Why Countries Revise Their Data*

Iasmin Goes†

July 2024

Abstract

Different sources of international statistics — like the World Development Indicators (WDI), the Penn World Table, or the Maddison Project — often provide conflicting information about a country's economic activity. Even different versions of the same data source contradict each other. Why? I use machine learning to understand what predicts variation in data coverage and revisions across all WDI releases from 1994 to 2021. Democracies with higher state capacity are less likely to report missing data and more likely to revise previous statistics. These findings suggest that revisions, unlike missing data, are not usually an ex-post attempt to manipulate existing information. Instead, revisions reflect a country's improving ability to quantify its economy. More broadly, these findings highlight the importance of disclosing the chosen data sources and versions, which might affect researchers' empirical results.

*Thanks to Brendan Apfeld, Sabrina Arias, Ryan Briggs, Terry Chapman, Andrés Cruz, Kerice Doten-Snitker, Timon Forster, Matt Hitt, Andrew Kerner, Anna Minasyan, Saliha Metinsoy, Carolina Moehlecke, Gaurav Sood, Martina Müller, and Daniel Weitzel for sharing data and/or providing feedback on various iterations of this paper.

†Assistant Professor, Colorado State University. Contact: iasmin.goes@colostate.edu.

1 Introduction

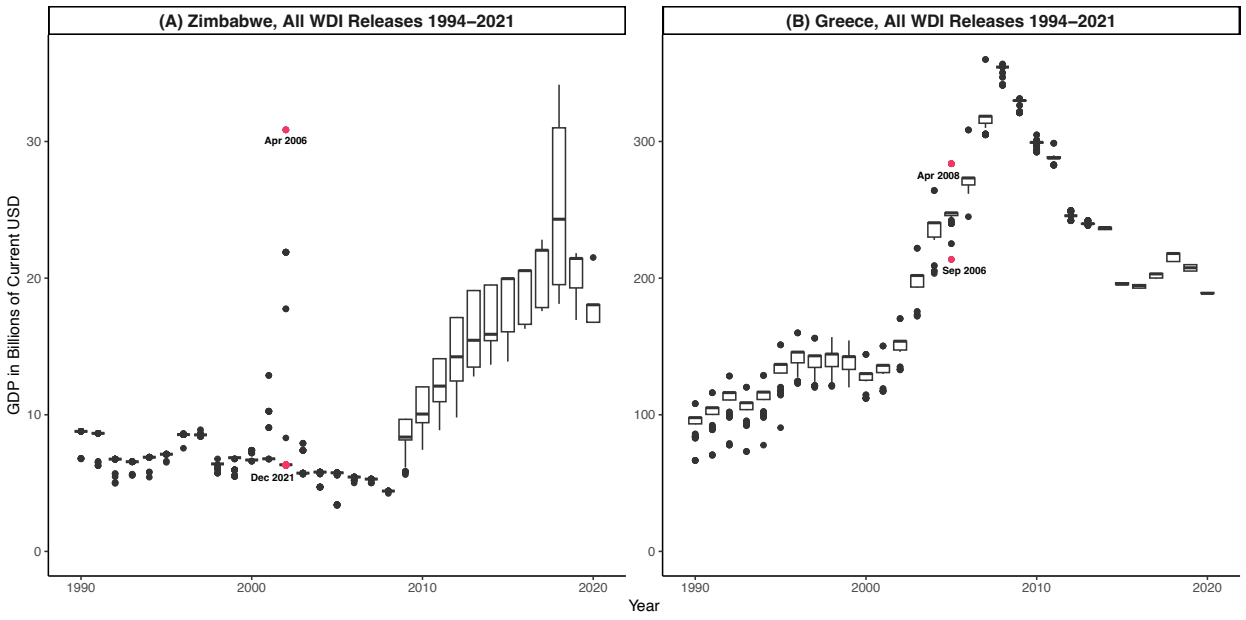
Performance indicators like the Ease of Doing Business index, the Millennium Development Goals, or the Freedom in the World report define global standards and rank countries according to their ability to meet such standards (Doshi, Kelley and Simmons, 2019; Bisbee et al., 2019). These indicators have recently come under fire for their inconsistency and politicization; in September 2021, for instance, the World Bank announced that it would discontinue the Ease of Doing Business index after an internal audit found irregularities in the coding process (World Bank, 2021). These and other latent measures of election integrity, state capacity, or democratic consolidation rely on expert coding, raising concerns that experts might be ideologically biased or improperly aggregate different ratings into one single measure (Bollen and Paxton, 2000; Giannone, 2010; Martínez i Coma and van Ham, 2015; Hanson and Sigman, 2021; McMann et al., 2022).

One might think that national accounts data are less controversial; in measuring observable and not latent concepts, they deliver a seemingly "neutral, sanitized, and objective expression of an unseen truth" (Ward, 2004, 25). Gross domestic product (GDP), the value of all final goods and services produced in a country during a specific period, is the most ubiquitous measure of national wealth: in 2020, 205 out of 206 economies surveyed by the International Monetary Fund (IMF) compiled annual GDP statistics (Baer, Guerreiro and Silungwe, 2022).[1] By comparison, only 109 compiled institutional sector accounts, such as deficit, debt, trade, and foreign direct investment (FDI).

GDP calculations have faced a fair share of criticism for delivering a biased oversimplification of the world (Mügge, 2022; Hoekstra, 2019; Fioramonti, 2013; Merry, 2011) — for example, by excluding unpaid household services, which are disproportionately performed by women (DeRock, 2021). But even if we take this indicator at face value and assume it is conceptually valid (that is, it accurately captures the underlying theoretical concept of national wealth), GDP measurements are not as reliable as they might seem. National

[1] Eritrea was the lone exception.

Figure 1: Current GDP of Zimbabwe and Greece, 1990–2020



These boxplots present the distribution of current GDP estimates for (A) Zimbabwe and (B) Greece from 1990 to 2020, using data drawn from the 104 WDI releases from April 1994 to December 2021. The estimate reported for Zimbabwe in 2002 is 24.5 billion dollars larger in the April 2006 WDI than in the December 2021 WDI. The estimate reported for Greece in 2005 is 58.5 billion dollars larger in the April 2008 WDI than in the April 2007 WDI. Section 3 discusses the data in more detail.

accounts data are not fixed data points: they are preliminary estimates that are constantly revised, and revisions might provide conflicting information. The two most common data sources in political science and economics — the World Development Indicators (WDI) and the Penn World Table (PWT), respectively (Goes, 2023; Johnson et al., 2013)[2] — often contradict each other. Indeed, different versions *of the same source* often provide conflicting information (see Figure 1). According to WDI figures released in April 2006, Zimbabwe's GDP in 2002 was around 30.8 billion current US dollars; the December 2021 WDI release reduced this number to just 6.3 billion. Industrialized democracies are not immune to this problem: Greece's reported GDP for 2005 increased nearly 26 percent in one year, from 225.2 billion in the April 2007 WDI to 283.7 billion in the April 2008 WDI. Even different

[2]According to these authors, other sources — like the Maddison Project, the IMF World Economic Outlook Database, and the UN National Accounts Main Aggregates Database — are considerably less common.

departments *within the same organization*, like the IMF, report different data (Pellechio and Cady, 2006). Why do these discrepancies exist?

Researchers do not know how far each measurement is from the *true* GDP (the measurement error). Still, they can quantify the reliability of GDP data (the measurement uncertainty) by comparing different sources or releases, as Fariss et al. (2022) do. Beyond that, it is crucial to understand why data are revised because national income estimates condition access to finance and voting power in international organizations. For example, countries with a per capita income above a certain threshold are not eligible for interest-free loans from the World Bank (Kerner, Jerven and Beatty, 2017). Data on GDP and international reserves are used to calculate member quotas that determine how much say countries have within the IMF (Pellechio and Cady, 2006). Withholding estimates can reduce bureaucratic quality (Williams, 2009), and uncertain estimates can compromise academic research. In a series of replications, Goes (2023), Johnson et al. (2013), and Croushore and Stark (2003) show that published research is not always robust to data changes; replacing the 2006 WDI with, say, the 2021 WDI might lead to very different empirical conclusions. At a minimum, researchers who know the origin of these discrepancies can better interpret their empirical results.

I begin by reviewing a rich literature that pinpoints several drivers of revisions, including low statistical capacity, large informal economies, and political incentives to misrepresent statistics. With this research as a starting point, I use machine learning to identify the systematic predictors of WDI revisions, missing data, and outliers. Democracies with high state capacity are less likely to report extreme values and more likely to disseminate data but also more likely to revise existing statistics. Given that these countries have the resources and political incentives to improve their data, revisions are plausibly reducing the measurement error but increasing the measurement uncertainty: they are getting closer to the truth at the expense of consistency. Still, not all data issues can be systematically predicted; many are distinctive to specific countries and years.

Just as it is common to present robustness checks with alternative measures of regime type (like Polity or Polyarchy), researchers should estimate separate models with alternative measures of GDP, exports, foreign aid, FDI, or economic growth from different sources and vintages. This should be the case even for studies focusing on industrialized democracies with purportedly high-quality data. To better allow for comparisons and robustness checks, the online appendix of this study provides GDP data (in both current and constant dollars) for all available WDI vintages since 1994, consolidated into one single file. This reflects the need to be transparent about the data sources and vintages. In addition, researchers should not draw conclusions based on recent years alone; information for these years is particularly noisy and susceptible to revisions. Ultimately, scholars should be modest when interpreting empirical findings, particularly if these findings are seemingly novel and counter-intuitive: one cannot trust the results of empirical models unless one can trust the underlying data.

2 Why Revise?

Different data sources can diverge significantly, even if the underlying information is the same. Ram and Ural (2014) identify 33 cases for which GDP estimates from the WDI and the PWT differ by over 25 percent. This issue goes beyond GDP: exporters and importers record the same bilateral trade flows differently (Linsi, Burgoon and Mügge, 2023), and a comparison of export data from two sources — the IMF and the UN Commodity Trade Statistics — concludes that "the data are neither comparable nor in a number of cases, correlated" (Amin Gutiérrez de Piñeres, 2006, 35). Foreign aid (Michaelowa and Michaelowa, 2011; Weikmans and Roberts, 2019), FDI (Kerner, 2014), and population data (Devarajan, 2013) face similar measurement issues.

Revisions happen for five primary reasons. The first is lack of statistical capacity. National statistical offices (NSOs) are supposed to collect data in line with a global standardization framework, the System of National Accounts (SNA). However, communist countries

did not begin to adopt the SNA until 1993; North Korea still appears to use a Marxism-inspired alternative, the Material Product System (Herrera 2010, 23n8; van Heijster and DeRock 2022, 84n1). Even NSOs that adopted the SNA might be underfunded, understaffed, use outdated methods, or experience frequent turnover. A 2005 survey by the UN Economic Commission for Africa found that most NSOs in the continent had three to 12 national accountants (United Nations Economic Commission for Africa, 2005). A 2023 survey of 14 NSOs, conducted by the Inter-American Development Bank, found that only half of the employees working with statistical analysis displayed basic competence in probability, descriptive statistics, survey sampling, and arithmetic (Mejía Guerra et al., 2023, 14). Population figures tend to be extrapolated from the last census; since censuses are expensive, countries like Lebanon (which last conducted a census in 1932) rely on extrapolations that grow progressively inaccurate over time (Devarajan, 2013). And different agencies within one country — say, the Korean Economic Planning Board and the Korean Central Bank — might contradict each other (Pellechio and Cady, 2006). Consequently, governments either fail to report estimates altogether or report inaccurate estimates. While this problem is prevalent in Africa (Jerven, 2010, 2013, 2018, 2019), agencies in Latin America and elsewhere also struggle with difficult-to-measure concepts like imputed rent, thus underestimating household final consumption expenditure — an important component of GDP (Olinto Ramos, Pastor and Rivas, 2008).

International organizations are not involved in data collection, only standardization (Ward, 2004, 98). This is the second driver of revisions. Every five to ten years, the International Comparison Program (ICP) surveys how much the same basket of goods costs in different currencies and constructs purchasing power parity (PPP) exchange rates. These exchange rates, in turn, are used to convert SNA data from nominal (current) to PPP terms, which are comparable across borders. Until 1996, ICP price surveys only covered the developed world, making less accurate extrapolations for the developing world (Deaton and Aten, 2017). ICP rounds in 2005, 2011, and 2017 reduced uncertainty by including large developing

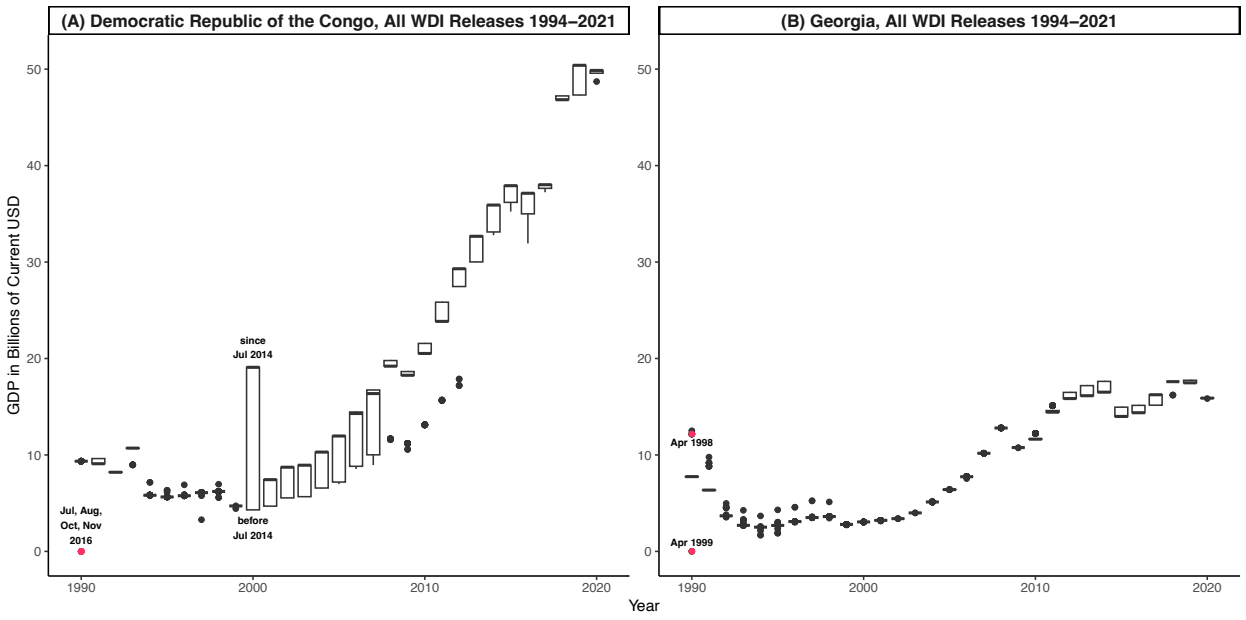
countries, but price surveys in China were only conducted in urban areas, introducing yet another potential source of error (Bolt and van Zanden, 2024). Overall, ICP rounds disagree with each other due to differences in relative prices, consumption patterns, region-specific PPP adjustments, and accounting or reporting practices (Deaton and Aten, 2017).

Data might also be revised for political reasons. Autocracies are less likely to report policy-relevant data (Hollyer, Rosendorff and Vreeland, 2011), and when they do, they overstate growth rates (Martínez, 2022; Magee and Doces, 2015), particularly in politically sensitive times (Wallace, 2014). Between 2000 and 2005, Burundi halted the collection of national accounts data due to civil war (Randriambolamanitra, Ligbet and Stalom Kamga, 2014, 6). In federations like Nigeria, states inflate population figures to receive higher fiscal transfers from the federal government (Devarajan, 2013). Aid-dependent countries systematically understate their finances to appear poorer and attract more aid (Kerner, Jerven and Beatty, 2017). Even industrialized democracies overstate how much climate aid they provide — particularly when domestic constituencies value environmental objectives (Michaelowa and Michaelowa, 2011) — and misrepresent public finance statistics to abide by the rules of the European Union, as Greece did (Alt, Lassen and Wehner, 2014).

The fourth driver of revisions is a large informal economy. In 2014, EU countries revised their GDP calculations to include drug trafficking and prostitution; as a result, the Italian and British economies increased by four percent each (Coyle, 2014, 110). In 2016, the Irish Central Statistics Office reported a GDP growth of 26 percent but refused to say why, citing statistical confidentiality rules. Instead, it introduced a new measure of economic output, Modified Gross National Income (GNI*), to remove "globalization-related" distortions. Years later, economists discovered what had prompted these distortions: Apple's decision to onshore intellectual property assets to Ireland in 2015 (Polyak, 2023). Measurement uncertainly is even higher in developing countries, where the informal economy accounts for up to 44 percent of the GDP (Coyle, 2014, 110).

Humans are a final driver of revisions: they commit coding errors, selectively exclude

Figure 2: Current GDP of the Democratic Republic of the Congo and Georgia, 1990–2020



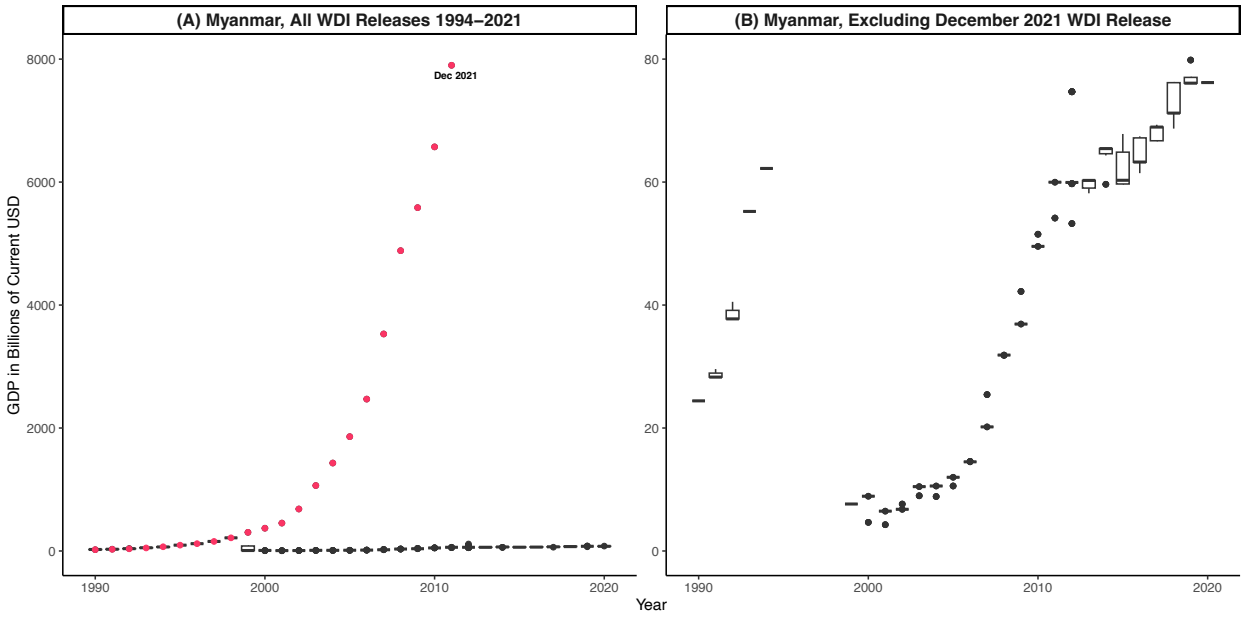
These boxplots present the distribution of current GDP estimates from 1990 to 2020 for (A) the Democratic Republic of the Congo and (B) Georgia, using data drawn from the 104 WDI releases from April 1994 to December 2021. Four WDI releases reported a GDP of zero for the Democratic Republic of the Congo in 1990. In addition, all 32 releases before July 2014 reported a GDP of 4.3 billion for 2000, a figure revised to 19.1 billion in July 2014. Georgia's GDP in 1990 was reported as 12.1707 *million* in some vintages and 12.1707 *billion* in others. Section 3 discusses the data in more detail.

available data, or weigh summary statistics inappropriately. As Figure 2 shows, four different WDI releases (in July, August, October, and November 2016) reported the GDP of the Democratic Republic of the Congo in 1990 as *zero*; two other releases (in December 2016 and April 2017) reported this value as missing. Georgia's 1990 GDP — reported to be around 12.1707 *billion* until April 1998 — "lost" three digits in the April 1999 and April 2000 vintages, shrinking to 12.1707 *million* before regaining its billionaire status in April 2003.[3] The December 2021 update contains a similar error for Myanmar, illustrated in Figure 3. In nearly all available vintages, Myanmar's GDP in 2011 ranged from 54 to 59 billion current US dollars. However, the December 2021 release reported a figure over 100 times as high: 7.899 *trillion*. The February 2022 update corrected this mistake. But

[3]Georgia only gained formal independence from the Soviet Union in December 1991, but its WDI coverage begins in 1990.

individuals who downloaded *any* WDI data in the preceding two months likely retrieved wrong numbers, as all GDP-based variables (including constant GDP, GDP in PPP, GDP per capita, and GDP growth) use current GDP as a starting point for calculations.

Figure 3: Current GDP of Myanmar, 1990–2020



These boxplots present the distribution of current GDP estimates from 1990 to 2020 for Myanmar, using data drawn from the 104 WDI releases from April 1994 to December 2021. The December 2021 WDI release (in pink) is included in (A), but not in (B). As the different y-axes show, the December 2021 release was an outlier, reporting exceptionally high values for the entire time series. Section 3 discusses the data in more detail.

These issues are not unsolvable. International organizations matter: subscribing to the IMF's Special Data Dissemination Standard (SDDS) increases transparency even after accounting for self-selection (Vadlamannati, Cooray and Brazys, 2018). Following the end of its civil war, Burundi received technical assistance from AFRISTAT and financing from the African Development Bank to resume its data collection and update its SNA (Randriambolamanitra, Ligbet and Stalom Kamga, 2014, 6). With support from the Danish International Development Agency and the IMF, the Ghana Statistical Service released new GDP estimates in 2010: after upgrading from the 1968 to the 1993 SNA, including new data disaggregated by economic sector, it concluded that the country's GDP was 60.3 percent

larger than previously thought (Jerven and Ebo Duncan, 2012). Likewise, Nicaragua revised its national accounts in 2003, changing the base year from 1980 to 1994 and implementing the 1993 SNA; as a result, the country's current GDP for the year 2000 increased by 70 percent (Olinto Ramos, Pastor and Rivas, 2008, 9). Political leadership can also make a difference: Greece revised its finances after Prime Minister George Papandreou came to power in 2009 and requested help from Eurostat and the IMF (Aragão and Linsi, 2022). These revisions increased the accuracy of Ghanaian, Nicaraguan, and Greek statistics, but also reduced their reliability, given the gap between old and new estimates. Finally, data transparency and replication can identify human errors. A replication exercise led Herndon, Ash and Pollin (2014) to identify serious miscalculations in a famous study connecting higher sovereign debt to lower GDP growth.

Revisions are consequential for both policy and research. In 2010, the World Bank upgraded Ghana from low income to lower middle income economy, a shift associated with less generous lending terms. That same year, Greece was downgraded by credit rating agencies and requested multiple IMF and EU loans to avoid default. Given the discrepancies between WDI and PWT vintages, replacing one source or vintage with another can also significantly alter published research findings (Goes, 2023; Johnson et al., 2013; Croushore and Stark, 2003). Overall, there is widespread heterogeneity in data quality: researchers can make more precise inferences about some countries and years than others.

