Measurement Uncertainty in International Statistics[*]

Iasmin Goes[†]

January 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 WDI provide conflicting information. Why? I use machine learning to understand what predicts variation in data coverage and comparability across all WDI releases from 1994 to 2022. Authoritarianism, corruption, low state capacity, and recent independence strongly predict whether observations are missing, but not whether they are comparable. Rather, different WDI releases become more comparable as they adopt newer standardization frameworks or simply as time goes by. These findings highlight the importance of disclosing the chosen data sources and releases, which might affect researchers' empirical results.

[*] For constructive feedback, I thank Brendan Apfeld, Sabrina Arias, Ryan Briggs, Terry Chapman, Kerice Doten-Snitker, Timon Forster, Matt Hitt, Andrew Kerner, Anna Minasyan, Saliha Metinsoy, Carolina Moehlecke, Gaurav Sood, Daniel Weitzel, and participants of the University of Vienna's Government Department Workshop, the 2022 ACUNS-UN Workshop, the 2022 Joint IPE Conference of the Austrian, German, and Swiss Political Science Associations, and the 2023 EWIS Workshop.

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

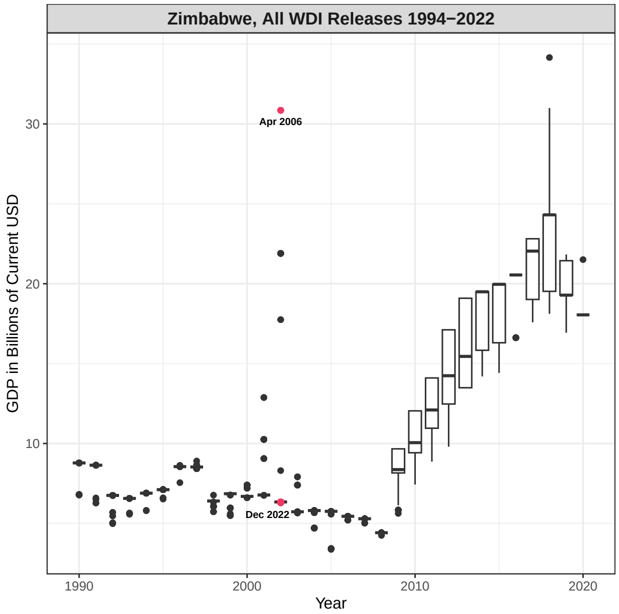
1 Introduction

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 many other 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 economic indicators are less controversial. Gross domestic product (GDP) captures the value of all final goods and services produced in a country during a specific period; it does not rely on the construction of subjective categories to measure latent concepts like business regulation, human development, or freedom. Despite growing criticism (Mügge, 2022; Merry, 2011), GDP is still widely used in the social sciences, suggesting that researchers consider it a valid and reliable indicator: it accurately captures its underlying theoretical concept (the size of a country's economy) and provides consistent information across repeated measurements (Gerring, 2012).

Still, GDP measurements are not as consistent as they might seem. The three main data sources — the World Bank's World Development Indicators (WDI), the Penn World Table (PWT), and the Maddison Project — often provide conflicting information. In fact, different versions *of the same data source* often provide conflicting information, as Figure 1 shows (see also Goes 2023). According to WDI figures released in April 2006, Zimbabwe's GDP in 2002 was around 30.8 billion current US dollars; the December 2022 WDI release reduced this number to just 6.3 billion. It is difficult, if not impossible, to assess the accuracy of these

Figure 1: Current GDP of Zimbabwe, 1990–2020



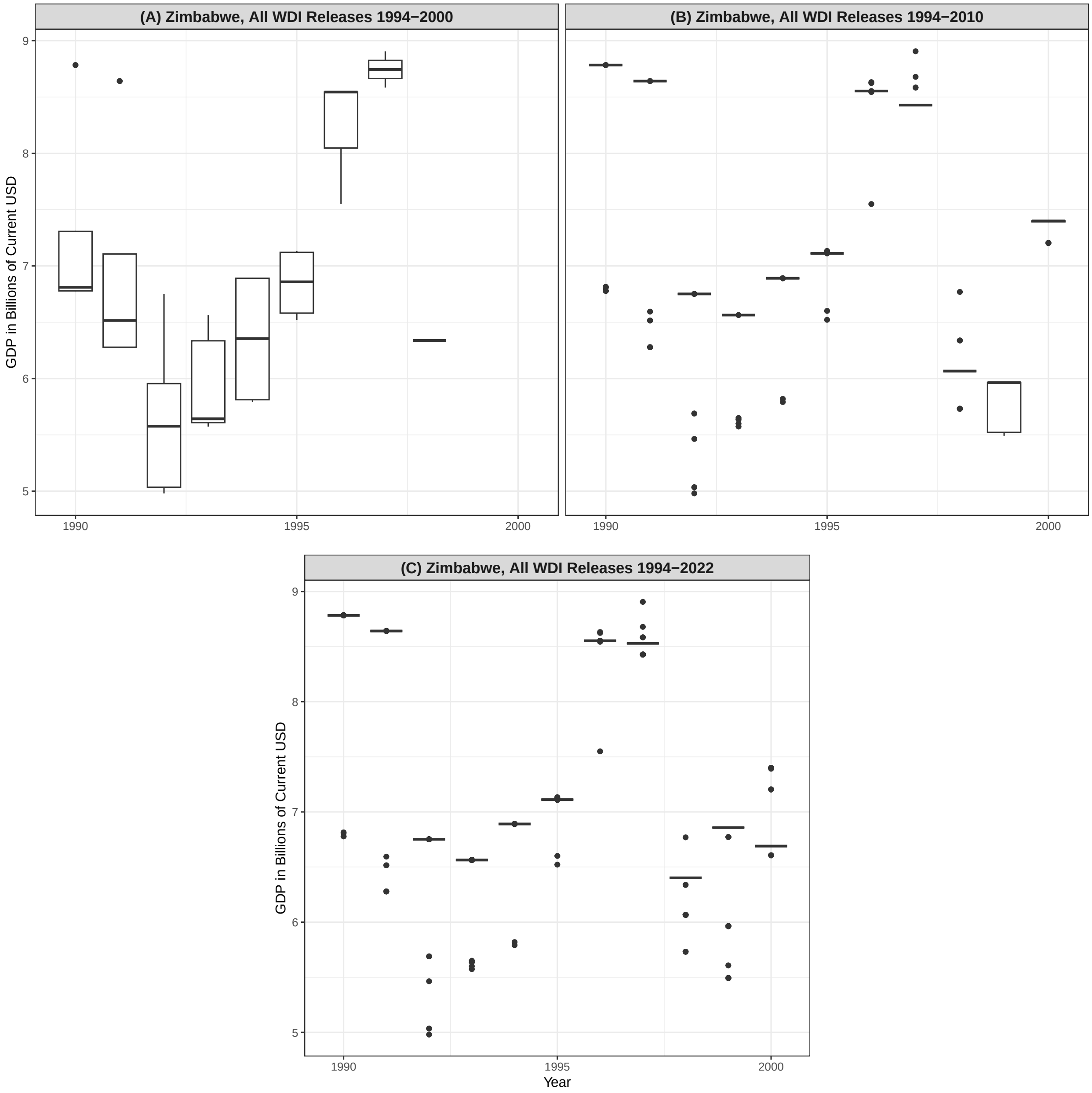
This boxplot presents the distribution of current GDP estimates for Zimbabwe from 1990 to 2020, using data drawn from the 109 WDI releases from April 1994 to December 2022. The estimate reported for 2002 is 24.5 billion dollars larger in the April 2006 WDI than in the December 2022 WDI. Section 3 discusses the data in more detail.

estimates: researchers do not know how far each measurement is from Zimbabwe's true GDP in 2002 (the measurement error). But it is possible to compare GDP measurements across different WDI releases (also called *vintages*) to quantify their reliability (the measurement uncertainty).[1] This is what the present study aims to do.

I begin by reviewing a rich literature that identifies several sources of measurement uncertainty in economic data. Autocracies (Hollyer, Rosendorff and Vreeland, 2011), islands (Ram and Ural, 2014), and African states (Devarajan, 2013) disclose statistics less frequently, and their statistics tend to be of lower quality. With this research as a starting point, I use machine learning to identify the systematic predictors of measurement uncertainty across different vintages of the WDI, the most prevalent source of economic data in political science research (Goes, 2023). Zooming in on Zimbabwe, Figure 2 already identifies one source of uncertainty: time. Data for older periods tend to be less noisy than data for recent periods.

[1] This is "the problem of consistency, comparability, or reliability across countries" outlined by Herrera and Kapur (2007, 368); I use the three words interchangeably.

Figure 2: Current GDP of Zimbabwe, 1990–2000



These boxplots present the distribution of current GDP estimates for Zimbabwe from 1990 to 2000, using data drawn from (A) the 8 WDI releases from April 1994 to April 2000, (B) the 26 releases from April 1994 to December 2010, and (C) the 109 releases from April 1994 to December 2022. As the number of available releases goes up, the uncertainty goes down: different releases tend to coalesce around one value for each country-year pair. Section 3 discusses the data in more detail.

The more time has elapsed, the more vintages are available, which means there are more measurements of each country-year pair — and these tend to coalesce around one value. In addition, results coincide with Hollyer, Rosendorff and Vreeland (2011) that uncertainty is best predicted by a mix of bureaucratic capacity and political will: different vintages often provide no GDP information, or conflicting GDP information, when the country in question lacks the resources or political incentives to provide accurate data. Still, not all data issues can be systematically predicted; many are idiosyncratic to specific countries and years.

Fariss et al. (2022) have previously quantified uncertainty in GDP and population data, whereas Johnson et al. (2013) and Goes (2023) showed that this uncertainty affects the replicability of published studies in economics and political science, respectively. Building on their work, I provide descriptive evidence with important empirical implications. First, researchers should be transparent about their data sources and vintages. Second, researchers should not draw conclusions based on recent years alone, since data for these years are noisy and susceptible to revisions. Third, those who study autocracies, islands, or African states should measure the size of a country's economy using multiple indicators, not just GDP, given the prevalence of missing or unreliable data. Finally, scholars should be modest when interpreting empirical findings: one cannot trust empirical findings unless one can trust the underlying data.

2 The Sources of Uncertainty

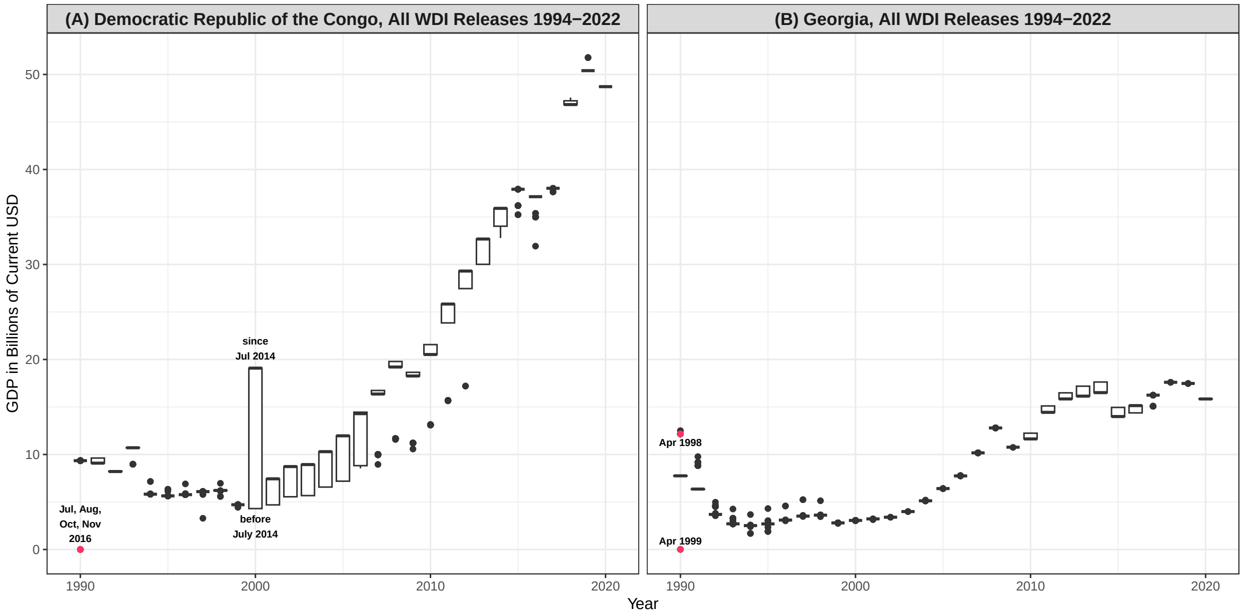
WDI, PWT, and Maddison estimates can diverge significantly, even if the underlying data are the same. Ram and Ural (2014) identify 33 cases (typically island nations or countries in Sub-Saharan Africa) for which GDP estimates from the WDI and the PWT differ by over 25 percent. This issue goes beyond GDP data: 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 International Monetary Fund (IMF) and the United Nations Commodity Trade Statistics (Comtrade) — 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), foreign direct investment (Kerner, 2014), and population data (Devarajan, 2013) face similar measurement issues.

Previous research has identified four primary sources of measurement uncertainty. The first is statistical capacity, or lack thereof. International organizations do not compile statistics. The World Bank, for example, "was never involved in actual basic data collection for the national accounts" (Ward, 2004, 98). Instead, it disseminates data from national statistical agencies, which report data in line with a global standardization framework — the System of National Accounts (SNA) — developed by the International Comparison Program (ICP) to enable cross-country comparisons. Many national statistical agencies are underfunded, understaffed, use outdated methods, and do not coordinate their statistical activities (Devarajan, 2013). Population figures tend to be extrapolated from the last census; the more time has elapsed since the last census, the larger the uncertainty included in these extrapolations (Devarajan, 2013). As a result, statistical agencies either fail to report estimates altogether or report inaccurate estimates, a problem that is particularly prevalent in Africa (Jerven, 2010, 2013, 2018, 2019).

Measurement uncertainty can also exist for political reasons. Autocracies are less likely to report policy-relevant data (Hollyer, Rosendorff and Vreeland, 2011), and when they do, they tend to overstate annual growth rates (Magee and Doces, 2015; Martínez, 2022), particularly at politically sensitive times (Wallace, 2014). In federations like Nigeria, states inflate population figures to receive higher fiscal transfers from the federal government (Devarajan, 2013). Aid-dependent countries systematically underreport economic data 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 in order to abide by the rules of the European Union, as Greece did (Alt, Lassen and Wehner, 2014).

A third source of uncertainty is the ICP standardization framework. Every five to ten years, the ICP surveys how much the same basket of goods costs in different currencies, using this information to construct purchasing power parity (PPP) rates that enable the comparison of living standards across borders. Until 1996, these price surveys collected data only for the developed world, making considerably less accurate extrapolations for the developing world (Deaton and Aten, 2017). ICP rounds in 2005, 2011, and 2017 suffer from smaller uncertainty because they include China and other large developing countries, but still 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).

