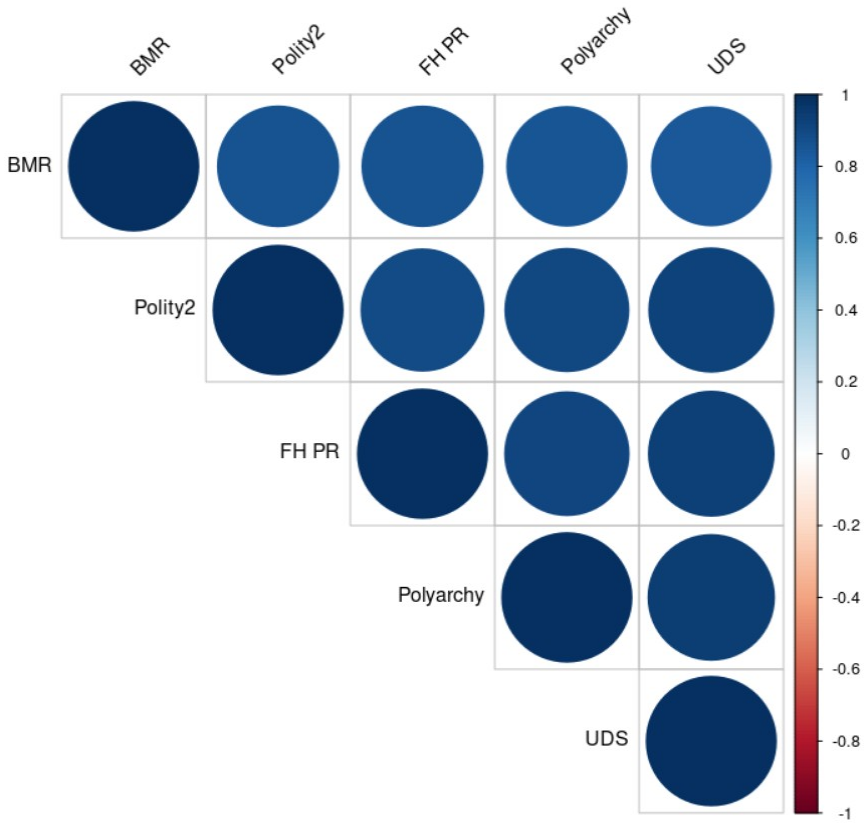

Figure 5: Correlation of Democracy Indices

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. We can see that all indices are positively correlated with each other.

III. Discussion

Indices of important concepts do not arise spontaneously from the ether. Authors construct them, and this involves judgments. When one says that a judgment is "subjective" one implies that other authors (or observers) might have made a different decision. When one says that a decision is "objective" one implies that others would have made the same choice. So, the notions of subjectivity/objectivity are similar to reliability in measurement. Since reliability is a matter of degree, our use of these terms must be understood accordingly. No index is perfectly objective or subjective.

It is important to recognize that *many* decisions are required in order to compose an index for a complex, latent concept like democracy. At the

very least, one must define it, measure its components, and aggregate the resulting indicators (if more than one).

Our approach offers an objective strategy to measurement while side-stepping questions of conceptualization and aggregation. Thus, if the reader is reasonably content with the way in which an existing index (e.g., Polyarchy) defines and measures democracy we offer a way to purge that index of coder bias, while also providing broader coverage.

There is, unfortunately, a cost. This can be represented formally in a simple model:

$$I_S = I_O + \varepsilon$$

where I_S = a subjective index, I_O = an objective index derived from observables through our suggested procedure, and ε = error. The tricky aspect of this model 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 way of distinguishing between error and information loss.

Moreover, information loss is bound to affect 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. It would be nice if we could identify those "missing" elements; if so, we could reframe our index in a narrower fashion, avoiding mis-interpretation.

Our commonsense conclusion is that an objective index arrived at in the fashion outlined in this paper is correctly regarded as superior to the original (subjective) version in some respects (absence of coder error) 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.

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 mean many things. 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 concept of democracy that is unbiased, or less biased. We must consider whether the observable aspects of democracy bias the project in a particular way, and in what direction this might be.

