USC University of Southern California

Predictive Analysis of Emissions Growth within United States

Mark Mann - MS Candidate in Applied Data Science (2021)
Javier Blandon - MS Candidate in Systems Architecting and Engineering (2020)

Introduction

Environmental implications of human actions are continually being discussed in hot-button topics such as global warming, greenhouse gases (GHG), and air quality. As technologies have progressed, viable solutions have begun to penetrate various industries and change the metrics and relationships between environmental impact and it's associated sources. As an example, research has proved that a positive regression relationship between GDP and GHG emissions is no longer the case for 33 U.S. states. The majority of the U.S. has managed to increase their GDP while decarbonizing, or decreasing their GHG emissions [Ref 1]. Research herein seeks to expand on this discovery by first detailing how the energy landscape has changed over the past two decades and secondly by exploring the relationships between GHG emissions, energy consumption, and energy production behavior.

The motivation for this topic exploration is driven by the desire to increase awareness and knowledge on this real world issue. The urgency of greenhouse gas emissions and societal responsibility loses its momentum in critical political discussions, as the supporting data can be overwhelming. This presents an opportunity to further the conversation of GHG emissions, and find new insights that may more effectively resonate in today's political climate. Questions include the following:

1. Of the main sources of emissions, which is the best indicator of overall state emissions?
2. Which is a better indicator of emissions, a state's energy consumption or production?

This report will give a foundational knowledge of greenhouse gases and their sources, describe the progression of energy sources over the last two decades, and attempt to find new relationships between human behaviors and their impacts on GHG emissions.

Emissions Domain Knowledge

In order to properly understand the significance of the data being observed, a baseline knowledge was developed regarding the subject matter of the atmosphere and greenhouse gases. The atmosphere is a layer of gases that acts as a filter, shielding the earth from harmful ultraviolet (UV) rays; it also serves as an insulator, trapping in infrared radiation (i.e. heat). This phenomenon is referred to as the *greenhouse effect*.

The greenhouse effect insulates the earth and helps minimize the earth's surface temperature variance. The solar energy that makes it through the atmosphere reflects off of earth's surface and projects outward as heat. While some of this heat continues on into space, the majority of the heat is absorbed and reflected back by greenhouse gases. Greenhouse gases are unique in this ability, and are enabled to do so by having an atomic structure of three or more atoms. This structure allows these gases to vibrate in an electrically unbalanced manner. Oxygen and Nitrogen are examples of gases that are abundant in the atmosphere but do not possess the same characteristics as GHGs. These gases have a two atom structure that cannot behave with electric potential even if they are in motion, or vibrating.

So how exactly are emissions measured? GHG emissions are measured in units of mass; the industry standard is metric tons. The amount, or mass, of GHGs being exhausted is calculated using chemical equations to convert amounts of fossil fuels burned, into the various gases emitted in the process of combustion. This data is then normalized by using a conversion called Global Warming Potential (GWP). This normalization acknowledges the different radiative efficiency, or its ability to absorb energy, as well as its lifetime, or how long it stays in the atmosphere. All GHGs are then normalized to the GWP of 1 metric ton of CO_2 over the time period of 100 years.

The breakdown by major source and examples of each are below [Ref 2]:

- 29% Transportation: cars, trucks, trains, planes
- 28% Electricity Production: Coal, Natural Gas
- 22% Industry: Chemical Reactions and fossil fuels
- 12% Commercial/Residential: Heating/Cooling, waste handling
- 9% Agriculture: cows, soil, rice production

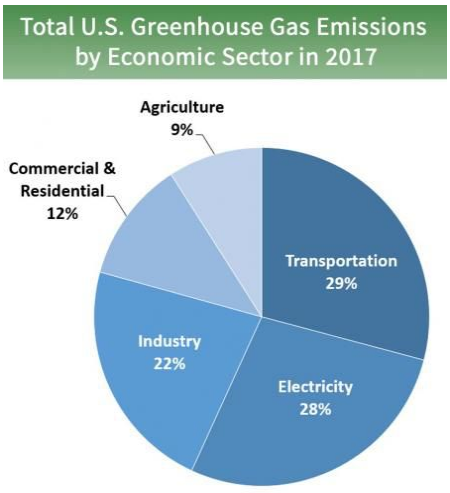

Figure 1 - Emission Source Breakdown

Process

Data Sources & Acquisition

This analysis considered state historic emissions, energy production, energy consumption, and economic data. This data was retrieved from two U.S. agency websites - the Energy Information Administration (EIA) and the Bureau of Economic Analysis (BEA). The EIA describes itself as an entity that "...collects, analyzes, and disseminates independent and impartial energy information to promote sound policymaking, efficient markets, and public understanding of energy and its interaction with the

economy and the environment." Similarly, the BEA is an organization of economists tracking United States statistics, one being gross domestic product (GDP).

