PROJECT REPORT

On

Analysis of Global Greenhouse Gas Emission

By

Shweta Dutta

Registration Number: 18MSMS19

School of Mathematics and Statistics

UNIVERSITY OF HYDERABAD

हैदराबाद विश्वविद्यालय

Contents

About the Project

Project Name	**Analysis of Global Greenhouse Gas Emission**
Paper	Multivariate Analysis
Project Guide	Prof. M. Bhattacharjee
Department	School of Mathematics and Statistics
Project By	Shweta Dutta
Registration Number	18MSMS19
Semester	3
Course	M.Sc. Statistics – Operations Research
Close Date	05-11-2019

1. Background

Climate Change

The climate we live in is changing. The changes that are observed over the 20th century include increases in global average air and ocean temperature, rising global sea levels, long-term sustained widespread reduction of snow and ice cover, and changes in atmospheric and ocean circulation and regional weather patterns, which influence seasonal rainfall conditions.

These changes are caused by extra heat in the climate system due to the **addition of greenhouse gases** to the atmosphere. The additional greenhouse gases are primarily input by human activities such as the burning of fossil fuels (coal, oil, and natural gas), agriculture, and land clearing. These activities increase the amount of heat-trapping greenhouse gases in the atmosphere. The pattern of observed changes in the climate system is consistent with an increased greenhouse effect. Other climatic influences like volcanoes, the sun and natural variability cannot alone explain the timing and extent of the observed changes.

The science behind climate change is supported by extensive scientific research performed and reported across the world. Past and present climate information is collected from observations and measurements of our environment, including trapped air in ice from thousands of years ago. Climate models are used to understand the causes of climate change and to project changes into the future.

Greenhouse effect

The **greenhouse effect** is a natural process that warms the Earth's surface. When the Sun's energy reaches the Earth's atmosphere, some of it is reflected back to space and the rest is absorbed and re-radiated by greenhouse gases. The absorbed energy warms the atmosphere and the surface of the Earth. This process maintains the Earth's temperature at around 33 degrees Celsius warmer than it would otherwise be, allowing life on Earth to exist.

Greenhouse gases include water vapour, carbon dioxide, methane, nitrous oxide, ozone and some artificial chemicals such as chlorofluorocarbons (CFCs).

Source of greenhouse gases: Energy consumption is the major contributor of GHGs (61%). Within energy consumption, 40% is electricity and heat generation, another 20% is transportation and the remainder is building heat and industry. But energy consumption is by far not the only GHG source. Land-use change is the second largest contributor globally. Land-use change includes deforestation, reforestation (replanting in existing forested areas) and afforestation (creating new forested areas). Together, the activities under land-use change can be either a source or a sink of greenhouse gases; they can either contribute GHGs to or remove them from the atmosphere. Agriculture is another significant GHG source.

The major greenhouse gas is of course carbon dioxide (CO2) and nearly all CO2 comes from fossil fuels and land-use change. But methane (CH4) and nitrous oxide (N2O), which mostly come from agriculture and waste, are also significant GHGs and shouldn't be discounted.

Problems due to Greenhouse gas emissions: The problem we now face is that human activities – particularly burning fossil fuels (coal, oil and natural gas), agriculture and land clearing – are increasing the concentrations of greenhouse gases. This is the enhanced greenhouse effect, which is contributing to warming of the Earth.

Process of Greenhouse effect:

Step 1: Solar radiation reaches the Earth's atmosphere - some of this is reflected back into space.
Step 2: The rest of the sun's energy is absorbed by the land and the oceans, heating the Earth.
Step 3: Heat radiates from Earth towards space.
Step 4: Some of this heat is trapped by greenhouse gases in the atmosphere, keeping the Earth warm enough to sustain life.

Step 5: Human activities such as burning fossil fuels, agriculture and land clearing are increasing the amount of greenhouse gases released into the atmosphere.
Step 6: This is trapping extra heat, and causing the Earth's temperature to rise.

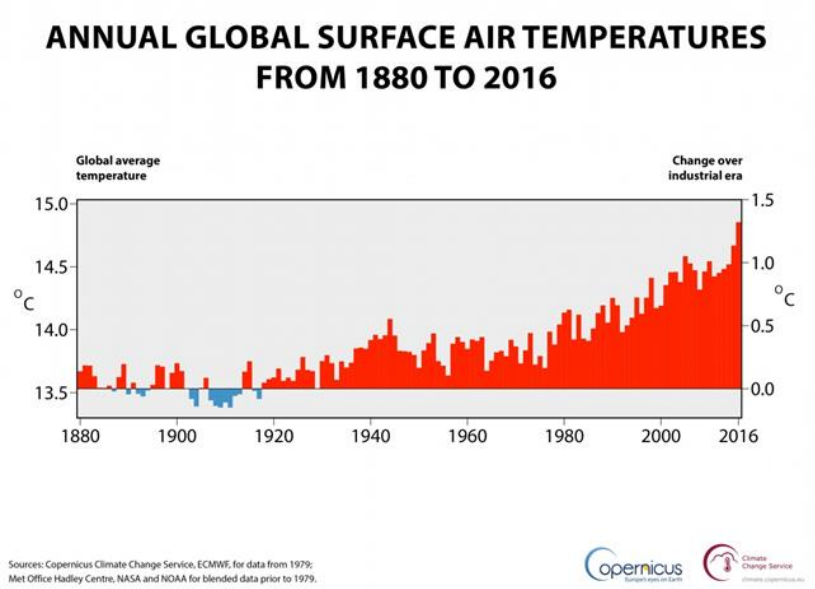
Impact of Climate Change

Many of the impacts of climate change pose risks to human and natural systems, in the form of more frequent and severe heat waves, coastal inundation due to sea level rise, disruptions to rainfall patterns and other effects. Analyses of a range of climate scenarios indicate the most severe risks of climate change can largely be mitigated if carbon dioxide emissions are reduced to the point where carbon dioxide is no longer accumulating in the atmosphere.

Following are some of the major impacts of climate change:

(i) Increased Air Temperatures:

Air temperatures have increased globally by around 1.1 degrees Celsius since the late 1800s, from which time modern meteorological record keeping is judged widespread and reliable enough to include in global syntheses. Most of the warming from 1880 to present has occurred since the 1970s. The observed increase in temperatures has occurred across the globe, with rising temperatures recorded on all continents and in the ocean. 2016 was the world's warmest year since 1880. The years 2015 and 2017 were the equal second warmest years on record globally.



(ii) Ocean Warming and sea level rise:

One of the strongest indicators of climate change is the amount of heat stored in the world's oceans. The heat content of oceans has increased during recent decades and accounts for more than 90 per cent of the total heat trapped by added greenhouse gases and accumulated by the land, air and ocean since the 1970s. Ocean warming is continuing, especially in the top several hundred metres of the ocean.

(iii) Extreme Weather events:

Extreme weather and climate events have serious impacts on our economy, society and environment. Extreme weather events include heat waves, bushfires, tropical cyclones, cold snaps, extreme rainfall including flash flooding, and droughts.

(iv) Rainfall Patterns:

Rainfall patterns are changing around the world. Research shows the global water cycle is intensifying with a warming climate, which means wet areas are likely to get wetter and dry regions are likely to be drier in response to climate change.

Further evidence of climate change

There are multiple lines of evidence that show the climate system is changing. These include: record high surface air temperatures, increased average number of hot days per year, decreased average number of cold days per year, increasing intensity and frequency of extreme events (e.g. fires, floods), changing rainfall patterns, increasing sea surface temperatures, rising sea levels, increasing ocean heat content, increasing ocean acidification, changing Southern Ocean currents, melting ice caps and glaciers, decreasing Arctic sea ice.

2. Purpose of the Project

The rising levels of manmade greenhouse gases (GHGs) in the atmosphere and their resulting impact on climate is now one of the single biggest technological and environmental challenges facing the world. This makes intensified monitoring of these gases more critical than ever in order to better quantify the role of the numerous natural and manmade sources, sinks and buffers involved in the cycles of GHGs. It also enables us to objectively audit GHG fluxes at the factory, city, country and continental level. Such objective auditing is ultimately essential to facilitate effective enforcement and compliance with any regulations, laws, treaties and trading agreements based on GHG metrics such as carbon footprints.

