The End of History? Predicting Future Political Regimes

12/21/21

Note:

This R Markdown notebook is best viewed as an HTML, as it contains animated visualizations that will not appear in a PDF file. R Markdown will not be able to compile this file into a PDF unless the animated code block is suppressed with `eval=FALSE`; in such a case, the block below should be set to `eval=TRUE`.

Abstract

This study attempts to tackle the question: what historical factors influence the form of political regime that a country takes on, and can those factors be used to predict future trends? To this end, I collected internal data (GDP, life span, etc.) from over 100 countries from the past 200 years, in addition to data on their corresponding political regimes from the Polity5 project. The data was staggered in the fashion of a time series, using the previous ten years of data to predict a current year's political regime; after cleaning, this presented almost 8000 data points. To accomplish this task, I built a linear model using lasso and cross validation to choose the most relevant predictors. Results were ambivalent, with predictions for the year 2017 falling within 6 points of the actual value, on average. As such, it appears that these predictors and their trends are not sufficient to get a clear picture of political regime development.

Introduction

In his 1992 book, *The End of History and the Last Man*, Francis Fukuyama famously predicted that with the dissolution of the Soviet Union, the world was reaching an equilibrium in which Western liberal democracy would become the dominant—and final—form of government across the world. This raises an interesting question: what historical factors influence the form of political regime that a country takes on, and can those factors be used to predict future trends? My goal was to answer this question on the basis of data collected on countries' political regimes over the past 200 years.

For the predictors, data was collected from Gap Minder and Our World in Data. As predictors, I used only internal and objective numeric data from each country; these included: GDP per capita, Gini Coefficient, and poverty rate; infant mortality and average lifespan; population; and average number of years in school. Together, these can be taken to represent the society's economic stability and opportunity, quality of life, and education–all of which are commonly thought to be the backbone of democracy.

The previous decade's worth of these features were used to predict the country's political regime in a given year, as measured by the Polity5 Regime Assessment, which ranks political regimes on an integer scale of -10 (complete autocracy) to 10 (complete democracy). This ranking does not rely on any of the aforementioned predictors, but rather arises from a qualitative evaluation of the regime itself.

Section 1 contains examples from the data set, exploration of what it contains, and visualizations to assist in perceiving the data with regards to time and space. In Section 2, I analyze this data and build several linear models on the basis of this analysis, then select the best one, based on its high R^2. However, in Section 3, I conclude that in spite of this, the model does not work well for prediction, having a huge RMSE on both

randomized and sequential test sets; this is emphasized by a visual presentation of the model's predictions. Finally, in section 4, I discuss why this might be and suggest future avenues of study.

Section 1: Data Exploration and Visualization

Below is an example of five randomized rows from the data set. As one can see, each row contains: (1) Entity and Year identifiers; (2) the Polity5 regime assessment; (3) the previous decade's predictors, with a 1 marking the first year of that decade (i.e. 10 years before the prediction year), and a 10 marking the last. The data contains roughly 8000 rows.

```
##                        Entity Year Regime DecadePoverty GDPPerCapita1
## 7866                    Yemen 2006     -2   37.67003919          4520
## 6847              Switzerland 1977     10    0.02869855         41300
## 269                 Argentina 1993      7    6.85262091         15800
## 1336 Central African Republic 1994      5   87.04405191          1340
## 4977                    Nepal 1998      5   80.78517595          1390
##      GDPPerCapita2 GDPPerCapita3 GDPPerCapita4 GDPPerCapita5 GDPPerCapita6
## 7866          4650          4690          4850          4890          4940
## 6847         43000         45200         46500         47500         48400
## 269          14600         15500         15800         15300         14300
## 1336          1360          1270          1260          1260          1200
## 4977          1410          1460          1480          1500          1580
##      GDPPerCapita7 GDPPerCapita8 GDPPerCapita9 GDPPerCapita10 GiniCoef1
## 7866          4980          5030          5160           5170      35.3
## 6847         48800         45000         44200          45300      38.5
## 269          14100         15200         16200          17300      42.1
## 1336          1170          1060          1030           1050      64.3
## 4977          1590          1640          1680           1700      32.3
##      GiniCoef2 GiniCoef3 GiniCoef4 GiniCoef5 GiniCoef6 GiniCoef7 GiniCoef8
## 7866      35.0      35.0      34.9      34.9      34.9      34.8      34.8
## 6847      38.4      38.3      38.2      38.0      37.8      37.5      37.3
## 269       42.5      42.8      45.3      45.6      46.0      46.4      46.8
## 1336      63.8      63.3      62.8      62.3      61.8      61.7      61.3
## 4977      32.8      33.2      33.7      34.2      34.7      35.2      36.2
##      GiniCoef9 GiniCoef10 InfantMortality1 InfantMortality2 InfantMortality3
## 7866      34.7       34.9            107.0            103.0             99.3
## 6847      36.9       36.7             19.9             19.2             18.4
## 269       45.5       44.9             32.1             30.6             29.6
## 1336      61.0       60.7            179.0            178.0            177.0
## 4977      37.2       38.2            147.0            140.0            133.0
##      InfantMortality4 InfantMortality5 InfantMortality6 InfantMortality7
## 7866             94.9             90.3             85.6             81.1
## 6847             17.6             16.7             15.8             14.8
## 269              29.1             28.9             28.8             28.6
## 1336            178.0            179.0            180.0            180.0
## 4977            126.0            119.0            113.0            107.0
##      InfantMortality8 InfantMortality9 InfantMortality10 LifeSpan1 LifeSpan2
## 7866             76.7             72.5              68.4      62.0      62.5
## 6847             13.8             12.9              12.1      73.0      73.1
## 269              28.1             27.3              26.3      70.8      71.7
## 1336            179.0            179.0             178.0      49.5      49.5
## 4977            101.0             96.0              90.8      57.6      58.3
```

```
##      LifeSpan3 LifeSpan4 LifeSpan5 LifeSpan6 LifeSpan7 LifeSpan8 LifeSpan9
## 7866      63.0      63.5      64.0      64.5      64.9      65.3      65.8
## 6847      73.3      73.6      74.0      74.3      74.7      75.0      75.3
## 269       72.0      72.1      72.1      72.3      72.5      72.7      72.8
## 1336      49.4      49.0      48.7      48.1      47.5      46.8      46.0
## 4977      58.9      59.8      60.3      61.2      61.8      62.6      63.4
##      LifeSpan10 Population1 Population2 Population3 Population4 Population5
## 7866       66.2    16000000    16500000    16900000    17400000    17900000
## 6847       75.5     6040000     6100000     6150000     6200000     6250000
## 269        73.0    29700000    30200000    30700000    31200000    31700000
## 1336       45.3     2540000     2600000     2650000     2690000     2750000
## 4977       64.0    18400000    18900000    19400000    19900000    20500000
##      Population6 Population7 Population8 Population9 Population10
## 7866    18400000    19000000    19500000    20100000     20700000
## 6847     6290000     6320000     6340000     6340000      6320000
## 269     32100000    32600000    33100000    33500000     34000000
## 1336     2810000     2880000     2960000     3050000      3140000
## 4977    21000000    21600000    22100000    22600000     23100000
##      YearsSchooling1.5 YearsSchooling6.10
## 7866              0.80               1.30
## 6847              9.76              10.34
## 269               7.45               7.88
## 1336              1.49               2.06
## 4977              2.02               2.20
```

As a result of the data cleaning that resulted in this data, thousands of rows needed to be thrown away. Thus, not every country is represented equally in the data, nor is every year represented equally. Below are tables with the 5 least represented and 5 most represented countries and years in the data, accompanied by the mean and median frequencies of each.

```
## [1] "Median frequency (country): 49"
```

```
## [1] "Mean frequency (country): 56"
```

```
##             Entity Freq
## 1       Montenegro    2
## 2     Turkmenistan    2
## 3          Lebanon    3
## 4    Guinea-Bissau    7
## 5          Comoros    8
## 137    Switzerland   97
## 138 United Kingdom   97
## 139  United States   97
## 140        Uruguay   97
## 141      Venezuela   97
```

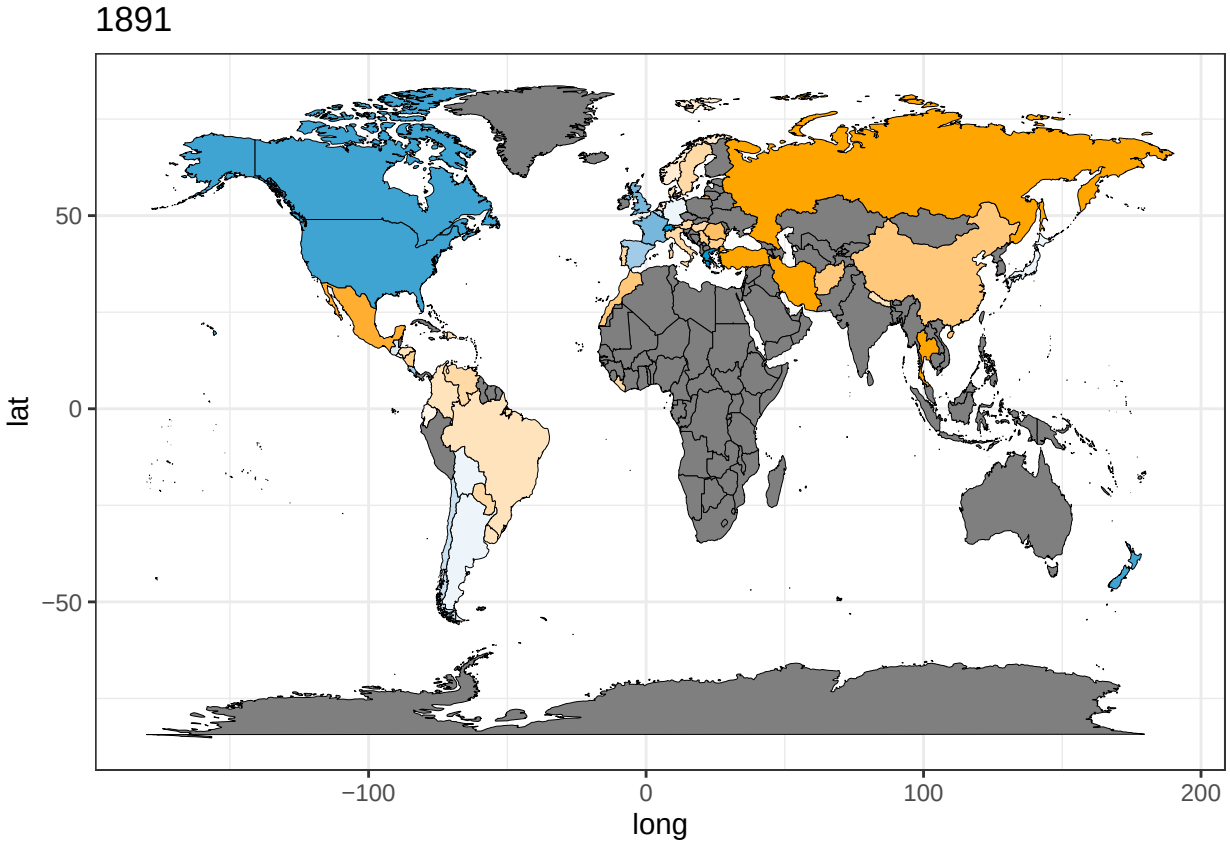
```
## [1] "Median frequency (year): 72"
```

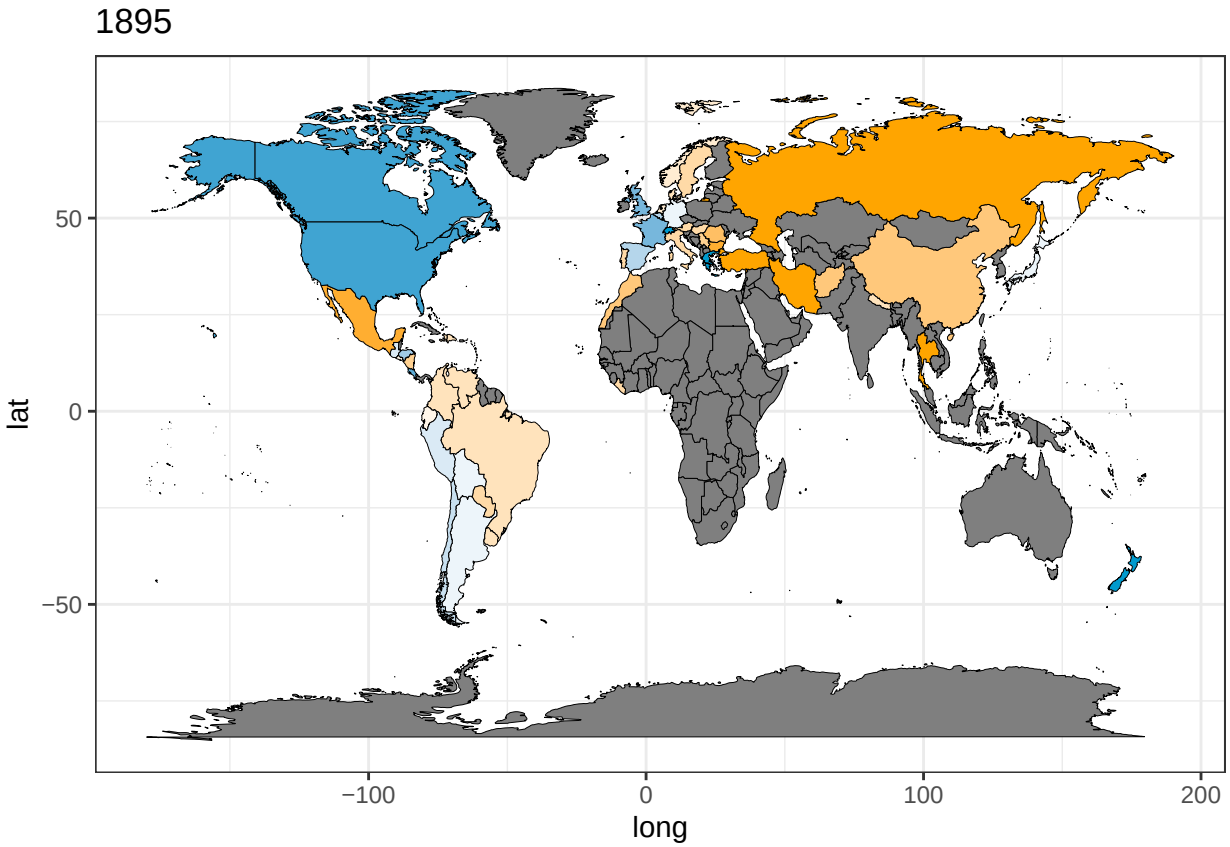
```
## [1] "Mean frequency (year): 81"
```

