Data Building

Project: Dominant Party Regimes in Sub-Saharan Africa

Stepan Polikanov

This file serves as a research notebook, detailing sequential data building steps for the project. It aims to provide a transparent view of coding practices, analytical decisions and inherent data limitations.

I first describe the building of key explanatory variables and then the controls.

I start with reading in the data.

```
party_class <- read_excel(here("data", "data_raw", "partyreg.xlsx"))
colonial <- read_csv(here("data", "data_raw", "COLDAT_colonies.csv"))
vdem <- readRDS(here("data", "data_raw", "V-Dem-CY-Full+Others-v10.rds"))
epr <- read_csv(here("data", "data_raw", "EPR.csv"))
dpi <- read_dta(here("data", "data_raw", "DPI2020.dta"))
vparty <- read_rds(here("data", "data_raw", "V-Dem-CPD-Party-V2.rds"))
wb_oil <- read_csv(here("data", "data_raw",
                        "API_NY.GDP.PETR.RT.ZS_DS2_en_csv_v2_2446738.csv"))
```

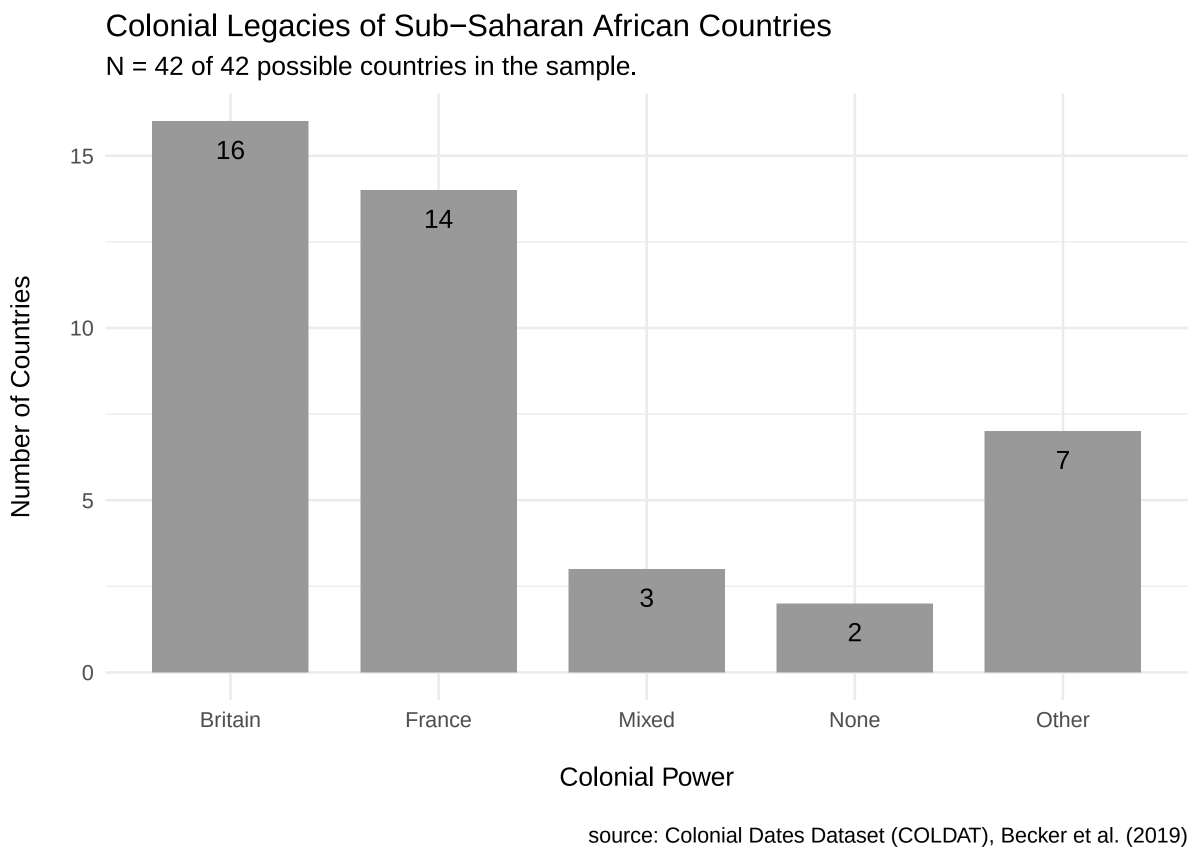
Create a character vector of country-year pairs for subsetting.

```
cy <- str_c(party_class$ccodecow, party_class$year, sep = "_")
```

I construct dummies for the colonial legacies of sub-Saharan African countries from the Colonial Dates dataset (Becker, 2019).

```
col_legacy <- colonial |>
  mutate(ccodecow = countryname(country, destination = "cown", warn = F)) |>
  filter(ccodecow %in% party_class$ccodecow) |>
  reframe(ccodecow, col.britain, col.france,
          col.other = if_else((col.britain == 0 & col.france == 0)
                              & (col.belgium == 1 | col.germany == 1
                                 | col.italy == 1 | col.netherlands == 1
                                 | col.portugal == 1 | col.spain == 1), 1, 0),
          col.none = if_else(col.other == 0 & col.britain == 0
                             & col.france == 0, 1, 0),
          col.type = case_when(col.britain == 1 & col.france == 0
                               & col.other == 0 ~ "Britain",
                               col.france == 1 & col.britain == 0
```

```
                                & col.other == 0 ~ "France",
                                col.other == 1 & col.britain == 0
                                & col.france == 0 ~ "Other",
                                col.none == 1 ~ "None",
                                .default = "Mixed"))

ggplot(col_legacy, aes(x = col.type)) +
  geom_bar(fill = "grey60", width = 0.75) +
  geom_text(aes(label = after_stat(count)), stat = "count", vjust = 2) +
  labs(title = "Colonial Legacies of Sub-Saharan African Countries",
       subtitle = paste0("N = ", sum(!is.na(col_legacy$col.type)), " of ",
                         length(col_legacy$col.type),
                         " possible countries in the sample."),
       caption =
         "\nsource: Colonial Dates Dataset (COLDAT), Becker et al. (2019)",
       x = "\nColonial Power",
       y = "Number of Countries\n")
```

