SUPPORTING INFORMATION

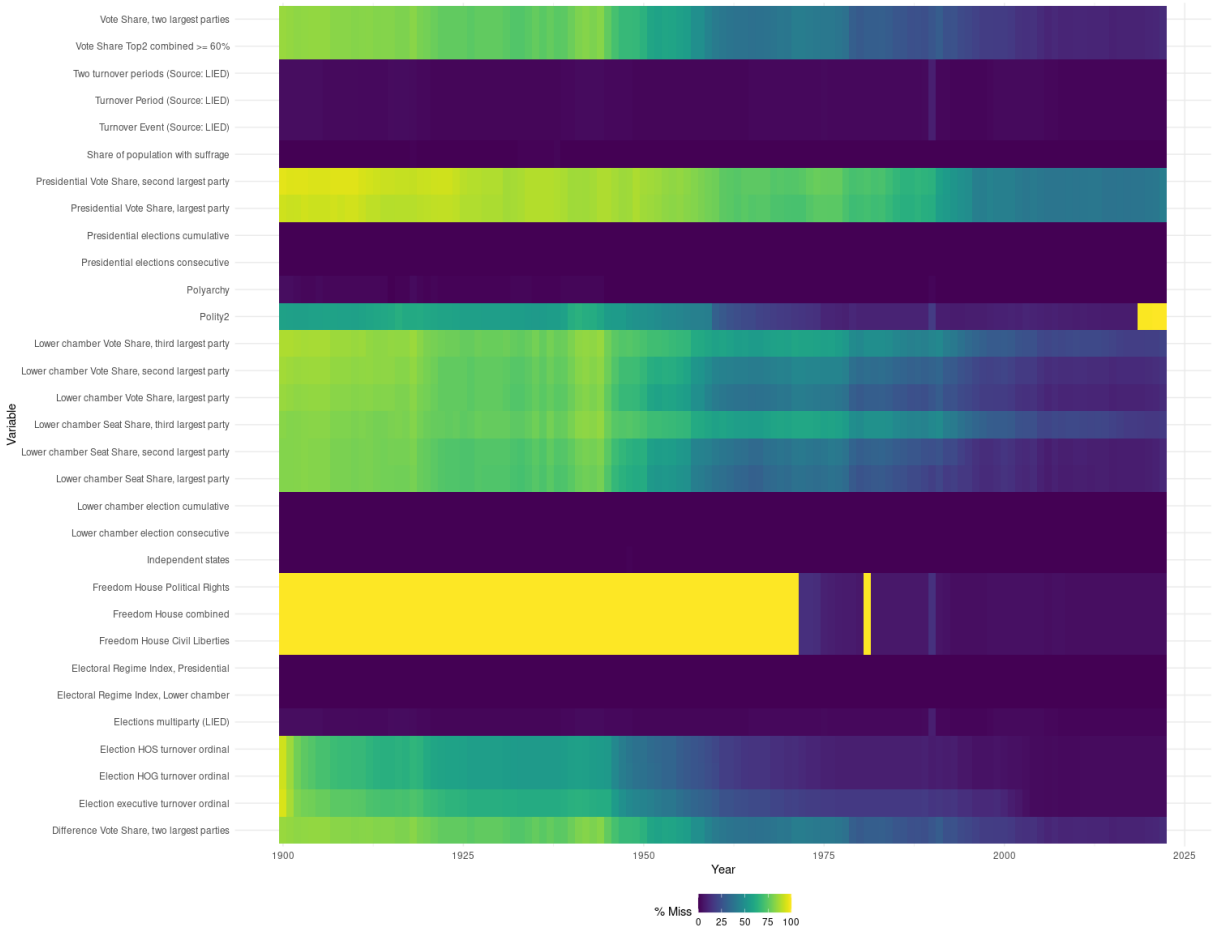
SI 1 Variables

Table SI 1.1: Outcomes and predictors

Democracy Indicators	
Electoral democracy index (D)	*v2x_polyarchy*
Freedom House, combined	*e_fh_combined*
Polity revised combined score (E)	*e_polity2*
Predictors	
Lower chamber election vote share of largest vote-getter (A)	*v2ellovtlg*
Lower chamber election vote share of second-largest vote-getter (A)	*v2ellovtsm*
Lower chamber election vote share of third-largest vote-getter (A)	*v2ellovttm*
Lower chamber election seat share won by largest party (A)	*v2ellostsl*
Lower chamber election seat share won by second largest party (A)	*v2ellostss*
Lower chamber election seat share won by third largest party (A)	*v2ellostts*
Presidential election vote share of largest vote-getter (A)	*v2elvotlrg*
Presidential election vote share of second-largest vote-getter (A)	*v2elvotsml*
Executive electoral regime index (A)	*v2xex_elecreg*
Legislative electoral regime index (A)	*v2xlg_elecreg*
Elections multiparty (LIED)	*multi_party_elections*
Share of population with suffrage (D)	*v2x_suffr*
Dummy for legislative elections	*v2eltype_legislative*
Dummy for presidential elections	*v2eltype_presidential*
Difference in vote share of top two parties	*top2_difference*
Combined vote share of top two parties	*top2_combined*
Top two parties have vote share larger than 59.99	*top2_monopoly*
Legislative elections, consecutive	*v2ellocons*
Legislative elections, cumulative	*v2ellocumul*
Presidential elections, consecutive	*v2ellocons*
Presidential elections, cumulative	*v2elprescumul*
Head of government turnover	*v2elturnhog*
Head of state turnover	*v2elturnhos*
Executive turnover	*v2eltvrexo*
Turnover period (LIED)	*turnover_period*
Turnover event (LIED)	*turnover_event*
Two turnover period (LIED)	*two_turnover_period*

SI 2 Missing Data Plot

Figure SI 2.1: Missing Data Plot



SI 3 Variable Importance Plot

In this appendix section we present the Variable Importance Plots for each model presented in the main manuscript. A variable importance plot in a Random Forest provides insights into the importance of different features (variables) in making accurate predictions. It quantifies how much each feature contributes to the overall predictive power of the model. This information is valuable for understanding which features are most influential in driving the model's decisions and can help guide feature selection, data analysis, and problem understanding. The variable importance is not necessarily a measure of causality. A feature might be important for prediction without directly causing the predicted outcome. In theory, variables could even achieve high VIP scores through spurious correlation or a confounding factor that leads to high importance. More generally, random forests are based on decision trees, which generate a tree-like structure that makes decisions by recursively splitting the data based on the values of input features. Each internal node of the tree represents a decision based on a particular feature, each branch represents an outcome of that decision, and each leaf node represents a predicted class (in classification) or a predicted value (in regression). It is possible that the features mentioned by the reviewer are very important for the model because they allow it to reliably split the data into high- and low-scoring democracies and they receive the high variable importance score because they do so across a lot of different model specifications.

It is important to clarify here that the VIP (Variable Importance Plot) does not show how much variation can be explained by a specific variable.