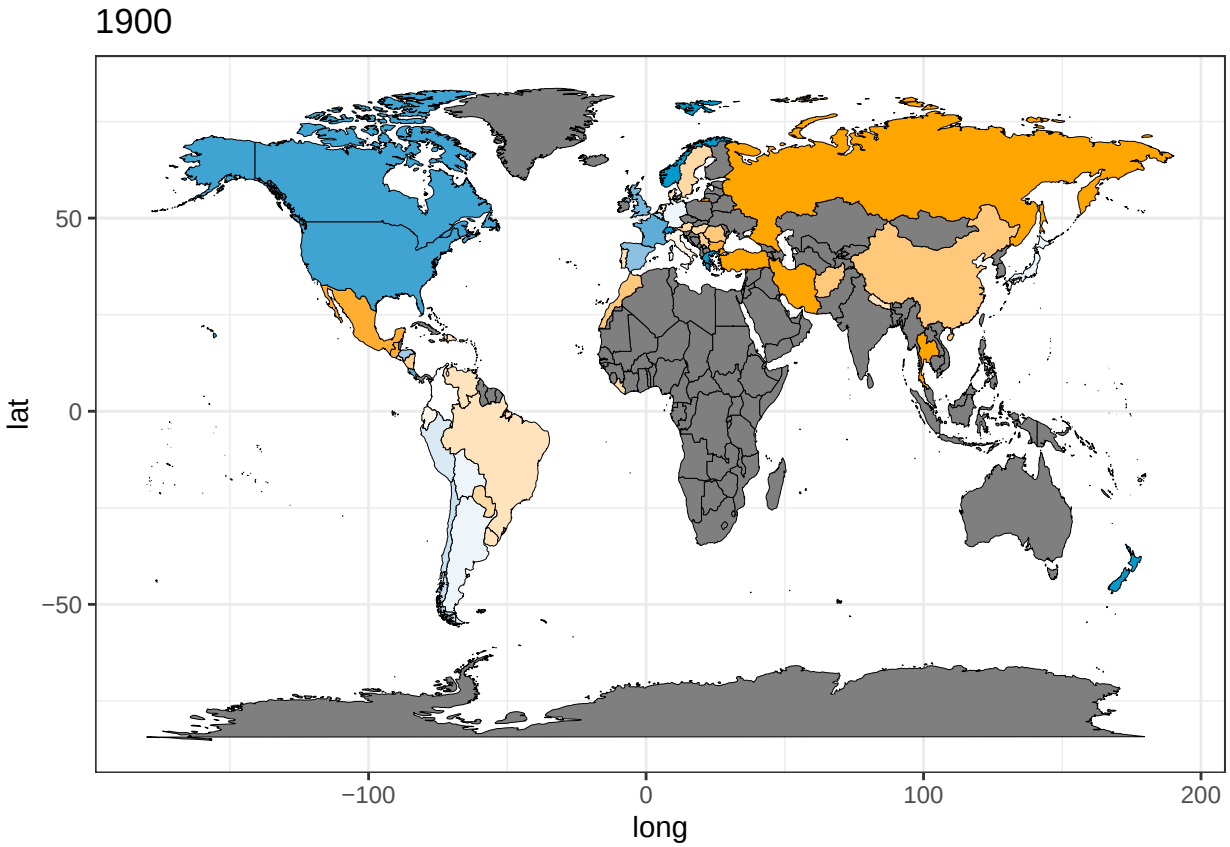
```
##    Year Freq
## 1  1918   43
```

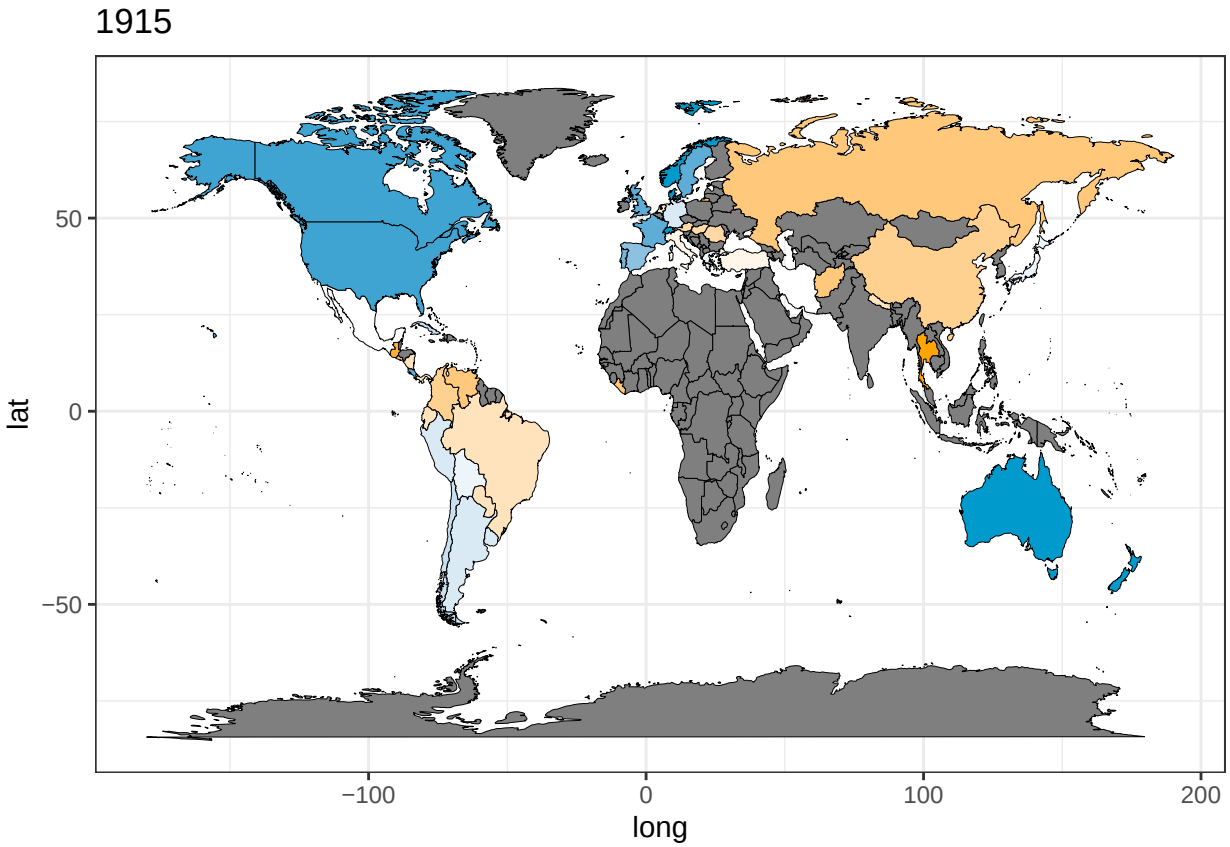
```
## 2  1919   43
## 3  1916   44
## 4  1917   44
## 5  1920   44
## 93 2013  134
## 94 2014  134
## 95 2015  135
## 96 2016  138
## 97 2017  138
```

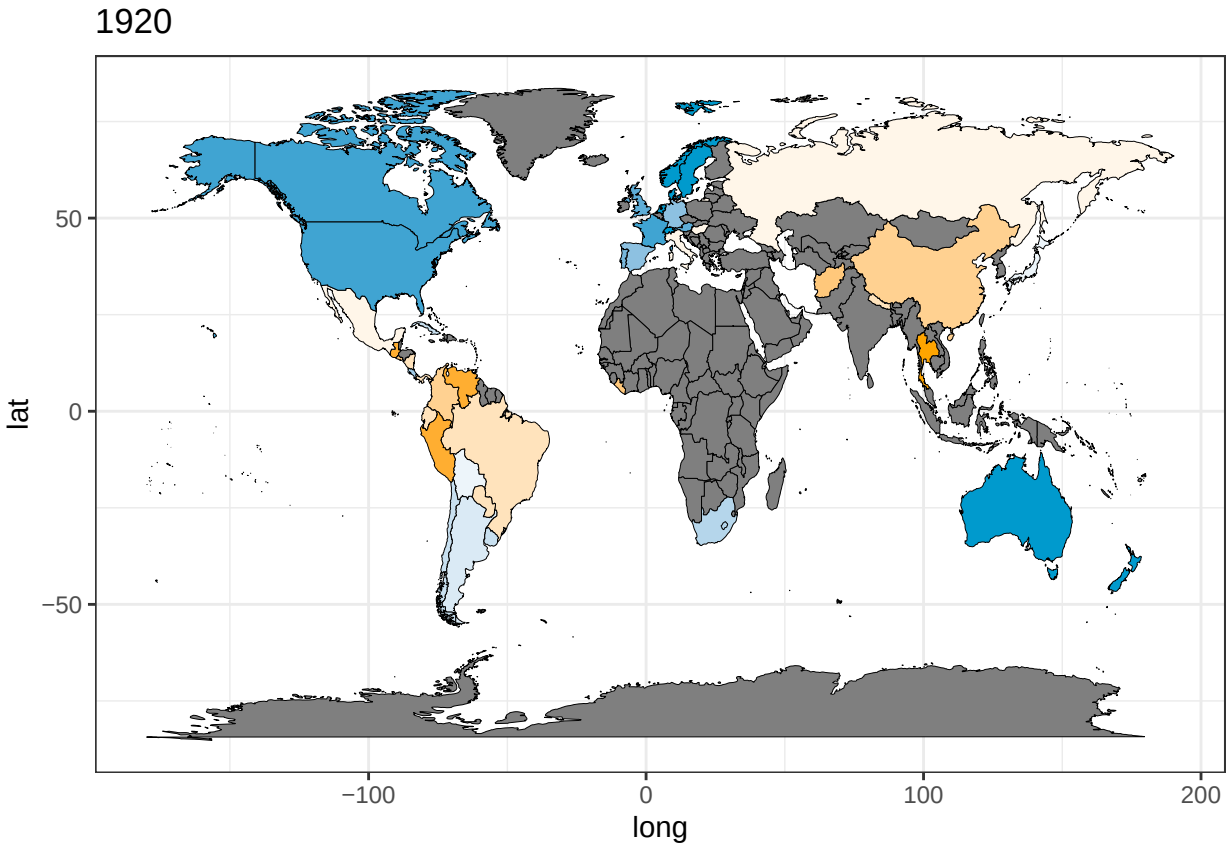
It is difficult, however, to conceptualize this and the impact it may have on our model. To assist in this, below is an animated world map contained 20 evenly spaced sample years from the data. Through it, one can see the changes in the data set from year to year, and whether it will have a major impact.

