Democratic Spaces Dashboard: 2021 Update and Accuracy Assessments

Andreas Beger^*

2021-04-07

Contents

This report was produced during the spring 2021 update of the Democratic Spaces dashboard and forecasts, on behalf of V-Dem for IRI.

Introduction

The Democratic Spaces forecasting project measures six democratic spaces using indicators selected from the Varieties of Democracy (V-Dem) project. For each of the six spaces, we are interested in significant improvement (up) or deterioration (down) from one year to another. "Significant changes" are operationalized as …

Space	Indicator	Description
Electoral	v2x_veracc_osp	Vertical accountability index
Associational	v2xcs_ccsi	
Individual	v2xcl_rol	Rule of law index
Informational	v2x_freexp_altinf	
Governing	v2x_horacc_osp	
Economic	v2x_pubcorr	Absence of public corruption. Note that this is inverted from the original V-Dem variable so that high values indicate absence of public corruption.

There are thus 12 outcomes that we are trying to model in total: 6 spaces and for each space whether a shift in the up or down direction occurred.

While the outcomes are yearly in nature, the forecasts themselves cover 2-years ahead. We aggregate the yearly democratic space changes data to a 2-year target for the forecasting model using logical "or" relationships. Thus for example the target that the electoral space "up" model is trying to predict indicates whether an "up" movement occurred in the electoral space of a country in at least 1 year during the 2-year window, but it could also have happened twice in succession. Logically this also means that a country could experience both an "up" *and* "down" shift in the same 2-year window, which would warrant high values in both the "up" and "down" forecasts, but this doesn't happen very often. Geographically, the forecasts cover 169 countries.

The project was initially developed in 2019 and the first set of forecasts were made in late 2019 with the V-Dem version 9 data, covering the 2-year window from 2019 to 2020. The forecasts have since been updated twice, in the spring of 2020 and now in the spring of 2021. There are thus now in total three sets of forecasts, indexed by the V-Dem data version they were based on:

- v9: covering 2019-2020
- v10: 2020-2021
- v11: 2021-2022

V-Dem version 11 now has data through to 2020, so we can fully score the first forecasts, and partially score the v10 forecasts.

The rest of this note will show the (partial) scoring of the first two forecasts, discuss how improvements in the V-Dem data over time impact forecast accuracy, and then finally review the latest set of forecasts for 2021–2022.

Outline:

- v9 forecasts we can assess
- show fit values
- one issue is ground shifting
- show case overlap between v9 and v11
- show decrease in nominal test accuracy due to ground shifting
- v10 forecasts we can partially assess
- show case overlap between v9, v10, v11
- show test forecast loss due to ground shifting in v10
- v11 forecasts

Table 1: Accuracy of the v9 forecasts for 2019–2020

Space	Cases	In_top20	AUC-ROC	AUC-PR	Pos_rate
Downards movement					
Associational	29	5	0.65	0.25	0.17
Economic	31	5	0.72	0.30	0.18
Electoral	10	3	0.72	0.12	0.06
Governing	21	3	0.62	0.15	0.12
Individual	34	6	0.70	0.36	0.20
Informational	30	6	0.67	0.26	0.18
Upwards movement					
Associational	16	5	0.75	0.22	0.09
Economic	42	11	0.67	0.42	0.25
Electoral	4	0	0.64	0.03	0.02
Governing	26	4	0.72	0.24	0.15
Individual	19	8	0.82	0.29	0.11
Informational	19	8	0.77	0.35	0.11
Average					
	23	5	0.71	0.25	0.14

- nominal test accuracy
- how much is that likely to decrease?

Scoring past forecasts

Scoring the v9 2019-2020 forecasts

The first set of forecasts were done in late 2019 using V-Dem version 9 and for years 2019–2020. Table 1 shows their accuracy when scored using the V-Dem v11 data. There are in total 12 different outcomes we forecast: downwards (worse) and upwards (better) movements (i.e. 2 directions) for each of the 6 spaces. The forecasts cover 169 countries and the first column (“Cases”) shows the number of corresponding events recorded in the new V-Dem data. The last column, “Pos_rate”, show the rate, i.e. the number of cases divided by 169 countries. The first metric we look at, “In top20”, simply counts how many of the highest 20 forecasts for each outcome had an actual event. Ideally this would be 20, or the number of cases if it is lower than that, out of 20. In practice, more like 1 in 4 of the top 20 highest forecasts had an actual event, if we average across all outcomes.