3 Predicting Revisions

3.1 GDP Data

The WDI first appeared as a printed annex to the 1978 World Development Report and became a standalone publication in 1997 (World Bank, 2018). In 2018, the World Bank discontinued print reports and launched a data portal that includes the WDI Database

Archives, providing 104 electronic WDI releases from 1994 to 2021.[4] I focus on the indicator *GDP in current US dollars* (ID `NY.GDP.MKTP.CD`), the annual "sum of gross value added by all resident producers in the economy." The production approach is the most widely compiled and disseminated approach to GDP estimation (Baer, Guerreiro and Silungwe, 2022, 12). Current GDP enables comparisons across vintages, though not across countries or over time, as it does not make PPP or inflation adjustments.[5] I examine GDP in billions, rounded to two decimal places. Zimbabwe's 1990 GDP, for example, was reported as 8,783,816,666 dollars in all vintages from April 2011 to November 2014 and 8,783,816,700 dollars in all vintages since December 2014. This difference of 34 dollars is negligible and increases computational demands without any substantive gain in meaning, so I treat both values as 8.78 billion.

3.2 Operationalizing Revisions

For *GDP in current US dollars*, consider each observation x_{itk} for country i, year t (the *reported* date), and WDI release k (the *reporting* date), with $N = 429,261$. The main outcome of interest is *Revision*, coded one if x_{itk} is different from x_{itk+1} (that is, if the value reported for country i and year t differs between two consecutive vintages) and zero otherwise. Revisions are not intrinsically problematic. Changes between vintages might reflect ex post data manipulation or an improvement in statistical capacity, though it is often difficult to distinguish between both.

Besides *Revision*, I examine two related outcomes.[6] *Missingness*, a well-studied proxy for transparency (Hollyer, Rosendorff and Vreeland, 2011), is coded one if x_{itk} is missing from release k and zero otherwise. Missingness is intrinsically problematic because the WDI only

[4] Though all releases since 1989 are available, the indicator of interest is missing from all releases before 1994, and the WDI released no data updates in 1996.

[5] *GDP, PPP (current international $)* (ID `NY.GDP.MKTP.PP.CD`) allows for comparisons across countries, but not across vintages, as the PPP conversion factor changes from one ICP round to another. *GDP in constant US dollars* (ID `NY.GDP.MKTP.KD`), calculated using the GDP deflator (the ratio of GDP in current local currency to GDP in constant local currency) to account for inflation, allows for comparisons over time, but not across vintages.

[6] Appendix E presents results for additional outcomes: the z-score, the percentage change from the median, and the speed with which data are reported, also a proxy for transparency (Islam, 2006).

omits "questionable" observations (Hollyer, Rosendorff and Vreeland, 2014, 414).

For all non-missing values ($N = 404,480$), *Outlier* is coded one if x_{itk} falls outside of the typical ranges for country i and year t and zero otherwise. To identify outliers, the Tukey rule (used to construct boxplots) leverages the Interquartile Range (IQR), the difference between the third quartile (Q3) and the first quartile (Q1); x_{itk} is an outlier if it falls below $Q1 - 1.5 \times IQR$ or above $Q3 + 1.5 \times IQR$.

3.3 Modeling Strategy

This study is exploratory. I do not know the nature of the underlying data generating process and do not expect one predictor to matter more than others in explaining data revisions. Absent a strong theory driving the selection of predictor variables, tree-based models tend to outperform linear regression. Researchers can include any number of variables; trees choose relevant predictors and filter out irrelevant ones. These models make no assumptions about functional form and are robust to including predictors with outliers or long-tailed distributions. Instead of using listwise deletion or imputation, the algorithm interprets missing predictor values as a separate category containing information, assuming that values are missing not at random. This is desirable, as missing predictors could be related to measurement uncertainty in GDP data. As with linear regression, tree-based models do not identify causal relationships; they merely show whether variation in one indicator is associated with variation in another indicator.

Both classification trees (with categorical outcomes) and regression trees (with continuous outcomes) assume that all observations are part of one covariate space (Montgomery and Olivella, 2016). The model splits this covariate space into non-exhaustive and overlapping regions, each corresponding to a unique covariate combination, and makes one prediction for all observations falling within one region. To ensure that the data are not fragmented too quickly, with too many regions, the model grows trees through sequential binary splits (rather than multiway splits) and follows the best split at each step, without looking ahead.

Since a single tree can be sensitive to data changes, most researchers grow tree ensembles to reduce variance. Two tree-based ensemble models — random forests and gradient boosting machines (GBM) — tend to outperform other tree-based or non-tree-based models in predicting US Supreme Court rulings (Kaufman, Kraft and Sen, 2019), civil war onset (Muchlinski et al., 2016), allocation of government expenditures (Funk, Paul and Philips, 2022), regime type (Weitzel et al., 2023), and other "complicated" data generating processes with nonlinearities, discontinuities, additive terms, or interactions (Montgomery and Olivella, 2016). Random forests are *forests* because they build an ensemble of trees and *random* because each binary split of a tree makes predictions using a random sample of covariates, aggregating the results based on the prediction made by most trees. Even if there is a strong predictor in the dataset, not all trees use this strong predictor in the first split. The resulting trees are less correlated with each other, with more reliable average results (Breiman, 2001). While random forests build trees simultaneously, GBMs build trees sequentially, with each new tree designed to rectify the mistakes of its predecessors. This sequential refinement, driven by gradient descent optimization, enables GBMs to capture complex relationships in the data, though it also risks overfitting (Cook, 2017, 147). I use random forests to understand how GDP varies across different WDI vintages; Appendix D presents the results of alternative models.

Following conventions in machine learning, I split the data into training, validation, and test sets accounting for 60, 20, and 20 percent of all observations, respectively, stratified by World Bank income group to ensure that all income groups are represented proportionally across sets.[7] One concern is that unintentional information might leak across related observations. For instance, the model might use information from Zimbabwe's past to predict Zimbabwe's future (temporal leakage) or from Zimbabwe in 1990 to predict outcomes for

[7]The World Bank classifies countries into four groups: low income, lower middle income, upper middle income, or high income. Based on the classification for the 2024 fiscal year, these groups account for 15.30, 29.03, 26.45, and 28.65 percent of the dataset, respectively. The remaining 0.57 percent correspond to Venezuela, which has been temporarily unclassified since July 2021 due to lack of revised national accounts statistics.

other countries in 1990 (spatial leakage). Either way, the model would return predictions that are too good to be true (Kaufman et al., 2012). To address leakage, I use group-based splitting and leave-one-group-out cross-validation (LOGOCV). Group splitting means that all observations for a specific country are assigned to the same set, such that Zimbabwe's past cannot be used to predict Zimbabwe's future. LOGOCV means that I train each iteration of the model on the entire training set minus one country, then evaluate how well the model generalizes to the left-out country. After iterating through all countries, the algorithm builds a final model for the entire training set, without partitions, comparing this model's performance to the average performance of the cross-validation models. Based on several metrics, the algorithm selects the model that best explains variation in the training data while making accurate predictions for the new data. This helps ensure that the final model does not overfit to the patterns specific to, say, Zimbabwe.

I use the training and validation sets to iteratively calibrate the model, adjusting hyperparameters like the number of trees or the number of splits per trees.[8] Once I am satisfied with the results, I use the chosen model to make out-of-sample predictions for the test set. This final evaluation on unseen data provides a reliable measure of the model's predictive capability and its real-world applicability.

3.4 Predictors

Though tree-based models can handle several predictors, there is a trade-off: the model should include enough predictors to capture important patterns without being overly complex and fitting noise. With this in mind, I collect 46 variables that fall under the five primary drivers of revisions outlined in previous sections (see Appendix C for full list). To measure the first driver of revisions, statistical capacity, I collect information about census frequency, freedom of information laws, statistical agencies, participation in IMF data dissemination initiatives (like the aforementioned SDDS), and SNA in use. The driver of

[8] See Appendix F for a discussion of the chosen hyperparameters and a description of `H2O`, the machine learning platform used to implement this algorithm.

revisions is standardization. Current GDP does not require PPP conversions; according to the WDI Metadata, "dollar figures for GDP are converted from domestic currencies using single year official exchange rates." However, if "the official exchange rate does not reflect the rate effectively applied to actual foreign exchange transactions," the World Bank applies an alternative exchange rate. Correspondingly, *Alternative Conversion Factor* takes the value of one whenever this alternative exchange rate is used.

The third driver of revisions is politics, operationalized in various ways: regime type, Polyarchy scores, ideology of the executive, occurrence of specific events (like elections, financial crises, or civil wars), and International Country Risk Guide (ICRG) expert ratings (such as law and order, ethnic tensions, and investment profile). The main models do not include the size of the informal sector (the fourth driver of revisions) or economic and demographic predictors (such as FDI flows, inflation, unemployment, population, or urbanization rates) because most of these predictors are also reported by the WDI, strongly correlate with GDP data, and tend to be revised just as frequently. Appendix E reports models with these additional predictors.

The last driver of revisions, and perhaps the most difficult to predict, is human error. *Coding Error* takes the value of one for the errors identified in Figures 2 and 3, in Appendix B, or listed by the World Bank in its Data Updates and Errata website.[9] Yet many less conspicuous errors go unnoticed and are impossible to predict on a systematic basis.

About 39.8 percent of all economies surveyed by the IMF in 2020 disseminated their annual GDP data within 90 days of the reference period, whereas 54.4 percent did so within 91 to 365 days, with no available information for the remaining 5.8 percent (Baer, Guerreiro and Silungwe, 2022, 17). Considering this trend, and the fact that several predictors are not available before 1990 or after 2021, I examine each reported year t from 1990 to 2020 for each reporting year k from 1994 to 2021.[10] Most predictors are included for both k and

[9]For example: "The September update of the WDI 2009 database contained an error for China's current U.S. dollar GDP for 2007 and 2008" (World Bank, 2023).

[10]Since most predictors are only available for sovereign states, Georgia, Timor-Leste, South Sudan, and others that did not exist in 1990 only enter the analysis after independence, though WDI coverage often

t, as current circumstances might motivate retroactive changes to older data. For example, the Greek government revised existing statistics after Prime Minister Papandreou came to power in 2009, so Greek statistics with $k \geqslant 2009$ could be different from previous vintages.

3.5 Main Results

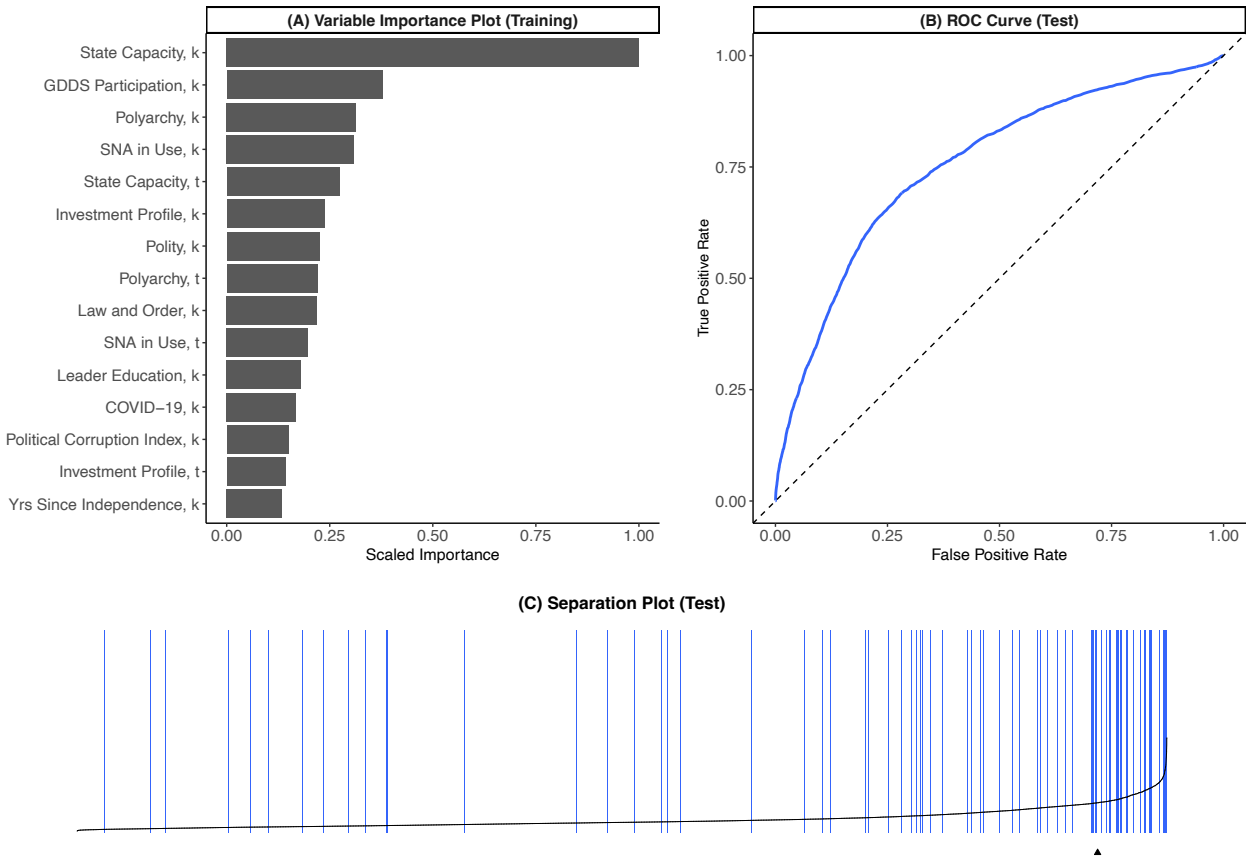
Revision is a binary outcome, and most observations belong in one class: only 7.42 percent record a substantial change. Thus, I follow Muchlinski et al. (2016) and balance the majority and minority classes in the training set. In Figure 4, panel (A) presents the relative importance of the 15 most important predictors for the training set. The least important predictor equals zero, while the most important predictor equals one. The importance of each predictor is a function of whether it was selected to create a binary split, and if so, how much the squared error (averaged over all trees) increased or decreased because of said split.

The four most important predictors of revision all refer to the reporting year k: state capacity (Hanson and Sigman, 2021), participation in the IMF's General Data Dissemination System (GDDS), SNA in use, and the *Polyarchy* score. Tree-based models do not identify causal relationships, so it is inaccurate to say that a country's extractive, coercive, and administrative capacity *causes* revisions. But partial dependence plots, presented in Appendix D, allow me to identify the direction of each effect: more democratic countries, with higher state capacity, are more likely to revise their data. This is reflected by the fact that the GDDS, called e-GDDS (enhanced GDDS) since 2015, is open to all IMF member countries and merely provides recommendations for data dissemination, whereas the SDDS (mentioned in previous sections) has stricter standards and is available only to member countries that have or seek access to international capital markets. As of 2024, 190 countries participate in these initiatives, which are not mutually exclusive: China, Kazakhstan, Mongolia, Romania, Senegal, and others progressed from the GDDS to the SDDS over time. But the average GDDS participant is a developing country that does not yet meet a high-quality data stan-

begins before independence (see Footnote 3).

dard (Vadlamannati, Cooray and Brazys, 2018); unsurprisingly, this country tends to be associated with fewer revisions. In contrast, oil discoveries, military regimes, and coups are not strongly associated with revisions and nor is the existence of a statistical agency (see Appendix D for full variable importance plot).

Figure 4: Assessing the Performance of a Model Predicting Data Revisions



These figures assess the fit of a random forest predicting the outcome *Revision*. Panel (A) is a variable importance plot, indicating the relative importance of the 15 most important predictors; t or k denotes the predictor's value for the reported or reporting year, respectively. Panel (B) is a Receiver Operating Characteristic (ROC) curve, illustrating the trade-off between the true positive rate and the false positive rate across different probability thresholds. Panel (C) is a separation plot that organizes the predicted probabilities for each observation in ascending order, highlighting whether each observation corresponds to an instance of *Revision*.

My primary goal is to uncover relationships between variables, not optimize predictive power, but I assess the quality of the out-of-sample predictions as a final step. Following Muchlinski et al. (2016), panels (B) and (C) in Figure 4 present a Receiver Operating Char-

acteristic (ROC) curve and a separation plot, respectively. In panel (B), the ROC curve illustrates the trade-off between the true positive rate and the false positive rate across different probability thresholds. The y-axis represents the true positive rate (the proportion of revised observations correctly classified as revised), whereas the x-axis represents the false positive rate (the proportion of non-revised observations incorrectly classified as revised). A random model would produce a diagonal line from the bottom-left corner to the top-right corner, whereas a perfect classifier would achieve a true positive rate of 1 and a false positive rate of 0, corresponding to the top-left corner of the plot. These figures are paired with a performance metric, the Area Under the ROC Curve (AUC), which ranges from 0 to 1, with 0.5 denoting random guessing and 1 denoting a perfect classifier. The AUC value for the test set (0.75) indicates that the model makes good out-of-sample predictions: it can typically distinguish between true positives and false positives, between observations that are truly revised and observations that are not.

Moving to panel (C), the separation plot organizes the predicted probabilities for each observation in ascending order, highlighting whether each observation corresponds to an instance of *Revision*. The plot also provides information about the predicted probabilities (a black line) and the expected number of events (a black triangle). When the model makes perfect predictions, there is a clear separation between zeroes and ones: lower probabilities (in white) are always associated with no event (left of the triangle) and higher probabilities (in blue) are always associated with an event (right of the triangle). Deviations from this ideal pattern highlight areas where the model struggles to distinguish between the classes. The separation plot confirms that the model does a good — if imperfect — job of predicting revisions. Appendix D presents additional performance metrics.

In sum, data revisions are more frequent as countries become more democratic and improve their statistical capacity (by updating their SNA, for example). Assuming that democracies with high statistical capacity are more committed to high-quality data, these results suggest that revisions are generally desirable: most of them reflect a country's ability

to measure its GDP more precisely, not a deliberate attempt to retroactively falsify data. But not all revisions are the same. Below, I explore the drivers of missing or extreme values, both of which are unequivocally problematic.

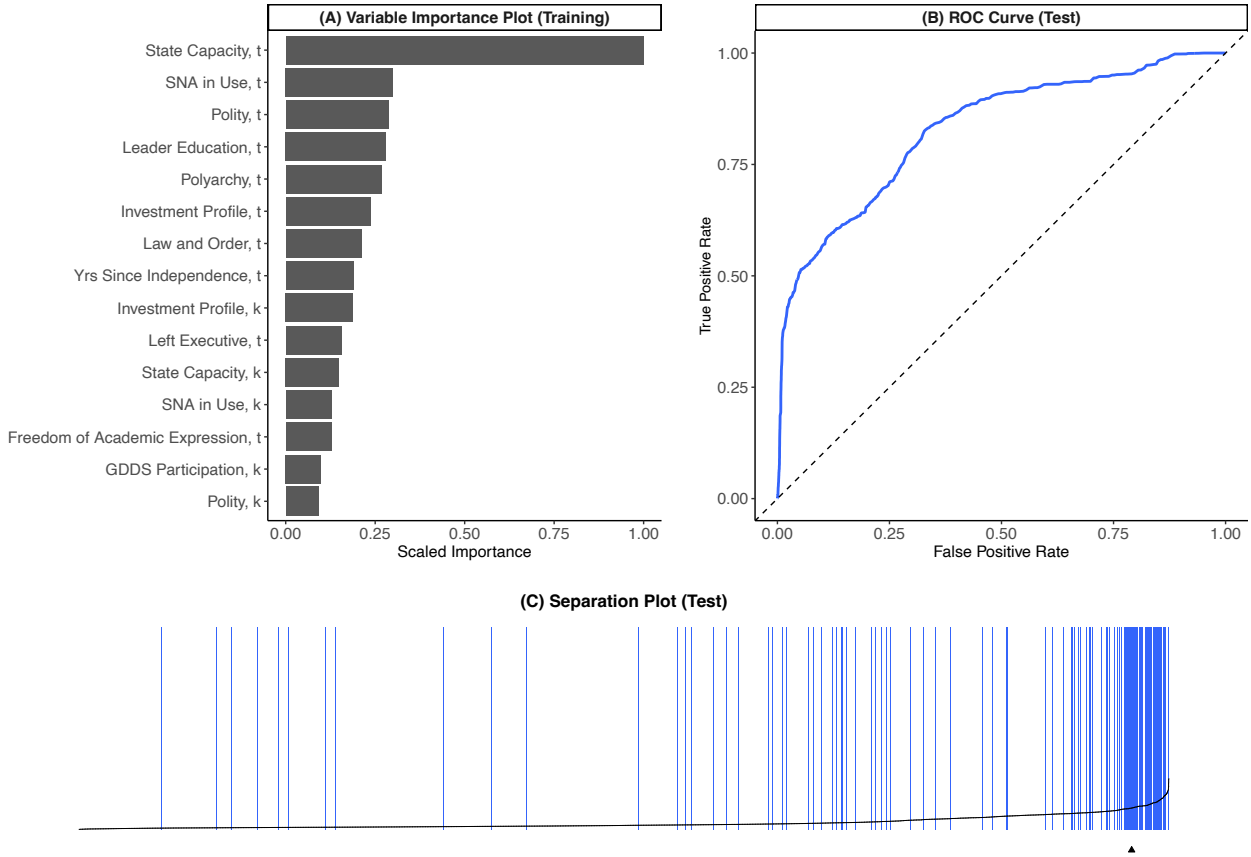
3.6 Additional Results

Missing and *Outlier* are also binary outcomes with most observations belonging in one class: 94.22 percent are *not* missing and 91.8 percent of all non-missing observations are *not* outliers. Correspondingly, I balance the majority and minority classes, as before.

Hollyer, Rosendorff and Vreeland (2014, 417) show that WDI data disclosure is a "political decision, not simply a reflection of bureaucratic capacity." In Figure 5, panel (A) reflects a mix of both. Hanson and Sigman's state capacity measure is the most important predictor of *Missing*, but observations are also more likely to be missing for countries with less educated leaders or lower Polity scores, consistent with previous findings that autocrats are more prone to withholding data (Hollyer, Rosendorff and Vreeland, 2011). Whereas revisions are best explained by information from reporting year k, missingness is best explained by information from reported year t. And higher state capacity, while associated with *more* revisions, is associated with *less* missing data. Again, these relationships are not causal; other studies show that release of information *causes* better bureaucratic capacity, not necessarily the reverse (Williams, 2009).