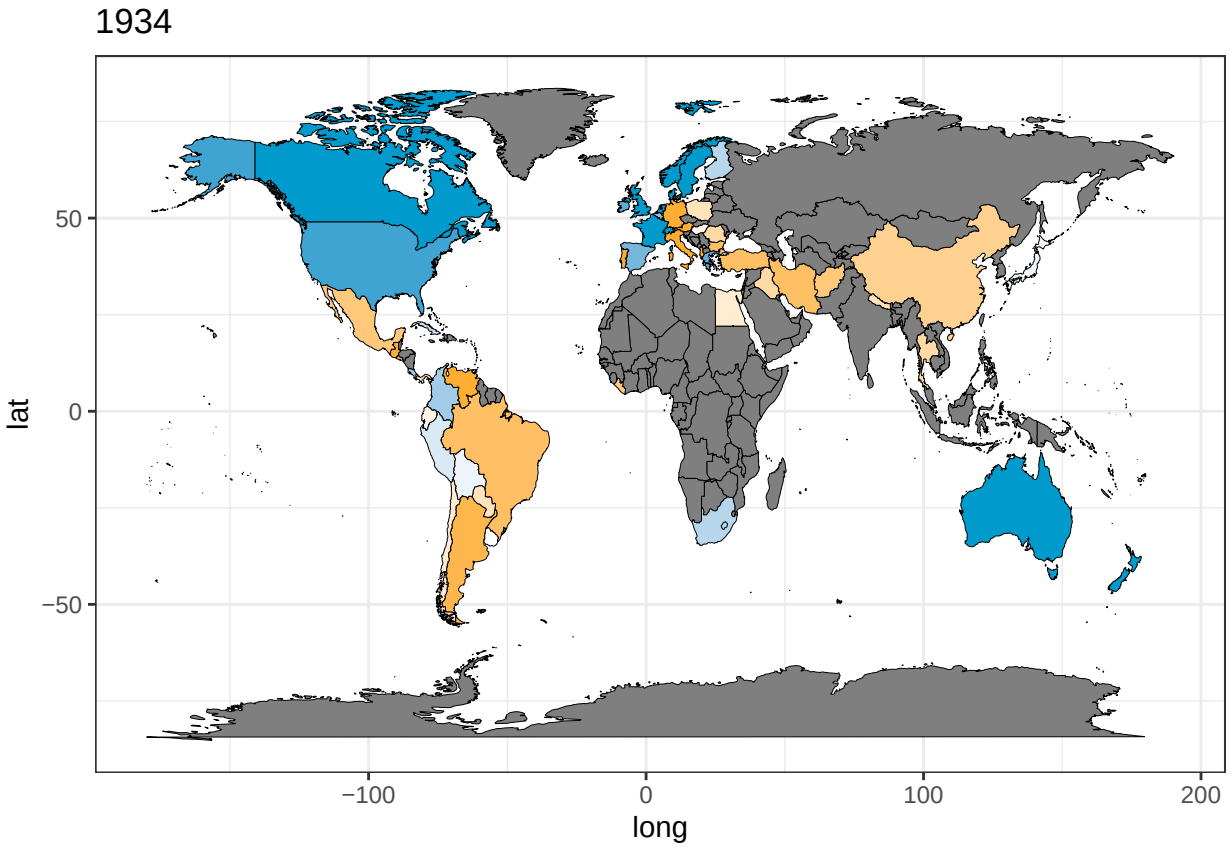


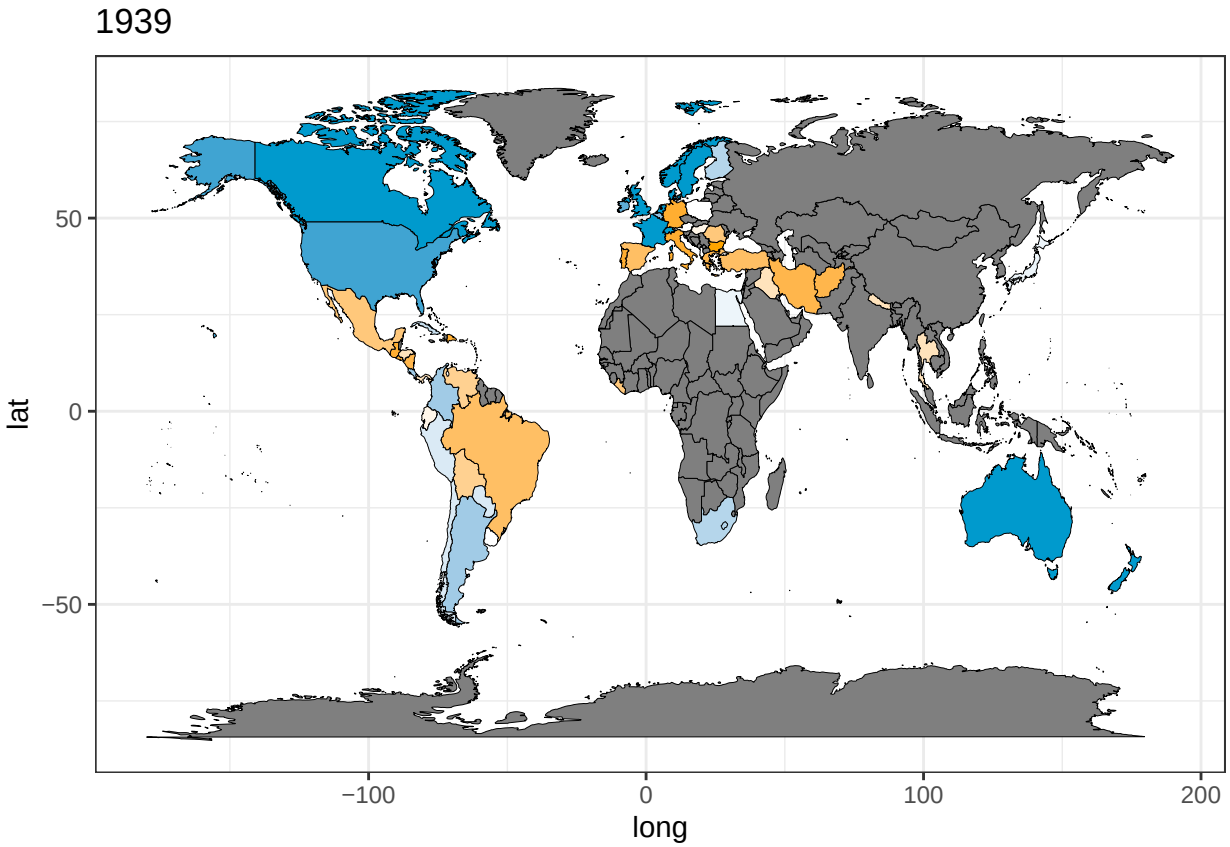

1895
50
0
−50
lat
−100
0
100
200
long

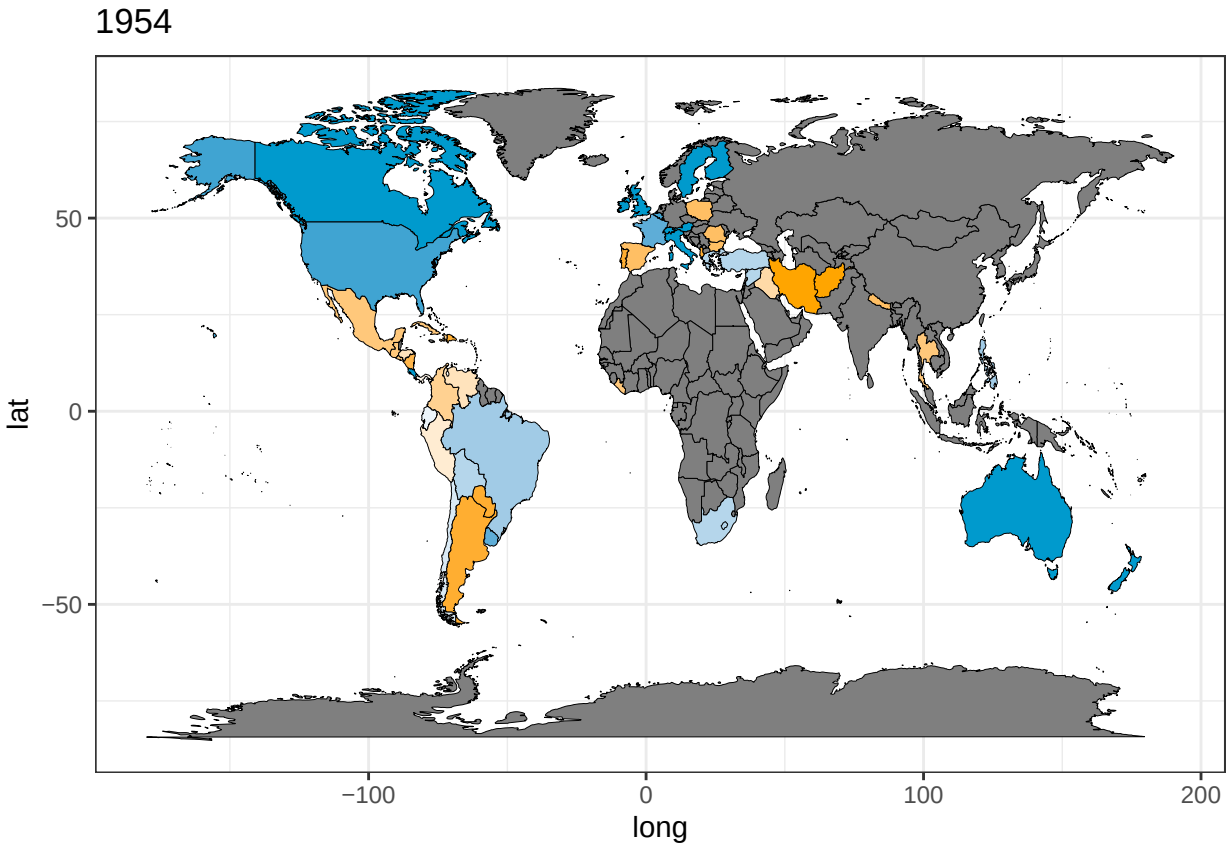

1900
50
0
−50
lat
−100
0
100
200
long

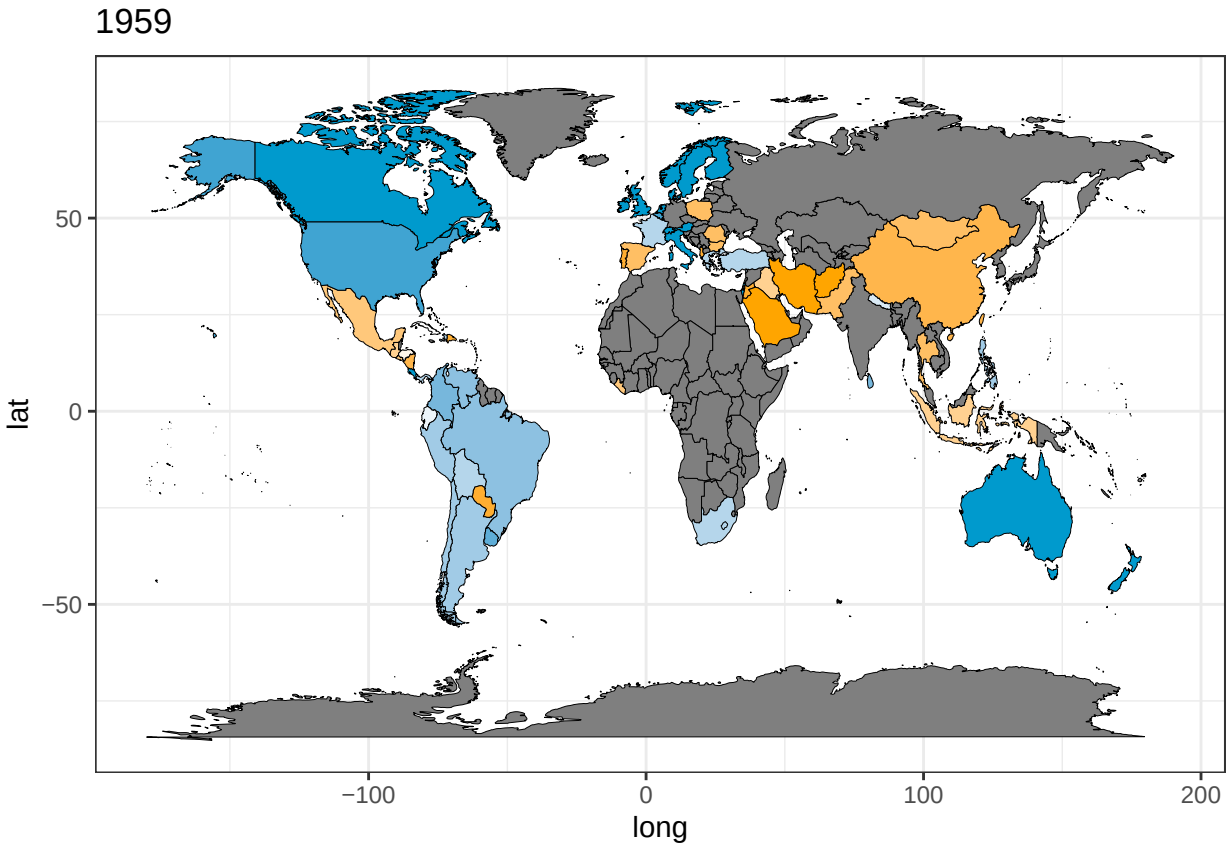

1915
50
0
−50
lat
−100
0
100
200
long

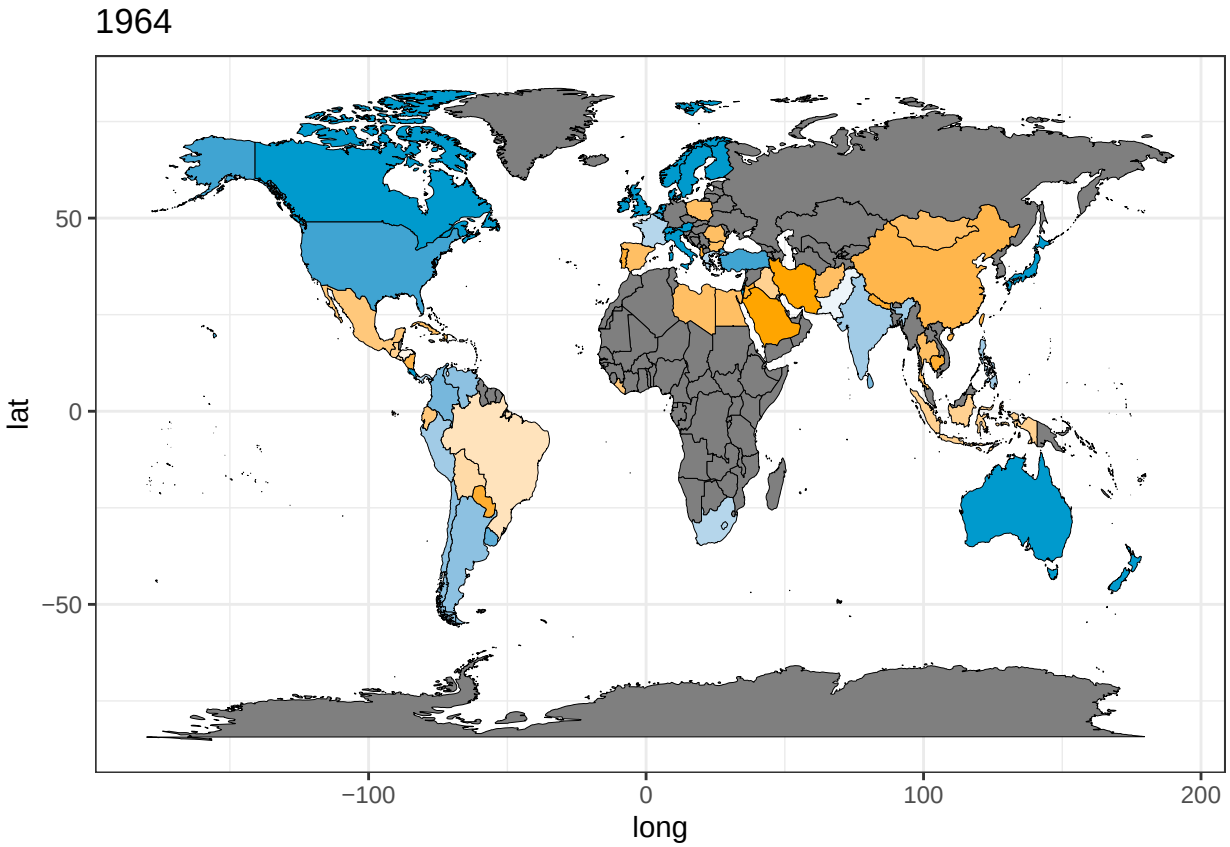

1920
50
0
−50
lat
−100
0
100
200
long

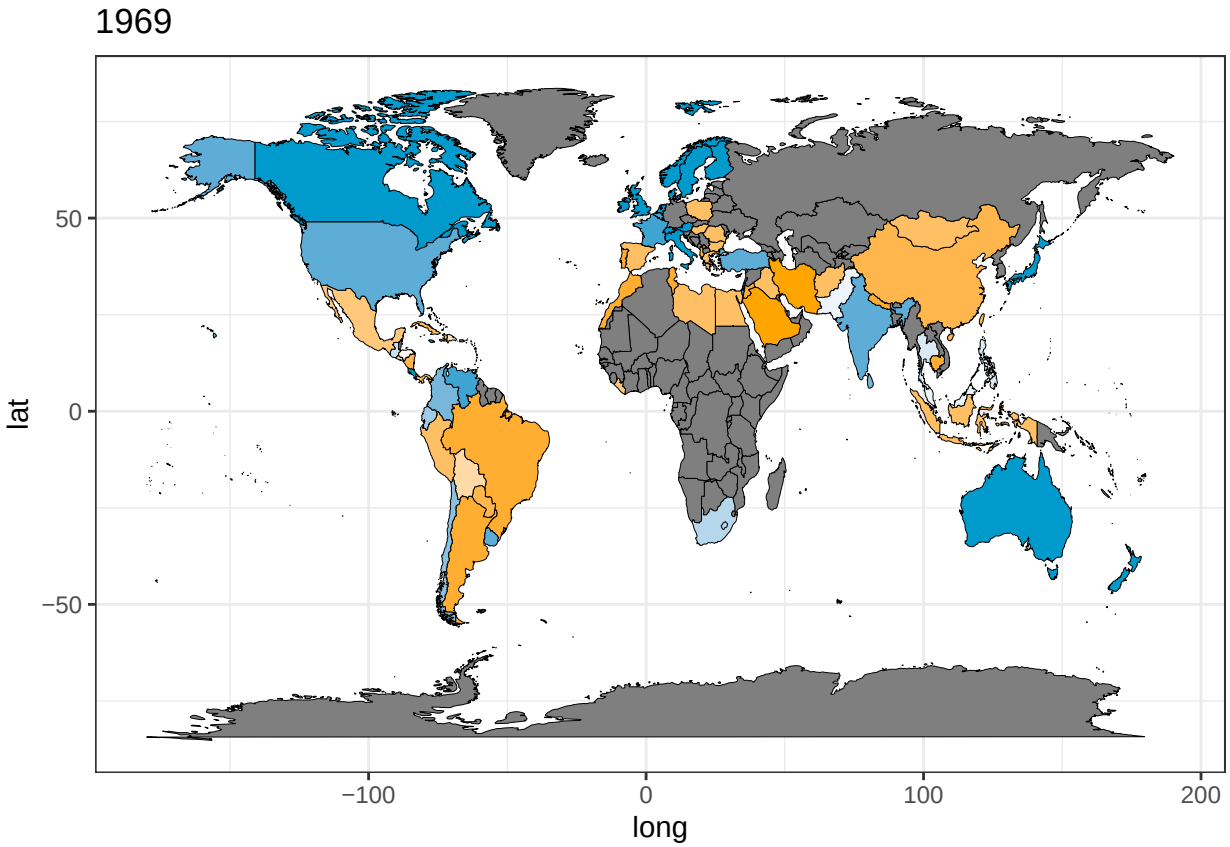

1934
50
0
−50
lat
−100
0
100
200
long

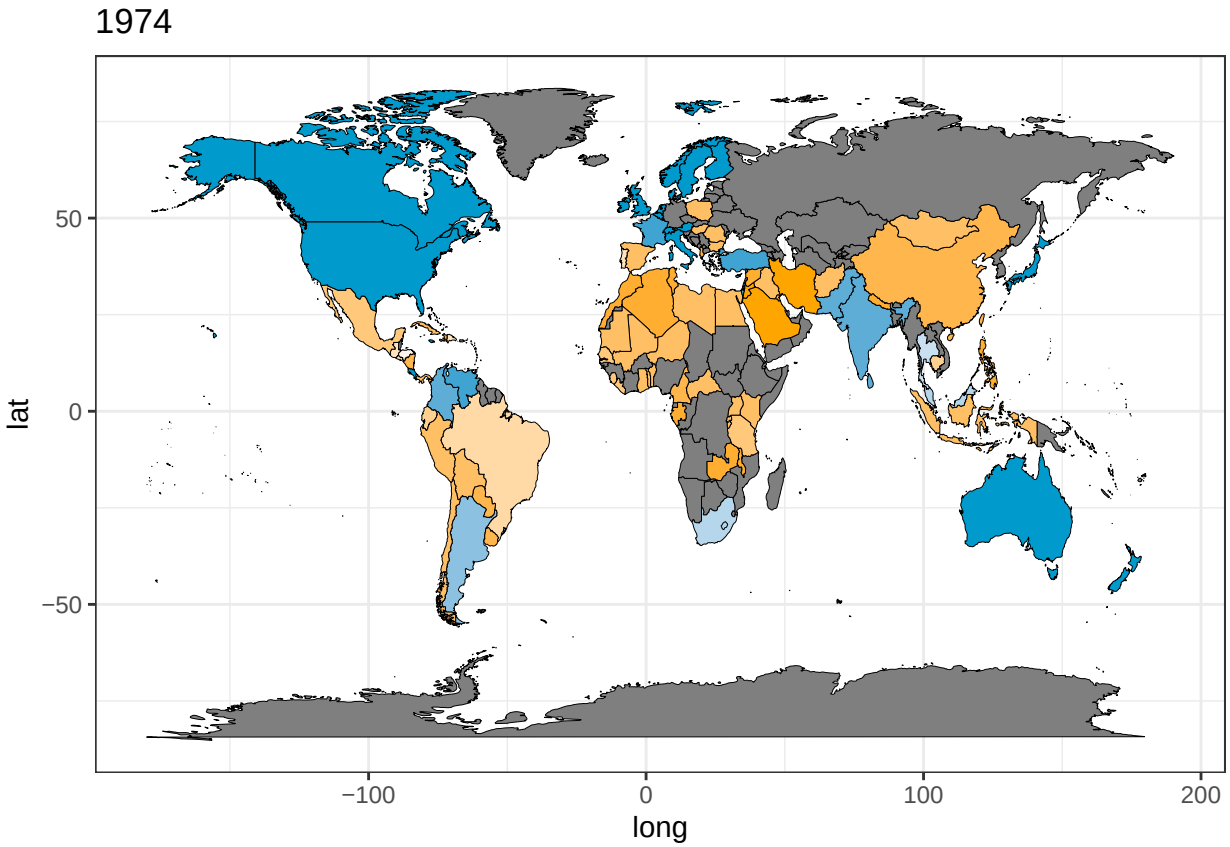

