A Prediction on CO2 Emissions in Different Countries with Multiple Factors

Luo Cao, Ashwin Nair, Vishesh Singh Thakur

December 12 2019

Abstract

Countries, whether it being undeveloped, developing, or developed, all have emitted global carbon dioxide in one way or another. Factors causing these emissions differ per country, which is the focus in this paper. We analyzed multiple countries with different predictors of carbon dioxide (CO2), such as gross domestic product (GDP) per capita, country development, population, etc. Using these factors,we built and ran several predictive models and collected the root mean square error for each of the models to determine the best model, which is the boosting model. This leads to the findings of the significant factors of Carbon Dioxide emissions, which includes urban population and total energy consumption. This paper will discuss different factors of CO2 emission, the building of the predictive models, significant predictor variables of CO2 emission, and the interpretation of the Environmental Kuznets Curve (EKC) given our predictive models.

1 Introduction

Global carbon dioxide emissions have been increasing in the last few years. One of the issues that arise due to the level of CO2 emissions are increases in temperature and in air pollution. As fossil fuel is being burned and CO2 is released and being absorbed into the environment, there has been increases in atmospheric temperature [1]. According to the Environmental Protection Agency, CO2 accounted for approximately 65 percent of the overall gas emissions in 2010 [2]. The increase of this emission across countries is due to multiple factors, which includes the growth in GDP of the country, different industry sectors, the population of people, whether they import or export products etc.

With these factors, we can create a predictive model to forecast the CO2 emissions for each country. The basic idea of this predictive model is how any country can curb its carbon footprint without sacrificing the economic growth. The model will check the different sectors of economic growth and what their contribution is to the carbon emission and how we can either increase or decrease the growth of some sectors to be more eco-friendly without losing the economic advantage.

There are previous research done on CO2 emission and its factors, with a focus on GDP of one specific country. The model these research have been based off was the Environmental Kuznets Curve (EKC) hypothesis. This model states that as a country is developing and its GDP is low, the CO2 footprint is at its maximum, and once the country is developed, the carbon footage decreases for the country [3].

Our paper proposes to collect data on several countries over the span of 18 years along with multiple factors that may cause CO2 increase. With this data, we found the factors that most influences CO2 increase and built multiple predictive models to determine the best model for predicting causes of CO2 emissions. We then selected the best model to determine the most significant factors for CO2 emissions and compared it to the EKC model, to test its accuracy.

This paper is structured in the following: Section 2 will discuss the overview of the EKC hypothesis in research papers, different factors in regards to CO2 emissions from several research and gaps in the research. Our research framework follows in Section 3, then lead by our exploratory analysis in Section 4, methodologies are then explained in Section 5 followed by our results in Section 6 along with a summary of our prediction models.

2 Literature Review

2.1 Understanding the EKC Hypothesis

The EKC model is in a bell shaped curve, where it depicts that as the growth of a country increases with the GDP per Capita, the CO2 emission increase. As the country becomes developed and GDP is high, the CO2 emission decreases [2]. There has been several literature which studies the effect of GDP per Capita of a country on CO2 emissions which checks the accuracy of this model. Figure 1 shows the graphical representation of the EKC

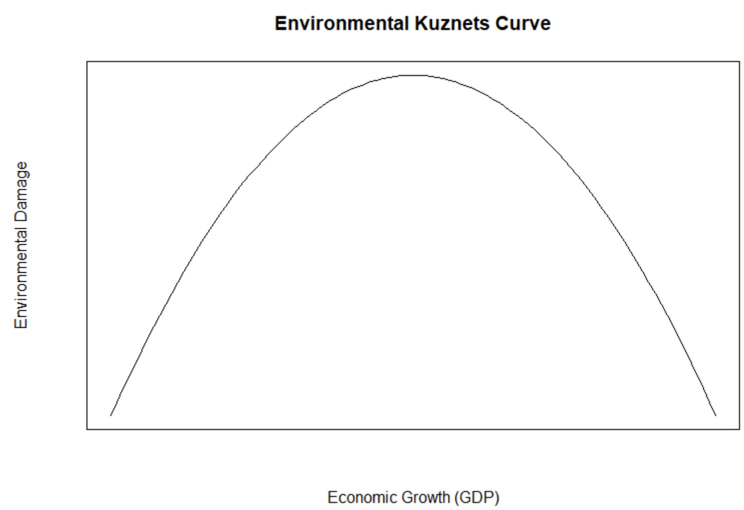

Figure 1: The Environmental Kuznets Curve

Hsiao-Tien Pao and Chung-Ming Tsai [4] investigated the relationship between carbon emissions, energy consumption, foreign direct investment (FDI), and GDP for four countries, Brazil, India and China and China using a panel cointegration framework. This framework tests to see if there is a relationship between the variables over a long period of time [4]. Their findings shows support of the EKC hypothesis, with bi-directional causalities between emissions and FDI, emission and output, energy and FDI, and energy and output [5].

With some research validating the EKC hypothesis, there are a handful that do not validate this curve. J.W.Sun's literature, The Nature of CO2 Emission Kuznets Curve, his findings show that not all countries follow the this hypothesis as some countries have not yet gone through industrialization [6]. Jenny Cederborg and Sara Snobohm [7] analyzes multiple curve hypotheses in which they reject the EKC hypothesis as there is not a lot of evidence indicating the GDP per capita in which the curve turns to a decrease in CO2 emissions.

