DATA 607 Data Project Final

Ben Wolin

Abstract

Climate change is one of the biggest existential challenges we face today. It is fundamentally reshaping our world in ways that cannot always be predicted. This has the potential to lead to massive consequences such as food insecurity, housing insecurity due to rising sea levels, mass extinction, etc. One of the best ways we as people could help slow down this shift is by reducing our emissions. This study looks at one possible way of doing that. The meat industry has a massive footprint on our emissions. These emissions stem from every aspect of the industry, from transportation of products to the actual animals themselves. What this study attempts to do is show a correlation between a country's overall meat consumption and the number of agricultural emissions that country produces. Using data from the World Bank Data Catalog we were able to take match 27 years of agricultural emissions and meat consumption data to create one data set to use in this study. What we found was that these two factors have a correlation coefficient of 0.9523607, showing a strong positive correlation. This correlation was then shown to have an extremely small p-value of <2.2e-16, showing that it is not only strong, but statistically significant as well. However, there were some issues with this study. The countries present in the data tended to be part of 2 groups, either they were economically very strong or modest. There were not many countries in the middle ground. This could potentially skew our result, however, with the data at hand we did see that positive, strong correlation that we expected. This means that in the end our hypothesis was confirmed.

Introduction

In this project we will be taking data from the World Bank Data Catalog. This data provides with information on agricultural emissions and meat consumption by country by year. We will be using this data to answer the question, does the amount of meat eaten in a country on a given year correlate with the amount of agricultural CO2 that is released in that year?

Data Loading & Manipulation

First we will load in the necessary packages for this project.

```
library(tidyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(aws.s3)
```

```
## Warning: package 'aws.s3' was built under R version 4.4.2
```

```
library(ggplot2)
```

Next we will load in our data. The emissions data is taken from an AWS S3 bucket, while the consumption data is a CSV we are loading in from GitHub.

```
Sys.setenv("AWS_ACCESS_KEY_ID" = "AKIAVY2PGVWISHYZ56BY","AWS_SECRET_ACCESS_KEY" = "VhjexYyYvZ7SYNJmWqNL/

Emissions_raw <-
    aws.s3::s3read_using(read.csv, object = "s3://data-607/Emissions.csv")

Consumption_raw <- read.csv('https://raw.githubusercontent.com/bwolin99/TestRepo/refs/heads/main/607%20P
```

Now we will use the gather function to create useful columns for our analysis. The resulting clean data from our two sources are then joined together. This will allows us to create linear models from the data located in these 2 sources.