1939
50
0
−50
lat
−100
0
100
200
long

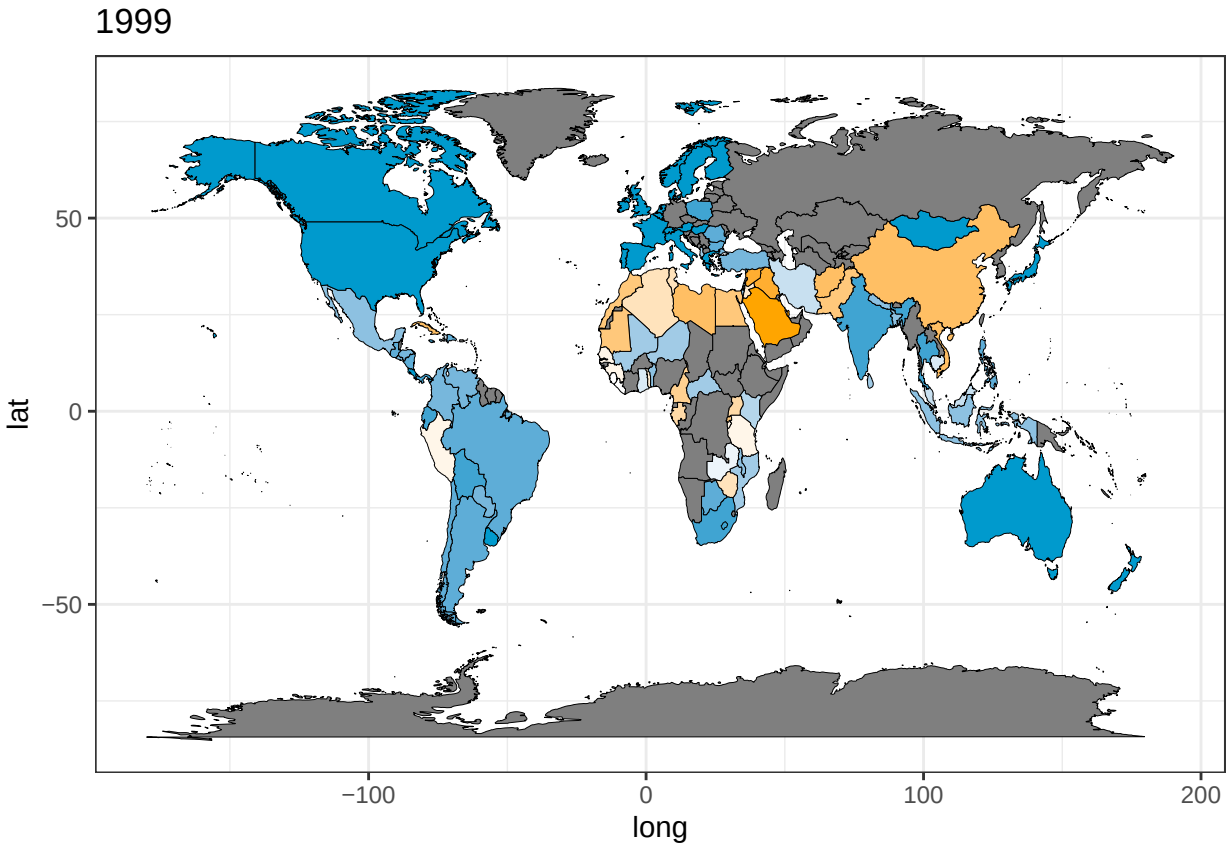

1954
50
0
−50
lat
−100
0
100
200
long

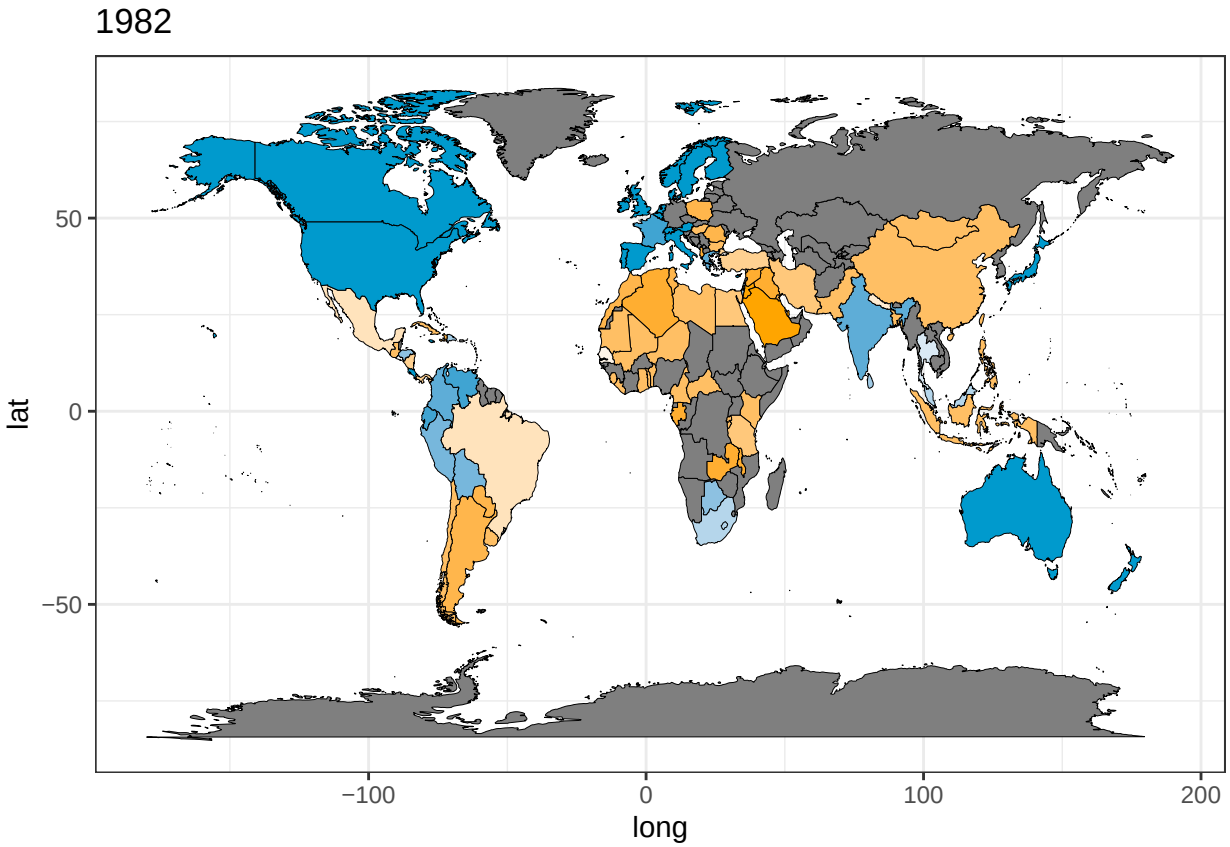

1959
50
0
−50
lat
−100
0
100
200
long

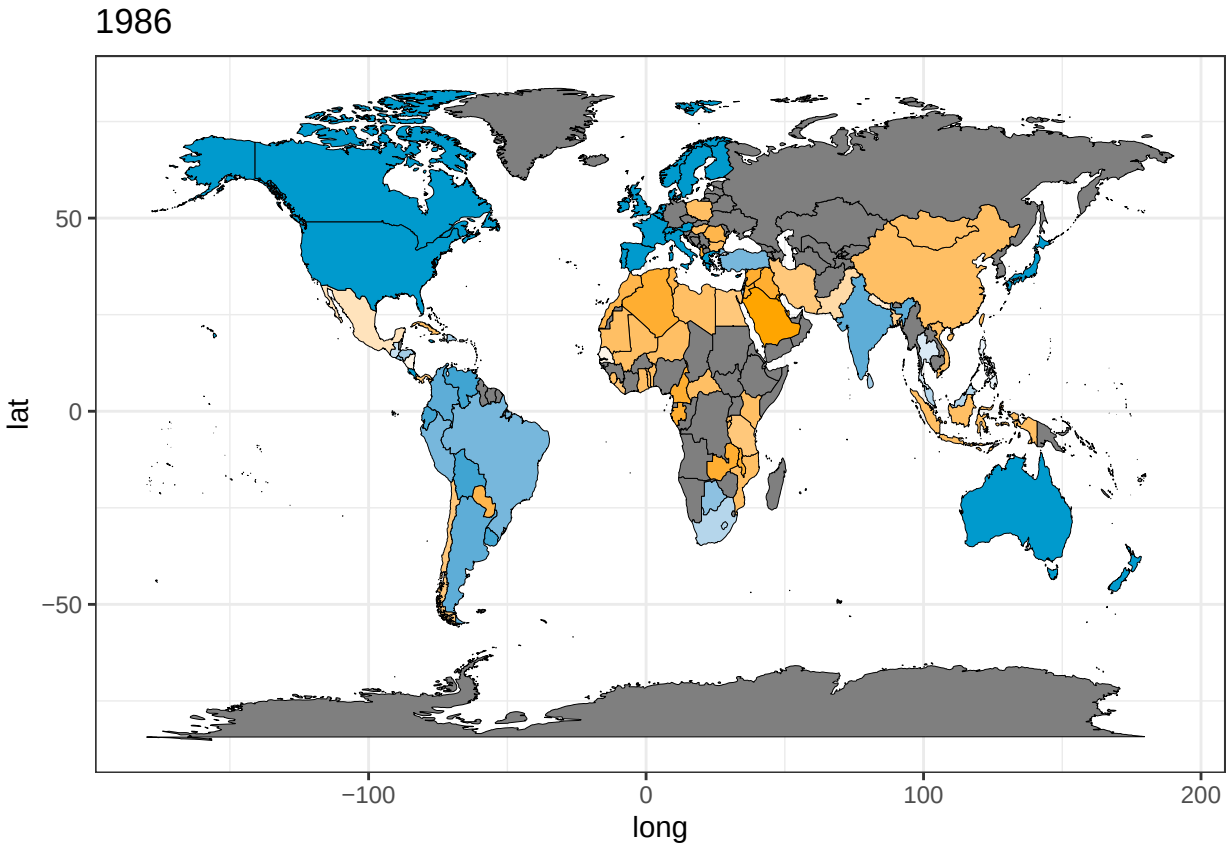

1964
50
0
−50
lat
−100
0
100
200
long

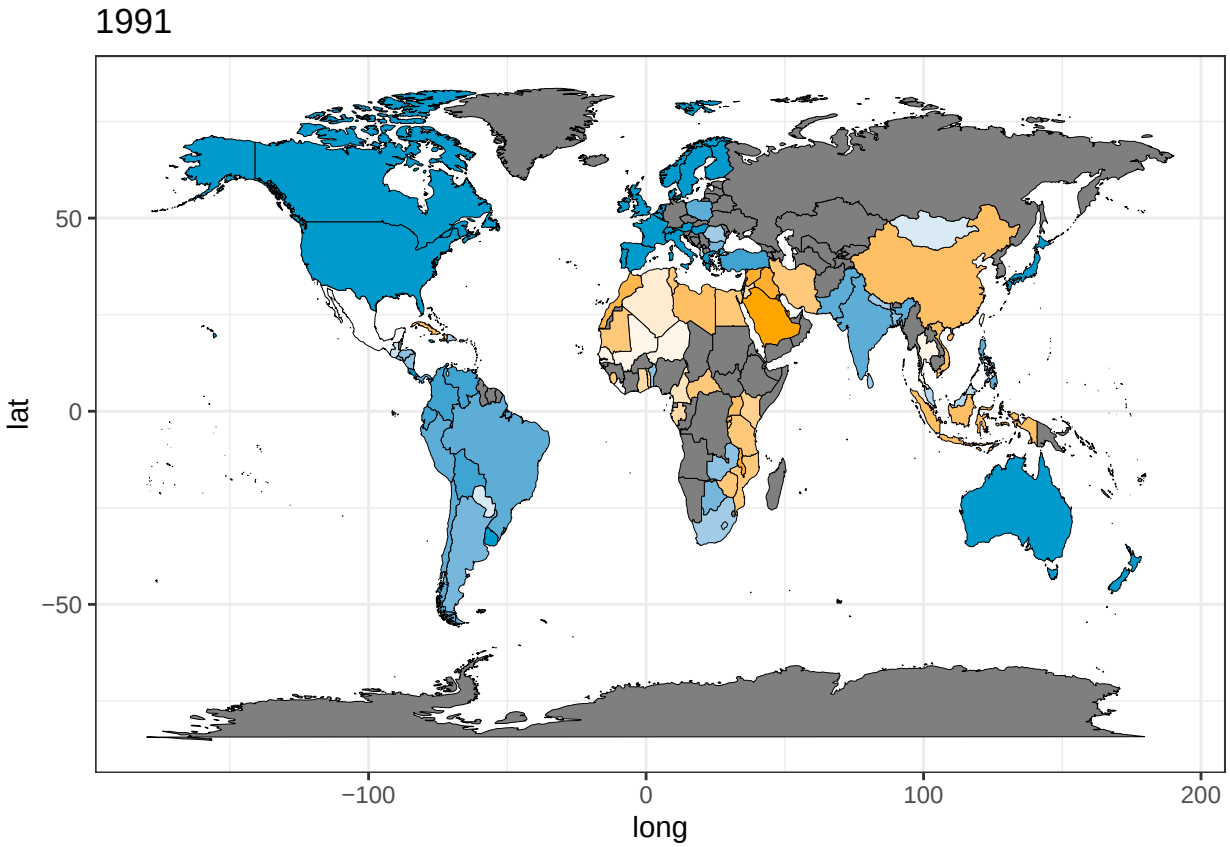

1969
50
0
−50
lat
−100
0
100
200
long

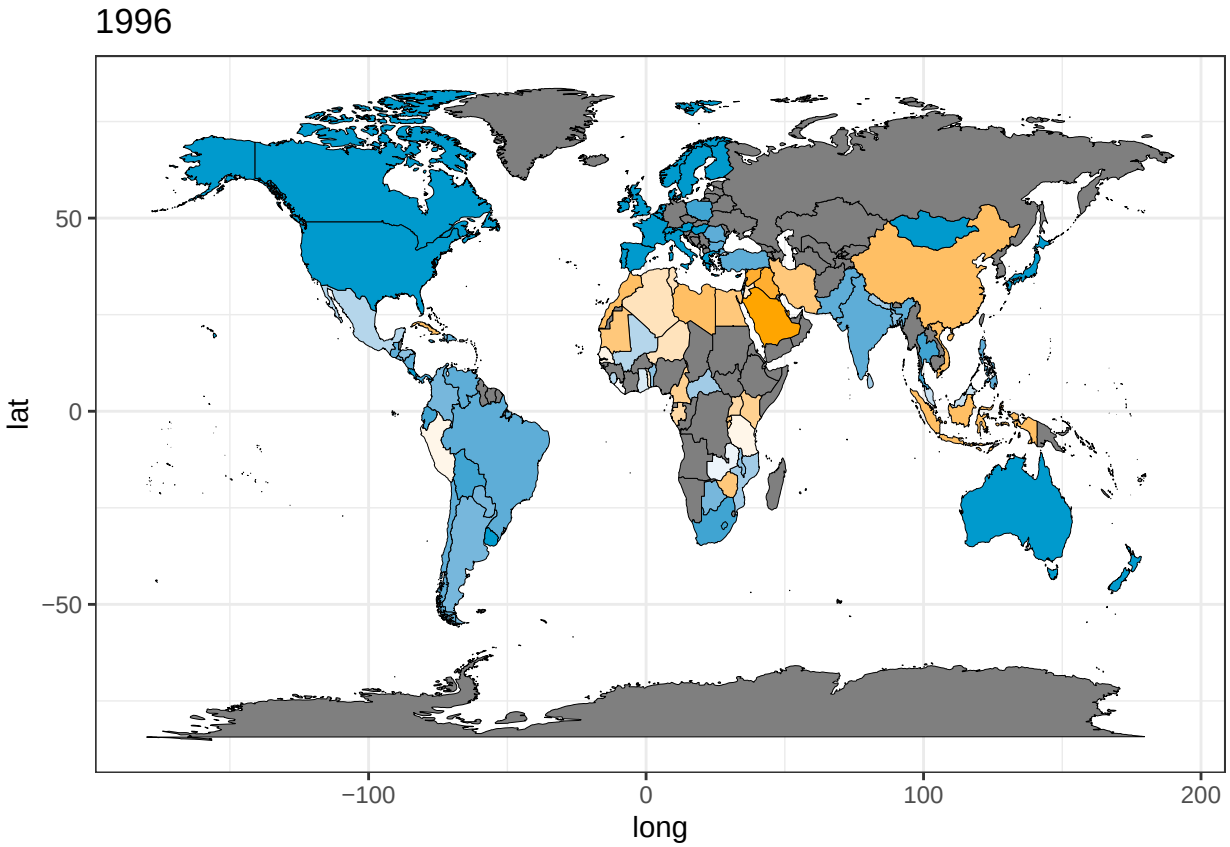

1974
50
0
−50
lat
−100
0
100
200
long

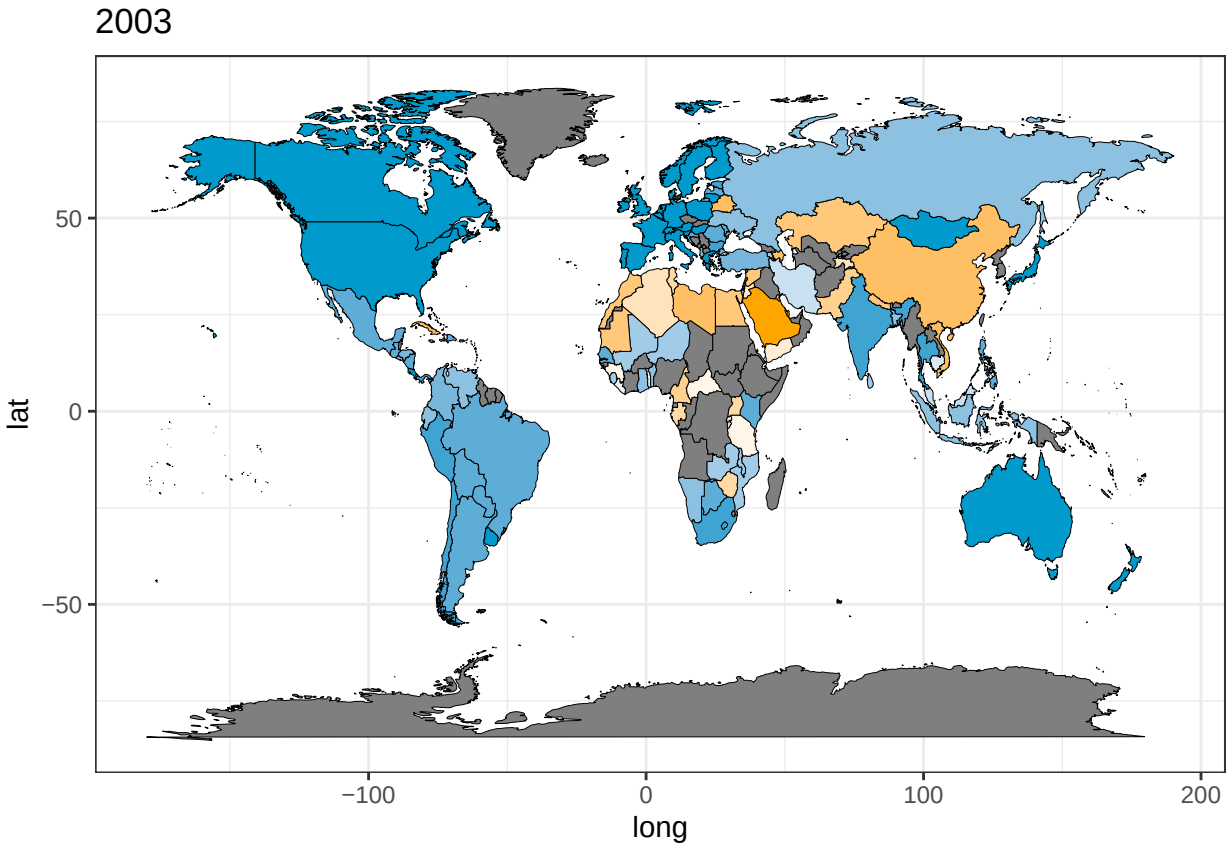

1999
