Appendix

Table of Contents

Appendix A: **Codebook**

Democracy Indices

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

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

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

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

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

Variables in the Full OSM

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Lower chamber election vote share, third-largest party. In this election to the lower (or unicameral) chamber of the legislature, what percentage (%) of the vote was received by the third largest party in the first/only round? *Scale:* interval. *v2ellostts*

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

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

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

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

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

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

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

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

Turnover HOG, cumulative. Was there turnover in the office of the head of government (HOG) as a result of this national election? This variable counts the number of turnovers. Source(s): Henisz (2000; 2002); Lentz (1994; 1999); worldstatesmen.org; V-Dem Country Coordinators. *Scale:* interval. *v2elturnhog_cum*

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

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

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

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

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

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

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

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

Appendix B: Serial Omission

In this section we report the effect of excluding individual variables on the model performance in the training, validation, and cross-validation data. We report variables based on their names in the data set. Reported are key metrics, such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and R2.

Table B-1: Serial Omission

	Training			**Validation**			**Cross-validation**		
Excluded	*MSE*	*RMSE*	*R2*	*MSE*	*RMSE*	*R2*	*MSE*	*RMSE*	*R2*
turnover_period	0.003	0.058	0.953	0.003	0.054	0.956	0.003	0.058	0.952
years_turnover_period_cont	0.003	0.057	0.953	0.003	0.055	0.958	0.003	0.058	0.951
top2_difference	0.003	0.058	0.952	0.003	0.056	0.953	0.003	0.058	0.952
top2_combined	0.003	0.057	0.953	0.003	0.059	0.949	0.003	0.058	0.952
v2ellovtlg	0.003	0.058	0.951	0.003	0.058	0.951	0.003	0.058	0.951
v2ellostsm	0.004	0.060	0.948	0.004	0.061	0.944	0.004	0.061	0.947
v2x_suffr	0.004	0.062	0.944	0.003	0.059	0.949	0.004	0.063	0.943
v2x_elecreg_jg	0.003	0.059	0.950	0.004	0.059	0.947	0.004	0.059	0.949
top2_monopoly	0.003	0.056	0.955	0.003	0.057	0.952	0.003	0.056	0.955
turnover_total_ln	0.004	0.060	0.948	0.004	0.061	0.946	0.004	0.060	0.947
multi_party_leg_elec	0.004	0.061	0.945	0.004	0.062	0.945	0.004	0.062	0.944
years_exec_elec_continuous	0.004	0.061	0.947	0.003	0.058	0.953	0.004	0.061	0.947
female_suffrage	0.003	0.057	0.953	0.004	0.059	0.948	0.003	0.058	0.952

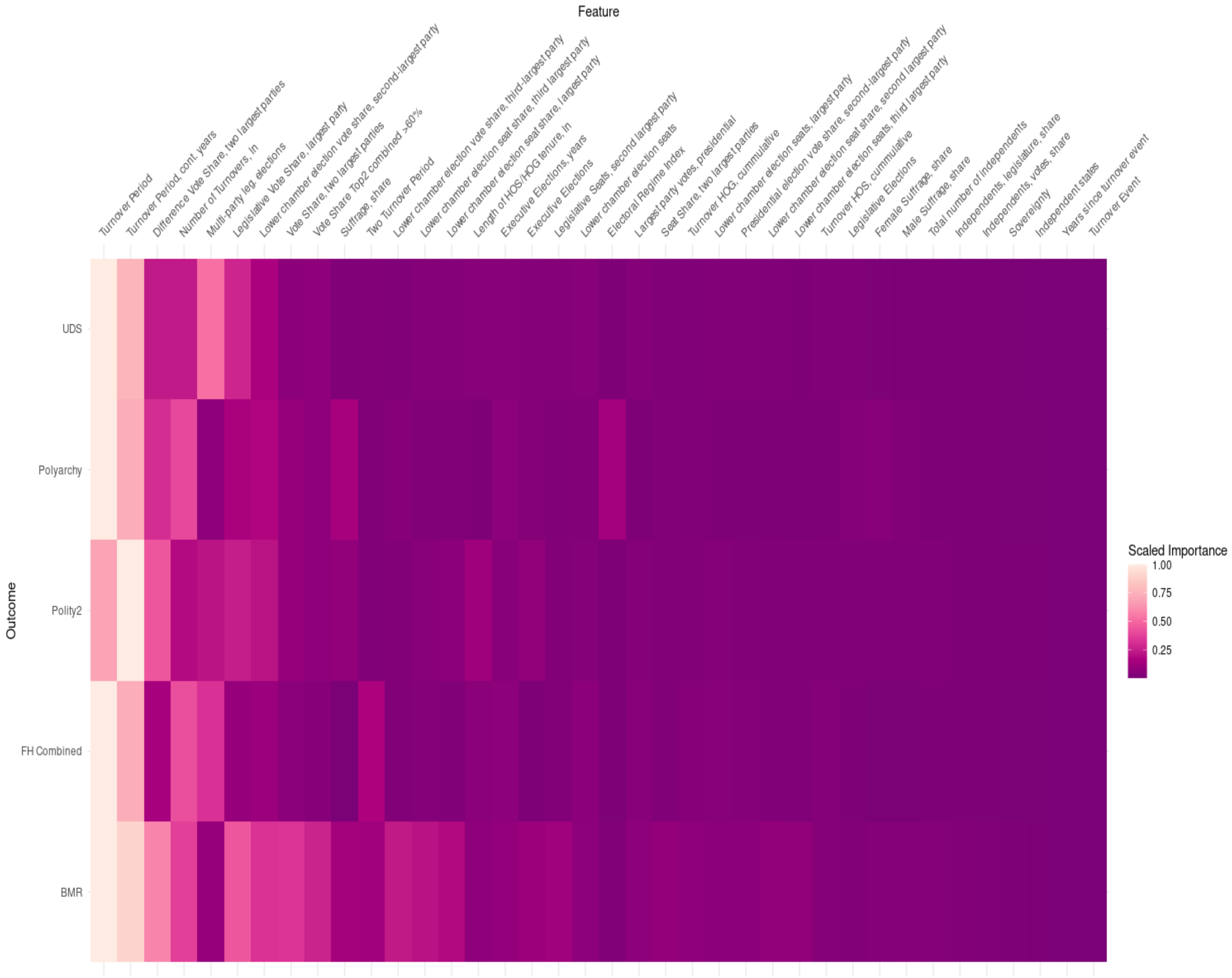
Appendix C: Goodness of Fit of Full and Reduced Models

We use both a full and a reduced list of predictors in our models. We introduce the reduced list in order to show the performance of the model with a easy and cheap to collect set of predictors. In Figure C-1 we show the heatmap of the Variable Importance Plots produced for different democracy outcomes. Across all democracy indicators a very similar set of variables is highly influential.

