Automated Democracy Scores*

Thiago Marzagão**

Abstract

This paper uses natural language processing to create the first machine-coded democracy index, which I call Automated Democracy Scores (ADS). The ADS is based on 42 million news articles from 6,043 different sources and cover all independent countries in the 1993-2012 period. Unlike the democracy indices we have today the ADS is replicable and has standard errors small enough to actually distinguish between cases.

Keywords: Democracy; Measurement; Natural language processing.

JEL Codes: C44, C63, C80.

*Submitted in May 2016. Revised in March 2017.

**Data scientist at the Office of the Comptroller-General, Brazil. The entirety of this work was done when the author was a Ph.D. student at the Department of Political Science of the Ohio State University. E-mail: thiago.marzagao@gmail.com

1. Introduction

Democracy is a central variable in economics and other social sciences.[1] For instance, Barro (1999) argues that "Democracies that arise without prior economic development [...] tend not to last." (160). Rodrik (1999) argues that wages are higher under democracies. Bueno de Mesquita et al. (2006) argue that democracy makes the government supply more public goods. Persson and Tabellini (2006) argue that premature democratization leads to lower economic growth. To carry out such investigations we need democracy measures. But can we rely on the measures we have today? And if not, how can we improve on them?

There are at least twelve democracy indices today (Pemstein, Meserve, and Melton 2010), the most popular of which is the Polity (Marshall, Gurr, and Jaggers 2013), which assigns a score from –10 to +10 to each of 167 countries in each year of the 1946-2013 period. Another popular democracy index is the Freedom House one (Freedom House, 2013), which assigns a score from 1 to 7 to each of 195 countries in each year of the 1972-2013 period.

All democracy indices draw to some extent from Dahl's (1972) conceptualization: democracy as a mixture of competition and participation. Yet none are replicable or provide adequate measures of uncertainty. All democracy indices we have today rely directly or indirectly on country experts checking boxes on questionnaires. We do not observe what boxes they are checking, or why; all we observe are the final scores. The process is opaque and at odds with the increasingly demanding standards of openness and replicability of academia. More importantly – and potentially fatal for economic analyses –, opacity makes it easy for country experts to boost the scores of countries that adopt the 'correct' policies.

Coding rules help, but still leave too much open for interpretation. For instance, consider this excerpt from the Polity handbook (Marshall, Gurr, and Jaggers 2013): 'If the regime bans all major rival parties but allows minor political parties to operate, it is coded here. However, these parties must have some degree of autonomy from the ruling party/faction and must represent a moderate ideological/philosophical, although not political, challenge to the incumbent regime.' (73). How do we measure autonomy? Can we always observe it? What is 'moderate'? Clearly it is not hard to smuggle ideological contraband into democracy scores.

Ideological biases, in turn, make empirical tests circular. If we find an association between democracy and some economic policy x is that a genuine association

[1]This work was funded by the Fulbright (grantee ID 15101786), by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES (BEX 2821/09-5), and by the Ministério do Planejamento, Orçamento e Gestão – MPOG (proceed n. 03080.000769/2010-53). This work was supported in part by an allocation of computing time from the Ohio Supercomputer Center. I thank Sarah Brooks, Irfan Nooruddin, Marcus Kurtz, Janet Box-Steffensmeier, Philipp Rehm, Paul DeBell, Margaret Hanson, Carolyn Morgan, Peter Tunkis, Vittorio Merola, Raphael Cunha, and Marina Duque for their helpful feedback. All errors are mine.

or an artifact of human coders' preferences regarding x? With the democracy measures we have today it is hard to know. (Naturally, having an unbiased measure of democracy does not solve the reverse causality problem: the ADS is purged of ideological bias, it is not purged of any effects x may have on democracy.)

Another problem with existing indices is the lack of proper standard errors. The two most popular indices – the Polity and the Freedom House – only give us point estimates, without any measure of uncertainty. That prevents us from knowing, say, whether Uruguay (Polity score = 10) is really more democratic than Argentina (Polity score = 8) or whether the uncertainty of the measurement process is sufficient to make them statistically indistinguishable. In other words, we cannot do descriptive inference.

