2019

Carbon Footprints

USING DATA TO PREDICT CARBON EMISSIONS BY COUNTRY

HEATHER EWTON

Table of Contents

Carbon Footprints

Problem

Every person, community, and country contributes to a change in ecosystems and climate through an 'ecological footprint'. An ecological footprint is a measure of how much a person, community, or country affects the biosphere, hydrosphere, and atmosphere through their daily activities. Every single activity impacts the biosphere, hydrosphere, and/or atmosphere in some way and because of this impact, it is essential that we understand how to minimize harmful impacts from ecological footprints.

One of the current concerns that humans are facing is an increase in global temperatures. Even a 1.5^{o} Celsius change in average global temperatures could lead to catastrophic results, which led 195 nations to come together in late 2015 to sign the Paris Agreement[1]. As part of the Paris Agreement, nations agreed to combat climate change by tackling known climate change contributors, such as greenhouse gases, or gases that trap heat in the atmosphere[2]. Of the many greenhouses gases, there is one atom that is known to be a problem: carbon.

Due to its atomic structure (4 valence electrons), carbon is rarely found in the atmosphere by itself; rather, the molecules that most concern scientists are carbon dioxide (CO_2) and methane (CH_4). These two gases are responsible for 92% of greenhouse gas emissions in the United States in 2017, with carbon dioxide alone accounting for 82% of emissions[2]. Given the large percentage of carbon emissions, it is important to understand where these emissions come from so that we are better able to reduce these emissions (also known as a carbon footprint).

Hypothesis

Can a country's carbon footprint be predicted by its other footprints (cropland footprint, grazing footprint, forest footprint, buildup land, and fish footprint)?

Data

The data set used was composed by the National Footprint Accounts, and measures the ecological resource use and resource capacity of nations from 1961 to 2016[3]. This data set is based on calculations of data from several United Nations data sets, including data sets published by the Food and Agriculture Organization, United Nations Commodity Trade Statistics Database, the United Nations Statistics Division, and the International Energy Agency.

The dataset consists of 72,186 observations from 1961-2016, with multiple observations per country per year. The 12 columns contain country, year, country_code, record, and columns with individual footprint measures for each country in global hectares (gha) per

person (crop_land, grazing_land, forest_land, fishing_ground, built_up_land, carbon, total). The last column, q score, is a data accuracy score assigned by the National Footprint Accounts that is based on the omission of unreliable data.

Data Processing

The first step in processing the data was to convert the categorical "record" column into land types with each respective footprint value. This was accomplished through the use of a pivot table and join. This was a necessary first step as each row needed to represent a single observation rather than having 8 rows for one observation. Once the table was altered, the dataset consisted of 9,024 observations with 59 columns.

The next step in processing the data was to address the missing values. Although the dataset now contained 9,024 observations, only 6,462 of these observations had non-null values. A closer look at the data revealed that the observations with null values were from countries that were listed but had no data due to the country's independent status. For example, 'Russia' had observations listed going back to 1961 but Russia did not exist as a country until December 1991, so any observations prior to 1991 were filled with null values. Similarly, the Soviet Union has data in the data set starting in 1961 and ending in 1991, with all observations after 1991 filled with null values. Due to this variation in country data, the observations with null values and missing data were deleted from the set, leaving 6,462 observations across 59 columns.

Once the missing values were addressed, the next step was to add columns for analysis. A 'population' column was created for each observation by dividing the built up land biocapacity total by the build up land biocapacity per capita. Total summative columns were created for each footprint type (crop land, grazing land, forest land, fishing ground, built up land, carbon, ecological footprint) by summing the different total parts columns (production, consumption, and biocapacity). Once these total columns were calculated, the total footprints per capita columns were also created by dividing the total footprint columns by the population. The end result of this transformation was a data set consisting of 6,462 observations with 74 columns.

