



College Dropouts

Project Proposal

03/08/2024

—

Kayla Casey
Joshua Carlson
Cora McAnulty
Claudia Dare

Data


We will be drawing our data from several sources for this project and synthesizing it into a cohesive project that aims to describe the ways in which our climate is changing. Below are the descriptions of the data sources we are using right now, but we may add more later.

- [FAOSTAT Dataset](#)
 - This dataset is from the Food and Agricultural Organization of the United Nations, and it contains yearly overall temperature change data and temperature standard deviation for every country. For every country (with the exception of a few where data was not collected), there is data from the years of 1961 to 2022 with the change in average temperature and standard deviation compared to the year before. This data is also separated by month, making it applicable to compare how climate change has affected the average temperature in different months/seasons.
- [NFA Dataset](#)
 - This dataset is from the National Footprints Accounts which is a dataset provided by the United Nations, culminating data from the Food and Agriculture Organization, United Nations Commodity Trade Statistics Database, and the UN Statistics Division, as well as the International Energy Agency. This dataset features information on most countries to measure the ecological resource use and resource capacity of nations spanning from 1961-2014, given that data was recorded in each of these nations for this entire range of years. This dataset takes various climate information into account, such as carbon emission, fishery & aquaculture, and livestock products, in order to produce an ecological footprint measurement and footprint rating for each year in each nation.

Motivation/Goals

After forming our group, we discussed what interests we all had in common seeing that we all come from different majors and different backgrounds. We all agreed that we had a strong interest in environmental science and how we could use machine learning techniques to explore research in this field. Predictive analysis was an interest we also shared, and in general how machine learning could be used to predict and provide solutions to the climate crisis.

Our goals for this project are to be able to explore how different machine learning techniques such as regression, time series analysis, and ensemble methods can be used to create meaningful predictions on temperature change and deviation on a country, regional,



and worldwide level. We are hoping to provide statistically significant results on temperature change and deviation with the additional data we found on climate footprints to be able to meaningfully validate our results.

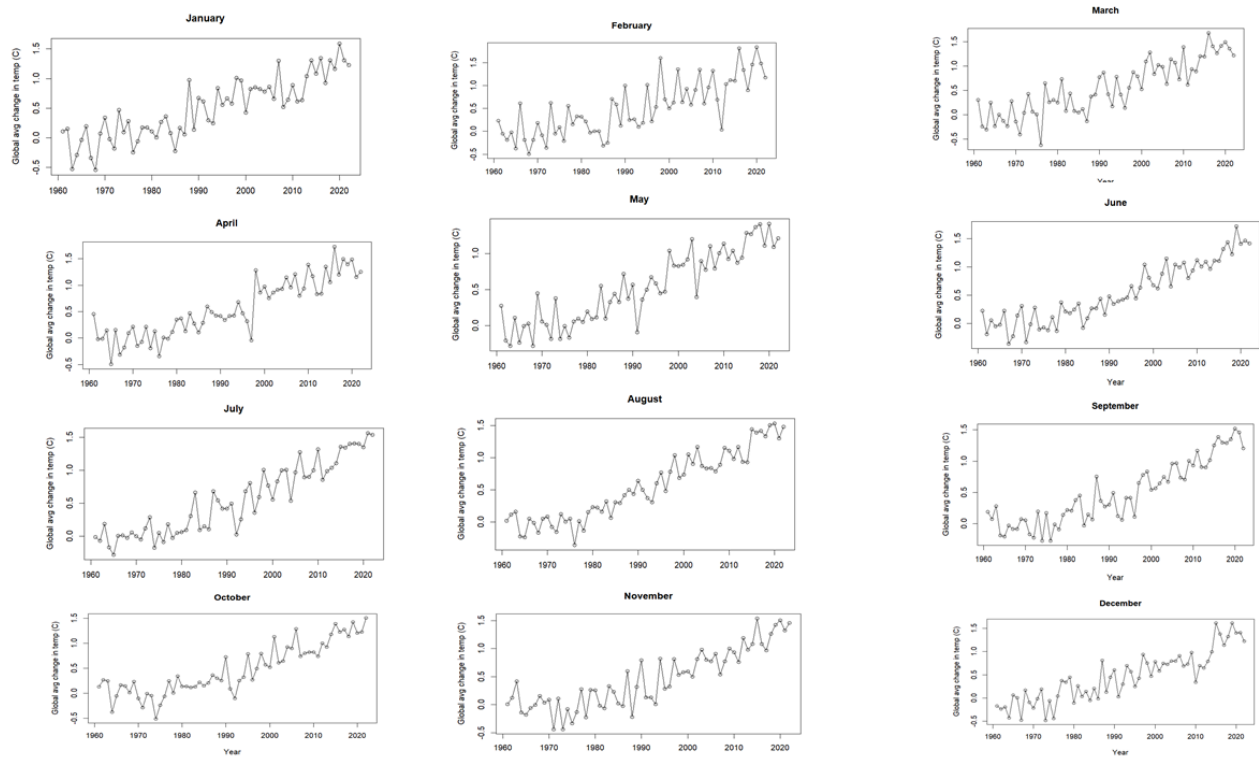
Exploratory Data Analysis

We began our exploratory data analysis in R as this is the language we will be using predominantly for the project. Before making any plots of the data, we cleaned our two datasets to ensure that they only contained countries found in the intersection of the two data sources. We did this step to ensure that when we began to analyze the data, we would not run the risk of outlier countries having an impact on our results. This step would make it easier for us to use the two datasets concurrently as well.

