Energy consumption and CO2 emission

Madhu Sudha Subramani
Department of Computer Science
City, University of London

Abstract

This paper analyses the trend of energy consumption by people over the period and studies the relationship between the growth of a country's economy and also the CO2 emissions. Among the many sources of CO2 emission, the CO2 emitted from production of electricity from different sources are analysed in particular to determine which ones contribute more CO2 emission using prediction models.

I. Introduction

In today's world, humanity cannot survive without the use of electricity and other forms of energy. We need energy everywhere and for everything, from households to various industries such as agriculture, manufacturing, transportation, and electricity production, among many others. Following the industrial revolution, developed countries began to use more energy, which was soon followed by developing countries, and now every country with a human population requires electricity. With new inventions happening every other day, the demand for electricity and other fuels is constantly increasing. While energy consumption has numerous advantages, it also has numerous disadvantages. Global warming is a well-known threat to the environment's sustainability. Carbon dioxide (CO2), which is emitted in large quantities due to energy consumption, is the most significant contributor to global warming. CO2 emissions are rising due to a variety of factors, including the use of fossil fuels to generate energy, natural or anthropogenic deforestation, cement production, and so on. While we understand the issue, we cannot stop using energy because of our living standards and the benefits mentioned above. However, in this analysis, let us look at the trend of energy consumption and its relationship to both CO2 emissions and a country's growth.

II. Analytical questions

We chose data from the World Bank's World Development Indicators for our analysis. The analysis is carried out to discover the answers to the following research questions

1. What is the average energy consumption per capita in the whole world, and how has it changed over the years?
2. How and to what degree does energy consumption affect the growth of a country's economy? We know that the use of energy in households and various industries plays a major role in improving a country's economy, but how strong is the correlation between them?
3. What is the energy consumption trend of countries based on income groups, namely high, upper middle, lower middle- and low-income groups?
4. What has been the trend of energy consumption and CO2 emissions over the years in the top 10 biggest economies in the world?
5. Electricity is produced from various sources, like burning fossil fuels, nuclear sources, renewable sources, etc. For any particular country, can we predict which source of electricity production contributes more or less to the CO2 emissions?

III. DATA

The data for this analysis is taken directly from the World Development Indicators from the World Bank. There are a total of 1442 indicators and their values from the year 1960, belonging to different categories. Though data can be filtered based on these categories and their sub-categories from the website itself, the data required for this analysis needs indicators from different categories, so the entire dataset was downloaded and then the required indicators were selected manually as part of the data preparation.

On downloading, there were a total of 6 csv files, namely, WDICountry-Series.csv, WDICountry.csv, WDIData.csv, WDISeries-Time.csv, WDISeries.csv

Out of these 6 csv files, the main information is in the WDIData.csv file with 383572 rows and 67 columns.
The indicators contained not only the values of all the countries but also different aggregates of countries based on region, income group, or some other category. The dataset is not in a format that can be easily used; all the indicators were listed as values of single column while the years from 1960 to 2021 were listed in separate columns. There is a lot of data preparation required to convert them into an easily usable format. Some of the data we are analysing does not have values before 1990 or after 2014 or 2015. Therefore, most of our analysis would be based on this period.

IV. Analysis

1. Data preparation

a. Data pre-processing

The columns with redundant information are removed, like Country Code and Country Name provides the same information, so Country Code is removed. Similarly, Indicator Code is removed, keeping Indicator Name as the name provides more meaning than the code.
For the ease of use, converted the rows from years '1960' to '2021' into a single column named 'Year' and converted

each value of 'Indicator Name' column to separate columns after choosing the required indicators.

b. Feature selection

For the ease of data retrieval and analysis, only 12 indicators are selected and derived a dataset that contained values from 1960 to 2021 for all the countries and aggregate groups. This is assumed to be the superset of our analysis and for each of our analysis, a subset of this superset is extracted with required indicators and then few preprocessing is done separately for each analysis. The 12 indicators are manually selected based on the research questions as few of them have to be taken from different categories from the original raw dataset.