As before, the ROC curve in panel (B) and the separation plot in panel (C) confirm that the model makes good predictions; the AUC value for the test set is 0.83. Yet some instances of missingness cannot be predicted on a systematic basis. As an illustration, consider New Zealand's 2012 GDP. Once it enters the analysis in October 2013, its predicted probability of missingness is zero for all vintages, a correct prediction for all but one vintage: December 2015. The February 2016 WDI explains: "Corrections have been made to ... GDP-related data for New Zealand from 2012-15" (World Bank, 2023). New Zealand's 2012 GDP is missing from the December 2015 WDI for idiosyncratic reasons that the model is unable to

Figure 5: Assessing the Performance of a Model Predicting Missing Data

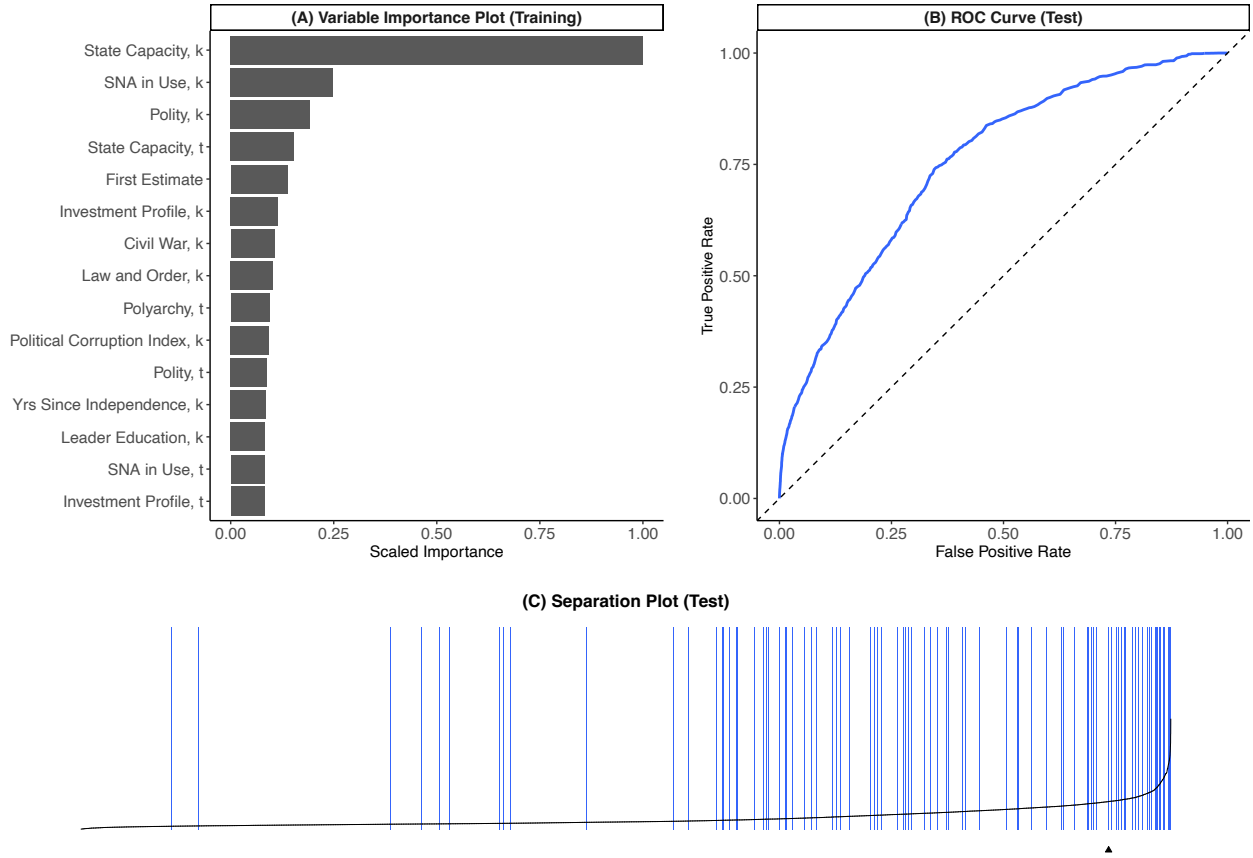


These figures assess the fit of a random forest predicting the outcome *Missing*. Panel (A) is a variable importance plot, indicating the relative importance of the 15 most important predictors; t or k denotes the predictor's value for the reported or reporting year, respectively. Panel (B) is a Receiver Operating Characteristic (ROC) curve, illustrating the trade-off between the true positive rate and the false positive rate across different probability thresholds. Panel (C) is a separation plot that organizes the predicted probabilities for each observation in ascending order, highlighting whether each observation corresponds to an instance of *Missing*.

predict correctly. And some observations might be missing due to human error, as was likely the case for the four different WDI releases that reported the 1990 GDP for Democratic Republic of the Congo as zero (see Figure 2).

Turning to all non-missing observations, Figure 6 shows — yet again — that state capacity and regime type are important predictors of *Outlier*. Though the model says nothing about the *direction* of the outlier, this is consistent with research showing that autocrats overstate GDP growth rates (Magee and Doces, 2015; Martínez, 2022). Among the five

Figure 6: Assessing the Performance of a Model Predicting Outliers



These figures assess the fit of a random forest predicting the outcome *Outlier*. Panel (A) is a variable importance plot, indicating the relative importance of the 15 most important predictors; t or k denotes the predictor's value for the reported or reporting year, respectively. Panel (B) is a Receiver Operating Characteristic (ROC) curve, illustrating the trade-off between the true positive rate and the false positive rate across different probability thresholds. Panel (C) is a separation plot that organizes the predicted probabilities for each observation in ascending order, highlighting whether each observation corresponds to an instance of *Outlier*.

most important predictors is *First Estimate*, denoting whether a WDI release is the first to include the country-year pair in question. This first attempt is often a preliminary best guess corrected in subsequent vintages. For example, the first estimate for Estonia's 1995 GDP appeared in the April 1997 WDI: 60.8 billion. All subsequent vintages reported a value 13 to 16 times smaller, reflecting the fact that WDI vintages tend to coalesce around one single value over time — presumably because measurement improves over time.[11] In

[11] See Appendix B for a discussion of Estonia and other countries with extreme values.

addition, observations following more recent versions of the SNA are less likely to have an extreme value. As before, coups, military regimes, and the COVID-19 pandemic explain little variation in the outcome of interest. The models predictions are adequate if imperfect, with an AUC of 0.75 for the test set.

These intuitive results show that discrepancies between WDI vintages are not restricted to individual cases. By and large, these issues are *systematic*: we can trust a single measurement of GDP for some countries and years far more than for others, and we can identify ahead of time which measurements are most trustworthy. At the same time, models might not correctly predict extreme cases due to the idiosyncrasies of each country, year, and vintage. This might be due to a specific event (like Zimbabwe's 2002 election, which was neither free nor fair) or due to the government's deliberate choice to misrepresent GDP data (as in Greece).

4 Conclusions

Political scientists have long debated how to measure latent concepts like democracy (Munck and Verkuilen, 2002; Giannone, 2010; Coppedge and Gerring, 2011) without devoting as much attention to the measurement of observable concepts, which can be just as difficult to quantify. GDP is the foundation for calculating multiple variables in social science research, including foreign aid, FDI, and trade flows. Even when GDP is "just" a control variable, it is important to take its measurement seriously, as its inclusion might affect the sample size and shape researchers' conclusions about the relationship between other variables (Goes, 2023).

Two common solutions to the problem of missing data are listwise deletion (excluding cases with missing values) and multiple imputation (generating multiple plausible values for the missing observations). Both provide unbiased estimates if data are missing completely at random: missingness is not related to observed or unobserved factors. Multiple imputation also provides unbiased estimates if data are missing at random: missingness is related to

observed but not unobserved factors. Yet GDP is missing not at random: missingness is related to observed factors (which models can predict) *and* unobserved factors (which models cannot). Consequently, multiple imputation and listwise deletion would be biased (Pepinsky, 2018). The same applies to revised and extreme values: a GDP of zero for the Democratic Republic of the Congo is wrong, but deleting such observation or replacing it with plausible values (perhaps borrowed from other WDI releases) could generate bias. Researchers do not know the true data-generating process underlying national accounts statistics. There are significant regional disparities in statistical capacity. When statistical capacity is high, there might be political interest in reporting biased (or no) data. Even when statistical capacity is high *and* there is political interest in reporting accurate data, GDP is intrinsically difficult to quantify. Outliers might be honest mistakes, and revisions might be a sincere attempt to fix them.

Besides listwise deletion and multiple imputation, a more realistic solution is resampling. While traditional bootstrap methods involve random sampling with replacement from the entire dataset, a leave-one-group-out bootstrap can systematically exclude one country at a time during resampling iterations, allowing researchers to assess whether results are robust to omitting individual countries.[12] Even if someone working with the December 2021 WDI is not aware of Myanmar's extreme values, resampling will ensure that the empirical results are not driven by such outliers — provided the goal is to make generalizations across countries. Making specific statements about Myanmar might prove more challenging, hence the importance of also using alternative GDP measures and examining trends over time. This study is accompanied by a GitHub repository, updated every month, that consolidates GDP data from all available WDI vintages since 1994. Yet scholars do not need to peruse 104 WDI vintages to recognize that there is something wrong with a GDP of zero. They only need to take their data seriously — and this includes control variables.

As Herrera and Kapur (2007, 381) state, "the penalties for using low-quality data are

[12]This is similar to LOGOCV performed in the random forest, except LOGOCV leaves out one group at a time without replacement, whereas a leave-one-group-out bootstrap resamples with replacement.

small." Still, researchers should be transparent about the data origins and research implications, acknowledging that the choice of one source or vintage over another can affect the empirical conclusions. In particular, researchers should use recent data releases (recent values of k). Newer vintages, which rely on more recent SNA versions and (in the case of PPP data) more recent ICP rounds, provide more precise information for developing countries and are more consistent. Researchers should also consider dropping recent years (recent values of t) from the analysis, if only in robustness checks. For example, GDP estimates for 2018, 2019, and 2020 were first available in the February 2020, February 2021, and February 2022 WDI releases, respectively. Someone using the February 2022 WDI might not want to include 2019 and 2020 in their analysis, as the numbers reported for these years are preliminary, possibly extreme, and bound to change in subsequent data releases. These revisions can happen for good reason — perhaps countries are improving their data collection process and correcting previous mistakes, or the World Bank is refining its data standardization tools. Either way, scholars who eliminate more recent observations ensure that their empirical results are not just the product of unstable measurements that have not yet coalesced around a single value.

References

Alt, James, David Dreyer Lassen and Joachim Wehner. 2014. "It Isn't Just about Greece: Domestic Politics, Transparency and Fiscal Gimmickry in Europe." *British Journal of Political Science* 44(4):707–716.

Amin Gutiérrez de Piñeres, Sheila. 2006. "What a Difference a Source Makes! An Analysis of Export Data." *Applied Economics Letters* 13(1):35–39.

Aragão, Roberto and Lukas Linsi. 2022. "Many Shades of Wrong: What Governments Do When They Manipulate Statistics." *Review of International Political Economy* 29(1):88–113.

Baer, Andrew, Vanda Guerreiro and Anthony Silungwe. 2022. "2020 Global Stocktaking of National Accounts Statistics: Availability for Policy and Surveillance." *IMF Working Papers* (29):1–25.

Bisbee, James H., James R. Hollyer, B. Peter Rosendorff and James Raymond Vreeland. 2019. "The Millennium Development Goals and Education: Accountability and Substitution in Global Assessment." *International Organization* 73(3):547–578.

Bollen, K. A. and P. Paxton. 2000. "Subjective Measures of Liberal Democracy." *Comparative Political Studies* 33(1):58–86.

Bolt, Jutta and Jan Luiten van Zanden. 2024. "Maddison-Style Estimates of the Evolution of the World Economy: A New 2023 Update." *Journal of Economic Surveys* pp. 1–41.

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

Cook, Darren. 2017. *Practical Machine Learning with H2O: Powerful, Scalable Techniques for Deep Learning and AI.* Sebastopol, CA: O'Reilly.

Coppedge, Michael and John Gerring. 2011. "Conceptualizing and Measuring Democracy: A New Approach." *Perspectives on Politics* 9(2):247–267.

Coyle, Diane. 2014. *GDP: A Brief But Affectionate History.* Princeton and Oxford: Princeton University Press.

Croushore, Dean and Tom Stark. 2003. "A Real-Time Data Set for Macroeconomists: Does the Data Vintage Matter?" *Review of Economics and Statistics* 85(3):605–617.

Deaton, Angus and Bettina Aten. 2017. "Trying to Understand the PPPs in ICP 2011: Why Are the Results So Different?" *American Economic Journal: Macroeconomics* 9(1):243–64.

DeRock, Daniel. 2021. "Hidden in Plain Sight: Unpaid Household Services and the Politics of GDP Measurement." *New Political Economy* 26(1):20–35.

Devarajan, Shantayanan. 2013. "Africa's Statistical Tragedy." *Review of Income and Wealth* 59(S1):9–15.

Doshi, Rush, Judith G. Kelley and Beth A. Simmons. 2019. "The Power of Ranking: The Ease of Doing Business Indicator and Global Regulatory Behavior." *International Organization* 73(3):611–643.

Fariss, Christopher J., Therese Anders, Jonathan N. Markowitz and Miriam Barnum. 2022. "New Estimates of Over 500 Years of Historic GDP and Population Data." *Journal of Conflict Resolution* 66(3):553–591.

Fioramonti, Lorenzo. 2013. *Gross Domestic Problem: The Politics Behind the World's Most Powerful Number*. London: Zed Books.

Funk, Kendall D., Hannah L. Paul and Andrew Q. Philips. 2022. "Point Break: Using Machine Learning to Uncover a Critical Mass in Women's Representation." *Political Science Research and Methods* 10(2):372–390.

Giannone, Diego. 2010. "Political and Ideological Aspects in the Measurement of Democracy: The Freedom House Case." *Democratization* 17(1):68–97.

Goes, Iasmin. 2023. "New Data, New Results? How Data Vintaging Affects the Replicability of Research." *Research and Politics* (April-June):1–13.

Hanson, Jonathan K. and Rachel Sigman. 2021. "Leviathan's Latent Dimensions: Measuring State Capacity for Comparative Political Research." *Journal of Politics* 83(4):1–16.

Herndon, Thomas, Michael Ash and Robert Pollin. 2014. "Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff." *Cambridge Journal of Economics* 38(2):257–279.

Herrera, Yoshiko M. 2010. *Mirrors of the Economy: National Accounts and International Norms in Russia and Beyond*. Ithaca: Cornell University Press.

Herrera, Yoshiko M. and Devesh Kapur. 2007. "Improving Data Quality: Actors, Incentives, and Capabilities." *Political Analysis* 15(4):365–386.

Hoekstra, Rutger. 2019. *Replacing GDP by 2030: Towards a Common Language for the Well-Being and Sustainability Community*. Cambridge: Cambrdge University Press.

Hollyer, James R., B. Peter Rosendorff and James Raymond Vreeland. 2011. "Democracy and Transparency." *Journal of Politics* 73(4):1191–1205.

Hollyer, James R., B. Peter Rosendorff and James Raymond Vreeland. 2014. "Measuring Transparency." *Political Analysis* 22(4):413–434.

Islam, Roumeen. 2006. "Does More Transparency Go Along With Better Governance?" *Economics and Politics* 18(2):121–167.

Jerven, Morten. 2010. "Accounting for the African Growth Miracle: The Official Evidence – Botswana 1965–1995." *Journal of Southern African Studies* 36(1):73–94.

Jerven, Morten. 2013. "Comparability of GDP Estimates in Sub-Saharan Africa: The Effect of Revisions in Sources and Methods Since Structural Adjustment." *Review of Income and Wealth* 59(S1):1–21.

Jerven, Morten. 2018. "Controversy, Facts and Assumptions: Lessons from Estimating Long Term Growth in Nigeria, 1900–2007." *African Economic History* 46(1):104–136.

Jerven, Morten. 2019. "The History of African Poverty By Numbers: Evidence and Vantage Points." *Journal of African History* 59(3):449–461.

Jerven, Morten and Magnus Ebo Duncan. 2012. "Revising GDP Estimates in Sub-Saharan Africa: Lessons from Ghana." *African Statistical Journal* 15:13–24.

Johnson, Simon, William Larson, Chris Papageorgiou and Arvind Subramanian. 2013. "Is Newer Better? Penn World Table Revisions and Their Impact on Growth Estimates." *Journal of Monetary Economics* 60(2):255–274.

Kaufman, Aaron Russell, Peter Kraft and Maya Sen. 2019. "Improving Supreme Court Forecasting Using Boosted Decision Trees." *Political Analysis* 27:381–387.

Kaufman, Shachar, Saharon Rosset, Claudia Perlich and Ori Stitelman. 2012. "Leakage in Data Mining: Formulation, Detection, and Avoidance." *ACM Transactions on Knowledge Discovery from Data* 6(4):1–21.

Kerner, Andrew. 2014. "What We Talk About When We Talk About Foreign Direct Investment." *International Studies Quarterly* 58(4):804–815.

Kerner, Andrew, Morten Jerven and Alison Beatty. 2017. "Does It Pay to Be Poor? Testing for Systematically Underreported GNI Estimates." *Review of International Organizations* 12(1):1–38.

Linsi, Lukas, Brian Burgoon and Daniel Mügge. 2023. "The Problem with Trade Measurement in IR." *International Studies Quarterly* 67(2):1–18.

Magee, Christopher S.P. and John A. Doces. 2015. "Reconsidering Regime Type and Growth: Lies, Dictatorships, and Statistics." *International Studies Quarterly* 59(2):223–237.

Martínez i Coma, Ferran and Carolien van Ham. 2015. "Can Experts Judge Elections? Testing the Validity of Expert Judgments for Measuring Election Integrity." *European Journal of Political Research* 54(2):305–325.

Martínez, Luis R. 2022. "How Much Should We Trust the Dictator's GDP Growth Estimates?" *Journal of Political Economy* 130(10):2731–2769.

McMann, Kelly, Daniel Pemstein, Brigitte Seim, Jan Teorell and Staffan Lindberg. 2022. "Assessing Data Quality: An Approach and An Application." *Political Analysis* 30(3):426–449.

Mejía Guerra, José Antonio, Christian Schuster, Magdalena Rojas Wettig, Kim Sass Mikkelsen and Jan Meyer-Sahling. 2023. *Making National Statistical Offices Work Better:*

Evidence from a Survey of 13,300 National Statistical Office (NSO) Employees in 14 Latin American and Caribbean Countries. Washington, D.C.: Inter-American Development Bank.

Merry, Sally Engle. 2011. "Measuring the World: Indicators, Human Rights, and Global Governance." *Current Anthropology* 52(S3):S83–S95.

Michaelowa, Axel and Katharina Michaelowa. 2011. "Coding Error or Statistical Embellishment? The Political Economy of Reporting Climate Aid." *World Development* 39(11):2010–2020.

Montgomery, Jacob M. and Santiago Olivella. 2016. "Tree-Based Models for Political Science Data." *American Journal of Political Science* 62(3):729–744.

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(1):87–103.

Mügge, Daniel. 2022. "Economic Statistics as Political Artefacts." *Review of International Political Economy* 29(1):1–22.

Munck, Gerardo L. and Jay Verkuilen. 2002. "Conceptualizing and Measuring Democracy: Evaluating Alternative Indices." *Comparative Political Studies* 35(1):5–34.

Olinto Ramos, Roberto, Gonzalo Pastor and Lisbeth Rivas. 2008. "Latin America: Highlights from the Implementation of the System of National Accounts 1993 (1993 SNA)." *IMF Working Paper* 08(239):1–51.

Pellechio, Anthony and John Cady. 2006. "Differences in IMF Data: Incidence and Implications." *IMF Staff Papers* 53(2):326–349.

Pepinsky, Thomas B. 2018. "A Note on Listwise Deletion versus Multiple Imputation." *Political Analysis* 26(4):480–488.

Polyak, Palma. 2023. "Jobs and Fiction: Identifying the Effect of Corporate Tax Avoidance Inflating Export Measures in Ireland." *Journal of European Public Policy* 30(10):2143–2164.

Ram, Rati and Secil Ural. 2014. "Comparison of GDP Per Capita Data in Penn World Table and World Development Indicators." *Social Indicators Research* 116(2):639–646.

Randriambolamanitra, Samuel, Magloire Ligbet and Alain Magloire Stalom Kamga. 2014. *Revue par les pairs de la fiabilité de comptes - Cas du Burundi*. Tunis: African Development Bank.

United Nations Economic Commission for Africa. 2005. Assessment of the Implementation of the 1993 System of National Accounts in Africa. In *Fourth Meeting of the Committee on Development Information (CODI-IV)*.

Vadlamannati, Krishna Chaitanya, Arusha Cooray and Samuel Brazys. 2018. "Nothing to Hide: Commitment to, Compliance With, and Impact of the Special Data Dissemination Standard." *Economics and Politics* 30(1):55–77.

van Heijster, Joan and Daniel DeRock. 2022. "How GDP Spread to China: The Experimental Diffusion of Macroeconomic Measurement." *Review of International Political Economy* 29(1):65–87.

Wallace, Jeremy L. 2014. "Juking the Stats? Authoritarian Information Problems in China." *British Journal of Political Science* 46(1):11–29.

Ward, Michael. 2004. *Quantifying the World: UN Ideas and Statistics*. Bloomington and Indianapolis: Indiana University Press.

Weikmans, Romain and J. Timmons Roberts. 2019. "The International Climate Finance Accounting Muddle: Is There Hope on the Horizon?" *Climate and Development* 11(2):97–111.

Weitzel, Daniel, John Gerring, Daniel Pemstein and Svend-Erik Skaaning. 2023. "Measuring Electoral Democracy with Observables." *American Journal of Political Science* (forthcoming).

Williams, Andrew. 2009. "On the Release of Information by Governments: Causes and Consequences." *Journal of Development Economics* 89(1):124–138.