Moreover, without standard errors we cannot do causal inference when democracy is one of the regressors. As Treier and Jackman (2008) warn, 'whenever democracy appears as an exploratory variable in empirical work, there is an (almost always ignored) errors-in-variables problem, potentially invalidating the substantive conclusions of these studies' (203).

Only one (publicly available) measure has standard errors: the Unified Democracy Scores (UDS), created by Pemstein, Meserve, and Melton (2010). To produce the UDS Pemstein, Meserve, and Melton (2010) treated democracy as a latent variable and used a multirater ordinal probit model to extract that latent variable from twelve different democracy measures (among which the Polity and the Freedom House). The UDS comes with point estimates (posterior means) and confidence intervals (posterior quantiles).

The UDS is a big improvement on all other measures, but its standard errors are too large to be useful. 70% of the countries are all statistically indistinguishable from each other (in the year 2008 – the last year in the UDS dataset at the moment of writing); pairs as diverse (regime-wise) as Denmark and Suriname, Poland and Mali, or New Zealand and Mexico have overlapping confidence intervals.

Hence we need a better way to measure democracy. In this paper I argue that natural language processing can help us achieve that.

2. Method

2.1 Overview

The basic idea is simple. News articles on, say, North Korea or Cuba contain words like 'censorship' and 'repression' more often than news articles on Belgium or Australia. Hence news articles contain quantifiable regime-related information that we can use to create a democracy index.

I adopt a *supervised learning* approach. In supervised learning we feed the machine a number of pre-scored cases – the training data. The machine then 'learns' from the training data. In text analysis that means learning how the frequency of each word or topic varies according to the document scores. For

instance, the algorithm may learn that the word 'censorship' is more frequent the lower the democracy score of the document. Finally, the algorithm uses that knowledge to assign scores to all other cases – i.e., to the test data.

We can use a number of supervised learning methods. Here I use the Wordscores algorithm, created by Laver et al. (2003), which I explain in a moment.

2.2 Corpus

I use a total of 6,043 news sources. These are all the news sources in English available on LexisNexis Academic, which is an online repository of journalistic content. The list includes American newspapers like The New York Times, USA Today, and The Washington Post; foreign newspapers like The Guardian and The Daily Telegraph; news agencies like Reuters, Agence France Presse (English edition), and Associated Press; and online sources like blogs and TV stations' websites.

I use LexisNexis' internal taxonomy to identify and select articles that contain regime-related news. In particular, I choose all articles with one or more of the following tags: 'human rights violations' (a subtag of 'crime, law enforcement and corrections'); 'elections and politics' (a subtag of 'government and public administration'); 'human rights' (a subtag of 'international relations and national security'); 'human rights and civil liberties law' (a subtag of 'law and legal system'); and 'censorship' (a subtag of 'society, social assistance and lifestyle').

LexisNexis' news database covers the period 1980-present (though actual coverage varies by news source), so in principle the ADS could cover that period as well. In practice, however, LexisNexis does not provide search codes for countries that have ceased to exist, so we cannot reliably retrieve news articles on, say, the Soviet Union or East Germany (we could search by the country's name but that yields unreliable results – think of Turkey, for instance). Hence I limit myself to the 1992-2012 interval.

That selection – i.e., regime-related news, all countries that exist today, 1992-2012 – results in a total of about 42 million articles (around 4 billion words total), which I then organize by country-year. To help reduce spurious associations I remove proper nouns (that should help prevent, for instance, 'Washington' being associated with high levels of democracy just because the word appears frequently on news stories featuring a democratic country), in a probabilistic way (if all occurrences of the word are capitalized then that is probably a proper noun and therefore it is removed).

For each country-year I merge all the corresponding news articles into a single document and transform it into a term-frequency vector. I then merge all vectors together in a big term-frequency matrix.