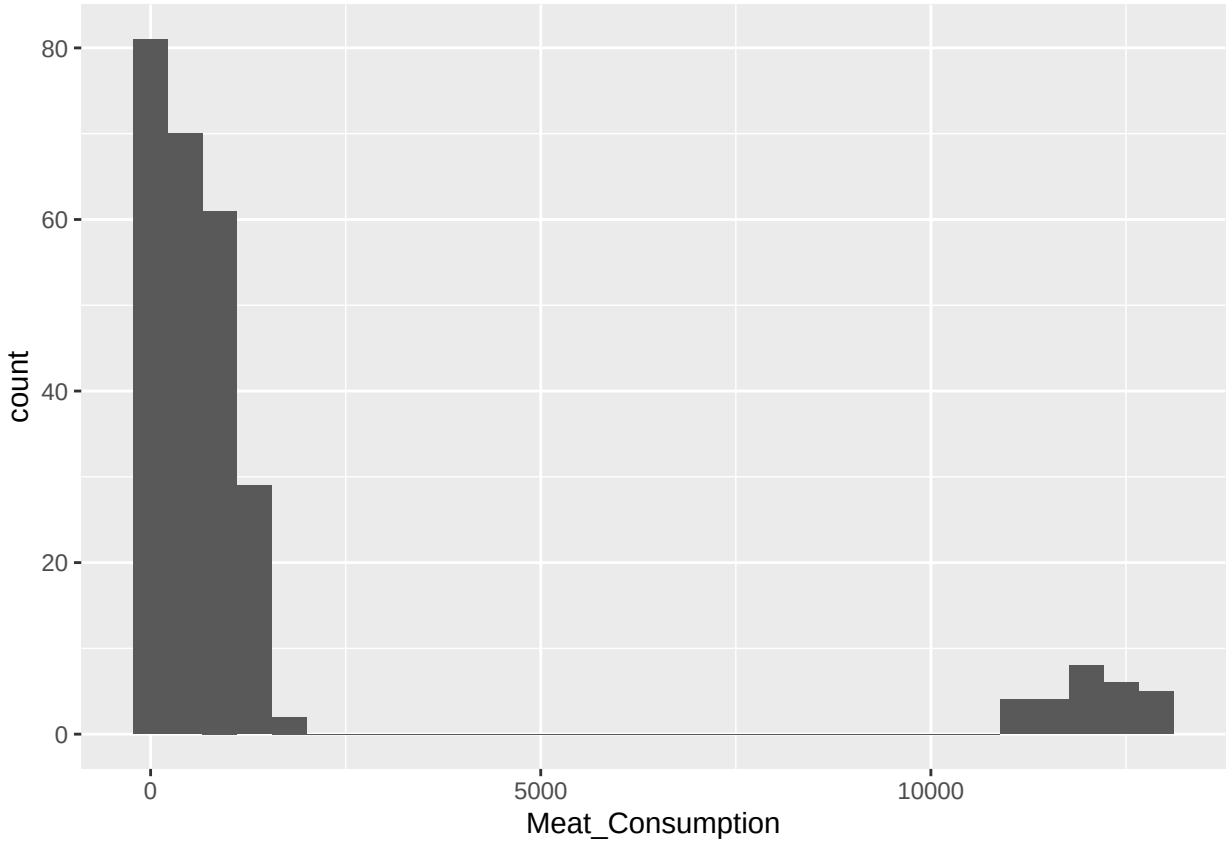
```
Emissions <- gather(Emissions_raw, key = 'year', value = 'Aggricultural_Emissions',3:31)

Consumption <- gather(Consumption_raw, key = 'year', value = 'Meat_Consumption',2:28)

data_final <- Consumption %>% inner_join(Emissions,
        by=c('year'='year', 'country'='country'))
head(data_final)
```

```
##   country  year Meat_Consumption gas Aggricultural_Emissions
## 1     UKR X1991        1878.0000 CO2               1.5810500
## 2     NOR X1991          78.0000 CO2               0.2011950
## 3     KAZ X1991         724.0000 CO2               0.0913600
## 4     USA X1991       11076.0477 CO2               6.9605966
## 5     NZL X1991         118.3013 CO2               0.4405406
## 6     TUR X1991         373.0000 CO2               0.4361977
```

Next we will visualize our data using a few plots. Here we can see 2 histograms that are very similar, with a lot of data points on both the low end and high end of the spectrum, with few in the middle. This may point to an income inequality for the countries included in this analysis. Since meat and power used for agriculture are costly this could be showing a large gap between 2 levels of economy.