These indicators are,
GDP per capita (current US$)
Energy use (kg of oil equivalent per capita)
Fossil fuel energy consumption (% of total)
Renewable energy consumption (% of total final energy consumption)
Electric power consumption (kWh per capita)
Electricity production from coal sources (% of total)
Electricity production from hydroelectric sources (% of total)
Electricity production from natural gas sources (% of total)
Electricity production from nuclear sources (% of total)
Electricity production from oil sources (% of total)
Electricity production from renewable sources, excluding hydroelectric (% of total)
CO2 emissions (metric tons per capita)

c. Handling missing data

There are many missing values in the original dataset across the years for each indicator. These are not removed in the superset dataset, instead it will be handled in the subset created for each analysis. In each of the subsets, rows containing the null values are removed from the dataset as data cannot be imputed.

2. Data derivation

a. Outlier detection

In figure 1, the correlation between Energy per capita and GDP per capita is established and there seems to be many outliers and also all the datapoints are clustered together in a corner which is difficult to understand. Therefore, to analyse the distribution of both these columns, the datapoints are plotted in Q-Q plot (figure 2) against normal distribution. It is observed that the datapoints are not normally distributed.

Figure 1

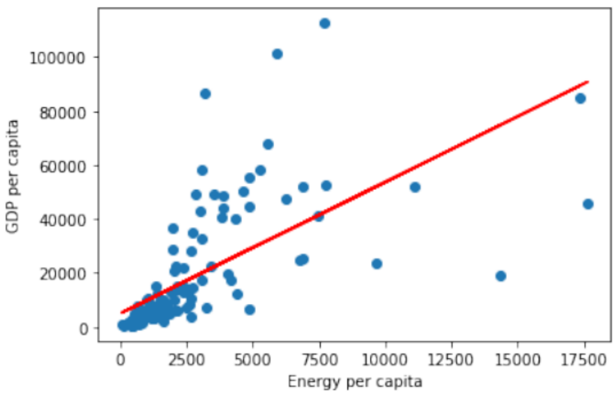
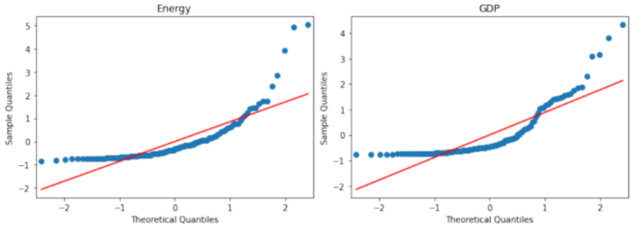


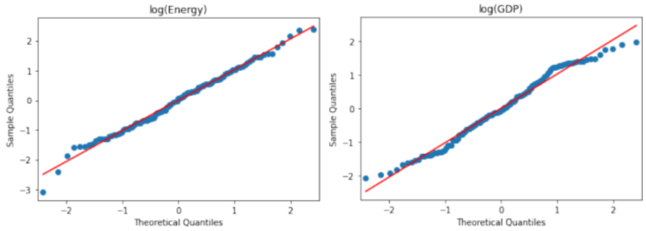
Figure 2



b. Data Transformation

However, the core value of 66 between these columns suggests that there is linearity which is not clearly visible from the first figure. So, a log transformation is applied on these columns and then Q-Q plots are plotted. Now, both these column values fit perfectly in the normal distribution except few datapoints, which we choose to keep it.

