

Data Building

Stepan Polikanov

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"))
```

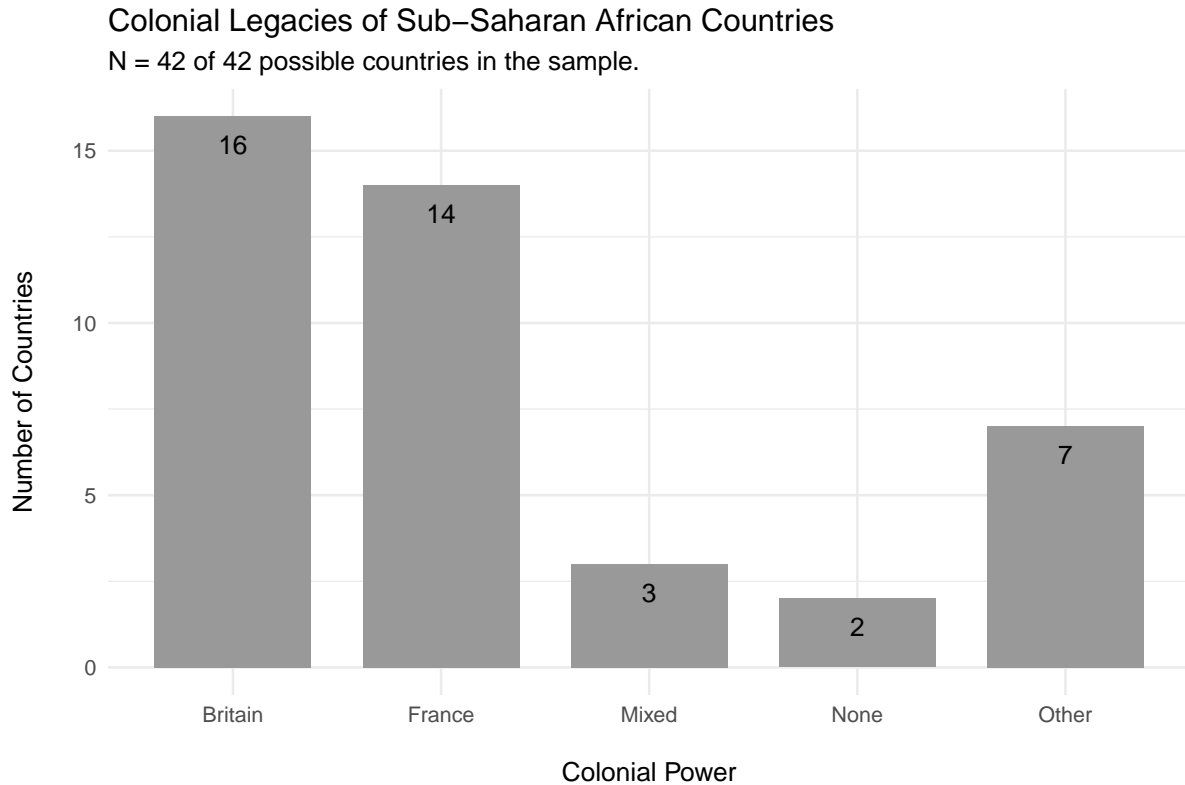
```
cy <- str_c(party_class$ccodecow, party_class$year, sep = "_")
```

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")
```



source: Colonial Dates Dataset (COLDAT), Becker et al. (2019)

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.

