Democratic Spaces Dashboard: 2021 Update and Accuracy Assessments

Andreas Beger*

2021-04-14

Contents

Summary **1**

Introduction **2**

Scoring past forecasts **4**
Scoring the v9 2019-2020 forecasts 4
Partial scoring of the v10 2020-2021 forecasts 7
Discussion 9

Changes in V-Dem data over time and impact on forecast accuracy **11**
How much does this impact the "ground truth" with any given data version? 12

Forecasts for 2021-2022 **14**

Conclusion **19**

References **20**

Summary

The 2021 forecast updates are the third round of democratic spaces forecasts, made using V-Dem version 9, 10, and 11 data, respectively, and covering 2019-2020, 2020-2021, and 2021-2020. The first set of forecasts can now be fully assessed, and the second set of forecasts partially, with observed data for one of the two year forecast period.

The accuracy results show that the forecasts are informative and more accurate, by a significant margin, than a naive base-rate forecast. However, the available results for the live forecasts are not as good as those for the much more extensive test forecasts, where we go back in time to 2005 and replicate the live forecasting process for each year until we reach the end of fully observed outcomes, two years ago. The decline in expected versus actual accuracy is likely driven by two factors:

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

1. The Covid-19 pandemic, which has been used by some governments as cover to enact anti-democratic policies that exceed the needs of pandemic response and instead empower and consolidate executive power, for example recent restrictions on media freedom in Viktor Orban's Hungary.[1]

2. Changes between V-Dem data versions, in part due to continual improvements in the quality of the data, but also inherent variation in the Bayesian models used to create the data. This results in changes in the sets of opening and closing events for the democratic spaces between the data version used to create a forecast and the subsequent data versions used to score it.

The impact of Covid-19 is difficult to directly incorporate into the models as this is the first pandemic in recent history. However, one can try to assess the forecasts in combination with other information on the Covid-19 impact so far, like V-Dem's pandemic backsliding project (Lührmann and Rooney 2020; Edgell et al. 2020). The changes in the V-Dem data between versions are mostly inherent in the nature of the project. However, further examination also reveals that in the vast majority of cases, the disagreement between data versions is about the magnitude, not the direction of change. It might be possible to ameliorate the impact of the changes between data versions by changing the way the forecast models work.

To be clear though: the accuracy results so far indicate that the forecasts are informative and add substantial value over naive base rate-anchored forecasts. Any changes to try to address the two factors above would simply be an attempt to further improve accuracy.

Introduction

The Democratic Spaces forecasting project measures six aspects of democratic governance—"democratic spaces"—using indicators selected from the Varieties of Democracy (V-Dem) project. The spaces, the corresponding V-Dem indicator, and short descriptions are listed in Table 1. For each of the six spaces, we are interested in significant opening (improvement) or closing (deterioration) movements compared to last year. "Significant changes" are operationalized as year-to-year changes that exceed a certain threshold that is specific to each space and based on the range of past fluctuations that are normal (for details, see the original project report, Beger, Morgan, and Maxwell 2020). There are thus a total of 12 outcomes to forecast: 6 spaces and for each space whether a shift in the opening or closing direction occurred.

While the outcomes are yearly in nature, the forecasts themselves cover a period of two 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 opening model is trying to predict indicates whether an opening 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 opening *and* closing shifts in the same 2-year window, which would warrant high values in both the opening and closing 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:

[1] https://ipi.media/hungary-seeks-power-to-jail-journalists-for-false-covid-19-coverage/

Table 1: Democratic spaces and corresponding V-Dem indicators

Space	Indicator	Description
Electoral	v2x_veracc_osp	The ability of the population to hold their government accountable through elections and political parties. Measured using the vertical accountability index
Associational	v2xcs_ccsi	The degree of CSO autonomy from the state and citizens' ability to freely and actively pursue their political and civic goals. Measured with the core civil society index.
Individual	v2xcl_rol	The extent to which the laws are transparent and rigorously enforced and public administration impartial, and the extent to which citizens enjoy access to justice, secure property rights, freedom from forced labor, freedom of movement, physical integrity rights, and freedom of religion. Measured with the rule of law index
Informational	v2x_freexp_altinf	The degree of media censorship, harassment of journalists, media bias, media self-censorship, whether the media is critical and pluralistic, as well as the freedom of discussion and academic and cultural expression. Measured with the freedom of expression and alternative sources of information index.
Governing	v2x_horacc_osp	The degree to which the legislative and judicial branches can hold the executive branch accountable as well as legislative and judical oversight over the bureaucracy and security services. Measured using the horizontal accountability index.
Economic	v2x_pubcorr	Absence of public corruption. The extent to which public sector employees grant favors in exchange for bribes (or other material inducements), and how often they steal, embezzle, or misappropriate public funds or other state resources for personal or family use. Note that this is inverted from the original V-Dem variable so that high values indicate absence of public corruption.

Table 2: Accuracy of the v9 forecasts for 2019–2020

Space	Cases	In_top20	AUC-ROC	AUC-PR	Pos_rate
Closing 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
Opening 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

- 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 Covid-19 and changes in the V-Dem data between versions likely impacted it, and then briefly go over the new 2021-2022 forecasts.

Scoring past forecasts

Scoring the v9 2019-2020 forecasts

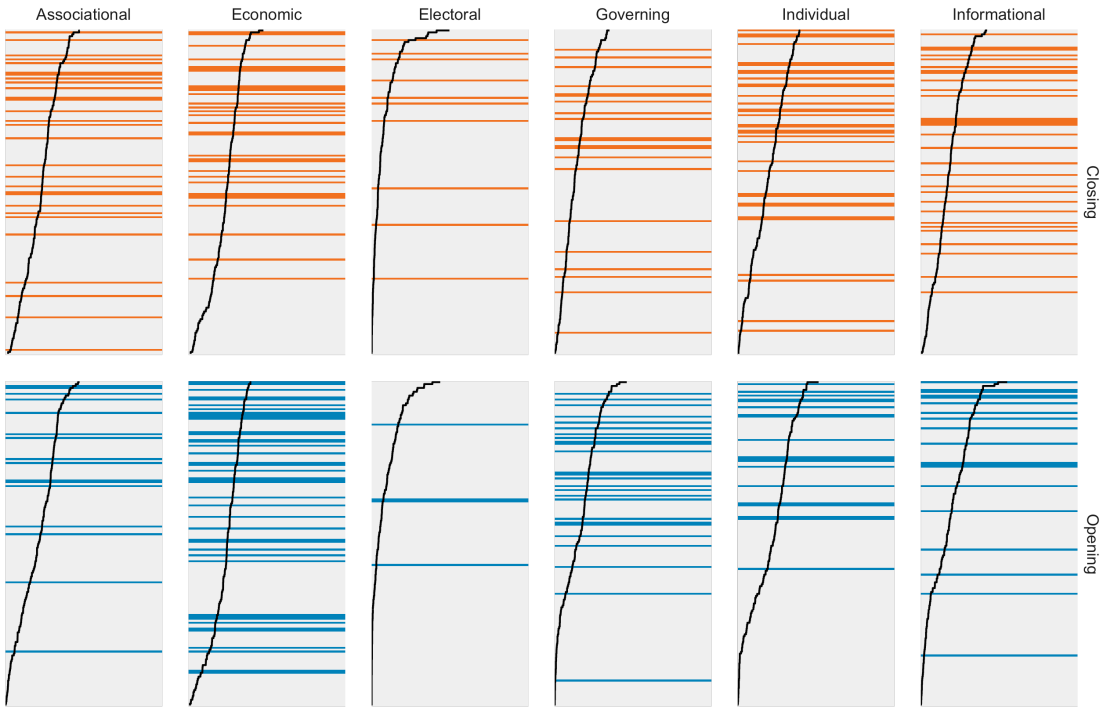
The first set of forecasts were done in late 2019 using V-Dem version 9 and for years 2019–2020. Table 2 shows their accuracy when scored using the V-Dem v11 data. There are in total 12 different outcomes we forecast: closing (worse) and opening (better) movements (i.e. 2 directions) for each of the 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 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 (i.e. positive 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.

