Machine Learning for Earth System Sciences

Identifying Causal factors behind Global Warming

Under the guidance of

Prof. Adway Mitra

Indian Institute of Technology Kharagpur

Spring 2022

Submitted By:

Paras Jain (18IM10018)

Nitish Kumar (18EE35011)

Sarthak Vijay (18NA30018)

Ravindra Patel (18NA30014)

Ishan Bhattacharya (18AG10013)

Amber Shekhar Shrivastava (18NA30003)

Abstract

This paper tries to pinpoint the exact likely cause of global warming by analysing the 60 years data on various metrics. When looking at all the evidence, there is a large scientific consensus that humans are the leading cause of climate change. In their latest report, the Intergovernmental Panel on Climate Change(IPCC) stated unequivocally that human activity is the cause of global warming. Natural climate cycles can change the temperature of Earth, but the changes we are seeing are happening at a scale and speed that natural cycles cannot explain. These cycles affect the global temperature for years, or sometimes just months, not the 100 years that we have observed. Human activity, such as burning fossil fuels and changing how we use the land, is the leading cause of climate change. We have therefore extracted many Indicators that could likely be the cause of temperature rise in India in the subsequent years using RA and XGB models' correlation and we have determined casualties by the Granger causality tests and Pearl Causality tests respectively.

Contents

1

Introduction

Does population change cause global warming or is it the lifestyle of those people that determines Global warming? For some time, we have known that CO2 is the likely cause of global warming and have tried to curb it as much as possible.

First, let us understand what is causing the climate to change. The climate on Earth has been changing since it formed 4.5 billion years ago. Until recently, natural factors have been the cause of these changes. Natural influences on the climate include volcanic eruptions, changes in the orbit of the Earth, and shifts in the Earth's crust.

Over the past one million years, the Earth has experienced a series of ice ages, including cooler periods (glacials) and warmer periods (interglacials). Glacial and interglacial periods cycle roughly every 100,000 years, caused by changes in Earth's orbit around the sun. For the past few thousand years, Earth has been in an interglacial period with a constant temperature.

However, since the Industrial Revolution in the 1800s, the global temperature has increased at a much faster rate. By burning fossil fuels and changing how we use the land, human activity has quickly become the leading cause of changes to our climate. Some gases in the Earth's atmosphere trap heat and stop it escaping into space. We call these 'greenhouse gases'. These gases act as a warming blanket around the Earth, known as the 'greenhouse effect'.

Greenhouse gases come from both human and natural sources. Gases like carbon dioxide, methane, and nitrous oxide naturally occur in the atmosphere. Others, such as chlorofluorocarbons (CFCs), are only produced by human activity. When short-wave radiation from the sun reaches Earth, most of it passes straight through and hits the surface. The Earth absorbs most of this radiation and gives off longer-wavelength infrared radiation.The greenhouse gases absorb some of this infrared radiation, instead of it passing straight out into space. The atmosphere then emits radiation in all directions, sending some of it back to the surface, causing the planet to heat up. This process is known as the 'greenhouse effect'.

The greenhouse effect is critical to our survival. In fact, without greenhouse gases, Earth would be about 30 degrees colder than it is today. Without greenhouse gases and their warming effect, we wouldn't be able to survive.

However, since the Industrial Revolution, we've been adding more and more greenhouse gases into the air, trapping even more heat. Instead of keeping Earth at a warm, stable temperature, the greenhouse effect is heating the planet at a much faster rate. We call this the 'enhanced greenhouse effect' and it's likely the main cause of climate change.

2

Literature Review

In this project, the factors causing global warming are identified using Regression Analysis and XGBoost and the casualties are found using Pearl and Granger causality tests. For that the weather prediction-based ideas and projects are focused mainly.

“Analysis of Global Warming Using Machine Learning” by Harvey Zheng is also used as a reference. In this project Decision Tree, XGB, random forest is used. The idea has focused on global warming but the explanation of this idea is more complex and not so clear.

Microsoft's paper on dowhy’s CausalModel was also a huge help in getting to understand the know hows of the model and the Granger causality tests of the Statsmodels library also played an important role to complete our model.

“Monthly prediction of air temperature in Australia and New Zealand with machine learning algorithms” by S. Salcedo-Sanz, R. C. Deo, L. Carro-Calvo, B. Saavedra Moreno is also a prediction-based idea. In this idea, the entire focus lies on temperature. The physical factors are not focused which are responsible for Global Warming. In this project SVR and multilayer-perceptron methods are used.

“An Integrated Approach for Weather Forecasting based on Data Mining and Forecasting Analysis” by G.Vamsi Krishna is also used as a reference for this project. The explanation of the paper is very good but it is focused on the overall weather, not a particular thing.

“An interactive predictive system for weather forecasting” by Ayham Omary, Ahmad Wedyan, Ahmed Zghoul, Ahmad Banihai, Izzat Alsmadi is also used as a reference in our project. The explanation is also based on the overall weather not in a particular field.

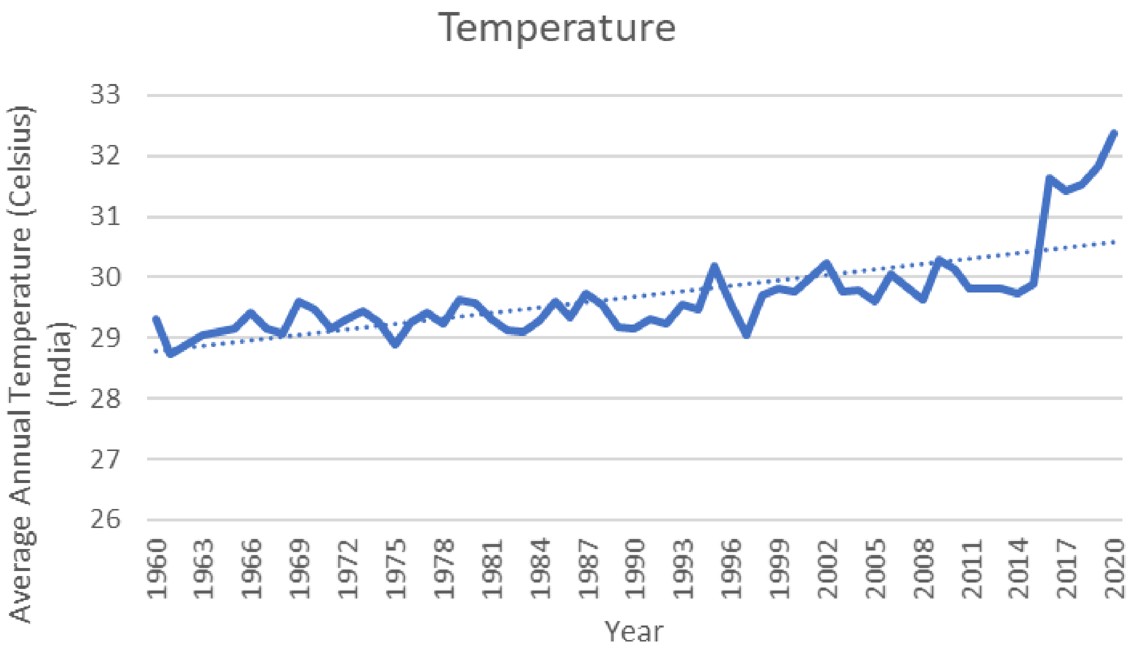
“Multiple regression and Artificial Neural Network for long- term rainfall forecasting using large scale climate modes” by F. Mekanik, M.A. Imteaz, S. Gato-Trinidad, A. Elmahdi is also another prediction-based idea. In this project regression and artificial neural network are used to predict the rainfall. This idea is focussing on only the rainfall. It is not focussing on temperature nor greenhouse gases.

