Climate Change Markers and Refugee Flow

HARP: Hadeel Hamoud, Alexis Bianco, Ryan Lee, Pierce Hollier

November 20, 2020

INTRODUCTION AND DATA

As the world struggles to convince so many people about the urgency of climate change it needs to be known that the threat we face is not just that of rising sea levels or CO2 levels—it is that of losing our homes. It is that of entire cities having to uproot and move elsewhere because they can no longer sustain themselves. Far from just a small increase in temperature, but a disruption of our lives as we know it.

Each year, tens of millions of people are driven from their homes by floods, storms, and droughts. The Ecological Threat Register, conducted by The Sydney-based Institute for Economics and Peace (IEP), measures ecological threats over 157 independent states and territories. The report projects that as many as 1.2 billion people around the world could be displaced by 2050 (Institute for Economics and Peace, 2020). The adverse effects of global climate change will induce more extreme weather, growing food and water insecurity, and rising sea levels which will cause the number of displaced people to rise (UNHCR, 2019). The IEP report additionally identifies three clusters of ecological hotspots: the Sahel-Horn belt of Africa, from Mauritania to Somalia; the Southern African belt, from Angola to Madagascar, and the Middle East and Central Asian belt, from Syria to Pakistan.

The intersection of climate change and migration requires comprehensive data analysis and solutions to the multidimensional challenges it creates (Podesta, 2019). Therefore, analyzing the dynamics between climate change predictors and refugee flow from a country can reveal opportunities for interventions. Our primary goal in this project is to understand if and how climate change indicators correlate with refugee flow. In order to focus our analysis and delineate a more specific model, we will focus on relationship between climate change indicators and refugee data in the Middle East and Central Asian belt region, an at-risk region identified by The Ecological Threat Register.

Our dataset comes from the World Bank Development Indicators Databank. According to the World Bank, most of the data comes directly from each country in the World Bank Group's national statistical systems. The raw data itself contains many development indicators, and the series name tells us the metric or variable for which we are getting data. Within each series, the data is broken down into the data for each nation for each year between 1960-2019. The dataset additionally contains all the markers that the WDB has tracked in association with climate change in almost every country on Earth. This includes variables such as CO2 emissions levels of every country and agricultural output of each country. Unfortunately, the WDB does not have very complete data for some of the variables. However, we will focus on variables that have sufficient data, unless the variable is unlikely to change much over time.

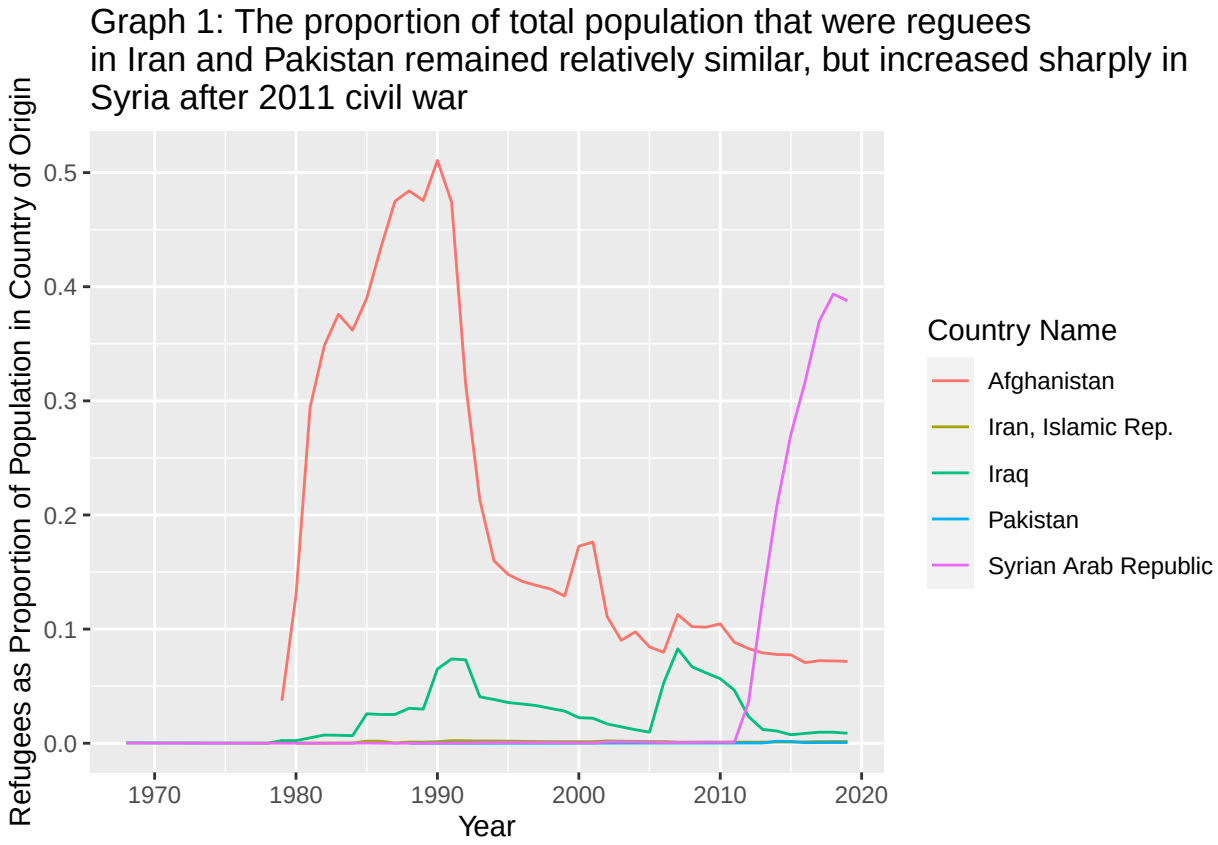
We will start by examining variables in the WDB data that scientific literature identifies as climate change predictors such as CO2 emissions (as measured in tons per capita), other greenhouse emissions (in kilotons), land under cereal production, and percentage of arable land. We will also examine refugee flow out of the countries relative to each country's population. Then, we will analyze the correlation between refugee flow and climate predictors. The predictors will be used to build and evaluate a linear model that attempts to demonstrate if there is a relationship between predictors and refugee flow as well as explain some of the variance in said refugee flow. While refugee flow from a country can be influenced by an almost innumerable amount of variables, the hope for the project is not predict all of that variance. We hypothesize that at least one of the climate predictors (CO2 emissions in tons per capita, other greenhouse emissions in kilotons,

and land under cereal production) in our linear model will have a statistically significant relationship with refugee flow from the countries in the Middle East and Central Asian Belt. We will test this hypothesis using a multiple linear regression model and associated tests such as a mutlicollinearity test.

METHODOLOGY

Visualizations and Exploratory Data Analysis

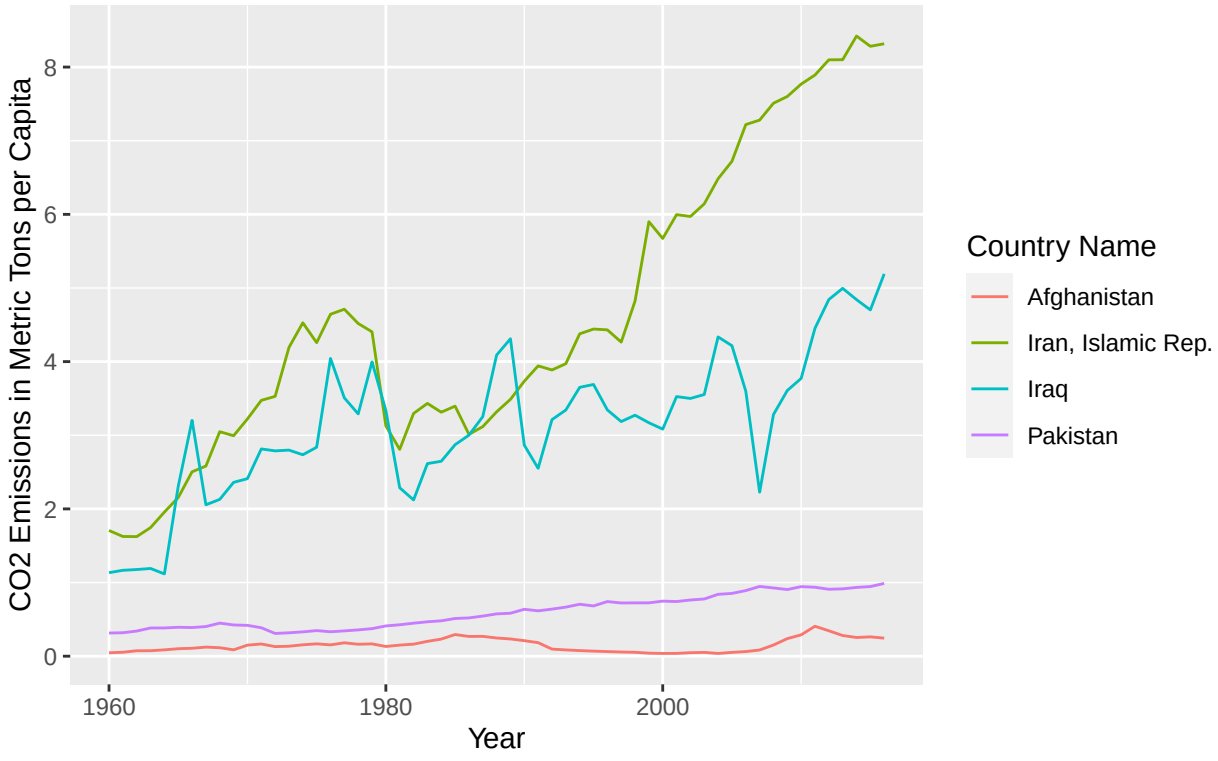
To visualize climate change predictors within each Middle Eastern country over time, we plotted the yearly proportion of population leaving as refugees, CO2 emissions, N2O emissions, methane emissions, percent of country's land that was arable, and hectares of arable land used for cereal cultivation over time.