World Bank. 2018. *World Development Indicators: The Story.*

URL: *https://datatopics.worldbank.org/world-development-indicators/stories/world-development-indicators-the-story.html*

World Bank. 2021. *World Bank Group to Discontinue Doing Business Report.*

URL: *https://www.worldbank.org/en/news/statement/2021/09/16/world-bank-group-to-discontinue-doing-business-report*

World Bank. 2023. *Data Updates and Errata.*

URL: *https://datahelpdesk.worldbank.org/knowledgebase/articles/906522-data-updates-and-errata*

Appendix for
Why Countries Revise Their Data

July 2024

Contents

A Countries Included in the Analysis **2**

B Additional Descriptive Information **2**

C List of Predictors **6**

D Performance Metrics and Alternative Models **14**
D.1 Revisions 14
D.2 Missingness 19
D.3 Outlier 22

E Alternative Predictors and Outcomes **24**
E.1 Alternative Predictors: WDI 24
E.2 Alternative Outcome: Z-Score 28
E.3 Alternative Outcome: Percentage Deviation From the Median 30
E.4 Alternative Outcome: Reporting Speed 32

F Hyperparameters **34**
F.1 Classification Trees 34
F.2 Regression Trees 35

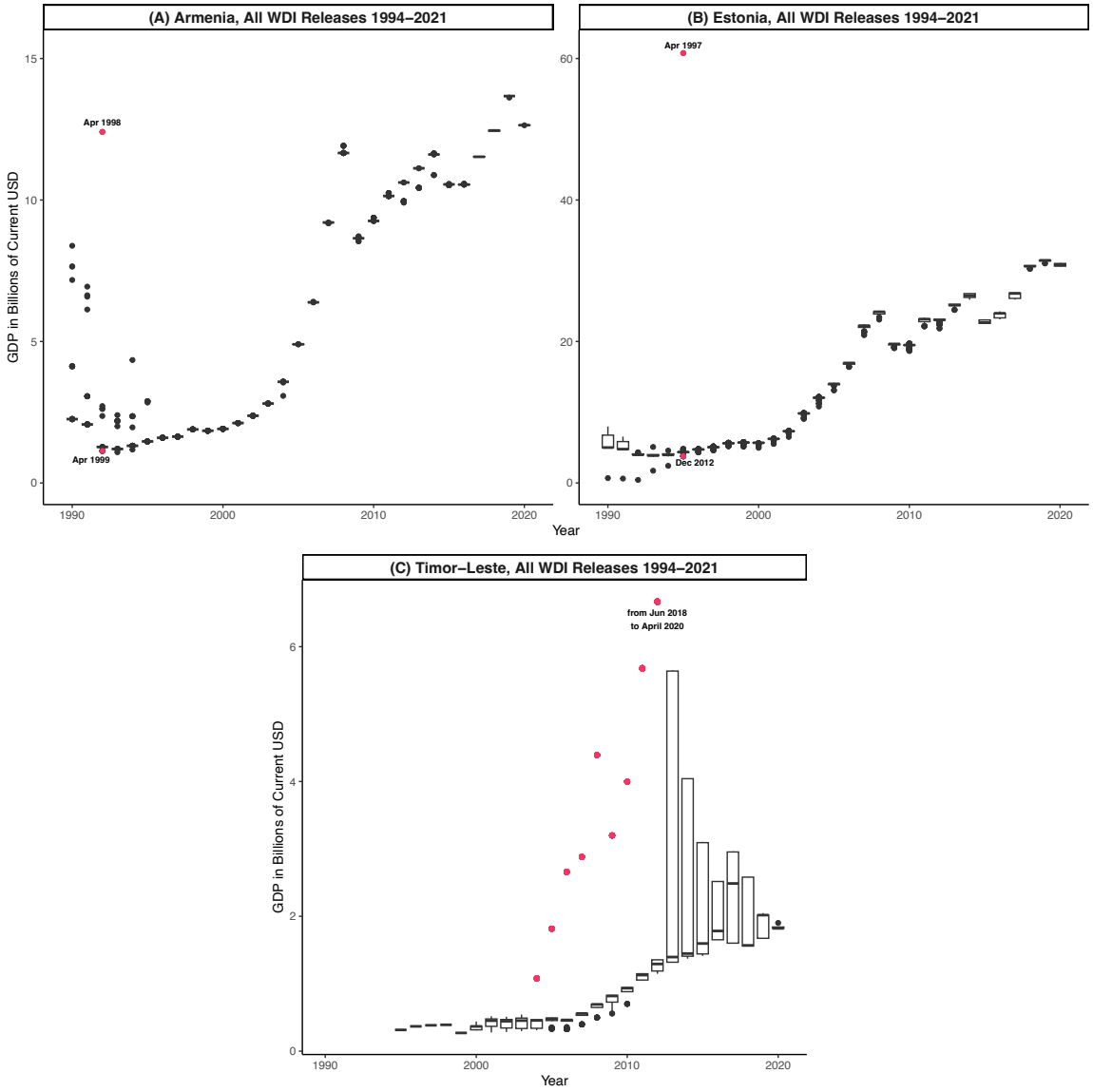
A Countries Included in the Analysis

Afghanistan, Albania, Algeria, Angola, Argentina, Armenia, Australia, Austria, Azerbaijan, Bahamas, Bahrain, Bangladesh, Barbados, Belarus, Belgium, Belize, Benin, Bhutan, Bolivia, Bosnia and Herzegovina, Botswana, Brazil, Brunei Darussalam, Bulgaria, Burkina Faso, Burundi, Cabo Verde, Cambodia, Cameroon, Canada, Central African Republic, Chad, Chile, China, Colombia, Comoros, Congo, Costa Rica, Cote d'Ivoire, Croatia, Cuba, Cyprus, Czech Republic, Democratic Republic of the Congo, Denmark, Djibouti, Dominican Republic, Ecuador, Egypt, El Salvador, Equatorial Guinea, Eritrea, Estonia, Eswatini, Ethiopia, Fiji, Finland, France, Gabon, Gambia, Georgia, Germany, Ghana, Greece, Grenada, Guatemala, Guinea, Guinea-Bissau, Guyana, Haiti, Honduras, Hungary, Iceland, India, Indonesia, Iran, Iraq, Ireland, Israel, Italy, Jamaica, Japan, Jordan, Kazakhstan, Kenya, Kiribati, Kuwait, Kyrgyzstan, Laos, Latvia, Lebanon, Lesotho, Liberia, Libya, Lithuania, Luxembourg, Madagascar, Malawi, Malaysia, Maldives, Mali, Malta, Mauritania, Mauritius, Mexico, Moldova, Mongolia, Montenegro, Morocco, Mozambique, Myanmar, Namibia, Nepal, Netherlands, New Zealand, Nicaragua, Niger, Nigeria, North Korea, North Macedonia, Norway, Oman, Pakistan, Panama, Papua New Guinea, Paraguay, Peru, Philippines, Poland, Portugal, Qatar, Romania, Russia, Rwanda, Saint Vincent and the Grenadines, Samoa, São Tomé and Príncipe, Saudi Arabia, Senegal, Serbia, Sierra Leone, Singapore, Slovakia, Slovenia, Solomon Islands, Somalia, South Africa, South Korea, South Sudan, Spain, Sri Lanka, Sudan, Suriname, Sweden, Switzerland, Syria, Tajikistan, Tanzania, Thailand, Timor-Leste, Togo, Tonga, Trinidad and Tobago, Tunisia, Turkey, Turkmenistan, Uganda, Ukraine, United Arab Emirates, United Kingdom, United States, Uruguay, Uzbekistan, Vanuatu, Venezuela, Vietnam, Yemen, Zambia, Zimbabwe.

B Additional Descriptive Information

To give readers a clearer grasp of the variation in the data, Figure B.1 presents the GDP of two former Soviet republics, Armenia and Estonia (both of which gained independence in 1991), and one country that gained independence in 2002, Timor-Leste. According to the April 1998 WDI, Armenia had a GDP of 12.4 billion in 1992 — a number over four times as large as what any other WDI release reports. According to the April 1997 WDI, Estonia had a GDP of 60.8 billion in 1995 — a number at least 13 times as large as what other releases report. And all vintages report a GDP between 1.1 and 1.2 billion for Timor-Leste in 2012, with the exception of 16 vintages between June 2018 and April 2020 that report a number six times larger. Indeed, these 16 vintages appear to misreport Timor-Leste's GDP from 2004 until 2012. These unique and extreme values are either the product of human error or an honest guesstimate corrected in subsequent vintages.

Figure B.1: Current GDP of Armenia, Estonia, and Timor-Leste, 1990–2020

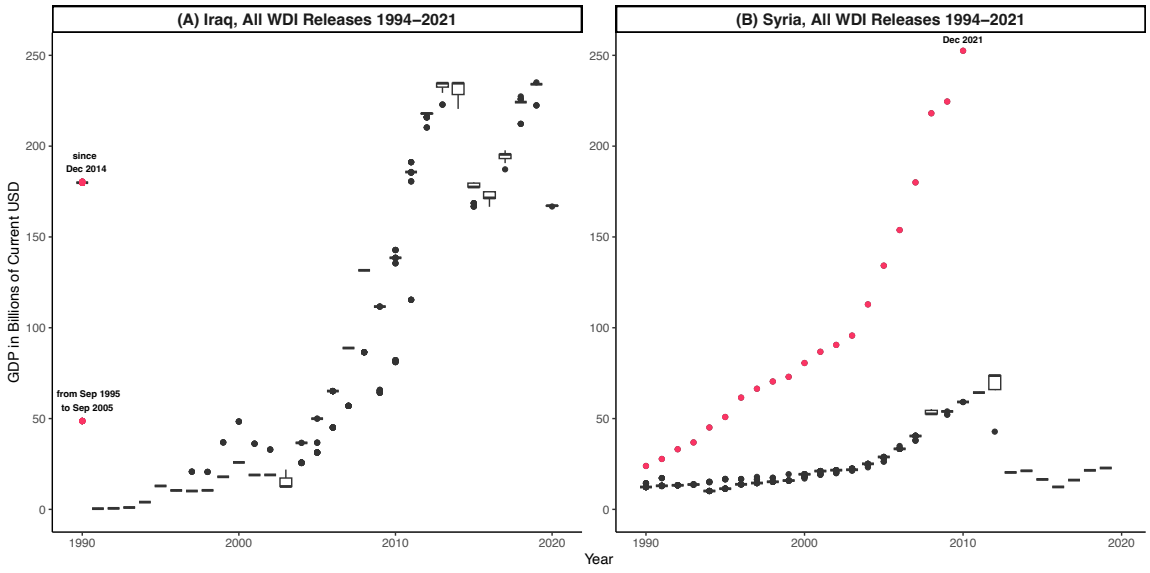


These boxplots present the distribution of current GDP estimates for (A) Armenia, (B) Estonia, and (C) Timor-Leste, from 1990 to 2020, using data drawn from the 104 WDI releases from April 1994 to December 2021. Section 3 discusses the data in more detail.

Figure B.2 presents the GDP of two war-plagued countries in the Middle East, Iraq and Syria. Iraq in 1990 is an interesting case: this country-year pair first enters the WDI in September 1995 and takes the value of 48.66 billion until September 2005, at which point it ceases to be included. It reappears in the December 2014 WDI, at which point it is reported to be nearly four times as large: 179.91 billion. Syria's GDP from 1990 to 2010 is considerably larger in the December 2021 WDI than in other vintages. Iraq's and Timor-Leste's outliers appear in multiple vintages, as do Syria's (which remained unchanged in 2022 vintages, though these are not included in the analysis). As of 2024, these values have not been revised; they

are the most up-to-date values, suggesting that the WDI has not recognized them as erroneous. They are not listed by the World Bank in its Data Updates and Errata website, and the WDI team did not respond to my inquiries about these specific observations. Thus, I do not code them as an error.

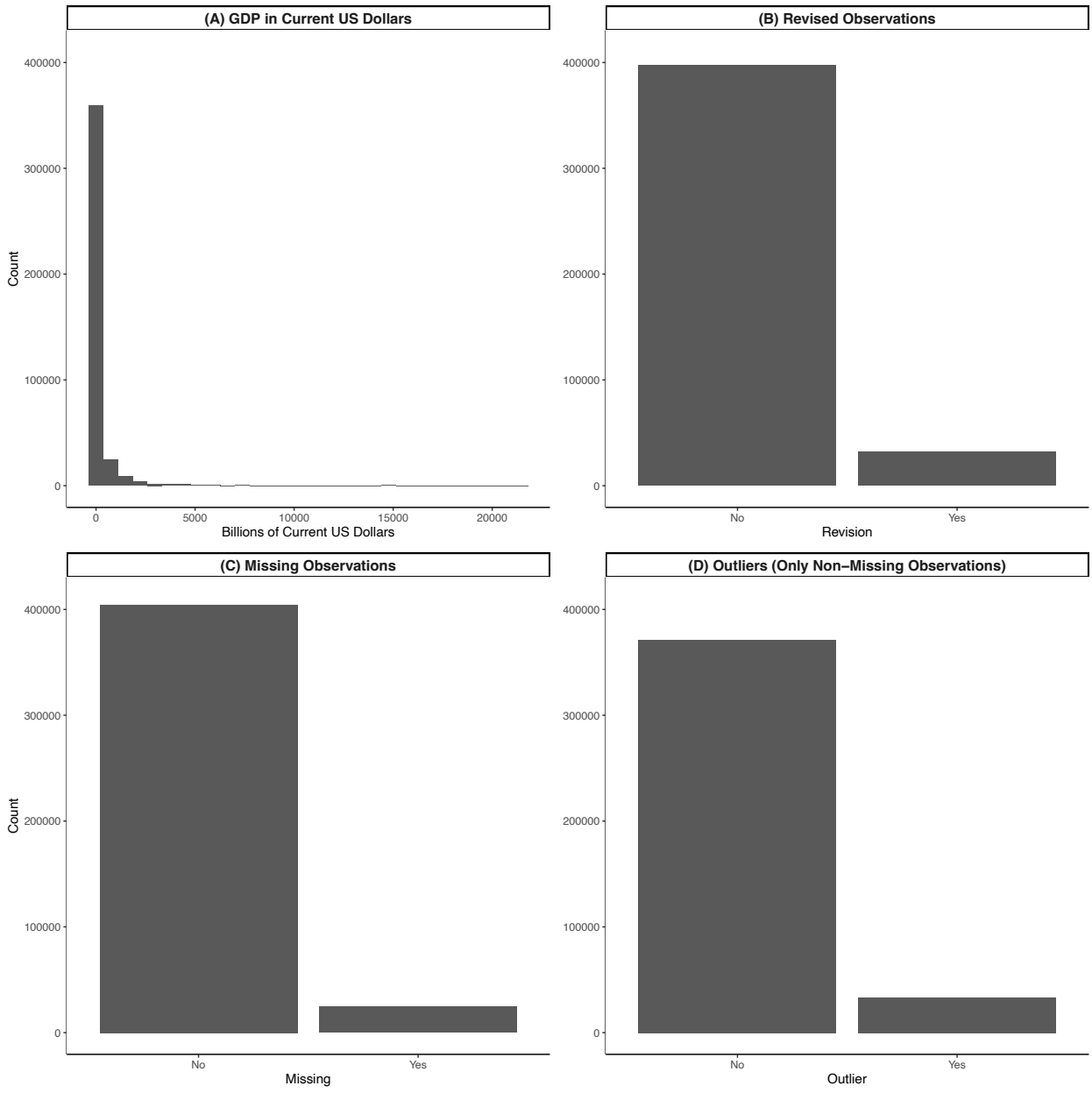
Figure B.2: Current GDP of Iraq and Syria, 1990–2020



These boxplots present the distribution of current GDP estimates for (A) Iraq and (B) Syria from 1990 to 2020, using data drawn from the 104 WDI releases from April 1994 to December 2021. Section 3 discusses the data in more detail.

In Figure B.3, panel (A) shows the distribution of *GDP in constant US dollars* (ID `NY.GDP.MKTP.CD`). This WDI variable, rounded to two decimal places, is used to generate the outcomes in panels (B), (C), and (D).

Figure B.3: Distribution of the Outcome Variables



Panel (A) shows the distribution of the WDI variable *GDP in constant US dollars* (ID `NY.GDP.MKTP.CD`. According to (B), 7.42 percent of all observations record a change from x_{itk} to x_{itk+1}. According to (C), 5.77 percent of all observations are missing. According to (D), 8.20 percent of all non-missing observations are outliers, as defined by the Tukey rule.

C List of Predictors

The main analysis includes all predictors listed in Tables C.1 and C.2. Those in Table C.1 are only included once, either because they are time-invariant (like *Former European Colony* and *Island*) or because they refer exclusively to the reporting year k (like *Alternative Conversion Factor* and *Coding Error*). Meanwhile, those in Table C.2 are included twice, for both the reported year t and the reporting year k.

In robustness checks (see Appendix E.1), I include additional economic and demographic predictors (listed in Table C.3) that are highly correlated with the outcome of interest, and thus likely suffer from the same measurement errors. For each source, I used the most recent release as of 1 June 2024. I downloaded all WDI data using Vincent Arel-Bundock's `WDI` package for `R`.

Table C.1: Predictors Included Once

Variable	Description	Coverage	Source
Alternative Conversion Factor	Does the WDI team use an alternative conversion factor for this country-year pair? Yes = 1	1994–2021	WDI Metadata
Autonomous Regions	Are there autonomous regions? Yes = 1	1990–2020	Cruz, Keefer and Scartascini (2021)
Coding Error	Coded 1 for the following observations: Armenia, 1992, April 1998 WDI; China, 2007 and 2008, September 2009 WDI; Democratic Republic of the Congo, 1990, July to November 2016 WDI; Estonia, 1995, April 1997 WDI; Georgia, 1990, April 1999 and April 2000 WDI; Myanmar, all years, December 2021 WDI; Timor-Leste, 2004 to 2012, June 2018 to April 2020 WDI	1990–2021	Own Coding
First Estimate	Is this the first WDI release to include this country-year pair? Only included in models with the outcome *Outlier* or with continuous outcomes in Appendix E	1990–2021	Own Coding
Former European Colony	Is this country a former colony of Belgium, France, Germany, Great Britain, Italy, Netherlands, Portugal, or Spain? Yes = 1	1990–2021	Becker (2019)
Island	Is the country an island? Yes = 1	1990–2021	Own coding

OECD Membership	Was this country a member of the Organization for Economic Co-Operation and Development at the time of reporting? Yes = 1	1990–2021	Dreher et al. (2022)
Post-Soviet State	Former Republic of the Union of Soviet Socialist Republics	1990–2021	Own coding
Tax Haven	Does the US Department of Treasury consider this country a tax haven? Yes = 1	1990–2021	Graham et al. (2018); Graham and Tucker (2019)

Table C.2: Predictors Included Twice, for Reported Year t and Reporting Year k

Variable	Description	Coverage	Source
Armed Conflict	Was any armed conflict recorded? Yes = 1	1990–2021	Gleditsch et al. (2002); Pettersson et al. (2021)
Bureaucratic Quality	To what extent does the country's bureaucracy have the strength and expertise to govern without drastic changes in policy or interruptions in government services? Low = 0, High = 4	1990–2021	The PRS Group (2022)
Bureaucratic Remuneration	To what extent are state administrators salaried employees? None = 0, Small Share = 1, Half = 2, Substantial Number = 3, All = 4	1990–2021	Coppedge et al. (2023)
Census in Previous 10 Years	Was there a national census in the previous 10 years? Yes = 1	1990–2021	Coppedge et al. (2023); Dang et al. (2023)
Civil War	Was there a civil war? Yes = 1	1990–2018	Marshall (2019)
Coup	Did a coup d'etat occur? Yes = 1	1990–2021	Coppedge et al. (2023)
COVID-19	Coded 1 for 2020 and after, the years of the COVID-19 pandemic	1990–2021	Own Coding
Disaster	Was there a biological (epidemic), climatological (drought, wildfire), meteorological (storm, extreme temperature), hydrological (flood, landslide), or geophysical (earthquake, volcanic activity) disaster? Yes = 1	1990–2021	Centre for Research on the Epidemiology of Disasters (2020)
Ethnic Tensions	Degree of tension within a country attributable to racial, nationality, or language divisions. High Tension = 0, Low Tension = 0	1990–2021	The PRS Group (2022)

Executive Tenure So Far	Number of years a leader has been in power during their current tenure	1990–2020	Bell, Besaw and Frank (2021)
Executive Was Elected	Was the executive leader elected to office? Yes = 1	1990–2020	Bell, Besaw and Frank (2021)
Financial Crisis	Was there a banking, currency, or debt crisis? Yes = 1	1990–2019	Nguyen, Castro and Wood (2022)
Freedom of Academic Expression	Is there academic freedom and freedom of cultural expression related to political issues? Yes = 1	1990–2021	Coppedge et al. (2023)
GDDS Participation	Does the country participate in the IMF's (Enhanced) General Data Dissemination System (GDDS or e-GDDS)? Yes = 1	1996–2021	IMF Dissemination Standards Bulletin Board
IMF Program	Was there an IMF program? Yes = 1	1990–2021	Kentikelenis, Stubbs and King (2016), IMF MONA Database
Investment Profile	Measure consisting of three components: contract viability/expropriation, profits repatriation, and payment delays. Each component ranges from very low risk = 4 to very high risk = 0	1990–2021	The PRS Group (2022)
Law and Order	Measure consisting of two components: law (strength and impartiality of the legal system) and order (popular observance of the law). Each component ranges from poor = 0 to high = 3	1990–2021	The PRS Group (2022)
Leader Education	Leader's level of education summarized in eight categories	1990–2020	Dreher et al. (2020)
Left Executive	Party orientation of the executive with respect to economic policy. Left = 1	1990–2020	Cruz, Keefer and Scartascini (2021)
Military	Direct or indirect military regime. Yes = 1	1990–2020	Bell, Besaw and Frank (2021)
Monarchy	Monarchy. Yes = 1	1990–2020	Bell, Besaw and Frank (2021)
Number of Protests	Number of recorded protests	1990–2020	Clark and Regan (2020)
Oil Discovery	Did this country discover a giant, megagiant, or supergiant oil or gas field? Yes = 1	1990–2020	Horn (2014); Cust, Mihalyi and Rivera-Ballesteros (2021)

Parliamentary Election Year	Did a legislative or constituent assembly election take place? Yes = 1	1990–2021	For Brunei and Belize, Cruz, Keefer and Scartascini (2021); for all other countries, Coppedge et al. (2023)
Political Corruption Index	On an interval scale, how pervasive is political corruption? Low = 0, High = 1	1990–2021	Coppedge et al. (2023)
Polity	Revised combined Polity score, from –10 (hereditary monarchy) to +10 (consolidated democracy)	1990–2018	Marshall and Gurr (2020)
Polyarchy	Electoral democracy index	1990–2021	Coppedge et al. (2023)
Presidential Democracy	Presidential democracy. Yes = 1	1990–2020	Bell, Besaw and Frank (2021)
Presidential Election Year	Did a presidential election take place? Yes = 1	1990–2021	For Brunei and Belize, Cruz, Keefer and Scartascini (2021); for all other countries, Coppedge et al. (2023)
Right to Information Law	Does the country have a Freedom of Information law (also known as a Right to Information law)? Yes = 1	1990–2021	Global Right to Information Rating
SDDS Compliance	Does the state comply with the IMF's Special Data Dissemination Standard (SDDS) specifications for the coverage, periodicity, and timeliness of data dissemination? Yes = 1	1996–2021	IMF Dissemination Standards Bulletin Board
SNA in Use	What System of National Accounts (SNA) is in use?	1998–2021	UN National Accounts Statistics, complemented by WDI Metadata and IMF International Financial Statistics

SNA Update	Was the SNA in use updated this year?	1998–2021	UN National Accounts Statistics, complemented by WDI Metadata and IMF International Financial Statistics
State Capacity	Estimate of state capacity by Hanson/Sigman	1990–2015	Hanson and Sigman (2021)
Statistical Agency	Is there a national statistical agency? Yes = 1	1990–2022	Coppedge et al. (2023); UN Statistics Division
World Bank Statistical Project	Is there a World Bank project in place that relates to data, surveys, censuses, or overall statistical capacity development?	1990–2021	World Bank Projects Portal
Years Since Independence	How many years have passed since this country's most recent foundation, independence, or reunification?	1990–2021	Own coding

Table C.3: Predictors Included Twice, for Reported Year t and Reporting Year k, in Robustness Checks

Variable	Description	Coverage	Source
Agriculture	GDP, share of value added by kind of economic activity: agriculture, hunting, forestry, fishing	1990–2021	UNCTAD
Central Government Debt	Central government debt, share of GDP	1990–2020	IMF
Diversification Index	Merchandise: product diversification index of exports	1995–2021	UNCTAD
Fertility Rate	Fertility rate, total (births per woman). WDI ID: `SP.DYN.TFRT.IN`	1990–2021	WDI
Foreign Aid	Net official development assistance and official aid received (current US dollars). WDI ID: `DT.ODA.ODAT.CD`	1990–2021	WDI
Imports	Imports of goods and services, share of GDP	1990–2021	UNCTAD
Income Share Top 10%	Share of pre-tax national income held by the top 10%	1990–2021	World Inequality Database
Industry	GDP, share of value added by kind of economic activity: industry	1990–2021	UNCTAD
Inflation	Inflation, consumer prices (annual %). WDI ID: `FP.CPI.TOTL.ZG`	1990–2021	WDI