In Wang's [8] findings for his research in the relationship between CO2 from oil and economic growth for 98 countries, the EKC hypothesis cannot hold. This is due to that in countries which the economic growth is of a high level, it does not effect how much CO2 is emitted.

Findings in Ameyaw and Yao's [9] literature suggests that the countries in Africa shows that there is also no relation with the EKC hypothesis with their research. There is no evidence between CO2 and gross fixed capital information (GFCF). They suggest that a new evaluation system should be created to analyze CO2 and its causes.

Analysis of EKC hypothesis is validated for certain countries but not all countries. Reasons for this could include that there are other factors including GDP that play a large role in the changes of global carbon dioxide. While the EKC hypothesis is widely used as a comparison for research models, there are among many literature that explores the CO2 emissions without exploring and validating this curve.

2.2 Analysis of Factors causing CO2 Emissions

The gross domestic product (GDP) per capita is known as a large factor of CO2 emissions, and is used analyzed in literature while studying CO2. Hatzigeorgiou, Polatidis, and Haralambopoulos [10] analyzed the causality among energy consumption, GDP and carbon emissions in Greece with data from 1977-2007. They used the Granger-causality test, which tests to see if one time period can be used to predict the next time period. Another method was the cointegration test used by Govindaraju and Tang [11] to investigate the relationship between CO2 emissions, economic growth and coal. However, with the analysis of coal in CO2 emissions, Ramseur [12] found that the there has been a 20 percent decrease in coal as a electricity, which contributes to the decrease in CO2 emissions. He suggests that the predicted emissions when compared from 1990 to 2017, has been higher than the actual emissions.

Following coal for CO2 emissions, Wang, Li, and Zhang [13] focuses on the CO2 emission based on energy, and GDP in China. This research found that there was a negative decoupling between energy consumption and GDP. It also indicated an increase of energy emission is based in two stages: slow growth and rapid growth. With Wang, Li, and Zhang [13] focusing on all energy based emissions, Usama Al-mulali [14], focused on energy emissions from nuclear energy. Models of CO2 emission and nuclear energy consumption along with GDP and nuclear energy consumption showed that nuclear consumption does not play a big role in CO2 emission as much as fossil fuels.

Another method to determine CO2 emissions is the extreme learning machine (ELM) used by Sun and Sun [15]. ELM is a method used to choose nodes randomly and it chooses the weights of these layers based on analyzing the nodes chosen [16]. It is a type of learning algorithm that has a fast learning speed and can give a small training error. In doing so, it gives a better performance and can predict the response variable, the CO2 emissions, more accurately. The research on different factors that impact CO2 along with EKC hypothesis shows gives insight on variations due to different methods.

2.3 Gap on Research and our Contribution

The current research on CO2 gives a lot of focus on economic growth, GDP, and electricity consumption based CO2 emissions. These focuses are usually based on one country over a span of several years, which do not correlate well with other countries. Kasperowicz [17] takes into account data from 18 European Union countries, however it does dot include the countries development, whether it is underdeveloped, developing, or developed. Similarly, Usama Al-mulali [14] does not consider the different CO2 emissions and country development. Following these few gaps, our contribution follows different factors and multiple countries with this paper.

Our contribution lies with using not only GDP, and energy consumption, but other factors that are not as noticeable. This research focuses on CO2 emissions in different factors, which include: GDP, population, per capita income, labour force, literacy, rate, tourists per year, etc. with data from 1999-2007. We are also measuring CO2 from different countries and taking account of their development status. We are analyzing and determining the major influences on this data, and comparing the predicted CO2 emission for each of our models. Based on this, we will be comparing the EKC curve with our models to test its accuracy.

3 Research Framework and Data Collection

Before collecting data observations, we first identified our predictor variables for our models. We selected these predictors based on past literature reviews and collected variables of significance in predicting CO2 emissions. After determining the variables needed, we created a framework as seen in Figure 2, which shows the process in which our research would flow.

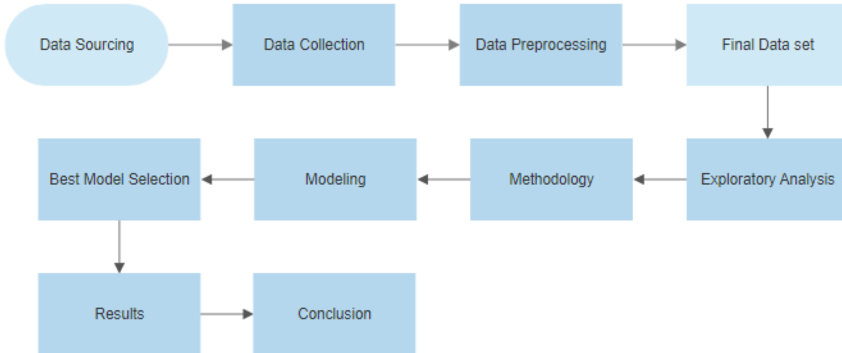

Figure 2: Research Framework

The flowchart shows the start of our research with the data source. We collected the observations from four difference sources on the Worldbank Data [18] and collected data for each of our predictor variables. We then combined all observations into one file to create one full set of data. The final data set collected consists of 576 observations and 12 variables from 32 countries for a period of 18 years.