There may be biases associated with an observable index when it is used as a predictor or an outcome in a causal model. For example, if one is using such an index to test the relationship between democracy and growth, and the index 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

index 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 index the greater the prospect for circularity between a predictor and an outcome. That is why we intend to produce several versions of the index, one of which will be a parsimonious model including only the most important factors. 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 conscious. By contrast, with an objective index 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).

References

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

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* 21, no. 1: 143-152.

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

Boese, Vanessa A. 2019. "How (not) to measure democracy." *International Area Studies Review* 22.2: 95-127.

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. 1993. “Liberal democracy: Validity and method factors in cross-national measures.” *American Journal of Political Science*, 37(4), 1207–1230.

Bollen, Kenneth, Pamela Paxton. 1998. “Detection and determinants of bias in subjective measures.” *American Sociological Review* 63(3), 465–478.

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. "Random forests." Machine learning 45.1 (2001): 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.

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." *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.

Elff, Martin, Sebastian Ziaja. 2018. “Method factors in democracy indicators.” *Politics and Governance* 6(1): 92–104.

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

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, Wouter Veenendaal. 2020. *Population and politics: The impact of scale*. 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.

Giebler, Heiko. 2012. "Bringing methodology (back) in: Some remarks on contemporary democracy measurements." *European Political Science* 11.4: 509-518.

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

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. (2013). The elements of statistical learning. Vol. 1. No. 10. New York: Springer series in statistics.

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* 97:311-331.

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

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.

Marzagão, Thiago. 2017. "Automated democracy scores." *Brazilian Review of Econometrics* 37.1: 31-43.

Norris, Pippa, Richard W. Frank, Ferran Martinez i Coma. 2013. "Assessing the quality of elections." *Journal of Democracy* 24.4: 124-135.

Office of the High Commissioner for Human Rights. 2012. *Human rights indicators: A guide to measurement and implementation*. Geneva: OHCHR.

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

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.

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, no. 1: 71-95.

Vanhanen, Tatu. 1990. *The process of democratization: A comparative study of 147 states, 1980-88*. New York: Crane Russak.

Vanhanen, Tatu. 1997. *Prospects of democracy: A study of 172 countries*. London: Routledge.

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

Vanhanen, Tatu. 2003. *Democratization: A comparative analysis of 170 countries*. London: Routledge.

Vanhanen, Tatu. 2011. "Measures of democracy 1810–2010." *FSD1289, version* 5.

Appendix A: Codebook

Democracy Indices

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

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 et al. (2013), Boix et al. (2018). *Scale:* binary. *e_boix_regime*

Freedom House Political Rights. 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. *Source:* Freedom House (2018). *Scale:* ordinal. *e_fh_pr*

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 and Jaggers 2020). *Scale:* ordinal. *e_polity2*

UDS. Mean estimate from IRT model combining…. *Source:* Pemstein et al. (2010). *Scale:* continuous. *e_uds_mean*

Variables based on election results

Sources: Nohlen et al. (1999, 2002, 2005, 2010); Chronicle of Parliamentary Elections (IPU), Wikipedia entries.

Largest party seats. Share (%) of seats in the lower (or unicameral) chamber of the legislature obtained by the largest party. *Scale:* Interval. *v2ellostsl*

Largest party votes, legislature. 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*

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*

Total number of independents v2elinds

Independents, legislature. 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. 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*

Top two parties, monopoly. Dummy variable indicating whether the top two parties in the lower house gain more than 2/3 of the votes. *Scale:* dichotomous. *top2_monopoly*

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

Age, largest party. Age of the largest party in the lower or unicameral chamber of the national legislature. Party size is determined by vote share in the last election. Age is measured as the length of time (years) a party is continuously (without interruption) among the top three vote-getters. *v2lpname_age*

Age, second largest party. Age of the second largest party in the lower or unicameral chamber of the national legislature. Party size is measured on vote share in the last election. Age is measured as the length of time (years) a party is continuously (without interruption) among the top three vote-getters. *v2slpname_age*

Age, third largest party. Age of the third largest party in the lower house. Party size is measured on vote share in the last election. Age is measured as the length of time (years) a party is continuously (without interruption) among the top three vote-getters. *v2tlpname_age*

Age, top two parties, combined. Combined age of the top two parties. Party size is measured on vote share in the last election. Age is measured as the length of time (years) a party is continuously (without interruption) among the top three vote-getters. *top2_age_combined*

Age, top two parties, mean. Mean age of the top two parties in the last election to the lower house. Age is measured as the length of time (years) a party is continuously (without interruption) among the top three vote-getters. *top2_age_mean*

Lexical variables

Sources: PIPE (Przeworski 2013), Skaaning et al. (2015), with additional coding by the authors.