Inward FDI, Flows	Inward foreign direct investment flows, share of GDP	1990–2021	UNCTAD
Inward FDI, Stock	Inward foreign direct investment stock, share of GDP	1990–2021	UNCTAD
KAOPEN	Normalized Chinn-Ito index, ranging from zero to one	1990–2020	Chinn and Ito (2006)
Military Expenditure	Military expenditure per capita, in current US dollars	1990–2021	SIPRI Military Expenditure Database
Service	GDP, share of value added by kind of economic activity: service	1990–2021	UNCTAD
Tax Revenue	Total tax revenue, excluding social security contributions, share of GDP	1990–2021	Government Revenue Dataset
Total Population	Population, total. WDI ID: `SP.POP.TOTL`	1990–2021	WDI
Unemployment	Unemployment (% of total labor force), modeled ILO estimate	1991–2021	WDI
Urban Population	Urban population (% of total population). WDI ID: `SP.URB.TOTL.IN.ZS`	1990–2021	WDI

Figure C.1: Missingness Map: Predictors, 1990–2021

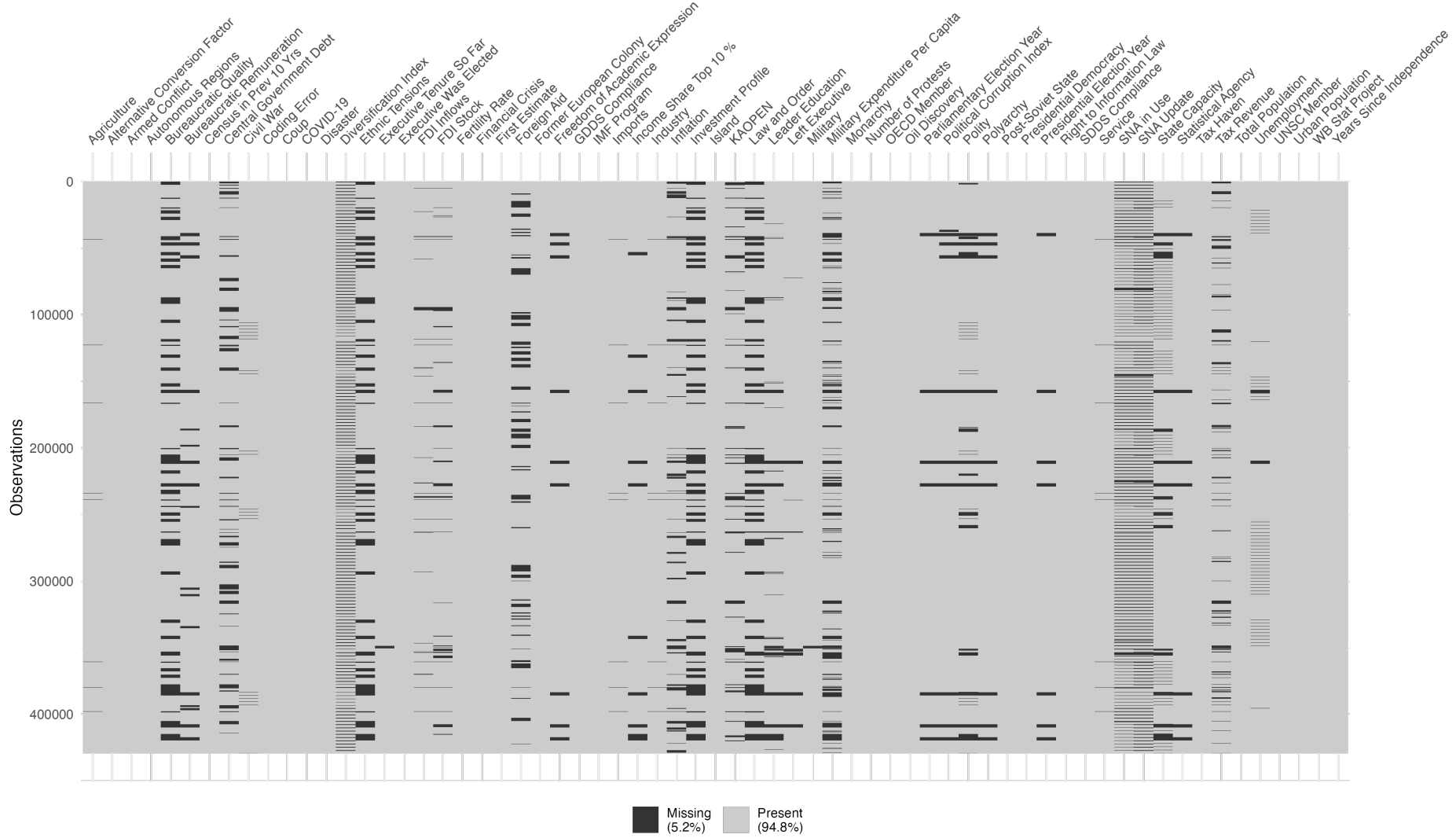
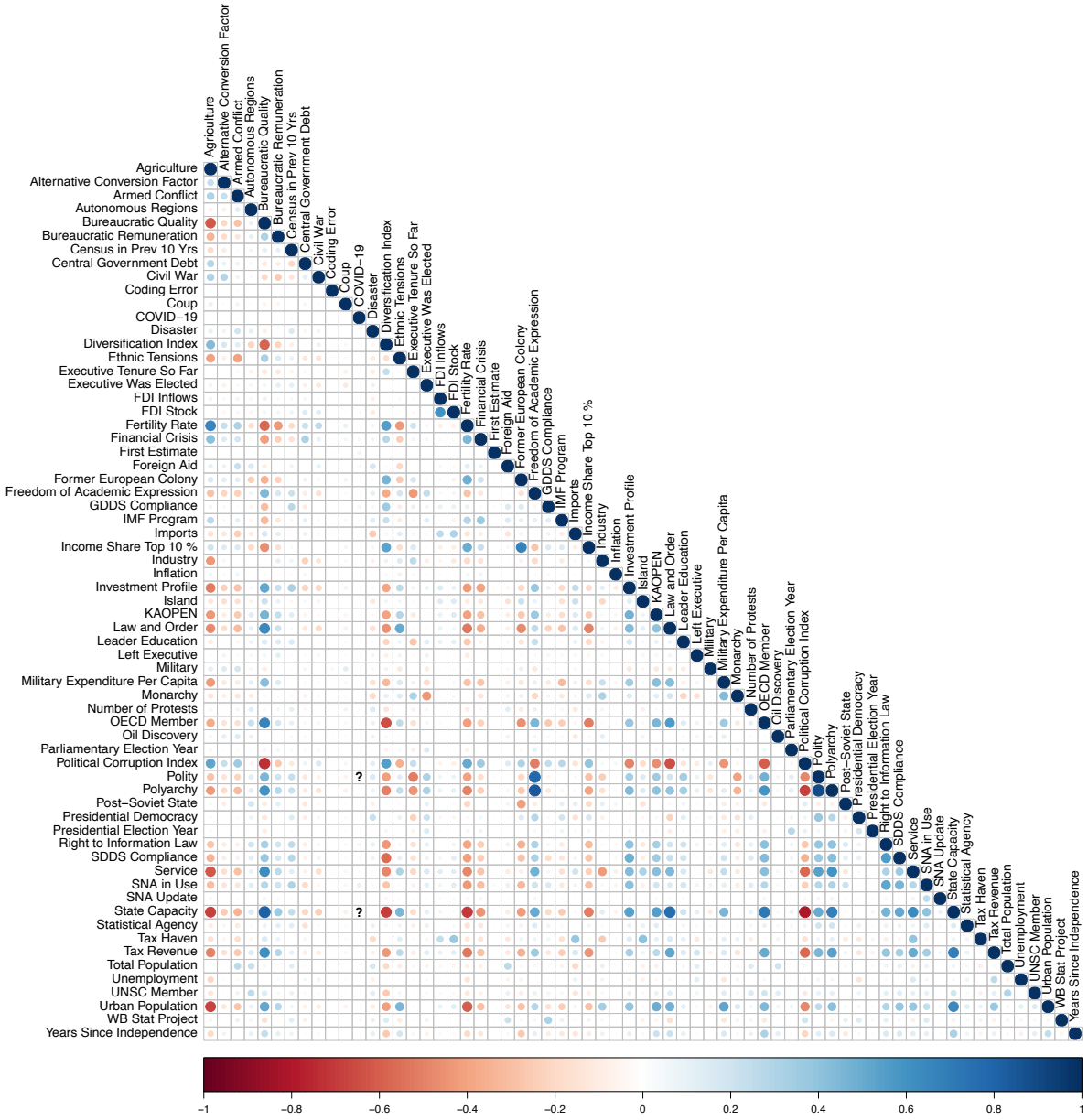


Figure C.2: Correlation Matrix: Predictors, 1990–2021

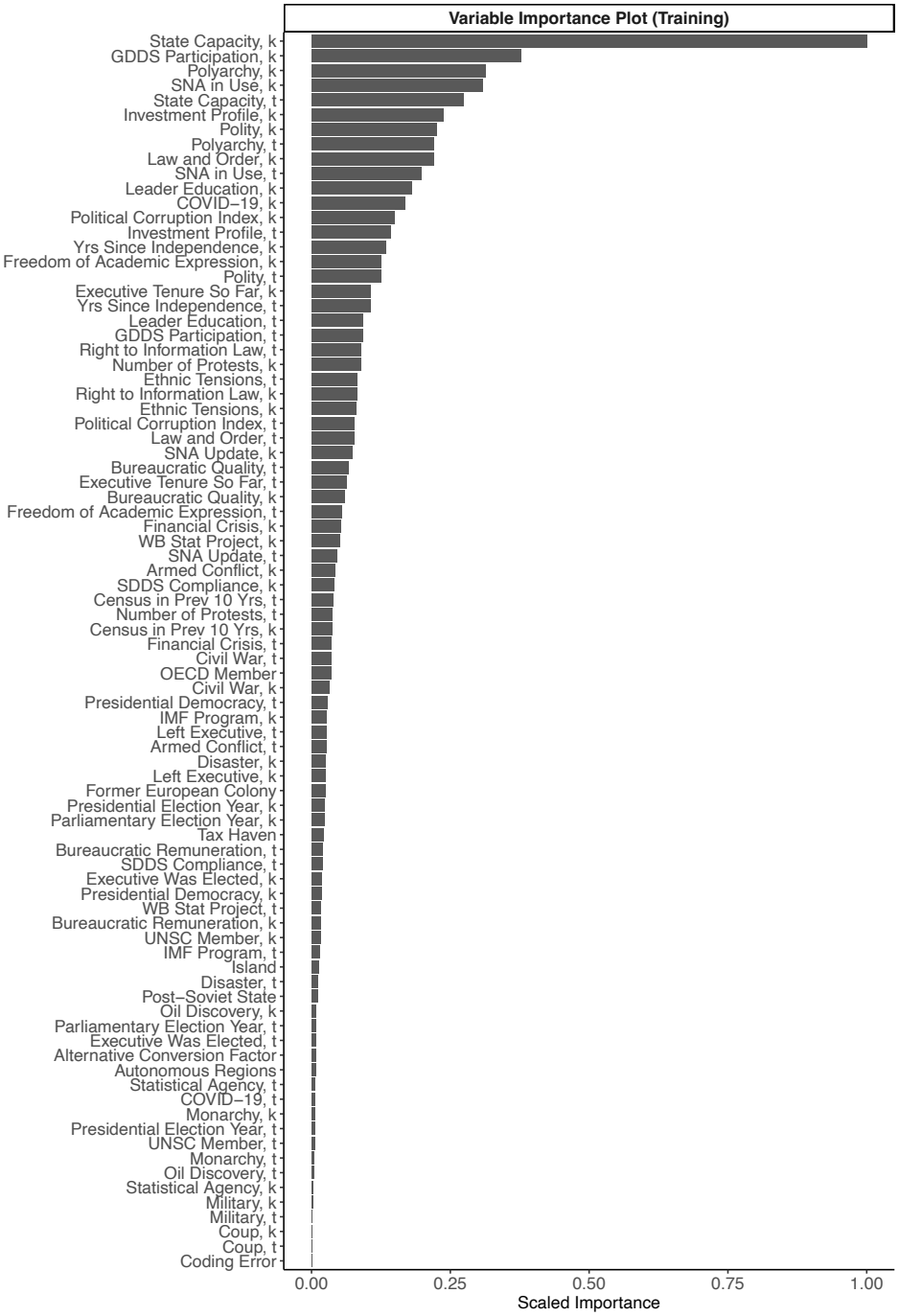


The pairwise correlation between *Polity* and *COVID-19* cannot be computed because the former ends its coverage in 2015, whereas the latter is a vector of zeroes until 2020. The same applies to the pairwise correlation between *State Capacity* and *COVID-19*.

D Performance Metrics and Alternative Models

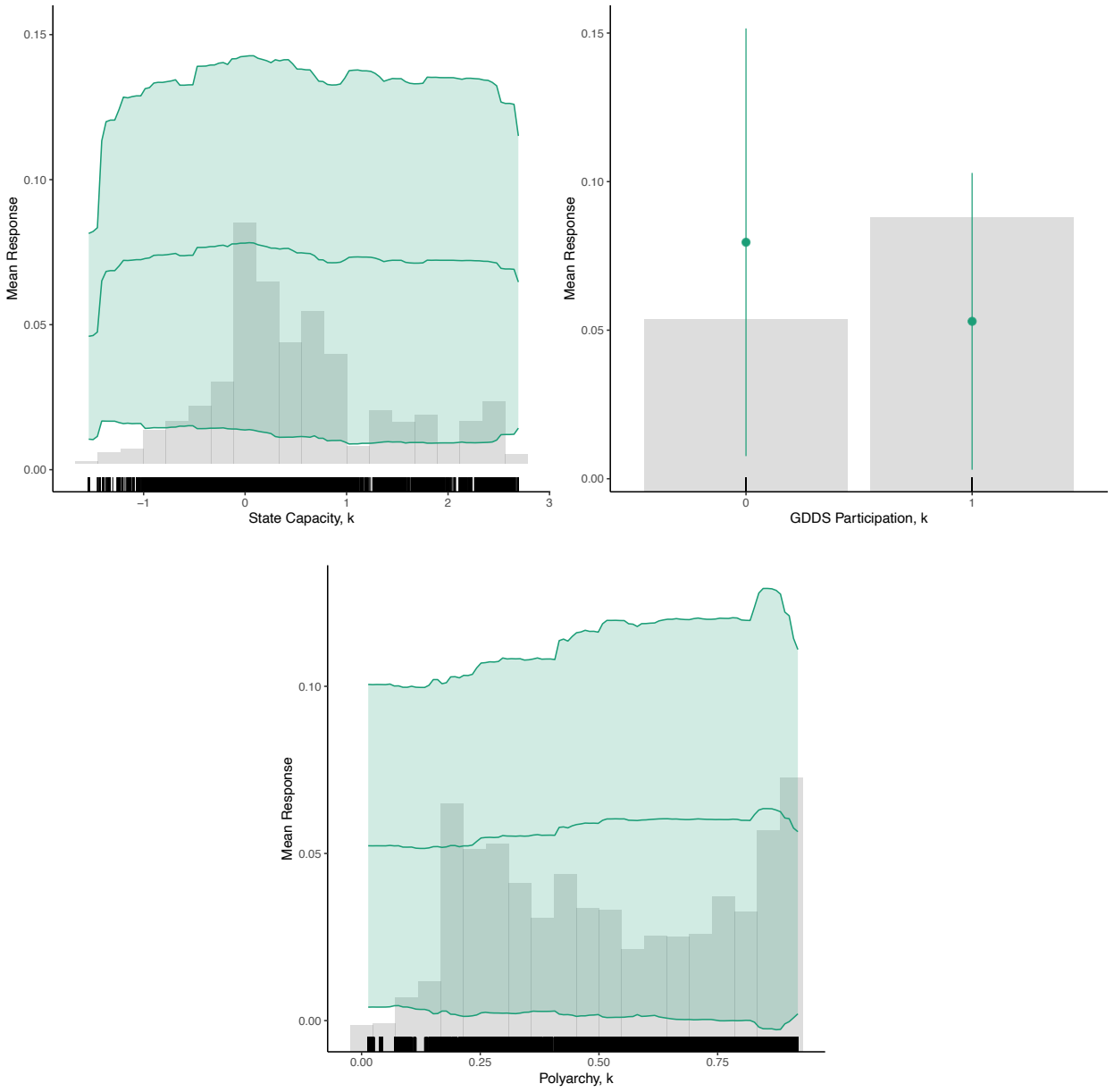
D.1 Revisions

Figure D.1: Variable Importance Plot for a Random Forest Predicting Data Revisions (Training Set)



This variable importance plot indicates the relative importance of all predictors; t or k denotes the predictor's value for the reported or reporting year, respectively.

Figure D.2: Partial Dependence Plots for a Random Forest Predicting Data Revisions (Training Set)



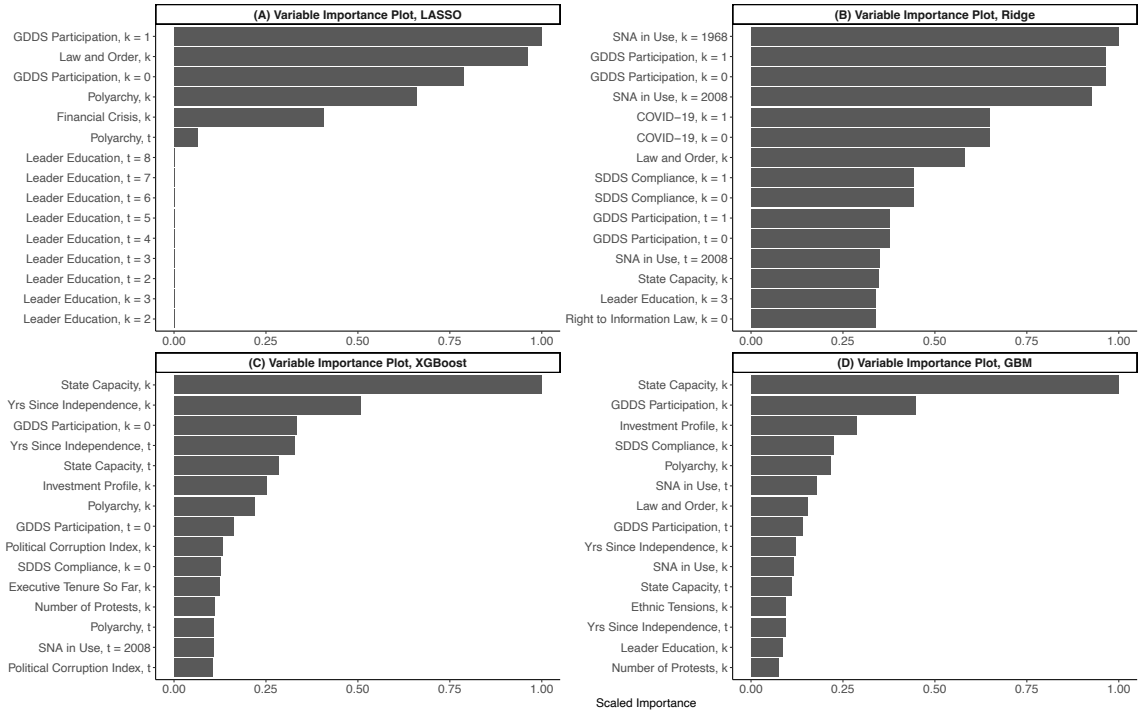
These figures assess the relative importance of the three most important variables (*State Capacity*, *GDDS Participation*, and *Polyarchy*, all at time k) in predicting *Revision*. These are equivalent to marginal effects plots, isolating each variable's impact while holding others constant.

Figure D.1 presents the full variable importance plot for a random forest with all predictors. Figure D.2 shows partial dependence plots for the three most important variables: *State Capacity*, *GDDS Participation*, and *Polyarchy*, all at time k.

To predict *Revision*, I estimate not only a random forest but also two penalized generalized linear models (GLM) — least absolute shrinkage and selection operator (LASSO) and ridge regression — and two tree-

based models – eXtreme Gradient Boosting (XGBoost) and gradient boosting machine (GBM). GLMs are based on linear relationships between the predictors and the response variable; they involve straightforward mathematical operations that are computationally efficient. A random forest is computationally more demanding but can be parallelized to some extent, because it is a bagging algorithm: it builds multiple trees independently. XGBoost and GBM are the deepest, most complex, and most computationally intensive algorithms of all five. This is because they are boosting algorithms: they builds trees sequentially, with each tree trying to correct the errors of the previous ones, aiming to reduce bias.

Figure D.3: Variable Importance Plot for Alternative Models Predicting Data Revisions (Training Set)



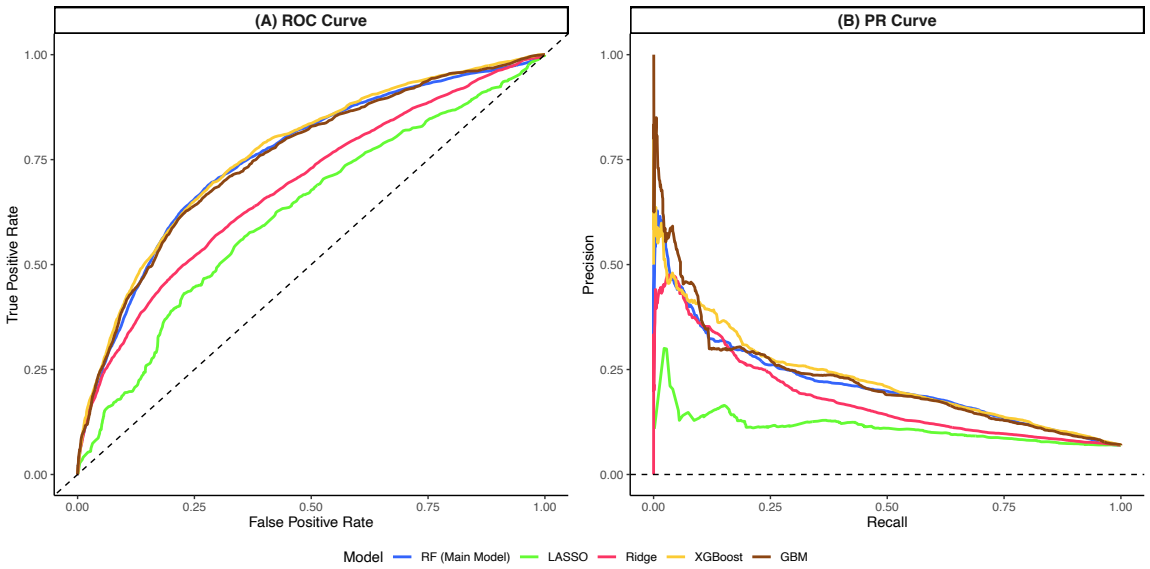
These panels show the relative importance of the top 15 variables, by model. The least important variable equals zero, while the most important variable equals one. LASSO adds a penalty to the absolute values of the coefficients (L1 regularization) that encourages most coefficients — like that of all other variables not depicted here — to become exactly zero.