Figure SI 3.1: Variable Importance Plot for Polyarchy

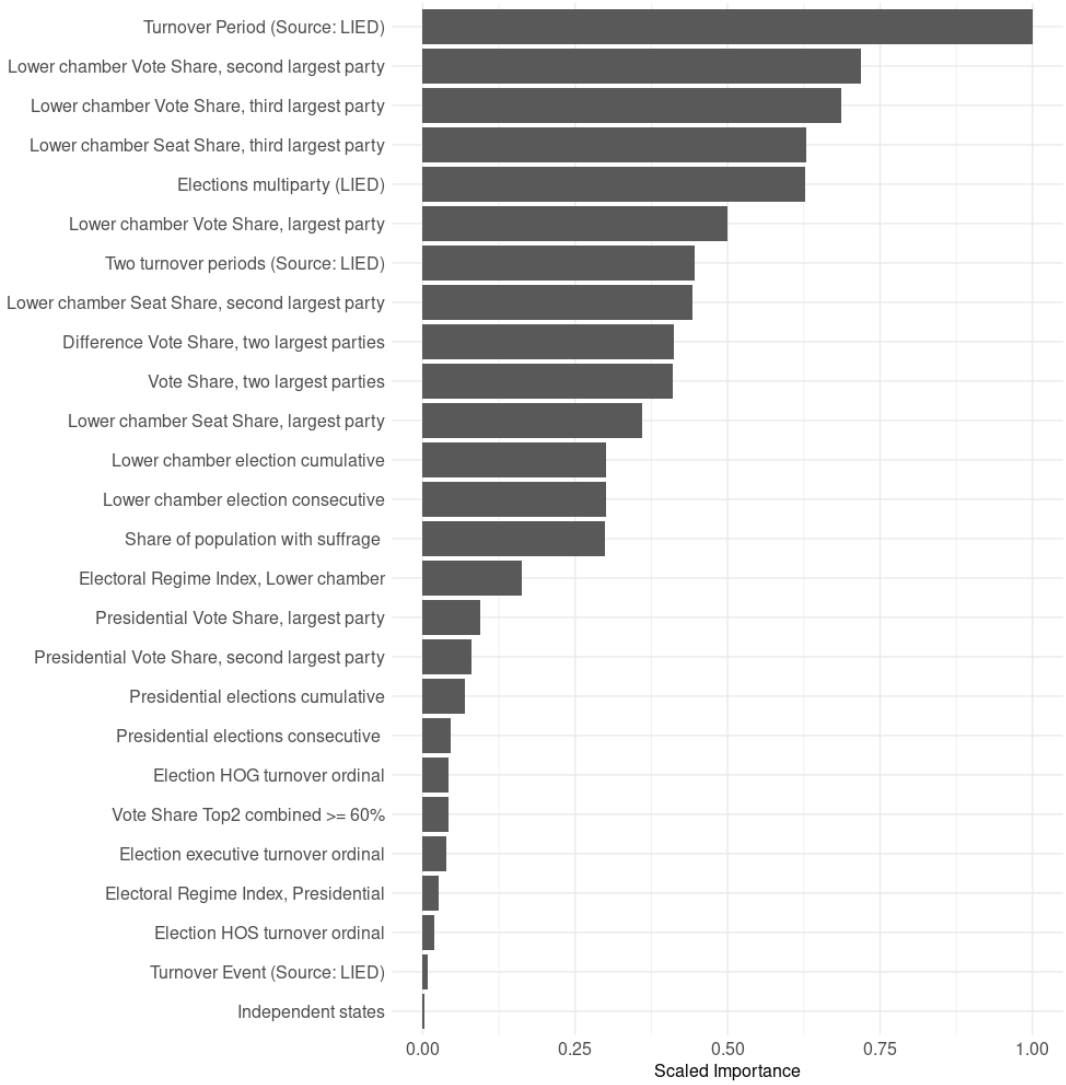


Figure SI 3.2: Variable Importance Plot for Polity2

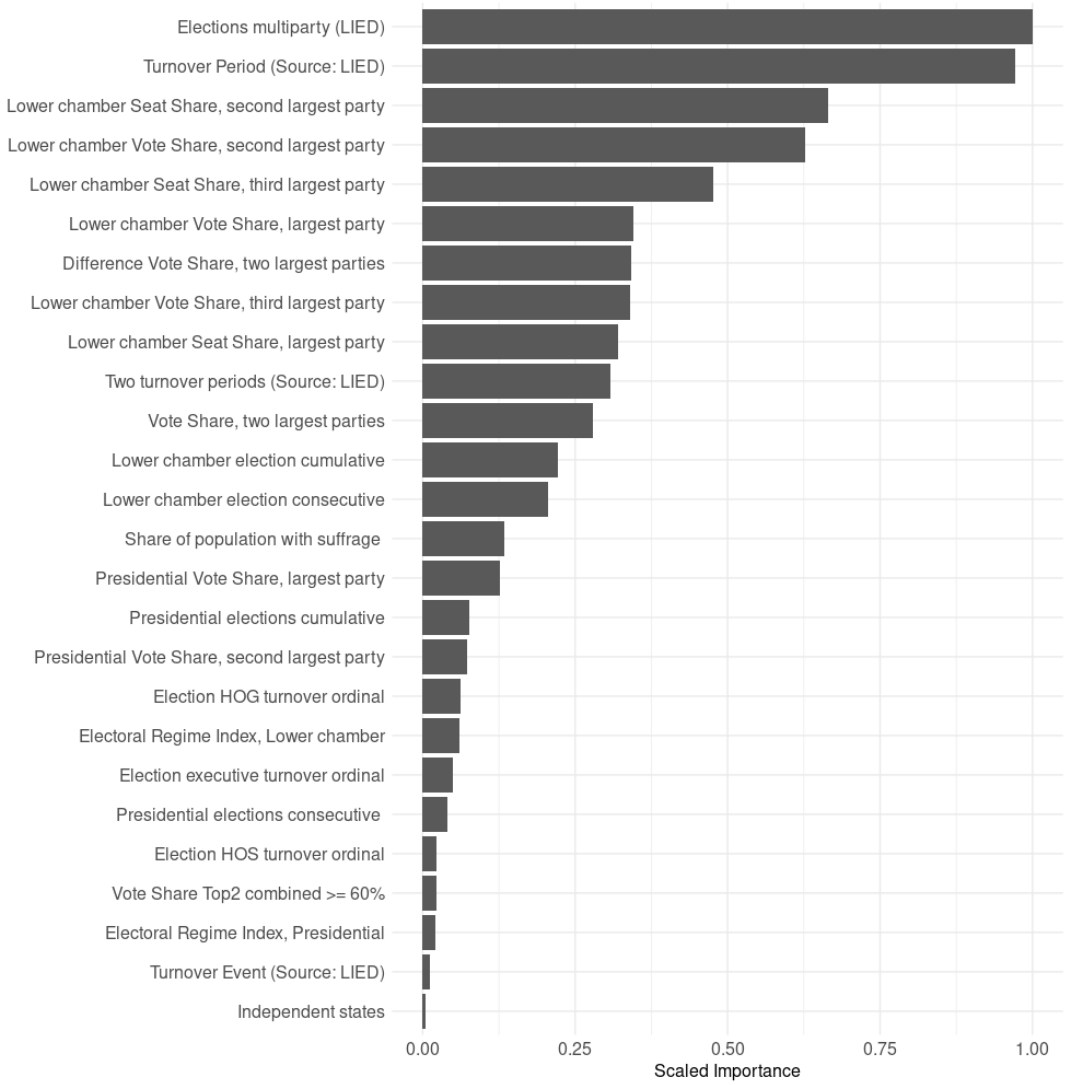
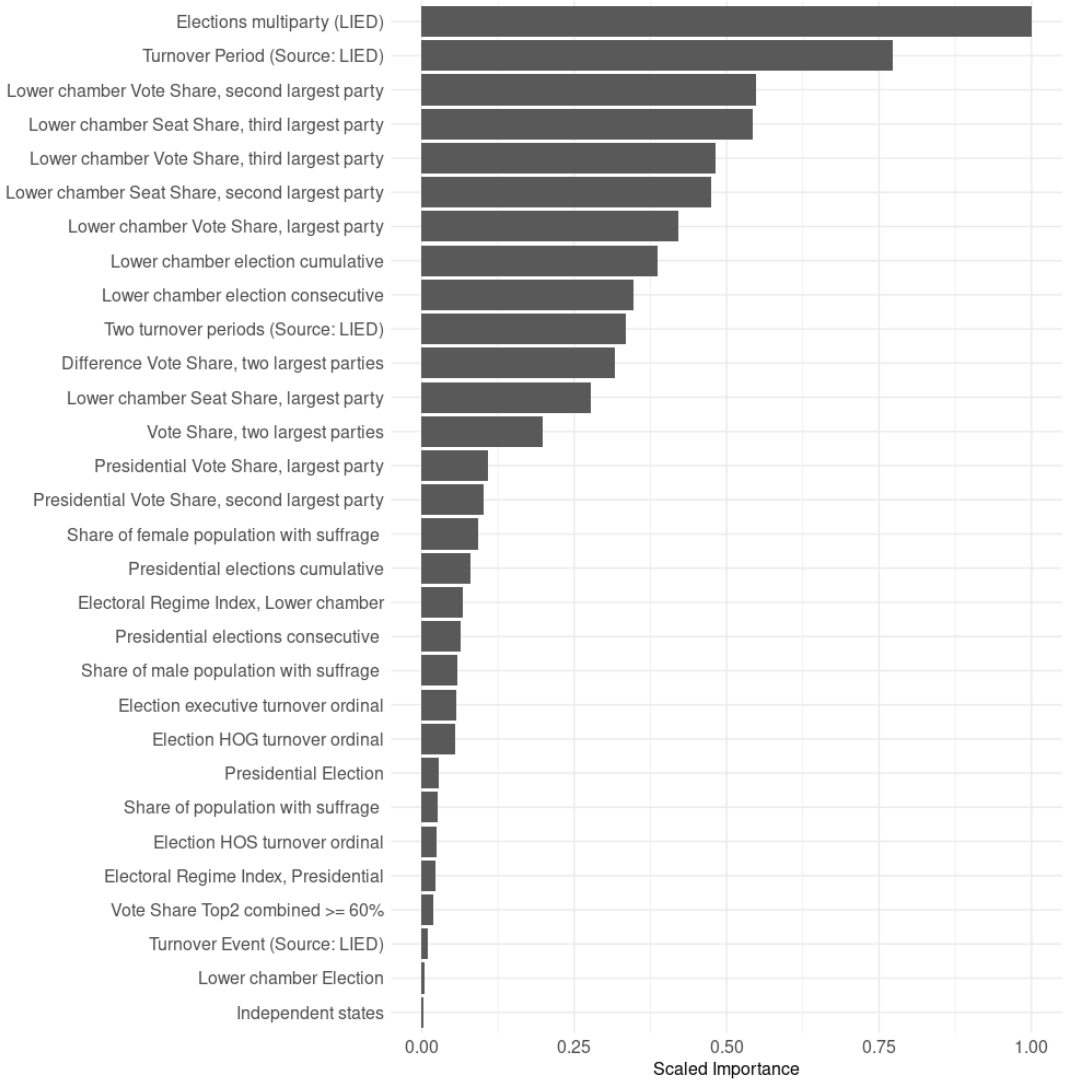