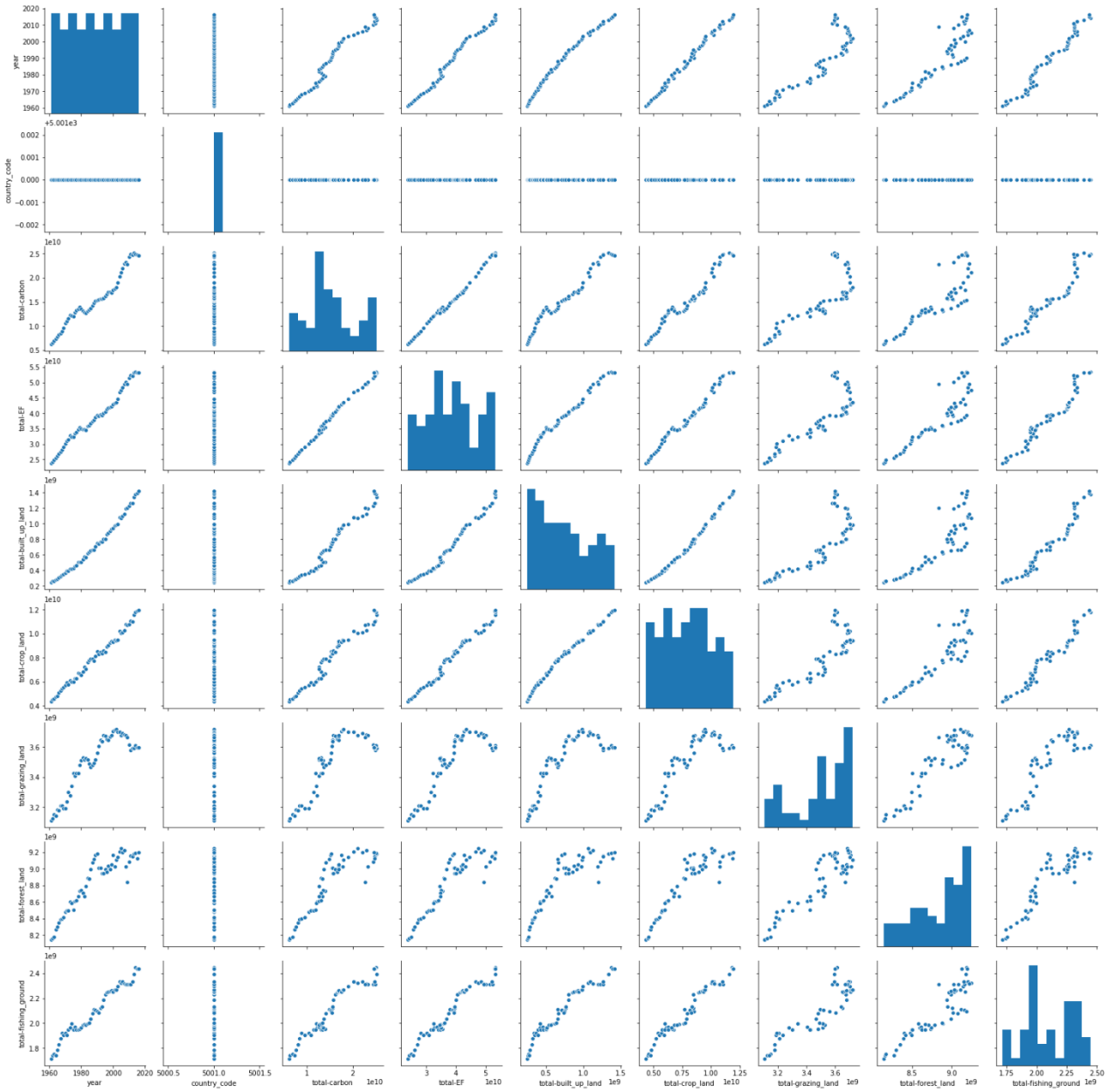
The last data processing step performed on this data set was to separate the world data from the country data. During an initial search for outliers, it was found that the data set contained summative data for the world for each year. Since these values were substantially larger than the country data for each year, the world data was pulled into its own dataset and analyzed separately from the country data (the remainder of the dataset).

Exploratory Data Analysis

World data

The first step in analyzing the world data was to establish what categories are correlated. To do this, total columns for each footprint type were selected and then charted via pair plots as seen in Figure 1 below.

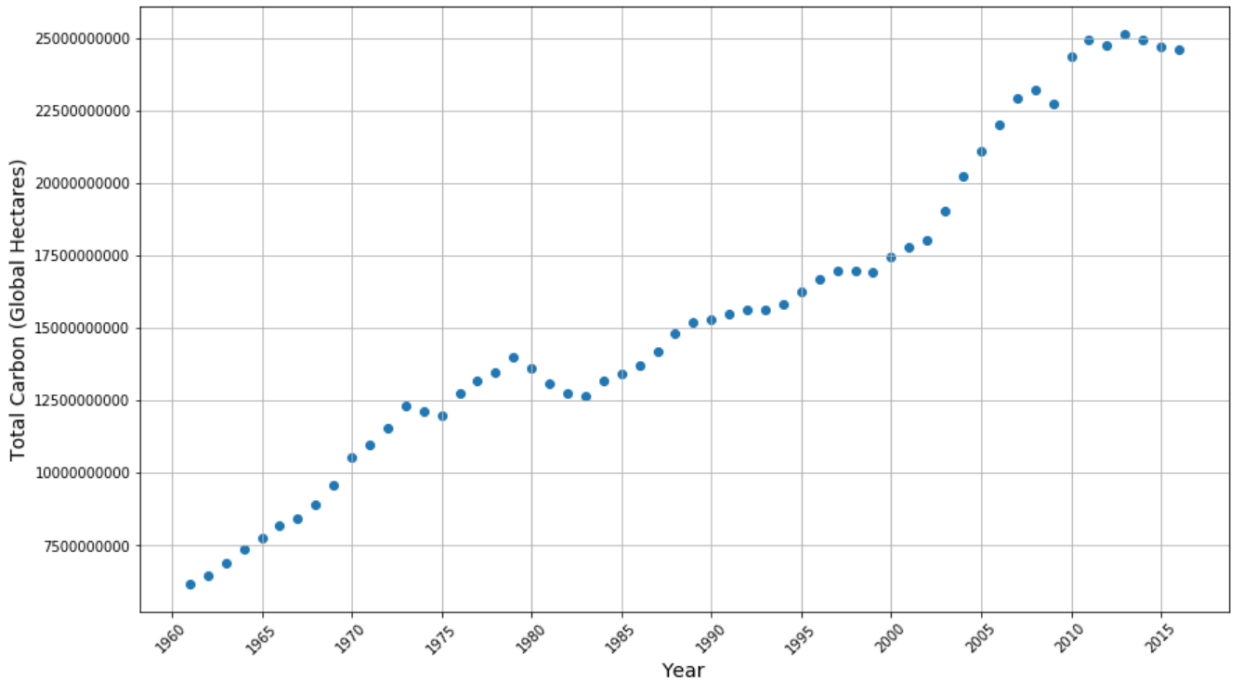
Figure 1: Pair plots of footprints for world data since 1961



As we see in the pair plots in Figure 1, all of the footprints are highly, positively correlated with each other. This makes sense as the total land area within countries does not change much but the uses for that land can change as consumer needs change. Interestingly, the types of land and total carbon all increased fairly linearly up until the last 20 years or so when the growth patterns started changing. Built up land and grazing land became negatively correlated as did grazing land and the total ecological footprint of a country.

With the correlations established above, the next step in analyzing the world data was to look at the total carbon footprint of the world by year, as shown in Figure 2 below.