Figure 1 shows another way to visually evaluate the forecasts, using separation plots (Greenhill, Ward, and Sacks 2011). Imagine we listed all forecasts in a table, in order of highest to lowest probability, and then colored each row that had an actual case. That's essentially what these plots are. Each plot has 169 bars–for the 169 countries–and is colored based on whether a closing (orange) or opening (blue) event occurred or not (gray). The black line shows the original forecast probabilities. In a good forecast, most or all of the colored lines would be clustered in a solid block at the top. Gray bars that are high up indicate false positive forecasts, i.e. a high forecast but no event. The limitations of the forecasts are quite easy to see: while the positives generally tend to cluster somewhat towards the top, only a few forecasts do not also have at least a few positives cases at the bottom ranks.

Figure 1: Separation plots for the v9 forecasts covering 2019-2020



Finally, here are lists of all positives for the 12 outcomes, along with the forecast rank (1 = highest risk, 169 lowest) and probability:

Closing

Associational: Benin, 7, 0.41; Nigeria, 13, 0.38; Colombia, 17, 0.37; Guatemala, 19, 0.37; Iraq, 21, 0.36; Sri Lanka, 23, 0.35; Mexico, 26, 0.35; Thailand, 36, 0.32; Algeria, 52, 0.3; Mauritania, 65, 0.27; Madagascar, 67, 0.27; Cote D'Ivoire, 71, 0.27; Belarus, 82, 0.24; Bhutan, 93, 0.22; Panama, 94, 0.22; Slovenia, 109, 0.18

Economic: India, 2, 0.43; Lesotho, 8, 0.41; Colombia, 11, 0.4; El Salvador, 12, 0.4; Armenia, 16, 0.39; Sri Lanka, 27, 0.37; Indonesia, 36, 0.35; Ecuador, 37, 0.34; Ukraine, 39, 0.34; Nigeria, 43, 0.33; Comoros, 46, 0.32; Namibia, 49, 0.32; Kenya, 66, 0.3; Bosnia-Herzegovina, 69, 0.29; Mauritania, 70, 0.29; Nicaragua, 71, 0.28; Sudan, 73, 0.28; Somalia, 80, 0.26; CAR, 104, 0.22

Electoral: Guinea, 5, 0.31; CAR, 23, 0.21; Cote D'Ivoire, 61, 0.09; Iran, 65, 0.09

Governing: Benin, 5, 0.34; Kyrgyzstan, 8, 0.32; Philippines, 13, 0.3; Venezuela, 19, 0.28; Sri Lanka, 30, 0.24; Serbia, 42, 0.22; Mali, 43, 0.22; Cote D'Ivoire, 45, 0.22; Yemen, 61, 0.19; Laos, 84, 0.15; Botswana, 86, 0.14; Vietnam, 92, 0.13; Burundi, 96, 0.13; Mauritius, 106, 0.11; Malaysia, 117, 0.09; Spain, 152, 0.03

Individual: Bolivia, 2, 0.45; Uganda, 8, 0.42; Zambia, 10, 0.42; Guinea, 13, 0.4; Pakistan, 17, 0.38; Mexico, 22, 0.36; Kyrgyzstan, 23, 0.36; Cote D'Ivoire, 29, 0.35; Ethiopia, 46, 0.31; Sri Lanka, 47, 0.31; Thailand, 63, 0.27; El Salvador, 72, 0.26; Belarus, 78, 0.25; Rwanda, 79, 0.25; Romania, 89, 0.22; Moldova, 95, 0.22; East Timor, 119, 0.15; Montenegro, 120, 0.15; Mauritius, 127, 0.11; Portugal, 130, 0.09; Slovenia, 142, 0.06

Informational: Bolivia, 7, 0.36; Philippines, 20, 0.32; Comoros, 26, 0.29; Iraq, 33, 0.28; CAR, 38, 0.27; Haiti, 53, 0.26; El Salvador, 70, 0.23; Mauritania, 73, 0.22; Slovenia, 91, 0.18; Mauritius, 96, 0.17; Malaysia, 99, 0.16; Belarus, 102, 0.15; Yemen, 103, 0.14; Cyprus, 115, 0.12

Opening

Associational: Ethiopia, 5, 0.44; South Sudan, 7, 0.43; Congo, 22, 0.36; Azerbaijan, 26, 0.36; Uganda, 42, 0.32; Nicaragua, 53, 0.31; Romania, 71, 0.28; Bangladesh, 85, 0.25; Botswana, 97, 0.22; South Korea, 105, 0.19

Economic: Pakistan, 7, 0.4; Romania, 10, 0.39; Benin, 12, 0.38; Dominican Republic, 18, 0.38; Gabon, 32, 0.34; Senegal, 36, 0.33; Zambia, 37, 0.33; Tajikistan, 38, 0.33; Sierra Leone, 39, 0.33; Cameroon, 42, 0.33; Tanzania, 43, 0.33; Nepal, 47, 0.32; South Africa, 66, 0.29; Mauritius, 67, 0.28; Laos, 81, 0.27; Botswana, 87, 0.26; Haiti, 91, 0.25; Israel, 106, 0.23; East Timor, 123, 0.19; Saudi Arabia, 124, 0.19; Slovakia, 130, 0.17

Electoral: Malawi, 10, 0.31; Niger, 12, 0.27

Governing: Uzbekistan, 6, 0.39; Mauritania, 14, 0.36; Congo, 23, 0.32; Bolivia, 25, 0.31; Romania, 31, 0.3; Dominican Republic, 36, 0.29; Turkey, 43, 0.28; Syria, 51, 0.26; Ecuador, 57, 0.25; Togo, 63, 0.24; Moldova, 65, 0.24; Liberia, 68, 0.23; Cyprus, 74, 0.19

Individual: Afghanistan, 21, 0.38; Algeria, 30, 0.35; Dominican Republic, 34, 0.32; Congo, 43, 0.31; Myanmar, 49, 0.3; Egypt, 53, 0.28; Zimbabwe, 69, 0.25; Lesotho, 88, 0.22; Malawi, 100, 0.19

Table 3: Partial accuracy of the v10 forecasts for 2020–2021 with outcomes for 2020 outcomes

Space	Cases	In_top20	AUC-ROC	AUC-PR	Pos_rate
Closing 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
Opening 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

Informational: Sudan, 1, 0.51; South Sudan, 4, 0.43; Ukraine, 14, 0.33; Afghanistan, 19, 0.3; Somalia, 33, 0.27; Guatemala, 73, 0.19; Romania, 79, 0.17; Dominican Republic, 92, 0.15; Argentina, 131, 0.04

Partial scoring of the v10 2020-2021 forecasts

The forecasts made in the spring of 2020 using V-Dem version 10 data cover 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. Table 3 shows the resulting accuracy metrics, and again corresponding separation plots in Figure 2 and lists of the positive cases are below. The values are overall on track to match the v9 accuracy results above. AUC-PR is lower, but this is to be expected since we only have about half the number of positives; note that the AUC-PR values still on average are roughly twice the positive rate in the data, as was the case with the v9 forecast accuracy in Table 2.