The period 1992-2012 gives us a total of 4,067 country-years. I choose the year 1992 for the training data and extract the corresponding scores from the Unified Democracy Scores – UDS (Pemstein, Meserve, and Melton 2010). The

UDS have data on 184 countries for the year 1992. Hence we have 184 samples in the training data and 3,883 (4,067 – 184) samples in the test data. I select the year 1992 simply because it is the first year in our dataset. I select the UDS because it is an amalgamation of several other democracy scores, which reduces measurement noise.

2.3 Wordscores

The Wordscores algorithm was created by Laver et al. (2003) – henceforth LBG. We begin by computing scores for each word. Let F_{wt} be the relative frequency of word w on training document t. The probability that we are reading document t given that we see word w is then $P(t|w) = F_{wt}/\sum_t F_{wt}$. We let A_t be the *a priori* score of training document t and compute each word score as $S_w = \sum_t (P(t|w) \cdot A_t)$.

A concrete example may help. Suppose that we choose North Korea 2012 and Belgium 2012 as our reference cases and assign them democracy scores of 0 and 10 respectively. We merge all news articles on North Korea in 2012 into a single document and merge all news articles on Belgium in 2012 into another document. Suppose now that the word "censorship" accounts for 15% of all the words in the North Korea document and for 1% of the words in the Belgium document. If we see the word "censorship" the probability that we are reading the North Korea document is $0.15/(0.15 + 0.01) = 0.9375$ and the probability that we are reading the Belgium document is $0.01/(0.15 + 0.01) = 0.0625$. The score of the word "censorship" is thus $(0.9375 \cdot 0) + (0.0625 \cdot 10) = 0.625$.

The second step is to use the word scores to compute the scores of the test documents (also called 'virgin' documents). Let F_{wv} be the relative frequency of word w on virgin document v. The score of virgin document v is then $S_v = \sum_w (F_{wv} \cdot S_w)$. To score a virgin document we simply multiply each word score by its relative frequency and sum across.

The third step is the computation of uncertainty measures for the point estimates. LBG propose the following measure of uncertainty: $\sqrt{V_v}/\sqrt{N^v}$, where $V_v = \sum_w F_{wv}(S_w - S_v)^2$ and N^v is the total number of virgin words. The V_v term captures the dispersion of the word scores around the score of the document. Its square root divided by the square root of N^v gives us a standard error, which we can use to assess whether two cases are statistically different from each other.

The fourth and final step is the re-scaling of the test scores. In any given text the most frequent words are stopwords ('the', 'of', 'and', etc). Because stopwords have similar relative frequencies across all reference texts they will have centrist scores. That makes the scores of the virgin documents 'bunch' together around the middle of the scale; their dispersion is just not in the same metric as that of the training documents.

To correct for the 'bunching' of test scores LBG propose re-scaling these as

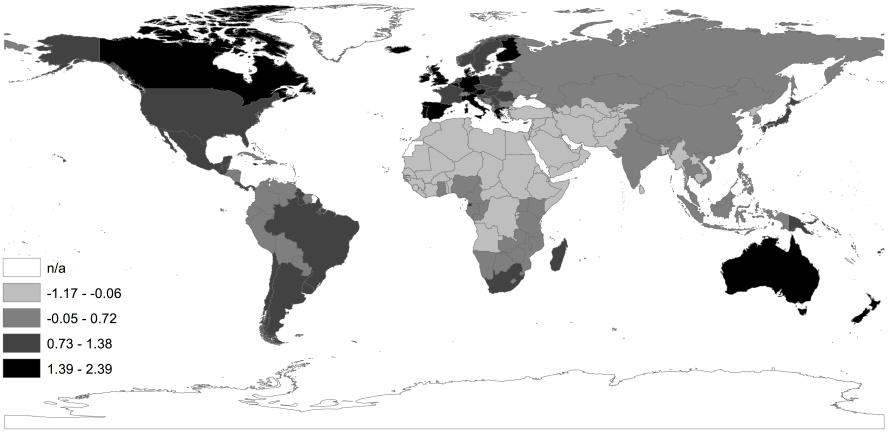
follows: $S_v^* = (S_v - S_{\bar{v}})(\sigma_t/\sigma_v) + S_{\bar{v}}$, where S_v is the raw score of virgin document v, $S_{\bar{v}}$ is the average raw score of all virgin documents, σ_t is the standard deviation of the training scores, and σ_v is the standard deviation of the virgin scores. This transformation expands the raw virgin scores by making them have the same standard deviation as the training scores. Martin and Vanberg (2008) propose an alternative re-scaling formula, but Benoit and Laver (2008) show that the original formula is more appropriate when there are many test samples and few training samples, which is the case here.

