Global CO2 Emissions: Prediction and Mitigation

By Olia Javidi, Samantha Lee, and Katie Li

The Problem

With rising CO2 emissions across the globe, it's more important now than ever to try to reduce our global emission levels. By 2050 the world civilization will be at risk with flooded cities, forced migration, and the Amazon turning into the savannah (Watts, 2019). In our project, we aim to identify countries with the highest CO2 emissions or increasing trends in CO2 emissions in order to make recommendations based on the factors that correlate with higher emission levels.

The Data Set

To answer these questions, we used two datasets. Our initial dataset was organized so that a row consisted of a country, an indicator name, a year, and a value **(Table A)**. We thought it would be more helpful for the dataset to use indicator names as columns with the values being the entry. The first was a dataset of 5,656,458 rows. We then used pandas to transpose this dataset into the workable version we envisioned by taking the entries for a given IndicatorName and left joining based on CountryName and Year **(Table B)**.

Table A

	CountryName	CountryCode	IndicatorName	IndicatorCode	Year	Value
0	Arab World	ARB	Adolescent fertility rate (births per 1,000 wo...	SP.ADO.TFRT	1960	1.335609e+02
1	Arab World	ARB	Age dependency ratio (% of working-age populat...	SP.POP.DPND	1960	8.779760e+01
2	Arab World	ARB	Age dependency ratio, old (% of working-age po...	SP.POP.DPND.OL	1960	6.634579e+00

Table B

	CountryName	CountryCode	Year	Adolescent fertility rate (births per 1,000 women ages 15-19)	Population density (people per sq. km of land area)	GDP at market prices (current US$)	CO2 emissions (metric tons per capita)	Arms exports (SIPRI trend indicator values)	Birth rate, crude (per 1,000 people)	Death rate, crude (per 1,000 people)
0	Arab World	ARB	2000	53.829472	20.651464	7.307833e+11	3.717403	NaN	28.282837	6.465841
1	Caribbean small states	CSS	2000	71.619280	15.943475	3.220670e+10	6.617366	NaN	20.044584	7.146089

From there, we noticed our data was organized such that each year, the first 33 entries are regions (i.e. Arab World, East Asia, European Union) or descriptors (i.e. low and middle income, fragile and conflict affected situations, heavily indebted poor countries) and the next 194 entries for that year are countries. For this reason we split our dataset into a regions/characteristics set and a countries set. This will allow us to help understand what's happening geographically, but keep our data from being interdependent (because the country data is represented in the region data).

We first built the regional dataset. There were 1,348 features once we removed entries that had a NaN value for our target column - the CO2 emissions (metric tons per capita). From there we built models and evaluated the mean squared error with different thresholds of NaN values, ultimately getting the lowest error when we required that there was a minimum of 28 real values. We felt this was low, but agreed that in our final feature selection if it was made off of features with primarily imputed values, we would change our threshold approach. With this metric in place we had 771 features with a total of 90,021 NaN values out of 252,888 total values **(Table C)**.

Table C

	Original	Filtered 1990 - 2011, transposed	Filtered on region	Threshold of NaN	Final - with random forest
Rows x Columns	5,656,458 × 6	4994 × 1349	330 × 1349	328 x 771	328 x 23
NaN values	N/A	N/A	274,433	90,021	0

(original structure didn't provide NaN for missing values, just omitted those rows)

The Data Set: CO2 Emissions (Metric Tons Per Capita)

To figure out what areas in the world have the fastest and highest rising CO2 emission levels, we first focused on analyzing the regions and then proceeded to look into the countries within each region. We decided to use CO2 emissions (metric tons per capita) to analyze the data trends for each region and country. In the Indicators file, there are six columns: CountryName, CountryCode, IndicatorName, IndicatorCode, Year, and Value.

