Renewable Energy and Greenhouse Gas Emissions in Europe

Matthew Harrison
Roderick Fan
Chengxi Wu
Ze Yu

DSCI 310
8/12/2022

Executive Summary

The European Union has stated a goal of being climate-neutral with net zero greenhouse gas emissions by 2050[1] to combat global climate change. Central to this goal is the reduction of power generated by fossil fuels such as coal, oil, and gas and increasing the deployment of renewable energy sources such as solar and wind. Our goal was to use machine learning techniques to observe the progression of the energy sector in Europe and predict if this goal will be met.

Our first dataset[2] contained the annual electricity generated measured in Terawatt hours in 27 European countries from different fossil fuel and renewable energy sources from 2000 to 2020. Visualization and calculations performed on this dataset showed a trend of decreasing energy generation from fossil fuels and an increase in energy generated by wind and solar production, especially between 2010 and 2020. The average share per country of energy generated from wind power increased from 0.9% in 2000 to 14.5% in 2020; for solar power, it increased from 0 in 2000 to 4.0% in 2020. However, some countries were more dependent on certain energy sources than others, for example, Sweden with hydropower and Malta with gas, and the dataset had a limited number of data points for a time series analysis and machine learning modeling.

We obtained a second dataset[3] to model carbon dioxide emissions in European countries. This dataset included the annual emissions of carbon dioxide measured in metric tons for 23 European countries from 1900 to 2020 and included 2,754 rows with 9 attributes. The attributes included country, year, annual carbon dioxide emissions per capita, and annual emissions from gas, oil, cement, coal, flaring, and an "other" emissions category. This dataset was cleaned from a larger dataset that contained the annual carbon dioxide emissions for countries around the globe with historical data dating back to the 1700s. It was found that annual emissions from gas, oil, cement, and coal shared the strongest correlation with total annual carbon dioxide emissions with a Pearson correlation between 65-87%. A linear regression model was built to predict the total annual carbon dioxide emissions for a country in a given year based on these four attributes. Cross-validation was used in this model with holdout data being a 20 year period and training data being the remaining 100 years between 1900-2020. On average, the regression model performed with 99% accuracy on the holdout data.

In conclusion, we were able to build a machine learning model with linear regression wherein you can predict the total annual carbon dioxide emissions for a country in a given year based on the emissions from the key attributes of gas, coal, oil, and cement. This will be useful for European governments and private companies in the energy sector to set emissions goals in the coming years in order to meet the target of carbon neutrality by 2050.

---

[1]https://ec.europa.eu/clima/eu-action/climate-strategies-targets/2050-long-term-strategy_en
[2]https://data.world/makeovermonday/2021w5/workspace/file?filename=Data-file-Europe-Power-Sector-2020.xlsx
[3]https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions

Technical Report

Introduction

This report aimed to build machine learning models to predict if the European Union will meet its goal of carbon neutrality by 2050. To do so, we would like to obtain an overview of energy production, measured in electricity generated, by fossil fuel sources and renewable energy sources. Related to this, we would like to measure the rate at which European countries emit greenhouse gases. These areas should give us an understanding of how the European Union is progressing towards its goal of slowing climate change and allow us to model future greenhouse gas emissions.

Data Understanding – Cleaning and Visualization

Our first dataset contained the electricity generated for 27 European countries measured in Terawatt per hour annually from 9 different energy sources.