“Development and Analysis of ANN Models for Rainfall Prediction by Using Time-Series Data” by Neelam Mishra, Hemant Kumar Soni, Sanjiv Sharma, AK Upadhyay is also used as a reference. In this project regression, mean square error and MRE are used. This idea also focused only on rainfall not on temperature nor greenhouse gases.

3

Problem Definition

This project is aimed to identify the causal factors that contribute most to global warming using the existing set of tools. The vast number of variables make the problem much more complex. Not just that, but the lack of data in the early years combined with the missing values add more to the problem.
The correlation and causation indicators that lead to the rise in temperature in India is to be determined by using the available set of tools.



As we can see from the above graph, the average annual temperature is rising after each passing year and thus it is important for us to investigate what are the factors that are leading to this cause. It is also equally important to pinpoint the causal factors so as to mitigate the disaster that is unfolding before us.

A lot of work has already been done forecasting the rise of temperature and the sea levels which will be caused by global warming but very little has been done to understand and to investigate the factors that are leading this change. Therefore, it was necessary to have a detailed review on this specific problem based on the available data.

4

Data Collection

The data pertaining to ‘Average annual temperature’, ‘population’, ‘manufacturing’, ‘fuel imports’, consumption of various items for India were retrieved from the World bank database for the year 1960 to 2020.

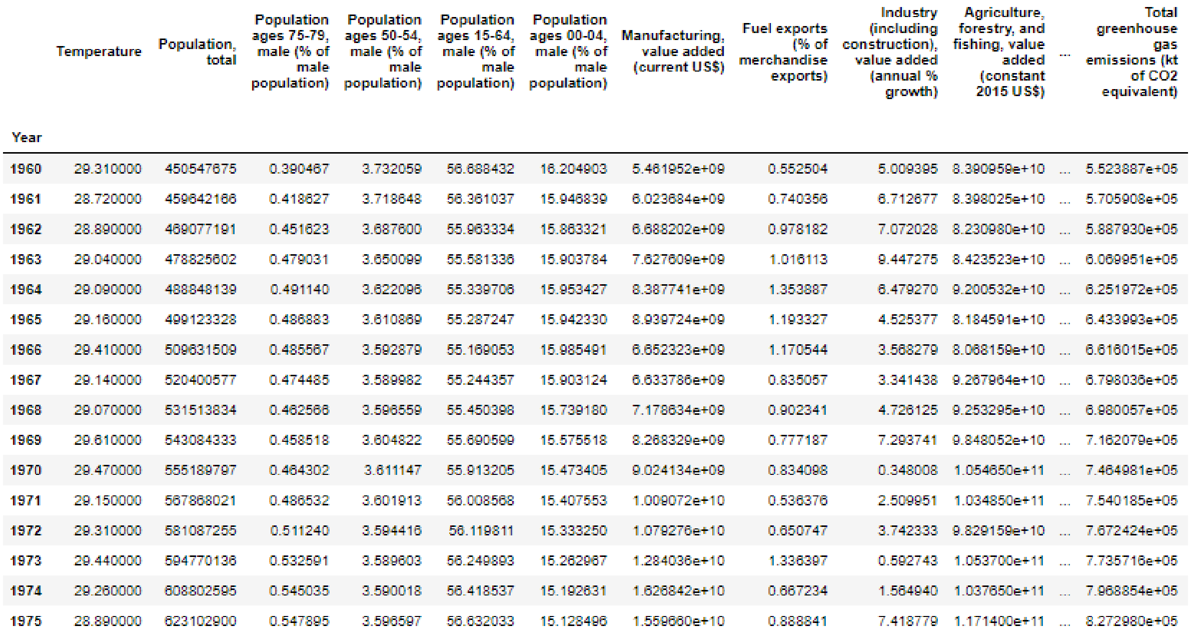
The necessary information was gathered from the accompanying sources-

- World Bank, India Data, 1960-2020
- Temperature Data set-https://www.kaggle.com/datasets/venky73/temperatures-of-india

Limitation of the data -

i) There was some missing data for some variables that were predicted by linear interpolation technique.

ii) Data is available till 2020 only. Therefore, our model did not incorporate trends from 2021. Thus, the projections and estimates will vary slightly with the revised estimates.



	Temperature	Population, total	Population ages 75-79, male (% of male population)	Population ages 50-54, male (% of male population)	Population ages 15-64, male (% of male population)	Population ages 00-04, male (% of male population)	Manufacturing, value added (current US$)	Fuel exports (% of merchandise exports)	Industry (including construction), value added (annual % growth)	Agriculture, forestry, and fishing, value added (constant 2015 US$)	...	Total greenhouse gas emissions (kt of CO2 equivalent)
Year												
1960	29.310000	450547675	0.390467	3.732059	56.688432	16.204903	5.461952e+09	0.552504	5.009395	8.390959e+10	...	5.523887e+05
1961	28.720000	459642166	0.418627	3.718648	56.361037	15.946839	6.023684e+09	0.740356	6.712677	8.398025e+10	...	5.705908e+05
1962	28.890000	469077191	0.451623	3.687600	55.963334	15.863321	6.688202e+09	0.978182	7.072028	8.230980e+10	...	5.887930e+05
1963	29.040000	478825602	0.479031	3.650099	55.581336	15.903784	7.627609e+09	1.016113	9.447275	8.423523e+10	...	6.069951e+05
1964	29.090000	488848139	0.491140	3.622096	55.339706	15.953427	8.387741e+09	1.353887	6.479270	9.200532e+10	...	6.251972e+05
1965	29.160000	499123328	0.486883	3.610869	55.287247	15.942330	8.939724e+09	1.193327	4.525377	8.184591e+10	...	6.433993e+05
1966	29.410000	509631509	0.485567	3.592879	55.169053	15.985491	6.652323e+09	1.170544	3.568279	8.068159e+10	...	6.616015e+05
1967	29.140000	520400577	0.474485	3.589982	55.244357	15.903124	6.633786e+09	0.835057	3.341438	9.267964e+10	...	6.798036e+05
1968	29.070000	531513834	0.462566	3.596559	55.450398	15.739180	7.178634e+09	0.902341	4.726125	9.253295e+10	...	6.980057e+05
1969	29.610000	543084333	0.458518	3.604822	55.690599	15.575518	8.268329e+09	0.777187	7.293741	9.848052e+10	...	7.162079e+05
1970	29.470000	555189797	0.464302	3.611147	55.913205	15.473405	9.024134e+09	0.834098	0.348008	1.054650e+11	...	7.464981e+05
1971	29.150000	567868021	0.486532	3.601913	56.008568	15.407553	1.009072e+10	0.536376	2.509951	1.034850e+11	...	7.540185e+05
1972	29.310000	581087255	0.511240	3.594416	56.119811	15.333250	1.079276e+10	0.650747	3.742333	9.829159e+10	...	7.672424e+05
1973	29.440000	594770136	0.532591	3.589603	56.249893	15.262967	1.284036e+10	1.336397	0.592743	1.053700e+11	...	7.735716e+05
1974	29.260000	608802595	0.545035	3.590018	56.418537	15.192631	1.626842e+10	0.667234	1.564940	1.037650e+11	...	7.968854e+05
1975	28.890000	623102900	0.547895	3.596597	56.632033	15.128496	1.559660e+10	0.888841	7.418779	1.171400e+11	...	8.272980e+05

A small subset of the input data set.