Male suffrage. *male_suffrage*

Female suffrage. *female_suffrage*

Executive elections. *executive_elections*

Legislative elections. *legislative_elections*

Multi-party elections. Dummy variable indicating whether there were multi-party elections. *multi_party_leg_elec*

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.

Misc variables

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_Cojocaru*

Sovereignty, first decade. Coded 1 if the first decade of sovereignty. *Source:* authors. *Scale:* dichotomous. *sovereign_Cojocaru_firstdecade*

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*

Election HOG turnover. 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:* Continuous. *v2elturnhog_cum*

Election HOS turnover. 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. Source(s): Henisz (2000; 2002); Lentz (1994; 1999); worldstatesmen.org; V-Dem Country Coordinators. *Scale:* Continuous. *v2elturnhos_cum*

Suffrage. The share (%) of enfranchised adults older than the minimal voting age who are legally allowed to vote. *Sources:* Bilinski (2015); Chronicle of Parliamentary Elections (IPU); Nohlen et al. (1999, 2002, 2005, 2010); constituteproject.org. *Scale:* Continuous. *v2asuffrage*

Appendix B: Variable Importance Scores for Other Indices

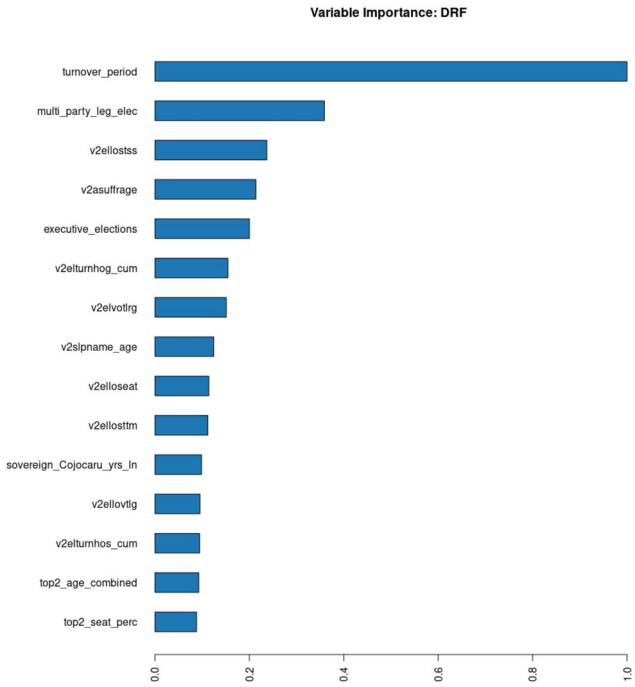

Figure A2-1: BMR Variable Importance

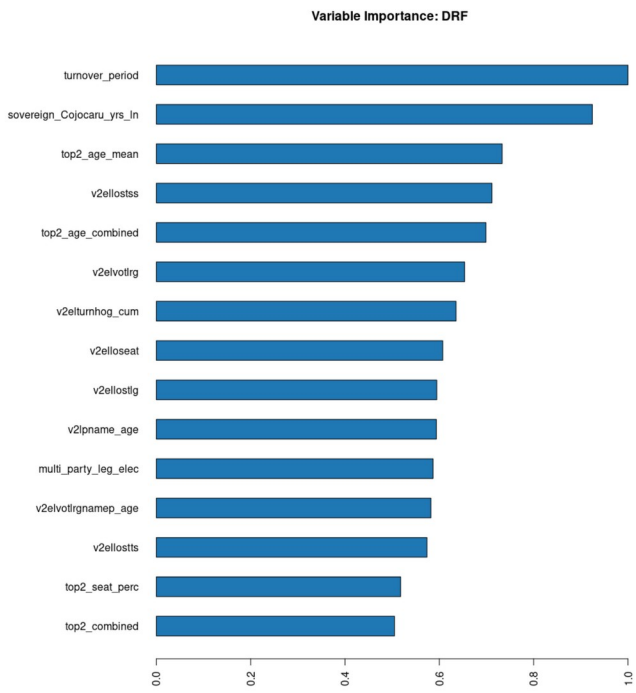

Figure A2-2: Freedom House Political Rights Variable Importance

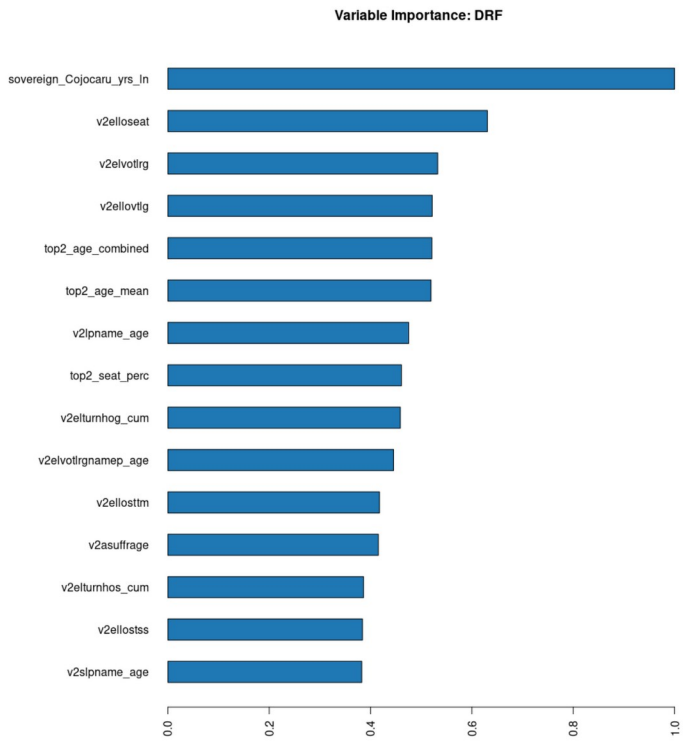

Figure A2-3: Polity2 Variable Importance

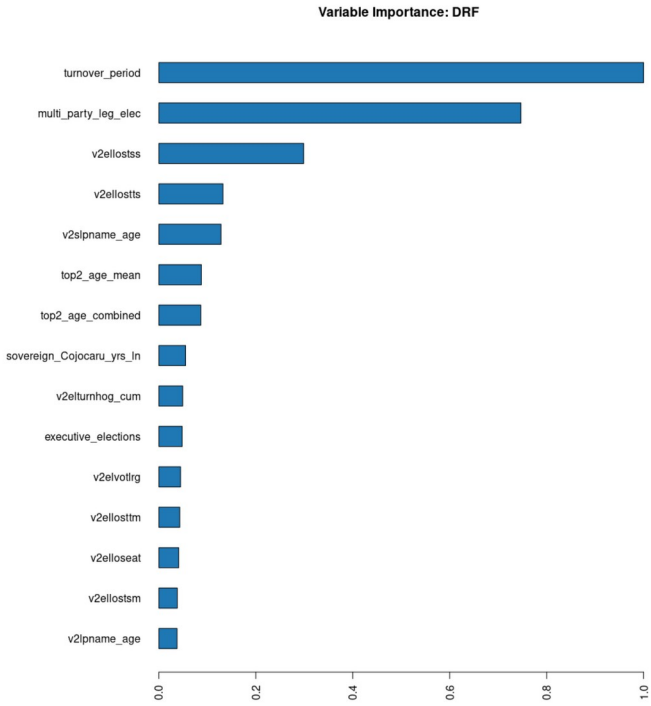

Figure A2-4: UDS Variable Importance

Appendix C: Treating Categorical Outcomes as Continuous

In the main text we treat the Polity2 and the Freedom House Political Rights measures as categorical. In this appendix we replicate the models from the main text by treating the two measures as continuous. As can be seen in Figures A3-1 and A3-2 below the results do not change much and the patterns in the predictions are fairly similar.