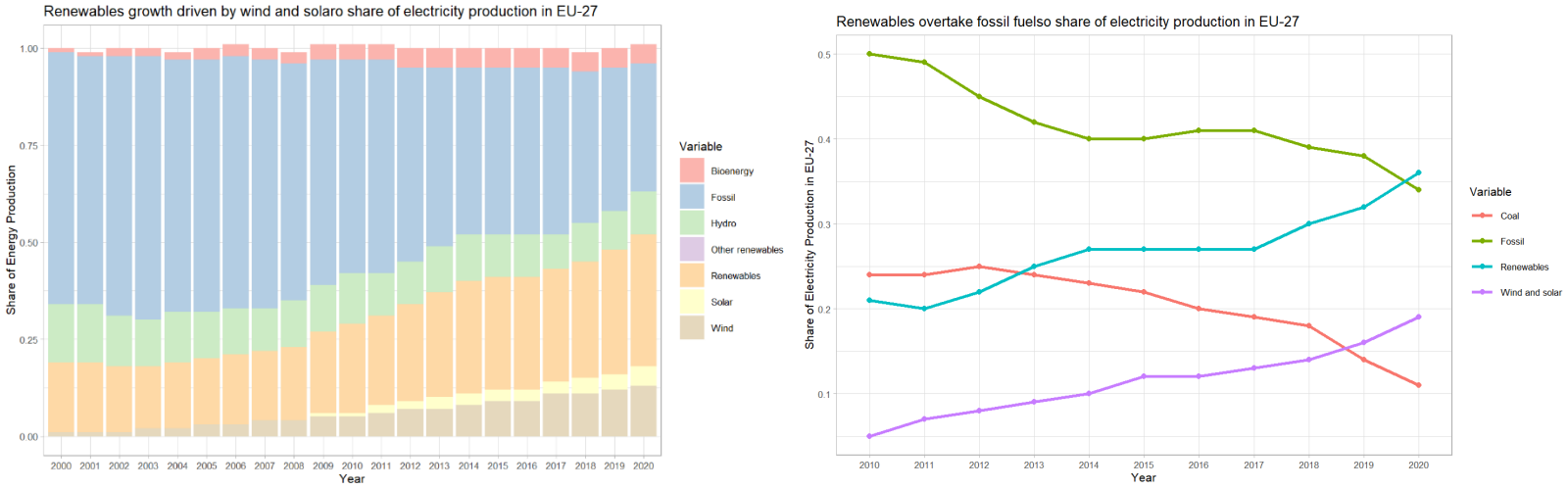
These power sources or attributes included wind, solar, other renewables, nuclear, lignite, hydropower, hard coal, gas coal, and bioenergy. This data was cleaned so that all attributes were given to each country (NA's replaced with a 0 measure) and each country was read as a factor in the data frame in R. From this first dataset, we were able to draw some conclusions about the energy sector in Europe and visualize certain trends. It was noticed that across the countries, the share at which different sources contributed to energy production varied. Or, different countries are dependent on different energy sources. For example, France as of 2020 has a much larger dependency on nuclear power at 67% of its total production of electricity. In 2020, on average across all European countries, nuclear power was only a 10% of total electricity generated. The mean share of production summarized from all countries for these categories followed this trend seen in Table 1. Additionally, some countries are still very dependent on fossil fuels.

The ten largest countries for total electricity generated in terawatt hours as of 2020 were Germany, France, United Kingdom, Italy, Spain, Poland, Sweden, Netherlands, Czech Republic, and Belgium. The energy production of these ten countries showed the trend that fossil fuels were decreasing as a proportion of energy generated, versus renewable energy increasing as a share of energy produced. However, gas as an energy source seemed to have leveled off and increased in recent years.

Energy Source	Mean Share of Production per Country in 2000	Mean Share of Production per Country in 2020
Gas	16.6%	23.8%
Coal	25.4%	10.7%
Wind	0.9%	14.5%
Solar	0.0%	4.0%

Table 1

In 2020, we can see that renewable energy overtook fossil fuels as the EU's main source of electricity for the first time, as shown in Figure 1. The major source of fossil energy was coal, and the share of electricity production of coal dropped from 23% to 10.7% from 2010 to 2020. The major sources of renewable energy are wind and solar and their share of electricity production of them increased from 21% to 36% in the same period.