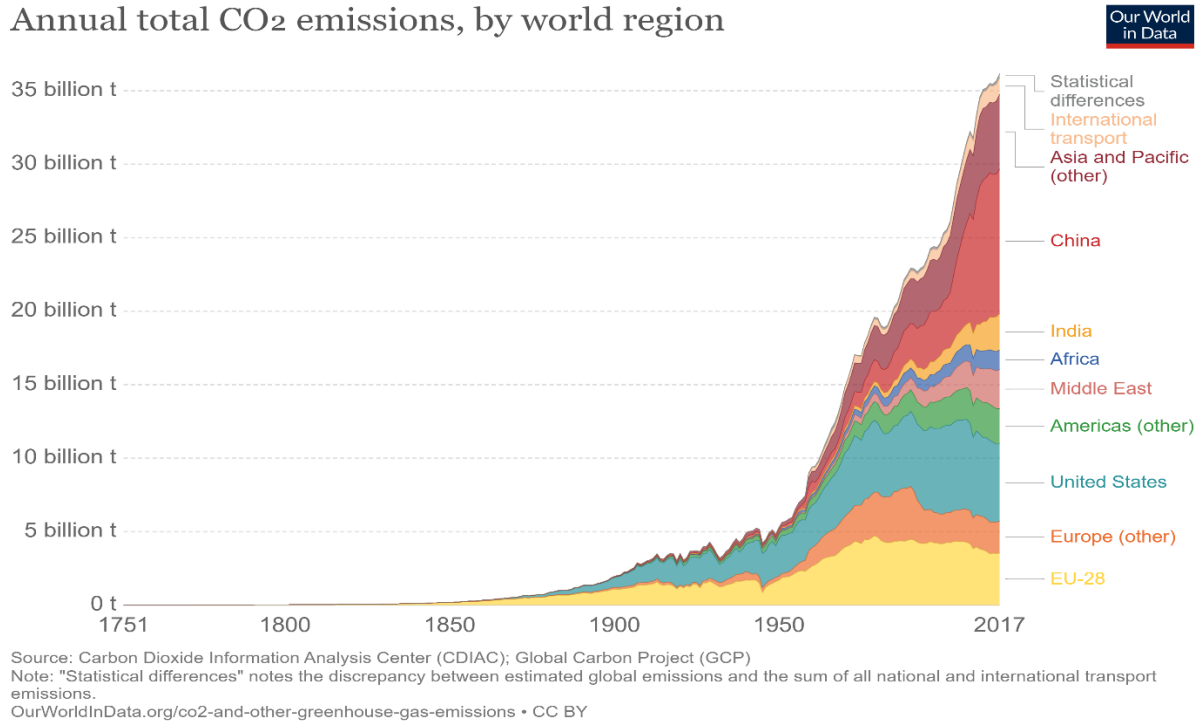
Sources of Greenhouse gases: The following sectors are the main sources of greenhouse gas emissions:

(i) **GHG emissions from energy sector –** The major sectors for which greenhouse gases are assessed under electricity consumption are consumption in domestic sector, commercial sector, industrial sector and other (public lighting, advertisement hoardings etc.)
(ii) **GHG emissions from transportation sector –** In major countries, transportation sector is one of the anthropogenic contributors of greenhouse gases. Emissions resulting from road transportation including CNG vehicles.
(iii) **GHG emissions from domestic sector –**The major sources include electricity consumption for lightning and other household appliances and consumption of fuel for cooking, etc.
(iv) **GHG emissions from industrial sector –** Iron and steel industry, cement industry, fertilizer plants are the few major industries which releases huge amount of greenhouse gases into the atmosphere.
(v) **GHG emissions from agricultural sector –** Methane emissions from paddy cultivation, nitrous oxide emissions from soil management are the major sectors responsible for greenhouse gas emissions from this sector.
(vi) **GHG emissions from waste sector –** Methane and nitrous oxide are the major two greenhouse gases that are emitted from solid waste and industrial waste water.

The possible questions now arise are:

1. **How have global emissions of CO2 changed over time?**
2. **How have global emissions of CH4, N2O and F gases changed over time?**
3. **Which sector yield maximum greenhouse gas?**
4. **Which country/ continent is responsible for maximum emission of greenhouse gases?**

The visualization below presents the long-run perspective on global CO_2 emissions. Global emissions increased from 2 billion tonnes of carbon dioxide in 1900 to over 36 billion tonnes 115 years later.



Thus, the main purpose of the project will be to:

- **Identify the main source of greenhouse gas**
- **Find how the emission values have changed over the years**
- **Predict the emission values of the greenhouse gases when the emission values from sources are know**
- **Forecast how much greenhouse gases will be emitted in future**

3. Analysis

3.1. Data Description

Food and Agriculture Organization of the United Nations provides data on emissions of greenhouse gases (GHG) by gas, economic sector, country and year, from 1990-2010.

Source and Description

- **Source**: http://www.fao.org/faostat/en/#data/EM
- **Description**: The Emissions by sector domain of the FAOSTAT Agri-Environmental Indicators section contains data on emissions of greenhouse gases (GHG) by gas, economic sector, country and year. It also displays the shares of each sector in the total emissions of each gas (e.g. share of agriculture in total CH4 emissions) and the shares of each gas in the emissions from each sector (e.g. share of CH4 in the emissions from Agriculture). The data provided are based on FAOSTAT (FAO, 2016) for Agriculture total, Land Use sources and Forest, and on the EDGAR Database (JRC/PBL, 2016) for the other sectors. The aim of this domain is to provide a global database of reference data to support countries in addressing statistical data gaps and exploring policy-relevant emissions indicators.

Original Structure

- The dataset originally had **40,586** total number of observations with **49** total columns.
- This is a **time series dataset** containing **numerical values** on **14 factors of elements** from **10 items/sources** for **231 countries**, for **21 years** from the year **1990-2010**.
- **Factor variables**:
    - **Area (Region)** with **268** levels –
        - 231 countries – viz., Afghanistan, Kenya, Spain, Thailand etc.
        - 37 aggregate levels – viz., World, Africa, Asia, Net Food Importing Developing Countries, and Low Income Food Deficit Countries etc.
    - **Item (Source)** with **14** levels –
        - Energy
        - Transport
        - Residential, Commercial, Institutional
        - Industrial Processes
        - Agriculture Total
        - Land use sources
        - Other sources
        - Waste
        - Energy Total
        - Land use total
        - Forest
        - International Bunkers
        - Sources total excl. AFOLU
        - Sources total
    - **Element (Emission value/share)** with **14** levels –
        - Total Emissions
        - Emissions from CO2
        - Emissions from CH4
        - Emissions from N2O
        - Emissions from F gases
        - Share of CO2 in sector emissions
        - Share of CH4 in sector emissions
        - Share of N2O in sector emissions
        - Share of F gases in sector emissions
        - Share of sector in total CO2 emission

        - Share of sector in total CH4 emission
        - Share of sector in total N2O emission
        - Share of sector in total F gases emission
        - Share of sector in total emission
- **Area code**, **Item Code** and **Element Code** columns contain the unique codes of the levels of Area, Item and Element, respectively.
- Elements are expressed in Gigagram and % units, denoted in **Unit** columns.
    - Emission values are expressed in Gigagram
    - Emission shares and sector shares are expressed in % units
- **Missing Values:** Dataset contains multiple missing values for each year.
    - viz., 4,904 rows have missing values for the year 1990, whereas 1,191 rows have missing values for the year 2010.

Reshaped Structure for project purpose

For project purposes the data was reshaped into a simpler structure, where the emission values from main independent sources are considered for **7** areas – 4 continents and 3 parts of America. Thus, the data is now in an aggregated form, since instead of each individual countries, the aggregate of emission values from the continents are considered.

- The dataset now has **147** total number of observations with **16** total columns.
- This is a **time series dataset** containing **numerical values** for **21 years** from the year **1990-2010** of **7 levels of Area** on –
    - **Total Emission values of GHG** from **10 items/sources.**
    - **Emission values of 4 main greenhouse gases.**
- **Factor variables**:
    - **Area** with **7** levels – Africa, Asia, Europe, Oceania, North America. Central America, South America.
    - **Item/Sources** with **10** levels –
        - Energy
        - Transport
        - Residential, Commercial, Institutional
        - Industrial Processes
        - Agriculture Total
        - Land use sources
        - Other sources
        - Waste
        - Energy Total
        - Land use total
    - **Element** with **4** levels –
        - Emissions from CO2
        - Emissions from CH4
        - Emissions from N2O
        - Emissions from F gases
- **Missing Values:** The data set is now free from any missing value.
  The figure below shows a glimpse of the structure of the reshaped data.

