
Codebook & Working Notes for the Comparative Democracy Index

Cyrus Gazdar

Abstract This is both the codebook and working notes for the Comparative Democracy Index (CDI). The paper will cover all the variables and the methodology of coding them. I have also included the sources and the original location of my databases.

Introduction I can point to the exact point where I realized I wanted to start this project. I was writing a paper on the onset of civil wars, and while replicating the model, I realized that the operationalization of certain variables used was not accurately describing the outcome they were looking for. In an attempt to criticize the methodology, I wanted to compare a couple of data indices, but quickly realized that since they were all scaled differently, it became difficult to put them all into a table to compare them. So, it got me thinking, why not just make a new dataset that scales a couple of these important indices while also treating it as a way to do some new empirical testing?

The purpose of this dataset is easy to understand: it will serve as a common platform to find comparative democracy scores, and also as a way to obtain z-scores for empirical testing.¹ This book will do a very few things. First, I will break down the various variables used in this dataset. While the variables themselves are pretty self-explanatory, they still had to undergo a good amount of mutation to shape this data. Additionally, a handful of the variables were extrapolated to Frankenstein to combine the data. Furthermore, a few of the variables were inspired by certain indices, so covering this is important. Finally, there is also a bit of variance in the category of the variables, so that will need to be looked at.

After this, I will describe a couple more coding rules. Despite my best efforts to get only relevant cases from all indices, it was not possible to get results for every

1. I want to foreword this by saying this process was not without help from open source organizations. While data for the four indices can be found on their respective websites, I found data from Freedom House, Polity Project, and the EIU from <https://datafinder.qog.gu.se/>. Data for V-Dem was found on their website at <https://v-dem.net/data/the-v-dem-dataset/>. More info will be provided in the "Work Notes" section at the end of the codebook.

observation, so I will discuss how widespread this issue was. Once that is done, I will then talk about the math behind the scaling. As will be shown in the next section, the dataset operated on a 0-10 scale, which is the scale adopted by the EIU, so there was no math required in this case. The other three, however, did require changes, so the math is important. I will also touch on the math used for other variables, how I derived and extrapolated my normative variables.

The last part of this paper will be split up into two parts: the first will briefly go over version information, both for this dataset and the obtained data for it. The second part will discuss plans for this dataset. This is the prototype of it, so there are not too many variables to look at at the moment. However, I hope to build on it in the future, while also having to keep it updated in the future.

Hope you enjoy and can derive something good from it!

I. Dataset Overview

The Comparative Democracy Index (CDI) is a dataset which compiles cross-national measures of international democracy by integrating democratic indicators from major datasets. The purpose of the dataset is not only to compile the data, but also to provide standardized and comparative measurements of democracy over time and countries, demonstrating the strength of institutionalized liberalism and democracy internationally. The unit of analysis for this dataset is country-year. There is coverage for over 200 UN-recognized countries and sovereign territories between the years 1940 and 2025.

Scraped variables are derived from four sources: *Freedom House*, *Varieties of Democracy*, the *Polity Project*, and the *Economist Intelligence Unit*. All datasets provide rich coverage of liberal institutions and have strong metrics for measuring democratic institutions, alongside helpful geographic indicators and some normative variables. There are, however, several variables that have been created. Adding onto the raw score provided by each index, there is also a scaled score, allowing for easy comparative analysis. Additionally, z-scores are included to allow for regression testing.

II. Variable Breakdown

2.1 Variables

2.1.1 Country Code (*ccode*) The respective ISO for each country included in this dataset. The tags were taken from the EIU database and were extrapolated and tied to country IDs to ensure that ISO's were given to countries even if there was not an EIU entry for that given observation.