Figure SI 3.3: Variable Importance Plot for Freedom House



SI 4 Histogram and Density Plots for out of sample prediction

Figure SI 4.1: Histogram and Density for Polyarchy

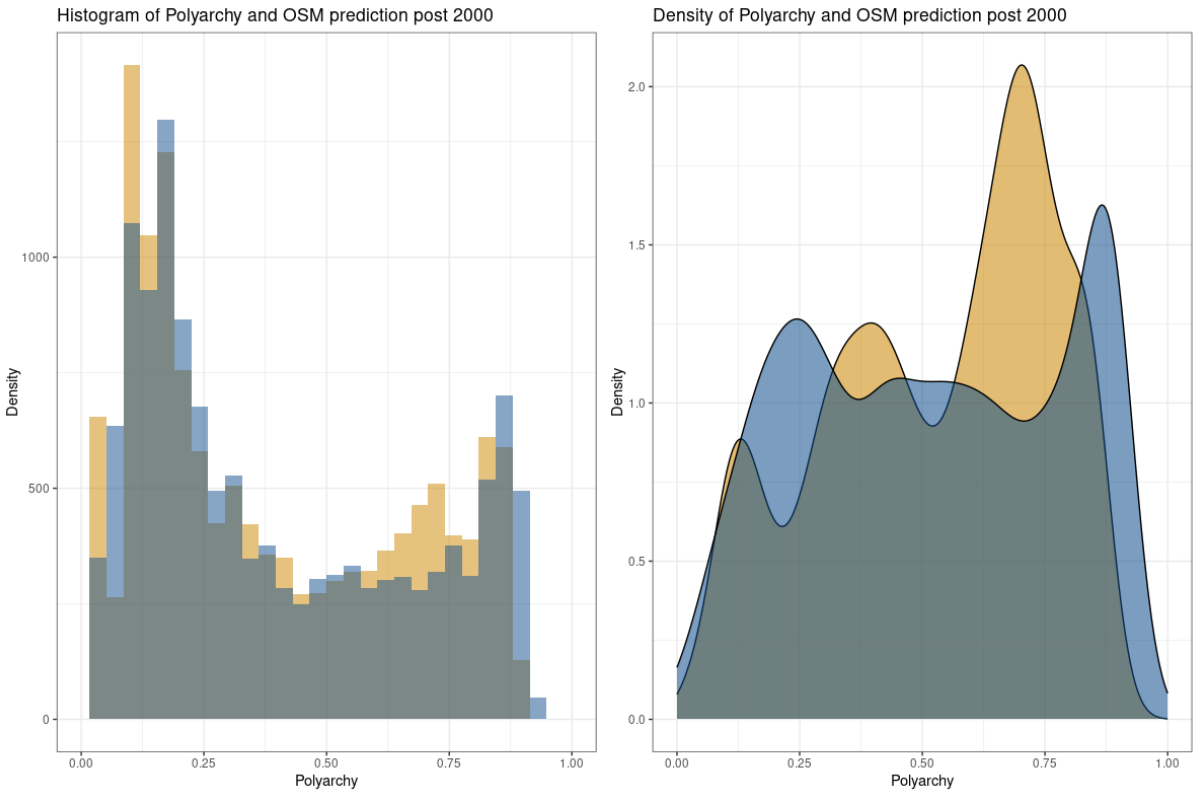


Figure SI 4.2: Histogram and Density Plot for Polity2

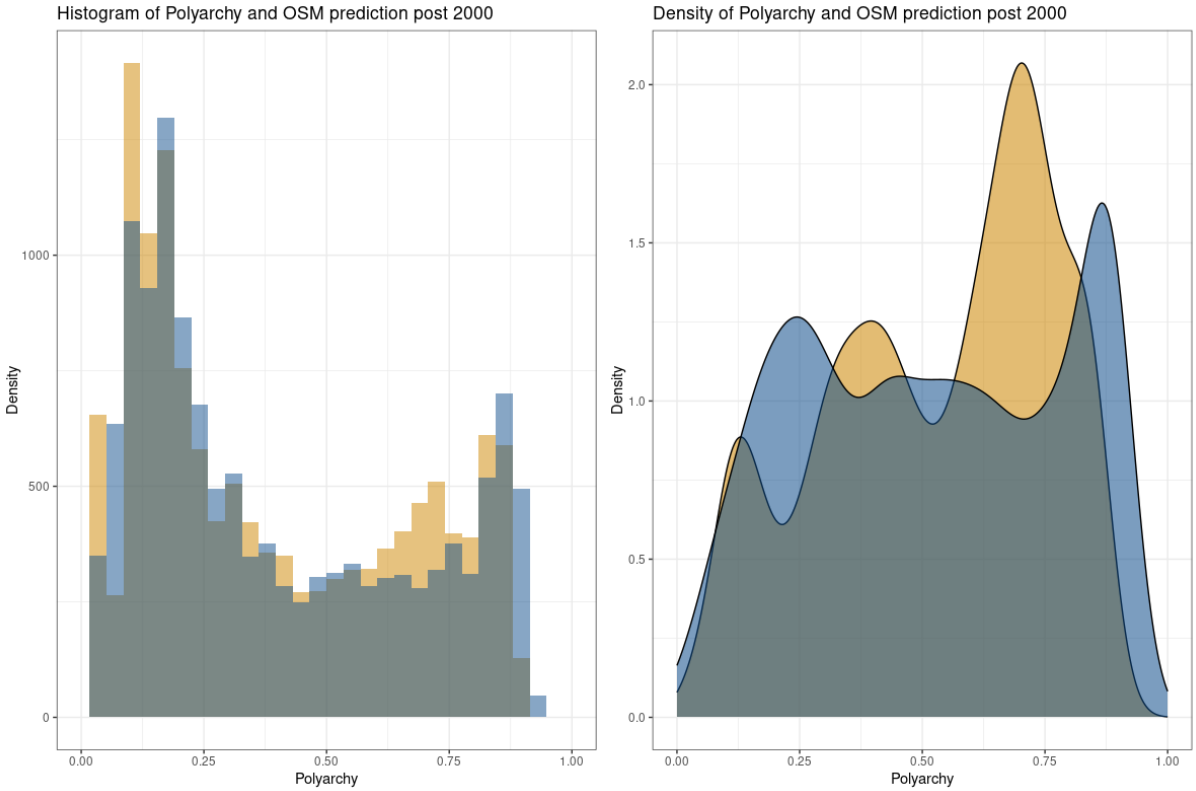
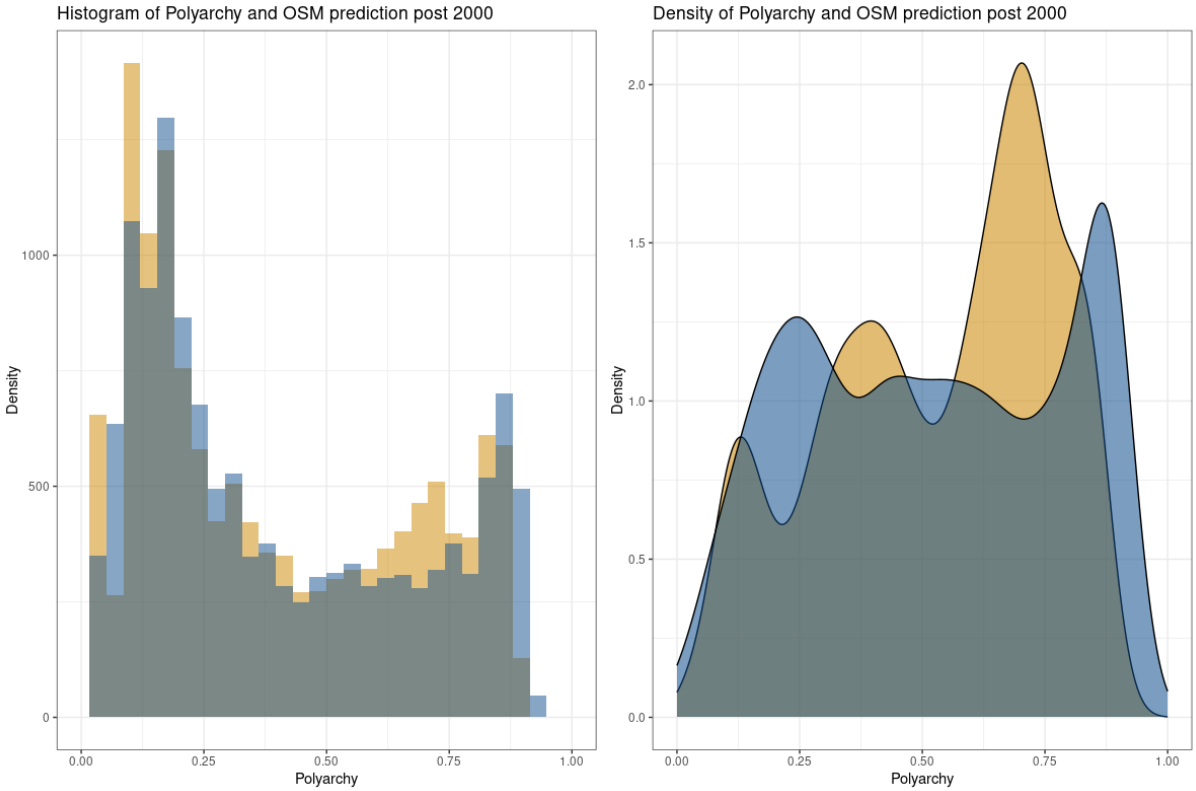