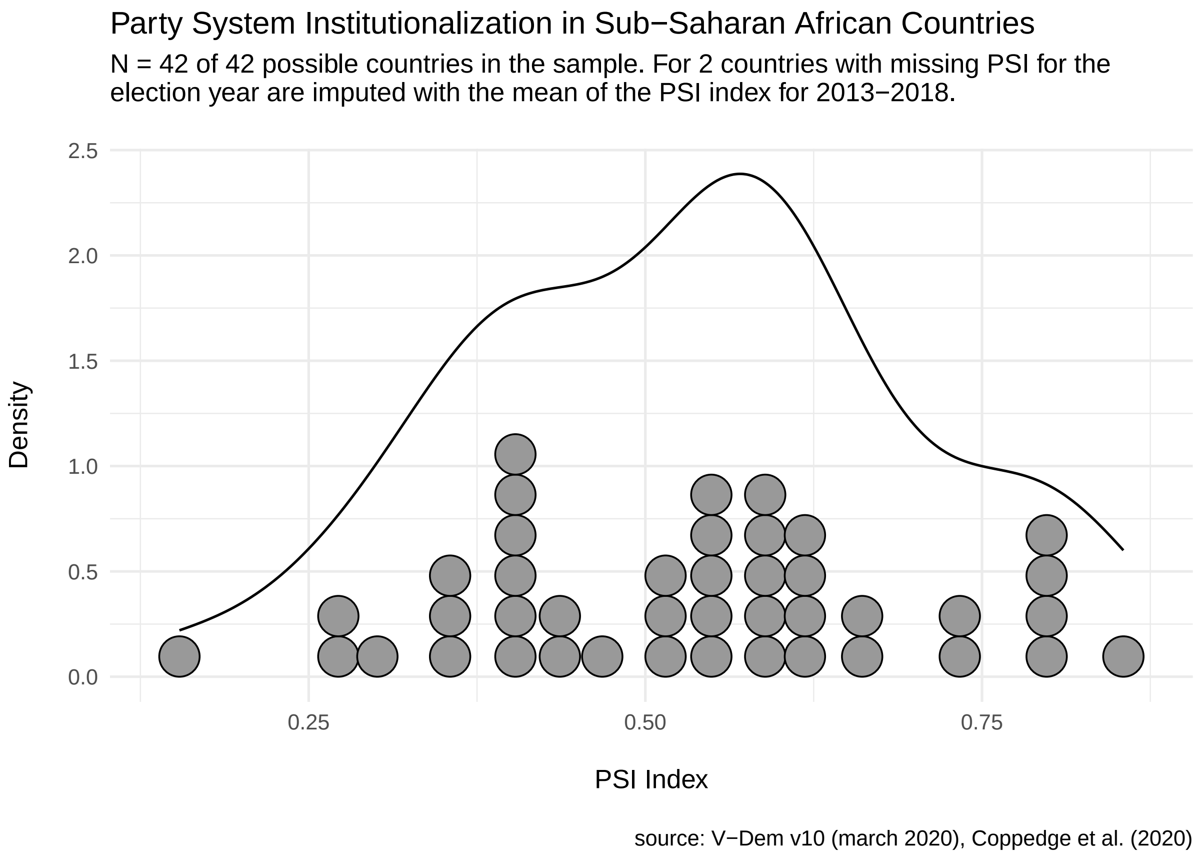


The result is that some countries had colonial histories with both Britain and France, either of those, a different metropole, or were not colonized at all.

The variable is not ideal: one obvious issue is the three "misc" categories, with countries being either under both French and British rule, not colonized or under control of other colonial powers. The original paper collapsed these categories into a single one, but the experience of "trading hands" of different powers is likely to be extremely different from not being colonized at all. I will keep the variable as is, but I will also provide a sensitivity analysis where I collapse the categories into a single one. The latter option allows us to zero in on the differences between the British and the French, which is of course the motivating factor between the variable.

Party system institutionalization data comes from V-Dem v10 (march 2020) (Coppedge et al., 2020).

```
psi <- vdem |>
  select(country_name, ccodecow = COWcode, year, v2xps_party) |>
  mutate(v2xps_party_mean = if_else(is.na(v2xps_party)
                                    & str_c(ccodecow, year, sep = "_") %in% cy,
                                    round(mean(v2xps_party[year >= 2013
                                                           & year <= 2018],
                                         na.rm = T), 3), v2xps_party),
         v2xos_party_next = if_else(is.na(v2xps_party)
                                    & str_c(ccodecow, year, sep = "_") %in% cy,
                                    lead(v2xps_party, 1), v2xps_party)) |>
  filter(str_c(ccodecow, year, sep = "_") %in% cy)

ggplot(psi, aes(x = v2xps_party_mean)) +
  geom_density() +
  geom_dotplot(fill = "grey60", binwidth = 0.03) +
  labs(title =
         "Party System Institutionalization in Sub-Saharan African Countries",
       caption = "\nsource: V-Dem v10 (march 2020), Coppedge et al. (2020)",
       subtitle = paste0("N = ", sum(!is.na(psi$v2xps_party_mean)), " of ",
                         length(psi$v2xps_party_mean),
                         " possible countries in the sample. For ",
                         sum(!is.na(psi$v2xps_party_mean))
                         - sum(!is.na(psi$v2xps_party)),
                         " countries with missing PSI for the\n",
                         "election year are imputed with the mean of the PSI ",
                         "index for 2013-2018.\n"),
       x = "\nPSI Index",
       y = "Density\n")
```