Figure D.3 presents the variable importance plots for each model. LASSO, ridge, and XGBoost do not drop the baseline category of a categorical variable, as a traditional regression would do to avoid multicollinearity; this is why the corresponding panels include specific levels of categorical variables. In addition, LASSO adds a penalty to the absolute values of the coefficients (L1 regularization) that encourages most coefficients to become exactly zero, effectively performing feature selection by eliminating irrelevant variables. Only the first few variables displayed in panel (A) have any importance; the remaining ones have zero importance. In contrast, ridge regression adds a penalty to the squared values of the coefficients (L2 regularization) that discourages large coefficients but does not force any coefficients to become exactly zero.

Table D.1: Performance Statistics for Alternative Models Predicting Data Revisions

	Training	Validation	Test
Random Forest (Main Model)			
AUC	0.8830324	0.7519537	0.7546658
AUCPR	0.8655482	0.1988233	0.2160246
LASSO			
AUC	0.6490398	0.6134968	0.6217638
AUCPR	0.1255696	0.1031015	0.1122228
Ridge			
AUC	0.7527355	0.6948954	0.6840914
AUCPR	0.2256603	0.1579986	0.1806999
XGBoost			
AUC	0.8843092	0.7610988	0.7635452
AUCPR	0.4425107	0.2203509	0.2272135
GBM			
AUC	0.8475273	0.7495276	0.7518939
AUCPR	0.3845288	0.2072344	0.2237693

Figure D.4: ROC and PR Curves for Alternative Models Predicting Data Revisions (Test Set)



Panel (A) presents a Receiver Operating Characteristic (ROC) curve for the the test set using five models. Panel (B) presents a Precision-Recall (PR) curve, also for the test set and also using the same five models.

Table D.1 presents common performance metrics for each model. In general, these models make good out-of-sample predictions, as illustrated by the high Area Under the ROC Curve (AUC). This statistic ranges from 0 to 1, with 0.5 denoting random guessing and 1 denoting a perfect classifier. High AUC values indicate good discrimination ability between the positive and negative classes: the models are effective at ranking instances in terms of their likelihood of belonging to the positive class. Since the outcome is very imbalanced, I balance the majority and minority classes in the training set, but not the other sets. This is partly why all models perform best on the data they were trained on. To mitigate overfitting, all models use early stopping

and iterative tuning of hyperparameters. In Figure D.4, panel (A) plots the area under the ROC curve.

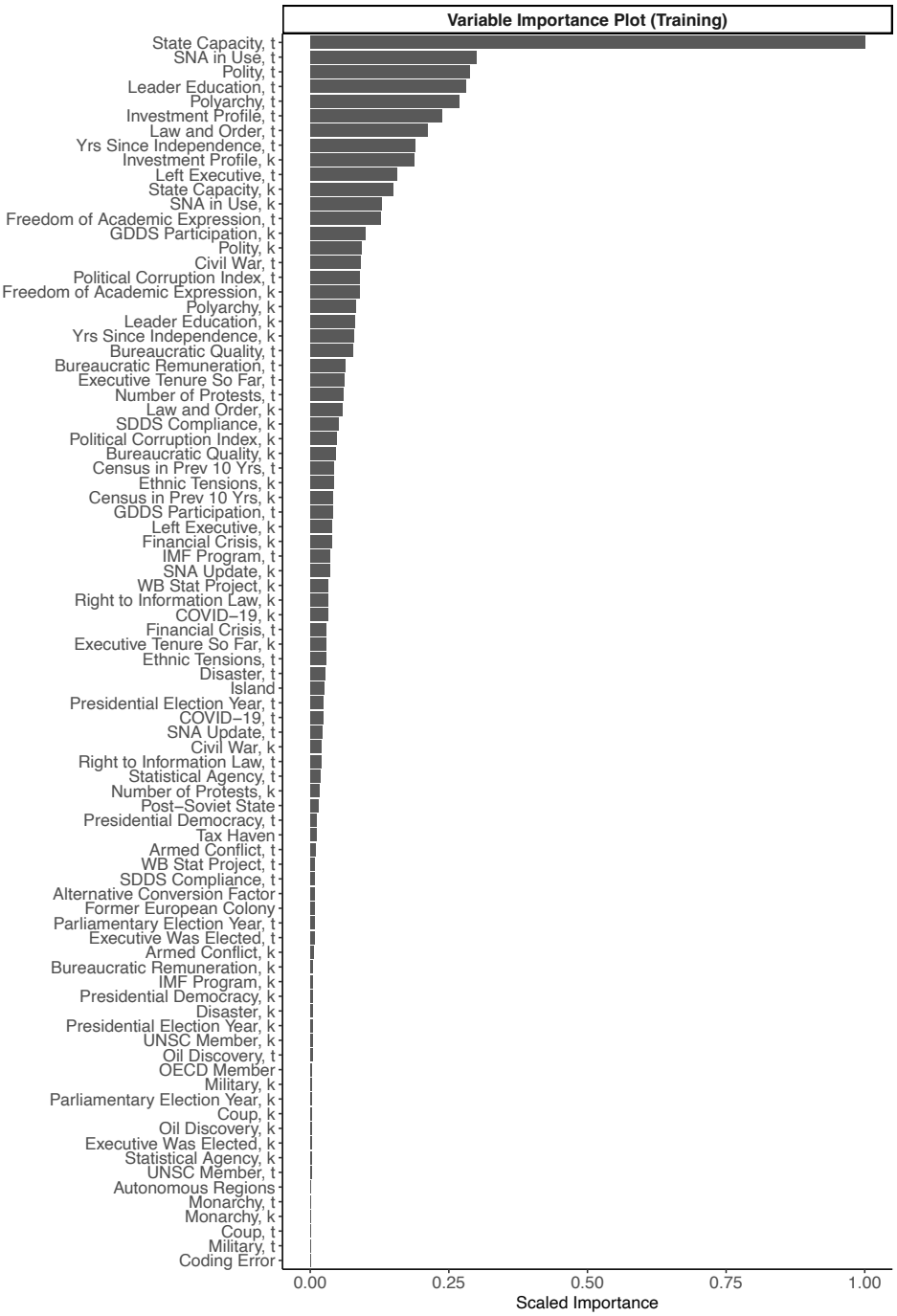
Another important performance metric for classification tasks is the Area Under the Precision-Recall (PR) Curve (AUCPR). This metric indicates the trade-off between precision — the missing observations (true positives) the model correctly identified from all the observations it labeled as missing (true positives plus false positives) — and recall — the missing observations (true positives) the model correctly identified from all the actual missing cases (true positives the false negatives). Like AUC values, AUCPR values range from 0 to 1, with 0.5 denoting random guessing and 1 denoting a perfect classifier. In Table D.1, the AUCPR for the test set is consistently below 0.5, which might appear modest, but it is crucial to contextualize this result within the unique challenges posed by the data. In cases of extreme class imbalance, achieving an AUCPR close to 1 is unrealistic, given how difficult it is to simultaneously optimize precision and recall. As the proportion of positive instances diminishes, the denominator in the precision calculation becomes small, amplifying the impact of false positives on the metric. Accordingly, the observed AUCPR values underscore the model's ability to discern positive instances amid a predominantly negative class distribution. In such imbalanced settings, where the random chance may hover around the proportion of positive instances, a model exhibiting substantial discrimination capability is promising. To illustrate this, panel (B) of Figure D.4 plots the AUCPR values; a random model would produce a horizontal line, whereas a perfect classifier would score 1 for both precision and recall, corresponding to the top-right corner of the plot. Though the models are not perfect, they perform considerably better than a random classifier.

Overall, Table D.1 and Figure D.4 show that a GBM and an XGBoost make better out-of-sample predictions than a random forest, but not by much. Since these models are more computationally intensive and less straightforward to interpret, I reported the results of a random forest in the main text. The other two models perform considerably worse, confirming Muchlinski et al.'s (2016) conclusion that tree-based models outperform logistic regressions when predicting class-imbalanced data in political science. I also considered other options, but they all had important shortcomings: Naïve Bayes Classifiers rely on strong assumptions about the independence of predictors; Support Vector Machines are sensitive to outliers; deep learning models are computationally challenging and difficult to interpret. Tree-based models are not perfect, but they do a better job of capturing the idiosyncrasies of GDP data than alternative models.

Below, I present the corresponding figures and tables for the outcomes *Missing* and *Outlier*; the discussion above applies to both of them as well.

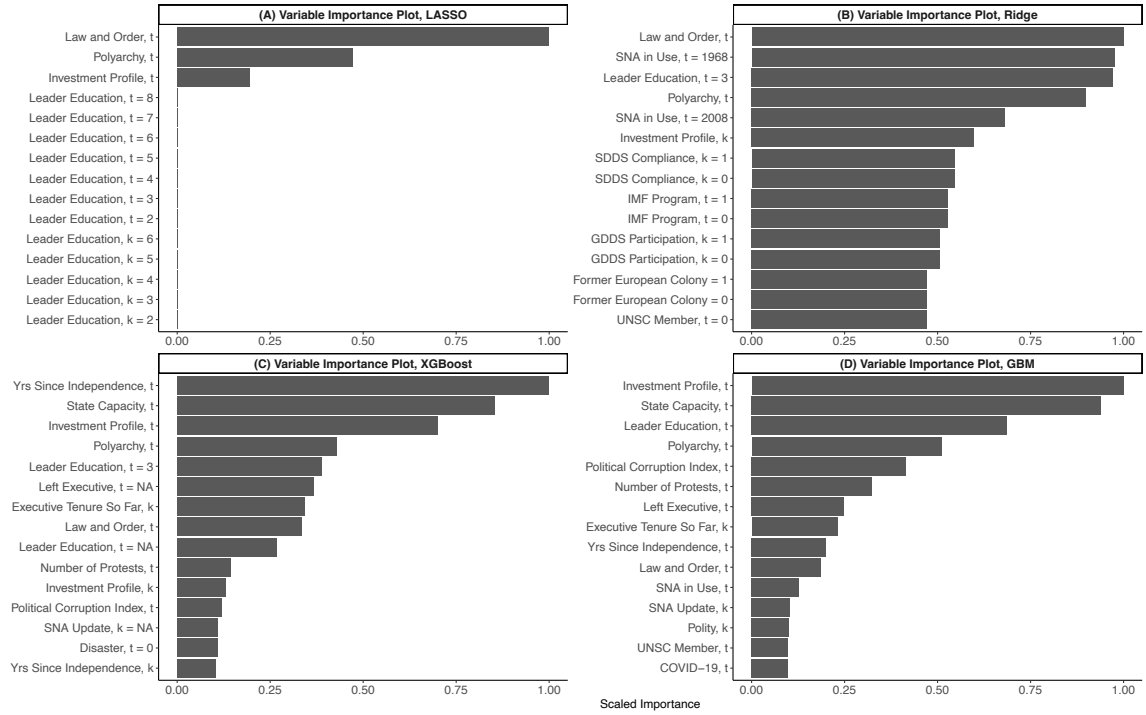
D.2 Missingness

Figure D.5: Variable Importance Plot for a Random Forest Predicting Missing Data (Training Set)



This variable importance plot indicates the relative importance of all predictors; t or k denotes the predictor's value for the reported or reporting year, respectively.

Figure D.6: Variable Importance Plot for Alternative Models Predicting Missing Data (Training Set)

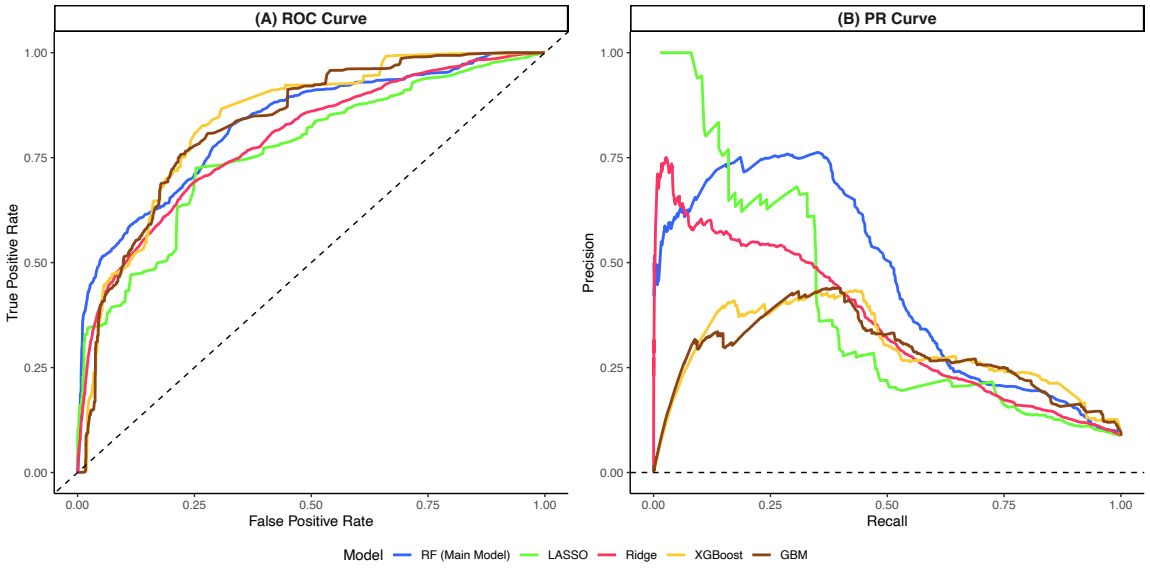


These panels show the relative importance of the top 15 variables, by model. The least important variable equals zero, while the most important variable equals one. LASSO adds a penalty to the absolute values of the coefficients (L1 regularization) that encourages most coefficients — like that of all other variables not depicted here — to become exactly zero.

Table D.2: Performance Statistics for Alternative Models Predicting Missing Data

	Training	Validation	Test
Random Forest (Main Model)			
AUC	0.9496553	0.8009512	0.8292276
AUCPR	0.9520517	0.3727305	0.453423
LASSO			
AUC	0.6368983	0.6076552	0.765502
AUCPR	0.16588557	0.12871057	0.3966254
Ridge			
AUC	0.8104001	0.7814685	0.7910807
AUCPR	0.38178204	0.22081129	0.354779
XGBoost			
AUC	0.9418934	0.7422108	0.8367934
AUCPR	0.7470930	0.2035123	0.290113
GBM			
AUC	0.9546643	0.7886464	0.8295033
AUCPR	0.7647020	0.2980293	0.2818602

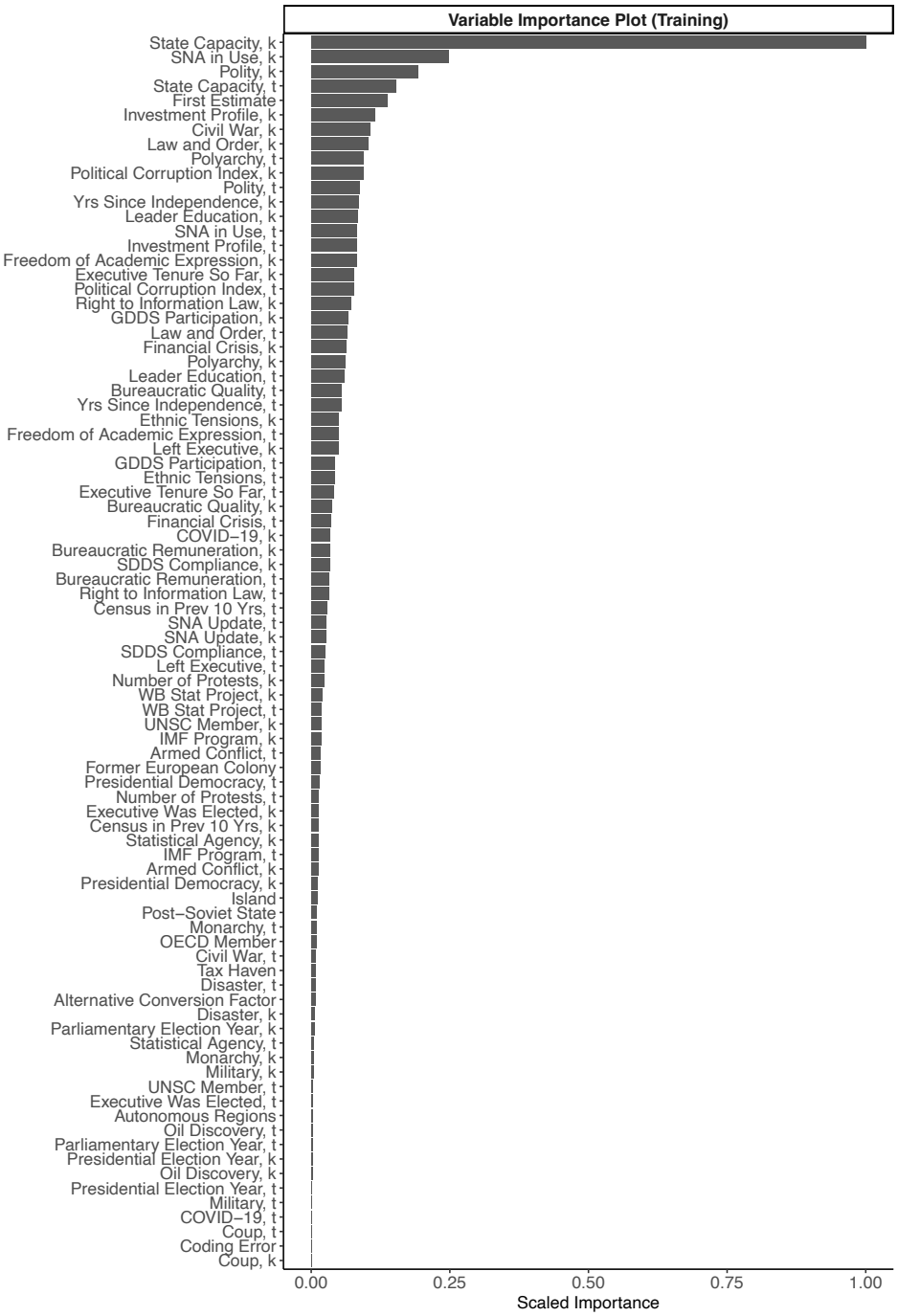
Figure D.7: ROC and PR Curves for Alternative Models Predicting Missing Data (Test Set)



Panel (A) presents a Receiver Operating Characteristic (ROC) curve for the the test set using five models. Panel (B) presents a Precision-Recall (PR) curve, also for the test set and also using the same five models.

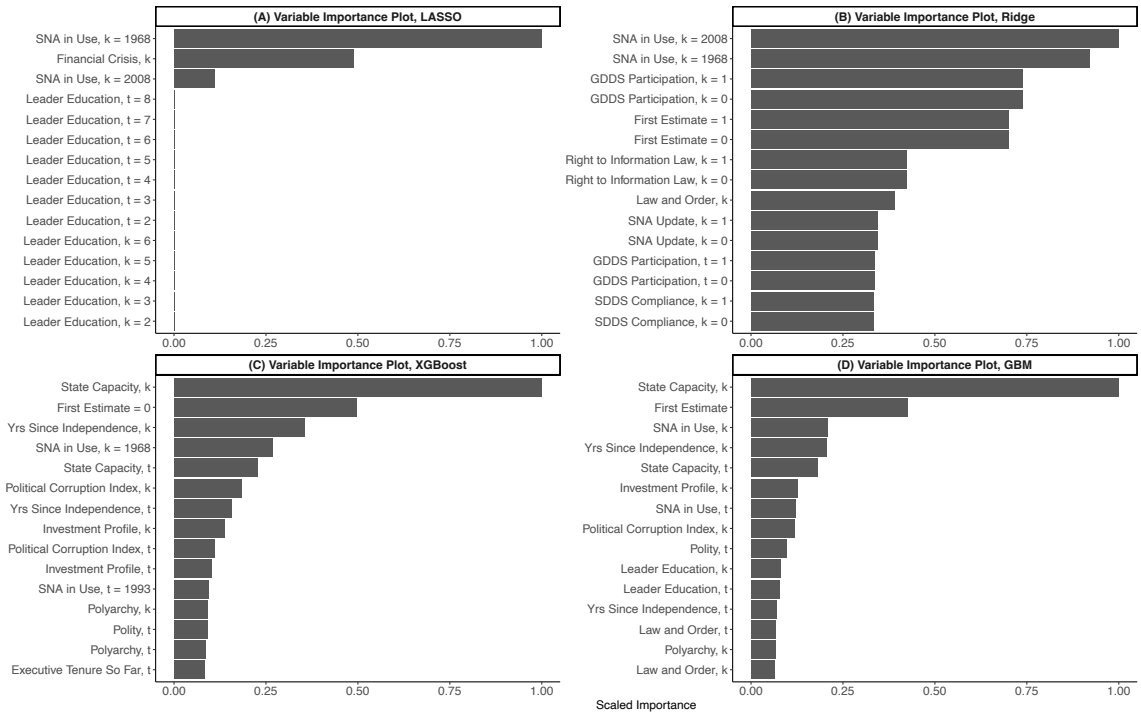
D.3 Outlier

Figure D.8: Variable Importance Plot for a Random Forest Predicting Outliers (Training Set)



This variable importance plot indicates the relative importance of all predictors; t or k denotes the predictor's value for the reported or reporting year, respectively.

Figure D.9: Variable Importance Plot for Alternative Models Predicting Outliers (Training Set)

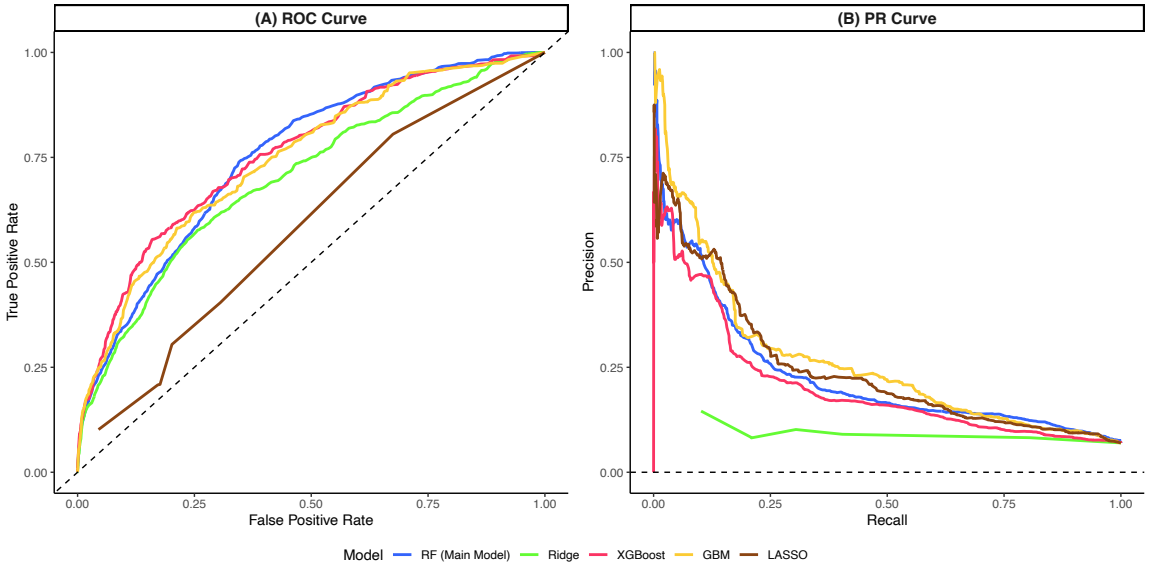


These panels show the relative importance of the top 15 variables, by model. The least important variable equals zero, while the most important variable equals one. LASSO adds a penalty to the absolute values of the coefficients (L1 regularization) that encourages most coefficients — like that of all other variables not depicted here — to become exactly zero.