Figure SI 4.3: Histogram and Density Plot for Freedom House



SI 5 Polities

Countries			
AFG	Afghanistan	LAO	Laos
AGO	Angola	LBN	Lebanon
ALB	Albania	LBR	Liberia
ARE	United Arab Emirates	LBY	Libya
ARG	Argentina	LKA	Sri Lanka
ARM	Armenia	LSO	Lesotho
AUS	Australia	LTU	Lithuania
AUT	Austria	LUX	Luxembourg
AZE	Azerbaijan	LVA	Latvia
BDI	Burundi	MAR	Morocco
BEL	Belgium	MDA	Moldova
BEN	Benin	MDG	Madagascar
BFA	Burkina Faso	MEX	Mexico
BGD	Bangladesh	MKD	North Macedonia
BGR	Bulgaria	MLI	Mali
BHR	Bahrain	MMR	Burma/Myanmar
BIH	Bosnia and Herzegovina	MNE	Montenegro
BLR	Belarus	MNG	Mongolia
BOL	Bolivia	MOZ	Mozambique
BRA	Brazil	MRT	Mauritania
BTN	Bhutan	MUS	Mauritius
BWA	Botswana	MWI	Malawi
CAF	Central African Republic	MYS	Malaysia
CAN	Canada	NAM	Namibia
CHE	Switzerland	NER	Niger
CHL	Chile	NGA	Nigeria
CHN	China	NIC	Nicaragua
CIV	Ivory Coast	NLD	Netherlands
CMR	Cameroon	NOR	Norway
COD	Democratic Republic of the Congo	NPL	Nepal
COG	Republic of the Congo	NZL	New Zealand
COL	Colombia	OMN	Oman
COM	Comoros	PAK	Pakistan
CPV	Cape Verde	PAN	Panama
CRI	Costa Rica	PER	Peru
CUB	Cuba	PHL	Philippines
CYP	Cyprus	PNG	Papua New Guinea
CZE	Czechia	POL	Poland
DDR	German Democratic Republic	PRK	North Korea
DEU	Germany	PRT	Portugal

Continued on next page...

Table SI 5.2 (Continued)

Countries			
DJI	Djibouti	PRY	Paraguay
DNK	Denmark	QAT	Qatar
DOM	Dominican Republic	ROU	Romania
DZA	Algeria	RUS	Russia
ECU	Ecuador	RWA	Rwanda
EGY	Egypt	SAU	Saudi Arabia
ERI	Eritrea	SDN	Sudan
ESP	Spain	SEN	Senegal
EST	Estonia	SGP	Singapore
ETH	Ethiopia	SLB	Solomon Islands
FIN	Finland	SLE	Sierra Leone
FJI	Fiji	SLV	El Salvador
FRA	France	SOM	Somalia
GAB	Gabon	SRB	Serbia
GBR	United Kingdom	SSD	South Sudan
GEO	Georgia	SUR	Suriname
GHA	Ghana	SVK	Slovakia
GIN	Guinea	SVN	Slovenia
GMB	The Gambia	SWE	Sweden
GNB	Guinea-Bissau	SWZ	Eswatini
GNQ	Equatorial Guinea	SYR	Syria
GRC	Greece	TCD	Chad
GTM	Guatemala	TGO	Togo
GUY	Guyana	THA	Thailand
HND	Honduras	TJK	Tajikistan
HRV	Croatia	TKM	Turkmenistan
HTI	Haiti	TLS	Timor-Leste
HUN	Hungary	TTO	Trinidad and Tobago
IDN	Indonesia	TUN	Tunisia
IND	India	TUR	Turkey
IRL	Ireland	TWN	Taiwan
IRN	Iran	TZA	Tanzania
IRQ	Iraq	UGA	Uganda
ISR	Israel	UKR	Ukraine
ITA	Italy	URY	Uruguay
JAM	Jamaica	USA	United States of America
JOR	Jordan	UZB	Uzbekistan
JPN	Japan	VDR	Republic of Vietnam
KAZ	Kazakhstan	VEN	Venezuela
KEN	Kenya	VNM	Vietnam
KGZ	Kyrgyzstan	XKX	Kosovo
KHM	Cambodia	YEM	Yemen

Continued on next page...

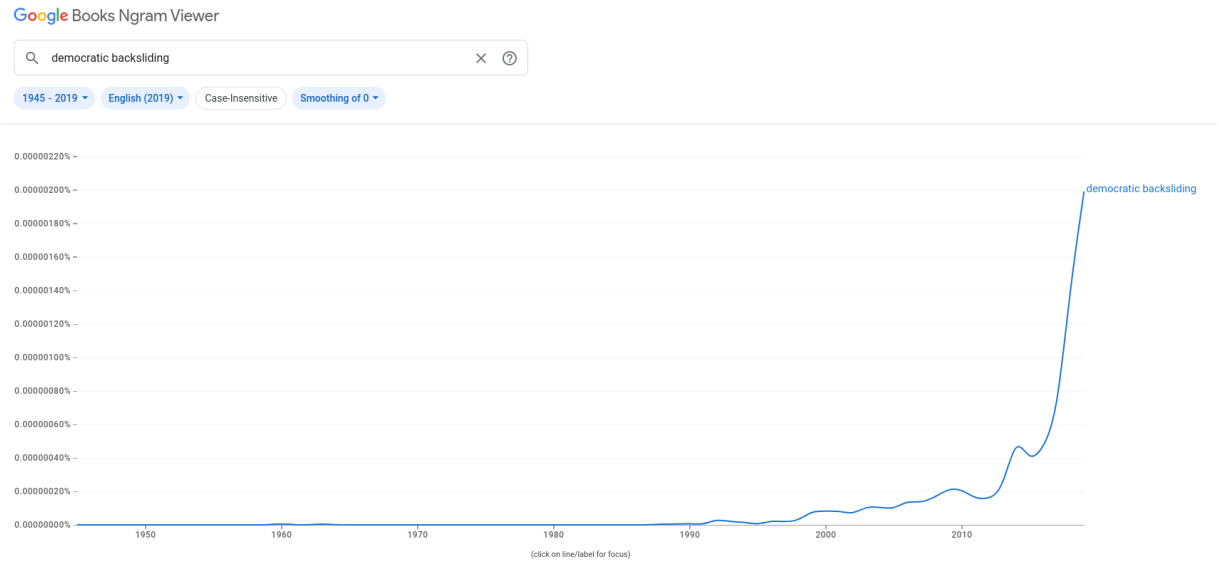
Table SI 5.2 (Continued)

Countries			
KOR	South Korea	YMD	South Yemen
KWT	Kuwait		

SI 6 Google Ngram for Democratic Backsliding

In order to determine a cut-off for for the backsliding period we need to determine when backsliding started. There are various time points mentioned in the literature. Usding a Google Ngram, we can see that the academic literature started to pick up the topic more frequently around the 2000s and had significant jumps in 2005 and 2010 (allowing for some lag in the time it takes to publish research). Based on this we have set the cut off in our manuscript for 2000 but also included replications with the year 2005 (see SI 7) and 2010 (see SI 8) .

Figure SI 6.1: Google Ngram for Democratic Backsliding



SI 7 Cut off at 2005

In this appendix we demonstrate that our conclusions are robust to using the year 2005 as a cut off for the pre-backsliding period.

Figure SI 7.1: Cut off at 2005 for Polyarchy

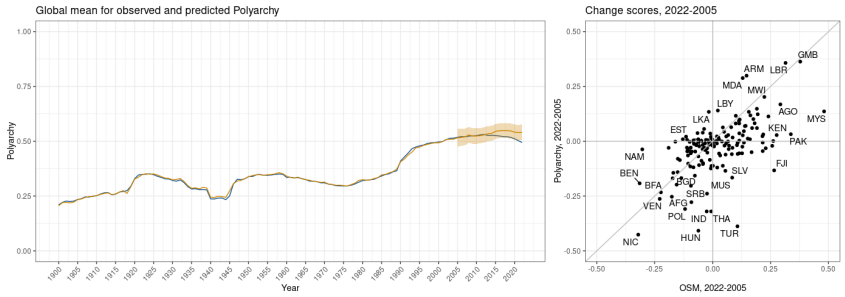


Figure SI 7.2: Cut off at 2005 for Polity2

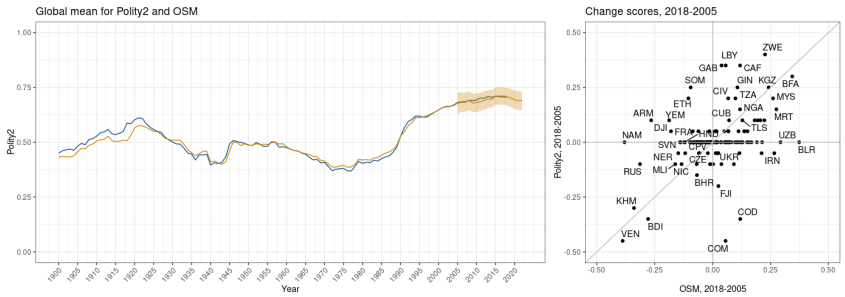
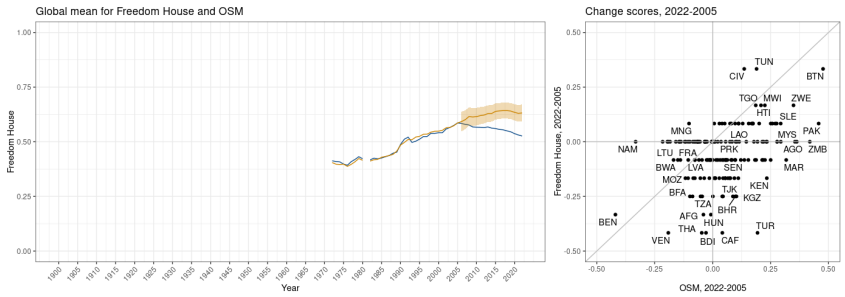


Figure SI 7.3: Cut off at 2005 for Freedom House



SI 8 Cut off at 2010

In this appendix we demonstrate that our conclusions are robust to using the year 2010 as a cut off for the pre-backsliding period.