From these web sources, six tables were extracted and normalized each one with its own python script. EIA provided data sets on state emissions, emissions by sector, energy production, and energy consumption covering years 2000-2016. This data is published in .csv format on the agency's website. BEA sourced two data sets containing state real GDP, income, and population. The BEA data sets were extracted using the BEA public API tool, allowing flexibility to request the exact data and attributes pertinent to the analysis.

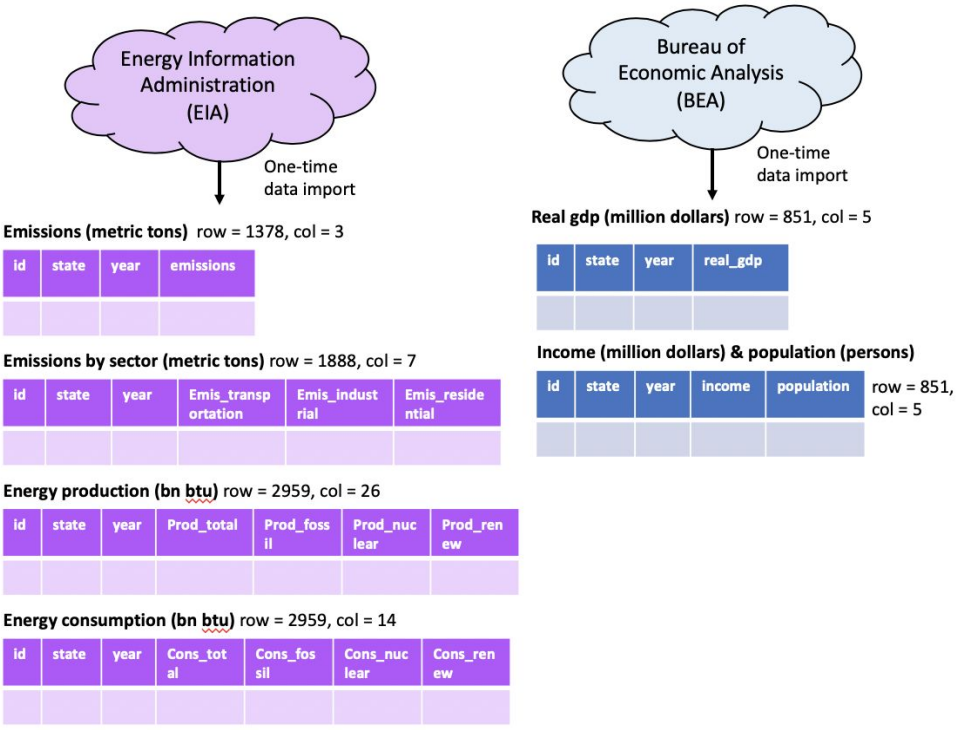

Figure 2 - Dataset Visual Depiction

Python scripts were developed to automate the data extraction process as well as transform the data for the joining of various data sets. The main functions of these scripts are as follows:

- **Extract data set from source -** for this function, a url is provided to the source website, such that the script can directly reference the dataset.
- **Transform data -** These datasets had varying years of coverage, and expanded in various dimensions. Some datasets had the attributes as rows and states as columns, others vice versa. Each script had to be tailored to the data source's original format, such that the proper transformation was completed. Steps were also taken to only extract data that were relevant to the research questions.
- **Assure quality of data -** Results of data transformation were checked against initial raw web dataset to confirm transformation was successful. Data quality was assured through this step.
- **Load .csv files for analysis** - Lastly, the scripts exported transformed tables into local .csv files. The cleaned .csv files were then stored in a shared data warehouse (Google Drive). The data warehouse

was then connected to a business intelligence software, Tableau. After Tableau was connected to the data, table joins and data exploration was completed within Tableau.