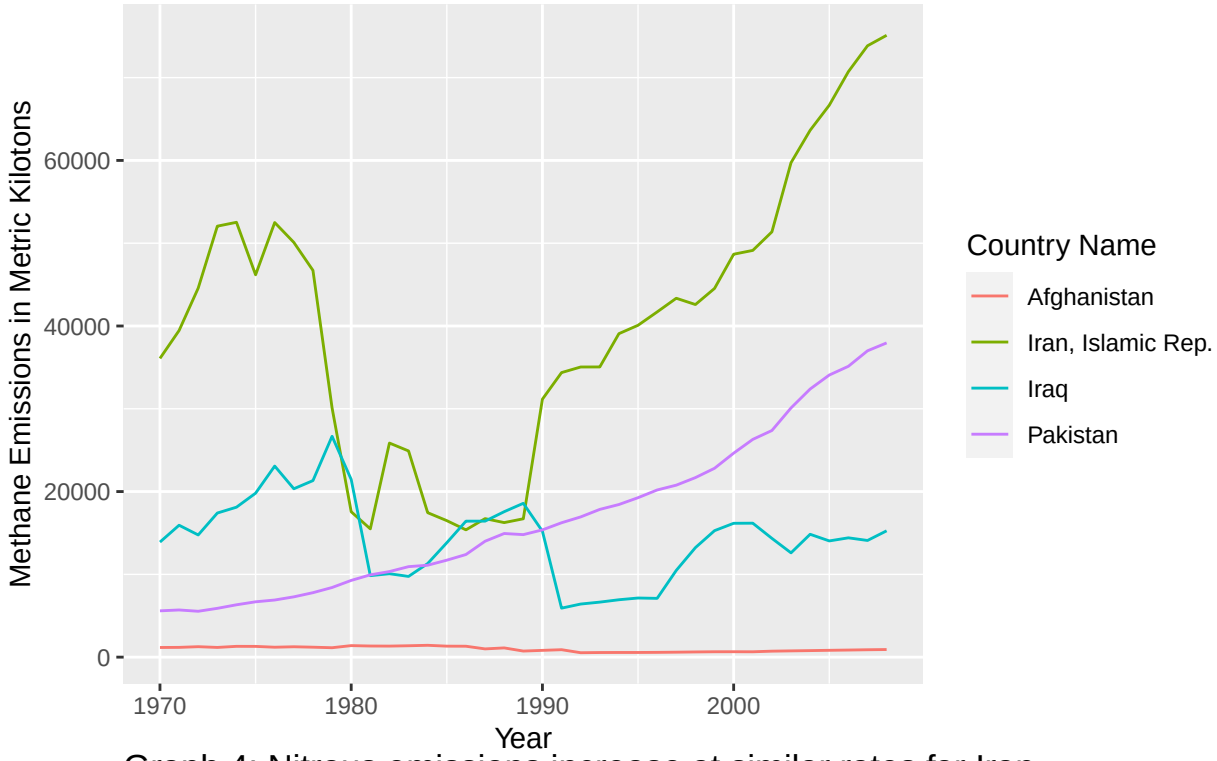
The Middle East and Central Asian belt region, as generally defined, includes 5 countries, Syria, Iran, Iraq, Afghanistan, and Pakistan. All of these were included in the modeling for our analysis except for Syria. Syria had barely any refugees leaving their country until the start of the Syrian Civil War in 2011. It is clear from this exploratory data analysis, specifically the above line graph of refugee flow (as proportion of overall population) over time, that there is a massive spike in 2011. Therefore, that massive increase in refugee population is likely due to conflict. Thus, data from Syria was excluded. One can also see significant spikes in refugee population fleeing Afghanistan. Despite this, Afghanistan was still included in the model. This is because the refugee population spikes do not correlate as well with periods of war/conflict. The number of refugees fleeing Afghanistan actually decreased dramatically during civil wars in the 1990's. While it is undeniable that Afghanistan did experience periods of civil wars and attacks by foreign powers, that alone is not enough to disqualify them. The refugee population fleeing Afghanistan has existed at significant levels regardless of period of conflict or not. This is also the case with Iraq. While the country experienced conflict during the period of study, the refugee trends do not correlate well with that conflict. In fact, the refugee population from Iraq was still slightly decreasing at the start of the American invasion. This was the justification for including Iraq in the data. More broadly, the mere existence of conflict does not justify

the exclusion of a country from our analysis. It is unlikely that climate factors alone would push someone to leave their country of residence. Moreover, the goal of this project is not to entirely explain the variance in refugee population with only these factors. That said, we wanted to observe how influential climate factors are when there are other factors (i.e. conflict) that would increase a person's desire to leave their country of residence. Thus countries that have had significant refugee populations both with and without conflict present are being included.

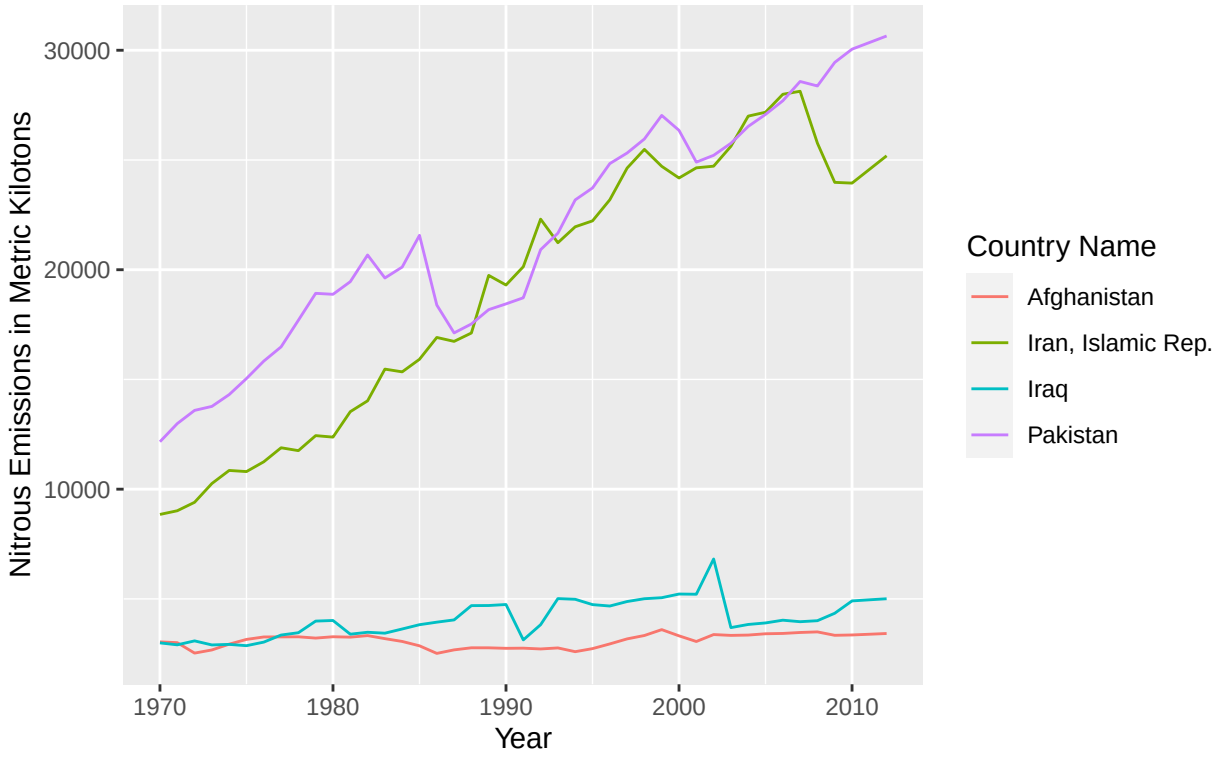
Graph 2: CO2 emissions in Iran and Iraq have higher relative increases over time compared to the other countries



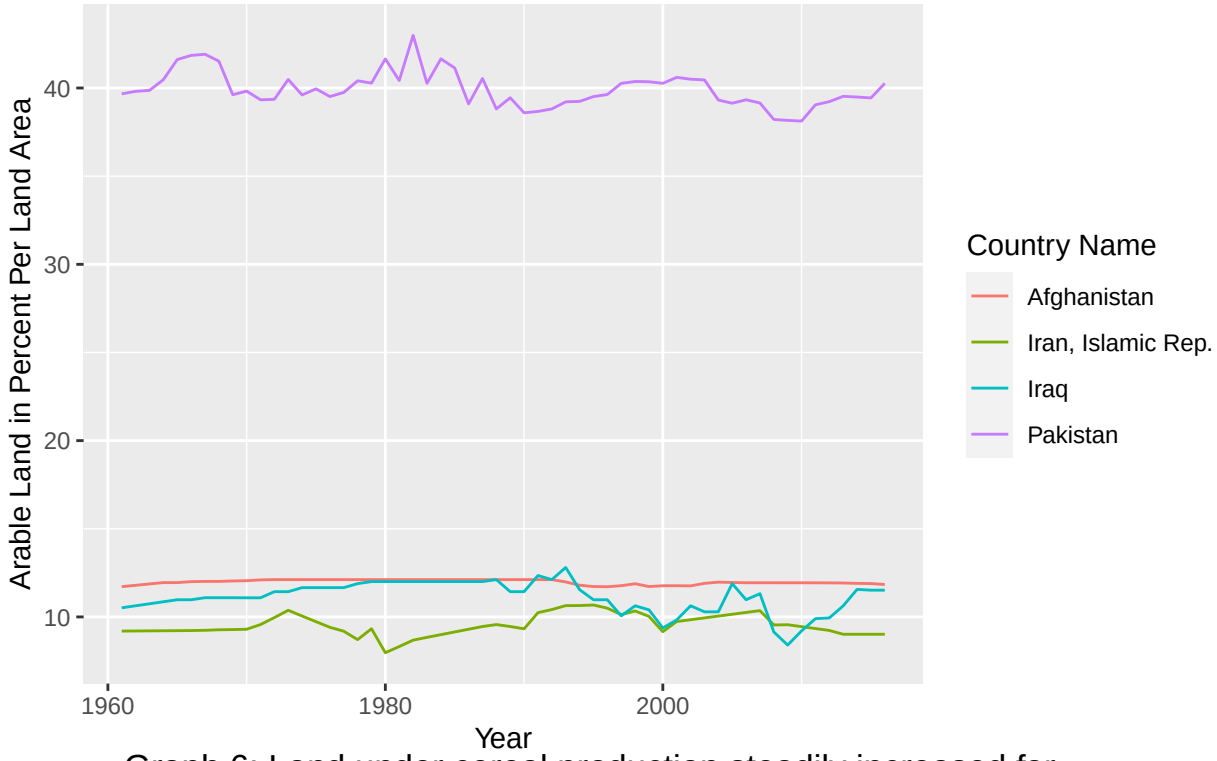
Graph 3: Methane emissions in Afghanistan remain relatively constant