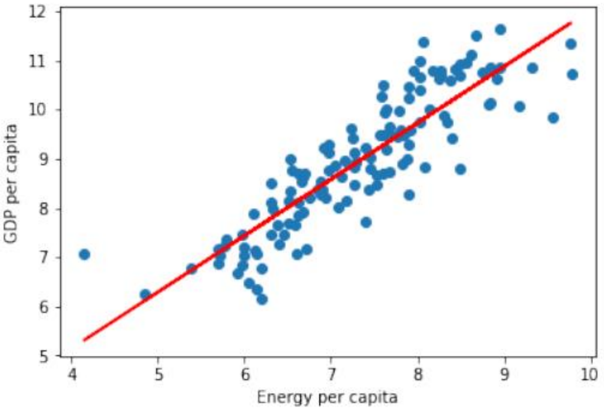
Figure 3



c. Exploratory Data Analysis

Data preparation and derivation are part of exploratory data analysis. The main aim of EDA is to analyse the data with simple transformation or with visual analysis. As part of our analysis, we try to find how much does energy consumption affect the growth of a country's economy. For this, a simple scatter plot is used and measured against a linear line drawn with the correlation coefficients which is shown in figure 3. It is observed that the correlation between Energy use and GDP is very strong and around 76 percent of the variance in GDP is based on the Energy consumption.

Figure 4



3. Construction of Models

To answer the research question of whether we can predict which source of electricity production contributes more or less to the CO2 emissions, we take the aid of machine learning models, multiple linear regression and random forest regression and compare the results of both. We aim to do this analysis with datapoints from one country, e.g., United States. So, there are very few datapoints to draw a result from single model and hence two models are used. The target value is Co2 emission per capita and the input features are percentage of different sources of energy used to produce electricity.

a. Finding correlation and co-linearity

The correlation and co-linearity between all the columns are detected from the figures 5 and 6. It seems that the use of Coal has high correlation of 94% with CO2 emission which is very evident. So, we try to choose 2 other features (Oil and Renewable sources) which has strong correlation with CO2 emission but not co-linear.

Figure 5

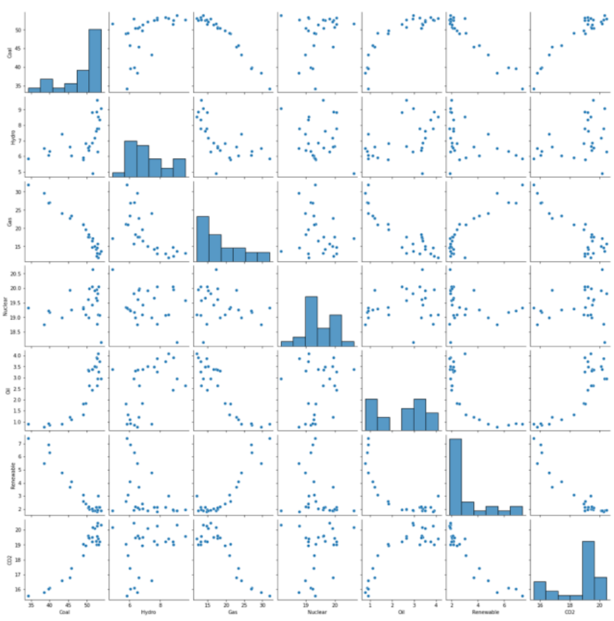
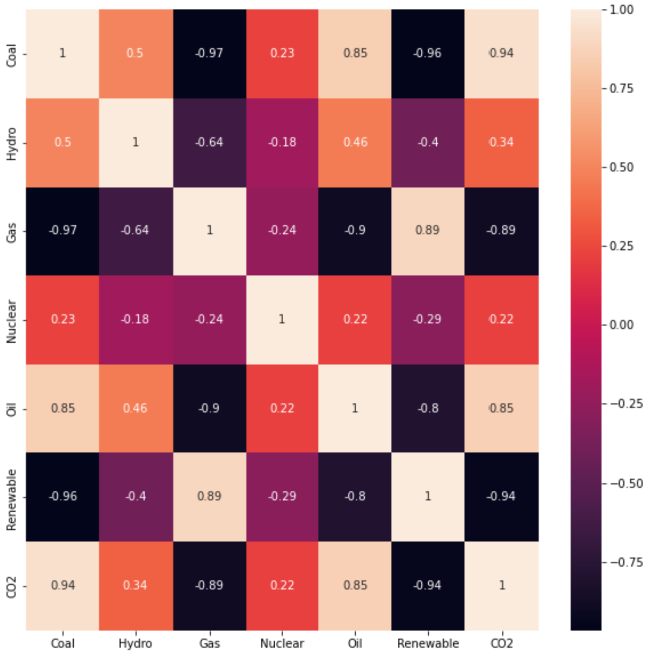