5

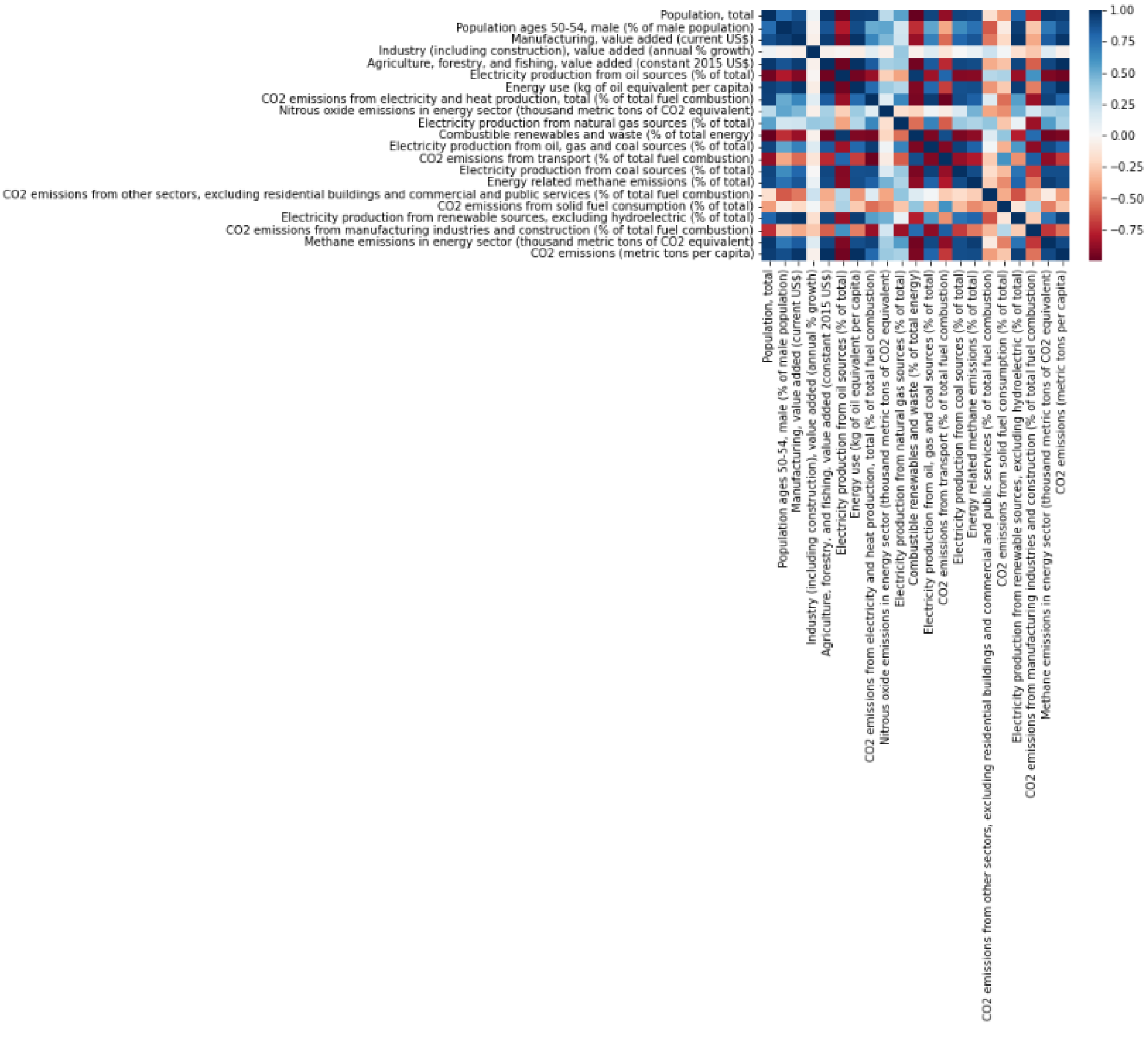
Variable Selection

Open source data on India contained a lot of indicators from the year 1960 to 2020 ranging from population to CO2 emissions(metric tonnes per capita). Out of those, 50 indicators were chosen for our model based on the following criteria -

i) R2 score and correlation matrix

ii) The abundance of data throughout the period.

iii) Emissions from different sources including Agriculture, transport, fishing and energy sector that seemed relevant for our study.



The following heatmap was generated in the process of Feature selection where higher contrast means higher correlation (positive or negative), thus features having high correlation amongst each other were dropped in the process and also those features that seemed irrelevant such as Net foreign assets, secure internet servers etc.

The following indicators seemed fit and thus were included in our model as features-

1. **Temperature**- Year-wise average temperature.
2. **Population**- Year-wise total population of the country for different categories.
3. **Manufacturing, value added (current US$)**- Manufacturing value added (MVA) of an economy is the total estimate of net-output of all resident manufacturing activity units obtained by adding up outputs and subtracting intermediate consumption.
4. **Fuel exports (% of merchandise exports)**- Percentage of fuel exports in total merchandise exports
5. **Industry (including construction), value added (annual % growth)**- Growth value added by industry(including construction) in percentage
6. **Agriculture, forestry, and fishing, value added (constant 2015 US$)**-Growth value added by Agriculture, forestry, and fishing in constant 2015 US$
7. **Agricultural methane emissions (% of total)**- percentage of methane emissions by agriculture in total methane emissions
8. **CO2 emissions from gaseous fuel consumption (% of total)**-percentage of CO2 emissions from gaseous fuel consumption in total CO2 emissions
9. **Electricity production from oil sources (% of total)**-percentage of Electricity production from oil sources in total Electricity production
10. **Energy use (kg of oil equivalent per capita)**-Energy used in kg of oil equivalent per capita
11. **CO2 emissions from electricity and heat production, total (% of total fuel combustion)**-percentage of CO2 emissions from electricity and heat production in total CO2 emissions
12. **Nitrous oxide emissions in energy sector (thousand metric tons of CO2 equivalent)**- Nitrous oxide emissions in energy sector in thousand metric tons of CO2 equivalent
13. **Natural gas rents (% of GDP)**- Natural gas rents in percentage of GDP
14. **Fuel imports (% of merchandise imports)**- Percentage of fuel imports in total merchandise exports
15. **Adjusted savings: energy depletion (current US$)**- Energy depletion is the ratio of the value of the stock of energy resources to the remaining reserve lifetime. It covers coal, crude oil, and natural gas. Its value is measured in US$ as adjusted savings.
16. **CO2 intensity (kg per kg of oil equivalent energy use)**- CO2 intensity in kg per kg of oil equivalent energy use
17. **Electricity production from natural gas sources (% of total)**- Percentage of Electricity production from natural gas sources compared to total electricity production
18. **Combustible renewables and waste (% of total energy)**- Percentage of Combustible renewables and waste in total energy