Figure 2: Total World Carbon Footprint by Year (1961-2016)

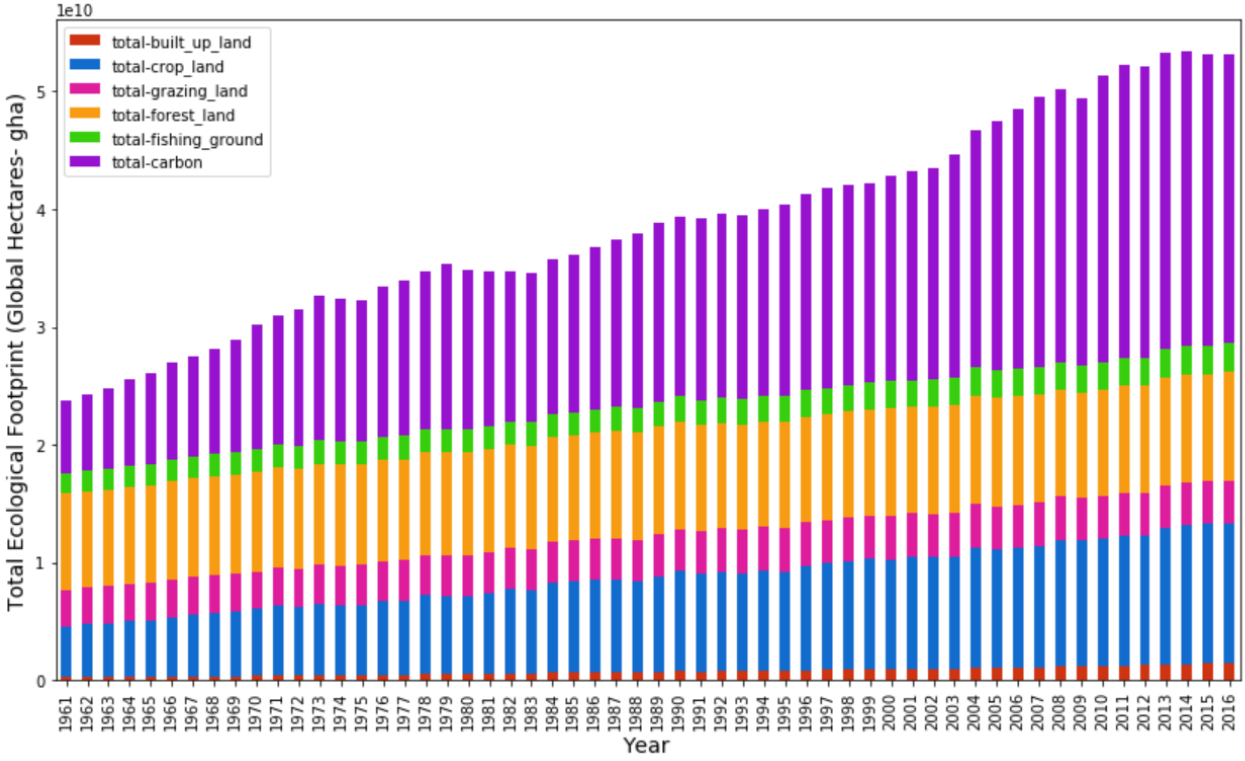


Looking at the total carbon footprint of the world starting in 1961, we see that the carbon footprint steadily increased throughout the 1960s with a brief drop in 1973 and 1979. Although these declines in carbon output may have been a response to the passing of several environmental regulation policies (such as the Clean Air Act in 1970 and the Clean Water Act in 1977 in the United States), there is not enough data in this particular data set to make the conclusion that the decline was directly a result of a change in environmental regulations.

Since the decline in the early 1980s, the total carbon footprint has continued to rise- sometimes sharply- until 2013. However, we see from the data that the carbon emissions

have increased from 6.18 x 10^9 to 2.45 x 10^{10} global hectares, an increase of 397.81%. Given this increase and the impact that carbon emissions have on the global temperature, the footprints that were positively correlated with carbon emissions were charted by year to determine how much of each footprint type contributes to the ecological footprint for the year, as seen in Figure 3.

Figure 3: World Ecological Footprint by Year (1961-2016)