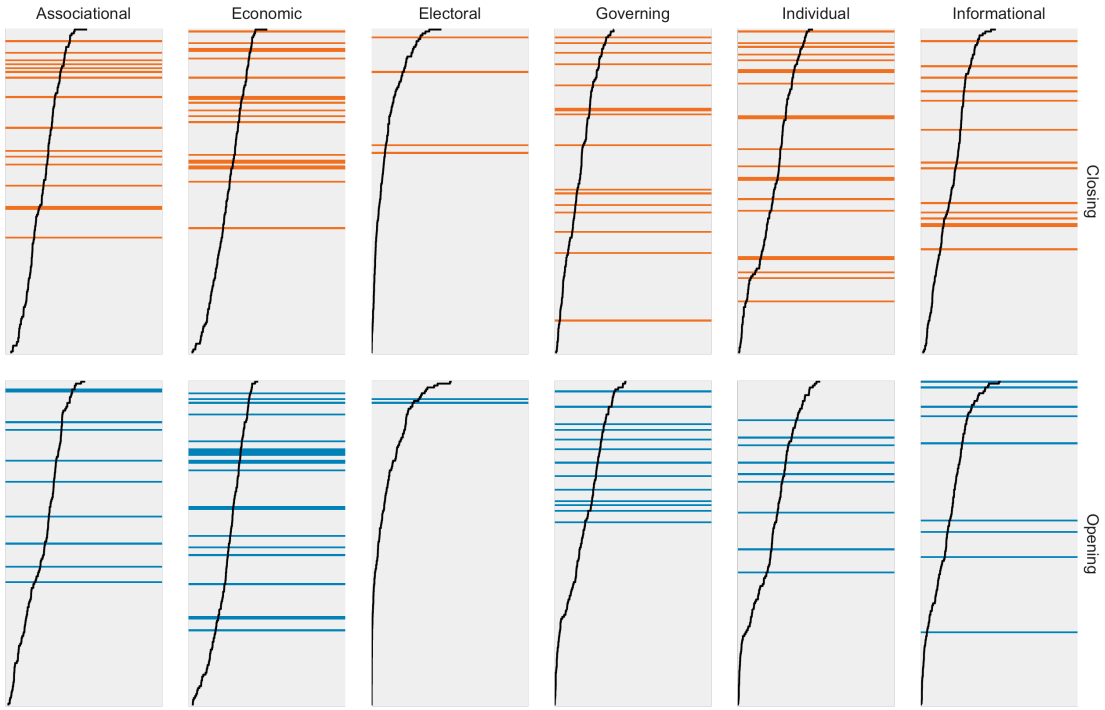
And again, here are all positive cases, due to opening or closing events in 2020:

Closing

Associational: Benin, 7, 0.41; Nigeria, 13, 0.38; Colombia, 17, 0.37; Guatemala, 19, 0.37; Iraq, 21, 0.36; Sri Lanka, 23, 0.35; Mexico, 26, 0.35; Thailand, 36, 0.32; Algeria, 52, 0.3; Mauritania, 65, 0.27; Madagascar, 67, 0.27; Cote D'Ivoire, 71, 0.27; Belarus, 82, 0.24; Bhutan, 93, 0.22; Panama, 94, 0.22; Slovenia, 109, 0.18

Economic: India, 2, 0.43; Lesotho, 8, 0.41; Colombia, 11, 0.4; El Salvador, 12, 0.4; Armenia, 16, 0.39; Sri Lanka, 27, 0.37; Indonesia, 36, 0.35; Ecuador, 37, 0.34; Ukraine, 39, 0.34; Nigeria, 43, 0.33; Comoros, 46, 0.32; Namibia, 49, 0.32; Kenya, 66, 0.3; Bosnia-Herzegovina, 69, 0.29; Mauritania,

Figure 2: Separation plots for the v10 forecasts covering 2020-2021, 2020 outcomes only



70, 0.29; Nicaragua, 71, 0.28; Sudan, 73, 0.28; Somalia, 80, 0.26; CAR, 104, 0.22

Electoral: Guinea, 5, 0.31; CAR, 23, 0.21; Cote D'Ivoire, 61, 0.09; Iran, 65, 0.09

Governing: Benin, 5, 0.34; Kyrgyzstan, 8, 0.32; Philippines, 13, 0.3; Venezuela, 19, 0.28; Sri Lanka, 30, 0.24; Serbia, 42, 0.22; Mali, 43, 0.22; Cote D'Ivoire, 45, 0.22; Yemen, 61, 0.19; Laos, 84, 0.15; Botswana, 86, 0.14; Vietnam, 92, 0.13; Burundi, 96, 0.13; Mauritius, 106, 0.11; Malaysia, 117, 0.09; Spain, 152, 0.03

Individual: Bolivia, 2, 0.45; Uganda, 8, 0.42; Zambia, 10, 0.42; Guinea, 13, 0.4; Pakistan, 17, 0.38; Mexico, 22, 0.36; Kyrgyzstan, 23, 0.36; Cote D'Ivoire, 29, 0.35; Ethiopia, 46, 0.31; Sri Lanka, 47, 0.31; Thailand, 63, 0.27; El Salvador, 72, 0.26; Belarus, 78, 0.25; Rwanda, 79, 0.25; Romania, 89, 0.22; Moldova, 95, 0.22; East Timor, 119, 0.15; Montenegro, 120, 0.15; Mauritius, 127, 0.11; Portugal, 130, 0.09; Slovenia, 142, 0.06

Informational: Bolivia, 7, 0.36; Philippines, 20, 0.32; Comoros, 26, 0.29; Iraq, 33, 0.28; CAR, 38, 0.27; Haiti, 53, 0.26; El Salvador, 70, 0.23; Mauritania, 73, 0.22; Slovenia, 91, 0.18; Mauritius, 96, 0.17; Malaysia, 99, 0.16; Belarus, 102, 0.15; Yemen, 103, 0.14; Cyprus, 115, 0.12

Opening

Associational: Ethiopia, 5, 0.44; South Sudan, 7, 0.43; Congo, 22, 0.36; Azerbaijan, 26, 0.36; Uganda, 42, 0.32; Nicaragua, 53, 0.31; Romania, 71, 0.28; Bangladesh, 85, 0.25; Botswana, 97, 0.22; South Korea, 105, 0.19

Economic: Pakistan, 7, 0.4; Romania, 10, 0.39; Benin, 12, 0.38; Dominican Republic, 18, 0.38;

Gabon, 32, 0.34; Senegal, 36, 0.33; Zambia, 37, 0.33; Tajikistan, 38, 0.33; Sierra Leone, 39, 0.33; Cameroon, 42, 0.33; Tanzania, 43, 0.33; Nepal, 47, 0.32; South Africa, 66, 0.29; Mauritius, 67, 0.28; Laos, 81, 0.27; Botswana, 87, 0.26; Haiti, 91, 0.25; Israel, 106, 0.23; East Timor, 123, 0.19; Saudi Arabia, 124, 0.19; Slovakia, 130, 0.17