19. **Electricity production from oil, gas and coal sources (% of total)**-Percentage of Electricity production from oil, gas and coal sources compared to total electricity production
20. **CO2 emissions from transport (% of total fuel combustion)-** percentage CO2 emissions from transport in total fuel combustion
21. **Other greenhouse gas emissions, HFC, PFC and SF6 (thousand metric tons of CO2 equivalent)**
22. **Electricity production from coal sources (% of total)**
23. **Agriculture, forestry, and fishing, value added (annual % growth) -** percentage of annual growth due to Agriculture , forestry and fishing
24. **Agricultural methane emissions (thousand metric tons of CO2 equivalent)** - Amount of methane emissions from Agriculture in tonne equivalent to CO2
25. **CO2 emissions from gaseous fuel consumption (kt)** - Amount of CO2 emission in kilo tonnes from the consumption of gaseous fuel
26. **Electricity production from renewable sources, excluding hydroelectric (kWh)**
27. **Energy related methane emissions (% of total)** - methane emission from energy source in percentage
28. **CO2 emissions from other sectors, excluding residential buildings and commercial and public services (% of total fuel combustion)** - Percentage of CO2 emission of total fuel combustion due to other sectors, excluding residential buildings and commercial and public services
29. **Nitrous oxide emissions (thousand metric tons of CO2 equivalent)** - NO emission in thousand metric tons as equivalent to CO2
30. **CO2 emissions from solid fuel consumption (% of total)** - Percentage of CO2 emission of total due solid fuel consumption
31. **Fossil fuel energy consumption (% of total)** - Amount of fossil fuel energy consumption of total energy consumption in percentage
32. **Agriculture, forestry, and fishing, value added (% of GDP)** - percentage of GDP due to Agriculture, forestry and fishing
33. **Methane emissions (kt of CO2 equivalent)** - Amount of methane emission in kilon tonne equivalent to CO2
34. **CO2 emissions from liquid fuel consumption (kt)** - CO2 emission from the use of liquid fuel in kilo tonnes
35. **Electricity production from renewable sources, excluding hydroelectric (% of total) -** Amount of electricity produced from renewable sources in percent of total excluding hydroelectric
36. **Total greenhouse gas emissions (kt of CO2 equivalent) -** Total greenhouse gas emission in kilo tonnes equivalent of CO2
37. **CO2 emissions from manufacturing industries and construction (% of total fuel combustion) -** percentage of CO2 emissions from construction and manufacturing industries
38. **Nitrous oxide emissions in energy sector (% of total) -** N2O emission from the energy sector in percentage compare to other gases

39. **CO2 emissions from solid fuel consumption (kt)** - CO2 emission from solid fuels usage in kilo tones
40. **Methane emissions in the energy sector (thousand metric tons of CO2 equivalent) -** CH4 emission from the energy sector equivalent to CO2.
41. **Agricultural nitrous oxide emissions (thousand metric tons of CO2 equivalent) -** NO emission from Agriculture as compared to CO2.
42. **CO2 emissions (metric tons per capita) -** Amount of CO2 emission in tons per person
43. **CO2 emissions from liquid fuel consumption (% of total) -** Amount of CO2 emission in percentage of total emission from liquid fuel
44. **Fertilizer consumption (kilograms per hectare of arable land) -** Amount of fertilizer consumed in kg per hectare of arable land.

6
Solution Methodology

In this work, the two popular methods used to identify causal factors behind global warming are RA and XGBoost. In this section, a brief introduction is given about RA and XGBoost.

6.1 Regression Analysis (RA)

RA is a strategy applied for modelling and examining numerical data. Regression analysis uses one or more independent variables to predict the target variable. Linear regression is a **linear model**, e.g. it assumes the target output variable(y) is linearly dependent on one or more input variable (x). More clearly, that the target variable can be predicted using a linear combination of independent variables. Linear regression uses a least squares function to model the observational data. The data model, which represents simple linear regression, can be written as

$$Y = a_1X_1 + a_2X_2 + \ldots + a_nX_n + e$$

where Y is the dependent variable (Average Annual Temperature), X_1, X_2, …., X_n are the independent variables mentioned above, a_1, a_2, …..., a_n are the regression coefficients and e is the error term. The error term represents the "noise" or the random disturbances in the dependent variable which could not be explained and is treated as a random variable. The coefficients of independent input variables are determined to give the best fit model of the data. Generally, the least squares method is adopted to evaluate the best fitting model.

6.1.1 Fitting the Regression Model

The initial input variable selection was based on correlation coefficient of the input variable and the average annual temperature. These correlation coefficients measure the degree of relation between the two variables. Variables having high correlation coefficient with average annual temperature were selected and used for the modelling purpose. MS Excel and Python are used to carry out RA. To select the appropriate set of final input variables, backward elimination technique was carried out iteratively until we get the most significant model.

6.1.2 Avoiding Multicollinearity

A linear regression model encounters the problem of Multicollinearity when some highly correlated independent variables are used in the regression model. Due to this, model interpretation becomes difficult and an overfitting problem arises. Highly correlated independent variables can cause significant change to one another with a small change in them and so the model results fluctuate significantly which may result in a highly unstable model and cannot be relied upon to get accurate predictions. A small change in the data or model will drastically change model outputs which is undesirable. So, in order to have a reliable prediction of causal factors, some independent variables which were highly correlated were discarded to avoid multicollinearity.

6.1.3 Final Regression Model

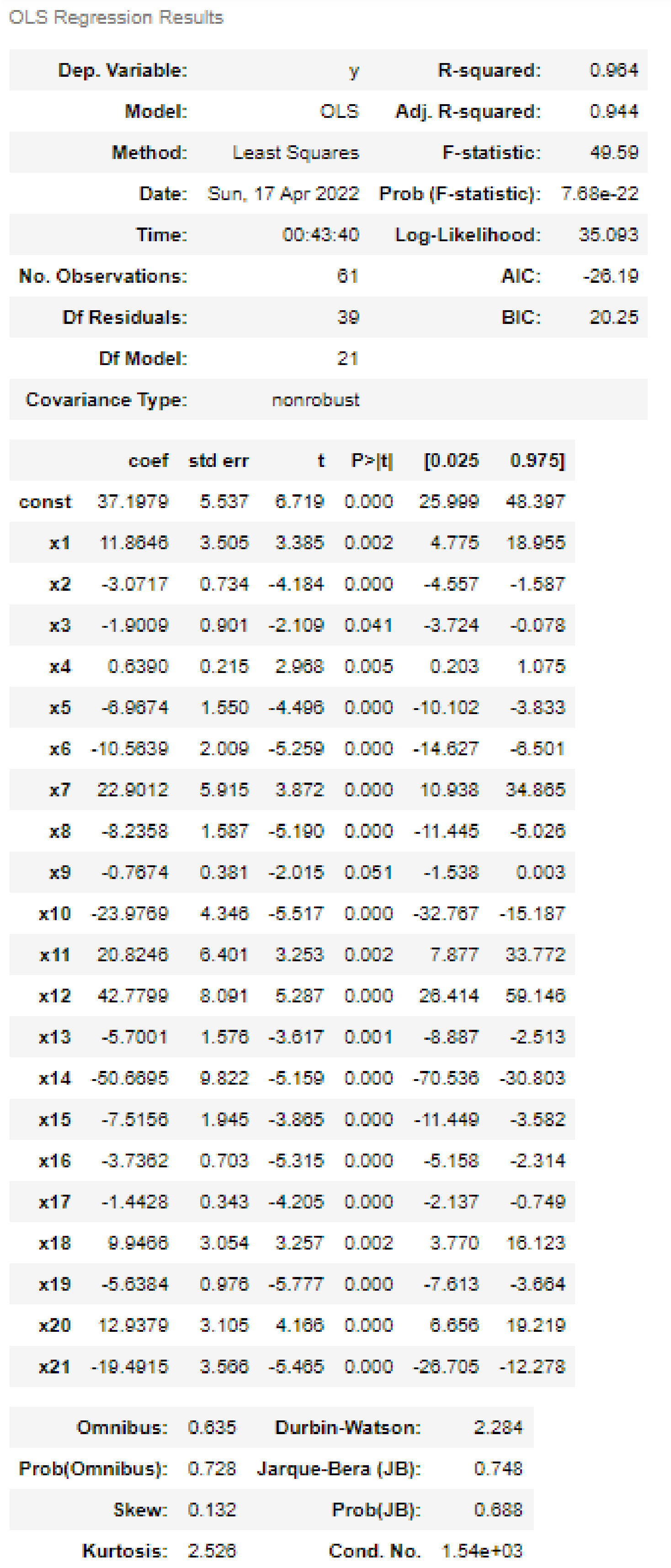
After the iterative process of eliminating variables, the twenty-one input variables that made final cut in the regression model are:

X_1: Population, total	X_2: Population ages 50-54, male (% of male population)
X_3: Manufacturing, value added (current US$)	X_4: Industry (including construction), value added (annual % growth)
X_5: Agriculture, forestry, and fishing, value added (constant 2015 US$)	X_6: Electricity production from oil sources (% of total)
X_7: Energy use (kg of oil equivalent per capita)	X_8: CO2 emissions from electricity and heat production, total (% of total fuel combustion)

X_9: Nitrous oxide emissions in energy sector (thousand metric tons of CO2 equivalent)	X_{10}: Electricity production from natural gas sources (% of total)
X_{11}: Combustible renewables and waste (% of total energy)	X_{12}: Electricity production from oil, gas and coal sources (% of total)
X_{13}: CO2 emissions from transport (% of total fuel combustion)	X_{14}: Electricity production from coal sources (% of total)
X_{15}: Energy related methane emissions (% of total)	X_{16}: CO2 emissions from other sectors, excluding residential buildings and commercial and public services (% of total fuel combustion)
X_{17}: CO2 emissions from solid fuel consumption (% of total)	X_{18}: Electricity production from renewable sources, excluding hydroelectric (% of total)
X_{19}: CO2 emissions from manufacturing industries and construction (% of total fuel combustion)	X_{20}: Methane emissions in energy sector (thousand metric tons of CO2 equivalent)
X_{21}: CO2 emissions (metric tons per capita)	

Table 3: Input Variables for Regression Analysis

Below is the summary of Regression Model:



OLS Regression Results

Dep. Variable:	y	R-squared:	0.964
Model:	OLS	Adj. R-squared:	0.944
Method:	Least Squares	F-statistic:	49.59
Date:	Sun, 17 Apr 2022	Prob (F-statistic):	7.68e-22
Time:	00:43:40	Log-Likelihood:	35.093
No. Observations:	61	AIC:	-26.19
Df Residuals:	39	BIC:	20.25
Df Model:	21		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	37.1979	5.537	6.719	0.000	25.999	48.397
x1	11.8646	3.505	3.385	0.002	4.775	18.955
x2	-3.0717	0.734	-4.184	0.000	-4.557	-1.587
x3	-1.9009	0.901	-2.109	0.041	-3.724	-0.078
x4	0.6390	0.215	2.968	0.005	0.203	1.075
x5	-6.9674	1.550	-4.496	0.000	-10.102	-3.833
x6	-10.5639	2.009	-5.259	0.000	-14.627	-6.501
x7	22.9012	5.915	3.872	0.000	10.938	34.865
x8	-8.2358	1.587	-5.190	0.000	-11.445	-5.026
x9	-0.7674	0.381	-2.015	0.051	-1.538	0.003
x10	-23.9769	4.346	-5.517	0.000	-32.767	-15.187
x11	20.8246	6.401	3.253	0.002	7.877	33.772
x12	42.7799	8.091	5.287	0.000	26.414	59.146
x13	-5.7001	1.576	-3.617	0.001	-8.887	-2.513
x14	-50.6695	9.822	-5.159	0.000	-70.536	-30.803
x15	-7.5156	1.945	-3.865	0.000	-11.449	-3.582
x16	-3.7362	0.703	-5.315	0.000	-5.158	-2.314
x17	-1.4428	0.343	-4.205	0.000	-2.137	-0.749
x18	9.9466	3.054	3.257	0.002	3.770	16.123
x19	-5.6384	0.976	-5.777	0.000	-7.613	-3.664
x20	12.9379	3.105	4.166	0.000	6.656	19.219
x21	-19.4915	3.566	-5.465	0.000	-26.705	-12.278

Omnibus:	0.635	Durbin-Watson:	2.284
Prob(Omnibus):	0.728	Jarque-Bera (JB):	0.748
Skew:	0.132	Prob(JB):	0.688
Kurtosis:	2.526	Cond. No.	1.54e+03

Figure 1: OLS Regression Results

6.1.4 Test for Homoscedasticity

The assumption of homoscedasticity (meaning "same variance") is integral to linear regression models. Homoscedasticity depicts a condition in which the error term (that is, the "noise" or random disturbance in the relationship between the independent variables and the dependent variable) do not change significantly with change in values of an independent variable. When the magnitude of the error term differs across values of an independent variable, the model is said to violate the assumption of homoscedasticity and is said to be heteroscedastic in nature.

Breusch-Pagan test and White's test were carried out for checking for homoscedasticity. "Statsmodels" library in Python is used to carry out this test.

Null Hypothesis (H_0): Homoscedasticity is present (the residuals are distributed with equal variance)
Alternative Hypothesis (H_A): Heteroscedasticity is present (the residuals are not distributed with equal variance)

Results for the White's Test:

The p-value was **0.4397.**
We fail to reject the null hypothesis, so there is no heteroscedasticity.

Results for the Breusch-Pagan's Test:

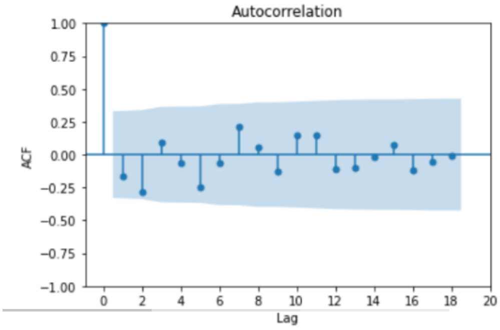
The p-value was **0.7279.**
We fail to reject the null hypothesis, so there is no heteroscedasticity.

6.1.5 Test for Auto-correlation:

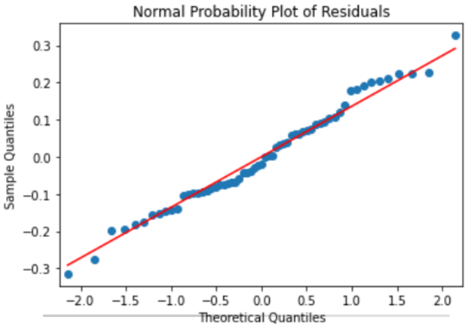
Autocorrelation occurs when a given time series values highly correlates with a lagged version of itself over successive time intervals. It's conceptually similar to the correlation between two different time series, but autocorrelation uses the same time series twice: once in its original form and once lagged one or more time periods.
"Statsmodels" library in python is used to check for Auto-correlation.

Durbin-Watson Test for autocorrelation:

From the regression model summary, the Durbin Watson statistic was calculated to be 2.284 which is between d_L (= 0.716) and d_U (= 2.338), slightly in the zone of indecision. However, the lowest p-value found was 0.03305. We fail to reject the null hypothesis, so there is no autocorrelation.