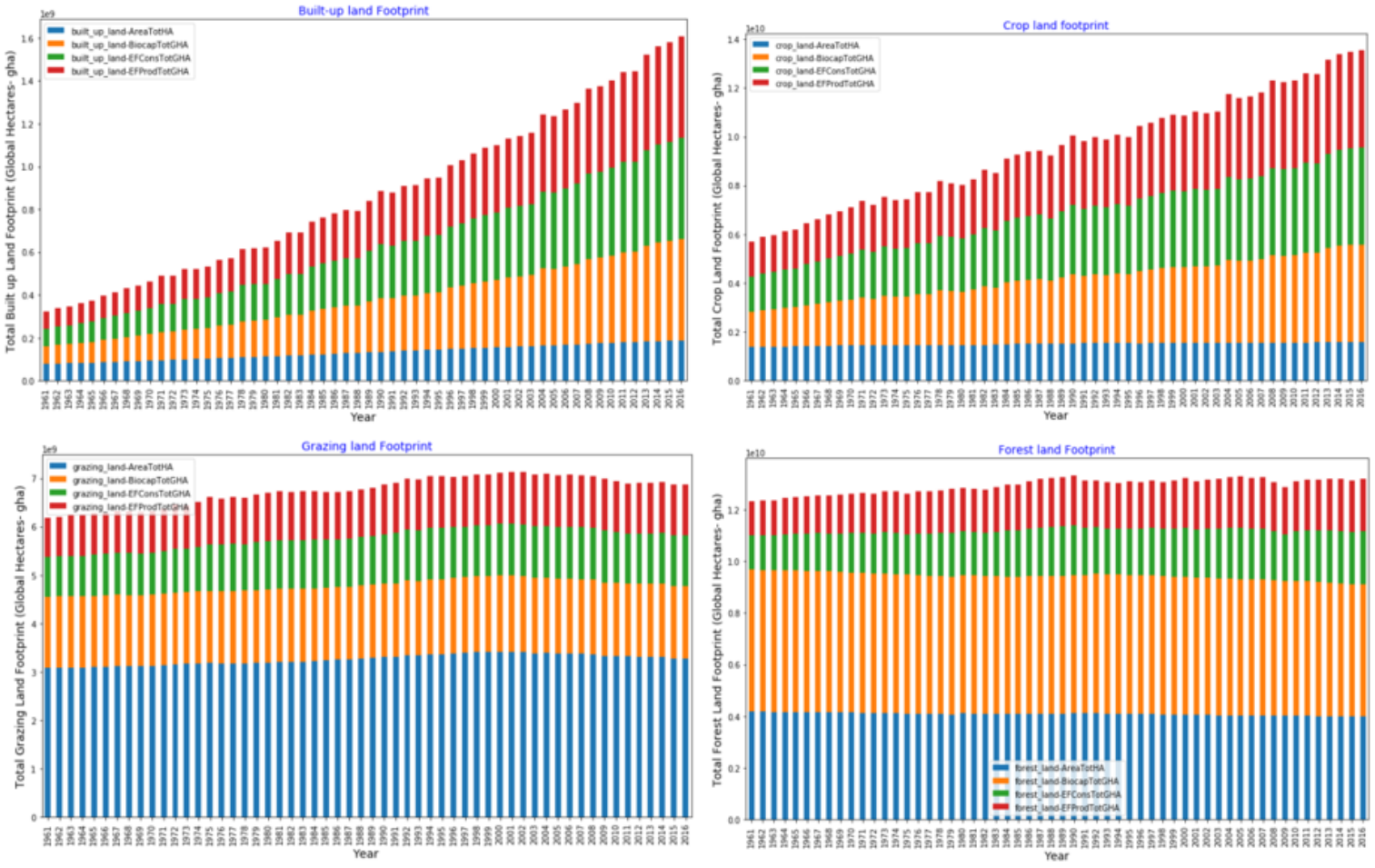
In Figure 3 above, we see that the total built up land (in red) has steadily but slowly increased while the total crop land has more than doubled in footprint size, which makes sense as the increasing world population needs increasing amounts of food. The total grazing land has remained approximately the same, with just slight increases as has the total forest land footprint and the total fishing ground footprint. Based on this, we can infer that the increase in carbon may be correlated with the increase in the crop land footprint and the built up land footprint.

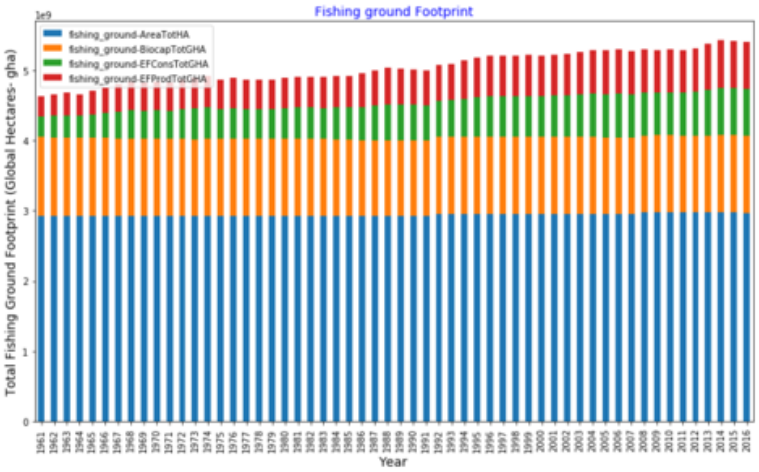
Each footprint- built up land, crop land, grazing land, forest land, fishing ground, and carbon- are composed of three different subcategories: biocapacity, consumption, and production. Since trade is global, the footprint for resources and manufactured goods is shared between countries. Countries that supply the raw resources have that part of the footprint count towards their production footprint in each land use category. Countries that import and purchase the resources or manufactured goods have the transport and use of those

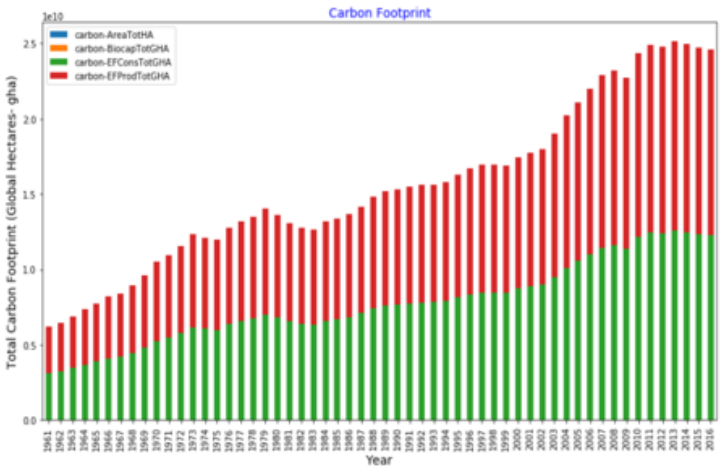
goods and resources count towards their consumption footprint. The biocapacity footprint for each land type is the amount of life that that area of that type of land can sustain- where life includes all living species from plants and animals to archaea.

To get a better idea of what each land type footprint was composed of, each land type footprint was next charted, broken down by land area, biocapacity, consumption, and production totals, shown in Figure 4. Note that the land area devoted to each footprint is also included to show fluctuations in area of each land type although the land area itself does not contribute to the total footprint for the land type.