50
0
−50
lat
−100
0
100
200
long

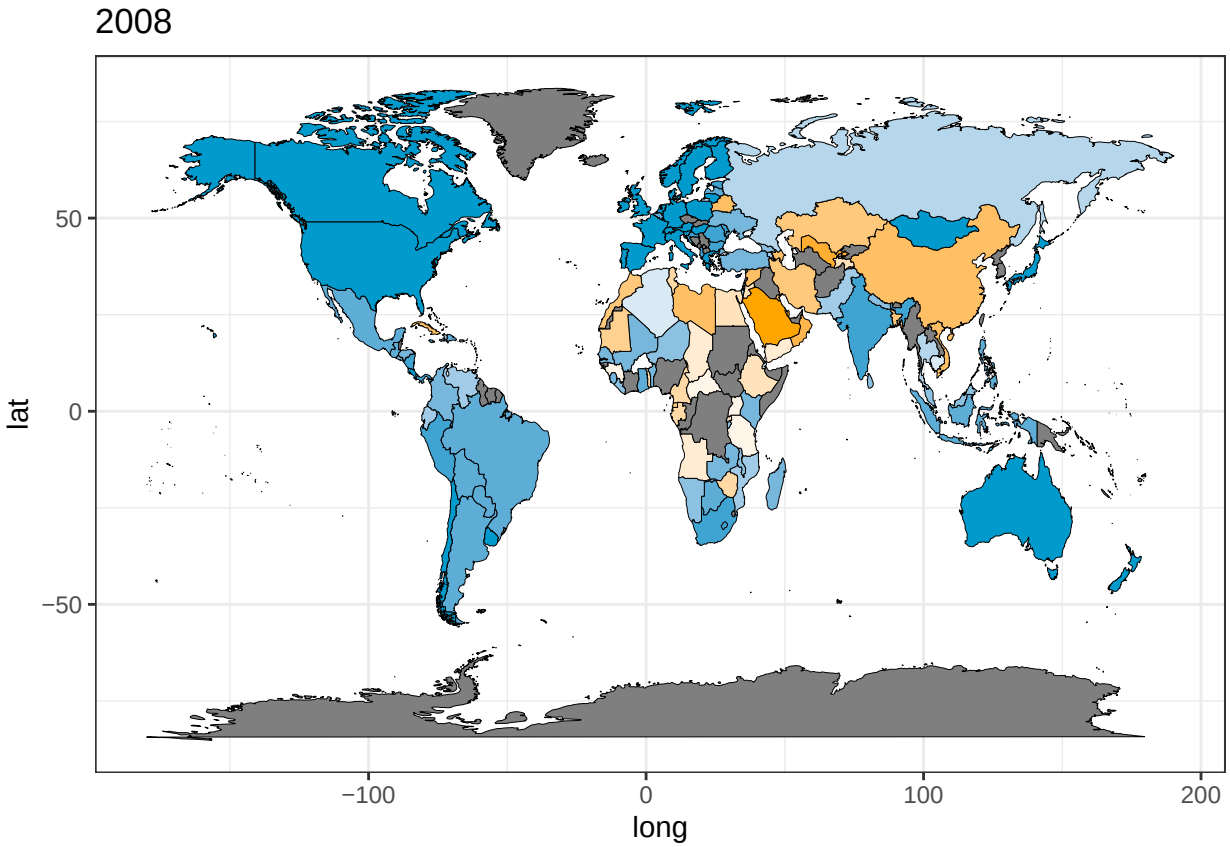

1982
50
0
−50
lat
−100
0
100
200
long

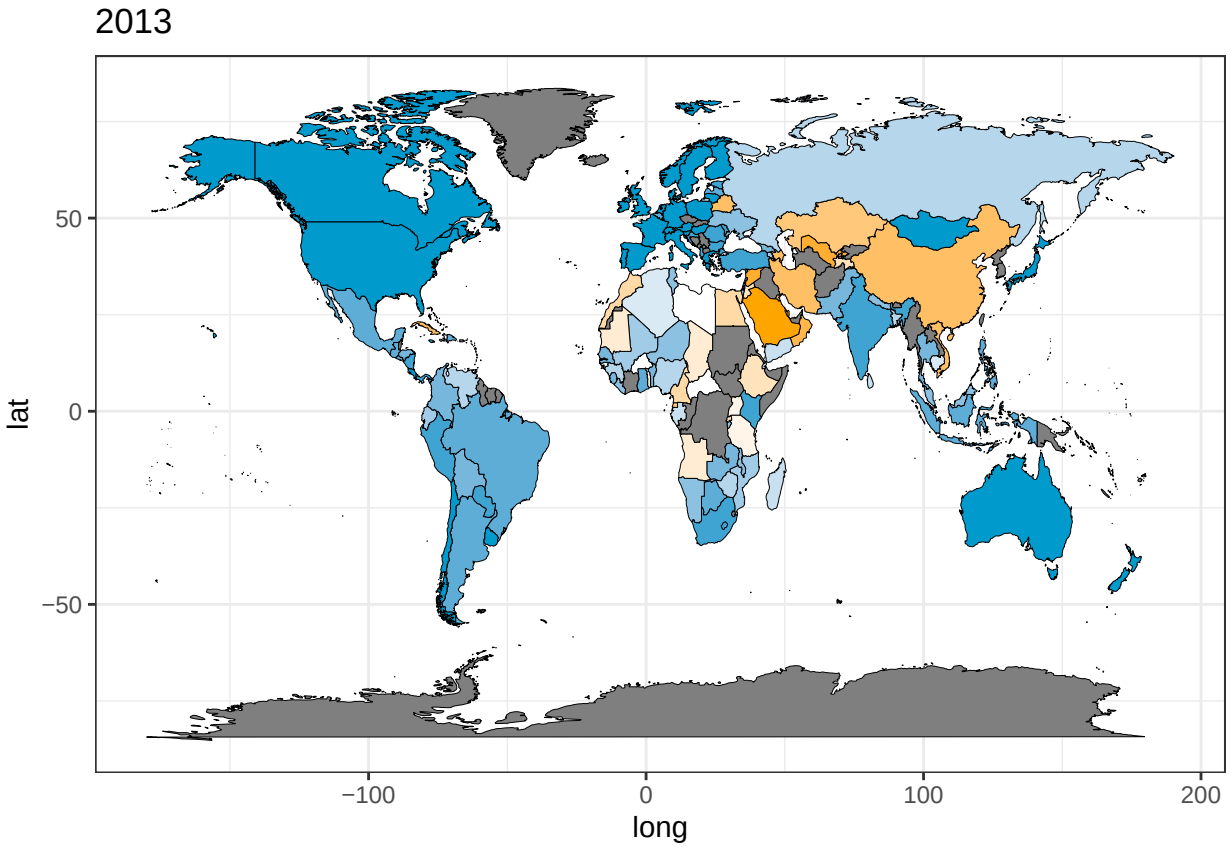

1986
50
0
−50
lat
−100
0
100
200
long


1991
50
0
−50
lat
−100
0
100
200
long

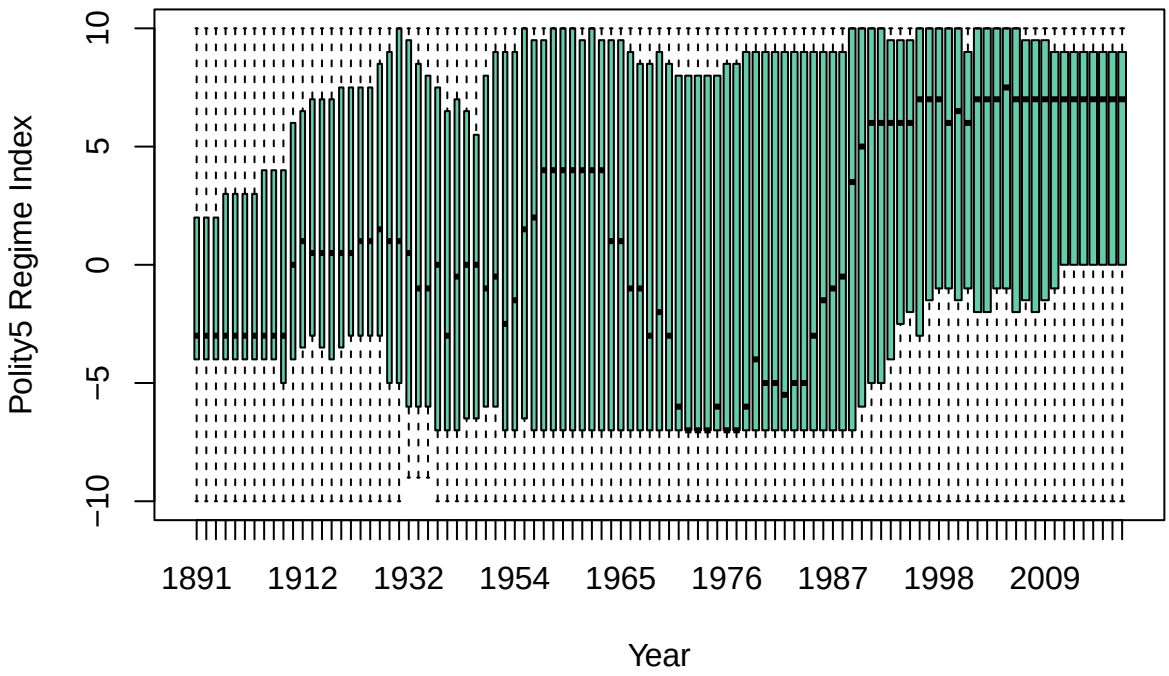

1996
50
0
−50
lat
−100
0
100
200
long


2003
50
0
−50
lat
−100
0
100
200
long


2008
50
0
−50
lat
−100
0
100
200
long



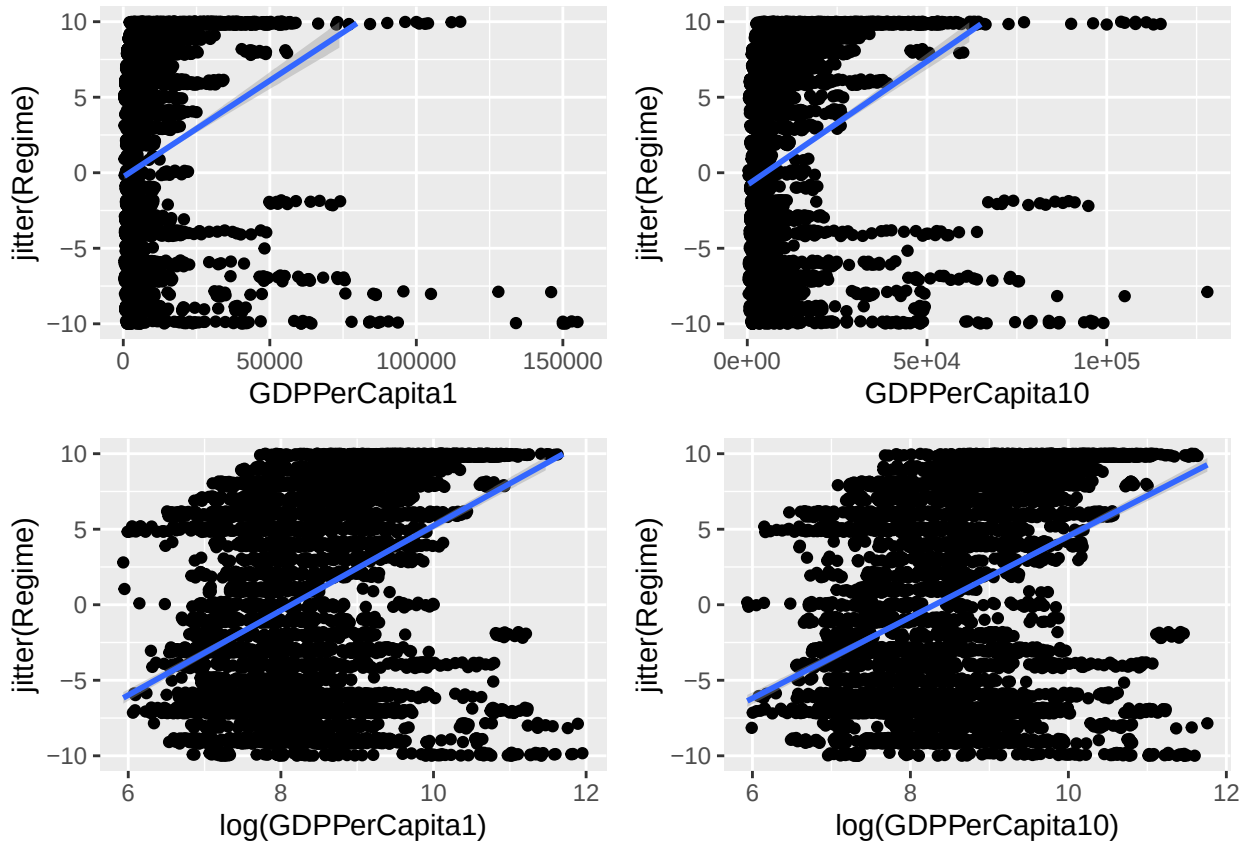
As one can see, the sampling is relatively sparse for the first several years represented, with almost all of the African continent entirely absent. As time progresses, however, more countries are represented. Moreover, large swaths of Europe, Asia, and North and South America are present throughout. While this should raise bias concerns about using "Year" as a prediction variable–it is not used in the model–it nevertheless appears reliable to gain insight about larger trends in the data over time. Below, I have done just that.



