**Exploratory data analysis of CO₂ emissions for countries per continent for identification of trends in human development index**

Arun Kaniyamattam1\*, Bolu Oluwalade1\*, Chinazom Chukwuemeka1\*

1 Department of BioHealth Informatics, Indiana University Purdue University, Indianapolis, Indiana, United States of America

\* [arkani@iupui.edu](mailto:arkani@iupui.edu) (AK); [boluwala@iupui.edu](mailto:boluwala@iupui.edu) (BO); [chichukw@iu.edu](mailto:chichukw@iu.edu) (CC)

**Abstract**

As developed countries around the world continue to put in efforts to reduce their carbon footprint by using renewable energy, research has shown the increase in carbon emissions in relation to human development index in developing countries. The purpose of this project is to explore the relationship between human development index and carbon emissions through an exploratory data analysis. This analysis is performed using R programming language in RStudio with the combination of data from 1990 to 2017 on human development index and carbon emissions per capita, sector and source. Countries were split into 3 groups based on human development index levels using a threshold of 0.8 for high level countries. Tests carried out on the dataset include linear regression, ANOVA, Kruskal Wallis, Shapiro Wilks normality test and Spearman’s correlation test. Through the Spearman’s correlation test, results show that there is a significant association between human development index and carbon emissions in low level countries. Sectors with the highest carbon emissions are heat and electricity for high and middle level countries while the transport sector stands prominent in low level countries. It was also discovered that the human development index does not fully explain the variations in carbon emissions per capita and therefore calls for the inclusion of other secluded variables in future research such as population, health outcomes and income.

**Background**

Prior studies have shown that the higher the level of development of a country, the higher the CO2 emissions produced by different industries [1]. Therefore, developing countries would need to strive to maintain minimal CO2 emissions as new technologies are implemented in such countries. However, in the last decade, developed countries have experienced a decline in CO2 emissions as a result of introduction of renewable energy to replace the use of fossil fuels [2].

**Aim and Objectives**

The aim of this study is to identify the associations between Human Development Index (HDI) and CO2 emissions over time in different countries via an exploratory data analysis.

The objective is to conduct an exploratory data analysis of CO2 emissions across sectors, its sources and CO2 per capita in countries around the world over 27 years and examine its relationship with HDI.

Through this study we aimed to answer the following questions:

1. Has there been an increase/decrease in CO2 emissions over time for the past 27 years?
2. Do countries with high HDI have higher CO2 emissions when compared to countries with low HDI?
3. What sectors result in higher CO2 emissions? Are there any other additional features that affect HDI?
4. What is the percentage of CO2 emissions per sector and how is this different in each country class?

**Hypothesis**

* H0: There is no significant association in amount of CO2 emissions in relation to the HDI of a country
* H1: There is a significant association in amount of CO2 emissions in relation to the HDI of a country

**Methodology**

**Dataset description**

The datasets used for this exploratory data analysis are HDI from the years 1997 – 2017 retrieved from the United Nations Development Program [3], and CO2 emissions per capita, sector and source [4] retrieved from our world in data organization’s website. The HDI dataset contained 189 rows for countries and HDI values for the years 1990 to 2017 while CO2 emissions contained over 6000 rows for each separate CO2 dataset from the years 1990 to 2017. HDI was calculated using the uniformly weighted sum with ​1⁄3 contributed by each of the following factor indices:

* Life Expectancy Index
* Education Index; comprising of Adult Literacy Index (ALI) and Gross Enrollment Index (GEI)
* Gross Domestic Product per capita [5]