4 Exploratory Data Analysis

An initial analysis of our data set is performed to explore the basic relationships and between the predictor variables and the response variable. The first relationship we analyze is relation between increase in Real GDP and Carbon emission. The relationship does not follow any specific trends. The carbon emission may increase with increase in Real GDP or will remain constant. But the significant thing we can take from this graph is at no point their is a decrease in carbon emission with increase in Real GDP. .Figure 4 shows the plot for GDP vs the CO2 emission

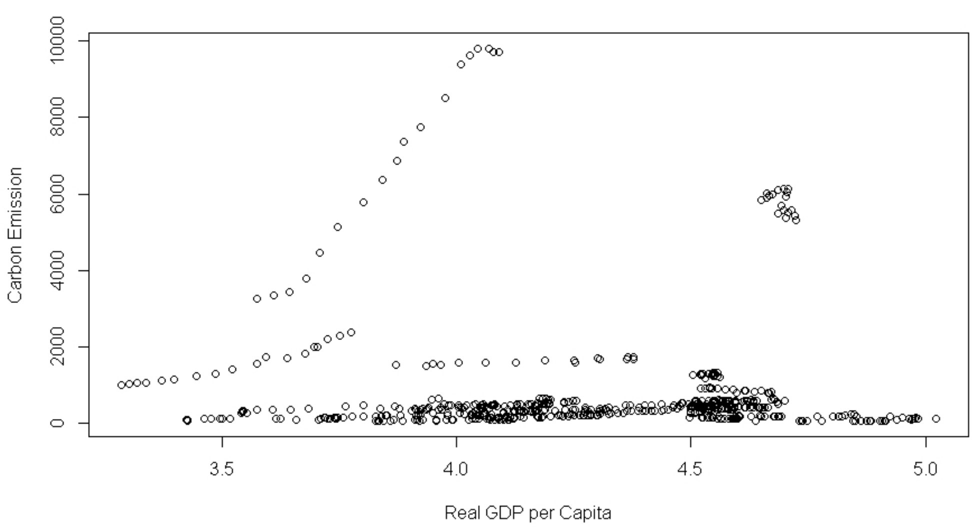

Figure 3: GDP vs CO2 emissions plot

We then take a look at how the carbon emission behave with increase in urban population. The relationship seems to increase suddenly after a certain point of time.

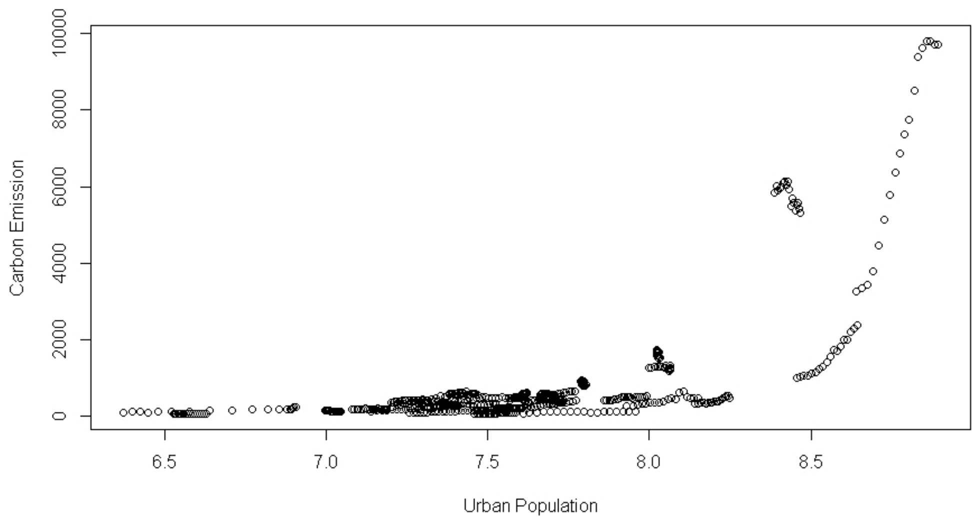

Figure 4: Urban Population vs CO2 emissions plot

We then compute a correlation plot for all the variables, which shows that the variables Labour Force, Total Energy Consumption, Urban Population and Oil Product Consumption per person have a high correlation with the CO2 emissions. This can be seen from Figure 5