Table C-1: Goodness of Fit

		Training			**Validation**			**Cross-validation**		
Measure	*OSM*	*MSE*	*RMSE*	*R2*	*MSE*	*RMSE*	*R2*	*MSE*	*RMSE*	*R2*
Polyarchy	Full	0.002	0.047	0.968	0.002	0.045	0.970	0.002	0.048	0.967
	Reduced	0.003	0.056	0.954	0.003	0.057	0.953	0.003	0.057	0.953
Freedom House	Full	0.008	0.090	0.928	0.007	0.086	0.936	0.008	0.091	0.926
	Reduced	0.008	0.090	0.928	0.007	0.086	0.936	0.008	0.091	0.926
Polity2	Full	0.015	0.123	0.878	0.015	0.122	0.879	0.015	0.123	0.877
	Reduced	0.015	0.123	0.878	0.015	0.122	0.879	0.015	0.123	0.877
UDS	Full	0.003	0.056	0.954	0.003	0.057	0.953	0.003	0.057	0.953
	Reduced	0.003	0.056	0.954	0.003	0.057	0.953	0.003	0.057	0.953
BMR	Full	0.018	0.135	0.916	0.016	0.125	0.927	0.019	0.137	0.914
	Reduced	0.018	0.135	0.916	0.016	0.125	0.927	0.019	0.137	0.914

Figure C-1: VIP Heatmap



Appendix D: Imputation

Table D-1: Different Imputation Strategies

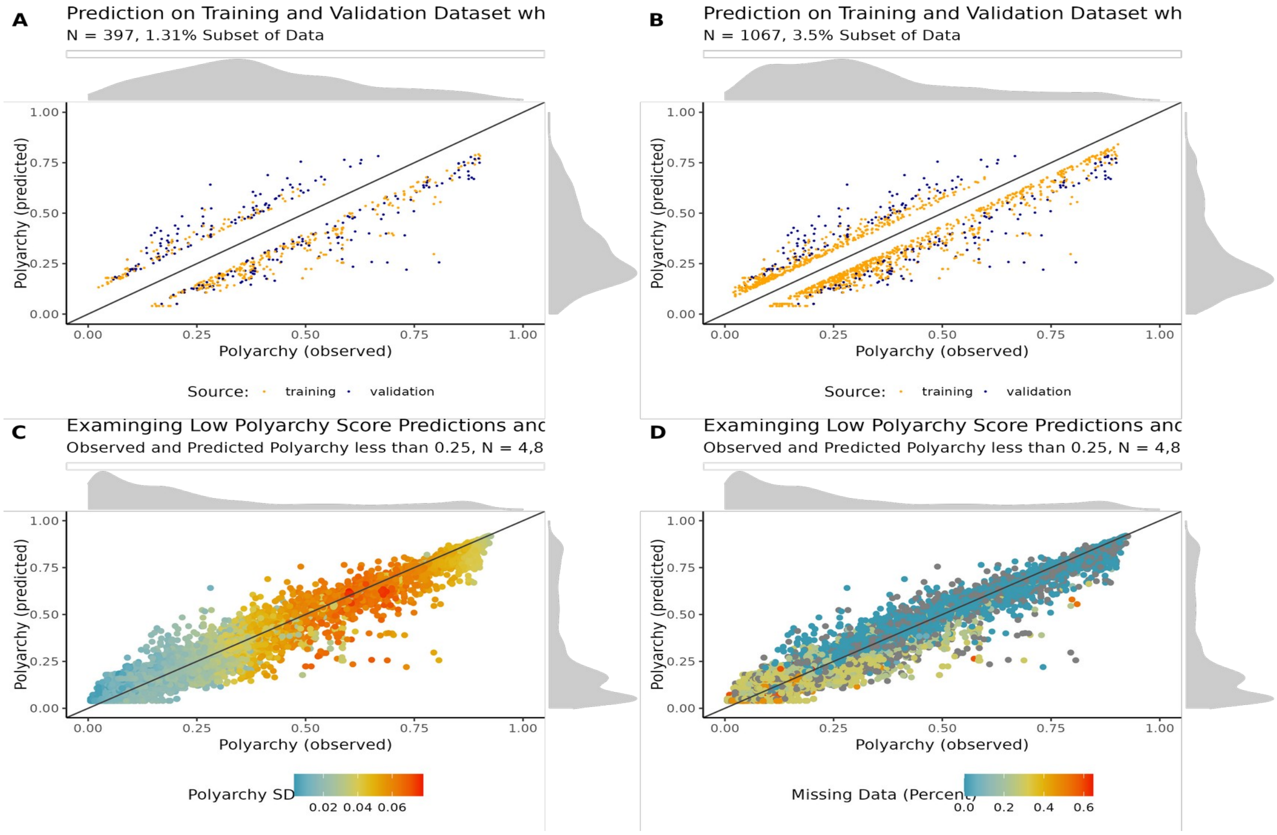
	Training			**Validation**			**Cross-Validation**		
Method	*MSE*	*RMSE*	*R2*	*MSE*	*RMSE*	*R2*	*MSE*	*RMSE*	*R2*
KNN	0.003	0.057	0.950	0.003	0.057	0.953	0.004	0.060	0.948
Bagged Trees	0.003	0.055	0.957	0.003	0.054	0.958	0.003	0.055	0.956
MICE	0.004	0.065	0.938	0.005	0.071	0.927	0.004	0.066	0.938

In order to assess the role of missingness we used three different imputation strategies (KNN, Bagged Trees, and MICE).In the models we report in the manuscript we treat missing data as its own class, assuming that there is a specific data generating process that leads to missing data about democracy related concepts.