Figure SI 8.1: Cut off at 2010 for Polyarchy

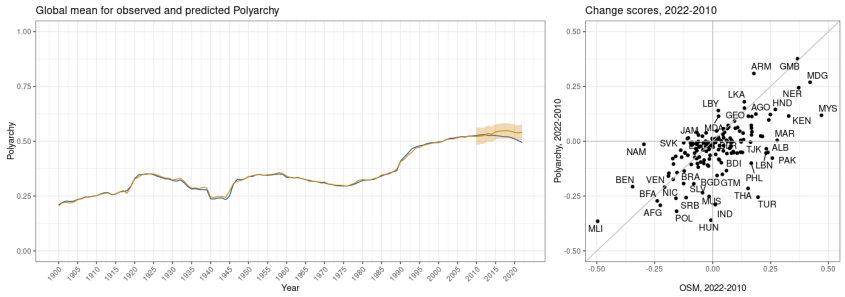


Figure SI 8.2: Cut off at 2010 for Polity2

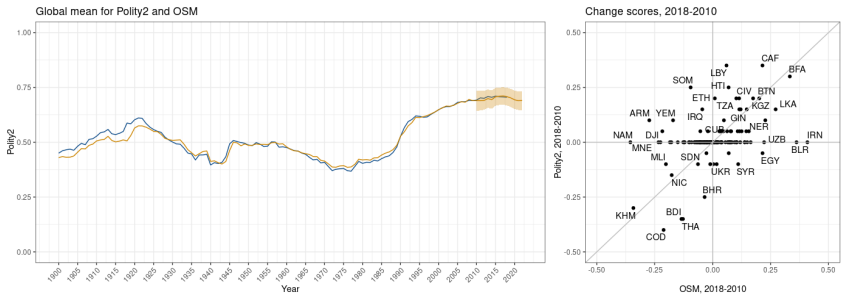
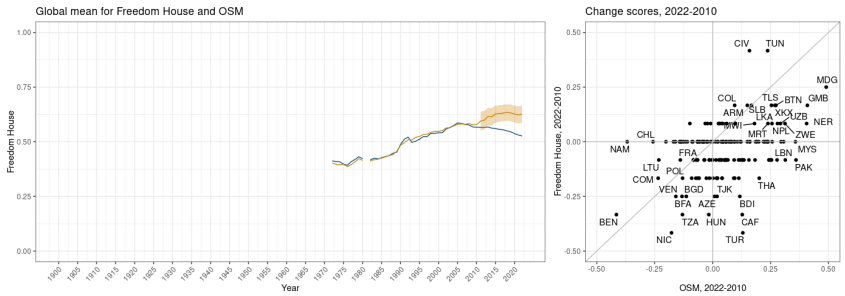
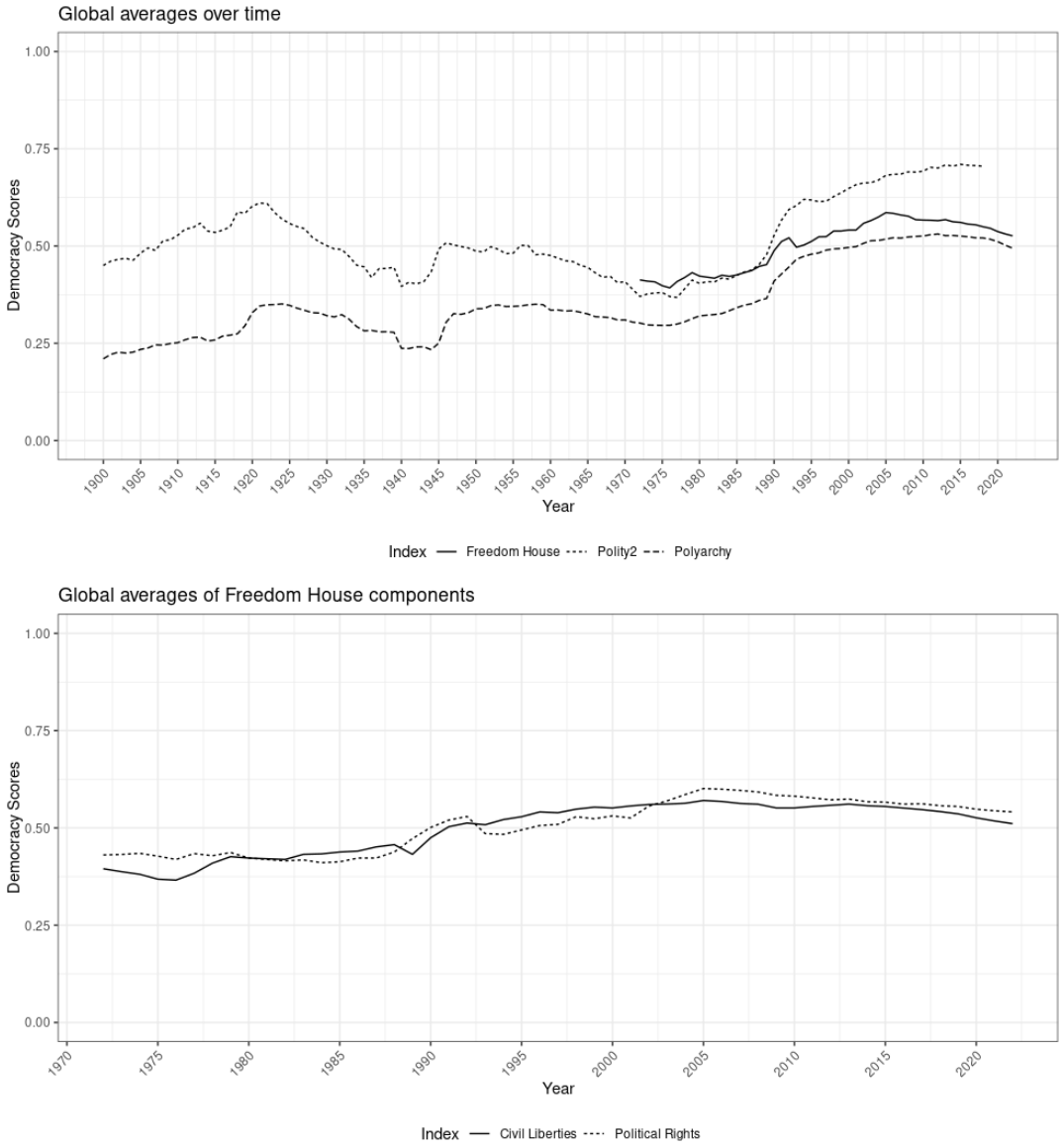


Figure SI 8.3: Cut off at 2010 for Freedom House



SI 9 Global means of democracy indicators

Figure SI 9.1: Global means of democracy indicators



SI 10 Annual changes in democracy scores of countries

Figure SI 10.1: Annual changes in democracy scores of countries

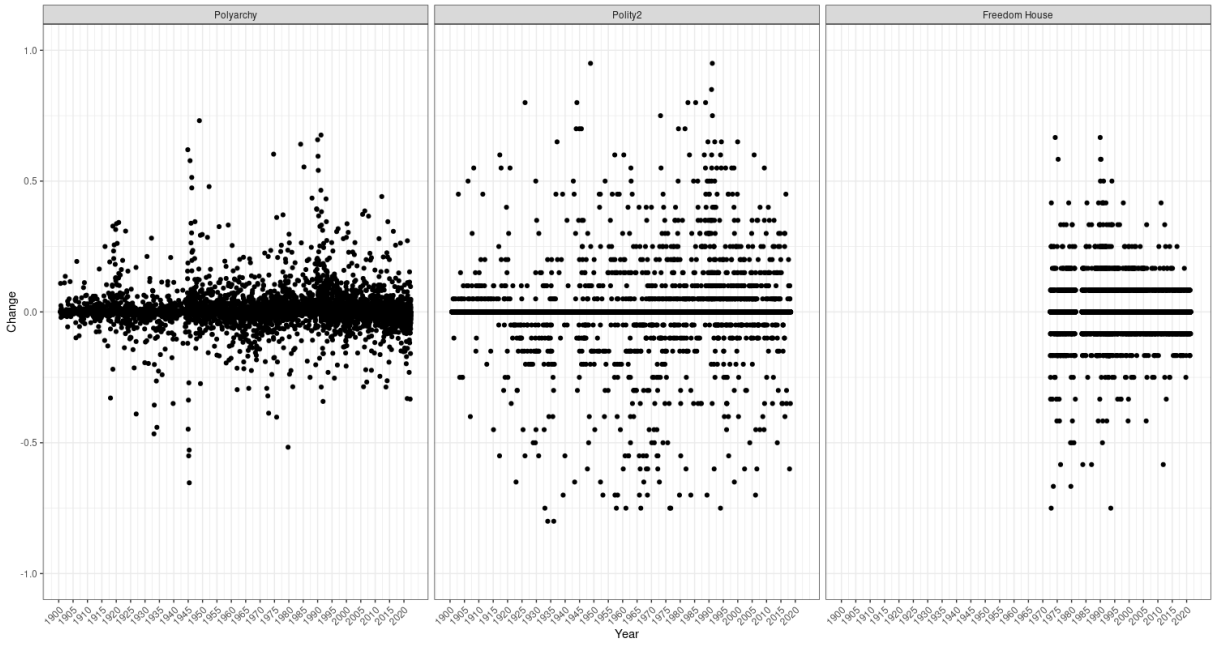


Table SI 10.3: Annual Change Statistics

	Polyarchy		Polity2		Freedom House
	1900-2022	1972-2022	1900-2018	1972-2018	1972-2022
Mean	0.003	0.004	0.003	0.006	0.002
Std. Dev.	0.046	0.045	0.082	0.079	0.067