2.1.2 Year The year of the observation. I have filtered for observations specifically between 1940 and 2025. This serves two fold. First, while the different indices have different observation starting dates, V-Dem, in many cases, starts much earlier than other sources, sometimes during the 1700s. I do not think it is necessary to have hundreds of increased observations specifically for V-Dem, especially since the value does not change much during these years. More importantly, having contemporary dates has the positive effect of cutting out countries that do not exist anymore. While several countries have come and gone since 1940, several observations, notably from V-Dem, cover data from countries that do not exist, with many of them being pre-confederation German States.² Filtering the dates as such provided an easy way to quickly cut out non-existing states.

2.1.3 Region Regional information about a given state. Each observation is given a continent from which the country is. This, like the ISO's, is taken from EIU and attached to a country name to ensure EIU-NA observations were still given a region. A number of countries did not have any tag assigned to them,³ so this had to be done manually. Additionally, EIU and V-Dem also made several observations that are not nation states at all, so I cut out observations that did not have an approved regional tag.⁴

2.1.4 Freedom House Democracy Score (*democracy.score.fh*) The Freedom House score given to a country in a given year. The scale goes from 0-100, and comes with an ordinal variable of the freedom status of that country. I will use this variable as well with Freedom House's measurement for this derived variable.

2.1.5 V-Dem Democracy Score (*democracy.score.vdem*) The democracy score given to a country from the V-Dem dataset. While V-Dem has a number of democracy and electoral variables,⁵ I picked the liberal democracy measurement (*v2x_libdem*) to stay in line with the stated goals of Freedom House to test civil and political liberties.

2.1.6 Polity Project Democracy Score (*democracy.score.polity*) The Democracy score from the Polity Project. The measurement scaled from -10 to 10.

2.1.7 Economist Intelligence Unit (*democracy.score.eiu*) The EIU democracy score. Scales from 0-10.

2. Some of the countries I had to get rid of manually included Saxony, Hessen-Darmstadt, & Brunswick.

3. The missing countries had to be manually added, and will be discussed later.

4. The approved continents were "Asia", "Africa", "Europe", "South America", "North America", & Oceania.

5. For all these variables, see section 2.1 in Coppedge, Michael, et. al. 2024. "V-Dem Codebook v14" Varieties of Democracy (V-Dem) Project.

2.1.8 Freedom House Scaled Democracy Score (*fh.scaled.score*) Freedom House's democracy score scaled to a 0-10 scale.

2.1.9 V-Dem Scaled Democracy Score (*vdem.scaled.score*) V-Dem's democracy score scaled to a 0-10 scale.

2.1.10 Polity Project Scaled Democracy Score (*polity.scaled.score*) Polity Project's democracy score scaled to a 0-10 scale.

2.1.11 Economist Intelligence Unit Scaled Democracy Score (*eiw.scaled.score*) EIU's democracy score scaled to a 0-10 scale.⁶

2.1.12 Average Democracy Score (*scaled.democracy.avg*) The average democracy between the four indices. Scored on a 0-10 scale.

2.1.13 Yearly Change (*yearly.change*) The year-over-year democracy score change for a given observation. The change will be taken from the change in (*scaled.democracy.avg*).

2.1.14 Freedom House Scaled z-score (*freedom.house.scaled.z.score*) The z-score for the Freedom House democracy ranking

2.1.15 V-Dem Scaled z-score (*vdem.scaled.z.score*) The z-score for the V-Dem democracy ranking

2.1.16 Polity Project Scaled z-score (*polity.scaled.z.score*) The z-score for the Polity Project democracy ranking

2.1.17 Economist Intelligence Unity Scaled z-score (*eiw.scaled.z.score*) The z-score for the EIU democracy ranking

2.2 Derived Variables

2.2.1 Freedom Status (*freedom.status*) A normative ranking of a nations democracy status. The ranking will use both the labels⁷ and methodology⁸ of Freedom House.