Appendix E: Further Examination of Predictive Performance

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

Figure E-1: **Predictive quality**



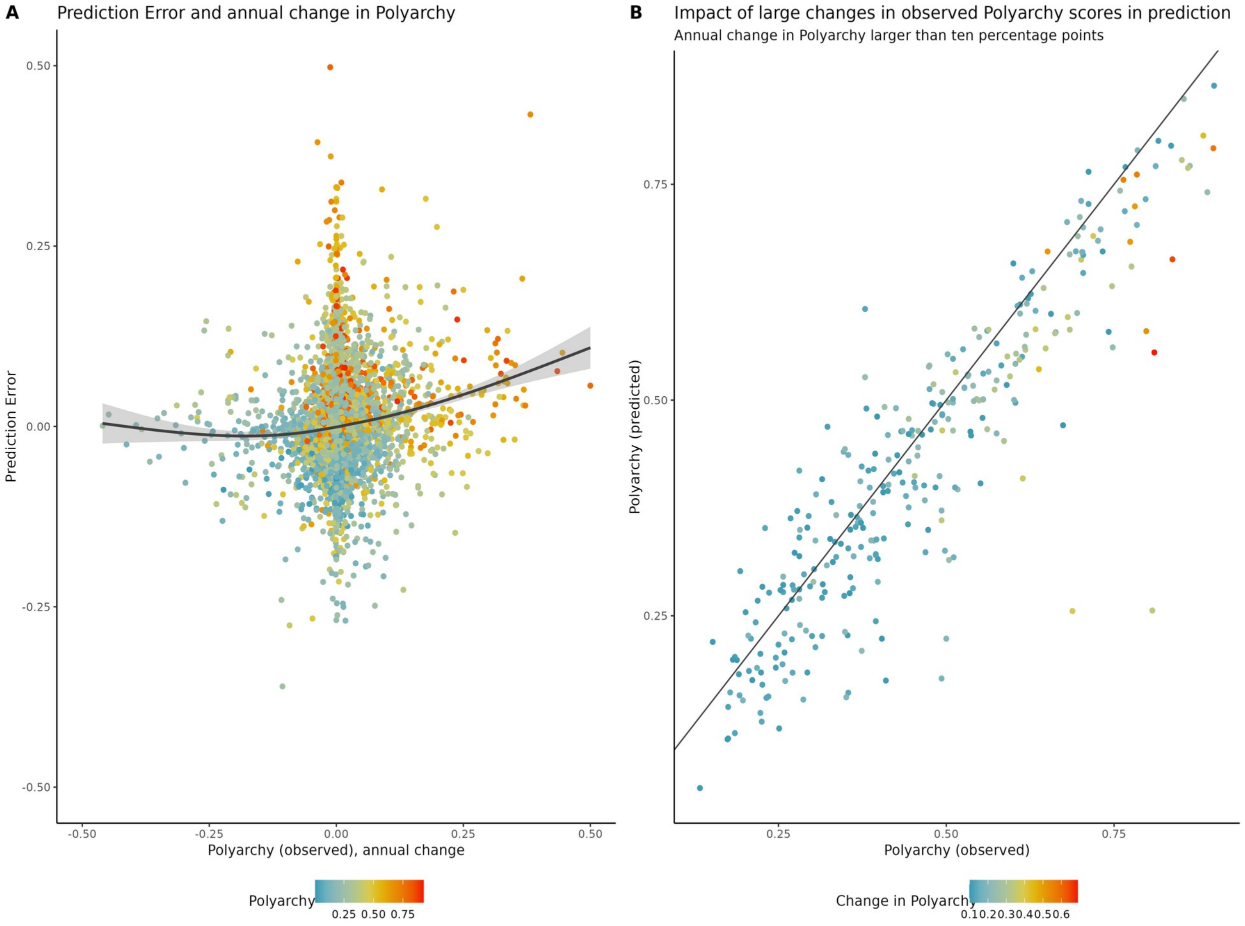
Panel A plots observed vs. predicted Polyarchy scores. Points are colored based on the standard deviation of the observed Polyarchy measure. Warmer colors indicate a higher standard deviation in the observed Polyarchy measure. Panel B also plots observed vs predicted Polyarchy scores. The color indicates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

A further source of potential problems could be in rapid changes in democracy scores. Countries that experience a sudden in- or decrease in the assigned democracy scores might be experiencing different dynamics and the learned relationship between our observed features and the democracy scores might not hold in these particular circumstances. Figure E-2 below plots the annual changes in a country's Polyarchy score. Panel A shows the relationship between the prediction error (observed Polyarchy score - predicted Polyarchy score) against the annual change in Polyarchy as coded by the V-Dem Project. It will be seen that increases in Polyarchy scores are associated with larger prediction errors. Note that the first quadrant of the cartesian plane has warmer colors than the third.

Panel B shows the relationship between the predicted and the observed values for instances where these changes are larger than 10 percentage points (>0.10 on the 0-1 Polyarchy scale). It will be seen that large changes are rather rare.

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

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



Panel A plots the annual change in the observed Polyarchy against the prediction error. The goal is to examine if large changes in the democracy score are hard to predict. This seems to be the case. Panel B shows this relationship for instances where the change was more than 10%.

Appendix F: Other ML Models

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

Table F-1: **Different Algorithms**

	Training			**Validation**			**Cross-Validation**		
Method	*MSE*	*RMSE*	*R2*	*MSE*	*RMSE*	*R2*	*MSE*	*RMSE*	*R2*
XGBoost	0.003	0.056	0.955	0.005	0.067	0.935	0.004	0.065	0.938
GBM	0.006	0.077	0.914	0.007	0.083	0.901	0.006	0.080	0.907
GLM	0.012	0.107	0.833	0.012	0.110	0.827	0.012	0.107	0.833

Appendix G: Country-Year Plots for Polyarchy

A full PDF with all countries is attached at the end of this document. The R script producing this PDF is called g_country_year_all.R.