6.1.6 Checking for normality of residuals:



Normality of residuals is the assumption that the underlying residuals are approximately normally distributed. This is more of a subjective test which depends on how one perceives the results. The Normal Probability Plot of residuals was obtained and the residuals were very close to the straight-line formation, which is enough to say that the residuals are normally distributed.

6.2 XGBoost

XGBoost stands for "Extreme Gradient Boosting". XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements Machine Learning algorithms under the Gradient Boosting framework. Boosting is an ensemble learning technique to build a strong classifier from several weak classifiers in series. Boosting algorithms play a crucial role in dealing with bias-variance trade-off. Unlike bagging algorithms, which only control for high variance in a model, boosting controls both the aspects (bias & variance) and is considered to be more effective.
"Xgboost" library in Python is used to carry this out.

6.2.1 Calculating Feature Importances:

There are several types of importance in the XGBoost - it can be computed in several different ways. The default type is gain. The gain type shows the average gain across all splits where the feature was used. Information Gain is simply the expected reduction in entropy caused by partitioning the examples according to this attribute.

IG(S, a) = H(S) – H(S | a)

where IG(S, a) is the information for the dataset S for the variable a for a random variable, H(S) is the entropy for the dataset before any change and H(S | a) is the conditional entropy for the dataset given the variable a.

The feature_importances_ attribute of the model after training returns the feature importances for the problem.

6.3 Causal Inference

Causal inference is the process of ascribing causal relationships to associations between variables. Statistical inference is the process of using statistical methods to characterize the association between variables. Causality is at the root of scientific explanation which is considered to be causal explanation. However, establishing causal relationships is extremely difficult in spite of substantial advancements made during the past decades. Statistical inference works like a black box and generates the best possible characterization of the relationships between variables. Statistical inference provides estimates of the associations between variables but of course, association does not imply causation, so there is little that statistical inference can provide to establish causation.

6.3.1 Granger Causality

Granger causality is a statistical concept of causality that is based on prediction. According to Granger causality, if a signal X_1 "Granger-causes" (or "G-causes") a signal X_2, then past values of X_1 should contain information that helps predict X_2 above and beyond the information contained in past values of X_2 alone. Its mathematical formulation is based on linear regression modeling of stochastic processes (Granger 1969). More complex extensions to nonlinear cases exist, however these extensions are often more difficult to apply in practice.

X is expressed as an autoregressive process (linear function of past values of itself)

$$X(t) \sim a1X(t-1) + a2X(t-2) + a3X(t-3) + \dots\dots$$

Again, the values of Y are used to estimate the value of X(t)

$$X(t) \sim a1X(t-1) + a2X(t-2) + a3X(t-3) + \dots\dots + b1Y(t-1) + b2Y(t-2) + b3Y(t-3) + \dots\dots$$

If the estimation improves, Y Granger-causes X.

“Statsmodels” library in Python is used to carry out this test.

6.3.2 Pearl Causality

Pearl's work is based on ‘Structural Causal Models’, which is a triple M = (U, V, F). In this model U is the collection of the exogenous (background, or driving) unobserved variables, V is the collection of endogenous (determined in some way by variables from U and V) variables, and F is a collection of functions f1, f2, ..., for each Vi in V. The variable Vi is fully determined as Vi = fi(U, V \ Vi), that is the arguments to fi are some of the variables in U, and some of the variables in V, but not Vi itself. In order to turn this into a probabilistic model, U is augmented with a probability distribution. An example is given where U1 is a court order for a man's execution, V are the actions of a captain (V1) and two riflemen (V2,V3) in a firing squad as well as the living/dead state of the person to whom the court order pertains (V3). If the judge orders the man shot (U1 = 'execute'), then this causes the captain to issue the order to fire, which causes the riflemen to shoot the prisoner, and hence causing his death. If the court order is not given, the captain remains silent, the riflemen don't shoot, and the prisoner is left alive.

To define the causal effect, consider two worlds: 1. World 1 (Real World): Where the action A was taken and Y observed 2. World 2 (Counterfactual World): Where the action A was not taken (but everything else is the same) Causal effect is the difference between Y values attained in the real world versus the counterfactual world.

In other words, A causes Y iff changing A leads to a change in Y, keeping everything else constant. Changing A while keeping everything else constant is called an intervention, and represented by a special notation do(A). Formally, causal effect is the magnitude by which Y is changed by a unit interventional change in A:

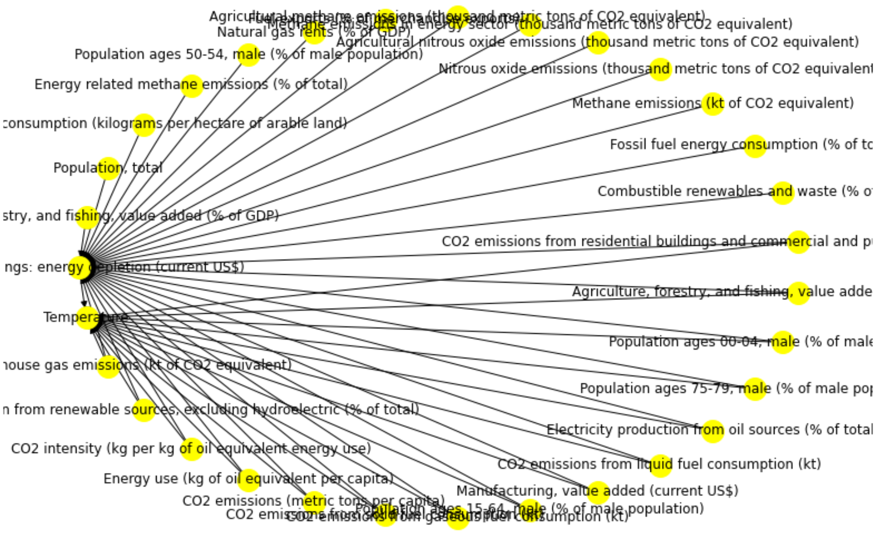
$$CE = E(Y|do(A=1)) - E(Y|do(A=0))$$

A CE value of >=0.05 is taken as the threshold for A to have a causal effect on Y for our analysis,

“Dowhy” library in Python is used to carry out this test.

6.3.3 Structural Causal Model

As defined above, the formation of the Structural Causal Model for Pearl Causality requires extensive domain knowledge. As it is beyond the scope of this project, the confounding variables for the model are estimated using the correlation between the treatment and outcome variables. Variables with high correlation with both the treatment and outcome variables are taken as the common cause variables and those with high correlation only with the treatment variable are taken as the instrument variables.