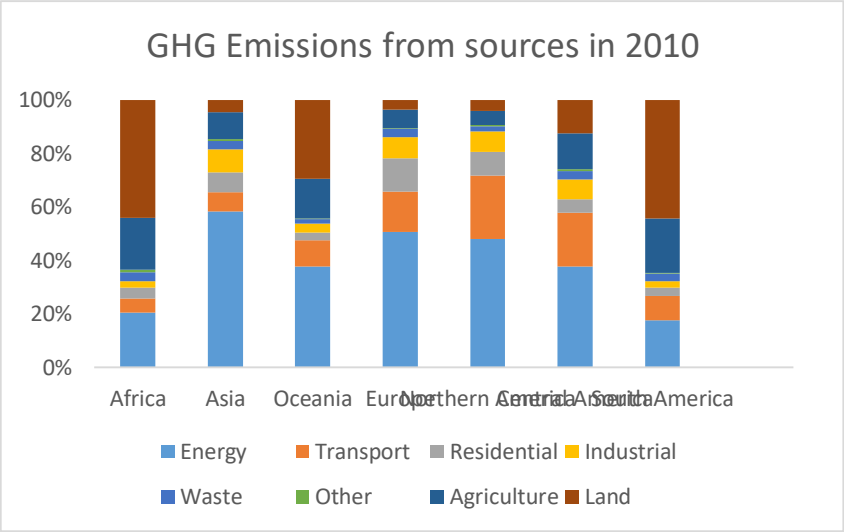
Area	Year	Energy	Transport	Residenti	Industrial	Waste	Other	Agricultur	Land	Energy tot	Land total	CO2_7194	CH4_7244	N2O_7243	F_7178
Africa	1990	602075.8	107165.3	102367.4	54363.75	77038.23	34644.05	569252.9	1877155	811608.5	1843262	2297726	711093.4	375628.6	5721.257
Africa	1991	610116.3	110440.2	106754.6	55673.58	79536.78	25348.8	571132.4	1877167	827311	1843262	2307888	721335.9	367675.8	5364.177
Africa	1992	634064.2	116649.2	110670.2	58085.37	82016.08	41186.77	578389.2	1877173	861383.5	1843262	2334274	735553.1	389325.9	5169.524
Africa	1993	629766.3	116310.3	118430.5	57605.77	84801.48	23963.41	578990.5	1875595	864507.1	1843262	2331314	742616.4	374293.5	4906.127
Africa	1994	620363.9	118201.4	121597.1	57879.11	87579.18	27290.73	586133.4	1875568	860162.4	1843262	2326096	749692.4	381947	4570.589
Africa	1995	658422	123964	124641.6	58831.54	90083.35	30081.23	595586.9	1875532	907027.5	1843262	2361843	769198.2	389704	4126.5

3.2. Data Exploration

Graphical Method

The following graphical methods are applied to the data for analysis, considering the values from recent observed **year 2010**.

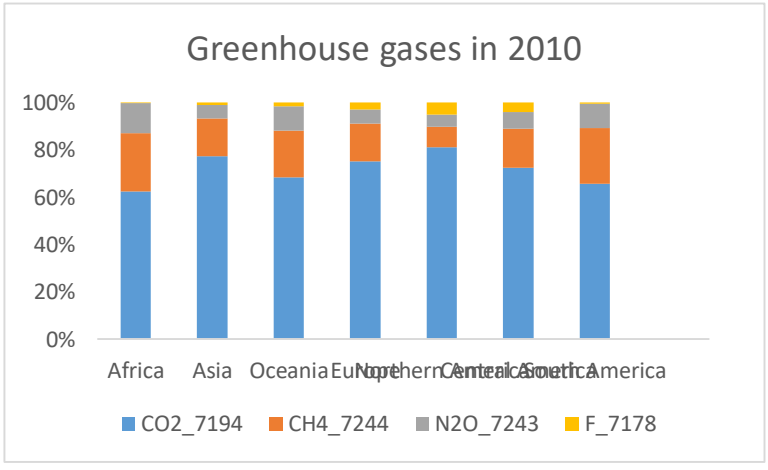
(i) The emission values from **8 sources from 7 levels** are plotted in a **sub-divided bar diagram.**



Conclusion: It is observed that most of the GHG emissions have the source

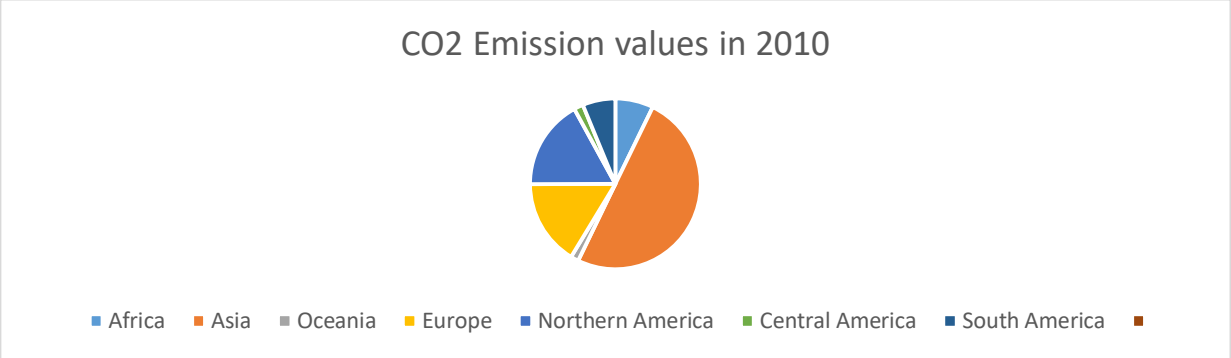
- "Energy", followed by "Land" for the levels Asia, Oceania, Europe, Northern and Central America.
- "Land", followed by "Energy" for the levels Africa and South America.

(ii) The data values of the **4 gases from 7 levels** are plotted in a **sub-divided bar diagram**.



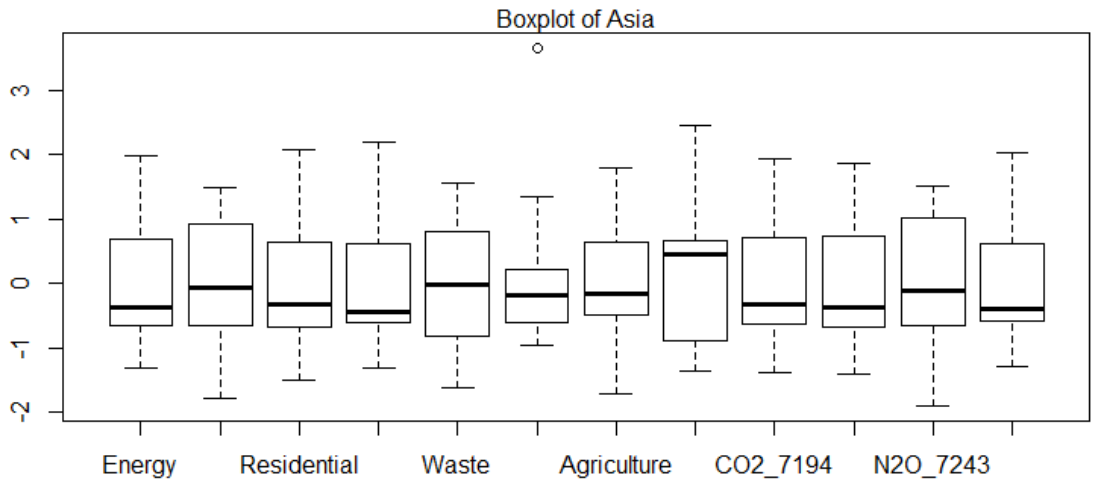
Conclusion: It was observed that **CO2** is the most emitted greenhouse gas for all the levels.

(iii) The data values of **CO2 gas from 7 levels** are plotted in a **pie diagram**.



Conclusion: It was observed that "**Asia**" is the main source of CO2 emission.

(iv) For **Asia**, the following box plot on emission values from sources and also that of share of GHG is obtained. Here, all the values have been **scaled** to have a proper diagram, within equal range of all variables.



Conclusion: 1 outlier is observed for emission values from 'Other' sources, and only emission values from Waste and Other sources seem to be distributed symmetrically. All other values are asymmetric.

Numerical Method

The summary statistics obtained are:

(i) Average emission values over 21 years of **each sources** under each level.

	Energy	**Transport**	**Residential**	**Industry**	**Waste**	**Other**	**Agriculture**	**Land**
Africa	731971	156195	140209	71920	104884	32305	662788	**1859776**
Asia	**7758113**	**1222278**	**1403996**	**1095005**	**578847**	**135060**	**1980573**	1675434
Europe	4563817	1203821	1147641	613158	280783	34642	680120	341171
Oceania	302526	87143	24244	23935	15500	5270	197339	150641
North America	3700251	1802025	724731	428098	189984	34880	402953	33526
Central America	253367	128883	40016	33097	25429	4577	106019	124145

South America	488883	264222	98696	76259	91267	12685	663433	1853615

(ii) Average emission values of **each GHG** over 21 years under each level.