6. The default EIU scale is already 0-10, so there was no value adjusting from (*democracy.score.eiu*)

7. Labels are broken down into "free", "partly free", and "not free."

8. The rankings are categorized based on the democratic score given to a country in a given year. When reduced to the 0-10 scale, "free" countries are those with a score ≥ 7 ; partly free countries require a score of ≥ 3.5 ; "not free" countries have a score < 3.5 .

2.3 Missing Data Rules

There are a number of missing observations throughout the dataset, especially for observations before the 1970s, and for countries that are not established nation-states. This, of course, is not ideal; there are enough full observations, especially from the start of the 21st century, to make the data valuable nevertheless. My advice would be to use this dataset in that capacity. The dataset is coded so that even in cases of missing observations, the score is still logged as all the others.

III. Data Overview

Here, I will break down the indices used in this dataset. As I mentioned earlier, this dataset was a combination of different variables taken from different organizations. This section will first discuss where I got the databases from and what original variables were downloaded. I will also talk about which variables were selected and incorporated into the final dataset.

3.1 Dataset Information All datasets have public domains that have the option to download their data for research. For simplicity reasons, however, I installed datasets for Freedom House and EIU from *Our World in Data*, an open-source repository with up-to-date datasets.⁹ A positive from downloading here was the ability to filter through and select which variables to include in the original datasets. This makes it much easier to exclude unnecessary variables right off the bat.¹⁰ Data from the Polity Project was downloaded from the *Quality of Government Data Finder*, another open-source repository with countless datasets.¹¹ Finally, the V-Dem dataset was installed directly from its website. To ensure we did not have duplication issues, the majority of variables were cut from each one to ensure everything stayed as lean as possible. Below, I will include what variables were included for each dataset.

3.1.2 V-Dem V-Dem is different from the others in that it was installed directly from the V-Dem website, and since you cannot selectively crop out variables, I pulled the full dataset with 4370 variables. As mentioned earlier, since I wanted my ranking to be in line with Freedom House measurements, the variable to measure democracy score is *v2x_libdem*, which measures the liberal democracy of a nation. V-Dem has the advantage over other datasets in its vast coverage and observations. Therefore, the additional variables are also taken from V-Dem into the dataset: country name (*country_name*), country code (*country_text_id*), and year (*year*)

9. Access to the website can be found at: <https://ourworldindata.org/>.

10. This was especially helpful for dealing with the Freedom House variables, which include civil liberties and press freedoms.

11. The website for QoG is <https://datafinder.qog.gu.se/>

3.1.3 Freedom House The original Freedom House dataset came with the following variables: Country name (*entity*), country code (*code*), observation year (*year*), Freedom House democracy score (*total.democracy.score*), and regional affiliation (*world.region.according.to.OWID*). All variables were cut except for (*total.democracy.score*).

3.1.4 Polity Project The Polity Dataset came with a series of variables, including observation number *X*, country code (*ccode*), country ID (*cname*), year of observation (*year*), country ID according to QoG (*cname_qog*), country name according to QoG (*cname_qog*), country code (*ccodealp*), country code according to the Correlates of War Project (*ccodecow*), country name for a given year (*cname_year*), country code for a given year (*ccodealp_year*), and the democracy score (*p_polity2*). The dataset kept (*p_polity2*), and also the *ccode*.

3.1.5 EIU Since both EIU and Freedom House were installed from OWID, they both had more or less the same variables in country name (*entity*), country code (*code*), observation year (*year*), the EIU democracy score (*democracy.index*), and regional affiliation (*world.region.according.to.OWID*). The variables kept included (*democracy.index*), and after comparing the total coverage on regional affiliation between EIU and Freedom House, the former dataset has much better regional coverage, so (*world.region.according.to.OWID*) was also kept.