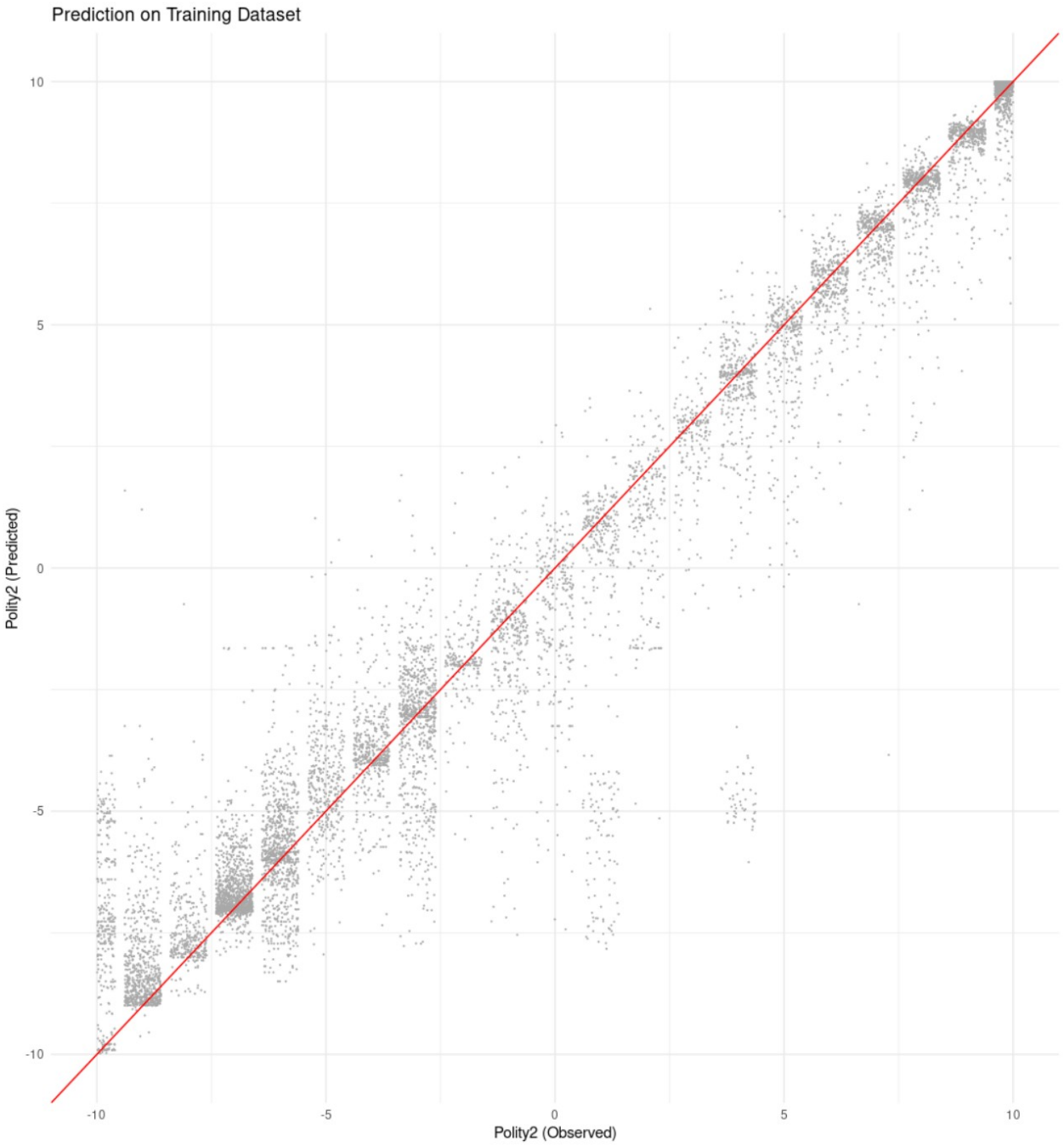

Figure A3-1: Treating Polity2 as Continuous

Note: Shown are predicted vs actually observed Polity2 values as coded by the POLITY Project. We show values for all observations in our training dataset. The red line indicates a perfect match between predicted and coded scores. The further points are from the line the more are they under- or overpredicted. The Polity2 variable is treated as continuous and random forest regression is used to predict. Points are jittered for better visibility.