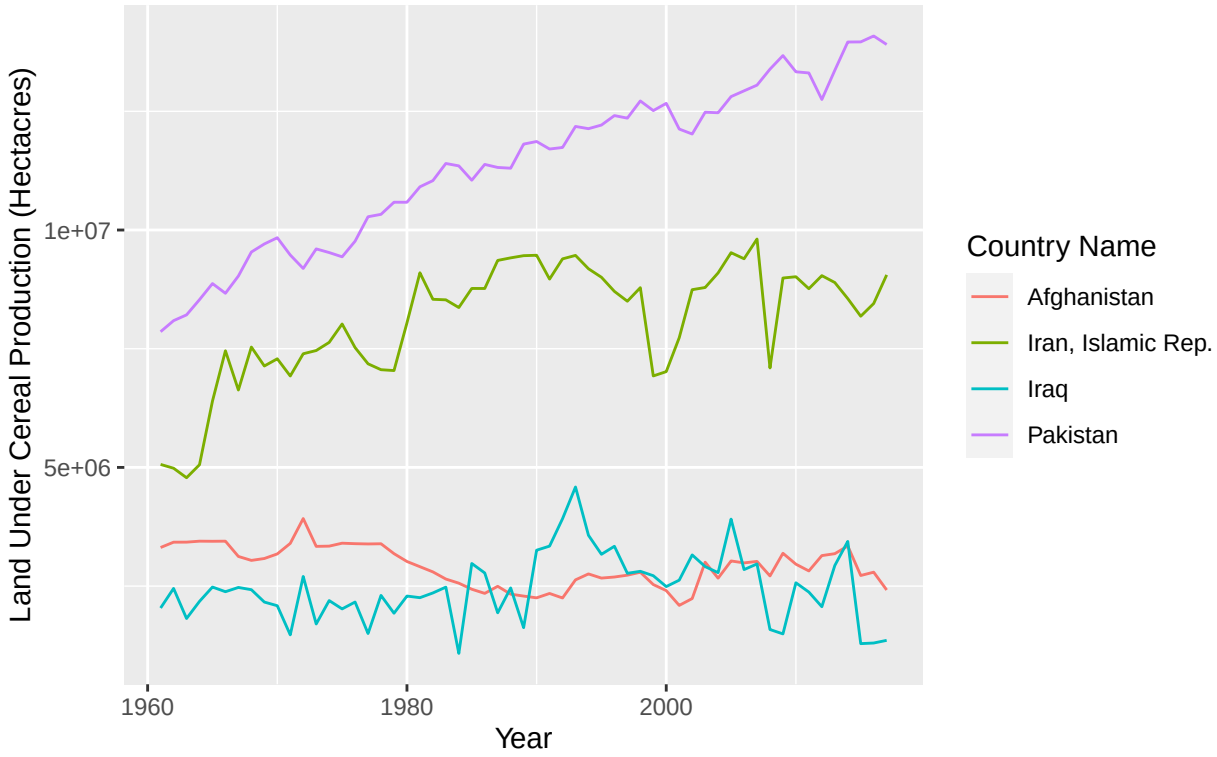
Graph 4: Nitrous emissions increase at similar rates for Iran and Pakistan



Graph 5: Percentage of arable land remained relatively constant for all four countries



Graph 6: Land under cereal production steadily increased for Pakistan and Iran



Graphs 2, 3, and 4 demonstrate that the trend for greenhouse gas emissions differs for each country. Carbon

dioxide, nitrous oxide, and methane emissions generally increase in Iran over time. Nitrous oxide emissions for Iraq and Afghanistan increased slightly and percentage of arable land remained relatively constant for all four countries. The percentage of land under cereal production (Graph 6) steadily increased for Pakistan and Iran while it fluctuated for Iran and Iraq and dipped slightly for Afghanistan.

The percentage of arable land remained relatively constant for all four countries. Thus, we will not use it as a predictor variable in our model as it won't have a significant relationship with the size in refugee population.

Main Effects Linear Regression Model with Log Transformation and Assessing Quality of Fit

In order to analyze our data, we will construct a multiple linear regression model with the amount of land under cereal production, methane and nitrous emissions (both measured in kilotons), and CO2 emissions (measuered in tons per capita) as predictor variables and refugees as a proportion of the overall population as the response variable. The model was constructed using data from Afghanistan, Iraq, Iran, and Pakistan. Constructing this model will allow us to investigate the existence and nature of any relationship between the climate change indicators and the refugee flow. It will also allow us to discern which climate change indicators have the most significant impact on refugee flow from a country. Since the distribution of the response variable, proportion of refugees, is skewed, we used a logarithmic transformation in order to more accurately construct a linear model.

From this information, we will now construct a confidence interval to provide insight into the validity of the above coefficients. Additionally, the values obtained will be exponentiated in order to determine the factor by which each variable affects refugee flow.

```
## # A tibble: 5 x 7
##   term            estimate   std.error statistic  p.value    conf.low  conf.high
##   <chr>              <dbl>       <dbl>     <dbl>    <dbl>       <dbl>      <dbl>
## 1 (Intercept)   0.700      0.389            1.80 7.42e- 2 -0.0698        1.47e+0
## 2 land_cereal  -0.00000158 0.000000163     -9.69 2.75e-16 -0.00000191   -1.26e-6
## 3 co2_tons     -0.372      0.150           -2.49 1.45e- 2 -0.669        -7.55e-2
## 4 nitrous_tons  0.000450   0.0000951        4.73 6.93e- 6  0.000261      6.39e-4
## 5 methane_tons -0.0000545  0.0000329       -1.66 1.00e- 1 -0.000120      1.07e-5

##
## Call:
## lm(formula = log(prop_refugee) ~ land_cereal + co2_tons + nitrous_tons +
##     methane_tons, data = climate_filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2600 -0.8001  0.2136  1.1190  2.7183
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.005e-01  3.885e-01   1.803   0.0742 .
## land_cereal  -1.585e-06  1.635e-07  -9.693 2.75e-16 ***
## co2_tons     -3.723e-01  1.497e-01  -2.486   0.0145 *
## nitrous_tons  4.500e-04  9.512e-05   4.731 6.93e-06 ***
## methane_tons -5.452e-05  3.289e-05  -1.658   0.1003
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.532 on 106 degrees of freedom
## Multiple R-squared:  0.7998, Adjusted R-squared:  0.7923
## F-statistic: 105.9 on 4 and 106 DF,  p-value: < 2.2e-16
```

```
## # A tibble: 5 x 2
##   term          estimate
##   <chr>            <dbl>
## 1 (Intercept)     2.01
## 2 land_cereal     1.00
## 3 co2_tons        0.689
## 4 nitrous_tons    1.00
## 5 methane_tons    1.00
```

The model that we constructed can be seen here:

Predicted proportion of refugees = e^(7.004603e-01 + -1.584535e-06 * land_cereal + -3.723390e-01 * co2_tons + 4.499857e-04 * nitrous_tons + -5.452494e-05 * methane_tons)

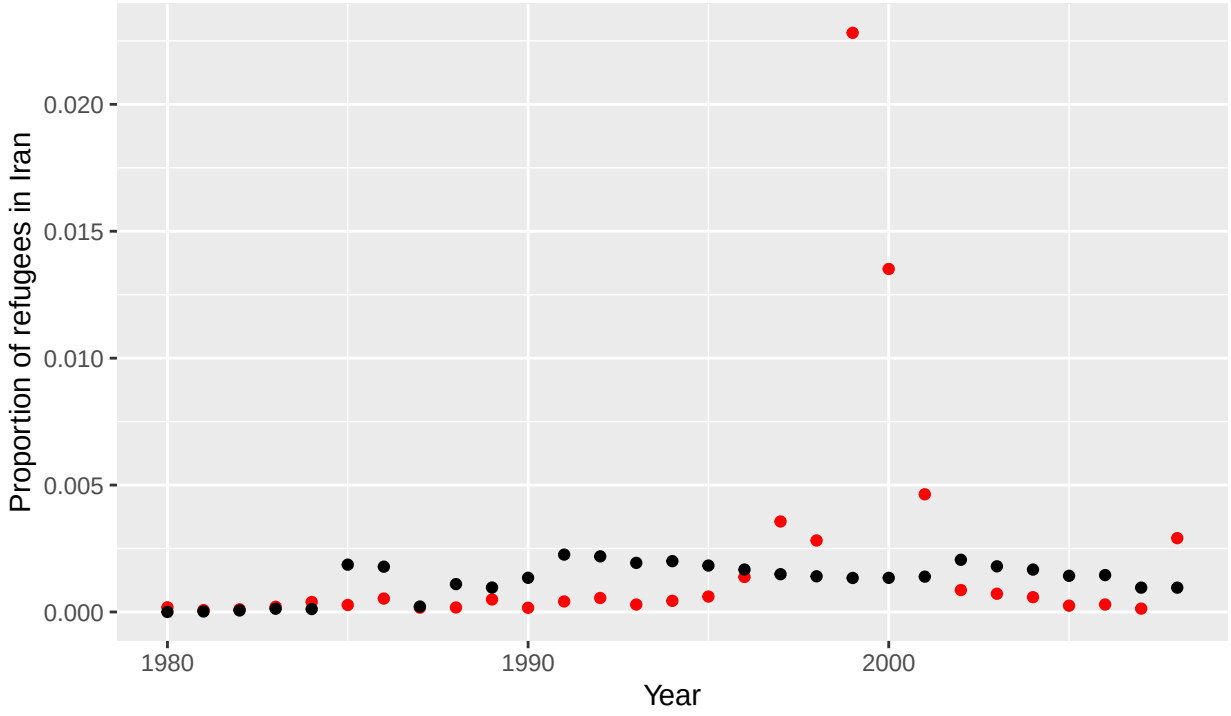
We obtained the following 95% of confidence intervals for the slope corresponding to each predictor variable conditional on all other predictors in our model:

Land under cereal production: (0.9999981, 0.9999987) CO2 emissions: (0.5121042, 0.9273253) N20 emissions: (1.0002614, 1.0006388) Methane emissions: (0.9998803, 1.0000107)

We performed a hypothesis test to determine the existence of statistically significant between a predictor (CO2 emissions in tons per capita) and refugee flow as a proportion of population. It is discussed in detail in the results section. CO2 emissions were chosen due to it having the largest impact on refugee flow in terms of its estimated β. Next we constructed visualizations for the predicted values given by our model as compared to the observed values for refugee flow from each country.

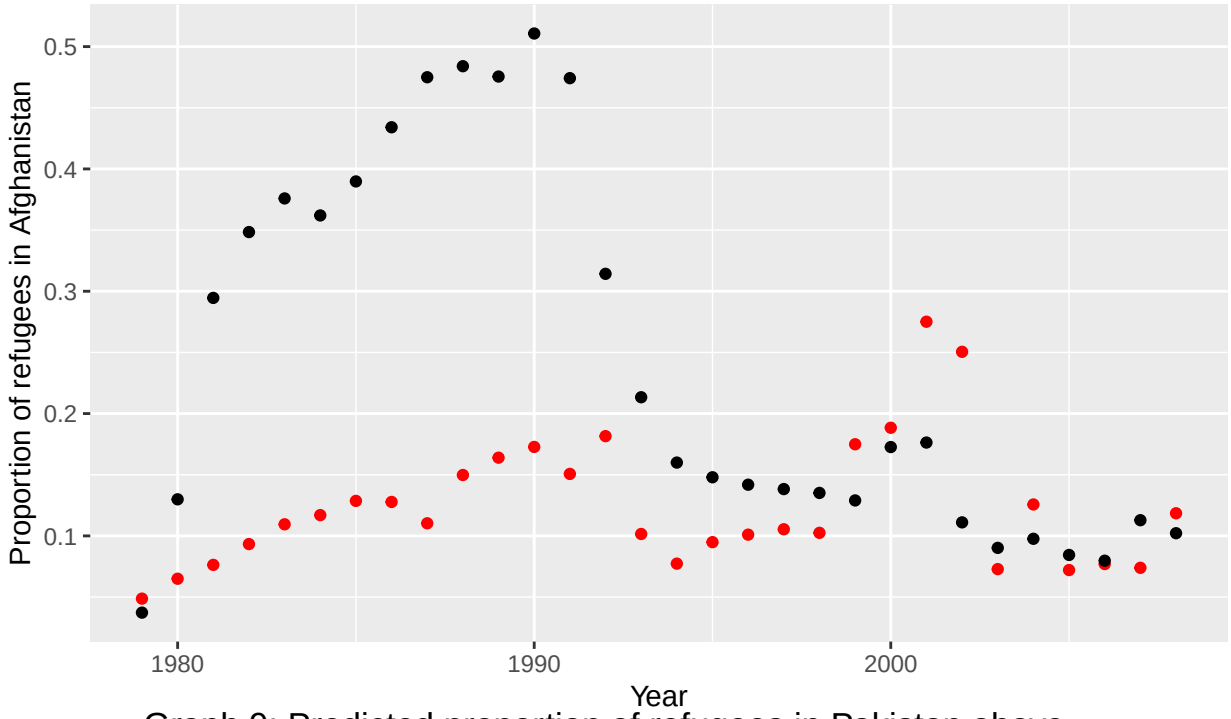
Graph 7: Predicted proportion of refugees in Iran relatively similar to observed proportion aside from 2 major outliers

Red points are predicted values and black points are observed values



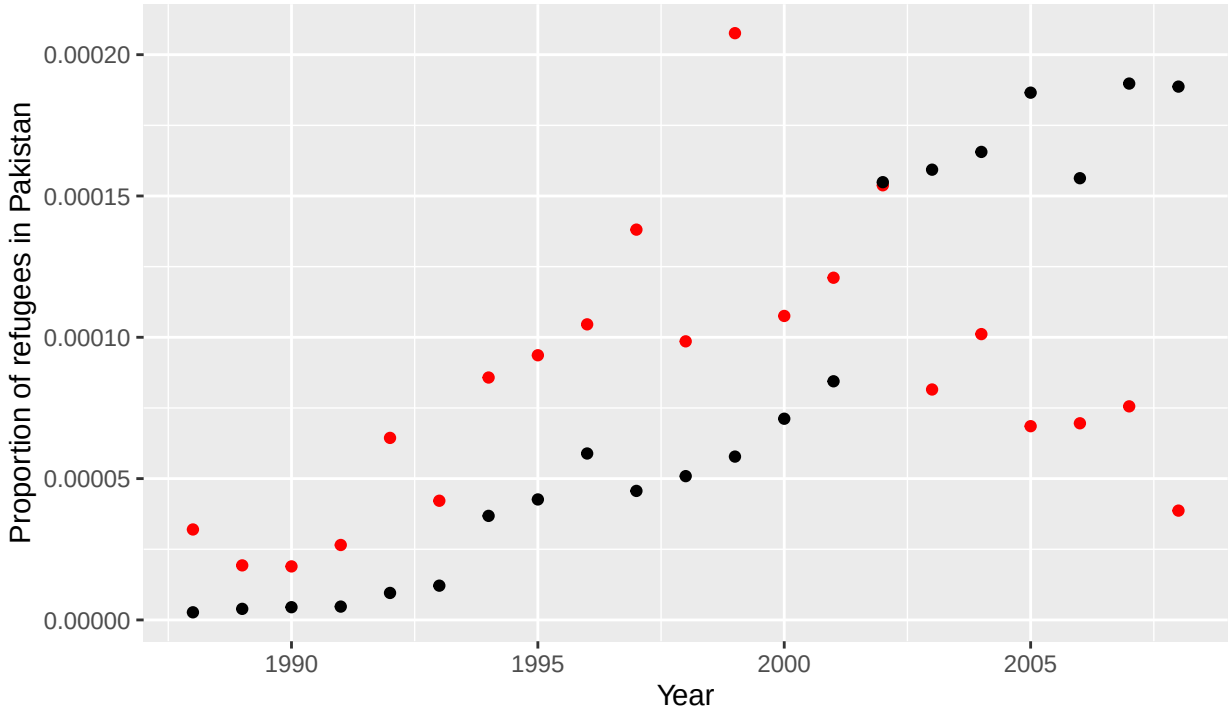
Graph 8: Predicted proportion of refugees in Afghanistan largely below observed proportion until 1995

Red points are predicted values and black points are observed values



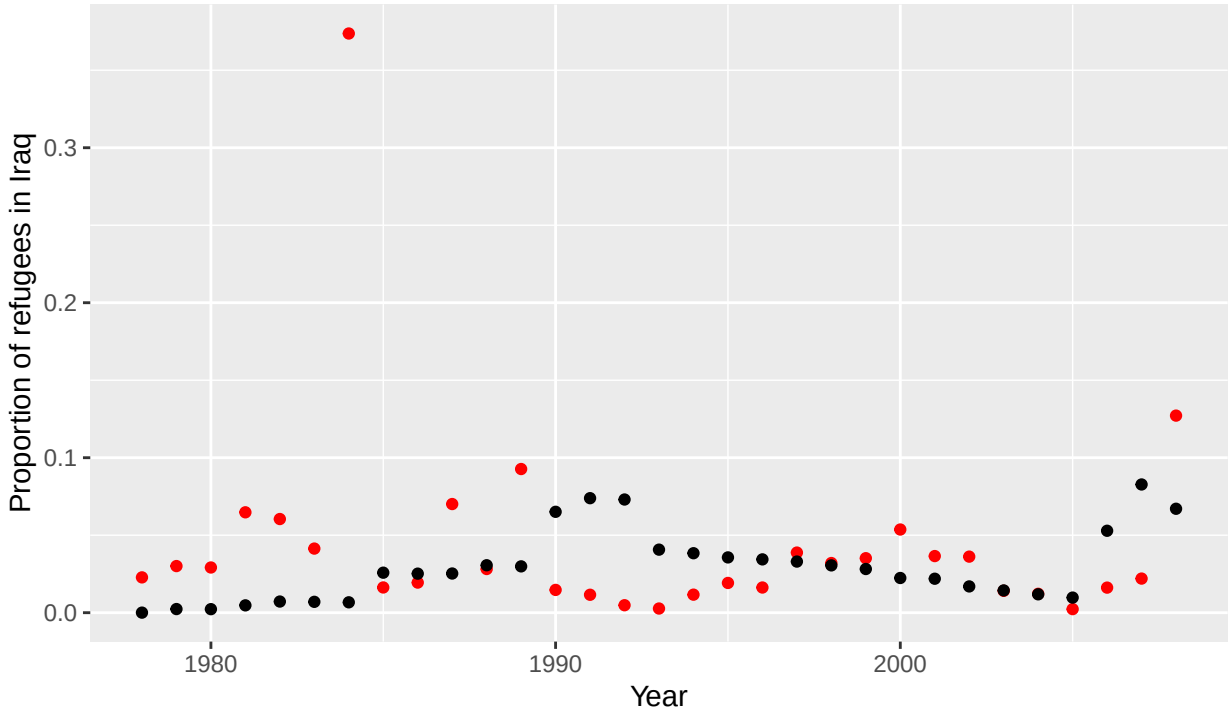
Graph 9: Predicted proportion of refugees in Pakistan above observed value until 2002