	CO2	CH4	N2O	F gases
Africa	2428487	843398	422783	6346
Asia	**11119256**	**2763313**	**1083262**	**111063**
Europe	6301067	1276561	535055	122032
Oceania	437358	161708	106603	6142
North America	5850307	678912	376084	192410
Central America	520405	122368	58050	4056.1
South America	851801	649098	271016	9696

Conclusion: **Asia** being the greatest continent yields maximum all GHG under all sectors, except 'Land'. From the sector Land, Africa yields the maximum GHG.

(iii) Mean emission values from **each source** is

	Mean
Energy	2542704
Transport	**694938.2**
Residential	511361.9
Industry	334496.1
Waste	183813.5
Other	37059.96
Agriculture	670476.5
Land	905758.3

Conclusion: "**Transport**" is the source from which overall highest average amount of GHG are emitted, considering all the continents.

(iv) Mean emission values of **each gas** is

	Mean
CO2	4122492.41
CH4	927908.09
N2O	407550.36
F gases	64535.07

(v) The **correlation matrix of Emission** from sources is

	Energy	Transport	Residential	Industry	Waste	Others	Agriculture	Land
Energy	8.25E+12	1.47E+12	1.48E+12	1.15E+12	5.13E+11	1.04E+11	1.35E+12	1.63E+11
Transport	1.47E+12	4.34E+11	2.87E+11	1.94E+11	8.03E+10	1.43E+10	1.51E+11	-8.9E+10
Residential	1.48E+12	2.87E+11	2.96E+11	2.04E+11	9.39E+10	1.86E+10	2.39E+11	2.26E+10
Industry	1.15E+12	1.94E+11	2.04E+11	1.62E+11	7.24E+10	1.47E+10	1.94E+11	3.22E+10
Waste	5.13E+11	8.03E+10	9.39E+10	7.24E+10	3.5E+10	7.49E+09	1.01E+11	4.32E+10

Other	1.04E+11	1.43E+10	1.86E+10	1.47E+10	7.49E+09	1.93E+09	2.36E+10	1.51E+10
Agriculture	1.35E+12	1.51E+11	2.39E+11	1.94E+11	1.01E+11	2.36E+10	3.42E+11	2.71E+11
Land	1.63E+11	-8.9E+10	2.26E+10	3.22E+10	4.32E+10	1.51E+10	2.71E+11	6.44E+11

(vi) The **correlation matrix of Emission of GHG** is

	CO2	CH4	N20	F
CO2	1.43E+13	3.01E+12	1.14E+12	2.06E+11
CH4	3.01E+12	7.31E+11	2.74E+11	2.86E+10
N20	1.14E+12	2.74E+11	1.05E+11	1.18E+10
F	2.06E+11	2.86E+10	1.18E+10	6.49E+09

Conclusion: The correlation matrix gives the values that how the variables are correlated between themselves.

3.3. Basic Analysis

The purpose of the basic analysis includes

1. Dimensionality reduction of the data
2. Prediction

Keeping in mind the above mentioned purposes, the following basic analysis were applied to the data:

1. Principal Component Analysis
2. Linear Discriminant Analysis
3. Multivariate Analysis of Variance
4. Multiple Regression

3.3.1. Principal Component Analysis

Background:

- Principal Component Analysis is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of linearly uncorrelated variables. These new set of uncorrelated variables are called **Principal Components.**
- This method is concerned with explaining the variance - covariance structure of a set of variables through a few linear combinations of these variables.
- The main objectives are: Data reduction and Interpretation.
- In other words, the purpose of principal component analysis is to find the best low dimensional representation of variation in the multivariate data set.
- Suppose there are p components that produces the total system variability, often it is observed that much of this variability can be accounted for by a small number of principal components, say k<p. In such case, the k principal components can replace the initial p variables and the original data set of n observations on p variables reduces to a data set consisting of n observations on k principal components.
- **Rotation Matrix:** The rotation matrix is a matrix with the loadings of each principal components where the first column in the matrix contains the loadings for the first principal component, and so on.

Application on the dataset:

In the given data set, we have emission values from **8** different resources (- Energy, Transport, Residential, Industrial, Agriculture, Land, Waste, Others) that are obtained from **7** different levels (- Africa, Asia, Europe, Oceania, Northern America, Central America and South America) in the world.

By carrying out a principal component analysis after **scaling** the data, it was found that most of the variation in the emission values can be captured using the first **2** principal components, where each of the principal component is a particular linear combination of the **8** resource values.

Here, we are considering the total emission values from 8 resources only since the other two resources – Energy Total and Land Total is a more aggregate form of values from Energy and Land.

Findings:

The principal components can be predicted after applying PCA to the dataset. The rotation matrix obtained after applying principal component analysis is:

Variable	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**
Energy	0.397	-0.152	0.105	0.201	-0.329	0.312	-0.101	-0.742
Transport	0.296	-0.436	-0.758	-0.205	-0.181	-0.25	0.079	0.049
Residential	0.385	-0.185	-0.051	0.177	0.843	0.223	0.146	-0.013
Industrial	0.397	-0.116	0.181	0.274	-0.378	0.366	0.215	0.629
Waste	0.404	0.051	0.140	0.121	0.050	-0.382	-0.785	0.183
Others	0.371	0.221	0.184	-0.859	0.0168	0.196	0.015	0.0189
Agriculture	0.366	0.342	0.171	0.146	-0.043	-0.625	0.536	-0.128
Land	0.105	0.754	-0.543	0.185	-0.010	0.281	-0.1	0.012

Conclusion:

Table of variation:

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**
Standard Deviation	2.4474	1.2094	0.56614	0.34824	0.26846	0.15032	0.8662	0.06263
Proportion of variance	0.7487	0.1828	0.04006	0.01516	0.00901	0.00282	0.00094	0.00049
Cumulative Proportion	0.7487	0.9315	0.97158	0.98674	0.99575	0.99857	0.99951	1

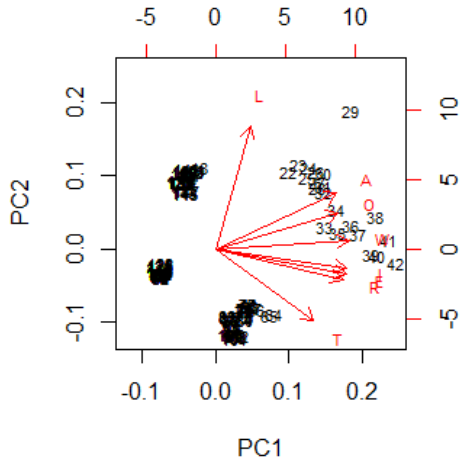
The total variance explained by the components is the sum of the variances of the components which is **8**.

From the table above, it is observed that the first principal component explains 74.87% of the total sample variance. The first two principal components collectively explain **93.15%** of the total sample variance. Consequently, sample variation is summarized by two principal components and therefore the data from 147 observations on 8 variables reduced to 147 observations on 2 principal components.

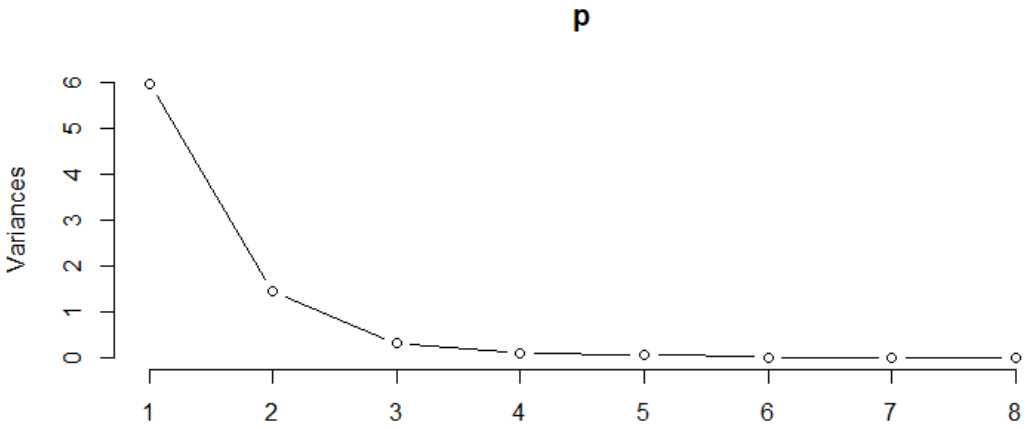
Visualization:

(i) A **biplot** displays the data points along the principal components and adds arrows to indicate the contributions of each of the variables to these principal components. A biplot