Figure 3: Curent GDP of the Democratic Republic of the Congo and Georgia, 1990–2020



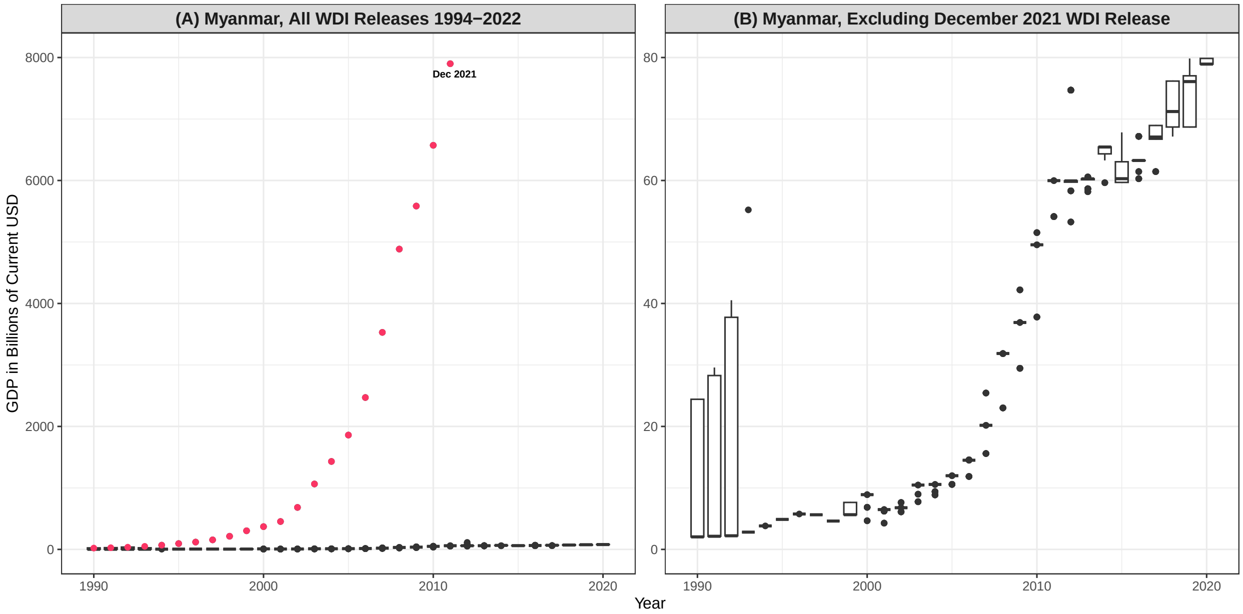
These boxplots present the distribution of current GDP estimates from 1990 to 2020 for (A) the Democratic Republic of the Congo and (B) Georgia, using data drawn from the 109 WDI releases from April 1994 to December 2022. Four WDI releases reported a GDP of zero for the Democratic Republic of the Congo 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.

Humans are a final source of measurement uncertainty: they can commit coding errors, selectively exclude available data, or weigh summary statistics inappropriately (Herndon, Ash and Pollin, 2014). As Figure 3 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. Relatedly, 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.[2] "Losing" three digits is not the product of low statistical capacity, political interference, or faulty standardization; it is the product of human error, as is a GDP of zero. The December 2021 update contains a similar error for Myanmar, illustrated in Figure 4. According to all other 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 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.

These issues are not unsolvable. With support from the Danish International Development Agency and the IMF, the Ghana Statistical Service released new GDP estimates in 2010: after updating the base year from 1993 to 2006 and 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). 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 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

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

Figure 4: 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 109 WDI releases from April 1994 to December 2022. 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.

to lower GDP growth.

Revisions are consequential for both policy and research. In 2010, the World Bank updated Ghana's classification from low income to lower middle income economy, and the government suddenly became eligible to apply for loans from the International Bank for Reconstruction and Development. Greece's revisions had a less fortunate effect: the country was downgraded by credit rating agencies and requested multiple IMF and EU loans to avoid default. Given the data discrepancies across WDI and PWT vintages (Goes, 2023; Johnson et al., 2013), replacing one vintage with another can also significantly alter published research findings. Overall, low statistical capacity, deliberate political choice, imperfect standardization practices, and human error lead to heterogeneity in data quality: researchers can make more accurate and precise inferences about some countries and years than others.

3 Predicting Measurement Uncertainty

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 109 available electronic WDI releases from 1994 to 2022.[3] I focus on the indicator *GDP in current US dollars* (ID `NY.GDP.MKTP.CD`), defined as "the sum of gross value added by all resident producers in the economy."[4] This indicator 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 to filter out the noise. 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; accounting for it would increase computational demands without any substantive gain in meaning.

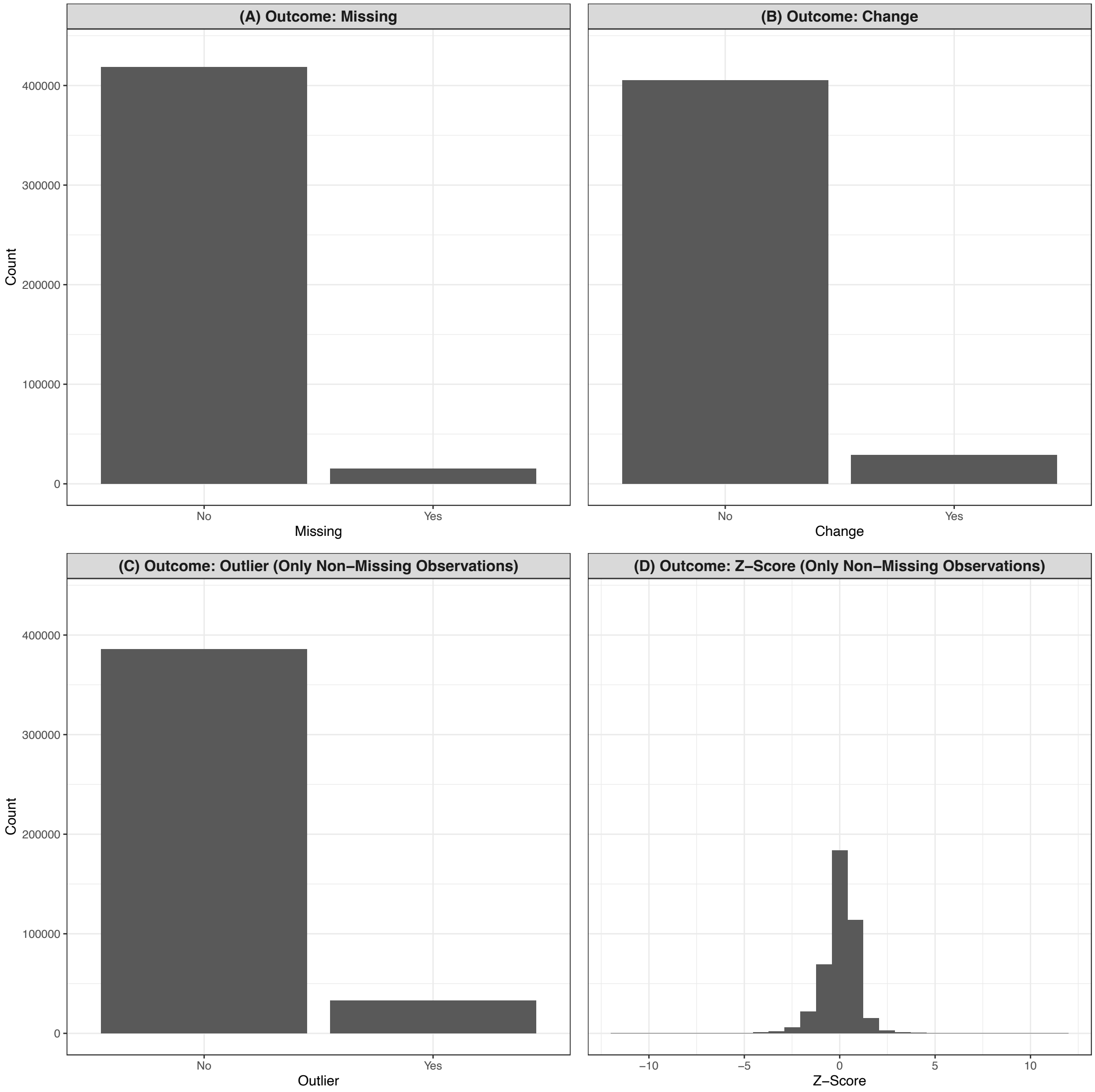
I use this indicator to generate four outcomes that capture different degrees of measurement uncertainty; Figure 5 shows their distribution. The first outcome measures data coverage, whereas the remaining three measure data comparability across vintages. Consider each observation x_{itk} for country i, year t (the *reported* date), and WDI release k (the *reporting* date), with $N = 433,803$. The first outcome is missingness, coded one if x_{itk} is missing from WDI release k and zero otherwise. Of all observations, 15,160 (about 3.5 percent) are

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

[4] GDP is "converted from domestic currencies [into US dollars] using single year official exchange rates." In the rare cases "where the official exchange rate does not reflect the rate effectively applied to actual foreign exchange transactions," the World Bank applies an alternative exchange rate, the so-called DEC conversion factor.

[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.

Figure 5: Distribution of the Outcome Variables



These histograms show the distribution of the four outcomes of interest. According to (A), 3.5 percent of all observations are missing. According to (B), 9.9 percent of all observations record a change from x_{itk} to x_{itk+1}. According to (C), 9.3 percent of all non-missing observations are outliers, as defined by the Tukey rule. Finally, (D) shows the distribution of the z-score, indicating that 96 percent of all non-missing observations fall within two standard deviations of the mean.

missing. The second outcome is change, 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. There are 42,814 instances of change (9.9 percent).

Turning to the non-missing values ($N = 418,643$), the third outcome indicates the presence of an outlier, coded one if x_{itk} falls outside of the typical ranges for country i and year t. To identify outliers, the Tukey rule (also employed in the construction of boxplots) leverages the Interquartile Range (IQR), which is 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 $Q3 + 1.5 \times IQR$. About 9.3 percent of all non-missing observations are outliers (39,035).

The fourth outcome also focuses on non-missing values and operationalizes uncertainty as distance from the mean: how many standard deviations is a non-missing observation away from its country-year mean μ_{ij}? This outcome corresponds to the z-score,

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

which divides the raw difference between x_{itk} and μ_{ij} by the country-year standard deviation, σ_{ij}. This metric indicates a direction (whether an observation is above or below the mean) and standardizes the data (allowing for meaningful comparison between countries). While the z-score ranges from –10.34 to 10.34, about 96 percent of all non-missing observations fall within two standard deviations of the mean.

3.2 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 measurement uncertainty. 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[6]

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

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 (GBMs) — 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 GBMs to understand how GDP figures vary across different WDI vintages and present the results of random forests, LASSO, and ridge regressions in the appendix.

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 all sets.[7] One concern is that unintentional information might leak across related observations. For instance, the model might use information from Zimbabwe's future to predict Zimbabwe's past (temporal leakage), or it might use information from Zimbabwe in 1990 to predict outcomes for 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 future cannot be used to predict Zimbabwe's past. 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 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 (see Appendix F). Once I am satisfied with the results, I use the chosen model to make out-of-sample predictions 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 approximately 14.7, 27.7, 25.8, and 31.3 percent of the dataset, respectively. The remaining 0.5 percent of observations correspond to Venezuela, which has been temporarily unclassified since July 2021 due to lack of revised national accounts statistics.

the test set. This final evaluation on unseen data provides a reliable measure of the model's predictive capability and its real-world applicability.

3.3 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 37 variables that plausibly explain measurement uncertainty. Some of these variables are political (regime type, election year, Polyarchy scores, ideology of the executive), others indicate the occurrence of specific events (like elections, financial crises, or climate disasters), and others, still, are V-Dem indices (Coppedge et al., 2023) measuring freedom of academic expression or bureaucratic remuneration (see Appendix C). I restrict the analysis to all values of t from 1990 to 2020, as different data sources cover different periods: while V-Dem includes all years since 1789, the International Disaster Database (Centre for Research on the Epidemiology of Disasters, 2020) and the Mass Mobilization Protest Data (Clark and Regan, 2020) begin their coverage in 1988 and 1990, respectively.

In addition to the 37 variables, models include a vintage identifier (*Vintage ID*) and the difference between the reporting and reported years, $k - t$ (*Time Between Vintage and Year*). There is typically a two-year lag between the reporting year and the most recent reported year. For example, GDP estimates for 2018, 2019, and 2020 were first available in the February 2020, February 2021, and February 2022 WDI releases, respectively. Thus, estimates for year t only enter the analysis at year $t + 2$.

The main models include all predictors for year t. Additional models in Appendix E include each variable twice, both for t and for k, 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. Still, these cases are rare. Since reporting-year characteristics strongly correlate with reported-year characteristics, their inclusion leads to unstable and redundant models with worse fit. Variable importance plots confirm that the *reported* year matters most.

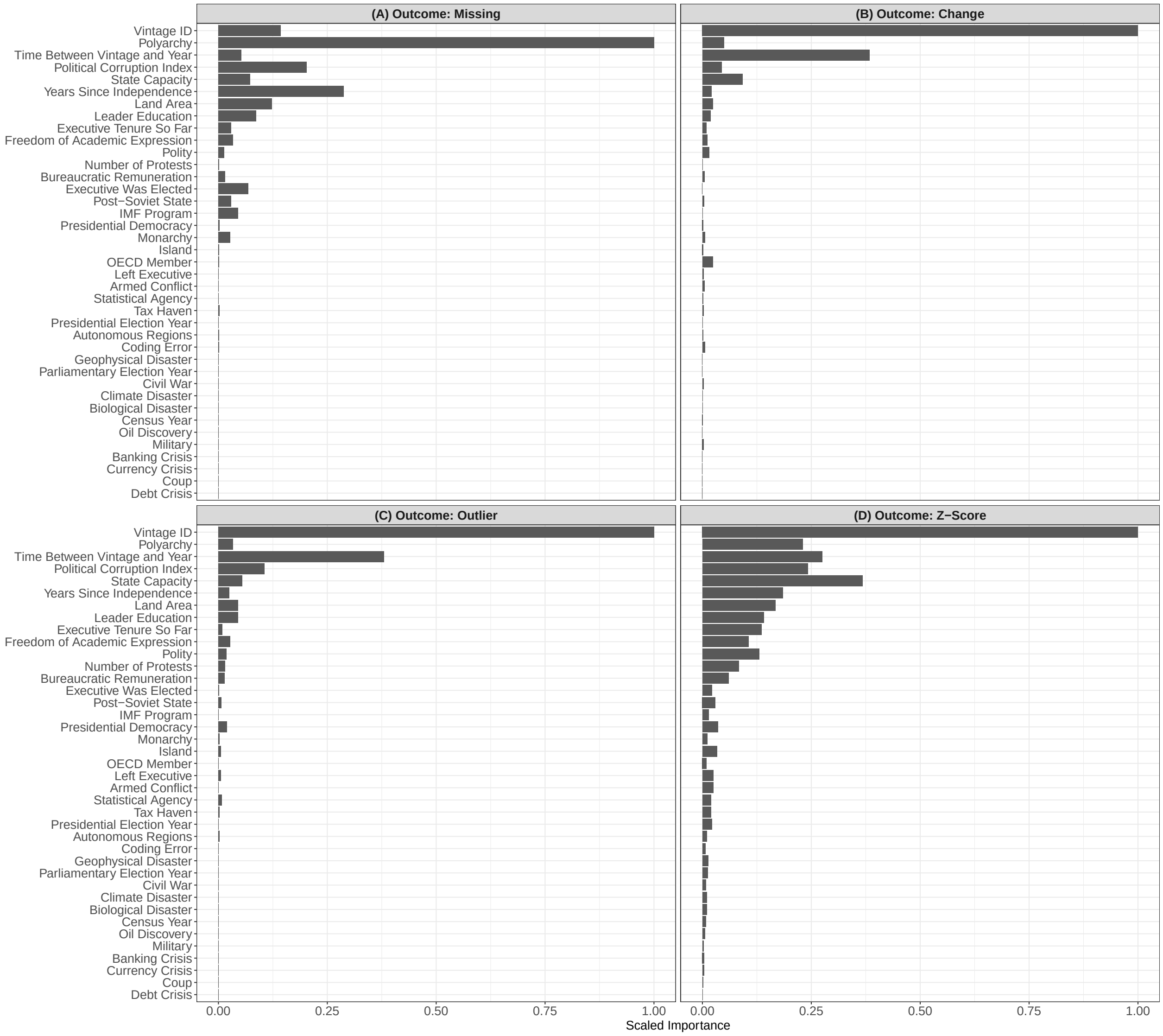
Other robustness checks in Appendix E control for economic and demographic predictors (such as FDI flows, inflation, unemployment, population, or urbanization rates). Several of these predictors are also reported by the WDI, strongly correlate with GDP data, and likely suffer from the same measurement uncertainty.