I also tried several other algorithms – different combinations of topic-extraction methods (Latent Semantic Analysis and Latent Dirichlet Allocation) and tree-based regression. These results are available online.[2] Wordscores outperforms all alternatives though.

3. Results: the Automated Democracy Scores

The full 1993-2012 dataset is available for download.[3] Figure 1 gives an idea of the ADS distribution in 2012.

Figure 1
Automated Democracy Scores, 2012
(range limits are Jenks natural breaks)



As expected, democracy is highest in Western Europe and in the developed

[2] `https://s3.amazonaws.com/thiagomarzagao/Marzagao_AlternativesToWordscores.pdf`

[3] `https://s3.amazonaws.com/thiagomarzagao/ADS.csv` There is also a web app where people can replicate and tweak the data-generating process: `http://democracy-scores.org` For now the app is restricted to pre-authorized users, as it runs on a commercial cloud plataform and therefore costs money every time someone hits "submit". I intend to secure funding to lift that restriction in the future, so that anyone can use the app.

portion of the English-speaking world, and lowest in Africa and in the Middle East.

Figure 2 shows that the ADS follow a normal distribution.

Figure 2
Automated Democracy Scores, 1993-2012
(with normal distribution)

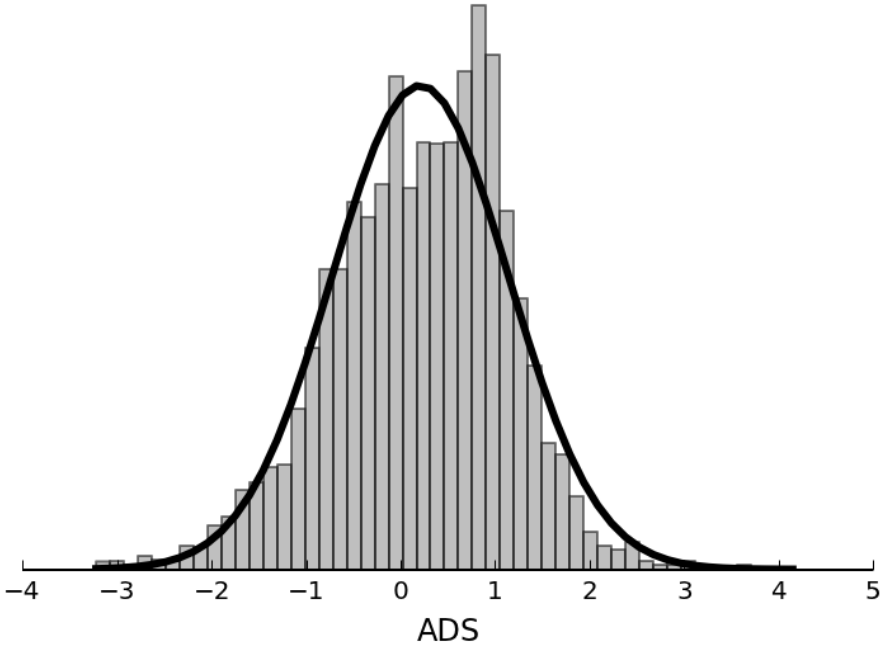


Table 1 shows the ADS summary statistics by year.