Figure 4: Land Type Footprints by Year for World Data (1961-2016)







Looking at each subplot, the first thing that immediately jumps out is that the carbon footprint is only a result of production and consumption. This is because a carbon footprint is a measure of carbon emissions that occur as a result of production and consumption. Another thing to note from this subplot is that the production and consumption values are approximately equal although in the last five years of data, emissions from consumption are slightly higher than from production and we see that the production emissions actually decline slightly.

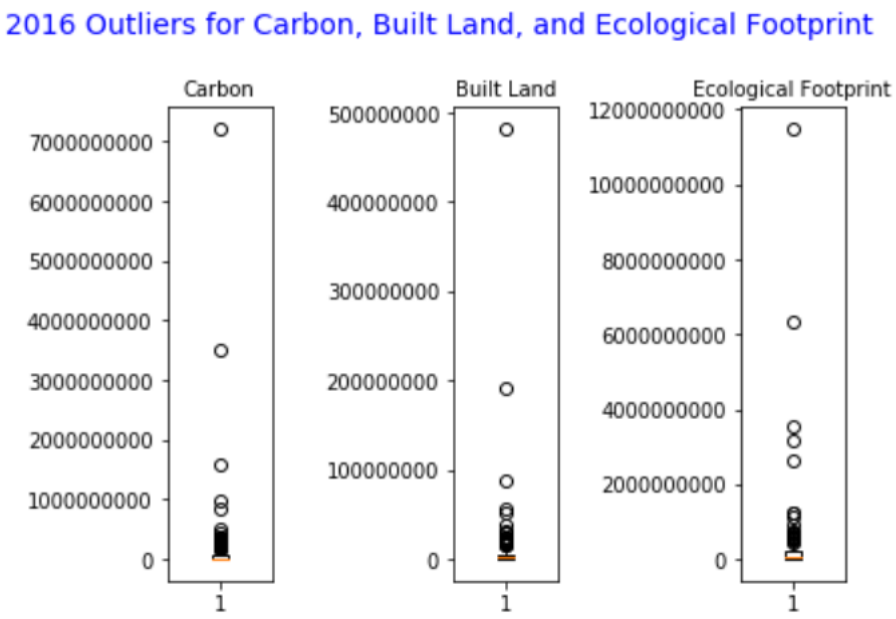
In the remaining five subplots, the blue represents the area devoted to each type of land footprint. We see that as the forest land is on a scale of 1 x 10^{10} that forest land has the largest amount of area followed by crop land, then grazing land, then fishing ground, and finally built up land. Of these land types, crop land is the only land type to have experienced a steady decline in area while simultaneously experiencing an increase in biocapacity (orange). Although forest land area has only experienced a marginal decline in area, it had a steep decrease in biocapacity. Opposite this, built up land increased in size and had a noticeable increase in biocapacity. The fishing ground footprint experienced more fishing resulting in an increased footprint but area and biocapacity remained relatively unchanged.

After looking at the world data and the relationships between each type of footprint and its components, the exploratory data shifted to looking at the data from the countries.

Countries data

The first step in analyzing the dataset for the countries is to identify outliers within the dataset. This was first done for the most recent year in the data (2016) to assess individual country outliers, as seen in Figure 5.

Figure 5: 2016 Outliers for Carbon, Built Land, and Ecological Footprint



As we see from the boxplots above, the median carbon footprint, built land footprint, and ecological footprint are all very low (median is indicated by the red line). There are very clearly outliers above even the third quartile, with three outliers having extreme footprints. Looking into the dataset, we see that the median carbon footprint for a country in 2016 was 15789598.8380 global hectares of carbon, or approximately 1.9637 global hectares of carbon per person, per year. However, the country with the most extreme values, China, had a carbon output of 7206757941.00 hectares, or approximately 25.3086 global hectares per person- an increase of 12.88 times the carbon footprint of the median value.

A look into the top ten countries provided a glimpse into the outliers for these footprints in Figure 6:

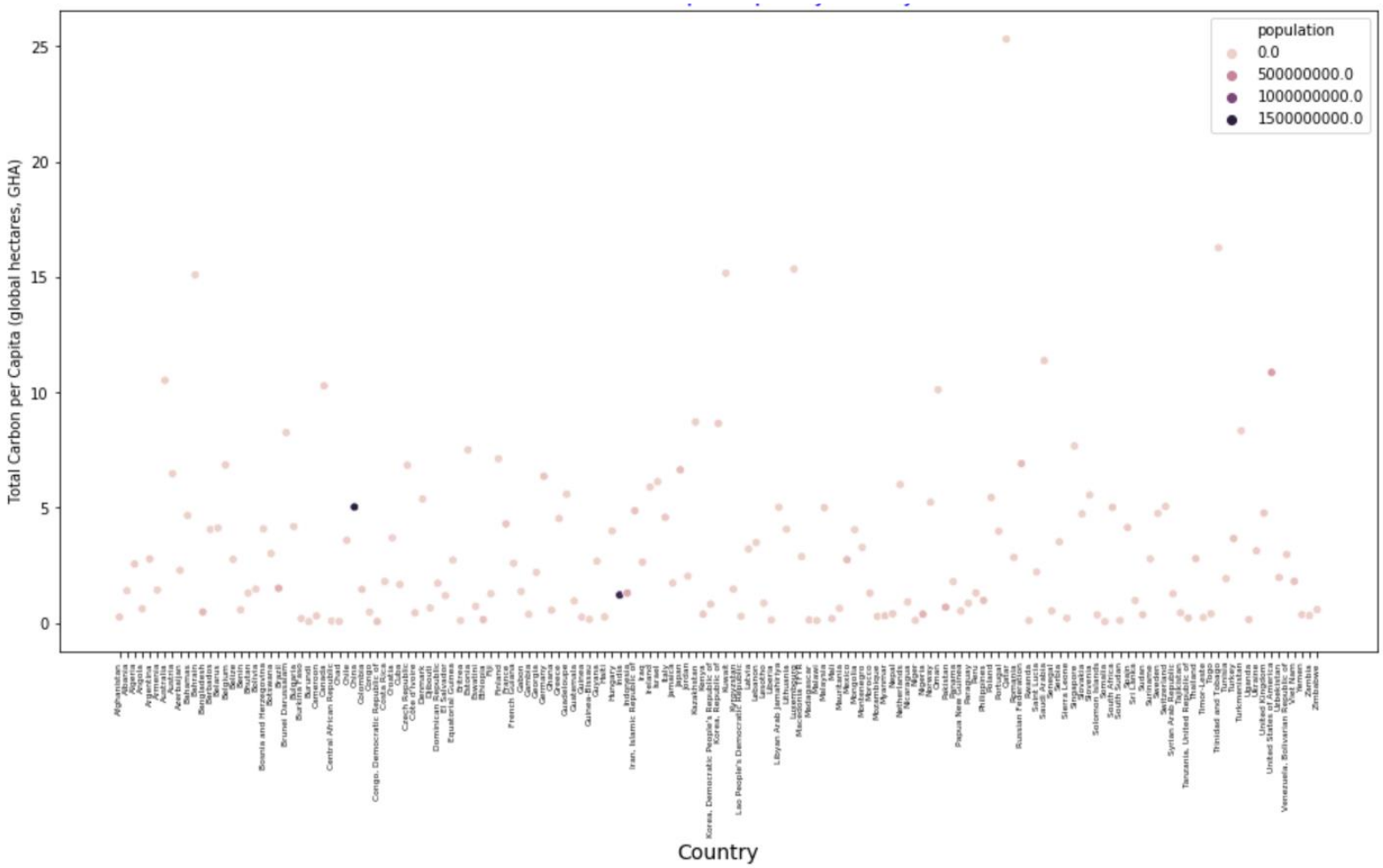
Figure 6: 2016 Top 10 Outliers for Carbon Emissions