SI 11 Freedom House predictions for different periods

Figure SI 11.1: Predicting 1988-2004

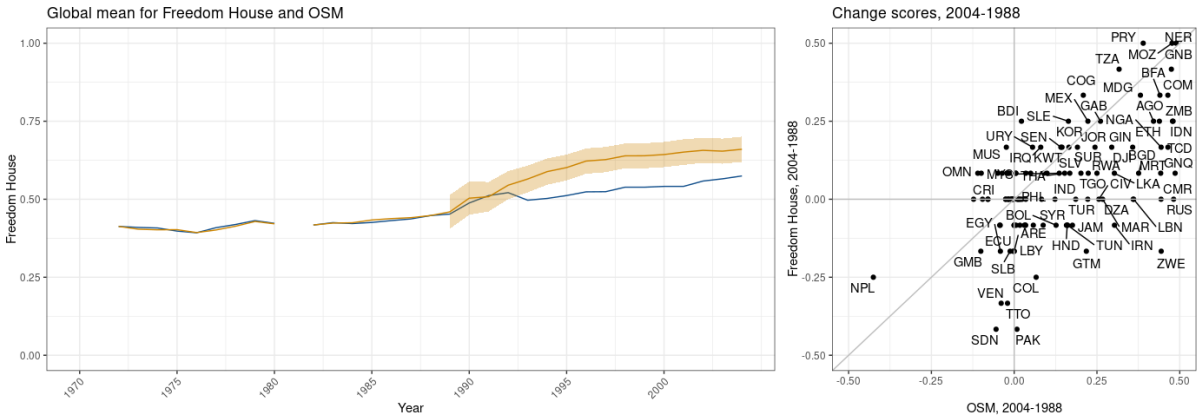
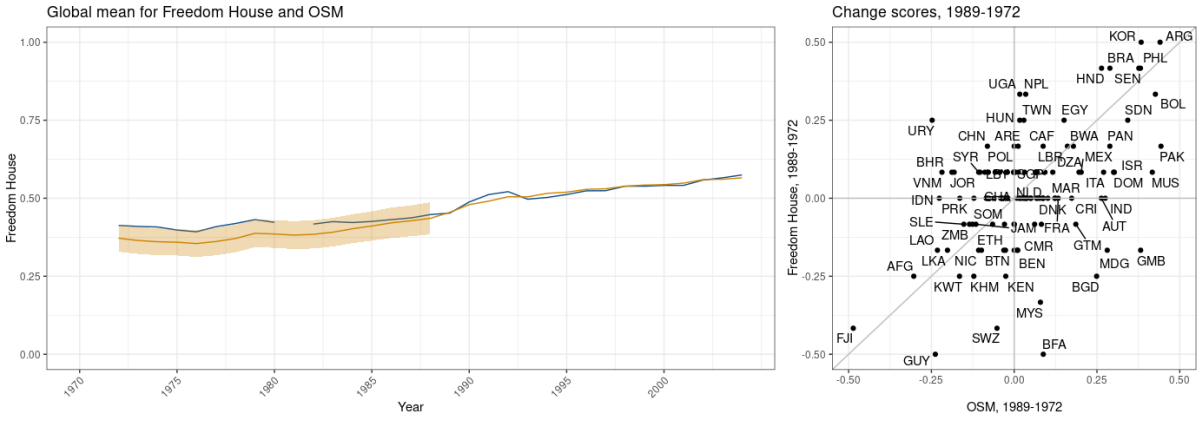


Figure SI 11.2: Predicting 1972-1988



SI 12 Predicting the interwar period

Figure SI 12.1: Polyarchy for the interwar period

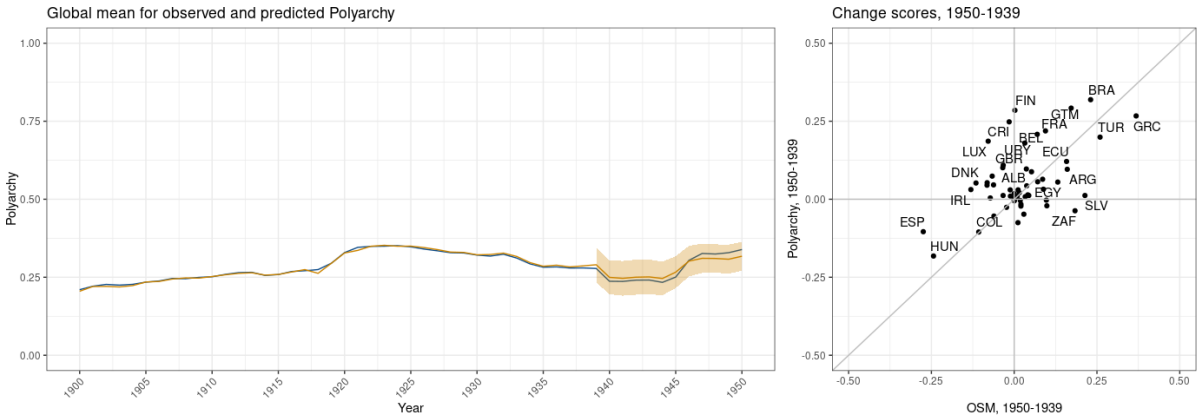
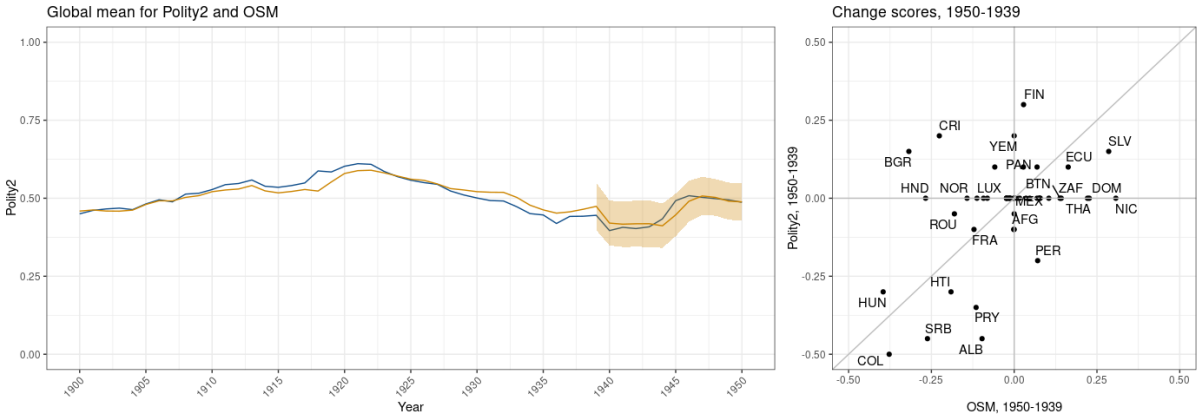


Figure SI 12.2: Polity2 for the interwar period



SI 13 Predicting the 1970s

Figure SI 13.1: Polyarchy for the 1970s

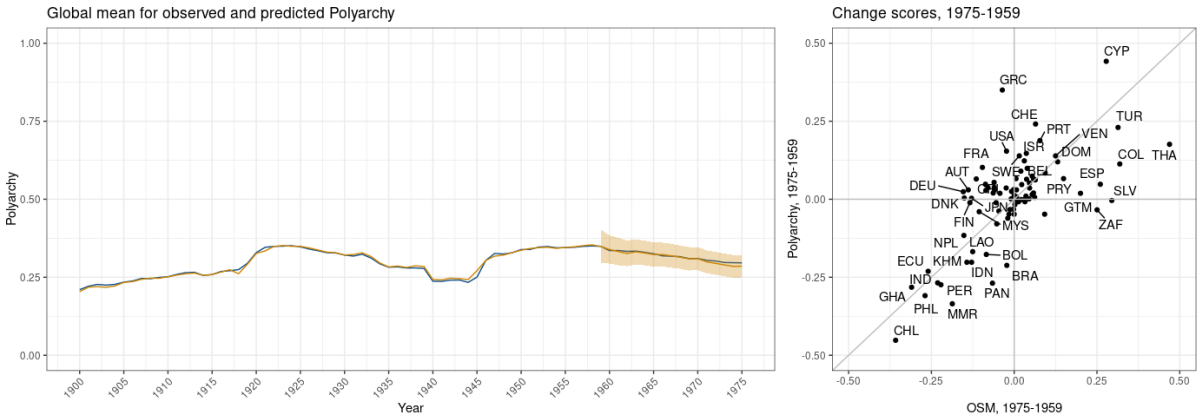
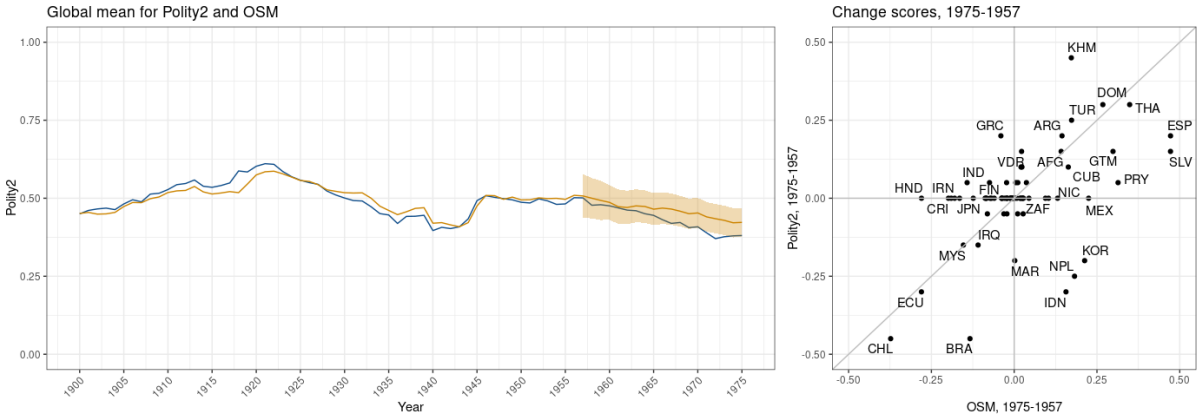