Figure 6



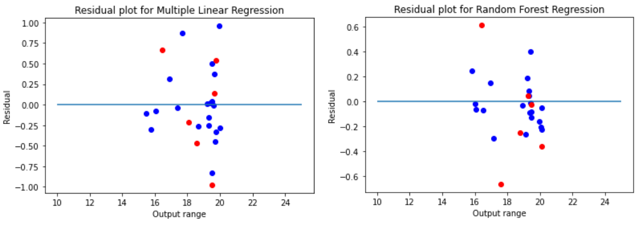
4. Validation of results

Both models are applied on the dataset with 80% of the data divided to train the models and 20% to test. A random state variable is set to 1, as this shuffle of dataset has good mix of datapoints. The results of these models are shown in Table 1. On comparison, Random Forest regressor predicts better with 92% R2 value and RMSE of .40 against the linear regression results of 84% R2 value and RMSE of .57. Figure 7 shows difference in the residuals.

Table 1

Models	**R2 value**	**MSE**	**RMSE**
Multi linear	0.84	0.32	0.57
Random forest	0.92	0.16	0.40

Figure 7

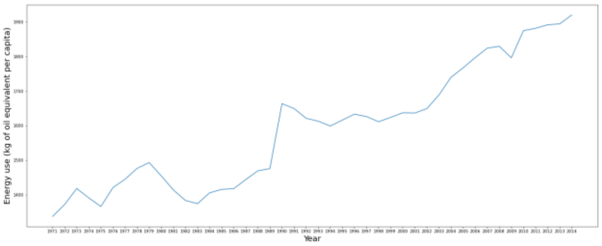


V FINDINGS

1. What is the average energy consumption per capita in the whole world, and how has it changed over the years?

From the figure 8, the energy consumption saw a tremendous increase from below 1400 kg of oil equivalent per capita in early 1970s to more than 1900 in 2014. There was a sudden spike in the consumption in early 1990s and then constantly kept increasing except for a small dip in 2008 and 2009.

Figure 8



2. What is the energy consumption trend of countries based on income groups, namely high, upper middle, lower middle- and low-income groups?

Out of 4 income groups, the low-income group had many null values, so this group is removed from our analysis. The total energy consumption is analysed along with how much is from non-renewable and renewable sources. The high-income group has always been consuming more than 4000 kg of oil equivalent per capita from 1990s, with most of it sourced from non-renewable energy while very small amount sourced from renewables. But we can see the increase in the use of renewables from mid-2000s which is a good sign. The lower and upper middle income has seen an increase in the total energy consumption from 2000s but it seems that the energy is majorly sourced from non-renewables while the renewables had stayed constant over the years.