Table D

	CountryName	CountryCode	IndicatorName	IndicatorCode	Year	Value
0	Arab World	ARB	CO2 emissions (metric tons per capita)	EN.ATM.CO2E.PC	1990	3.203907
1	Caribbean small states	CSS	CO2 emissions (metric tons per capita)	EN.ATM.CO2E.PC	1990	5.367886
2	Central Europe and the Baltics	CEB	CO2 emissions (metric tons per capita)	EN.ATM.CO2E.PC	1990	8.847908
3	East Asia & Pacific (all income levels)	EAS	CO2 emissions (metric tons per capita)	EN.ATM.CO2E.PC	1990	2.600991
4	East Asia & Pacific (developing only)	EAP	CO2 emissions (metric tons per capita)	EN.ATM.CO2E.PC	1990	1.803359

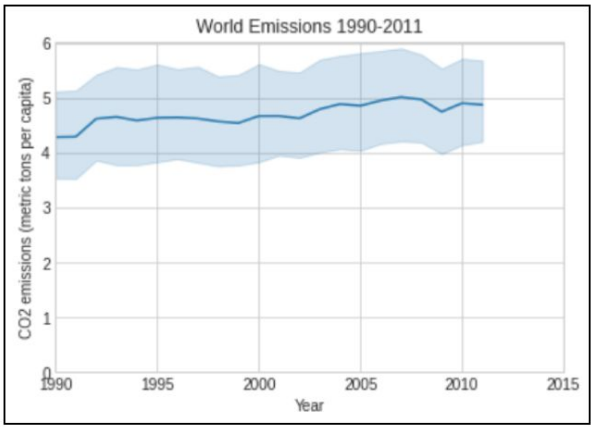
(table after filtering out all entries for CO2 emissions that are from 1990-2011 and grouping by CountryName)

We first filtered to values between 1990 to 2011 (Table D). This reduced the data subset from 5,656,458 rows down to 3,776,056 rows. Next, we filtered based on the indicator being CO2 emissions (metric tons per capita). This filtered the data subset even further to 4929 rows. After selecting distinct values from CountryName, we found that there were 247 rows of data. Then, we grouped by region from the Country file and ordered them in descending order of number of countries. We found that there were 7 regions: Europe &

Table E

	Region	Count
0	Europe & Central Asia	57
1	Sub-Saharan Africa	48
2	Latin America & Caribbean	41
3	East Asia & Pacific	36
4		33
5	Middle East & North Africa	21
6	South Asia	8
7	North America	3

Figure A



Central Asia (57 countries), Sub-Saharan Africa (48 countries), Latin America & Caribbean (41 countries), East Asia & Pacific (36 countries), Middle East & North Africa (21 countries), South Asia (8 countries), and North America (3 countries) (**Table E**). We noticed that there were 33 rows of data that did not belong to a region, and upon closer inspection we found that they were correlated with general regional categories, such as Arab World, Caribbean small states, East Asia & Pacific (developing only), European Union, and High Income. These rows were then filtered out of the dataset. To visualize global trends, we plotted world emissions and found that CO2 emissions are rising, albeit steadily (**Figure A**). To seek out which regions were the highest emitters of CO2, we plotted line plots for each of the 7 regions. Further, we plotted the countries within each region, and then found the following countries as the top emitters of CO2 within each region (**Table F**).

Table F

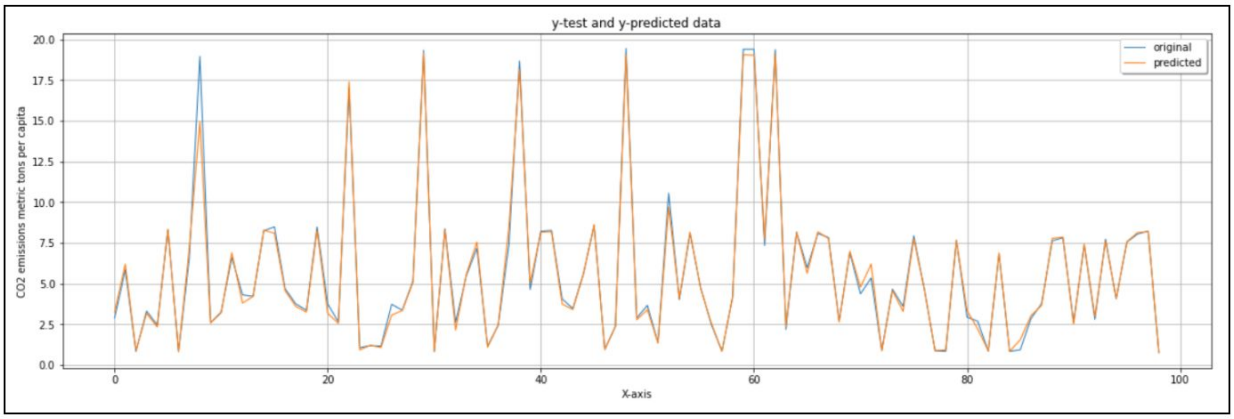
Region	Top 3 Countries (descending order by metric tons per capita)
Europe & Central Asia	Luxembourg, Kazakhstan, Estonia
Sub-Saharan Africa	South Africa, Equatorial Guinea, Seychelles
Latin America	Trinidad and Tobago, Aruba, Cayman Islands
East Asia & Pacific	Australia, New Caledonia, Palau
Middle East & North Africa	Qatar, Kuwait, United Arab Emirates
South Asia	Maldives, India, Pakistan
North America	United States, Canada, Bermuda