3.4 Results

Missing, *Change*, and *Outlier* are binary outcomes, and most observations belong in one class: 96.5 percent are *not* missing, 83 percent record *no* change, and 89.5 percent of all non-missing observations are *not* outliers. For each model, I thus follow Muchlinski et al. (2016) and balance the majority and minority classes in the training set. Figure 6 presents the 37 predictors, plus *Vintage ID* and *Time Between Vintage and Year*, ranked by their importance for each outcome.

In Figure 6, Panel (A) reinforces Hollyer, Rosendorff and Vreeland's (2014, 417) finding that WDI data disclosure is a "political decision, not simply a reflection of bureaucratic capacity." The variable that explains the most variation in missingness is, by far, the electoral democracy index *Polyarchy*, ranging from zero (low) to one (high). Data are less likely to be missing for countries with higher Polyarchy scores. Tree-based models do not identify causal relationships, so it is inaccurate to say that regime type *causes* missingness. But this aligns with existing causal evidence that autocrats are less likely to report GDP information (Hollyer, Rosendorff and Vreeland, 2014) and often overstate GDP growth rates (Magee and Doces, 2015; Martínez, 2022). In addition, data are more likely to be missing when the *Political Corruption Index* is high or for newly independent countries, like Timor-Leste, Montenegro, and South Sudan (founded in 2002, 2006, and 2011, respectively), which are still in the process of developing institutions that collect and disseminate high-quality data.

Figure 6: Variable Importance Plot (Training Set)



This figure shows the relative importance of each predictor, by outcome. 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.

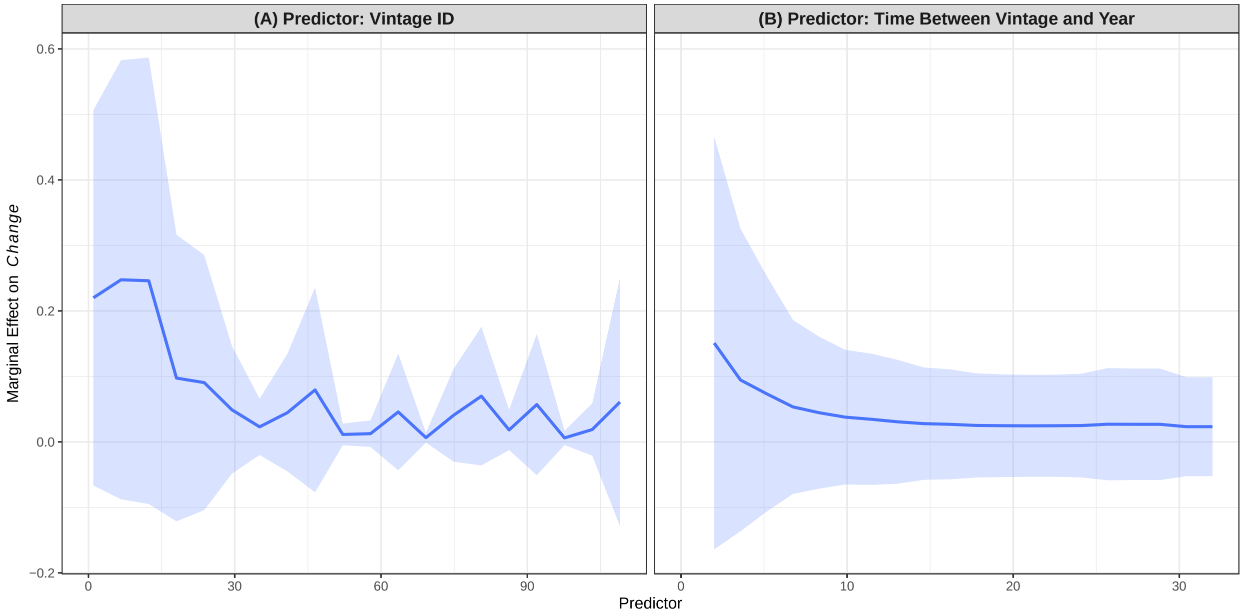
And there is evidence that the WDI backfills data: the larger the gap between reporting and reported year, $k - t$, the less likely estimates will be missing. In contrast, economic crises (banking, debt, or currency), election years, and coups are not strongly associated with variation in coverage and nor is OECD membership, suggesting that wealthier countries are not intrinsically more likely to disseminate data.

Even when data are not missing, the 109 available WDI releases often provide conflicting information. As Panels (B), (C), and (D) show, the other three outcomes — which examine the comparability of data across vintages — share the same two top predictors: *Vintage ID* and *Time Between Vintage and Year*. To illustrate these predictions, Figure 7 presents two partial dependence plots for the second model. These plots show how variation in *Vintage ID* and *Time Between Vintage and Year* relates to variation in the outcome *Change*, without assuming causality. Across all models, *Vintage ID* indicates that the first 30 vintages (from April 1994 to December 2011) are associated with more frequent change and outliers than the subsequent ones, which follow newer ICP benchmarks that include more precise information for developing countries. For example, Zimbabwe's 1990 GDP is more likely to depart from previous values or be an extreme value in the April 1994 WDI than in the April 2014 WDI. *Time Between Vintage and Year* indicates that data become more comparable as $k - t$ increases, regardless of vintage: changes and extreme values become less frequent, whereas the z-score nears zero. Put differently, WDI vintages usually coalesce around one value over time: in the April 2014 WDI, information about Zimbabwe's 2012 GDP is more likely to depart from previous values or be an extreme value than information for Zimbabwe's 1990 GDP. Since researchers do not know Zimbabwe's true GDP in 1990, they cannot say whether more recent WDI vintages are closer to the truth, but they can say that these vintages are less likely to change or report extreme values, instead converging to the mean.

Other than *Vintage ID* and *Time Between Vintage and Year*, the most important predictors of variation in *Change*, *Outlier*, and *Z-Score* are the same: higher state capacity or Polyarchy scores and lower values of the political corruption index are associated with more consistent and comparable GDP data. As before, natural disasters, economic crises, and coups explain practically no variation in the outcome of interest.

In sum, data coverage (measured as *Missing*) is best predicted by a mix political will and state capacity. Withholding data is often a political choice: data coverage tends to decline in contexts of authoritarianism and widespread corruption. Beyond political choices,

Figure 7: Partial Dependence Plots, Outcome: Change (Training Set)



These partial dependence plots illustrate the relationship between a specific predictor — *Vintage ID* in Panel (A), *Time Between Vintage and Year* in Panel (B) — and the predicted outcome *Change* while holding all other predictors constant.

low state capacity and recent independence pose a hard barrier to a country's ability to disseminate information. In contrast, data comparability (measured as *Change*, *Outlier*, and *Z-Score*) is less tied to country-specific characteristics. Intuitively, one would expect low-income countries to produce less comparable data. But OECD membership is not a strong predictor of variation in any outcome of interest, suggesting that different measurements of high-income countries are just as likely to be comparable. Rather, low data comparability is a technical problem best predicted by changes in the ICP standardization framework or simply the passage of time. Individual measurements become more reliable as they age, though it is impossible to say whether they get any closer to the truth.

3.5 Assessing Model Performance

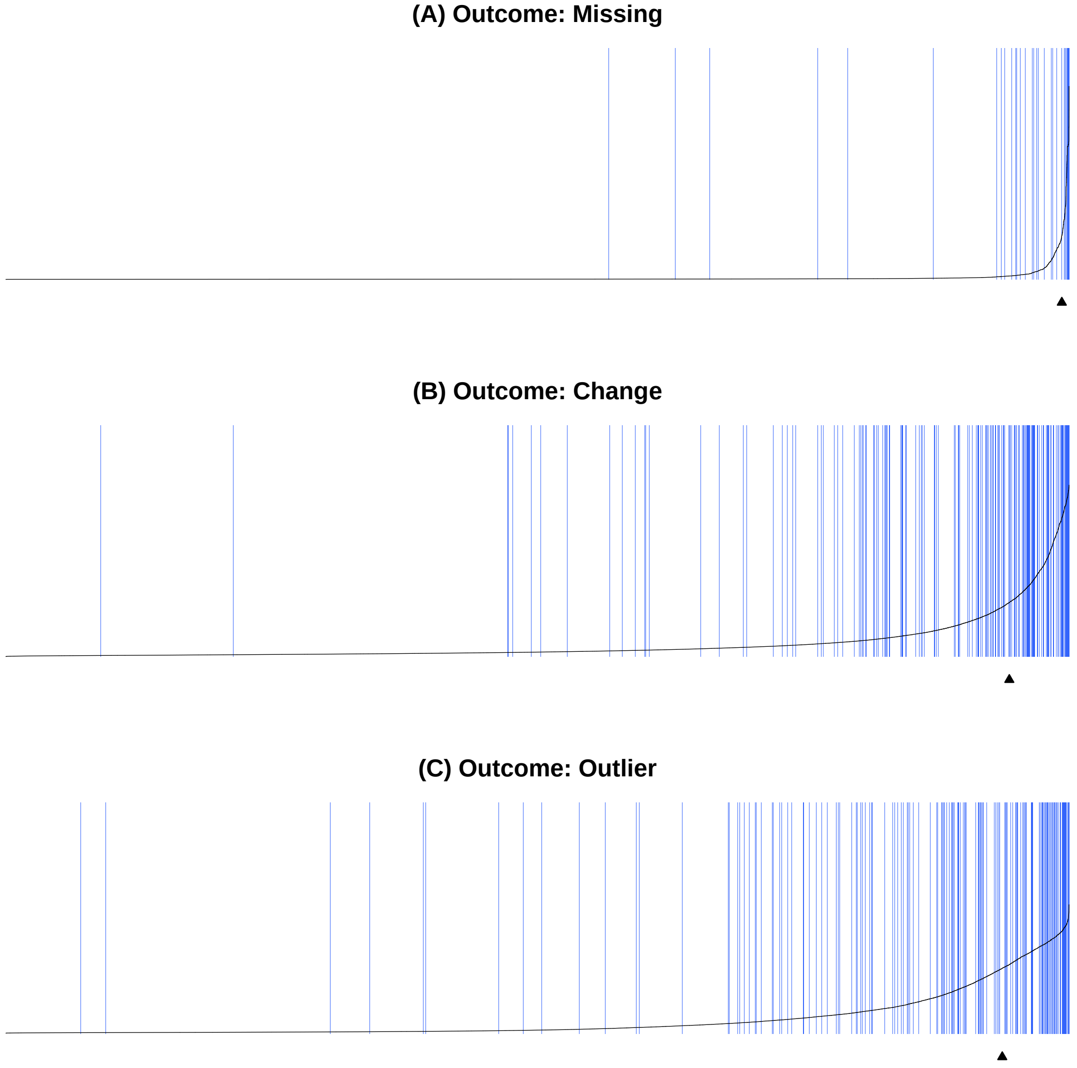
My primary goal is to uncover relationships between variables, not optimize predictive power, but I assess the quality of the out-of-sample predictions as a final step. For the three binary variables, I follow Muchlinski et al. (2016) in presenting separation plots and Receiver Operating Characteristic (ROC) curves for the test sets. Three separation plots in Figure 8 organize the predicted probabilities for each observation in ascending order, highlighting whether each observation corresponds to an actual event: missingness, change, or outlier (Greenhill, Ward and Sacks, 2011). These plots also provide information about the predicted probabilities (a line) and the expected number of events (a triangle). When the model makes perfect predictions, the plot showcases a clear separation between the zeroes and ones: lower probabilities are always associated with no event (left of the triangle) and higher probabilities 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.

While the model predicting coverage has the best fit, some events cannot be predicted on a systematic basis. As an illustration, consider New Zealand's 2012 GDP, which enters the analysis in 2014. 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 first model is unable to predict; the second model is similarly unable to predict the resulting changes.[8]

Figure 9 presents ROC curves for the three models with binary outcomes. In each panel, the y-axis represents the true positive rate (the proportion of missing observations that are correctly classified as missing), whereas the x-axis represents the false positive rate (the proportion of non-missing observations that are incorrectly classified as missing). 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

[8] Random forests, LASSO, and ridge regressions face similar issues (see Appendix E).

Figure 8: Separation Plots (Test Set)



(A) Outcome: Missing

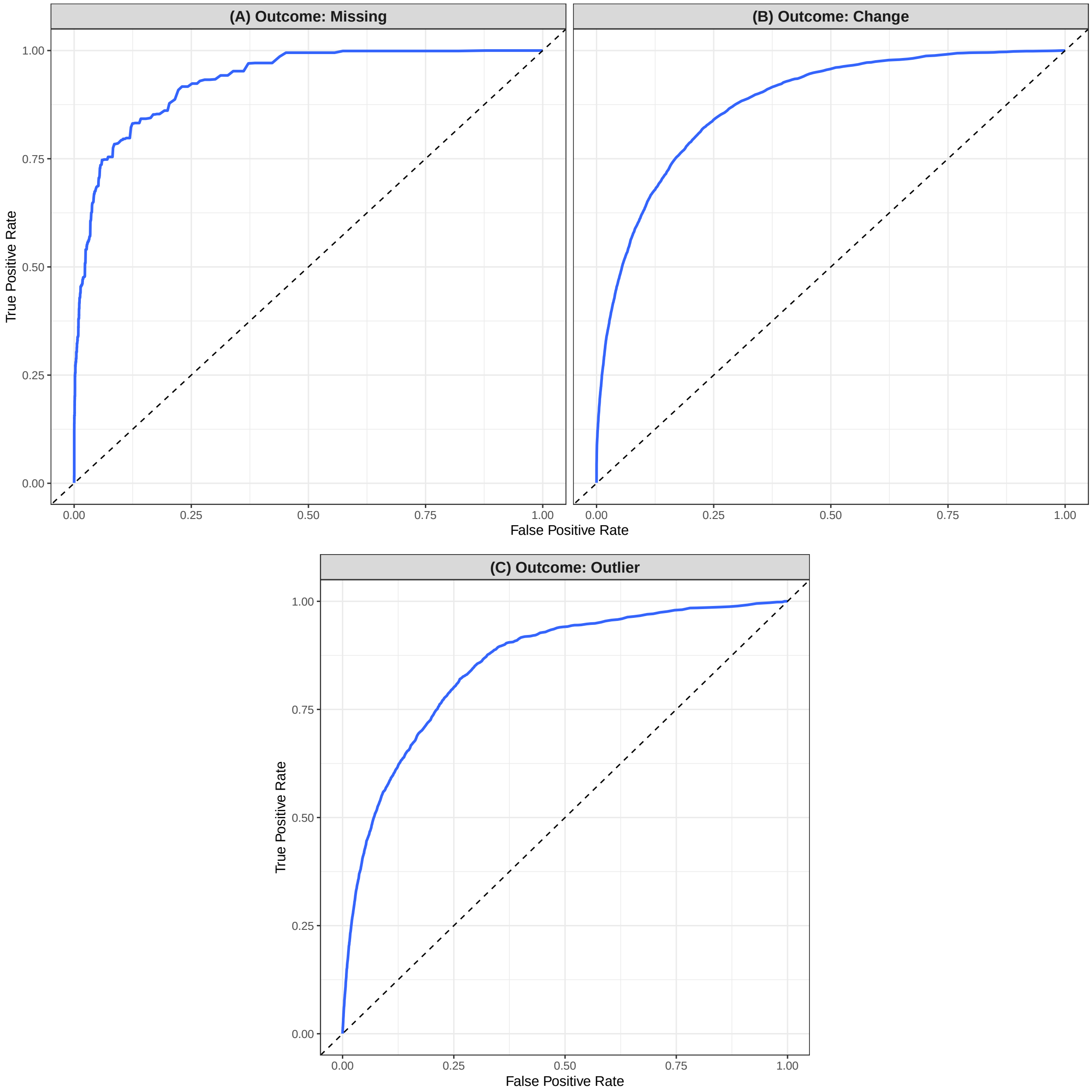
(B) Outcome: Change

(C) Outcome: Outlier

Separation plots organize the predicted probabilities for each observation in ascending order, highlighting whether each observation corresponds to an actual event: (A) missingness, (B) change, or (C) outlier. Separation plots also provide information about the predicted probabilities (a line) and the expected number of events (a triangle). If the model makes perfect predictions, the plot will showcase a clear separation between the zeroes and ones: lower probabilities (in white) will always be associated with no event (left of the triangle) and higher probabilities (in blue) will always be associated with an event (right of the triangle).