Another important acknowledgment we made was that the two datasets cover different ranges in years. The FAOSTAT dataset has data ranging from 1961-2022, whereas the NFA dataset only has data ranging from 1961-2014. When we found these datasets, we agreed that the majority of our analysis would be on the FAOSTAT dataset, and the NFA dataset would serve as supplementary information. Although these ranges in years are similar, we still agreed to use the FAOSTAT dataset for the bulk of our project, and to use the NFA dataset to give necessary context into the results we find from the FAOSTAT dataset.

Graphs

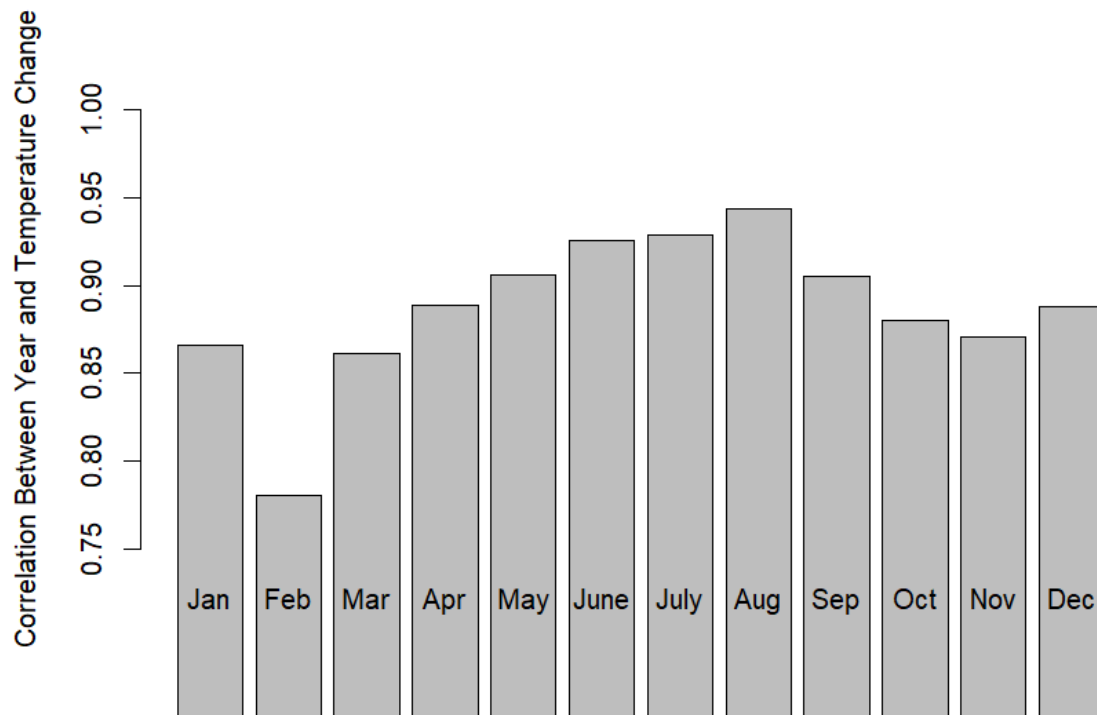
Average Temperature Change by Month



As seen in this set of plots, the year over-year average global temperature has tended to increase in every single month for the past several decades. These plots tell us several things

- Although there does appear to be a strong overall increase in temperature change over time, the data is by no means smooth. Yearly global climates are affected by a variety of large-scale weather events that are difficult to predict and account for, which means that there is quite a bit of noise in our data.
- The noticeable recent increases in temperature change are not distributed universally over all of the months, and further analysis is needed to identify trends on this subject.
- Temperature changes have not always been increasing. Looking at the plots for many of the months above, limiting our analysis to the 60s and 70s shows a relative flatness before a sharp increase beginning in the 80s. This could reflect the beginning of a period of increased industrialization and/or pollution, but other variables from different data sets will need to be examined to make a better guess at the cause of this trend.