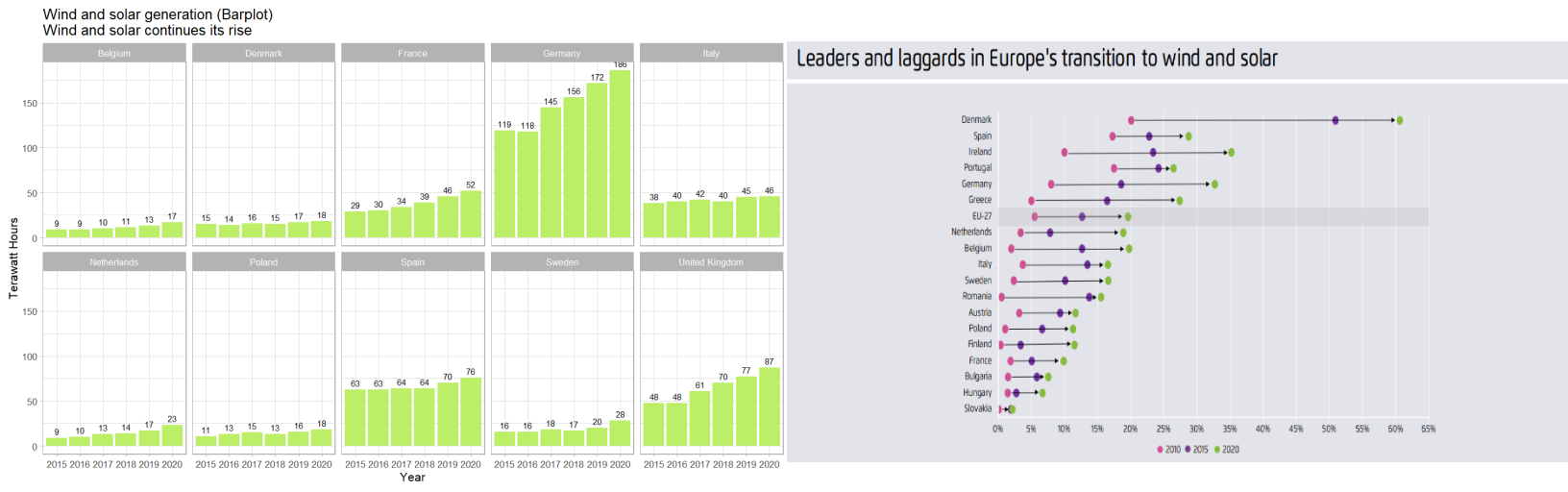


Wind and solar as a percentage of total energy generation.

Figure1

However, while some countries make advancements in renewable energy, others continue to lag. Denmark generated 62% of its electricity from wind and solar in 2020, which was almost double as compared to Ireland. Germany was third, and then Spain overtook Portugal into fourth place. Many of the smaller countries not on the graph continued to see near-zero growth in wind and solar. Seven countries - some with excellent conditions for solar and wind - have barely seen any growth since 2015 - Portugal, Romania, Austria, Italy, Czechia, Slovakia, and Bulgaria.

Figure 2



Although this dataset showed trends of the energy sector as a whole in the last twenty years, we found it had insufficient data for machine learning modeling. Additionally, we would not be able to effectively name a goal for where we hoped renewable energy production levels to be in coordination with carbon dioxide emissions. We found a second dataset that showed total greenhouse emissions for Europe.

Our second dataset was obtained online and showed the annual emissions of carbon dioxide measured in metric tons for 23 European countries from 1900 to 2020 and included 2,754 rows with 9 attributes. The attributes included country, year, annual carbon dioxide emissions per capita, and annual emissions from gas, oil, cement, coal, flaring, and an "other" emissions category. This dataset was cleaned from a larger set that included many more countries from around the globe as well as historical data dating back to the 1700s. Countries were set as a factor in R. Any years with unavailable data (NA) were set to a 0 value.