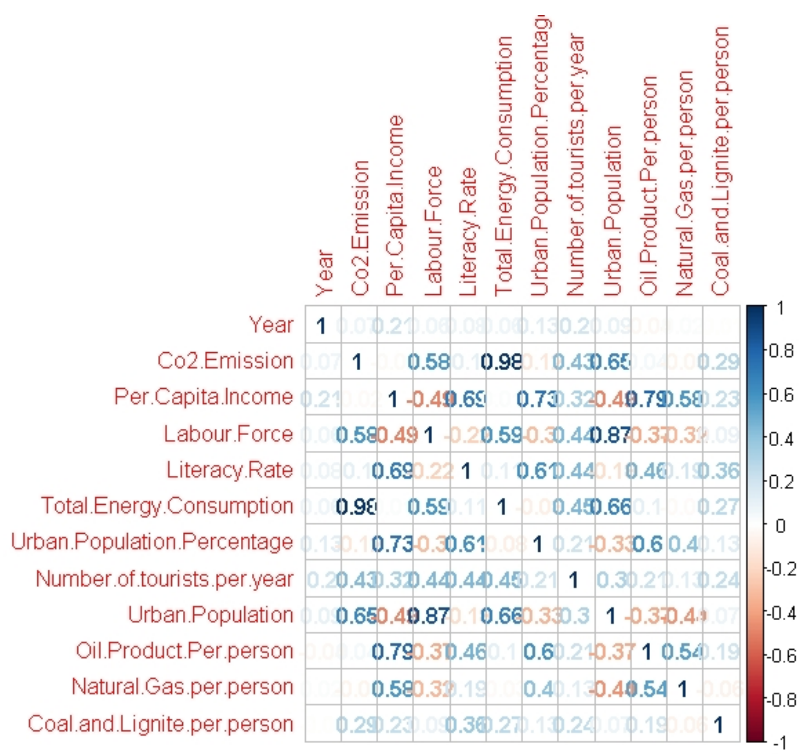

Figure 5: Correlation plot between all the variables

5 Research Methodology

The data is gathered for 32 countries from the year 1999-2016. To find the best predictive model for our data, we implore various predictive models to deduce the model which fits best based on the mean square errors that the model produces. The variables for our data are as follows

- Real GDP Per Capita
- CO2 emissions in metric tons per year
- Urban Population
- Labour force
- Literacy rate in percentage
- Total Energy Consumption in Millions of tons of oil equivalent (Mtoe)
- Urban Population percentage
- Number of tourists per year
- Oil Product Consumption per person in tons
- Natural Gas Domestic Consumption per person in cubic metres
- Coal and Lignite Domestic Consumption per person in tons

Models used

5.1 Linear Regression

This model assumes a linear relationship between the response variable (CO2 emissions in our case), and the independent variables (such as GDP, population, per capita income, total energy consumption, etc). This model uses the least squares approach which chooses values of the unknown constants to minimize the Residual Sum of Squares (RSS). [19] Linear regression uses the form

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_n x_{in} + \epsilon_i \qquad\qquad i = 1, ..., n$$

Where Y is our response variable and x_1, x_2, . . . , x_n are predictor variables or independent variables. The terms $\beta_0, \beta_1, \beta_2, ..., \beta_n$ are unknown constants representing the intercept and slopes and ϵ is known as the error term

5.2 Polynomial Regression

This type of regression model analyses the relationship between independent variables x and the dependent variable Y, and is modelled as the n^{th} degree polynomial in x. This is similar to linear regression, but instead creates new variables such as $x_1 = x$, $x_2 = x^2$, . . . , $x_n = x^n$, and then treats it like a multiple linear regression.

$$y = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + ... + \beta_n x_i^n$$

This regression is not concerned with the coefficients but is interested in the fitted function values at any value x0. To pick the degree of the polynomial function, we either pick a reasonably low value, or use the method of cross-validation to select a suitable value of n

$$fx_0 = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 + \hat{\beta}_3 x_i^3 + ... + \hat{\beta}_n x_i^n$$

5.3 Generalized Additive Models

This model is a generalized linear model in which the linear predictors depend linearly on unknown smooth functions of the predictor variables, and focus on the inference about these smooth functions. To allow for non-linear relationship, each linear component $\beta_j x_{ij}$ with a smooth non-linear function $f_j(x_{ij})$ [20]. GAM allows for flexible nonlinearities in several variables while retaining the additive structures of the linear models.

$$y_i = \beta_0 + \sum_{j=1}^{P} f_j x_{ij} + \epsilon_i$$

5.4 Ridge and Lasso Regression

Ridge and Lasso are some techniques to reduce model complexity and prevent over-fitting which may result from simple linear regression

- Ridge regression shrinks the coefficients and it helps to reduce the model complexity and multi-collinearity. It alters the ordinary least squares function by adding a penalty equivalent to the square of the magnitude of the coefficients [21]

$$\sum_{i=1}^{M}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{M}(y_i - \sum_{j=0}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=0}^{p}{\beta_j}^2$$

- The Lasso minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. Because of the nature of this constraint it tends to produce some coefficients that are exactly 0 and hence gives interpretable models [22]

$$\sum_{i=1}^{M}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{M}(y_i - \sum_{j=0}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=0}^{p}|\beta_j|$$

5.5 Multivariate Adaptive Regression Splines

MARS is a form of non-parametric regression analysis which can be seen as an extension of linear models that automatically models nonlinearities and interactions between models. The model is a weighted sum of basis functions Bi(x) [23]. Each ci is a constant coefficient. Each basis function takes the form of either a constant, or a hinge function, or a product or two or more hinge functions that define the intercepts and knots for the values of variables.