As expected, the average ADS increases over time, from 0.006 in 1993 to 0.495 in 2012. That reflects the several democratization processes that happened over that period. We observe the same change in other democracy indices as well. Between 1993 and 2012 the average Polity score (polity2) increased from 2.24 to 4.06 and the average Freedom House score (civil liberties + political rights) decreased from 7.46 to 6.63 (Freedom House scores decrease with democracy); the average UDS score increased from 0.21 to 0.41 between 1993 and 2008, the last year in the UDS dataset.

Also as expected, the standard errors decrease with press coverage. The larger the document with the country-year's news articles, the narrower the corresponding confidence interval. As Figure 3 shows, that relationship is not linear though: after 500KB or so the confidence intervals shrink dramatically and do not change

Table 1
ADS summary statistics, by year

	N	mean	std. dev.	min.	max.
1993	193	0.0061666	1.40437	–3.20916	3.81217
1994	193	0.0939503	1.36697	–2.6985	4.14979
1995	193	0.1005004	1.073329	–2.99738	2.36396
1996	193	-0.1076104	1.128553	–3.22593	2.26484
1997	193	-0.0159435	1.25768	–2.93822	3.03361
1998	193	0.0088406	1.150099	–2.54625	2.7043
1999	193	-0.0999732	1.134464	–2.9453	2.63257
2000	193	0.2312175	0.7445582	–1.31987	2.66054
2001	193	0.2222522	0.7182253	–1.29777	1.92263
2002	194	0.2400814	0.735135	–1.18534	2.33285
2003	194	0.2121506	0.7185639	–1.3477	2.50623
2004	194	0.2213473	0.645	–1.69878	2.03608
2005	194	0.3315942	0.6461306	–1.08297	2.19639
2006	195	0.2869473	0.6760403	–1.28804	2.18348
2007	195	0.3678394	0.7192703	–1.11441	2.4193
2008	196	0.3860345	0.7002583	–1.11659	2.58216
2009	196	0.3212706	0.6923328	–1.487	2.34994
2010	196	0.4233154	0.6748002	–1.08075	2.29522
2011	196	0.4015369	0.7163083	–1.15564	2.38172
2012	196	0.4958635	0.7909505	–1.16859	2.38636
all	3883	0.2073097	0.9338698	–3.22593	4.14979

much afterwards, not even when the document has 15MB or more.

The ADS has much smaller standard errors than the UDS (the only other democracy index that also comes with standard errors). On average, each country in the ADS dataset in the year 2008 overlaps with other 4.49 countries; in the UDS dataset that average is 99.67. The ADS confidence intervals tend to be larger the less press coverage the country gets, but in all cases they are smaller than the corresponding UDS ones.[4]

3.1 The ADS vs other indices

The ADS point estimates correlate 0.7439 with the UDS' (posterior means), 0.6693 with the Polity's (polity2), and –0.7380 with the Freedom House's (civil liberties + political rights). Table 2 below breaks down these correlations by year.

As we see, the correlations do not vary much over time. This is a good sign: it means that the ADS is not overly influenced by the idiosyncrasies of the year 1992, from which we extract the training samples. Otherwise we would see the

[4]For instance, in the UDS the United States is statistically indistinguishable from 80 other countries, whereas in the ADS the United States is statistically indistinguishable from only one other country (Solomon Islands, which rarely appears in the news and thus has a wide confidence interval). The country with most overlaps in the UDS data is Sao Tome and Principe, which is statistically indistinguishable from 135 other countries. That makes 70% of the UDS scores (for 2008) statistically the same. The worst case in the ADS is Czech Republic, which overlaps with 25 other countries (in the UDS Czech Republic overlaps with 110 other countries).

Figure 3
ADS range and press coverage
(ADS range = 95% upper bound minus 95% lower bound.)