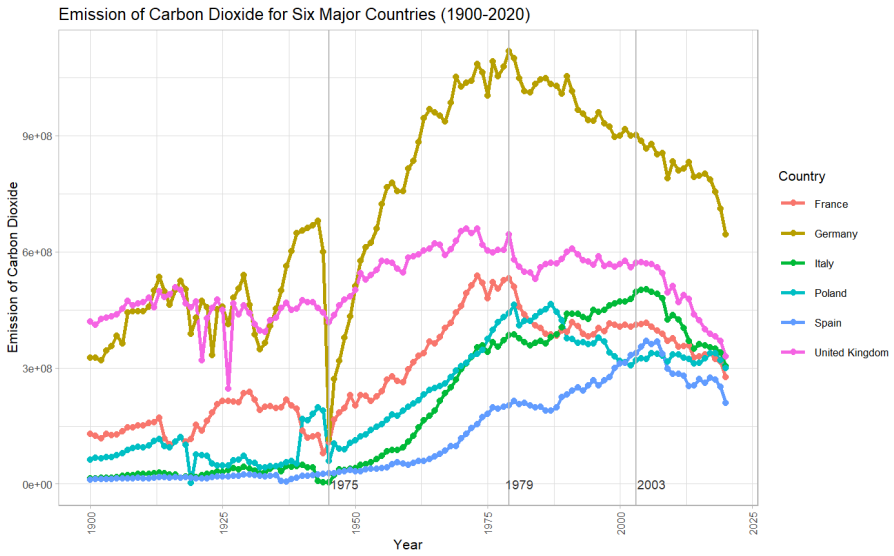


Figure 3

From our second dataset, we selected data from the top six countries, France, Germany, Italy, Poland, Spain, and United Kingdom, and plotted the trend of emission of carbon dioxide for these six major countries. The total carbon emissions for these countries is plotted versus time in Figure 3. We can see there was a huge drop in carbon emissions starting from the 1970s. Germany's carbon dioxide emission showed a progressive decline, and until 2003, the total emissions for the rest of the countries started to decline as well.

Figure 4 shows the major carbon dioxide sources for these six major countries. We can see that before 1949, coal was the only major source of carbon dioxide emission. Starting in 1949, new energy such as oil, gas, and cement gradually appeared. Until the 1980s, oil and coal were still the major energy source of creating emissions for EU countries.

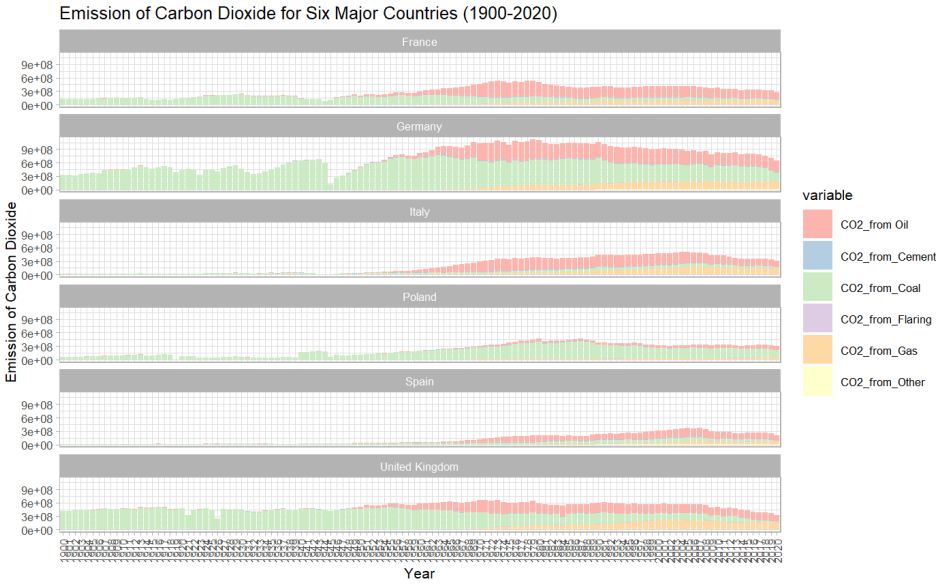


Figure 4

Modeling

The second dataset was used for a linear regression model. This model was used to predict annual carbon dioxide emissions for any European country in a given year as a function of annual emissions from gas, coal, cement, and oil. These attributes were chosen because they best correlated with the target variable, total annual carbon emissions. Their calculated Pearson Correlation Coefficients with total annual emissions is shown in Table 2. In Figure 5, you can see a visualization of the correlation between the attributes and the total annual emissions.