$$f(X) = \beta_0 + \sum_{m=1}^{M}\beta_m h_m(X)$$

We used MARS as it effectively uncovers important data patterns and relationships which are too difficult to reveal for other regression models. MARS builds model by piecing together a series of straight lines with each line allowed its own slope. This point where the linear regression model line is shifted to a different regression line is defined by hinge functions. MARS forms the model in two phases: Forward and Backward pass, where the forward pass generally overfits the data, and the backward prunes the model. The backward pass uses generalized cross validation to compare the performance of model subsets to choose the best subset.

$$GCV(\lambda) = \frac{\sum_{i=1}^{N}(y_i - f_\lambda(x_i))^2}{(1 - \frac{M(\lambda)}{N})^2}$$

5.6 Decision Trees

A decision tree model uses a tree-like model of decisions and their possible consequences. In these tree structures, leaves represent class labels and branches represent conjugations of features that lead to those class labels. The two types of decision trees are-

- **Regression Trees-** The tree model where the response variable can take continuous set of values. This method [24] divides the predictor variables x1, x2, . . . , xn into J distinct and non-overlapping regions R1, R2, . . . , Rj. For every observation in the region Rj the model makes the same prediction, which is the mean of the response value for the observations in training set. We choose to divide the predictor shape into high dimensional boxes for ease of interpretation. The goal is to find the boxes that minimize the Residual Sum of Squares (RSS), given by

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - y_{R_j})^2 + \alpha||$$

The tree building uses recursive binary splitting approach, it begins at the top of the tree and then successively splits the predictor space. The predictor xj and cut point s is selected such that it leads to the greatest possible reduction in RSS. Once an overfitting tree is created, pruning is done to reduce the tree size and number of splits. This offers a lower variance and better interpretation. For pruning the model considers a sequence of trees indexed by a non-negative tuning parameter alpha which acts as a penalty for large tree size. Mod T indicates the number of terminal nodes of the tree T, Rm is the subset of predictor space of mth terminal node, and YhatRM is the mean of training observation in Rm. K-fold cross-validation is used to choose alpha that minimizes the average error.

- **Classification Trees-** The tree model where the response variables can take discrete set of values. Similar as in regression trees, this model also uses recursive binary splitting. In classification trees, instead of RSS, classification error rate is used for making binary splits. Classification error is not sufficient for tree growing, so instead two other measures are used

1. **Gini Index:** It is a measure of total variance across K classes [25]

$$G = \sum_{k=1}^{K} p_{mk}(1 - p_{mk})$$

2. **Cross entropy:** Also known as log loss, measures the performance of the model whose output is a probability value [26]

$$D = -\sum_{k=1}^{K} p_{mk} log(p_{mk})$$

Methods used in Classification and Regression to reduce the variance in the model

1. **Bagging:** We take repeated samples (subsets) from the training data and generate B different bootstrapped training datasets. We then train the bth bootstrapped training set to get f*b(x), the prediction at x, and then we take the average of all predictions [27]

$$f_{bag}(x) = \frac{1}{B} \sum_{b=1}^{B} f^{*}b(x)$$

2. **Random Forest:** Similar to bagging method, but instead it decorrelates the trees to reduce the variance at averaging. After building a number of trees on bootstrapped training samples, at every split, a random selection of m predictors is chosen from a set of p predictors, and only one of those m predictors is used. [28] Typically $m \approx \sqrt{p}$

3. **Boosting:** Bagging creates multiple copies of the training data using bootstrap, and fits a separate tree to each copy, and then combines all of the trees to create a single predictive model. The output boosted model is given by [27]

$$f(x) = \sum_{b=1}^{B} \lambda f^b(x)$$

6 Results

After building and running all the models, we collected the root mean square errors for all as seen in Table 1. It shows that Boosting gives us the lowest value of RMSE, which means it is our best fit model.

Models	RMSE Values
Null	273.5251
Boosting	9.31011
Random Forest	41.07046
GAM	58.41482
Bagging	38.4400
MARS	84.56537
Best subset	211.46
Lasso	211.9223
Ridge	295.18
Linear Regression	210.4044

Table 1: RMSE Values Across All Models

The results also give us how the models are working. At the beginning the variables used were population, coal, oil consumption etc of a whole country. The results of the linear models where pretty good but there was correlation between increase in population and other variables. So, rather than staying with whole populations aggregate we switched the data set to per persons contribution form. The GDP was converted to Real GDP per Capita, Coal and lignite consumption was converted into per persons coal and lignite consumption in tons similarly oil and natural gas consumption's were also converted. After making this changes the results gave us a sight of what effects the carbon emission the most. When we check the Bagging and Random Forest models we can clearly see that the carbon