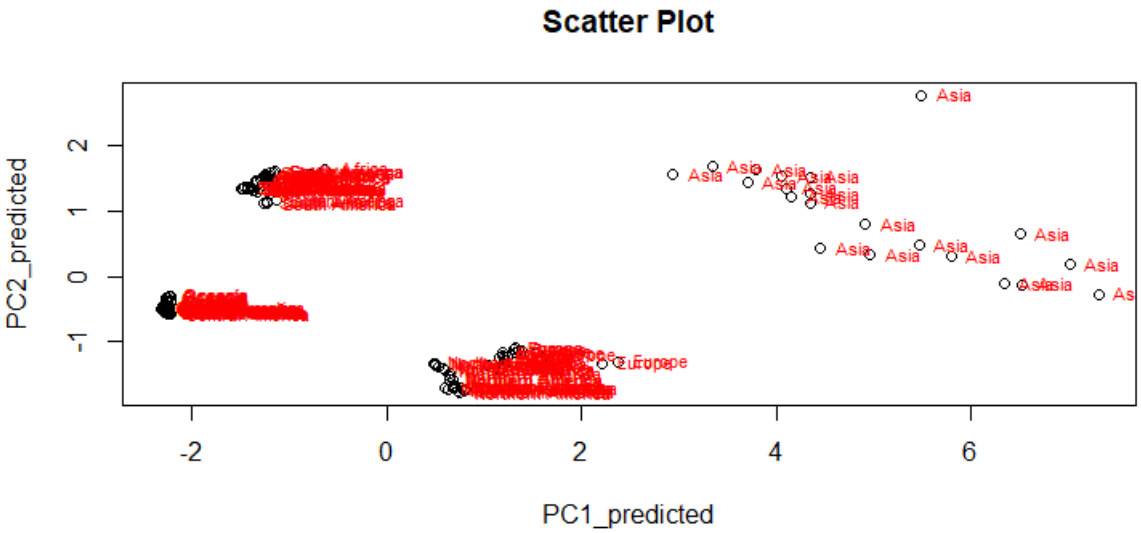
of 1st and 2nd Principal components is given below:



(ii) A **scree plot** plots the variances against the number of the principal components.



(iii) A **scatter plot** of the first 2 principal components, with labelling the data points with the levels (continents) that the emission values are coming from is given below:



The graph becomes messy since there are 7 levels of area.

3.3.2. Linear Discriminant Analysis:

Background:

- Linear discriminant function analysis is used to find axes (linear combinations of variables) that best separate predefined groups. The axes maximize variation between groups relative to variation within groups.
- In the previous method, principal components analysis pays no attention to groupings in the data and finds axes that maximize total variation. But, LDA is used for modeling differences in groups i.e. separating two or more classes.
- Linear discriminant analysis is also known as "**canonical discriminant analysis**", or simply "**discriminant analysis**".
- The purpose of linear discriminant analysis (LDA) is to find the linear combinations of the original variables that gives the best possible separation between the groups in the data set.
- **Loadings for the Discriminant Functions**: A matrix, in which the first column contains the loadings for the first discriminant function, the second column contains the loadings for the second discriminant function and so on, is the matrix of loadings of the discriminant functions.

Application on the dataset:

If we want to separate the emission values by levels (continents), which come from 7 different levels, so the number of groups (G) is 7, and the number of variables is 8 (8 sources of emission; p = 8).

The maximum number of useful discriminant functions that can separate the wines by cultivar is the minimum (G-1, p) = minimum (6, 8) = **6**. Thus, we can find at most 6 useful discriminant functions to separate the emissions by levels, using the 8 source variables.

Findings:

(i) Group Means:

The summary statistics obtained is the following table of group means of each variable, which shows the mean value for each of the independent variables for each group.

Levels	Energy	Transport	Residential	Industry	Waste	Others	Agriculture	Land
Africa	731970.6	156194.8	140209.2	71920.39	104884.1	32305.02	662787.5	1859776

Asia	7758113	1222278	1403996	1095005	578847.4	135060.3	1980573	1675434
Central America	253367.1	128883	40016.43	33097.08	25429.14	4577.293	106019.1	124144.5
Europe	4563817	1203821	1147641	613158.4	280782.9	34642.06	680230.4	341171
Northern America	3700251	1802025	724731.3	428098.1	189984.4	34880.17	402953	335526
Oceania	302526.1	87143.05	24243.55	23935.07	15499.59	5270.147	197339.1	150641
South America	488883.2	264222	98695.72	76259.01	91266.75	12684.75	663433.3	1853615

**The similar result was obtained under numerical analysis of data exploration.*

(ii) Matrix of loadings of discriminant functions:

	LD1	**LD2**	**LD3**	**LD4**	**LD5**	**LD6**
Energy	2.06E-06	3.18E-06	-3.32E-06	3.41E-07	1.97E-06	-3.25E-06
Transport	-1.35E-05	-1.64E-06	-9.60E-06	4.75E-06	-2.54E-06	9.30E-07
Residential	-2.51E-05	-3.18E-06	1.01E-05	-9.93E-06	-5.68E-07	-4.00E-06
Industry	1.34E-06	-1.47E-05	2.12E-05	-8.57E-06	-1.07E-05	1.73E-05
Waste	-3.94E-05	-3.26E-05	1.32E-05	-1.12E-05	1.91E-05	4.30E-05
Other	-7.86E-06	2.34E-05	3.57E-05	5.90E-05	4.50E-05	-1.61E-06
Agriculture	2.57E-05	-1.00E-05	-3.58E-06	1.22E-05	-1.05E-05	-7.98E-06
Land	5.66E-07	-5.46E-06	-4.39E-06	-5.10E-06	1.93E-06	7.32E-07

Conclusion:

We can predict the value of all 6 discriminant functions. The "**proportion of trace**" is the percentage separation achieved by each discriminant function, i.e., it describes the proportion of between-class variance that is explained by successive discriminant functions, which is given by:

	LD1	LD2	LD3	LD4	LD5	LD6
Proportion of trace	0.5766	0.3640	0.046	0.0114	0.0007	0.0004

It can be concluded that LD1 explains 57.66% of the total variance.

Visualization:

A **Stacked Histogram** is a way of displaying the results of a linear discriminant analysis (LDA) where a stacked histogram of the values of the discriminant function for the samples from different groups (different levels, here) is made. Thus, a stacked histogram of the all the discriminant function's values for emission values from the 7 different levels can be obtained.

**Due to high margin values, the histogram cannot be shown here.*

3.3.3. Multivariate Analysis of Variance

Background:

- **ANOVA** or Analysis of Variance is a group of statistical models to test for significant difference between means. It tests whether the means of various groups are equal or not.
- A Multivariate analysis of Variance is called MANOVA, which is similar to ANOVA except that there are more than one variable or factors involved. This is used in studies where more than one factors affect the dependent variable.
- Thus, MANOVA is an extension of the ANOVA that allows taking a **combination of dependent variables** into account instead of a single one.
- With MANOVA, explanatory variables are often called **factors**.

Assumptions:

- **Independent Random Sampling**: MANOVA assumes that the observations are independent of one another, there is not any pattern for the selection of the sample, and that the sample is completely random.
- **Level and Measurement of the Variables**: MANOVA assumes that the independent variables are categorical and the dependent variables are continuous or scale variables.
- **Absence of multicollinearity**: The dependent variables cannot be too correlated to each other.
- **Normality**: Multivariate normality is present in the data.
- **Homogeneity of Variance**: Variance between groups is equal

Application on the dataset:

In the dataset, "**Area**" is the factor variable, and so there are **7 different factors**.

(i) Suppose we want to test whether the means of the various groups of Area of the 8 variables "**Emission from Sources**" are equal or not, MANOVA is carried out and the following **coefficients** are obtained.

	Energy	Transport	Residential	Industry	Waste	Others	Agriculture	Land
(Intercept)	731970.6	156194.8	140209.2	71920.39	104884.1	32305.02	662787.5	1859776
Asia	7026143	1066083	1263786	1023084	473963.3	102755.3	1317785	-184342
Central America	-478604	-27311.8	-100193	-38823.3	-79455	-27727.7	-556768	-1735632
Europe	3831846	1047626	1007432	541238	175898.9	2337.043	17442.9	-1518605
Northern America	2968280	1645830	584522.1	356177.8	85100.32	2575.157	-259835	-1524250
Oceania	-429445	-69051.7	-115966	-47985.3	-89384.5	-27034.9	-465448	-1709135
South America	-243087	108027.3	-41513.5	4338.617	-13617.3	-19620.3	645.8432	-6161.48

Pillai's Approximation is 4.185 and p value<0.05 and <0.01. Thus, we **reject** the null hypothesis that the mean of the groups are same at 95% as well as 99% level of significance and conclude that the mean of the variables of the groups are **unequal**.