**Data exploration**

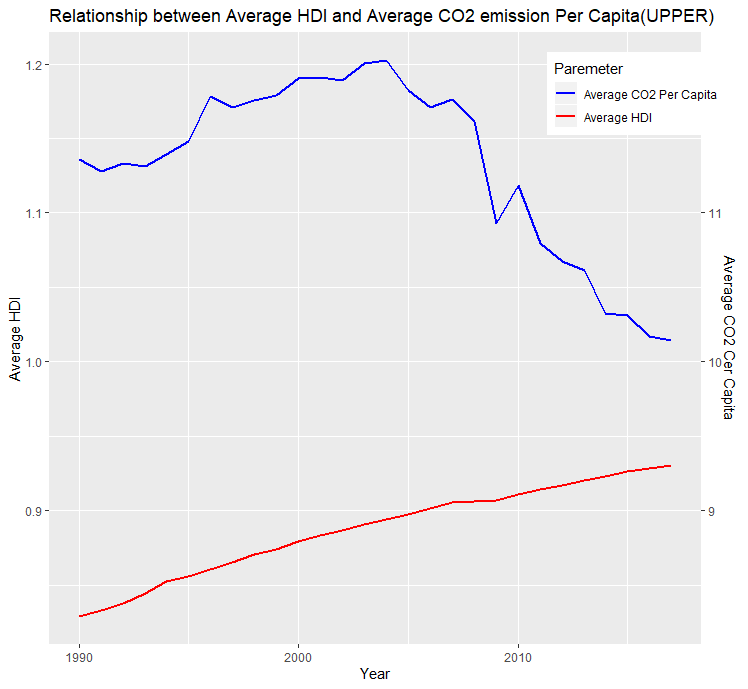
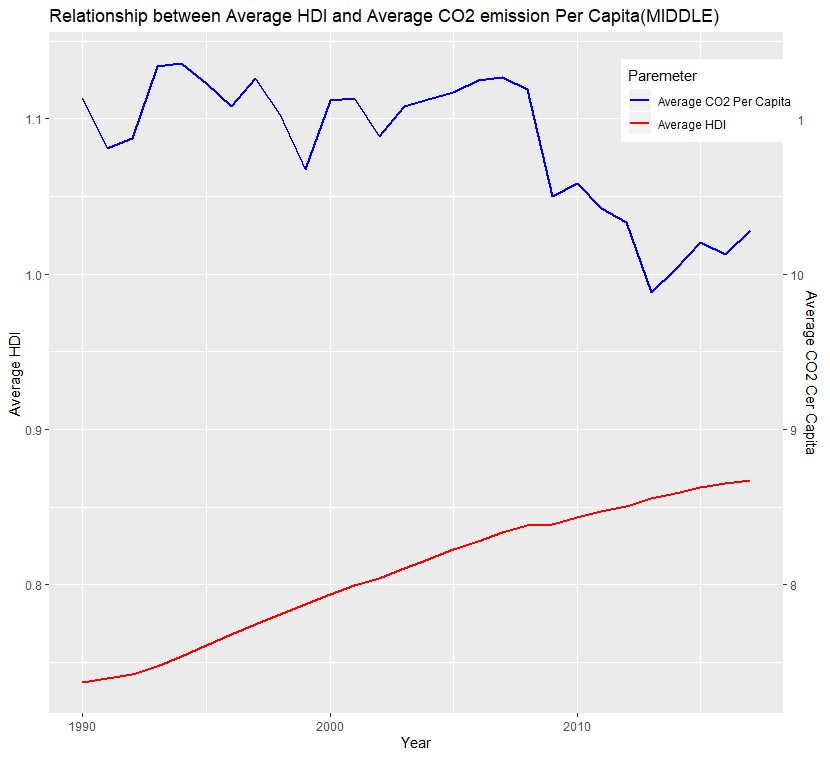
Based on HDI, we split the countries into three groups – upper, middle and lower class. The upper level countries have >0.8 HDI throughout from 1990 to 2017, middle level countries have HDI > 0.8 at any point from 1990 to 2017 and the lower level countries never had HDI > 0.8 from 1990 to 2017. The threshold of 0.8 HDI was selected based on the United Nation’s Development Program definition of developed countries as those with an HDI value of at least 0.8 [5]. Based on the groups we had 12 upper countries, 38 middle and 88 lower HDI level countries. We then developed a sample dataset containing all the countries of interest from the main dataset and checked for missing values. Next, we merged CO2 per capita based dataset with HDI based on the countries for the selected years. We computed the combined average for HDI and CO2 for each country within the period and checked the correlation between CO2 per capita and HDI using Spearman’s method because the dataset was non-parametric (which was confirmed with Shapiro wilk’s test). This was done separately for each group. We also created visualizations for the dataset across each HDI group to assess the relationship between HDI and CO2. Next, we created another subset based on the year intervals 1997, 2000, 2010 and 2017 with the intention to understand if the differences in CO2 emissions per capita can be explained by HDI levels and country groups at the four time periods. In addition to normality and correlation tests, we conducted the following statistical tests:

* **One-way Anova**: In CO2 emissions per sector (in tonnes) category, we aim to understand if the differences in the levels of CO2 emitted from each sector is statistically significant. Anova was chosen because the data met the Anova assumptions including being normally distributed. The sectors include transportation, electricity and heat, manufacturing & construction, residential building, commercial and public services.
* **Kruskal Wallis ranked test**: Because the data was non-parametric, this was appropriate for determining if there is a significant difference in the total CO2 emissions by source across the three different HDI levels. The sources include Cement, Flaring, Oil, coal and Gas.
* **Linear regression:** to understand if all the differences in CO2 emmsion per capita could be explained with HDI and HDI levels. The value of the adjusted R-squared would help provide a better understanding of this.

**Results**

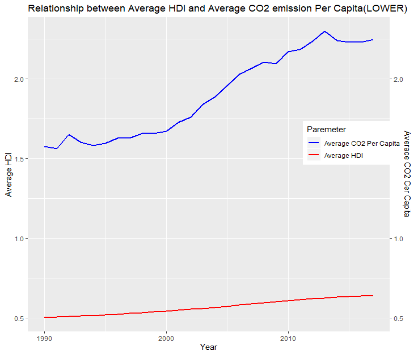
Results show that there is a strong positive correlation (Spearman’s rho=0.9) between the average HDI and average CO2 per capita for all the countries in our dataset. On classification into the different categories, we found that the lower HDI countries have the strongest positive correlation between the average HDI and Average CO2 per capita (Spearman’s rho= 0.83), middle had 0.21 while the high HDI countries had a weak negative correlation (Spearman’s rho= -0.07).

This correlation can also be shown in the figures below. For upper level countries as seen in Fig 1, there is a constant increase in average CO2 emissions as the HDI steadily rises from the year 1997 to 2013. However, from 2013, there is a drop in CO2 emissions, and this continues for the subsequent years. In Fig 2 for middle level countries, a similar trend is seen in but not as pronounced as the upper level countries.

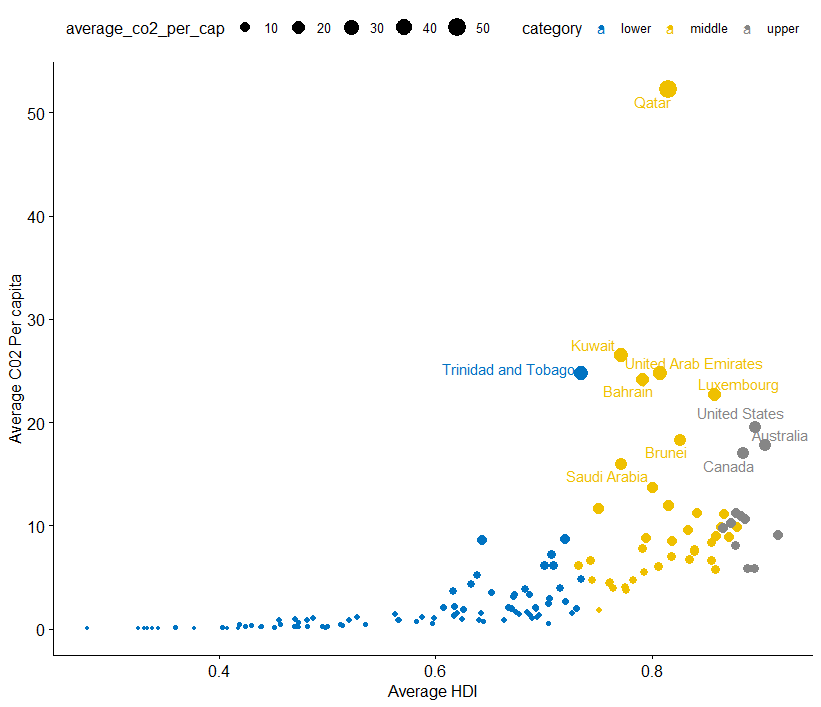
 

**Fig 1: Average HDI and Average CO2 per capita Fig 2: Average HDI and Average CO2 for middle level countries from 1997 -2017. (Upper level) for middle level countries from 1997-2017**

However, in Fig 3 for lower level countries, CO2 emissions steadily rises as HDI rises and this continues throughout the year period.