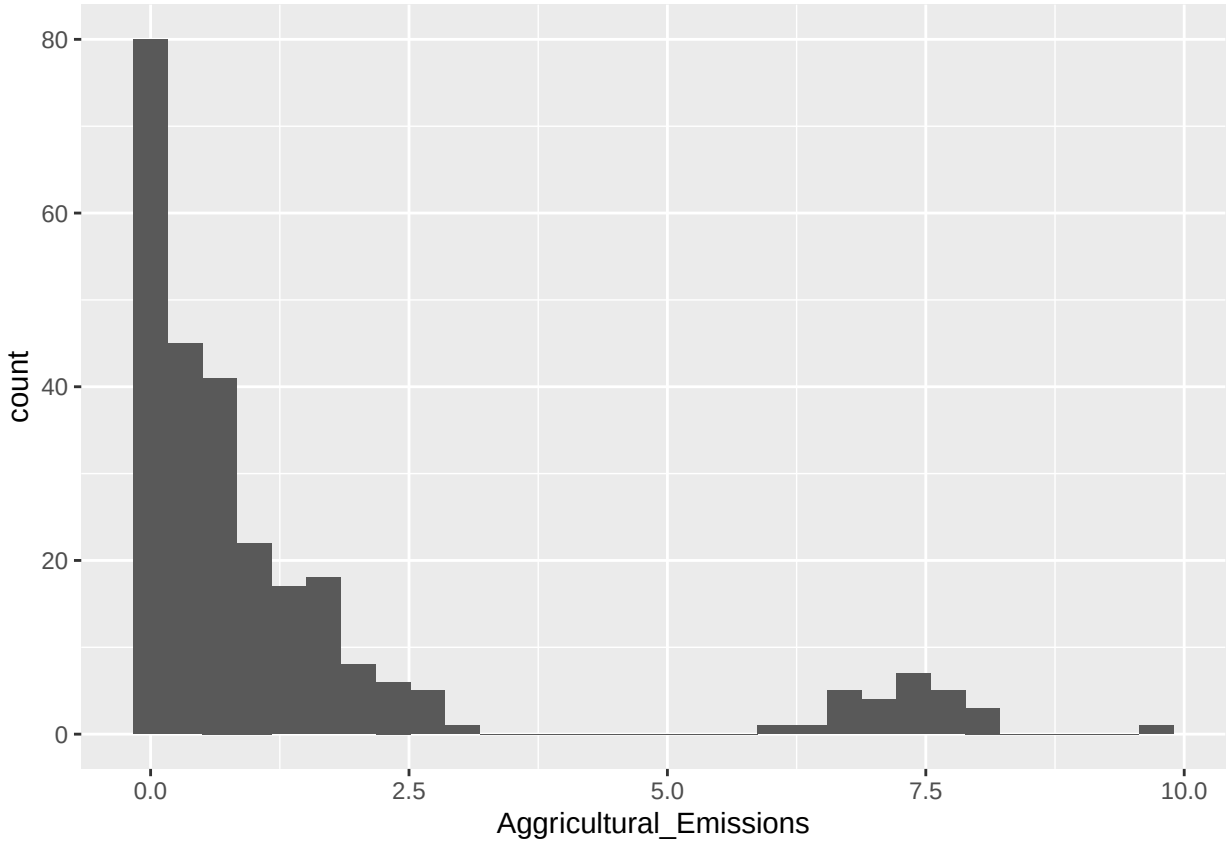
```
library(ggplot2)
ggplot(data_final, aes(x=Meat_Consumption