Appendix H: Coder Judgment

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

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

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

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

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

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

The DD Index regards a polity as democratic if four conditions hold: (1) the chief executive is chosen (directly or indirectly) by popular election, (2) the legislature is popularly elected, (3) more than one party competes in elections, (4) an alternation in power occurs under electoral rules identical to the ones that brought the incumbent to office (Cheibub et al. 2010: 69). These rules are fairly clear in most instances but encounter ambiguity in others. Condition (1) is unclear where unelected and elected officials share power, as in many constitutional monarchies or polities where the military exercises power sotto voce behind the throne. Condition (2) is complicated if there are multiple chambers or legislatures, some of which are elective and others appointive. Condition (3) is ambiguous in cases where the independence of "opposition" parties is in doubt.

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

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

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

Appendix I: Bias Reduction

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

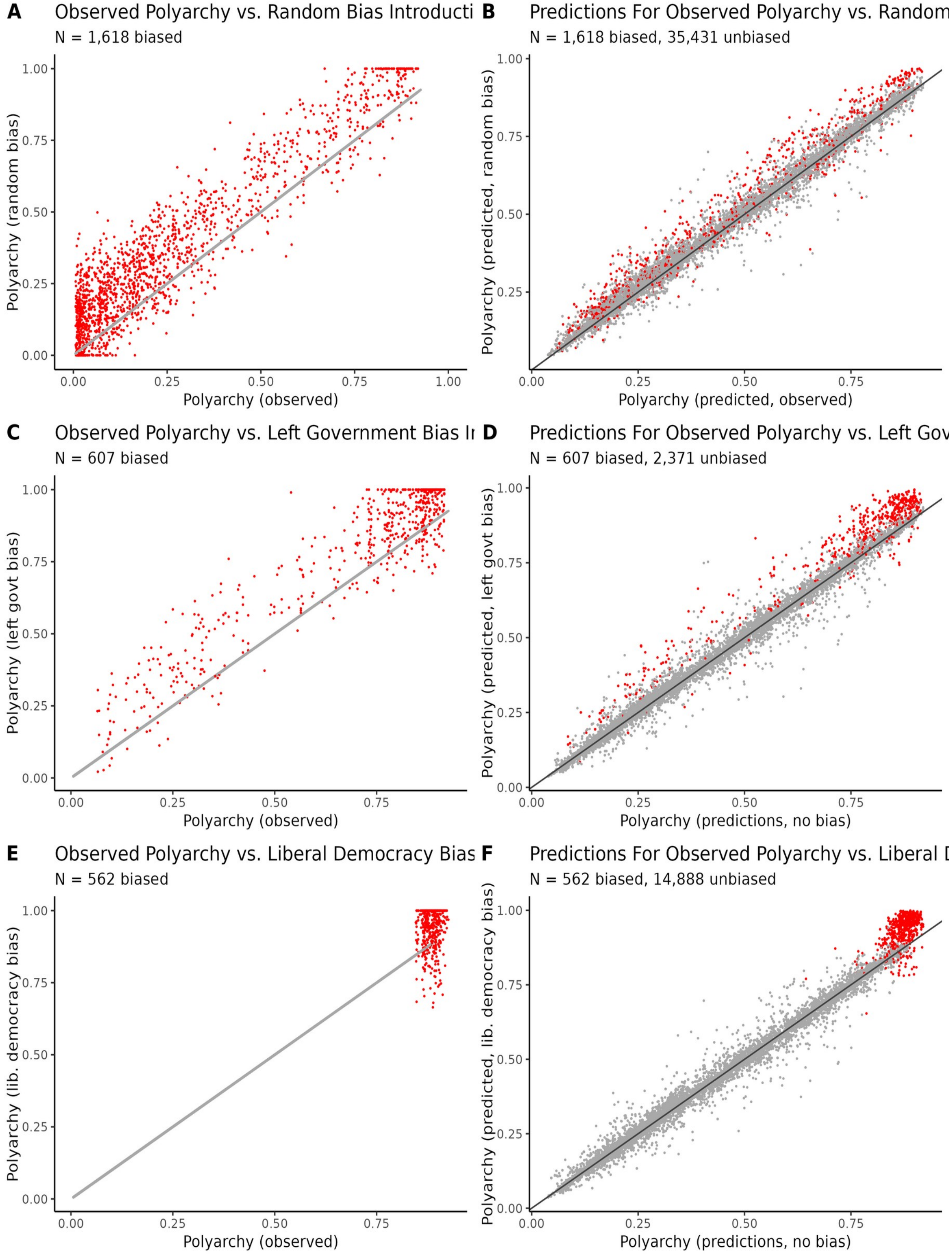
Across all tests we always introduce bias generated with a draw from a normal distribution with mean .1 and variance .1. As a reminder, the Polyarchy index ranges from 0 to 1 and has a variance of 0.069, a N(.1,.1) bias introduction is hence always substantial in size. The number of biased observations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

***Table I-1:* Bias introduced**

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

We first introduce random error (Figure I-1, Panel A and B). This error is not correlated to our outcome or our predictors and is not clustered in the data distribution. We introduce this bias for 1,997 out of 20,020 observations. This bias introduction is 1.445 times the variance of the Polyarchy measure and hence a substantial introduction of upward bias. Panel A in Figure I-1 shows the relationship between the original Polyarchy measure and the Polyarchy measure with bias. The gray observations on the 45-degree line are the unchanged Polyarchy variables, the red observations are the biased Polyarchy scores. In panel B we plot the predictions of a model that was trained on the unbiased Polyarchy data against the predictions of a model that was trained on the biased Polyarchy data. It is immediately visible that the biased observations in red are sitting much tighter to the 45-degree line in panel B than in panel A. Calculating the average distance between biased and unbiased observations we can see that there was a drop from 0.0987 in the original data to 0.0162 in the prediction. This came at an increase of the mean error in the prediction from 0.0096 in the unbiased model to 0.0103 in the biased model. We can hence conclude that our random forest approach is able reduce significant biases that are uncorrelated with any of our measures.