Agricultural methane emissions (thousand metric tons of CO2 equivalent)
Methane emissions in energy sector (thousand metric tons of CO2 equivalent)
Natural gas rents (% of GDP)
Agricultural nitrous oxide emissions (thousand metric tons of CO2 equivalent)
Population ages 50-54, male (% of male population)
Nitrous oxide emissions (thousand metric tons of CO2 equivalent
Energy related methane emissions (% of total)
Methane emissions (kt of CO2 equivalent)
consumption (kilograms per hectare of arable land)
Fossil fuel energy consumption (% of tc
Population, total
Combustible renewables and waste (% o
stry, and fishing, value added (% of GDP)
CO2 emissions from residential buildings and commercial and p
ngs: energy depletion (current US$)
Agriculture, forestry, and fishing, value adde
Temperature
Population ages 00-04, male (% of male
house gas emissions (kt of CO2 equivalent)
Population ages 75-79, male (% of male po
n from renewable sources, excluding hydroelectric (% of total)
Electricity production from oil sources (% of total
CO2 intensity (kg per kg of oil equivalent energy use)
CO2 emissions from liquid fuel consumption (kt)
Energy use (kg of oil equivalent per capita)
Manufacturing, value added (current US$)
CO2 emissions (metric tons per capita)
Population ages 15-64, male (% of male population)
CO2 emissions from gaseous fuel consumption (kt)

7

Results and Evaluation

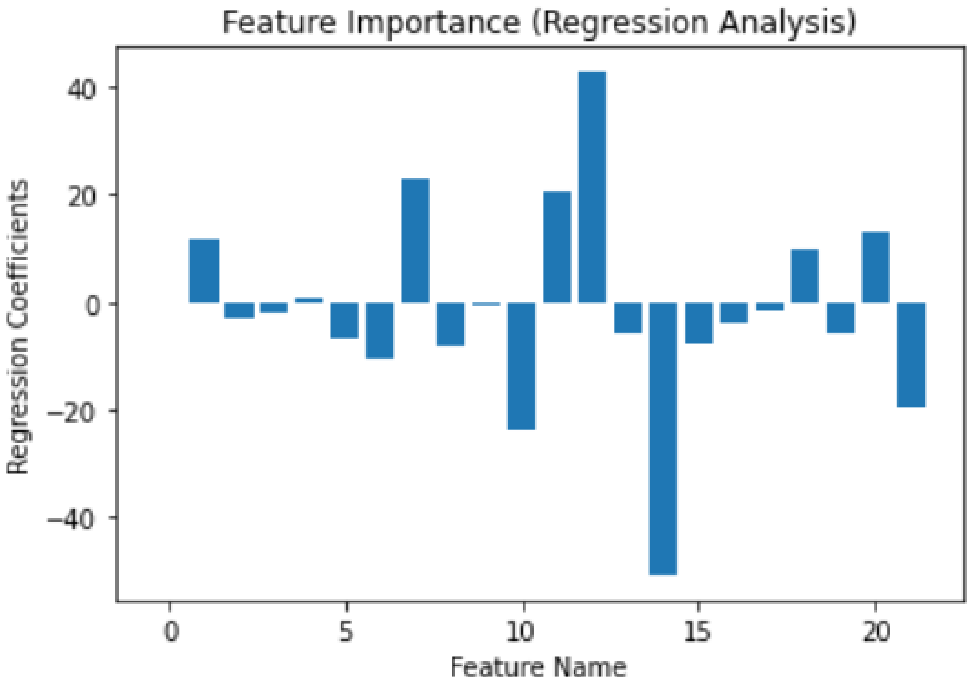
7.1 Regression Analysis(RA) Results

The results that were obtained from Regression Analysis (RA) is as follows :

Variable	**Variable Name**	**Regression Coefficient**	**P-value**
X_1	Population, total	11.8646	0.002
X_2	Population ages 50-54, male (% of male population)	-3.0717	0.000
X_3	Manufacturing, value added (current US$)	-1.9009	0.041
X_4	Industry (including construction), value added (annual % growth)	0.639	0.005
X_5	Agriculture, forestry, and fishing, value added (constant 2015 US$)	-6.9674	0.000
X_6	Electricity production from oil sources (% of total)	-10.5639	0.000
X_7	Energy use (kg of oil equivalent per capita)	22.9012	0.000
X_8	CO2 emissions from electricity and heat production, total (% of total fuel combustion)	-8.2358	0.000
X_9	Nitrous oxide emissions in energy sector (thousand metric tons of CO2 equivalent)	-0.7674	0.051
X_{10}	Electricity production from natural gas sources (% of total)	-23.9769	0.000
X_{11}	Combustible renewables and waste (% of total energy)	20.8246	0.002

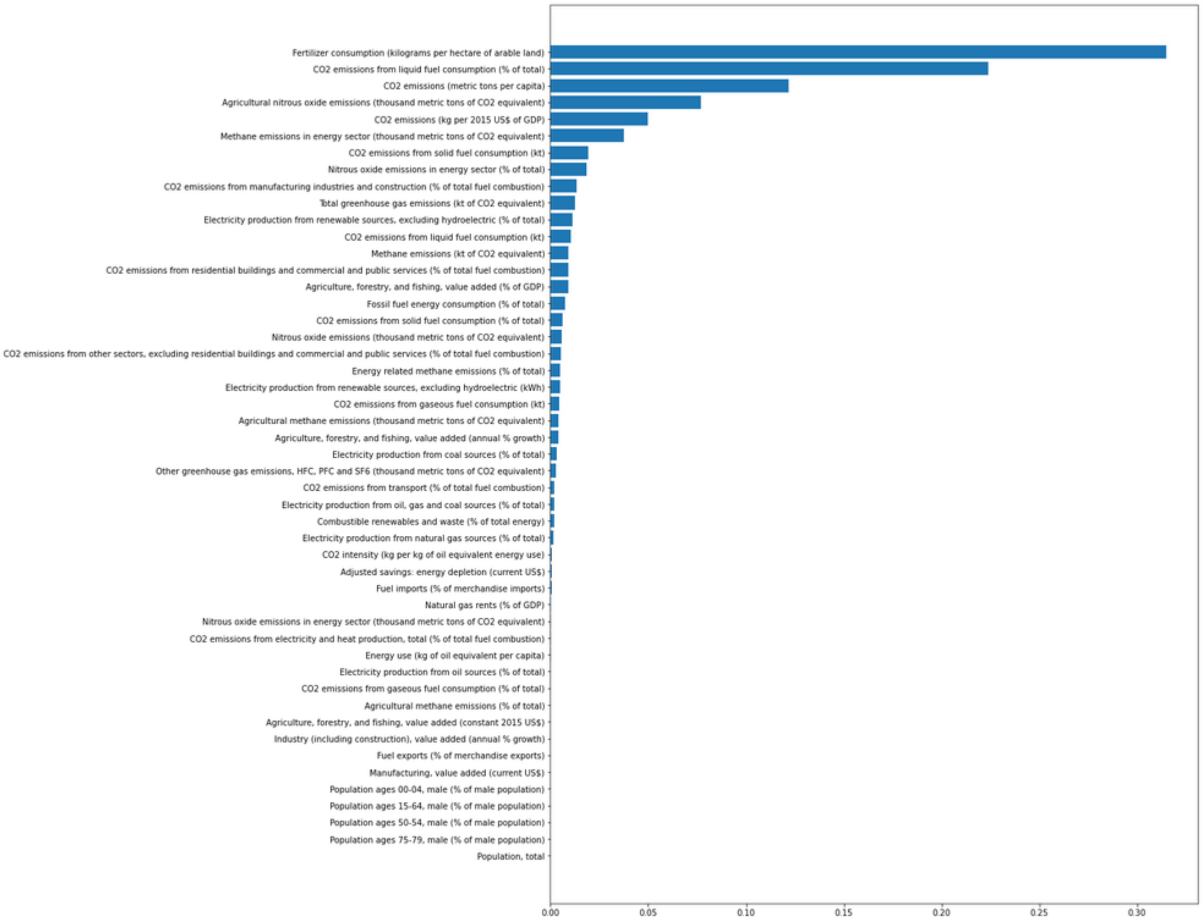
X_{12}	Electricity production from oil, gas and coal sources (% of total)	42.7799	0.000
X_{13}	CO2 emissions from transport (% of total fuel combustion)	-5.7001	0.001
X_{14}	Electricity production from coal sources (% of total)	-50.6695	0.000
X_{15}	Energy related methane emissions (% of total)	-7.5155	0.000
X_{16}	CO2 emissions from other sectors, excluding residential buildings and commercial and public services (% of total fuel combustion)	-3.7362	0.000
X_{17}	CO2 emissions from solid fuel consumption (% of total)	-1.4428	0.000
X_{18}	Electricity production from renewable sources, excluding hydroelectric (% of total)	9.9466	0.002
X_{19}	CO2 emissions from manufacturing industries and construction (% of total fuel combustion)	-5.6384	0.000
X_{20}	Methane emissions in energy sector (thousand metric tons of CO2 equivalent)	12.9379	0.000
X_{21}	CO2 emissions (metric tons per capita)	-19.4915	0.000