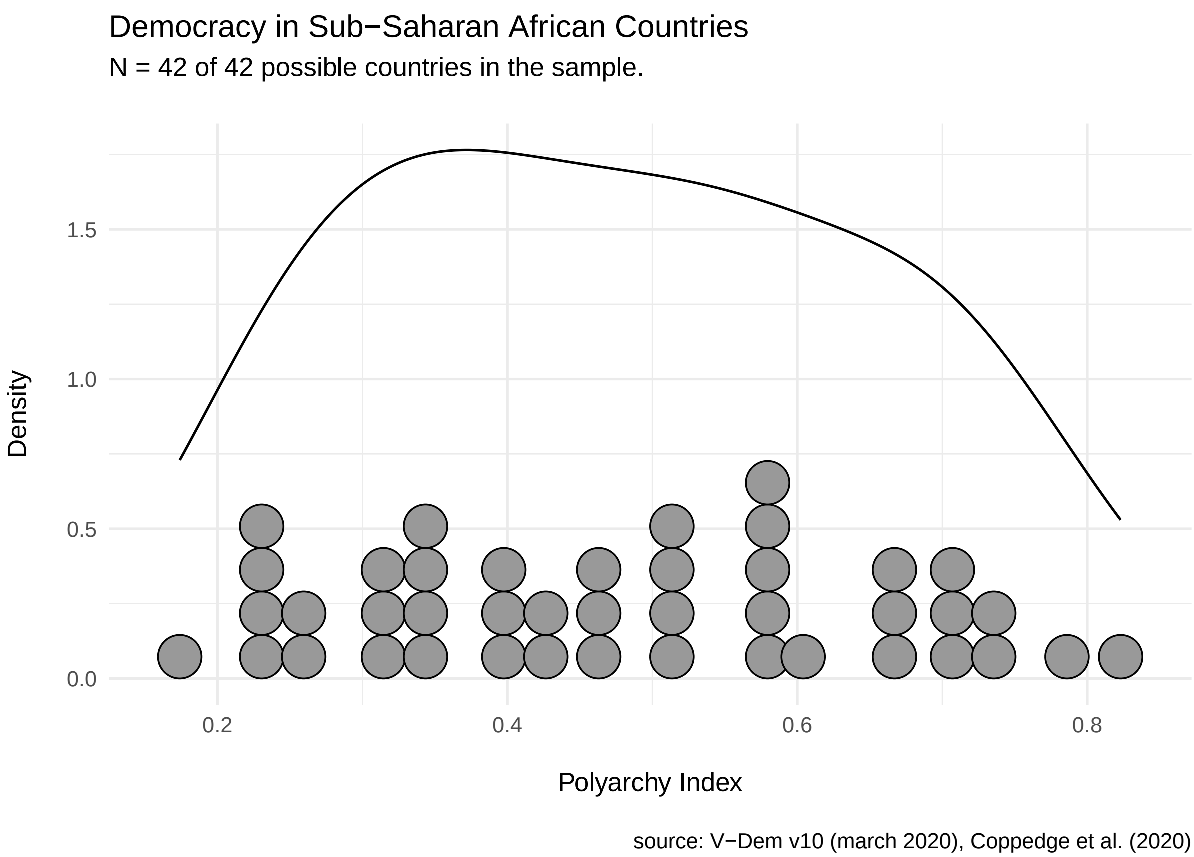


There are two instances when the PSI index is not recorded by V-Dem in the delineated geographical and temporal scope: Mali in 2013 and Guinea in 2013. I fill these with the mean of the PSI index for these countries for the years 2013-2018. The decision is aimed at minimizing missing values in an already very small sample. This may be suboptimal, so I also provide a sensitivity analysis to this choice. I compare the mean solution with imputing the next value or discarding the observations altogether.

The data on the level of democracy also comes from V-Dem v10 (Coppedge et al., 2020). I use the V-Dem polyarchy measure.

```
dem <- vdem |>
  select(ccodecow = COWcode, year, v2x_polyarchy) |>
  filter(str_c(ccodecow, year, sep = "_") %in% cy)

ggplot(dem, aes(x = v2x_polyarchy)) +
  geom_density() +
  geom_dotplot(fill = "grey60", binwidth = 0.03) +
  labs(title =
         "Democracy in Sub-Saharan African Countries",
       caption = "\nsource: V-Dem v10 (march 2020), Coppedge et al. (2020)",
       subtitle = paste0("N = ", sum(!is.na(dem$v2x_polyarchy)), " of ",
                         length(dem$v2x_polyarchy),
                         " possible countries in the sample.\n"),
```

```
    x = "\nPolyarchy Index",
    y = "Density\n")
```



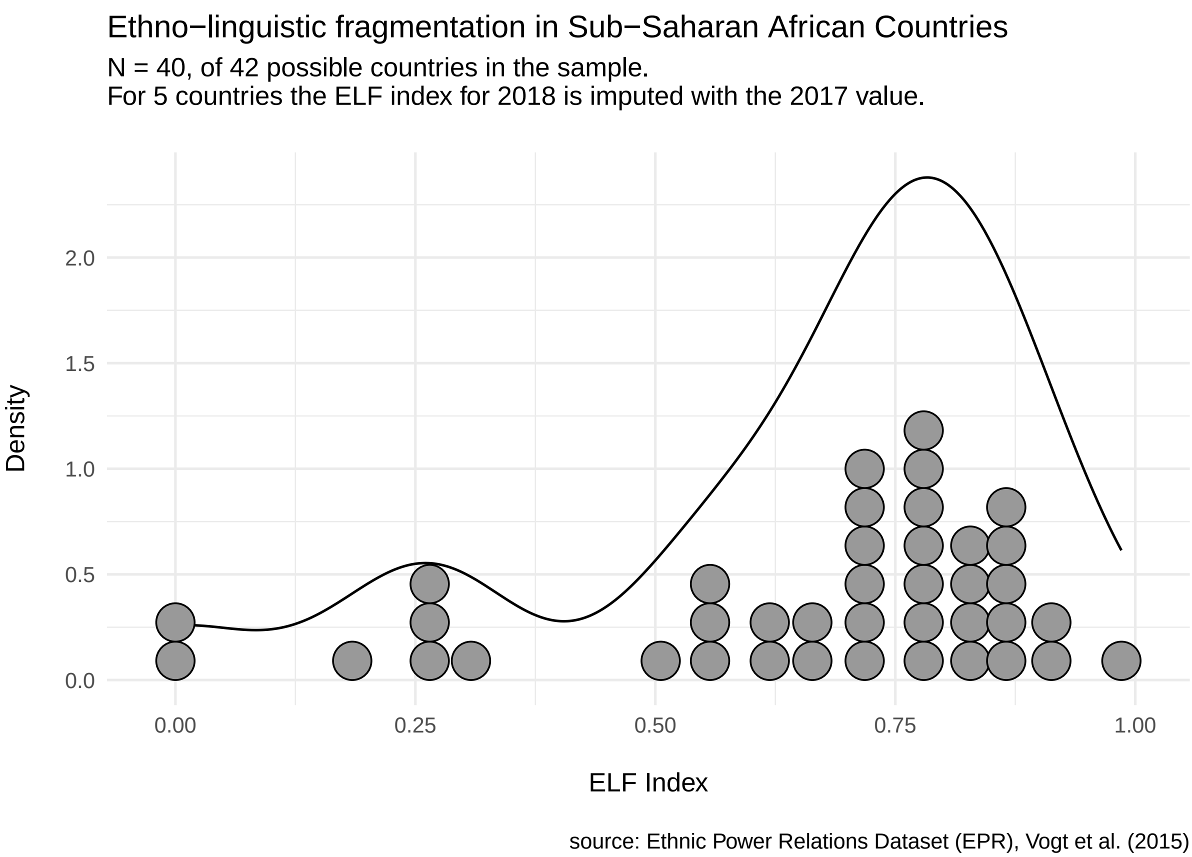
I use Ethnic Power Relations core dataset to construct a measure of ethno-linguistic fragmentation (ELF) (Vogt et al., 2015). The ELF index is a measure of the probability that two randomly selected individuals in a country belong to different ethno-linguistic groups. The calculation is straightforward:

$$ELF_i = 1 - \sum_{j=1}^{K} s_j^2$$

where for to get the ELF index for country i, I deduct from 1 the sum of squared proportions s of ethnic groups j across K groups in a country.

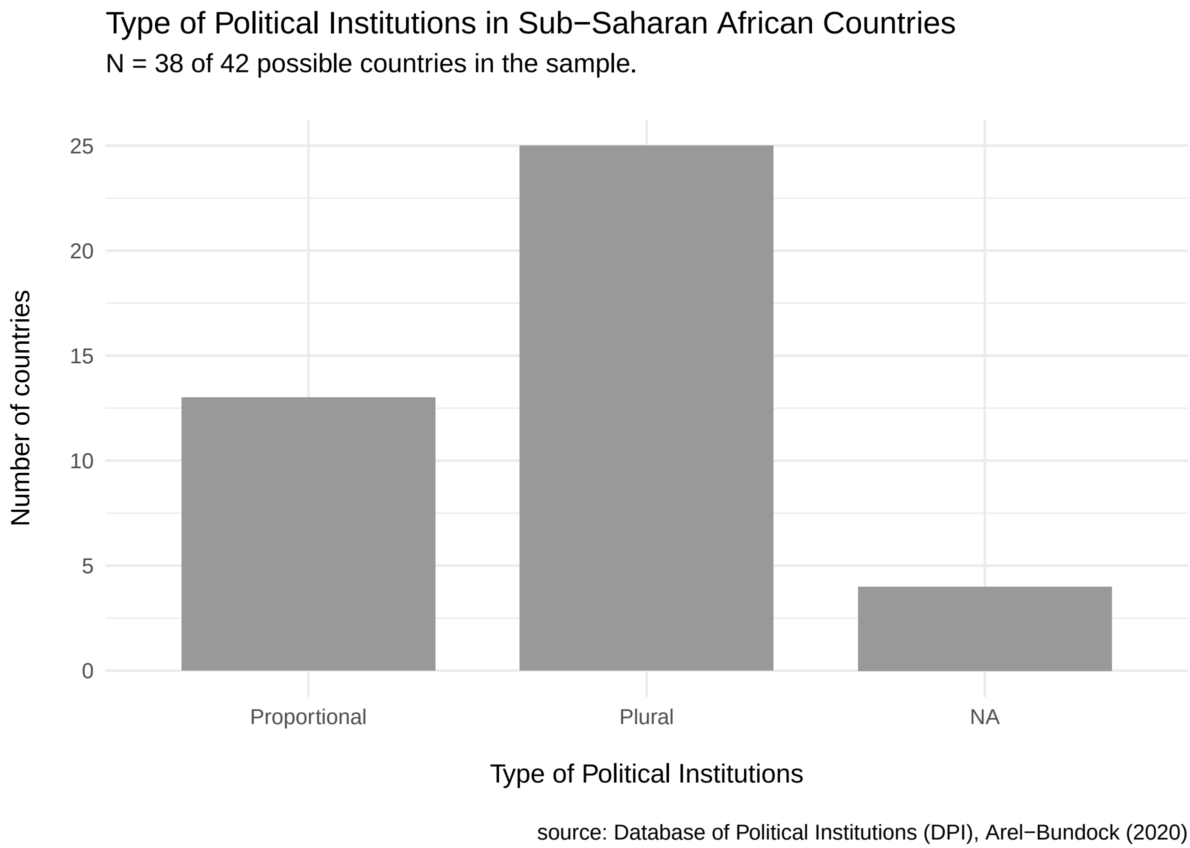
```
elf <- epr |>
  mutate(ccodecow = countryname(statename, destination = "cown")) |>
  group_by(statename, ccodecow, from, to) |>
  summarize(elf = 1 - sum(size^2)) |>
  mutate(year = list(seq(from, to))) |>
```

```
  unnest(year) |>
  ungroup() |>
  select(-from, -to) |>
  full_join(select(party_class, ccodecow, year)) |>
  arrange(ccodecow, year) |>
  group_by(ccodecow) |>
  mutate(elf_fill = if_else(is.na(elf), lag(elf, 1), elf)) |>
  right_join(select(party_class, ccodecow, year)) |>
  ungroup()

ggplot(elf, aes(x = elf_fill)) +
  geom_density() +
  geom_dotplot(fill = "grey60", binwidth = 0.04) +
  labs(title = "Ethno-linguistic fragmentation in Sub-Saharan African Countries",
       subtitle = paste0("N = ", sum(!is.na(elf$elf_fill)), ", of ",
                         length(elf$elf_fill), " possible countries in the sample.\nFor ",
                         sum(!is.na(elf$elf_fill)) - sum(!is.na(elf$elf)),
                         " countries the ELF index for 2018 is imputed with the 2017 value.\n"),
       caption = "\nsource: Ethnic Power Relations Dataset (EPR), Vogt et al. (2015)",
       x = "\nELF Index",
       y = "Density\n")
```



