Cleaning Document

CKP

Read Me

Initially, we proposed to use an R^2 metric for evaluation, and we were going to plot a sample of R^2 values from the model space. To do this, we were going to iterate through combinations of predictor variables in the models.

After looking through the files and reading through the paper more carefully, I realize that this won't work because the predictions are created from an average of 9 multinomial logit models.

My new proposal is that we use RMSE as the metric for evaluation (comparing the real values of conflict between 2010-2018 to the predictions). For our extension, instead of iterating through predictors, we can iterate through different (unweighted) combinations of the 9 models.

For our evaluation and extension, we can limit our analysis to only be looking at *any* type of conflict (=1), or no conflict (=0) (instead of the "minor" or "major" conflict disaggregation). We will also be concerned with *global* levels of conflict in any given year (instead of country level disaggregation). In part, this is to make our lives easier, but in others it's because of what is already pre-processed between the data frames.

Data Cleaning

There will be two datasets that we will have to use, one from 2013 that has all of the global **predictions** that Hegre used, by model. The other dataset will have all of the **actual values** of conflict which we can turn into a metric of global shares of conflict.

Data Set 1: Predictions from 2013, by model

```
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --
v ggplot2 3.4.1     v purrr   1.0.1
v tibble  3.2.1     v dplyr   1.1.2
v tidyr   1.3.0     v stringr 1.5.0
v readr   2.1.3     v forcats 1.0.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

```r
library(haven)
setwd("/Users/cp5045/Library/CloudStorage/OneDrive-PrincetonUniversity/limits_assignment1"

predictions <- read_dta("/Users/cp5045/Library/CloudStorage/OneDrive-PrincetonUniversity/l

Hegre_predictions <- predictions %>%
  select(year, sh_w_c, model) %>%
  filter(year >= 2010 & year <= 2018)

write.csv(Hegre_predictions, "predictions_by_model.csv")
```

Data Set 2: Actual Values of Conflict (2010 to 2018)

```r
actual_conflict <- read.csv("/Users/cp5045/Library/CloudStorage/OneDrive-PrincetonUniversi

yearly_conflict <- actual_conflict %>%
  group_by(year) %>%
  summarize(num_conflicts = sum(either_actual))

yearly_conflict_2010_2018 <- yearly_conflict %>%
  filter(year >= 2010 & year <= 2018)

actual_conflict_denoms <- read.csv("/Users/cp5045/Library/CloudStorage/OneDrive-PrincetonU

real_prop_conflict <- left_join(yearly_conflict_2010_2018, actual_conflict_denoms)
```

```
Joining with `by = join_by(year)`
```

```
real_prop_conflict <- real_prop_conflict %>%
  mutate(prop_conflict = num_conflicts/n_countries)

write.csv(real_prop_conflict, "real_prop_conflict.csv")
```

My Next Proposed Steps

For the python people!

1. Make an empty data frame which we can fill with a) the RMSE of Hegre et al.'s actual predictions and b) our iterations of the predictions based on different model combinations.
2. Fill the data frame accordingly (via nested loop or function that goes through the different iterations of model combinations). The variable `sh_w_c` in the predictions data set is the share of world conflicts (by each model), which Hegre et al. averaged across all 9 models to get their annual predicted share.
3. Plot the distribution of the RMSEs from the model combination space? Would be nice to plot this in relation to the RMSE of their actual model.