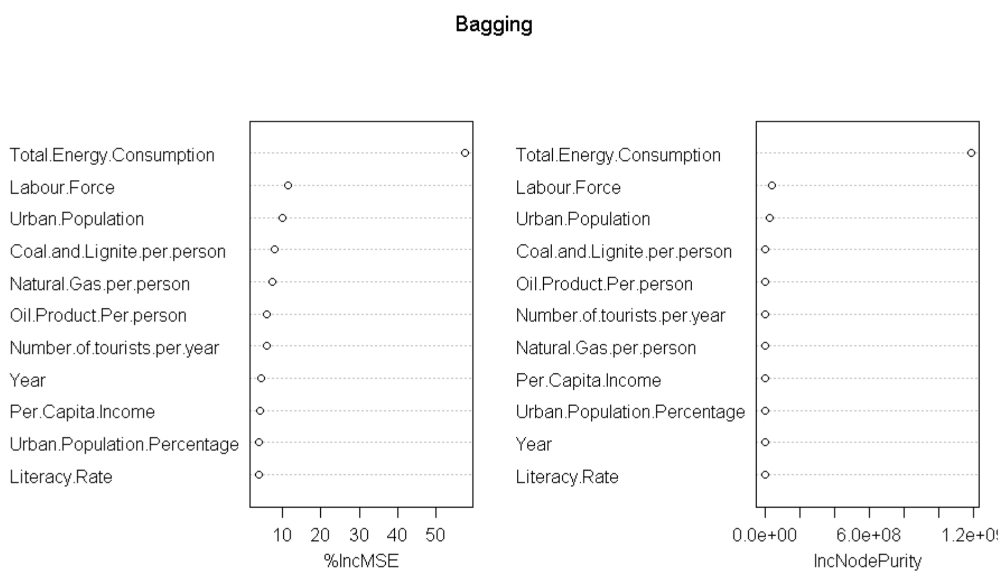

Figure 6: Important Variables for Bagging Model

emission is highly dependent on Total Energy Consumption and Urban Population which is indicative of how the carbon emission is a function of population rather than of economic growth or how much energy a single person uses in an year. This is again validated by the best performing model, the boosting model.

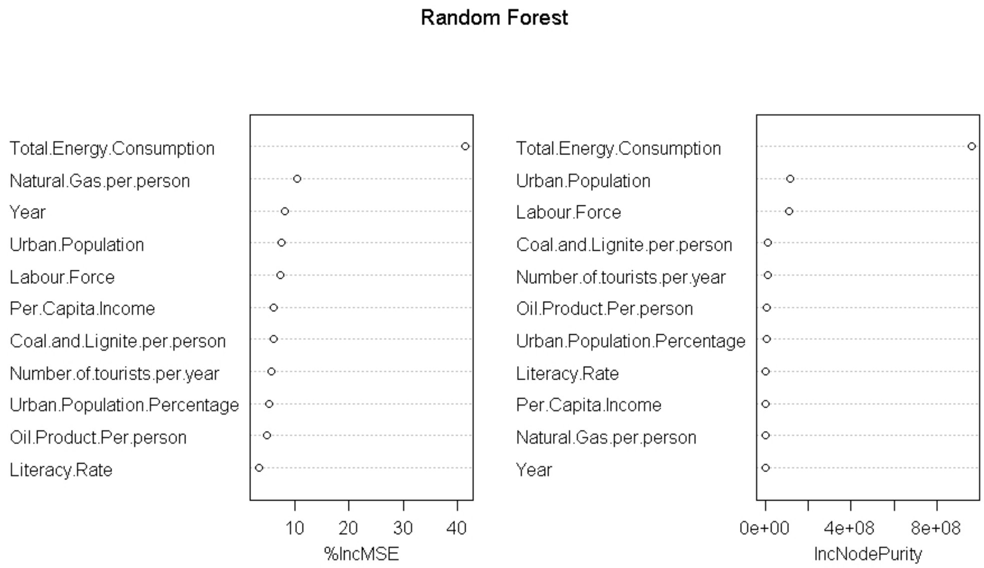

Figure 7: Important Variables for Random Forest Model

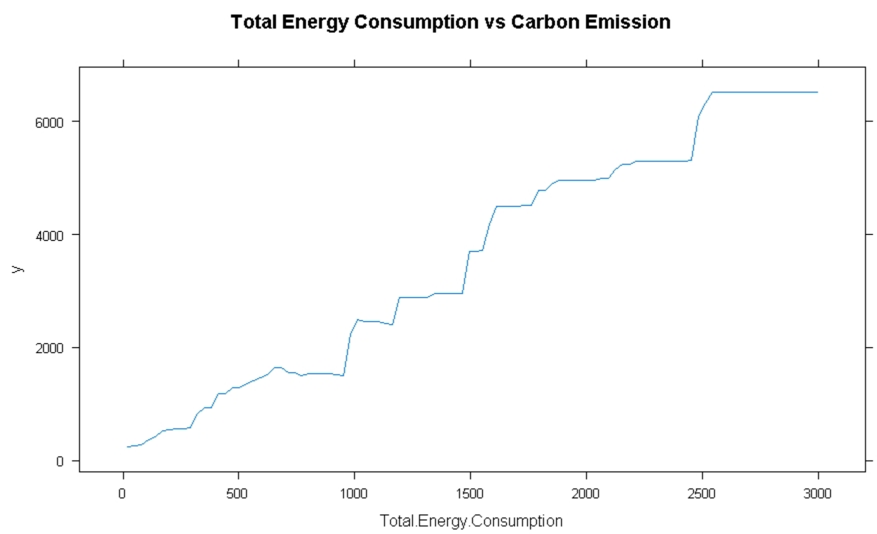

Figure 8: Partial Dependency Plot

Figures 8, 9 and 10 are partial dependencies of the boosting model. When we look closely we can see a trend of how population is effecting the carbon emission. Both the labour force and urban population is a derivative of population and it can be seen that carbon emission only rises with increase in both the variables. The figure 7 shows the relation between Total Energy Consumption and Carbon Emission, the trend is the same as labour force and urban population, with increase in energy consumption the carbon emission is increasing but after certain time we can see that the carbon emission goes constant. The reason behind change in behaviour is maybe an indication of EKC in someway. But to validate that theory, the number of years that is taken into account should be increased.