Red points are predicted values and black points are observed values



Graph 10: Predicted proportion of refugees in Iraq relatively similar to observed proportion aside from 1 major outliers

Red points are predicted values and black points are observed values

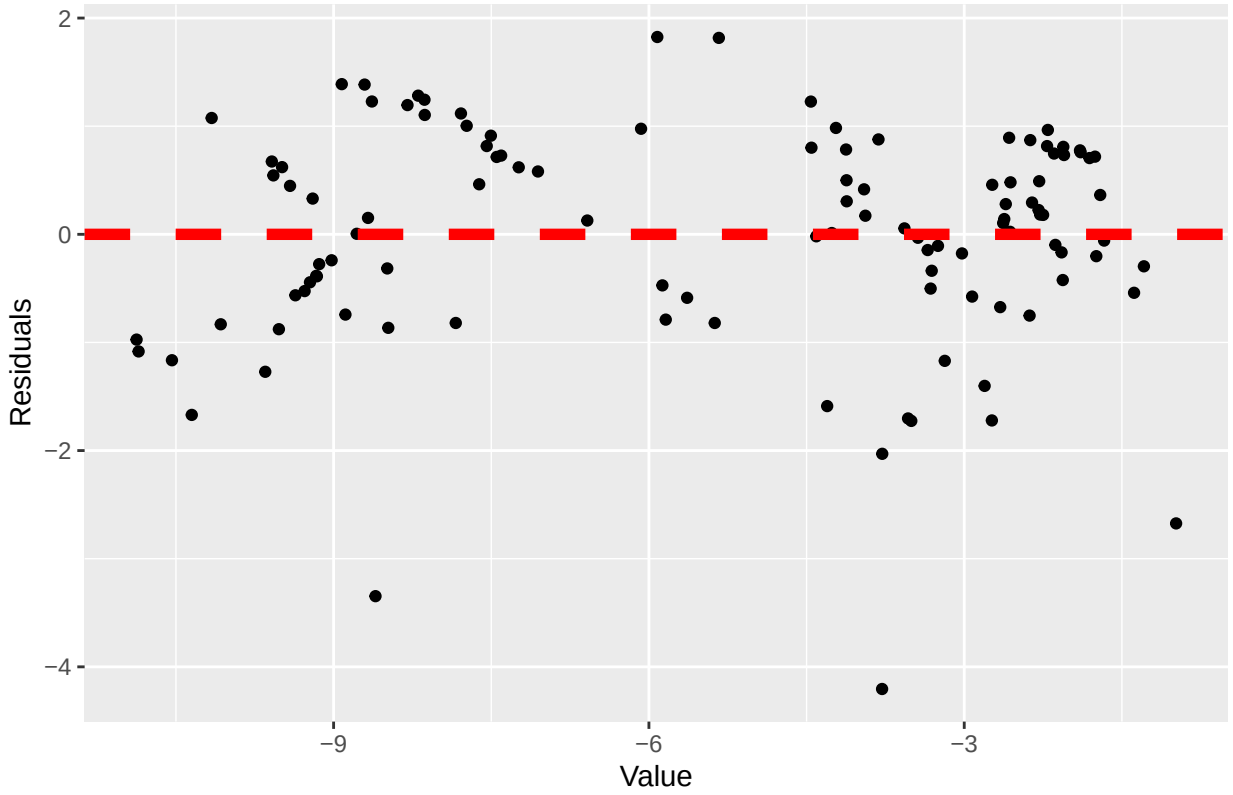


Diagnostic Plots for Main Effects Linear Regression Model with Log Transformation

Since the dataset includes many different predictor variables, it was first necessary to understand which would be the most powerful holding the others constant. In order to make inferences in regression, certain conditions must be met. The first two conditions that must be met are linearity which means the relationship between the variables and the predictors be linear and equal variance which means that residuals have relatively constant variance. These conditions were checked by creating a basic linear model and plotting the residuals such that we could examine their variance and linearity.

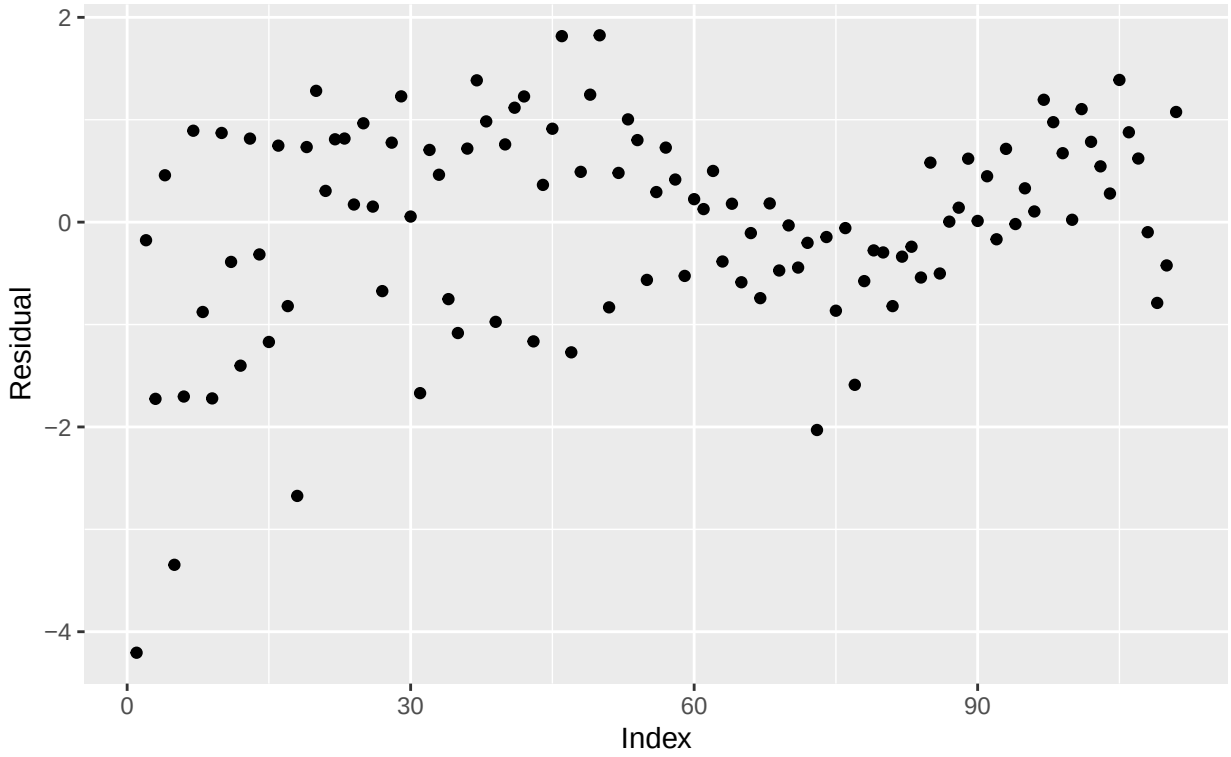
The first condition being tested for is residual variance to be 0. In order to verify whether this condition is met, residuals will be plotted alongside the line $y = 0$, and their spread will be evaluated both above and below the line in question.

Graph 11: Equal variance condition not met



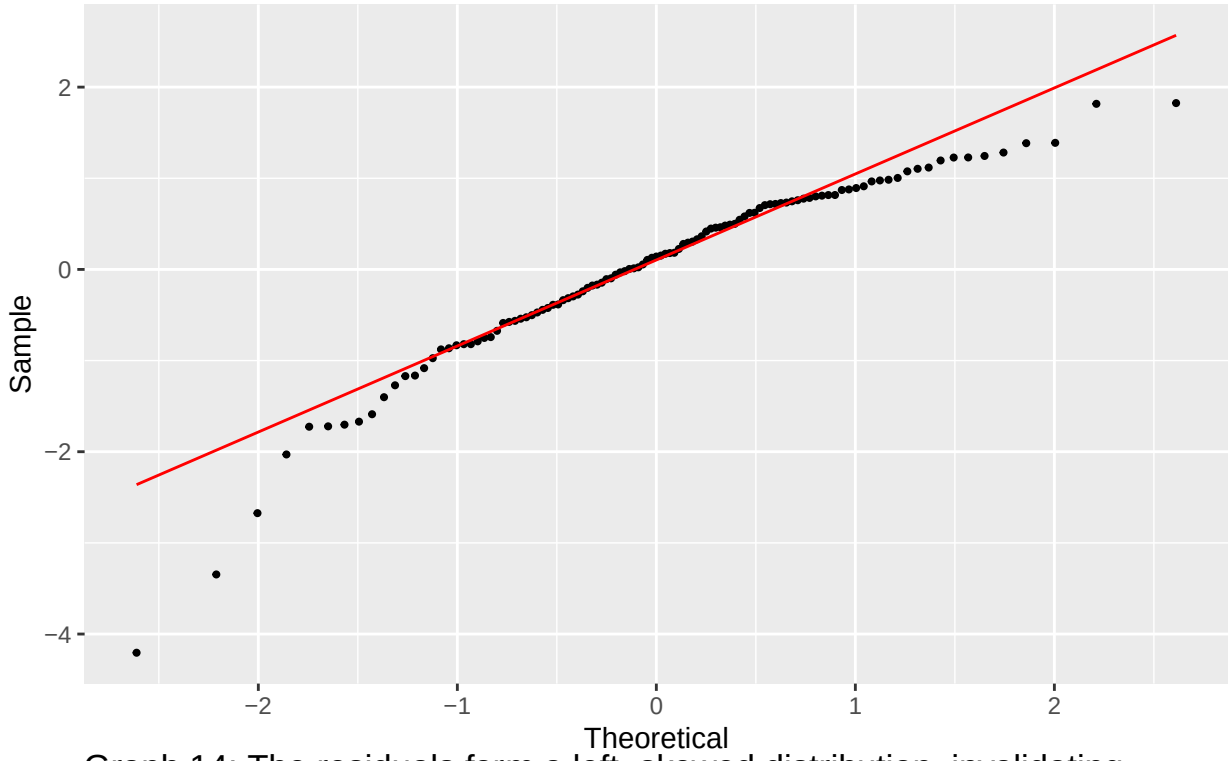
The next condition is independence which requires that the residuals be independent. We will evaluate independence by simply plotting all the residuals on a graph and checking for abnormal patterns or clusters.