The next two measures are the areas under the receiver operating characteristic (ROC) and precision-recall curves—AUC-ROC and AUC-PR. Both of these are based on the forecasts’ ability to correctly rank countries so that countries that experience an event are ranked higher than those that do not. Where they differ is that the AUC-ROC measures the trade-off between true positive (predicted and actual event) and true negative rates (predicted and actual non-event), while the AUC-PR measures the trade-off between the true positive rate (also called recall) and the precision of the forecasts (how many positive predictions actually had an event). Both are among the standard measures for this kind of prediction problem, but unlike other measures like Brier scores or average log loss, they have natural reference values that make them easier to

Table 2: Partial accuracy of the v10 forecasts for 2020–2021 with 2020 outcomes only

Space	Cases	In_top20	AUC-ROC	AUC-PR	Pos_rate
Downards movement					
Associational	16	4	0.73	0.17	0.09
Economic	19	5	0.77	0.22	0.11
Electoral	4	1	0.78	0.07	0.02
Governing	16	4	0.65	0.14	0.09
Individual	21	5	0.65	0.20	0.12
Informational	14	2	0.62	0.10	0.08
Upwards movement					
Associational	10	2	0.71	0.12	0.06
Economic	21	4	0.68	0.18	0.12
Electoral	2	2	0.94	0.09	0.01
Governing	13	2	0.77	0.14	0.08
Individual	9	0	0.69	0.08	0.05
Informational	9	4	0.72	0.25	0.05
Average					
	13	3	0.73	0.15	0.08

interpret. Both theoretically can range from 0 to 1. However a naive forecast that, for example, randomly guesses positive and negative predictions using the base rate, will on average have an AUC-ROC score of 0.5 and an AUC-PR score equal to the base rate (shown in the last column). To be useful, a forecast should exceed these reference values. This is the case for both measures and all 12 outcomes, and thus we can conclude that the forecasts are informative; they add a signal over the naive base rate.

Partial scoring of the v10 2020-2021 forecasts

The forecasts made in the spring of 2020 using V-Dem version 10 data covered 2020-2021. We have data for 2020 in V-Dem version 11 and can thus partially score the forecasts using observed positive outcomes in the first year of the forecast time period.

Expected test forecast accuracy versus actual accuracy

That said, the absolute values for both measures are not particularly good, even if we keep in mind that these scores are only for a single set of forecasts and there is some natural year to year variation in accuracy. The AUC-ROC scores range from 0.62 to 0.82 with an average of 0.7. For AUC-PR the range is 0.03 to 0.42 with an average of 0.25; for comparison, the average positive rate is 0.14. Similar forecasting applications with other forms of political instability typically achieve AUC-ROC values in the 0.8 to 0.9 range, and similar AUC-PR scores as here but with much lower base rates, on the order of a handful per hundred, not a dozen per hundred like here.

Furthermore, the test forecasts we produce as part of these forecasts specifically in order to get a sense of the likely accuracy of the live forecasts had much higher accuracy. Namely, the test forecasts are where we pretend to go back in time to 2005, make a 2-year forecast, then move a year up and do it again, etc., and then at the end use our knowledge of the actual historical outcomes to score them. For v9, this gave us 12 distinct 2-years-ahead forecasts from 2005 to 2016, and for which we already knew the actual outcomes, because they were in the v9 V-Dem data. The average AUC-ROC and AUC-PR scores from those were around 0.83 and 0.35, which is noticeably higher than the accuracy of the live forecasts.

It turns out that the accuracy of the forecasts is negatively affected by changes in the underlying V-Dem data itself between different versions of the data.

Changes in V-Dem data over time and impact on forecast accuracy

Changes in events between different V-Dem version

The V-Dem data are continually improving, which leads to changes in the data between different versions of the data, even for past, historical values. To understand why these shifts occur it helps to review how the V-Dem data are created (the methodology for the current V-Dem version 11.1 on which this summary is based is outlined in Coppedge et al. 2021). Ultimately, the top-level indices that measure different aspects of democracy like liberal or electoral democracy, rule of law, vertical and horizontal accountability, etc. are built up from survey questionnaires in which country experts rate some aspect of a country's political situation in a given year. There are to date hundreds of thousands of such survey responses to more than 470 questions for different countries in different years, submitted by more than 3,000 country experts. These raw survey responses–usually at least 5 per question for a given country and year–are then passed through Bayesian measurement models and Bayesian factor analysis in order to extract underlying latent "signals". These latent signals are what the top-level indices are. That is, unlike older coding projects like Polity, the top-level indices are not straightforward mathematical transformations based on specific values that are coded by only one person or where inter-coder disagreement is resolved "behind the scenes" to still produce a single value. Rather, all the indices that we use to measure the democratic spaces are *estimates* that are *uncertain*, i.e. explicitly include measurement error.