Figure SI 13.2: Polity2 for the 1970s



SI 14 Restricting the Polyarchy and Polity2 sample to 1972

Figure SI 14.1: Polyarchy since 1972

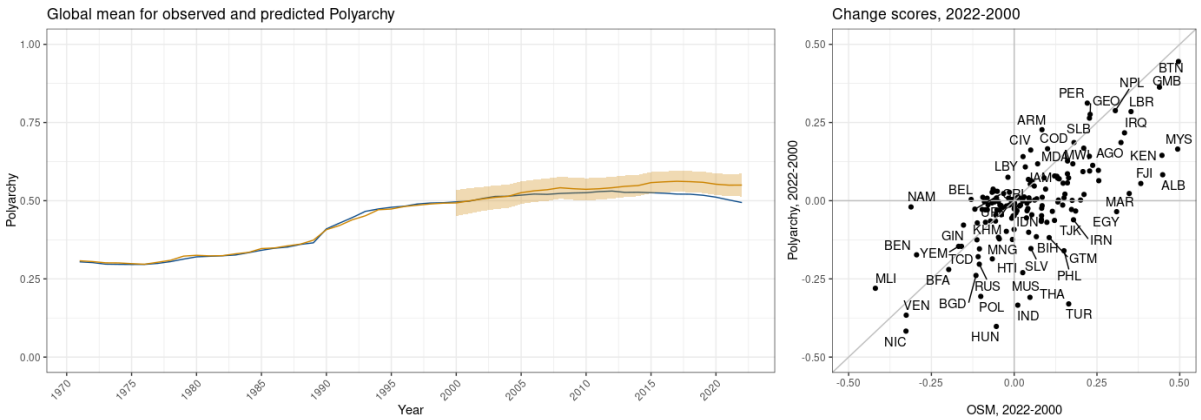
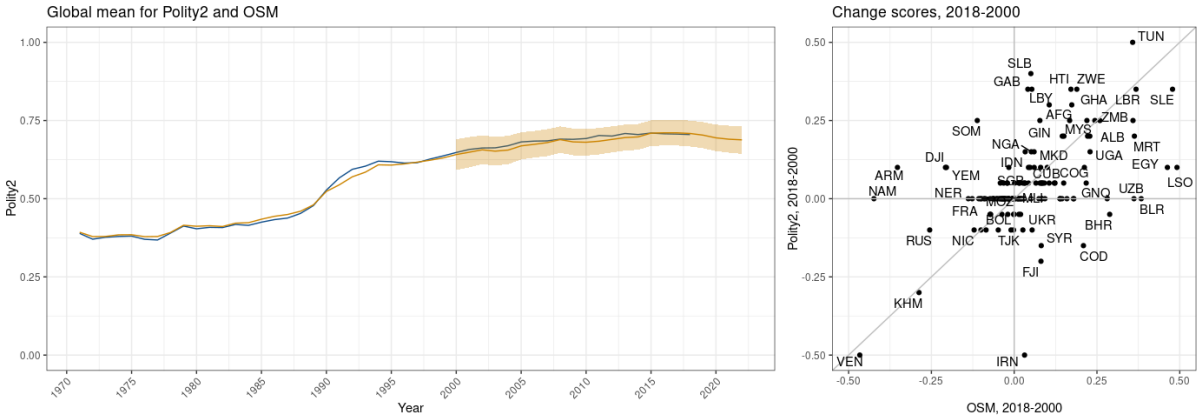


Figure SI 14.2: Polity2 since 1972



SI 15 Reversed prediction

In this appendix we are training our model on the (potential) backsliding period (2000-) and predict the pre-backsliding period (1979-1999).

Figure SI 15.1: Polyarchy prediction reversed

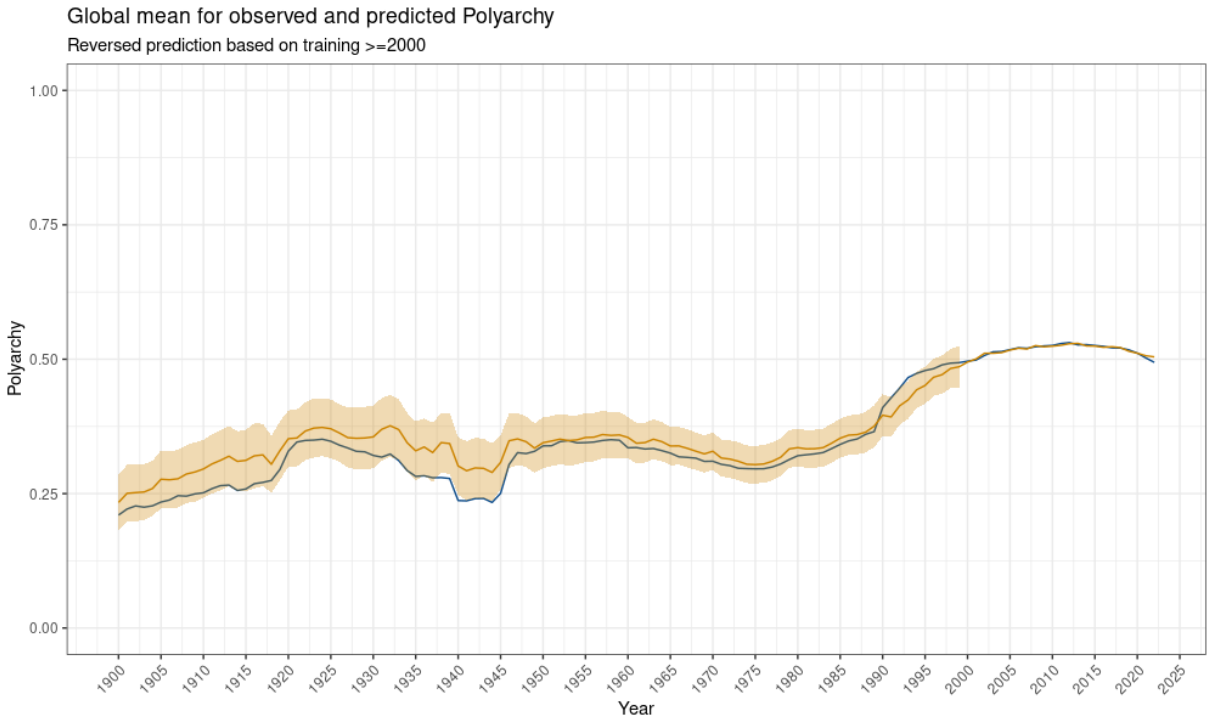
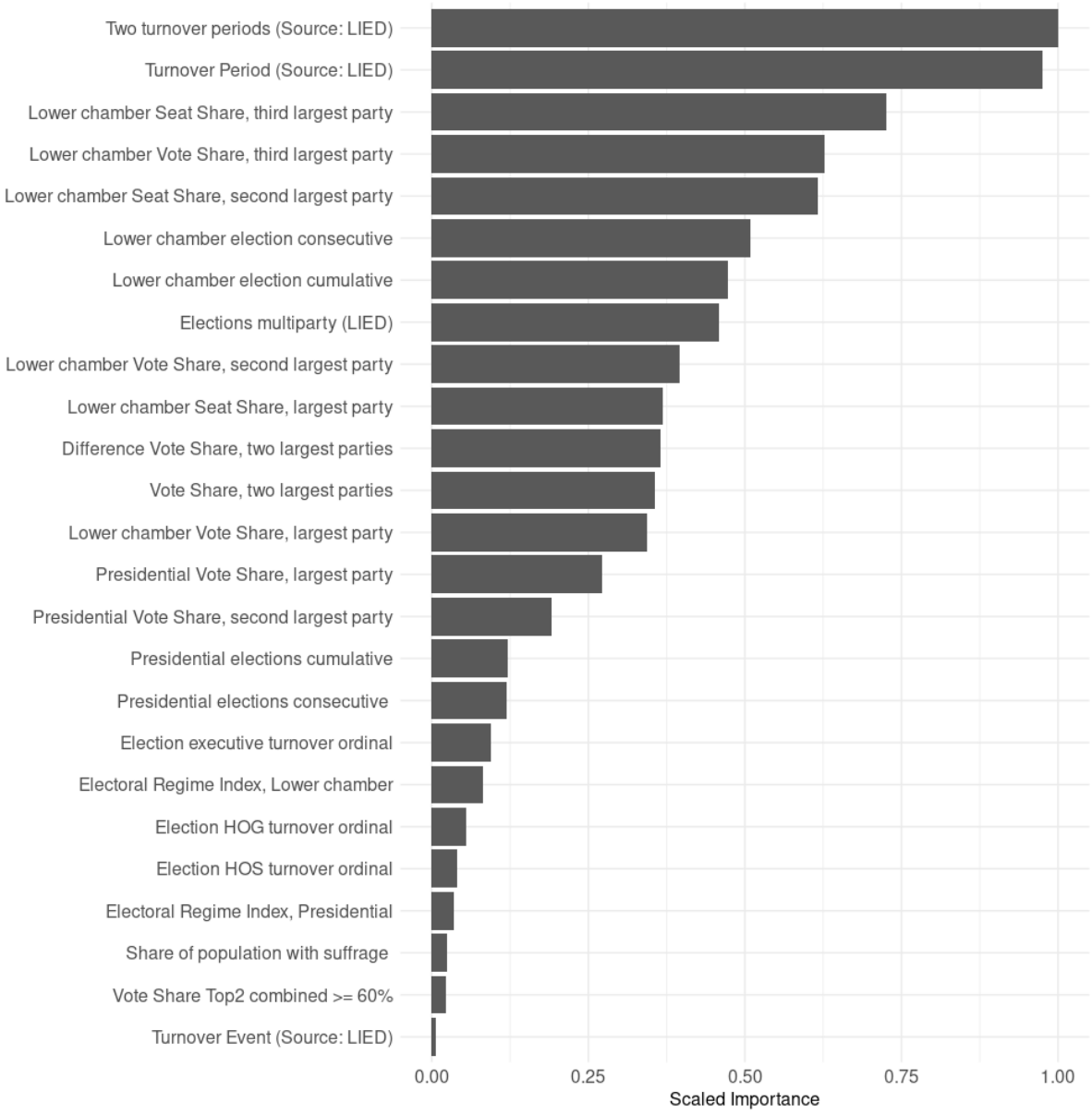