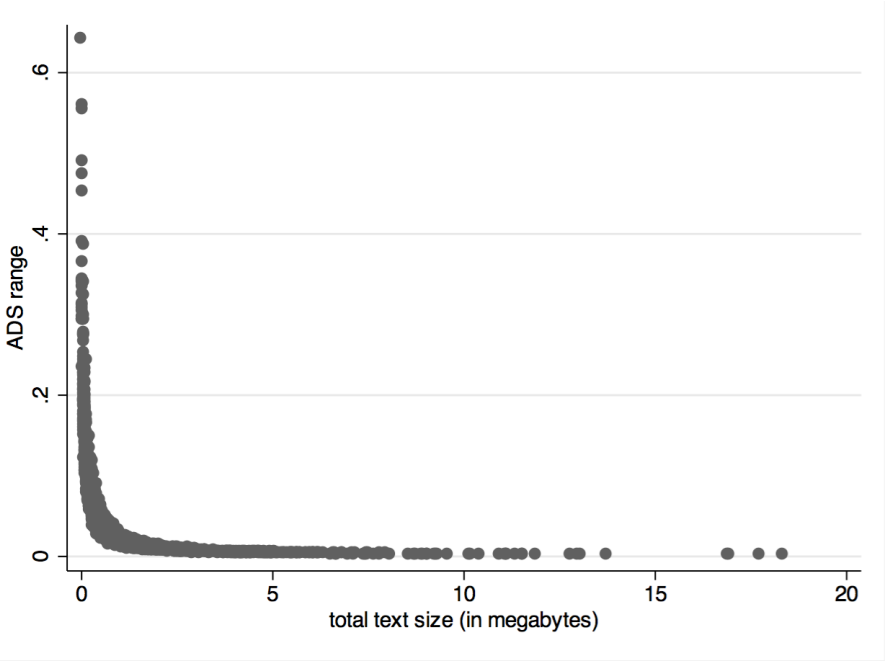


Table 2
Correlation between ADS and other indices, by year

	UDS	Polity	FH		UDS	Polity	FH
1993	0.8021	0.7279	-0.7677	2003	0.7470	0.6610	–0.7445
1994	0.7921	0.6947	-0.7574	2004	0.7493	0.6635	–0.7553
1995	0.7797	0.7221	-0.7650	2005	0.7702	0.6833	–0.7632
1996	0.7783	0.7457	-0.7812	2006	0.7140	0.6458	–0.7596
1997	0.8059	0.7647	-0.8001	2007	0.6982	0.6207	–0.7413
1998	0.8052	0.7355	-0.7864	2008	0.7377	0.6363	–0.7506
1999	0.7729	0.7260	-0.7714	2009	n/a	0.6353	–0.7627
2000	0.7491	0.6794	-0.7579	2010	n/a	0.6467	–0.7791
2001	0.7641	0.6881	-0.7948	2011	n/a	0.6472	–0.7661
2002	0.7668	0.6793	-0.7875	2012	n/a	0.6155	–0.7603

correlations decline sharply after 1993. The correlations do not vary much across indices either, other than being somewhat weaker for the Polity data. This is also a good sign: it means that the ADS is not overly influenced by the idiosyncrasies of the UDS, from which we extract the training scores.

I also ran the algorithm using other years (rather than 1992) for the training data, using UDS as well. I also ran the algorithm using multiple years (up to all years but one) for the training data, again using UDS. Finally, I also ran the algorithm using not the UDS but the Polity and Freedom House indices for the training data. In all these scenarios the correlations remained in the vicinity of 0.70. This corroborates Klemmensen, Hobolt, and Hansen's (2007) finding that Wordscores' results are robust to the choice of training data.

(The samples we use for the training data cannot be used for the test data. For instance, in one scenario I used every other year for the training data, starting with 1992. In that scenario the training data was thus [1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012] and the test data was [1993 1995 1997 1999 2001 2003 2005 2007 2009 2011]. To compute the correlations with other indices I only used the test data.)

Country-wise, what are the most notable differences between the ADS and the UDS? Table 3 shows the largest discrepancies.