The process of normalizing the table formats was an unavoidable hurdle. However, this step was an essential prerequisite for exploratory data analysis (Tableau) and machine-learning model creation (Python). With normalized tables, the data could now be joined together to investigate relationships between economic and environmental metrics. For simplicity, the normalized tables were designed to use state and year as inner join keys. As a result, data between tables were only joined when there was a match of both state and year. Once all tables were joined, the full dataset could be described as approximately 860 rows, with 36 attributes.

Exploratory Analysis

After joining the tables, Tableau was used to explore visualizations of the connected datasets against the hypothesis questions. In alignment with the hypothesis, the goal was to uncover predictable relationships between economic activity, energy consumption, and emissions. In addition, one of the first steps was to create a visualization to define 'Top Performing' states who reduce their emissions while growing GDP. Therefore, focus was placed on these questions when creating visualizations - to steer efforts towards a model that would help answer these questions. The main methods used during data exploration included the following:

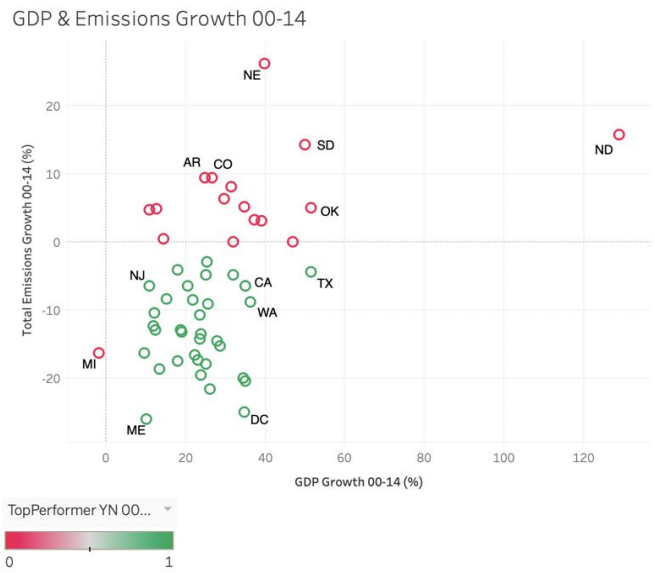

Figure 3 - 'Top Performing' States

Variable Creation - a state 'Top Performer' status variable was created that flagged states who decreased emissions while increasing GDP. To do this calculation, growth rate variables for GDP and Emissions were created from year 2000 to 2014. Then another variable considered these growth rate variables and flagged a state as 'Top Performer' status if the GDP Growth metric was positive, and the Emissions Growth metric was negative. This visual is displayed in Figure 3. Some high-level observations were: 1) 34 states were 'Top Performers' from 2000-2014; 2) Maine had the highest emissions decrease (26.09%); 3) Nebraska had the highest emissions growth (26.19% increase).

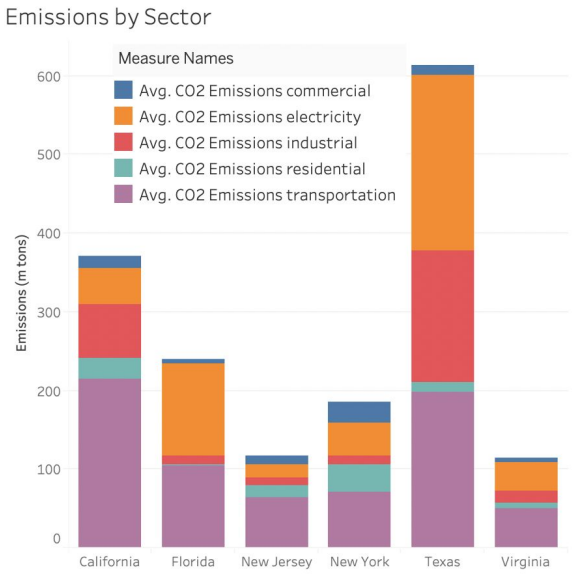

Figure 4 - State Emissions Source Breakdown

Univariate Analysis - Several other univariate visualizations were generated to better understand the make-up of the data. These included: CO_2 growth, CO_2 emissions by source sector (Figure 4), Energy consumption by source, Energy fossil fuel consumption by source. Better understanding was gained through these visualizations, which helped provide inspiration for model creation. Univariate analysis divulged that from 2000-2016, across all states, Transportation Emissions was the highest contributor to total emissions at 43% of total. Understanding the behavior of these various sectors at a state level was key to the research effort, as it continued to identify which attributes of a state would provide the most information gain during model creation.