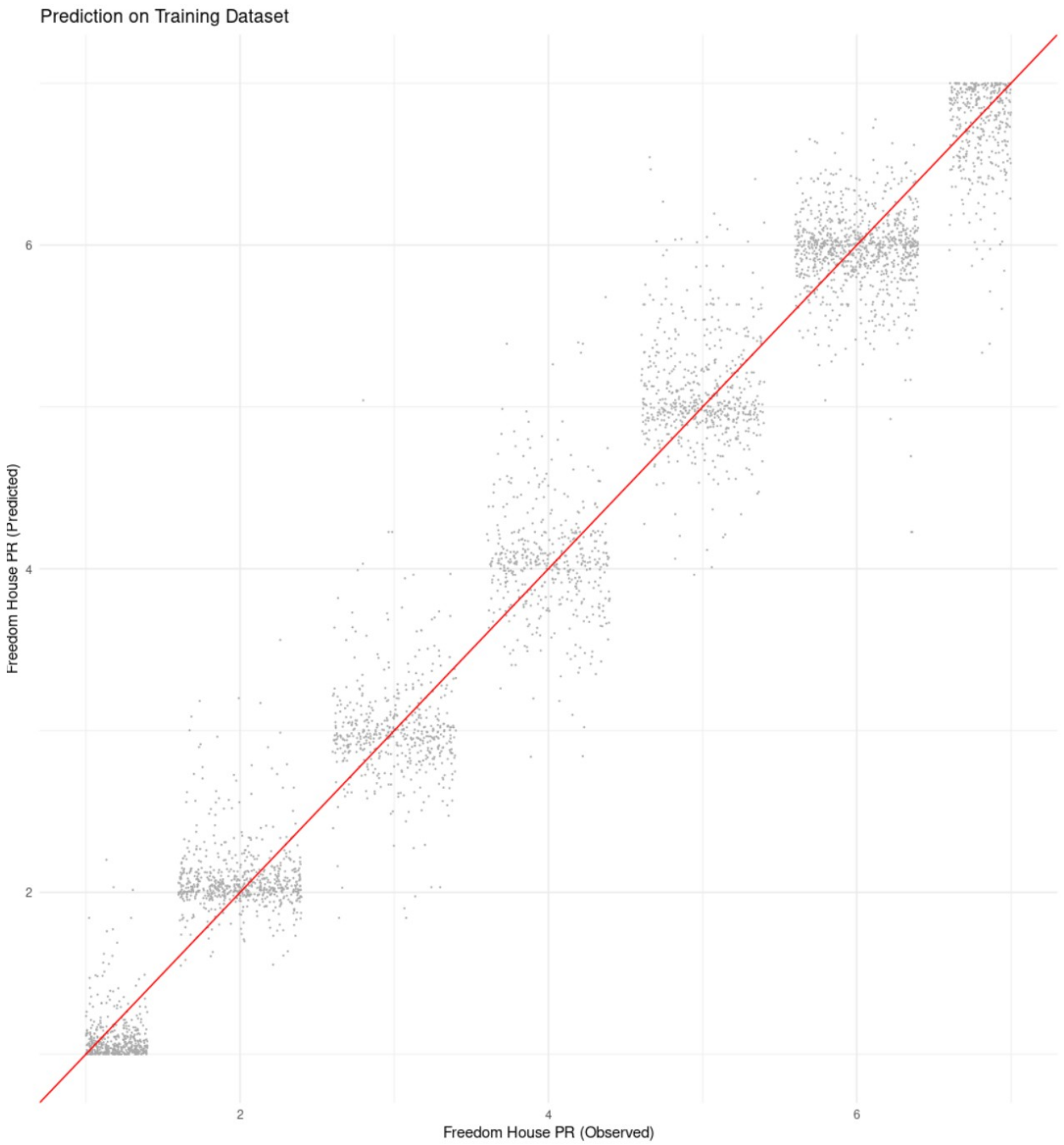

Figure A3-2: Treating Freedom House Political Rights as Continuous

Note: Shown are predicted vs actually observed Freedom House Political Rights values as coded by the Freedom House Project. We show values for all observations in our training dataset. The red line indicates a perfect match between predicted and coded scores. The further points are from the line the more are they under- or overpredicted. The Freedom House Political Rights variable is treated as continuous and random forest regression is used to predict. Points are jittered for better visibility.