Electoral: Malawi, 10, 0.31; Niger, 12, 0.27

Governing: Uzbekistan, 6, 0.39; Mauritania, 14, 0.36; Congo, 23, 0.32; Bolivia, 25, 0.31; Romania, 31, 0.3; Dominican Republic, 36, 0.29; Turkey, 43, 0.28; Syria, 51, 0.26; Ecuador, 57, 0.25; Togo, 63, 0.24; Moldova, 65, 0.24; Liberia, 68, 0.23; Cyprus, 74, 0.19

Individual: Afghanistan, 21, 0.38; Algeria, 30, 0.35; Dominican Republic, 34, 0.32; Congo, 43, 0.31; Myanmar, 49, 0.3; Egypt, 53, 0.28; Zimbabwe, 69, 0.25; Lesotho, 88, 0.22; Malawi, 100, 0.19

Informational: Sudan, 1, 0.51; South Sudan, 4, 0.43; Ukraine, 14, 0.33; Afghanistan, 19, 0.3; Somalia, 33, 0.27; Guatemala, 73, 0.19; Romania, 79, 0.17; Dominican Republic, 92, 0.15; Argentina, 131, 0.04

Discussion

Although the accuracy metrics for both forecast rounds indicate that the forecasts are informative, the absolute values should be better. For the v9 forecasts, 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.

What is particularly interesting is that the test forecasts have much higher accuracy values. We make these specifically to get a sense of what accuracy we can expect. 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. Their accuracy is summarized in Table 4, which has the same format as the other score tables but the values here are average performance over the 12 test forecast years we have, not a single live forecast set.[2] The average AUC-ROC and AUC-PR scores from those were around 0.83 and 0.40, which is noticeably higher than the accuracy of the live forecasts (0.71 and 0.73; 0.25 and 0.15, respectively).

Investigating this discrepancy, it seems that two factors are at work:

First, Covid-19-related government policies. Looking through the lists of cases above, and specially cases where the forecast did poorly, a couple of relatively wealthy and stable European countries stand out. For example, Slovenia experienced closing events for several spaces, but was missed by both the v9 and v10 forecasts. In the v9 forecasts for closing events, it was ranked 150, 157, and 137 for associational, individual, and informational, despite experiencing closing events for those spaces. Similarly in the v10 forecasts: ranks 109, 142, and 91, respectively. Spain also shows up as one of the particularly poor forecasts.

[2]Specifically, we calculate accuracy for each of the 12 sets of test forecasts, then average the accuracy scores. As opposed to pooling all 12 years of test forecasts into one set that we score.

Table 4: Average accuracy of the v9 test forecasts from 2005–2016, scored with v9 V-Dem data

Space	Cases	Top 20	AUC-ROC	AUC-PR	Pos. rate
Closing movement					
Associational	24.8	9.0	0.82	0.42	0.15
Economic	21.4	7.5	0.77	0.35	0.13
Electoral*	4.4	3.2	0.85	0.36	0.03
Governing	12.2	6.5	0.84	0.39	0.07
Individual	21.1	8.2	0.84	0.42	0.13
Informational	17.1	7.7	0.82	0.42	0.10
Opening movement					
Associational	18.6	7.8	0.82	0.38	0.11
Economic	25.8	7.8	0.75	0.37	0.15
Electoral	5.8	4.2	0.88	0.46	0.03
Governing	11.2	6.2	0.85	0.43	0.07
Individual	18.8	8.6	0.84	0.40	0.11
Informational	12.2	6.2	0.86	0.43	0.07
Average					
	16.1	6.9	0.83	0.40	0.10

Note:
All values are averages over performance scores for the 12 distinct test forecasts between 2005–2016.

* There were 0 closing events in 2010; this year is not figured in the average performance calculation.

V-Dem has documented government responses to Covid-19, and identified instances where restrictions have been used to unduly empower the executive, in what they term "pandemic backsliding" (see Lührmann and Rooney 2020; Edgell et al. 2020). It seems very likely that this dynamic is responsible for at least some of the closing events during the period covered by the forecasts. Both Spain and Slovenia were coded as "medium risk" in a related risk assessment for pandemic backsliding[3], and many more countries, mostly outside of Europe, were identified as high risk. Since the pandemic is a one-off event without prior instances to search for patterns, it is however hard to directly incorporate its impact into the forecast data and models. The two forecasts above were in any case made before and during the very early stages of the pandemic, respectively.

The second factor is improvements in the V-Dem data over time, which lead to changes between different versions of the data. These changes between data versions have a large impact on the cases we are trying to forecast. To be clear, the changes to a large extent are due to continual improvements in the quality of the V-Dem data as more country expert survey responses are accumulated. But what this means for the forecasts is that they, essentially, have to be made with data that is fuzzier, less clear, or less accurate than the future version of the V-Dem data with which it is scored. This as a result reduces the accuracy of the forecasts.

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

To understand why changes in V-Dem data versions 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 simple 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 based on *estimates* that are *uncertain*, i.e. explicitly include measurement error and disagreement between country experts.

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. Country experts can also change previous values if warranted. 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-

[3] https://www.v-dem.net/media/filer_public/52/eb/52eb913a-b1ad-4e55-9b4b-3710ff70d1bf/pb_23.pdf

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 (Bayesian uncertainty). 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 deterministically optimized statistical models. But this also means that even with identical data, code, and models, re-running the models 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.

To summarize, the way the V-Dem data are generated impacts the forecast outcomes in two ways:

1. **Data improvements**: Each data update does not just add new year to the data, but also adds survey responses or changes values in existing repsonses/values for past country-years that are used to improve the quality of *all* the data. Thus some of the opening or closing changes identified with one data version may turn out to have been incorrect once more accurate country expert opinions have been incorporated into the final data.
2. **MCMC sampling variation / Bayesian uncertainty**: There is inherent sampling variation in the Bayesian measurement and factor analysis models that are used to create the V-Dem measures, which means that a re-generation of the data, even with identical inputs and models, will produce slightly different outputs. Although these changes are small and not important in a substantive sense, we use thresholds to code opening and closing movements, as a result of which even small perturbations can put a case above or below a threshold.

How much does this impact the "ground truth" with any given data version?

Instances of opening or closing events are overall the exception rather than the norm, in other words most of the data consists of "no large change". As a result, the agreement in the overall data between V-Dem data versions is high, above 96% when we compare v9 to v10 or 11, and v10 to v11. However, it's the positive cases of shifts that are more important, and there the agreement rates are lower.

Table 5 shows how cases that were positive in either the v9 or the v11 data show up in the other data version, respectively. The v9 data had about 1,700 closing events in total, and 68% of those are coded the same way when we use the v11 data. Conversely, with the v11 data there are and additional 276 (292 + 14) closing events that are not in the v9 data, i.e. from the perspective of the v11 data, 80% of the closing events are in the other data version. The overall agreement rate for

Table 5: Comparison between v9 and v11 data positive cases

	v11		
v9	Closing	No change	Opening
Closing	1176	544	9
No change	262	0	300
Opening	14	625	1766

Table 6: Agreement rate for positive cases in different V-Dem data versions

Comparison	Agreement %
v9-v11	62.6
v9-v10	66.0
v10-v11	67.9

positive cases–the number of cases on the diagonal, where both v9 and v11 agree on the change–is 63%. That is a quite dramatic *disagreement* rate. Table 6 shows the overall agreement rates when we compare the v9 to v10 and v10 to v11 data as well. The rates are similar.