Attribute	Pearson Correlation Coefficient with Total Emissions
Emissions per capita	0.03
Coal	0.87
Cement	0.79
Gas	0.65
Flaring	0.51
Other	0.52
Oil	0.80

Table 2

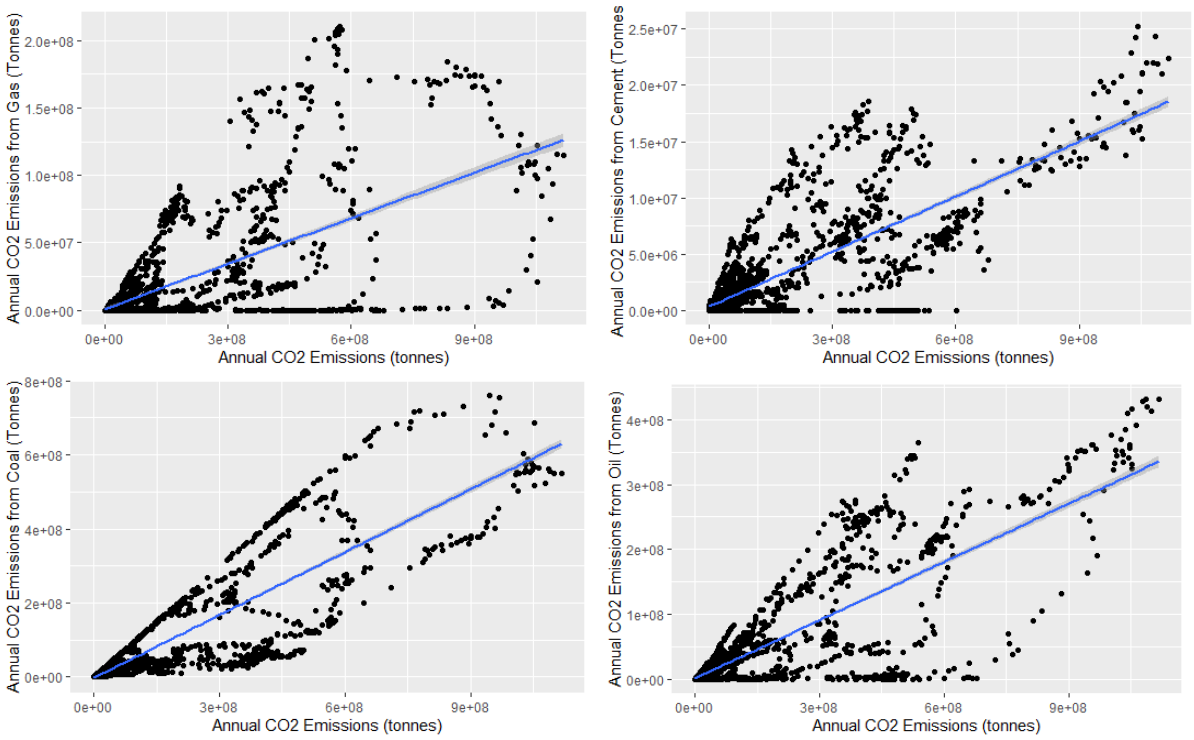


Figure 5

In the building of our model, we originally used 6-fold validation. For each fold, the holdout set was 20 years and the training set was the remaining years between 1900-2020. For example, a training set for Fold 1 was 1920 to 2020 and the test set was 1900 to 1919. However, we noticed that for most countries prior to 1940 the only data recorded for annual carbon dioxide emissions were for the attribute of coal emissions. Therefore, any training data that used the years from 1900-1940 would be biased towards coal as an attribute. This original model was discarded and we then used 4-fold validation with each year from 1940-2020 in the training data, and a holdout set of a twenty year period. The training data and holdout set for each fold are shown in Table 3. The glm() function was used in R for the building of this model.

Analysis

Each fold performed exceptionally well on the test set. The prediction accuracy was calculated for each fold as well as the mean absolute percentage error, as summarized in Table 3. The prediction accuracy is the total number of correct predictions divided by the total number of predictions. The calculation for the mean absolute percentage error is shown in Figure 6[4]. A sample of fold performance is shown in Figure 7, as a graph of predicted outcomes versus actual outcomes on the test set.

Overall, we would trust this model to accurately predict the total annual emissions for a country based on these four attributes. Choosing any of the four folds would work in the deployment of the model, however, the best to use would be a training set that uses the years 2000-2020. This is because much can change in the energy sector in twenty years, and it is important to show the most recent trends while also factoring in the growth of renewable energy. As we saw in our

earlier dataset, much has changed in the past decade regarding how countries produce energy. And certainly, much has changed in the past 100 years.