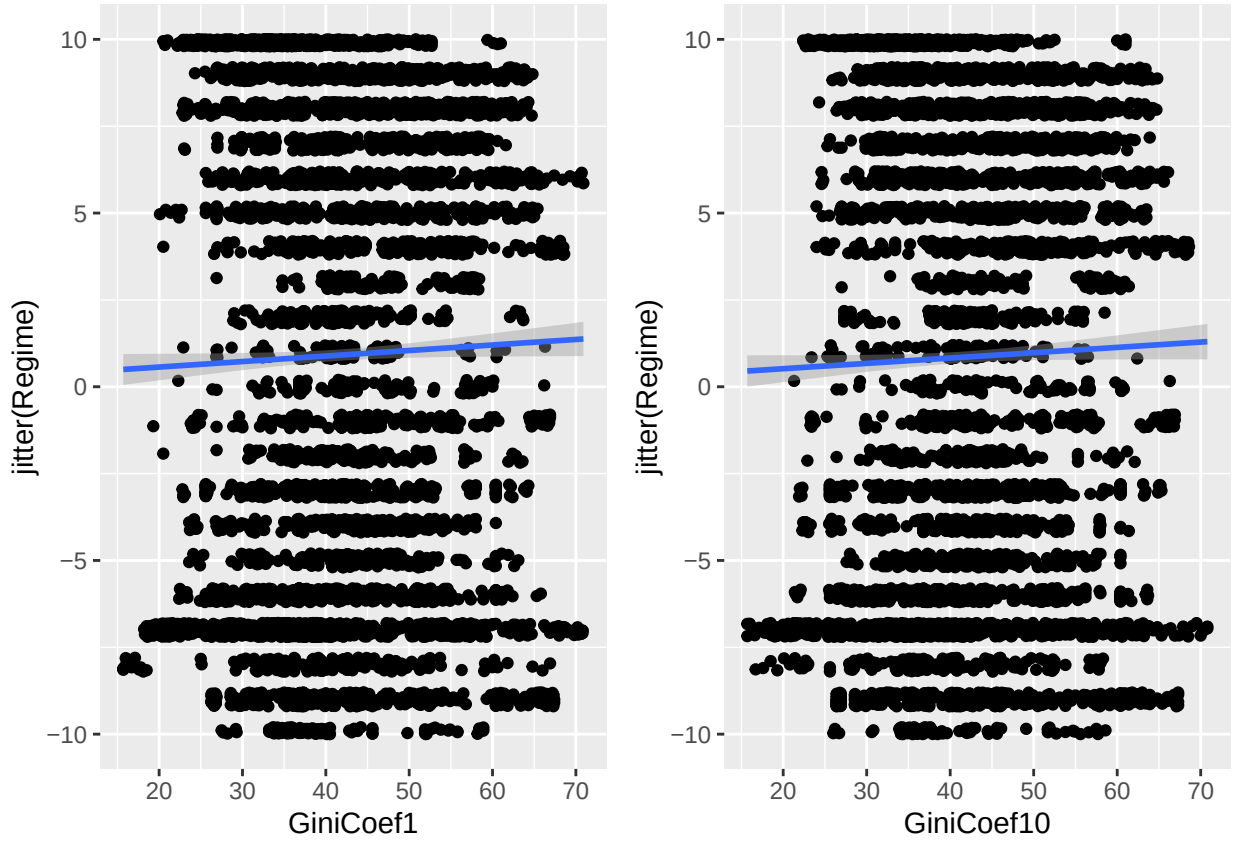
This graph could effectively be called a visualization of "Fukuyama's hypothesis." Indeed, at least based on this subjective regime assessment index, he appears to be correct in asserting that democracy is ascendant and the world is nearing an equilibrium. In this series of boxplots, the median reaches a sort of equilibrium in 1993, with the median firmly in the region of democracy. Moreover, about 75% of the data is higher than a 0, and thus also to some extent "democratic." However, the large valley between 1960 and 1993 should be noted, in spite of the fact that there is less data in this time period (represented by the thinner boxes). This dip and rebound within a 40 year span indicates that massive changes can occur within decades, and thus any hypothesis should take into account large amounts of data from large time spans.

The following plots examine the relation of the possible features to the Regime target variables. For space constraints, only years 1 and 10 are plotted. In cases where the log values represent better predictors than the actual values, all four plots are shown.

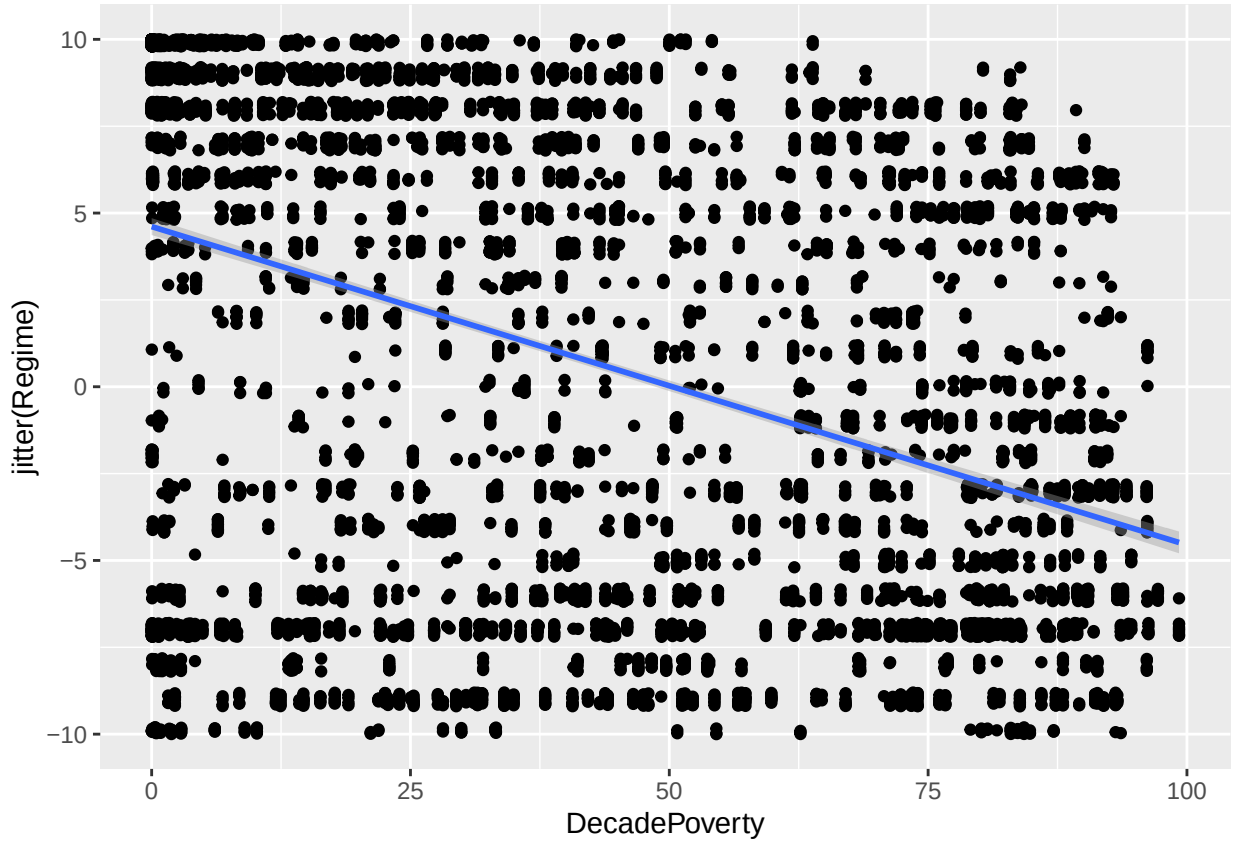
```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



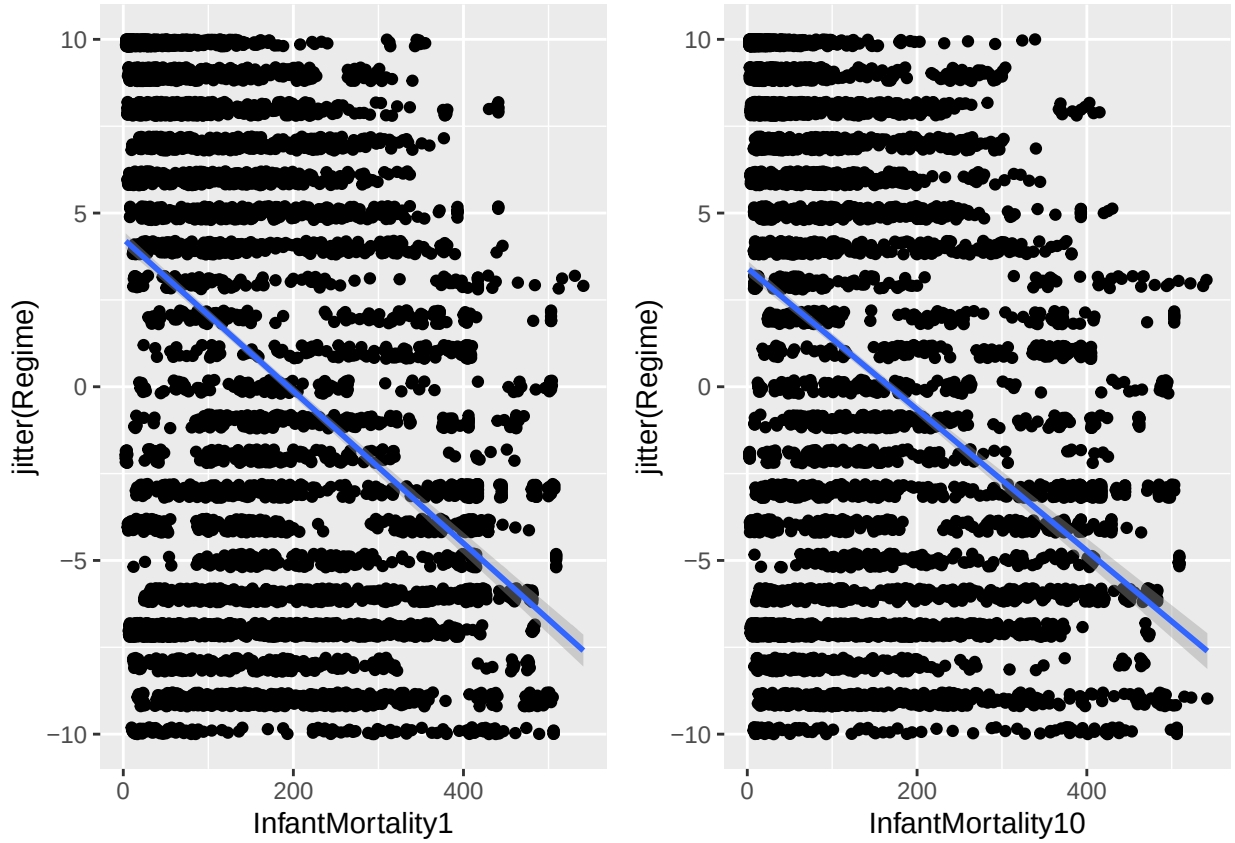

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



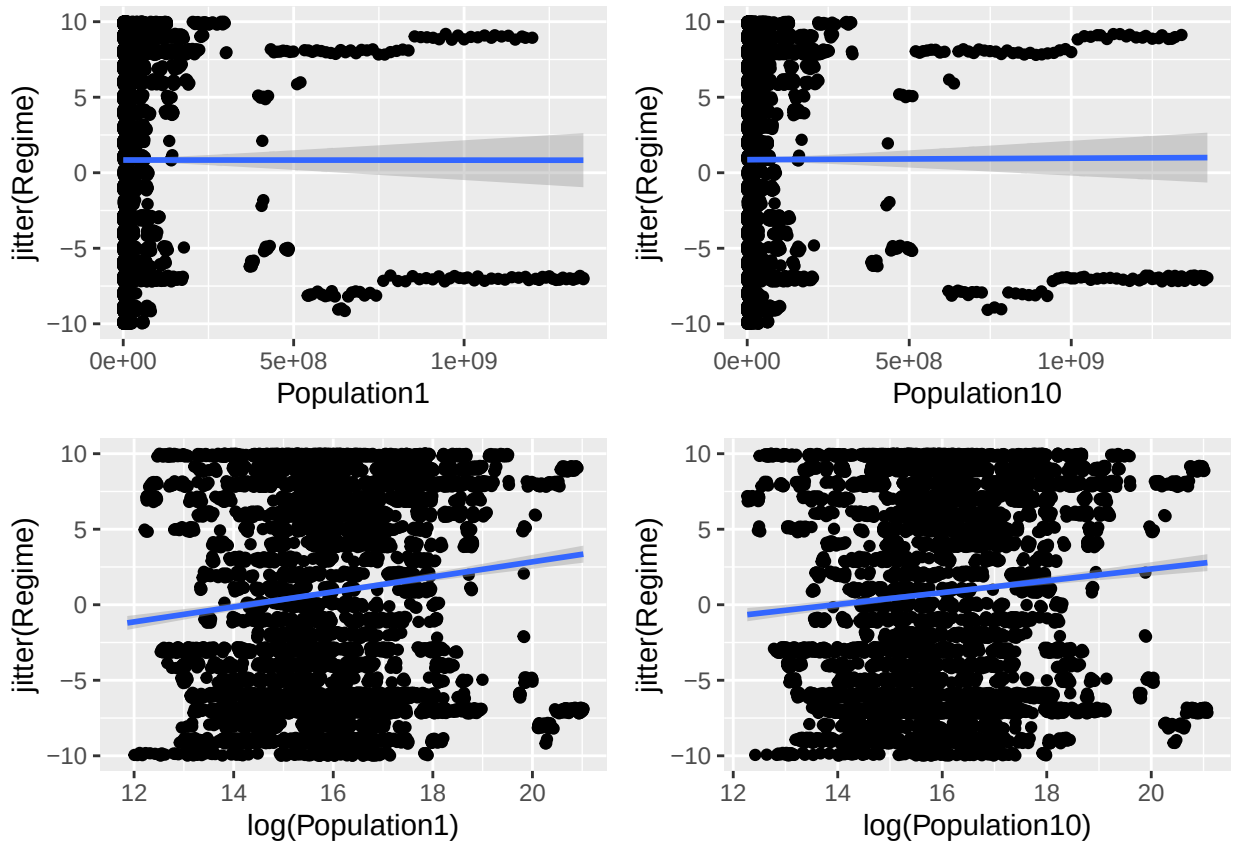