Figure SI 15.2: VIP Polyarchy prediction reversed



SI 16 Methodology

Our research methodology revolves around the utilization of random forests, which represent an ensemble learning approach. Unlike traditional single-model methods such as decision trees, random forests amalgamate multiple models to enhance robustness and accuracy. Random forests build upon decision trees, forming a collective of individual trees that work in synergy. In this section, we delve into a comprehensive explanation of decision trees and random forests, contextualizing their application in the realm of political science. We initiate by elucidating the fundamental concepts behind decision trees, the transition from individual trees to a forest, and reference seminal work by Guyon et al. (1997) along with subsequent research that has evolved around her contributions in the field of random forest analysis.

SI 16.1 The starting point: Decision Trees

Random forests serve as a machine learning algorithm rooted in the principles of decision trees. Decision trees, a category of supervised learning algorithms, find utility in both classification and regression tasks. They operate by recursively segmenting input data into subsets based on specific feature values, constructing a tree-like model that maps decisions to their potential outcomes. Within this framework, internal nodes within the tree represent features, while leaf nodes correspond to either a class or a regression value. The process of making predictions for new data points commences at the root node, with traversal along the appropriate branches guided by the values of relevant features, ultimately concluding at a leaf node that furnishes the associated class or regression value (Hastie, Tibshirani, and Friedman 2013; also see McAlexander and Mentch 2020 or Hill and Jones 2014 for applications in political science).

The essence of decision trees lies in their ability to construct a sequence of binary decisions predicated on input features (independent variables), culminating in predictions of the output class (dependent variable). Each decision manifests as a split in the tree, giving rise to distinct branches that lead to different sets of decisions or predictions. In essence, a decision tree can be conceptualized as a succession of if-then statements, culminating in a final prediction.

Decision trees share a foundational data structure with conventional statistical methods in the social sciences, involving an outcome variable and a collection of predictor or explanatory variables. The decision tree algorithm scrutinizes this data to identify the optimal data split (Step 1), selects the input feature that most effectively segregates the data, typically based on criteria such as information gain, gain ratio, or the Gini index (Step 2). Following the selection of the optimal split, the algorithm generates a new node within the tree (Step 3) that encapsulates the decision based on the chosen input feature. Subsequently, the data is partitioned into two branches, each representing one of the possible outcomes of the decision.

This process iterates recursively, with split selection and node creation repeated for each of the two branches generated in Step 3. This continues until predefined stopping criteria are met, which might include reaching a maximum tree depth, achieving a minimum number of data points in a leaf node, or obtaining a minimum information gain. Each branch in the tree encapsulates a sequence of decisions that collectively yield a prediction of the output variable.

Once the decision tree is fully constructed, making predictions for new data points involves traversing the tree from the root node to a leaf node, guided by the values of the input

features. At each node, a decision based on the input feature is made, and traversal continues down the corresponding branch until a leaf node is reached. The value contained in the leaf node represents the predicted output variable value (Hastie, Tibshirani, and Friedman 2013; Greenwell 2022).

SI 16.2 From a tree to a forest

Decision trees are susceptible to a phenomenon known as overfitting, particularly when they become excessively deep or intricate. In such cases, decision trees tend to capture the idiosyncrasies of the training data too closely, hindering their ability to generalize effectively to new, unseen data. To mitigate this challenge, random forests employ an ensemble approach, leveraging multiple decision trees to yield more robust and accurate predictions when applied to data not seen during training.

Within a random forest, each tree is trained using a randomly selected subset of the training data. Additionally, only a random subset of features is considered at each split. This strategy reduces correlation between the trees and diversifies their predictions while still preserving their individual strengths. The final prediction of the random forest is determined by either majority voting (for classification tasks) or the mean (for regression tasks) of the predictions from all constituent trees (Greenwell 2022).

In summary, decision trees constitute the foundational elements of random forests, serving as the building blocks for making individual predictions. Random forests, in turn, aggregate predictions from multiple decision trees to yield a more dependable and generalizable prediction (e.g., Muchlinski et al. 2016; also see responses by Wang 2019 and Muchlinski et al. 2019).

SI 16.3 Random Forest model application

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

1. Training set: This subset of data is utilized to train the random forest model. The model leverages the training set to comprehend the relationships between input features and target variable(s). The size of the training set should be sufficient to encompass data variability but not so extensive as to hinder training or induce overfitting.

2. Validation set: A subset of the data is earmarked for evaluating the model's performance during training. This set plays a crucial role in tuning the model's hyperparameters, such as the number of trees, maximum tree depth, and minimum samples required for node splitting. The validation set should be substantial enough to provide a reliable performance estimate without overfitting hyperparameters.

3. Cross-validation: Cross-validation is a technique for estimating model performance by dividing data into multiple folds. The model is trained on each fold and evaluated on the remaining folds. Cross-validation serves to estimate the model's generalization error and facilitates the selection of the best model from a set of candidate models. The number

of folds used in cross-validation depends on data size and available computational resources. Typically, 5-10 folds are used for small to medium-sized datasets, while 3-5 folds suffice for larger datasets. Cross-validation is conducted exclusively within the training set.

4. Test set: This subset of the data is reserved for evaluating the final model's performance after training and hyperparameter tuning. It serves to estimate the model's generalization error and allows for performance comparison against other models. The test set should be substantial enough to yield a reliable performance estimate without causing computational issues or significantly diminishing the size of the training set.

The determination of data split sizes in random forest analysis depends on several factors, including data size, model complexity, and computational resources. It is worth noting that split selection decisions are also contingent on the data distribution and necessitate ensuring that key data features are adequately represented in training, validation, and test data subsets.

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.

SI 16.4 Our model specification

Our model specifications are based on a hyper-parameter grid-search. We build 300 decision trees in the random forest, select three features randomly to be considered at each split of the tree, and use 70% of all available features for each tree. Across all models this specification has the highest predicitive power in the cross-validation. An important aspect of our data are temporal connections between country-years of a given country. We stratify our data by country, meaning that all country-year observations of a specific country are assigned either to the training, cross-validation, or test set. We do this in order to prevent "leakage" from the training data into the cross-validation or test data set. Leakage occurs when information from the training data set can "leak" into the other data sets and influence the model performance. Since country observations are likely highly correlated over the years having some country-years in the training and other country-years in the cross-validation or test set could be a problem. We use six-fold cross-validation for our training data (-2012).

We split the data into a training and a test set based on theoretical expectations. We want to train our model on data that is very unlikely to be influenced by potential backsliding biases by coders. In order to do so we set the cut-off at 2000. All country-years before 2000 are assumed to be free of backsliding coder bias and in this data our model learns the relationship between the democracy scores and our features. We then use this model, that was trained on the pre-backsliding period, to predict the backsliding period and examine the differences between observed and predicted democracy scores. We argue that small differences should be due to uncertainty in the predictions of the random forest and larger differences should be due to changes in the underlying data generating process between the training and the test data. Hence, if we observe large changes we take it to mean that the way democracy scores were coded before and after 2000 has changed in some systematic way. In order to make sure that our results are not a feature of our cut-off at 2000 we implement a series of robustness checks by setting the cut-off at 2005 and 2010. The results remain very similar and our conclusions remain.