How you tried to solve the problem

For our midterm report, we selected an initial feature pool to work with based on intuition, and then examined their individual correlation plots. We were advised that this could definitely be an inputting of bias and result in overfitting. For this reason, this time we ran a random forest regression on the features to determine which ones we wanted to consider.

Because we still had many NaN values even with the threshold, we imputed the NaN values and filled in the column mean in its place. With this feature selection using random forest regression, our R-squared score using those features hovered at 0.9977 (**Figure B**).

Figure B



Above, our x-axis is different indices in y-test, with the orange line being the predicted value of that specific set of features. This displays that our model with all the features has the potential to make really good predictions on our dataset with the model we eventually choose.

We wanted to reduce the number of features from 328 because if we wanted to run predictions on a single region or country we would have more features than entries if we kept all 328 features. Using our random forest regression results, we selected 23 features out of the top 60 that were not interdependent (i.e. rural population percentage and urban population percentage add up to 100). We also left out features like land under cereal production and rural population because they were not adjusted for population like our target feature (CO2 emissions metric tons per capita).

This left us with Energy use (kg per capita), employment to population ratio, lifetime risk of maternal death ratio, % of renewable energy consumption, age dependency ratio, % of male unemployment, % of forest area, GDP per capita, fossil fuel energy consumption (% of total), employment to population ratio, death rate, % of population ages 65 and above, % of area that's terrestrial & marine protected areas, % of total energy that is alternative or nuclear energy, % of arable land, fixed telephone subscriptions per 100 people, changes in inventories, military expenditure, % of population with access to improved sanitation facilities, female life expectancy at birth, % of total merchandise exports to developing economies in Middle East & North Africa, agriculture value added per worker, population density per sq. km

After using this process of refining our region dataset to determine the features for our model, we then ran this same process of filtration, transposing, and imputing on our country data set limiting it to the chosen features for algorithmic use.

Algorithms used

By knowing the 23 features that are important to our dataset, we would be able to create a linear regression model without personal bias in influencing which features we selected. However, this doesn't tell us which and how many of the 23 features will create an optimal linear regression model. In order for us to create the best possible linear regression model, we employed best subset selection along with cross validation. Best subset selection tests all permutations of models for each possible number of predictors {1,2,...,p} and it selects the best model (largest R^2 value) for each k in {1,2,...,p}. For example, in the case of our 23 possible features, let's say that we want to make a linear model that only uses 3 features. In this case, that gives us ${}_{23}C_3$ or 1771 combinations of potential models, with best subset selection choosing the best combination. In the case with our 23 features, there are 23 models provided with best subset selection (ISLR Ch.6).

We started by comparing the Adjusted R^2 values from best subset selection and found that the values plateaued around the 15-feature model, with very small fluctuation (**Figure C**). Although this gave us an even better sense of what features are important to our data, we needed to train, test, and validate to understand how our 23 models would perform on new data. To achieve this, we used 10-fold cross validation to understand how our models would perform on new data. After performing 10-fold cross validation, we found that our 21-feature model had the lowest error. To confirm that we had not overfit or underfit our model, we compared our training MSE with our CV Error (**Figure D**).