Graph 12: Residuals
do not form any patterns or clusters, indicating independence

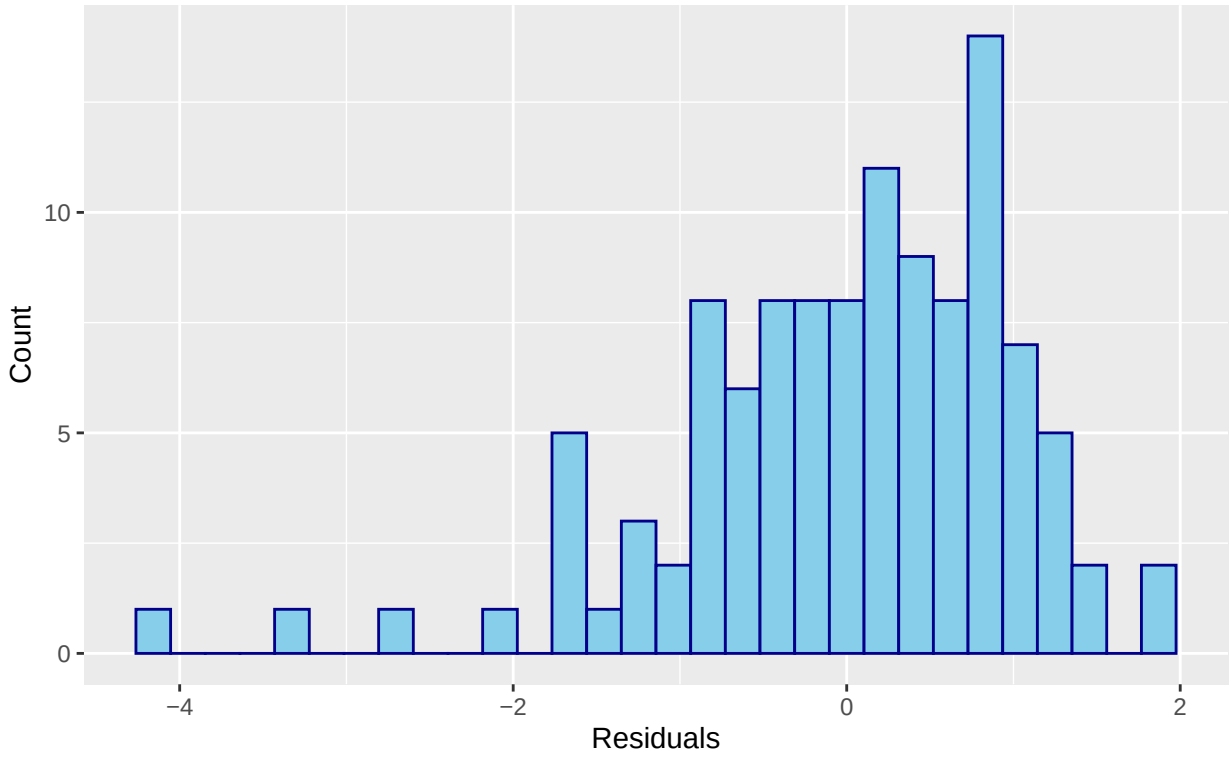


The next requirement is that of normality. The checks for this involved both a plotting of the residuals on a histogram as well as the creation of a Q-Q plot. The histogram was examined for a potential relationship between the residuals and the fit of the model to the qq line was checked too.

Graph 13: The residuals do not form a normal distribution, invalidating the normality condition



Graph 14: The residuals form a left−skewed distribution, invalidating the normality condition



Multiple Linear Regression Model with Log Transformation and Interactions

Next, we constructed a linear regression model with a logarithmic transformation and with interactions between the three greenhouse gas emissions and between CO2 emissions and land under cereal production to better understand the relationship between the predictor variables.

```
## # A tibble: 9 x 7
##   term               estimate   std.error statistic  p.value conf.low  conf.high
##   <chr>                 <dbl>       <dbl>     <dbl>    <dbl>    <dbl>      <dbl>
## 1 (Intercept)         2.05e+0     3.76e-1     5.44  3.73e- 7  1.30e+0    2.79e+0
## 2 land_cereal        -1.85e-6     2.62e-7    -7.05  2.17e-10 -2.37e-6   -1.33e-6
## 3 co2_tons           -1.06e-1     2.42e-1    -0.439 6.62e- 1 -5.87e-1    3.74e-1
## 4 nitrous_tons        4.09e-4     1.61e-4     2.53  1.29e- 2  8.83e-5    7.29e-4
## 5 methane_tons       -2.80e-4     5.27e-5    -5.31  6.44e- 7 -3.84e-4   -1.75e-4
## 6 co2_tons:nitro~    -9.92e-5     3.69e-5    -2.69  8.35e- 3 -1.72e-4   -2.60e-5
## 7 co2_tons:metha~    -5.25e-6     9.70e-6    -0.541 5.90e- 1 -2.45e-5    1.40e-5
## 8 nitrous_tons:m~     1.19e-8     1.87e-9     6.38  5.37e- 9  8.20e-9    1.56e-8
## 9 land_cereal:co~     2.57e-7     6.79e-8     3.78  2.67e- 4  1.22e-7    3.91e-7
```

```
##
## Call:
## lm(formula = log(prop_refugee) ~ land_cereal + co2_tons + nitrous_tons +
##     methane_tons + co2_tons * nitrous_tons + co2_tons * methane_tons +
##     nitrous_tons * methane_tons + co2_tons * land_cereal, data = climate_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5013 -0.4023  0.0801  0.6416  2.1880
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 2.045e+00  3.762e-01   5.436 3.73e-07 ***
## land_cereal                -1.848e-06  2.620e-07  -7.051 2.17e-10 ***
## co2_tons                   -1.064e-01  2.424e-01  -0.439 0.661722
## nitrous_tons                4.086e-04  1.615e-04   2.531 0.012917 *
## methane_tons               -2.799e-04  5.270e-05  -5.310 6.44e-07 ***
## co2_tons:nitrous_tons      -9.918e-05  3.687e-05  -2.690 0.008352 **
## co2_tons:methane_tons      -5.250e-06  9.701e-06  -0.541 0.589602
## nitrous_tons:methane_tons   1.190e-08  1.866e-09   6.376 5.37e-09 ***
## land_cereal:co2_tons        2.565e-07  6.792e-08   3.777 0.000267 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.135 on 102 degrees of freedom
##   (133 observations deleted due to missingness)
## Multiple R-squared:  0.8943, Adjusted R-squared:  0.886
## F-statistic: 107.8 on 8 and 102 DF,  p-value: < 2.2e-16
```

```
## # A tibble: 9 x 2
##   term                      estimate
##   <chr>                        <dbl>
## 1 (Intercept)                  7.73
## 2 land_cereal                  1.00
```

```
## 3 co2_tons                    0.899
## 4 nitrous_tons                1.00
## 5 methane_tons                1.00
## 6 co2_tons:nitrous_tons       1.00
## 7 co2_tons:methane_tons       1.00
## 8 nitrous_tons:methane_tons   1
## 9 land_cereal:co2_tons        1
```

The model that we constructed can be seen here:

Predicted proportion of refugees = e^(2.045090 - 1.847728e-06 * land_cereal - 1.063508e-01 * co2_tons + 4.085926e-04 * nitrous_tons - 2.798573e-04 * methane_tons - 9.918373e-05 (co2_tons * nitrous_tons) - 5.249514e-06 (co2_tons * methane_tons) + 1.189883e-08 (nitrous_tons * methane_tons) * 2.565460e-07 (land_cereal * co2_tons))

We obtained the following 95% of confidence intervals for the slope corresponding to each predictor conditional on all other predictors in our model:

Land under cereal production: (0.9999981, 0.9999987) CO2 emissions: (0.5121042, 0.9273253) N20 emissions: (1.0002614, 1.0006388) Methane emissions: (0.9998803, 1.0000107) CO2 emissions:N20 emissions: (0.9998277, 0.999974) CO2 emissions:methane emissions:(0.9999755, 1.000014) N2O emissions:methane emissions:(1, 1) CO2 emissions:cereal land: (1.0000001, 1.0000004)

Testing for Multicollinearity

To further analyze the relationship between the predictor variables, we tested for mutlicollinearity. This was to determine if two or more explanatory variables in our multiple regression model were highly linearly related. VIF scores were calculated for each predictor used in the main effects model.