)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



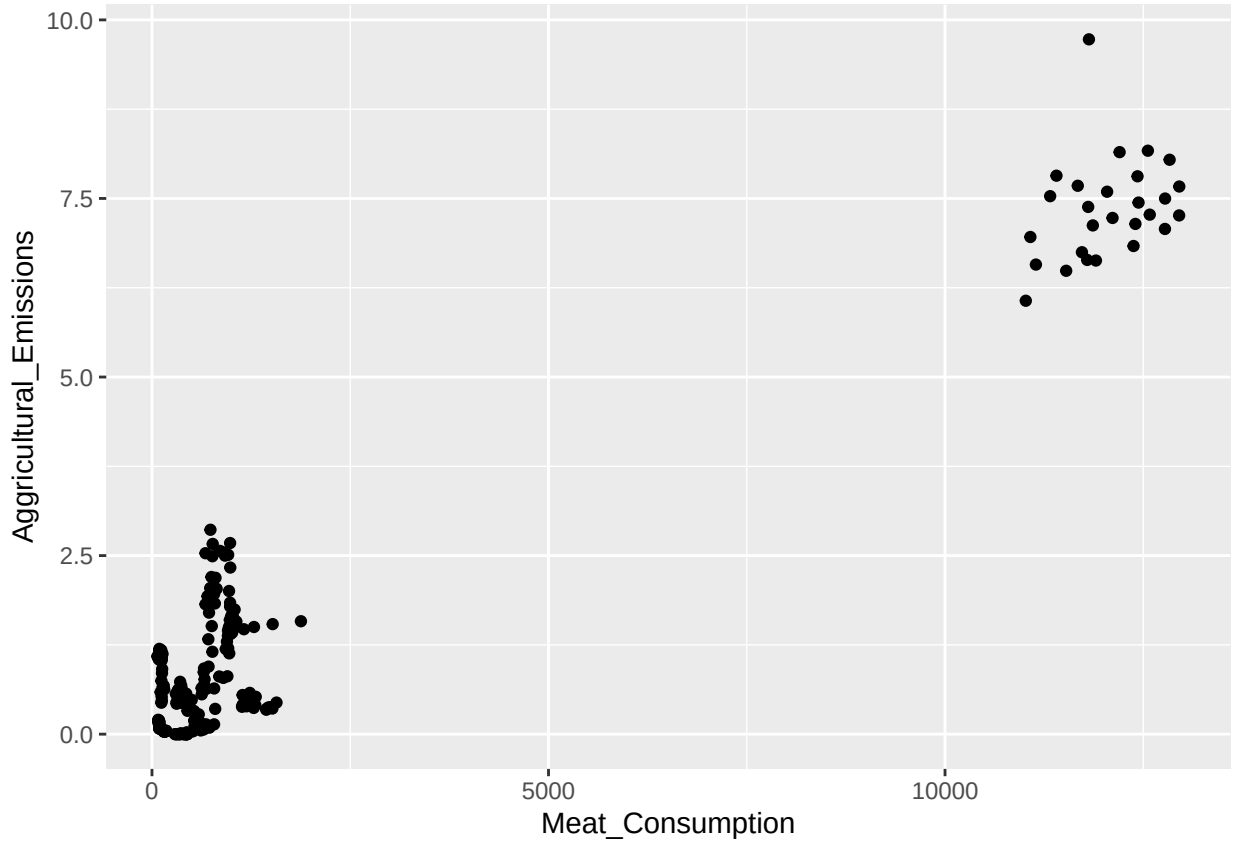