IV. Coding Rules

Here, I will cover a series of coding rules I implemented to ensure our findings are accurate.¹²

4.1.1.1 Mutated Datasets All datasets were first mutated individually before being merged into a single dataset. Below was the script used to construct these new dataframes. The first dataframes that the "selected" dataframes are based on are the original datasets mentioned in part I and section 3.1.1. The selected datasets (ex, "Freedom.House.Selected.Data) are skimmed datasets based on the conditions mentioned in sections 3.1.2-3.1.5.¹³

4.1.1.2 Freedom House

```
1 Freedom.House.Selected.Data <- Freedom.House.Data %>%  
2   select(
```

12. All replicating data is available in the "replication data" folder. The script for part 4.1.1 is titled "Data Set Filing.R", and the script for part 4.2.1 is called "Aggregating Data.R."

13. The only package used for the following code is "dplyr"

```

3     Entity,
4     Year,
5     Code,
6     Total.democracy.score
7   )
8   Freedom.House.Mutated.Data <- Freedom.House.Selected.Data %>%
9     mutate(source = "Freedom House") %>%
10    rename(democracy.score.fh = Total.democracy.score,
11           ccode = Code,
12           cname = Entity,
13           year = Year)

```

4.1.1.3 EIU

```

1   EIU.Selected.Data <- EIU.Data %>%
2     select(
3       Entity,
4       Code,
5       Year,
6       Democracy.Index,
7       World.region.according.to.OWID
8     )
9   EIU.Mutated.Data <- EIU.Selected.Data %>%
10    mutate(source = "EIU") %>%
11    rename(democracy.score.eiu = Democracy.Index,
12           ccode = Code,
13           cname = Entity,
14           year = Year,
15           region = World.region.according.to.OWID)

```

4.1.1.4 V-Dem

```

1   VDem.Selected.Data <- VDem.Data %>%
2     select(
3       country_name,
4       country_text_id,
5       year,
6       v2x_libdem
7     )
8   VDem.Mutated.Data <- VDem.Selected.Data %>%
9     mutate(source = "VDem") %>%
10    rename(democracy.score.vdem = v2x_libdem,

```

```
11     ccode = country_text_id,  
12     cname = country_name)
```

4.1.1.5 Polity Project

```
1 Polity.Selected.Data <- Polity.Data %>%  
2   select(  
3     cname,  
4     year,  
5     ccodealp,  
6     ccodealp_year,  
7     p_polity2  
8   )  
9 Polity.Mutated.Data <- Polity.Selected.Data %>%  
10  mutate(source = "Polity") %>%  
11  rename(democracy.score.polity = p_polity2,  
12         ccode = ccodealp)
```

The mutate function was done more or less for simplicity, and was removed for the final dataset. The rename function was used to standardise variable names, which, when paired with the distinct function, cuts out harmful duplicates and keeps unique observations, and will continue to be used throughout the process. We now have all the datasets needed to construct the first dataframe.

4.1.2.1 Constructing the First Dataframe The first step in constructing the dataframe was combining our mutated datasets.

4.1.2.2 Combining Datasets

```
1 MDFBind <- bind_rows(  
2   Freedom.House.Mutated.Data %>% select(cname, ccode, year),  
3   VDem.Mutated.Data %>% select(cname, ccode, year),  
4   Polity.Mutated.Data %>% select(cname, ccode, year),  
5   EIU.Mutated.Data %>% select(cname, ccode, year)  
6 ) %>%  
7   distinct()
```

4.1.2.3 Separating Regions and Re-adding Since the regional affiliation for the dataset was taken from EIU, not all observations had a regional tag to it, notably those that occurred before EIU observations took place. The following code took regional tags and extrapolated them based on country tags. First, I made a new dataset with just regional tags and country codes, and the second step assigned the country codes to