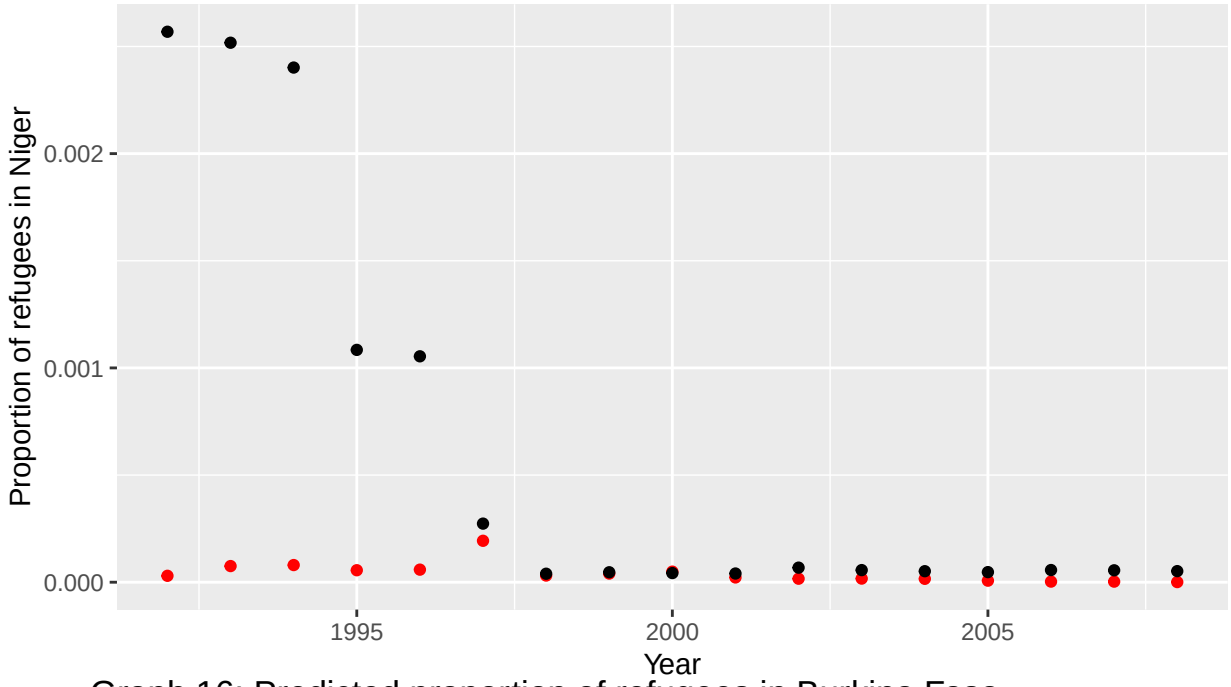
```
##  land_cereal     co2_tons nitrous_tons methane_tons
##    20.132819     4.343368    41.210613    16.117866
```

Applying the Main Effects Model to Another Country

To further evaluate our model, we applied it to another at risk region as identified by the Ecological Threat Register, the Sahel Belt of Africa. The Sahel Belt includes Senegal, Mauritania, Mali, Burkina Faso, Niger, Chad, Sudan, Eritrea. These countries have similar economic situations thus lending it to another application of our model. However, some do not have similar conflict levels so we chose to apply the model to one country that did experience significant amounts of armed conflict and one that did not. Burkina Faso did not experience significant levels of armed conflict while Niger has and continues to experience high amounts of armed conflict.

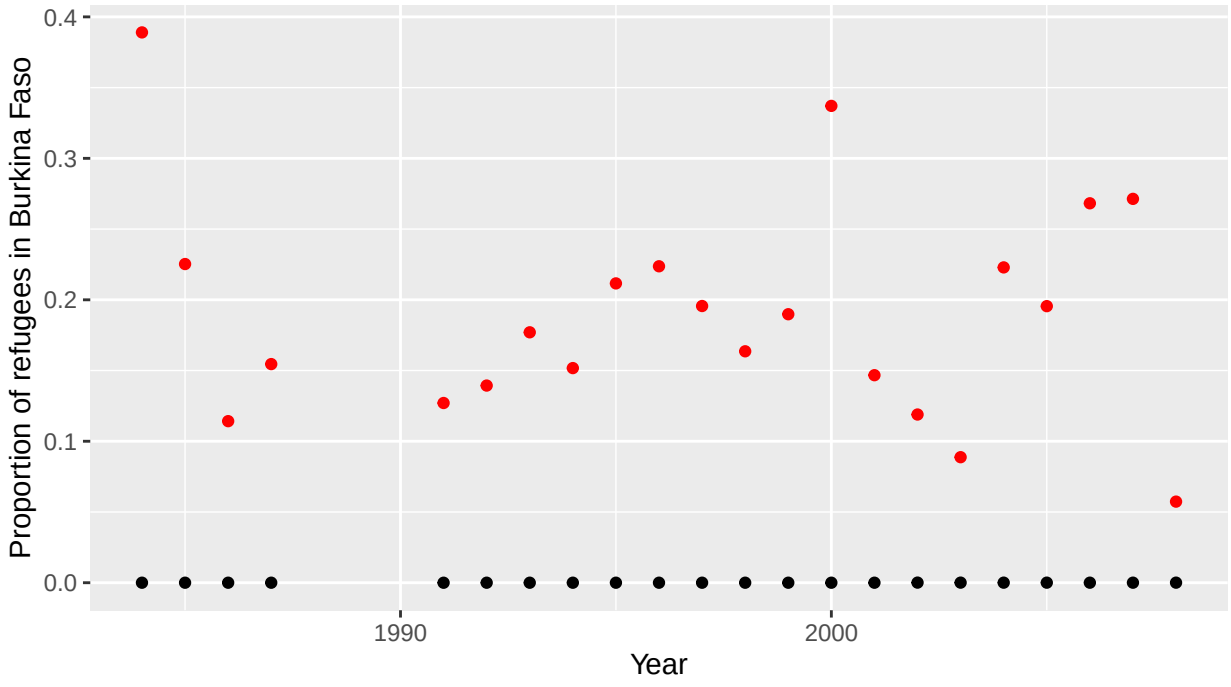
Graph 15: Predicted proportion of refugees in Niger mostly fit observed values after 1995

Red points are predicted values and black points are observed values



Graph 16: Predicted proportion of refugees in Burkina Faso higher than observed proportion of refugees

Red points are predicted values and black points are observed values



RESULTS

Main Effects Linear Regression Model with Log Transformation and Assessing Quality of Fit

Our fitted main-effects multiple linear regression model with a logarithmic transformation is as follows:

Predicted proportion of refugees = eˆ(7.004603e-01 + -1.584535e-06 * land_cereal + -3.723390e-01 * co2_tons + 4.499857e-04 * nitrous_tons + -5.452494e-05 * methane_tons))

At the $\alpha = 0.05$ significance level, the fitted coefficients for each of our predictors in the log transformation relative to the baseline category of standard refugee rate were statistically significant except for methane emissions. Therefore, there is sufficient evidence to suggest that the true coefficients of the log transformation corresponding to the predictor variables hectares of land for cereal production, metric tons of CO2 produced per capita, and kilotons of nitrous oxide produced compared with the log of the proportion of refugees in the country is not equal to 0. There is some relationship between these predictor variables and change in the proportion of refugees in the countries studied.

We are particularly interested in testing to see whether there is a relationship between carbon dioxide emissions and the log proportion of refugees in a population. β_{CO2} is the coefficient in the linear model with log transformation for CO2 emissions per capita effect on refugee flow.

We conducted a hypothesis test at the $\alpha = 0.05$ level.

H_0: There is no relationship between carbon emissions per capita and the log proportion of refugees. β_{CO2} is equal to 0.

H_1: There is a relationship between carbon emission per capita and the log proportion of refugees. β_{CO2} is not equal to 0.

Under the null hypothesis, our test statistic follows a standard normal distribution. The t-statistic is equal to approximately -2.486452 with 106 degrees of freedom, which corresponds to a p-value of 1.446306e-02. Thus, at the $\alpha = 0.05$, we reject our null hypothesis; we have sufficient evidence to suggest that there is a relationship between the metric tons of CO2 produced per capita and the log of the proportion of refugees. The estimated log transformation coefficient was -3.723390e-01 for CO2 emissions. Therefore, we are 95% confidence that the proportion of refugees to decreases by a factor of 0.5121042 - 0.9273253, holding all other variables constant. The 95% confidence intervals for nitrous oxide emissions and methane emissions respectively are (1.0002614 - 1.0006388) and (0.9998803 - 1.0000107), which suggests that compared carbon dioxide emissions, nitrous and methane emissions have a lesser impact on the proportion of refugees in a population.

The adjusted coefficient of determination (r squared) for the linear regression model was 0.792. This means 79.2% of the variability in the proportion of refugees can be explained by the model. The adjusted coefficient of determination was used since our model uses multiple predictors.

Diagnostic Plots for Main Effects Linear Regression Model with Log Transformation

Independence: The residuals do not exhibit any pattern or abnormal clusters.

Linearity: There is an approximately symmetric distribution above and below the y=0 line.

Equal Variance: The vertical variation or spread of the residuals is not approximately equal. There is less variance at the extremes of the plot indicating that there is not equal variance.

Normality: The residuals do not appear to have a normal distribution in either the histogram and the Q-Q plot.

The conditions of normality and equal variance were violated, indicating that the linear model with log transformation was does not fully capture the relationship between the climate change indicators and the log of the proportion of refugees. However, given the very low p-values for some of the coefficients in the

model, we can make some inferential statements from our model. However, of the statistically significant predictors, the CO2 emissions coefficient had the highest p-value of 1.446306e-02. Since this is close to the alpha, we should be more careful about inferential statements with CO2 emissions. In the future we would look to use a different model that can better capture the sophisticated relationship between these climate change predictors and the proportion of refugees in a country.

Multiple Linear Regression Model with Log Transformation and Interactions

We also constructed a linear model with interactions between each of the predictor variables and a logarithmic transformation. However, we found that this model had limited usefulness compared to the model that analyzed main effects.

Testing for Multicollinearity

The VIF scores of the variables output above indicate high degree of multicollinearity with several of the variables. Everything expect for CO2 emissions had a VIF greater than 5 which was our established level of concern. This not surprising given that we were including variables that measured emissions. Something that is a source of emissions is likely not a source for only a single type of emissions. This means that as one type of emissions increases, the other kinds of emissions will likely also increase. Thus there was a great deal of correlation between our variables. That interaction between the variables made the point predictions in our test data from the Sahel very inaccurate. Methods for resolving these problems are addressed in the limitations and concerns sections.