```
ggplot(data_final, aes(x=Aggricultural_Emissions)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



When plotted on a scatter plot it seems like there is a positive linear relationship between agricultural emissions and meat consumption at first glance. Again, we are also see 2 distinct cluster, presumably between wealthier and poorer nations.

```
ggplot(data_final, aes(x=Meat_Consumption,y = Aggricultural_Emissions)) + geom_jitter()
```



Data Analysis

Now we'll move onto our data analysis. First we will calculate the pearson's product-moment correlation of this data set.

```
cor.test(x = data_final$Meat_Consumption, y=data_final$Aggricultural_Emissions)
```

```
##
##  Pearson's product-moment correlation
##
## data:  data_final$Meat_Consumption and data_final$Aggricultural_Emissions
## t = 51.122, df = 268, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9398335 0.9623303
## sample estimates:
##       cor
## 0.9523607
```

This shows us that we have a correlation coefficient of about 0.95. This confirms our suspicions that there is a large, positive correlation between these two factors. Next we will create a linear regression model of this data to observe the p-value and get a equation.

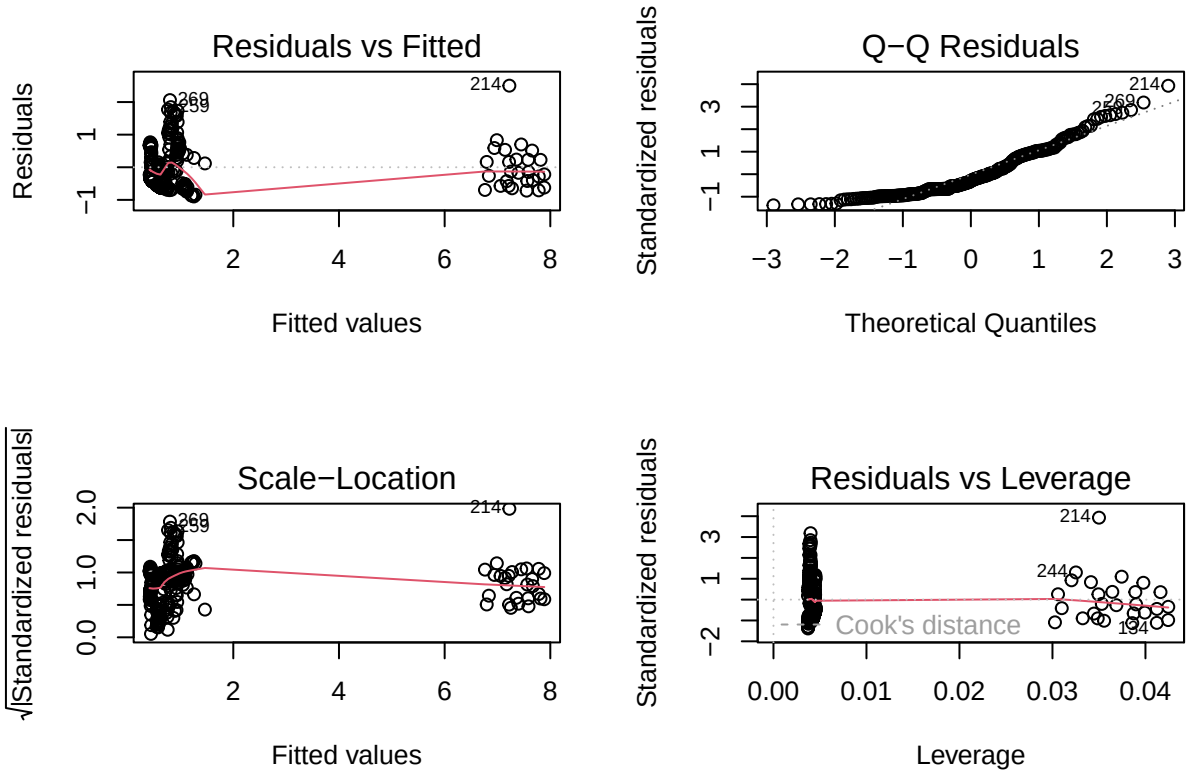
```r
Emission_lm = lm(data_final$Aggricultural_Emissions ~ data_final$Meat_Consumption)
summary(Emission_lm)
```

```
##
## Call:
## lm(formula = data_final$Aggricultural_Emissions ~ data_final$Meat_Consumption)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8949 -0.4867 -0.1864  0.4593  2.5024
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  3.737e-01  4.393e-02   8.507 1.28e-15 ***
## data_final$Meat_Consumption 5.799e-04  1.134e-05  51.122  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6484 on 268 degrees of freedom
## Multiple R-squared:  0.907,  Adjusted R-squared:  0.9066
## F-statistic:  2613 on 1 and 268 DF,  p-value: < 2.2e-16
```

As we can see here the p-value for this relationship is below the lower threshold of this command which is <2.2x10^-16. This is far below the threshold of 0.05, showing that the large correlation we saw before is statistically significant. Also, the adjusted R squared is 0.9066, which means this model is a very strong fit for the data. The linear model generated is as follows: Aggricultural_Emissions = 0.3737 + 0.0005799(Meat_Consumption)

Now we will take a look at the residuals plots to determine if the conditions of least squares are reasonable.