0.5 denoting random guessing and 1 denoting a perfect classifier. The AUC values (ranging from 0.854 to 0.932) indicate that all three models make good out-of-sample predictions: they can typically distinguish between true positives and false positives, between observations

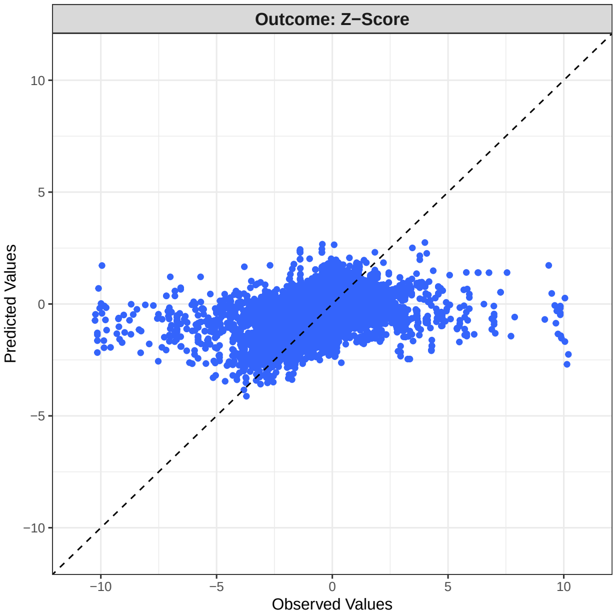
Figure 9: Receiver Operating Characteristic Curves (Test Set)



These Receiver Operating Characteristic (ROC) curves illustrate the trade-off between the true positive rate and the false positive rate across different probability thresholds. In each curve, the y-axis represents the true positive rate (the proportion of missing observations that are correctly classified as missing), whereas the x-axis represents the false positive rate (the proportion of non-missing observations that are incorrectly classified as missing). 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 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. For the ROC curves above, the corresponding AUC values are (A) 0.932, (B) 0.878, and (C) 0.854, indicating that all models make good out-of-sample predictions.

that are truly missing and observations that are not. In Appendix D, I present additional performance metrics confirming that these models do a good — if not perfect — job of predicting missingness, change, and outliers.

Figure 10: Predicted Versus Observed Values (Test Set)



This figure plots the observed values on the x-axis against the predicted values on the y-axis. Each point represents an observation, and the diagonal line represents perfect predictions. The closer the points are to the diagonal line, the better the model's predictions align with the actual values.

Since the fourth model has a continuous outcome, I use a different metric to assess its performance. The R^2 indicates the correlation between predicted and observed values, from 0 (no correlation) to 1 (complete correlation). The R^2 for the test set is 0.221: only 22.1 percent of the out-of-sample variation in z-scores can be systematically explained. To better grasp this statistic, Figure 10 plots the observed values on the x-axis against the predicted values on the y-axis. Each point represents an observation, and the diagonal line represents perfect predictions. The closer the points are to the diagonal line, the better the model's predictions align with the actual values. The model consistently makes predictions that are up to two standard deviations above or below the mean, an accurate prediction for 96 percent of all non-missing observations. Models 3 and 4 jointly indicate that 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).

4 Conclusions

Political scientists have long debated how to measure abstract concepts like democracy (Munck and Verkuilen, 2002; Giannone, 2010; Coppedge and Gerring, 2011) without devoting as much attention to the measurement of seemingly concrete concepts. GDP is generally considered a valid and reliable measure of national wealth, but national accounts data are not fixed data points: they are preliminary estimates that are constantly revised, and revisions might provide conflicting information. GDP is the foundation to calculate a number of variables in social science research, including foreign aid, foreign direct investment, and trade flows. Even when GDP is "just" a control variable, 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 when missingness cannot be predicted by observed or unobserved factors; multiple imputation also provides unbiased estimates when missingness can be predicted by observed factors. But missingness in GDP data cannot be fully predicted by observed factors, as I showed. Researchers do not know the true data generating process underlying national accounts data; there are significant regional disparities in statistical capacity, and even where statistical capacity is high, there might be political interest in reporting biased (or no) data. GDP is missing not at random, so both multiple imputation and listwise deletion would be biased (Pepinsky, 2018). The same applies to other data issues: a GDP of zero for the Democratic Republic of the Congo is clearly wrong, but deleting such observation would generate bias.

There are three straightforward ways to address the potential influence of discrete errors. The first 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. Someone working with the December 2021 WDI might not be aware of the extreme values for Myanmar but will at least confirm that their empirical results are not driven by such outliers.

Second, just as it is common to present robustness checks with alternative measures of regime type (like Polity or Polyarchy), researchers can estimate separate models with alternative measures of GDP, exports, foreign aid, foreign direct investment, or economic growth from different sources and vintages. To facilitate this, 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. Alternative measures are particularly relevant for studies focusing on non-democracies, recently independent countries, and settings with high corruption and low state capacity. Under these circumstances, it is safe to assume that national accounts data are flawed: observations are more likely to be inconsistent or missing.

As Herrera and Kapur (2007, 381) state, “the penalties for using low-quality data are small.” But at the very minimum, 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 ICP rounds, provide more precise information for developing countries and are more consistent, as my analysis shows. 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 and will likely 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.

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.

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

Centre for Research on the Epidemiology of Disasters. 2020. *EM-DAT: The International*

Disaster Database.
URL: *https://public.emdat.be*

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

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.*

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.

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.

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.

Gerring, John. 2012. *Social Science Methodology. A Unified Framework*. Cambridge: Cambridge University Press.

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

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

Greenhill, Brian, Michael D. Ward and Audrey Sacks. 2011. "The Separation Plot: A New Visual Method for Evaluating the Fit of Binary Models." *American Journal of Political Science* 55(4):991–1002.

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

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.

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.

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.

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

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.

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).

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
Measurement Uncertainty in International Statistics

Iasmin Goes*

January 2024

Contents

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

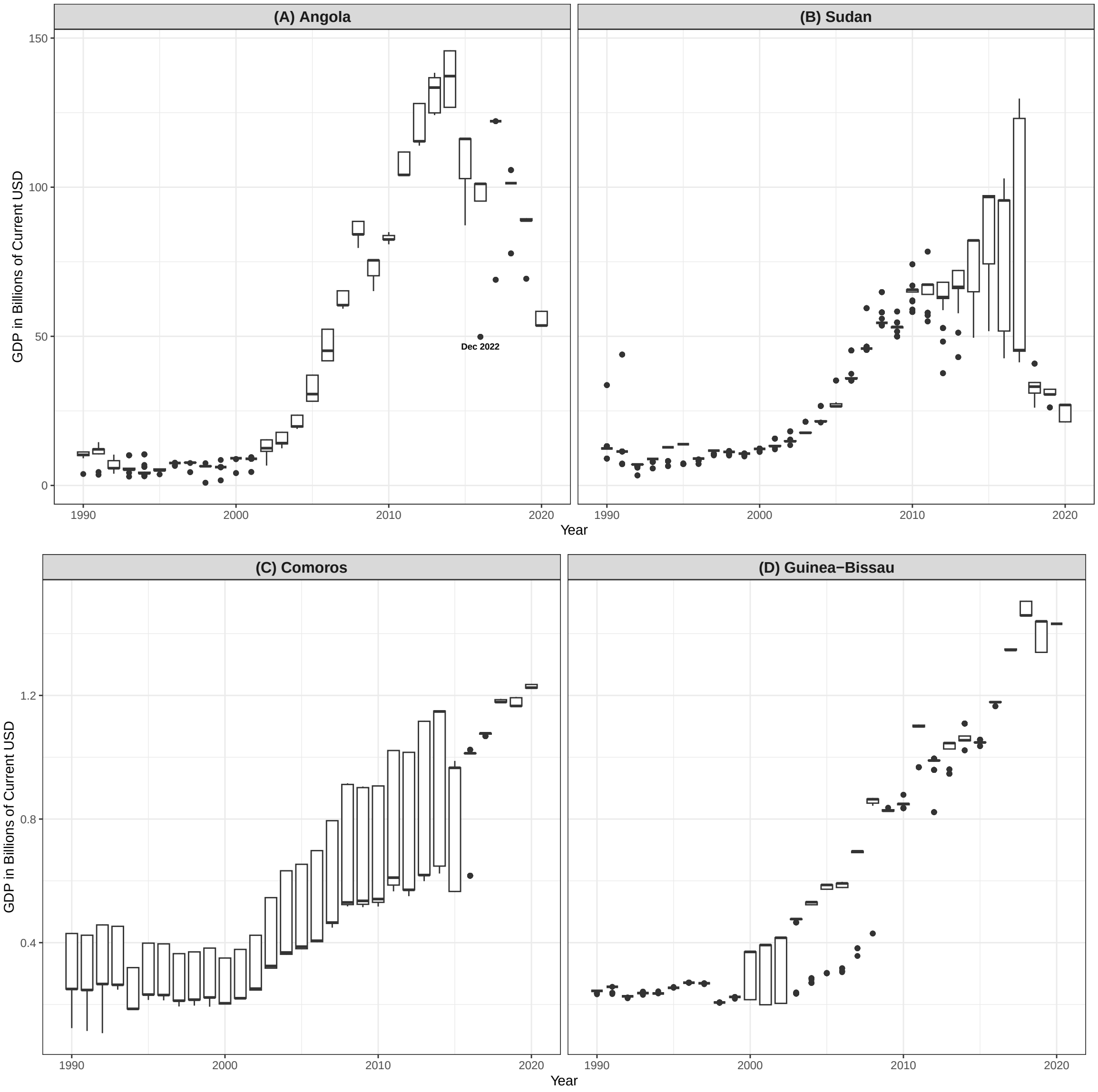
A Countries Included in the Analysis

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

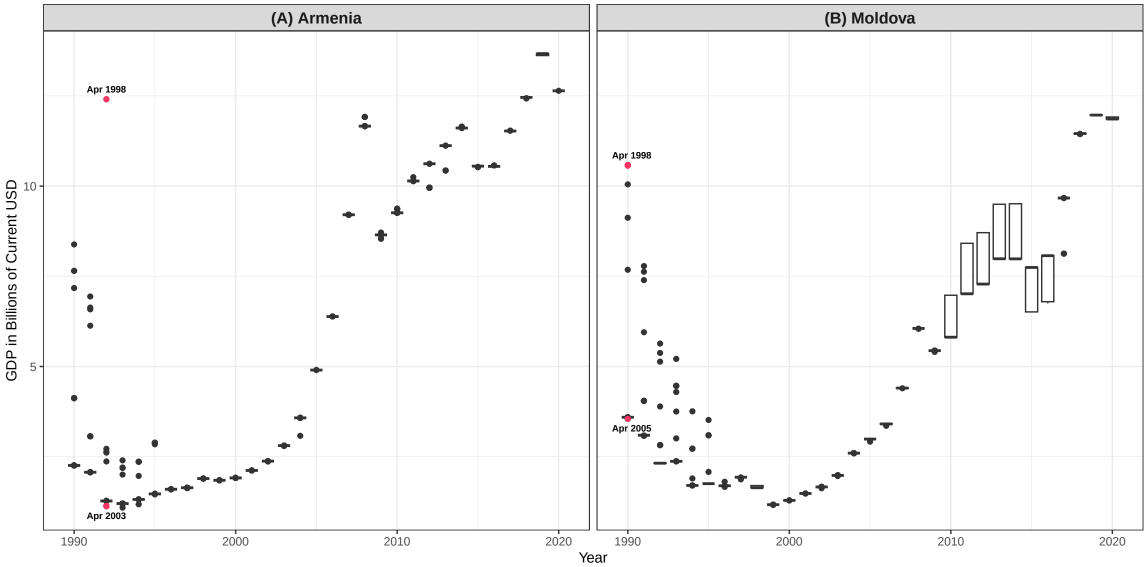
To give readers a clearer grasp of the measurement uncertainty in the data, Figure B.1 presents GDP data for two large African oil producers (Angola and Sudan) and two small African economies (Comoros and Guinea-Bissau), whereas Figure B.2 does the same for two former Soviet republics (Armenia and Moldova). More generally, Figure B.3 shows the distribution of the WDI variable *GDP in constant US dollars* (ID `NY.GDP.MKTP.KD`), which is used to generate four outcomes: *Missing*, *Change*, *Outlier*, and *Z-Score*.

Figure B.1: Current GDP of Angola, Sudan, Comoros, and Guinea-Bissau, 1990–2020



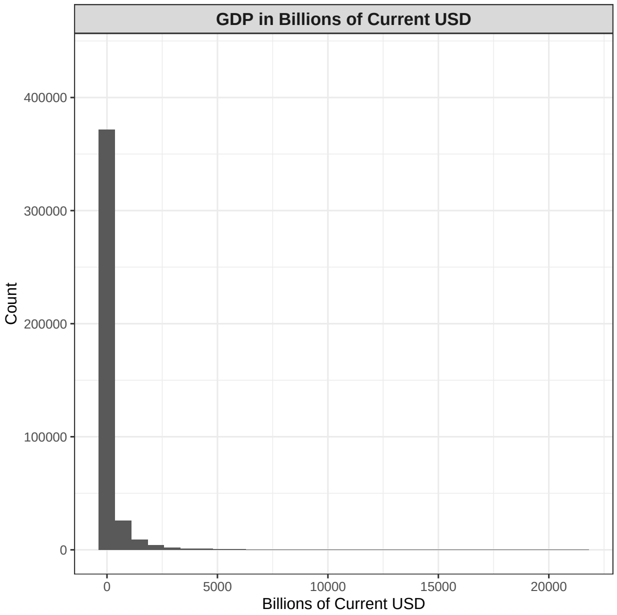
These boxplots present the distribution of current GDP estimates for (A) Angola, (B) Sudan, (C) Comoros, and (D) Guinea-Bissau from 1990 to 2020, using data drawn from the 109 WDI releases from April 1994 to December 2022. Section 3 discusses the data in more detail.

Figure B.2: Current GDP of Armenia and Moldova, 1990–2020



These boxplots present the distribution of current GDP estimates for (A) Armenia and (B) Moldova from 1990 to 2020, using data drawn from the 109 WDI releases from April 1994 to December 2022. Section 3 discusses the data in more detail.

Figure B.3: Distribution of the WDI Variable *GDP in Constant US Dollars*



This histogram shows the distribution of the WDI variable *GDP in constant US dollars* (ID `NY.GDP.MKTP.KD`), which is used to generate four outcomes: *Missing*, *Change*, *Outlier*, and *Z-Score*.