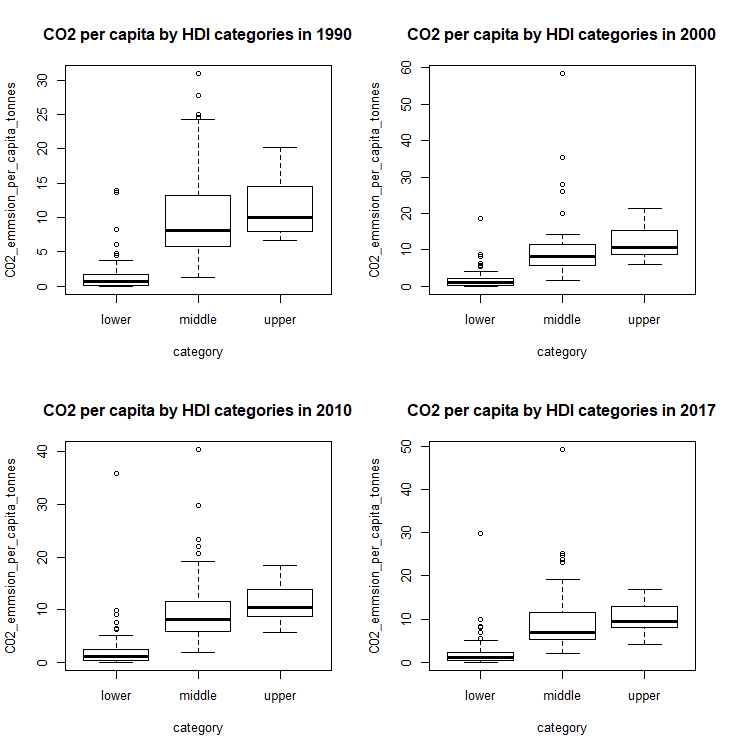
**Fig 3: Relationship between Average HDI and CO2 emission for lower level countries 1997 -2017.**



**Fig 4: Combined HDI and CO2 per capita for all countries**

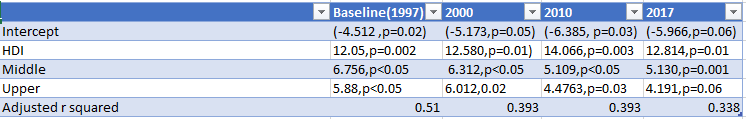
From the trend plots above, it is worth noting that even with the increase in Average CO2 per capita in lower level countries and the downward trend of average CO2 per capita for upper and middle level countries, the maximum average for the lower level(2.4) countries still significantly dwarfs the middle and higher level countries (>10). As a result of the visualizations, we created a scatterplot to see the specific countries with the highest CO2 emissions per capita. In Fig 4, the labelled countries are shown to have a CO2 per capita average of at least 15 CO2 per capita. The colors show the categories of countries based on the HDI (see legend) and the size of the dots corresponds to the size of the average HDI. From the figure, we could see that the countries with the highest average CO2 are in all the HDI categories. One common factor shared by these countries is that they are all oil producing countries.

Visualizations are shown in Fig 5 for the years 1990, 2000, 2010 and 2017 for each HDI level.



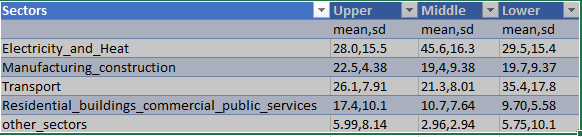
**Figure 5: Boxplots showing the CO2 emitted per capita at baseline (1997), 2000, 2010 and 2017** **for each HDI category.**

**Regression models for selected years**



**Table 1: The linear regression result showing the interaction between HDI and HDI categories with the CO2 per capita as outcome variables.**

The model shows that the differences seen in CO2 emission per capita can be explained with HDI and HDI categories for about 51 percent of the data (adjusted R squared) for the baseline model. The adjusted R squared value continuously reduced in 2000, 2010 and the 2017 model. This shows that the HDI and HDI categories (low, middle and upper) are not the main reason for the different levels of CO2 emissions per capita.