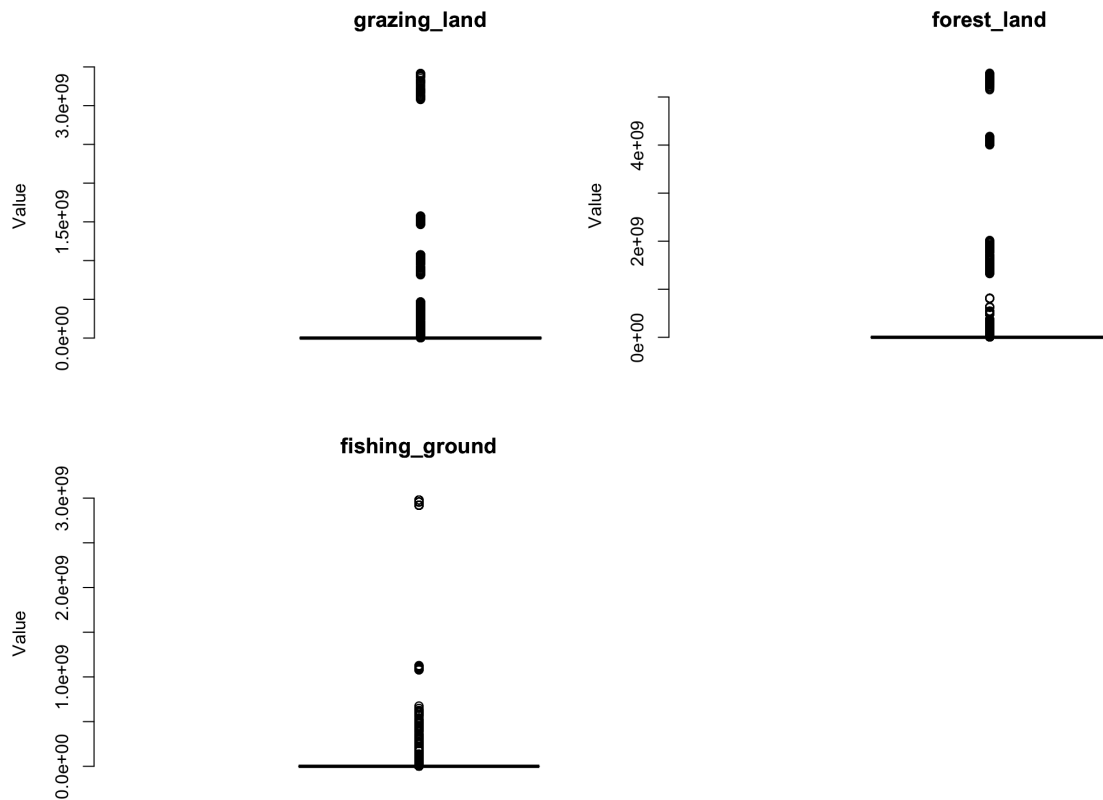
As mentioned in the second bullet point, we can do further analysis to see how these noticeable increases in temperature change are distributed over the months of the year.



This plot, again separated by month, shows the correlation between time and global average temperature change. The fact that all three of the highest values on this bar plot occur in the summer months shows that temperatures have tended to increase more in the summer over time. However, it is important to note the scaling of the plot: all of the correlations are quite close to 1, meaning that every month has a strong upwards trend in temperature change, so the hypothesis that summer temperatures have tended to increase by more may not be significant.

Boxplots of footprint numerical data

The NFA dataset has different measures of a country's ecological footprint in areas such as grazing land and fishing ground. Before we began our analysis using this data, we decided it would be important to visualize the spread of the numerical data that we had. We made boxplots for each of the numerical columns in this dataset. Below, you will see the boxplots for `grazing_land`, `forest_land`, and `fishing_gound`.



From these boxplots, we are able to conclude that many of the countries in this dataset do not have data on some of the predictors. This may be due to countries not recording this data for a period of time, not having any quantifiable measure of the predictors (i.e. a country does not have a fishing industry so they have no fishing_ground), or other reasons unknown to us. In knowing this, we will be mindful of how we use this data to draw conclusions as the entire story to the data is not fully explained.

A short-term goal for the project would be to find other datasets that, when combined with this analysis, could lead to more information about the actual causes of these changes. For example, we will probably look for general country-information datasets and perhaps use dimension reduction techniques and clustering to search for trends and possibly make predictions.

Techniques

Some techniques that we are interested in using include regression, time series analysis, ensemble methods, and clustering.

I. Regression

Our FAOSTAT dataset is well suited for regression, as is the NFA data. We could use regression to create models to predict temperature change or ecological footprint change.

II. Time Series Analysis

Our FAOSTAT dataset seems to be a great candidate for time series analysis. We believe we could use this entire dataset, removing trends and seasonality, and then creating predictive models to approximate future temperature change and deviation.

III. Ensemble Learning

Given that we have many categories within both the NFA and FAOSTAT dataset, we could use ensemble methods to create multiple models and then combine them into a cohesive, well rounded model used for predicting ecological footprint or temperature change in countries.

IV. Clustering

Clustering could be a good technique to use on our data in order to lump countries into groups based on similarities in their data, which could then influence the models we make.