Fortunately, it seems that most of the dropoff in agreement is from one year to another, and that the data version two years from now (v11) is not as dramatically different from the current data (v9) as the first update (v10) had been. In other words, the agreement rate seems to stabilize quite quickly around a core of cases that are indisputably significant opening or closing events.

Furthermore, in almost all cases of disagreement, the difference is between an opening or closing event and "no change"; only rarely is there complete disagreement like an "opening event" in one data version and a "closing event" in the other. This only happens 21 times in the roughly 4,700 cases shown in Table 5, for data spanning almost 5 decades.

The underlying reason is that even if two V-Dem data versions have slightly different values for an indicator value, they usually are not *that* different. In terms of the year-to-year changes we use to identify opening or closing changes, even when a change is not large enough for us to code an opening or closing event, more often than not the direction of change is the same anyways.

Figure 3 plots the raw year-to-year changes in the indicator variables that we use to code opening and closing events. Each point is the change in a V-Dem indicator for a country compared to the previous year. On the x-axis we use the v9 data to calculate it, on the y-axis the v11 data. They are highly correlated, as one can see. The gray shaded areas (crosses) in each plot show the thresholds above or below which we would code a change as an opening or closing event. For example, looking at the "Governing" plot, cases to the *right* of the vertical shaded area are opening events in the v9 data; cases *above* the horizontal shaded line are opening events in the v11 data. Thus cases in the top right quadrant are opening events in both the v9 and v11 data. Similarly for closing events. Cases that fall within a shaded area are coded as "no event" in the corresponding data version. Note that the shaded center square where the horizontal and vertical shaded areas overlap would be cases coded as "no event" in both data versions. They are empty because we have left out all of those cases in the plots.

Cases/points in which the data versions disagree on the event coding are colored in red. The worst possibility are cases falling into the top-left or bottom-right quadrants. Here the data versions completely disagree on what happened: one tells us opening, the other closing, like point "1: Gambia 1995" in the "Governing" plot. As we mentioned earlier these are rare. Instead, the most likely disagreement is that one version codes an opening or closing event, but for the other we have "no event", i.e. a point that falls into the gray area. But there is still a meaningful distinction for these cases, namely whether the direction of change was still the same or different. Point 2, the DRC in 1992, is coded as a closing event in the v9 data and "no event" in v11. Furthermore, while the year-to-year change in v9 is negative, in v11 it's slightly positive. Thus there is fundamental disagreement on the direction of change between data versions. One other other hand, point 3, Pakistan in 2002, shows a case where both data versions agree that it was a negative (closing) change, but differ in the magnitude of change. Most cases of disagreement are like point 3, i.e. agreement on the direction of change but in only one is the threshold for our coding crossed. Specifically, generally more than 80% of cases are like that, except for the economic space, where it's 57%.

What this shows is that most of the time when there is disagreement over whether an opening or closing event occurred in different data versions, the underlying movement in the democratic space was still the same–just a little bit less or more of it, but in the same direction. So although these kinds of difference between data versions are in a technical sense mis-predictions, substantively they are generally still somewhat correct.

In the conclusion we will discuss some possibilties for getting around this issue.

Forecasts for 2021-2022

Figures 4 through 9 show the 30 highest closing and opening forecasts for each space. The full set of forecasts can be explored at the dashboard at https://www.v-dem.net/en/analysis/DemSpace/ as well. Given the problem in using the test forecasts as a measure of expected accuracy, it is not included in this note. However, it does match the test forecast accuracy from the previous forecast rounds, and actually slightly improved as a result of streamlining the dataset of predictors during the course of the 2021 update.[4]

There are three general points about the forecasts to note:

Some countries have high forecasts for both the possibility of opening and closing events. The overall correlation in the forecasts for opening and closing events is around 0.56—not high, but also not low. This is related to a country's past history of opening or closing events, and specifically how variable the corresponding indicators have been. For example, Thailand has high forecasts in the associational space (Figure 5) for both opening and closing events: in the last 10 years it has also experienced 3 opening and 4 closing events (using the v11 data).

Some spaces are more stable than other. The forecasts thus are correspondingly also more or less dramatic in response. This accounts in part for the differences in the highest probabilities forecasted in the different spaces, i.e. while the highest closing forecast in the electoral space is at a probability of around 0.3, some forecasts in the other spaces exceed 0.7. Table 7 shows the total number of

[4]The goal of the data streamlining was to make future updates easier and quicker. The number of predictors was cut in half, from more than 400 to slightly more than 200. This was done on the basis of random forecast variable importance score. More details are available at https://github.com/vdeminstitute/demspaces/blob/main/2021-update/variable-importance.md.

Figure 3: Comparison of year-to-year changes in each space in the v9 and v11 data versions