Table D.3: Performance Statistics for Alternative Models Predicting Outliers

	Training	Validation	Test
Random Forest (Main Model)			
AUC	0.9228975	0.7744063	0.752992
AUCPR	0.9151335	0.2418411	0.2297873
LASSO			
AUC	0.6311433	0.6027568	0.5829745
AUCPR	0.1387587	0.1060097	0.09296241
Ridge			
AUC	0.7627921	0.6871322	0.7020276
AUCPR	0.2890987	0.2195110	0.2045901
XGBoost			
AUC	0.9030125	0.7500928	0.759284
AUCPR	0.6194107	0.2376396	0.2643427
GBM			
AUC	0.8955522	0.7686128	0.7464724
AUCPR	0.6132619	0.2546080	0.2431034

Figure D.10: ROC and PR Curves for Alternative Models Predicting Outliers (Test Set)



Panel (A) presents a Receiver Operating Characteristic (ROC) curve for the the test set using five models. Panel (B) presents a Precision-Recall (PR) curve, also for the test set and also using the same five models.

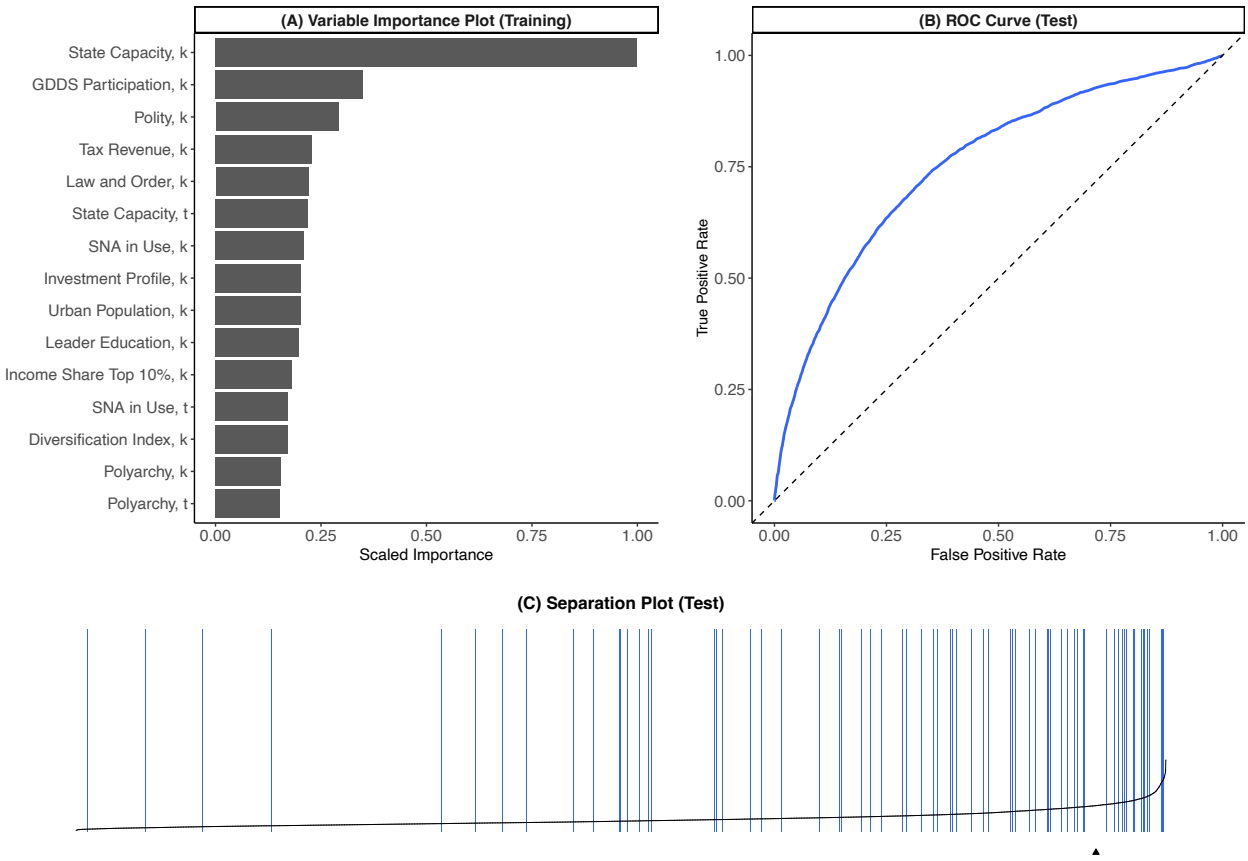
E Alternative Predictors and Outcomes

E.1 Alternative Predictors: WDI

Table C.2 lists 18 economic and demographic predictors that are not included in the main analysis. Below, I present the results of additional models including these 18 predictors (in addition to the original predictors). Including more predictors increases a model's complexity; when the model is too complex, it fits the training data too closely, capturing noise and outliers. This might result in a lower performance on new data, as the model fails to generalize effectively. Indeed, models including these 18 predictors tend to perform no better than the main models, and sometimes in fact slightly *worse*, as Tables E.1 to E.3 show.

As Figures E.1 to E.3 show, the main predictor of *Revision*, *Missing*, and *Outlier* in an expanded model continues to be *State Capacity*. Given that these models increase complexity and computational needs without improving performance, I opted to present the more parsimonious models in the main text.

Figure E.1: Assessing the Performance of a Model Predicting Data Revisions (With WDI Predictors)

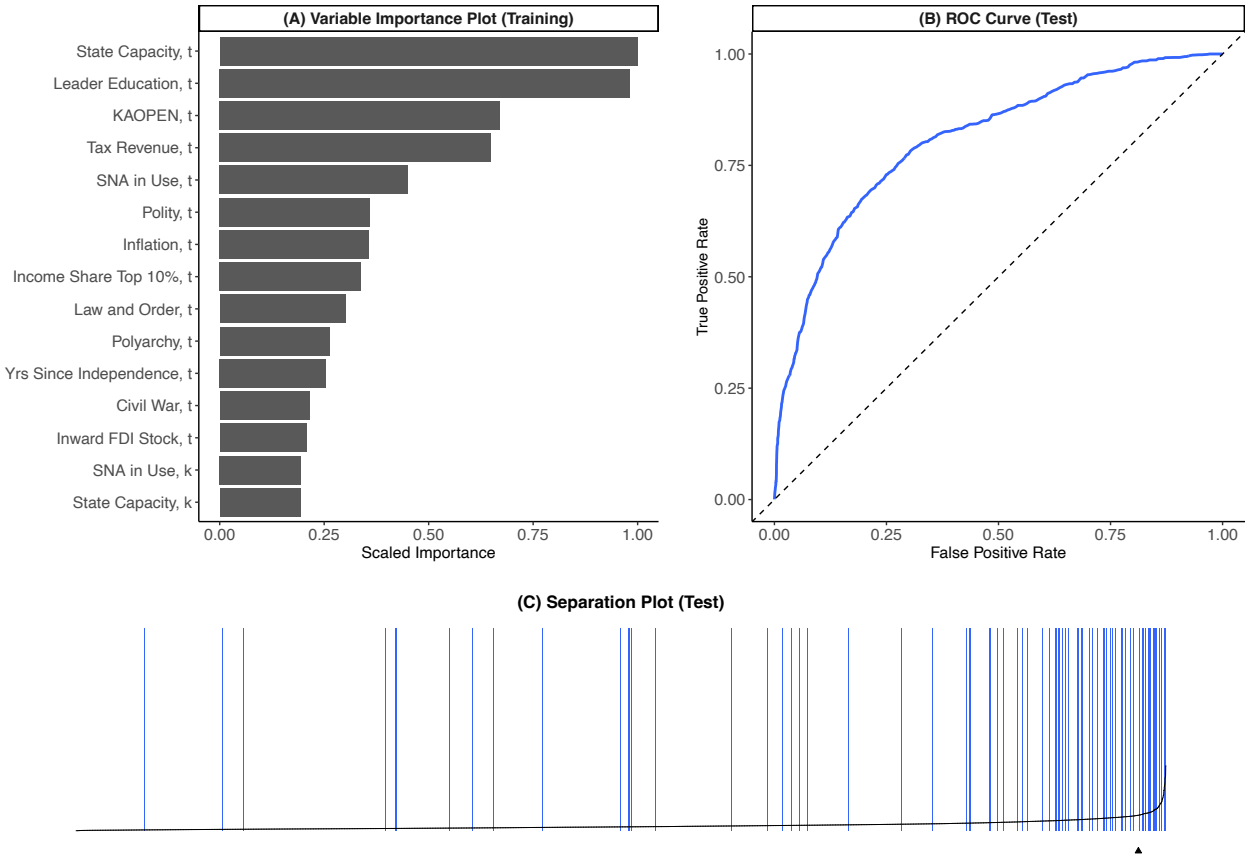


These figures assess the fit of a random forest predicting the outcome *Revision*. Panel (A) is a variable importance plot, indicating the relative importance of the 15 most important predictors; t or k denotes the predictor's value for the reported or reporting year, respectively. Panel (B) is a Receiver Operating Characteristic (ROC) curve, illustrating the trade-off between the true positive rate and the false positive rate across different probability thresholds. Panel (C) is a separation plot that organizes the predicted probabilities for each observation in ascending order, highlighting whether each observation corresponds to an instance of *Revision*.

Table E.1: Performance Statistics for Alternative Models Predicting Data Revisions (With WDI Predictors)

	Training	Validation	Test
Main Model			
AUC	0.9496553	0.8009512	0.8292276
AUCPR	0.9520517	0.3727305	0.453423
Model With WDI Variables			
AUC	0.8951852	0.7551928	0.7536769
AUCPR	0.8782806	0.2086043	0.2113169

Figure E.2: Assessing the Performance of a Model Predicting Missing Data (With WDI Predictors)

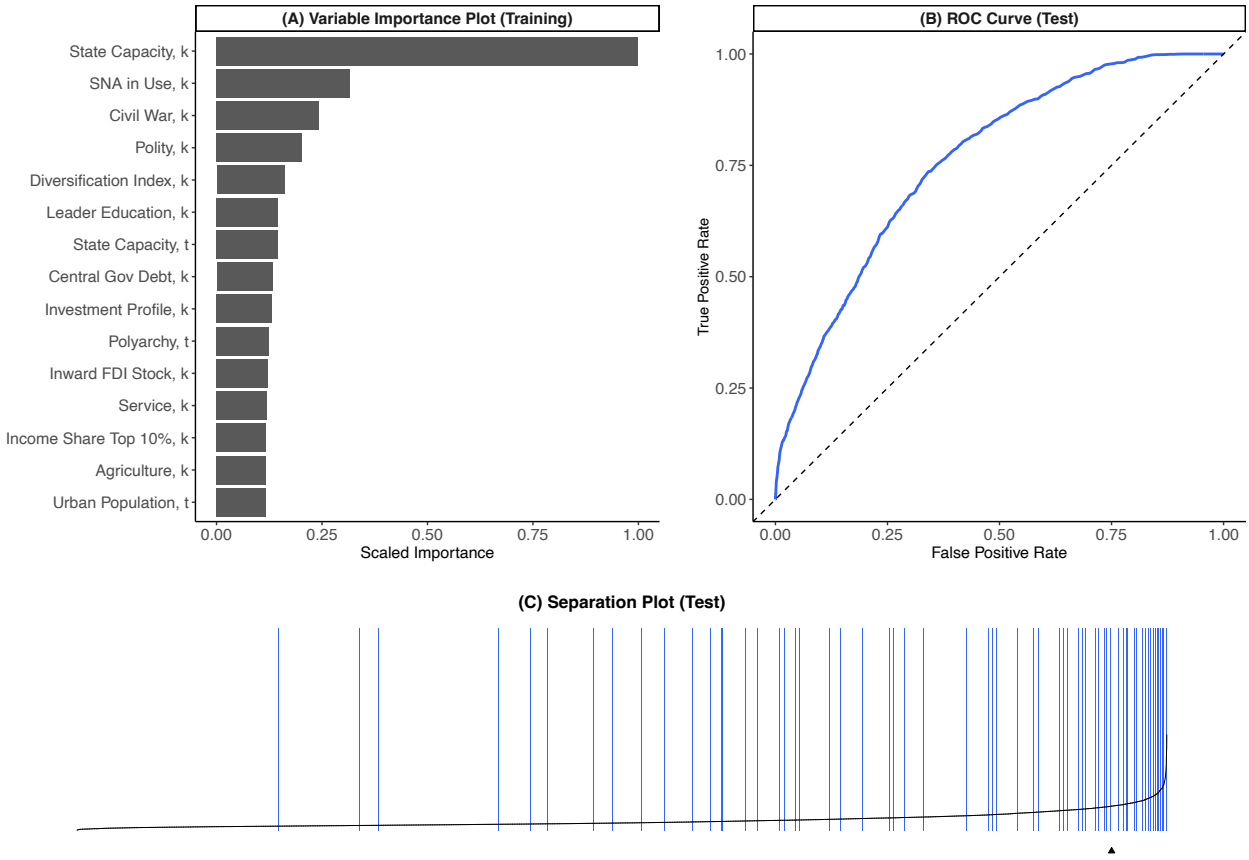


These figures assess the fit of a random forest predicting the outcome *Missing*. Panel (A) is a variable importance plot, indicating the relative importance of the 15 most important predictors; t or k denotes the predictor's value for the reported or reporting year, respectively. Panel (B) is a Receiver Operating Characteristic (ROC) curve, illustrating the trade-off between the true positive rate and the false positive rate across different probability thresholds. Panel (C) is a separation plot that organizes the predicted probabilities for each observation in ascending order, highlighting whether each observation corresponds to an instance of *Missing*.

Table E.2: Performance Statistics for Alternative Models Predicting Missing Data (With WDI Variables)

	Training	Validation	Test
Main Model			
AUC	0.9496553	0.8009512	0.8292276
AUCPR	0.9520517	0.3727305	0.453423
Model With WDI Variables			
AUC	0.9829094	0.8585561	0.8082296
AUCPR	0.9823812	0.3791382	0.3512808

Figure E.3: Assessing the Performance of a Model Predicting Outliers (With WDI Predictors)



These figures assess the fit of a random forest predicting the outcome *Outlier*. Panel (A) is a variable importance plot, indicating the relative importance of the 15 most important predictors; t or k denotes the predictor's value for the reported or reporting year, respectively. Panel (B) is a Receiver Operating Characteristic (ROC) curve, illustrating the trade-off between the true positive rate and the false positive rate across different probability thresholds. Panel (C) is a separation plot that organizes the predicted probabilities for each observation in ascending order, highlighting whether each observation corresponds to an instance of *Outlier*.

Table E.3: Performance Statistics for Alternative Models Predicting Outliers (With WDI Variables)

	Training	Validation	Test
Main Model			
AUC	0.9228975	0.7744063	0.752992
AUCPR	0.9151335	0.2418411	0.2297873
Model With WDI Variables			
AUC	0.9707273	0.7691793	0.7611753
AUCPR	0.9663670	0.2667292	0.214751

E.2 Alternative Outcome: Z-Score

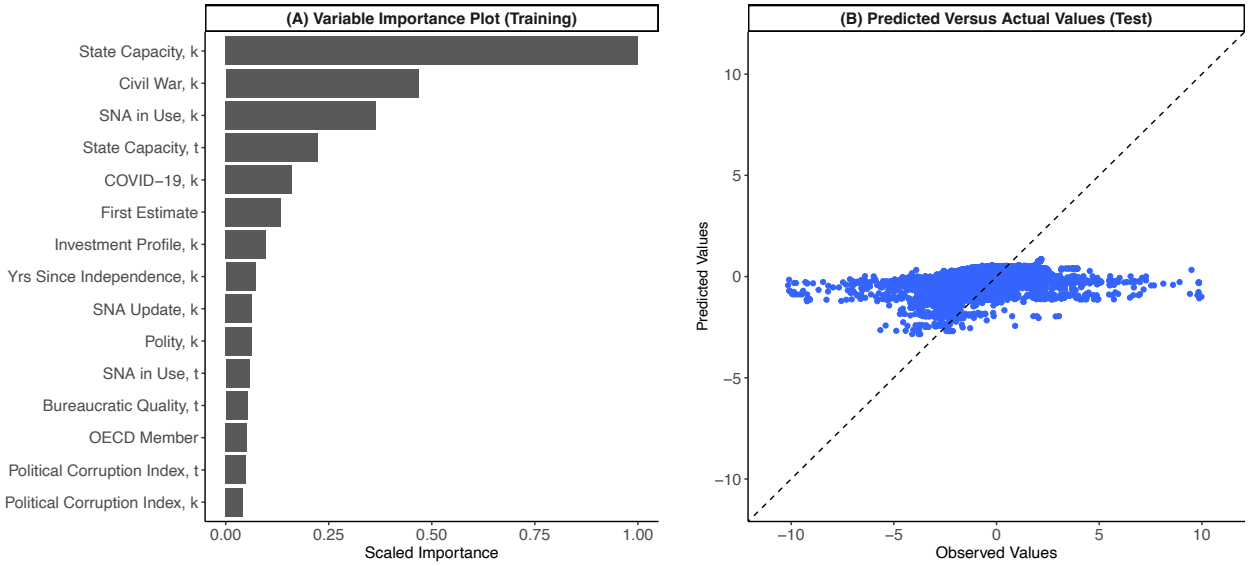
Focusing on non-missing values, the z-score

$$z_{itk} = \frac{x_{ijk} - \mu_{ij}}{\sigma_{ij}}$$

divides the raw difference between x_{itk} and the country-year mean μ_{ij} by the country-year standard deviation, σ_{ij}. Put differently, the z-score indicates how many standard deviations one single measurement is from its country-year mean. The GDP z-score ranges from –10.34 to 10.34, though 95.7 percent of all non-missing observations fall within two standard deviations of the mean.

In Figure E.4, panel (A) presents the relative importance of the 15 most important predictors for the training set. The least important predictor equals zero, while the most important predictor equals one; as before, *State Capacity* is the most important predictor.

Figure E.4: Assessing the Performance of a Model Predicting the Z-Score



These figures assess the fit of a random forest predicting the outcome *Z-Score*. Panel (A) is a variable importance plot, indicating the relative importance of the 15 most important predictors; t or k denotes the predictor's value for the reported or reporting year, respectively. Panel (B) is a scatter plot of predicted values (x-axis) against actual values (y-axis), along with a line indicating perfect predictions (x = y). Points that are far from the line indicate large errors.

Since this outcome is continuous, I use different metrics to assess its performance. Panel (B) of Figure E.4 plots predicted values (x-axis) against actual values (y-axis), along with a line indicating perfect predictions (x = y). The closer the points are to the diagonal line, the better the model's predictions align with the actual values. The R^2 indicates the correlation between predicted and observed values, from 0 (no correlation) to 1 (complete correlation). As Table E.4 shows, R^2 for the test set is 0.22: only 22 percent of the out-of-sample variation in z-scores can be systematically explained. The model consistently makes predictions that are

up to two standard deviations above or below the mean, an accurate prediction for 95.7 percent of all non-missing observations. The predictors can systematically identify extreme values but tend to underestimate their magnitude. As a result, models are unable to correctly predict z-scores of –10.34 (for Argentina's 1991 GDP, according to the April 1999 WDI) or 10.34 (for Zambia's 1990 GDP, according to the April 1994 WDI).

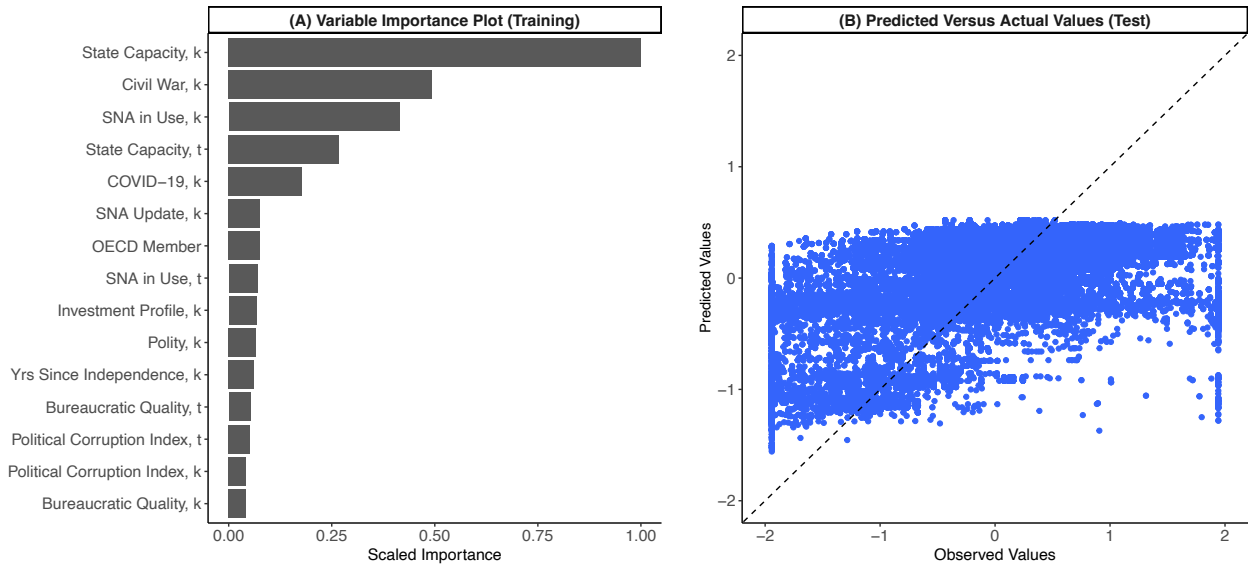
Still, the R^2 is not the ideal metric to assess outlier prediction: it generally reflects how well the model fits the majority of the data rather than extreme values. It emphasizes the model's ability to predict values near the mean, which is not useful when the focus is on outliers. Thus, Table E.4 also reports the results of the Mean Squared Error (MSE), which measures the average squared difference between the observed and the predicted values. The measurement unit for the MSE is the same as the unit for the outcome of interest (in this case, from –10.34 to 10.34), with smaller values reflecting more accurate predictions. The MSE for the training set is 0.75 and declines to 0.72 in the test set. This corresponds to a Root Mean Squared Error (RMSE) of 0.86 and 0.85, respectively: on average, the model's predictions deviate from the actual z-scores by about 0.85 standard deviations.

To reduce the influence of outliers, I also estimate models predicting the winsorized z-score. Since 95.7 percent of all non-missing observations fall within two standard deviations of the mean, I choose this as the cutoff. This means that values *more* than two standard deviations away from the mean are adjusted (or "shrink") to *exactly* two standard deviations away from the mean. Figure E.5 presents the results. The most important predictors are the same, but as the x-axis in panel (B) indicates, the observed values are now capped at –2 and 2. The y-axis indicates that the model continues to make conservative predictions (between –1.5 and 0.5), but now these predictions are slightly closer to the truth, hence the larger R^2 and lower MSE/RMSE in Table E.4.