(ii) Suppose we want to test whether the means of the various groups of Area of the 4 variables "**Emission of GHG**" are equal or not, MANOVA is carried out and the following coefficients are obtained.

	CO2	CH4	N2O	F gases
(Intercept)	2428487	843398.2	422783.4	6346.391
Asia	8690769	1919914	660478.3	104716.5
Central America	-1908081	-721030	-364733	-2290.28
Europe	3872580	433162.3	112271.1	115685.7
Northern America	3421820	-164487	-46699.1	186063.9
Oceania	-1991129	-681690	-316180	-204.814
South America	-227920	-194301	-151767	3349.714

Pillai's Approximation is 2.38 and p value<0.05 and <0.01. Thus, we **reject** the null hypothesis that the mean of the groups are same at 95% as well as 99% level of significance and conclude that the mean of the variables of the groups are **unequal**.

3.3.4. Multiple Regression

Background:

- **Multiple regression** is an extension of simple **linear regression**. It is used when we want to predict the value of a variable based on the value of two or more other variables.
- The variable we want to predict is called the **dependent** variable (or sometimes, the outcome, target or criterion variable).
- Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data.

Application on the dataset:

In the given data set, the variables "Emission of CO2, CH4, N2O, F gases" are dependent in nature with that of the variables of "Emission from sources". Since there share of values are already included in the other variables of "Emission from sources".

Thus, we can regress the dependent variables on the independent variables, keeping an **area fixed**, and predict the future values of gases when the emissions from sources are already know to us.

Suppose, the area "**Asia**" is selected.

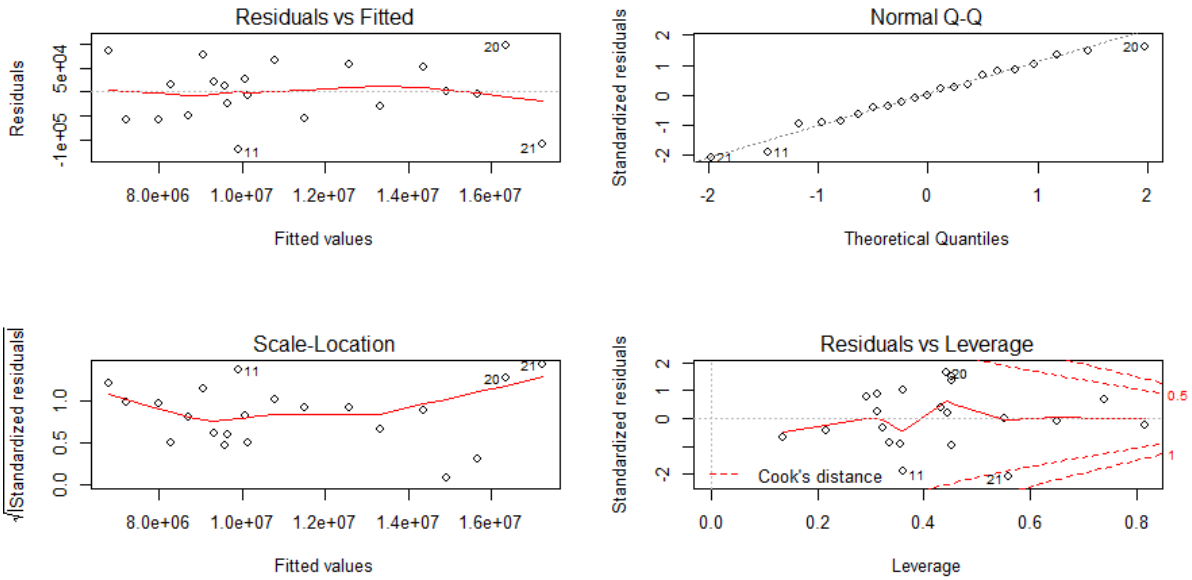
Sl. No.	Regression type	Adjusted R square
1	Regressing "Emission of CO2" on other variables.	99.94%
2	Regressing "Emission of N2O" on other variables.	99.75%
3	Regressing "Emission of CH4" on other variables.	99.65%
4	Regressing "Emission of F gases" on other variables.	99.19%

And, therefore for the continent Asia, the **coefficients** of the variables (Emission from Source) are obtained for each of the regression of the dependent variables (Emission values of GHG).

	1	2	3	4
(Intercept)	1008467	1242230	-449516	-22965.2
Energy	0.853012	0.130972	-0.00457	0.009425
Transport	0.899115	0.108148	0.062399	-0.00468
Residential	1.191189	0.29185	-0.08859	0.009574
Industry	1.136006	0.009758	-0.01141	0.039151
Waste	3.461359	1.166339	0.208982	0.020158
Other	4.081592	0.969888	1.151361	-0.01879
Agriculture	-1.65281	-0.48611	0.667088	-0.00013
Land	0.117201	0.065061	0.018571	0.000858

Visualization:

The following plots are obtained from "**Regressing "Emission of CO2" on other variables**."



- The first plot (corner left) gives the plot of "Residual v/s Fitted values"
- The second plot (corner right) gives the Normality Q-Q plot. Most of the values fall on the straight line and thus it can be commented that the values follow Normality assumption.
- The observations with the index 11, 20, 21 i.e., on the years 2001, 2009, 2010 are considered as **outliers**.

Note: Similarly, for other areas also, similar results can be obtained.

3.4. Advanced Analysis

Under advanced analysis, we want to predict the future values of emission.

Time Series Analysis

Background:

- Time series analysis comprises methods for analysing time series data in order to extract meaningful statistics and other characteristics of the data.
- **Forecasting** is a method for predicting probable amount in the future, based on the analysis of past amount for that product under the present condition. When the past data is used to predict the future amounts, then this method is referred as **time series method**.
- **Time series forecasting** is the use of a model to predict future values based on previously observed values.
- **Examples** of method of Time Series Forecasting include Moving Average, Exponential Smoothing, and ARMA/ARIMA etc.
- **Autocorrelation** measures correlation between observations as a function of the time lag between them.

Application of the dataset:

The dataset contains the emission values from different sources and of different gases from 7 levels, over a span of 21 years from 1990-2010.

(i) The following table contains the **auto correlation values** of the **8 source's emission** values from the levels, at **lag 1**.

	Energy	Transport	Residential	Industry	Waste	Others	Agriculture	Land
Africa	0.967174	0.981478	0.982412	0.990474	0.998955	0.605546	0.978536	0.357826
Asia	0.995904	0.982886	0.928002	0.992864	0.999569	-0.116	0.981666	0.677687
Central America	0.971903	0.98741	0.577941	0.978927	0.993661	-0.14193	0.853245	0.824975
Europe	0.923848	0.923762	0.851262	0.83303	0.983568	0.547286	0.98028	0.051689
Northern America	0.858009	0.96374	0.382999	0.982151	0.983591	0.931453	0.8421	0.21518
Oceania	0.990657	0.983885	0.964113	0.976753	0.618878	0.216733	0.577111	0.806849
South America	0.972018	0.979646	0.943554	0.92416	0.99317	0.244655	0.984081	0.712497

(ii) The following table contains the **auto correlation values** of the **4 gas's emission** values from the levels, at **lag 1**.

	CO2	CH4	N2O	F Gases
Africa	0.98857	0.612024	0.682564	0.966863
Asia	0.99569	-0.14508	0.992559	0.978971
Central America	0.98586	0.847365	0.957301	0.91031
Europe	0.905117	0.834655	0.916922	0.952547
Northern America	0.914595	0.692872	0.919237	0.914853
Oceania	0.992613	0.26299	0.920495	0.086818
South America	0.977339	0.729262	0.737465	0.960946

It is observed that over all the variables have high auto correlation values and in some cases the correlation values are quite significant, which implies that the data set can be considered as a **time series data**, i.e., the values of emission on a year significantly depends on that of the past years.

Forecasting:

The continent "**Asia**" is considered to find out the future forecast values of emission.

Firstly, an **ARIMA** model is fitted to the data, and based on that, forecast is generated for all the variables for 10 years. Since the dataset had the data till 2010, now we obtain the data for the next 10 years.