Bivariate Analysis - Many combinations of two variables were explored during this step in attempts to uncover predictive relationships. Tableau was used to build the bivariate visualizations quickly. During this step, the focus was also on answering two questions related to the initial hypothesis:

1. Does energy production or consumption have greater correlation with emissions? - To explore this question, 2016 data was analyzed and built a scatterplot matrix, setting Production and Consumption as the independent variable and Emissions by Source (Total & Transportation) as the dependent variable (Figure 5). This visualization clearly identified the stronger relationship that exists between Consumption and Emissions. Because consumption has a strong relationship, a deeper dive was taken into consumption types (fossil fuel, nuclear, renewable) later in the analysis.

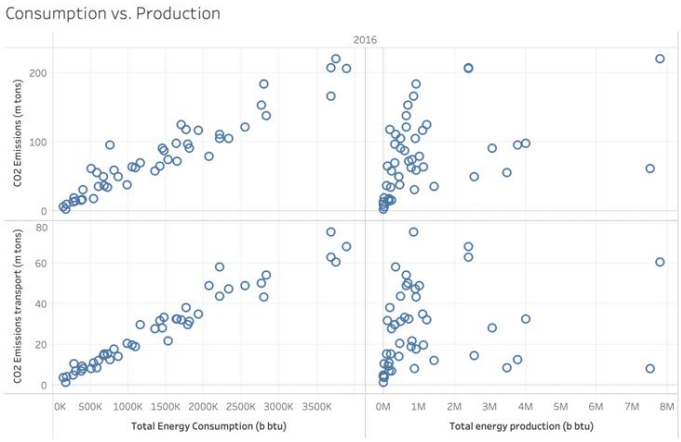

Figure 5 - Bivariate Analysis Visualizations

2. What economic factors can be used to predict emissions? - To understand this question, year 2016 data was selected to create another scatterplot matrix in hopes of observing a linear relationship. This scatterplot

used economic metrics (Real GDP, Population, Personal income, Energy Consumption) as the independent variable and Emission by Source (Total, Electricity, Transportation) as the dependent variable. Each observation reflects a specific state during 2016. To preserve the relationship, outliers were removed including larger states that must be studied separately. These state outliers were CA, FL, LA, NY, TX.

In reviewing Figure 6, it is distinguishable that economic variables have less correlation with Total Emissions and Electricity Emissions. However, the correlations with **Transportation Emissions** are striking.

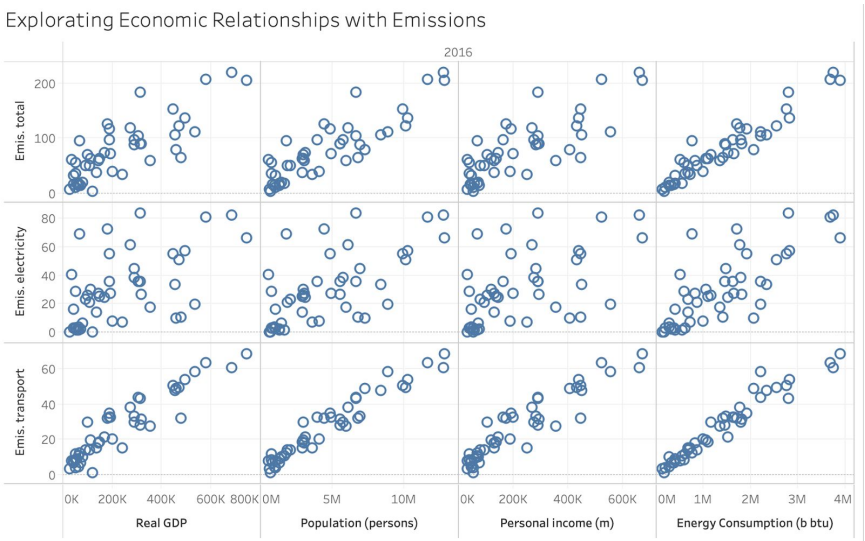

Figure 6 - Bivariate Analysis Visualizations

This strong relationship may indicate that Transportation Emissions are more driven by economic factors than other Emissions by Source (i.e. Electricity). Because of the strong linear relationship, this indicated that Transportation Emissions may be a strong candidate for prediction, using a linear regression line.