To update the data for a new year, V-Dem polls the country experts in order to obtain question responses for the last year, but potentially also for previous years. The *entire* set of indicators and top-level indices that are derived from the raw survey responses–not just for the new year but for all years–are then re-created by running the relevant models and factor analyses. There are several aspects of this process that can lead to changes in historical values, compared to previous versions of the V-Dem data:

- The overall pool of survey responses changes from year to year. Country experts do not only code the new year, but can also adjust previous responses, e.g. if new information has come to light. There is churn in the country-expert pool as well, and both old and new coders may add additional question responses for previous years or other countries (these kinds of cross-coding are used to help adjust for inter-coder and inter-country variation and coder-specific idiosyncracies).
- As the entire data, including historical data, are re-generated on update, these general changes in the pool of survey responses will also alter historical values that were already recorded in previous versions of the V-Dem data.
- Minor changes in the data and modeling process itself, e.g. to adjust for newly discovered issues or fix bugs.
- Inherent sampling variation in the models. The Bayesian measurement and factor analysis models work through Markov chain Monte Carlo (MCMC) sampling, which gives them more

flexibility and other advantages compared to more conventional deterministicially optimized statistical models. But this also means is that even with identical data, code, and model, re-running the model will produce slightly different estimates each time.

Aside from the inherent sampling variation, these sources of change are not arbitrary or random. They represent improvements in V-Dem's *measures* of different aspects of democratic governance. Events that led a country expert to answer a question with a specific value in one year maybe be interpreted in a different way in light of subsequent events, or maybe previously unknown information has now publicly become available; there may now be additional survey responses to cover a given country in a given year than had been available last year, etc. In essence, the picture that version 9 of V-Dem portrayed of democratic governance in the world is not as clear as the picture now provided by version 11, which in turn is not as clear and accurate as that given by the next version of V-Dem, etc.

threshold based outcomes, small perturbations matter

show a table comparing positives in one year against another

How much does accuracy decrease due to shifting ground truth?

So, to summarize the discussion so far, although we specifically had (and have) test forecasts as a means of getting a sense of how accurate the live forecasts are, now that we have enough updated V-Dem data to score the first set of forecasts made in 2019, we can see that the refinements to the V-Dem data, as they are updated each year, represent an additional limitation on our ability to accurately forecast changes in the democratic spaces. Essentially we have to develop our models and thus the forecasts with data that from the perspective of two years from now, will be outdated and not as accurate as the more recent data versions.

How to overcome this issue by adjusting the models that produce the forecasts is a challenging questions without an immediate, straightforward answer. What we can try to do however is to quantify how much changes in the underlying V-Dem data impact our initial sense of accuracy derived from the test forecasts. We looked at one set of forecasts that we made using the version 9 of the V-Dem data, where we had an expectation of their accuracy that were based on test forecasts that were also made with the v9 data, and then we were able to generate an acutal score for the single year of live forecasts using the v11 V-Dem data. We can do the same thing with the v9 test forecasts themselves however, i.e. score them as well using the v11 data. That gives us more than one year of accuracy metrics, and thus a better sense of how much the expected and actual accuracy (with updated data) compare to each other. And even though we do not yet have enough data to score the 2020 *live* forecasts made with V-Dem v10 data, we can score the v10 *test* forecasts. That gives us not only even more years, but now we can also see how the changes from v9 to v10 compare to the changes from v10 to v11 V-Dem data.

Forecasts for 2021-2022

Expected accuracy

Nominal test accuracy

How much is that likely to decrease?

Show nominal test accuracy.