C List of Predictors

In addition to *Time Between Vintage and Year*, the main analysis includes 37 predictors, listed in Table C.1 (along with their respective description, coverage, and source). In robustness checks (see Appendix E.3), I include additional economic and demographic predictors (listed in Table C.2) 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 July 2023. I downloaded all WDI data using Vincent Arel-Bundock's `WDI` package for `R`. As Figure C.1 shows, 3.7 percent of the data are missing.

Table C.1: Main Predictors

Variable	Description	Coverage	Source
Armed Conflict	Was any armed conflict recorded? (yes = 1)	1939–2021	Gleditsch et al. (2002); Pettersson et al. (2021)
Autonomous Regions	Are there autonomous regions? (yes = 1)	1970–2020	Cruz, Keefer and Scartascini (2021)
Banking Crisis	Was there a banking crisis this year? (yes = 1)	1970–2017	Laeven and Valencia (2020)
Biological Disaster	Occurrence of a biological (epidemic) disaster (yes = 1)	1988–2021	Centre for Research on the Epidemiology of Disasters (2020)
Bureaucratic Remuneration	To what extent are state administrators salaried employees? (none = 0, small share = 1, half = 2, substantial number = 3, all = 4)	1789–2022	Coppedge et al. (2023)
Census Year	Was there a national census in this year? (yes = 1)	1789–2020	Coppedge et al. (2023)
Civil War	Was there a civil war this year? (yes = 1)	1946–2018	Marshall (2019)
Climate Disaster	Occurrence of a climatological (drought, wildfire), meteorological (storm, extreme temperature), or hydrological (flood, landslide) disaster (yes = 1)	1988–2021	Centre for Research on the Epidemiology of Disasters (2020)
Coup	Did a coup d'etat occur? (yes = 1)	1789–2020	Coppedge et al. (2023)

Coding Error	Coded 1 for all Myanmar observations in the December 2021 WDI as well as for the Democratic Republic of the Congo–1990 observation in the July, August, October, and November 2016 WDI	1990–2022	Own Coding
Currency Crisis	Was there a currency crisis this year? (yes = 1)	1970–2017	Laeven and Valencia (2020)
Debt Crisis	Was there a debt crisis this year? (yes = 1)	1970–2017	Laeven and Valencia (2020)
Executive Tenure So Far (Years)	Number of years that a leader has been in power during their current tenure period	1950–2020	Bell, Besaw and Frank (2021)
Executive Was Elected	Executive leader was elected to office (yes = 1)	1950–2020	Bell, Besaw and Frank (2021)
Freedom of Academic Expression	Is there academic freedom and freedom of cultural expression related to political issues? (yes = 1)	1789–2022	Coppedge et al. (2023)
Geophysical Disaster	Occurrence of a geophysical (earthquake, volcanic activity) disaster (yes = 1)	1988–2021	Centre for Research on the Epidemiology of Disasters (2020)
IMF Program	Participation in an IMF program (yes = 1)	1978–2022	Kentikelenis, Stubbs and King (2016), IMF MONA Database
Island	Is the country an island? (yes = 1)	1990–2020	Own coding
Land Area	Land area (sq. km)	1961–2020	WDI
Leader Education	Leader's level of education summarized in eight categories	1948–2020	Dreher et al. (2020)
Left Executive	Party orientation of the executive with respect to economic policy (left = 1)	1970–2020	Cruz, Keefer and Scartascini (2021)
Military	Direct or indirect military regime (yes = 1)	1950–2020	Bell, Besaw and Frank (2021)
Monarchy	Monarchy (yes = 1)	1950–2020	Bell, Besaw and Frank (2021)
Number of Protests	Number of recorded protests	1990–2020	Clark and Regan (2020)
OECD Membership	Membership in the Organization for Economic Co-Operation and Development	1950–2020	Dreher et al. (2022)

Oil Discovery	Discovery of a giant, megagiant, or supergiant oil or gas field (yes = 1)	1868–2020	Horn (2014); Cust, Mihalyi and Rivera-Ballesteros (2021)
Parliamentary Election Year	Did a legislative or constituent assembly election take place? (yes = 1)	1789–2022	For Brunei and Belize, Cruz, Keefer and Scartascini (2021); for all other countries, Coppedge et al. (2023)
Political Corruption Index	How pervasive is political corruption? (low = 0, high = 1, on an interval scale)	1789–2022	Coppedge et al. (2023)
Polity	Revised combined Polity score, from –10 (hereditary monarchy) to +10 (consolidated democracy)	1800–2018	Marshall and Gurr (2020)
Polyarchy	Electoral democracy index	1789–2022	Coppedge et al. (2023)
Post-Soviet State	Former Republic of the Union of Soviet Socialist Republics	1990–2022	Own coding
Presidential Democracy	Presidential democracy (yes = 1)	1950–2020	Bell, Besaw and Frank (2021)
Presidential Election Year	Did a presidential election take place? (yes = 1)	1789–2022	For Brunei and Belize, Cruz, Keefer and Scartascini (2021); for all other countries, Coppedge et al. (2023)
State Capacity	Estimate of state capacity by Hanson/Sigman	1960–2015	Hanson and Sigman (2021)
Statistical Agency	Is there a national statistical agency? (yes = 1)	1789–2022	Coppedge et al. (2023)
Tax Haven	Is this state considered a tax haven? (yes = 1)	1983–2020	US Department of Treasury, via Graham et al. (2018); Graham and Tucker (2019)
Years Since Independence	Most recent date of foundation, independence or reunification	751–2022	Own coding

Table C.2: Additional Predictors

Variable	Description	Coverage	Source
Agriculture (% GDP)	GDP, share of value added by kind of economic activity: agriculture, hunting, forestry, fishing	1970–2021	UNCTAD
Central Government Debt (% GDP)	Central government debt, share of GDP	1950–2020	IMF
Diversification Index	Merchandise: product diversification index of exports	1995–2021	UNCTAD
Fertility Rate	Fertility rate, total (births per woman)	1960–2021	WDI
Foreign Aid	Net official development assistance and official aid received (current US dollars)	1960–2021	WDI
Imports (% GDP)	Imports of goods and services, share of GDP	1970–2021	UNCTAD
Income Share Top 10%	Share of pre-tax national income held by the top 10%	1820–2021	World Inequality Database
Industry (% GDP)	GDP, share of value added by kind of economic activity: industry	1970–2021	UNCTAD
Inflation	Inflation, consumer prices (annual %)	1960–2022	WDI
Inward FDI, Flows (% GDP)	Inward foreign direct investment flows, share of GDP	1970–2021	UNCTAD
Inward FDI, Stock (% GDP)	Inward foreign direct investment stock, share of GDP	1970–2021	UNCTAD
KAOPEN	Normalized Chinn-Ito index, ranging from zero to one	1970–2020	Chinn and Ito (2006)
Military Expenditure Per Capita	Military expenditure per capita, in current US dollars	1949–2022	SIPRI Military Expenditure Database
Service (% GDP)	GDP, share of value added by kind of economic activity: service	1970–2021	UNCTAD
Tax Revenue (% GDP)	Total tax revenue, excluding social security contributions, share of GDP	1980–2021	Government Revenue Dataset
Total Population	Population, total	1960–2022	WDI

Unemployment	Unemployment (% of total labor force), modeled ILO estimate	1991–2022	WDI
Urban Population	Urban population (% of total population)	1960–2022	WDI

Figure C.1: Missingness Map: Predictors, 1990–2020

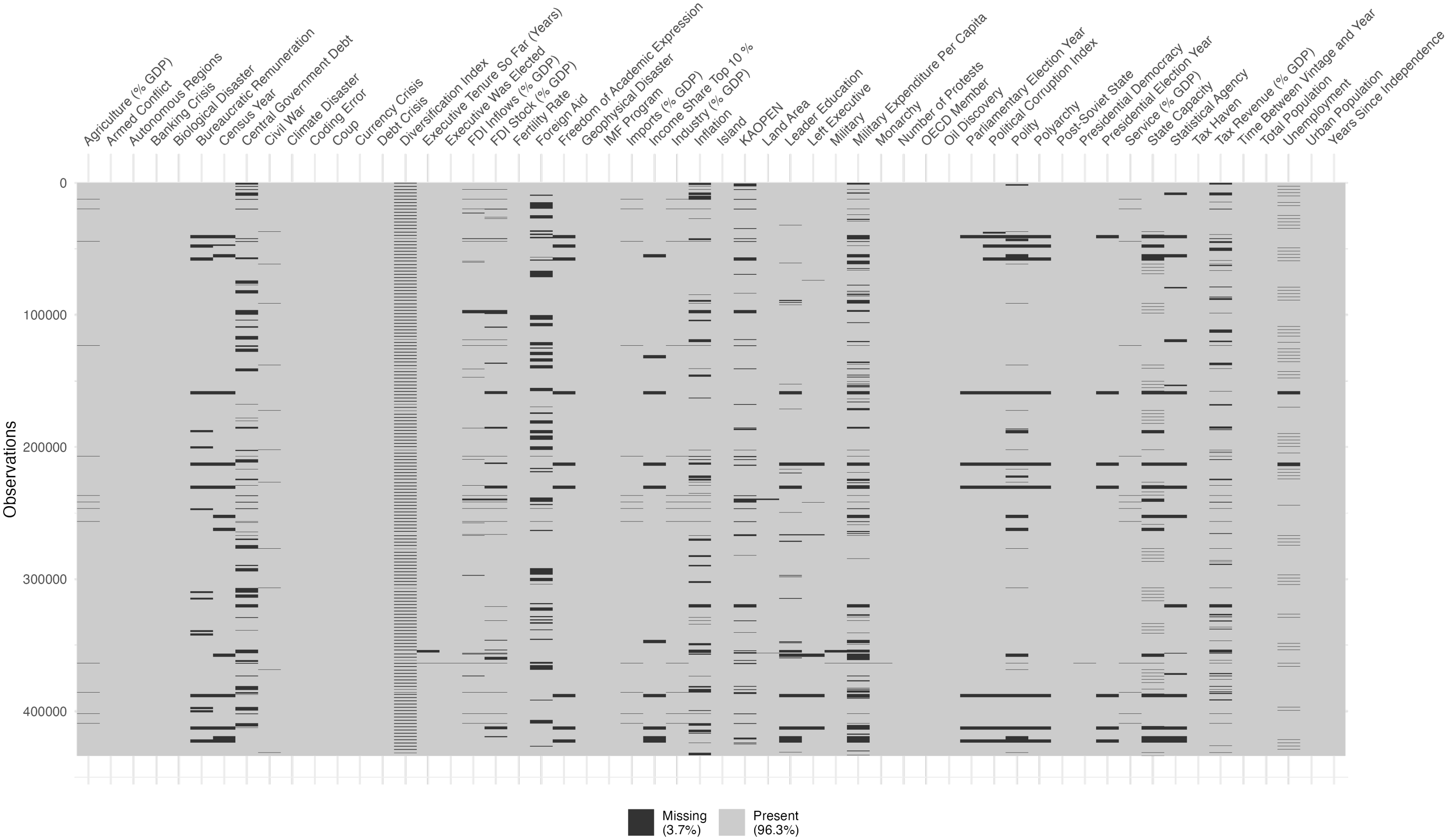
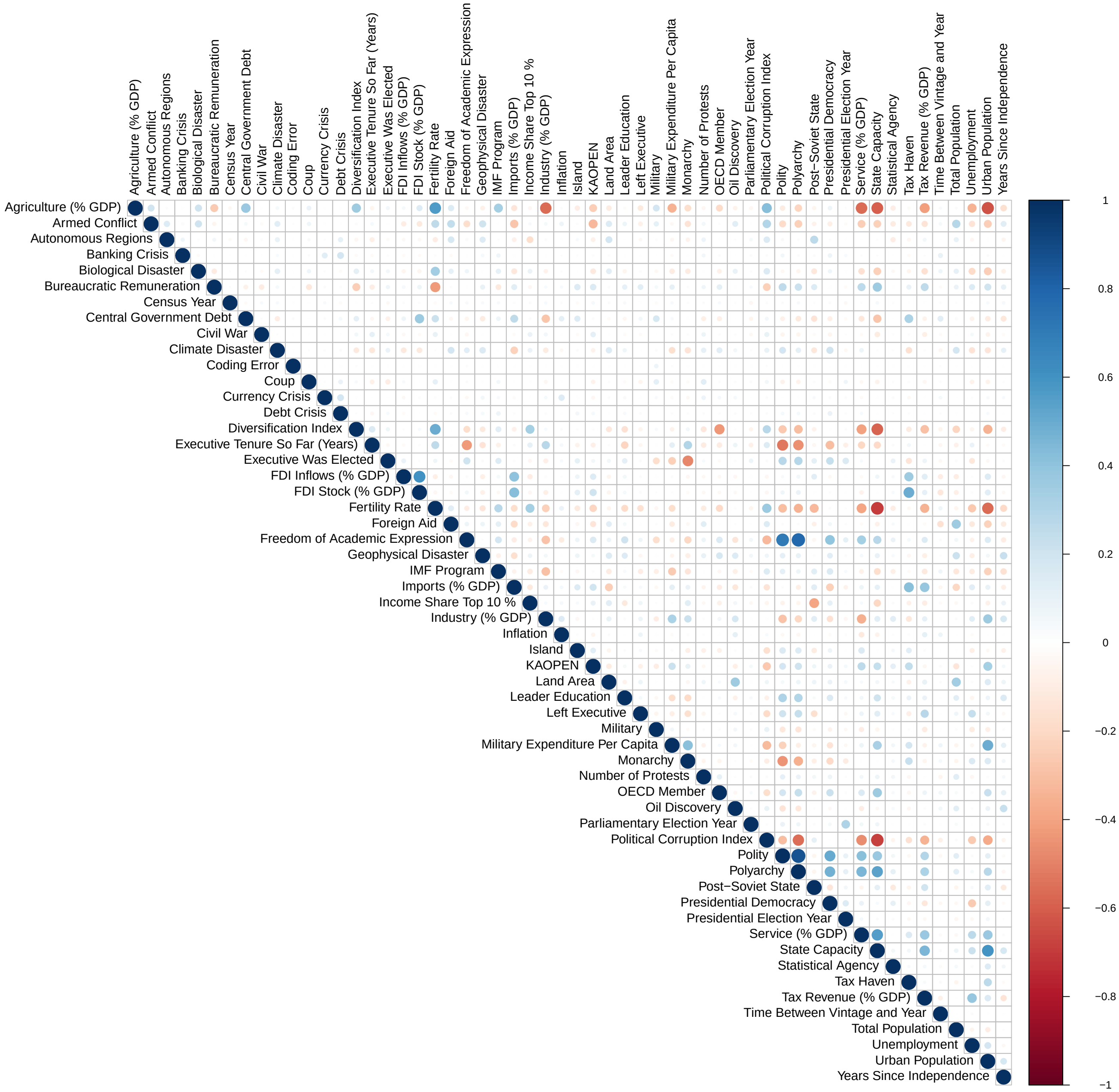