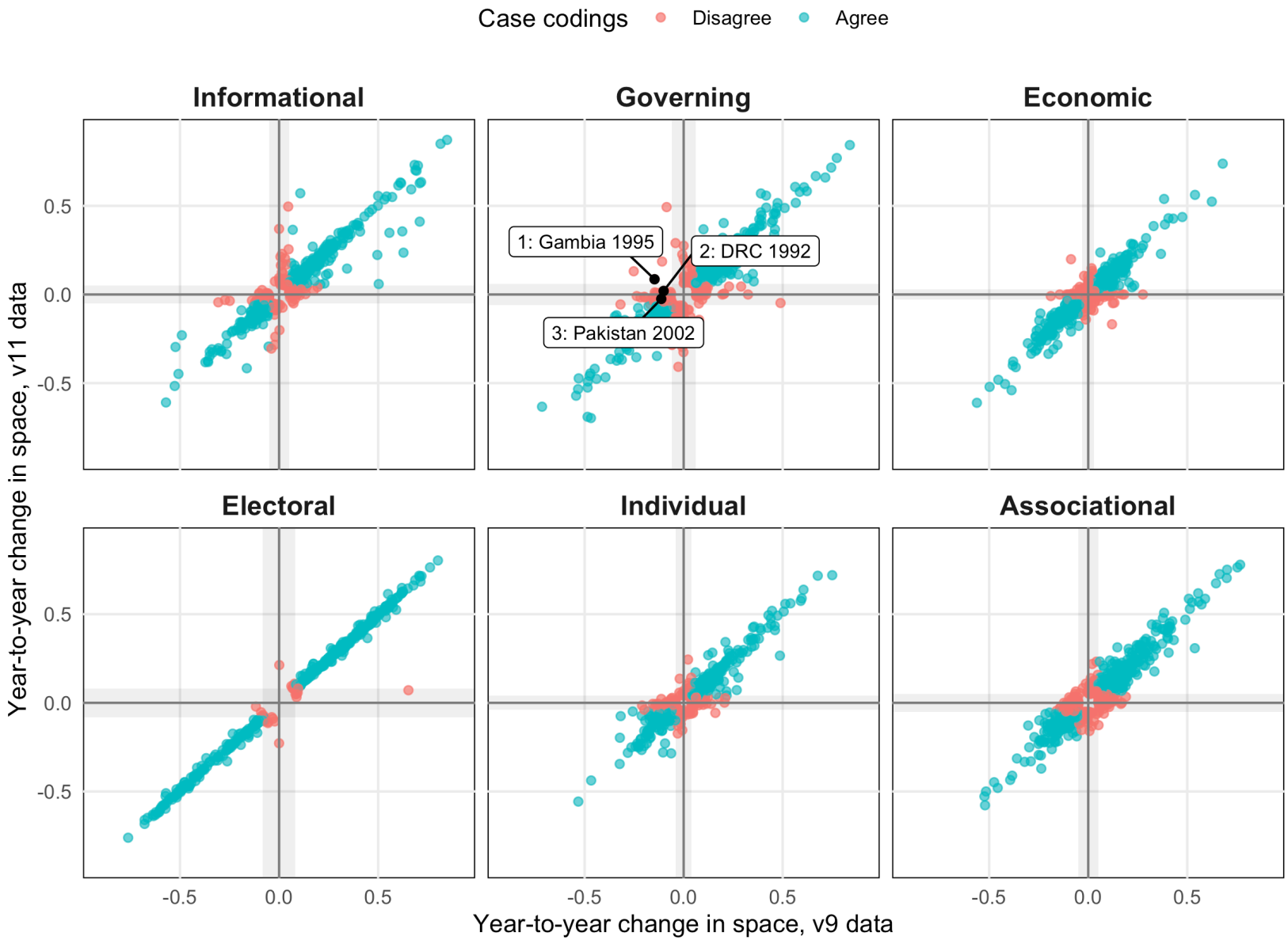


Figure 4: Electoral space

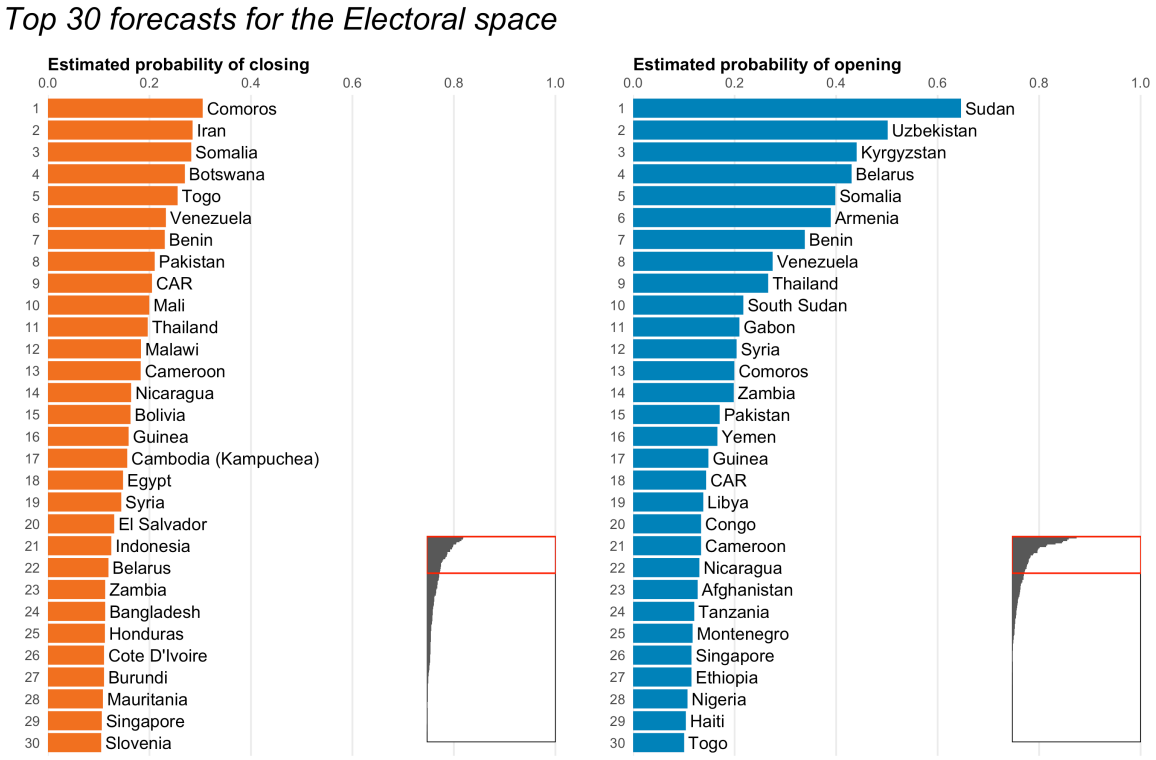


Figure 5: Associational space

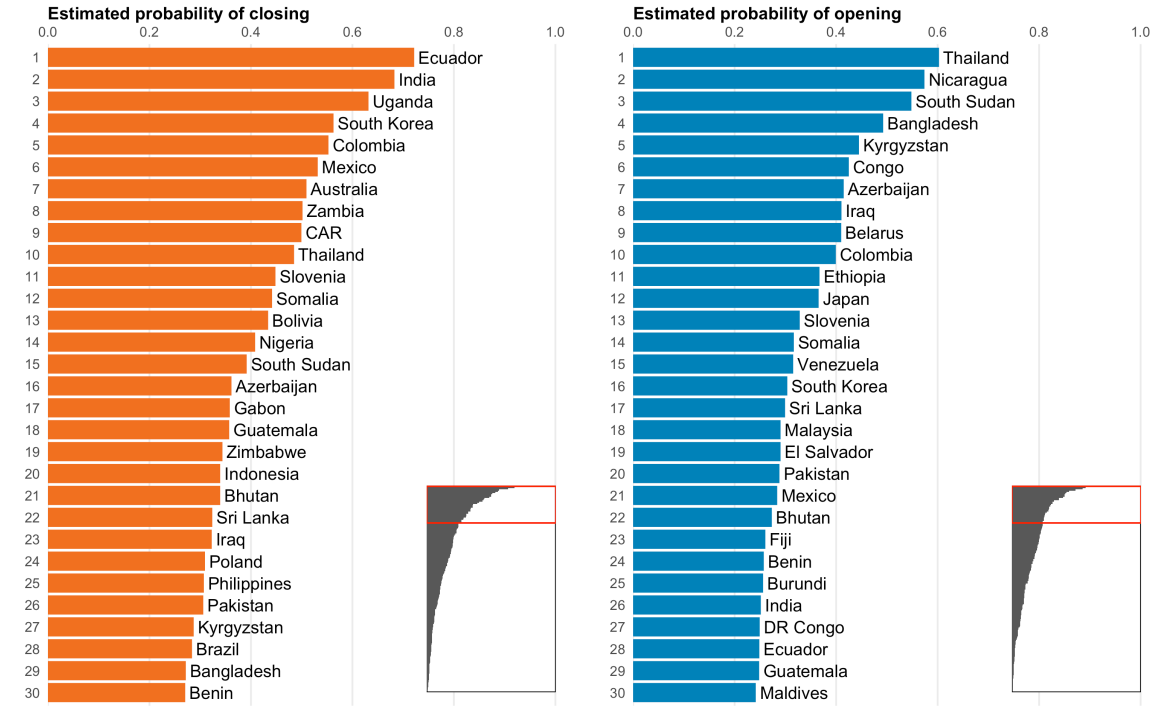


Figure 6: Individual space

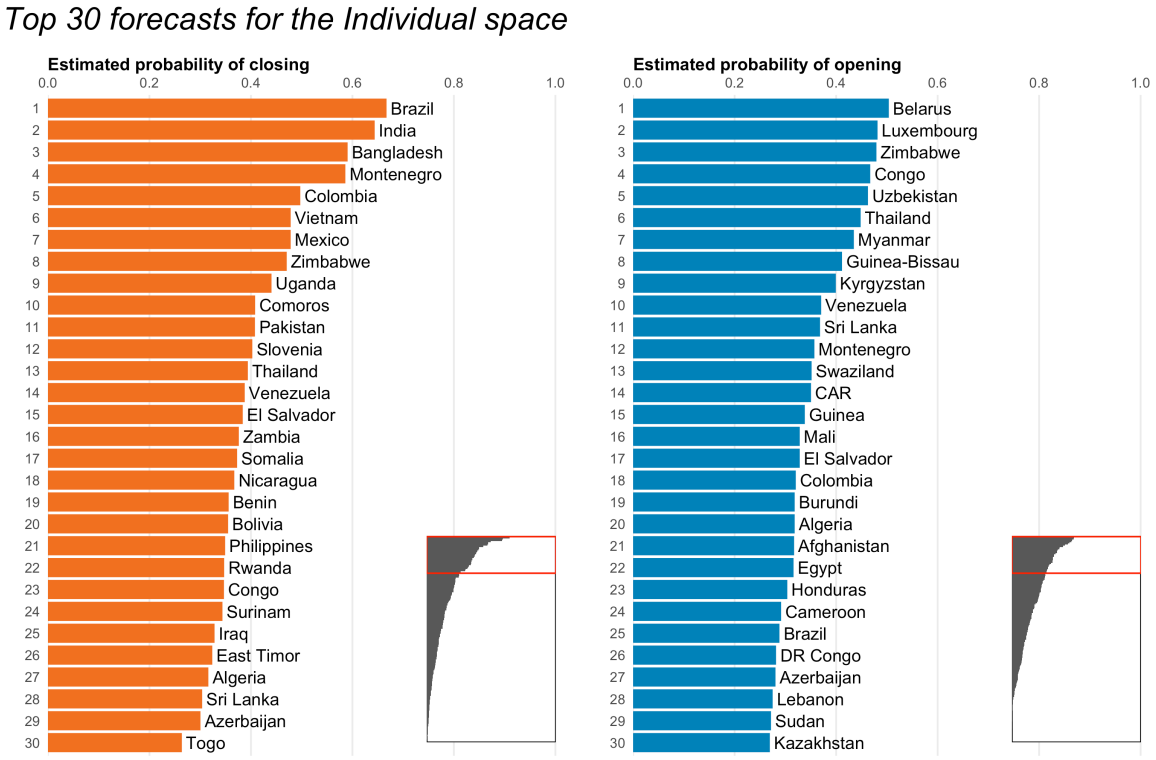


Figure 7: Informational space

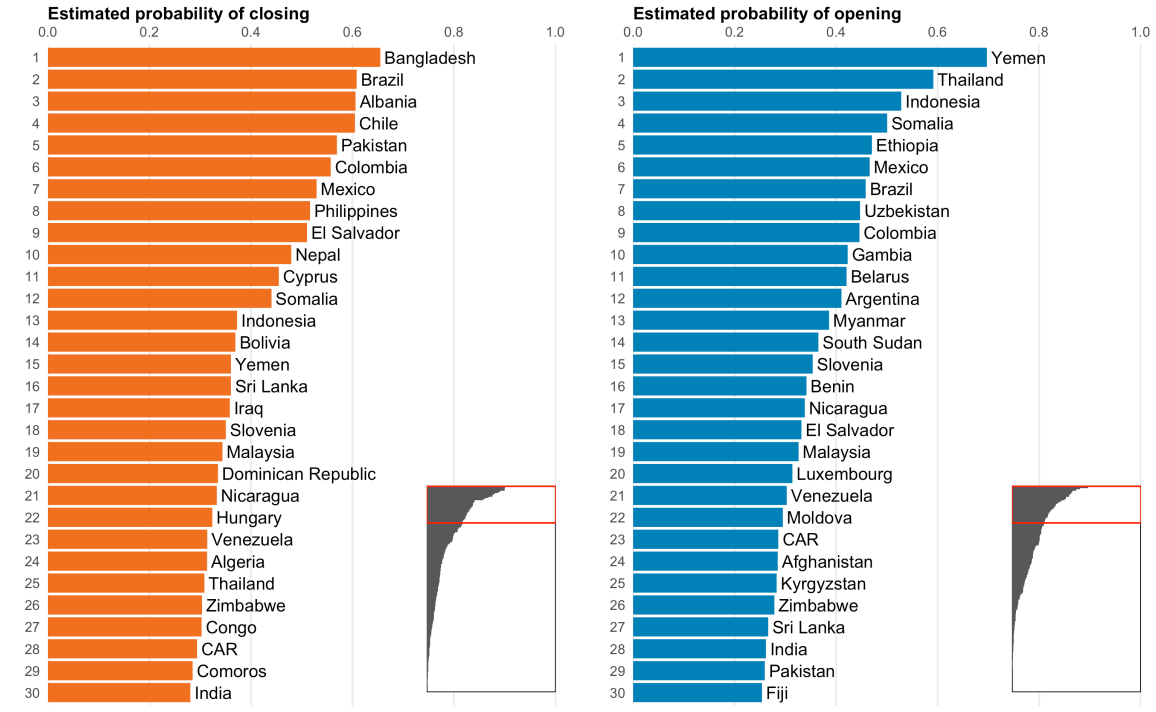


Figure 8: Governing space

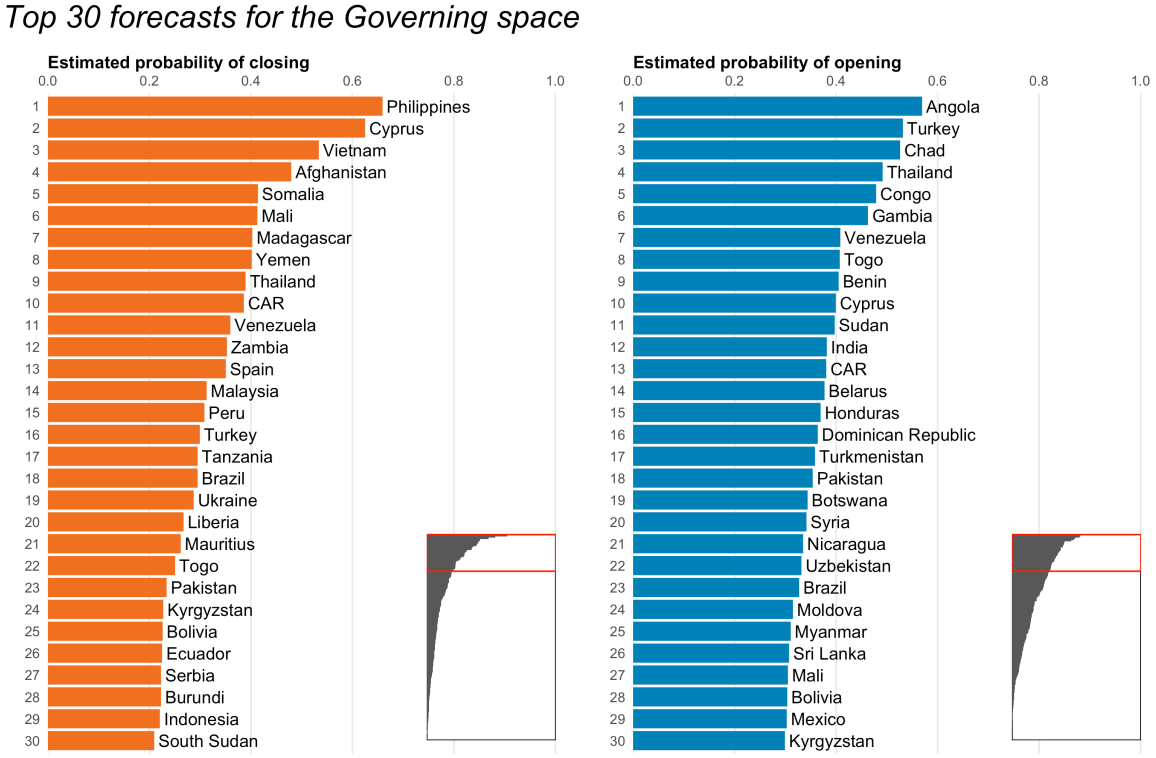


Figure 9: Economic space

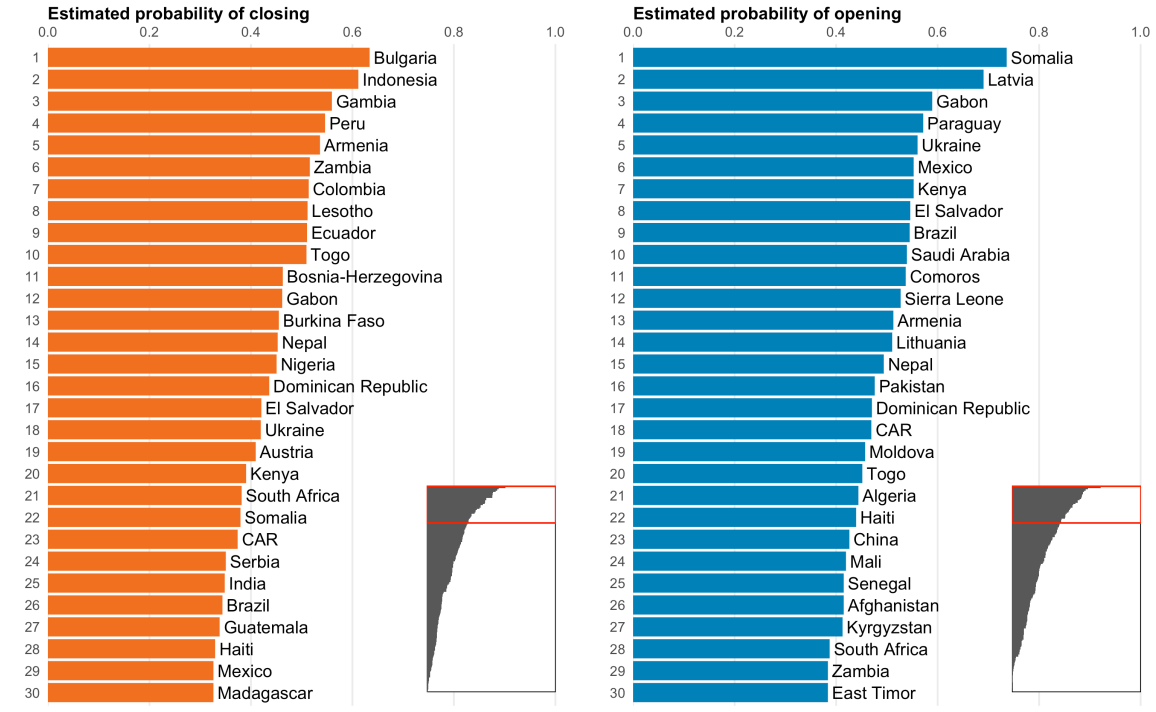


Table 7: Opening and closing events and event rates per thousand, 1970–2020, using v11 data

	Number			Rater per 1,000		
Space	Open	Close	Total	Open	Close	Total
Associational	451	285	736	56.2	35.5	91.8
Economic	371	441	812	46.3	55.0	101.3
Electoral	247	170	417	30.8	21.2	52.0
Governing	374	234	608	46.6	29.2	75.8
Individual	380	245	625	47.4	30.6	77.9
Informational	389	242	631	48.5	30.2	78.7
Average						
	369	270	638	46.0	33.6	79.6

events for each space, as well as the rate of events per 1,000 cases, for the spaces. The associational and economic spaces have notably more fluctuation, while the electoral space is much more stable. These patterns roughly are reflected also in the spread of forecast probabilities across the different spaces in the figures above.

Thirdly, it is very likely that Covid-19 plays some role in movement within the democratic spaces, and closing movements specifically, as discussed above. Given the nature of pandemic response policies, the associational, individual, and informational spaces should be those most impacted by the pandemic; the electoral and governing spaces depend more on structural factors that should not change that fast, while the economic space, regarding the absence of public corruption, maybe is impacted somewhere