	country	year	country_code	built_up_land-AreaPerCap	built_up_land-AreaTotHA	built_up_land-BiocapPerCap	built_up_land-BiocapTotGHA	built_up_land-EFConsPerCap	built_up_land-EFConsTotGHA	built_up_land-EFProdPerCap	...
1783	China	2016	351	0.0221	31691300.7800	0.1119	160538830.6000	0.1119	160538830.6000	0.1119	...
8578	United States of America	2016	231	0.0272	8761080.0780	0.0907	29215883.2200	0.0907	29215883.2200	0.0907	...
3955	India	2016	100	0.0187	24802199.2200	0.0479	63470660.5800	0.0479	63470660.5800	0.0479	...
6804	Russian Federation	2016	185	0.0334	4810729.9800	0.0418	6014960.1840	0.0418	6014960.1840	0.0418	...
4376	Japan	2016	110	0.0193	2470149.9020	0.0542	6930153.6600	0.0542	6930153.6600	0.0542	...
3327	Germany	2016	79	0.0369	3020124.8740	0.1328	10878955.9000	0.1328	10878955.9000	0.1328	...
4625	Korea, Republic of	2016	117	0.0148	749236.0229	0.0588	2987266.0920	0.0588	2987266.0920	0.0588	...
4067	Iran, Islamic Republic of	2016	102	0.0415	3334580.0780	0.0700	5617958.8500	0.0700	5617958.8500	0.0700	...
1538	Canada	2016	33	0.0292	1061449.9510	0.0695	2523558.4810	0.0695	2523558.4810	0.0695	...
7120	Saudi Arabia	2016	194	0.0548	1767329.9560	0.0348	1124531.4590	0.0348	1124531.4590	0.0348	...

Looking at the results above for carbon footprint, it's easy to see why it was important to separate out one year of data; because the top three outliers are so far above the other countries, a list of the top 10 carbon footprints from the full data set would yield the same 3 countries over and over. Looking at the total carbon footprint from these top 10 countries reveals that the top carbon emissions come from China at a rate of more than double the emissions from the second highest country, the United States. However, the carbon footprint totals do not tell the entire story. Once population is factored into the total (resulting column is total carbon per capita), we see that the United States has the highest carbon emissions per capita at 10.8560 global hectares per person while China has less than half that at 5.0222 global hectares per person. Comparatively, India, which has the third highest carbon footprint, has the *lowest* carbon footprint per capita of the top ten outliers at just 1.2040 global hectares per person.

Using the above metrics, it becomes clear that population contributes to total carbon footprint but that population does not tell the entire story; after all, the United States had a very small population in 2016 compared to both India and China yet still ranked second in the world for carbon emissions. To look into this further, a scatterplot of total carbon per capita was charted against country using color to show population size in Figure 7.