Table 3
Largest discrepancies between ADS and UDS

largest positive differences				largest negative differences			
	ADS	UDS	Δ		ADS	UDS	Δ
Swaziland2007	1.53	–1.13	2.66	Israel1994	–1.71	0.97	–2.69
Liechtenstein1994	4.14	1.56	2.58	Israel1993	–1.70	0.97	–2.67
Liechtenstein1993	3.81	1.57	2.23	Israel1999	–1.20	1.44	–2.65
Ireland1994	3.08	1.17	1.90	Israel1997	–1.36	1.07	–2.44
Andorra1993	2.41	0.60	1.80	Benin1993	–1.79	0.49	–2.28
Luxembourg1994	3.29	1.51	1.77	Israel1998	–1.17	1.08	–2.26
Bhutan1996	–0.21	–1.97	1.75	Yemen1993	–2.60	–0.40	–2.20
Ireland1993	2.90	1.16	1.73	Israel1996	–1.08	1.08	–2.16
Finland1994	3.67	2.00	1.67	Tunisia1993	–2.71	–0.55	–2.15
China2008	0.68	–0.97	1.65	Oman1996	–3.18	–1.12	–2.05

The largest positive differences – i.e., the cases where the ADS is higher than the UDS – are mostly found in small countries with little press coverage. That is as expected: the less press attention, the fewer news articles we have to go by, and the harder it is to pinpoint the country's 'true' democracy level.

The largest negative differences, however, tell a different story. It seems as if either the ADS repeatedly underestimates Israel's democracy score or the UDS repeatedly overestimates it (and not only for the years shown in Table 3). We do not observe a country's 'true' level of democracy, so we cannot know for sure whether the ADS or the UDS is biased (though of course these two possibilities

are not mutually exclusive) but the ADS should be unbiased to the extent that we managed to filter out news articles not related to political regime; whatever biases exist in the UDS should become, by and large, random noise in the ADS.

For instance, imagine that the UDS is biased in favor of countries with generous welfare, like Sweden. The UDS of these countries will be 'boosted' somewhat. But to the extent that the news articles we selected are focused on political regime and not on welfare policy, Wordscores will not associate those boosted scores with welfare-related words and hence the ADS will not be biased. The ADS will be less precise, as (ideally) no particular words will be associated with those boosted scores, but that is it.

The UDS, on the other hand, relies on the assumption that 'raters perceive democracy levels in a noisy but unbiased fashion' (Pemstein et al., 2010, :10), which as Bollen and Paxton (2000) have shown is simply not true: raters have policy preferences and these preferences influence their ratings. Hence whatever biases exist in the Polity, Freedom House, etc, wind up in the UDS as well. The data-generating process behind the UDS does not mitigate bias in any way.

In other words, it seems more likely that the UDS is overestimating Israel's democracy scores than that the ADS is underestimating them. This pro-Israel bias is interesting in itself, but it also raises the more general question of whether the UDS might have an overall conservative bias. To investigate that possibility I performed a difference-of-means test, splitting the data in two groups: country-years with left-wing governments and country-years with right-wing governments (I used the EXECLRC variable from Keefer's (2002) Dataset of Political Institutions for data on government ideological orientation.)

The test rejected the null hypothesis that the mean ADS-UDS difference is the same for the two groups: the mean ADS-UDS difference for left-wing country-years (–0.127, std. error = 0.024, n = 802) is statistically smaller than the mean ADS-UDS difference for right-wing country-years (–0.328, std. error = 0.025, n = 603), with $p < 0.00001$. As both means are negative, it seems that the UDS tend to reward right-wing governments.

I also checked whether the UDS may be biased toward economic policy specifically. I split the country-years in the Index of Economic Freedom (Heritage Foundation 2014) dataset into two groups: statist (IEF score below the median) and non-statist (IEF score above the median). The difference-of-means test shows that the mean ADS-UDS difference for statists (–0.132, std. error = 0.0196, $n = 1057$) is statistically lower than that of non-statists (–0.215, std. error = 0.015, $n = 1977$), with $p < 0.0006$. Both means are negative here as well, so it seems that the UDS somehow rewards free market policies.