**Table 2: This shows the percentage of CO2 emission by each sector for the HDI categories.**

**The one-way ANOVA**

For all three HDI levels, the ANOVA p-values were less than 0.05 which means there was a statistically significant difference in the average CO2 emissions for the sectors. Post hoc test (Tukey test) was performed for all three to identify the specific sectors that have statistically significant difference in average CO2. The outcome of the Tukey results is:

* Upper Level Countries- Only the differences in the other sectors and electricity and heat, other sector and manufacturing and construction and other sector and transportation are statistically significant (p<0.05).
* Middle Level Countries: all the differences in the sectors are statistically significant (p<0.05) except the difference between transport and manufacturing (p=0.9)
* Lower Level Countries: all the differences in the sectors are statistically significant (p<0.05) except the difference between residential building, commercial public services and other sectors (p=0.3)

**The Kruskal Wallis**

For all three HDI levels, the p-value was less than 0.05 which means there was a statistically significant difference in the average CO2 for each source. Post-Hoc (Pairwise Wilcox test with Bonferroni correction) test was performed for all three levels and it showed the following results:

* Upper Level Countries: only the differences the CO2 emissions in tonnes in oil and cement, coal and cement, coal and flaring gas and flaring and oil and flaring sources are statistically significant (p<0.05).
* Middle Level Countries: all the differences the CO2 emissions in tonnes except in gas and coal, coal and oil and oil and gas sources are statistically significant (p<0.05).
* Lower Level Countries: all the differences the CO2 emissions in tonnes except in gas and coal, gas and cement and coal and cement sources are statistically significant (p<0.05).

**Discussion:**

Human development index has often been linked to the levels of CO2 production. In this work we aimed to examine this association. From the results, the levels of CO2 produced is higher in countries with HDI above 0.8 from 1997 to 2017 and HDI of 0.8 at any year between the time period produced between an average 10 and 12 CO2 per capita. These countries belong to the upper and middle categories respectively. There has been a decrease in the CO2 per capita from 2013 for upper level countries. This could be attributed to the push towards a global greener environment by these countries. This trend is also noticed in the middle level countries. In the lower level countries, there is a strong association between the HDI and CO2 per capita as both exhibit an increasing trend from 1997 to 2017. Although, this trend exists, the quantity of CO2 per capita is significantly smaller (less than 3) than those produced by middle and upper level countries. We also discovered that the countries with the highest CO2 per capita are oil producing countries and this cuts across the lower, upper and middle HDI categories. Our analysis also shows that there are significant differences between the CO2 produced in tonnes across the different sources (oil, gas, coal, cement and flaring) with coal producing on average the highest total CO2 emissions from 1997 to 2017. There is also a statistically significant difference in the percentage of CO2 emitted across the different sectors such as manufacturing and construction, electricity and heat, other sources, residential building, commercial and public services and transport. Heat and Electricity was the highest sector for percentage CO2 emission in upper and middle HDI countries while the transport sector produces the highest percentage CO2 for the lower HDI countries. Across all the categories of HDI, Coal was the highest source of CO2 emission (tonnes). This is followed by oil, gas, cement and flaring respectively

**Conclusion:**

This work shows that there is an association between HDI and CO2 emission. This association is also determined by the level of development of each country. The source of the CO2 emissions is also different based on the HDI of each country. We therefore reject the null hypothesis that there is no association between HDI and CO2 emissions.

**Challenges**:

Selection of the appropriate statistical method proved initially challenging but we were able to select the required ones based on our understanding of the data.

**References**

1. Costa L, Rybski D, Kropp JP. A Human Development Framework for CO2 Reductions. PLoS ONE [Internet]. 2011 Dec 21 [cited 2019 Dec 12];6(12). Available from: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0029262>
2. Quéré CL, Korsbakken JI, Wilson CJ, Tosun JG, Andrew RP, Andres RPV, et al. Drivers of declining CO2 emissions in 18 developed economies. Nature Climate Change [Internet]. 2019 Feb 25 [cited 2019 Dec 12];9(3):213–7. Available from: <https://www.nature.com/articles/s41558-019-0419-7>
3. Human Development Reports [Internet]. Human Development Data (1990-2018) | Human Development Reports. [cited 2019 Dec 12]. Available from: <http://hdr.undp.org/en/data>
4. Roser M. Human Development Index (HDI) [Internet]. Our World in Data. 2014 [cited 2019 Dec 12]. Available from: <https://ourworldindata.org/human-development-index>
5. World population review. [Internet] Human Development Index (HDI) by Country 2019; 2019 [Accessed 12 Dec. 2019]. Available at: <http://worldpopulationreview.com/countries/human-development-index-(hdi)-by-country/>