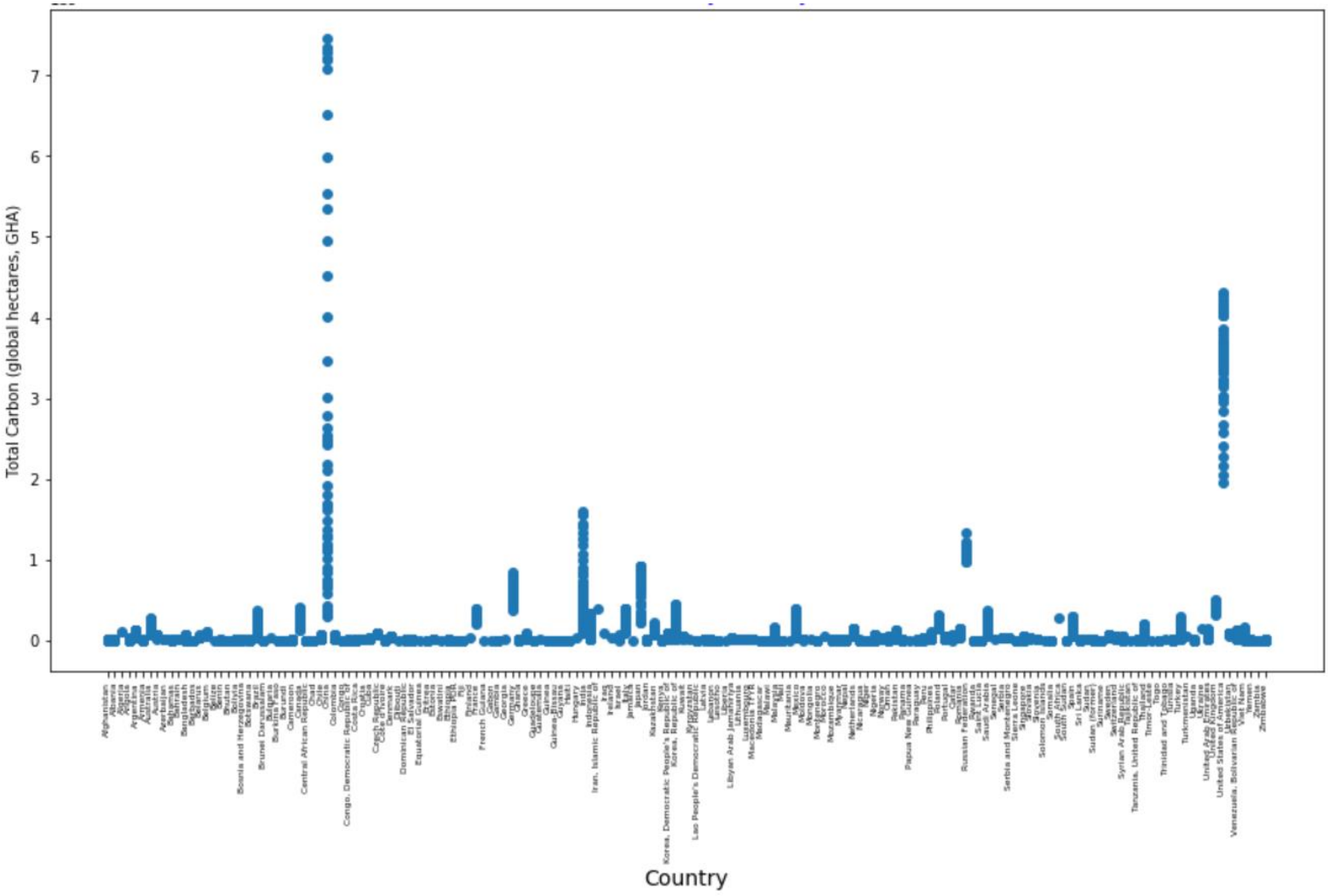
Figure 7: Total Carbon Emissions per Capita by Country, 2016



Looking at Figure 7 above, we see that the two countries with the largest populations- China and India- have significantly lower carbon footprints per person than countries that have much smaller populations. So what countries ranked in the top 10 carbon emissions per person? A quick resorting of the data revealed the ten countries with the highest carbon emissions per person to be (in order from greatest to least): Qatar, Trinidad and tobago, Luxembourg, Kuwait, Bahrain, Saudi Arabia, United States of America, Australia, Canada, and Oman.

After looking at the data for 2016, the next step in exploring the data was to chart the total carbon emissions for the full countries data set (Figure 8).

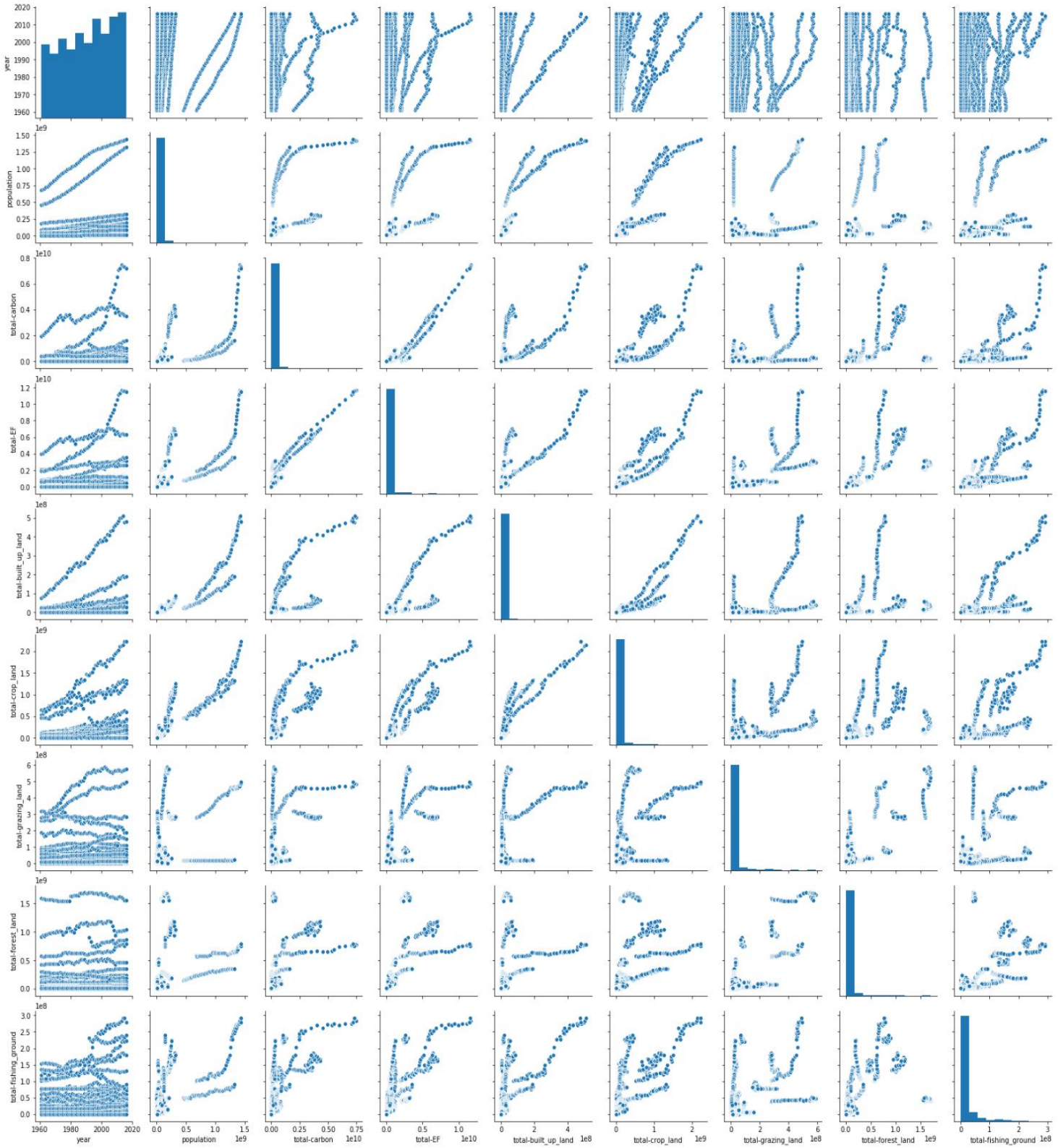
Figure 8: Total Carbon Emissions by Country since 1961