Figure C.2: Correlation Matrix: Predictors, 1990–2020



D Main Models: Diagnostics

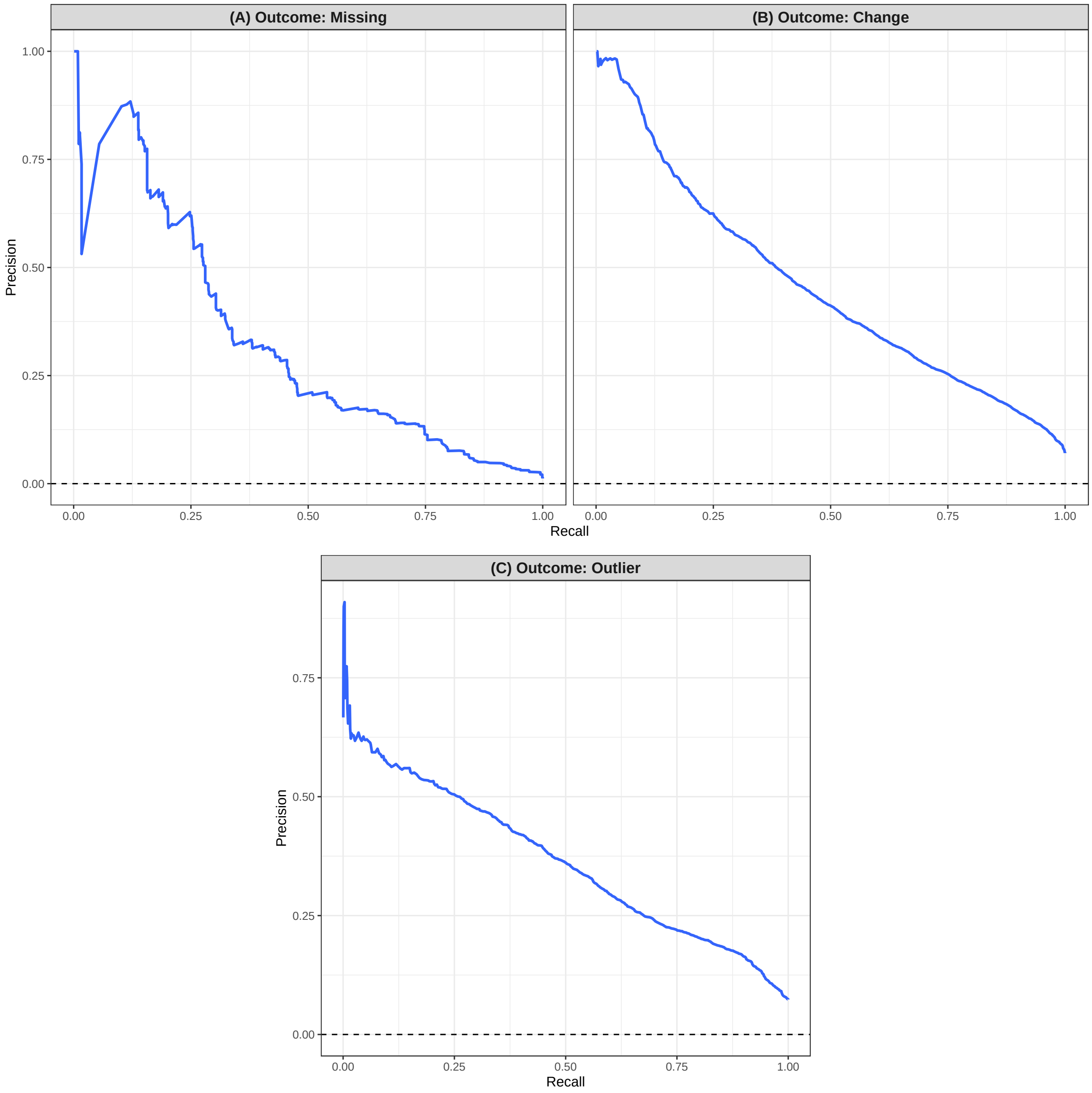
Table D.1: Performance Statistics, Main Models

	Training	Validation	Test
Outcome: Missingness			
AUC	0.9978304	0.7662927	0.9317631
AUCPR	0.9976250	0.2921954	0.3297605
Outcome: Change			
AUC	0.9293748	0.8933057	0.8783696
AUCPR	0.9188185	0.4705715	0.4557725
Outcome: Outlier			
AUC	0.7401503	0.7441611	0.7255345
AUCPR	0.2015761	0.2042317	0.172373
Outcome: Z-Score			
R^2	0.7123624	0.3019717	0.2208295
MSE	0.2561792	0.6169476	0.6936291

Table D.1 presents common performance metrics for the main models. I begin by discussing the three models with binary outcomes (*Missing*, *Change*, and *Outlier*). 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. Recall that the outcomes are very imbalanced: most observations are *not* missing, do *not* change from one vintage to another, and are *not* outliers. To address this issue, thee majority and minority classes are balanced 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 address overfitting concerns, all models use early stopping and iterative tuning of hyperparameters.

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. The AUCPR for all test sets in Table D.1 is 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

Figure D.1: Precision-Recall Curves, Main Models



These panels present Precision-Recall (PR) curves for the test set. In each panel, the y-axis represents the precision, which is the proportion of missing observations (true positives) the model correctly identified from all the observations it labeled as missing (true positives plus false positives). The x-axis represents the recall, which is the proportion of missing observations (true positives) the model correctly identified from all the actual missing cases (true positives plus false negatives). 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.

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 (3.5, 9.9, and 9.3 percent, respectively), a model exhibiting substantial discrimination capability is promising. To illustrate this, Figure D.1 plots the AUCPR values for the three classification tasks; 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.

Turning to the fourth model, I use two different performance metrics because the outcome (the z-score) is continuous. The R^2 indicates the correlation between predicted and observed values, from 0 (no correlation) to 1 (complete correlation). According to Table D.1, the explains 71.2 percent of the variation in the training set, but only 22.1 percent of the variation in the test set. A second statistic, the Mean Squared Error (MSE), 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 0.256, but increases to 0.694 in the test set. These discrepancies partly reflect the fact that GBMs tend to overfit on the training data in the presence of outliers.

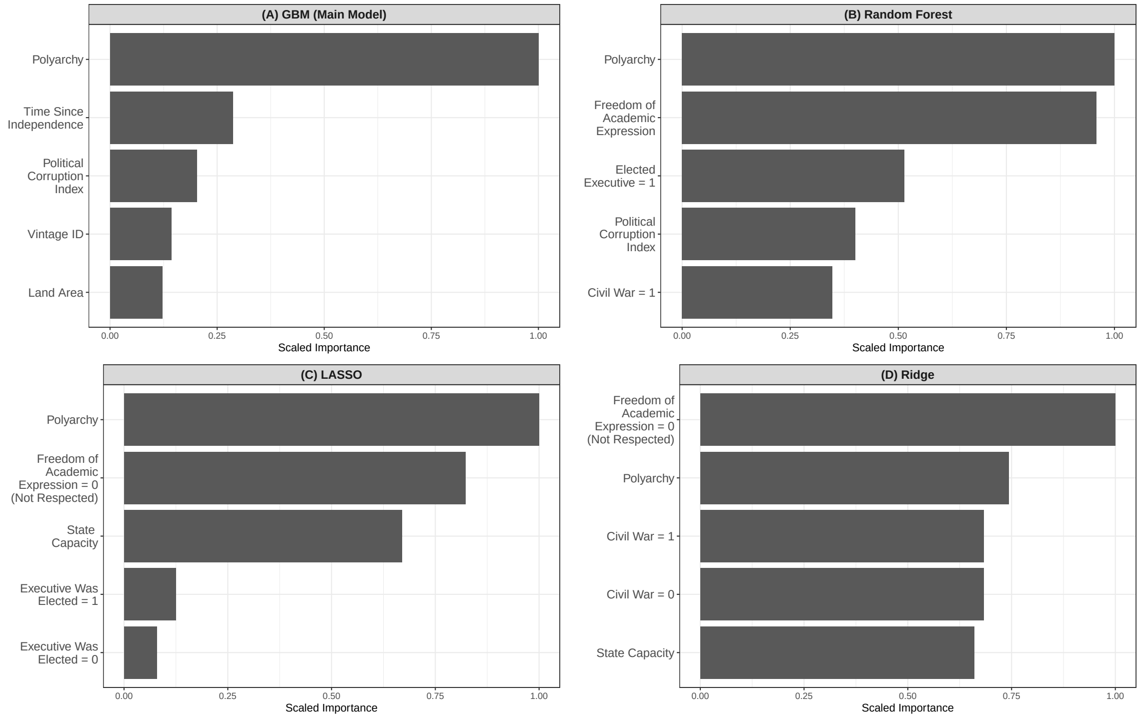
E Robustness Checks

E.1 Alternative Models: Random Forests and Generalized Linear Models

To predict each outcome, I estimate not only a GBM but also a second tree based model — a random forest — as well as two penalized generalized linear models (GLM) — least absolute shrinkage and selection operator (LASSO) and ridge regression. I begin by examining how different models predict the outcome *Missing*. Figure E.1 presents the variable importance plots for all models predicting missingness. LASSO and ridge do not drop the baseline category of a categorical variable, as a traditional regression would do to avoid multicollinearity; this is why the two bottom panels in Panel (C) include specific levels of the variable *Executive Was Elected*, for example. 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. For this specific model, only the five variables displayed in Panel (C) 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.

Figure E.2 and Table E.1 show that a GBM makes better out-of-sample predictions than a random forest, though not by much. 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

Figure E.1: Predicting Missingness: Variable Importance Plot, All Models (Training Set)



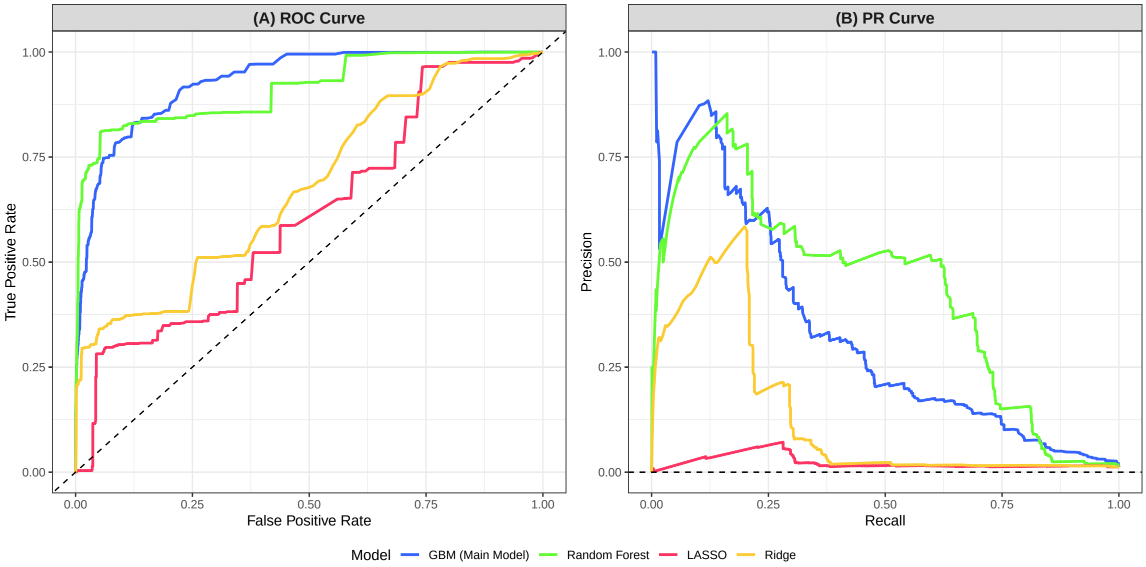
These panels show the relative importance of the top five variables, by model. The least important variable equals zero, while the most important variable equals one. LASSO adds a penalty to the absolute values of the coefficients (L1 regularization) that encourages most coefficients — like that of all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

Table E.1: Predicting Missingness: Performance Statistics, All Models

	Training	Validation	Test
GBM (Main Model)			
AUC	0.9978304	0.7662927	0.9317631
AUCPR	0.9976250	0.2921954	0.3297605
Random Forest			
AUC	0.9998257	0.6969753	0.9125035
AUCPR	0.9188185	0.1691374	0.4266687
LASSO			
AUC	0.8785639	0.6440872	0.6104736
AUCPR	0.4569622	0.1046754	0.02249346
Ridge			
AUC	0.9144365	0.5869146	0.6799721
AUCPR	0.6919903	0.0565727	0.1277197

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.

Figure E.2: Predicting Missingness: ROC and PR Curves, All Models (Test Set)



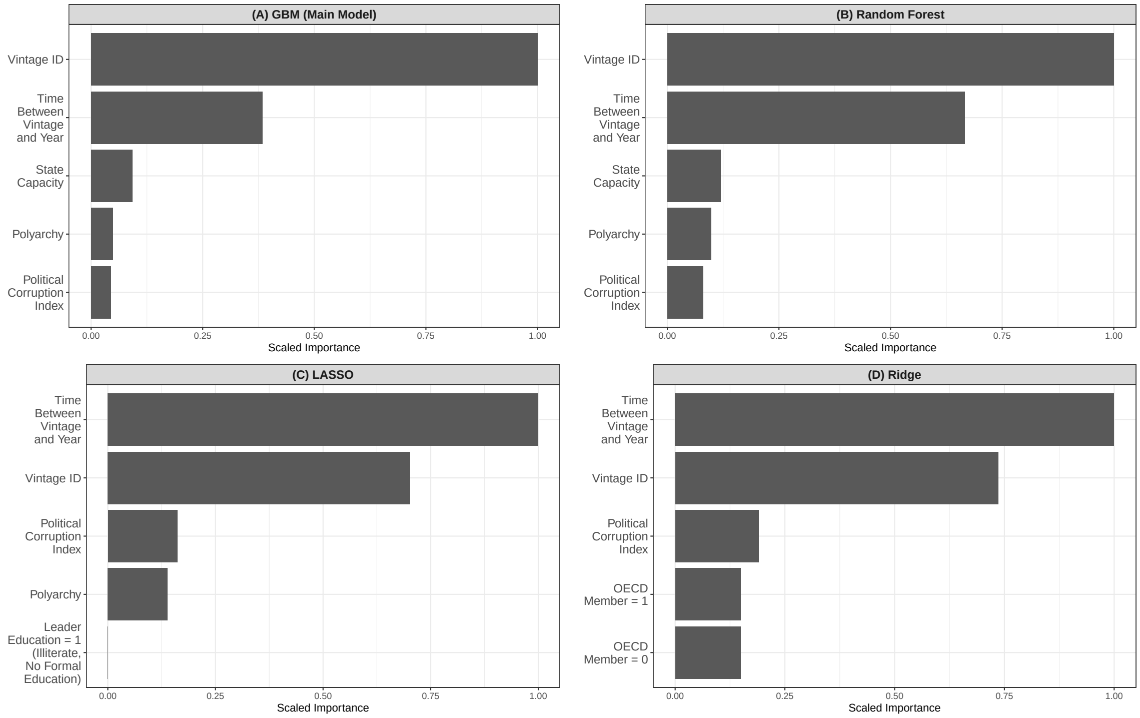
Panel (A) presents a Receiver Operating Characteristic (ROC) curve for the the test set using four models: GBM (presented in the main text), random forest, LASSO, and ridge regression. Similarly, Panel (B) presents a Precision-Recall (PR) curve, also for the test set and also using the same four models.

Below, I present similar results for the other outcomes. Figure E.3, Table E.2, and Figure E.4 confirm that a GBM makes better out-of-sample predictions for *Change*, whereas Figure E.5, Table E.3, and Figure E.6 support the same conclusion for the outcome *Outlier*.