between those two groups. It is difficult to directly incorporate any effects of the pandemic into the existing forecasting process, as we are just learning what those impacts are. But it probably is useful to analytically assess the democratic spaces forecasts in combination with the pandemic response violations of democratic standards index V-Dem is collecting.[5] For example, the two highest forecasts for a closing event in the associational space are Ecuador and India, with similar forecast probabilities. However, Ecuador has so far had only minor pandemic response violations while India has had major violations. Thus we should probably weigh the closing forecast for India higher and discount the one for Ecuador.

Conclusion

The picture emerging so far is that the democratic spaces forecasts are an informative addition to the kind of naive base rate forecasts that might anchor a judgment otherwise. At the same time, the accuracy results so far are notably lower than those we had expected based on more extensive test forecasts conducted along with each forecast set. One reason for the discrepancy probably is the impact that Covid-19 has had both on autocratization, and breaking historical patterns in general. This is idiosyncratic and difficult to incorporate into the forecasting process until some time has passed for a clearer picture of the pandemic's impact to crystallize, but we made some suggestions for how to incorporate information on pandemic backsliding risks into an assessment of the democratic spaces forecasts.

The other challenge arises due to the continuing improvements in the V-Dem over time, which

[5]Dashboard available at https://www.v-dem.net/en/analysis/PanDem/.