Appendix B: Predicted and V-Dem Polyarchy Values for Denmark

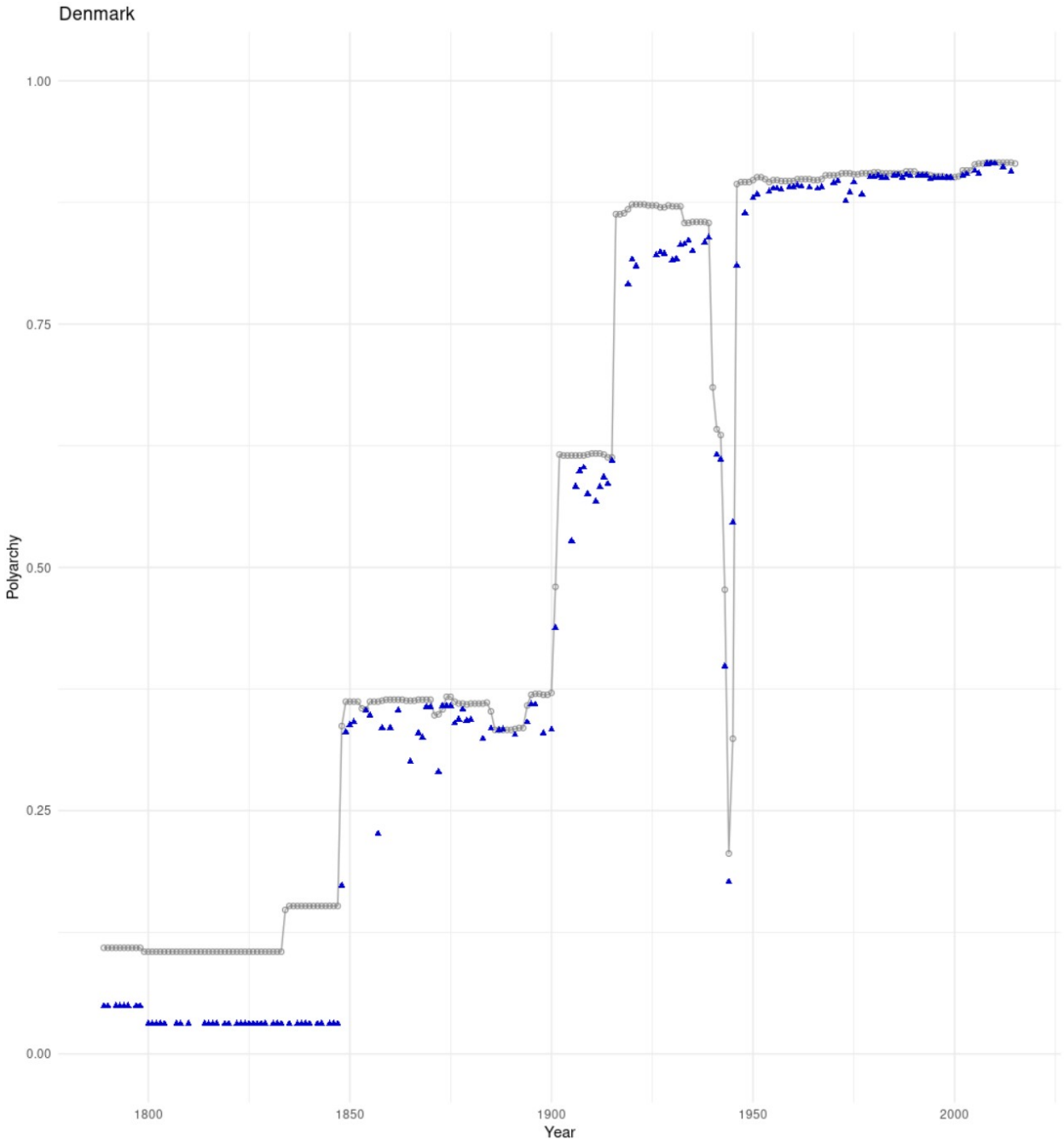

Figure A4-1: Predicted and V-Dem Polyarchy Values for Denmark

Note: Predicted and actual Polyarchy scores as coded by V-Dem for Denmark. The gray circles are the V-Dem Polyarchy values connected by a line. Blue triangles indicate predicted Polyarchy values based on objective measures of democracy in the training dataset.