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



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

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

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

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

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

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

Appendix J: Decision Trees and Random Forest

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

The starting point: Decision Trees

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

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

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

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

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

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

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

From a tree to a forest

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

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

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

Random Forest model application

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

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

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

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

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

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

Appendix K: Out of Sample Application

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

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

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

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

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

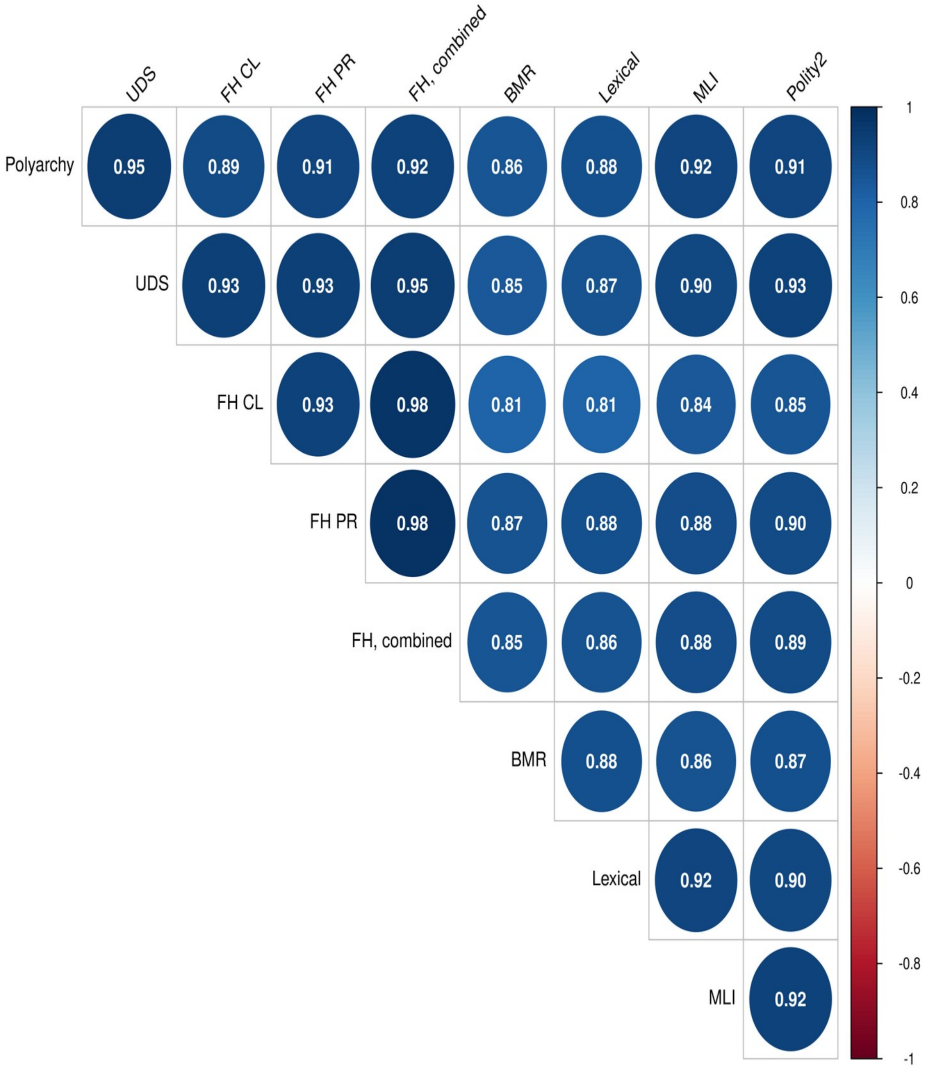
Table K-1: **Cross-Validation Results**

	Data	
Metric	**Training**	**Cross-validation**
Mean Squared Error	0.003	0.010
Root MSE	0.051	0.099
Mean Absolute Error	0.032	0.068
Root Mean Squared Log Error	0.039	0.071
Mean Residual Deviance	0.003	0.010
R2	0.964	0.8561