Table E.2: Predicting Change: Performance Statistics, All Models

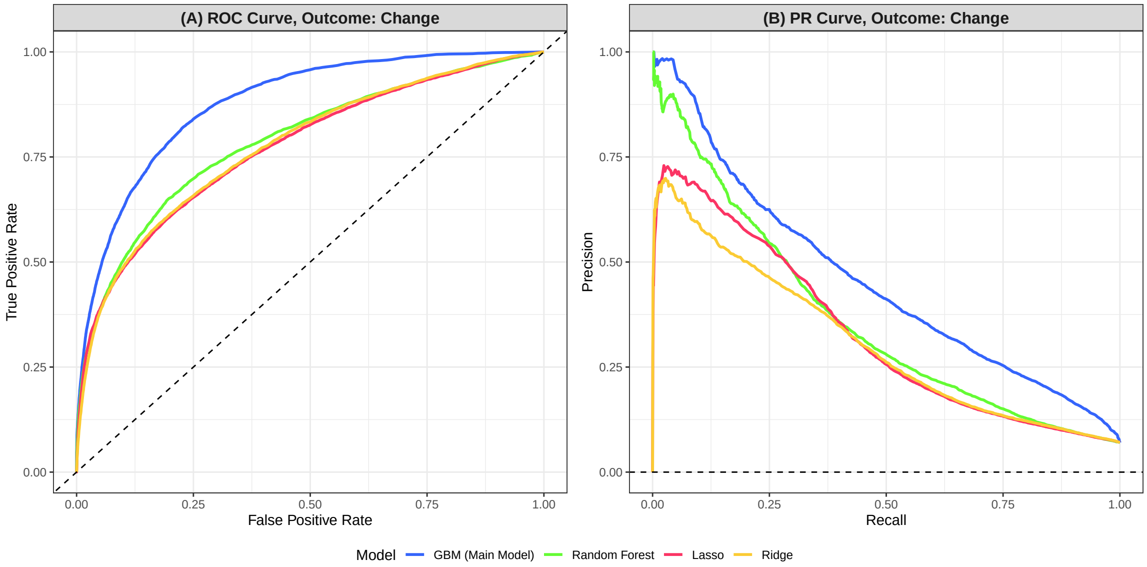
	Training	Validation	Test
GBM (Main Model)			
AUC	0.9293748	0.8933057	0.8783696
AUCPR	0.9188185	0.4705715	0.4557725
Random Forest			
AUC	0.9625070	0.8052770	0.7879675
AUCPR	0.9544986	0.3814282	0.3609348
LASSO			
AUC	0.7819766	0.7924853	0.7725845
AUCPR	0.3238795	0.3434476	0.3273591
Ridge			
AUC	0.7884536	0.7945496	0.7757873
AUCPR	0.2975521	0.3217342	0.3064647

Figure E.3: Predicting Change: Variable Importance Plot, All Models (Training Set)



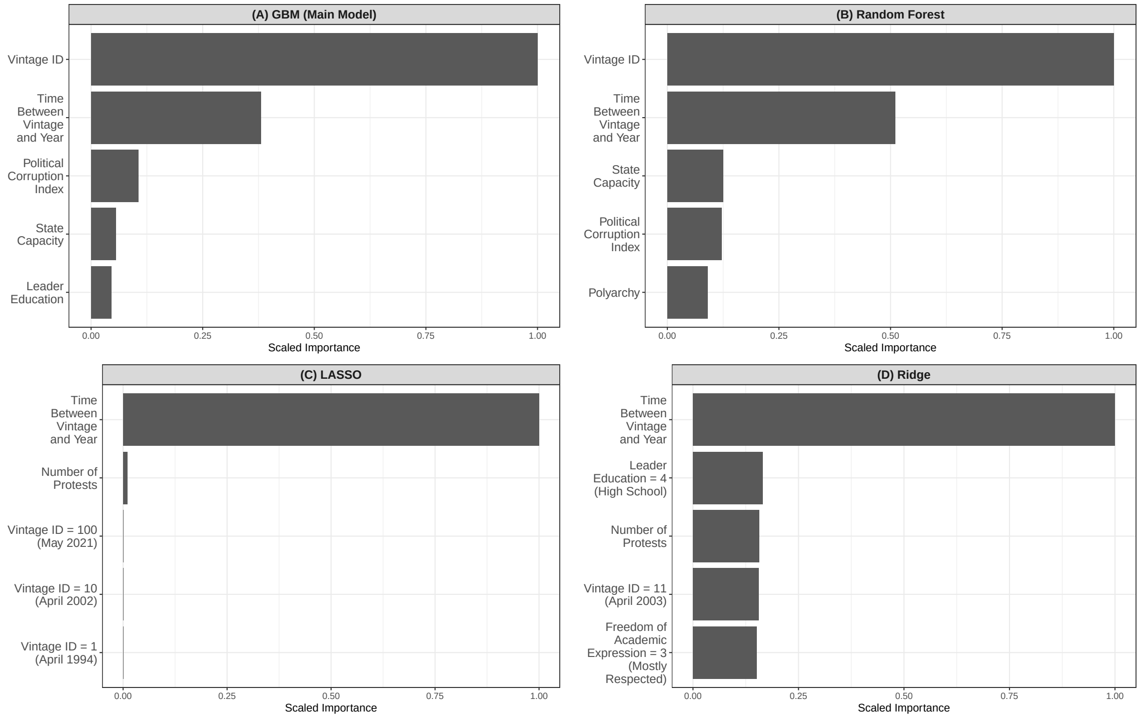
These panels show the relative importance of the top five variables, by model. The least important variable equals zero, while the most important variable equals one. LASSO adds a penalty to the absolute values of the coefficients (L1 regularization) that encourages most coefficients — like that of all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

Figure E.4: Predicting Change: ROC and PR Curves, All Models (Test Set)



Panel (A) presents a Receiver Operating Characteristic (ROC) curve for the the test set using four models: GBM (presented in the main text), random forest, LASSO, and ridge regression. Similarly, Panel (B) presents a Precision-Recall (PR) curve, also for the test set and also using the same four models.

Figure E.5: Predicting Outliers: Variable Importance Plot, All Models (Training Set)

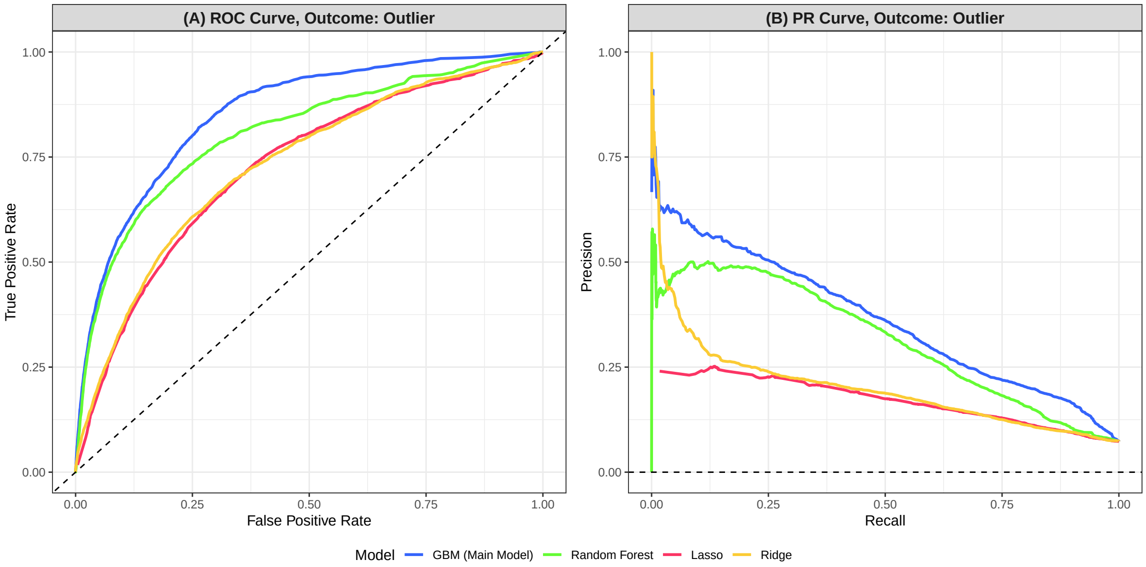


These panels show the relative importance of the top five variables, by model. The least important variable equals zero, while the most important variable equals one. LASSO adds a penalty to the absolute values of the coefficients (L1 regularization) that encourages most coefficients — like that of all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

Table E.3: Predicting Outliers: Performance Statistics, All Models

	Training	Validation	Test
GBM (Main Model)			
AUC	0.9354866	0.8695898	0.8536371
AUCPR	0.9212325	0.3784833	0.3641515
Random Forest			
AUC	0.9990912	0.8245832	0.8062617
AUCPR	0.9987284	0.3076397	0.315212
LASSO			
AUC	0.7401503	0.7441611	0.7255345
AUCPR	0.2015761	0.2042317	0.172373
Ridge			
AUC	0.7696063	0.7670707	0.7298865
AUCPR	0.2556064	0.2308627	0.2013158

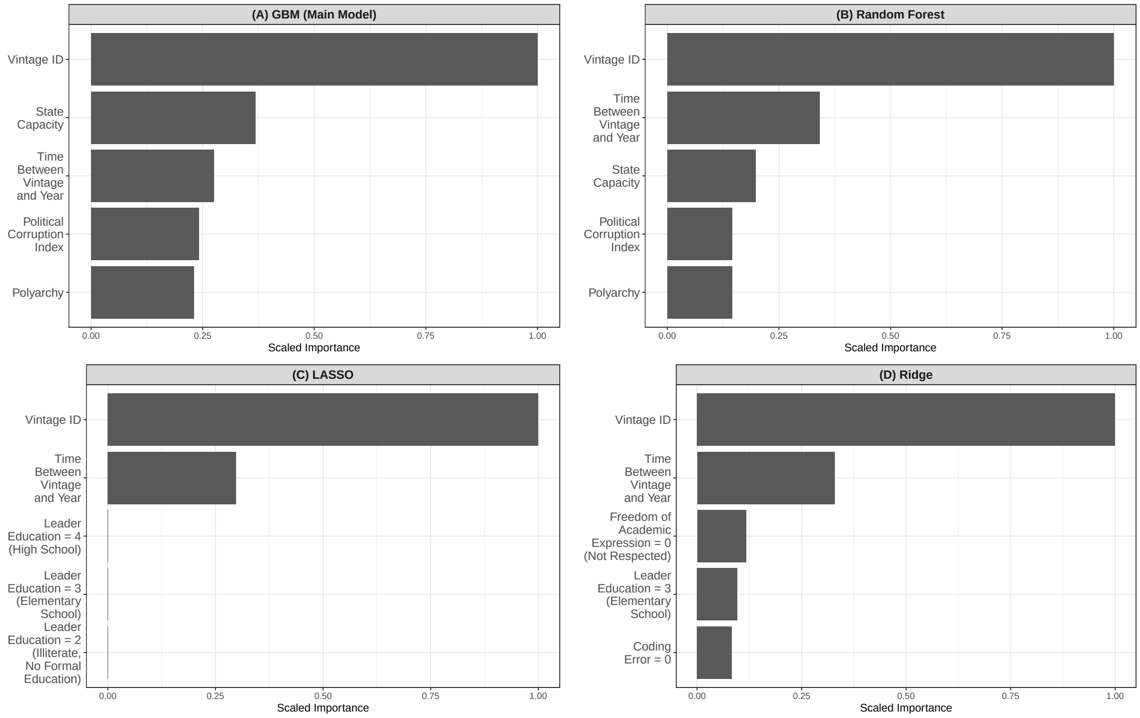
Figure E.6: Predicting Change: ROC and PR Curves, All Models



Panel (A) presents a Receiver Operating Characteristic (ROC) curve for the the test set using four models: GBM (presented in the main text), random forest, LASSO, and ridge regression. Similarly, Panel (B) presents a Precision-Recall (PR) curve, also for the test set and also using the same four models.

Lastly, I examine the relative performance of a GBM predicting the *Z-Score*. As Figure E.7, Table E.4, and Figure E.8 show, all four models (and the two GLMs in particular) struggle to predict extreme values of this continuous outcome.

Figure E.7: Predicting Z-Scores: Variable Importance Plot, All Models (Training Set)

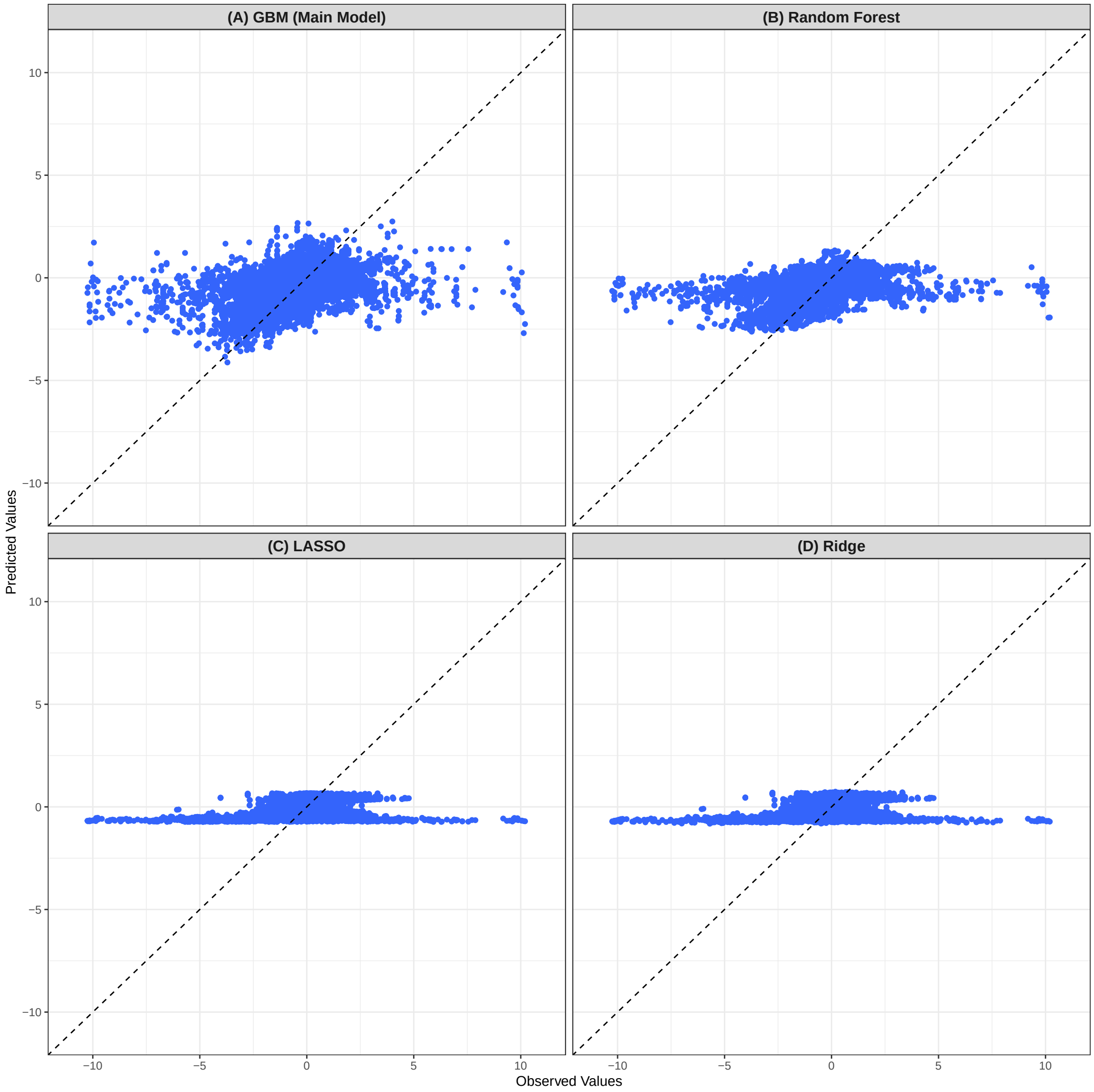


These panels show the relative importance of the top five variables, by model. The least important variable equals zero, while the most important variable equals one. LASSO adds a penalty to the absolute values of the coefficients (L1 regularization) that encourages most coefficients — like that of all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

Table E.4: Predicting Z-Scores: Performance Statistics, All Models

	Training	Validation	Test
GBM (Main Model)			
R^2	0.7123624	0.3019717	0.2208295
MSE	0.2561792	0.6169476	0.6936291
Random Forest			
R^2	0.4145735	0.3173313	0.2521846
MSE	0.5213995	0.6033722	0.6657163
LASSO			
R^2	0.1216233	0.2230289	0.2023817
MSE	0.7823102	0.6867207	0.7100516
Ridge			
R^2	0.1223481	0.2259166	0.2050922
MSE	0.7816648	0.6841684	0.7076387

Figure E.8: Predicted Versus Observed Values (Test Set)



This figure plots the observed values on the x-axis against the predicted values on the y-axis. Each point represents an observation, and the diagonal line represents perfect predictions. The closer the points are to the diagonal line, the better the model's predictions align with the actual values. Note that Panels (C) and (D) are nearly identical: LASSO and ridge models struggle to predict extreme values at a similar rate.