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

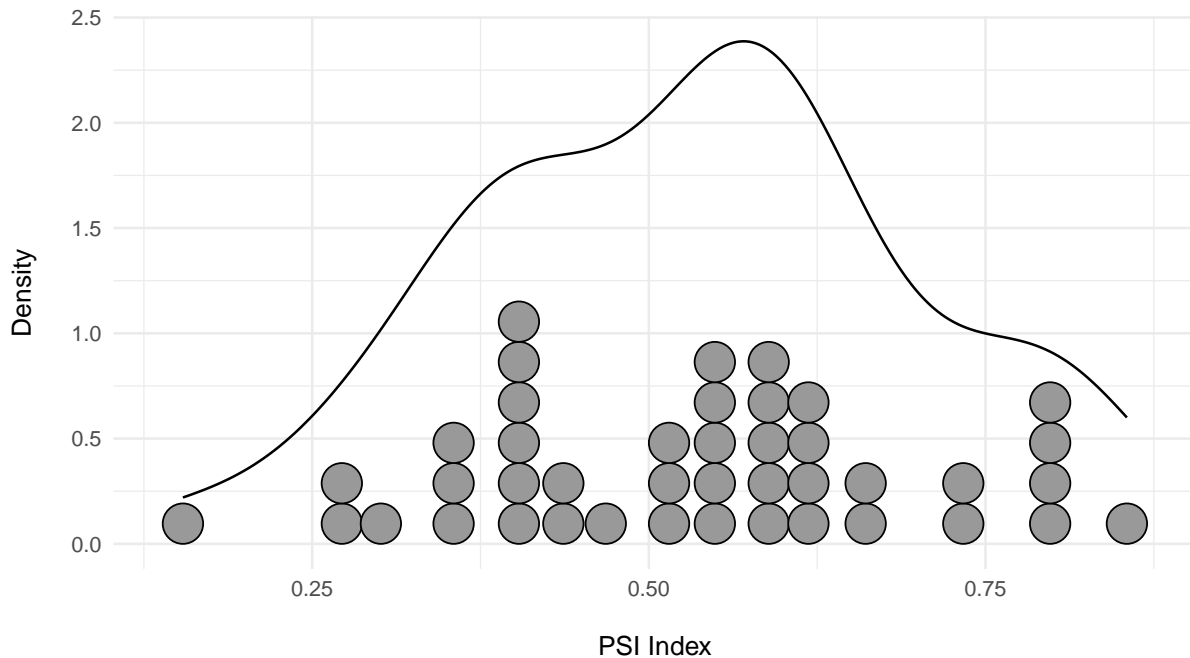
```
psi <- vdem |>
  select(country_name, ccodecow = COWcode, year, v2xps_party) |>
  mutate(v2xps_party_mean = if_else(is.na(v2xps_party)
    & str_c(ccodecow, year, sep = "_") %in% cy,
    round(mean(v2xps_party[year >= 2013
    & year <= 2018],
    na.rm = T), 3), v2xps_party),
    v2xps_party_next = 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")
```

Party System Institutionalization in Sub-Saharan African Countries

N = 42 of 42 possible countries in the sample. For 2 countries with missing PSI for the election year are imputed with the mean of the PSI index for 2013–2018.



source: V-Dem v10 (march 2020), Coppedge et al. (2020)

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.

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.