Figure 1: Electoral space

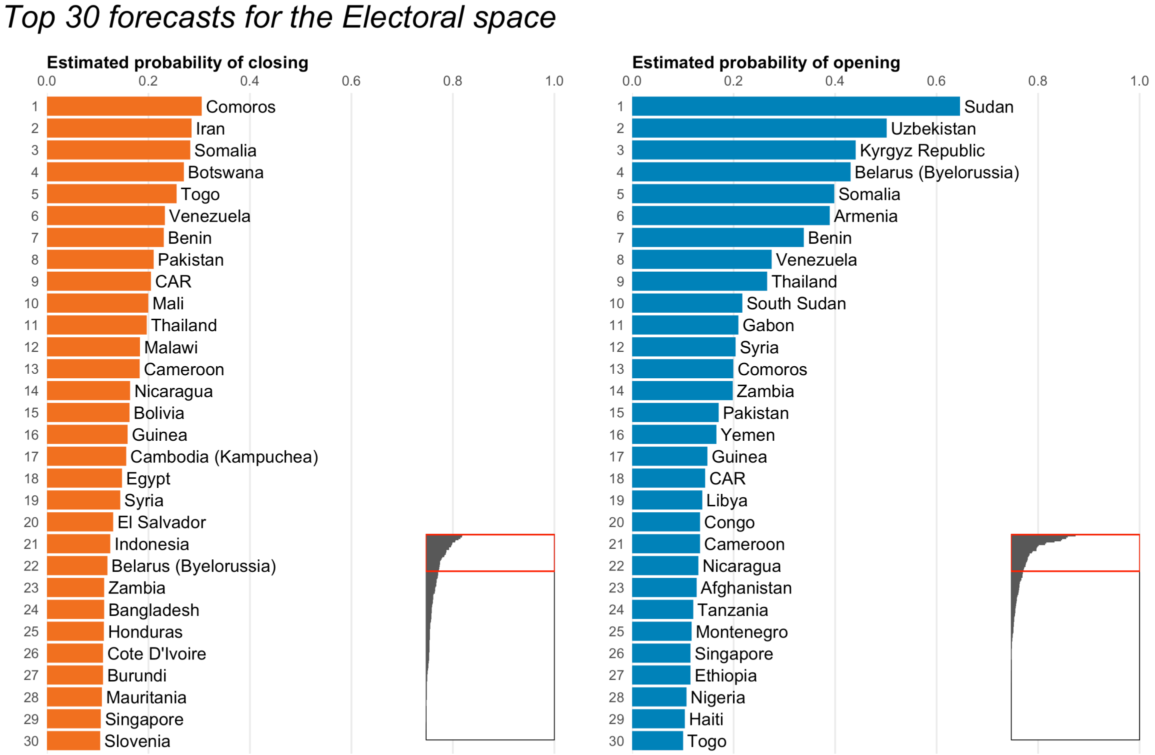


Figure 2: Associational space

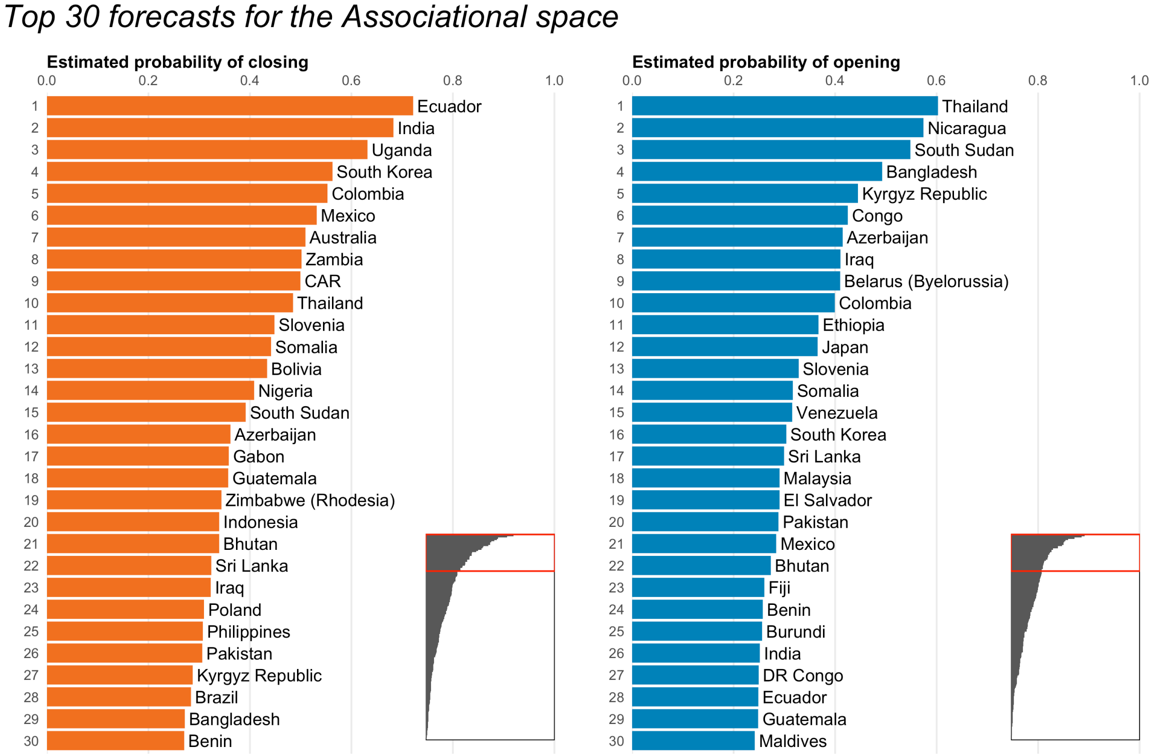


Figure 3: Individual space

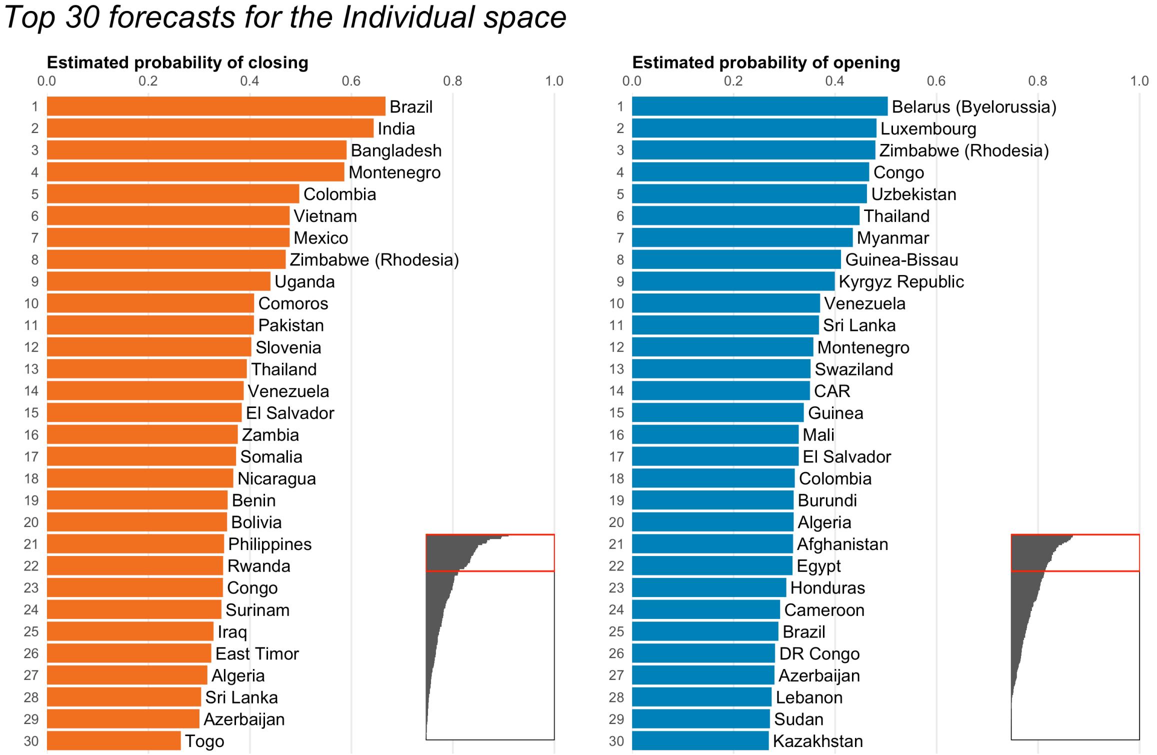