E.2 Alternative Predictors: Reported and Reporting Years

The main analysis only includes predictors for the *reported* year t. I also estimate models including each variable twice, both for year t and for year k, as current circumstances might drive changes in 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. Still, these cases are rare. Since reporting-year characteristics are highly correlated with reported-year characteristics, the inclusion of the former leads to an unstable and redundant models that overfit the training data, without meaningful gains in the predictions for the test set, as Tables E.5 to E.8 show. Figure E.9 confirms that variables from the *reported* year matter most: the top predictor of variation in *Missing* continues to be *Polyarchy (Reported)*, whereas the top predictor of variation in *Change*, *Outlier*, and *Z-Score* continues to be *Vintage ID*.

Table E.5: Predicting Missingness: Performance Statistics, Including Reported and Reporting Years

	Training	Validation	Test
GBM (Main Model)			
AUC	0.9978304	0.7662927	0.9317631
AUCPR	0.9976250	0.2921954	0.3297605
GBM (Including Reported and Reporting Years)			
AUC	0.9999704	0.7691764	0.8692491
AUCPR	0.9999551	0.1079385	0.3108512

Table E.6: Predicting Change: Performance Statistics, Including Reported and Reporting Years

	Training	Validation	Test
GBM (Main Model)			
AUC	0.9293748	0.8933057	0.8783696
AUCPR	0.9188185	0.4705715	0.4557725
GBM (Including Reported and Reporting Years)			
AUC	0.9958117	0.8954220	0.8715448
AUCPR	0.9941449	0.4715352	0.3861584

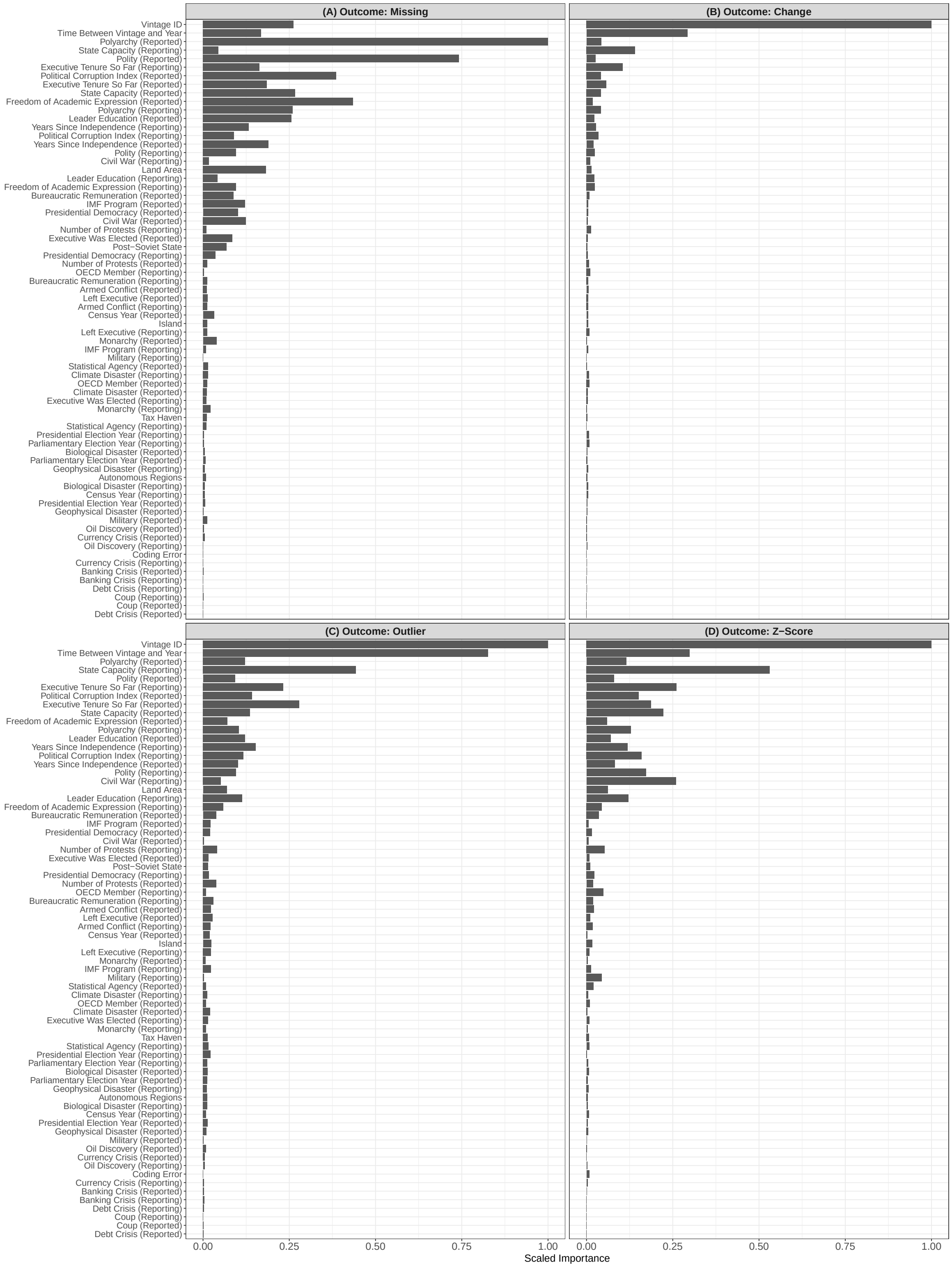
Table E.7: Predicting Outliers: Performance Statistics, Including Reported and Reporting Years

	Training	Validation	Test
GBM (Main Model)			
AUC	0.9354866	0.8695898	0.8536371
AUCPR	0.9212325	0.3784833	0.3641515
GBM (Including Reported and Reporting Years)			
AUC	0.9980507	0.8157824	0.8243962
AUCPR	0.9975861	0.2992289	0.2953656

Table E.8: Predicting Z-Scores: Performance Statistics, Including Reported and Reporting Years

	Training	Validation	Test
GBM (Main Model)			
R^2	0.7123624	0.3019717	0.2208295
MSE	0.2561792	0.6169476	0.6936291
GBM (Including Reported and Reporting Years)			
R^2	0.4550616	0.3020185	0.2509994
MSE	0.4853395	0.6169063	0.6667714

Figure E.9: Variable Importance Plot (Training Set)



This figure shows the relative importance of each predictor, by outcome. 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.

E.3 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 38 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.9 to E.12 show.

Table E.9: Predicting Missingness: Performance Statistics, Including WDI Variables

	Training	Validation	Test
GBM (Main Model)			
AUC	0.9978304	0.7662927	0.9317631
AUCPR	0.9976250	0.2921954	0.3297605
GBM (Including WDI Variables)			
AUC	0.9992545	0.8261948	0.8727175
AUCPR	0.9991468	0.2379344	0.2521146

Table E.10: Predicting Change: Performance Statistics, Including WDI Variables

	Training	Validation	Test
GBM (Main Model)			
AUC	0.9293748	0.8933057	0.8783696
AUCPR	0.9188185	0.4705715	0.4557725
GBM (Including WDI Variables)			
AUC	0.9448182	0.8027452	0.7843408
AUCPR	0.9350744	0.3717094	0.3406877

Table E.11: Predicting Outliers: Performance Statistics, Including WDI Variables

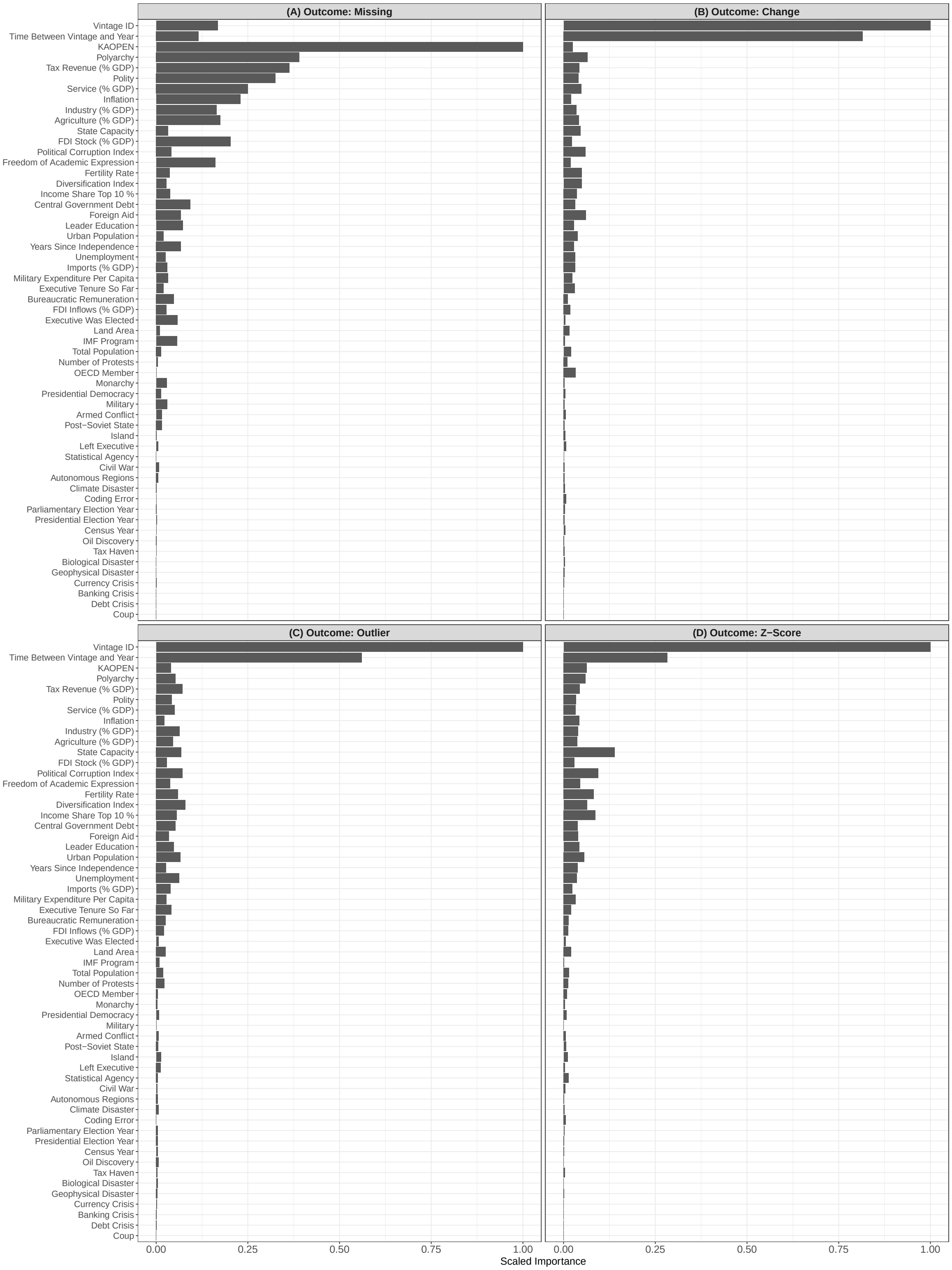
	Training	Validation	Test
GBM (Main Model)			
AUC	0.9354866	0.8695898	0.8536371
AUCPR	0.9212325	0.3784833	0.3641515
GBM (Including WDI Variables)			
AUC	0.9989961	0.8391951	0.8195714
AUCPR	0.9985814	0.3473658	0.3283745

Table E.12: Predicting Z-Scores: Performance Statistics, Including WDI Variables

	Training	Validation	Test
GBM (Main Model)			
R^2	0.7123624	0.3019717	0.2208295
MSE	0.2561792	0.6169476	0.6936291
GBM (Including WDI Variables)			
R^2	0.4619303	0.3230428	0.2692702
MSE	0.4792220	0.5983240	0.6505065

As Figure E.10 shows, the main predictor of missingness in an expanded model is KAOPEN, Chinn and Ito's (2006) capital openness index, constructed using data from the IMF's Annual Report on Exchange Arrangements and Exchange Restrictions (see Chinn and Ito 2008 for more information). The main predictor of *Change*, *Outlier*, and *Z-Score* continues to be *Vintage ID*. 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.10: Variable Importance Plot (Training Set)



This figure shows the relative importance of each predictor, by outcome. 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.

F Model Specification

F.1 Classification Trees

I estimate all models using the open source machine learning platform `H2O`, implemented via `R`. To predict missingness (a classification task), I estimate a GBM 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). 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.

`ntrees = p × 20`. This is the number of trees, with $p = 40$ in this case. Higher values are computationally intensive and do not perform better.

`sample_rate = 1`. Each tree is trained on 100 percent of the training data, drawn at random and without replacement (default value).

`col_sample_rate = 1`. 100 percent of the $p = 40$ columns are randomly selected and used for building each tree in the ensemble (default value).

`col_sample_rate_per_tree = 0.8`. 80 percent of the $p = 40$ columns are used for each individual tree (default value is 1). This allows for different columns to be selected for different trees.

`max_depth = 15`. The maximum tree depth is specified as 15 (default value is 5), which means that each tree has up to 15 splits. Higher values (as in, more complex trees) are computationally intensive and can lead to overfitting.

`min_rows = 1`. This parameter specifies the minimum number of observations for a terminal node (default value). The default value indicates that there might be a combination of splits that explains something seen only once in the training data: there might be a path through the tree that leads to only one observation.

`learn_rate = 0.1`. Rate at which the algorithm learns (default value). Lower learning rates are better, but more computationally intensive.

`stopping_rounds = 5`. The model uses early stopping: it stops training when the option selected for stopping˙metric does not improve for 5 training rounds, based on a simple moving average (default value is 0, without early stopping).

`stopping_metric = ‘‘AUTO’’`. The default stopping metric for classification tasks is Log Loss.

`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.

`class_sampling_factors = c(0.8, 1)`. This hyperparameter only exists for classification tasks; it tells the model to specifically undersample the majority class.

F.2 Regression Trees

As before, I estimate all models using the open source machine learning platform `H2O`, implemented via `R`. To predict deviation (a regression task), I estimate a GBM with the following hyperparameters:

`ntrees = p` × `20`. This is the number of trees, with $p = 40$ in this case. 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 1).

`col_sample_rate = 0.6`. 60 percent of the $p = 40$ columns are randomly selected and used for building each tree in the ensemble (default value is 1).

`col_sample_rate_per_tree = 1`. 100 percent of the $p = 40$ columns are used for each individual tree (default value).

`max_depth = 5`. The maximum tree depth is specified as 5 (default value), which means that each tree has up to 5 splits.

`min_rows = 10`. I increased the minimum number of observations for a terminal node from 1 to 10 (default value is 1).

`learn_rate = 0.05`. Rate at which the algorithm learns (default is 0.1). Lower learning rates are better, but more computationally intensive.

`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-4`. This is the tolerance value by which a model must improve before training ceases (default value is 1e-3).

References

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.

Chinn, Menzie D. and Hiro Ito. 2008. "A New Measure of Financial Openness." *Journal of Comparative Policy Analysis: Research and Practice* 10(3):309–322.

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.

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*

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.

Laeven, Luc and Fabian Valencia. 2020. "Systemic Banking Crises Database II." *IMF Economic Review* 68:307–361.

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.

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.