Figure C

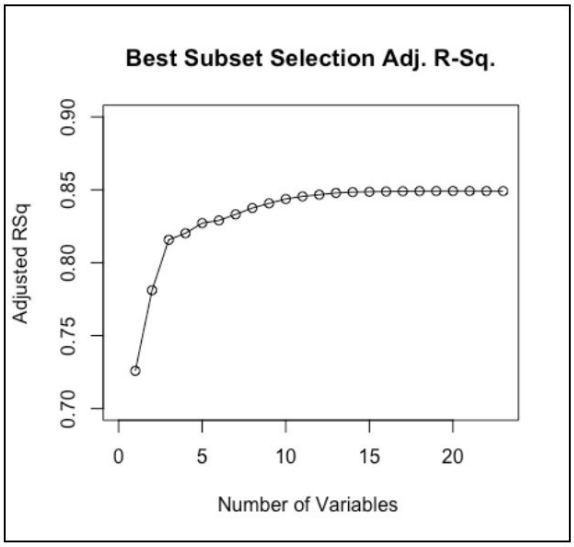


Figure D

This revealed that our training error is slightly smaller than the CV error for all 23 models. Specifically, for the 21-feature model we have the following metrics:

- Training MSE: 6.77

- Mean CV MSE: 6.96
- Training Adjusted R^2: 0.857
- Mean CV Adjusted R^2: 0.849

With this knowledge, we can confidently choose the 21-feature best subset model. The features used with this model along with their coefficients are in **Table G**.

Table G.

Feature	Coeff.
`Population ages 65 and above (% of total)`	1.57E+00
`Unemployment, male (% of male labor force)`	3.31E-02
`Improved sanitation facilities (% of population with access)`	2.41E-02
`Employment to population ratio, 15+, total (%) (modeled ILO estimate)`	2.39E-02
`Terrestrial and marine protected areas (% of total territorial area)`	2.26E-02
`Life expectancy at birth, female (years)`	1.61E-02
`Employment to population ratio, 15+, male (%) (modeled ILO estimate)`	1.13E-02
`Energy use (kg of oil equivalent per capita)`	2.04E-03
`GDP per capita, PPP (constant 2011 international $)`	5.44E-05
`Changes in inventories (current LCU)`	-3.29E-15
`Agriculture value added per worker (constant 2005 US$)`	-4.24E-05
`Lifetime risk of maternal death (1 in: rate varies by country)`	-1.06E-04
`Population density (people per sq. km of land area)`	-4.15E-04
`Fixed telephone subscriptions (per 100 people)`	-6.68E-03
`Arable land (% of land area)`	-1.02E-02
`Merchandise exports to developing economies in Middle East & North Africa (% of total merchandise exports)`	-1.67E-02
`Forest area (% of land area)`	-1.96E-02
`Fossil fuel energy consumption (% of total)`	-2.97E-02
`Renewable energy consumption (% of total final energy consumption)`	-3.46E-02
(Intercept)	-1.04E-01
`Alternative and nuclear energy (% of total energy use)`	-1.20E-01
`Age dependency ratio, old (% of working-age population)`	-1.03E+00

Results

We found that amongst all of the countries in the world, the three countries that emitted the most CO2 (metric tons per capita) during 1990-2011 were Qatar (~44 metric tons per capita), Trinidad and Tobago (~35-40 metric tons per capita), Kuwait (~30 metric tons per capita), Aruba (~22 metric tons per capita), and Luxembourg (~20-22 metric tons per capita). We found that Maldives had the fastest growing rate of CO2 compared to its region.

Therefore we should make recommendations to these countries that align with our most important features to our model -- the age dependency ratio (population age 65 & above), renewable energy consumption percentage, and unemployment ratio. For age dependency ratio, higher percentages of populations age 65+ aligned with higher CO2 emission levels, so potentially recommending policies (i.e. better retirement benefits) that ease the financial burden that comes with caring for elders would be helpful in increasing financial freedom of the working class. In renewable energy consumption percentage, the recommendation would just be aligned with switching over to or increasing use of renewable energy resources either through government incentives or industry incentives. The unemployment ratio is also positively correlated with CO2 emissions per capita so policies targeting lowered unemployment could help to lower CO2 emissions.