Data on the electoral rules for the lower house of the national legislature comes from the Database of Political Institutions (DPI) (Beck et al., 2010). I focus on the type of electoral system used in the most recent election year. The data is coded as 0 for proportional representation and 1 for plurality.

```
housesys <- dpi |>
  select(countryname, ifs, year, housesys) |>
  mutate(ccodecow_1 = countrycode(ifs, origin = "iso3c", destination = "cown"),
         ccodecow_2 = countryname(countryname, destination = "cown"),
         ccodecow = case_when(ccodecow_1 == ccodecow_2 ~ ccodecow_1,
                              is.na(ccodecow_1) ~ ccodecow_2,
                              is.na(ccodecow_2) ~ ccodecow_1,
                              !is.na(ccodecow_1)
                              & !is.na(ccodecow_2) ~ ccodecow_1),
         housesys = factor(if_else(housesys == -999, NA, housesys),
                           levels = c(0, 1),
                           labels = c("Proportional",
                                      "Plural"))) |>
  filter(str_c(ccodecow, year, sep = "_") %in% cy) |>
  right_join(select(party_class, ccodecow, year))

ggplot(housesys, aes(x = housesys)) +
  geom_bar(fill = "grey60", width = 0.75) +
  labs(title = "Type of Political Institutions in Sub-Saharan African Countries",
       subtitle = paste0("N = ", sum(!is.na(housesys$housesys)), " of ",
                         length(housesys$housesys),
                         " possible countries in the sample.\n"),
       caption = "\nsource: Database of Political Institutions (DPI), Arel-Bundock (2020)",
       x = "\nType of Political Institutions",
       y = "Number of countries\n")
```



There are around twice as many countries with plurality as with with proportional representation in the sample. For 4 countries data is unavailable. For Sao Tome and Principe and the Seychelles, the project does not collect data, and for Guinea and Sudan, the data is missing in source.

The data on effective number of parties comes from the V-Party dataset (Lindberg et al., 2022). I calculate both ENPP (effective number of parliamentary parties) and ENEP (effective number of presidential candidates) indices. The ENPP index is calculated as:

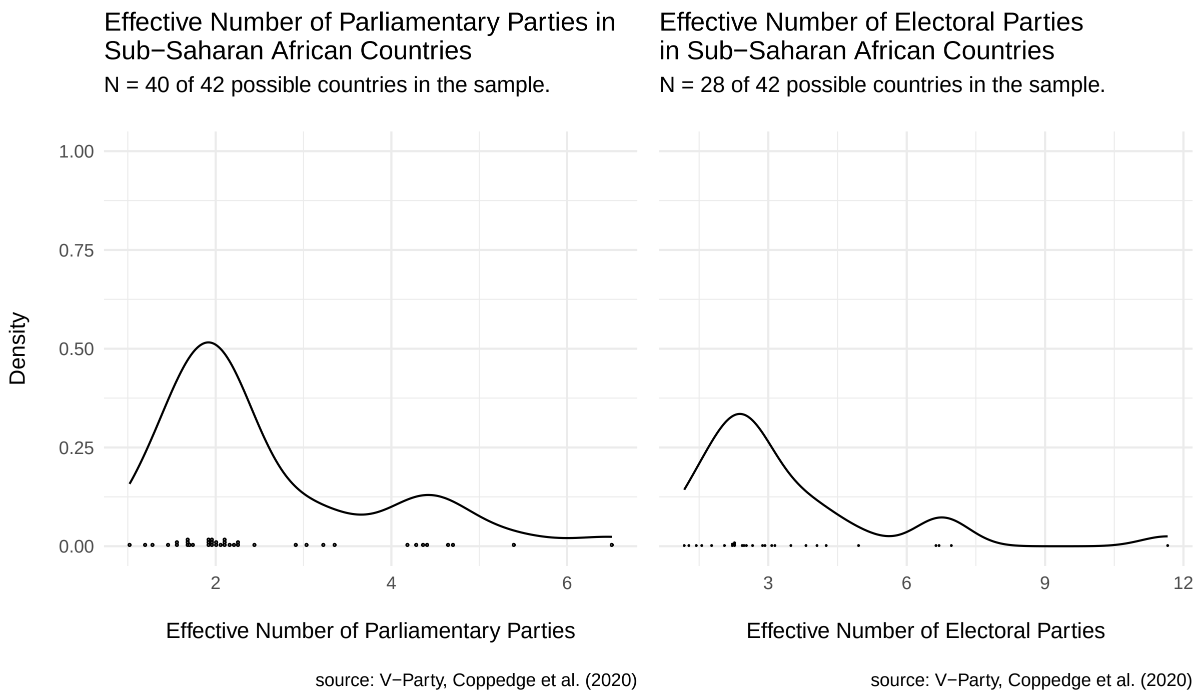
$$ENPP_i = \frac{1}{\sum_{p=1}^{L} ss_p^2}$$

which is the inverse of the sum of squared seat shares of parties in the lower house of the national legislature and where ss_p is the seat share of party p. The ENEP index is calculated as:

$$ENEP_i = \frac{1}{\sum_{p=1}^{L} sv_p^2}$$

which is the inverse of the sum of squared vote shares of presidential candidates, with sv_p being that share.

```r
enp <- vparty |>
  select(country_name, year, ccodecow = COWcode, v2paseatshare,
         v2pavote) |>
  filter(str_c(ccodecow, year, sep = "_") %in% cy) |>
  mutate(across(c(v2paseatshare, v2pavote), ~ ./100)) |>
  group_by(ccodecow, year) |>
  summarize(enpp = 1/sum(v2paseatshare^2, na.rm = T),
            enep = 1/sum(v2pavote^2, na.rm = T),
            enep = if_else(is.infinite(enep), NA, enep),
            .groups = "drop") |>
  right_join(select(party_class, ccodecow, year))

enp_graph_1 <- ggplot(enp, aes(x = enpp)) +
  geom_density() +
  geom_dotplot(fill = "grey60", binwidth = 0.03) +
  labs(title = "Effective Number of Parliamentary Parties in\nSub-Saharan African Countries",
       subtitle = paste0("N = ", sum(!is.na(enp$enpp)), " of ",
                         length(enp$enpp),
                         " possible countries in the sample.\n"),
       caption = "\nsource: V-Party, Coppedge et al. (2020)",
       x = "\nEffective Number of Parliamentary Parties",
       y = "Density\n")

enp_graph_2 <- ggplot(enp, aes(x = enep)) +
  geom_density() +
  geom_dotplot(fill = "grey60", binwidth = 0.03) +
  labs(title = "Effective Number of Electoral Parties\nin Sub-Saharan African Countries",
       subtitle = paste0("N = ", sum(!is.na(enp$enep)), " of ",
                         length(enp$enep),
                         " possible countries in the sample.\n"),
       caption = "\nsource: V-Party, Coppedge et al. (2020)",
       x = "\nEffective Number of Electoral Parties",
       y = "Density\n")

enp_graph_1 + enp_graph_2 + plot_layout(guides = 'collect',
                                        axes = "collect")
```

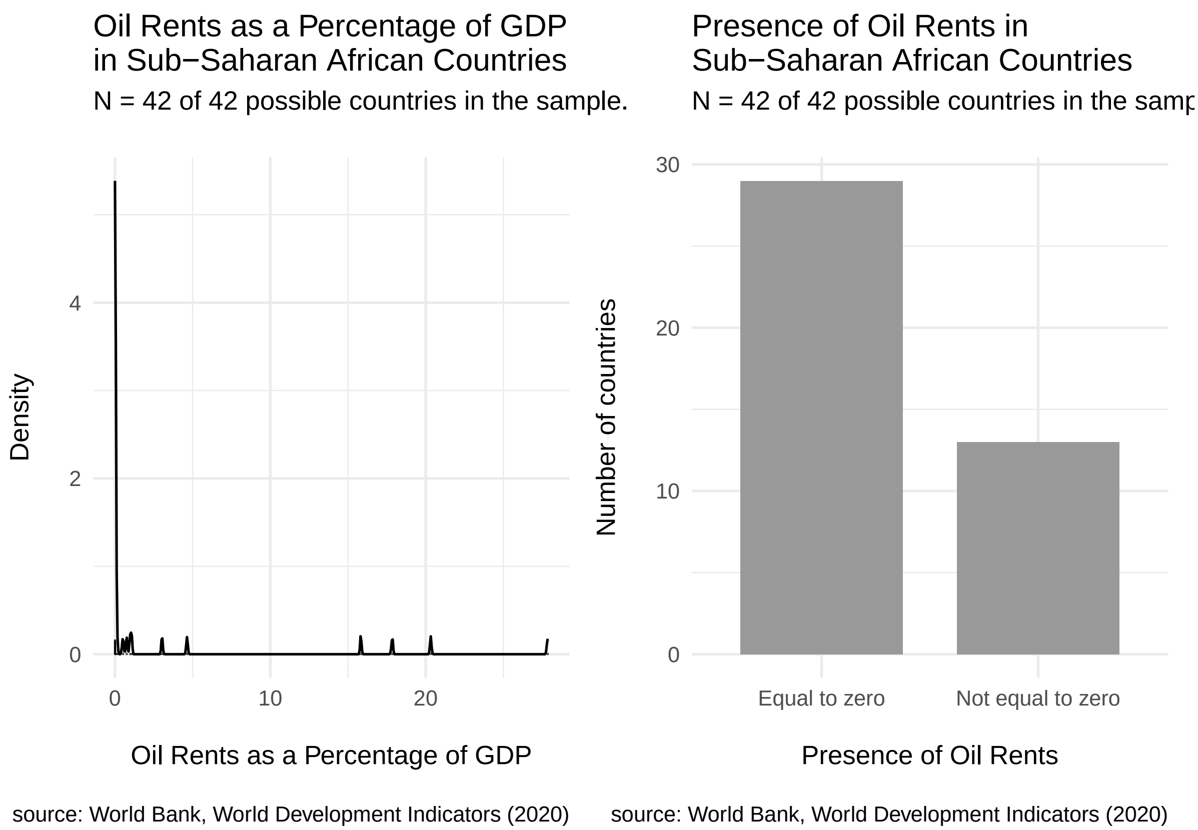


The ENPP index is calculated for each country-year pair in the sample, but for the ENEP, there are a multiple missing values. The preferred measure to validate the analytical coding is therefore ENPP.

Data on oil rents as a percentage of GDP comes from the World Bank (Bank, 2020). I construct a dummy variable for the presence of oil rents in a country-year pair.

```
oil <- wb_oil |>
  pivot_longer(cols = -c("Country Name", "Country Code", "Indicator Name",
                        "Indicator Code"),
              names_to = "year", values_to = "oil_rent_perc_gdp") |>
  mutate(ccodecow = countryname(`Country Name`,
                              destination = "cown"),
        oil_rent_dummy = factor(if_else(oil_rent_perc_gdp != 0, 1, 0),
                               levels = c(0, 1),
                               labels = c("Equal to zero",
                                          "Not equal to zero")),
        year = as.numeric(year)) |>
  filter(str_c(ccodecow, year, sep = "_") %in% cy)

oil_graph_1 <- ggplot(oil, aes(x = oil_rent_perc_gdp)) +
  geom_density() +
  geom_dotplot(fill = "grey60", binwidth = 0.03) +
  labs(title = "Oil Rents as a Percentage of GDP\nin Sub-Saharan African Countries",
      subtitle = paste0("N = ", sum(!is.na(oil$oil_rent_perc_gdp)), " of ",
                        length(oil$oil_rent_perc_gdp),
                        " possible countries in the sample.\n"),
      caption = "\nsource: World Bank, World Development Indicators (2020)",
      x = "\nOil Rents as a Percentage of GDP",
      y = "Density\n")
```

```
oil_graph_2 <- ggplot(oil, aes(x = oil_rent_dummy)) +
  geom_bar(fill = "grey60", width = 0.75) +
  labs(title = "Presence of Oil Rents in\nSub-Saharan African Countries",
       subtitle = paste0("N = ", sum(!is.na(oil$oil_rent_dummy)), " of ",
                         length(oil$oil_rent_dummy),
                         " possible countries in the sample.\n"),
       caption = "\nsource: World Bank, World Development Indicators (2020)",
       x = "\nPresence of Oil Rents",
       y = "Number of countries\n")

oil_graph_1 + oil_graph_2
```