	E	T	R	I	W	O	A	L	CO2	CH4	N20	F
2011	13637868	1635349	1691301	2063799	743215.9	135060.3	2285727	1182920	17626860	3632986	1284551	204649.5
2012	14272404	1676057	1714074	2220663	757985	135060.3	2312877	1114675	18138830	3702826	1303691	214730.4
2013	14883419	1716764	1736847	2375029	773374.9	135060.3	2340028	1147105	18650800	3772665	1322831	224811.3
2014	15504723	1757472	1759620	2530857	788513.4	135060.3	2367178	1131694	19162770	3842504	1341971	234892.1
2015	16121526	1798179	1782393	2685830	803753.7	135060.3	2394328	1139018	19674740	3912344	1361111	244973
2016	16740298	1838887	1805166	2841303	818952.7	135060.3	2421479	1135538	20186710	3982183	1380251	255053.8
2017	17358209	1879594	1827940	2996484	834168.4	135060.3	2448629	1137191	20698680	4052023	1399391	265134.7
2018	17976497	1920302	1850713	3151835	849377.4	135060.3	2475779	1136406	21210650	4121862	1418531	275215.6
2019	18594619	1961009	1873486	3307087	864589.1	135060.3	2502930	1136779	21722620	4191701	1437671	285296.4
2020	19212814	2001716	1896259	3462397	879799.8	135060.3	2530080	1136602	22234591	4261541	1456810	295377.3

Note: Similar forecasts can be generated for the other levels or continents, also.

4. Conclusion

Identifying the main source:

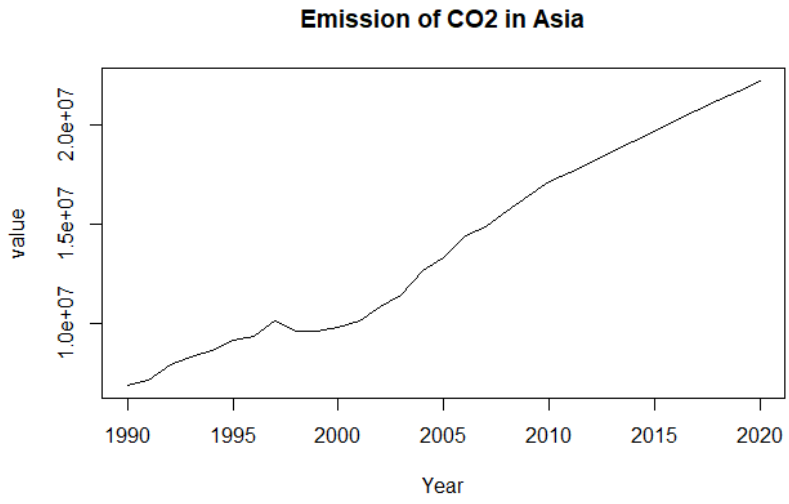
- Since Asia is the largest continent, thus it was observed that the maximum of GHG emission occurs in Asia.
- The overall main source of emission of GHG can be considered as "**Transport**", since the highest average emission is obtained from this sector, for all continents.
- But for each level of continent, it can be seen that the main sectors from which maximum emission of GHG occurs are "**Energy**" and "**Land**".
- The main GHG emitted is **CO2** and Asia is the continent which is responsible for most of its emission.

Modelling and Prediction:

- Fitting the model of "each GHG emission" by Multiple Regression Method will now give the predicted values of each of the main GHG, when the total value of all the gases under each source is known.

Forecasting:

- From the forecast values obtained, a graph can be plotted between years and the emission values to notice where the future is heading towards.



From the graph above, it can be seen that there has been extreme increase in emission of CO2 gas in Asia, and the forecast also shows that the amount will keep on increasing in future. Thus, it is high time when each one of us should be aware of the current and upcoming environment situation and start taking steps to save the planet earth and make the environment a better place to live.

Further Studies

Further analysis on the reshaped data:

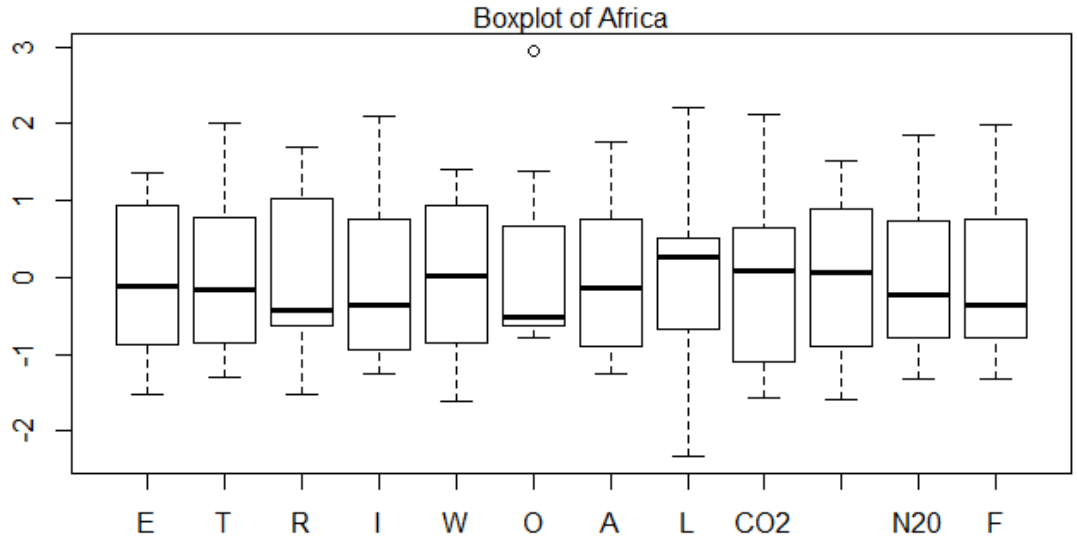
- While applying MANOVA, the validity of the assumptions were not checked which should be done.
- Multiple Regression is applied only to the observations under the continent "Asia" and thus the model of "GHG emissions" is obtained only for Asia. It can be checked for other continents as well.
- The method used here to generate forecast values is "Auto Regressive Integrated Moving Average method (ARIMA)". The other methods "Holt Winter's Exponential Smoothing" or "Moving Average" are not considered. Thus, it may be so that some other method gives better forecast values.
- The forecast values generated by using the "auto.arima" function in R. This function automatically chooses the best 'p', d', 'q' values. Also, during forecast the whole of the dataset is considered and the usual procedure of dividing the data into "training" and "testing" part and then applying the method is not considered. Thus, this technique can be made more detailed, so that better forecasts are generated.
- Further graphs can also be plotted to notice the future values of different gases, or values of emissions from different sources and be aware of the situation.

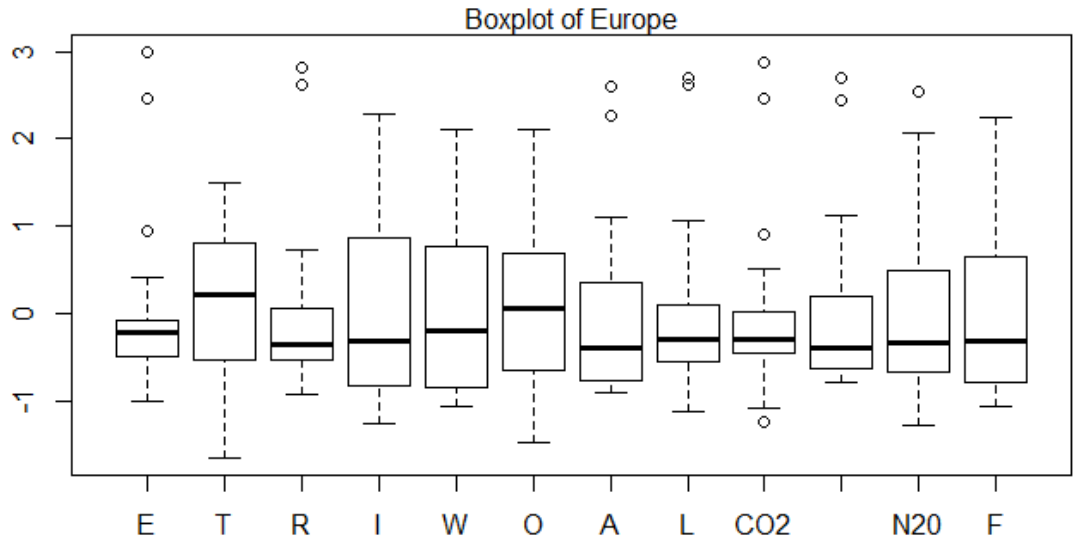
Considering the detailed data:

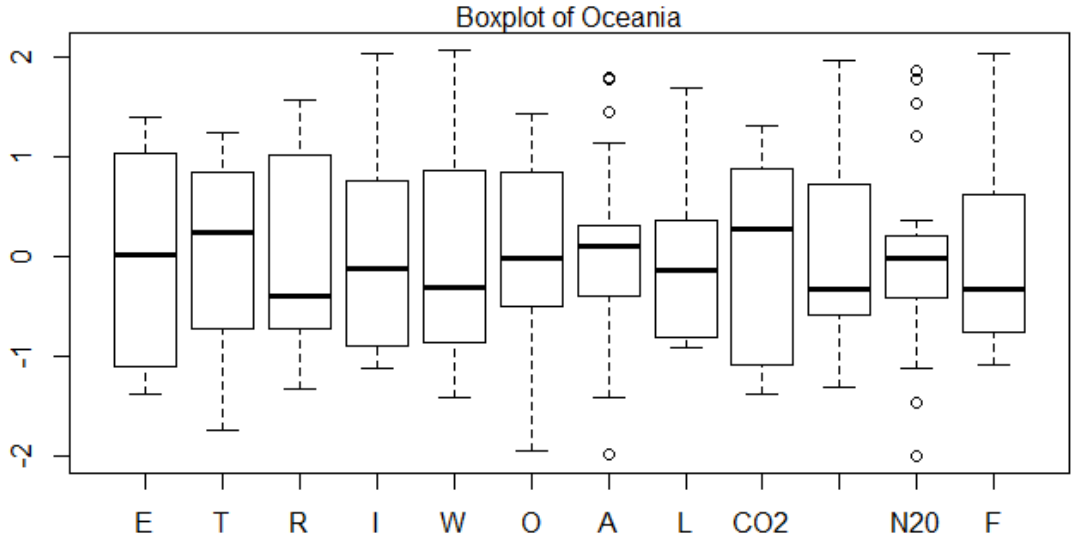
- Since, the original dataset provides value for 231 countries over the world, thus further and similar analysis can be carried out for each countries, which will give a better and detailed idea about the GHG emission scenario, and thus the country that contributes maximum GHG to the environment can take precautionary steps, beforehand.

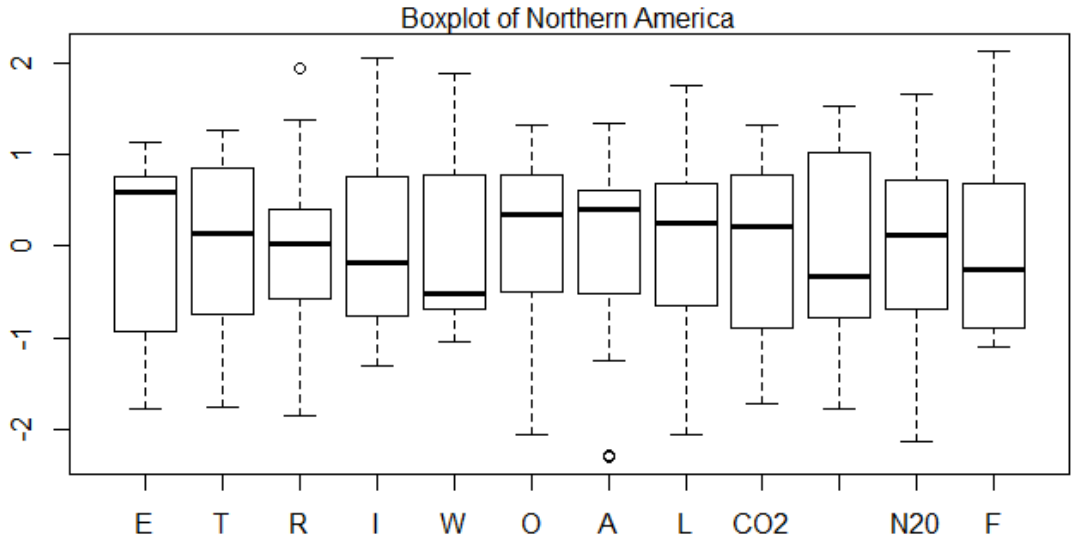
Appendices

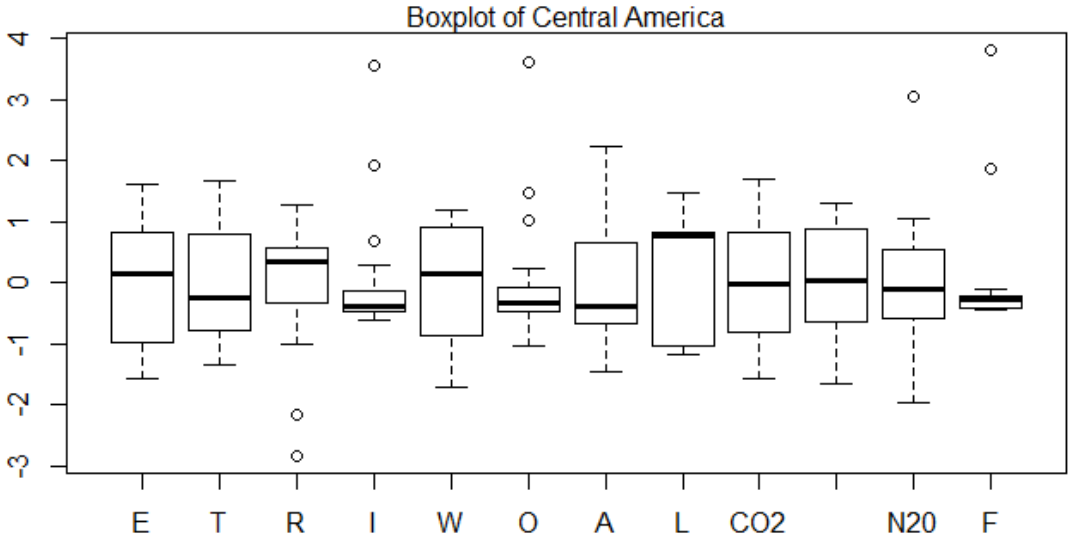
Graphs - Boxplots for other levels:

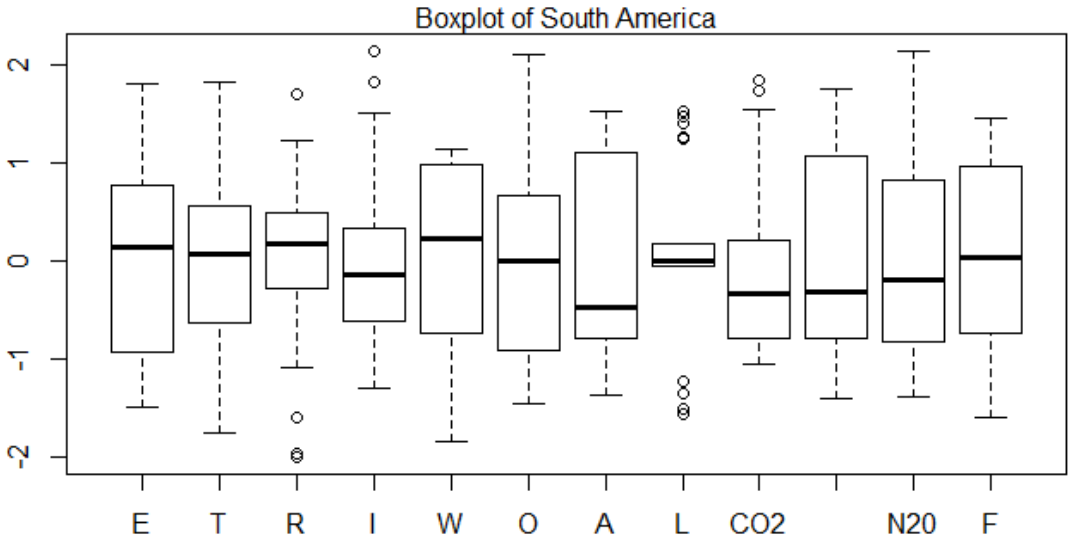





Boxplot of Oceania
2
1
0
-1
-2
E
T
R
I
W
O
A
L
CO2
N20
F


Boxplot of Northern America
2
1
0
-1
-2
E
T
R
I
W
O
A
L
CO2
N20
F





Packages used in R for analysis

Purpose	Packages and Functions
Reshaping the data	dplyr, tidyr, reshape2, DataCombine, raster
Principal Component Analysis	stats: prcomp
Linear Discriminant Analysis	MASS: lda
Manova and Multiple Regression	stats: manova, lm
Forecasting	Forecast: auto.arima, forecast

References

Johnson, Wichern: Pearson Prentice Hall, 'Applied Multivariate Statistical Analysis'

Mardia, Kent, Bibby: Academic Press 1995, 'Multivariate Analysis'

Research Paper by Sridevi H., Shreejith K., T.V. Ramachandra 2014, 'Comparative Analysis of greenhouse gas emissions from major cities'

Aaron French, Marcelo Macedo, John Poulsen, Tyler Waterson and Angela Yu, 'MANOVA'

NCSS, 'Multiple Regression'

Links:

https://www.environment.gov.au/climate-change/climate-science-data/greenhouse-gas-measurement

https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c

https://www.analyticsvidhya.com/blog/2018/09/multivariate-time-series-guide-forecasting-modeling-python-codes/

https://www.geeksforgeeks.org/ml-linear-discriminant-analysis/