the regional tags, and the third step built a new dataframe from the regional dataframe by adding in the other variables.

```

1  1. Building a regional dataframe
2  regions <- EIU.Mutated.Data %>%
3    select(ccode, region) %>%
4    distinct()
5  2. Adding codes to regions
6  MDFRegion <- MDFBind %>%
7    left_join(regions, by = "ccode")
8  3. Adding the remaining variables to the dataset
9  MDF <- MDFR %>%
10    left_join(Freedom.House.Mutated.Data %>% select(ccode, year,
11      ↪ democracy.score.fh), by = c("ccode", "year")) %>%
12    left_join(VDem.Mutated.Data %>% select(ccode, year,
13      ↪ democracy.score.vdem), by = c("ccode", "year")) %>%
14    left_join(Polity.Mutated.Data %>% select(ccode, year,
15      ↪ democracy.score.polity), by = c("ccode", "year")) %>%
16    left_join(EIU.Mutated.Data %>% select(ccode, year,
17      ↪ democracy.score.eiu), by = c("ccode", "year"))
18  MDF <- MDF %>%
19    arrange(cname, year)

```

4.1.2.4 Addressing Countries without regional tags While most countries were assigned a regional tag, some country codes were not caught or logged by other datasets,¹⁴ so I first built another dataframe that includes only these nations, and after this, manually assigned regional/continental tags to each nation based on their country code.

```

1  NARegions <- MDF %>%
2    filter(is.na(region))
3
4  MDF <- MDF %>%
5    mutate(
6      region = case_when(

```

14. The countries that were not included were Andorra, Antigua and Barbuda, Bahamas, Barbados, Belize, Brunei Darussalam, Dominica, Grenada, Kiribati, Kosovo, Liechtenstein, Maldives, Marshall Islands, Micronesia (Federated States of), Monaco, Nauru, Palau, Palestine/Gaza, Saint Kitts and Nevis, Saint Lucia, Saint Vincent and the Grenadines, Samoa, San Marino, Sao Tome and Principe, Serbia and Montenegro, Seychelles, Solomon Islands, Somalia, Somaliland, South Sudan, Tonga, Tuvalu, Vanuatu

```

7   ccode %in% c("AND", "KKX", "LIE", "MCO", "SMR", "SCG") ~
   ↪ "Europe",
8   ccode %in% c("ATG", "BHS", "BRB", "BLZ", "DMA", "GRD",
   ↪ "KNA", "LCA", "VCT") ~ "North America",
9   ccode %in% c("BRN", "MDV", "PSE", "PSG") ~ "Asia",
10  ccode %in% c("KIR", "MHL", "FSM", "NRU", "PLW", "WSM",
   ↪ "SLB", "TON", "TUV", "VUT") ~ "Oceania",
11  ccode %in% c("STP", "SYC", "SOM", "SSD", "ZZB", "SML") ~
   ↪ "Africa",
12  TRUE ~ region
13 )
14 )

```

4.1.2.5 Cropping Dates and Corrupted Observations The last part of the first dataframe is to crop dates and also get rid of observations that are still not tied to a nation or territory. In the case of the latter, this was affiliated more with EIU and V-Dem, which have observations for world or continental populations.

```

1  1. Cropping Dates
2  MDF <- MDF %>%
3    filter(year >= 1940 & year <= 2026)
4  2. Getting rid of corrupted observations
5  MDF <- MDF %>%
6    distinct(ccode, year, .keep_all = TRUE)
7  MDF %>%
8    filter(is.na(region) | region == "" | trimws(region) == "")
   ↪ %>%
9    distinct(cname, ccode)
10 MDF <- MDF %>%
11   mutate(region = trimws(region)) %>%
12   filter(!is.na(region) & region != "")

```

4.2.1.1 Adding in New Variables The next part of this project was to incorporate new variables for comparative analysis and empirical study. The first part will cover the former topic. The packages used for the following code are a combination of *"dplyr"* and *"tidyr."*

4.2.1.2 Establishing a 0-10 Standard Scale To set up a 0-10 scale, I first made a function that establishes the rules of the function. I then applied this function to the variables in the mutated datasets.

```

1  1. Establishing the Code
2  rescaled.democracy <- function(x) {