Figure 9

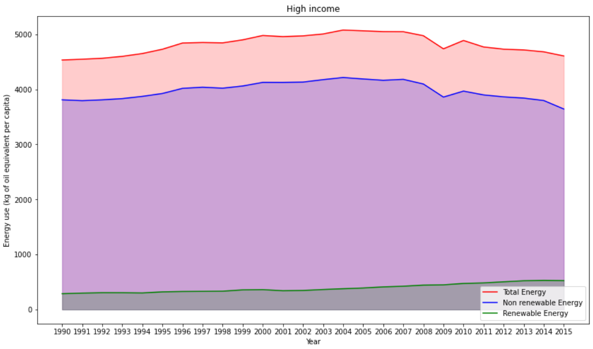


Figure 10

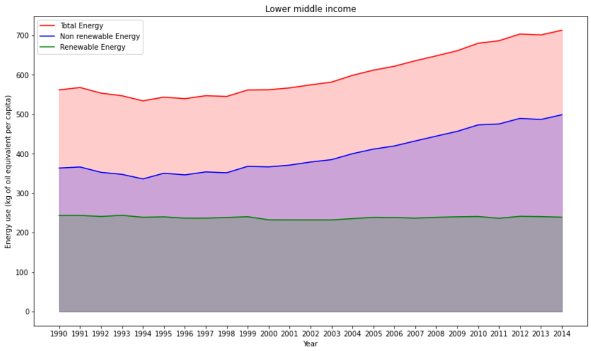
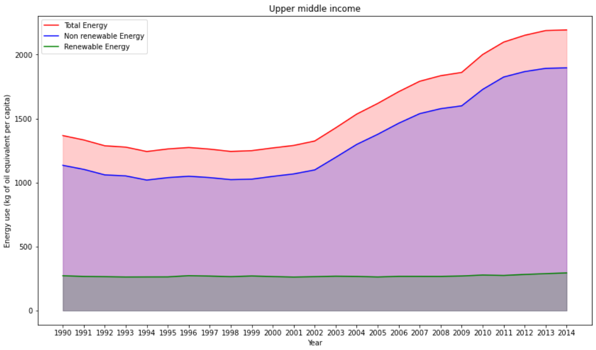


Figure 11



3. What has been the trend of energy consumption and CO2 emissions over the years in the top 10 biggest economies in the world?

The figure 12 and 13 shows the trend of the energy consumption and CO2 emission looks similar for most of the top 10 biggest economies in the world, except China where a small increase in the energy consumption reflects tremendous increase in CO2 emission. Among the developed nations, the United States seemed to have reduced the energy consumption in late 2000s which brought down the CO2 emission significantly.

Figure 12

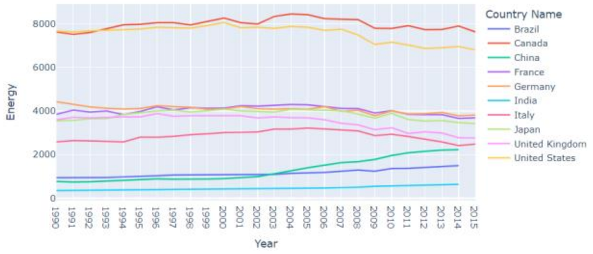
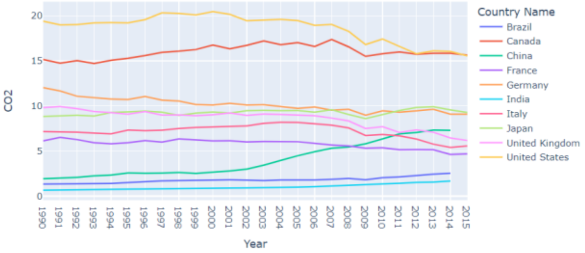


Figure 13



4. Electricity is produced from various sources, like burning fossil fuels, nuclear sources, renewable sources, etc. For any particular country, can we predict which source of electricity production contributes more or less to the CO2 emissions?

From our analysis using prediction models, it seems that electricity from burning coal releases more CO2 while the electricity from renewables controls the CO2 emission to a very great extent. Also, the second most contributor is oil combustion. While electricity production from gas also controls the CO2 emission, the hydro and nuclear sources had very little effect.

V. Reflections and further work

The datapoints used for modelling was very less as we analysed for only one country. There was a challenge in considering all the countries, as indicators of each country showed different distribution. For generalization, a great amount of manipulation and transformation is required, which is very difficult to achieve within the limited time. As the dataset was very small, the feature selection for multiple linear regression was also limited to two. For future work, these limitations can be overcome by adding more features when we have much larger dataset.

References

https://databank.worldbank.org/source/world-development-indicators

https://www.rockefellerfoundation.org/wp-content/uploads/2020/12/Modern-Energy-Minimum-Sept30.pdf

https://www.sciencedirect.com/topics/economics-econometrics-and-finance/renewable-energy-consumption