Table E.4: Performance Statistics for a Model Predicting the (Winsorized) Z-Score

	Training	Validation	Test
Z-Score			
R^2	0.2138257	0.1892496	0.2202714
MSE	0.7478067	0.7591492	0.7231198
RMSE	0.8647582	0.8712917	0.8503645
Z-Score (Winsorized)			
R^2	0.2637889	0.2433378	0.2691619
MSE	0.4623150	0.4516757	0.4762011
RMSE	0.6799375	0.6720682	0.6900733

Figure E.5: Assessing the Performance of a Model Predicting the Winsorized Z-Score



These figures assess the fit of a random forest predicting the outcome *Winsorized Z-Score*. Panel (A) is a variable importance plot, indicating the relative importance of the 15 most important predictors; t or k denotes the predictor's value for the reported or reporting year, respectively. Panel (B) is a scatter plot of predicted values (x-axis) against actual values (y-axis), along with a line indicating perfect predictions (x = y). Points that are far from the line indicate large errors.

E.3 Alternative Outcome: Percentage Deviation From the Median

Again, I turn to the non-missing values and examine their percentage deviation from the country-year median, $\tilde{X}_{ij}$:

$$dev_{itk} = \frac{x_{ijk} - \tilde{X}_{ij}}{\tilde{X}_{ij}} \times 100$$

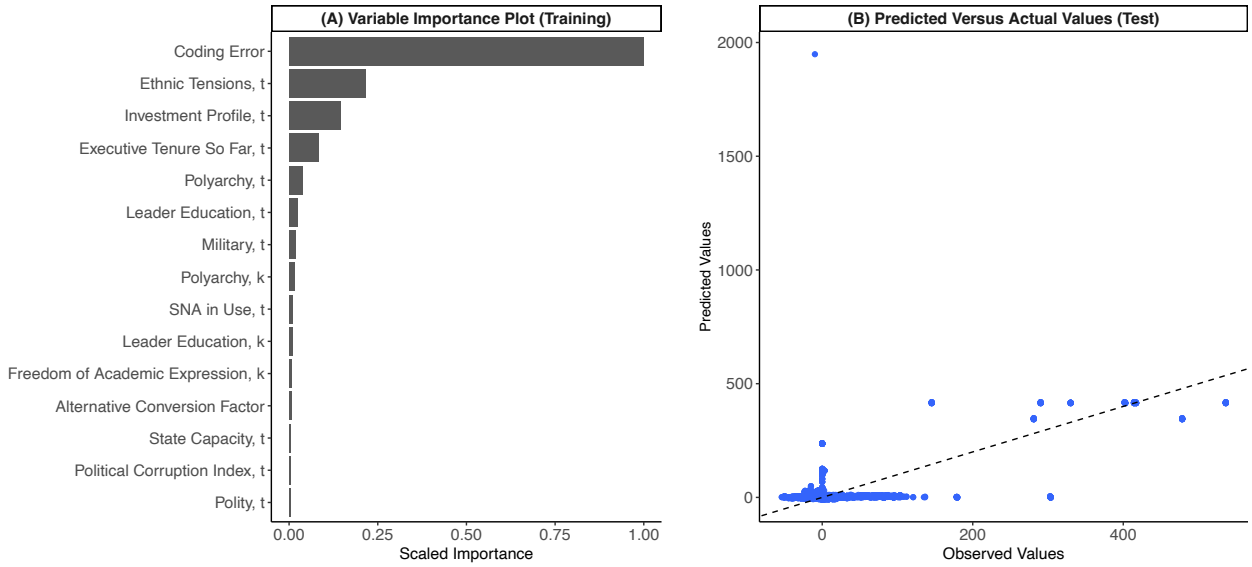
Compared to the mean, the median provides a more accurate central tendency measure in the presence of outliers. In distributions with a long tail or multiple peaks, the median can provide a better indication of where the center of the data lies. Unlike the *absolute* deviation, the *percentage* deviation standardizes the data and allows for cross-country comparisons.

In Figure E.6, panel (A) presents the relative importance of the 15 most important predictors for the training set. The least important predictor equals zero, while the most important predictor equals one. This time, *Coding Error* is the most important predictor: mistakes in data reporting tend to be associated with very large percentage deviations from the median. To illustrate this, consider panel (B), which again plots predicted values (x-axis) against actual values (y-axis), along with a line indicating perfect predictions (x = y). The extreme values at the bottom right correspond to 16 vintages between June 2018 and April 2020 that reported very large values for Timor-Leste's GDP and were flagged as erroneous (see Figure B.1). For Timor-Leste in 2008 GDP, for example, these vintages reported a GDP that deviated from the median by

536.23 percent; the model predicts a deviation of 418.90 percent, which is relatively close.

The observed value in the upper left corner corresponds to China in 2007, as reported by a 2009 WDI vintage: its predicted deviation from the median is 1,948.49 percent, and its actual deviation from the median is –9.71 percent. This aligns with an error identified by the World Bank: "The September update of the WDI 2009 database contained an error for China's current U.S. dollar GDP for 2007 and 2008" (World Bank, 2023).

Figure E.6: Assessing the Performance of a Model Predicting the Z-Score



These figures assess the fit of a random forest predicting the outcome *Percentage Deviation From the Median*. Panel (A) is a variable importance plot, indicating the relative importance of the 15 most important predictors; t or k denotes the predictor's value for the reported or reporting year, respectively. Panel (B) is a scatter plot of predicted values (x-axis) against actual values (y-axis), along with a line indicating perfect predictions (x = y). Points that are far from the line indicate large errors.

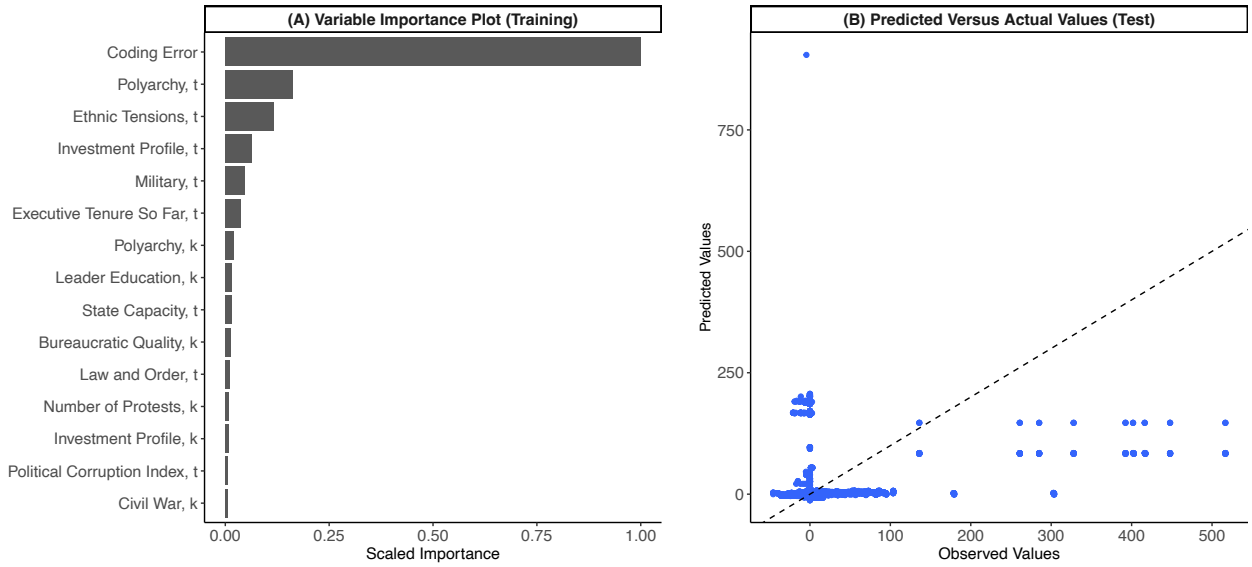
Table E.5: Performance Statistics for a Model Predicting the (Winsorized) Percentage Deviation From the Median

	Training	Validation	Test
Perc. Deviation From the Median			
R^2	0.466992791	–1.825295882	0.4644636
MSE	4443.1755	174.2011	192.9489
RMSE	66.65715	13.19853	13.8906
Perc. Deviation From the Median (Winsorized)			
R^2	0.48781738	–0.47548318	–0.1199223
MSE	577.97033	59.76702	373.6962
RMSE	24.041014	7.730913	19.33122

As Table E.5 shows, this model does not have a good fit. This is unsurprising, as the most important predictor is a dichotomous indicator of data idiosyncrasies that simply cannot be predicted on a systematic

basis. Several observations with coding errors (like the Chinese and Timorese observations discussed above) appear to have been assigned to the test set, which would explain this set's comparatively high R^2 and low RMSE. As before, I also estimate models predicting the winsorized percentage deviation from the median to reduce the influence of outliers, again using two standard deviations from the mean as a cutoff. While the top predictor continues to be *Coding Error* (see Figure E.7), the model fit is even worse, again reflecting the problem that idiosyncratic errors cannot be predicted on a systematic basis.

Figure E.7: Assessing the Performance of a Model Predicting the Winsorized Percentage Deviation From the Median



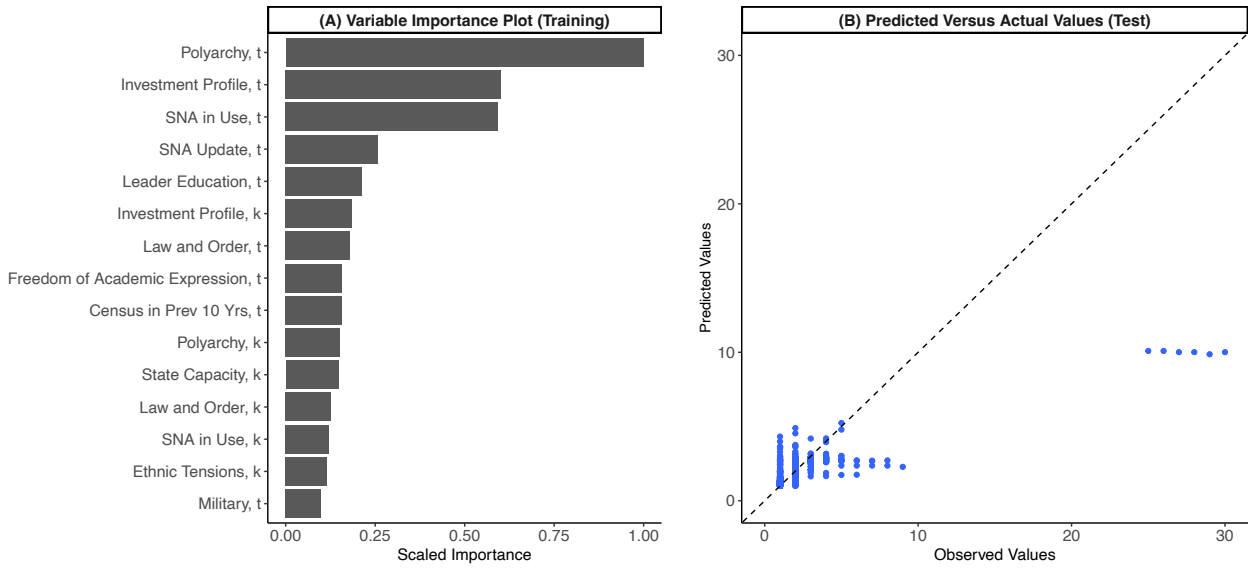
These figures assess the fit of a random forest predicting the outcome *Winsorized Percentage Deviation From the Median*. Panel (A) is a variable importance plot, indicating the relative importance of the 15 most important predictors; t or k denotes the predictor's value for the reported or reporting year, respectively. Panel (B) is a scatter plot of predicted values (x-axis) against actual values (y-axis), along with a line indicating perfect predictions (x = y). Points that are far from the line indicate large errors.

E.4 Alternative Outcome: Reporting Speed

About 39.8 percent of all economies surveyed by the IMF in 2020 disseminated their annual GDP data within 90 days of the reference period, whereas 54.4 percent did so within 91 to 365 days, with no available information for the remaining 5.8 percent (Baer, Guerreiro and Silungwe, 2022, 17). What predicts this variation? The final outcome I examine is the *Reporting Speed*, the time elapsed between a reported year t and its first estimate, which previous research has shown to be positively related to better governance (Islam, 2006). This model is restricted to the first estimates, assuming one exists ($N = 5,416$).

As Figure E.8 shows, the most important predictor of *Reporting Speed* is the Polyarchy score: democracies

Figure E.8: Assessing the Performance of a Model Predicting the Reporting Speed



These figures assess the fit of a random forest predicting the outcome *Reporting Speed*. Panel (A) is a variable importance plot, indicating the relative importance of the 15 most important predictors; t or k denotes the predictor's value for the reported or reporting year, respectively. Panel (B) is a scatter plot of predicted values (x-axis) against actual values (y-axis), along with a line indicating perfect predictions (x = y). Points that are far from the line indicate large errors.

report data more quickly than autocracies. The second most important predictor is the ICRG investment profile measure, reflecting a connection between investment attraction and up-to-date statistics. Turning to the scatter plot in panel (B), the five observations on the right correspond to Iraq for all reported years between 1991 and 1997. For these observations, the model predicted a ten-year gap between reported and reporting years, but the truth is even worse: these seven observations were reported for the first time in the July 2021 WDI. As Table E.6 shows, this model does not make good predictions for the training set, let alone for the validation and test sets, reflecting how difficult it is to systematically predict unsystematic events.

Table E.6: Performance Statistics for a Model Predicting the Reporting Speed

	Training	Validation	Test
R^2	0.6870651	0.4153909	0.5016181
MSE	0.8665914	0.4027653	2.326748
RMSE	0.9309089	0.6346379	1.525368

F Hyperparameters

F.1 Classification Trees

I estimate all models using the open source machine learning platform `H2O`, implemented via `R`. To predict the dichotomous outcomes *Revision*, *Missing*, and *Outlier* (a classification task), I estimate a random forest with the hyperparameters described below; the description draws heavily from the `H2O.ai` user documentation (available under `https://docs.h2o.ai/`) as well as from Cook (2017, 117-125). Using a cartesian grid search, I trained a random forest for every possible combination of the hyperparameter values, sorted the resulting models according to their performance (as measured by the mean squared error), and chose the model that returned the lowest mean squared error. I maintained several of the default values provided by `H2O`, because there are so many available observations that not much additional calibration is needed to improve performance.

`nfolds = ''fold''`. The model performs leave-one-group-out cross-validation, with groups (countries) indicated in the column "fold."

`ntrees = 500`. This is the number of trees. Higher values are computationally intensive and do not perform better.

`sample_rate = 0.7`. Each tree is trained on 70 percent of the training data, drawn at random and without replacement (default value is 0.6320000291).

`col_sample_rate_per_tree = 0.7`. 70 percent of all columns are used for each tree, without replacement (default value is 1). This allows for different columns to be selected for different trees.

`col_sample_rate = 1`. Out of the 70 percent of columns used for each tree (`col_sample_rate_per_tree`), 100 percent are used for each split decision (default value).

`max_depth = 12`. Tree depth is the number of edges from the root node to the deepest leaf. The maximum tree depth is specified as 12 (default value is 5). This means that the longest path from the root to any leaf node in the tree can have at most 12 edges, resulting in up to 12 levels.

`min_rows = 12`. This parameter specifies the minimum number of observations for a terminal node (default value is 1).

`min_split_improvement = 1e-4`. This option specifies the minimum relative improvement in squared error reduction in order for a split to occur (default is 1e-5).

`stopping_rounds = 10`. The model uses early stopping: it stops training when the option selected for stopping_metric does not improve for 10 training rounds, based on a simple moving average (default value is 0, without early stopping).

`stopping_metric = ''AUC''`. The default stopping metric for classification tasks is the AUC.

`stopping_tolerance = 1e-3`. This is the tolerance value by which a model must improve before training ceases (default value).

`balance_classes = T`. This hyperparameter only exists for classification tasks; it balances the class distribution, either by undersampling the majority class or by oversampling the minority class.

F.2 Regression Trees

As before, I estimate all models using the open source machine learning platform `H2O`, implemented via `R`. To predict the outcomes in Appendix E (a regression task), I estimate a random forest with the following hyperparameters:

`nfolds = ''fold''`. The model performs leave-one-group-out cross-validation, with groups (countries) indicated in the column "fold."

`ntrees = 200`. This is the number of trees. Higher values are computationally intensive and do not perform better.

`sample_rate = 0.8`. Each tree is trained on 80 percent of the training data, drawn at random and without replacement (default value is 0.6320000291).

`col_sample_rate_per_tree = 0.8`. 80 percent of all columns are used for each tree, without replacement (default value is 1). This allows for different columns to be selected for different trees.

`col_sample_rate = 1`. Out of the 80 percent of columns used for each tree (`col_sample_rate_per_tree`), 100 percent are used for each split decision (default value).

`max_depth = 6`. Tree depth is the number of edges from the root node to the deepest leaf. The maximum tree depth is specified as 6 (default value is 5). This means that the longest path from the root to any leaf node in the tree can have at most 6 edges, resulting in up to 6 levels.

`min_rows = 10`. This parameter specifies the minimum number of observations for a terminal node (default value is 1).

`min_split_improvement = 1e-3`. This option specifies the minimum relative improvement in squared error reduction in order for a split to occur (default is 1e-5).

`stopping_rounds = 10`. The model uses early stopping: it stops training when the option selected for stopping_metric does not improve for 10 training rounds, based on a simple moving average (default value is 0, without early stopping).

`stopping_metric = ''AUTO''`. The default stopping metric for regression tasks is the mean residual deviance.

`stopping_tolerance = 1e-3`. This is the tolerance value by which a model must improve before training

ceases (default value).

References

Baer, Andrew, Vanda Guerreiro and Anthony Silungwe. 2022. "2020 Global Stocktaking of National Accounts Statistics: Availability for Policy and Surveillance." *IMF Working Papers* (29):1–25.

Becker, Bastian. 2019. "Introducing COLDAT: The Colonial Dates Dataset." *SOCIUM/SFB1342 Working Paper Series* (2):1–17.

Bell, Curtis, Clayton Besaw and Matthew Frank. 2021. *The Rulers, Elections, and Irregular Governance (REIGN) Dataset.*
URL: *https://oefdatascience.github.io/REIGN.github.io/*

Centre for Research on the Epidemiology of Disasters. 2020. *EM-DAT: The International Disaster Database.*
URL: *https://public.emdat.be*

Chinn, Menzie D. and Hiro Ito. 2006. "What Matters for Financial Development? Capital Controls, Institutions, and Interactions." *Journal of Development Economics* 81(1):163–192.

Clark, David and Patrick Regan. 2020. *Mass Mobilization Protest Data.*
URL: *https://massmobilization.github.io/*

Cook, Darren. 2017. *Practical Machine Learning with H2O: Powerful, Scalable Techniques for Deep Learning and AI.* Sebastopol, CA: O'Reilly.

Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell, David Altman, Michael Bernhard, M. Steven Fish, Agnes Cornell, Sirianne Dahlum, Haakon Gjerløw, Adam Glynn, Allen Hicken, Joshua Krusell, Anna Lührmann, Kyle L. Marquardt, Kelly McMann, Valeriya Mechkova, Juraj Medzihorsky, Moa Olin, Pamela Paxton, Daniel Pemstein, Josefine Pernes, Johannes von Römer, Brigitte Seim, Rachel Sigman, Jeffrey Staton, Natalia Stepanova, Aksel Sundström, Eitan Tzelgov, Yi-ting Wang, Tore Wig, Steven Wilson and Daniel Ziblatt. 2023. *V-Dem Country-Year Dataset v13.*

Cruz, Cesi, Philip Keefer and Carlos Scartascini. 2021. *Database of Political Institutions 2020.*

Cust, James, David Mihalyi and Alexis Rivera-Ballesteros. 2021. The Economic Effects of Giant Oil and Gas Discoveries. In *Giant Fields of the Decade: 2010-2020*, ed. Charles A. Sternbach, Robert K. Merrill and John C. Dolson. Tulsa: AAPG pp. 21–36.

Dang, Hai Anh H., John Pullinger, Umar Serajuddin and Brian Stacy. 2023. "Statistical Performance Indicators and Index—a New Tool to Measure Country Statistical Capacity." *Scientific Data* 10(1):1–14.

Dreher, Axel, Andreas Fuchs, Andreas Kammerlander, Lennart Kaplan, Charlott Robert and Kerstin Unfried. 2020. *The Political Leaders' Affiliation Database*.

Dreher, Axel, Valentin F. Lang, B. Peter Rosendorff and James Raymond Vreeland. 2022. "Bilateral or Multilateral? International Financial Flows and the Dirty-Work Hypothesis." *Journal of Politics* 84(4):1932–1946.

Gleditsch, Nils Petter, Peter Wallensteen, Mikael Eriksson, Margareta Sollenberg and Håvard Strand. 2002. "Armed Conflict 1946–2001: A New Dataset." *Journal of Peace Research* 39(5):615–637.

Graham, Benjamin A. T. and Jacob R. Tucker. 2019. "The International Political Economy Data Resource." *Review of International Organizations* 14:149–161.

Graham, Benjamin A.T., Raymond Hicks, Helen Milner and Lori D. Bougher. 2018. *World Economics and Politics Dataverse*.
URL: *https://ncgg.princeton.edu/wep/dataverse.html*

Hanson, Jonathan K. and Rachel Sigman. 2021. "Leviathan's Latent Dimensions: Measuring State Capacity for Comparative Political Research." *Journal of Politics* 83(4):1–16.

Horn, Myron K. 2014. *Giant Oil and Gas Fields of the World*.
URL: *https://edx.netl.doe.gov/dataset/aapg-datapages-giant-oil-and-gas-fields-of-the-world*

Islam, Roumeen. 2006. "Does More Transparency Go Along With Better Governance?" *Economics and Politics* 18(2):121–167.

Kentikelenis, Alexander E., Thomas H. Stubbs and Lawrence P. King. 2016. "IMF Conditionality and Development Policy Space, 1985–2014." *Review of International Political Economy* 23(4):543–582.

Marshall, Monty G. 2019. *Major Episodes of Political Violence, 1946–2018*.

Marshall, Monty G. and Ted Robert Gurr. 2020. *Polity5: Political Regime Characteristics and Transitions, 1800–2018*.
URL: *http://www.systemicpeace.org/inscrdata.html*

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(1):87–103.

Nguyen, Thanh Cong, Vítor Castro and Justine Wood. 2022. "A New Comprehensive Database of Financial Crises: Identification, Frequency, and Duration." *Economic Modelling* 108:105770.

Pettersson, Therése, Shawn Davies, Amber Deniz, Garoun Engström, Nanar Hawach, Stina Högbladh and Margareta Sollenberg Magnus Öberg. 2021. "Organized Violence 1989–2020, With a Special Emphasis on Syria." *Journal of Peace Research* 58(4):809–825.

The PRS Group. 2022. *International Country Risk Guide (ICRG)*.

World Bank. 2023. *Data Updates and Errata*.
URL: *https://datahelpdesk.worldbank.org/knowledgebase/articles/906522-data-updates-and-errata*