Applying Model to Another Country

The application of the multiple linear regression model with logarithmic transformation main effects to the Sahel Belt region countries yielded mixed results. While the model closely predicted the proportion of refugees in Niger after 1995, it overestimated the proportion of refugees in Burkina Faso over all years.

This discrepancy may be due to the low number of refugees in Burkina Faso. In fact, the proportion of refugees in Burkina Faso ranges from 6.386 * 10ˆ-6 to 2.69 * 10ˆ-5. Meanwhile, Niger has generally higher proportion of refugees. This indicates that the model may be best applied to countries with significant refugee populations. This relates to earlier discussion over the inclusion of Syria. Broadly, the existence of conflict does not immediately justify the exclusion of a country from the construction of our model. Climate factors alone would not push someone to leave their country of residence. Instead, we aim to observe how influential climate change indicators are when there other factors such as conflict which would increase a person's desire to leave their country. Thus countries that have had significant refugee populations both with and without conflict present are being included in the model and the model is consequently best applied to countries with significant refugee populations.

DISCUSSION

Given the rapidly accelerating nature of climate change, it is imperative that more data analysis be done on the massive impacts it has. Our goal was to see if and how climate change factors like CO2 emissions and land under cereal production would influence refugee flow from a country as a terminal measure of climate change impacts. With so many people denying that climate change impacts their lives, we decided to explore how climate change predictors may have influenced the lives of millions of people. We constructed a linear model which allowed us to understand not only the direct effects of several climate change indicators on refugee flow, but also the interactions and relationships among climate change indicators. Our linear model provided insight on the existence of a relationship between nitrous oxide, CO2, and methane emissions alongside cereal land on refugee flow in Middle Eastern countries as well as the extent of their impact. Some variables like cereal land available and CO2 emissions were chosen because they are very obvious markers of

human impact on the climate. Additionally, nitrous oxide and methane emissions were chosen because they are downstream markers of anthropogenic climate change that are closer to affecting human lives. There were also interactions effects measured between carbon, nitrous oxide, and methane emissions due to the facts that they all contribute to the greenhouse effect which has a direct impact on global climate change.

We hypothesized that at least one of the climate predictors (CO2 emissions in kilotons per capita, other greenhouse emissions in kilotons, and land under cereal production) in our linear model would have a statistically significant relationship with refugee flow from the countries in the Middle East and Central Asian Belt. Through our analysis on the Middle East and Central Asian belt, after excluding Syria due to the clear correlation between refugee flow and the 2011 Civil War, we observed that there is a relationship between certain climate change indicators and refugee. The p-values for CO2 emissions, nitrous oxide emissions, methane emissions, and cereal land in our regression model are 1.446306e-02, 6.929930e-06, 1.003421e-01, and 2.749387e-16 respectively. Since the p-values for CO2 emissions, nitrous oxide emissions, and land under cereal cereal production are less than 0.05, we found sufficient evidence to suggest a statistically significant relationship between each climate change predictors and refugee flow.

In particular, we conducted a hypothesis test at the $\alpha = 0.05$ level to understand if there is a relationship between carbon dioxide emissions and the log proportion of refugees in a population. We concluded that there was sufficient evidence to suggest that the β_{CO2} coefficient for carbon dioxide emissions was statistically significant. We found that on average, a one unit increase in carbon dioxide (metric tons per capita) resulted in a decrease in refugee flow by a factor of 0.689 holding all else constant. This is a significantly large effect on refugee flow. However, on average, we found that a one unit increase in land under cereal production, resulted in an increase in refugee flow by a factor of about 1 holding all other variables constant. The same was found for nitrous oxide emissions. For example, when we exponentiate the coefficients for these variables in our model, we find that the the proportion of refugees changes, on average, by factors of 0.999998 and 1.000450 for each hectare increase for cereal-production land and for each additional kiloton of nitrous oxide produced, respectively, when we keep all other variables the same. Although the coefficients for nitrous oxide emissions and percentage of land allocated to cereal production are significant at the alpha = 0.05 level, their actual effects according to the model are extremely small.

We also found that a multiple regression model incorporating the interaction between the explanatory variables was no more useful than a multiple regression model analyzing main effects. Nevertheless, when testing for multicollinearity to further understand the relationship between explanatory variables, the variance inflation factor (VIF) score for every variable except for carbon dioxide emissions was greater than 5. This indicates a high degree of multicollinearity with several of the variables. These may have made the model less applicable to other region and accounted for some of the inaccuracies of the model.

While we did not formally predict a direction for the relationship between our predictors and refugee flow, our general thought was that as markers of climate change worsened (i.e. emissions increased), the refugee flow from a country would increase. However, the opposite was true for some emissions markers, most clearly for CO2 emissions in kilotons per capita. This was likely the case because CO2 emissions are not just a marker of a worsening climate, but also a marker of the economic development of a country. As almost every country in the world has industrialized, their CO2 emissions have increased. Additionally, while the impacts of climate change can take many years to accumulate to a disastrous level, the impacts of rapid economic development can be felt much faster. Those impacts are usually more job opportunities and decreases in poverty, both of which would make someone more likely to stay in the current country of residence. Overall, this data signifies a general relationship all over the world which is that rapid economic development may make immediate living situations easier, but it will also have impacts on the environment such as increased emissions.

Through our analyses, we were able to obtain an adjusted R squared of 0.7923 which means our model explains 79.2% of the proportion refugee flow. When applied to countries in other at-risk ecological hotspots (in accordance with the Ecological Threat Register by the IEP) with high proportion of refugees and some conflict such as Niger, the model proved to be more accurate compared to an application on a country with low conflict and low proportion of refugees. This demonstrates that variables not included in our model such as conflict could be variable which are associated with refugee flow. This is further discussed in the limitations and concerns section below.

Limitations, Concerns, and Future Investigation

Our adjusted r squared value was surprising and much higher than we expected. Although our analyses show significant relationships between climate change indicators with the proportion of refugee flow, we cannot conclude with certainty that climate change indicators directly impact (and increase) refugee flow or that these relationships are equally applicable in other regions and countries. It may be that climate change indicators coincided with major political shifts within a country which caused an increase (or decrease) in the proportion of refugee flow. The data on refugee populations was a raw sum and doesn't disaggregate the data based on reason for refugee status. Further studies on the relationship between climate predictors and displaced populations should adjust for other push factors such as conflict, poverty, socio-economic factors, and civil rights violations which may be associated with the proportion of refugees. Though the World Bank data includes data for internally displaced people due to disasters, this metric lacks enough data for analysis. Further investigation should continue to measure and increase data on the number of "climate refugees" or people displaced due to natural disasters.

Moreover, a major concern with the validity of our model was the lack of consistent, regularly recorded data. The World Bank data which was used was missing for several years and we had to develop our regional groupings due to lack of collation on the part of the World Bank. Moreover, robust, comprehensive data on refugee populations from each country doesn't begin until the late 1980s and early 1990s whereas other metrics such as carbon dioxide emissions begin in the 1960s.

Additionally, and perhaps most importantly, although our model seems to satisfy the conditions of linearity and independence, it does not satisfy the conditions of normality and equal variance. Because of this, the linear model with log transformation is likely not the best choice for modeling the data. Our low p-values and enabled us to evaluate and make some statements on construction of a linear model to study if and how a relationship between climate. However, due to these concerns and violations of conditions required for inference in our linear regression model, we cannot make extensive inferential statements using the model and further studies should test other models to understand any relationships between climate indicators and response variables.

Were we to repeat our analysis, we would include and adjust for more potential confounders and factors associated with proportion of refugee flow to improve the prediction accuracy of our model. The test for multicollinearity demonstrated high correlations between some of the predictor variables, especially related to greenhouse gas emissions. We can address multicollinearity by linearly combining the highly correlated independent variables. In context, having a variable for greenhouse gas emissions makes sense within the literature since greenhouse gas emissions are often measured together. Moreover, relying on more robust and extensive data will help bolster the accuracy of any model. Additionally, other investigations should test and analyze the relationship between other climate change indicators such as PM2.5 pollution levels, sea level rise, temperature increases, and agricultural output and with refugee flow. Notably, we focused on factors which cause climate change but analyzing the relationship of variables that correspond to the consequences of climate change and may have a more acute effect on refugee flow. Lastly, testing other nonlinear models may produce more accurate and insightful results.

Overall, this not only served as an excellent exercise for us in applying linear regression to the real world, but also can act as the basis for future work trying to develop a quantitative prediction of refugee flow in the future.

REFERENCES

Aye, Goodness C., et al. "Effect of Economic Growth on CO2 Emission in Developing Countries: Evidence from a Dynamic Panel Threshold Model." Cogent Economics & Finance, vol. 5, no. 1, 2017, p. 1379239., doi:10.1080/23322039.2017.1379239.

BhandariI, Aniruddha. "Multicollinearity: Detecting Multicollinearity with VIF." Analytics Vidhya, 16 Apr. 2020, www.analyticsvidhya.com/blog/2020/03/what-is-multicollinearity/.

Brown, O. (2008). Migration and Climate Change. IOM Migration Research Series Migration and Climate Change, (31). https://doi.org/10.18356/bd790a56-en

Institute for Economics and Peace (2020). Ecological Threat Register. http://www.activist360.co/wp-content/uploads/2020/09/ETR_2020_web-1.pdf

Podesta, John (2019). The Climate Crisis, Migration, and Refugees. Brookings Institution. https://www.brookings.edu/research/the-climate-crisis-migration-and-refugees/

The World Bank (2020). Data Bank: World Development Indicators. https://databank.worldbank.org/source/world-development-indicators#

United Nations High Commissioner for Refugees. Climate change and displacement. https://www.unhcr.org/en-us/news/stories/2019/10/5da5e18c4/climate-change-and-displacement.html.