We cannot conclusively indict the UDS or its constituent indices though. Perhaps democracy and right-wing government are positively associated and the ADS is somehow less effective at capturing that association. This is consistent with the Hayek-Friedman hypothesis that left-wing governments are detrimental to democ-

racy because economic activism expands the state's coercive resources (Hayek, 1944, Friedman, 1962). As we do not observe a country's true level of democracy, it is hard to know for sure what is going on here.

At least until we know whether the UDS is biased or the ADS is imprecise, the ADS is the conservative choice. Say we regress economic policy on the UDS and find that more democratic countries tend to have less regulation. Is that relationship genuine or is it an artifact of the UDS being biased in favor of free market policies? With biased measures our tests become circular: we cannot know the effect of x on y when our measure of x is partly based on y. Imprecision, on the other hand, merely makes our tests more conservative. The ADS does not solve the reverse-causality problem – maybe free market policies do cause democracy –, but at least we remove ideological bias from the equation.

4. Conclusion

The ADS addresses important limitations of the democracy indices we have today. The ADS is replicable and have standard errors narrow enough to distinguish cases. The ADS is also cost-effective: all we need are training documents and training scores, both of which already exist; there is no need to hire dozens of country experts and spend months collecting and reviewing their work.

It would be interesting to replicate substantive work on democracy but using the ADS instead, to see how the results change. The ADS comes with standard errors, so we could incorporate these in the regressions, perhaps using errors-in-variables models.

We could extend the method here to produce a daily or real-time democracy index. Existing indices are year-based, so we do not know how democratic a country is *today* or how democratic it was, say, on 11/16/2006. Automated text analysis can help us overcome those limitations. We cannot score the news articles from only one or two days, as there would not be enough data to produce meaningful results, but we can pick, say, the 12-month period immediately preceding a certain date – for instance, 11/17/2005-11/16/2006 if we want democracy scores for 11/16/2006.

References

Barro, R. (1999). Determinants of democracy. *Journal of Democracy*, 107:158–183.

Benoit, K. & Laver, M. (2008). Compared to what? A comment on 'a robust transformation procedure for interpreting political text' by Martin and Vanberg. *Political Analysis*, 16:101–111.

Bollen, K. & Paxton, P. (2000). Subjective measures of liberal democracy. *Comparative Political Studies*, 33:58–86.

Bueno de Mesquita, B., Morrow, J., Diverson, R., & Smith, A. (2006). Political competition and economic growth. *Journal of Democracy*, 12:58–72.

Dahl, R. (1972). *Polyarchy*. Yale University Press.

Freedom House (2013). *Freedom In the World*. Freedom House.

Friedman, M. (1962). *Capitalism and Freedom*. University of Chicago Press.

Hayek, F. v. (1944). *The Road to Serfdom*. University of Chicago Press.

Keefer, P. (2002). *DPI2000 Database of Political Institutions: Changes and Variable Definitions*. Development Research Group, The World Bank.

Klemmensen, R., Hobolt, S. B., & Hansen, M. E. (2007). Estimating policy positions using political texts: An evaluation of the wordscores approach. *Electoral Studies*, 26:746–755.

Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review*, 97:311–331.

Marshall, M., Gurr, T., & Jaggers, K. (2013). *Polity IV Project: Political Regime Characteristics and Transitions, 1800-2012, Dataset Users' Manual*. Center for Systemic Peace, Viena, VA.

Martin, L. W. & Vanberg, G. (2008). A robust transformation procedure for interpreting political text. *Political Analysis*, 16:93–100.

Pemstein, D., Meserve, S., & Melton, J. (2010). Democratic compromise: A latent variable analysis of ten measures of regime type. *Political Analysis*, 18:426–449.

Persson, T. & Tabellini, G. (2006). Democracy and development: the devil in the details. *The American Economic Review*, 96:319–324.

Rodrik, D. (1999). Democracies pay higher wages. *The Quarterly Journal of Economics*, 114:707–738.

Treier, S. & Jackman, S. (2008). Democracy as a latent variable. *American Journal of Political Science*, 52:201–217.