practically for this forecasting process means that the outcome data used to steer the forecasting models suffer from inaccuracies in respect to the subsequent data version that two years down the road is used to assess the forecasts.

There are two potential avenues one could explore for ameliorating this issue. First, one of the source of data version changes—Bayesian uncertainty—can in principle be reduced at the expense of computational time and resources. It is not clear to which extent data version differences are due to MCMC sampling variation (Bayesian uncertainty) rather than changes in the set of survey responses, and this question should thus be answered first. The second possibility is to directly model changes in the V-Dem indicators that we use to measure the spaces. This could be either internally at the model level, while still retaining discrete "opening/closing/no event" outcomes, or it could be more fundamentally by focusing on raw changes as the outcome of interest. It's likely that there would be substantively important tradeoffs and limitations with such an approach that need to be taken into consideration.

References

Beger, Andreas, Richard K. Morgan, and Laura Maxwell. 2020. "Democratic Spaces Barometer."

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_v11.1.pdf.

Edgell, Amanda B., Anna Lührmann, Seraphine F. Maerz, Jean Lachapelle, Sandra Grahn, Ana Flavia Good God, Martin Lundstedt, et al. 2020. *Pandemic Backsliding: Democracy During Covid-19 (Pandem), Version 5*.

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

Lührmann, Anna, and Bryan Rooney. 2020. "Autocratization by Decree: States of Emergency and Democratic Decline." https://www.v-dem.net/media/filer_public/31/1d/311d5d45-8747-45a4-b46f-37aa7ad8a7e8/wp_85.pdf.