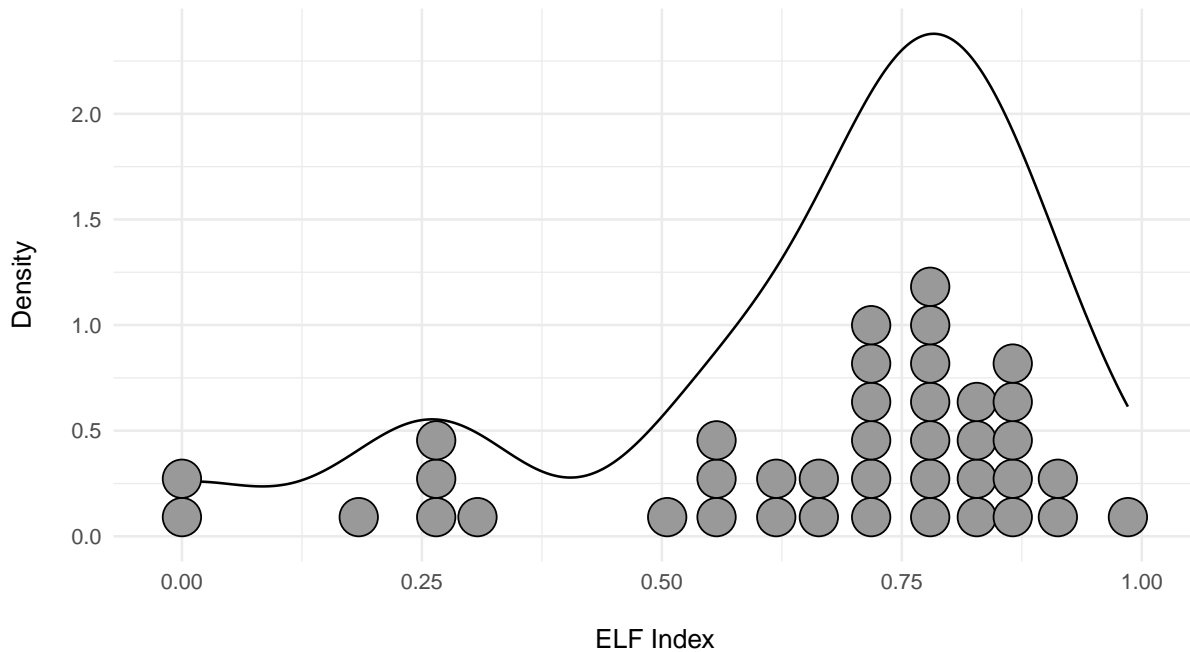
```
elf <- epr |>
  mutate(ccodecow = countryname(statename, destination = "cown")) |>
  group_by(statename, ccodecow, from, to) |>
  summarize(elf = 1 - sum(size^2)) |>
  mutate(year = list(seq(from, to))) |>
  unnest(year) |>
  ungroup() |>
  select(-from, -to) |>
  full_join(select(party_class, ccodecow,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

Ethno-linguistic fragmentation in Sub-Saharan African Countries

N = 40, of 42 possible countries in the sample.

For 5 countries the ELF index for 2018 is imputed with the 2017 value.



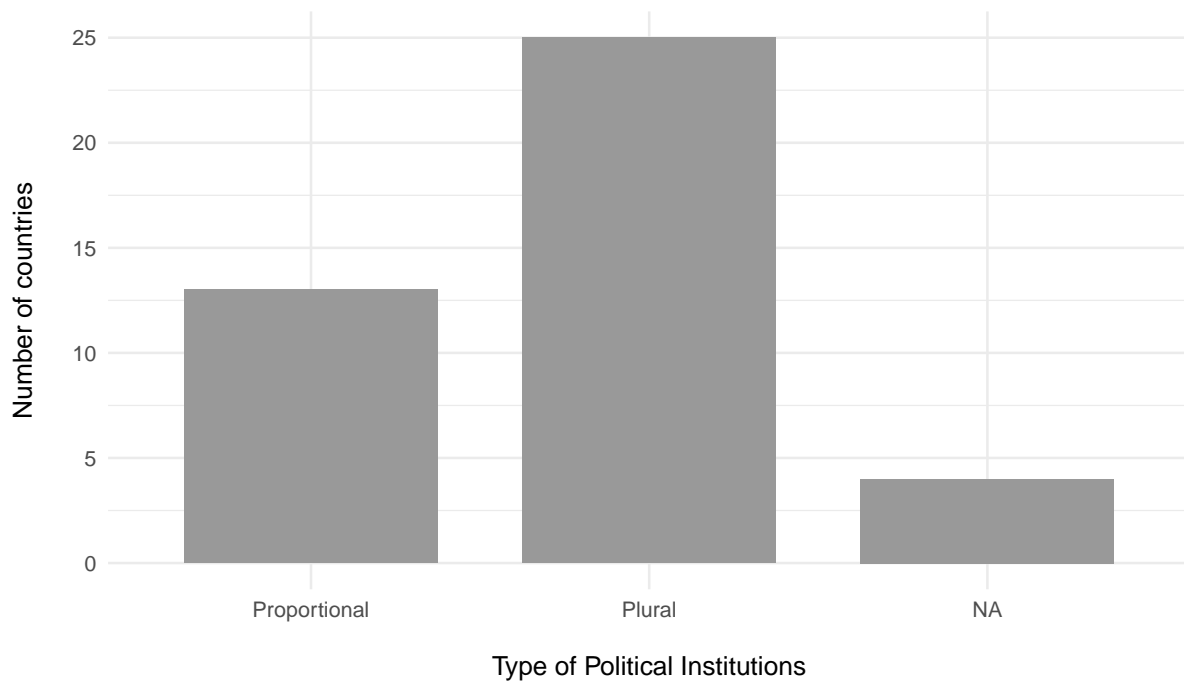
source: Ethnic Power Relations Dataset (EPR), Vogt et al. (2015)

```
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")
```

Type of Political Institutions in Sub-Saharan African Countries

N = 38 of 42 possible countries in the sample.



source: Database of Political Institutions (DPI), Arel-Bundock (2020)

References

- 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>
- 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>