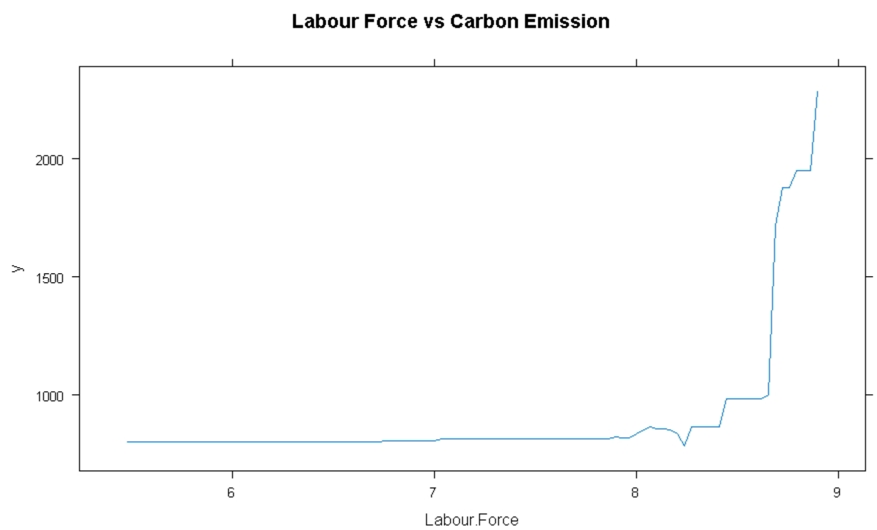

Figure 9: Partial Dependency Plot

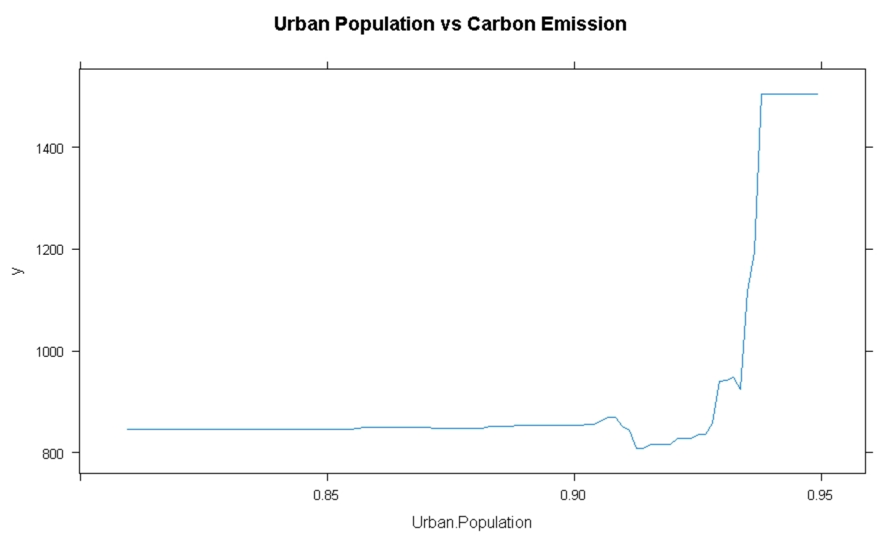

Figure 10: Partial Dependency Plot

7 Conclusion

In this study an attempt was made to investigate relation between Carbon Emission, Real Gross Domestic Production Per Capita, Urban Population, Total Energy Consumption, Coal, Oil and Natural Gas use per person in tons, Number of tourists per year and literacy rate of 32 countries for a span from 1999 to 2016. The study applied different models like Random Forest, Linear model, CART, GAM etc. The Study also tried to check the validation of the Environment Kuznets Curve. Many studies done by researchers such as World Bank [29], Shafik [30], Roberts and Grimes [31], and Unruh and Moomaw [32] testified for the validation of the EKC curve.

The results were not very positive for the validation of EKC curve but it gave us a sight of what the carbon emission is mostly dependent on. The dependency of Carbon Emission is more on the population of a country rather than its economic growth. This hypothesis gets even stronger when we take into account countries like China and India where the Economic Growth is pretty high, but the carbon Emission is also high and there is no downward curve that can be seen. If taken a closer look, they are the two most populated countries in the world. Also, the dependency plots show us after a certain period the carbon emission goes constant with increase in energy consumption and increase in urban population. This may be indicative of some curve like relation between population and carbon emission, which can be validated only by further studies.

Also, the studies showed us how the countries carbon emission is also slightly dependent on the number of tourists that visit that country in per year. This study is also testified by Eyup Dogan and Alper Aslan [33] in their paper.

Finally, all the models indicate a relationship between carbon emission and population growth hence, not validating the Environment Kuznets Curve.