```

```

3  10 * (x - min(x, na.rm = TRUE)) /
4    (max(x, na.rm = TRUE) - min(x, na.rm = TRUE))
5  }
6  2. Applying the Code to the Datasets
7  Dataframe <- Dataframe %>%
8    mutate(
9      fh.scaled.score = rescaled.democracy(democracy.score.fh),
10     vdem.Scaled.Score =
11       ↪ rescaled.democracy(democracy.score.vdem),
12     polity.scaled.score =
13       ↪ rescaled.democracy(democracy.score.polity),
14     eiu.scaled.score =
15       ↪ rescaled.democracy(democracy.score.eiu))

```

4.2.1.3 Averages The last part of this dataset was to add democracy aggregates and z-scores. There are two variables made for the averages:

```

1  1. Average Standardised Democracy Score
2  Dataframe <- Dataframe %>%
3    mutate(
4      scaled.democracy.avg = rowMeans(
5        select(., fh.scaled.score, vdem.Scaled.Score,
6          ↪ polity.scaled.score, eiu.scaled.score),
7        na.rm = TRUE
8      )
9    )
10  2. Year-Over-Year Change in Score
11  Dataframe <- Dataframe %>%
12    group_by(ccode) %>%
13    mutate(
14      yearly.change = scaled.democracy.avg -
15        ↪ lag(scaled.democracy.avg)
16    ) %>%
17    ungroup()

```

4.2.1.4 Freedom Status I wanted to add a normative variable assigned to a country's democracy score. I decided to take inspiration from Freedom House in this regard due to their easy-to-recognise labels and simple scaling. The following code takes their rankings and scales them down to 0-10:

```

1  Dataframe <- Dataframe %>%
2    mutate(

```

```
3 freedom.status = case_when(  
4   scaled.democracy.avg >= 7 ~ "f",  
5   scaled.democracy.avg >= 3.5 ~ "pf",  
6   scaled.democracy.avg < 3.5 ~ "nf",  
7   TRUE ~ NA_character_  
8 )  
9 )
```

4.2.1.5 Z-Scores The following code adds the z-scores to the dataset:

```
1 Dataframe <- Dataframe %>%  
2   mutate(  
3     freedom.house.z.score =  
4       ↪ as.numeric(scale(democracy.score.fh)),  
5     vdem.z.score = as.numeric(scale(democracy.score.vdem)),  
6     polity.z.score =  
7       ↪ as.numeric(scale(democracy.score.polity)),  
8     eiu.z.score = as.numeric(scale(democracy.score.eiu))  
9   )
```

V. Version Information & Future Plans

5.1 Version Information All datasets taken for this dataset are up-to-date. While some datasets, such as V-Dem, release editions more than once a year, this does not matter as all scores are lagged by one year, cutting out the most recent (in this case, 2026).

5.2.1 Future Updates I plan on updating this dataset at least once a year when data for the previous year comes out. While I do think this dataset does have a good purpose, I also derive good personal use (as I hope you can as well!), so I am incentivised to keep this going. I will likely update once the repositories I pulled these from are also updated, but if need be, I will scrape the data directly from the websites themselves.

5.2.2 Future Variables While there are some good uses for this dataset, it is lacking in a large number of variables. This is something that I plan to do by September 2026. The first variable I want to include is more in-depth civil liberties, such as freedom of the press, association, and religion. Freedom House covers this very well, so it is something that should not be too hard to incorporate.

I am also interested in adding new indices that are not directly related to democratic institutions but may contribute to their strength. I can first think of the *Fragile State Index* (FSI), but I want to first find other repositories that can be included alongside it.

I am also keen on adding socio-economic and socio-political variables that may also contribute to democratic legitimacy. To this end, I am curious about engaging with the *Cross-National Time Series* (CNTS) dataset, which contains the needed variables.

This is my first ever real project outside of work for classes, so I know there might be issues. I am open to any criticisms and all ears if you have any suggestions. I hope this dataset can serve you well, and thank you for your engagement.