In the chart above, we see that the outliers have consistently been outliers since the establishment of their country (if occurred during the period of this dataset) or have been outliers since the beginning of the data collection (1961). There are large gaps in some of the data, such as the data for China, where the amount of total carbon emissions significantly increased from the prior year. Other gaps in the data are a result of missing data/no data for that country for that year.

The last step in analyzing the 2016 data was to use the total columns for each footprint category to create pair plots showing the correlation between each set of variables (Figure 9).

Figure 9: Pair Plots for Countries Data



Looking at the pair plots above, we see there is a very strong positive correlation between total Ecological Footprint and total carbon footprint as well as between total crop land and total built up land. Interestingly, we see that the relationships between population and the footprints are not linear but rather taper off after quickly increasing; this same type of correlation pattern is found between total carbon and built up land, total carbon and crop land, and across all variables interacting with total grazing land.

Predictive Models

Since there was a high correlation found between 'total-EF' and 'total-carbon', 'total-EF' was removed from the data set. The columns 'carbon-BiocapTotGHA', 'carbon-EFConsTotGHA', and 'carbon-EFProdTotGHA' were also removed as these features created the 'total-carbon' column when summed together. After removing these features, the country column was transformed into binomial values using one hot encoding. Once these two data transformations were complete, the resulting data set contained 6,369 observations across 232 columns.

Given that the goal of the model was to predict total carbon, three different regressions were used to make the prediction: multiple linear regression, Ridge regression, and Random Forest regression. Each regression was run against a training set of data as well as a testing set of data. After running each regression, the R^2 value, feature coefficients, mean squared error, mean absolute error, and root mean squared error were calculated for each model. Of the three models used, the multiple linear regression and the Random Forest regression each produced an R^2 value of 1.00 while the Ridge regression produced a value of 0.99. While at first glance these appear to be accurate models, the error and coefficient values indicate otherwise. Both the Ridge regression and the Random Forest regression produced sizable errors (for all error types calculated) while the multiple regression model produced very small, nearly zero errors and coefficients.

Given the variation in error and coefficient size, a Variance Inflation Factor (VIF) test was run against each feature used in the data set. The results of the VIF were astounding: every factor (excluding year) was highly correlated with every other factor. In an attempt to combat the multicollinearity, the data was then scaled and run through Principal Component Analysis (PCA) to reduce the features and their effect on each other. Unfortunately, once the models were run with the scaled data, the models became ineffective at calculating the total carbon. As a result, the scaling and PCA were removed and the models were left as they originally were.

Conclusion and Recommendations

From the above models, it is clear that the features in this data set have a very high degree of multicollinearity; therefore the amount of total carbon produced by a country could not be realistically predicted using this dataset. Although scaling, PCA, and feature removal were all used to try to reduce the multicollinearity, the data set used in this analysis was simply too correlated to reach a conclusion about which features can be used to predict total carbon.

Although initially unexpected, the high levels of correlation within the data make sense. For example, as forested land decreases, it stands to reason that carbon production and built up land would increase. Similarly, as population increased, built land and crop land increased in area.

Going forward, I would recommend two different approaches to using this data. One of the ways that is suggested to determine total carbon output for a country would be to include additional outside data; more detailed data with additional features could potentially reduce the multicollinearity within the dataset and allow for a more accurate prediction. The other suggestion would be to analyze the highest carbon emission countries separately; the extreme outliers found in this dataset are not helpful when running regressions where the majority of countries have exceptionally low carbon emissions. However, without one or both of these suggestions being implemented, it is not recommended that countries use only this data to predict their carbon emissions.

Acknowledgements

I would like to acknowledge and thank my mentor, Devin Cavagnaro, for his guidance in analysis and predictive modeling on this project.

References

[1] "What Is the Paris Agreement?" *UNFCCC*, United Nations Climate Change, unfccc.int/process-and-meetings/the-paris-agreement/what-is-the-paris-agreement.

[2] "Overview of Greenhouse Gases." *EPA*, Environmental Protection Agency, 11 Apr. 2019, www.epa.gov/ghgemissions/overview-greenhouse-gases.

[3] "NFA 2019 Edition - Dataset by Footprint." *NFA 2019 Edition*, Data.world, 2 July 2019, data.world/footprint/nfa-2019-edition.