8 Future Research

The Environmental Kuznets Curve is focused on the overall environmental degradation and its relationship with economic growth. Our current work focuses on only one form of environmental degradation, with is Carbon dioxide emissions, which may not fully encapsulate the EKC. There may be environmental causes that relates more to the EKC curve compared to CO2 emissions.

The future work should be something that encapsulates two areas. First one being the addition of more variables such as deforestation, natural calamities, number of vehicles on road etc. This will help us also understand how the other types of environmental degradation are also occurring. The second part being using more number of countries and a bigger time span to check if the relation between carbon emission and population growth still holds true.

References

[1] The National Academy of Sciences NAS. The cost of energy - environmental impact, n.d.

[2] David I Stern. International society for ecological economics internet encyclopaedia of ecological economics the environmental kuznets curve. *Department of Economics, Rensselaer Polytechnic Institute*, 2003.

[3] Ottmar Edenhofer. *Climate change 2014: mitigation of climate change*, volume 3. Cambridge University Press, 2015.

[4] Hsiao-Tien Pao and Chung-Ming Tsai. Multivariate granger causality between co2 emissions, energy consumption, fdi (foreign direct investment) and gdp (gross domestic product): evidence from a panel of bric (brazil, russian federation, india, and china) countries. *Energy*, 36(1):685–693, 2011.

[5] Peter Pedroni. Panel cointegration: asymptotic and finite sample properties of pooled time series tests with an application to the ppp hypothesis. *Econometric theory*, 20(3):597–625, 2004.

[6] JiWu Sun. The nature of co2 emission kuznets curve. *Energy policy*, 27(12):691–694, 1999.

[7] Jenny Cederborg and Sara Snöbohm. Is there a relationship between economic growth and carbon dioxide emissions?, 2016.

[8] Kuan-Min Wang. Modelling the nonlinear relationship between co2 emissions from oil and economic growth. *Economic Modelling*, 29(5):1537–1547, 2012.

[9] Bismark Ameyaw and Li Yao. Analyzing the impact of gdp on co2 emissions and forecasting africa's total co2 emissions with non-assumption driven bidirectional long short-term memory. *Sustainability*, 10(9):3110, 2018.

[10] Emmanouil Hatzigeorgiou, Heracles Polatidis, and Dias Haralambopoulos. Co2 emissions, gdp and energy intensity: a multivariate cointegration and causality analysis for greece, 1977–2007. *Applied Energy*, 88(4):1377–1385, 2011.

[11] VGR Chandran Govindaraju and Chor Foon Tang. The dynamic links between co2 emissions, economic growth and coal consumption in china and india. *Applied Energy*, 104:310–318, 2013.

[12] Jonathan L Ramseur. Us carbon dioxide emissions trends and projections: Role of the clean power plan and other factors. *US Congressional Research Service*, 2017.

[13] Wenwen Wang, Man Li, and Ming Zhang. Study on the changes of the decoupling indicator between energy-related co2 emission and gdp in china. *Energy*, 128:11–18, 2017.

[14] Usama Al-Mulali. Investigating the impact of nuclear energy consumption on gdp growth and co2 emission: A panel data analysis. *Progress in Nuclear Energy*, 73:172–178, 2014.

[15] Wei Sun and Jingyi Sun. Prediction of carbon dioxide emissions based on principal component analysis with regularized extreme learning machine: The case of china. *Environmental Engineering Research*, 22(3):302–311, 2017.

[16] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme learning machine: theory and applications. *Neurocomputing*, 70(1-3):489–501, 2006.

[17] Rafał Kasperowicz. Economic growth and co2 emissions: The ecm analysis. *Journal of International Studies*, 8(3):91–98, 2015.

[18] Data.Worldbank. World bank open data.

[19] Jeffrey M Stanton. Galton, pearson, and the peas: A brief history of linear regression for statistics instructors. *Journal of Statistics Education*, 9(3), 2001.

[20] Trevor Hastie and Robert Tibshirani. Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398):371–386, 1987.

[21] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[22] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[23] Jerome H Friedman et al. Multivariate adaptive regression splines. *The annals of statistics*, 19(1):1–67, 1991.

[24] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[25] Max O Lorenz. Methods of measuring the concentration of wealth. *Publications of the American statistical association*, 9(70):209–219, 1905.

[26] Robert Tibshirani. *Bias, variance and prediction error for classification rules.* Citeseer, 1996.

[27] Zhi-Hua Zhou. *Ensemble methods: foundations and algorithms.* Chapman and Hall/CRC, 2012.

[28] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[29] World Bank. World development report 1992, Jan 1992.

[30] Nemat Shafik. Economic development and environmental quality: an econometric analysis. *Oxford economic papers*, 46(4):757–774, 1994.

[31] J Timmons Roberts and Peter E Grimes. Carbon intensity and economic development 1962–1991: A brief exploration of the environmental kuznets curve. *World development*, 25(2):191–198, 1997.

[32] Gregory C Unruh and William R Moomaw. An alternative analysis of apparent ekc-type transitions. *Ecological Economics*, 25(2):221–229, 1998.

[33] Eyup Dogan and Alper Aslan. Exploring the relationship among co2 emissions, real gdp, energy consumption and tourism in the eu and candidate countries: Evidence from panel models robust to heterogeneity and cross-sectional dependence. 2017.