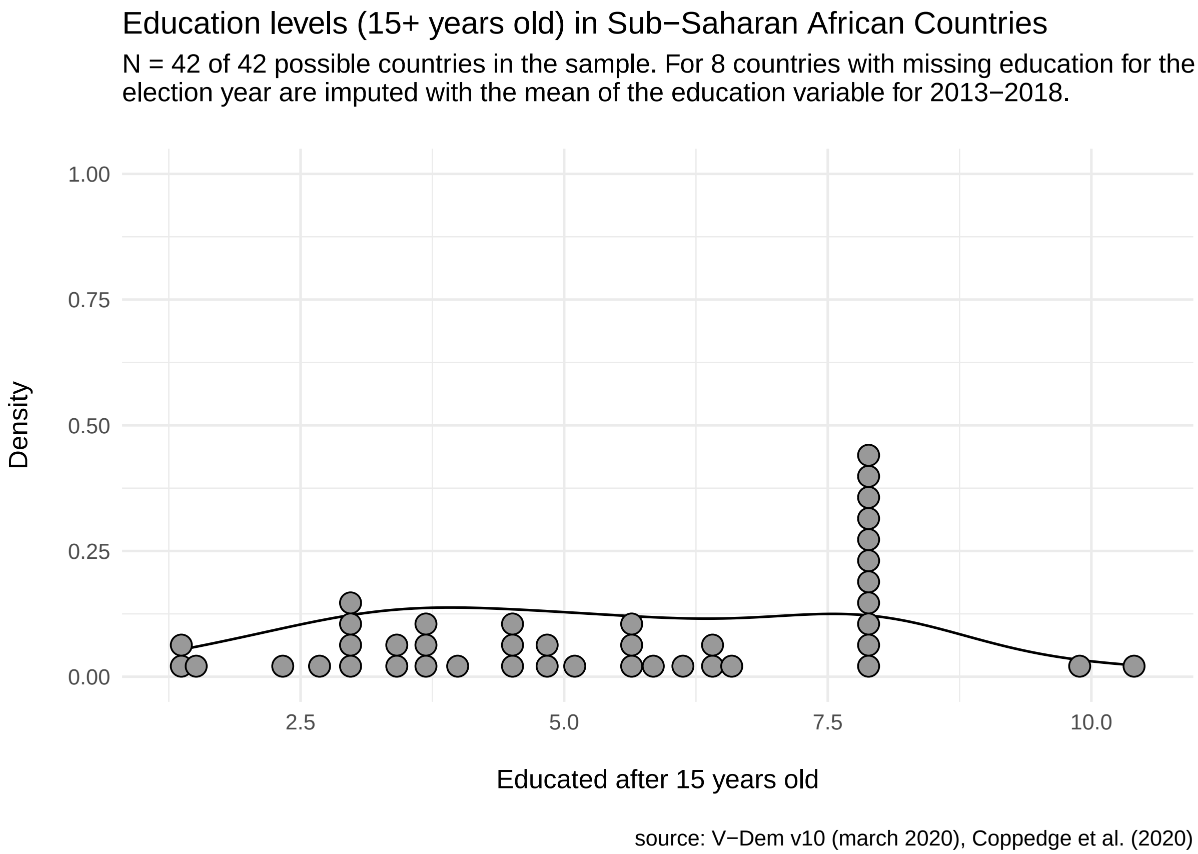


Since most of the countries in the sample do not export oil, I present both possible variables: the percentage of oil rents in GDP and a dummy variable for the presence of oil rents. Normally, a continuous measurement would be preferable, but the distribution is very skewed, hence the dummy variable seems to be a more elegant solution.

The data for education also comes from the V-Dem dataset (Coppedge et al., 2020). The variable is the average years of education for individuals aged 15 and older.

```
edu <- vdem |>
  select(country_name, ccodecow = COWcode, year, e_peaveduc) |>
  mutate(e_peaveduc_mean = if_else(is.na(e_peaveduc)
```

```r
                                    & str_c(ccodecow, year, sep = "_") %in% cy,
                                    round(mean(e_peaveduc[year >= 2013
                                                           & year <= 2018],
                                         na.rm = T), 3), e_peaveduc),
        e_peaveduc_next = if_else(is.na(e_peaveduc)
                                    & str_c(ccodecow, year, sep = "_") %in% cy,
                                    lead(e_peaveduc, 1), e_peaveduc),
        # Does not work
        e_peaveduc_prev = if_else(is.na(e_peaveduc)
                                    & str_c(ccodecow, year, sep = "_") %in% cy,
                                    lag(e_peaveduc, 1), e_peaveduc)
        # Does not work
        ) |>
  filter(str_c(ccodecow, year, sep = "_") %in% cy)

ggplot(edu, aes(x = e_peaveduc_mean)) +
  geom_density() +
  geom_dotplot(fill = "grey60", binwidth = 0.2) +
  labs(title = "Education levels (15+ years old) in Sub-Saharan African Countries",
       subtitle = paste0("N = ", sum(!is.na(edu$e_peaveduc_mean)), " of ",
                         length(edu$e_peaveduc_mean),
                         " possible countries in the sample. For ",
                         sum(!is.na(edu$e_peaveduc_mean))
                         - sum(!is.na(edu$e_peaveduc)),
                         " countries with missing education for the\n",
                         "election year are imputed with the mean of the education ",
                         "variable for 2013-2018.\n"),
       caption = "\nsource: V-Dem v10 (march 2020), Coppedge et al. (2020)",
       x = "\nEducated after 15 years old",
       y = "Density\n")
```