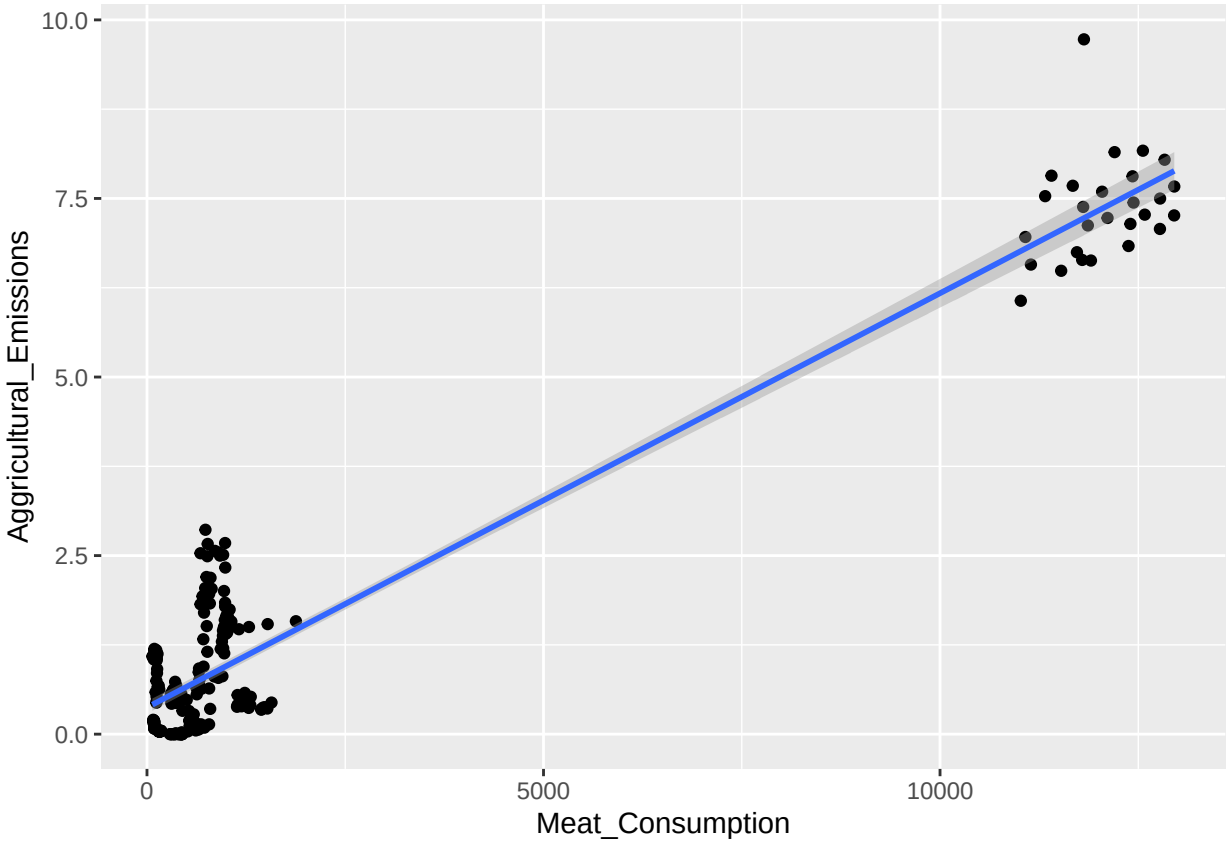
```r
par(mfrow = c(2,2))
plot(Emission_lm)
```



The residuals vs fitted graph is what you would expect. Since the data is skewed towards 2 extremes with not many data points in the middle it does make these graphs slightly less useful. However we do see the values scattered across the fitted line, meaning this should meet the conditions of least squares. The QQ plot does have a slightly rightward skew, meaning that the data is not completely evenly distributed throughout it's quartiles. However, this could be due to us having a small sample of countries to work with. Overall, these graphs do highlight the issues with using a smaller data set skewed toward 2 sides of the spectrum.

```
ggplot(data_final, aes(x=Meat_Consumption,y = Aggricultural_Emissions)) + geom_jitter() + geom_smooth(m
```

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



Conclusion

In conclusion, this project found that there is a correlation between the amount of meat a country consumes and its agricultural emissions. This correlation was show to be positive and strong, with a correlation coefficient of 0.9523607. It was also shown to be a statistically significant correlation, with a p-value of <2.2e-16. The relationship was mapped out with a linear regression model following the formula Agricultural_Emissions = 0.3737 + 0.0005799(Meat_Consumption). However, there were some issues with this study. First, we were only able to collect the data for 10 different countries over a 27 year time frame. This gave us a decent amount of data to work with, however, these countries typically fell into 2 groups. It would greatly benefit the validity of the study to have more data from countries that have moderate level of consumption, as opposed to the 2 extreme groups that are present. This was most present in the scatter plot and the residual diagnostics graphs. In these graphs we saw how the 2 extremes affected the QQ plot, skewing it and leading to lopsided quartiles. Despite this drawback, we are still able to conclude that a higher level of meat consumption will lead to greater agricultural emissions.

References

This data comes from the world bank data catalog, and is a combination of data from these two different datasets. https://energydata.info/dataset/world-climate-watch/resource/47608ee8-92b9-4611-ba6f-720662193e25/

https://energydata.info/dataset/world-climate-watch/resource/1631a4e8-a59a-4026-aa36-162df9b15340