Appendix L: Correlations among Democracy Indicators

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

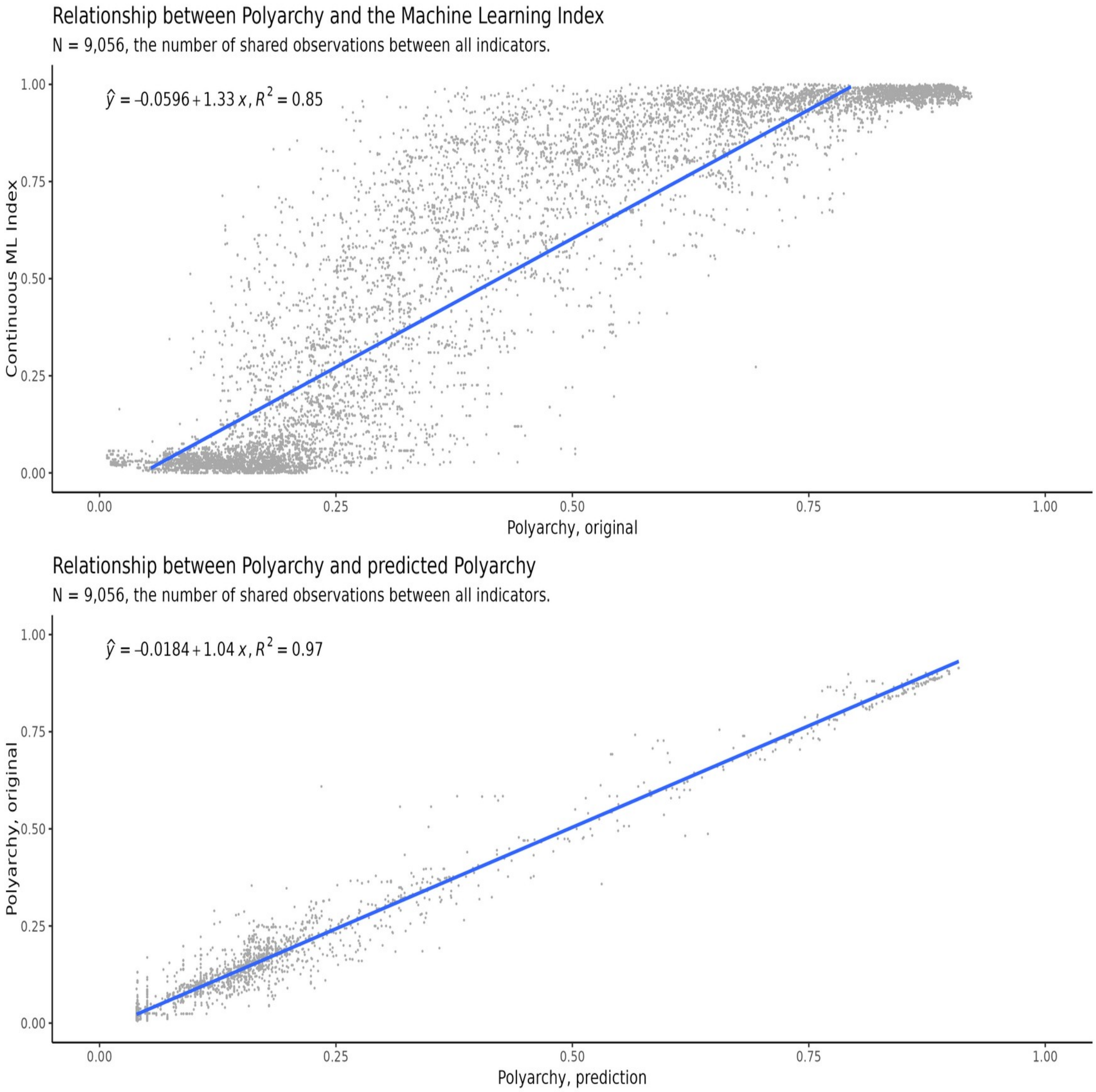
Figure L-1: **Correlation of Democracy Indices**



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

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

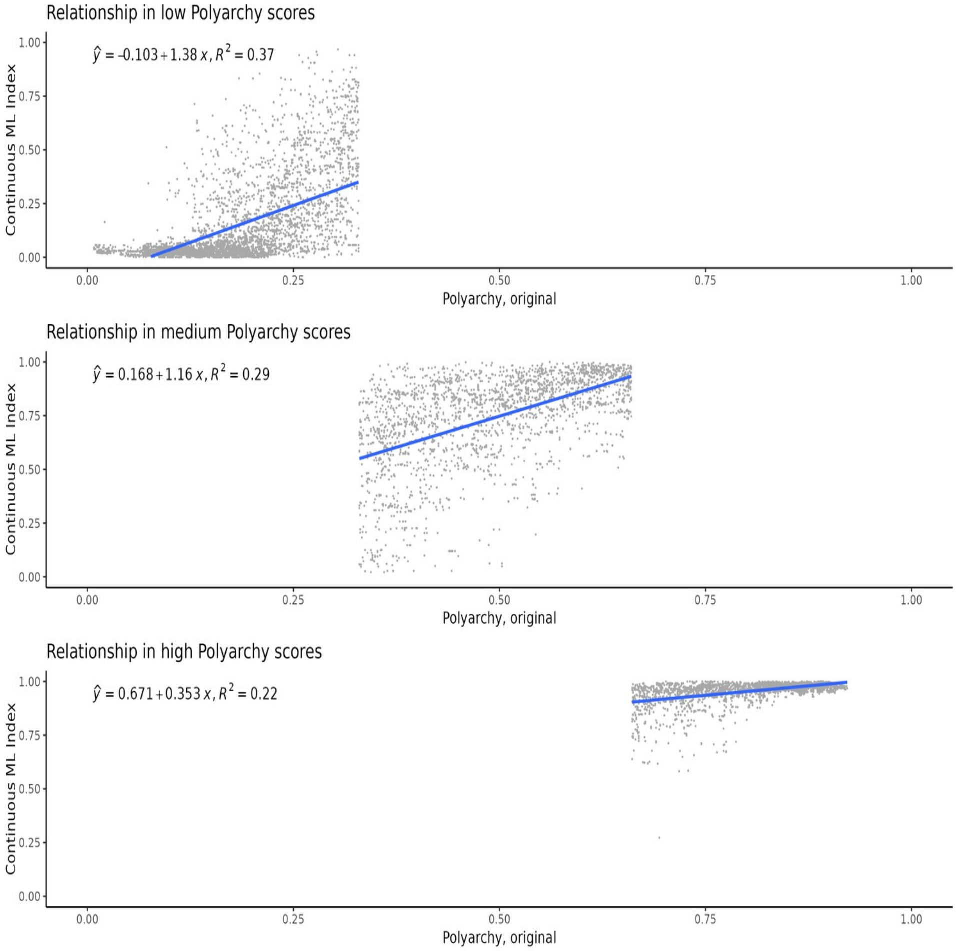
***Figure L-2:* Relationship between MLI and Polyarchy**



Note: Comparison of observed Polyarchy with Gründler and Krieger’s Machine Learning Index (Figure L-2, top) and our predicted (bottom) Polyarchy scores.

In Figure L-3 we take a closer look at the fit between the observed Polyarchy scores and the MLI. As can be seen, the relationship between the two variables varies considerably across the three slices of data. We split the Polyarchy score in the lower (0-0.33), middle (above 0.33-0.66), and the upper third (above 0.66-1). The overall fit and slope of bivariate regression lines varies across all subsets of the data. Although the overall correlation between the MLI and Polyarchy is 0.92, in the lower quarter it is 0.35, in the middle it is 0.29, and in the top quarter it is 0.25.[2] (Analogous plots for Polyarchy and OSM are displayed in Appendix M.)