Model Creation
Linear Regression

- **Outliers removed:** CA, NY, TX
- **Tools:** python sklearn, pandas, matplotlib, pyplot

After observing strong linear relationships between economic variables and Transportation Emissions, bivariate linear regression relationships were formally defined. These regressions used 2016 data, and again removed the state outliers that must be studied individually (CA, NY, TX). Python scripts were connected to the .csv tables and used to build the regression models. Python packages pandas, pyplot, matplotlib, and scikit learn were used for the model and visualization.

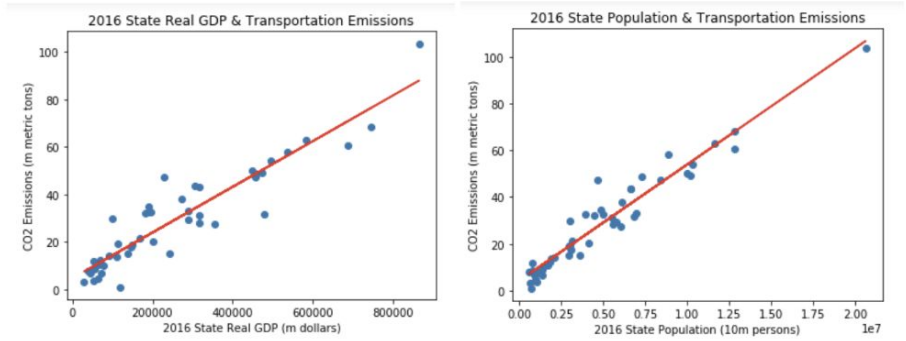

Figure 7 - Linear Regression Model

Each regression line had an R^2 score of at least 0.8659, entailing that the lines will explain at least 0.8659 variation in the dependent variable. Without domain knowledge of industry standard, it is difficult to deem the quality of this measure. However, the regression model provides a relatively accurate tool for predicting transportation emissions. A multivariate regression line, utilizing all independent economic variables (Real GDP, Population, Income, Consumption) was not created due to collinearity between one or more of the independent variables. Although possible, this was outside of the intended scope of this effort.

Decision Tree

Table 1 - Decision Tree Performance Results

Consumption Attributes (X)	Emissions Class (Y)	Training Years	Test Years	Decision Tree Accuracy
Fuel types	All Emissions	1990-2015	2015-2016	0.94
Fossil Fuel types	All Emissions	1990-2015	2015-2016	0.58
Fossil Fuel Types	Transportation Emissions	1980-2015	2015-2016	0.7

After identifying a relationship between Total Emissions and Energy Consumption, efforts were shifted to further develop this relationship by seeking what energy consumption types (Fossil, Nuclear, Renewable) and fossil fuel consumption types (Coal, Natural Gas, Petroleum) are most related to increases in Emissions. Python pandas was used for data transformations and scikit-learn for the algorithm and visualization. As a prerequisite for this analysis, the data had to be transformed to fit the decision tree categorical input requirements. For this analysis, the Class was set to Emissions and Attributes to Coal, Natural Gas, and Petroleum Consumption. Then, the year-over-year growth was computed for each class and attribute from 1990-2016. Finally, each growth metric was categorically encoded as 0 - Decrease, 1 - No Change, 2 - Increase. After encoding the data, three different decision trees were **trained** on data from 1990-2015 (rows = 1250), and limited growth to three nodes to avoid overfitting. Each tree was then **tested** with data from 2015-2016 (rows = 50) and the results analyzed. Within the python scikit-learn package, DecisionTreeClassifier object was used to grow the tree and partition data nodes using the Gini impurity index. Gini impurity shows the frequency that a randomly chosen element would be incorrectly labelled. Similar to node splitting using highest information gain, the smallest gini impurity is used for splitting the tree. Also for further nodes, the DecisionTreeClassifier object considers gini impurity of smaller partitions of the same attribute, to maximize the accuracy