Confidence

We are confident in our results. We have shown through our low training and CV errors and high training and CV Adjusted R^2 values that our model performs well for all the countries and years in our data set. As with all data analysis projects, we feel that if we had a more complete dataset with fewer missing values, then we could have created an even better model. Additionally, if we wanted to create a model specifically for each country, then our model would likely employ different features between the different models which could make the models even more accurate. This, however, would not have been possible because there would only one datapoint per year per country, which would give us 21 data points and 328 features per country on our 1990-2011 dataset.

Next Steps

After evaluating our results and analyzing trends in our data, we would be willing to use them in recommendations to change how countries make decisions on how to decrease their CO2 emission levels. However, a one-size fits all policy would not work as policies must take into account the area's context before recommendations. For example, for a country that has a low percentage of improved sanitation facilities in urban areas, it would not be as important to a government to focus on their terrestrial and marine protected areas because they should be focusing on increasing the standard of living for their denizens. In 2015, countries signed the Paris Agreement, a landmark environmental agreement created by the United Nations to address climate change and its negative impacts. However, global carbon emissions surged the year after it was signed, with energy-related emissions increasing 1.4% to 32.5 gigatons in 2017, the equivalent of adding 170 million cars to the road. The Paris Agreement may have brought together all of the nations in the world to settle on a target, but the question still remains of how governments will enforce effective policies. Nonetheless, prioritizing a plan to implement more renewable energy infrastructure in all countries by 2050 is crucial for preventing disastrous environmental effects ("Paris Climate Agreement Countries", 2020).

Weapon of math destruction

On an individual country level or an individual city level, this model has potential to become a weapon of math destruction. For example, if we advised cities that plant species had nothing to do with CO2 emissions, and countries/cities began cutting down all of their trees without hazard because our model said they weren't correlated, it could cause CO2 emissions to increase because they started destroying CO2 consumers. The model didn't show trees were important because many of them exist and are protected/conserved to begin with. However, if we told cities not to care about them, it would undermine our model and prior assumptions, creating higher CO2 emissions.

Fairness

When making recommendations to various countries on the actions they should take to minimize their per capita CO2 emissions, it is important to understand metrics that can influence the quality of life of its citizens. For example, it would be hard for a country to justify increasing their investments in nuclear and alternative energy if their unemployment rates for men aged 15+ is high, because reducing unemployment would be the priority for the country, especially if jobs created with nuclear and alternative energy required skilled labor that was not represented in unemployed populations. Additionally, in countries that have high life expectancy at birth for females and high percentages of improved sanitation facilities (both features with positive coefficients) we would not recommend that they take measures to reduce life expectancy or decrease sanitation in order to decrease CO2 emissions per capita because both are metrics of good quality of life for its citizens.

We can use the country Luxembourg as an example of how we might make a recommendation on how to reduce per capita CO2 emissions. Luxembourg's per capita CO2 emissions are almost 5x the world average. When looking at the features in our model, we see that Luxembourg has an elderly population percentage double the global average and that renewable and alternative/nuclear energy only make up 4% of the country's energy supply (compared to 45% on average globally). Elderly population percentage is positively correlated with per capita CO2 emissions and renewable and alternative/nuclear energy are negatively correlated with per capita CO2 emissions. Because Luxembourg is a wealthy country, we could recommend that they invest more into renewable and alternative/nuclear energy sources, including hydropower on the Moselle or Sûre Rivers or nuclear energy. This would also have the effect of the country needing a larger population of young people to work in these energy systems and thus encouraging immigration or incentivizing working-age citizens of Luxembourg to stay in the country.

Bibliography

Watts, Jonathan. "The Environment in 2050: Flooded Cities, Forced Migration – and the Amazon Turning to Savannah." *The Guardian*, Guardian News and Media, 30 Dec. 2019, www.theguardian.com/environment/2019/dec/30/environment-2050-flooded-cities-forced-migration-amazon-turning-savannah.

"Paris Climate Agreement Countries 2020." *World Population Review*, 2020, worldpopulationreview.com/country-rankings/paris-climate-agreement-countries.

Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Best subset selection algorithm from the ISLR 7th edition textbook Chapter 6.