Figure 4: Informational space

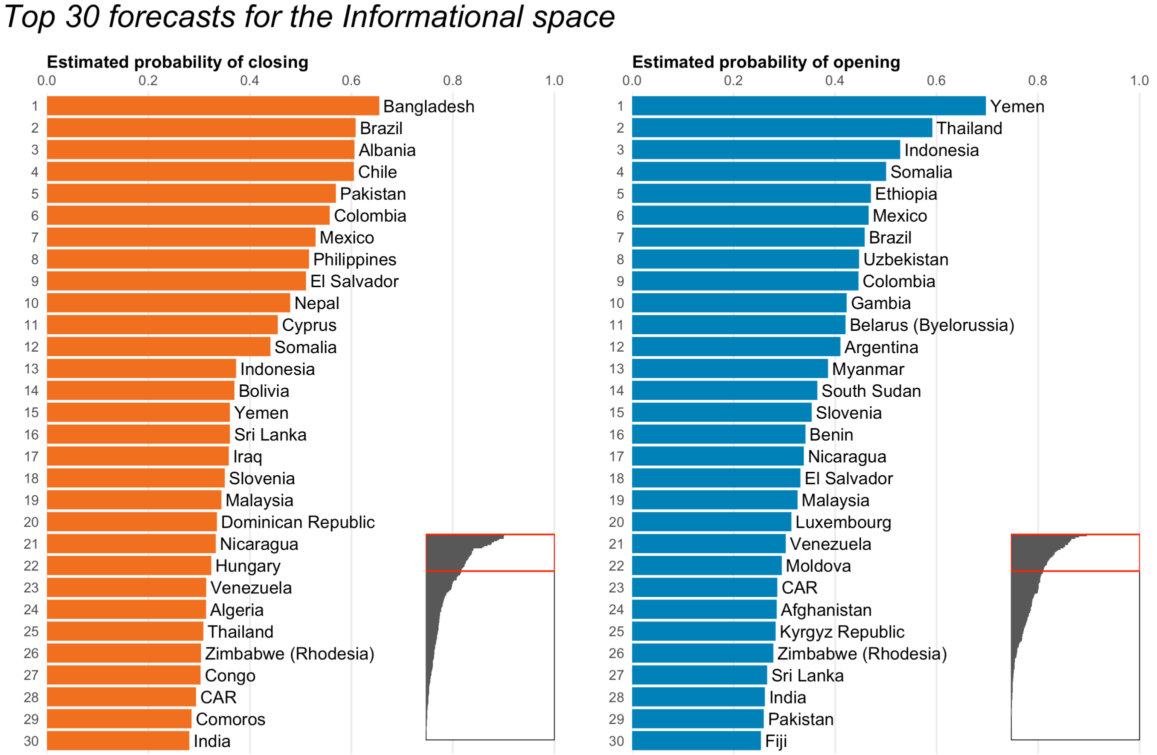


Figure 5: Governing space

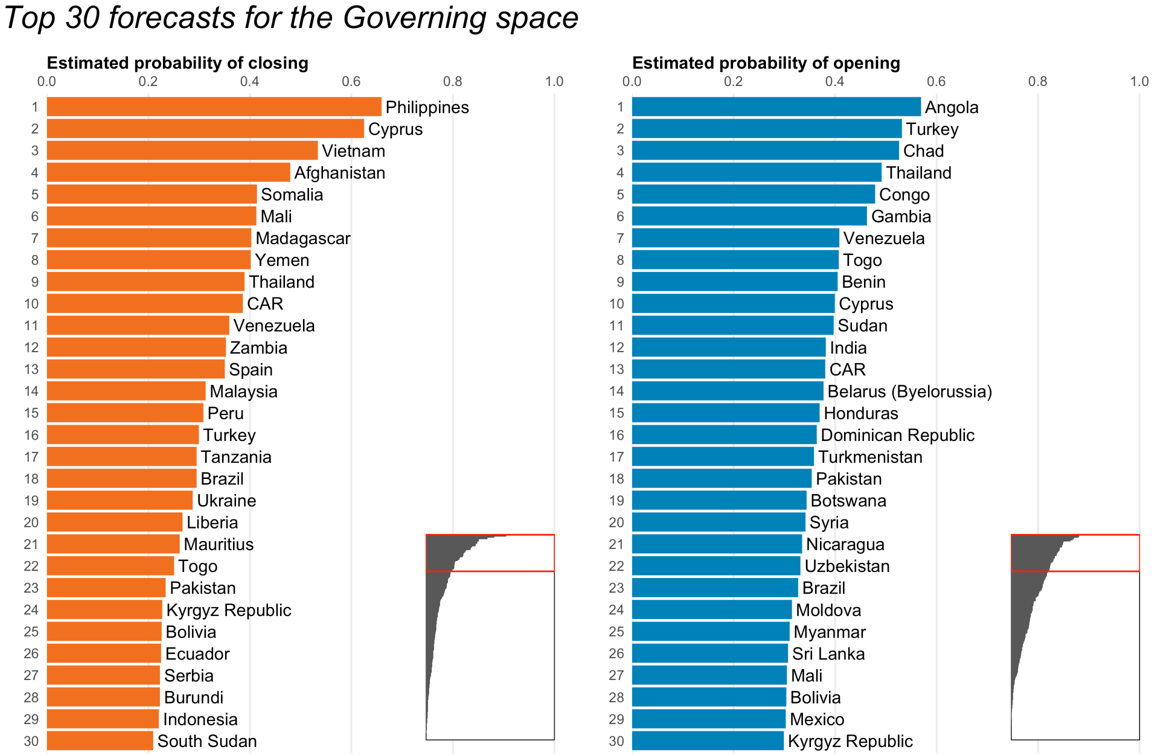
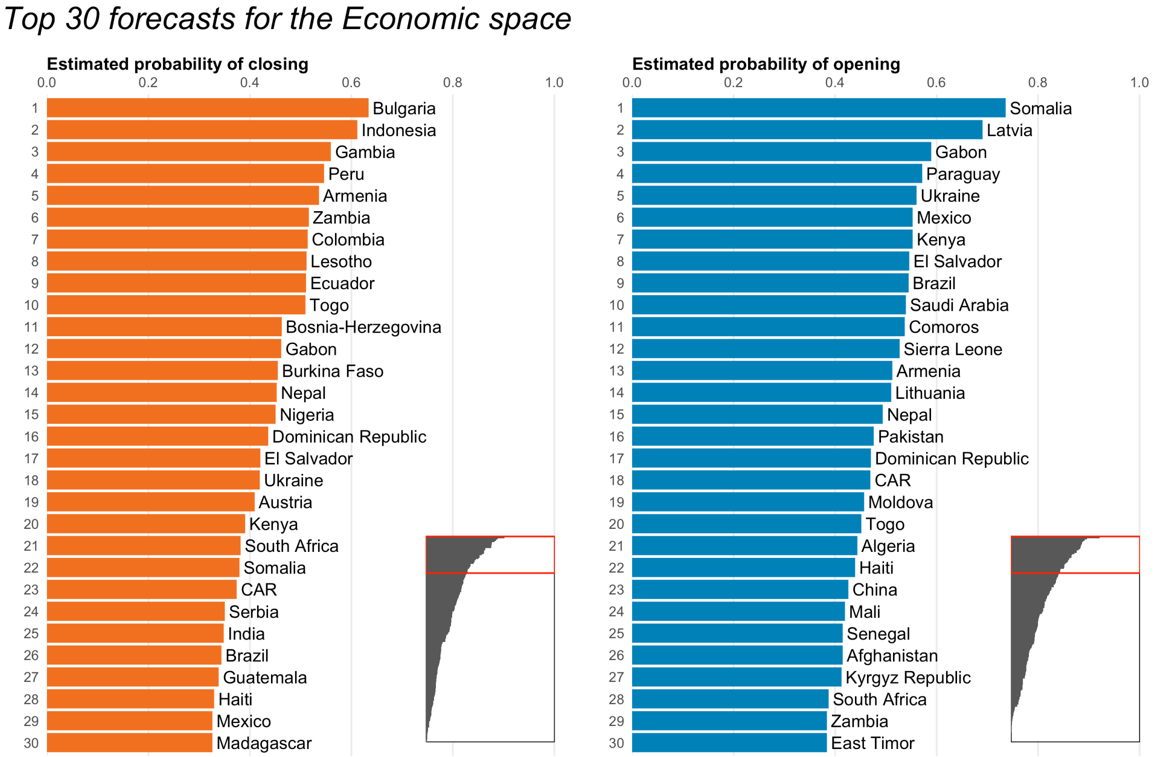


Figure 6: Economic space



Show test accuracy decreases from v9 with v10 data, v10 with v11 data

Conclusion

One final note: model raw indices not discretized outcomes. Maybe that helps get around the shifting data problem.

References

Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Jan Teorell, Kyle L. Marquardt, Juraj Medzihorsky, et al. 2021. "V-Dem Methodology V11.1." https://www.v-dem.net/media/filer_public/4e/1c/4e1c47ae-4800-436a-bbf1-c5fb50798bd3/methodology_v111.pdf.