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



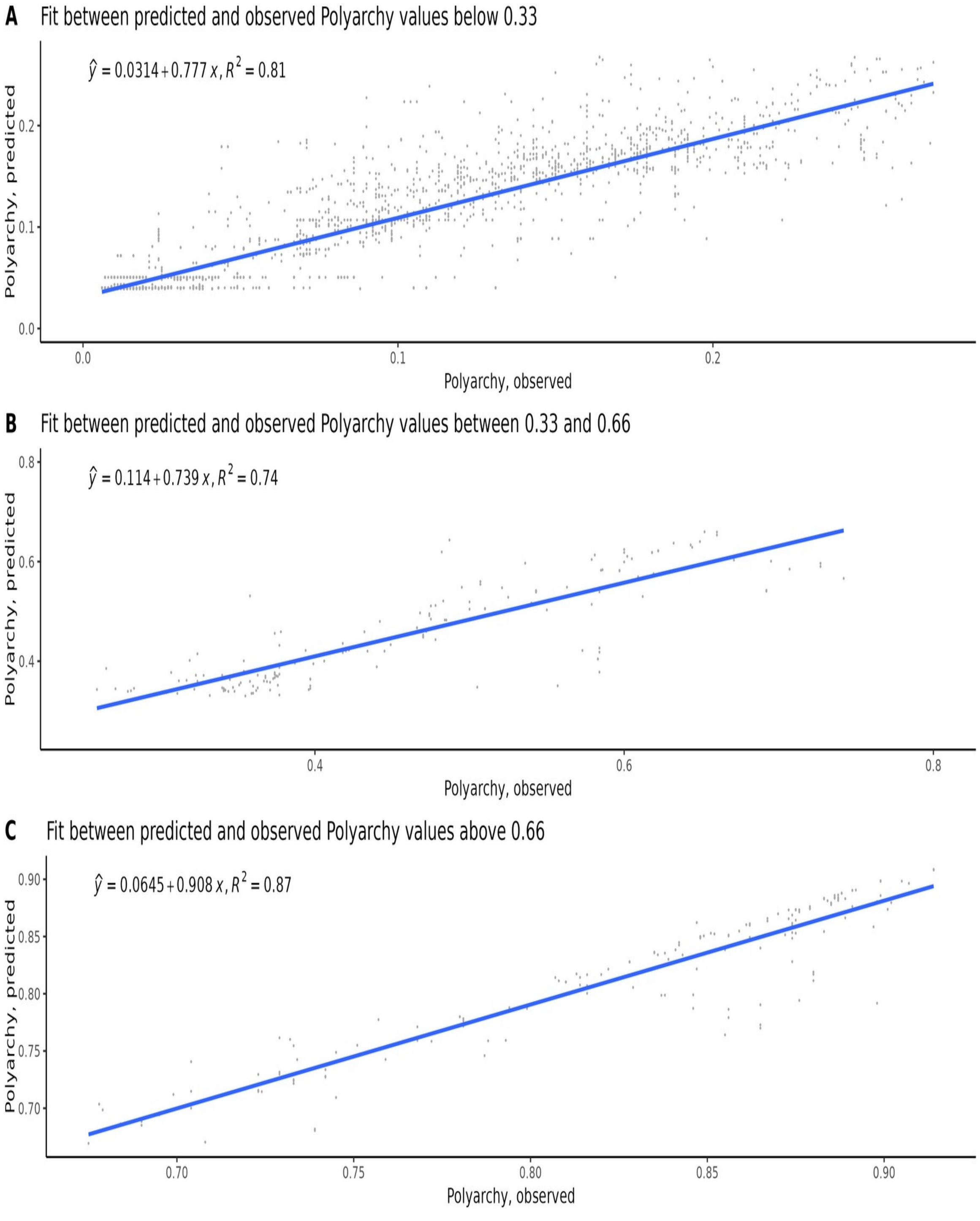
Note: The relationship between the MLI (Gründler and Krieger 2021) and Polyarchy across three equal-sized sub-sections of the Polyarchy index. The blue line is a bivariate linear regression line. The relationship between the two indicators varies across subsets of the data.

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

Appendix M: Conditionality of fit

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

Figure M-1: Fit across different values



Note: Figure M-1 examines the relationship between predicted and observed Polyarchy scores in different ranges of the observed Polyarchy scores. It can be seen that the fit is very similar across the range and there is no easy classification range that drives the performance of the random forest.

Appendix N: Linear Regression Model

Table N-1: Linear Regression Model

	1	2
(Intercept)	-0.05 ***	0.13
	(0.01)	(0.12)
Turnover Period	0.16 ***	0.12 ***
	(0.00)	(0.02)
Turnover Period, cont. years	-0.00	0.00
	(0.00)	(0.00)
Difference Vote Share, two largest parties	-0.00 ***	
	(0.00)	
Vote Share, two largest parties	-0.00 ***	
	(0.00)	
Legislative Seats, second largest party	-0.00 ***	0.00 *
	(0.00)	(0.00)
Suffrage, share	0.17 ***	-0.02
	(0.01)	(0.12)
Electoral Regime Index	0.16 ***	0.07
	(0.01)	(0.05)
Vote Share Top2 combined >60%	0.04 ***	0.08 ***
	(0.01)	(0.02)
Number of Turnovers, ln	0.06 ***	0.02
	(0.00)	(0.01)
Multi-party leg. elections	0.15 ***	0.16 **
	(0.01)	(0.05)
Executive Elections, years	0.00 ***	0.00 ***
	(0.00)	(0.00)
Female Suffrage, share	0.13 ***	0.11
	(0.01)	(0.07)
Sovereignty	-0.02 **	-0.01
	(0.01)	(0.08)
Lower chamber election seats		-0.00 *
		(0.00)
Lower chamber election seat share, largest party		-0.00 *
		(0.00)
Legislative Vote Share, largest party		-0.00 **
		(0.00)
Lower chamber election seat share, third largest party		-0.00
		(0.00)
Lower chamber election vote share, second-largest party		-0.00 ***
		(0.00)
Lower chamber election seats, third largest party		0.00
		(0.00)
Lower chamber election seat share, second largest party		0.00
		(0.00)