Fold	Test Set Correlation Accuracy	Test Set Mean Absolute Percentage Error
Test Set (1940-1959) Training Set (1960-2020)	99%	0.01
Test Set (1960-1979) Training Set (1940-1959 & 1980-2020)	99%	0.06
Test Set (1980-1999) Training Set (1940-1979 & 2000-2020)	99%	0.02
Test Set (2000-2020) Training Set (1940-1999)	99%	0.04

Table 3

$$M = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

M = mean absolute percentage error
n = number of times the summation iteration happens
A_t = actual value
F_t = forecast value

Figure 6

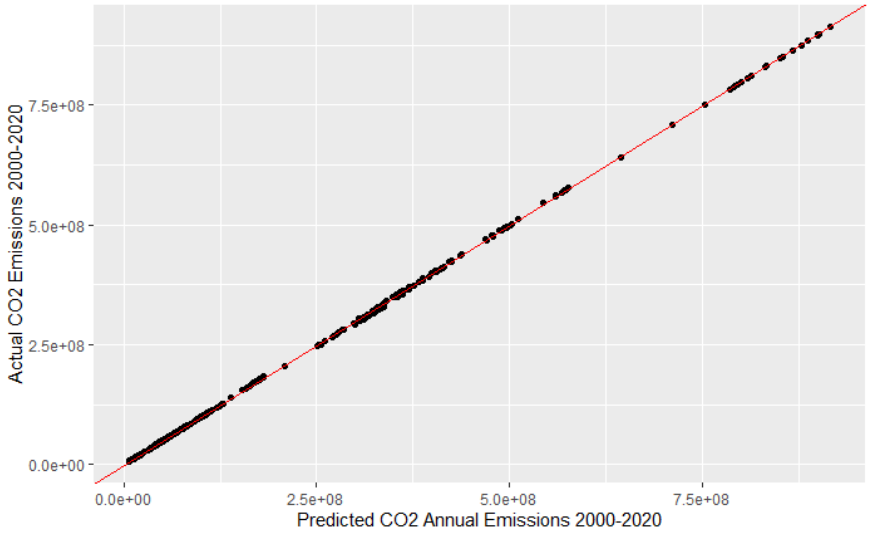


Figure 7

[4] https://en.wikipedia.org/wiki/Mean_absolute_percentage_error

Conclusions

After analysis of our test sets, we believe we have a model that will accurately predict the total annual carbon dioxide emissions for a country in a given year based on four key attributes – annual emissions of gas, coal, oil, and cement. This model will be important for European governments to calculate their emissions in future years. If they can have an understanding of where their emissions will be in a given year for these key attributes, they can ignore other sectors and calculate their expected total carbon emissions. This model would be most useful if private companies in the energy sector, namely in oil, gas, cement, and coal, were to report their expected production and emissions levels for a given year. The model could then be used with their total predicted emissions for the total overall emissions for a country.

We believe that this model will be important to use with the observations we have made in the renewable energy sector as well to predict future greenhouse gas emissions. As renewable energy technology continues to advance, we expect energy generation from fossil fuel sources to continue to decrease, as well as total overall carbon emissions as they both seem to follow a similar trend. We hope that a combination of the observations made in the first dataset, as well as the use of the model from the second dataset, will allow countries to have a better picture and be able to more accurately predict if they will be able to reach their goal of carbon neutrality by 2050.

Appendix

Matt – Wrote project description. Cleaned datasets 1 and 2. Made calculations on dataset 1. Helped write report. Wrote code for regression model for dataset 2 and analysis on results of a regression model. Visualized analysis of model. Helped create slides for presentation. Presentation(Modeling)

Roderick – Data Visualization for datasets 1 and 2. Briefly analyzed the data visualization part of the report. Wrote code for the data visualization. Helped group member to create and modify final report and presentation slides. Presentation(Data Visualization)

Chengxi Wu – Presentation (Introduction)

Ze Yu – Presentation (Conclusion)