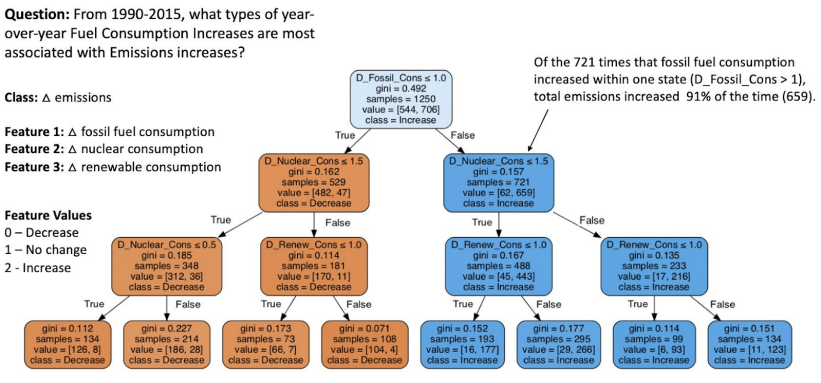

Figure 8 - Fuel Types/All Emissions Decision Tree

of classification. Per Table 1, the highest decision tree class labelling accuracy of 94% was obtained by setting Class to 'All Emissions' and Attributes to 'Fuel Types' (i.e. Fossil, Nuclear, Renewable) Within said decision tree (Figure 8), the first node suggests that fossil fuel consumption increases have the highest impact on emissions increases. Even after increases in nuclear and renewable consumption are considered (second and third nodes), the emissions increase prediction mostly mirrors fossil fuel increase status. This observation reflects that fossil fuel consumption is highly tied to emissions. However, inclusion of the second and third nodes (Nuclear and Renewable) increase the model prediction accuracy. Figure 9 displays a different decision tree Class 'All Emissions' and Attributes to 'Fossil Fuel Types' (row 2 of Table 1) showing a more balanced decision tree. The contrast between these two trees depicts the various forms a model can take based on the question asked; in most instances the question will drive the attributes used, and as a result the predictions and accuracy will differ.

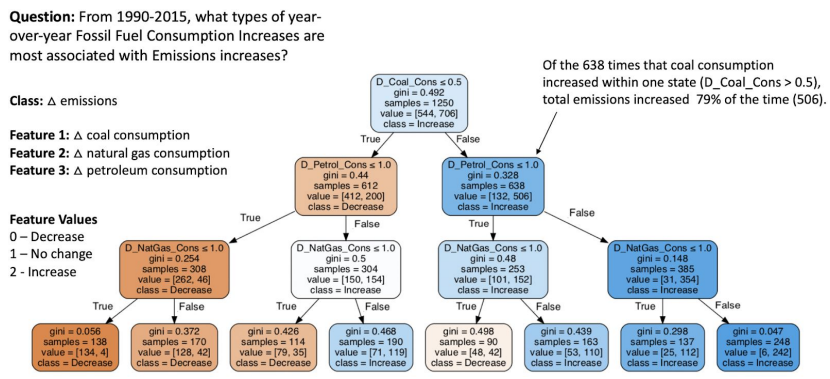

Figure 9 - Fossil Fuel Types/All Emissions Decision Tree

Results and Conclusion

This research effort proved that 1) State transportation emissions can be predicted using Population, GDP, and Consumption [Linear Regression] and 2) Decreasing fossil fuels has a greater impact on emissions than renewable energy consumption [Decision Tree]. These are important insights because they provide a data driven basis for particular courses of action. As an example, states looking to decrease their overall emissions may consider incentivizing decreased fossil fuel consumption, instead of renewable source consumption such as solar panels. This topic starts the conversation on devising creative ways of thinking and approaching data on GHG emissions. With a better understanding of energy behavior data, there will come a greater ability to address the issue in a more systematic and effective manner.

References

1) Devashree & Muro. "Economic Growth from Emissions Growth", Brookings, 8 Dec. 2016, [Link].
2) "Sources of Greenhouse Gas Emissions." Environmental Protection Agency, 13 Sept. 2019, [Link].
3) U.S. Energy Information Administration (EIA) - Independent Statistics and Analysis, [Link].
4) U.S. Bureau of Economic Analysis (BEA), [Link].