	1	**2**
Lower chamber election vote share, third-largest party		-0.00 **
		(0.00)
Total number of independents		0.00
		(0.00)
Independents, legislature, share		0.00
		(0.00)
Independents, votes, share		-0.01 ***
		(0.00)
Seat Share, two largest parties		0.42 **
		(0.15)
Largest party votes, presidential		-0.00
		(0.00)
Presidential election vote share, second-largest party		0.00 **
		(0.00)
Male Suffrage, share		0.12 **
		(0.05)
Executive Elections		0.04
		(0.05)
Legislative Elections"		0.02
		(0.07)
Turnover Event		-0.05 **
		(0.02)
v2elturnhog_cum		0.01 ***
		(0.00)
v2elturnhos_cum		-0.00
		(0.00)
Turnover HOG, cumulative		0.01
		(0.02)
Turnover HOS, cumulative		-0.04 ***
		(0.01)
R-squared	0.71	0.80
Adj. R-squared	0.71	0.79
Observations	7956	506

A simple linear regression model produces an adjusted R^2 of 0.71 for the reduced set of objective indicators and 0.79 for the full set. The analysis with the full set of indicators is not able to include all variables we use in the random forest model due to collinearity. This is not an issue for the random forest but the linear regression model drops variables that are derivatives of the election results (the combined vote share of the top two parties and the difference in vote share between the top two parties). The linear regression model drops all observations with missing values and we therefore have a considerably lower N in the two models (7,956 in the reduced model and 506 in the full model). The random forest model does not drop missing data but treats missing values as their own category. This is superior to imputation (especially for the pre-1900 years data is sparse in general and imputation can become difficult) and replacing missing values with 0s (since 0s can have their own meaning: An election result of 0 can be very different from an election result that is missing).

Appendix O: Test set evaluation

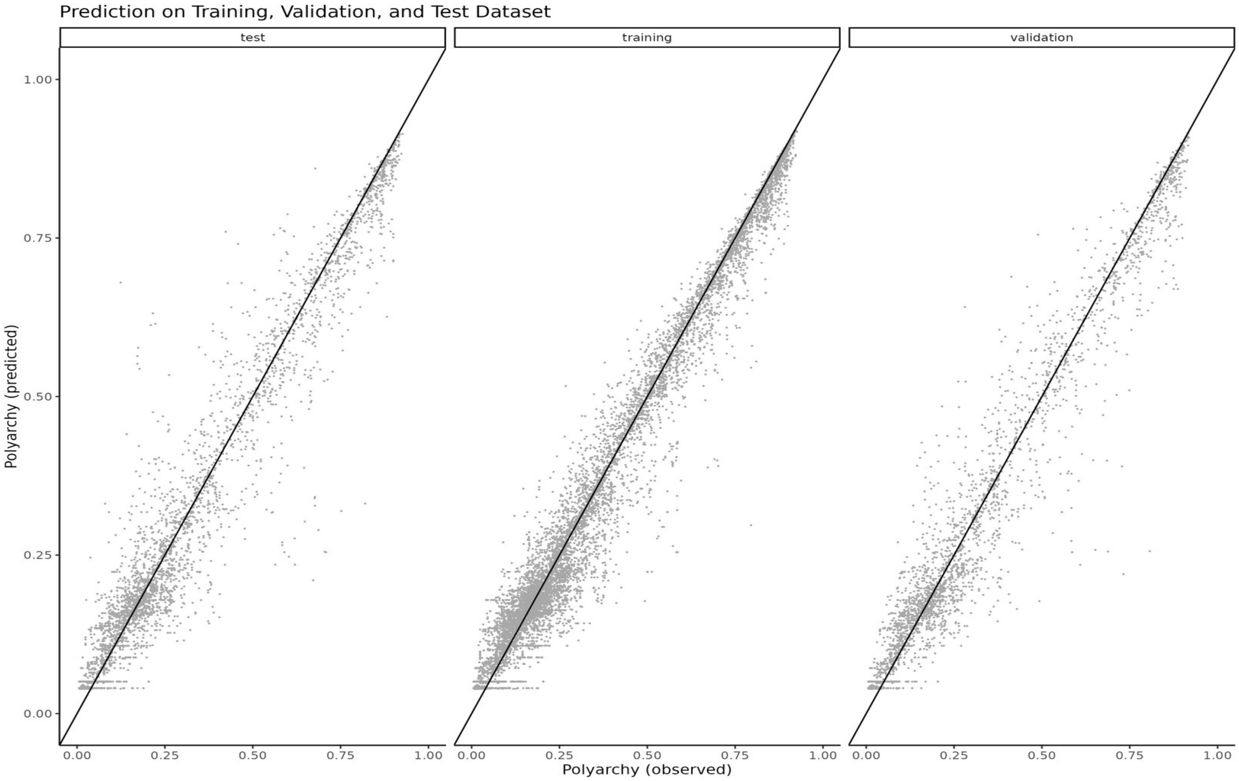
In our initial analysis we excluded a test set to assess model performance at the end of the project. This was done in order to prevent overfitting of our model to the data. We only conducted this analysis after the manuscript was accepted for publication. Table O-1 and Figure O-1 show that our predictions are very similar across training, validation, cross-validation, and test set. Across all data sets the MSE and RMSE are nearly identical and the relationship between observed and predicted values in Figure O-1 are very similar.

Table O-1: **Goodness of Fit for Polyarchy**

		Performance	
Dataset	*OSM*	*MSE*	*RMSE*
Training	Full	0.002	0.047
	Reduced	0.003	0.056
Validation	Full	0.002	0.045
	Reduced	0.002	0.057
Cross-Validation	Full	0.002	0.048
	Reduced	0.003	0.057
Test	Full	0.002	0.049
	Reduced	0.003	0.059

Note: Performance of the model for Polyarchy in different data sets. Shown are training, validation, cross-validation, and test data set mean squared error as well as root mean squared error.

Figure O-1: Observed vs predicted Polyarchy scores across different data sets



Note: Shown are predicted and observed Polyarchy scores for training, validation, and test data. Predictions are done with the reduced model.

Appendix References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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