```
## `geom_smooth()` using formula 'y ~ x'
```



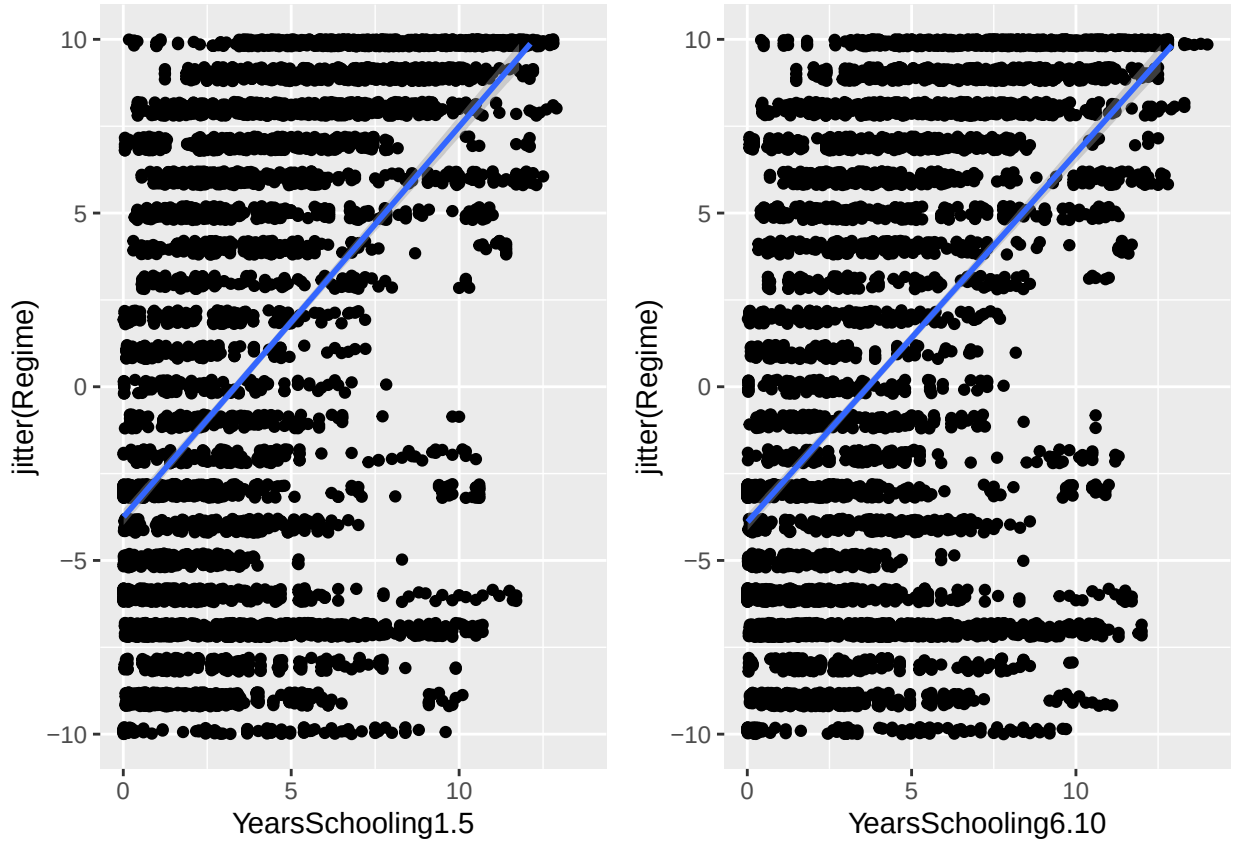

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```




```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```




```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



While it is not immediately apparent whether all of these variables can be correlated with the Regime response variable, several show at least a degree of linear relationship. Soem of these variables also may be colinear. Both of these will be examined in the model building process, below.

Section 2: Modeling/Analysis

For my model, I chose to use a linear regression model, rather than a multinomial logistic regression model. The latter may have been the more obvious choice for discrete response variables, but the former provides the advantage that, since these responses are numerically hierarchical, they have an order that should be taken into account. (I originally attempted to use numerous R packages that perform logistic regression on ordinal factor variables; however, R threw errors in numerous places, and after working through them for a long time to no avail, I decided to abandon this approach.)

In order to appraoch this problem in numerous ways, I used subsets of the predictive data that seemed reasonable; the years of predictors from the data that are used in these subsets appear below. Within each of these subsets, I used lasso and cross validation in order to whittle down the number of predictors within each of these subsets, and remove any colinearity (since lasso forces colinear predictors to zero). The output below represents the subset of years, followed by the predictors chosen by the lasso and cross validation, and finally the R^2 value of the model that uses those predictors.

```
## [1] "-----------Years 1, 10-----------"
##  [1] "DecadePoverty"      "GDPPerCapita1"      "GiniCoef1"
##  [4] "GiniCoef10"         "InfantMortality1"   "InfantMortality10"
##  [7] "LifeSpan1"          "LifeSpan10"         "Population1"
## [10] "Population10"       "YearsSchooling1.5"  "YearsSchooling6.10"
```

```
## [1] "R-squared: 0.843288331619858"
## [1] "-----------Years 9, 10-----------"
## [1] "DecadePoverty"      "GDPPerCapita10"     "GDPPerCapita9"
## [4] "GiniCoef9"          "Population9"        "YearsSchooling6.10"
## [1] "R-squared: 0.834747594550606"
## [1] "-----------Years 8, 10-----------"
## [1] "DecadePoverty"      "GDPPerCapita10"     "GDPPerCapita8"
## [4] "GiniCoef8"          "Population8"        "YearsSchooling6.10"
## [1] "R-squared: 0.834831276711371"
## [1] "-----------Years 5-10-----------"
## [1] "DecadePoverty"      "GDPPerCapita5"      "GiniCoef10"
## [4] "GiniCoef5"          "Population5"        "YearsSchooling6.10"
## [1] "R-squared: 0.834973317759709"
## [1] "-----------Years 1-10-----------"
##  [1] "DecadePoverty"      "GDPPerCapita1"      "GiniCoef1"
##  [4] "GiniCoef10"         "GiniCoef5"          "GiniCoef6"
##  [7] "GiniCoef7"          "GiniCoef9"          "InfantMortality1"
## [10] "InfantMortality10"  "LifeSpan10"         "LifeSpan8"
## [13] "LifeSpan9"          "Population1"        "Population10"
## [16] "Population4"        "Population5"        "Population6"
## [19] "Population7"        "Population8"        "Population9"
## [22] "YearsSchooling1.5"  "YearsSchooling6.10"
## [1] "R-squared: 0.843063273452441"
## [1] "-----------Years 1-5-----------"
##  [1] "DecadePoverty"     "GDPPerCapita1"     "GiniCoef1"
##  [4] "GiniCoef2"         "GiniCoef3"         "GiniCoef4"
##  [7] "GiniCoef5"         "InfantMortality1"  "LifeSpan5"
## [10] "Population1"       "YearsSchooling1.5"
## [1] "R-squared: 0.837078933815094"
```

The two sets of predictors with the highest R^2 value are the first one, that uses years 1 and 10, and the second to last, which uses years 1 *to* 10. Because of how close their values are, it is worth going with the first, for the sake of model simplicity and interpretability. Examining the coefficients of this model (coefficients marked with a '.' are 0), we find:

```
## 14 x 1 sparse Matrix of class "dgCMatrix"
##                               s0
## (Intercept)         .
## DecadePoverty      -0.016002429
## GDPPerCapita1       0.705534980
## GDPPerCapita10      .
## GiniCoef1           0.047931760
## GiniCoef10          0.062487982
## InfantMortality1   -0.009768679
## InfantMortality10   0.003434244
## LifeSpan1          -0.006112891
## LifeSpan10         -0.028662206
## Population1         4.230515465
## Population10       -4.066242290
## YearsSchooling1.5   1.368990550
## YearsSchooling6.10 -0.467806272
```

On first glance, several of these variables appear either useless or redunant. DecadePoverty appears too small to be significant; GDPPerCapita10 is intuitively more worthwhile than GDPPerCapita1; and GiniCoef,

InfantMorality, and LifeSpan appear to be redunant (especially if one looks at the plots above.) Removing these variables will leave us with a more intuitive model, giving the following:

```
##
## Call:
## lm(formula = Regime ~ . - 1, data = new.table)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.5097  -4.1127   0.8774   4.0439  14.8595
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## GDPPerCapita10       1.085794   0.092388  11.753  < 2e-16 ***
## GiniCoef10           0.131402   0.006784  19.370  < 2e-16 ***
## InfantMortality10   -0.009953   0.001194  -8.338  < 2e-16 ***
## LifeSpan10          -0.021118   0.012931  -1.633    0.102
## YearsSchooling1.5    1.721872   0.185801   9.267  < 2e-16 ***
## YearsSchooling6.10  -1.061485   0.184436  -5.755 8.97e-09 ***
## Population1         17.337089   0.656753  26.398  < 2e-16 ***
## Population10       -17.308626   0.660822 -26.193  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.672 on 7941 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8495
## F-statistic:  5610 on 8 and 7941 DF,  p-value: < 2.2e-16
```

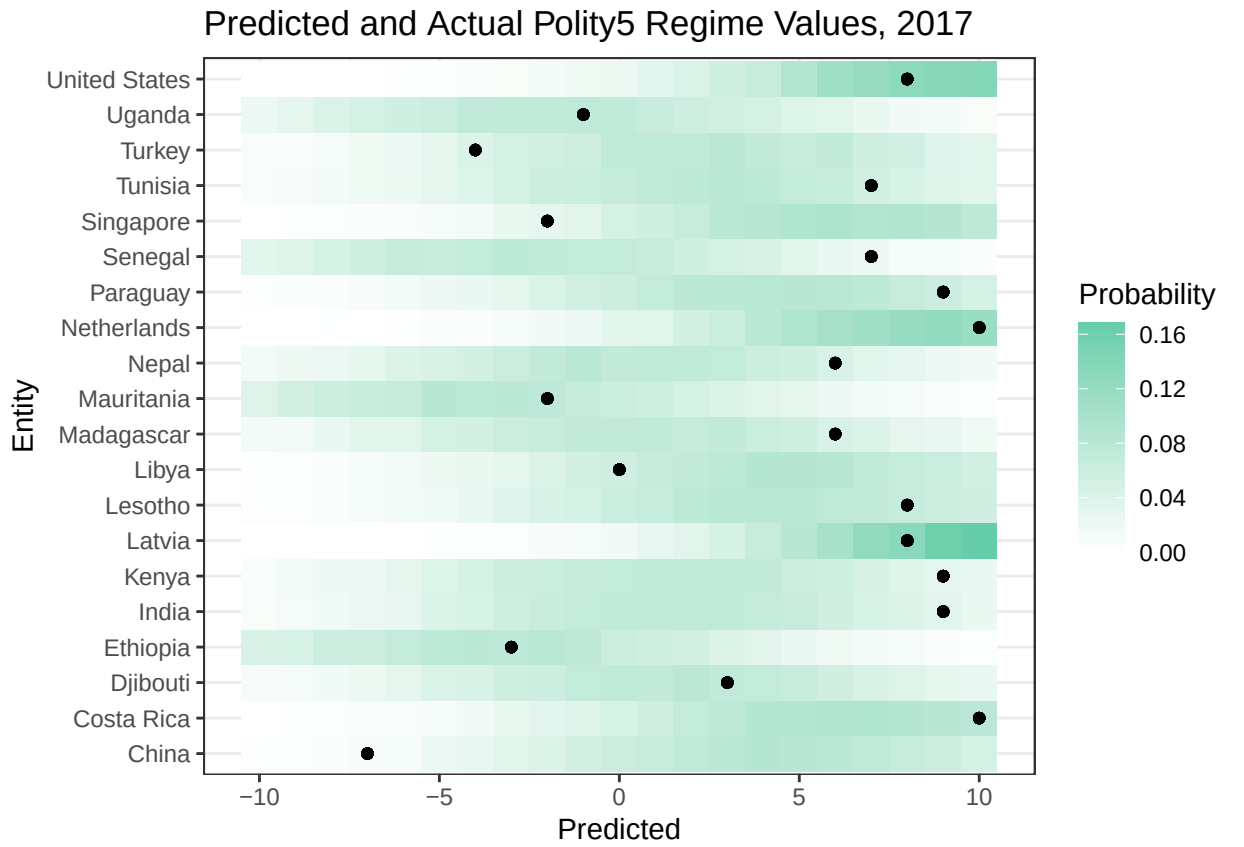
This model actually has a higher R^2 than the previously observed one. It also also more intuitive, as the 10th year of data is preferred in cases when the first year of data is insignificant, since it is closer to the event at hand and therefore more likely to impact the results. Finally, the coefficients themselves are more interpretable. As in the graphs in Section 1, GDP per capita is the most significant predictor, as it the most linearly related; the next three predictors have significantly smaller, but somewhat noticeable, impact. Finally, the years of schooling and population variables from both ranges are included, representing that the variables themselves are less important than the change that occurs between them over time.

However, the R^2 is only a measure of how well these variables explain the variability in the data, not a measure of how well they can predict this data. In order to test this, I evaluate the model by separating it into a training (all years before 2017) and test (data from the year 2017) set, and measure the RMSE, which comes out to:

```
## [1] 5.942356
```

Section 3: Visualization and interpretation of the results

It is easy enough to see from the RMSE value that this model is not good at predicting the actual outcome to a reasonable degree. Since there are only 21 possible categories for the Regime response variable (-10 to 10), and RMSE of approximately 6 is not sufficient to be a "good" prediction range, as it is too broad. To visualize this, we may create a visualization that displays a selection of the predicted values from the year 2017 based on our above model against their actual values.



In this visualization, the black points represent the countries' actual regime assessments, while the sliding color scale represents the range of possible values from the model and their associated probabilities. This has been done by simulating 10,000 model results. As one can see, the points frequently fall within the "green zone," indicating that they are within the realm of possibility according to the model. However, they sometimes fall in the lighter zones as well, reflecting the instability of the model as described earlier. In other words, the conclusion to be drawn is that these predictors are insufficient for creating a reliable model on the basis of linear regression.

Section 4: Conclusions and recommendations

As stated in the previous section, due to the high RMSE of the model, one cannot draw reliable conclusions from this. As such, I propose two avenues for further work and improvements. First, the data set that I have used is clearly insufficient to "predict" regime changes–this is true both quantitatively, as this study has shown, but also intuitively, as sudden regime changes, such as those resulting from coups, are inherently more difficult to predict using this data. However, this is not to say that such data does not exist–only that, due to lack of access, I have been unable to incorporate it into my model. Secondly, as I noted from the outset, I built a linear model, rather than an ordinal logistic regression model; the latter model might prove more effective than the one I chose, and would be possible to create with more time.

Finally, it is worth noting the inherent problem with the target data, that may make any study such as this difficult–namely, that it is more subjective than the predictive data. While the Polity5 regime assessments are relatively well-regarded, performing quantitative analysis with it thus poses problems to any researcher.