The feature importance is generated by Regression Analysis is shown below where the higher the value the more is the correlation with the ‘Average annual temperature’



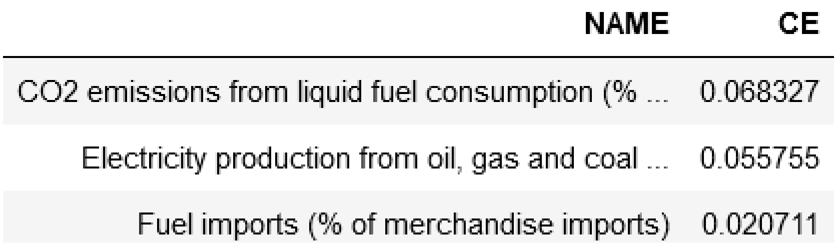
7.2 XGB Results

Our XGBoost model with n_estimators = 1000 gives us the following output for Feature Importances



7.3 Pearl Causality Results

When we used the Pearl Causal Method we get the following indicators that seemed to contribute most to the ‘Average temperature rise’ i.e. most causal along with their CE Score respectively.



NAME	CE
CO2 emissions from liquid fuel consumption (% ...	0.068327
Electricity production from oil, gas and coal ...	0.055755
Fuel imports (% of merchandise imports)	0.020711

7.4 Granger causality test Results

As for the Granger causality test, it produced the following output along with their lag(years) by which the indicator is leading the ‘Annual temperature’

1. ['Fossil fuel energy consumption (% of total)', 2],
2. ['CO2 emissions from liquid fuel consumption (kt)', 2],
3. ['CO2 emissions from manufacturing industries and construction (% of total fuel combustion)', 1],
4. ['CO2 emissions from solid fuel consumption (kt)', 5],
5. ['Methane emissions in energy sector (thousand metric tons of CO2 equivalent)', 5],
6. ['CO2 emissions (kg per 2015 US$ of GDP)', 4],
7. ['CO2 emissions from liquid fuel consumption (% of total)', 5]

Therefore, the Granger causality test found 7 indicators which it thought contributed most to the ‘Annual temperature rise’ along with the lag (in years) in between the effect and the cause.

8

Conclusion

It was observed that our RA model and XGB model gave out a list of indicators that were highly correlated with 'Annual average temperature'. For the RA it lists out Electricity production from oil and gas as the major cause for 'Annual temperature rise' and as for XGBoost, these variables include Fertilizer consumption, CO2 emissions, agricultural NO emissions etc. But as we know, correlation is not a true metric of causality. For e.g.

1. It is shown in the XGB result that fertilizer consumption is the most prevalent factor for the annual temperature rise which we know is not correct and it just happened to coincide with the increasing yearly temperature.
2. In the LA result, Electricity production from coal sources is shown to have a highly negative correlation with the 'annual average temperature' which again is not correct. Rather, it may have high positive correlation with a lag of 'π' giving it a resultant negative correlation.

Therefore, it was important to perform the causality tests as well. After performing the 2 tests available to us, we found that CO2 emissions from liquid fuel topped the chart in both the tests. It has to be noted that domain specific knowledge is required for Pearl causality tests for which we assumed the respective values. Therefore, the results of the Pearl test are subjected to change. Whereas, the Granger causality tests do not require any such things and are thus more concrete. Therefore, we can safely assume that Fossil Fuel Energy Consumption is the main cause of annual temperature rise followed by CO2 emission from liquid fuel and CO2 emissions from manufacturing industries and construction with the lag of 2, 2 and 1 years respectively. That means that the higher consumption of Fossil fuel will likely affect the average temperature after 2 years.

9

Future Work

As we know, volcanic eruptions in one part of the world can cause weather imbalance in some other parts of the world. Therefore, it is necessary to analyse the trends all over the world to come to any accurate conclusion about temperature rise. Thus, it is necessary to take a holistic view of the problem rather than viewing it as a country or a region specific problem. It would be quite a task to collect data on each country and mapping it in context of the whole world and it would require many more days to accomplish it but the result would be astonishing nonetheless.

Use of Neural networks, better causality tests after bringing domain specific knowledge and time series analysis is much welcomed in this type of study. For that, the need to collect relevant data across diverse fields along with the cleaning and taking care of missing data could also yield positive results.

10

References

1. *NASA (n.d.) How Is Today's Warming Different from the Past? - https://earthobservatory.nasa.gov/Features/GlobalWarming/page3.php*

2. *Pachauri, R.K. and Meyer, L. (2015) Climate Change 2014: Synthesis Report. https://www.ipcc.ch/pdf/assessment-report/ar5/syr/SYR_AR5_FINAL_full_wcover.pdf*

3. *Funk, C. and Kenendy, B. (2016) Americans' Views on Climate Change and Climate Scientists.-http://www.pewinternet.org/2016/10/04/public-views-on-climate-change-and-climate-scientists/*

4. *Environmental Protection Agency (n.d.) Overview of Greenhouse Gases.-https://www.epa.gov/ghgemissions/overview-greenhouse-gases*

5. *NASA (n.d.) The Consequences of Climate Change.-https://climate.nasa.gov/effects/*

6. *Crowley, T.J. (2000) Causes of Climate Change over the Past 1000 Years.Science,289, 270-277. https://doi.org/10.1126/science.289.5477.270*

7. *Chen, X.Y. and Chau, K.W. (2016) A Hybrid Double Feedforward Neural Network for Suspended Sediment Load Estimation.Water Resources Management,30,2179-2194. https://doi.org/10.1007/s11269-016-1281-2*

8. *Olyaie, E.et al. (2015) A Comparison of Various Artificial Intelligence Approaches Performance for Estimating Suspended Sediment Load of River Systems: A Case Study in United States.Environmental Monitoring and Assessment, 187, 189.https://doi.org/10.1007/s10661-015-4381-1*

9. *Wang, W.-C.et al. (2014) Assessment of River Water Quality Based on Theory of Variable Fuzzy Sets and Fuzzy Binary Comparison Method. Water Resources Man-agement, 28, 4183-4200. https://doi.org/10.1007/s11269-014-0738-4*

10. *https://www.metoffice.gov.uk/weather/climate-change/causes-of-climate-change*

11. *https://ec.europa.eu/clima/climate-change/causes-climate-change_en*

12. *https://towardsdatascience.com/implementing-causal-inference-a-key-step-towards-agi-de2cde8ea599*

13. *https://microsoft.github.io/dowhy/*

14. *https://data.worldbank.org/*