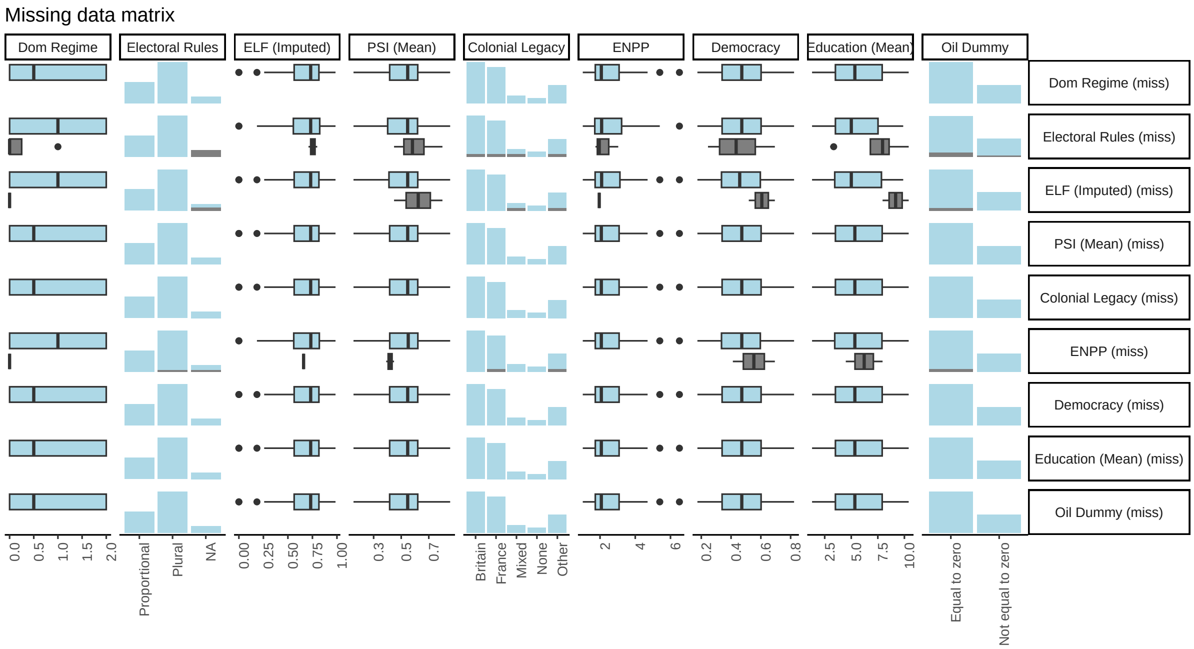
As with the PSI index, there are missing values for some of the specific country-years. I impute these with the mean of the education variable for the years 2013-2018. There is no alternative imputation, as imputing with a lag or a lead still does not alleviate the issue for the missing country-years.

Lastly, I combine the data into a single dataset.

```
data_built <- party_class |>
  left_join(col_legacy) |>
  left_join(psi) |>
  left_join(dem) |>
  left_join(elf) |>
  left_join(housesys) |>
  left_join(enp) |>
  left_join(oil) |>
  left_join(edu)

write_rds(data_built, here("data", "data_built", "data.built.rds"))
```

As the data has missing values, I present a missing data matrix.

```
var_label(data_built) <- list(
  country_name = "Country Name",
  ccodecow = "Country Code",
  year = "Year",
  partyreg = "Dom Regime",
  col.britain = "Colonized by Britain",
  col.france = "Colonized by France",
  col.other = "Colonized by Other",
  col.none = "Not Colonized",
  col.type = "Colonial Legacy",
  v2xps_party_mean = "PSI (Mean)",
  v2xos_party_next = "PSI (Next Year Imputed)",
  v2x_polyarchy = "Democracy",
  elf = "ELF",
  elf_fill = "ELF (Imputed)",
  housesys = "Electoral Rules",
  enpp = "ENPP",
  enep = "ENEP",
  e_peaveduc_mean = "Education (Mean)",
  e_peaveduc_next = "Education (Next Year Imputed)",
  oil_rent_perc_gdp = "Oil Rents (%)",
  oil_rent_dummy = "Oil Dummy"
)

missing_pairs(data_built, dependent = "partyreg",
              explanatory = c("housesys", "elf_fill", "v2xps_party_mean",
                              "col.type", "enpp", "v2x_polyarchy",
                              "e_peaveduc_mean", "oil_rent_dummy")) +
    theme(axis.text.x = element_text(angle = 90, hjust = 1),
          strip.text.y = element_text(angle = 0))
```



References

Bank, W. (2020). Oil rents (% of GDP). Retrieved August 12, 2024, from https://data.worldbank.org

Beck, T., Clarke, G., Groff, A., Keefer, P., & Walsh, P. (2010). *New tools and new tests in comparative political economy - the database of political institutions*. Retrieved August 11, 2024, from https://documents.worldbank.org/en/publication/documents-reports/documentdetail/870551468766532480/New-tools-and-new-tests-in-comparative-political-economy-the-database-of-political-institutions

Becker, B. (2019). Colonial dates dataset (COLDAT). https://doi.org/10.7910/DVN/T9SDEW

Coppedge, M., Gerring, J., Knutsen, C. H., Lindberg, S. I., Teorell, J., Altman, D., Bernhard, M., Fish, M. S., Glynn, A., Hicken, A., Lührmann, A., Marquardt, K. L., McMann, K., Paxton, P., Pemstein, D., Seim, B., Sigman, R., Skaaning, S.-E., Staton, J., … Ziblatt, D. (2020). V-dem country-year dataset v10. https://doi.org/https://doi.org/10.23696/vdemds20

Lindberg, S. I., Düpont, N., Higashijima, M., Kavasoglu, Y. B., Marquardt, K. L., Bernhard, M., Döring, H., Hicken, A., Laebens, M., Medzihorsky, J., Neundorf, A., Reuter, O. J., Ruth–Lovell, S., Weghorst, K. R., Wiesehomeier, N., Wright, J., Alizada, N., Bederke, P., Gastaldi, L., … Seim, B. (2022). Varieties of party identity and organization (v–party) dataset v2. https://doi.org/https://doi.org/10.23696/vpartydsv2

Vogt, M., Bormann, N.-C., Rüegger, S., Cederman, L.-E., Hunziker, P., & Girardin, L. (2015). Integrating data on ethnicity, geography, and conflict: The ethnic power relations data set family. *Journal of Conflict Resolution*, *59*(7), 1327–1342. https://doi.org/10.1177/0022002715591215